Journal of Statistics Applications & Probability
*An International Journal*

http://dx.doi.org/10.18576/jsap/050201

# A Parameter Leveraging Method for Unsupervised Big Data Modelling

*Kassim Mwitondi*[1] *and Eman Khorsheed*[2,*]

[1] Faculty of Arts, Computing, Engineering and Sciences, Sheffield Hallam University, Sheffield, United Kingdom.
[2] Department of Mathematics, University of Bahrain, Kingdom of Bahrain.

Received: 3 Jun. 2016, Revised: 10 Jun. 2016, Accepted: 12 Jun. 2016
Published online: 1 July 2016

**Abstract:** Increasingly sophisticated methods and tools are needed for tracking the dynamics and detecting inherent structures in modern day highly voluminous multi-faceted. Data scientists have long realized that tackling global challenges such as climate change, terrorism and food security cannot be contained within the frameworks and models of conventional data analysis. For example, separating noise from meaningful data in even a low-dimensional data with heavy tails and/or overlaps is quite challenging and standard non-linear approaches do not always succeed. Tracking the dynamics of multi-faceted data involving complex systems is tantamount to tracking agent-based complex systems with many interacting agents. Dimensional-reduction methods are commonly used to try and capture structures inherent in data but they do not generally lead to optimal solutions mainly because their optimisation functions and theoretical methods typically rely on special structures. We propose a parameter leveraging method for unsupervised big data modelling. The method searches for structures in data and creates a series of sub-structures which are subsequently merged or split. The strategy is to present the algorithm with a set of periodic data as one complex system. It then uses the patterns in the sub-structures to determine the overall behaviour of the complex system. Applications on solar magnetic activity cycles and seismic data show that the proposed method out-performs conventional unsupervised methods. We illustrate how the method can be extended to supervised modelling.

**Keywords:** Big Data, Clustering, Data Mining, Data Visualisation, k-Means, Optimisation, Seismic Signals, Sunspots, Unsupervised Modelling.

## 1 Introduction

Extracting knowledge from data continues to stimulate interdisciplinary research across the world mainly because the complex nature of global challenges such as climate change, terrorism and food security can no longer be tackled in isolation. In the big data era, data scientists are embroiled in tracking data dynamics, volume and variety as most multi-faceted data applications depart from the conventional models of data analysis. For example, separating noise from meaningful data in even a low-dimensional data with heavy tails and/or overlaps is quite challenging and standard non-linear approaches do not always succeed in detecting naturally arising structures in such circumstances. Identifying natural structures becomes even more challenging under the big data scenario in which data systems become increasingly complex and attribute relationships less obvious. Applications of mathematical structures in describing general behaviours of complex systems are well-documented. In such applications the main goal is typically to investigate how inter-relationships among attributes of partial systems lead to generalisations about aspects of broader systems [1]. A wide range of unsupervised and supervised modelling methods are used across applications. Dimensional-reduction methods are commonly used to try and capture structures inherent in data but they do not generally lead to optimal solutions mainly because their optimisation functions and theoretical methods typically rely on special structures. Large volumes of multi-faceted data related to global challenges - climate change, terrorism and food security - continue to flow in the big data era. Tracking their dynamics, volume and variety inevitably entails more sophisticated methods and tools for detecting inherent structures.

[*] Corresponding author e-mail: ekhorsheed@uob.edu.bh

Tracking the dynamics of multi-faceted data involving complex systems is tantamount to tracking agent-based complex systems with many interacting agents. We propose a parameter leveraging method for unsupervised big data modelling. The method sequentially searches for structures in data and creates a series of sub-structures which are subsequently merged or split. We follow [2] who used pattern-oriented modelling framework to design, test and analyse bottom-up models. The strategy is to present the algorithm with a set of periodic data as one complex system. It then uses the patterns in the sub-structures to determine the overall behaviour of the complex system. Applications on solar magnetic activity cycles and seismic data show that the proposed method out-performs conventional unsupervised methods. The algorithm inherently illustrates how it can be extended to supervised modelling applications. The paper is organised as follows. An overview of unsupervised modelling is given in Section 2; methods and data description in Section 3 are followed by data analyses, results and discussions in Section 4 and concluding remarks in Section 5.

## 2 BACKGROUND OF UNSUPERVISED MODELLING

The main idea of extracting knowledge from data relies on addressing the two main data mining problems - unsupervised and supervised modelling which, in a conventional statistical jargon, can be described as data clustering and classification/regression. This section provides focuses on the former. Under unsupervised modelling data points are allocated to, a priori, unknown groups (clusters) with those in each cluster being as homogeneous as possible while those between clusters being as heterogeneous as possible. The allocation rule is based on some measure of similarity - typically, the distance between data points. Unsupervised modelling is based on the well-known finite mixtures model [3] and [4] which constitutes a set of probability distributions each associated with membership to one of K defined clusters. Its mechanics can be illustrated by the category utility [5] which provides a measurement of partition quality as data are allocated to different clusters (categories). For instance, given clusters $C_{i=1,2,\ldots,K-1,K}$the category utility is define as

$$C_{uK} = \frac{\sum_l P(C_l)\sum_i \sum_j (P[x_i = \xi_{ij}|C_l]^2 - P[x_i = \xi_{ij}|C_i]^2)}{K} \tag{1}$$

where the outer summation is over the clusters $C_{i=1,2,\ldots,K-1,K}$ and the first inner summation is over the data attributes $x_i$ as they assume specific values $\xi_{i1,2,3,\ldots}$ summed over $j$. The main idea of Equation (1) is that the probability of a particular attribute assuming a specific value within a given cluster provides a better estimation than just the probability of an attribute assuming a specific value. Thus, the difference between the squared probabilities over all attributes and values is crucial in determining the usefulness of the clusters. The denominator gives per cluster measure to avoid overfitting [6]. Equation (1) can be extended to continuous variables by assuming a Gaussian model,

$$f(x,\mu,\sigma) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

with the analogy

$$P[x_i = \xi_{ij}|C_l]^2 \Leftrightarrow \int f(x_i)^2 dx_i = \frac{1}{2\sqrt{\pi\sigma_i}}.$$

Variation is estimated within cluster $(\sigma_{il})$ and over all clusters $(\sigma_i)$ as follows

$$C_{uK} = \frac{\sum_l P(C_l)\frac{1}{2\sqrt{\pi}}\sum_i(\frac{1}{\sigma_{il}} - \frac{1}{\sigma_i})}{K} \tag{2}$$

Typical unsupervised modelling algorithms such as the k-Means [7] and the EM [8], typically manage inherent data randomness and dependency on starting points via cross validation. This commonly accepted practice still leaves open the optimality challenge [9], effectively emphasising the need for model assessment such as the one in Equation (1).

## 3 METHODS AND DATA DESCRIPTION

The proposed strategy is inspired by the Gaussian mixture model approach to density estimation in which data are viewed as coming from a mixture of probability Gaussian distributions, each representing a different cluster [3]. Given data, the strategy is to sequentially search for structures, creating a series of sub-structures in the process which are subsequently merged or split. It follows [2] who used pattern-oriented modelling framework to design, test and analyse bottom-up models.

### 3.1 Data Description

Two time-series datasets with clear periodic patterns shown in Fig. 2 are used. The first is a set of 3160 average monthly sunspot records for the period from mid-18$^{th}$ century to early 2012 (NOOA, 2012) of which the 5$^{th}$ and 6$^{th}$ had the lowest activity and 19 and 21 had the highest. The second set consists of 65436 data seismic signals (861x76) obtained from the Department of Geophysics at the University of Leeds.

The left hand side panel in Fig.1 shows the individual densities of four cycles - two shortest and two longest - with a superimposed density of all cycles between them. The right hand side panel exhibits four sets of signal readings 1, 25, 50 and 76 with a superimposed overall density estimate curve for all signals. Identifying the essence of sub-structures in both cases is crucial to understanding numerous phenomena. For instance, sunspots numbers are known to be strongly correlated with modern measures of solar activity which we know that can interfere with power grids and communication satellites [10] and [11].

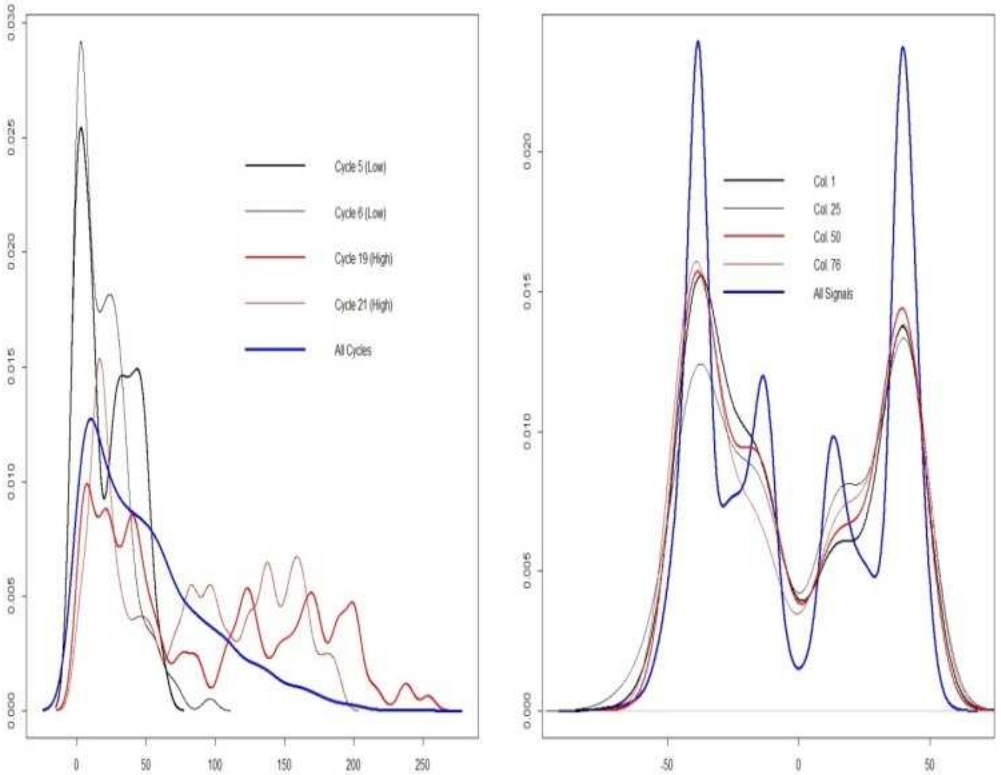

**Fig. 1:** Densities for selected sunspot cycles (LHS) and selected seismic columns (RHS)

Further, recent studies have closely associated sunspots with space weather [12], [14], and [20]. Similarly, segmentation of the earth is useful in many ways. Homogeneity/heterogeneity of sub-regions within the same geological structure may guide searches for natural resources - oil, minerals or water. The framework and mechanics of our proposed Data-Split-Merge (DSM) algorithm are described below.

### 3.2 The DSM Algorithm

The algorithm reads data as one complex system, determines its overall behaviour before carrying out parameter estimation. Its general mechanics can be summarised as follows.

```
$\mathbf{X} \leftarrow \mathbf{D}$ : Read Data into a Processable Medium
$\boldsymbol{EDA}(X)$ : Initial Exploratory Analysis (Explore Distributional Behaviour)
$\boldsymbol{F(X)}$ : Initial Density Estimation
$\phi$ : Initialise Density Variations
$SS_{cum}$ : Initialise a Substructure Cumulative Variable
$\theta_{\mathbf{cum}}$ : Initialise a Parameter Cumulating Variable
For   $i = 1$   to Length   $(\mathbf{X})$   Do
    For   $j = 1$   to Length   $(\mathbf{x_n} \in \mathbf{X})$   Do
        $\theta_{\mathbf{j}} = \{\theta_1, \theta_2, \ldots, \theta_{\gamma-1}, \theta_\gamma$ : Initial Parameters
        $SS_j$ : Determine Initial Substructure
$\theta_{\mathbf{n}} = \{\theta_{n_1}, \theta_{n_2}, \ldots, \theta_{n_s}\}$ : Substructure Parameters
          $F_j(x_n)$ : Density Estimation
            While   $\mathbf{i} \leq \mathbf{j}$   Do
               $SS_{cum} \leftarrow SS^*_{ij} = \sum_i \sum_j SS_j$
               $\theta_{\mathbf{cum}} \leftarrow \theta^*_{ij} = \{\theta_1, \theta_2, \ldots, \theta_{\vartheta \leq \gamma}\}$
               $|\phi| = F(X) - F_{SS_{cum}}(x_n)$ : Density Variation
               $\boldsymbol{E}(\theta^* || \theta|)$ : Conditional Parameter Update
               $\mathbf{CV(}||\phi|\mathbf{)}$ : Assess Estimation Quality via Cross Validation
               $\phi \leftarrow |\phi|$
            End While
          End For
        $\boldsymbol{E}(\theta^* | \phi)$ : Conditional Parameter Update
        $\boldsymbol{CV}(\phi)$ : Assess Estimation Quality via Cross Validation
End For
```

The algorithm's novelty derives from its dependency on data behaviour and modularity. The parameters $\theta$ can be adapted to specific datasets. For instance, $f(x) = x^T \beta, \theta = \beta$ while for the density in Equation (3), $\{\theta = \pi_k, \mu_k, \sigma_k\}$. Thus, for a general Gaussian model,

$$p(x_n, \theta_n) = \sum_{j=1}^{\gamma} \pi_j p(x_{nj} | \phi_{nj}) \Leftrightarrow \sum_{j=1}^{\gamma} \frac{\pi_j}{(2\pi)^{\frac{d}{2}}} exp - \frac{1}{2}(x_n - \mu_{nj})^T \sum_{nj}^{-1}(x_n - \mu_{nj}) \tag{3}$$

$\pi_j s$ are the mixing parameters; $p(x_{nj}|\theta_{nj})$ is the pdf corresponding to the distribution $F_j(x_n)$ and $\theta_n$ denotes the vector of all unknown parameters associated with the parametric forms adopted for these $j$ component densities. In the case of multivariate Gaussian components, theta consists of the elements of the mean vectors $\mu_{nj}$ and the covariance matrices $\sum_{nj}$. The conditional parameter update is a function of the adopted measure of fit quality such as cross validation or one in the form of Equation (2). Since $x_n$ are effectively random samples, maximisation of the likelihood of resulting density is generally very awful. Thus, we can treat group membership as missing data and use an adapted version of the EM algorithm, say, to estimate the MLEs for the vector of unknowns, $\pi$ and $\theta$ [15]. Continuous data can be clustered around medoids - a group of data objects having minimal average within-group dissimilarity [16]. The random samples $x_n$ can be selected using any robust sampling method or a sequential selection method as in [17] to form initial "medoids". Distances to all other points can then be computed on the basis of which data points are allocated to clusters. Finally, iteratively, the clusters can be optimised with the minimal average within-group dissimilarity being measured by the silhouette width [18] as follows

$$\triangle_i = \frac{\delta_i^{out} - \delta_i^{in}}{max(\delta_i^{out}, \delta_i^{in})} \Leftrightarrow \tag{4}$$

$$\triangle_i = \begin{cases} 1-\frac{\delta_i^{in}}{\delta_i^{out}} & \text{if} \quad \delta_i^{in} < \delta_i^{out} \\ 0 & \text{if} \quad \delta_i^{in} = \delta_i^{out} \\ \frac{\delta_i^{out}}{\delta_i^{in}} - 1 & \text{if} \quad \delta_i^{in} > \delta_i^{out} \end{cases} \quad \Leftrightarrow \quad -1 \le \triangle_i \le 1$$

where $\delta_i^{in}$ is the average dissimilarity of the $i^{th}$ observation with the other data points within the cluster, $\delta_i^{out}$ is the minimal average dissimilarity of the $i^{th}$ observation to any other cluster not containing it- i.e., $i's$ next best fit cluster.

## 4 DATA ANALYSES, RESULTS AND DISCUSSIONS

Analyses proceed in accordance with the algorithm above which selects a sample $x_n \in X$ of size $m$ - a partial or full cycle or signal, each time incrementing it by one unit or more. Initial EDA results for selected sample sizes for the two datasets are given in Fig.2. The top two panels highlight the presence of natural groupings within cycles and signals. The bottom two panels are results from a standard cluster-searching algorithm on the same data based on the assumption that 76 different signals and 23 cycles (omitting the incomplete $24^{th}$ cycle) constitute clusters. The algorithm therefore searched for clusters above and below these values. Here, the seismic signals exhibit a very high within group variation for less than 30 clusters which decreases with an increasing number of clusters. The pattern is repeated for the sunspots except that now extremely few clusters exhibit extreme rates of internal variation. It is these variations that we need to monitor and control.

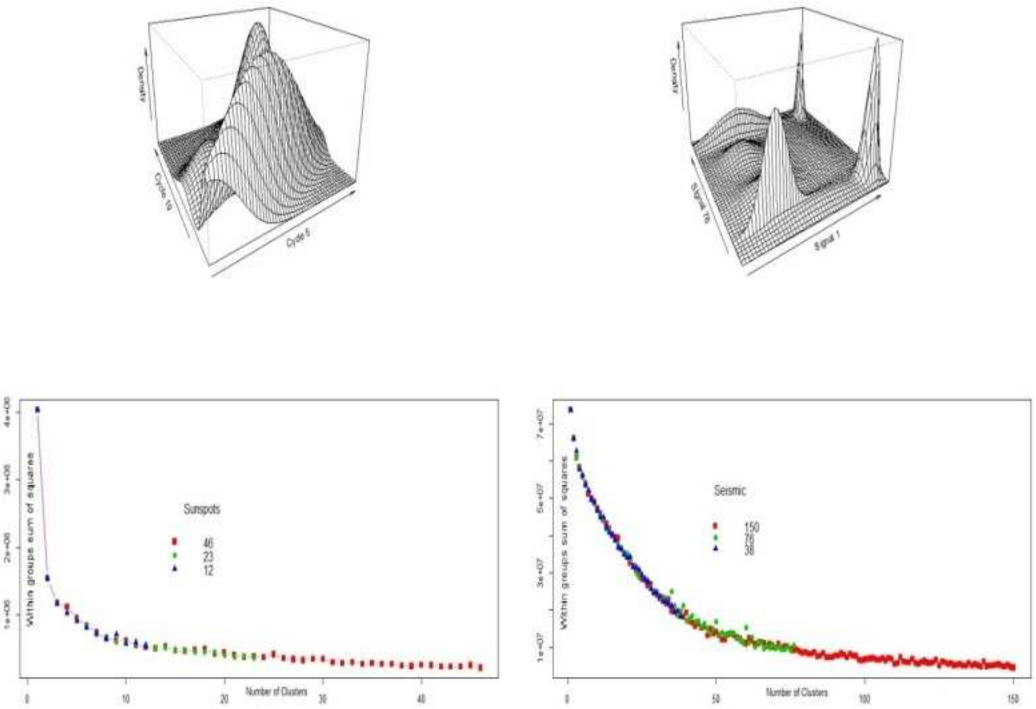

**Fig. 2: Visual natural structures in sunspots and seismic data (top panels) and results from a standard cluster-searching algorithm (bottom panels)**

Without loss of generality, we apply our proposed algorithm on continuous data and on the assumption that clustering is around medoids and that the number of clusters is estimated on the basis of optimum average silhouette width which as described in Equation (4). The plots in Fig.3 are generated via different data models and in each case the most likely model and number of clusters are determined by the maximum likelihood estimation and some Bayesian criteria.

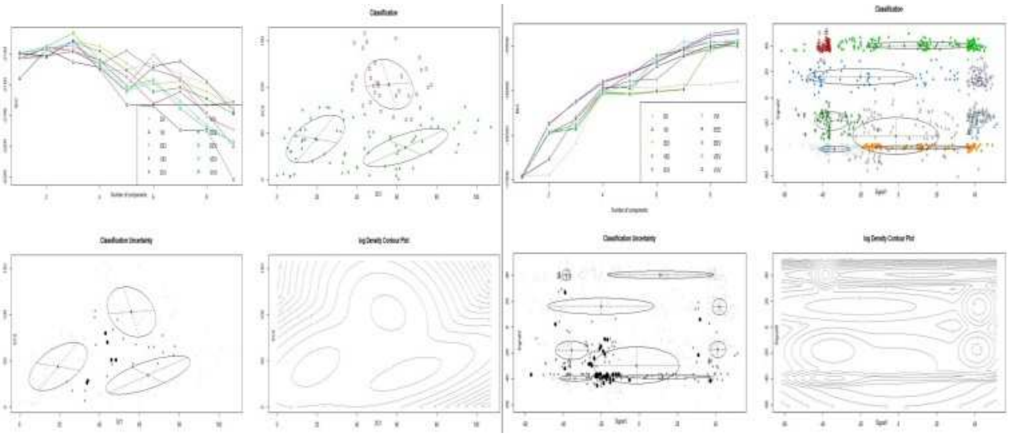

**Fig. 3: Sunspots and seismic signal analyses (LHS and RHS respectively) based on an arbitrary choice of data**

For the sunspots data, the optimal model according to the Bayesian Information Criterion (BIC) was a Gaussian ellipsoidal, equal volume and shape (EEV) with 3 components. For the seismic signal, it was a diagonal, varying volume and shape (VVI) with 9 components. Our adopted rule here follows [19] and chooses the model and number of clusters with the largest BIC. The choice of optimal clusters in both Fig.2 and Fig.3 is hugely affected by the starting point. The four panels in Fig.4 exhibit multiple simulations of criterion value densities for numbers of clusters in the first to the $23^{rd}$ sunspots cycles. Approximation of the average silhouette width is done by breaking the dataset into subsets 5 and taking averages. The numbers of clusters to be compared by the average silhouette width criterion are shown in the legend. The criterion varies inversely with the compared number of clusters. Here, the p-value for against the null hypothesis of similarity between clusters is extremely low. Also as the number of averaged subsets decreases, the criterion plot becomes spikier.

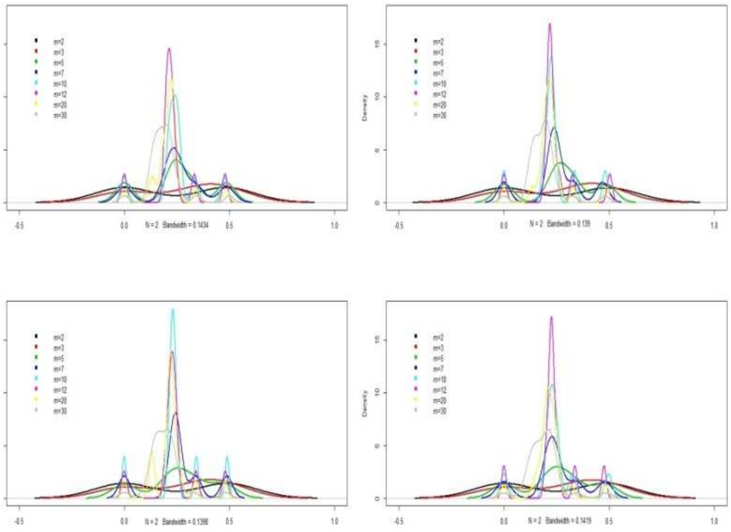

**Fig. 4: Multiple simulations of criterion value densities for numbers of clusters in sunspots data**

For the seismic data, averaging over 8 subsets yielded the best results. The four panels in Fig.5 exhibit a more consistent behaviour with bimodality being sustainable over all comparisons. Like in Fig.5 the criterion varies inversely with the compared number of clusters. Unlike in the sunspots example, the number of averaged subsets does not affect stability of inherent structures. Averaging over 2 to 20 subsets yielded similar results. In both cases, the density variation $|\phi| = F(X) - F_{SS_{cum}}(x_n)$ in the algorithm above can be measured using any appropriate criterion such as $C_{uK}$ in Equation

(1) or $\triangle_i$ in Equation (4). Spiky plots imply swamping as spurious clusters emerge warning against data over-fitting while unimodality may suggest masking or under-fitting. By cross-validating multiple runs of the algorithm the decision to merge or split clusters can be made.

## 5 CONCLUDING REMARKS

The complex nature of global challenges such as climate change, terrorism and food security can no longer be tackled in isolation and as Big Data becomes an increasingly household concept, we are all called upon to engage into interdisciplinary research initiatives. This paper focused on the knowledge extraction from data. Based on a general purpose data clustering algorithm, we were able to use silhouette plots and averaging over data subsets to determine the natural number of clusters within the two datasets. The paper's finds are readily extendable to classification and/or regression.

The algorithm's mechanics derive from its dependency on data behaviour as demonstrated by the different behaviour of the two datasets (Fig.4 and Fig.5). We sought to identify distinctive data sub-structures, verify model robustness via data reconstruction using a combination of $\phi$, the conditional parameter update function and cross-validation to determine sub-structure. Implementations on sunspots and seismic data revealed more stable structures than those obtained by conventional methods like PCA or $k$-Means. For example, neither principal component analysis nor data clustering tells us how to relate the sunspots variables or data to a particular characteristic and forming a new variable for future analysis. While plotting two components in a 2-D space may reveal a number of clusters which may guide future hypotheses, specifying criteria for robustly defining potential clusters in this case is a major challenge and it is what this paper sought to address. Apparently, more tests are necessary to verify the algorithm's robustness. It is expected that the proposed methods will contribute towards unifying algorithmic theories on adaptive behaviour and model complexity.

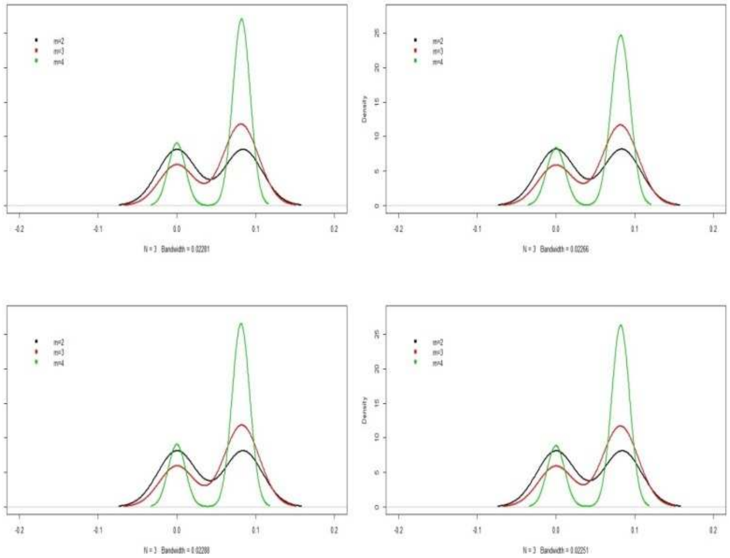

**Fig. 5: Multiple simulations of criterion value densities for numbers of clusters in seismic data**

## Acknowledgement

We would like to acknowledge the support from the project 2013/28 of Deanship of Scientific Research, University of Bahrain, Kingdom of Bahrain.

## References

[1] C. Cattani, A. Ciancio, Hybrid Two Scales Mathematical Tools for Active Particles Modelling Complex Systems with Learning Hiding Dynamics, Mathematical Models and Methods in Applied Sciences, **17 (2)**, 171-187 (2007).

[2] V. Grimm, E. Revilla, U. Berger, F. Jeltsch, W. Mooij, S. Railsback, H-H Thulke, J. Weiner, T. Wiegand, and D. DeAngelis, Pattern-Oriented Modelling of Agent-Based Complex Systems: Lessons from Ecology; Science, **310 (5750)**, 987-991 (2005).

[3] G. J. McLachlan, and D. Peel, Finite Mixture Models. Wiley, New York, 2000.

[4] S. Frühwirth-Schnatter, Finite Mixture and Markov Switching Models, Springer, Heidelberg, 2006.

[5] J. Corter, and M. Gluck, Explaining Basic Categories: Feature Predictability and Information, Psychological Bulletin, **111 (2)**, 291303 (1992).

[6] I. Witten, E. Frank, and M. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Elsevier, 2011.

[7] J. B. MacQueen, Some Methods for classification and Analysis of Multivariate Observations, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 281297 (1967).

[8] A. Dempster, N. Laird, and D. Rubin,Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society, **39 (1)**138 (1977).

[9] J. Bugrien, K. Mwitondi, and F. Shuweihdi, A Kernel Density Smoothing Method for Determining an Optimal Number of Clusters in Continuous Data, Risk Analysis IX, WIT Transactions on Information and Communication Technologies, **47**,WlT Press, 2013.

[10] D. Hathaway, and R. Wilson, What the Sunspot Record Tells Us About Space Climate, Solar Physics, **224 (1-2)**, 5-19 (2004).

[11] E. Khorsheed, M. Hurn, and C. Jennison, Mapping electron density in the ionosphere: a principal component MCMC algorithm; Computational Statistics & Data Analysis, **55 (1)**, 338352 (2011).

[12] A. Hanslmeier, The Sun and Space Weather, Heliophysical Processes Astrophysics and Space Science Proceedings, 233-249 (2010).

[13] W. Weber, Strong Signature of the Active Sun in 100 Years of Terrestrial Insolation Data, Annalen der Physik (Berlin), **522 (6)**, 372-381 (2010).

[14] G. Feulner, and S. Rahmstorf, On the Effect of a New Grand Minimum of Solar Activity on the Future Climate on Earth; Geophysical Research Letters, **37 (5)**, Article ID: L05707 (2010).

[15] K. S. Mwitondi, and R. A. Said, A step-wise method for labelling continuous data with a focus on striking a balance between predictive accuracy and model reliability, international Conference on the Challenges in Statistics and Operations Research (CSOR), 08th -10th March - 2011, Kuwait City.

[16] M. Van Der Lann, K. Pollard, and J. Bryan, A New Partitioning Around Medoids Algorithm; Journal of Statistical Computation and Simulation, **73 (8)**, 575584 (2003).

[17] A. Atkinson, Very Fast Robust Methods for Detection of Multiple Outliers, Journal of American Statistical Association, **89**, 1329-1339 (1994).

[18] P. Rousseeuw, Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis, Computational and Applied Mathematics, **20**, 5365 (1987).

[19] C. Fraley, and A. Raftery, Bayesian regularization for normal mixture estimation and model-based clustering, Journal of Classification, **24**, 155-181 (2007).

[20] W. Weber, Strong Signature of the Active Sun in 100 Years of Terrestrial Insolation Data, Annalen der Physik (Berlin), **522(6)**, 372-381 (2010).

**Kassim Mwitondi** is a Senior Lecturer in Applied Statistics/Data Mining, and has been a full-time member of staff of the Faculty of Arts, Computing, Engineering and Sciences at Sheffield Hallam University since January 2004. His research interests are in developing enhanced data mining methods for detecting patterns in space-terrestrial phenomena and uncovering their potential impact on human livelihood. He works in interdisciplinary research consortia with a general focus on knowledge discovery from multi-faceted data (KDMD) for tackling global challenges and sustainability. Mwitondi has active and long established research collaborations with researchers from various institutions across the world and he is on editorial boards of several international journals and data repositories. Mwitondi has published extensively in peer-reviewed journals and presented papers at numerous international conferences on topics related to KDMD applications for global sustainability.

**Eman Khorsheed** PhD, is an Assistant Professor at University of Bahrain Department of Mathematics. In 2011 she became a Fellow of the Higher Education Academy (HEA), UK. Her primary interests include Bayesian Statistics including spatial, spatio-temporal and hierarchical models, Markov chain Monte Carlo (MCMC) techniques, image analysis, Tomography, Demography, Biostatistics, Applied Statistics, Data Analysis, and Operations Research. Dr. Eman holds an MSc in Statistics & Operations Research from University of Bahrain, a Postgraduate Certificate in Academic Practice (PCAP) from York St. John University and a PhD from University of Bath, Uk. In 2013, Khorsheed founded a new international Journal under the umbrella of University of Bahrain: Journal of Data Analysis and Operations Research (JDAOR). She is also the managing editor of JDAOR. Khorsheed has several scientific papers, reports and projects presented at either local forums or International conferences or published in scientific journals. She is also a member of both the International Society for Bayesian Analysis and the Tunisian Decision Aid Society. She has served as an invited member of several International conference committees. In May 2016, she has been appointed Ambassador of the Asian Council of Science Editors (ACSE).