Efficient Inference in Dual-emission FHMM for Energy Disaggregation

Henning Lange
Aalto University
Otakaari 1 B
Finland, Espoo 02150

Mario Bergés
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, Pennsylvania 15213

Abstract

In this paper an extension to factorial hidden Semi Markov Models is introduced that allows modeling more than one sequence of emissions of the individual HMM chains, as well as a joint emission of all chains. Since exact inference in factorial hidden Markov Models is computationally intractable, an approximate inference technique is introduced that reduces the computational costs by first constraining the successor state space of the model, allowing state changes at statistically significant points in time (events) and by discarding low probability paths (truncating). Furthermore, by being agnostic about state durations the computational costs are further decreased. These assumptions allow for efficient inference that is less susceptible to local minima and allows one to specify the computational burden a priori. The performance of the inference technique is evaluated empirically on a synthetic data set whereas incorporating the feature emissions is evaluated on real world data in the context of energy disaggregation. Energy disaggregation tackles the problem of decomposing whole home energy measurements into the power traces of constituent appliances, and is a natural application for this type of models.

Introduction

Factorial hidden Markov models (FHMMs) have been used extensively in a variety of applications including speech recognition and pitch tracking [13], energy disaggregation [9] and motion tracking [12]. Although exact inference is computationally intractable in most practical cases, approximate algorithms have been introduced (e.g., Gibbs sampling [2] and variational methods [3]) which, in some cases, exploit the nature of the particular phenomena being modelled in order to relax the problem. Common relaxations include constraining the number of hidden states that can change at any one time [8] or the time points at which these changes can occur [5], introducing generic mixture components regularized to encourage piece-wise constant outputs [8], introducing signal aggregate constraints [15], and others. In other cases, the inference procedure is made tractable by trading off accuracy and computational complexity, such as by pruning likelihoods [13].

The model introduced in this paper combines and extends some of these relaxations. Specifically, our model incorporates the one-at-a-time constraint used in [8] and exploits changepoint side-information as presented in [5]. In addition to these, we allow individual HMMs to emit characteristic observation sequences (from here on called features) at changepoints, which may be observed independently of the sequence of joint emissions. Furthermore, our model allows for semi-Markovian state durations but, because it is discriminative, we do not need to explicitly define state duration parameters.

Finally, when performing inference our approach considers only the m most probable hidden state sequences that follow at any changepoint, using a modified version of the Viterbi algorithm similar to the one presented in [13]. By combining these restrictions, the computational complexity of inference can be brought down to $\mathcal{O}(Nm\sum_a k_a)$ from $\mathcal{O}(T(\prod_a k_a)^2)$ with N being the number of detected change points, k_a being the number of states of the ath HMM, T being the number of points in time and $N \ll T$ and $m \ll \prod_a k_a$.

Just as in most of the extensions to FHMMs that we build on, our model and the associated inference technique are motivated by the problem of electricity disaggregation or non-intrusive load monitoring (NILM) [16]. In NILM the objective is to infer the power consumption time-series for individual electrical appliances in a building from observations of the total power consumed by the premise.[1] Latent variable models, such as FHMM, have been used extensively in recent years [9, 8, 11, 7], yet the majority of the earlier work utilized what the community categorizes as "event-based" approaches, which focus on changepoints in the total power time-series and for the most part solve a separate classification problem at each of those points. The model presented in this paper can be understood as a hybrid between these two approaches, as it uses FHMMs in combination with changepoint side-information and classifiers for features emitted at those points.

For a simple visual representation of our graphical model, the reader can refer to Figure 1. A detailed description of the

[1] There are, of course, privacy implications related to electricity disaggregation. Though this is not the topic of the paper, some researchers have explored this problem in recent years [14].

model will follow in the next section.

Current FHMM approaches to energy disaggregation can capture side-information at changepoints only to small degree: the most prominent FHMM approaches also model the difference signal in order to capture transient information. These models do not explicitly incorporate the information that event-based approaches have been relying on to classifiy appliance state transitions that occur at changepoints. For example, start-up power transients [10] and high-frequency electromagnetic interference in the voltage line [4] have been shown to have high discriminatory power for classifying state changes. The first-order Markov assumption prevents this information from being used fully. Thus, from an application perspective, the idea behind the model is to capture more information about the transient behavior of appliance state transitions and augment existing state-based approaches that have been proposed to solve the problem.

From an algorithmic perspective, this paper shows how the computational costs can be kept low while still allowing semi-Markovian state durations in a factorial model. Hidden Semi Markov Models (HSMM), an extension to HMMs that allow for non-geometric state durations [5], model the state duration with a parameter D. HSMMs require modeling the distribution over D, which might be undesirable in some scenarios (e.g., if the goal is to detect anomalous state durations). A discriminitive version of a factorial HSMM that is agnostic about state durations will be discussed in this paper.

Dual-Emission FHMM

In our dual-emission FHMM, two separate sequences of different lengths and dimensions, namely $p \in \mathbb{R}^{T\times q}$ and $f \in \mathbb{R}^{N\times r}$, are observed. Here, T and N are the lengths of the aggregate power sequence p and the changepoint feature vector f, respectively, whereas q and r are the dimensions of each element in the respective sequences. These two sequences are conditionally independent given the hidden states of all of the HMMs in the factorial model. This is better illustrated in Figure 1. The motivation behind the two emissions is that in addition to the main observations of power p, there can be other (conditionally) independent observations occurring at changepoints such as environmental sensor measurements (e.g., sound, light intensity, etc.) or additional power measurements at higher resolution than p.

Unlike general FHMMs, the individual HMMs in our model are only allowed to transition one at a time, and only at changepoints of the p sequence. Thus, the model only makes sense once the observation sequence has been segmented by a changepoint detector. Let $E = \{0, e_1, ..., e_N, T\} \subset \{0, ..., T\}$ denote the ordered set of changepoints (0 and T are added simply for algorithmic convenience). The hidden states of each HMM independently emit features f at every changepoint depending on the previous and current state, and they jointly emit a sequence of power observations p up until the next changepoint depending solely on the current state. It is important to note that because the segmentation occurs beforehand, our model is no longer generative but rather discriminative (i.e., we do not model the length of the individual power segments emitted at each state).

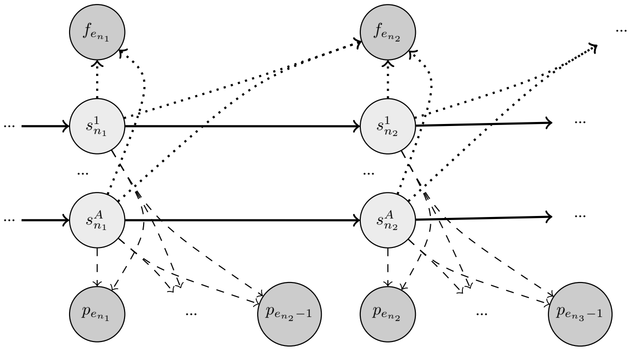

Figure 1: A graphical representation of the model. The light-grey nodes depict hidden states of the corresponding appliances. The upper and lower dark-grey nodes depict feature emissions and the observed power readings respectively. The number of observed power readings in between change points is variable.

Let $s \in \mathbb{N}^{N\times A}$ be a state matrix with $N = |E|$ being the number of changepoints and A the number of HMM chains. s_n^a denotes the state of HMM a after event e_n. Then the conditional probability $P(s|f,p)$ has the following proportionality:

$$P(s|f,p) \propto P(s_0)\prod_{n=1}^{N-1}(\prod_{a=0}^{A}\underbrace{(P(s_n^a|s_{n-1}^a)}_{\text{state transition}}\underbrace{P(f_{e_n}|s_n^a,s_{n-1}^a))}_{\text{event emission}}\underbrace{P(p_{e_n},...,p_{e_{n+1}-1}|s_n))}_{\text{joint emission}}$$

$P(s_0)$ is a probability distribution over initial joint states. $P(s_n^a|s_{n-1}^a)$ denotes the probability of HMM a transitioning from state s_{n-1}^a into state s_n^a. $P(f_{e_n}|s_n^a, s_{n-1}^a)$ models the probability of the features f_{e_n} observed at the change point detected a time point e_n given a state transition from s_{n-1}^a to state s_n^a. $P(f_{e_n}|s_n^a, s_{n-1}^a)$ allows the model to examine characteristics of the state change and make a guess about its nature[2]. Inferring a state transition rather than just the current state is motivated by the application domain: if, for example, a microwave oven turns on the magnetron to begin cooking after the plate is spinning, commonly used features such as the increase in power consumption will be very different from those found when the appliance is completely off and only the plate begins to spin. $P(f_{e_n}|s_n^a, s_{n-1}^a)$ can be interpreted as the Bayesian inversion of the output of a probabilistic classifier trying to infer the state transition given

[2] Note that during a state transition, the model assumes all HMMs emit the same feature. In the setting of energy disaggregation this is of course non-sensical: One appliance state change is responsible for features and other appliances do not emit these features by remaining in the same operational state. But modeling feature emission like this allows for plugging in probabilistic classifiers from the event-based energy disaggregation community.

the observed features. Thus, e.g. the likelihood term of a Naïve Bayes classifier could in principle be used as a model. $P(p_{e_n}, ..., p_{e_{n+1}}|s_n)$ holds the joint state s_n accountable for the observed features from the current possible change point up until the next. Let ρ^a_i be the estimated power consumption of appliance a in state i. Sensible models will penalize the deviation of $\sum_a^A \rho^a_{s^a_n}$ from the observed power features of the segment $p_{e_n}, ..., p_{e_{n+1}}$.

Efficient approximate inference

Exact inference in Factorial Hidden Markov Models would require to consider all joint states which are $\prod_a^A k_a$ for A many HMMs with k_a many states. Thus, for every point in time, computations in $\mathcal{O}(\prod_a k_a^2)$ would need to be carried out.[3] This problem is tackled by making three restriction: 1) only allowing a single appliance to change its state during each event (*one-at-a-time-constraint*), 2) only considering the successors of the m most likely paths (*cut-off constraint*) 3) allowing state changes only at points in time with significant change in the power sequence.

The *one-at-a-time* constraint allows for reducing the number of successor states of a joint state: The number of successor states of s_n without the constraint is $\prod_a^A k_a$ but with the constraint the number can be reduced to $(1 - a) + \sum_a^A k_a$. The successors of s_n can be computed by changing a single component in s_n: $suc(s) = \{s_*|s_* : s^a_* = s^a, s^i_* \in \{0, ..., k_i\}, \forall a \neq i\}$.

The *one-at-a-time* constraint allows us to reduce the number of successor states of a given state. If however, at every event all successors of all joint states are computed, the number of possible joint states is still growing exponentially. The *cut-off constraint* enforces that at every event only the successors of the m most probable paths so far are further considered which ultimately bounds the computational cost into quadratic time.

Both restrictions allow for an efficient approximate inference procedure in FHMM which is described in Algorithm 1. From a search perspective, the algorithm is a type of beam search, i.e. a breadth-first search with a limited agenda.

The Viterbi algorithm can be slightly reformulated as a forward-looking algorithm that updates the probabilities of all successor states given a previous state. Let

$$t(s \to s_*, n) = \prod_a^A P(s^a_*|s^a) P(f_{e_n}|s^a_*, s^a) P(p_{e_n}, ..., p_{e_{n+1}}|s_n)$$

At any given point in time, the algorithm computes the successors of the m most probable paths and computes their probabilities.

```
Input: $s_0$
Output: most probable sequence $s_1, ..., s_N$
$agenda = \{s_0\}$;
initialize $\phi(s, n) = 0$ for all $s$ and $n$;
$\phi(s_0, 0) = 1$;
for $n \in \{0, ..., N-1\}$ do
    for $s \in agenda$ do
        // one-at-a-time assumption
            incorporated into $suc(s)$
        for $s_* \in suc(s)$ do
            $\phi(s_*, n+1) =$
            $max[\phi(s_*, n+1), \phi(s, n)t(s \to s_*, n)]$;
        end
    end
    // cut-off restriction
    $agenda = b$-best states in $\phi(*, n+1)$
end
```

Algorithm 1: Truncated inference in FHMM

Parameter Estimation

The model can, in principle, be used in an un-, semi- and fully-supervised fashion. In the un- and semi-supervised fashion, the model can operate fully Bayesian: priors over the model parameters can be defined (that may depend on some training data) and since an approximation of forward probabilities can be obtained by substituting the max-operator by a sum in algorithm 1, joint states can be sampled efficiently. Note that in comparison to [5], this technique avoids local minima by sampling joint states instead of successively sampling states of a single HMM. The additive joint emission and having to sample states one after the other makes sampling individual state instead of joint states particularly vulnerable to local minima: the increase in energy consumption may be partially explained away by a state transition of an appliance and if in a subsequent step the state of the actual consumer of this power is sampled, it is unlikely to sample the correct state since some the power is already explained away. Consider the following example: a kettle consuming 1000W was turned *on* and the state of a hair-dryer either consuming 250W or 750W is sampled. It is very likely that the 750W state will be sampled since it explains away more of the unexplained power. If in the next step the kettle is sampled, it is very likely that it will be assumed to be *off* since only 250W are left unexplained.

Independent of the models for $P(f_{e_n}|s^a_n, s^a_{n-1})$ and $P(p_{e_n}, ..., p_{e_{n+1}}|s_n)$ information about the initial state distribution and state transitions matrices for each appliance needs to be provided either in the form of prior distributions in the semi- and unsupervised way or as hard distributions in the supervised way.

Estimating Transition Probabilities

Estimating state transition probabilities might not always be trivial since they are context dependent: The estimated probability of an appliance staying in an operational state $P(s^a_n = i|s^a_{n-1} = i)$ depends on the specific event detector used to segment the joint observations (i.e. the power time series) and on the nature of the other appliances present in the experiment. For instance, if a sensitive event detector is employed the number of events that do not correspond to real appliance state transitions will increase, which will in

[3] Exact inference with the Viterbi algorithm requires computations in $\mathcal{O}(\prod_a k_a^2)$. The Lauritsen & Spiegelhalter algorithm can reduce the computational cost to $\mathcal{O}(\sum_a k_a \prod_a k_a)$ [3]

turn affect the estimated probabilities associated with appliances remaining in their state. On top of that, appliances in the context of an appliance that changes its state rapidly will have to assign more probability mass to the diagonal of the state transition matrix.

There are two possible ways of estimating state transition probability matrices for each appliance. If a data set is available that contains the aggregate signal as well as temporally aligned single appliance measurements, an event detector can be applied to the aggregate signal and these events will then be imposed on the single appliance measurements. A quantization scheme can be applied to the single appliance measurements and if the level of the single appliance readings change at an event detected in the aggregate signal, a state transition for this appliance was detected. If however, the level did not change, the appliance remained in its state. However, this kind of ground truth data might not always be available.

We will show how the information from single appliance measurements can be transferred into different contexts of other event detectors and other appliances to obtain state transition probabilities: The diagonal of state transition matrices is dependent on the false positive rate of the event detector as well as their duration. If there is a two-state appliance whose *on*-state is short in comparison to its *off*-state, then $P(s_n^a = \mathit{off}|s_{n-1}^a = \mathit{off}) > P(s_n^a = \mathit{on}|s_{n-1}^a = \mathit{on})$. $P(s_n^a = i|s_{n-1}^a = i)$ terms need to be robust to the sensitivity of the event detector used, as well as the context in which an appliance is present. In other words, it would be useful to be able to estimate state transition probabilities given the false positive rate of an event detector as well as information about the duration that the appliances spend in each state. Let us assume that single appliance measurements are segmented by a perfect event-detector[4] (i.e. no false negatives or positives). Since there are no false positives, then $P(s_n^a = i|s_{n-1}^a = i) = 0$. Let us also assume that state duration information was extracted from these single appliance measurements, and that τ_i^a denotes the mean duration of state i for appliance a. For simplicity, let π^a denote the state transition matrix of appliance a, and ϕ denote the false positive rate of the event detector, i.e. the number of false positives per time step.

Using this information, we first compute the expected state duration for each appliance. In other words, we compute $\mathbb{E}[\tau^a]$, where this expectation is taken over the probability of appliance a being in any state i (i.e., $P(s_i^a)$). This probability is equal to the stationary distribution of π^a, i.e. the eigenvector of π^a with length 1. Thus, the expected state duration of appliance a is $\mathbb{E}[\tau^a] = \sum_i P(s_i^a)\tau_i^a$. We can now use this information to estimate the number of events that we expect to see in the aggregate observation.

Assuming that appliance state distributions are uncorrelated, then for a state with the duration τ_i^a, we expect to see $\sum_{j\neq a} \frac{\tau_i^a}{E[\tau^j]}$ events caused by appliances other than a and $\phi\tau_i^a$ false positive events. Finally, let $p^a(i|j) = P(s_n^a = i|s_{n-1}^a = j)$ and

$$\kappa_i^a = \sum_{j\neq a} \frac{\tau_i^a}{E[\tau^j]} + \phi\tau_i^a$$

Now we can renormalize the transition probabilities using all this information as follows:

$$\hat{P}(s_n^a = i|s_{n-1}^a = j) = \begin{cases} \frac{p^a(i|j)}{\kappa_i^a+1} & \text{if } i \neq j \\ \frac{\kappa_i^a}{\kappa_i^a+1} & \text{if } i = j \end{cases}$$

This shows how the state transition probabilities absorb information about the false positive rate of the event detector and implicitly incorporate state duration information without having to explicitly model a distribution over state durations.

Experiments

This paper introduces two novelties: first it shows how feature emissions can be incorporated into a FHMM and second an efficient approximate inference technique is introduced. Two experiments were conducted to evaluate the performance of each novelty individually. The first experiment shows the performance of the model in a fully supervised way with and without the additional feature emission. For this, the power consumption of 8 appliances were logged at a sampling rate of roughly 1.5 Hz over the span of $\approx 600,000s$. This dataset is part of the raw measurements collected for BLUED [1]. For more details on how the data was obtained see [1]. The sum of the individual appliances was then disaggregated using our model. In a second experiment different inference techniques are compared. For this, the experiments described in [8] were recreated and the performance of the inference technique introduced here is compared to the performance of inference techniques based on variational methods (Structured Mean Field [6]) and integer programming techniques (AFAMAP [8]). As a performance criteria the disaggregation error, also used in [8], was used (lower is better):

$$de = \sqrt{\frac{\sum_{a,t} \|y_t^a - \hat{y}_t^a\|_2^2}{\sum_{a,t} \|y_t^a\|_2^2}}$$

Experiment 1

The goal of the first experiment is to show that incorporating transient or change point feature information can improve the disaggregation performance. The data used for this experiment contains the power measurements of single appliances at a rate of roughly 1.5Hz. These measurements were approximated by a piece-wise constant time series in such a way that the number of different values the time series can take and the deviation to the original time series is minimized. The number of values the quantized time series takes

[4] This can easily be achieved by applying an event detector to single appliance measurements and then removing all false positive events.

	Actual	Dual emission			Single emission		
Appliance	Observed E	Estimated E	F1	de	Estimated E	F1	de
Compressor	104.9Wh	109.6Wh	0.73	0.75	104.77Wh	0.96	0.27
A-V	2321.7Wh	2240.2Wh	0.98	0.21	2301.4Wh	0.90	0.44
Desk lamp	220.3Wh	280.5Wh	0.87	0.53	219.4Wh	0.80	0.64
DVR/Blueray	3861.2W	3840.7Wh	0.96	0.26	3589Wh	0.94	0.32
Garage door	659.4Wh	591.5Wh	0.84	0.54	1899.8	0.50	1.38
Iron	142.0Wh	131.1	0.65	0.81	142Wh	0.97	0.24
Fridge	3546.8Wh	3535.8	0.96	0.27	3654Wh	0.87	0.51
Tall desk lamp	450.6Wh	575.8	0.87	0.53	511Wh	0.75	0.72
TV	3697.4Wh	3695.0	0.98	0.2	2583Wh	0.80	0.58
total	15004.6	15000.6		**0.3**	15006.7		**0.56**

Table 1: The appliance level performance of the disaggregation system with and without a feature emission.

constitute the number of states and each state is associated with a power level ρ_a^i. The data that was disaggregated is the sum of the raw unfiltered power signals of each appliance.

A change-point detector that identifies significant changes in the first difference of the signal was used to extract events in the aggregate signal. In order to obtain state transition probabilities, events detected in the aggregate signal were imposed onto the quantized appliance power traces and annotated according to the pre- and post-event power levels of that appliance during training.

Because of the rather low sampling rate of the data and the absence of any external features, rather inexpressive features were extracted at every event, namely the first difference of the aggregate signal in a symmetric window of 20 data points. The feature dimensionality was reduced by PCA retaining 93% of the variance.

As a model of the event emissions the likelihood term of a Gaussian Naïve Bayes classifier is used, thus conditional independence between the feature elements is assumed. The parameters of that feature distributions were then estimated using a maximum likelihood estimation approach.

The model of $P(p_{e_n}, ..., p_{e_{n+1}}|s_n)$ penalizes the deviation of the sum of estimated power consumptions from the observed signal. The difference between the sum of estimates and the observed signal contains highly structured artifacts caused by the transient behavior of the individual appliances and the model tries to relax this penalty by modeling these artifacts. For this, on the training set, the quantized appliances are summed up and subtracted from the aggregate signal. A slack function $\zeta(r_{t-1}, ..., r_{t-n})$, abbreviated as ζ from here on, is introduced to predict this residual by fitting an auto-regressive model of order 10.

$$P(p_{e_n}, ..., p_{e_{n+1}}|s_n) = \prod_{t=e_n}^{e_{n+1}} \mathbf{L}(p_t - \zeta | \sum_a^A \rho^a_{s^a_n}, b)$$

with r_t being the residual (i.e. $r_t = p_t - \sum_a^A \rho^a_{s^a_n}$) and $\mathbf{L}(x|\mu, b) = \frac{1}{2b} exp[-\frac{|x-\mu|}{b}]$ being the Laplace density. This model effectively introduces a temporal dependence between observations and can in principle also alleviate the impact of unmodeled appliances. The results were obtained by 5-fold cross-validation. The initial state is assumed to be known but this had little impact on the performance in comparison to a naïve initialization, i.e. all appliances are turned off.

Results Table 1 show the disaggregation performance of the model with (Dual emission) and without (Single emission) the feature emission. As can be seen, the overall performance increased by 0.26 as measured by de. However, the performance on the Iron and the Compressor decreased. This can be explained by the nature of the data: the activity of the Iron and Compressor are sparse and temporally clustered in the dataset. When performing cross-validation with temporally clustered activity the problem arises that most events are either in the training set or in the test set. In this case the training set sometimes contains only a single event for the compressor which makes parameter estimation particularly hard. For appliances whose activity is not temporally clustered the performance increases substantially. Incorporating the second emission, thus, shows overall promising results.

Experiment 2

The goal of the second experiment is to show that the inference technique introduced in this work can enable a trade-off between computational time and disaggregation accuracy. The performance of the modified Viterbi algorithm (TruncTerbi) is compared to Structured Mean Field and AFAMAP. AFAMAP is an inference introduced in [8] based on integer programming. For this, the experiments conducted in [8] are recreated. One experiment is selected for which AFAMAP and SMF perform sub-optimally because of the high but realistic number of HMMs included. The experiments are based on a synthetic dataset.

The dataset contains 20 sets that each comprise 10 cyclic HMMs, each with 4 states and 4 dimensional output. The initial state of each HMM is drawn from a uniform distribution. The mean output of each HMM in every state is drawn from a uniform distribution in the range $[0, 2]$. The observation is the sum of all HMMs plus zero-mean Gaussian noise with a covariance of $0.01I$.

Experimental setup For the second experiment, no transient information is incorporated. Thus the model of $P(s_n, s_{n-1}|f_{e_n}) = 1$ for all f_{e_n}, s_n and s_{n-1}. The model $P(p_{e_n}, ..., p_{e_{n+1}}|s_n)$ simply penalizes any deviation of the observed from the inferred aggregate power:

$$P(p_{e_n},...,p_{e_{n+1}}|s_n) = \prod_{e_n}^{e_{n+1}} N(p_{e_n} - \sum_{a}^{A} \rho_{s_n}^{a}|0, 0.01I)$$

with $N(x|\mu,\Sigma)$ being the multivariate Gaussian distribution. Since *AFAMAP* incorporates some transient information in the difference model, the fairest comparison is to *AFAMAP no diff* where the difference model is deactivated. The model parameters for the model introduced in this work were obtained on a held out training set. However, it is important to note that the model parameters for *SMF* and *AFAMAP no diff* are the true distribution in the data, which should give these approaches an advantage.

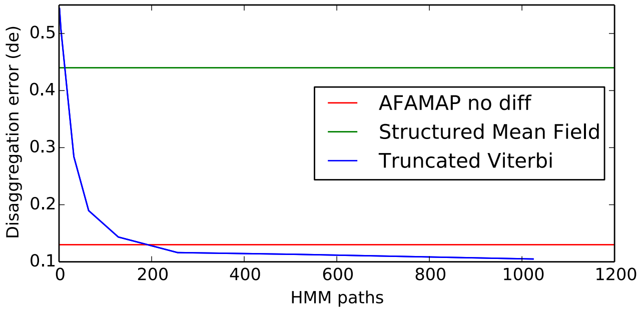

Figure 2: Even in a scenario where *SMF* and *AFAMAP no diff* are provided with a higher amount of ground truth, the inference technique introduced here outperforms its competitors

The parameter that controls how many HMM paths are cut off is varied in the experiment. The results are depicted in Figure 2. Even though the competitors have an advantage due to having access to the true underlying distributions during inference, the truncated Viterbi still outperforms them given that enough HMM paths are kept. Our inference technique allows us to control the computational expense and can, given unlimited computational power, approximate exact inference. The computational time increases linearly with the number of retained HMM paths.

The truncated Viterbi in conjunction with dual emission model introduced in this paper also has advantages from a modeling point-of-view: it is in a sense a meta-model. Different models for the feature emission and joint emission can be plugged in in order to adjust the model to different scenarios. If, for example, the model has to cope with unmodeled devices, the joint emission can be extended with a component that soaks up residual power that cannot be explained by the devices that are modeled. Variational methods, as well as Integer Programming techniques for inference require reformulating the inference problem, i.e. for variational methods the analytical solution to minimizing the Kullback-Leibner divergence needs to be calculated, whereas the inference technique based on the truncated Viterbi algorithm allows one to change the underlying model components without changing the inference technique.

Conclusion

This work made two contributions. First, we showed how event-based and event-less approaches can be fused together by incorporating feature emissions of the individual HMM chains along with a joint emission. Second, a generic inference technique is introduced, which allows for efficient inference in that model. This inference technique allows one to change the underlying model components of the feature and joint emission without having to change the inference technique.

The performance of the individual components was tested on a synthetic and a real world data set and both experiments yielded positive and promising results. One should note that the experiment on the real world dataset circumvented some NILM problems such as calibration difficulties by disaggregating the artificial sum rather than an actual measurement on the whole home or sub-metered signals. However, the model could not unfold its full potential since the data set does not provide external or high level features.

There is however still much room for improvement: One problem that arises is that state sequences with minimal differences might occupy the most probable paths. There are sometimes surges in the power line that only last for less than one second. The energy these spikes consume is negligible. A problem with simply retaining the most probable paths is that the most probable paths might actually be very similar and their only difference is their explanation of this single spike. A system that identifies maximally similar state sequences and removes or merges those paths might boost the performance of the system substantially.

Another point that can be improved is the approach for explaining-away the variance. The data is segmented by an event-detector and only the portion of data segment is explained. So far, the variance of the segment is not explained and the power traces of some appliances show significant differences in variance which might help discriminating appliances. Also, when an appliance that is known to have high variance is believed to have been turned on, this variance should explain away changes in the observed signal. The power consumption of a television, for example, might vary quickly due to different brightness or volume. When the TV is believed to be running, small variations in the observed signal are more likely caused by the television than by a low power device.

Finally, more work can also be devoted to improve the features: a promising approach could be to employ deep learning. Training a deep feature extractor on high-frequency current readings would be very time consuming but once the feature transformation is learned, computing the deep features is computationally not very expensive and could potentially be carried out by the smart meter itself. This technique could dramatically reduce the data sent out by the smart meter while still maintaining the high frequency information since data would only need to be transmitted when the power line features change. In a perfect world, this would abolish the need for an event detector if a change of the deep features marks a change point of an appliance.

References

[1] Kyle D Anderson. “Non-Intrusive Load Monitoring: Disaggregation of Energy by Unsupervised Power Consumption Clustering”. In: (2014).

[2] Stuart Geman and Donald Geman. “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 6 (1984), pp. 721–741.

[3] Zoubin Ghahramani and Michael I Jordan. “Factorial Hidden Markov Models”. In: *Machine learning* 29.2-3 (1997), pp. 245–273.

[4] Sidhant Gupta, Matthew S Reynolds, and Shwetak N Patel. “ElectriSense: single-point sensing using EMI for electrical event detection and classification in the home”. In: *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM. 2010, pp. 139–148.

[5] Matthew J Johnson and Alan S Willsky. “Bayesian Nonparametric Hidden Semi-Markov Models”. In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 673–701.

[6] Michael I Jordan et al. “An introduction to variational methods for graphical models”. In: *Machine learning* 37.2 (1999), pp. 183–233.

[7] Hyungsul Kim et al. “Unsupervised Disaggregation of Low Frequency Power Measurements.” In: *SDM*. Vol. 11. SIAM. 2011, pp. 747–758.

[8] J Zico Kolter and Tommi Jaakkola. “Approximate Inference in additive Factorial HMMs with application to Energy Disaggregation”. In: *International conference on artificial intelligence and statistics*. 2012, pp. 1472–1482.

[9] J Zico Kolter and Matthew J Johnson. “REDD: A public data set for energy disaggregation research”. In: *Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA*. Vol. 25. Citeseer. 2011, pp. 59–62.

[10] Steven B Leeb and James L Kirtley Jr. *Transient event detector for use in Non-Intrusive Load Monitoring systems*. US Patent 5,483,153. 1996.

[11] Oliver Parson et al. “Non-Intrusive Load Monitoring Using Prior Models of General Appliance Types”. In: *Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012.

[12] Deva Ramanan and David A. Forsyth. “Automatic Annotation of Everyday Movements”. In: *Advances in Neural Information Processing Systems 16*. Ed. by S. Thrun, L.K. Saul, and B. Schölkopf. MIT Press, 2004, pp. 1547–1554. URL: `http://papers.nips.cc/paper/2370-automatic-annotation-of-everyday-movements.pdf`.

[13] M. Wohlmayr and F. Pernkopf. “Model-Based Multiple Pitch Tracking Using Factorial HMMs: Model Adaptation and Inference”. In: *Audio, Speech, and Language Processing, IEEE Transactions on* 21.8 (2013), pp. 1742–1754. ISSN: 1558-7916. DOI: `10.1109/TASL.2013.2260744`.

[14] Lei Yang et al. “Optimal privacy-preserving energy management for smart meters”. In: *INFOCOM, 2014 Proceedings IEEE*. IEEE. 2014, pp. 513–521.

[15] Mingjun Zhong, Nigel Goddard, and Charles Sutton. “Signal Aggregate Constraints in Additive Factorial HMMs, with Application to Energy Disaggregation”. In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 3590–3598. URL: `http://papers.nips.cc/paper/5526-signal-aggregate-constraints-in-additive-factorial-hmms-with-application-to-energy-disaggregation.pdf`.

[16] Ahmed Zoha et al. “Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey”. In: *Sensors* 12.12 (2012), pp. 16838–16866.