This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2016.2614366, IEEE Transactions on Power Systems

C-Vine Copula Mixture Model for Clustering of Residential Electrical Load Pattern Data

Mingyang Sun, *Member, IEEE*, Ioannis Konstantelos, *Member, IEEE*, Goran Strbac, *Member, IEEE*

***Abstract*—The ongoing deployment of residential smart meters in numerous jurisdictions has led to an influx of electricity consumption data. This information presents a valuable opportunity to suppliers for better understanding their customer base and designing more effective tariff structures. In the past, various clustering methods have been proposed for meaningful customer partitioning. This paper presents a novel finite mixture modeling framework based on C-vine copulas (CVMM) for carrying out consumer categorization. The superiority of the proposed framework lies in the great flexibility of pair copulas towards identifying multi-dimensional dependency structures present in load profiling data. CVMM is compared to other classical methods by using real demand measurements recorded across 2,613 households in a London smart-metering trial. The superior performance of the proposed approach is demonstrated by analyzing four validity indicators. In addition, a decision tree classification module for partitioning new consumers is developed and the improved predictive performance of CVMM compared to existing methods is highlighted. Further case studies are carried out based on different loading conditions and different sets of large numbers of households to demonstrate the advantages and to test the scalability of the proposed method.**

***Index Terms*—Clustering, customer classification, C-vine, decision trees, mixture models, pair-copula construction, smart meters.**

I. Introduction

Electricity market liberalization has largely unbundled the distribution and supply services in many jurisdictions, providing customers with the freedom to select their electricity supplier. In this competitive environment, retail companies can improve the commercial attractiveness of their product by formulating tariffs aimed at different customer types. An important part of the tariff design process is the identification of meaningful customer classes that exhibit different consumption patterns, enabling the development of diversifiable products. Moreover, electrical customer classification can also play a crucial role in load forecasting [1][2] and modeling [3], electricity market development [4], energy system planning and operation [5] and theft detection [6]. Naturally, information on customer type (e.g. industrial, commercial, residential) provides important information regarding the likely electricity consumption pattern and intensity. However, for further partitioning and exploratory analysis to be carried out effectively, high-frequency demand measurements are necessary [7]. As such, the advent of smart metering has led to large-scale availability of consumption data that render clustering analysis increasingly possible.

Load profile clustering aims to allocate consumers into a small number of homogeneous groups, ensuring that elements of the same cluster are similar between them, while being dissimilar to elements of different clusters. A large number of clustering techniques have been proposed in the past and applied to electrical load data. Examples of centroid-based approaches include k-means [8], [9], fuzzy k-means [10], k-medoids [11], and modified follow-the-leader [12]. Connectivity-based approaches have also been applied, focusing on hierarchical clustering methods with different linkages [10],[12]. Other studied techniques include ant colony clustering [13], self-organizing maps [14] and neural networks [10],[15]. Typically, the number of clusters is determined using clustering validity indices as discussed in [5] and [16].

An alternative approach, beyond centroid and connectivity models, which is starting to gain interest among practitioners, is the use of distribution mixture models. One main advantage of distribution-based clustering is the ability to systematically select the number of clusters while penalizing model complexity to construct parsimonious models [17]. Another advantage is the ability to obtain a probability measure regarding the classification of each consumer, as opposed to the 'hard' clustering that characterizes other methods. Another practical advantage relates to reduced memory requirements; by relying on parametric functions it is possible to effectively compress large amounts of data in a reduced-sized model.

In light of the above, a number of model-based clustering techniques have been used in the past to tackle customer clustering. Gaussian mixture model (GMM) is one of the most widely used model-based clustering approaches. It has been applied for time-of-use tariff design in [18] and clustering households in [19]. GMM has also been used in [20] to identify suitable data clusters for training Markov demand models. Other model types can also be deployed; for example, a multivariate Dirichlet process mixture model has been used to cluster electricity profiles in [21]. However, all such techniques suffer from the inherent limitations of standard multivariate functions; all mixture model components follow a pre-specified marginal distribution function and dependence structure. Given the highly non-linear dependence structures and non-standard marginal distributions observed in demand datasets, developing high-quality mixture models for clustering purposes is a substantial challenge.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2016.2614366, IEEE Transactions on Power Systems

In response to this, copulas can be utilized as a powerful tool to capture more complex dependency structures between variables. Gaussian Mixture Copula Models (GMCMs) were proposed in [22], where a number of multivariate Gaussian copulas are fitted to a range of data sets. In [23], a Gaussian copula mixture model was developed for dependency-seeking clustering tasks for both synthetic and real data found in biological systems. In [24], the copulas-based mixture model clustering algorithm was extended beyond the Gaussian paradigm to also accommodate other copula families such as Gumbel and Clayton. Although the existing copula-based mixture model clustering techniques have demonstrated good performance, accurately capturing the complicated dependency structures exhibited by electrical system variables, such as loads, cannot be adequately described by solely relying on multivariate copula building blocks; more flexible modelling structures are required. Vine copula models make use of pair copula construction (PCC) schemes to decompose a high-dimensional copula into a cascade of bivariate copula functions [25]. This substantially increases the flexibility of the model by being able to capture complex dependencies across a large number of variables. In addition, vine copula models have been shown not to suffer from the curse of dimensionality that characterizes other high-dimensional models [26]. Authors in [27] have shown that vine copulas are a class of density functions whose convergence rate does not depend on the number of dimensions. Given that a high-dimensional vine copula model is a cascade of bivariate functions, its convergence rate is equal to the rate of a two-dimensional estimator, thus evading the curse of dimensionality.

In this paper, a specific pair-copula construction scheme known as C-Vine is used to perform a novel type of distribution-based clustering. It is important to note that an indirect clustering approach is followed; a transformation to a lower-dimensional feature space is first applied instead of the model being fit directly to the consumption data set. Many variants of indirect clustering have been applied in the past. For example, authors in [28] perform clustering on household occupancy states which have been inferred using a Hidden Markov model. In a similar vein, authors in [29] apply Fast Search and Find of Density Peaks [30], a novel density-based clustering method, on occupancy state transition matrices. Locality- sensitive hashing is used in [31] to substantially speed up subsequent similarity comparisons for clustering. Indirect clustering is also combined with GMM in [32]. A discussion on the selection of possible features takes place in [33]. In our work, by combining vine copulas with dimension reduction, we are capable of addressing the drawback of increased computational burden while also harvesting the synergy between the C-Vine's hierarchical structure and the ordered variables, as discussed in [34].

This paper proposes a C-vine copulas based mixture model clustering (CVMM) algorithm for grouping the load pattern data. For CVMM, the parameters of the constructed mixture model are estimated via the expectation-maximization (EM) algorithm. The clustering quality of CVMM is evaluated and compared to other classical methods by using a set of selected clustering validation indicators based on real demand measurements. In addition, a decision tree based classification module for assigning new consumers to the existing classes is developed to further assess the results of CVMM clustering.

This paper is structured as follows. Section II recalls the concept of copulas and introduces pair-copula construction. Section III illustrates the procedure of load pattern data processing and defines the proposed C-vine Mixture Model Clustering (CVMM) algorithm with EM estimation. Section IV introduces the selected clustering validation indicators and their expressions. In Section V, based on real residential load pattern data from London, the performance of the proposed CVMM algorithm is assessed and compared with other widely-used load pattern clustering algorithms via comparing various relevant metrics. Furthermore, the clustering results are analyzed using demographic metadata. In addition, the accuracy of a decision-tree based customer classification model is used as an additional performance metric. The clustering performance of CVMM is is shown to be superior to exsiting methods when applied to the autumn dataset and combined weekday/weekend dataset. Finally, a scaling analysis for different sets of large numbers of households is presented. Section VI contains the concluding remarks.

II. Copulas and Pair-copula Construction

Copulas are a powerful tool for modeling data that exhibit a complex dependence structure. The basic concept of copulas is illustrated by Sklar's theorem [35]. Let f, F, c and C denote probability density function, cumulative distribution function, copula density function and copula cumulative distribution function, respectively. Consider m random variables $X = (X_1, \dots, X_m) \in \mathbb{R}^m$ with marginal cumulative distribution functions $F_i(x_i)$ and marginal density function $f_i(x_i)$, for $i = 1, .., m$. The joint density function can be expressed as:

$$f(x_1, \dots, x_m) = \left(\prod_{i=1}^{m} f_i(x_i) \right) \times c_{1\dots m}\big(F_1(x_1), \dots, F_m(x_m)\big) \quad (1)$$

where the function $c_{1\dots d}: [0,1]^m \to \mathbb{R}$ is an m-dimensional copula with uniform marginals $U = \{U_1, U_2, \dots, U_m\} = \{F_1(X_1), F_2(X_2), \dots, F_m(X_m)\}$. Equation (1) demonstrates that a joint density function can be represented by a product of its margins and a multivariate copula density function. Sklar's theorem states that if all marginal distribution functions are continuous then the copula coupling all variables is unique. Thus, an m-dimensional copula is a parametric function defined on the $[0,1]^m$ space, describing the dependency between m variables. Moving beyond the above definition, copula functions can be described by different copula families and corresponding parameters. For example, as shown in Fig.1 Clayton and Gumbel are two types of Archimedean copulas that have lower-tail dependence and upper-tail dependence with the parameters defined on the range $(0, \infty)$ and $(1, \infty)$ respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2016.2614366, IEEE Transactions on Power Systems

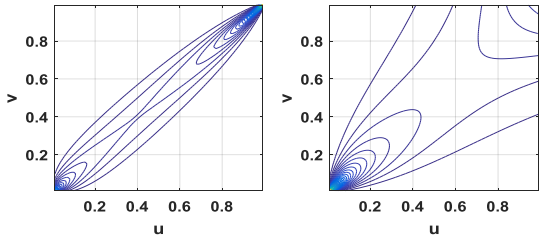

Fig. 1. Examples of a bivariate Gumbel copula with parameter $\vartheta = 5$ (left panel) and a bivariate Clayton copula with parameter $\rho = 2$ (right panel).

For the bivariate case, there are Gaussian copula, Student-t copula, Frank copula, Joe copula, BB1 copula, BB6 copula, BB7 copula, BB8 copulas, etc. [36]. Although such a rich variety is available, only very few copula families (e.g. Gaussian) can be extended to a high-dimensional version. Consequently, the performance of a single multivariate copula is limited when modeling a high-dimensional dataset with complex inter-dependencies. To handle this problem, the pair-copula construction method was proposed in [37] and developed in [25], [38] to decompose a high-dimensional copula function into a cascade of bivariate copulas. It is constructive to note that there is a vast number of possible pair-copula constructions for a high-dimensional distribution. To this end, a particular type of graphical model, Canonical vines (C-vines) was introduced in [39]. In general, the C-vine representation for the density function $f(x_1, \dots, x_m)$ can be expressed as follows:

$$f(x_1, \dots, x_m) = \left(\prod_{l=1}^{m} f_l(x_l)\right) \times \prod_{j=1}^{m-1} \prod_{i=1}^{m-j} c_{j,j+i|\omega_{i,j}}\left(F_{j|\omega_{i,j}}, F_{j+i|\omega_{i,j}}\right) \quad (2)$$

where $\omega_{i,j} = \{1, \dots, j-1\}$. In equation (2), the conditional distribution function can be denoted by *h*-functions [25] as follows:

$$h(U_i, U_j, \Theta) = F(U_i|U_j) = \frac{\partial C_{u_i,u_j}(U_i, U_j, \Theta)}{\partial U_j} \quad (3)$$

where Θ is the set of the parameters for the bivariate copula C fitted to two uniformly distributed variables U_i and U_j. For example, the general expression for the C-vine structure in the $m = 3$ case is given by:

$$\begin{aligned} f(x_1, x_2, x_3) = f(x_1) \cdot f(x_2) \cdot f(x_3) \cdot c_{12}(F(x_1), F(x_2)) \\ \cdot c_{13}(F(x_1), F(x_3)) \\ \cdot c_{23|1}(F(x_2|x_1), F(x_3|x_1)) \end{aligned} \quad (4)$$

Graphically, the C-vine construction can be represented as a sequence of dependency trees $T = [T_1, \dots, T_{m-1}]$. An example of a C-vine for $m = 3$ is shown in Fig.2.

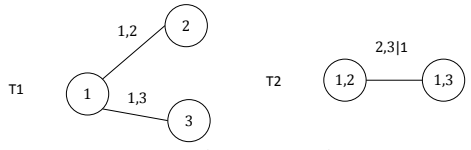

Fig. 2. Example of C-vine trees for $m = 3$.

As shown above, there is a total of $m(m-1)/2$ edges, where *m* is the number of variables being modeled. Each edge corresponds to a pair-copula density function fitted to the connected nodes. For each pair-copula, the best-fitting family, along with the optimal corresponding parameter(s) must be identified. To this end, a series of criteria can be employed such as goodness-of-fit (GOF) test, the Vuong test, the Akaike's information criterion (AIC), and the Bayesian inference criterion (BIC). Among them, we choose the AIC which has been shown to perform well in the past [40]. The optimal parameters of each candidate copula family are estimated via the Maximum Likelihood Estimation (MLE) method as implemented in [25] and [36]. In the context of a real data set, a large number of variables renders the C-vine construction process computationally expensive. To this end, dimensionality reduction technique such as Principal component analysis (PCA) can be exploited to decrease the computational complexity by fitting the parametric model in a lower-dimensional space.

III. C-vine Mixture Model Clustering Algorithm

A. Load Pattern Data Processing

Given a set of *N* customers, the electrical consumption of each customer is typically represented via a daily load pattern, calculated by averaging the measured load data over a period of time (typically in the order of a few months or a year). Subsequently, the representative load pattern (RLP) defined in [12] can be obtained as the normalized daily load pattern in [0,1] with regards to a reference value (i.e. peak value of the daily load pattern). In most of the existing literature on load pattern clustering, it has been demonstrated that it is effective to group customers into different classes via some appropriate clustering techniques on the basis of their RLPs.

Mathematically, let $X = [x_1, \dots, x_N] \in \mathbb{R}^{P \times N}$ denote the whole set of historical load data measurements, where *P* is the total number of data points measured during a determined period of time, each vector $x_n = [x_{n,1}, \dots, x_{n,P}] \in \mathbb{R}^P$ represents the monitored load data for the n^{th} customer, for $n = 1, \dots, N$. Assuming that m is the number of samples characterizing each daily load pattern, for each customer, the daily load pattern $\bar{x}_n = [\bar{x}_{n,1}, \dots, \bar{x}_{n,m}] \in \mathbb{R}^m$ can be calculated by averaging every m data points of x_n. Furthermore, the data set D of the N RLPs is represented by $D = [d_1, \dots, d_N]^T \in \mathbb{R}^{N \times m}$ in which each vector $d_n = [d_{n,1}, \dots, d_{n,P}] \in \mathbb{R}^P$ can be calculated as:

$$d_n = \bar{x}_n / \max(x_n) \quad (5)$$

After building the *N* RLPs via the above data processing procedure, the next step is to group the *N* RLPS into a pre-determined number of clusters by using the proposed CVMM clustering algorithm.

B. CVMM Clustering Algorithm

Let K denote the number of customer classes to be partitioned; the density of the C-vine mixture model (CVMM) for the data set D of the N RLPs is given by:

$$f(D|\Theta) = \sum_{k=1}^{K} \pi_k f_k(D|\Theta_k) \quad (6)$$

where $\pi_k \in [0,1]$ with $\sum_{k=1}^{K} \pi_k = 1$ and $\Theta = [{\Theta_1}^T, \dots, {\Theta_K}^T]^T$ represent the mixing proportions and the parameters of the component densities $f_k(D|\Theta_k)$, for $k = 1, \dots, K$, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2016.2614366, IEEE Transactions on Power Systems

Note that each parameter set Θ_k of Θ contains all the pair-copula parameters for the k^{th} component in the case of C-vine representation. According to equation (2), the multivariate distribution $f(D|\Theta)$ can be expressed as:

$$\sum_{k=1}^{K}\pi_k\prod_{l=1}^{m}f_k(d_{1:N,l})\prod_{j=1}^{m-1}\prod_{i=1}^{m-j}c_{j,j+i|v_{i,j}}^{k}\left(F_{j|v_{i,j}}^{k},F_{j+i|v_{i,j}}^{k}\right) \quad (7)$$

where $v_{i,j}=\{1,\dots,j-1\}$ and $c_{\cdot,\cdot}^{k}$ represents a pair-copula of the k^{th} component. Let $S=[s_1,\dots,s_N]^T$ denote the latent random variables whose element s_n indicates the label for d_n, the n^{th} observation of data set D, for $n=1,\dots,N$, the main aim of the proposed unsupervised CVMM clustering algorithm is to maximize the complete data log-likelihood function:

$$L=\sum_{n=1}^{N}\sum_{k=1}^{K}w_{n,k}\times\left(\sum_{j=1}^{m-1}\sum_{i=1}^{m-j}\ln c_{j,j+i|v_{i,j}}^{k}\left(F_{j|v_{i,j}}^{k},F_{j+i|v_{i,j}}^{k}|\theta_{j,i}^{k}\right)+\sum_{l=1}^{m-1}\ln f_k\left(d_{S=k,l}\right)+\ln\pi_k\right) \quad (8)$$

where $w_{n,k}$ represents the posterior probability of the n^{th} measurement that was drawn from the k^{th} component, calculated by:

$$w_{n,k}=\frac{\pi_k f_k(d_n|\Theta_k)}{\sum_{i=1}^{K}\pi_i\, f_i(d_n|\Theta_i)}\ (k=1,\dots,K). \quad (9)$$

In addition, the prior probability (mixing proportion) of the k^{th} component is denoted by π_k:

$$\pi_k=\frac{\sum_{i=1}^{N}w_{i,k}}{N}\ \ (k=1,\dots,K). \quad (10)$$

The expectation-maximization (EM) algorithm is used to estimate the maximum likelihood parameters Θ and the latent variables S of equation (8), given measurement data D. The conventional idea of the EM algorithm can be deciphered as an alternate between estimating the latent variables and the parameters. Beyond that, a more comprehensive understanding is to consider the EM algorithm as a lower bound maximization procedure [41]. Accordingly, the expectation step (E-step) can be regarded as a process to build a local lower bound to the posterior distribution, while the maximization step (M-step) aims to optimize the constructed bound. In general, after initializing the EM algorithm, the t^{th} iteration of the EM algorithm for CVMM can be summarized as:

- **E-step:** Calculate ${w_{n,k}}^{(t)}$ using equation (9), for $n=1,\dots,N, k=1,\dots,K$.
- **M-step:** Update ${\pi_k}^{(t)}$ according to equation (10). Maximize the likelihood function $L^{(t)}$ to get the parameters $\Theta^{(t+1)}$ and the labels $S^{(t)}$.

A detailed explanation of the above EM estimation for CVMM is given as follows. Firstly, regarding the initialization step, the superiority of the hierarchical clustering with average distance linkage criterion has been demonstrated in [12] when compared to k-means, fuzzy k-means, modified follow-the-leader, hierarchical clustering with other types of linkages, and self-organizing maps. Therefore, in this case, the first step is to initialize the EM algorithm by pre-grouping the input data D into K clusters via the hierarchical clustering with average distance linkage criterion. Note that the optimal number of clusters K is determined according to the clustering validity indicators, detailed in Section IV. Subsequently, the E-step of EM algorithm consists of the calculation of the posterior probabilities $\mathrm{W}=[w_{1:N,1},\dots,w_{1:N,K}]\in\mathbb{R}^{N\times K}$ and the update of labels S for N customers. The label s_n for each customer is set to the index of the maximum value in $[w_{n,1},\dots,w_{n,K}]$, for $n=1,\dots,N$. After updating the prior probabilities $\Pi=[\pi_1,\dots,\pi_K]\in\mathbb{R}^{K\times 1}$ using equation (10), the M-step maximize the likelihood function L by estimating the copula parameters set $\Xi=[\Theta_1,\dots,\Theta_K]$ for all the K C-vine models. During the procedure of C-vine construction, the first step is to transform the historical data to uniform margins through their corresponding empirical cumulative functions (ECDFs). Afterwards, model selection (via AIC) and parameter estimation (via MLE) are performed on each pair-copula to identify the best-fitting function. In each iteration, a total of $Km(m-1)/2$ pair copulas are fitted. The E-step and M-step are carried out iteratively until the relative change e in log-likelihood value L is less than a predefined threshold ε. The proposed algorithm is outlined below.

Algorithm 1 CVMM Clustering

Input: data set $\boldsymbol{D}=[\boldsymbol{d_1},\dots,\boldsymbol{d_N}]^{\boldsymbol{T}}$, number of clusters K.
Output: clustering labels $\boldsymbol{S}=[\boldsymbol{s_1},\dots,\boldsymbol{s_N}]^{\boldsymbol{T}}$.

Initialize customer labels $S^{(0)}=[{s_1}^{(0)},\dots,{s_N}^{(0)}]^T$ via an initial clustering technique (e.g. hierarchical clustering).
Initialize the prior probabilities $\Pi^{(0)}=[{\pi_1}^{(0)},\dots,{\pi_K}^{(0)}]\in\mathbb{R}^{K\times 1}$ and the posterior probabilities $W^{(0)}=[{w_{1:N,1}}^{(0)},\dots,{w_{1:N,K}}^{(0)}]\in\mathbb{R}^{N\times K}$.
$L^{(0)}=-\infty$
$t=1$
repeat
  for $n\leftarrow 1,..,N$
    Update ${w_{n,k}}^{(t)}\ \forall k$ using equation (9)
    $s_n^{(t)}=\arg\max\left\{[w_{n,1}^{(t)},\dots,w_{n,K}^{(t)}]\right\}$
  end for
  Update ${\pi_k}^{(t)}\ \forall k$ using equation (10)
  Transform ${f_k}^{(t)}\left(d_{S^{(t)}=k,l}\right)$ to ${F_k}^{(t)}\left(d_{S^{(t)}=k,l}\right)$ through its ECDF, $\forall l\in 1,..,m$
  for $j\leftarrow 1,\dots,m-1$
    for $i\leftarrow 1,\dots,m-j$
      Perform family selection and parameter estimation on the pair-copula ${c_{j,j+i|v_{i,j}}^{k}}^{(t)}({F_{j|v_{i,j}}^{k}}^{(t)},{F_{j+i|v_{i,j}}^{k}}^{(t)}|{\theta_{j,i}^{k}}^{(t)})$
    end for
  end for
  Calculate the log-likelihood $L^{(t)}$ using equation (8)
  $e^{(t)}=L^{(t)}-L^{(t-1)}$
  $t=t+1$
until $e^{(t-1)}<\varepsilon$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2016.2614366, IEEE Transactions on Power Systems

IV. Clustering Quality Evaluation

Different clustering validity indicators have been proposed in the past as a comprehensive tool to quantitatively analyze the quality of clustering schemes [1], [5], [12], [16]. As such, they can substantially inform the choice on the number of clusters to be applied in each case and enable the comparison between different clustering methods. Consider that the clustering results, K partitioned customer classes with computed centroids $C = \{c_1, \dots, c_K\}$, are obtained by applying the proposed CVMM algorithm on the processed data set D of the N RLPs. Let D_k denote the customers that belongs to k^{th} class, for $k = 1, \dots, K$, the four indicators taken into account in this paper are calculated as follows. Note that all distances are Euclidean.

1) The clustering dispersion indicator (CDI), defined as the ratio of the mean intraset distance between the RLPs in the same cluster and the intraset distance between the centroids of the K clusters [12], [10]:

$$CDI(D,K) = \hat{d}(C)^{-1}\sqrt{K^{-1}\sum_{k=1}^{K}\hat{d}^2(D_k)} \qquad (11)$$

2) The modified Dunn index (MDI), is computed as follows. For $i, j = 1, \dots, K$:

$$MDI(D,K) = \max_{1\le k\le K}\{\hat{d}(D_k)\} / \min_{i\neq j}\{d(c_i,c_j)\} \qquad (12)$$

where $d(.,.)$, $\hat{d}(.)$ represent cluster-to-cluster distance and intraset distance, respectively, as defined in [12].

3) The Davies-Bouldin index (DBI) [12], [10] is expressed as the system-wide average of the maximum ratio of the within cluster scatter to the between cluster separation, for $i, j = 1, \dots, K$:

$$DBI(D,K) = \frac{1}{K}\sum_{k=1}^{K}\max_{i\neq j}\{\frac{\hat{d}(D_i)+\hat{d}(D_j)}{d(c_i,c_j)}\} \qquad (13)$$

4) The mean index adequacy (MIA) [42] represents the average of the distances between the centroid of a cluster and each RLP in this cluster:

$$MIA(D,K) = K^{-1}\sum_{k=1}^{K} d^2(c_k, D_k) \qquad (14)$$

For all four indicators, a lower value indicates better clustering performance. By applying a clustering algorithm a number of times, we can obtain the curve that describes the relationship between the indicator values and the number of clusters K. Subsequently, the optimal number of clusters can be obtained either by seeking the first knee of this curve, considering the Bayesian information criterion (BIC), or the informational complexity criterion (ICOMP) [12]. In this paper, we determine the optimal value of K by utlizing the AIC. This way, the performance of the proposed CVMM clustering algorithm can be evaluated and compared to other existing technique (e.g. k-means, hierarchical clustering, etc.) through the comparison of their respective indicator values.

V. Case Study Application

A. LCL Load Dataset and Data Processing

The Low Carbon London (LCL) smart meter trials were designed to characterize the residential consumer demand of London and to assess the benefits from employing smart metering for distribution network operation [43]. In the project, Landis and Gyr (L+G) E470 electricity meters were installed in 2,613 residential homes across the Mayor of London's Low Carbon Zones and the London Power Networks distribution network license area operated by UK Power Networks. Specifically, the Engineering Instrumentation Zones of the LCL trial contain Queen's Park, Merton, and Brixton. In this paper, the LCL load dataset consists of 17,520 half-hourly electrical load consumption measurements across 2,613 customers in kW for a full calendar year from 1st January 2013 to 31st December 2013. The RLPs of all $N = 2{,}613$ customers were calculated according to the data processing procedure described in Section III. All RLPs are shown in Fig.3, each RLP consisting of 48 half-hourly normalized demand values. The mean of all RLPs, indicated in Fig.3 by a bold black line, follows the typical 'duck curve' shape indicating increased consumption levels during afternoon hours with a peak at around 7:00 pm. However, individual RLPs exhibit very high variation, with some consumers having a profile very different to the average RLP pattern.

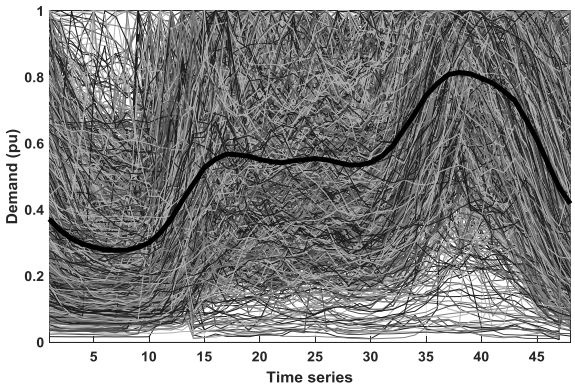

Fig.3. RLPs for all 2,613 customers of the LCL smart meter trials. Thick black line shows the mean RLP.

In this part, the proposed CVMM algorithm introduced in Section III is applied to group the customers in the LCL dataset. For the proposed CVMM clustering algorithm, it is required to fit $(48 \times 47)/2 \times t \times \varphi = 1128\varphi$ pair-copulas, where φ is the number of candidate bivariate copula families. The candidate families considered in this application are Gaussian, Student-t, Clayton, Frank, Gumbel as well as their 90°,180° and 270° rotated versions i.e. φ = 20. As such, the presented modelling task will involve fitting 22,560*K copulas, repeated a number of times until convergence is achieved. Given that a single copula parameterization may take a few milliseconds, it is clear that the proposed procedure can involve a high computational burden.

To accelerate this procedure, a widely-used linear dimensionality reduction technique, principal component analysis (PCA), is used to construct a lower dimensional representation of the 48-dimensional RLPs named principal components (PCs), that retains as much of the variance as possible. The principal components are listed in descending order of the eigenvalues of the RLPs' covariance matrix. Note that the eigenvalues can be treated as a metric of the information contained within each principal component. To this end, we use a user-defined information retainment threshold to determine the number of PCs to be retained. In this case, we set the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2016.2614366, IEEE Transactions on Power Systems

information retainment threshold to 97.5%, which corresponds to 12 PCs being retained. The size of the reduced dataset is 2,613 observations of 12 variables. The convergence threshold ε has been set to 10^{-5}. We indicatively mention that, for the CVMM algorithm with $K = 8$, convergence was achieved within 3 iterations, totaling CPU time of 3 minutes when running on 8 cores. Note that the CVMM algorithm can be parallelized so that each candidate copula family fit is carried out independently, resulting in significant speed-ups. In addition, a smaller number of retained PCs for the CVMM methods can also reduce the computational burden. A more extensive analysis of the method's computational performance and how it compares with other methods is shown in Section V.C.

B. Determining the Optimal Number of Clusters for CVMM

As stated in [44], one of the most important advantages of model-based clustering methods is rendering the issue of determining the optimal number of clusters to a statistical model selection problem based on information criteria such as AIC and BIC. For the proposed CVMM algorithm, the calculated AIC and BIC values for $K = 5$ to 20 are shown in Fig. 4. Accordingly, the optimal number of mixture models (clusters) for the tested dataset is equal to eight ($K_{opt} = 8$) indicated by the first local maximum in both of the AIC and BIC curves. Although AIC and BIC differ in the way they are not guaranteed to agree in their selection, in this case both methods indicate the same optimal number of clusters.

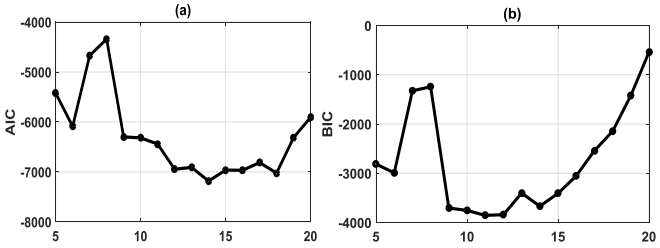

Fig.4. Calculated AIC (a) and BIC (b) values for $K = 5$ to 20.

C. Clustering Validity Assessment

In this section the performance of CVMM is evaluated and compared with other well-known clustering techniques used in the past; k-means, hierarchical clustering (complete linkage), hierarchical clustering (average linkage), hierarchical clustering (weighted linkage), hierarchical clustering (Ward linkage), and Gaussian mixture model clustering (GMM). Note that all the above approaches are applied to the original dataset of RLPs (48 variables), whereas the CVMM algorithm includes a dimension reduction stage that carries out clustering on a lower-dimensional dataset (12 variables in this case). As discussed in Section IV, four types of indicators, CDI, MDI, DBI and MIA, are chosen for carrying out the clustering validity assessments for different numbers of clusters *K*, ranging from 5 to 20.

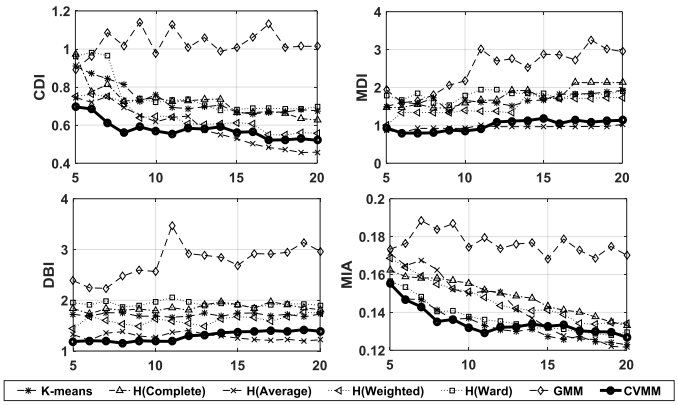

Fig. 5. Comparisons among the clustering techniques by using (upper-left) the CDI indicator, (upper-right) the MDI indicator, (lower-left) the DBI indicator, and (lower-right) the MIA indicator for $K = 5$ to 20.

The results illustrated in Fig.5 indicate that, for the tested dataset, the proposed CVMM algorithm exhibits superior performance to the other techniques. This is evidenced by the fact that classification via CVMM consistently results in lower values across all four indicators in the range up to 12 clusters. Although the H(Average) hierarchical method does result in lower indicator values for larger numbers of K we are interested in comparing values in the range where the knee for each method occurs. Overall it is constructive to note that the hierarchical methods H(Average) and H(Weighted) perform relatively better than H(Complete), k-means and H(Ward). The limitations of GMM is indicated by the large values across all four indicators, highlighting the concern with standard multivariate model-based methods. As such, the proposed CVMM clustering algorithm is shown to form well-separated customer classes and to detect outlier load patterns.

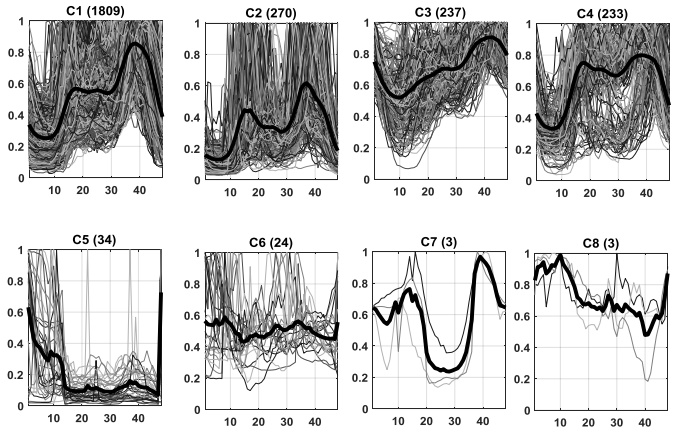

Fig. 6. Clustering results for the proposed CVMM clustering algorithm with 8 clusters. Thick black lines show mean RLP for each cluster. For each subplot, horizontal axis represents time series index (hour); and vertical axis indicates demand (pu).

Fig.6 shows the load patterns groups that were classified by the proposed CVMM clustering algorithm with $K = 8$. The number of consumers categorized in each cluster is shown in brackets next to each panel's title. Within each subplot, the average load pattern of all the customers in this class is plotted using a thick black line. From visually inspecting the plots, it is evident that the average load patterns of different classes have radically different shapes. In addition, there are clusters that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2016.2614366, IEEE Transactions on Power Systems

have a very large number of customers, while some clusters include outlying consumption behavior.

Cluster 1 (denoted C1) is the most populous cluster with over 69% of all customers. The mean profile exhibits the typical 'duck curve' shape, with peak consumption occurring in the afternoon hours. Also, mean profiles of C2 (10% of all consumers) and C4 (9%) have the similar shape as C1, but with slightly lower and higher consumption levels respectively. C3 (9%) and C8 (0.1%) contain the residents with relatively high consumption especially in the evening hours and the midnight hours, respectively, whereas most of the residents in C5 (1%) exhibit low electrical energy consumption during the daytime and have a sudden increase in midnight hours. In C6 (0.9%), mean consumption has less fluctuation than other clusters. On the contrary, large volatility is shown in C7 (0.1%) with abruptly low RLP values between 10:00 am and 17:00 pm. C8 exhibits high overall consumption with particularly high values during nighttime.

It is constructive to highlight that only 3 consumers have been placed in C7 and C8. By closer inspection of the individual patterns clustered in C7 and C8 we see that these profiles are indeed very similar between them and yet substantially different to other groups. Efficient detection of outliers is an important feature of well-performing clustering schemes; CVMM clearly succeeds in this task, presenting superior outlier detection capability. In addition, an example of the clustering results of k-means method is shown in Fig. 7. It can be observed that the simpler method, k-means, fails to correctly identify the outliers identified in C7 and C8, clustered via CVMM.

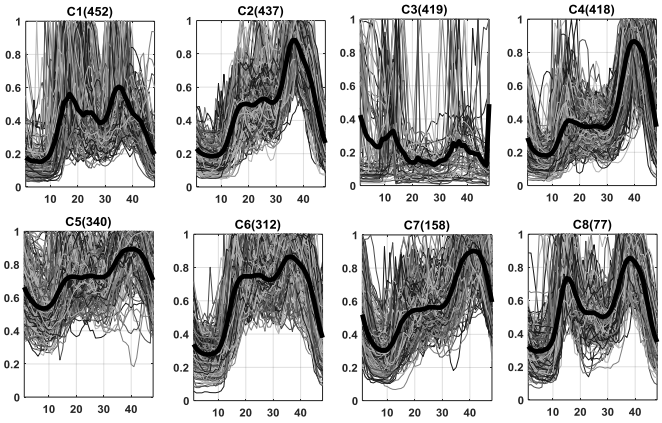

Fig. 7. Clustering results for k-means with 8 clusters. Thick black lines show mean RLP for each cluster. For each subplot, horizontal axis represents time series index (hour); and vertical axis indicates demand (pu).

A way to further analyze the average load profiles of each cluster is to calculate the normalized load shape indices that involve the information of the daily load curve shape during different periods of time. We calculate three attributes of interest as presented in [42]; load factor (d1), night impact (d3), and lunch impact (d5). In particular, d1 represents the full day consumption behavior, while d3 and d5 capture the load shape from 11:00 pm to 7:00 am (8 hours) and from 12:00 pm to 15:00 pm (3 hours), respectively. For the different clusters obtained by CVMM algorithm, the calculated d1, d3 and d5 values of the mean RLP of each cluster are shown in TABLE I. In terms of d1, C3 and C8 have relatively high values compared to other clusters. C5 exhibits the lowest d1 value but also the highest d3 value, which means it has a very high consumption during the night. In addition, C7 exhibits the lowest d5 value, which means severely reduced electricity usage around noon, as also evidenced in Fig.4.

TABLE I. LOAD ATTRIBUTES FOR DIFFERENT CLUSTERS

	C1	C2	C3	C4	C5	C6	C7	C8
d1	0.534	0.345	**0.706**	0.625	**0.193**	0.499	0.566	**0.723**
d3	0.197	0.190	0.287	0.221	**0.622**	0.345	0.379	0.413
d5	0.127	0.115	0.124	0.135	0.062	0.128	**0.052**	0.116

D. Clustering Results Analysis Using Metadata

In the LCL smart meter trials, in addition to consumption measurements, various socio-economic conditions of the participating household were also recorded. In this paper, we focus on the data pertaining to household occupancy and wealth level. The former relates to the number of people living in the property. The latter has been drawn on the basis of mapping all participating households to ACORN groups [43]. Subsequently, three wealth classes have been defined: Adverse, Comfortable, and Affluent in increasing order. A demographic group is defined as a combination of occupancy and wealth level. The number of customers belonging to each of the nine groups is shown in Table VII. The clusters generated via the CVMM algorithm are analyzed in Fig. 8 using the metadata in Table II.

TABLE II. NUMBER OF HOUSEHOLDS ACROSS DEMOGRAPHIC GROUPS

	1 occupant	2 occupants	3+ occupants
Adverse	312	275	233
Comfortable	237	300	209
Affluent	428	398	221

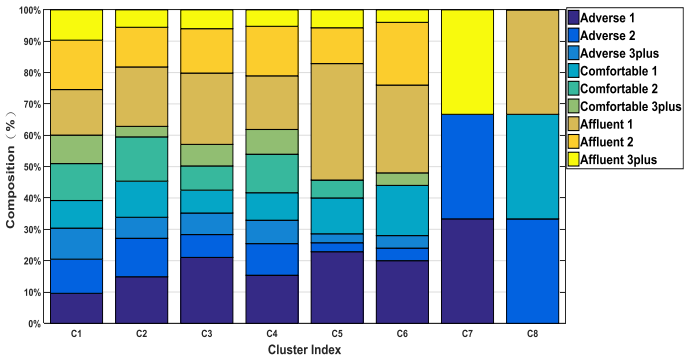

Fig. 8. Cluster composition analysis using demographic metadata.

For C2, C5, and C6, there is no wealth category that dominates; these three clusters consist of Adverse, Comfortable, and Affluent households in almost equal distribution. The number of people living in the property also cannot be evidently distinguished in this case. On the other hand, in C6, households with one occupant account for the highest percentage of about 46%. C4 consists mostly of adverse wealth households (67%), with the rest of customers belonging to the Affluent group. In C6, groups of 1 and 2 occupants account for 67% and 23%, respectively. In addition, approximately 50%, 43%, and 53% of the RLPs in C1, C3, and C8 are in the Affluent

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2016.2614366, IEEE Transactions on Power Systems

group respectively. Meanwhile, households with 1 occupant account for 67%, 50% and 74% of households, respectively. The above results demonstrate that there are no clear nor consistent discrepancies across different wealth levels and number of occupants when comparing residential RLPs. This highlights the importance of performing customer partitioning on the basis of actual consumption measurements.

E. New Customer Characterization

An electrical customer characterization framework was proposed in [42] to generate customers classes based on the existing consumers' RLPs (*Clustering Module*) and subsequently build a classification model using decision trees (DTs) that assist system operators to assign new consumers to the previously constructed classes (*Classification Module*). In [42] the k-means clustering algorithm was applied. In this paper, customer clustering is carried out using the proposed CVMM clustering algorithm. In the *Classification Module*, d1, d3, and d5 are chosen as classifying variables. A DT model is trained on the load shape indices (d1, d3, and d5) and the corresponding cluster indices. Using the constructed DT model, we can accomplish the assignment of new customers to the classes obtained by the CVMM algorithm.

For the purpose of evaluating the constructed DT model, ten-fold cross-validation is used to get an estimate of classification accuracy. To this end, the calculated load shape indexes are randomly partitioned into 10 test datasets. In fact, the customers in the test dataset can be treated as new customers but with known labels. Each test data set accounts for 10% of the total customers, and the remaining 90% customers are attributed to the training data set. For each fold, the DT model is trained using one of the training data sets and then tested with the corresponding test data set. This process is presented in Fig. 9.

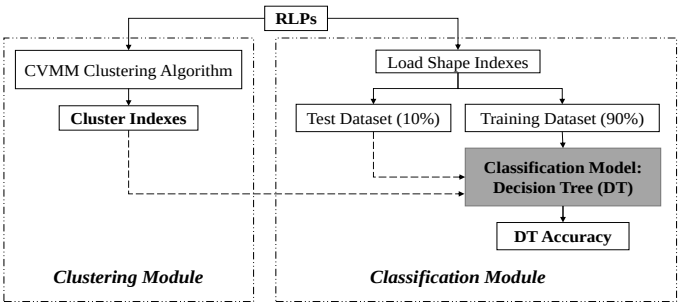

Fig.9. The structure of the modified customer characterization framework with the proposed CVMM clustering algorithm.

The overall accuracies of the constructed DT models with different clustering methods are listed in Table VIII. In this case, accuracy is defined as the ratio of correctly classified households, averaged across all ten test sets. As can be seen below, the DTs trained on the classes generated via the proposed CVMM algorithm have the highest accuracy followed by hierarchical clustering. This analysis simulates a challenge that suppliers have to tackle in practice; classification of a new customer for which they may have limited information (in this case we assume we have an approximation of their overall consumption level, as well as an estimate of consumption intensity at night and noon times). The fact that CVMM exhibits the best predictive capability means that the proposed algorithm has partitioned customer profiles in such a way so that accurate conclusions about detailed usage patterns can be drawn in the absence of full information. As shown in Table VIII, other methods such as k-means and GMM perform poorly in this respect, rendering them unfit for classifying new consumers. The hierarchical clustering method with average linkage, which was used to initialise the CVMM algorithm performs well, but it is evident that the proposed method enhances clustering performance.

It is also important to note that the order of the DTs' performance across the different clustering methods is highly consistent with the results of the clustering validation assessment shown in Fig.3. For example, H(Average) is the second-best performing algorithm after CVMM, whereas the DT accuracy of GMM is only 51.93%, which is the worst performing method as shown in Fig. 3.

Note that when including ACORN demographic data in the training attribute set, the classification error improves only marginally. This highlights the significance of d1, d3, d5 in terms of clustering new customers and shows that the availability of demographic data may convey little information regarding daily demand consumption patterns of a prospective user.

TABLE III. CLASSIFICATION MODULE DT ACCURACIES

Method	Accuracy
CVMM	93.65%
Hierarchical(Average)	89.82%
Hierarchical(Weighted)	87.83%
Hierarchical(Complete)	62.31%
K-means	56.90%
Hierarchical(Ward)	55.42%
GMM	51.93%

F. Clustering Across Different Seasons and Day Types

In all the preceding analysis we have analysed the full dataset corresponding to an entire year without differentiating between different calendar seasons nor between weekends and weekdays. Naturally, the season and day type has a substantial impact on consumer behaviour as shown in Fig. 10 (note that seasons are defined in accordance to the specification of the UK market operator [45]). In Fig. 10. (a), average residential electricity consumption begins later in the morning and leads to a small mid-day peak instead of the valley experienced in weekdays when residents are typically away from their home. The impact of season is also substantial as expected; for example, the 7pm consumption peak is higher in winter than the other seasons. To this end, it is of interest to evaluate the performance of CVMM when dealing with the datasets under different loading conditions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2016.2614366, IEEE Transactions on Power Systems

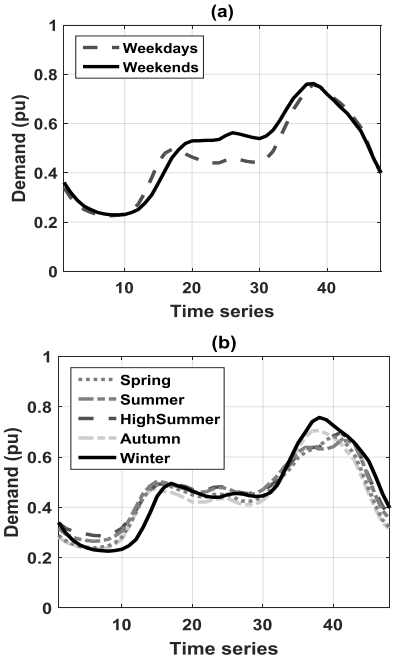

Fig. 10. Average RLPs for all 2,613 customers of the LCL smart meter trials under the loading conditions: (a) Weekdays/Weekends; (b) Seasons.

For the purpose of determining tariffs for different customer types, the consumption behaviors of each customer in weekdays and weekends could be considered simultaneously. Therefore, we combine the calculated RLPs under loading conditions, weekdays and weekends, together to construct a 96-dimensional dataset. Based on this dataset, the performance of the proposed CVMM method is evaluated and compared with the other candidate methods, using the above-mentioned four types of indicators, the CDI, the MDI, the DBI, and the MIA, for different numbers of clusters, ranging from 5 to 20.

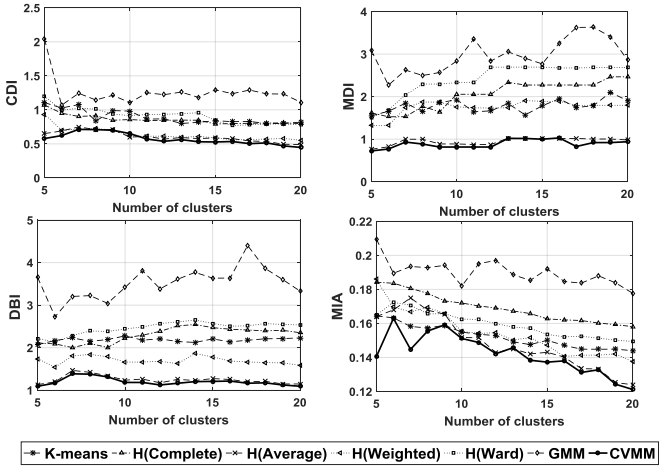

Fig. 11. Comparison among the clustering techniques based on weekdays and weekends RLPs using CDI, MDI, DBI and MIA indicators for $K = 5$ to 20.

Similar to Fig. 5, the results illustrated in Fig.11 also show better performance of the proposed CVMM algorithm comparing to the other methods according to their indicator values. It is imperative to note that the proposed CVMM method has the advantage of detecting outlier load patterns, although H(Average) method has a competitive performance in terms of the indicator values. In this case, the optimal number of classes is equal to twelve ($K_{opt}^{week} = 12$) according to the calculated AIC and BIC values.

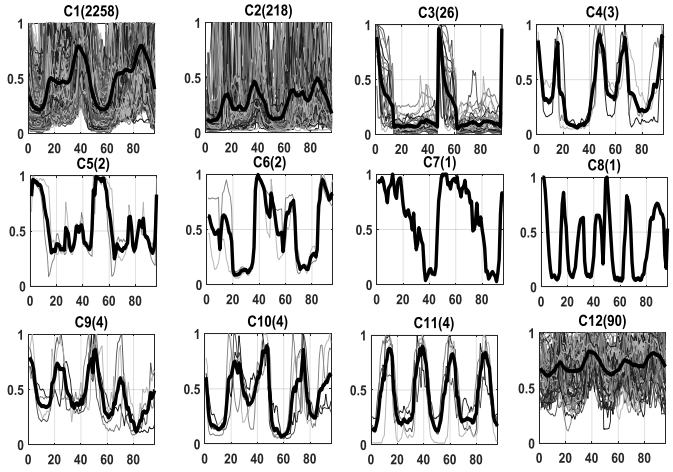

Fig. 12. Clustering results of the combined weekdays and weekends dataset for the proposed CVMM clustering algorithm with 12 clusters. Thick black lines show mean RLP for each cluster. For each subplot, horizontal axis represents time series index (hour); and vertical axis indicates demand (pu).

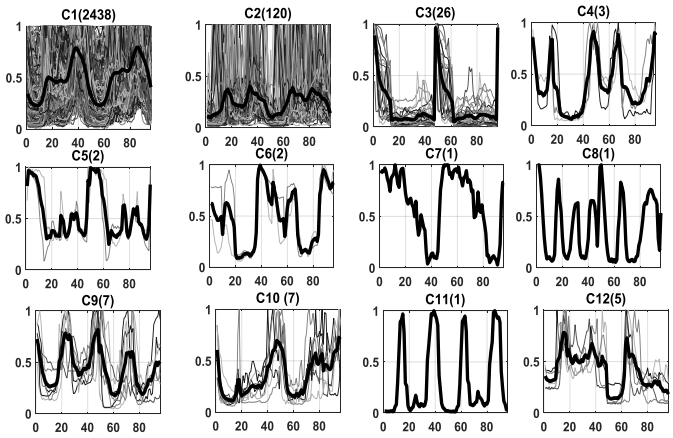

Fig. 13. Clustering results of the combined weekdays and weekends dataset for the H(Average) clustering algorithm with 12 clusters. Thick black lines show mean RLP for each cluster. For each subplot, horizontal axis represents time series index (hour); and vertical axis indicates demand (pu).

Fig.12 and Fig.13 show the load patterns groups that were classified by the proposed CVMM clustering algorithm and H(Average) algorithm with $K = 12$, respectively. The average load pattern of all the customers in this class is plotted using a red bold line within each subplot. Visually inspecting indicates that both of these methods have the capability to identify outlying consumption behaviors. As H(Average) is the initial clustering algorithm of the proposed CVMM method, it is reasonable that the first nine clusters in Fig.12 and Fig.13 have similar average load patterns. However, comparing the last three clusters, more important load patterns are successfully detected by the proposed CVMM algorithm. In particular, the customers in Cl2 of Fig.12 exhibit consistent high consumptions during the whole week, which is important to be distinguished for the purpose of designing the tariffs.

Besides the influence of days, the impact of different calendar seasons can also be a significant consideration. As shown in Fig.10 (b), during the peak time between 17:00 pm

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2016.2614366, IEEE Transactions on Power Systems

and 22:00 pm, the average RLP of winter exhibits the highest normalized peak demand, whereas customers in high-summer include relatively low consumption. In addition, the average consumption autumn, spring, and summer decrease progressively. According to the previous results, H(Average) method has the most competitive performance compared with the proposed CVMM algorithm. Therefore, for different seasons' datasets, we only perform these two clustering methods with the optimal number of clusters K=8. The results of CVMM and H(Average) are evaluated by the aforementioned four indicators, given in Table V and Table VI, respectively. Comparing the criterions values across all seasons and all types of indicators, the proposed CVMM method always exhibit lower values which mean better clustering results than the H(Average) method.

TABLE IV. PERFORMANCE OF CVMM ACROSS DIFFERENT SEASONS

	CDI	MDI	DBI	MIA
Winter	0.5976	0.8383	1.0912	0.1470
Spring	0.7173	0.8916	1.4450	0.1465
Summer	0.5486	0.8387	1.2494	0.1329
High Summer	0.5287	0.8952	1.2183	0.1257
Autumn	0.5388	0.8516	1.3003	0.1364

TABLE V. PERFORMANCE OF H(AVERAGE) ACROSS DIFFERENT SEASONS

	CDI	MDI	DBI	MIA
Winter	0.6322	0.8384	1.2380	0.1557
Spring	0.7212	0.9187	1.4571	0.1598
Summer	0.5601	0.8954	1.2870	0.1381
High Summer	0.5450	0.9052	1.2227	0.1340
Autumn	0.6104	0.9483	1.3942	0.1487

An example of the Autumn season is given to visually demonstrate the superior performance of the proposed CVMM method. In Fig. 14 and Fig. 15, the first five clusters have the similar load patterns. However, more customers within C3 and C5 are detected by the CVMM method. In addition, another type of important load pattern is identified by the proposed CVMM method, which exhibits a low consumption during the daytime and a sudden high consumption in the midnight. Furthermore, C7 in Fig. 14 is also an outlier which is not contained in the results of H(Average). On the other hand, the outliers detected by H(Average) method include some similar shapes such as C1 and C6, C3 and C7, C4 and C8.

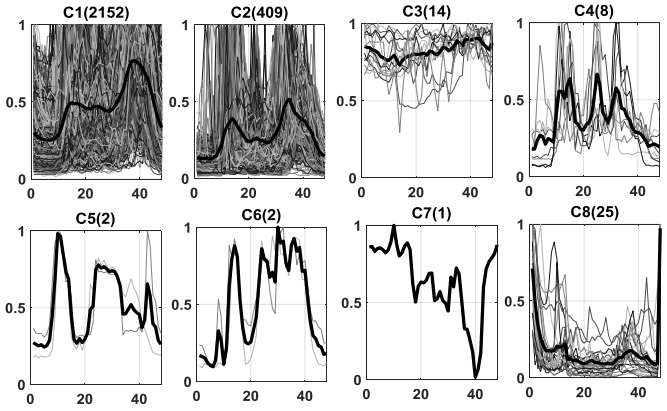

Fig. 14. Clustering results of the Autumn dataset for the CVMM clustering algorithm with 8 clusters. Thick black lines show mean RLP for each cluster. For each subplot, horizontal axis represents time series (hour); and vertical axis indicates demand (pu).

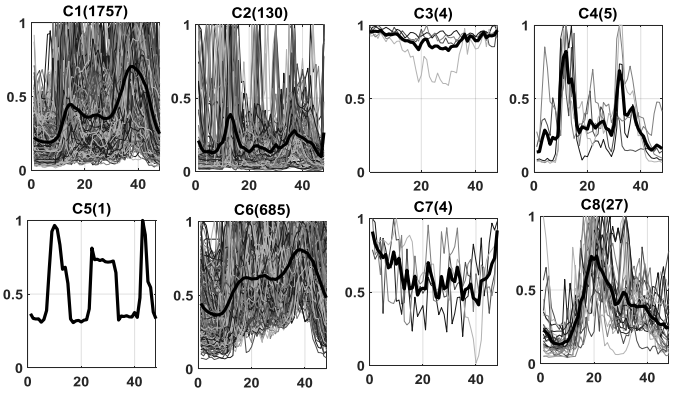

Fig. 15. Clustering results of the Autumn dataset for the H(Average) clustering algorithm with 8 clusters. Thick black lines show mean RLP for each cluster. For each subplot, horizontal axis represents time series (hour); and vertical axis indicates demand (pu).

G. Performance for Large Data Sets

As mentioned in [5], one major direction that warrants further exploration is the development of well-performing clustering techniques capable of dealing with large amounts of smart metering data. In this section we explore the proposed method's computational and performance scaling when dealing with a large number of customers, well above the size of the original dataset of 2,613 households. For this aim synthetic data sets of 5,000, 10,000, 20,000, and 50,000 customers have been constructed by sampling the initial population of households. Gaussian noise has been added to the measurements to approximate the variability that would exist in larger datasets. The CPU times and CDI for all different methods are shown in Table VI and Table VII respectively. Note that in the case of CVMM method two variants were carried out; 6 and 12 retained PCs to illustrate the trade-off between the computation time the clustering performance. Each method was carried out in parallel on 8 cores.

TABLE VI. COMPUTATION TIMES (SECONDS) FOR DIFFERENT METHODS

	Customer Population				
	2,613	**5,000**	**10,000**	**20,000**	**50,000**
CVMM-12PCs	178.55	310.71	680.26	1436.21	4027.38
CVMM-6PCs	24.50	38.85	176.95	358.6	1374.29
H(Average)	0.02	0.65	2.57	12.26	156.45
H(Weighted)	0.11	0.56	2.33	12.96	142.33
H(Complete)	0.17	0.58	2.41	13.81	143.44
H(Ward)	0.10	0.54	2.35	13.08	135.46
K-means	0.02	0.05	0.14	0.47	1.91
GMM	0.11	2.58	3.18	3.375	8.09

TABLE VII. CDI FOR DIFFERENT METHODS

	Customer Population				
	2,613	**5,000**	**10,000**	**20,000**	**50,000**
CVMM-12PCs	0.56	0.32	0.33	0.41	0.43
CVMM-6PCs	0.62	0.38	0.36	0.45	0.47
H(Average)	0.69	0.38	0.37	0.47	0.51
H(Weighted)	0.72	0.48	0.81	0.68	0.93
H(Complete)	0.73	0.62	0.99	1.05	1.10
H(Ward)	0.71	1.07	0.97	1.15	1.09
K-means	0.81	0.96	0.96	0.96	0.97
GMM	1.01	1.21	1.87	1.42	1.21

In terms of computation times, k-means and GMM are the fastest methods by a large margin. However, as evidenced by the high CDI criteria values shown in Table VII, these conventional techniques exhibit poor clustering performance. The four hierarchical methods were found to scale well computationally with little CPU time variation between them. However, H(Average) is the best performer in terms of clustering quality, as indicated by the lower indicator values.

On the other hand, CVMM is clearly the most computationally expensive method. When 12 PCs are retained, the clustering of 50,000 customers takes more than one hour. However, the clustering quality of CVMM is substantially higher as suggested by the lower CDI values. CVMM's superior performance persists even when only 6 PCs are retained, which involves a substantial decrease in computation times. It is also constructive to highlight that CVMM's computational burden increases linearly as a function of households, which indicates good scaling performance for practical applications. We finally note that the fully-parallelisable nature of the algorithm could be exploited further; deploying more CPU resources will result in a linear decrease of computation time, potentially allowing the clustering of millions of consumers within a few hours.

VI. Conclusion

This paper has proposed a novel copula-based mixture model clustering algorithm based on C-vine representations for grouping electrical customers according to consumption data. C-vine copulas, which can capture complex dependency structures by employing cascades of bivariate distribution, have been, for the first time, integrated into a mixture model-based clustering method. In order to address the high computational burden of the proposed scheme, model fitting takes place in a lower-dimension space where we exploit the synergy between PCA and the C-Vine structure's hierarchy. To evaluate the performance of the proposed CVMM algorithm, a set of well-established clustering validation indicators (CDI, MDI, DBI and MIA) have been used.

One major conclusion stemming from the analysis performed on a large dataset containing half-hourly smart meter recordings is that the CVMM method exhibits superior performance when compared to existing methods, indicated by the lower validation indicators' values. We also highlight the method's ability to detect outliers, in contrast to conventional techniques. The classification result of different techniques was further assessed according to a decision-tree based classification module. The result of the DTs' performance is highly consistent with the outcome of the clustering validation assessment. This result highlights the superiority of the proposed approach in a practical context when assigning new consumers to existing classes. In addition, the proposed method is used to analyze load partitioning behavior under different loading conditions such as different calendar seasons and days. This increases the dimensionality of the clustering problem but could uncover particular customer classes that have markedly daily and seasonal changes in their behavior, opening up the possibility for more tailored tariff structures. The superior performance of the proposed method is also well demonstrated under different conditions in terms of the indicator values and the visual comparisons.

Future research will focus on improving the proposed CVMM method by employing more possible types of vine constructions (e.g. R-vine), and exploring alternative dimension reduction techniques to further relieve the method's computational burdsen. Another topic of interest is the extension of the proposed model to mixed data sets so as to accommodate additional information such as load conditions; this is important to render the algorithm suitable for real-time pricing applications.

References

[1] M. Misiti, Y. Misiti, G. Oppenheim, and J.-M. Poggi, "Optimized clusters for disaggregated electricity load forecasting," *REVSTAT Statist. J.*, vol. 8, no. 2, pp. 105–124, Nov. 2010.

[2] M. Chaouch, "Clustering-Based Improvement of Nonparametric Functional Time Series Forecasting: Application to Intra-Day Household-Level Load Curves," *IEEE Trans. Smart Grid*, vol. 5, pp. 411-419, 2014.

[3] J. Nazarko, A. Jurczuk, and W. Zalewski, "ARIMA models in load modeling with clustering approach," in *Proc. IEEE Russia Power Tech, St. Petersburg, Russia*, Jun. 27–30, 2005, pp. 1–6.

[4] R. F. Chang and C. N. Lu, "Load profiling and its applications in power market," in *Proc. IEEE Power Eng. Soc. General Meeting*, Jul. 13–17, 2003, vol. 2.

[5] G. Chicco, "Overview and performance assessment of clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68–80, Jun. 2012.

[6] P. Jokar, N. Arianpoo and V. C. M. Leung, "Electricity Theft Detection in AMI Using Customers' Consumption Patterns," in *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216-226, Jan. 2016.

[7] R. Granell, C. J. Axon and D. C. H. Wallom, "Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles," in *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 3217-3224, Nov. 2015.

[8] F. L. Quilumba, W. J. Lee, H. Huang, D. Y. Wang and R. L. Szabados, "Using Smart Meter Data to Improve the Accuracy of Intraday Load Forecasting Considering Customer Behavior Similarities," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 911-918, March 2015.

[9] I. Benítez, A. Quijano, J.-L. Díez, and I. Delgado, "Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers," *International Journal of Electrical Power & Energy Systems*, vol. 55, pp. 437-448, 2014.

[10] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," *IEEE Trans. Power Syst.*, vol. 22, no. 3, pp. 1120–1128, Aug. 2007.

[11] F. McLoughlin, A. Duffy, M. Conlon, A clustering approach to domestic electricity load profile characterisation using smart metering data, *Applied Energy*, Vol. 141, pp. 190-199, Mar. 2015.

[12] G. Chicco, R.Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 933–940, May 2006.

[13] G. Chicco, O.-M. Ionel, and R. Porumb, "Electrical load pattern grouping based on centroid model with ant colony clustering," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1706–1715, May 2013.

[14] S.V. Verdu, M.O. Garcia, C. Senabre, A.G. Martin and F.J.G. Franco, "Classification, Filtering, and Identification of Electrical Customer Load Patterns Through the Use of Self-Organizing Maps," *IEEE Transactions in Power Systems*, vol. 21, no. 4, pp. 1672-1682, Nov. 2006.

[15] J. Nazarko and Z. A. Styczynski, "Application of statistical and neural approaches to the daily load profile modeling in power distribution systems," *Proc. IEEE Transm. Dist. Conf.,New Orleans*, LA,Apr. 11–16, 1999, vol. 1, pp. 320–325.

[16] Y. Wang; Q. Chen, C. Kang; M. Zhang; K. Wang and Y. Zhao, "Load profiling and its application to demand response: A review," *Tsinghua Science and Technology* , vol.20, no.2, pp.117-129, Apr. 2015.

[17] C. Fraley and A.E. Raftery, "How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis," *The Computer Journal*, vol. 41, no. 8, pp. 578-588, 1998.

[18] R .Li, Z. Wang, C. Gu, F. Li, H. Wu, "A novel time-of-use tariff design based on Gaussian Mixture Model," *Applied Energy*, vol. 162, pp. 1530-1536, Jan. 2016.
[19] B. Stephen, A. J. Mutanen, S. Galloway, G. Burt and P. Järventausta, "Enhanced Load Profiling for Residential Network Customers," in *IEEE Transactions on Power Delivery*, vol. 29, no. 1, pp. 88-96, Feb. 2014.
[20] W. Labeeuw , G. Deconinck, "Residential electrical load model based on mixture model clustering and Markov models," *IEEE Trans. Ind. Informat.*, vol. 9, no. 3, pp. 1561–1569, Aug. 2013.
[21] R. Granell, C. J. Axon, D. Wallom, "Clustering disaggregated load profiles using a Dirichlet process mixture model," *Energy Conversion and Management*, pp. 507 – 516, Mar. 2015.
[22] A. Tewari, M. J. Giering, and A. Raghunathan, "Parametric characterization of multimodal distributions with non-gaussian modes," *Proc. IEEE International Conference on Data Mining Workshops*, pp. 286-292, 2011.
[23] M. Rey, V. Roth, "Copula mixture model for dependency-seeking clustering," in: *Proceedings of the 29th International Conference on Machine Learning, ICML*, pp. 927–934, 2012.
[24] I. Kosmidis and D. Karlis, "Model-based clustering using copulas with applications", *arXiv:1404.4077*, 2014.
[25] K. Aas, C. Czado, A. Frigessi and H. Bakken, "Pair-copula constructions of multiple dependence," *Insurance: Mathematics and Economics*, vol. 44, no. 2, pp. 182-198, 2009.
[26] C.J. Stone, "Optimal rates of convergence for nonparametric estimators," The Annals of Statistics, vol. 8, pp. 1348-1360, 1980.
[27] T. Nagler and C. Czado, "Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas," arXiv:1503.03305v3, May 2016.
[28] A. Albert and R. Rajagopal, "Smart Meter Driven Segmentation: What Your Consumption Says About You," *IEEE Trans. Power Systems*, vol. 28, pp. 4019-4030, 2013.
[29] Y. Wang, Q. Chen, C. Kang and Q. Xia, "Clustering of Electricity Consumption Behavior Dynamics toward Big Data Applications," *IEEE Trans. Power Syst.*, vol. PP, no. 99, pp.1-11, Apr. 2016.
[30] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 334, pp. 1492-1496, 2014.
[31] E. D. Varga, S. F. Beretka, C. Noce and G. Sapienza, "Robust Real-Time Load Profile Encoding and Classification Framework for Efficient Power Systems Operation," *IEEE Transactions on Power Systems*, vol. 30, no. 4, pp. 1897-1904, July 2015.
[32] S. Haben, C. Singleton and P. Grindrod, "Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 136-144, Jan. 2016.
[33] R. Al-Otaibi, N. Jin, T. Wilcox and P. Flach, "Feature Construction and Calibration for Clustering Daily Load Curves from Smart-Meter Data," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 2, pp. 645-654, April 2016.
[34] M. Sun, I. Konstantelos, S. Tindemans and G. Strbac, "Evaluating composite approaches to modelling high-dimnesional stochastic variables in power systems," PSCC '16, Genoa, pp. 1-8, 2016.
[35] A. Sklar, "Fonctions de répartition à n dimensions et leurs marges, " *Publ. Inst. Statist. Univ. Paris* **8**: 229–231.
[36] E. C. Brechmann et al. "Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine", *Journ. of Stat. Soft.*, Vol. 52 (3), '13.
[37] H. Joe, "Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters*", Lecture Notes-Monograph Series*, vol. 28, pp. 120-131, 1996. [Online]. Available: http://www.jstor.org/stable/4355888
[38] J. Mai and M. Scherer, *Simulating Copulas: Stochastic Models, Sampling Algorithms and Applications*, World Scientific, 2012.
[39] T. Bedford and R. M. Cooke, "Probability density decomposition for conditionally dependent random variables modeled by vines", *Annals of Mathematics and Artificial intelligence*, vol.32, no.1-4, pp. 245-268, 2001.
[40] E. C. Brechmann, C. Czado, and K. Aas, "Truncated regular vines in high dimensions with application to financial data." *Canadian Journal of Statistics* 40, no. 1, pp. 68-85, 2012.
[41] F. Dellaert "The expectation maximization algorithm," College of Comput., GATECH, Atlanta, GA, Tech. Rep. GIT-GVU-02–20, 2002.
[42] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on datamining techniques," *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 596–602, May 2005.
[43] M. Sun, I. Konstantelos and G. Strbac, "Analysis of Diversified Residential Demand in London using Smart Meter and Demographic Data," *IEEE PES GM 2016*, Boston, Jul. 2016.
[44] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery and W.L. Ruzz, "Model-Based Clustering and Data Transformations for Gene Expression Data," *Bioinformatics*, vol. 17, pp. 977-987, 2001.
[45] Elexon, Definitions of seasons and day types: v.1.01, 2014.

BIOGRAPHIES

Mingyang Sun (SM'12) received the bachelor's degree in mechanical engineering from Dalian University of Technology, China, and the master's degree in control system from Imperial College London, UK. He is currently pursuing his Ph.D. in the Control and Power group at Imperial College London. His research focuses on modelling of high-dimensional stochastic variables in power systems.

Ioannis Konstantelos (M'12) received the MEng degree in Electrical and Electronic Engineering from Imperial College London in 2007. He obtained his PhD from the same university in 2013 in the field of electrical energy systems. His research interests include mathematical programming and statistical modeling techniques applied to the planning and operation of energy systems.

Goran Strbac (M'95) is a Professor of electrical energy systems at Imperial College, London, UK. His current research interests include electricity generation, transmission and distribution operation, planning and pricing, and integration of renewable and distributed generation in electricity systems.