On the Use of Information Theoretic Mean Shift for Electricity Load Patterns Clustering

Jean Sumaili, *Member, IEEE*, Hrvoje Keko, *Member, IEEE*,
Vladimiro Miranda, *Fellow, IEEE* and Gianfranco Chicco, *Senior Member, IEEE*

***Abstract* — This paper analyzes the application of the Information Theoretic (IT) Mean Shift algorithm for modes finding in order to provide the classification of Electricity Customer Load Patterns. The impact of the algorithm parameters is discussed and then clustering indices are used in order to make a comparison with the classical methods available. Results show a good capability of the modes found in capturing the data structure, aggregating similar load patterns and identifying the uncommon patterns (outliers).**

***Index Terms* — clustering, information theoretic learning, load patterns, mean shift, modes finding, outliers.**

I. Introduction

THE importance of correct consumer classification and identification of shapes of their consumption curves has increased with the emergence of time-dependent market prices. In the characterization of the power curves, clustering techniques have been successfully used, aiming to identify the consumer groups. Subsequently, dedicated tariff structures can be associated to the consumer groups identified by clustering, according to the regulatory rules and characteristics of a particular market.

Load pattern clustering is a challenging task due to relatively high dimensionality. In most cases in Europe, each daily consumption pattern is measured in 15-minute intervals which results in 96 daily values. Moreover, in general there is no a priori information on the possible structure of the clustered data or on the final number of clusters. If such descriptive information exists, it is often burdened with low information content and high content of erroneous values, so only the most basic information is actually usable [1].

Typically, clustering techniques may be divided in two groups: the methods that require the desired number of clusters as a parameter, and the methods that try to determine an optimal number of clusters given a certain set of clustering parameters. In addition, clustering methods may be divided into partitional and hierarchical algorithms: while the partitional algorithms obtain a single partition of the data, hierarchical clustering algorithms produce a multi-level hierarchy. In both cases, clustering results depend on the similarity measure adopted.

Clustering adequacy depends both on the characteristics of the dataset and on the capabilities of the algorithm to capture regularities in the data structure and isolate uncommon patterns (outliers). Several methods rely on the notion of Euclidean distance.

A typical example of a common partitioning algorithm that uses a sum of squared errors between patterns and cluster centroids is the K-means algorithm [2].

Besides the Euclidean distance-based metrics, similarity measures may be based on the concepts of entropy introduced in information theory [3], [4]. Such similarity measures have been successfully applied to identify natural structures, with particularly evident results in low dimension clustering problems such as image segmentation [5], [6].

For a higher dimension load pattern clustering problem, Euclidean distance is still a commonly adopted metric. In [7], the authors discuss the use of Renyi entropy for load pattern classification offering a good alternative approach aimed at avoiding attraction among extremes.

Mean shift is a non-parametric mode finding method using a kernel-based construction of a search space surface, presented in [8] in 1975. This technique was not applied widely until 1995 when a slightly updated version was presented in [9]. In most of cases Gaussian kernels are used to construct the surface so the algorithms have been known as Gaussian Blurring Mean Shift (GBMS) algorithm and Gaussian Mean Shift (GMS), respectively.

The mean shift algorithm has been successfully applied in various areas, due to its simplicity and ability to deal with clusters of arbitrary shape. Moreover, usually the only parameters the mean shift needs are the ones related to the underlying kernels. A generalization of mean shift method is presented in [10] and [11], the authors start from a generalized information theoretic point of view, and develop a broader class of clustering algorithms. The presented class of algorithms introduces a new cost function that tries to "tune" to principal curves and modes of the data and representing a constrained entropy minimization problem. The information theoretic mean shift has an additional parameter governing the algorithm behavior, and the two algorithms presented earlier (GBMS and GMS) are special cases of the newly presented algorithm class. Based on these observations, in [11] the authors discuss superior performance of GMS when compared to GBMS, on a variety of lower dimension clustering cases.

In this paper, an information theoretic approach to mean shift algorithm used to perform clustering of higher dimension samples – realistic 24-hour power curves, and the impact of the clustering parameters is discussed.

In order to compare the quality of clustering applied to

J. Sumaili, H. Keko and V. Miranda are with the Power System Unit of INESC Porto – Instituto de Engenharia de Sistemas e Computadores do Porto, Campus da FEUP, Rua Dr. Roberto Frias 378, 4100 – 465 Porto, Portugal Phone: +351-222-094-224, Fax: +351-222-094-050
(e-mail: jean.sumaili@inescporto.pt, hrvoje.keko@inescporto.pt, vmiranda@inescporto.pt)

G. Chicco is with the Dipartimento di Ingegneria Elettrica, Politecnico di Torino, Corso Duca degli Abruzzi 24, I-10129 Torino, Italy (e-mail: gianfranco.chicco@polito.it).

This research work has been in part founded by FCT – Fundação para a Ciência e a Tecnologia) within the program "Ciência 2008" (J. Sumaili), and by the PhD FCT scholarship SFRH/BD/43087/2008 (H. Keko).

customer characterization, the authors of [12] developed a set of clustering quality indices. These performance indices have been developed in order to have a suitable means of performance comparison of various clustering methods, when applied to a particular problem of customer characterization. In this paper, the performance of information theoretic learning mean shift is compared to the classical K-means clustering, using the performance indicators described in detail in [12].

II. Information Theoretic Learning and Mean Shift Algorithm

Considering a data set, consisting of independent and identically distributed (i.i.d.) samples $X = (x_i)_{i=1}^N \in \mathbb{R}^d$, using the Parzen window estimation technique, its probability density can be estimated based on a Gaussian kernel with a bandwidth $\sigma > 0$

$$p(x) = \frac{1}{N}\sum_{i=1}^{N} G_\sigma(\|x - x_i\|^2) \tag{1}$$

where $G_\sigma(t) = e^{-\frac{t}{2\sigma^2}}$

In order to find the modes of such estimated probability density function (pdf), the stationary point equation $\nabla p(x) = 0$ can be arranged into an iterative fixed point scheme as follows:

$$x^{(\tau+1)} = m\left(x^{(\tau)}\right) = \frac{\sum_{i=1}^N G_\sigma(\|x - x_i\|^2)\, x_i}{\sum_{i=1}^N G_\sigma(\|x - x_i\|^2)} \tag{2}$$

In [8], the initial presentation of this technique, the term m(x) – sample mean of all samples weighted by kernel centered at x – was named mean shift. Since an initial dataset is successively blurred using the above equation, in the later discussion the method was named Gaussian Blurring Mean Shift (GBMS). A modification named Gaussian Mean Shift (GMS) algorithm was proposed in [9]. In this version, the dataset is initialized to $X_o = X^{(\tau=0)}$, which is subsequently kept constant. This way a dataset $X^{(\tau+1)}$ is produced by comparing the present dataset $X^{(\tau)}$with the initial data set X_o, which made the algorithm more stable and it was successfully applied to a number of tasks.

In [11], an information theoretic learning point of view of both GMS and GBMS is presented. Given a random variable with $X = (x_i)_{i=1}^N \in \mathbb{R}^d$ with iid samples, the non-parametric density estimator using Parzen window [13] technique is:

$$p_{x,\Sigma}(x) = \frac{1}{N}\sum_{i=1}^{N} K_\Sigma(x - x_i) \tag{3}$$

where K_Σ is a kernel with covariance matrix Σ. In most cases only spherical covariance matrix in the form of $\Sigma = \sigma^2 I$ is used.

In [4], Renyi defines quadratic entropy of a dataset X as $H(X) = -\log\left(\int p^2(x)dx\right)$. Substituting the probability density function with its Parzen estimate based on Gaussian kernel and spherical covariance, one easily obtains a non-parametric entropy estimator

$$\begin{aligned} H(X) &= -log\big(V(X)\big) \\ &= -log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_\sigma\left(\|x_i - x_j\|^2\right)\right) \end{aligned} \tag{4}$$

The logarithm argument is a potential field over the space of the samples. Its derivative can be considered as an information force

$$F(x_i) = \frac{\partial}{\partial x_i}V(x_i) = \sum_{j=1}^{N} F\left(x_i|x_j\right) \tag{5}$$

The contributions of each sample can be easily seen from the formulation, and the notion of interaction between samples of the same dataset can be extended to quantifying interactions between two different datasets, employing the notion of cross-entropy

$$\begin{aligned} H(X;Y) &= -\,log\big(\textstyle\int p_X(t)p_Y(t)dt\big) \\ &= -log\,(\frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M} G_\sigma\left(\|x_i - y_j\|^2\right) \end{aligned} \tag{6}$$

These are the basic ideas of information-theoretic learning, and a more detailed summary may be found in [11].

The main motivation of using ITL measures is that by estimating the entropy non-parametrically, ITL manages to go above the second order statistics.

In [10] and [11], a connection between mean shift algorithms and Renyi entropy is developed. Considering an original dataset $X_o = (x_{oi})_{i=1}^{N_o} \in \mathbb{R}^d$ with i.i.d. samples and keeping this dataset constant, one may define a cost function $J(X) = \min_X H(X)$ that minimizes the Renyi entropy with regards to the dataset.

The dataset changes iteratively hence it is the parameter of cost function minimization. The idea of the minimization is to evolve the dataset, such that net force acting on each sample is equal to zero. In clustering, one aims to find a dataset $X = (x_i)_{i=1}^M \in \mathbb{R}^d$, consisting of $M \leq N$ samples, that captures the structure of the initial dataset.

Since it is $1 \leq M \leq N$, there are two extremes when representing the structure. If the whole structure is represented with a single sample ($M = 1$), the entropy of the dataset X is equal to zero. On the other hand, if the initial dataset is represented as itself (i.e. $M = N$), the entropy of X is obviously equal to initial dataset entropy $H(X_0)$ [10].

Having this in mind, the Renyi entropy minimization problem can be formulated as a constrained problem. The additional constraint is the Cauchy-Schwarz distance between the initial dataset X_o and the current dataset X. The constraint is to keep the Cauchy-Schwarz distance constant, at a value k between zero and initial entropy $H(X_0)$, $0 \leq k \leq H(X_0)$

This constrained optimization problem can be reformulated as an unconstrained problem, employing the Lagrangian multipliers. This gives the following cost function update rule:

$$J(X) = \min_X [-(1-\lambda)\log\big(V(X)\big) - 2\lambda\log\big(V(X;X_0)\big) + \lambda\log\big(V(X_0)\big) - \lambda k] \quad (7)$$

This rule can be differentiated with respect to each sample which then delivers a update rule parameterized only with λ, i.e. the constant parameter k disappears from the formulation.

$$\frac{\partial}{\partial x_i} J(X) = -\frac{1-\lambda}{V(X)} F(x_i) - \frac{2\lambda}{V(X;X_o)} F(x_i;X_0) \quad (8)$$

Including the Gaussian kernel formulation the update rule for the components of X easily follows:

$$x_i^{\tau+1} = \frac{c_1 \sum_{j=1}^{N} G\left(\left\|\frac{x_i^\tau - x_j^\tau}{\sigma'}\right\|^2\right) x_j^\tau + c_2 \sum_{j=1}^{N} G\left(\left\|\frac{x_i^\tau - x_{0j}}{\sigma'}\right\|^2\right) x_{0j}}{c_1 \sum_{j=1}^{N} G\left(\left\|\frac{x_i^\tau - x_j^\tau}{\sigma'}\right\|^2\right) + c_2 \sum_{j=1}^{N} G\left(\left\|\frac{x_i^\tau - x_{0j}}{\sigma'}\right\|^2\right)} \quad (9)$$

This is the foundation of information theoretic mean shift clustering. There are two special cases of the above. Firstly, for $\lambda = 0$ one directly obtains Gaussian blurring mean shift (GBMS) clustering: reformulation of the equations directly exposes that this is exactly equal to GBMS algorithm. This means the GBMS algorithm actually minimizes the Renyi entropy. For $\lambda = 1$, one obtains Gaussian mean shift (GMS). Further details and proofs may be found in [10] and [11].

In this paper the performance of the information theoretic mean shift applied to clustering of power curves is discussed. The influence of the parameter λ of the ITL-MS clustering, as well as the influence of the parameter σ of the underlying Gaussian kernel estimator are addressed.

III. Clustering Algorithm

The following analyses have been performed: given the parameters σ and λ, the ITL-MS algorithm has been run for a maximum number of iterations equal to 1000. The procedure has been stopped if for more than 200 iterations the maximum variation in X has been constantly lower than 10^{-3}. This additional stopping criterion avoids the instability that can result when λ is set to 0 as mentioned in [10].

The obtained different load curves are the modes of the probability density function of the initial target set. Each mode is used to represent a data set cluster. For cluster composition, each customer load curve is allocated to the cluster where the representative pattern (i.e. mode) has the lowest distance. As the data structure has been captured the found modes, the metric used to calculate distance is the classic Root Mean Squared Error (RMSE) in order to reduce the difference between each load curve and its representative mode.

IV. Case Study

The target set of data refers to 193 non-residential consumers connected to the medium voltage distribution network (Fig. 1). The daily load pattern data is converted from 15-minute data to hourly data, so that it contains the average power evaluated at one hour time steps ($D = 24$).

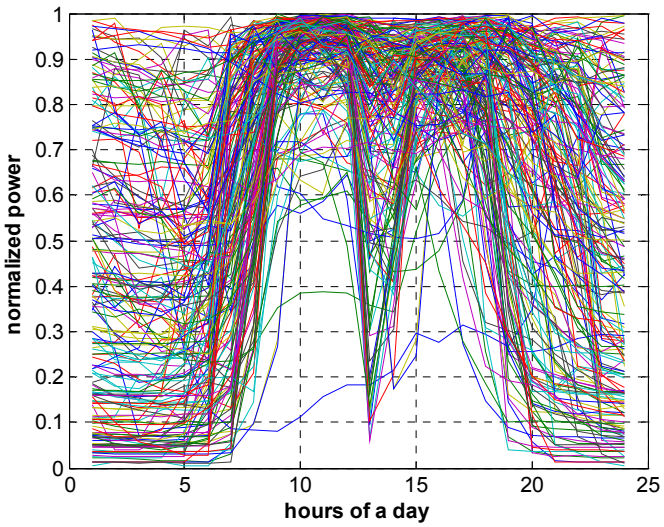

Fig. 1 Initial set of electricity load curves to be clustered

The clustering algorithm has been run for different values of σ and λ within the intervals [0.1, 0.25] and [0, 1], respectively. The results are presented below.

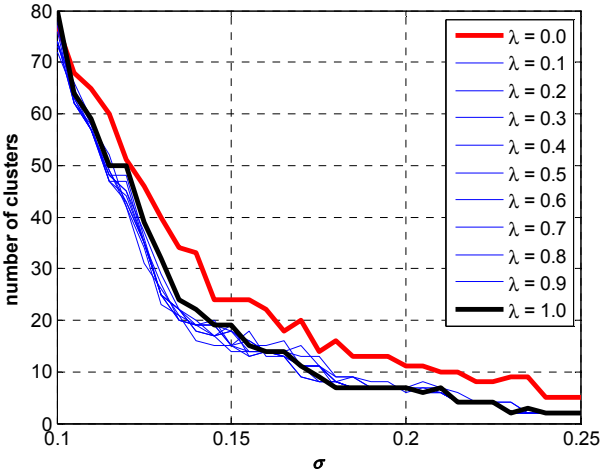

Fig. 2 Number of clusters as a function of the value of σ (the red line and the black line indicate $\lambda = 0$ and $\lambda = 1$)

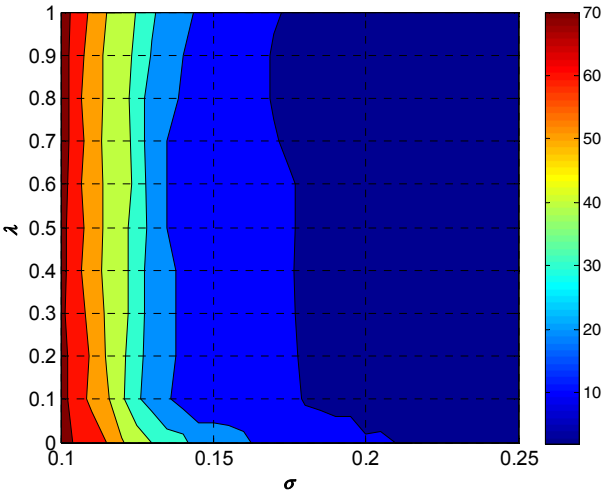

Fig. 3 Number of clusters in the plane (σ, λ)

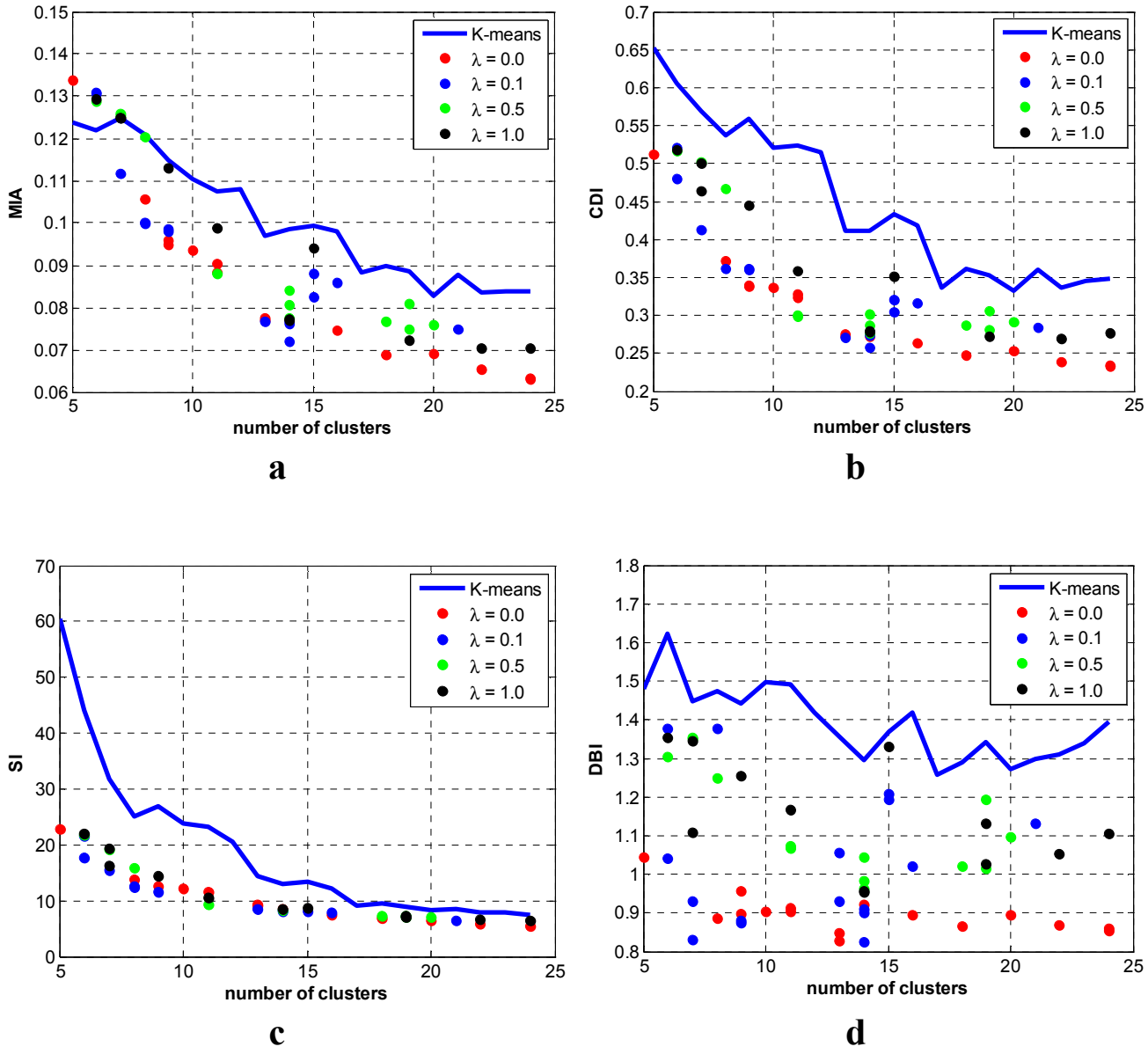

Fig. 4 Clustering validity indicators (a:Mean index adequacy – MIA, b: Clustering dispersion indicator – CDI, c: Scatter index – SI, d: Davies-Bouldin index – DBI)

Fig. 2 and Fig. 3 show how the number of cluster varies with respect to the variation of the parameters σ and λ. In fig. 2 each line corresponds to a different value of λ and the red line and the black line are related to $\lambda = 0$ and $\lambda = 1$, respectively. Notice that the number of clusters is more sensitive to the value of σ, but is quasi invariant to the value of lambda in the interval [0.1, 0.9]. For the same value of σ, it can be observed that $\lambda = 0$ commonly leads to higher number of clusters than other values of λ. For λ=0, the algorithm also stops earlier due to lack of improvement. Thus one may argue that for λ=0 algorithm may have stopped before the all the modes have been completely identified. On the other hand, even for low values of λ (e.g. λ=0.1), the algorithm behaves as when λ=1.0, maintaining higher speed of convergence of λ=0.

Fig.4 presents the comparison between K-means performance and the ITL-MS using some clustering validity indicators, namely the mean index adequacy (MIA), the clustering dispersion indicator (CDI), the Scatter index (SI) and the Davies-Bouldin index (DBI). For the clarity of the figure, only the results related to λ equal to 0, 0.1, 0.5 and 1 have been presented. Since lower index values represent better clustering performance, it can be observed that within the interval [5, 25] for the number of clusters, the ITL-MS clustering algorithm constantly presents better results than the K-means algorithm. Moreover, this is valid for all the presented indices of clustering validity.

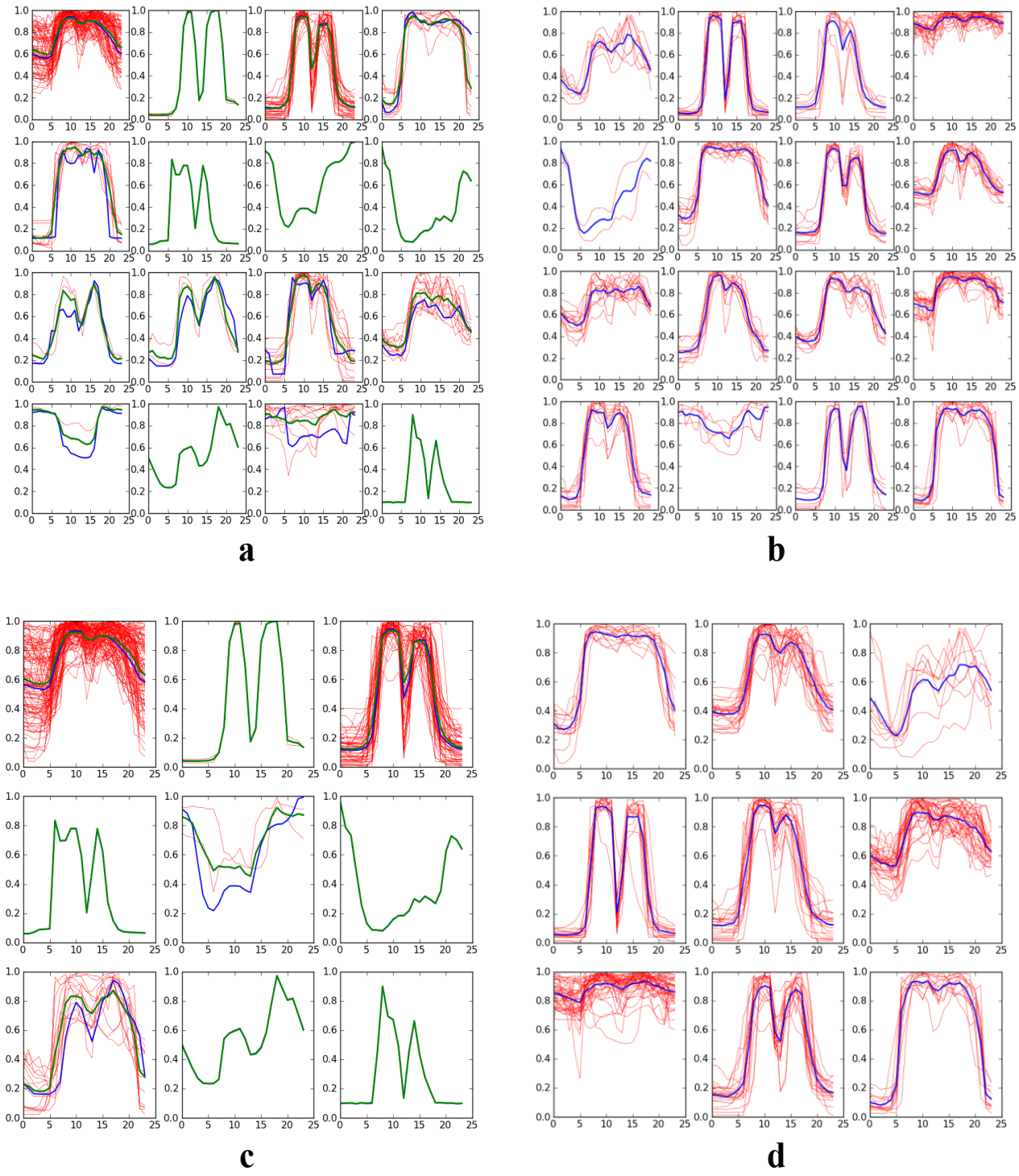

Fig. 5 Cluster composition (a: Mean Shift with $\lambda = 0.1$, $\sigma = 0.14$ leading to 16 clusters, b: K-means with 16 clusters, c: Mean Shift with $\lambda = 0.1$, $\sigma = 0.18$ leading to 9 clusters, d: K-means with 9 clusters). The red lines are the cluster members. For the K-means clustering, the blue line indicates the cluster centroid while for the ITL-MS, the blue line indicates the mode found and the green line, the centroid.

Fig. 5 shows the cluster composition for 16 clusters and 9 clusters making a comparison between ITL-MS algorithm with $\lambda = 0.1$ and the K-means algorithm. It clearly appears that the ITL-MS algorithm is better in identifying and isolating the outliers. The particular ability of the IT Mean Shift algorithm to identify the outliers can be of effective interest for distribution suppliers and aggregators, from different points of view [7].

The strong outlier isolation ability of ITL-MS algorithm might be used in order to filter the initial set of load patterns, before submitting them to classical clustering schemes. This way the ITL-MS algorithm may increase the performance of the clustering methods already being applied in practice, reducing the dispersion within the obtained clusters (and thus increasing the clusters' representativeness).

V. Conclusion

The Information Theoretic Mean Shift Algorithm has been applied for the clustering of electricity consumer load patterns. The results show good performance of the modes found in capturing the data structure and isolating the outliers.

The number of the clusters mainly depends on the value of the bandwidth σ: Larger bandwidth values result in lower number of found modes and consequently in lower number of clusters, however the uncommon patterns (outliers) are the latest to be attracted by different modes. Whereas the classical K-means algorithm in the presence of outliers tends to spread the total error among the clusters, ITL-MS algorithm seems to favorably concentrate undesired samples in a fewer number of clusters. Exploring this capability of ITL-MS algorithm may enable easier filtering of erroneous samples, and a filtered set of samples may then be used in "classic" clustering methods. Such a hybrid method would keep the advantageous characteristics of classic clustering methods (e.g. the number of clusters known in advance), while increasing the robustness of the method.

VI. References

[1] M. Matos, J. Fidalgo, and L. Ribeiro, "Deriving LV Load Diagrams for Market Purposes using Commercial Information," Proceedings of the 13th International Conference on Intelligent Systems Application to Power Systems, 2005, pp. 105-110, 2005.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," ACM Comput. Surv., vol. 31, no. 3, pp. 264-323, 1999.

[3] C.E. Shannon, "A mathematical theory of communication", Bell Syst. Tech. J., 1948, 27, pp. 370–423, 623–656

[4] A.Renyi, "On measures of entropy and information", Proc. Fourth Berkeley Symp. Mathematics, Statistics and Probability, 1960, pp. 547–561

[5] E. Gokcay and J. Principe, "Information theoretic clustering," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no. 2, pp. 158-171, 2002.

[6] R. Jenssen, K. Hild, D. Erdogmus, J. Principe, and T. Eltoft, "Clustering using Renyi's entropy," Proceedings of the International Joint Conference on Neural Networks, 2003., vol. 1, pp. 523-528 vol.1, 2003.

[7] G. Chicco and J. Sumaili Akilimali, "Renyi entropy-based classification of daily electrical load patterns," IET Generation, Transmission & Distribution, vol. 4, no. 6, pp. 736-745, 2010.

[8] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," IEEE Trans. on Information Theory, vol. 21, no. 1, pp. 32-40, 1975.

[9] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 17, no. 8, pp. 790-799, 1995

[10] S. Rao, Weifeng Liu, J. Principe, and A. de Medeiros Martins, "Information Theoretic Mean Shift Algorithm," Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, 2006, pp. 155-160, 2006.

[11] S. Rao, A. de Medeiros Martins, and J. C. Príncipe, "Mean shift: An information theoretic perspective," Pattern Recognition Letters, vol. 30, no. 3, pp. 222-230, Feb. 2009

[12] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer characterization options for improving the tariff offer," IEEE Trans. on Power Systems, vol. 18, no. 1, pp. 381-387, 2003.

[13] E. Parzen, "On Estimation of a Probability Density Function and Mode," The Annals of Mathematical Statistics, vol. 33, no. 3, pp. 1065-1076, 1962.