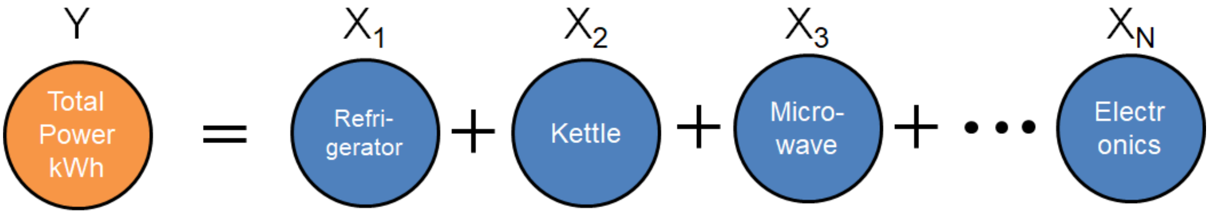
1 Month Progress Report
15 Feb 2018

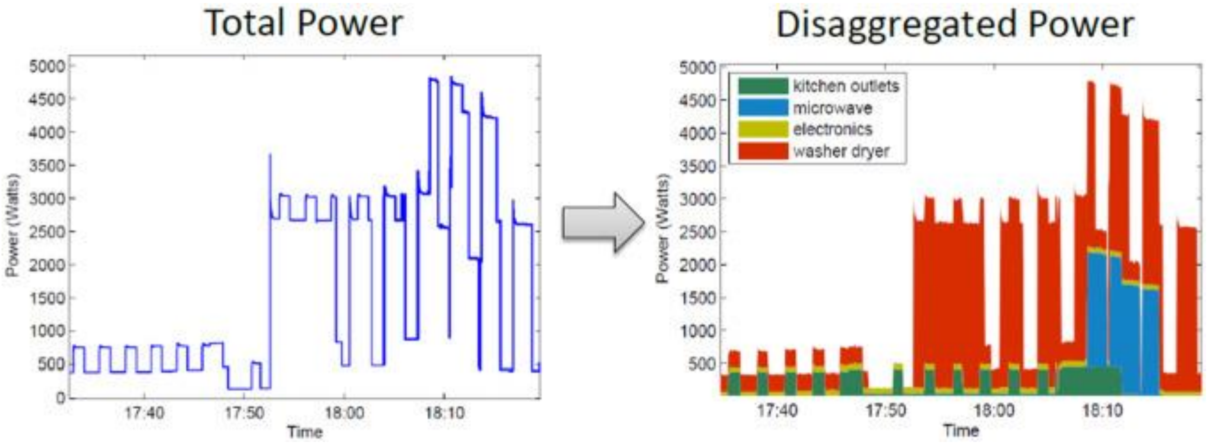
Non-Intrusive Load Monitoring (NILM)

Non-Intrusive Load Monitoring (NILM) is an algorithm to disaggregate the power from the household meter into appliance energy information.

The goal in the NILM is,
Given the number of appliances in a household (N), the power consumed in a household (Y) from each appliance (X_n) for a time period (t) will be aggregated at meter as,

$$Y(t) = X_1(t) + X_2(t) + \cdots + X_N(t)$$





Appliances

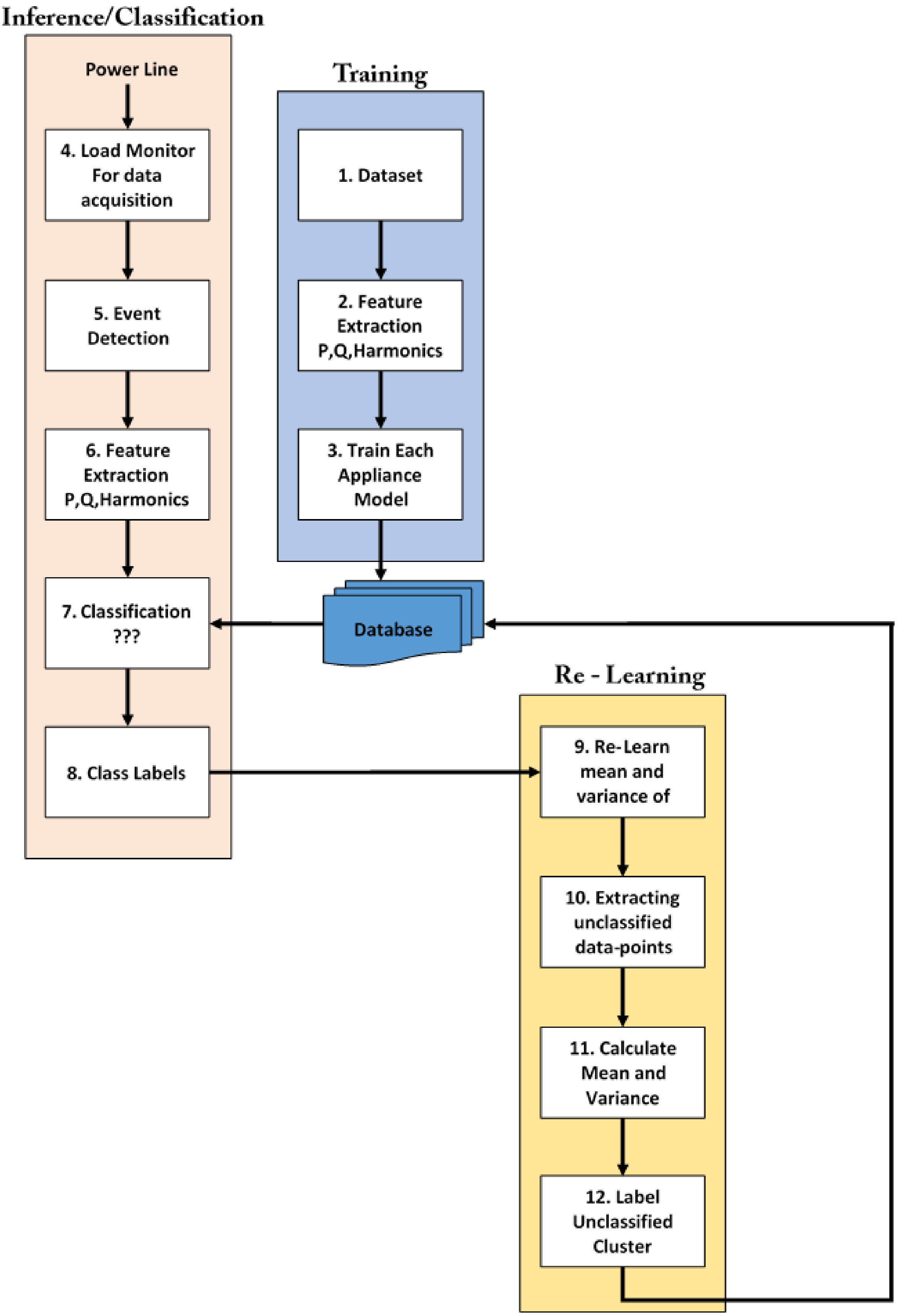
The appliance energy consumption (X_n) is unknown independent random variable.
X_n have characteristic distribution specific to appliance type, but then general enough to accommodate all the appliance of the same type.
For example

- A laptop from different manufacturer can have variation in load consumption.
- Air conditioner will have different load characteristics based on tonnage, area of cooling and temperature setting
-

As mentioned in multiple research, there are mostly 4 types of appliance in a residential setting.

- **Simple ON/OFF** – Eg. Table Lamp, Toaster
- **Multi State Appliance** – Appliance with finite number of operations Eg. Washing Machine, Refrigerator
- **Continuously Variable Devices** – The devices that have no fixed state, then change their load consumption based upon the usage. Eg Light Dimmer, Power Drill
- **Always ON** – Appliance that are always ON hence they cannot be detected in event based algorithm. Eg Smoke Detectors, Alarm clocks.

Proposed NILM Algorithm


Inference/Classification
Power Line
4. Load Monitor For data acquisition
5. Event Detection
6. Feature Extraction P,Q,Harmonics
7. Classification ???
8. Class Labels
Training
1. Dataset
2. Feature Extraction P,Q,Harmonics
3. Train Each Appliance Model
Database
Re - Learning
9. Re-Learn mean and variance of
10. Extracting unclassified data-points
11. Calculate Mean and Variance
12. Label Unclassified Cluster

Reason behind the Proposed Algorithm

Usually the NILM research does not focus on updating the data-base based on which the classification is done. There is an assumption that the limited dataset available for training represent all the appliance in the category. But that is not true in real-world-scenario. Each specific appliance in the same category can have different features. For example, a refrigerator from Samsung will have different Power rating then refrigerator from LG.

Another assumption that the dataset covers all the appliances that can be present in a household. But, that is also not true. A typical household can have up to 20-25 different appliances, and to create a database for all the appliances is difficult.

Then, if we can create a database, from a general appliance model, that can learn while doing the classification specific to the household, then we can have better prediction accuracy and overall estimation of power for that specific household.

The proposed algorithm tries to solve all of the above stated issues in current NILM algorithms.

The proposed algorithm covers 12 steps that are divided into three sections-
Training (Offline)
Inference/Classification (Online)
Re-Learning (Offline)

Inference methods

For the purpose of Load disaggregation, there are many different algorithm present, mostly divided across on the basis of dataset present for learning. For low frequency data, the most common approach is to use HMM or its variants.
The HMM models I checked (and understood), the usual algorithm is –
Create the models using the Baum Welch Algorithm, to learn the parameters. This is to create the general appliance model.
Either assume that only one appliance is working at a given time instant, or then use logic sequence to of ON/OFF combinations. For example, if there are 4 appliance then there are 16 possible combinations of ON/OFF sequence.
The HMM model can differentiate between the different states in the model.
For the load disaggregation, they take a lot of time series data, and then run Viterbi algorithm over it to find the accurate sequence of the states.

I feel that this particular algorithm is not good, as fist modelling large number of states/appliances are difficult.
It is not a real time inference algorithm, depends on the previous aggregated data.
There is lot of computation required to do, and is not possible on embedded solutions.

There are other variants for HMM, like FHMM and other, but not sure how they work.

I choose to do the GMM clustering for creating the general model and then use EM algorithm for inference.
The Gaussian cluster is fast to compute and I assume that it can be improved as we have more features. We can even identify different state of operation of an appliance based on their clusters.
Also, during the inference of the aggregated data, we can just check the expectation of each data across all the clusters. This is computationally lighter than HMM. Once, we have the classification, then can use this data point for re-learning.

Dataset

As we have no publically available dataset of aggregated power consumption and sub metered data a house in Sweden, we used the already available dataset from other countries.

The system developed in the Zyax can measure real power, reactive power, current and voltage for the aggregated power consumption. Also, the sampling frequency of the

From our side we need a dataset for training purposes that has –

1. Main Voltage – 240V (Sweden main is 240V/ 50Hz)
2. The aggregated data should have Real Power and Reactive Power information
3. The sub metered data from different appliances (fridge, dishwasher, coffee, HVAC, lightning)
4. Sub metered data of an appliance from different brands to create a generalizable model for the appliance.
5. Sub metered data of at least Real and Reactive power. More features then better.
6. The sampling frequency should be smaller (preferred smaller than 1 sec)

By searching, I was able to get hold of REDD and ECO dataset.
REDD dataset
Country: USA
Main Voltage: 120V / 50Hz
No of Houses: 6
No of Appliances: 24
Features: Voltage and Real power (Agg) / Real power (Sub)
Resolution: 15kHz (Agg) / 0.5Hz or 1Hz (Sub) based on the appliance

Can't use this dataset to create a general appliance model, as the main voltage is different from Sweden.

ECO dataset
Country: Switzerland
Main Voltage: 230V / 50Hz
No of Houses: 6
No of Appliances: ~7 appliances each house (different in each house)
Features: Real power, Current, Voltage, Power factor all phase (Agg) / Real Power (Sub)
Resolution: 1 sec

The dataset is good, as it covers lot of houses, has main voltage exactly like Sweden and have sub metered data for different appliances.
The only issue here it is that it has only one feature (Real Power) for sub metered data.

For now I have started using the ECO dataset.

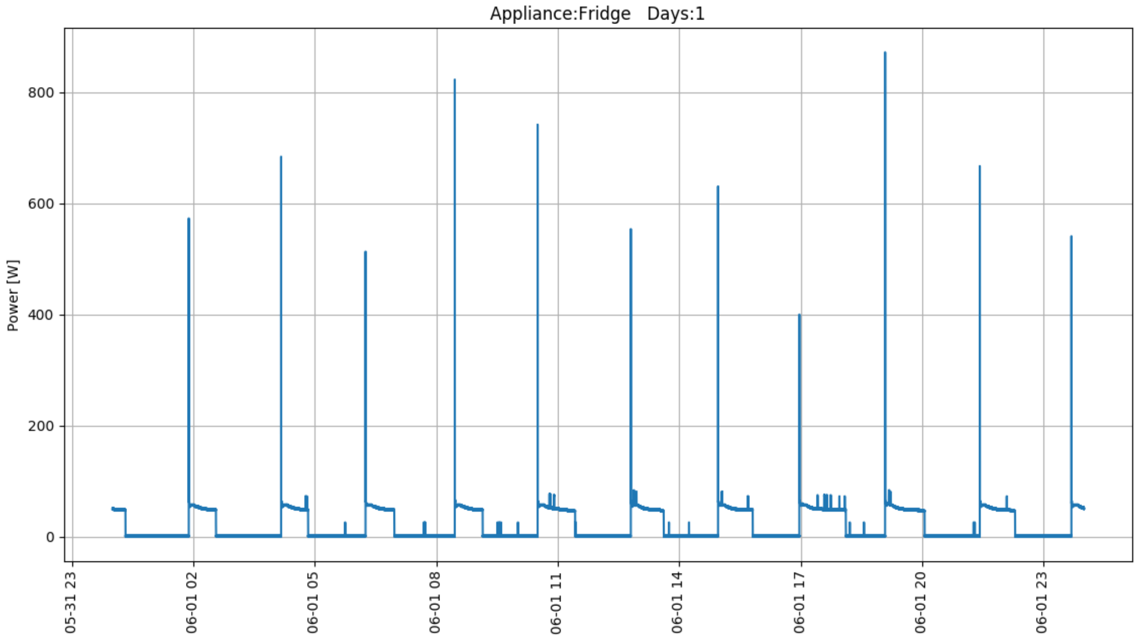
Data Wrangler

The ECO dataset is divided across folder for each house and then separate folder for each appliances measurement. A separate folder for aggregated data.
Each folder contains the measurement for each day, file named with date of the measurement. With sampling of 1sec for collection, there are 86400 data points in each file. I wrote a data wrangler for the ECO dataset.
I can select the specific house, the appliance and number of days for the data.
The class function written in Python.

Feature Extraction

Only Real Power

The idea of creating a generalizable model is to create a Gaussian Cluster for each appliances.

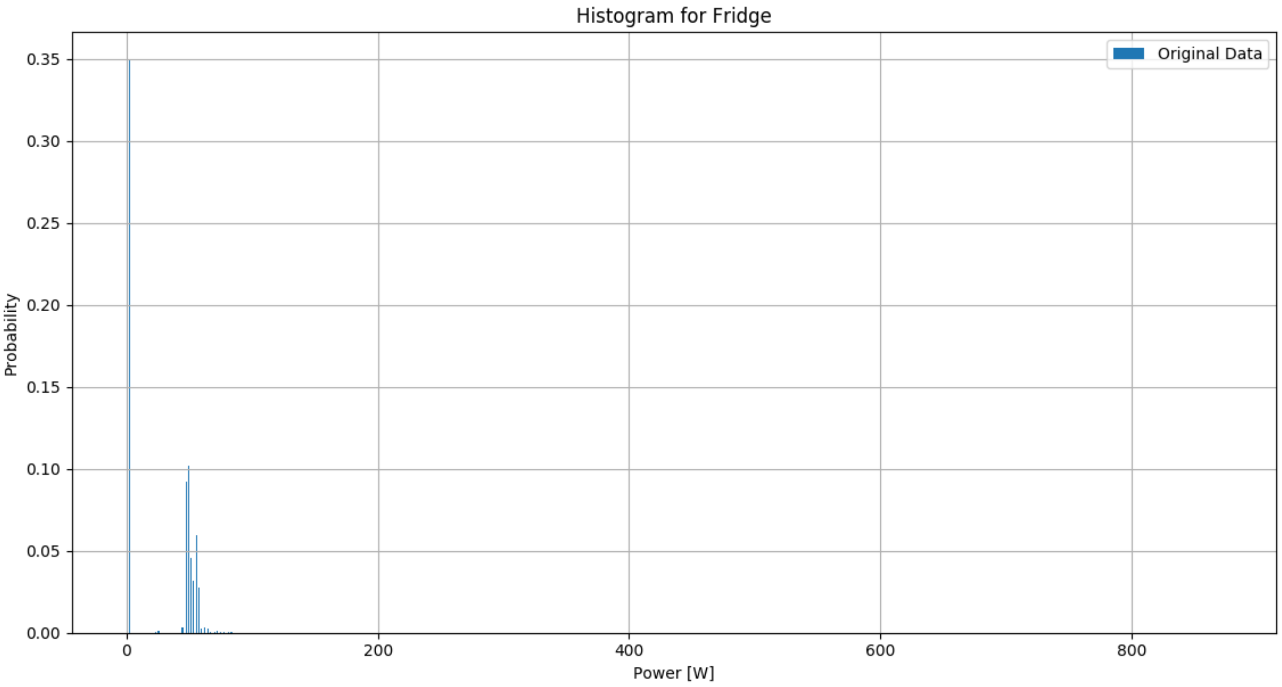


Because I have only one feature (Real Power) for the data, I tried to create Probability Density Function (PDF).

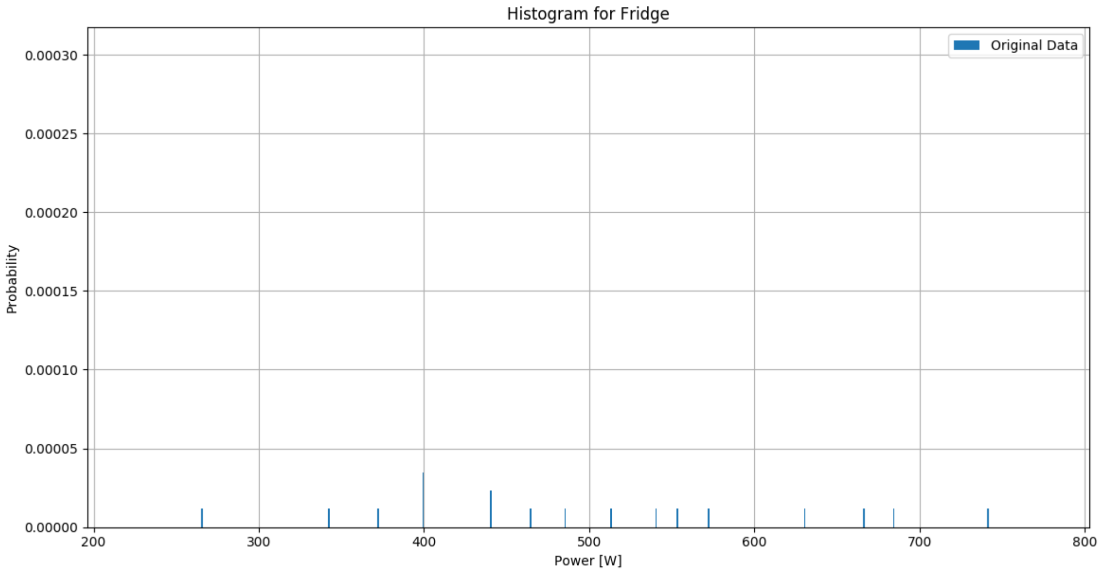
To calculate the PDF, I used probabilistic histogram with equal number of bins range of Power. The histogram will give us the probability of each power reading for the appliance.

In the histogram, we can see that there are distinct peaks. Each peak can be identified as a state of the appliance.

In the figure, we are seeing the Histogram for fridge for 1 days.

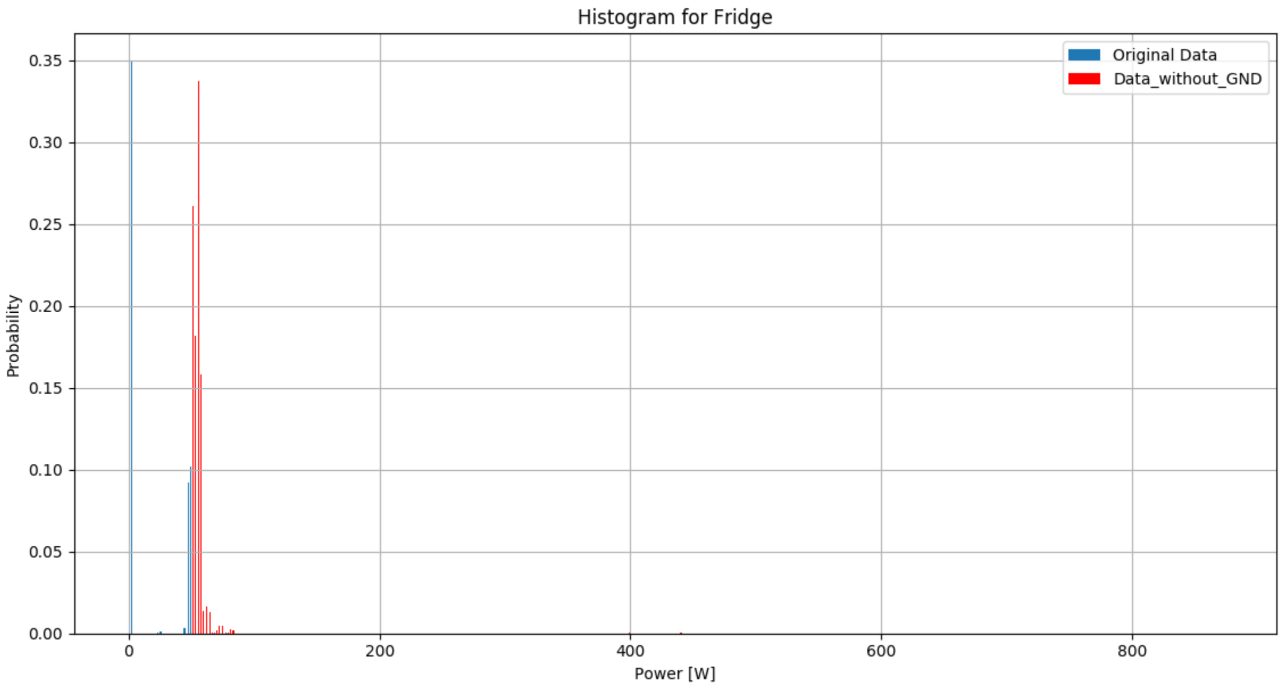


There are two prominent peaks for the histogram, one for the GND (and fluctuations) and other for the operating region of the fridge (around 50W). Also, there are small peaks ranging from 300W to 800W, which shows initial surge in the power for active region of fridge.

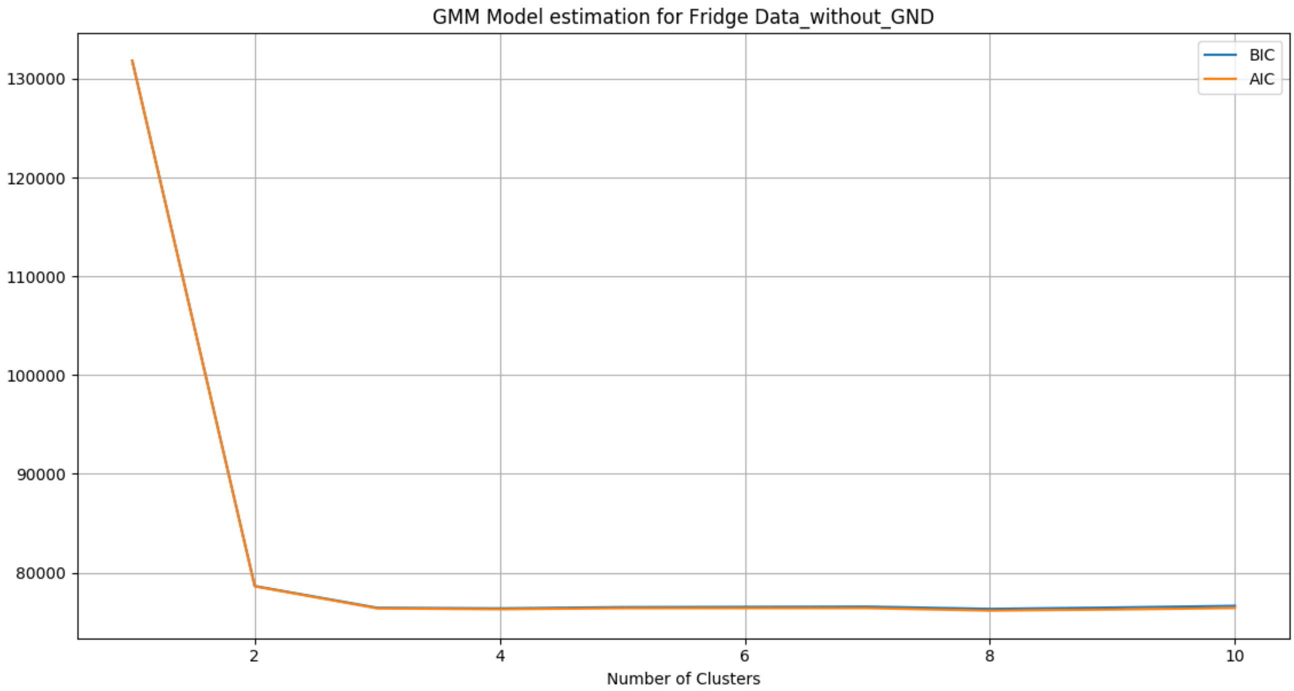


Once removing the GND (assuming anything less than 5W as GND), I again took the histogram and then using the Gaussian Mixture Model (GMM) found two clusters.

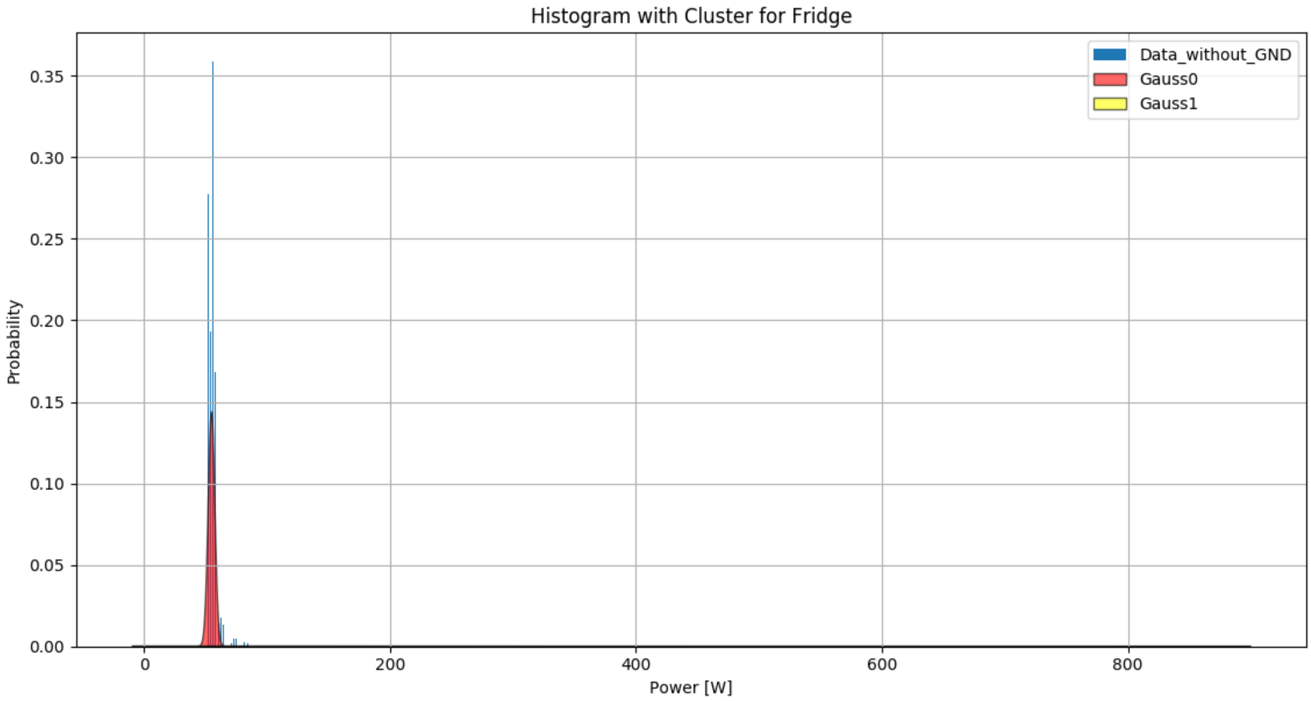
In order to find the number of clusters, I used Akaike information criterion (AIC) and Bayes information criterion (BIC) criteria to find the minima at the number of clusters.



The following figure shows the AIC and BIC for fridge and we can see 2 or 3 is a good number.



The 2 Gaussian Cluster for the fridge is then shown below.



I feel the clustering done here is bit misleading. According to this there are two distinct states for the fridge, whereas this is not true. The peak is only the starting value and then come the steady state. The complete signal is the active region.

Also, there is a big region of inactive phase where just GND signal is there. This also affects the probability and affects the Gaussian Mean and variance.

So, here I realized that I need to add time also as a feature to make the whole active period as one cluster. By this I can also eliminate the GND phase completely, hence will have model only for the active region.

Active Region Extraction

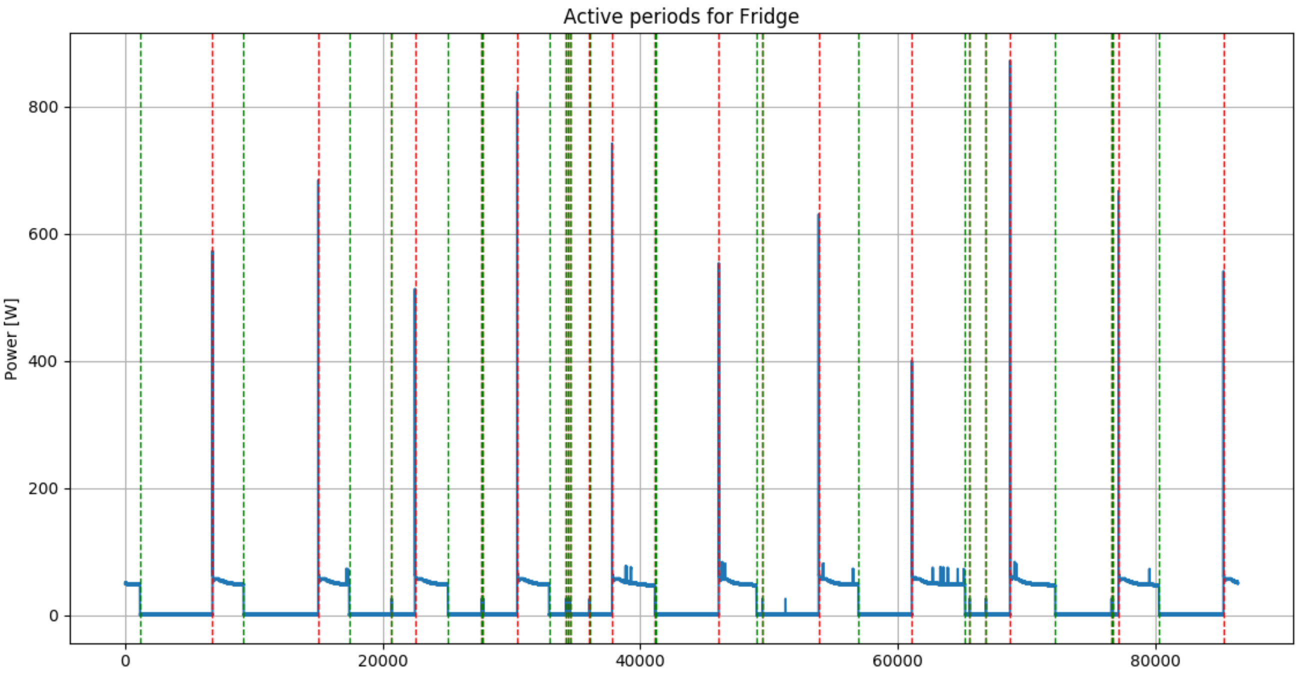
Now in order to extract only the active region, we have to identify when the signal is starting and when it is ending.
I used a simple thresholding criteria for defining the start and end of the signal.
Before doing the thresholding, I have to filter the signal to remove noises in the signal. In the power signal, the noise are fast changing with small peak.
If not removed, the noise will be treated as an active region in the simple thresholding criteria.

I used a Moving Average Filter with variable window to smoothen out the noise. The window currently is kept as 5 (so takes the average of 5sec data around the data-point)

Run through the complete data
Filter all the points that are more than the mean of the data.

Start Criteria: If the filtered data point one second before is less than 5W, then the point is start of active region
End Criteria: If the filtered data point one second after is less than 5W, then the point is end of active region.

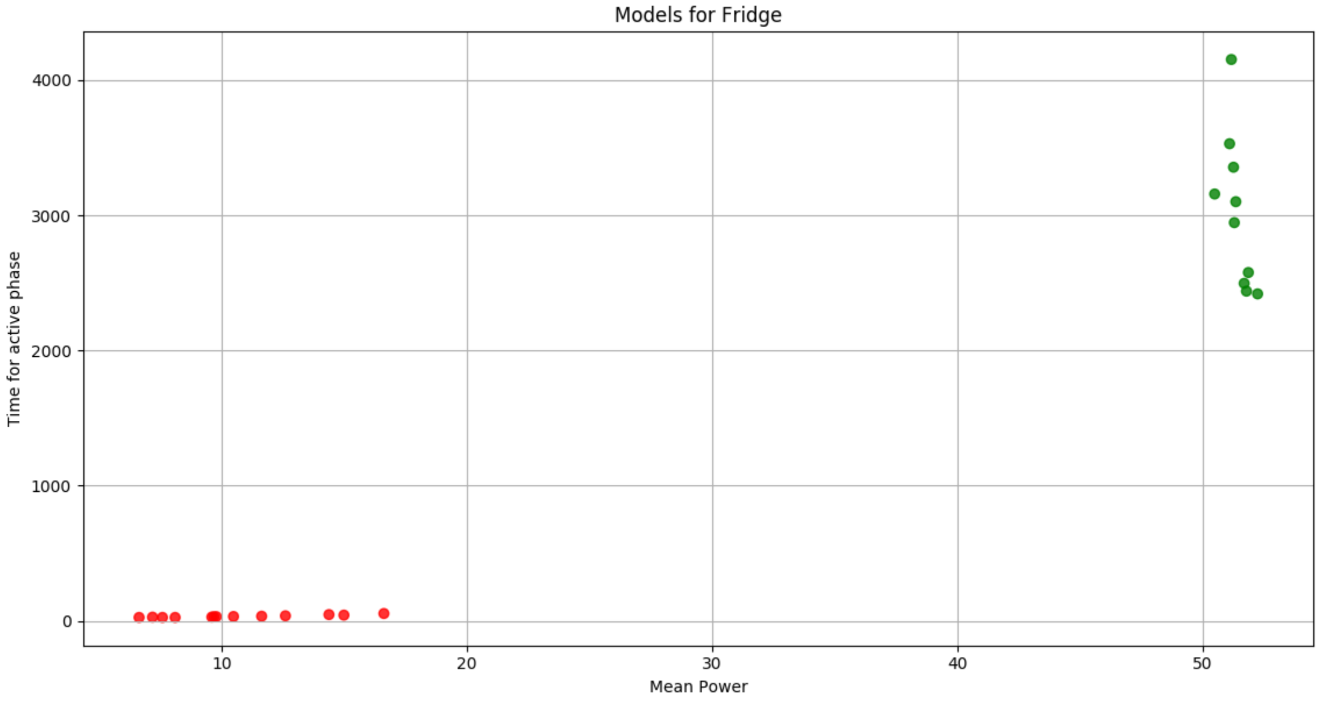
Now, just going from one start point till the closest end point, we extract the active region.
The figure below shows the active region for the fridge for 1 day period. The red line are the start line and the green line is the end line.



Real Power and Time

Doing this, we found that there are two types of active region in the fridge. One for the compressor and another for the refrigerator light.
The compressor active region is long with mean power of 50W, the light active region is small with mean power of 20W.

So now for the GMM clustering, we used Mean power and time period as features.
The AIC and BIC shows 2 clusters (not optimal but good enough).
Hence we are able to identify the two different states in the fridge and have mean and variance of both.



Currently, I have mean and variance of Fridge, dryer, kettle and coffee machine.

But again I feel that this is also not a good model to use, because we have time as a feature.
During the inference from the aggregated data, I have to capture a lot of data, then to the active phase matching in that to classify the points.
This is not real-time and one of the requirement is to do real time inference.

Next Steps

- We are currently searching for sub metered data for appliance that has more features, at least Real power and Reactive power
- Also we are looking into creating our own setup for collecting the data. We haven't figured out how to do this yet.
- Will start creating the event detection algorithm for the aggregated data.
- If we get sub metered data, then next step will be to create the Gaussian models for the appliance with features that are time independent.
- Start writing the literature review and background in the Thesis report