Temporal Association Rules For Electrical Activity Detection in Residential Homes

Hông-Ân Cao
Department of Computer Science
ETH Zurich, Switzerland
Email: hong-an.cao@inf.ethz.ch

Tri Kurniawan Wijaya, Karl Aberer
Department of Computer Science
EPFL, Switzerland
Email: {tri-kurniawan.wijaya, karl.aberer}@epfl.ch

Nuno Nunes
Madeira Interactive Technologies Institute
Funchal, Portugal
Email: njn@uma.pt

***Abstract*— Attaining energy efficiency requires understanding human behaviors triggering energy consumption within households. In conjunction to providing appliance-level feedback, targeting human activities that involve the usage of electrical appliances can provide a higher abstraction level to bring awareness to the electricity wastage. In this paper, we make use of a large dataset with appliance- and circuit-level power data and provide a framework for determining temporal sequential association rules. Sequences of time intervals where the appliances are in usage can vary in their order, duration and the time elapsed between these events. Our contribution consists in providing a full pipeline for mining frequent sequential itemsets and a novel way to discover the time windows during which these sequences of events occur and to capture their variance in terms of duration and order. Our method is data-driven and relies on the data's statistical properties and allows us to avoid an exhaustive search for the time windows' sizes, by relying instead on machine learning techniques to identify and predict those time windows.**

***Index Terms*—Time series analysis; Data mining; Information search and retrieval; Clustering; Smart energy; Smartmeters; Activity inference; Appliances states; Energy data analytics; Datasets; Algorithms**

I. Introduction

Increasing energy efficiency is part of the goals set by governments across the world to reduce the energy footprint and provide sustainable development to all. The advent of new technologies that permit the monitoring of the electrical consumption within households, such as with smartmeters and future smart appliances that are likely to report their own consumption, and progress in control technologies for actuating different components such as lights or thermometers, offer prospects to smarten homes by exploiting the large amounts of data available and derive processes to conserve energy. The opportunity to collect real-time consumption data prepares us to contemplate real-time feedback to inform the residents about their usage of energy [1]. However, the final link to this chain, from data to action, relies on households' residents to assimilate the feedback and to change their relationship towards their energy consumption. The failure of earlier energy conservation campaigns was due to the discrepancy between the residents' energy knowledge (such as energy units awareness or the evaluation of how much energy an appliance would consume) and the expected energy reduction that utilities were aiming by offering money incentives [2]–[5]. Adequate information has to be provided in assisting the decision making as has been shown in the process of acquiring new appliances to reduce future energy costs [6].

While feedback at the appliance level could be provided, given the unfamiliarity with the energy jargon and the overwhelming occurrences of when diverse appliances are used throughout the week, the residents might not be able to associate the triggering of an appliance to a behavior to address. By aggregating interactions with appliances and abstracting the underlying ongoing activity, the granularity can be reduced. Also, if a resident were to keep a diary of their daily activities, since most of them are essential (e.g. cooking), they would be salient in their memory and thus more easily associated to effective interactions with electrical devices. Beyond identifying and estimating the amount of energy that is used during specific human activities, this additional information could be used to build new strategies within a smart home to improve and offset energy-hungry behaviors by providing automation measures to reduce their footprint. This would first require us to learn what activities can be detected and their scheduling, and more specifically to predict the time windows where they might occur.

Our contribution is to provide temporal sequential association rules in a novel way, based on machine learning techniques, to learn time windows where a rule's head and body take place and to exploit historical data and their statistical properties. Given the variance in the usage of different appliances for completing specific tasks, such as cooking, where the diversity of the recipes in terms of preparation and cooking time contributes to the variance in what appliances and in which order and how long they are used, considering sequential frequent itemsets allows us to capture rules that still reflect the underlying behavior. We provide an analysis on a dataset with disaggregated energy consumption and show that rules can be learned that reflect expected activities that should take place within households. Our technique is not limited to energy data and is thus generalizable to datasets for which temporal sequential rules should be mined. In the following, we will review related work in Section II. Then, we will present the methodology for extracting temporal

978-1-4673-9005-7/16/$31.00 ©2016 IEEE

sequential association rules in Section III. Experimental results on a disaggregated dataset with several households will be evaluated in Section IV. Finally, we will discuss improvements and future work in Section V.

II. Related Work

In the frame of Demand Side Management (DSM), the shifting of select appliances usages and their optimal scheduling to enable the shaving of peak consumption was studied experimentally [7]–[9]. As more smartmeters are rolled out and equipped in households, large datasets with aggregate load consumption are released such as the Irish Commission for Energy Regulations' dataset with over 5000 households or the PG&E dataset, and serve as a basis for research in customer segmentation or demand forecasting for Demand-Response strategies [10], [11]. Due to the difficulty to collect single appliances' power consumption, previous work relied on activities as described by human beings to model and synthesize electrical loads in households [12]. DRSim was developed as a simulator for DSM systems that is aware of the current status of the grid and the activities carried out inside a household and attempts to estimate the potential savings for demand side policies [13].

Until all households are equipped with smart appliances that can communicate their consumption and internal states, determining when an appliance is *active* or *idle* (being powered off or being in standby-mode), requires either datasets with explicit labeling of these states or an algorithm to determine them automatically [14], [15]. The Dataport dataset is one of the largest dataset with 1-minute power data including disaggregated readings from single appliances and circuit (such as power strips or rooms) from over 800 households, but it is still lacking rich metadata such as appliances' states and activity annotations. Efforts were directed at expanding the datasets that are available to the community by learning from the computer vision community and using crowdsourced human labeling to acquire labels for providing richer data to develop and refine algorithms based on machine learning [16]. This retrofitting of existing datasets offers an alternative to acquiring new datasets, which is prohibitive in terms of costs and time [17], as consumer electronics is widespread and requires a large set of smart plugs to be installed and even special instrumentation for larger appliances with higher wattage. Preliminary attempts at inferring activities from households' non-synthetic electric load curves relied on aggregated electricity consumption, with few annotations [18]. Activity detection was previously performed by considering activities as events' streams and using symbolic analysis with the specific goal of shifting activities to a more convenient moment of the day [19] and using an HMM [20], however, both analyses relied on synthetic data. Attempts at using real data are linked to the CASAS project [21]–[25] and set in a students' apartment, the data were used to extract sensor data features to link the aggregate household consumption load to human activities, but failed to address the bias induced by the inability to discard energy-hungry appliances from the overall load curve [24], [26]. Another attempt at using real data was achieved through a push-system for recording user activities based on the identification of interactive loads by clustering the states of appliances [27], but still failing to recognize appliances. This work spawned the analysis of association rules [28], but considering fixed hourly windows, instead of mining for variable time intervals and not considering the time relationships between the time intervals.

The development of APRIORI [29] was followed by different sequential pattern mining algorithms such as GSP [30], WINEPI and MINEPI [31], SPADE [32] or PREFIX-SPAN [33], they provided sequential pattern analysis but considered the events to be instantaneous. Temporal pattern mining progressively included different time relationships between the sequences of events [34]–[37]. A framework for identifying sequential temporal intervals provided an algorithm based on APRIORI for learning the association rules by searching for frequent arrangements of sequences of events, extending and revising Allen's temporal relationships and allowing user-defined constraints for mining the rules, but did not offer the possibility to mine for the time windows during which the temporal association rules occurred [38]. Titarl was developed to learn temporal association rules on symbolic time sequences (where sensor data were binarized by introducing discretization of each variable representing a sensor), but considered uniformly distributed intervals for the time intervals where the rules occurred, instead of exploiting the statistical property of the data [39]. The work was then extended to forecast temporal intervals by providing a refinement procedure for first extracting temporal association rules, then merging them [40].

III. Methodology

If we consider the *cooking* activity, we expect different appliances to be used to fulfill this task such as an oven or a kitchen stove. The triggering of those appliances can then be followed by the usage of a dishwasher for cleaning the dishes. Due to the diversity of recipes that can be used for preparing a meal for example, defining temporal thresholds for the duration of events during which different appliances are used is too restrictive and will not capture the variance in the way corresponding activities are conducted. Thus, to learn human activities triggering electrical consumption, we identify co-occurring events and their respective association rules. We exploit previous work on sequential itemsets mining by considering temporal relationships between the events and their respective time intervals allows us to classify and order these events according to the sequence in which they occur [38]. This means that different events can *follow*, *contain* or *overlap* one another. The succession and merging of these events can be identified as *activities*. Additionally, we define a novel method to derive the time windows where these *activities* arise and learn the association rules between these *activities*. In the following, we describe the methodology for deriving temporal association rules for sequential events such as defined in Equation 1 and through our pipeline as can be seen in Figure 1.

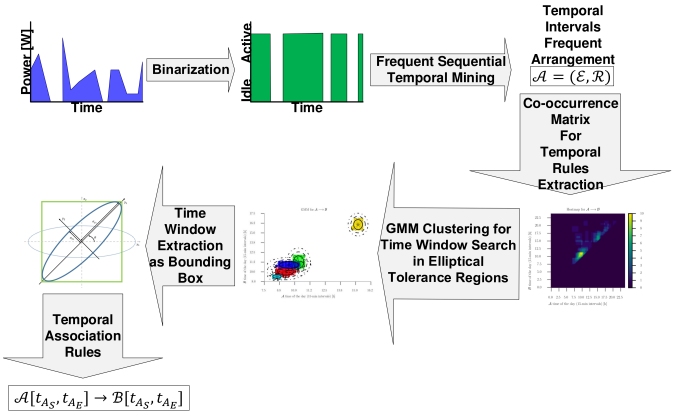


Fig. 1: Pipeline for deriving the temporal association rules

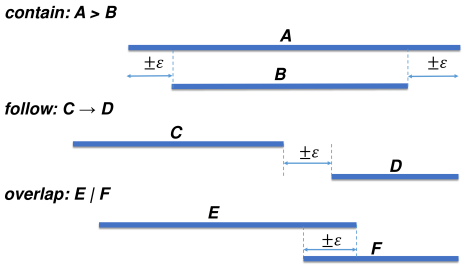


Fig. 2: Time relationships: *contain*, *follow* and *overlap*

A. Data Binarization

In the case of activities, the events consist in the triggering and *active* usage of appliances in residential homes. The data that are recorded consist in power data and do not contain information about when appliances are in usage, instead of being off or in standby-mode. The measurements are converted to a binary form, where *active* and *idle* states are determined using *GMMThresh* [14], [15]. This binarization method can be used for sensor data as well, to distinguish background noise (*idle* state) from meaningful readings (*active* state). For multi-state data, we advise to define a quantification scheme for the original data and to create a variable per quantification level and transform the data accordingly.

B. Sequential Association Rules Mining

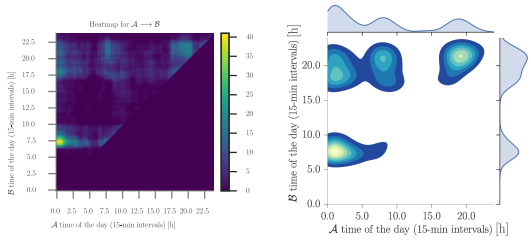
Since we want to learn daily temporal association rules, we can consider that each day of collected data represents a basket, in the traditional market basket analysis [29]. The intervals during which the appliances are *active* represent the items in each basket. Sequential association rules mining has been widely studied and defines how frequent sequential itemsets are extracted [38]. We briefly explain how these association rules are constructed. The events' intervals $\mathcal{E}$ maintain temporal relationships $\mathcal{R}$ and constitute arrangements $\mathcal{A} = (\mathcal{E}, \mathcal{R})$. An arrangement $\mathcal{A}$ defines the temporal relationships between the time intervals where different events take place. For n events E, $\mathcal{E} = (E_1, ..., E_n)$, $P_2^n = \frac{n!}{(n-2)!} = m$ permutations for the pair-wise temporal relationships $R(E_i, E_j)$ between E_i and E_j can be computed , where $i < j$ and $i, j \in \{1, ..., n\}$, thus we can define an m-tuple of pairwise relationships $\mathcal{R} = (R_1, ..., R_m)$. The temporal relationships that are selected in the case of the *activities* constitute a generalization of more refined ones [38] and in this paper, we restrict them to $R \in \{contain, follow, overlap\}$ as in Figure 2. The search for these arrangements is performed on an enumeration tree expanded breadth-wise and being pruned based on a minimum support value. The rules are determined for each arrangement by considering sub-arrangements and by extracting all partitions of the set of events into two subsets as the head and the body of the rule. In order not to repeat rules, the sub-arrangements are extracted in lexicographic order. The rules are accepted or discarded following APRIORI's strategy [38].

$$\mathcal{A}[t_{A_S}, t_{A_E}] \longrightarrow \mathcal{B}[t_{B_S}, t_{B_E}] \tag{1}$$

C. Time Windows

1) Bivariate histograms (or heatmaps): Having determined the body and head of the rules as two sub-arrangements, we derive a novel technique for extracting the time windows during which the rules hold. For this, we build a co-occurrence matrix for the head and the body of each rule based on the time intervals during which they occur. To preserve the order of the rule, i.e. the body and the head respectively, we only consider the cases where the head's time intervals are subsequent to the body's time intervals. In the reverse case, the rule with head and body inverted, should it have enough support, will be processed independently from another arrangement. In practice, we build a bivariate histogram (or heatmap) for each minute in a day for both the head and the body of the rule. For each day, co-occurring minutes for both the head and the body are marked as zones in the bivariate histogram. For example, for a particular day, if the head is active from 10 a.m. to 11 a.m. and the body is active from 11:30 a.m. to 2 p.m., the co-occurring region would be the rectangular area [10:00-11:00; 11:30-14:00], using the 24-hour notation, and would contain ones, while the other regions zeroes. The bivariate histogram is created by super-imposing the different co-occurring regions for each day. As can be seen in Figure 3a, each minute that has occurred more frequently throughout the dataset, will be more accounted for than minutes that have happened more irregularly. It can also be noticed that the matrix is upper triangular, as we are interested in events (the rule's head) appearing after the body's events.

2) Tolerance regions: The regions of interest for determining the time windows for the association rules are the temporal regions that occur the most often. They can be smoothed as can be seen in Figure 3b by using a kernel density estimation. Using the bivariate histogram concept allows us to conceptually assimilate each region as a trivariate Gaussian distribution. The regions of interest are then the projection of each trivariate



(a) Bivariate histogram (heatmap). (b) Gaussian Kernel Density Estimation.

Fig. 3: Heatmap and Gaussian Kernel Density Estimation for an association rule

Gaussian into the horizontal plane. Thus, identifying these regions can be undertaken by a Gaussian Mixture Model, where each Gaussian can represent a separate region or cluster of points. Equation 2 represents a k-dimensional Gaussian distribution, entirely defined by its mean $\vec{\mu}$ and covariance matrix Σ, defined separately in Equation 3. We will describe how to derive those regions in the following.

$$P(\vec{x}|\vec{\mu},\Sigma) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} exp(-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})) \quad (2)$$

$$\Sigma = E[(\vec{x}-\vec{\mu})(\vec{x}-\vec{\mu})^T]$$
$$\text{where } \Sigma_{ij} = cov(x_i, x_j) = E[(x_i-\mu_i)(x_j-\mu_j)]$$

The projection of a trivariate Gaussian on the horizontal plane is a bivariate Gaussian. The region they cover can be delimited by an isocontour [41] as defined in Equation 3. These isocontours can be defined by the mean $\vec{\mu}$ and the covariance matrix Σ of the data points clustered within, where the spread between the data points and the mean is represented by the Mahalanobis distance $d^2_\Sigma(\vec{x},\vec{\mu})$ as defined in Equation 4. We can define from a k-variate Gaussian distribution an ellipsoidal region as in Equation 5 where $\underline{\vec{\mu}}$ and $\underline{\Sigma}$ are the sample mean and the sample covariance matrix, respectively, obtained from the clustered data, where $c \rightarrow \chi^2_k(p)$, the chi-squared distribution with k degrees of freedom and for covering a p-percentage of the population as the population size $N \rightarrow \infty$ [42]. These regions are referred to as statistical tolerance regions [42]–[44]. They have been used in assembly tolerance for specifying the quality of production to be achieved [45].[1] The population coverage is thus a parameter for the size of the tolerance area. A closed form solution was defined for bivariate cases and approximations are available for higher dimensional distributions [42].

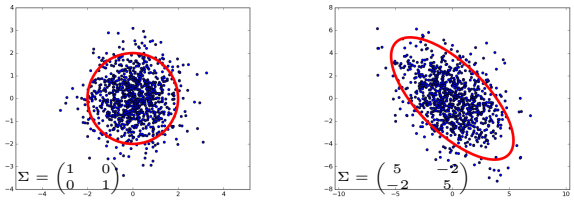
$$P(\vec{x}|\vec{\mu},\Sigma) = c \text{ with } c \in [0;1] \quad (3)$$

$$d^2_\Sigma(\vec{x},\vec{\mu}) = (\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu}) \quad (4)$$

[1]These should not be confused with confidence regions (or intervals), which yield the confidence for the sample mean and covariance matrix, as the experiment is repeated.

$$R(\underline{\vec{\mu}},\underline{\Sigma},c) = \{\vec{x} : d^2_{\underline{\Sigma}}(\vec{x},\underline{\vec{\mu}}) \leq c\} \quad (5)$$

To determine the closed form equation of the ellipsoid defined in Equation 5, we can recall the definition of the covariance matrix Σ as in Equation 3, which summarizes the spread of the data. Such observation is the basis to popular methods such as principal components analysis (PCA), to transform the data into an orthogonal basis set, where the first vector of this basis will have the direction of the largest variance of the data (this consists in performing the eigendecompostion of the covariance matrix $\Sigma = VLV^{-1}$, where L is the diagonal matrix of eigenvalues and V the respective eigenvectors). This change of coordinates operates under a linear transformation T and consists of a rotation through a matrix R and the scaling of the data points along each axis through a matrix S^2 where $T = RS$ [46], as illustrated in Figure 4 and $\Sigma = RSSR^{-1} = TT^T$, with $S = \sqrt{L}$ and $R = \sqrt{V}$ [47], the Cholesky decomposition of Σ. As can be seen in Figure 5, we can bound the ellipse by a box to get the approximation of the time intervals during which the events occur for both the head and the body of a rule, as the sides of the rectangle delimited by the ellipse's extremum points.



(a) The covariance matrix is the iden-(b) The covariance matrix is a full tity matrix, data contained in a circle.matrix. Notice the rotation and the spread of the data into an ellipse.

Fig. 4: Full and diagonal covariance matrices and corresponding data spread

Without loss of generality, if we consider the case of the bivariate Gaussian distribution, we can rewrite the density function as in Equation 7 by taking the covariance matrix Σ as in Equation 6. We can compute the change of coordinates as the linear transformation $Y = TX$. Since the new basis is orthogonal, the covariance matrix in that basis is a diagonal matrix as can be seen in Equation 8 as the variables are uncorrelated, and which can be solved to obtain the covariances $\sigma^2_{y_1}$, $\sigma^2_{y_2}$ and the rotation angle θ as in Equation 9. The tolerance regions, which we are interested in, are delimited by isocontours such that $f(x_1,x_2) = c$, where $c \geq 0$. $a = c\sigma_{y_1}$ is the the semi-major axis of the ellipse and $b = c\sigma_{y_2}$ the semi-minor axis, respectively.

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \text{ with } \rho = \frac{\sigma_{12}}{\sigma_1\sigma_2} \quad (6)$$

[2]The matrix is thus diagonal.

$$f(x_1,x_2) = \frac{1}{2\pi\sigma_{x_1}\sigma_{x_2}\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}[(\frac{x_1-\mu_{x_1}}{\sigma_{x_1}})^2 - 2\rho\frac{x_1-\mu_{x_1}}{\sigma_{x_1}}\frac{x_2-\mu_{x_2}}{\sigma_{x_2}} + (\frac{x_2-\mu_{x_2}}{\sigma_{x_2}})^2]) \quad (7)$$

$$SS = R^{-1}\Sigma R \implies \begin{pmatrix}\sigma_{y_1}^2 & 0\\ 0 & \sigma_{y_1}^2\end{pmatrix} = \begin{pmatrix}\cos(\theta) & \sin(\theta)\\ -\sin(\theta) & cos(\theta)\end{pmatrix}\begin{pmatrix}\sigma_{x_1}^2 & \sigma_{x_1x_2}\\ \sigma_{x_1x_2} & \sigma_{x_2}^2\end{pmatrix}\begin{pmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & cos(\theta)\end{pmatrix} \quad (8)$$

$$\begin{cases}\theta = \frac{1}{2}\arctan\frac{2\sigma_{x_1x_2}}{\sigma_{x_1}^2-\sigma_{x_2}^2}\\ \sigma_{y_1}^2 = \frac{\sigma_{x_1}^2+\sigma_{x_2}^2}{2} + \sqrt{\frac{(\sigma_{x_1}^2-\sigma_{x_2}^2)^2}{4}+\sigma_{x_1x_2}^2}\\ \sigma_{y_2}^2 = \frac{\sigma_{x_1}^2+\sigma_{x_2}^2}{2} - \sqrt{\frac{(\sigma_{x_1}^2-\sigma_{x_2}^2)^2}{4}+\sigma_{x_1x_2}^2}\end{cases} \quad (9)$$

Additionally, the size of the tolerance region also bears a statistical meaning that determines the value c. Indeed, the Mahalanobis distance r to the Gaussian can set the ellipse's size as for the bivariate case, it is dependent on the cumulative distribution of the Gaussian distribution. A closed form solution based on the cumulative distribution function as expressed in Equation 10, based on the parametrization of the ellipse and its Cholesky decomposition [48] and allows us to determine r based on the probability that an observation falls within the region delimited by the ellipse defined by the isocontour at value c. Additionally, the new coordinate system has an orthogonal basis and the variables y_1 and y_2 are thus uncorrelated and $\rho = 0$ and the ellipse's equation can be rewritten as $\left(\frac{y_1-\mu_1}{\sigma_{y_1}}\right)^2 + \left(\frac{y_2-\mu_2}{\sigma_{y_2}}\right)^2 \leq c^2$. Each term $\frac{y_1-\mu_i}{\sigma_{y_i}} \sim Z_i = \mathcal{N}(0,1)$ contributes as an i.i.d standard distribution, and is therefore equivalent to a chi-square distribution ($U^2 = \sum_{i=1}^{k} Z_i \sim \chi_k^2$) with k degrees of freedom. The isocontour can thus be computed for a specific proportion p of the population to be covered by the tolerance region as $c = \chi_2^2(p)$ or $c = \sqrt{-2\ln(1-p)}$, equivalently.

$$\begin{aligned} F(r) &= 1 - exp(-\frac{r^2}{2}) = p\\ r &= F^{-1}(p) = \sqrt{-2\ln(1-p)}\end{aligned} \quad (10)$$

To define the bounding box to the tolerance ellipse, we use the general form of the parametric equations of an ellipse as in Equation 11, obtained by rotating the polar coordinates of an ellipse through the rotation matrix R. The bounding box is delimited by lines passing through the extremum points of the ellipse and can thus be obtained by taking the partial derivatives of the general parametric equations as in Equation 12 to obtain the values t that should be set in the parametric equation.

$$\begin{cases}y_1 = \mu_1 + a\cos(t)\cos(\theta) - b\sin(t)\sin(\theta)\\ y_2 = \mu_2 + a\cos(t)\sin(\theta) + b\sin(t)\cos(\theta)\end{cases} \quad (11)$$

$$\begin{cases}\frac{\partial y_1}{\partial t} = 0 \iff t = -\frac{b}{a}\tan(\theta)\\ \frac{\partial y_2}{\partial t} = 0 \iff t = \frac{b}{a}\cot(\theta)\end{cases} \quad (12)$$

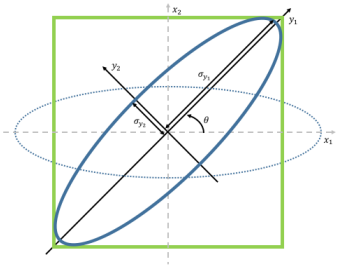


Fig. 5: Ellipse rotation and bounding box

Having derived how to obtain the bounding boxes for the time windows prediction, we use it in conjunction with a Gaussian Mixture Model that will identify clusters of data points. If this is successful, a temporal sequential rule as defined in Equation 1 is added to the set of association rules, if not it is discarded. If no rule can be determined for the current itemset, the node is discarded and is not expanded further. Our platform also includes constraints developed for an optimistic pruning of the frequent sequential itemsets [38], such as duration constraints for each arrangement or an ϵ time tolerance for the temporal relationships between the intervals representing the different variables (appliances) that are considered. It is also easily adapted to enforce constraints on the time windows for the predictions such that user-defined time of the day or durations δ as described in Equation 13 can prune out less relevant rules. The cases described in Equation 13 can be generalized to our general formulation in Equation 1 to perform an exhaustive search for all temporal sequential association rules.

$$\begin{aligned}\mathcal{A}\{t_1\} &\longrightarrow \mathcal{B}\{t_2\}\\ \mathcal{A}\{t_1\} &\longrightarrow \mathcal{B}[t_2,t_3]\\ \mathcal{A}\{t_1\} &\longrightarrow \mathcal{B}\{t_1+\delta\}\end{aligned} \quad (13)$$

IV. Empirical Evaluations

A. Dataset

We use the Dataport dataset, with data ranging from July 2012 until April 2015. The dataset contains 1-minute appliance-level (washing machines, ovens, etc.) and circuit-level (rooms, multiplugs for small appliances in the kitchen, etc.) power data for over 70 types of meters and more than 800 households located mainly in Texas and in California. We select 16 households with large numbers of appliances. The data are cleaned to contain only full days (discarding missing data and daylight saving time days). The measurements are binarized using *GMMThresh* [14], [15], which distinguishes when an appliance is *active* and thus triggered on to serve an activity, from when it is *idle*, being either off or in stand-by mode. Without loss of generality, we select January data for deriving the temporal association rules, with some households having 1, 2 or 3 months worth of data for that specific month. We remove appliances that are likely to always be on or exhibit a periodic behavior due to being controlled by

TABLE I: Households data details, with number of appliances per month for each household id

Household id	624	1464	1632	2018	2472	2974	3615	5568	6139	6348	6378	7510	7982	9776	9922	9926
2013-01	13	17	0	0	0	13	0	9	9	0	0	15	0	0	15	15
2014-01	12	15	14	14	15	11	22	15	14	0	17	12	15	15	15	12
2015-01	14	0	12	14	15	15	9	14	14	18	13	14	0	15	15	13

a timer or a thermostat, such as fridges, freezers, furnaces or air conditioning units. The selected households and the details about how many months of data and how many appliances are considered can be seen in Table I.

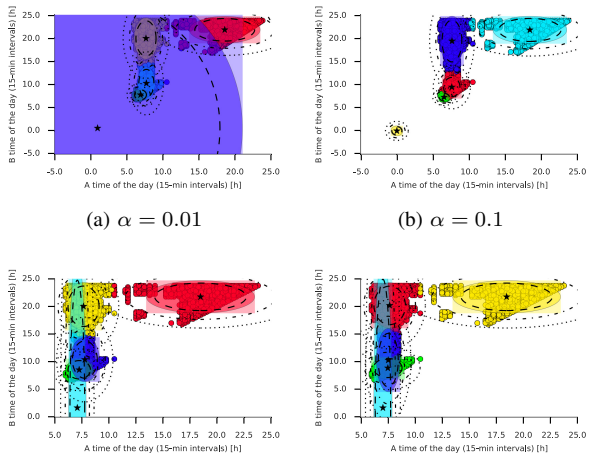
The scheduling and duration of usage of different appliances is expected to vary significantly. We are looking for general rules and this can be achieved by relaxing the conditions for the time and the duration of different events and instead considering sequences of events. While some activities such as textile care, would mostly have the washing machine first enabled, then would be followed by the dryer, activities such as cooking are less likely to preserve the order of different appliances as cooking relies on very different recipes, involving different subsets of appliances and durations for each instance. Thus, if we are interested in the sequence in which different groups of appliances are used, quantifying their total duration of usage and the exact time window during which they are triggered on is likely to fail. Instead, we are searching for intervals where the appliance is used and the time relationships between these intervals, accounting for some flexibility on the intervals' bounds. However, due to the fact that some appliances can be used for a very short time, while others are active for longer periods, we take the precaution to downscale the data to improve the detection of the time relationships between the time intervals where they take place. From the 1-minute data granularity we construct 15-minute intervals. We mine for the top 1000 (which means we can get more temporal rules as their number depends on the number of clusters that are detected) rules extracted from the arrangements in each household, they replace rules with lower scores.

B. Support and Interestingness Measures

Additional parameters can influence the search for the temporal sequential association rules platform, such as selecting the support threshold for the frequent itemsets filtering. The interestingness measures used for determining the association rules and the minimum thresholds for discarding or accepting them are quite diverse in the litterature. Two well-known measures are the *support* defined as $supp(X) = \frac{|\{t \in D; X \subseteq t\}|}{|D|} = P(X)$ and the *confidence* as $conf(X \implies Y) = \frac{supp(X \cup Y)}{supp(X)} = P(Y|X)$ [29]. The tolerance for the time relationships is represented by an ϵ slack on the bounderies of the intervals. Then, for determining the time windows, we choose a threshold for the bivariate histogram as a minimum support for how often each minute should have been marked as occurring, this allows us to discard noise and is similar to the support filtering when mining for the frequent itemsets.

C. GMM

The type of GMM method, i.e. a standard GMM, a GMM based on variational inference (VBGMM) and its infinite GMM counterpart based on a Dirichlet Process (DPGMM) influences the quality of the clustering and the proportion of the population that should be covered by the tolerance region impacts the size of the ellipse for the windows' prediction. The DPGMM and the VBGMM rely on a concentration parameter α as a DP can be described by a Chinese restaurant process where α is proportional to the probability to join a new table [49], [50]. A larger α will influence the clustering by assigning the data to more clusters. To approach the natural number of clusters in the data, we set α to the proportion $\alpha = \frac{\#days}{\#datapoints}$, which usually oscillates between 0.1 and 0.01 and as can be seen in Figure 6, the more natural clustering is achieved for $\alpha = 0.1$.



(a) $\alpha = 0.01$ (b) $\alpha = 0.1$

(c) $\alpha = 1$ (d) $\alpha = 10$

Fig. 6: Impact of the selection of the concentration factor α for the DPGMM. In Figures 6a, 6b, 6c, and 6d, α takes the values 0.01, 0.1, 1, and 10 respectively.

As can be seen in Figures 7 and 8, the quality of the predictions depends on how many clusters are detected and how precise the tolerance regions are. The DPGMM and VBGMM clustering methods overgeneralize the clustering, by merging smaller clusters into larger ones, often covering areas as large as a whole day. This in turn creates very large time windows. The choice of a full covariance matrix instead of a diagonal matrix also impacts the prediction as it will overfit the tolerance regions more and create larger time windows especially in the case of the DPGMM and VBGMM in Figure 8 as the estimated regions of interest are long tilted lines,

although the Kernel Density Estimation indicates large zones spawn by different Gaussians. As can be seen in both cases, the Kernel Density Estimation assigns similar concentrated regions as the GMM, which instead fits the data better. This is why we will consider the diagonal covariance matrices in more details. All evaluated parameters can be found in Table II.

D. Results

We summarize in Table II the total number of rules for all households. The number of association rules is the same for all methods depending on the chosen interestingness measure, as only the arrangements with enough support or confidence are selected, which guarantees enough data for the clustering. The number of distinct temporal rules varies based on the interestingness measure: the confidence creates significantly more rules due to its definition. Additionally, the number of distinct temporal rules also varies across the diverse clustering methods, due to the overgeneralization of DPGMM and VBGMM, which often creates tolerance regions covering the whole day. The threshold for the bivariate histogram should be chosen as a proportion on the dataset instead of an absolute value to reduce the noise incurred by the variance. Due to the type of appliances and circuits monitored in the dataset, most activities cannot be described in a detailed way.

As for the sequences of appliances or circuits that are aggregated in the frequent itemsets and eventually split into rules, we notice that interesting rules are inherent to the number of appliances that are available in each household and not so much about the number of days available for the clustering, as we obtained on average per number of months of data available per household about the same number of rules discovered. The configuration we tested was for a small tolerance to intervals' relationships misalignment by selecting an ϵ of 1 (15 minutes) and thus we are identifying itemsets where appliances are being used relatively closely in time, which is fitting for cooking for example. The parameter could be adapted to relax the constraint for time separation and to capture appliances' usages more disconnected in time.

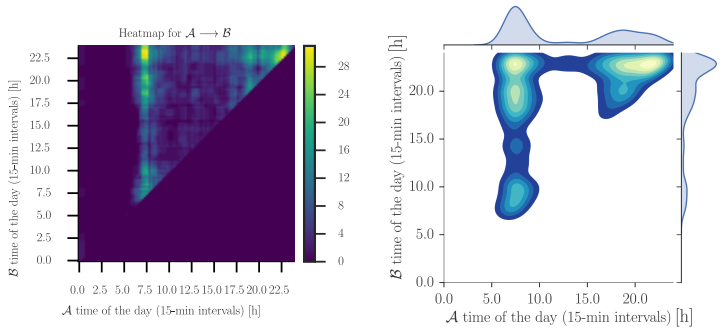
The number of interesting rules is also dependent on whether or not the household residents were at home and actually interacted with the devices and circuits, as some households that only had one month of data showed significantly more rules than households with three months of data. This is due to the fact that single appliances were aggregated such as kitchen appliances (that could contain toasters, coffee makers, etc.) and some houses were more instrumented than others and larger appliances such as ovens, cooking ranges, dishwashers or clothes washers are measured separately, but not available in all households.

We observe that appliances that are linked to cooking are identified in rules such as in Figure 9. Similar rules link ovens to ranges, or kitchen appliances to microwaves. Side interactions with cooking can be detected such as activities in bedrooms, bathrooms or living rooms. We also notice that the usage of cooking appliances can be preceded or followed by the usage of a dishwasher for cleaning the dishes. Additionally, dishwashers or clothes washers are used in conjunction with water heaters, triggered for warming the water (as it is common for such appliances in the US, where the appliances are connected to external cold and hot water sources). In some cases, we could suppose that the residents were preparing to leave as interactions with kitchen appliances were followed by activity in the garage, and conversely, the arrival of the residents could be detected as well. We verified with the surveys supplied with the Dataport dataset for some of the households and the rules that were mined were correlated with the residents' declarations about the rate of usage of different appliances (1-3 times per week), which influences how many rules can be detected. Additionally, in households where residents mentioned that they sometimes work at home during the week, more rules could be observed. The number of residents per dwelling also influenced the types of rules that could be discovered, due to having noisier rules due to activities being carried out by different people concurrently. However the time windows during which rules were discovered are meaningful when corroborated with the survey information and times where people can be expected to be at home.

V. Conclusions and Future Work

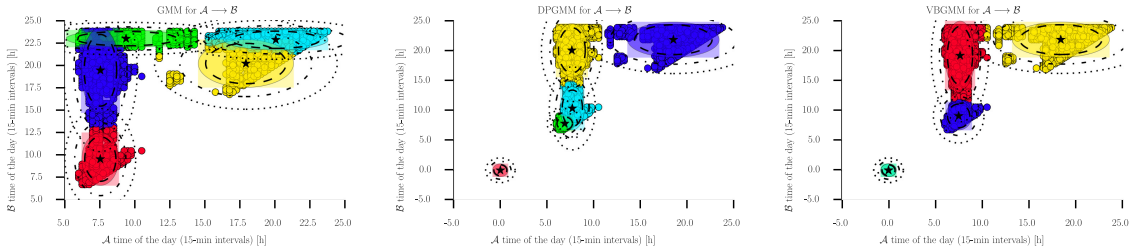
We have derived time windows for temporal sequential association rules based on the co-occurrence of time intervals through machine learning techniques. Our novel method uses the statistical properties of the data to efficiently identify time windows without having to perform an exhaustive search for their occurrence and duration. It is based on the co-occurrence of arrangements of sequential events that can be seen as a bivariate histogram (or heatmap), which can be adjusted to guarantee that events arise in a significant enough proportion by applying a support threshold for the co-occurrence matrix. Using a threshold on the co-occurrence frequency allows us to eliminate the noise from the variation of the time intervals and serves as the support filtering in the APRIORI algorithm when computing the frequent itemsets in a transaction. Each zone having strong co-occurrence can be approximated by a trivariate Gaussian distribution. We treat the planar projection of each Gaussian as a tolerance region, where a percentage of the population can be covered. As such, each region is an ellipse, whose area can be adjusted to the probability to cover a certain percentage of the population. The Gaussians are discovered by estimating them by a Gaussian Mixture Model, whose parameters can be used to determine the ellipses. Events that occur more often are concentrated in different temporal regions, this is captured by the clustering and the spread of the points around the mean of the Gaussians. The rules can be refined by adding more constraints on how the frequent itemsets are constructed (relaxing the time relationships) and the search can be parametrized.

Our predictions can be improved by selecting the features before computing the frequent itemsets selections (by observing the correlation, auto-correlation with time lags). But also,



(a) Bivariate histogram (heatmap).

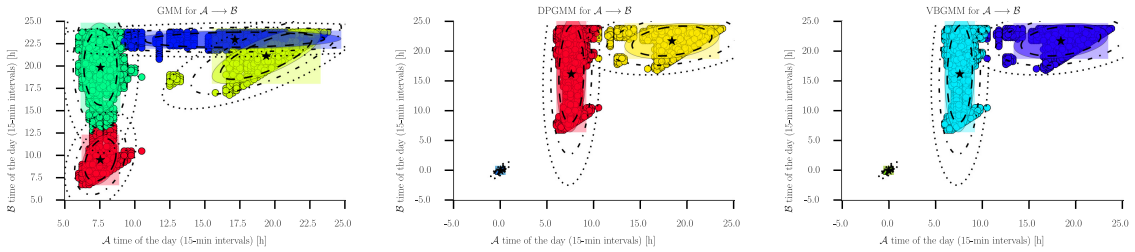
(b) KDE for the heatmap.



(c) Elliptical tolerance regions, GMM, diagonal covariance matrix Σ.

(d) Elliptical tolerance regions, DPGMM, diagonal covariance matrix Σ.

(e) Elliptical tolerance regions, VBGMM, diagonal covariance matrix Σ.



(f) Elliptical tolerance regions, GMM, full covariance matrix Σ.

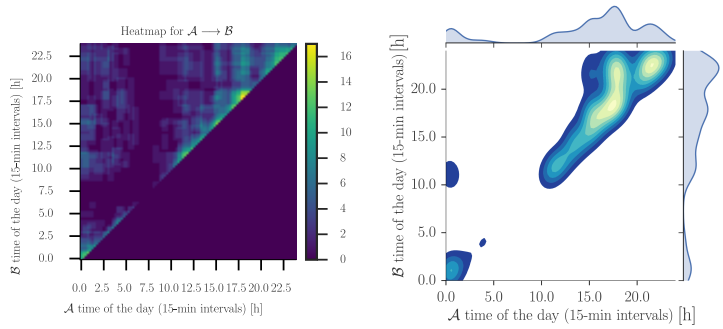
(g) Elliptical tolerance regions, DPGMM, full covariance matrix Σ.

(h) Elliptical tolerance regions, VBGMM, full covariance matrix Σ.

Fig. 7: Bivariate histogram and tolerance regions for household 624 where DPGMM overgeneralizes. In Figures 7c, 7d, 7e, 7f, 7g, and 7h, the $\star$ locates the center of the ellipse (as $\vec{\mu}$, the Gaussian's mean). The dashed, dash-dotted and dotted lines represent population coverage percentages at 1, 2 and 3 standard deviations σ_i from the means μ_i respectively. The colored tolerance regions show the ellipses and its rectangular bounding boxes.

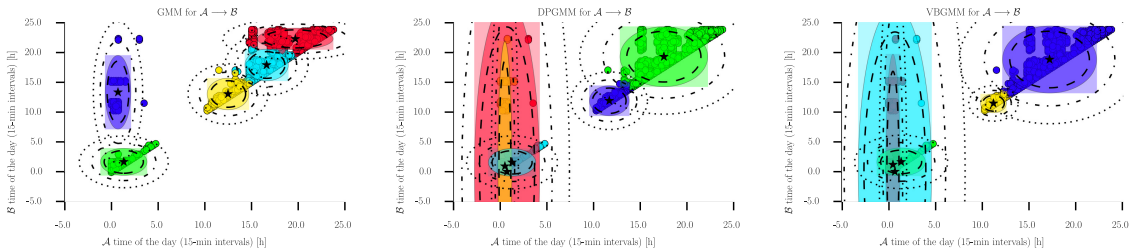
TABLE II: Parametrization for the temporal sequential association rules and results

GMM	Covar.	Freq. Supp.	Interestingness	Min Score	Prob. Ellipse	Time Supp.	Total Nb. Rules	Total Nb. Temp Rules
DPGMM	diag	0.1	confidence	0.4	0.8	5	14214	55705
VBGMM	diag	0.1	confidence	0.4	0.8	5	14214	52840
GMM	diag	0.1	confidence	0.4	0.8	5	14214	68493
DPGMM	diag	0.1	support	0.4	0.8	5	8173	35048
VBGMM	diag	0.1	support	0.4	0.8	5	8173	33140
GMM	diag	0.1	support	0.4	0.8	5	8173	40079
DPGMM	full	0.1	confidence	0.4	0.8	5	14214	45783
VBGMM	full	0.1	confidence	0.4	0.8	5	14214	46513
GMM	full	0.1	confidence	0.4	0.8	5	14214	67769
DPGMM	full	0.1	support	0.4	0.8	5	8173	28876
VBGMM	full	0.1	support	0.4	0.8	5	8173	28674
GMM	full	0.1	support	0.4	0.8	5	8173	39859



(a) Bivariate histogram (heatmap).

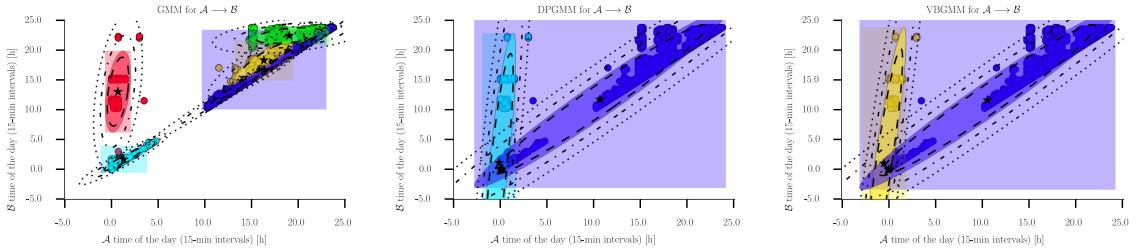
(b) KDE for the heatmap.



(c) Elliptical tolerance regions, GMM, diagonal covariance matrix Σ.

(d) Elliptical tolerance regions, DPGMM, diagonal covariance matrix Σ.

(e) Elliptical tolerance regions, VBGMM, diagonal covariance matrix Σ.



(f) Elliptical tolerance regions, GMM, full covariance matrix Σ.

(g) Elliptical tolerance regions, DPGMM, full covariance matrix Σ.

(h) Elliptical tolerance regions, VBGMM, full covariance matrix Σ.

Fig. 8: Bivariate histogram and tolerance regions for household 2974, where DPGMM and VBGMM overgeneralize and fail to capture smaller clusters. In Figures 8c, 8d, 8e, 8f, 8g and 8h, the $\star$ locates the center of the ellipse (as $\vec{\mu}$, the Gaussian's mean). The dashed, dash-dotted and dotted lines represent population coverage percentages at 1, 2 and 3 standard deviations σ_i from the means μ_i respectively. The colored tolerance regions show the ellipses and its rectangular bounding boxes.

Fig. 9: Kitchen and dishwasher rule, support: 0.548

we plan on collecting ground truth data and applying our analysis on a dataset with more appliances and activity labels for validation. Additionally, the temporal pattern analysis can be extended to accommodate different granularities, e.g., weekly rules can be mined by changing the data format to weekly data and thus deliver intervals across weeks instead of days. Our method is generalizable and can be applied to different datasets with time series to learn habits (mobility traces, smartphones' interaction, etc.) through temporal sequential rules and to use the time windows for scheduling and predictions.

References

[1] K. Carrie Armel, A. Gupta, G. Shrimali, and A. Albert, "Is Disaggregation the Holy Grail of Energy Efficiency? The Case of Electricity," *Energy Policy*, vol. 52, pp. 213–234, Jan 2013.

[2] W. Kempton and L. Montgomery, "Folk Quantification of Energy," *Energy*, vol. 7, no. 10, pp. 817–827, Oct 1982.

[3] W.-J. Schmidt-Küster, "Einfluß des Verbraucherverhaltens auf den Energiebedarf Privater Haushalte," in *Einfluß des Verbraucherverhaltens*

auf den Energiebedarf Privater Haushalte. Munich, Germany: Springer, 1982, vol. 15, pp. 3–6.

[4] J. Froehlich, K. Everitt, J. Fogarty, S. Patel, and J. Landay, "Sensing Opportunities for Personalized Feedback Technology to Reduce Consumption," in *Proc. CHI '09*. Boston, MA, USA: ACM, Apr 2009, pp. 1–8.

[5] J. Froehlich, L. Findlater, and J. Landay, "The Design of Eco-feedback Technology," in *Proc. CHI '10*. Atlanta, GA, USA: ACM, Apr 2010, pp. 1999–2008.

[6] J. Blasch, N. Kumar, M. Filippini, J. Blasch, N. Kumar, and M. Filippini, "Boundedly Rational Consumers, Energy and Investment Literacy, and the Display of Information on Household Appliances," Jun 2016.

[7] S. Barker, A. Mishra, D. Irwin, and E. Cecchet, "Smart*: An Open Data Set and Tools for Enabling Research in Sustainable Homes," in *Proc. SustKDD '12*. Beijing, China: ACM, Aug 2012, pp. 1–6.

[8] S. Barker, A. Mishra, D. Irwin, P. Shenoy, and J. Albrecht, "SmartCap: Flattening Peak Electricity Demand in Smart Homes," in *Proceedings of the 2012 IEEE International Conference on Pervasive Computing and Communications (Pervasive '12)*. Lugano, Switzerland: IEEE, Mar 2012, pp. 67–75.

[9] G. T. Costanzo, A. M. Kosek, G. Zhu, L. Ferrarini, M. F. Anjos, and G. Savard, "An Experimental Study on Load-peak Shaving in Smart Homes by Means of Online Admission Control," in *Proc. ISGT Europe '12*. Berlin, Germany: IEEE, Oct 2012, pp. 1–8.

[10] A. Albert and R. Rajagopal, "Building Dynamic Thermal Profiles of Energy Consumption for Individuals and Neighborhoods," in *Proc. BigData '13*. Santa Clara, CA, USA: IEEE, Oct 2013, pp. 723–728.

[11] T. K. Wijaya, M. Vasirani, S. Humeau, and K. Aberer, "Cluster-based Aggregate Forecasting for Residential Electricity Demand Using Smart Meter Data," in *Proc. BigData '15*. Santa Clara, CA, USA: IEEE, Oct 2015, pp. 879–887.

[12] I. Richardson, M. Thomson, D. Infield, and C. Clifford, "Domestic Electricity Ase: A High-resolution Energy Demand Model," *Energy and Buildings*, vol. 42, no. 10, pp. 1878–1887, Oct 2010.

[13] T. K. Wijaya, D. Banerjee, T. Ganu, D. Chakraborty, S. Battacharya, T. Papaioannou, D. P. Seetharam, and K. Aberer, "DRSim: A Cyber Physical Simulator for Demand Response Systems," in *Proc. SmartGridComm '13*. Vancouver, BC, Canada: IEEE, Nov 2013, pp. 217–222.

[14] H.-Â. Cao, T. K. Wijaya, and K. Aberer, "Estimating Human Interactions with Electrical Appliances for Activity-based Energy Savings Recommendations," in *Proc. BuildSys '14*. Memphis, TN, USA: ACM, Nov 2014, pp. 206–207.

[15] ——, "Estimating Human Interactions With Electrical Appliances for Activity-based Energy Savings Recommendations," in *Proc. BigData '16*. Washington, DC, USA: IEEE, Dec 2016.

[16] H.-Â. Cao, T. K. Wijaya, K. Aberer, and N. Nunes, "A Collaborative Framework for Annotating Energy Datasets," in *Proc. BigData '15*. Santa Clara, CA, USA: IEEE, Oct 2015, pp. 2716–2725.

[17] J. Feminella, D. Pisharoty, and K. Whitehouse, "Piloteur: A Lightweight Platform for Pilot Studies of Smart Homes," in *Proc. BuildSys '14*. Memphis, TN, USA: ACM, Nov 2014, pp. 110–119.

[18] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin, "Private Memoirs of a Smart Meter," in *Proc. BuildSys '10*. Toronto, ON, Canada: ACM, Nov 2010, pp. 61–66.

[19] P. Cottone, S. Gaglio, G. Lo Re, and M. Ortolani, "User Activity Recognition for Energy Saving in Smart Homes," *Pervasive and Mobile Computing*, vol. 16, no. Part A, pp. 156–170, Jan 2015.

[20] F. J. Ordonez, G. Englebienne, P. de Toledo, T. van Kasteren, A. Sanchis, and B. Krose, "In-home Activity Recognition: Bayesian Inference for Hidden Markov Models," *IEEE Pervasive Computing*, vol. 13, no. 3, pp. 67–75, Jul 2014.

[21] D. J. Cook and M. Schmitter-Edgecombe, "Assessing the Quality of Activities in a Smart Environment," *Methods of Information in Medicine*, vol. 48, no. 5, pp. 480–485, May 2009.

[22] D. Cook, "Learning Setting-generalized Activity Models for Smart Spaces," *IEEE Intelligent Systems*, vol. 27, no. 1, pp. 32–38, Jan 2012.

[23] P. Rashidi and D. J. Cook, "Activity Knowledge Transfer in Smart Environments," *Pervasive and Mobile Computing*, vol. 7, no. 3, pp. 331–343, Jun 2011.

[24] C. Chen and D. J. Cook, "Behavior-based Home Energy Prediction," in *Proc. IE '12*. Guanajuato, Mexico: IEEE, Jun 2012, pp. 57–63.

[25] N. C. Krishnan and D. J. Cook, "Activity Recognition on Streaming Sensor Data," *Pervasive and Mobile Computing*, vol. 10, no. PART B, pp. 138–154, Feb 2014.

[26] C. Chen, B. Das, and D. J. Cook, "Energy Prediction Based on Resident's Activity," in *Proc. SensorKDD '10*. Washington, DC, USA: ACM, Jul 2010, pp. 1–7.

[27] S. Rollins, N. Banerjee, L. Choudhury, and D. Lachut, "A System for Collecting Activity Annotations for Home Energy Management," *Pervasive and Mobile Computing*, vol. 15, pp. 153–165, Dec 2014.

[28] S. Rollins and N. Banerjee, "Using Rule Mining to Understand Appliance Energy Consumption Patterns," in *Proc. PerCom '14*. Budapest, Hungary: IEEE, Mar 2014, pp. 29–37.

[29] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases," in *Proc. SIGMOD '93*. Washington, DC, USA: ACM, May 1993, pp. 207–216.

[30] R. Agrawal and R. Srikant, "Mining Sequential Patterns: Generalizations and Performance Improvements," in *Proc. ICDE '95*. Taipei, Taiwan: IEEE, Mar 1995, pp. 3–14.

[31] H. Mannila, H. Toivonen, and A. Verkamo, "Discovery of Frequent Episodes in Event Sequences," *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 259–290, Sep 1997.

[32] M. J. Zaki, "Efficient Enumeration of Frequent Sequences Mohammed," in *Proc. CIKM '98*. Washington, DC, USA: ACM, Nov 1998, pp. 68–75.

[33] Jian Pei, Jiawei Han, B. Mortazavi-Asl, Jianyong Wang, H. Pinto, Qiming Chen, U. Dayal, and Mei-Chun Hsu, "Mining Sequential Patterns by Pattern-growth: The PrefixSpan Approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1424–1440, Nov 2004.

[34] R. Villafane, K. A. Hua, D. Tran, and B. Maulik, "Knowledge Discovery from Series of Interval Events," *Journal of Intelligent Information Systems*, vol. 15, no. 1, pp. 71–89, Jul 2000.

[35] F. Mörchen, "Unsupervised Pattern Mining From Symbolic Temporal Data," *ACM SIGKDD Explorations Newsletter*, vol. 9, no. 1, pp. 41–55, Jun 2007.

[36] F. Höppner and F. Klawonn, "Finding Informative Rules in Interval Sequences," in *Proc. DA '01*. Cascais, Portugal: Springer, Sep 2001, pp. 125–134.

[37] F. Höppner, "Discovery of Temporal Patterns," in *Proc. PKDD '01*. Freiburg, Germany: Springer, Sep 2001, pp. 192–203.

[38] P. Papapetrou, G. Kollios, S. Sclaroff, and D. Gunopulos, "Mining Frequent Arrangements of Temporal Intervals," *Knowledge and Information Systems*, vol. 21, no. 2, pp. 133–171, Nov 2009.

[39] M. Guillame-Bert and J. L. Crowley, "Learning Temporal Association Rules on Symbolic Time Sequences," in *Proc. ACML '12*. Singapore: JMLR, Nov 2012, pp. 159–174.

[40] M. Guillame-Bert and A. Dubrawski, "Learning Temporal Rules to Forecast Events in Multivariate Time Sequences," in *Proc. NIPS '14*. Montreal, QC, Canada: Curran Associates, Dec 2014, pp. 1–9.

[41] C. B. Do, "The Multivariate Gaussian Distribution," 2008. [Online]. Available: http://cs229.stanford.edu/section/gaussians.pdf

[42] M. Siotani, "Tolerance Regions for a Multivariate Normal Population," *Annals of the Institute of Statistical Mathematics*, vol. 16, no. 1, pp. 135–153, Dec 1964.

[43] K. Krishnamoorthy and S. Mondal, "Improved Tolerance Factors for Multivariate Normal Distributions," *Communications in Statistics - Simulation and Computation*, vol. 35, no. 2, pp. 461–478, Feb 2006.

[44] K. Krishnamoorthy and T. Mathew, *Statistical Tolerance Regions: Theory, Applications, and Computation*. Wiley, 2009.

[45] M. J. Law, "Multivariate Statistical Analysis of Assembly Tolerance Specifications," Master Thesis, Brigham Young University, 1996.

[46] D. C. Lay, *Linear Algebra and Its Applications*. Pearson, 2012.

[47] V. Spruyt, "A Geometric Interpretation of the Covariance Matrix," 2014. [Online]. Available: http://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix/

[48] M. Bensimhoun, "N-dimensional Cumulative Function, and Other Useful Facts About Gaussians and Normal Densities," Jerusalem, Israel, Tech. Rep., 2009.

[49] D. M. Blei and M. I. Jordan, "Variational Inference for Dirichlet Process Mixtures," *Bayesian Analysis*, vol. 1, no. 1, pp. 121–143, Mar 2006.

[50] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies," *Journal of the ACM*, vol. 57, no. 2, pp. 1–30, Jan 2010.