Statistics and Its Interface Volume 9 (2016) 159–170

Convergence and stability analysis of mean-shift algorithm on large data sets*

Xiaogang Wang†, Weiliang Qiu, and Jianhong Wu

We present theoretical convergent analysis of mean-shift type of clustering methods for large data sets. It is proved that correct convergence for unsupervised mean shift type of algorithms relies on its ability to successfully transform data points to be clustered into data patterns of a multivariate normal distribution. Our analytical stability analysis suggests that a judiciously chosen supervision mechanism might be essential for correct convergence in dynamical clustering. The proposed theoretical framework could be used to study other dynamical clustering methods.

Keywords and phrases: Anti-diffusion, Convergence, Conservation law, Dynamic clustering, Entropy, Partial differential equations.

1. INTRODUCTION

Fukunaga and Hostetler (1975) [1] propose the famous *mean-shift* algorithm. Given a kernel function K and a weight function w, the generalized mean-shift operation is given by

$$T(x) = \frac{\sum K(x,s)\, w(s)\, s}{\sum K(x,s)\, w(s)}. \tag{1}$$

This algorithm seems to originate from the intuitive idea of moving data points to denser regions by following estimated local gradient functions. There are many variations and applications of this algorithm in literature, including for example, Comaniciu and Meer (2002) [2], Virmajoki (2002) [3], Shi et al. (2005) [4], Woolfold and Braun (2007) [5], and Wang et al. (2007a) [6]. Choi and Hall (1999) [7] provide a mathematically rigorous demonstration of bias reduction in density estimation through one or more iterations of the mean-shift algorithm.

Although the algorithm is often found to be convergent in applications, there is little theoretical assurance, besides empirical or visual validations in low dimensions, that the result is indeed correct. Choi and Hall (1999) [7] suggest that too many iterations of the algorithm will lead to what they term as "over-sharpening". The choice and parameter of kernel functions are known to have significant impacts on the outcome and validity of the partition as suggested in Cheng (1995) [8]. Cheng (1995) [8] also points out that it is often very difficult to see where the mean-shift method leads to due to simultaneous movements of all the data points. It is well known in clustering and optimization that convergence does not mean *correct* convergence.

In this article, we provide theoretical convergence and stability analysis for mean-shift type of clustering algorithms for large data sets. When the sample size is large, it is impossible to have a complete description of the behavior of a large data set as in statistical mechanics. The successive movements of data points under the governance of the mean-shift clustering algorithm can be viewed collectively as an evolution process of a dynamical system. We follow the argument of Einstein (1956) [9] and employ a theoretical framework using partial differential equations that prescribes the collective or emergent behavior of large data sets. Consequently, the focus is on the macroscopic dynamics, *i.e.* large sample behavior of data points. Furthermore, it might provide an analytical framework to study any intrinsic instability or deterministic/Hamiltonian chaotic behavior defined in physics, see for example Pettini (2007) [10].

We prove that correct convergence for the mean-shift algorithm can only happen when the patterns to be clustered can be successfully transformed into those from a multivariate normal distribution with no dependence structure. Our stability analysis suggests that, in general, a supervised clustering using the mean shift method is preferred in order to ensure correct convergence. The proposed analytical framework for dynamical clustering is potentially useful to guide further research on choosing a supervision function.

The rest of this paper is organized as follows. We present an analytical framework for dynamical clustering in Section 2. Section 3 presents theoretical analysis of the unsupervised dynamical clustering. The stability and convergence of a supervised dynamical clustering is presented in Section 4. Section 5 contains demonstrative results from simulation studies and discussion on empirical supervision. The discussion is provided in Section 6.

2. FRAMEWORK FROM STATISTICAL MECHANICS

In traditional clustering analysis, a data set is a sample drawn from an underling probability density function. For

*The first and third author's research was supported in part by NSERC Discovery grants.
†Corresponding author.

each particular sample to be clustered, observations are all fixed and their values can not be altered in any way. In dynamical clustering methods such as the mean-shift algorithm, however, data points are no longer static. Their spatial locations undergo constant changes until a convergence is declared. Consequently, each data point can be viewed as a particle under a gravitational field or an autonomous agent governed by certain laws of attraction. To describe the location, speed and direction of all the movements of such a complex system, it would require many variables or parameters. Furthermore, such a full description of a complex system might not be necessary or even possible for the purpose of clustering since we are only concerned about emerging patterns at the macroscopic level.

When studying collective behavior of small particles suspended in a stationary liquid, Einstein (1956) [9] presents a well known comprehensive framework which utilizes a differential equation framework for diffusion based on an underlying probability distribution. Since then, it is now the standard analytical framework for statistical mechanics. Ellis (1985) [11] provides a measure-theoretical justification of such a probabilistic framework to study the macroscopic properties of a system with a large number of data points. From a statistical point of view, the underlying probability distribution can be estimated by an empirical or kernel density estimation with an arbitrary accuracy for a large sample. Detailed description and discussions of theoretical properties of kernel density estimation can be found in Simonoff (1996) [12]. We employ the analytical framework by Einstein (1956) [9] to study the mean shift clustering algorithm applied to a large number of data points. Since the data points are constantly moving, the associated probability density functions collectively define a spatial and temporal process:

$$\begin{aligned}\mathcal{E} =&\{f_t \,|f(\boldsymbol{x};t) \geq 0,\\ &\int f(\boldsymbol{x};t)d\boldsymbol{x} = 1,\ \ t \in N, \boldsymbol{x} \in R^n\}.\end{aligned} \tag{2}$$

At the macroscopic level, instead of modelling the individual movement of a particular data point, we examine patterns generated by these transformations of the underlying probability distribution. A dynamical clustering algorithm aims to identify the true underlying probability distribution through a dynamical self-organizing process.

In dynamical clustering, if an algorithm does not delete data points from the clustering process, the rate of change of the total number of particles or data points contained in a fixed volume is equal to the influx of particles or data points passing through the boundary. By using the conservation law, we now present a general differential form for high dimensional cases. Denote an influx vector by $\boldsymbol{q}(\boldsymbol{x};t)$ and the probability density function by $f(\boldsymbol{x};t)$. We then have

$$\frac{d}{dt}\int_V f(\boldsymbol{x};t)\, dV = -\int_S (\boldsymbol{q}.\boldsymbol{n})\, dS, \tag{3}$$

where dV is the volume element, dS is the surface element of the boundary surface S, and $\boldsymbol{n}$ denotes the outward unit normal vector to S with right-hand side measures to the *outward* influx indicated by the minus sign.

On applying the Gauss divergence theorem and taking d/dt inside of the integral on the left-hand side, we then have

$$\int_V \left(\frac{\partial f(\boldsymbol{x},t)}{\partial t} + \nabla\, \boldsymbol{q}(\boldsymbol{x},t)\right) dV = 0, \tag{4}$$

where ∇ is the divergence operator given by

$$\nabla \boldsymbol{q}(\boldsymbol{x},t) = \sum_{i=1}^{n} \frac{\partial^2 \boldsymbol{q}_i(\boldsymbol{x},t)}{\partial x_i}, \tag{5}$$

and $\boldsymbol{q}_i$'s are components of the vector $\boldsymbol{q}(\boldsymbol{x},t)$.

Since the result is valid for any arbitrary volume V, the integrand must be zero if it is continuous. The differential form of the general conservation law is then given by

$$\frac{\partial f(\boldsymbol{x},t)}{\partial t} + \nabla\, \boldsymbol{q}(\boldsymbol{x},t) = 0, \tag{6}$$

where $\boldsymbol{q}(x;t) = \boldsymbol{u}(x;t)\times \boldsymbol{f}(x;t)$. We refer the detailed derivations and discussions of conservation laws and associated differential forms to Debneth (2004) [13].

For supervised dynamical clustering, the trajectories of data points are influenced by supervision. There are many ways of injecting supervision in the process. One way is to impose a source or sink function in the dynamical process so data points will be *absorbed* into a given domain. Denote such a supervision function by $\psi(\boldsymbol{x},t)$. Consider an arbitrary small volume V_ϵ. By following the same argument as above, we then have

$$\frac{\partial f(\boldsymbol{x},t)}{\partial t} + \nabla q(\boldsymbol{x},t) = \psi(\boldsymbol{x},t). \tag{7}$$

When $f(\boldsymbol{x},t) = 0$, then it follows that $q(\boldsymbol{x},t) = 0$. Consequently, we have $\psi(\boldsymbol{x},t) = 0$ which implies that there is no need for supervision. When $f(\boldsymbol{x},t) > 0$ for this entire small volume, then we can further control the movement of particles by introducing a supervising function. This essentially changes the geometric properties of the probability density function which is the solution of the differential form. If we consider the underlying probability density function as a manifold, then the supervising function would change the intrinsic properties of the manifold such as the metric tensors and other determining features of the manifold. The exact impact of the supervising function, however, seems to be analytically intractable due to the complexity of the inhomogeneous PDE.

As an example, we consider the one dimensional case. Denote the one dimensional influx of data points by $q(x;t)$ and the probability density by $f(x;t)$ at the spatial location x and time t. We then have

$$q(x;t) = u(x;t) \times f(x;t), \tag{8}$$

where $u(x;t)$ is the speed of particles at location x and time t. In an unsupervised dynamical clustering, a constant flow of data points passes through an arbitrarily small interval using the laws of attractions is characterized by the speed $u(x,t)$ that will be specified in later sections. Data points are assumed to be incompressible, and hence a standard argument in fluid dynamics in one-dimension space yields the *conservation law* given by

$$\frac{d\, f(x;t)}{dt} + \frac{d\, q(x;t)}{dx} = 0. \tag{9}$$

This conservation law characterizes the functional connection between the probability density function and the influx function of data points at a given spatial location and time.

3. NECESSARY CONVERGENCE CONDITION WITHOUT SUPERVISION

In this section, we establish the necessary condition for *correct* convergence of the mean-shift type of algorithm. We show that correct convergence can only be achieved if the algorithm can transform each individual cluster into the shape of a cluster that is consistent with the patterns of a normal distribution with zero covariances. This, however, does not mean that the algorithm can only be applied to patterns of normal distributions. It means that the algorithm can be applied to any shape of cluster. However, correct convergence can only be achieved if the choice of parameter or kernel can be chosen to transform the initial data patterns into those from normal distributions.

3.1 Unsupervised mean shift type clustering

In mean shift type clustering methods, the movements are governed by a law such that a data point is forced to move to a local center along the direction of the gradient while the size of such a move is adjusted by the value of the current density function at current location. We now define a class of mean-shift type clustering methods satisfying the following assumption:

$$u(\boldsymbol{x};t) = a^2 \frac{\nabla f(\boldsymbol{x};t)}{f(\boldsymbol{x};t)}. \tag{10}$$

Cheng (1995) [8] shows that the mean-shift algorithm indeed belongs to this category. All other variations based on the mean-shift method are also embraced by this category. A more general formulation of these movements in the traditional mean-shift method was discussed in Wang et al. (2007b) [14], in which data points are pushed towards local centers given by the conditional mean:

$$\boldsymbol{x}^{k+1} = \frac{\int_{B(\boldsymbol{x}^k,d)} \boldsymbol{t}\, f(\boldsymbol{t})\, d\boldsymbol{t}}{\int_{B(\boldsymbol{x}^k,d)} f(\boldsymbol{t})\, d\boldsymbol{t}}, \tag{11}$$

where $B(\boldsymbol{x}^k,d)$ is a neighborhood with the center located at $\boldsymbol{x}^k$ and radius d. Wang et al. (2007b) [14] showed that, for any $\alpha > 0$ and d such that $\int_{B(\boldsymbol{x}^k,d)} f(\boldsymbol{t})d\boldsymbol{t} = \alpha$, $\alpha > 0$, we have

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \frac{n\, d^2}{n+2} \frac{\nabla f(\boldsymbol{x}^k;t)}{f(\boldsymbol{x}^k;t)} + O(d^3). \tag{12}$$

This is consistent with the assumption (10).

The gradient component forces each data point to follow a trajectory to a local cluster center. This is known as the *mode seeking* property in clustering. Any movement is also proportional to the reciprocal value of the current probability density function. This implies that data points in sparsely populated areas will travel much longer distances when compared with those in densely populated regions even if the gradient functions assume the same value at these two different locations. This is an important computational advantage to speed up convergence.

3.1.1 Anti-diffusion and convergence for one-dimensional case

Combining eqn (8) with the assumption (10), we obtain the corresponding differential form as follows

$$\frac{\partial}{\partial t} f(x;t) = -a^2 \frac{d^2 f(x;t)}{dx^2}, \tag{13}$$

where $a > 0$ is a constant, with $f(x,0) = \phi_0(x)$ the initial probability density function.

This is a one-dimensional anti-diffusion equation in Kaashoek (1990) [15]. In comparison with classical and popular diffusion where data points move from high density regions to lower ones, the movements in dynamical clustering are in the opposite directions. We now present the exact analytical solution to eqn (13).

Theorem 3.1. *Under the assumption (10), the one-dimensional anti-diffusion equation has one unique solution and takes the following form*

$$f(x;t) = \frac{1}{\sqrt{-4a^2\pi t}} \int_{-\infty}^{\infty} \phi_0(\xi) e^{-\frac{(\xi-x)^2}{-4a^2 t}} d\xi, t \le 0, \tag{14}$$

where $f_0(x) = \phi_0(x)$, the initial probability density function.

This theorem can be proved by considering the diffusion equation $u(x,-t)$ for forward time and applying well-known results in linear reaction-diffusion equations. It differs from the well-known theorem in diffusion by a sign. The fact that the solution is specified uniquely only for $t \le 0$ implies that the evolution of densities produces deterministic causal events. Given the current probability density, there is only one unique process or path in the functional space that led to the present patterns.

The previous result about unsupervised clustering leads naturally to the following convergence result for dynamical clustering.

Theorem 3.2. *Under the assumption (10), we have*

(i) convergence to a location μ_0 of the clustering algorithm can only occur at $t = 0$ for normal densities with mean μ_0 and variance; proportional to a^2.

(ii) the first order derivative of the variance with respect to time is given by

$$\frac{d\sigma_t^2}{dt} = -2a^2, \tag{15}$$

where σ_t^2 denotes the variance of the normal density at time t;

(iii) the converging rate of a data point at a location x at time t is

$$u(x;t) = \frac{x-\mu_0}{2a^2(-t)}, \quad t \leq 0. \tag{16}$$

Proof. If the convergence at time $t = 0$ at a location μ_0, it implies that $f_0 = \delta(x-\mu_0)$. By Theorem 3.1, it then follows that

$$f(x;t) = \frac{1}{\sqrt{-4a^2\pi t}}\ e^{-\frac{(x-\mu_0)^2}{-4a^2 t}}, \quad t \leq 0. \tag{17}$$

The variance takes the form $2a^2(-t)$. Therefore, $\frac{d\sigma_t^2}{dt} = -2a^2$. The rest of the result follows from the assumption (10). □

The conclusion of this theorem shows that, for one-dimensional data, convergence to a single point can only occur for a normal density. In the next section, we will prove that the same result holds true for higher-dimensional spaces when there are multiple cluster centers.

3.1.2 Convergence in multi-dimensional space without supervision

Under the assumption (10), it then follows that

$$\boldsymbol{q}(\boldsymbol{x},t) = u(\boldsymbol{x},t); \quad f(\boldsymbol{x},t) = a^2 \nabla f(\boldsymbol{x},t). \tag{18}$$

Consequently, equation (6) now becomes

$$\frac{\partial f(\boldsymbol{x},t)}{\partial t} + a^2\ \nabla^2 f(\boldsymbol{x},t) = 0, \tag{19}$$

where the Laplacian of f, $\nabla^2 f = \sum_{i=1}^n \partial^2 f/\partial x_i^2$, and the boundary condition is given by $f(\boldsymbol{x},0) = f_0(\boldsymbol{x})$.

The result for the one dimensional case can be generalized to multi-dimensional cases when there are multiple cluster centers.

Theorem 3.3. *Under the assumption (10), the anti-diffusion equation*

$$\frac{\partial f(\boldsymbol{x},t)}{\partial t} + a^2\ \nabla^2 f(\boldsymbol{x},t) = 0, \tag{20}$$

with the boundary condition $f|_{t=0} = \phi_0$ has the solution

$$f(\boldsymbol{x},t) = (-4a^2 t)^{-n/2} \int_{(\boldsymbol{\eta})} \phi_0(\boldsymbol{\eta})\ e^{-\frac{(\boldsymbol{\eta}-\boldsymbol{x})^2}{-4a^2 t}}\ d\boldsymbol{\eta}, \tag{21}$$

where $(\boldsymbol{\eta}-\boldsymbol{x})^2 = \sum_{i=1}^n (\eta_i - x_i)^2$.

Proof. The dynamical clustering can be characterized as

$$\frac{\partial f(\boldsymbol{x},t)}{\partial t} + a^2\ \nabla^2 f(\boldsymbol{x},t) = 0, \tag{22}$$

with boundary condition $f(\boldsymbol{x},0) = \phi_0(\boldsymbol{x})$, a probability density function.

Denote the n-dimensional Fourier transformation by

$$F_n(\boldsymbol{s};t) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\boldsymbol{x},t) e^{i\ \boldsymbol{s}\cdot\boldsymbol{x}}\ dx_1 dx_2 \ldots dx_n, \tag{23}$$

where $\boldsymbol{s}\cdot\boldsymbol{x} = \sum_{i=1}^n s_i\ x_i$.

Applying Fourier transformation on both sides of equation (22), we have

$$\partial F_n(\boldsymbol{s},t)/\partial t + a^2\ ||\boldsymbol{s}|| F_n(\boldsymbol{s},t) = 0 \tag{24}$$

where $||\boldsymbol{s}|| = \sum_{i=1}^n s_i^2$. The solution is given by

$$F_n(\boldsymbol{s},t) = F_n(\boldsymbol{s},0)\ e^{-a^2\ ||\boldsymbol{s}||\ t}. \tag{25}$$

where

$$F_n(\boldsymbol{s},0) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \int_{(\boldsymbol{x})} \phi_0(\boldsymbol{x})\ e^{i\ \boldsymbol{s}\cdot\boldsymbol{x}}\ d\boldsymbol{x}. \tag{26}$$

Using inverse Fourier transformation, we get

$$f(\boldsymbol{x},t) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \int_{(\boldsymbol{s})} F_n(\boldsymbol{s},0)\ e^{-a^2\ ||\boldsymbol{s}||\ t - i\ \boldsymbol{s}\cdot\boldsymbol{x}} d\boldsymbol{s}, \tag{27}$$

therefore

$$f(\boldsymbol{x},t) = \left(\frac{1}{2\pi}\right)^n \int_{(\boldsymbol{s})} \left(\int_{(\boldsymbol{\eta})} \phi_0(\boldsymbol{\eta}) e^{i\ \boldsymbol{s}\cdot\boldsymbol{\eta}} d\boldsymbol{\eta}\right) e^{-a^2||\boldsymbol{s}||t - i\ \boldsymbol{s}\cdot\boldsymbol{x}} d\boldsymbol{s} \tag{28}$$

where $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)$. By rearranging the order of integration and simplifying the same way in Theorem 3.1, we then have

$$f(\boldsymbol{x},t) = (-4a^2 t\pi)^{-n/2} \int_{(\boldsymbol{\eta})} \phi_0(\boldsymbol{\eta})\ e^{-\frac{(\boldsymbol{\eta}-\boldsymbol{x})^2}{-4a^2 t}}\ d\boldsymbol{\eta}, \tag{29}$$

where $(\boldsymbol{\eta}-\boldsymbol{x})^2 = \sum_{i=1}^n (\eta_i - x_i)^2$. □

The analytic form allows us to retract the probability density function in the past and determine convergence from a given initial probability density function. We now show that the family of probability density functions that guarantees correct convergence to distinct multiple cluster centres must be a multivariate normal distribution with independent correlation structures.

Theorem 3.4. *Under the assumption (10), a dynamical shrinking or clustering converges to m distinct cluster centers if and only if the density function is a mixture of normal distribution with equal variances, i.e.*

$$f(\boldsymbol{x},t)=\sum_{i=1}^{m}\lambda_i\,\phi_i,\quad t<0. \tag{30}$$

where ϕ_i is the normal density function with mean $\boldsymbol{\mu}_i$ and variance $-2a^2t$.

Proof. If a dynamical clustering process converges to a finite number of focal points, *i.e.*,

$$\phi_0(\boldsymbol{\eta})=\sum_{j=1}^{m}\lambda_j\delta(\boldsymbol{\eta}-\boldsymbol{\mu}_j),\lambda_j\geq 0\ and\ \sum_{j=1}^{m}\lambda_j=1, \tag{31}$$

where $\boldsymbol{\mu}_j=(\mu_{j1},\mu_{j2},\ldots,\mu_{jn})$, and $\delta(\boldsymbol{\eta}-\boldsymbol{\mu}_j)=\prod_{i=1}^{n}\delta(\eta_j-\mu_{ji})$, then

$$f(\boldsymbol{x},t)=\sum_{j=1}^{m}\lambda_j\left(\frac{1}{2\pi\sigma_t^2}\right)^{-n/2}e^{-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_j)^2}{2\sigma_t^2}}, \tag{32}$$

where $\sigma_t^2=-2a^2t$, $t<0$. □

This theorem also implies that the contraction rates of dynamical shrinking or clustering must be homogenous in all directions at some time point of the clustering process to ensure correct convergence. A persistent heterogeneous contraction pattern therefore implies non-convergent behaviors.

4. STABILITY AND CONVERGENCE WITH SUPERVISION

Despite the past success in applications of mean shift algorithm and its intuitively appealing nature, our theoretical analysis has shown that this type of dynamical clustering algorithm actually might not be able to converge correctly unless it can successively transform and arrange data into independent Gaussian patterns. The necessary condition established in the previous section is surprising since it is counter intuitive. This section provides a theoretical explanation of the source of deterministic chaotic behavior and some theoretical guidance to resolve the instability problem.

4.1 Instability of unsupervised dynamic shrinking

The instability of anti-diffusion for a one dimensional case has been established in literature, see Kaashoek (1990) [15]. However, to the best of our knowledge, the results for higher dimensions have not been established. In order to understand more precisely stability of mean shift method for higher dimensions, we now examine the temporal evolution of the system using a quantity called energy function or Lyapunov function. This function has been widely used in dynamical systems and partial differential equations to describe the decay or growth of a system's energy. Detailed discussions can be found in Sastry (1999) [16]. In our case, this is a functional of the following form:

$$H(f)=-\int f(\boldsymbol{x})\,\log f(\boldsymbol{x})\,d\boldsymbol{x}, \tag{33}$$

In information theory, this is also called entropy, a measure of uncertainty.

Theorem 4.1. *Under the assumption (10), if*

$$\lim_{x_i\to\infty}\log f(\boldsymbol{x})\frac{\partial f(\boldsymbol{x})}{\partial x_i}=0,\ i=1,2,\ldots,m, \tag{34}$$

then

$$\frac{dH(f_t)}{dt}<0. \tag{35}$$

Proof. Consider the first order derivative of $H(f_t)$ with respect to t. We then have

$$\begin{aligned}
&\frac{dH(f_t)}{dt}\\
&=-\int_{-\infty}^{\infty}\frac{\partial}{\partial t}\left(f\left(\boldsymbol{x}\right)\log f\left(\boldsymbol{x}\right)\right)d\boldsymbol{x}\\
&=-\int_{-\infty}^{\infty}\left(\frac{df(\boldsymbol{x})}{dt}\log f(\boldsymbol{x})+\frac{1}{f(\boldsymbol{x})}f(\boldsymbol{x})\frac{df(\boldsymbol{x})}{dt}\right)d\boldsymbol{x}\\
&=-\int_{-\infty}^{\infty}\left(1+\log f(\boldsymbol{x})\right)\frac{df(\boldsymbol{x})}{dt}d\boldsymbol{x}\\
&=\int_{-\infty}^{\infty}\left(1+\log f(x)\right)a^2\left(\nabla^2 f(\boldsymbol{x},t)\right)d\boldsymbol{x}\\
&=\int_{-\infty}^{\infty}\left(1+\log f(x)\right)a^2\left(\sum_{i=1}^{n}\frac{\partial^2 f(\boldsymbol{x},t)}{\partial x_i^2}\right)d\boldsymbol{x}\\
&=-\int_{-\infty}^{\infty}a^2\frac{1}{f(\boldsymbol{x})}\sum_{i=1}^{n}\left(\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right)^2 d\boldsymbol{x}\\
&<0.
\end{aligned}$$

□

So the Lyapunov function is non-increasing at all times and hence, the dynamical clustering process is in violation of the second law of thermodynamics. Thus the dynamical clustering does not correspond to any natural (physical) process and is unstable except for data with normal densities. To ensure correct convergence by using the mean shift algorithm, a suitable intervention or supervision should be implemented.

4.2 Convergence with supervision

Mathematically, one possible formulation of supervision is to impose the so called *sink* or *force* function into the PDE framework in the following form:

$$\frac{\partial f(\boldsymbol{x},t)}{\partial t} + a^2\,\nabla^2 f(\boldsymbol{x},t) = \psi(\mathbf{x},t), \tag{36}$$

where ψ is a continuous function.

We now show that correct convergence can be established through non-normal densities with the help of supervision function.

Theorem 4.2. *Under the assumption (10), we have*

(i) the PDE associated with supervised clustering has the following solution

$$\begin{aligned} & f(\boldsymbol{x},t) \\ & = (-4a^2t\pi)^{-n/2}\int_{(\boldsymbol{\eta})}\phi_0(\boldsymbol{\eta})e^{-\frac{(\boldsymbol{\eta}-\boldsymbol{x})^2}{-4a^2t}}\,d\boldsymbol{\eta} \\ & + \int_t^0\int_{(\boldsymbol{\xi})}\psi(\boldsymbol{\xi},\tau)[-4a^2(t-\tau)\pi]^{-n/2}e^{-\frac{(\boldsymbol{x}-\boldsymbol{\eta})^2}{-4a^2(t-\tau)}}\,d\boldsymbol{\xi}d\tau, \\ & t\le 0. \end{aligned} \tag{37}$$

(ii) if the clustering process converges to m distinct focal points, then

$$\begin{aligned} & f(\boldsymbol{x},t) \\ & = \sum_{j=1}^{m}\lambda_j\left(\frac{1}{2\pi\sigma_t^2}\right)^{-n/2}e^{-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_j)^2}{2\sigma_t^2}} \\ & + \int_t^0\int_{(\boldsymbol{\xi})}\psi(\boldsymbol{\xi},\tau)[-4a^2(t-\tau)\pi]^{-n/2}e^{-\frac{(\boldsymbol{x}-\boldsymbol{\eta})^2}{-4a^2(t-\tau)}}\,d\boldsymbol{\xi}d\tau, \\ & t\le 0. \end{aligned} \tag{38}$$

Proof. The general solution with the sink function in the PDE can be decomposed into two parts:

$$f(\boldsymbol{x},t) = g_1(\boldsymbol{x},t) + g_2(\boldsymbol{x},t), \tag{39}$$

where g_1 is the solution for the PDE eqn (20) with boundary condition $g_1(t=0)=\phi_0$ and g_2 satisfying the PDE eqn (36) with the boundary condition $g_{t=0}^2 = 0$. The function form of g_1 is given by Theorem 3.3. That is,

$$g_1(\boldsymbol{x},t) = (-4a^2t\pi)^{-n/2}\int_{(\boldsymbol{\eta})}\phi_0(\boldsymbol{\eta})\,e^{-\frac{(\boldsymbol{\eta}-\boldsymbol{x})^2}{-4a^2t}}\,d\boldsymbol{\eta},$$

where

$$(\boldsymbol{\eta}-\boldsymbol{x})^2 = \sum_{i=1}^{n}(\eta_i - x_i)^2.$$

To find g_2, we consider a nonhomogeneous differential equation of the form

$$L_{\boldsymbol{x}}\,u(\boldsymbol{x}) = \psi(\boldsymbol{x},t), \tag{40}$$

where

$$L_{\boldsymbol{x}}\,u(\boldsymbol{x}) = \frac{\partial u(\boldsymbol{x},t)}{\partial t} + a^2\,\nabla^2 u(\boldsymbol{x},t). \tag{41}$$

The Green function $G(\boldsymbol{x},\boldsymbol{\xi})$ of this problem satisfies the equation

$$L_{\boldsymbol{x}}\,G(\boldsymbol{x},\boldsymbol{\xi}) = \delta(\boldsymbol{x}-\boldsymbol{\xi})\delta(t-\tau),\ \ G_{t=0}=0. \tag{42}$$

The solution for the partial differential equation (40) is then given by

$$u(\boldsymbol{x}) = \int_t^0\int_{(\boldsymbol{\xi})}\psi(\boldsymbol{\xi},\tau)\,G(\boldsymbol{x},t;\boldsymbol{\xi},\tau)\,d\boldsymbol{\xi}d\tau, \tag{43}$$

where the Green function satisfying the following PDE

$$\frac{\partial\,G(\boldsymbol{x},t)}{\partial t} + a^2\,\nabla^2 G(\boldsymbol{x},t) = 0,\ \ G|_{t=\tau} = \delta(\boldsymbol{x}-\boldsymbol{\xi}).$$

By Theorem 3.3 and replacing t by $t-\tau$, the Green function is then given by

$$\begin{aligned} & G(\boldsymbol{x},t) \\ & = [-4a^2(t-\tau)\pi]^{-n/2}\int_{(\boldsymbol{\eta})}\delta(\boldsymbol{\eta}-\boldsymbol{\xi})e^{-\frac{(\boldsymbol{\eta}-\boldsymbol{x})^2}{-4a^2(t-\tau)}}\,d\boldsymbol{\eta} \\ & = [-4a^2(t-\tau)\pi]^{-n/2}e^{-\frac{(\boldsymbol{x}-\boldsymbol{\eta})^2}{-4a^2(t-\tau)}}, \end{aligned} \tag{44}$$

where $(\boldsymbol{x}-\boldsymbol{\eta})^2 = \sum_{i=1}^{n}(x_i-\eta_i)^2$.

It then follows that

$$\begin{aligned} & g_2(\boldsymbol{x},t) \\ & = \int_t^0\int_{(\boldsymbol{\xi})}\psi(\boldsymbol{\xi},\tau)[-4a^2(t-\tau)\pi]^{-n/2}e^{-\frac{(\boldsymbol{x}-\boldsymbol{\xi})^2}{-4a^2(t-\tau)}}\,d\boldsymbol{\xi}d\tau, \\ & t\le 0. \end{aligned} \tag{45}$$

Therefore,

$$\begin{aligned} & f(\boldsymbol{x},t) \\ & = \frac{1}{M}(-4a^2t\pi)^{-n/2}\int_{(\boldsymbol{\eta})}\phi_0(\boldsymbol{\eta})e^{-\frac{(\boldsymbol{x}-\boldsymbol{\eta})^2}{-4a^2t}}\,d\boldsymbol{\eta} \\ & + \frac{1}{M}\int_t^0\int_{(\boldsymbol{\xi})}\psi(\boldsymbol{\xi},\tau)[-4a^2(t-\tau)\pi]^{-n/2}\ e^{-\frac{(\boldsymbol{x}-\boldsymbol{\xi})^2}{-4a^2(t-\tau)}}\,d\boldsymbol{\xi}d\tau, \\ & t\le 0, \end{aligned} \tag{46}$$

where M is the normalizing constant to ensure that $f(\boldsymbol{x},t)$ is a proper probability density function. □

The fact that the original density function of the PDE is a function of the supervision function implies that correct convergence is dependent on the choice of the supervising function. The assertion of the theorem indicates that a universally effective supervising function might not exist. A supervising function then must be chosen judiciously to ensure correct convergence. A self-adaptive learning algorithm will also require that the sink function be a functional of the current and historical densities, and this issue is left for future studies. One such example is the crystallization processes

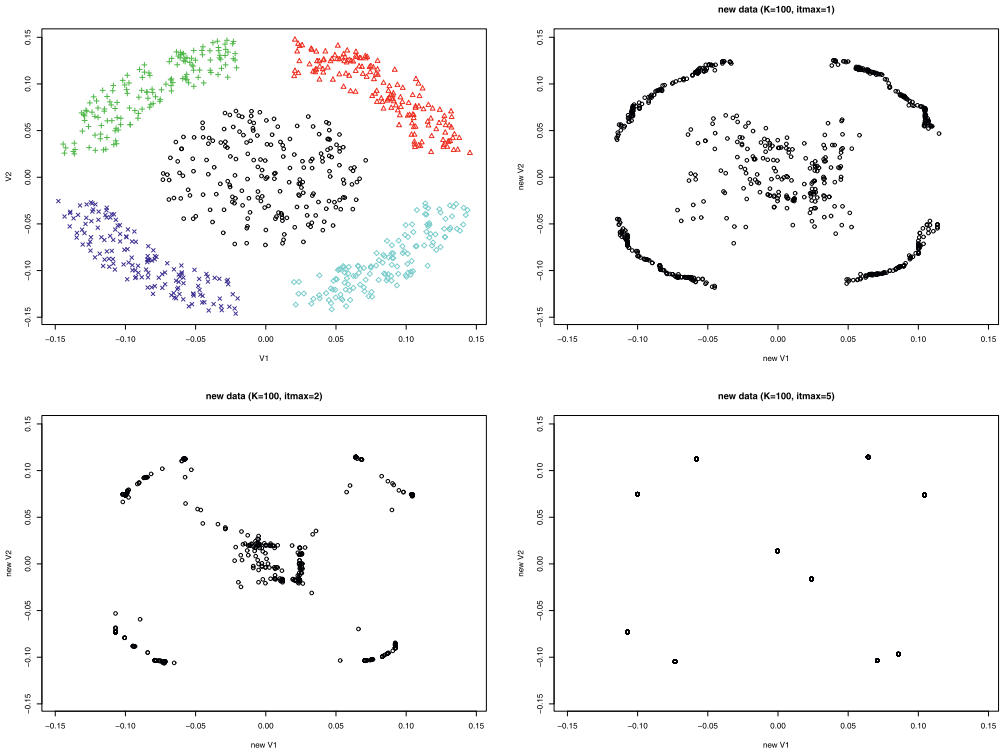

Figure 1. The upper-left plot of this figure displays the 5 clusters of the Broken Ring data set (cf. Wang et al. (2007a) [6]), which has 800 *data points measured by* 2 *variables. The upper-right plot, bottom-left plot, and bottom-right plot show the evolution of the convergence of the* 800 *data points towards the focal points after one, three and five iterations of the K-nearest-neighbor mean-shift algorithm. The number of nearest neighbors were set to be* $K = 100$.

as described in Teran and Bill (2010) [17]. It is stable due to the fact that particles are accumulating and transformed into solid with zero speed due to the crystallization. This might inspire a certain choice of supervision for dynamical clustering.

5. EMPIRICAL SUPERVISION AND SIMULATION STUDIES

There are many variants of the general mean-shift type of algorithms. In this section, we chose an adaptive version called *clues* proposed in Wang et al. (2007a) [6]. The algorithm *clues* employs K-nearest neighbour approach in the dynamic clustering with supervision function to be described in this section. To demonstrate the convergence and stability issues associated with unsupervised mean-shift algorithm, we present some clustering results on a simulated data used in Wang et al. (2007a) [6].

This simulated data set has five well separated clusters. Four clusters are in symmetric positions while one cluster sits in the middle region. This data is referred *Broken Ring* in Wang et al. (2007a) [6]. It is shown in Figure 1. Wang et al. (2007a) [6] show that classical methods of clustering have considerable difficulty to handle this simple data set. Wang et al. (2007a) [6] proposed a robust version of mean-shift algorithm while the local mean is replaced by the median in conjunction with the employment of K-nearest neighbors approach instead of a kernel function.

We present the impact of the robust version of the mean-shift algorithm in Figures 1, 2 and 3 which contain the original data set and positions of data points after one, three and five iterations respectively. By using $K = 100$ in the K-nearest neighbor approach, one can observe, in Figure 1, that the robust mean-shift algorithm is shrinking data towards several focal points. Although convergence is reached, the algorithm produced 10 small clusters by splitting each real cluster into two. By using $K = 200$, it can be seen from Figure 2, that the middle cluster was *incorrectly* absorbed by the two clusters. In both cases, convergence does not mean correct convergence and the instability is obvious.

To achieve a reliable and robust result, one would need a proper supervision. A theoretical supervision function is currently out of reach due to the complexity and uncertainty in clustering large data from non-normal densities. Therefore one would need an empirical supervisions or guidance to evaluate whether a generated clustering result is indeed reasonable. These kind of empirical supervision must be done without the knowledge of the true underlying cluster memberships. They are therefore exterior measures in this sense. These measures can be applied to evaluate clustering results

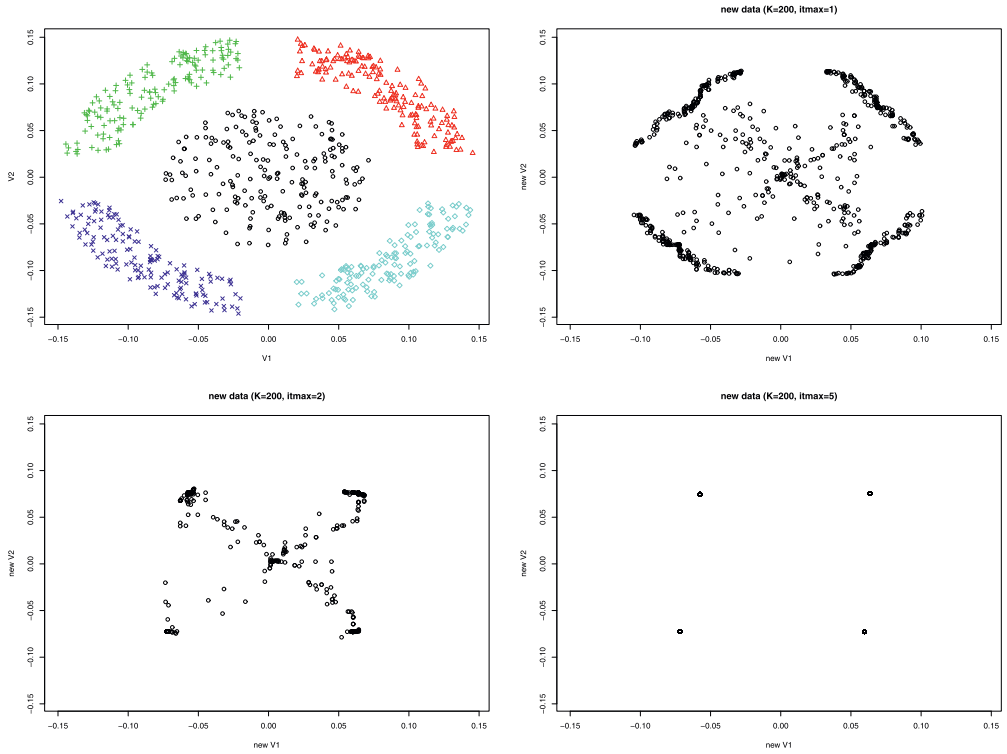

Figure 2. The upper-left plot of this figure displays the 5 clusters of the Broken Ring data set (cf. Wang et al. (2007a) [6]), which has 800 *data points measured by* 2 *variables. The upper-right plot, bottom-left plot, and bottom-right plot show the evolution of the convergence of the* 800 *data points towards the focal points after one, three and five iterations of the K-nearest-neighbor mean-shift algorithm. The number of nearest neighbors were set to be* $K = 200$.

and thus determine the path of clustering in the entire clustering process. There are many measures of the strength of clusters. For example Milligan and Cooper (1985) [18] compared 30 measures of the strengths of clusters for determining the number of clusters. Their simulation results show that the Calinski and Harabasz index (CH index) has the best performance. The CH index is defined as

$$CH(g) = \frac{B(g)/(g-1)}{W(g)/(n-g)} \tag{47}$$

where g is the number of clusters,

$$B(g) = \sum_{i=1}^{g} n_i(\bar{\boldsymbol{y}}^{(i)} - \bar{\boldsymbol{y}})(\bar{\boldsymbol{y}}^{(i)} - \bar{\boldsymbol{y}})^T \tag{48}$$

is the between-groups sum of squares and products matrix,

$$W(g) = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\boldsymbol{y}_j^{(i)} - \bar{\boldsymbol{y}}^{(i)})(\boldsymbol{y}_j^{(i)} - \bar{\boldsymbol{y}}^{(i)})^T \tag{49}$$

is the within-groups sum of squares and products matrix, n_i is the size of the i-th cluster, $\boldsymbol{y}_j^{(i)}$ is the j-th data point in the i-th cluster, $\bar{\boldsymbol{y}}^{(i)}$ is the mean vector of the i-th cluster, and $\bar{\boldsymbol{y}}$ is the overall mean of all data points.

Kaufman and Rousseeuw (1990) [19] proposed the Silhouette index to measure the strengths of clusters. The silhouette index is defined as

$$\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{50}$$

where

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{51}$$

and a_i is the average distance of the data point $\boldsymbol{y}_i$ to other points in the cluster A where $\boldsymbol{y}_i$ belongs to, i.e.

$$a_i = \frac{1}{(n_A - 1)}\sum_{j\in A, j\neq i} d(\boldsymbol{y}_i, \boldsymbol{y}_j),$$

and b_i is the average distance to points in the nearest neighbor cluster besides its own. Define

$$d(i, C) = \text{average dissimilarity of the data point } \boldsymbol{y}_i \text{ to all data points in Cluster } C.$$

Then

$$b_i = \min_{C\neq A} d(i, C).$$

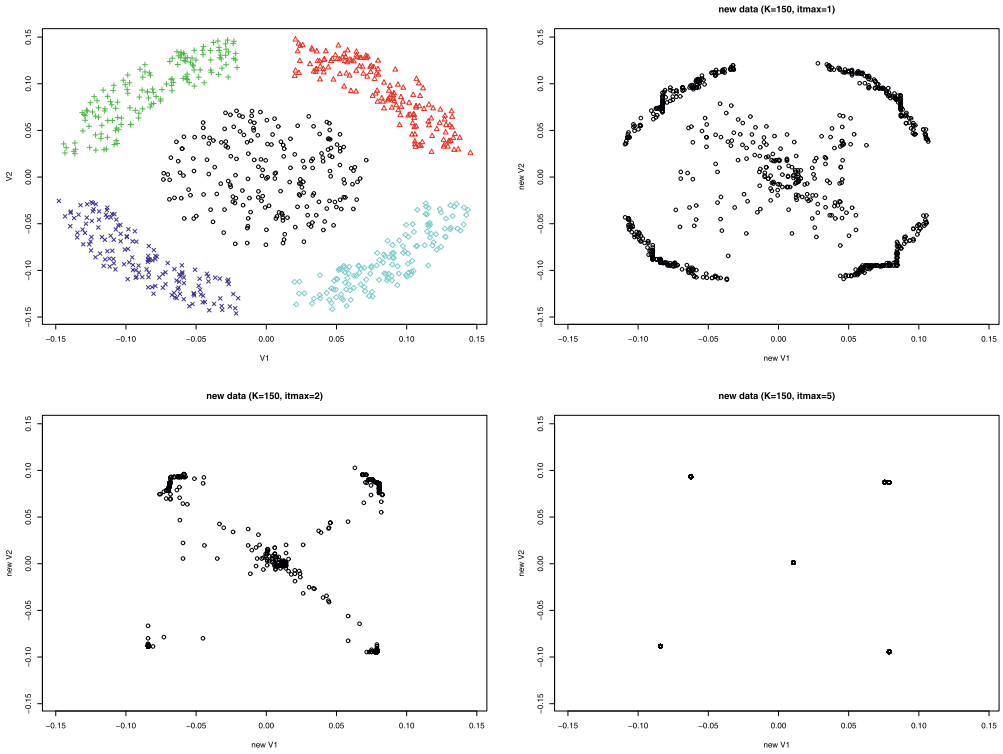

Figure 3. The upper-left plot of this figure displays the 5 clusters of the Broken Ring data set (cf. Wang et al. (2007a) [6]), which has 800 *data points measured by* 2 *variables. The upper-right plot, bottom-left plot, and bottom-right plot show the evolution of the convergence of the* 800 *data points towards to the focal points after one, three and five iterations of the K-nearest-neighbor mean-shift algorithm. The number of nearest neighbors were set to be* $K = 150$.

This index s_i can take values from -1 to 1. When the index is zero, the data point $\boldsymbol{y}_i$ has equal distance to its cluster and its nearest neighbor cluster. If the index is positive, then the data point $\boldsymbol{y}_i$ is closer to its cluster than other clusters. If the index is negative, then the data point $\boldsymbol{y}_i$ is wrongly assigned to the current cluster. Thus, if all data points are correctly assigned, then average of s_i's should be close to 1.

As one can see from Figure 3, with the help of supervision, the algorithm proposed in Wang et al. (2007a) [6] can select the correct clustering result among many candidates generated by different choices of K. The algorithm is implemented in the package *clues* in the open-source statistical software *R* [20]. Detailed instructions and description of this R package are available in Chang et al. (2010) [21]. However, this is an end-of-process supervision and is much more computationally intensive than in-process ones. To find a proper in-process supervising function remains an active research topic.

6. DETECTING PROSTATE CANCER SUBGROUPS BY USING CLUES

In this section, we demonstrate the convergence and stability issues associated with unsupervised mean-shift algorithm by using a real data set *prostate* in biomedical research field. The data set *prostate* is available in R [20] package *ElemStatLearn*.

Prostate cancer (PcA) is the most common cancer among men (after skin cancer). It may not cause signs or symptoms in its early stages. Prostate-specific antigen (PSA) test has been widely used to screen men for PcA. According to the National Cancer Institute[1], the PSA test measures the blood level of PSA, a protein that is produced by the prostate gland. The higher a man's PSA level, the more likely it is that he has PcA.

A PSA test is also used to monitor men who have been diagnosed with PcA to see if their cancer has recurred after initial treatment or is responding to therapy. Stamey et al. (1989) [22] examined the correlation between the level of post-operative PSA and a number of clinical measures (such as seminal vesicle invasion, cancer volume, gleason score, prostate weight, amount of benign prostatic hyperplasia, and capsular penetration) in 97 PcA patients measured before they received a radical prostatectomy. Seminal vesicle invasion is a binary variable and gleason score is an ordinal variable taking values 6, 7, 8, and 9. The post-operative PSA and the other clinical variables are continuous variables with

[1] http://www.cancer.gov/cancertopics/factsheet/detection/PSA

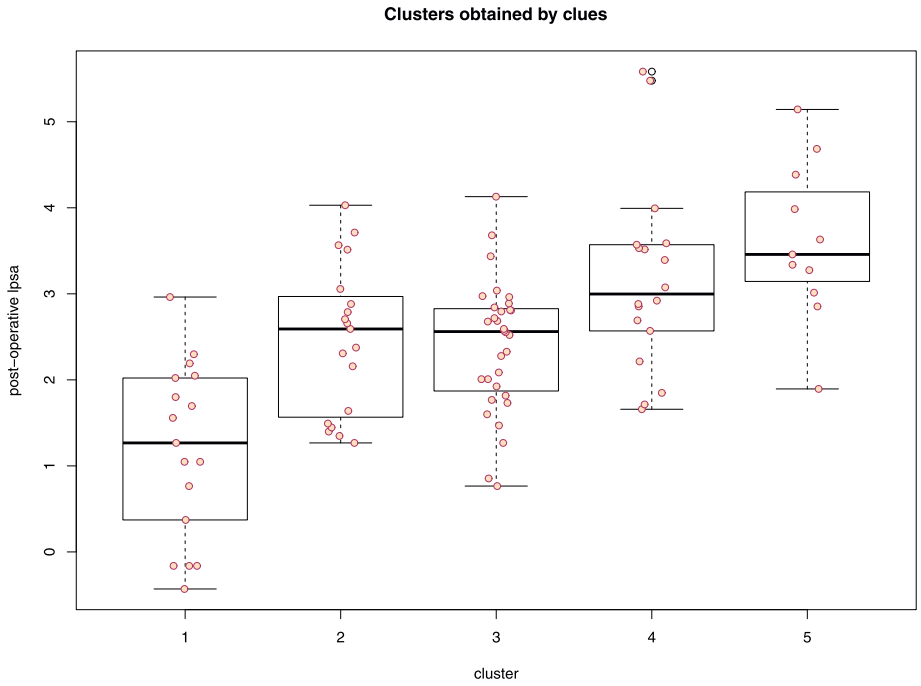

Figure 4. This figure shows the parallel boxplots of log post-operative PSA level (lpsa) across the 5 clusters detected by the clustering algorithm clues for the prostate cancer data set described in Section 6, in which four log transformed variables (cancer volume, prostate weight, amount of benign prostatic hyperplasia, and capsular penetration) were used to group the 97 prostate cancer patients. Based on the prostate cancer data set, Stamey et al. (1989) [22] examined the correlation between the level of post-operative PSA and a number of clinical measures, including the aforementioned 4 clinical variables. This figure shows that the log post-operative PSA levels are different among the 5 clusters that were formed based on the 4 clinical variables. In this figure, each boxplot corresponds to a cluster. The jittered red circles along the boxplot indicate the data points in the cluster. (Color figure online)

skewed distributions and were log transformed (denoted as lpsa, lcavol, lweight, lbph, and lcp, respectively).

We applied *clues* to the data set *prostate* to detect subgroups of the 97 PcA patients based on the 4 continuous clinical variables lcavol, lweight, lbph, and lcp. We then checked if the subgroups have different post-operative PSA levels. If radical prostatectomy is effective, the post-operative PSA levels would be low. Hence, subgroups having low post-operative PSA levels would benefit more from radical prostatectomy than subgroups having high post-operative PSA levels.

We used the default setting of the *clues* function in *clues* R package [21] to cluster the 97 PcA patients based on lcavol, lweight, lbph and lcp. Silhouette index was used to measure the compactness of the clusters. Five clusters were obtained. The cluster sizes are 17, 19, 32, 18, and 11, respectively.

Figure 4, which illustrates the parallel boxplots of the log post-operative PSA levels (lpsa) across the 5 clusters, shows that PcA patients in cluster 1 probably were more responsive to the radical prostatectomy therapy than PcA patients in cluster 5. Clusters 2 and 3 had similar lpsa level and the two clusters had higher lpsa than cluster 1, but had lower lpsa than cluster 4, which in turn had lpsa lower than cluster 5.

By examining the median levels of the 4 log transformed clinical variables for the 5 clusters (Table 1), we found that cluster 1 had much smaller median cancer volume and capsular penetration than cluster 5, but had similar median weight and amount of benign prostatic hyperplasia before the radical prostatectomy therapy. Although cluster 2 and cluster 3 had similar post-operative lpsa, PcA patients in cluster 2 had much smaller lbph and lcp than PcA patients in cluster 3. PcA patients in cluster 4 had smaller lcavol and lcp, but larger lweight and lbph than PcA patients in cluster 5.

We further examined if the 5 clusters detected by clues are separate from each other in the 4-dimensional space spanned by lcavol, lweight, lbph, and lcp. The separateness between each pair of clusters was measured by the separation index proposed by Qiu and Joe [23]. The separation index takes values from -1 (totally overlapping) to 1 (totally separated). A separation index value 0 indicates the 2 clusters are just "touching".

Table 2 shows the matrix of the pairwise separation indices. All pairwise separation indices are positive or

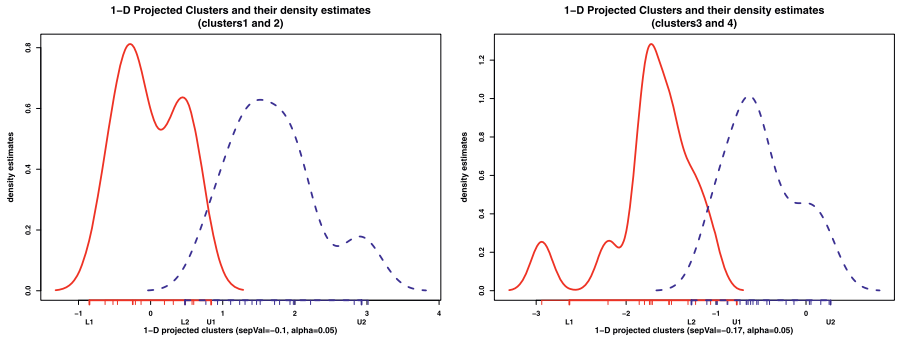

Figure 5. This figure shows the 1-dimensional projection of a pair of clusters along the optimal projection direction based on the method proposed by Qiu and Joe [23]. The colored ticks on x-axis indicate the positions of data points along the optimal projection direction. The red color ticks are for the 1st cluster in the pair, while the blue color ticks are for the 2nd cluster in the pair. L_i and U_i are the lower and upper 2.5-th sample percentile of the cluster i along the optimal projection direction, $i = 1, 2$. Density estimates of the pair of clusters are also shown in the figure. The 1-D projections in this figure are for 2 pairs of clusters among the 6-cluster partition of the prostate cancer data set (97 data points in 4 dimensional space) described in Section 6. The 6-cluster partition was obtained by the K-nearest-neighbor mean-shift algorithm, which moved the 97 data points in a 4-dimensional space toward 6 focal points ($K = 15$). The 2 plots in this figure are the 1-D projection for clusters 1 and 2 (left panel) and for clusters 3 and 4 (right panel), respectively, This figure shows that clusters 3 and 4 are overlapping (separation index value $= -0.17$), while clusters 1 and 2 are slightly overlapping (separation index value $= -0.10$). (Color figure online)

Table 1. Cluster medians for the 4 log transformed clinical variables

cluster	size	lcavol	lweight	lbph	lcp
1	17	−0.45	3.27	−1.39	−1.39
2	19	1.61	3.62	−1.39	−0.60
3	32	1.11	3.78	1.62	−1.39
4	18	2.03	3.75	1.42	1.40
5	11	2.83	3.58	−1.39	2.17

Table 2. Pairwise separation index for the 5 clusters obtained by clues. -1 indicates totally overlapping; 0 indicates just touching; 1 indicates totally separated.

	c1	c2	c3	c4	c5
c1	−1.00	−0.08	0.38	0.37	0.44
c2	−0.08	−1.00	0.36	0.22	−0.01
c3	0.38	0.36	−1.00	0.03	0.44
c4	0.37	0.22	0.03	−1.00	0.14
c5	0.44	−0.01	0.44	0.14	−1.00

close to zero, indicating the 5 clusters detected by clues are separated in the 4-dimensional space spanned by lcavol, lweight, lbph, and lcp. The online supplementary figure (http://www.intlpress.com/SII/p/2016/9-2/SII-9-2-WANG-supplement.pdf) illustrated the magnitude of separation along the estimated optimal one-dimensional projection for each pair of clusters.

In comparison, if we directly set the number of nearest neighbors as $K = 15$ and ignore the supervision guided by Silhouette index, we obtained 6 clusters. By examining the pairwise separation indices among the 6 clusters, we found that cluster 3 and cluster 4 are overlapped with separation index -0.17; and (2) cluster 1 and cluster 2 are slightly overlapped with separation index -0.10. Figure 5 illustrates the magnitudes of overlapping between clusters 1 and 2 and between clusters 3 and 4 along the optimal projection directions that separate the 2 clusters in each pair of clusters. The separation index value -0.17 and Figure 5 indicates that clusters 3 and 4 probably should be in one cluster. Hence, the 6-cluster partition of the *prostate cancer* data set obtained by using unsupervised mean-shift algorithm seems unreasonable.

7. DISCUSSION

We fill a critical theoretical gap in the literature between the reported successes in various applications and the lack of convergence and stability analysis of mean shift type of dynamical clustering algorithms. We employ the conservation law from physics and establish a general partial differential equation framework that prescribes the spatio-temporal evolution of dynamical clustering processes. We show that, in the absence of supervision, mean shift clustering and its variations may not result in correct convergence in general unless the underlying probability distribution can be transformed to normal densities. The non-decreasing backward in time of the Lyapunov function and the anti-diffusion nature of these dynamical clustering algorithms render them universally highly unreliable without a proper supervision. Therefore a supervised mean shift clustering should be preferred and a supervision function must be chosen carefully to ensure valid clustering results. The theoretical analysis of the role of a supervision function is useful on how to choose such a function in practice. This is our future research.

Received 14 October 2014

REFERENCES

[1] Fukunaga, K. and Hostetler, L. D. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21:32–40, 1975. MR0388638

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

[3] Virmajoki, O., Franti, P., and Kaukoranta, T. Iterative shrinking method for generating clustering. In *Proc. of the 6th European Conference on Computer Systemedings of the International Conference on Image Processing*, volume 2, pages 685–688. IEEE, 2002.

[4] Shi, Y., Song, Y., and Zhang, A. A shrinking-based clustering approach for multidimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 17:1389–1403, 1995.

[5] Woolfold, D. G. and Braun, W. J. Convergent data sharpening for the identification and tracking of spatial temporal centers of lightning activity. *Envirometrics*, 18:461–479, 2007. MR2380766

[6] Wang, X., Qiu, W., and Zamar, R. H. CLUES: a non-parametric clustering method based on local shrinking. *Computational Statistics and Data Analysis*, 52:286–298, 2007. MR2409982

[7] Choi, E. and Hall, P. Data sharpening as a prelude to density estimation. *Biometrika*, 86:941–947, 1999. MR1741990

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

[9] Einstein, A. *Investigations on the Theory of the Brownian Movement*. Dover Publications, Inc., 1956. MR0077443

[10] Pettini, M. *Geometry and Topology in Hamiltonian Dynamics and Statistical Mechanics*. Springer, New York, 2007. MR2331296

[11] Ellis, S. R. *Entropy, Large Deviations and Statistical Mechanics*. Springer, New York, 1985. MR0793553

[12] Simonoff, J. S. *Smoothing Methods in Statistics*. Springer-Verlag, New York, 1996. MR1391963

[13] Debnath, L. *Nonlinear Differential Equations for Scientists and Engineers*. Birkhauser, Boston, 2004.

[14] Wang, X., Liang, D., Feng, X., and Ye, L. A derivative-free optimization algorithm based on conditional moments. *Mathematical Analysis and Applications*, 331:1337–1360, 2007. MR2313717

[15] Kaashoek, J. F. *Modeling one Dimensional Pattern Formation by Anti-Diffusion*. CWI, 1990. MR1080355

[16] Sastry, S. *Nonlinear Systems: Analysis, Stability and Control*. Springer-Verlag, New York, 1999. MR1693648

[17] Teran, A. V., Bill, A., and Bergmann, R. B. Time-evolution of grain size distributions in random nucleation and growth crystallization processes. *Phys. Rev. B*, 81:075319, Feb 2010.

[18] Milligan, G. W. and Cooper, M. C. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–179, 1985.

[19] Kaufman, L. and Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990. MR1044997

[20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[21] Chang, F., Qiu, W., Zamar, R. H., Lazarus, R., and Wang, X. Clues: an R package for nonparametric clustering based on local shrinking. *Journal of Statistical Software*, 33(4):1–16, 2010.

[22] Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., and Yang, N. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii. Radical prostatectomy treated patients. *Journal of Urology*, 16:1076–1083, 1989.

[23] Qiu, W. and Joe, H. Separation index and partial membership for clustering. *Computational Statistics and Data Analysis*, 50:585–603, 2006. MR2196285

Xiaogang Wang
Department of Mathematics and Statistics
York University, Toronto
Canada
E-mail address: stevenw@mathstat.yorku.ca

Weiliang Qiu
Channing Division of Network Medicine
Brigham and Women's Hospital
Harvard Medical School
USA
E-mail address: stwxq@channing.harvard.edu

Jianhong Wu
Department of Mathematics and Statistics
York University, Toronto
Canada
E-mail address: wujh@mathstat.yorku.ca