UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL SCIENCE AND ENGINEERING

Electronics and Computer Science

Unsupervised Training Methods for Non-intrusive Appliance Load Monitoring from Smart Meter Data

by

Oliver Parson

Thesis for the degree of Doctor of Philosophy

April 2014

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL SCIENCE AND ENGINEERING
Electronics and Computer Science

Doctor of Philosophy

UNSUPERVISED TRAINING METHODS FOR NON-INTRUSIVE APPLIANCE LOAD MONITORING FROM SMART METER DATA

by Oliver Parson

Non-intrusive appliance load monitoring (NIALM) is the process of disaggregating a household's total electricity consumption into its contributing appliances. Smart meters are currently being deployed on national scales, providing a platform to collect aggregate household electricity consumption data. Existing approaches to NIALM require a manual training phase in which either sub-metered appliance data is collected or appliance usage is manually labelled. This training data is used to build models of the household appliances, which are subsequently used to disaggregate the household's electricity data. Due to the requirement of such a training phase, existing approaches do not scale automatically to the national scales of smart meter data currently being collected.

In this thesis we propose an unsupervised training method which, unlike existing approaches, does not require a manual training phase. Instead, our approach combines general appliance knowledge with just aggregate smart meter data from the household to perform disaggregation. To do so, we address the following three problems: (i) how to generalise the behaviour of multiple appliances of the same type, (ii) how to tune general knowledge of appliances to the specific appliances within a single household using only smart meter data, and (iii) how to provide actionable energy saving advice based on the tuned appliance knowledge.

First, we propose an approach to the appliance generalisation problem, which uses the Tracebase data set to build probabilistic models of household appliances. We take a Bayesian approach to modelling appliances using hidden Markov models, and empirically evaluate the extent to which they generalise to previously unseen appliances through cross validation. We show that learning using multiple appliances vastly outperforms learning from a single appliance by 61–99% when attempting to generalise to a previously unseen appliance, and furthermore that such general models can be learned from only 2–6 appliances.

Second, we propose an unsupervised solution to the model tuning problem, which uses only smart meter data to learn the behaviour of the specific appliances in a given household. Our approach uses general appliance models to extract appliance signatures from

a household's smart meter data, which are then used to refine the general appliance models. We evaluate the benefit of this process using the Reference Energy Disaggregation Data set, and show that the tuned appliance models more accurately represent the energy consumption behaviour of a given household's appliances compared to when general appliance models are used, and furthermore that such general models can perform comparably to when sub-metered data is used for model training. We also show that our tuning approach outperforms the current state of the art, which uses a factorial hidden Markov model to tune the general appliance models.

Third, we apply both of these approaches to infer the energy efficiency of refrigerators and freezers in a data set of 117 households. We evaluate the accuracy of our approach, and show that it is able to successfully infer the energy efficiency of combined fridge freezers. We then propose an extension to our model tuning process using factorial hidden semi-Markov models to model households with a separate fridge and freezer. Finally, we show that through this extension our approach is able to simultaneously tune the appliance models of both appliances.

The above contributions provide a solution which satisfies the requirements of a NIALM training method which is both unsupervised (no manual interaction required during training) and uses only smart meter data (no installation of additional hardware is required). When combined, the contributions presented in this thesis represent an advancement in the state of the art in the field of non-intrusive appliance load monitoring, and a step towards increasing the efficiency of energy consumption within households.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Oliver Parson, declare that the thesis entitled *Unsupervised Training Methods for Non-intrusive Appliance Load Monitoring from Smart Meter Data* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as: (Parson et al., 2011), (Parson et al., 2012) and (Parson et al., 2013).

Signed:..

Date:..

Acknowledgements

First and foremost, I would like to thank my supervisors, Alex Rogers and Mark Weal, for their excellent advice and constant support, and for being an endless source of motivation throughout the course of my PhD. Their commitment, not only to my PhD, but also to science in general, has been by far the largest source of inspiration for this thesis. I would also like to give thanks to my lab colleagues both past and present for their companionship and coffee stories, and although there are too many great people to individually name here, I am sure you all know who you are.

Second, I would like to thank Enrico Costanza, my internal examiner, and Gregory O'Hare, my external examiner, for somehow managing to balance the necessary rigour of a PhD viva with being a genuinely friendly pair of individuals. They succeeded in making the exam an enjoyable experience while also providing some excellent suggestions for improvement.

Third, I would like to thank Sara for her unwavering support when times were hard, and her ability to celebrate the good times even while juggling it with her own thesis. Completing a PhD feels like quite a journey, and I'm so glad it's one we could enjoy together.

Fourth, I would like to thank my family for putting up with the *occasional* conference trip, and also for mastering Skype to allow us to keep in contact while I was away.

Finally, I would like to acknowledge the financial support provided by the EPSRC through the ALADDIN and ORCHID projects, without which my PhD would not have been possible.

Nomenclature

$i:j$	Sequence of indicies $i,\dots,j$
$\mathbf{x}=\{x_1,\dots,x_T\}$	Sequence of aggregate power readings.
$\Delta\mathbf{x}=\{x_2-x_1,\dots,x_T-x_{T-1}\}$	Sequence of differences between aggregate power readings.
$\mathbf{x}^{(n)}=\{x_1^{(n)},\dots,x_T^{(n)}\}$	Sequence of appliance instance power readings.
$\mathbf{z}^{(n)}=\{z_1^{(n)},\dots,z_T^{(n)}\}$	Sequence of appliance instance operational states.
$n\in\{1,\dots,N\}$	An appliance.
$t\in\{1,\dots,T\}$	A time slice.
$k\in\{1,\dots,K\}$	An appliance's operational state.
$\boldsymbol{\theta}=\{\boldsymbol{\pi},\mathbf{A},\boldsymbol{\phi}\}$	Set of model parameters for a HMM.
$\boldsymbol{\pi}$	Vector of initial probabilities for a HMM.
$\mathbf{A}$	Matrix of transition probabilities for a HMM.
$\boldsymbol{\phi}$	Vector of emission densities for a HMM.
$\mathcal{N}(\mu,\tau)$	Gaussian distribution of mean μ and precision τ.
$\boldsymbol{\Theta}=\{\boldsymbol{\Theta}^{\alpha},\boldsymbol{\Theta}^{C},\boldsymbol{\Theta}^{\lambda},\boldsymbol{\Theta}^{r},\boldsymbol{\Theta}^{\beta},\boldsymbol{\Theta}^{w}\}$	General model of an appliance type.
$\boldsymbol{\alpha}$	Vector of concentration parameters of a Dirichlet distribution.
$\mathbf{C}$	Matrix where rows represent concentration parameters of a Dirichlet distribution.
$\mathcal{N}(\lambda,\ \mathrm{r})$	Gaussian distribution with mean λ and precision r over the mean of a Gaussian distribution.
$\mathrm{Gamma}(\beta,w)$	Gamma distribution with shape β and scale w over the precision of a Gaussian distribution.
D	Appliance specific likelihood threshold.

Acronyms

AMPds	Almanac of Minutely Power Data Set
BLUED	Building-Level fUlly-labeled data set for Electricity Disaggregation
CDA	conditional demand analysis
CFHMM	conditional factorial hidden Markov model
DHMM	difference hidden Markov model
EM	expectation maximisation
FHMM	factorial hidden Markov model
FHSMM	factorial hidden semi-Markov model
FN	false negative
FP	false positive
FPR	false positive rate
HMM	hidden Markov model
HSMM	hidden semi-Markov model
i.i.d.	independent and identically distributed
iHMM	infinite hidden Markov model
IOHMM	input-output hidden Markov model
NIALM	non-intrusive appliance load monitor
ROC	receiver operating characteristic
REDD	Reference Energy Disaggregation Data set
TN	true negative
TP	true positive
TPR	true positive rate

Chapter 1

Introduction

Governments around the world currently face three energy issues: decreasing reserves of fossil fuels, securing a sustainable energy supply and the effects of climate change (MacKay, 2008). Furthermore, global population growth will continue to raise the demand for energy for years to come. It is therefore necessary to investigate methods by which our energy infrastructure can meet such demand without causing catastrophic and irreversible damage.

In an effort to prevent dangerous climate change, many countries have committed to a reduction in greenhouse-gas emissions. For example, the UK has committed to a legally binding target to reduce greenhouse gas emissions by 80% by 2050 relative to their 1990 baseline (Department of Energy & Climate Change, 2008). To meet this target, various initiatives have been introduced to replace consumption of primary energy fuels with electric alternatives in order to reduce carbon emissions (e.g. electrification of transport and heating systems). However, such schemes will clearly increase the demand for electricity.

Domestic electricity consumption accounts for approximately 27% of worldwide electricity consumption (International Energy Agency, 2008). More specifically, in the UK during 2009, domestic electricity use accounted for approximately 126 TWh of energy; 24% of the country's overall electricity consumption (Department of Energy & Climate Change, 2010). In order to minimise the load placed on national infrastructure and resources, it is essential to optimise the efficiency of electricity consumption in homes by eliminating wasted energy. Such optimisations will contribute to a reduction in overall electricity demand so long as the consumption is not replaced by other forms of energy.

One approach to increase the efficiency of domestic electricity use is to inspire positive behavioural change in consumers (Ford, 2009). This can be achieved by providing feedback to a household's occupants indicating how much energy each home appliance has used (Ueno et al., 2006). A simple analogy can be found by comparing the bills received from a supermarket after a large grocery shop and from an energy supplier after a period of utility consumption (Froehlich, 2009). Whereas the supermarket bill breaks down the total price by individual items, the energy bill shows only the total price. There are many more energy or cost reduction methods that are enabled by disaggregated energy data, as discussed in Section 1.3. However, the primary application can be thought of as providing such a breakdown of energy by household appliances on an energy bill. We now describe the problem of breaking down household electricity in more detail.

1.1 Problem Statement

The research problem addressed in this work is concerned with how the total energy consumed in a household can be disaggregated into individual appliances. The aim is to calculate such information and provide it to the consumer through minimally intrusive methods. The approaches discussed in this work require that the aggregate power drawn by all appliances in a household is measured periodically. The term premises-level metering will be used to describe the aggregate metering of appliances, whether contained within a household, a workplace or any other building.

The literature around this subject defines a clear distinction between intrusive and non-intrusive metering (Hart, 1992). Intrusive metering refers to appliance-level metering; the deployment of one meter per appliance. Conversely, non-intrusive metering refers to premises-level metering; the deployment of one meter per premises. For consistency, this work will use these terms following the accepted definition in the literature.

Appliance energy disaggregation could be performed through intrusive metering. The deployment of one meter per appliance would allow each individual appliance's energy consumption to be communicated to a central hub. However, there are many practical disadvantages to this method that have prompted the study of non-intrusive metering. First, the financial cost of manufacturing and installing enough meters to match the number of domestic appliances would be considerable. Second, the installation of one meter per household appliance would clearly cause substantial inconvenience to the household's occupants. Third, the system would require additional meters to be deployed should the set of appliances change (e.g. appliance replacements or the introduction of new appliances). Therefore, until such appliance metering is available at scale at low cost, intrusive metering should not be considered as a practical or scalable solution to the appliance energy monitoring problem (Armel et al., 2013).

Alternatively, non-intrusive metering can be used to disaggregate appliance energy from a single point of measurement. Such a system is commonly referred to as a non-intrusive appliance load monitor (NIALM). One approach is to design a meter specifically for appliance energy disaggregation, which is able to sample the household's electricity demand thousands of times per second, therefore allowing multiple electrical features to be extracted. These features can be used to easily discriminate between appliance power demands, therefore simplifying the disaggregation task. However, the financial and convenience cost of installing a bespoke meter in each household is still substantial relative to the benefits of appliance energy disaggregation (Armel et al., 2013).

An alternative approach would be to use an existing premises-level meter instead of installing a custom meter, therefore eliminating any additional financial cost or physical intrusion. Smart meters are an example of such premises-level meters which are installed by energy providers primarily for the purpose of automated meter reading. Many countries have declared the deployment of one smart meter per household in an effort to increase the accuracy of electricity billing and reduce overall domestic energy consumption. However, since the primary function of smart meters is to provide automated meter reading, they typically only transmit 30 minute consumption data to the utility provider, which is too course for disaggregation purposes. In contrast, many smart meters also transmit higher granularity data, such as 10 second consumption, over a home area network, which is ideal data for a NIALM system.

For example, the UK Department of Energy & Climate Change (2009) has announced the mandatory deployment of smart meters to all households before the end of 2020. Such smart meters will be deployed to each household along with an in-home display, which will be used to provide real-time usage and pricing information to a household's occupants (Department of Energy & Climate Change, 2009). While smart meters are required to transmit only 30 minute consumption data to the utility provider for billing purposes, the smart meters are also required to transmit at least 10 second data to the in-home display in order to provide real-time feedback (Department of Energy & Climate Change, 2013a). However, both the 30 minute data and 10 second data collected by smart meters only represent each household's aggregate power demand, and as such the smart meters do not provide any insight into the energy consumption of individual appliances. Although the 30 minute consumption data is too course to disaggregate into individual appliances, the 10 second usage data constitutes a realistic input for a NIALM system. Therefore, although smart meters are not able to perform disaggregation themselves, they provide an ideal data collection platform to support the content of this thesis.

However, the roll out of smart meters in the UK has faced significant problems. The expansion of the original specification (Department of Energy & Climate Change, 2012a) motivated the publication of a second version of the Smart Metering Equipment Technical Specification (Department of Energy & Climate Change, 2013a). As a result, the main installation phase has been delayed and only 177,700 smart meters have been deployed at the time of writing, in contrast to the planned 2 million smart meters. Consequently, the main installation phase is now scheduled to begin in 2015 rather than 2014, and end during 2020 rather than 2019 (Department of Energy & Climate Change, 2012b, 2013b).

Another form of intrusion is the collection of appliance training data within a specific household. Such training data could be collected by either installing individual appliance sub-meters for a short period of time, or the individual operation of each appliance followed by the manual labelling of each operation. However, installing sub-meters is clearly costly and also manual appliance operation is intrusive upon the household occupants, and as a result either process decreases the potential scalability of disaggregation solutions. Therefore, scalable disaggregation solutions should be capable of operating in an unsupervised manner.

To summarise, a requirements list defining the problem is given below:

1. **Appliance load monitor:** The approach must estimate individual appliance loads at the same level of granularity as the premises-level monitor.
2. **Non-intrusive:** The approach must only require consumption data to be collected from a single point of measurement.
3. **Low granularity data:** The approach must be able to monitor appliances given aggregate power measurements at 10 second intervals.
4. **Unsupervised disaggregation:** The approach must not require training data to be collected from the houses in which disaggregation will be performed.

Having introduced the problem of energy disaggregation, we now describe the typical scenario in which the technology will be required to operate.

1.2 Scenario Description

The complexity of the NIALM task depends largely upon the target household, which is affected by many factors. The two most important of which are the appliances and occupants of the household. This section discusses a typical scenario in which a NIALM would be expected to operate, and the monitoring techniques which would be used.

Zeifman and Roth (2011) estimate that a typical household contains 30–50 appliances. These appliances draw a wide variation of power (0–3000 W) and are in operation for different durations of time (0–24 hours per day). As a result, domestic appliances can consume vastly different amounts of energy. Figure 1.1 displays approximate figures for the average energy consumption per day for the most common appliance types. The figure collects appliances of the same type (e.g. multiple light bulbs) as would be expected in households, and consequently shows fewer appliances than the estimate of Zeifman and Roth (2011). The estimates are calculated using power demands of household appliances (MacKay, 2008) and scaled up using approximate durations of use. A full breakdown of the figures used is given in Appendix A. The shape of the graph appears to roughly follow an exponential distribution, in which the majority of the household's energy is consumed by relatively few appliances, specifically those which perform heating or cooling tasks. Therefore, it is most important for a NIALM to successfully disaggregate such high energy consuming appliances. Having described the typical households in which a NIALM will be required to operate, we now describe the potential applications which are enabled by disaggregated energy consumption data.

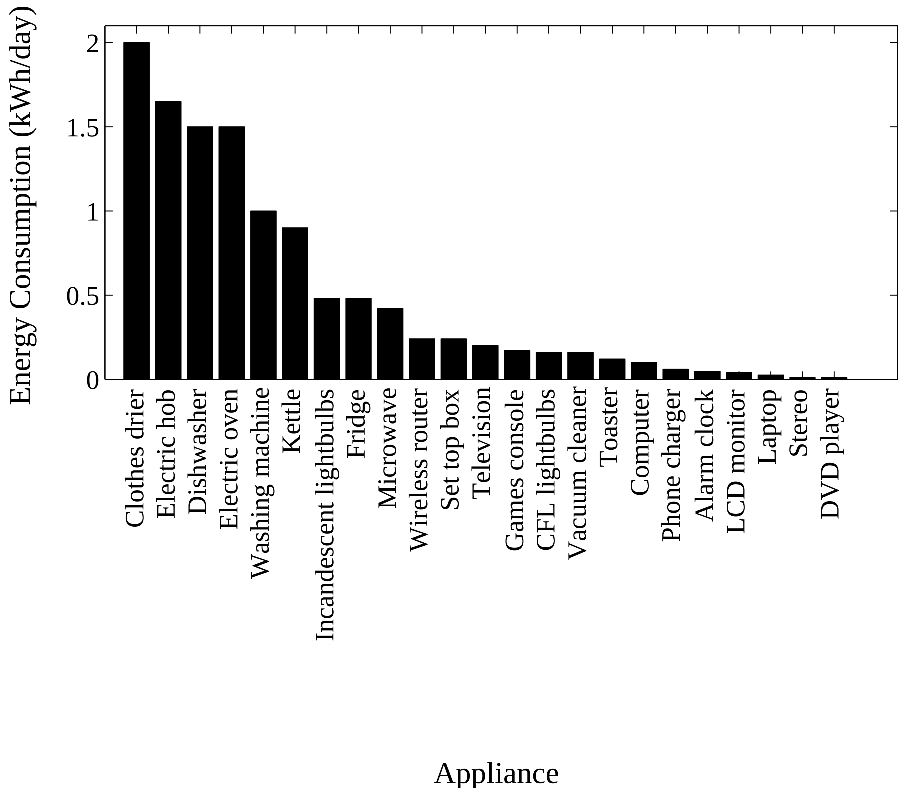


Figure 1.1: Average energy consumption of domestic appliances

1.3 Application Areas

Providing disaggregated real-time feedback has been found to reduce a household's electrical energy consumption by 9–18% (Ehrhardt-Martinez et al., 2010). Such a reduction in domestic energy consumption would clearly contribute to national goals of a reduction in carbon dioxide emissions. In addition, such an increase in the efficiency of domestic electricity consumption places a lower demand on electricity generation, and consequently a lower demand on the burning of fossil-fuels or international energy imports.

In addition to these national goals, individual consumers will also benefit financially from such reductions in electricity consumption. Furthermore, disaggregated energy feedback also has the potential to educate each household's occupants about the relative energy consumption of different appliances. This increase in awareness could also prompt energy savings in other domains, such as in commercial and industrial premises.

In order to realise such energy savings, disaggregated appliance data must be used to produce actionable suggestions which are then presented to a household's occupants. The remainder of this section discusses three types of such feedback.

First, disaggregated electricity consumption can be used to provide personalised suggestions regarding the mode of an appliance's use. The system would detect when an appliance is being used in a mode of poor energy efficiency and quantify the savings should the appliance be switched to a more energy efficient mode. Such suggestions do not prevent the household's occupants from carrying out their desired task, but instead allows them to make an informed decision regarding the mode of use of an appliance. Examples of such appliances are generally those with an economy or low power setting (e.g. an economy shower or a cool cycle of a washing machine).

Second, the disaggregated electricity consumption can also be used to provide automated load deferral suggestions. Since the mix of generators supplying electricity to the national grid varies with demand, so does the rate of carbon emissions. Load deferral is the act of delaying the use of electricity from a peak time to an off-peak time, therefore reducing the net carbon emissions despite the same amount of energy being used. An automated system could suggest the deferral of appropriate energy intensive appliances to off-peak times, and would therefore require information regarding the time of day when the appliance is used, the energy consumption of the appliance and the carbon intensity of the national grid throughout the day. Deferrable loads are generally appliances whose usage is not required to be performed immediately upon user interaction (e.g. running of a washing machine or dishwasher). Furthermore, with the introduction of time of use pricing or real-time pricing, load deferral can also decrease the overall cost of electricity for individual households.

Last, disaggregated electricity consumption could be used to detect faulty or deteriorating appliances (Laughman et al., 2003). Since the NIALM estimates the energy consumption of each appliance, a faulty appliance can be identified as either an appliance which draws significantly more power than the average for that appliance type or an appliance whose energy consumption increases over time. In such a situation, the automated system could suggest either a more energy efficient replacement for less expensive appliances or a repair for more expensive appliances. The system could even calculate how long it would take to break even after such a replacement or repair, compared to had the household occupants taken no action. Examples of such appliances are generally those which become less energy efficient throughout their lifetime (e.g. a refrigerator or oven with a deteriorating door seal). Having described the various applications which are enabled by disaggregated electricity data, we now consider existing solutions to the disaggregation problem.

1.4 Existing Solutions

A number of appliance monitoring methods exist which reduce the complexity of the disaggregation problem at the expense of a more substantial intrusion into the household. A summary of these methods is given below, and a full description is given in Chapter 2:

- **Electrical sub-metering:** Installing one electricity meter per appliance.
- **Smart appliances:** Appliances self-report energy consumption to a central hub.
- **Electrical probing:** Transmitting an electrical signal into the household mains circuit and analysing the return signal.
- **Appliance tagging:** Installing a low-cost appliance tag which detects usage through non-electrical methods.
- **Ambient sensors:** Using existing sensors such as occupancy, lighting and audio sensors to infer appliance usage.
- **Conditional demand analysis:** Using an appliance survey to estimate appliance energy consumption.

However, the intrinsically intrusive nature of these methods clearly violates Requirement 2, and as a result none of these methods constitute a solution for disaggregating smart meter data.

In an effort to avoid the deployment of such intrusive hardware throughout households, research has instead focused on single-point sensing, or non-intrusive monitoring. One non-intrusive approach is to sample a household's current and voltage at a high frequency (kHz), allowing various electrical features to be extracted, such as active and reactive power, current harmonics and voltage noise. These features provide a strong basis to discriminate between the usage of different appliances (Gupta et al., 2010; Kolter and Jaakkola, 2012). However, smart meters will not report such high frequency data, and therefore these methods violate Requirement 3.

An alternative type of non-intrusive monitoring is to use only low frequency data as the input to the appliance monitoring system. Due to the complexity of the problem, three common assumptions are often made regarding the knowledge about appliances within the household. First, some approaches assume it is possible to deploy individual appliance sub-meters for a short period of time to collect training data (Berges et al., 2008). Second, other approaches assume that the appliance models learned through an automatic training phase can subsequently be manually labelled (Kolter and Jaakkola, 2012). Third, other approaches assume knowledge of the number and types of appliances present in a household (Kim et al., 2011). None of this additional information will be collected during smart meter deployments, and such assumptions clearly violate Requirement 4.

In summary, existing research has ignored the scenario of disaggregating smart meter data through an entirely automated approach. This is exactly area of research we address in this work, and we discuss the potential research challenges in the following section.

1.5 Potential Challenges

There exist a number of technical and social challenges that must be overcome by any NIALM solution. This section will first discuss the technical issues which affect the performance of all energy disaggregation approaches. Second, the social challenges relating to data collection and transmission will be discussed.

Non-intrusive appliance monitoring uses unique appliance features detectable from a premises-level electrical meter to disaggregate a household's total energy consumption into individual appliances. However, in a typical household many appliances often operate simultaneously, resulting in the overlap of multiple appliances' features. As the number of frequently used appliances in a household increases, so does the complexity of the disaggregation problem as a result of the increased likelihood of feature overlap. Such feature overlap does not only obfuscate the features of low power appliances, but can also merge the features of multiple appliances to produce new features altogether.

Additional technical challenges exist due to the many variables affecting the energy consumption of appliances within a household. One such problem is that different households invariably contain a different set of appliances (e.g. electric or gas heating or cooking appliances). Furthermore, appliances demonstrate a considerable range even within the same appliance type (e.g. an old deteriorating refrigerator and a new energy efficient refrigerator). Last, the same models of appliances can consume different amounts of energy due to their usage by the household's occupants (e.g. washing machine used on 30° cycle and 60° cycle). Such variation can cause significant differences in the performance of NIALM systems between households. Consequently, it is a substantial challenge to build a NIALM that will perform satisfactorily in all environments.

Furthermore, the performance of NIALM solutions are also complicated by noise. Individual appliances contribute noise to the aggregate signal when their energy consumption deviates unpredictably from its expected value. In addition, variations in the supply voltage can contribute to inconsistent appliance features. Last, the premises meter contributes measurement noise when the aggregate energy measurement deviates from the sum of the energy consumption across all appliances. The compound effect each of these forms of noise further increases the difficulty of non-intrusive appliance monitoring.

In contrast to the technical challenges described so far, a number of social barriers have also prevented progress in the field of NIALM. One such barrier arises due to the intrinsic privacy concerns related to NIALM (Lisovich and Wicker, 2008). Given the time of use of each household appliance as produced by a NIALM, it might be trivial to infer an occupant's activities from this data. For instance, it might be possible to infer how often the occupants of a certain household take a shower or wash their clothes. Similar security concerns also arise due to the ease of occupancy inference from disaggregated appliance data. For instance, if a NIALM has detected that all the lights in a household are off during the night, it could be inferred that all occupants are either out of the house or are asleep. Some occupants might prefer that any information of this type could not be deduced from their energy data, motivating the use of privacy-preserving smart metering systems. However, these concerns can be minimised by requiring the consumer to either opt-in to the use of the NIALM system in the case that disaggregation is performed remotely, or even perform the appliance disaggregation at the same location as the smart meter. This work will consider the use case for an in-home NIALM system, therefore circumventing such privacy and security barriers. Having discussed the potential challenges within the domain of energy disaggregation, we now summarise the research contributions of this thesis.

1.6 Research Contributions

Against this background of existing work, our research objective is as follows:

To develop methods by which the total power demand of a household can be disaggregated into individual appliances. The granularity of the disaggregated power demand should match that of the readings from the single smart meter used to monitor the household's aggregate power demand. The disaggregation method should not require any further information regarding the appliances present in the household.

Our contributions towards this objective can be summarised as follows:

1. We adopt a hierarchical Bayesian framework for modelling appliances using hidden Markov models (HMMs), and show how existing appliance data sets, such as the Tracebase data set, can be used to learn appliance models which generalise across households. Our results show that generalisable appliance models can be learned from relatively few examples of an appliance type, and such generalisable models outperform models learned from a single appliance in cross validation tests.
2. We show how these general appliance models can be tuned to represent the appliance within a specific household using only smart meter data from that household. We evaluate this approach using the Reference Energy Disaggregation Data set (REDD), and our results show that tuned appliance models provide improved

performance over general appliance models, and furthermore that they can produce comparable performance to appliance models learned from sub-metered appliance data. Last, we show that our approach outperforms the state of the art which uses a factorial hidden Markov model (FHMM) to tune model parameters.

3. We present a large scale application of our approach to demonstrate its flexibility and give examples of the potential feedback that could be presented to the households' occupants. We use the Colden Common data set to evaluate the accuracy by which our approach can estimate the energy consumption of refrigerators and freezers across 117 households. Our results show that in most households the appliances' energy consumption can be estimated accurately enough to provide actionable energy saving suggestions to the household's occupants.

These contributions are also detailed in the following three papers:

1. Parson O, Ghosh S, Weal M, Rogers A. An Unsupervised Training Method for Non-intrusive Appliance Load Monitoring. In: Artificial Intelligence (under review). Springer; 2013.
2. Parson O, Ghosh S, Weal M, Rogers A. Non-intrusive Load Monitoring using Prior Models of General Appliance Types. In: 26th Association for the Advancement of Artificial Intelligence Conference. Toronto, Ontario, Canada; 2012.
3. Parson O, Ghosh S, Weal M, Rogers A. Using Hidden Markov Models for Iterative Non-intrusive Appliance Monitoring. In: Neural Information Processing Systems, Workshop on Machine Learning for Sustainability. Sierra Nevada, Spain; 2011.

Having summarised our research contributions, we now describe the structure of this thesis.

1.7 Thesis Structure

The remaining chapters of this thesis are structured as follows:

Chapter 2 provides a background of theoretical and empirical research relevant to non-intrusive appliance load monitoring. First, intrusive monitoring methods are discussed. Second, approaches based on high frequency electricity monitors are considered. Last, low frequency methods are introduced, with particular focus given to HMM-based approaches.

Chapter 3 gives a description of the various data sets available for evaluating NIALM systems. First, we describe the public data sets that are available, and highlight the advantages and disadvantages of each data set with respect to investigating the accuracy of disaggregation algorithms. Second, we give an overview of the private data sets which have been used in related work, and motivate the use of our own private data set to evaluate the robustness of our approach. Last, we provide a table which summarises the important attributes of each data set.

Chapter 4 proposes a method by which appliance models can be learned which generalise to new households. First, we describe a hierarchical Bayesian method for modelling appliance types using HMMs. Second, we show how generalisable appliance models can be learned using the Tracebase data set. Last, we use this data set to evaluate the performance of such generalisable appliance models.

Chapter 5 introduces a novel approach for tuning general appliance models to the appliances in a specific household using only smart meter data. First, we show how the hierarchical Bayesian framework provides an elegant approach to extract individual appliance signatures from an aggregate load. Second, we show how Bayesian inference can be used to update the general appliance models using the extracted signatures. We then evaluate our tuning approach against the state of the art using the REDD data set.

Chapter 6 demonstrates a case study application of the methods described in Chapter 4 and Chapter 5 to a large scale data set. We start by motivating the use of large scale data sets, such as the Colden Common data set, to evaluate energy disaggregation methods. We then describe how our methods can be used to infer the energy efficiency of combined fridge freezers for households containing a single cold appliance. Last, we extend our approach to allow it to be applied to households containing separate refrigerators and freezers, and evaluate the accuracy by which it can infer such appliances' energy efficiency.

Finally, Chapter 7 gives a summary of the research presented in this thesis and the conclusions that can be drawn from each chapter. We also discuss future extensions of the work presented in this thesis, with specific attention paid to realistic accuracy improvements and the potential for further user feedback. Having introduced the research presented in this thesis in this chapter, the following chapter gives a background of the existing relevant work.

Chapter 2

Background

This chapter gives a background of various existing approaches which aim to disaggregate a household's total energy consumption into individual appliances. We begin by describing intrusive methods for appliance monitoring, and highlight their intrinsic disadvantage of poor scalability. We then move on to non-intrusive monitoring methods and introduce high frequency based approaches. However, such methods require bespoke hardware to be installed within homes since smart meter data is of insufficient granularity. Next, we discuss low frequency event based methods, although such approaches inherently consider all appliance switch events to be independent, and as a result have poor sensitivity to errors. Last, we identify low frequency non-event based approaches as the most promising direction of NIALM research. We give a description of their theoretical foundation in temporal graphical models, before describing their previous applications to energy disaggregation. However, we conclude that their training requirements make many unrealistic assumptions regarding the available information about each household's appliances, and highlight this as the field of research explored in this thesis.

2.1 Intrusive Monitoring

Intrusive monitoring refers to the deployment of multiple hardware sensors throughout a household. Such intrusive methods can be further divided into direct and indirect monitoring methods. Direct monitoring methods measure the electrical characteristics of each appliance's power demand. In contrast, indirect methods measure non-electrical characteristics, from which each appliance's power demand is inferred. We give a discussion of both direct and indirect methods in Section 2.1.1 and Section 2.1.2 respectively, and highlight the reasons why neither category of methods is an appropriate solution for the disaggregation of smart meter data.

2.1.1 Direct Monitoring

This section describes three forms of direct intrusive monitoring: electrical sub-metering (Section 2.1.1.1), smart appliances (Section 2.1.1.2) and electrical probing (Section 2.1.1.3). We discuss the various costs involved with each method, and give reasons why each approach is not a suitable solution to the smart meter disaggregation problem.

2.1.1.1 Electrical Sub-metering

Electrical sub-metering refers to the installation of a system in which individual appliances are monitored directly using one meter per appliance. The appliance meters typically take the form of a plug-in meter or a clamp-on meter. Plug-in meters are installed by plugging the appliance into the meter, and plugging the meter into an electrical outlet. This allows the meter to both monitor the appliance and control the flow of electricity between the mains circuit and the appliance. Alternatively, clamp-on meters can be installed without breaking the electrical circuit, by attaching a clamp around a lightly insulated positive or neutral wire. The power drawn by the appliance can be calculated by measuring the electromagnetic field generated by the flow of current through the wire. The combination of plug-in and clamp-on meters allow appliances that are either plugged in to an electrical outlet or hard-wired in to the mains circuit to be monitored.

Although both plug-in and clamp-on meters allow accurate measurements to be made of the energy consumed by an appliance, they have many practical disadvantages. The significant cost and time required per installation are often cited as reasons why this approach is impractical to deploy for a large user base (Laughman et al., 2003; Berges et al., 2008; Yi-xin et al., 2008). Therefore, the use of electrical sub-meters for appliance monitoring will not be considered further in this work. We now discuss smart appliances as an alternative form of direct monitoring.

2.1.1.2 Smart Appliances

Smart appliances can be used to self-report their energy consumption to a central hub, therefore circumventing the issue of installing additional monitoring equipment. Such smart appliances would therefore need to be fitted with a wireless enabled energy monitoring module (e.g. the ZPM3570[1]). However, older appliances would need to be either replaced or retro-fitted in order to self-report their energy consumption. Replacing every domestic appliance is clearly prohibitively expensive, while retrofitting appliances incurs the same disadvantages as appliance sub-metering. The turnover of domestic appliances is generally quite slow, as most appliances can only be expected to be replaced if the

[1] `http://www.rfm.com/`

old appliance is faulty. Therefore, it would take many decades for most appliances to be replaced through this cycle. This is way beyond the 2020 target for the roll out of smart meters in the UK (Department of Energy & Climate Change, 2009), and as a consequence smart appliances will not be considered as a complete solution to the non-intrusive monitoring problem.

It is worth noting that both smart appliances and NIALM systems could cooperate. In such a scenario, each smart appliance could report its energy consumption to the NIALM system. The NIALM could then subtract each smart appliance's power demand from the household aggregate power demand prior to performing disaggregation, therefore simplifying the disaggregation task for the remaining appliances. However, this would require the standardisation of energy consumption reporting that does not yet exist. Having ruled out a complete deployment of smart appliances, and shown that a partial roll out would only slightly simplify the disaggregation problem, we now discuss disaggregation via electrical probing.

2.1.1.3 Electrical Probing

Electrical probing is the process of transmitting a signal into a household's electrical circuit and using features extracted from the returned signal to classify the loads currently in use (Hart, 1992). Electrical probing is not intrusive in the physical sense (as with sub-metering), but is instead intrusive upon the household's electrical circuit. However, electrical probing inherently adds interference to the electrical circuit, which can adversely affect the power quality delivered to each appliance. As a result, energy disaggregation by electrical probing has not been reported in the literature since it was first suggested by Hart (1992). For these reasons, electrical probing will no longer be considered as a solution to NIALM in this report. Having discussed three direct forms of monitoring, we now move on to indirect monitoring methods.

2.1.2 Indirect Monitoring

This section describes three forms of indirect intrusive monitoring: appliance tagging (Section 2.1.2.1), ambient sensors (Section 2.1.2.2) and conditional demand analysis (Section 2.1.2.3). We discuss the various costs involved with each method, and give reasons why each approach is not a suitable solution to the smart meter disaggregation problem.

2.1.2.1 Appliance Tagging

Appliance tagging refers to the modification of an appliance such that a tag emits a unique signal when the appliance turns on or off. These signals are detected by a central hub which estimates each appliance's energy consumption. McWilliam and Purvis (2006) demonstrate the use of transmitting RFID signals through the live mains circuit to a central recorder in order to uniquely identify appliances. However, this approach requires the customisation of each individual appliance in addition to the installation of a central signature detector. The installation time and cost per household of this method is considerable and will therefore not be considered further in this work. Having dismissed appliance tagging as a reasonable solution, we now consider the use of ambient sensors.

2.1.2.2 Ambient Sensors

Multiple wireless sensors could be used to monitor feeds other than electricity in order to disaggregate premises-level power measurements into individual appliances (Kim et al., 2009; Schoofs et al., 2010). Examples of such sensors include audio, temperature and light sensors, which could be used to monitor both human behaviour and appliance operation. As with appliance tagging, this approach requires the intrusive installation of multiple sensors throughout each household, and therefore will not be considered further in this work. Since ambient sensors do not provide a suitable solution, we now discuss the use of conditional demand analysis.

2.1.2.3 Conditional Demand Analysis

Unlike other approaches requiring the installation of additional meters, conditional demand analysis (CDA) uses only a household's billed energy consumption. In addition, CDA also requires information about the consumer, household and weather. Such data from many households are analysed using a multivariate regression technique to learn the typical contribution of individual appliances (Tiedemann, 2007). CDA can then be used estimate the energy consumption of domestic appliances. Again, the lack of equipment installation makes this a non-intrusive approach in the traditional metering sense. However, CDA requires a large participant base, in which each participant must complete a detailed questionnaire; an example of a social intrusion. Furthermore, CDA does not capture unusual cases which are not accounted for by such questionnaires, e.g. a day when the washing machine has been run three times. Therefore, this method will not be considered further in this work.

Having ruled out intrusive monitoring methods as appropriate solutions to the problem of smart meter disaggregation, we now turn to non-intrusive methods.

2.2 Non-intrusive Monitoring

We consider non-intrusive appliance monitoring as the disaggregation of a household's appliances from the total load through a single point of measurement. In this section, we first give a brief history of the field, before describing non-intrusive methods based on high frequency data, which are capable of disaggregating household energy consumption to a high degree of accuracy. However, smart meters are not capable of reporting such high frequency data, and as a result such methods would require the installation of additional hardware to each household. This is followed by a description of non-intrusive methods which make use of low frequency data, in which we highlight a direction of research with the potential to solve the smart meter disaggregation problem.

2.2.1 History

Hart (1992) first introduced the field in his seminal work, which outlined a set of principles NIALM algorithms should follow, a taxonomy of potential approaches, a set of features that such approaches could use to discriminate between appliances and the use of finite state machines to model appliances. Although Hart didn't pursue the problem of energy disaggregation much further, the concepts introduced in this work have since been consistently echoed by the literature.

Hart and Bouloutas later published an a theoretical method by which two appliances could be disaggregated via an approach based on the Viterbi algorithm (Hart and Bouloutas, 1993), although it was never applied to energy disaggregation in practice. This work laid the foundations for what would become known as non-event based monitoring, which describes the application of probabilistic temporal graphical models to the area of energy disaggregation, as discussed later in Section 2.2.3.3.

The field of energy disaggregation drew limited attention over the subsequent 15 years, until it received renewed interest as a result of decreasing hardware costs, expanding connectivity infrastructure, and most recently, national roll outs of smart electricity meters. Such factors have contributed to the formation of a community of researchers to establish the field in its own right. Since 2011, a number of public data sets designed specifically for energy disaggregation have been released (as described in Chapter 3), 2 international workshops have been held (NILM 2012, EPRI NILM 2013), and a toolkit has been released enabling the empirical comparison of various energy disaggregation algorithms (Batra et al., 2014). We now go on to describe developments in the field which rely on high frequency sampling.

2.2.2 High Frequency Sampling

We consider high frequency sampling as the measurement of electrical characteristics at a rate greater than 1 Hz. By sampling the current and voltage thousands of times per second, various electrical features can be calculated. Most commonly, real and reactive power are calculated from current and voltage readings over one cycle of the alternating current waveform. Hart (1992) first showed that such features could be used to discriminate between appliances of equal apparent power demand. Since, much research has applied various classification methods to such electrical features in order to disaggregate appliances (Sultanem, 1991; Shaw et al., 2008; Gonçalves et al., 2011; Marchiori et al., 2011).

In addition, Hart (1992) also demonstrated that certain appliances generate non-sinusoidal waveforms, and consequently create significant low-order odd harmonics. Such harmonic content of an aggregate load can also be used to accurately discriminate between appliances (Patel et al., 2007; Berges et al., 2010; Chang et al., 2010). Furthermore, Gupta et al. (2010) have shown that appliances' switch mode power supplies create frequency peaks at non-harmonic frequencies, referred to as switching frequencies. Appliance disaggregation based on switching frequencies can achieve even greater accuracy than harmonic based disaggregation, since switching frequencies are often unique to each appliance while harmonic frequencies are always multiples of the power line's fundamental frequency.

Last, Froehlich et al. (2011) have shown that the high frequency voltage noise generated by appliances as they switch on or off can be used to identify individual appliances. Since such transient voltage noise typically lasts only a few microseconds, these transients are unlikely to overlap, and as a result can discriminate between appliances with similar continuous power and frequency components. Furthermore, Sanquer et al. (2013) have shown that a hierarchical Bayesian framework can be used to extract features which generalise over multiple transient signals from a single appliance class.

However, although smart meters typically sample a household's current and voltage internally at a high frequency, only low frequency power is reported externally by the household's smart meter. As a result, each of these high frequency based approaches would require the installation of additional hardware into each household. This would clearly violate Requirement 3, and as a result will not be considered further in this work. We now move on to discuss approaches based on low frequency sampling.

2.2.3 Low Frequency Sampling

In contrast to high frequency sampling, we consider low frequency sampling as the reporting of household's electrical features at a rate less than 1 Hz. Smart meters belong to this category, since they will typically only report power at 10 second intervals. We now discuss low frequency methods in more detail, first covering event based methods in Section 2.2.3.1 and those based on blind source separation techniques, before giving an introduction to non-event based methods in Section 2.2.3.3.

2.2.3.1 Event Based Methods

Event based disaggregation methods aim to classify appliance switch events (e.g. a microwave turning on or off) using a set of features which can be immediately extracted from the power load. For low frequency methods, such features are generally the difference between the steady power demands before and after the switch event, and the duration of the switch event. However, since UK smart meters only report the power demand at 10 second intervals, the duration of each appliance's switch event will almost always be less than the sampling interval. As a result, the switch event duration cannot be used to discriminate between appliances, and therefore only the step change in power can be used.

Furthermore, event based approaches either consider each appliance switch event as independent, or make local classifications based on fixed previous classifications. In the first case of independent classification, the step change in power alone often does not provide enough information to produce an accurate classification. In the second case of local classifications, earlier incorrect classifications can 'lock' the algorithm into an incorrect event sequence (Kolter and Jaakkola, 2012).

As a result of these disadvantages, event based methods have focused only on the disaggregation of sequences of sampling rates of 1 Hz or greater (e.g. Berges et al., 2011), and have not been applied to power sequences of 0.1 Hz sampling rates as will be reported by UK smart meters. Therefore, we will not consider event based approaches further in this work, and move on to discuss methods based on blind source separation in the following section.

2.2.3.2 Blind Source Separation

Blind source separation aims to separate a set of mixtures of sources into a set of individual sources (Comon and Jutten, 2010). A classic example of blind source separation is that of speaker diarisation, in which multiple microphones are placed in a room containing multiple speakers, and the aim is to estimate when each speaker is speaking throughout the set of audio recordings. In the domain of energy disaggregation, the sources correspond to the appliances within a household and the mixtures correspond to electrical measurements taken at a single point of measurement. In the scenario in which smart meters are used as the measurement hardware, only a single mixture is observed (the household aggregate power demand), and as such the problem is severely underdetermined; there exist more unobserved sources than observed mixtures. This is in contrast to the typical scenarios in which blind source separation is applied, in which the number of mixtures is close to the number of sources, for example, the separation of two mixtures of three speech signatures (Lee et al., 1999). Furthermore, blind source separation techniques are typically applied to scenarios in which little or no information is available regarding the structure of sources or the mixing process. Again, this is in contrast to energy disaggregation, in which rich prior information is available regarding the behaviour of appliances and the mixing process is known, although sub-metered data from individual appliances in each household is rarely available to directly learn the structure of such appliances.

Kolter et al. (2010) proposed an approach for energy disaggregation via discriminative sparse coding, in which appliances are represented using a set of basis functions, and disaggregation is accomplished by finding a sparse set of activations which explain the household aggregate data. Crucially, the approach learns general appliance models from appliance data collected from households other than the test household in which disaggregation is performed. The authors then apply non-negative matrix factorisation to solve an optimisation problem in order to disaggregate appliances. This approach models sequential time slices independently, and as such this method is best applied to very low frequency data (e.g. data collected at 15 minute intervals). However, this approach is likely to ignore the strong dependency between sequential measurements taken at higher frequencies (e.g. 10 second intervals), and therefore will likely perform poorly when applied to the disaggregation of smart meter data in our scenario.

Dong et al. (2013) applied a similar approach based on discriminative sparse coding to the disaggregation of domestic water consumption data. However, the approach suffers from the same core disadvantage; that the approach does not exploit the strong dependencies between sequential readings taken at 10 second intervals, and as such is not well suited to the disaggregation of electrical data collected by a smart meter. However, it should be noted that the authors proposed a recursive technique, in which appliances are iteratively separated from the household aggregate data. Such an approach is particularly interesting to electricity disaggregation, given the complexity of modelling a large number of potentially unknown household appliances, but the vast majority of household energy can typically be accounted for by less than 10 appliances.

The approaches drawn from the blind source separation field discussed in this section share a common disadvantage; that such approaches do not exploit the dependencies between sequential measurements, and as such will not be considered further by this work. However, it is exactly this disadvantage that motivates the study of non-event based methods in the following section.

2.2.3.3 Non-event Based Methods

In contrast to event based methods, non-event based methods do not require a separate event detection process. Instead, event detection is integrated directly into the disaggregation model. All existing non-event based disaggregation methods use temporal graphical models to represent the event detection and disaggregation problems using a single probabilistic framework. Section 2.3 introduces the theory of relevant temporal graphical models, while Section 2.4 describes how related works have applied such models to energy disaggregation.

2.3 Temporal Graphical Models

This section introduces a class of probabilistic graphical models which address the shortcomings of the event based approaches discussed in Section 2.2.3.1. Such probabilistic graphical models have previously been applied to a number of real world problems, the prototypical example being speech recognition (Rabiner, 1989). Speech recognition shares a number of similarities with energy disaggregation, in that the aim is to identify the most likely sequence of discrete states (words) corresponding to a time series of continuous measurements (audio recordings).

However, with energy disaggregation, the aim is not to classify the operation of only a single appliance, but instead to classify the operation of a number of simultaneously operating household appliances given a time series of power measurements. The field of speech recognition has since expanded to address similar problems of simultaneous classifications, such as speech recognition with non-stationary noise or multiple speakers (Virtanen, 2006). A key difference between such domains and appliance monitoring is that these domains generally consider the classification of a small number of simultaneous sources of noise or speech (e.g. 2 or 3), whereas energy disaggregation methods must be robust to large numbers of simultaneous sources (e.g. 20 or more). As a result, similar assumptions of model scalability cannot be made, and consequently solutions to speech recognition problems are rarely applicable to the problem of energy disaggregation.

For each model discussed in this section, a figure will be given representing its graphical structure, where nodes represent random variables and directed edges represent a probabilistic dependency between two variables. In each graph, square nodes correspond to discrete random variables which can take on the value of one of a finite set of values, while circular nodes correspond to continuous random variables which can take on the value of any real number. In addition, shaded nodes correspond to observed random variables (system inputs), while unshaded nodes correspond to hidden (latent) variables (system outputs).

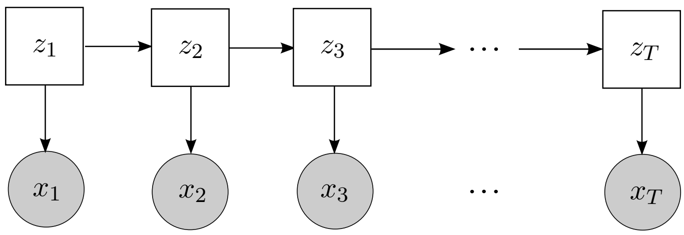


Figure 2.1: Hidden Markov model

Sections 2.3.1 – 2.3.6 define the models that have been successfully applied to speech recognition in abstract terms, while Section 2.3.7 describes three of the most common approximate inference methods used for temporal graphical models.

2.3.1 Hidden Markov Models

The simplest representation of sequential data is through the use of a Markov model. A Markov model is a sequence of random variables in which each variable is dependent upon only the variable immediately preceding it.

A hidden Markov model (HMM) is a Markov model in which the sequence is made up of discrete variables. In addition, each discrete variable emits a single continuous variable, which is dependent upon the value of the discrete variable. Furthermore, in a HMM the chain of discrete variables is not observed, while the continuous variables are observed. Figure 2.1 shows the graphical structure of a HMM, where the discrete, hidden variables are represented by the sequence $\mathbf{z} = z_1, \ldots, z_T$, and the continuous, observed variables are represented by the sequence $\mathbf{x} = x_1, \ldots, x_T$, where T is the length of the sequence (the number of time slices in the model). The value of each discrete variable z_t corresponds to one of K states, while each continuous variable can take on the value of any real number.

The behaviour of a HMM can be completely defined by three parameters. First, the probability of each state of the hidden variable at $t = 1$ can be represented by the vector $\boldsymbol{\pi}$ such that:

$$\pi_k = p(z_1 = k) \tag{2.1}$$

Second, the transition probabilities from state i at $t-1$ to state j at t can be represented by the matrix $\mathbf{A}$ such that:

$$A_{i,j} = p(z_t = j | z_{t-1} = i) \tag{2.2}$$

Third, the emission probabilities for $\mathbf{x}$ are described by a function governed by parameters $\boldsymbol{\phi}$, which is commonly assumed to be Gaussian distributed such that:

$$x_t|z_t, \boldsymbol{\phi} \sim \mathcal{N}(\mu_{z_t}, \tau_{z_t}) \tag{2.3}$$

where $\boldsymbol{\phi} = \{\boldsymbol{\mu}, \boldsymbol{\tau}\}$, and μ_{z_t} and τ_{z_t} are the mean and precision of a state's Gaussian distribution.

Equations 2.1, 2.2 and 2.3 can be used to calculate the joint likelihood of a HMM:

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = p(z_1|\boldsymbol{\pi}) \prod_{t=2}^{T} p(z_t|z_{t-1}, \mathbf{A}) \prod_{t=1}^{T} p(x_t|z_t, \boldsymbol{\phi}) \tag{2.4}$$

where the set of all model parameters is represented by $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\phi}\}$.

There exist two common goals when applying a HMM to a real world problem. First, one aim is to infer the model parameters $\boldsymbol{\theta}$ given a sequence of continuous variables $\mathbf{x}$. Second, another aim is to determine how the model parameters $\boldsymbol{\theta}$ and a sequence of continuous variables $\mathbf{x}$ can be used to infer the optimal sequence of discrete states $\mathbf{z}$. These problems will be referred to as learning (or training) and inference respectively, and the computational complexity of each will be discussed in the remainder of this section.

Learning in the context of HMMs refers to finding values for the model parameters which best explain the training data. The state of the art in terms of maximum likelihood estimation is the expectation maximisation (EM) algorithm (also known as the Baum-Welch algorithm) (Rabiner, 1989). However, approximate methods such as those described in Section 2.3.7 are more commonly used due to their support for fully Bayesian inference.

Inference in the context of HMMs refers to finding values for hidden variables which maximise the model's joint likelihood. The Viterbi algorithm is the state of the art in this context, and provides an efficient solution to the problem with complexity that scales linearly in the length of the chain (Viterbi, 1967). The Viterbi algorithm considers each time slice in sequence, and evaluates the probability of each transition from the previous time slice to the current time slice. However, only the transition with the maximum probability leading to each state in the current time slice is retained. By propagating the maximum probability of each state forwards through the subsequent time slices, the algorithm guarantees that the most probable sequence of states will be retained. When the maximum probabilities of each of the state has been propagated to the final time slice, the probability of each sequence is known. Since a unique transition leading to each state was retained, a single path exists from the most probable state in the final time slice backwards through each previous time slice representing the most probable joint sequence of states. The complexity of exact inference in a HMM is $O(K^2T)$, where K is the number of states and T is the number of time slices.

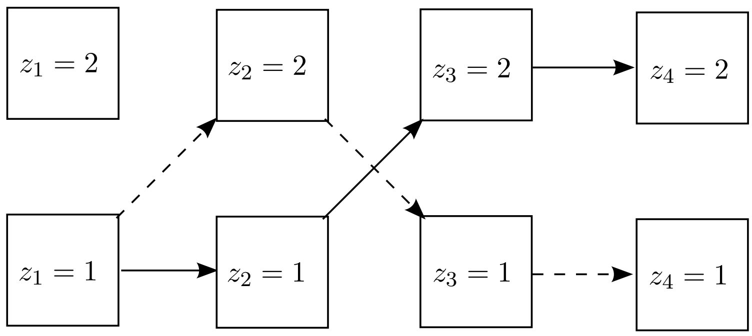


Figure 2.2: Sequences retained by Viterbi algorithm

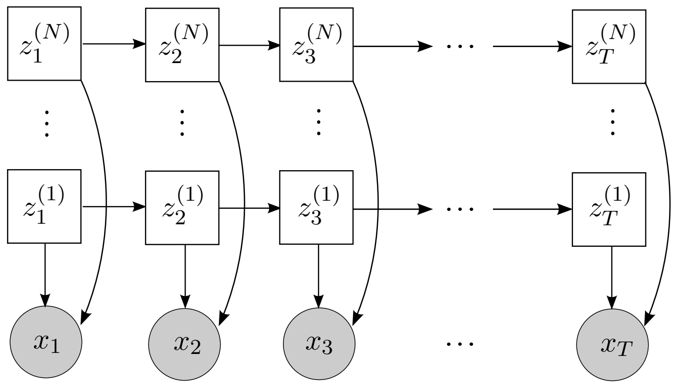


Figure 2.3: Factorial hidden Markov model

An example of the sequences (shown by dashed and dotted lines) that might be retained in a two state HMM by the Viterbi algorithm is shown by Figure 2.2. In the final time slice, $t = 4$, the sequence with the highest probability can be traced back to the first time slice, $t = 1$.

2.3.2 Factorial Hidden Markov Models

The factorial hidden Markov model (FHMM) is an extension of the HMM in which there are multiple independent Markov chains of hidden variables, $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(N)}$, where N is the number of chains. In this model, each observation is dependent upon multiple hidden variables (Ghahramani and Jordan, 1997). The graphical model of a FHMM is given by Figure 2.3.

Similar to Equation 2.4, the joint likelihood of a FHMM can be calculated by:

$$p(\mathbf{x}^{(1:N)}, \mathbf{z}|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(z_1^{(n)}|\boldsymbol{\pi}) \prod_{t=2}^{T} \prod_{n=1}^{N} p(z_t^{(n)}|z_{t-1}^{(n)}, \mathbf{A}) \prod_{t=1}^{T} p(x_t|z_t^{(1:N)}, \boldsymbol{\phi}) \tag{2.5}$$

where $1 : N$ represents a sequence of appliances $1, \dots, N$.

However, the computational complexity of both learning and inference is greater for FHMMs compared to HMMs. This is due to the conditional dependence of the Markov

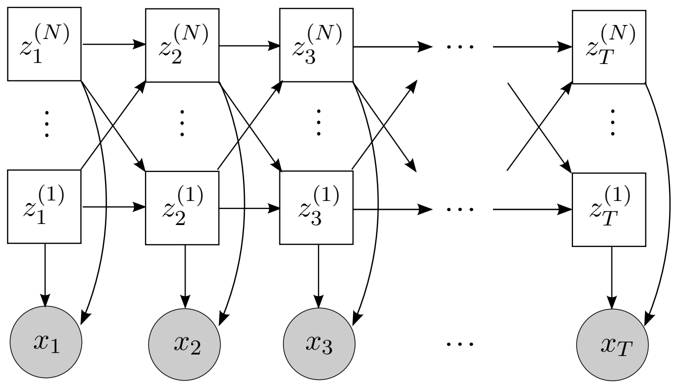


Figure 2.4: Conditional factorial hidden Markov model

chains given the observed variables. There are two possible solutions to perform learning and inference in a FHMM. The first is to transform the FHMM into a large HMM and perform learning and inference as discussed in Section 2.3.1. The alternative is to keep the factorial structure of the graphical model but use approximate techniques for inference. We now discuss the transformation to a HMM in more detail, while a full discussion of approximate techniques is given in Section 2.3.7.

The FHMM can be transformed into an equivalent HMM, which will allow standard HMM inference methods to be applied to the model. This can be achieved by using a single Markov chain with K^N states, one for each combination of states in the FHMM, resulting in a computational complexity of $O(K^{2N}T)$ for exact inference. Since the computational cost is clearly exponential in the number of chains, N, the model will therefore become computationally intractable for large N. Alternatively, approximate methods provide more tractable inference methods, as will be described in Section 2.3.7.

2.3.3 Conditional Factorial Hidden Markov Models

The conditional factorial hidden Markov model (CFHMM) is an extension of the FHMM, in which the state of each hidden variable is additionally dependent on the state of each variable of all other Markov chains in the previous time slice. For example, the variable $z_t^{(n)}$ would be dependent upon variables $z_{t-1}^{(\neg n)}$, in addition to $z_{t-1}^{(n)}$, where $\neg n$ represents the set of appliance indices excluding n. Figure 2.4 shows a graphical representation of the CFHMM.

Interestingly, learning and inference can be applied to a CFHMM in the same way as for a FHMM. Exact inference algorithms can still be applied by first transforming the CFHMM into an equivalent HMM with K^N states. Alternatively, approximate techniques such as sampling methods can be applied to CFHMMs at little extra computational cost to FHMMs.

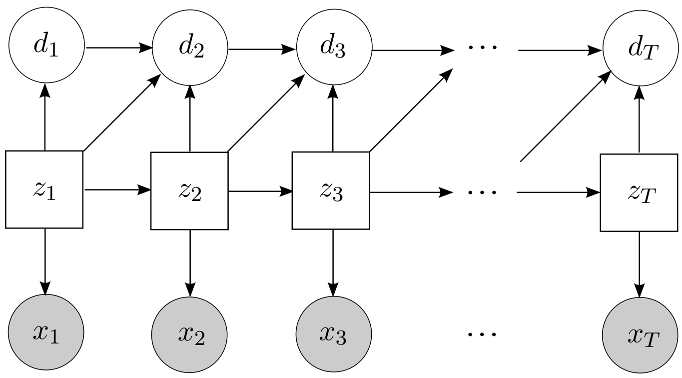


Figure 2.5: Hidden semi-Markov model

2.3.4 Hidden Semi-Markov Models

A hidden semi-Markov model (HSMM) is another extension of the HMM, in which each discrete variable is additionally dependent upon the number of time slices since it changed state (Yu, 2010). For example, variable z_t would be dependent upon the number of time slices since z_{t-1} changed state, in addition to variable z_{t-1} itself. One way to model a HSMM is to use a variable duration HMM, in which the duration of each state z_t is explicitly represented as an additional variable d_t, as shown by Figure 2.5.

The modelling of such temporal dependencies is often essential when accurately applying sequential data models to real world problems. For instance, the benefits of modelling state duration when performing parameter learning and inference tasks have been shown in both the areas of speech recognition (Ramesh and Wilpon, 1992) and handwriting recognition (Chen et al., 1995). Such methods are particularly preferable to traditional HMMs in these areas, as they allow chained variables to take on distributions other than geometric distributions (e.g. Poisson distribution), and therefore enable the state transition probabilities to depend on the current duration of that state.

Murphy (2002) and Yu (2010) give overviews of various representations of HSMMs. Both cite the computational cost of exact inference as $O(K^2DT)$, where K in the number of states, T is the number of time slices and D is the maximum duration of each state. However, such inference is clearly not tractable for HSMMs in which the duration of each variable's state is large, and instead approximate methods must be used.

2.3.5 Input-Output Hidden Markov Models

The input-output hidden Markov model (IOHMM) is another extension of the HMM. In an IOHMM, an additional sequence of variables $\mathbf{u}$ is introduced which the hidden Markov sequence $\mathbf{z}$ is dependent upon (Bengio and Frasconi, 1994). Figure 2.6 shows a graphical representation of an IOHMM.

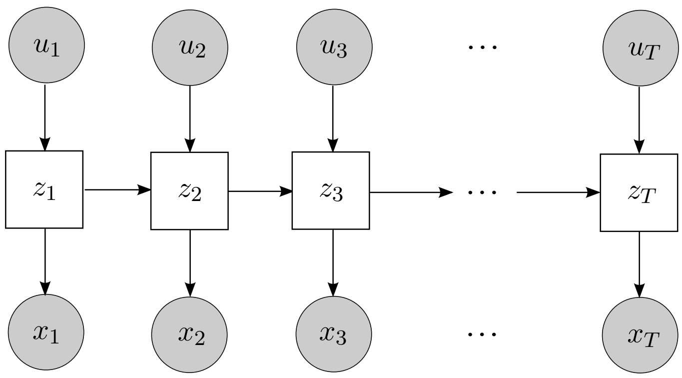


Figure 2.6: Input-output hidden Markov model

IOHMMs have the benefit that additional information that might have impacted upon the hidden variables' states can be integrated into the model at little extra computational cost. As a result, temporal dependencies external to the Markov chain (e.g. time of day) can be used during inference of the optimal hidden state sequence. Bengio and Frasconi (1994) describe how EM can be extended to allow the learning of an IOHMM's model parameters. In addition, the Viterbi algorithm can be trivially extended to IOHMMs by using the following joint likelihood function for such a model:

$$p(\mathbf{u}, \mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = p(z_1|u_1, \boldsymbol{\pi}, \mathbf{B}) \prod_{t=2}^{T} p(z_t|z_{t-1}, u_t, \mathbf{A}, \mathbf{B}) \prod_{t=1}^{T} p(x_t|z_t, \boldsymbol{\phi}) \tag{2.6}$$

where $\mathbf{B}$ is a vector of parameters governing the probability distribution $p(z_t|u_t)$. As with standard HMMs, the computational complexity is still $O(K^2T)$.

2.3.6 Difference Hidden Markov Models

The difference hidden Markov model (DHMM) is a trivial extension of the HMM, in which each observation is dependent upon the hidden variables in both the current time slice and the immediately preceding time slice (Kolter and Jaakkola, 2012). This is because each observation is derived from the difference between two consecutive readings, as given by $\Delta x_t = x_t - x_{t-1}$. For instance, in a DHMM, variable Δx_t would be dependent upon variables z_t and z_{t-1}, as illustrated by Figure 2.7. DHMMs have the advantage that a difference observation sequence $\Delta\mathbf{x}$ can be explicitly incorporated while the states of the hidden sequence $\mathbf{z}$ can be maintained. An example of such a difference observation sequence is the daily increase or decrease of a stock in a financial market.

As with standard HMMs, the computational complexity of exact inference in DHMMs is $O(K^2T)$. This is because no extra transitions are required to be evaluated during the forward pass of the Viterbi algorithm, although all K^2 transitions will need to be retained in each time slice rather than just the K transitions for a standard HMM.

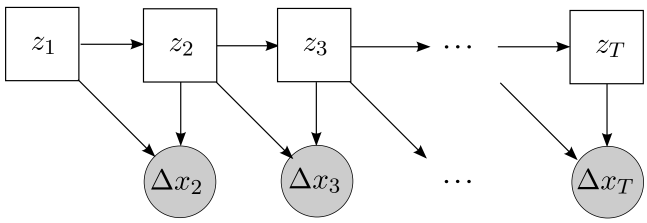


Figure 2.7: Difference hidden Markov model

Algorithm 1: Gibbs sampling.

initialise all variables;
while *number of iterations is less than maximum* **do**
    for *each variable* **do**
        sample new value conditional on all other variables;
    end
    increment number of iterations;
end

Having now introduced a range of temporal graphical models, in the following section we describe approximate methods which can be used to run inference over such models.

2.3.7 Approximate Inference in Temporal Graphical Models

Approximate methods are often used to perform inference over temporal graphical models as a result of the computational complexity of exact multivariate inference. This section discusses three common approximate inference methods, namely Gibbs sampling (Section 2.3.7.1), variational Bayes (Section 2.3.7.2), and variational message passing (Section 2.3.7.3), which are used by both the existing NIALM approaches discussed in Section 2.4 and the original work appearing in Chapters 4, 5 and 6 of this thesis.

2.3.7.1 Gibbs Sampling

Gibbs sampling is a Markov chain Monte Carlo algorithm which can be used to approximate samples drawn from a multivariate distribution (Casella and George, 1992). The algorithm works by repeatedly sampling a value for each latent variable of the model conditional on all other variables in the model, as shown by Algorithm 1.

Gibbs sampling is often used for inference due to the simplicity of its implementation. However, performing inference via Gibbs sampling can be prohibitively slow for two reasons. First, since the samples are drawn from a Markov chain, each sample is highly dependent upon the preceding sample. As a result, it is necessary to down-sample the data in order to assume independence between samples. Second, the Markov chain will often require a large number of iterations until convergence to the stationary distribution

Algorithm 2: Variational Bayes.

initialise variational distribution of all variables;
while *increase in lower bound is greater than threshold and number of iterations is less than maximum* **do**
    for *each variable* **do**
        update variable's variational distribution;
    end
    compute lower bound on joint likelihood;
    increment number of iterations;
end

is achieved. To control for this, the first m samples are normally discarded. For these reasons, Gibbs sampling is often used for inference when the simplicity of implementation is more important than computational efficiency.

2.3.7.2 Variational Bayes

Variational Bayes is an algorithm for approximate inference based on solving a set of interlocking update equations that cannot be solved analytically (Beal, 2003). The algorithm works by iteratively updating the variational distribution of all variables in the model, and therefore increasing the lower bound on the model likelihood. The algorithm continues to iterate until either convergence is achieved or the maximum number of iterations has been reached. This process is summarised in Algorithm 2.

Since a set of update equations are solved at each step, rather than sampling from a conditional distribution, the rate of convergence of variational Bayes is often much faster than for Gibbs sampling. However, the implementation of variational Bayes is more complex than Gibbs sampling, since the update equations for each variable must be derived by hand. As a result, variational Bayes is often used instead of Gibbs sampling when the speed of inference is more critical than the complexity of implementation.

2.3.7.3 Variational Message Passing

Variational message passing presents a generalisation of variational Bayes, which allows variational inference to be applied to arbitrary Bayesian networks without manually deriving each variable's update equation (Winn and Bishop, 2006). Instead, closed form update equations for each variable are obtained via a message passing scheme. Messages are passed between variables along the edges of the Bayesian network, and each variable is updated upon receiving messages from all variables within its Markov blanket.

Pseudocode for the variational message passing algorithm is given in Algorithm 3. First, the variational distribution of each variable is initialised. Next, each variable receives

Algorithm 3: Variational message passing.

initialise variational distribution of all variables;
while *increase in lower bound is greater than threshold and number of iterations is less than maximum* **do**
    for *each variable* **do**
        retrieve messages from all variable's parents and children;
        update variable's variational distribution;
    end
    compute lower bound on joint likelihood;
    increment number of iterations;
end

messages from each of its parents and children, before updating its variational distribution. Once all variables have updated their variational distribution, the lower bound on the model likelihood is calculated and the number of iterations is incremented. This process is repeated until either the increase in the lower bound is less than some threshold or the maximum number of iterations is reached.

Variational message passing requires that all distributions belong to the exponential family, and therefore all distributions can be expressed in a common exponential form. This common form allows a message passing scheme to be defined over a Bayesian network independent of the distributions over each variable, and as such removes the need to manually define update equations for each variable. As a result, variational message passing combines the ease of implementation of Gibbs sampling with the computational efficiency of variational Bayes, and therefore the algorithm can be applied in situations where both ease of implementation and computational efficiency are a priority.

Having introduced a number of temporal graphical models and three approximate inference methods in Section 2.3, in the next section we describe how they have been applied to the problem of energy disaggregation.

2.4 Temporal Graphical Models Applied to Non-intrusive Monitoring

In recent years, temporal graphical models have begun to be applied within the field of NIALM. This section discusses how the graphical models introduced in Sections 2.3.1 – 2.3.6 and the inference methods described in Section 2.3.7 have been used to disaggregate energy usage, along with their relative advantages and disadvantages.

Zaidi et al. (2010) demonstrated how standard HMMs can be used for appliance load recognition. However, it is important to distinguish between the fields of appliance load recognition and NIALM. Appliance load recognition corresponds to labelling an unknown appliance feed with the appliance's name, whereas NIALM aims to disaggregate and label appliance loads from a single aggregate feed. Zaidi et al. showed that HMMs are robust to various user behaviour for appliance load recognition, and suggested their application to the field of NIALM.

Kolter and Johnson (2011) showed how a FHMM, as described in Section 2.3.2, can be used as part of a NIALM system. The authors trained the model from sub-metered appliance feeds using EM and performed approximate inference through Gibbs sampling. However, the approach has two disadvantages. First, since sub-metered data is required for training, the resulting approach is both financially expensive and physically intrusive. Second, the FHMM is not tolerant to non-stationary noise, and as a result requires training data to be collected from all appliances in the household. Both of these disadvantages result in a disaggregation method which performs accurately in a controlled environment, but the required training phase is not practical enough for the approach to be widely applicable.

Kim et al. (2011) demonstrated how the CFHMM (Section 2.3.3) can be combined with the HSMM (Section 2.3.4) and the IOHMM (Section 2.3.5) to perform NIALM. The CFHMM allows the dependencies between appliances to be modelled (e.g. dependency between computer and monitor), while the HSMM allows appliance durations to be modelled explicitly (e.g. length of washing machine cycle), and finally the IOHMM allows additional observations which might influence appliance use to be built into the model (e.g. dependency of shower usage on time of day). However, such complexity results in a model in which inference is only possible through approximate methods. Interestingly, Kim et al. found that the performance gain of the factorial hidden semi-Markov model (FHSMM) was minimal in comparison to the CFHMM, and therefore explicitly modelling state durations resulted in little increase in accuracy. The authors show that the unsupervised learning approach can identify up to 10 appliances. However, as the number of appliances increases, the likelihood that each learned model will correspond to an individual appliance decreases, and therefore this unsupervised training method will not be applicable to households containing 20 or more appliances. In addition, the learned models also require manual labelling by a domain expert. Consequently, such approaches that do not incorporate any prior information are not suitable for realistic environments in which many appliances operate in parallel and it is infeasible to manually label appliances.

Kolter and Jaakkola (2012) demonstrate how the FHMM (Section 2.3.2) can be combined with the DHMM (Section 2.3.6) to perform NIALM. The FHMM structure ensures that the sum of inferred appliance power demands is equal to the measured aggregate power demand, while the DHMM structure enables step changes in the aggregate power to be attributed to specific appliance transitions. In addition, Kolter and Jaakkola extend the model to include a generic mixture component which enables the model to disaggregate a subset of all appliances. This ensures the model is robust to previously unseen observations (e.g. new or rarely used appliances). However, the described approach has only been evaluated using high-frequency data of the aggregate power demand. In addition, the described approach does not take into account prior knowledge about appliances, and therefore also requires the learned appliance models to be manually labelled by a domain expert. As a result, neither the assumed input data or the model training method are suitable for NIALM using low-granularity smart meter data.

Johnson and Willsky (2013) propose a Bayesian approach to the disaggregation problem using a FHSMM, in which prior appliance information can be incorporated into the graphical model, and therefore removing the need for a manual labelling process. Furthermore, Johnson and Willsky describe how change point detection can be incorporated into the temporal graphical model to greatly reduce the complexity of parameter learning and inference. However, the Bayesian approach requires prior models for each appliance, and therefore requires knowledge of the number and types of appliances present in each household. Unfortunately, this violates Requirement 4, and as a result does not constitute a realistic solution to the smart meter disaggregation problem.

In summary, temporal graphical models provide a promising potential solution to the problem of disaggregating smart meter data. However, all existing methods crucially require each appliance within a household to be identified manually either before or after an unsupervised training phase. Therefore, there is an area of unexplored research into methods which would allow general prior knowledge about appliance types to be integrated into such temporal graphical models, without requiring the number and types of each appliance in a household to be manually specified. We now summarise the related work discussed in this chapter.

2.5 Summary

This chapter has described various existing approaches to the energy disaggregation problem. We first introduced intrusive monitoring techniques, however they were dismissed due to the requirement to install multiple sensors throughout each household (Requirement 2). We then described a category of approaches based on the processing of high frequency data. However, such methods are not compatible with current smart meters, and would therefore require the installation of additional expensive metering hardware (Requirement 3). We also discussed how existing event based methods could be applied to smart meter data. However, such methods assume all appliance switch events to be independent, and as a result are unlikely to provide realistic solutions when applied to 10 second power data.

We then introduced a promising category of non-intrusive approaches which apply non-event based methods to low frequency data through the use of temporal graphical models.

The theory of such models is well studied in the field of machine learning, and these models have previously been applied to source separation problems such as speech recognition with multiple speakers. However, in the field of energy disaggregation, existing work has failed to provide realistic training methods for such models, and instead such training methods have relied on a manual labelling process of appliances by a domain expert (Requirement 4). Therefore, we have identified an important field of unexplored research, regarding how to incorporate prior appliance knowledge into the training process of such models without requiring manual intervention by a domain expert.

In order to objectively compare the accuracy and flexibility of various NIALM methods, it is essential to use power data collected from real households. Traditionally, existing work has often used small amounts of private data to evaluate their algorithms. However, in recent years a number of public data sets have been released which have been designed specifically for the evaluation of NIALM techniques. The following chapter summarises each data set, and gives reasons for the usage of certain data sets in this thesis.

Chapter 3

Household Energy Data Sets

Traditionally, NIALM techniques have been evaluated using either simulated data or private data sets. Although such data sets allow a single approach to be tested under a range of conditions, the lack of available code has made objective comparisons between approaches impossible. However, since 2011, a number of data sets have been publicly released, therefore allowing authors to independently evaluate their approaches under similar conditions. This chapter first describes the 8 public data sets that are currently available, and motivates the use of two of these data sets in this thesis. Furthermore, we highlight the lack of scale as a limitation of such public data sets, and give reasons for the use of large scale private data sets to evaluate the scalability of NIALM methods. Finally, we present a table of comparison of the various data sets available for training and evaluating energy disaggregation approaches.

3.1 Public Data Sets

This section discusses 8 public data sets which have been released in recent years. For each data set, we give the location from which it was collected, along with the number of meters used and the granularity of data which was recorded. See Appendix C for example data fragments of each data set.

3.1.1 Reference Energy Disaggregation Data Set

The Reference Energy Disaggregation Data set (REDD) was collected by a group at MIT from 6 households in the Greater Massachusetts area, MA, USA (Kolter and Johnson, 2011). The data set contains both household-level and circuit-level data over various durations (from a few weeks to several months). Each house had two current clamps monitoring the two-phase mains input, and 10-25 current clamps monitoring individual circuits. High-frequency (15 kHz) current and voltage data are available for both mains circuits, while low-frequency power measurements (3-4 second intervals) are available for the individual circuits. Since most households contain a large number of circuits, many circuits contain only one appliance, and therefore this data can be used as the ground truth for evaluating NIALM methods.

We use the REDD in Chapter 5 to demonstrate how general models of appliance types can be tuned to the specific appliance instances in a household using only aggregate data. This data set was chosen because it contains both household aggregate and individual appliance power measurements at 3 second intervals. Furthermore, at the time of writing, it was the most widely used data set for benchmarking NIALM methods.

3.1.2 Building-Level Fully Labeled Electricity Disaggregation Data Set

The Building-Level fUlly-labeled data set for Electricity Disaggregation (BLUED) was collected by Anderson et al. (2012) of CMU, from a single household in the Pittsburgh area, PA, USA. The data set contains only household-level aggregate power data, although labelled events are also reported for each individual appliance (e.g. microwave turns on at 10am). Household-level aggregate power data was collected from two current clamps monitoring both phases of its split phase power, over a period of 8 days. High frequency (12 kHz) current and voltage data are available for both phases.

We have chosen not to use BLUED in this report, since no sub-metered appliance power data was collected as part of the data set.

3.1.3 UMASS Smart* Home Data Set

The Smart* data set was collected by a group at the University of Massachusetts Amherst from three households (labelled A, B and C) in the Western Massachusetts area, MA, USA, over a period of 3 months (Barker et al., 2012). For household A, the data set contains household-level aggregate power data, circuit-level power data, and appliance-level power data. The aggregate-level and circuit-level data was collected at 1 second intervals, while appliance level data was collected for all appliances on a circuit each time the power demand of the circuit changed. However, only aggregate electrical data is available for households B and C.

We have chosen not to use the Smart* data set in this report since it only contains a single household in which individual circuits or appliances are monitored.

3.1.4 Tracebase Repository

Unlike the other data sets described in this section, the Tracebase repository was not collected for the purposes of disaggregation (Reinhardt et al., 2012). Instead, the repository was designed to aid the study of individual appliances. The Tracebase repository was set up by a group at Darmstadt University, and contains individual appliance data from an unspecified number of households in Germany. The repository contains a total of 1883 days of power readings, recorded at 1 second intervals, across 158 appliance instances (e.g. a Bosch Logixx KSV36AW41G refrigerator), of 43 different appliance types (e.g. refrigerator). Since the core aim was to create an appliance database, no household aggregate measurements were also collected.

Since the Tracebase repository contains many examples of different appliance instances of the same type, it provides an ideal data set from which to investigate the diversity of appliances within an appliance type. For this reason, we use data from the Tracebase repository in Chapter 4 to build general models of appliance types.

3.1.5 Individual Household Electric Power Consumption Data Set

EDF Energy released a data set in 2012 containing energy measurements made at a single household in France for a duration of 4 years. Average measurements were made at 1 minute resolution of the household aggregate active power, reactive power, voltage and current, in addition to the active power of 3 individual circuits. However, due to the relatively small number of household circuits, each circuit contains multiple appliances.

Although this data set spans an extensive period of time, it is unsuitable for evaluating the accuracy of NIALM algorithms as individual appliance data is not available.

3.1.6 Household Electricity Use Study

In 2012, the UK Energy Savings Trust, Department of Energy and Climate Change, and Department for Environment, Food and Rural Affairs published a report into electricity usage within UK households (Energy Saving Trust, 2012). This report summarises the Household Electricity Use Study (Zimmermann et al., 2012), which aims to better understand how electricity is consumed in UK households. As part of this study, 251 owner-occupier households were monitored across England between April 2010 and April 2011. Of these households, 26 were monitored for 12 months, and 225 were monitored for 1 month. For each household, the energy consumption of 13-51 appliances was monitored at 2 minute intervals.

However, since the focus of this study was to to investigate the energy consumption of individual appliances, the data set does not contain any household aggregate data.

As a result, any artificial aggregate calculated by summing the power demand of each appliance will exclude appliances which were not sub-metered, and also remove any additional measurement noise. Furthermore, the 2 minute sampling rate does not match the 10 second data that will be reported by UK smart meters. For these reasons, this data set was not used in this thesis.

3.1.7 Pecan Street Research Institute Sample Data Set

In 2013, the Pecan Street Research Institute released a sample data set designed specifically to enable the accuracy evaluation of electricity disaggregation technology. The sample data set contains 7 days of data from 10 houses in Austin, TX, USA, for which both aggregate-level and circuit-level data is available. At each measurement level, both apparent power and real power measurements are available at 1 minute intervals. In addition to common household loads, two houses also have photovoltaic systems and one house has an electric vehicle.

Although this data set contains sub-metered data on a comparable scale to the REDD data set, its recent release has not yet led to its use to benchmark any disaggregation algorithms. For this reason, we have chosen to use the REDD data set rather than this sample data set in this thesis.

3.1.8 The Almanac of Minutely Power Data Set

Most recently, Makonin et al. (2013) of the Simon Fraser University released the Almanac of Minutely Power Data Set (AMPds), which contains one year of data collected from a single household in the greater Vancouver area, BC, Canada. In addition to household-level readings, data was also collected from 19 individual circuits at 1 minute resolution. Each reading includes measurements of voltage, current, frequency, power factor, real power, reactive power and apparent power. Furthermore, the aggregate gas and water consumption was also measured at 1 minute intervals, in addition to 1 individual load for each utility.

We have chosen not to use the AMPds in this report since it only contains data from a single household. Having discussed 8 public data sets, we now move on to private data sets.

3.2 Private Data Sets

The previous section has discussed a range of public data sets which include data regarding individual appliance usage. However, the collection of both household aggregate and individual appliance data is inherently expensive and intrusive, since appliance meters are required to be installed on individual appliances. As a result, such data sets have thoroughly monitored a maximum of 10 households with both aggregate and appliance meters. The size of such data sets is insufficient to represent the variety of UK households, and furthermore such deployments are biased towards households for which appliance sub-metering is convenient. For these reasons, we turn to a private data set which better represents the diversity of households in the UK, as described in the following section.

3.2.1 Colden Common Data Set

The Colden Common data set was collected by a group at the University of Southampton from 117 households in the village of Colden Common, Hampshire, UK. The monitoring systems were installed in 2011, and data collection is currently ongoing. In each household, the aggregate power demand was measured at 1 second intervals.

Since this data set includes household aggregate data from over 100 households, it provides a far more representative sample of UK households than any of the public data sets discussed in Section 3.1. For these reasons, we use the Colden Common data set to evaluate the robustness and flexibility of our approach to NIALM in Chapter 6.

3.3 Table of Comparison

Table 3.1 gives a comparison of the 8 public data sets introduced in Section 3.1 and the Colden Common data set introduced in Section 3.2. Batra et al. (2014) provide further details regarding the completeness of such data sets. In the following section, we summarise the findings of this chapter.

Data set	Institution	Location	Duration	Number of houses	Number of sub-meters per house	Frequency of ground truth	Aggregate data available	Terms of use
REDD	MIT	MA, USA	3-19 days	6	9-24	3 second	Yes	Registration
BLUED	CMU	PA, USA	8 days	1	0	N/A	No	Registration
Smart*	UMass	MA, USA	3 months	1	54	1 second	Yes	Registration
Tracebase	Darmstadt	Germany	N/A	N/A	N/A	1-10 second	Yes	Public
Sample data set	Pecan Street	TX, USA	7 days	10	12	1 minute	Yes	Consortium
IHEPCDS	EDF R&D	France	4 years	1	3	1 minute	Yes	Public
HES UK	DECC	UK	3 months	250	13-51	2 minute	No	Academic
AMPds	Simon Fraser U.	BC, Canada	1 year	1	19	1 minute	Yes	Registration
Colden Common	U. of Southampton	UK	1 year	117	0	N/A	Yes	Private

Table 3.1: Table of comparison of household energy data sets.

3.4 Combining Data Sets

As shown by Appendix C, each public data set is structured according to a different data format. Consequently, a time-consuming engineering barrier exists when using the data sets. This has resulted in publications using only a single data set to evaluate a given approach, and consequently the generality of results over large numbers of households are rarely investigated. To address this challenge, Batra et al. (2014) introduced the Non-intrusive Load Monitoring Toolkit (NILMTK), which provides parsers for six publicly available data sets. This will allow energy disaggregation algorithms to be able to evaluated over a larger number of households from different locations with minimal engineering effort.

3.5 Summary

This chapter has presented 8 public data sets which have been released as a result of recent interest in the disaggregation of smart meter data. We have discussed their potential for use in the training and evaluation of NIALM algorithms, and motivated the use of the Tracebase repository and the REDD data set in this thesis. However, we have also identified such data sets as being a poor representation of the diversity of UK households. This has motivated our use of a larger scale private data set, the Colden Common data set, to evaluate the robustness and flexibility of energy disaggregation approaches.

In the following three chapters, we first propose a novel appliance generalisation method using the Tracebase repository (Chapter 4), followed by a new approach for tuning general appliance models using only aggregate data which is evaluated using the REDD data set (Chapter 5), and finally a large scale case study application of these approaches to cold appliances using the Colden Common data set (Chapter 6).

Chapter 4

Building Generalisable Appliance Models

This chapter proposes a novel method for learning general appliance models which will generalise to previously unseen appliance instances of the same type. We start by proposing a hierarchical model, which formalises the relationship between appliance types and the appliance instances operating in a specific household (Section 4.1). Next, we show how Bayesian inference can be used to infer the parameters of individual appliance instances given sub-metered appliance data from existing data sets (Section 4.2). We then go on to describe how the parameters of multiple appliance instances can be used to create general models representing the appliance type (Section 4.3). Next, we give an empirical comparison in which we investigate how many appliance instances are required to construct the general models of appliance types (Section 4.4). Furthermore, we use cross validation to compare the generality of models of appliance types to models learned from single appliance instances. Finally, Section 4.5 summarises the findings of this chapter.

4.1 Hierarchical Modelling of Appliance Types

The aim of this chapter is to learn distributions over the model parameters for each appliance type, such that both the mean and variance around each appliance parameter is derived from data. This process is effective as it allows tight distributions to be learned over appliance parameters which are similar for different appliance instances, and broad distributions to be learned when parameters vary greatly between different instances. In general, the most important factor is to ensure that the learned states align between different appliance instances, which we demonstrate through the Bayesian framework described in Section 4.2. Throughout this section, we use a running example

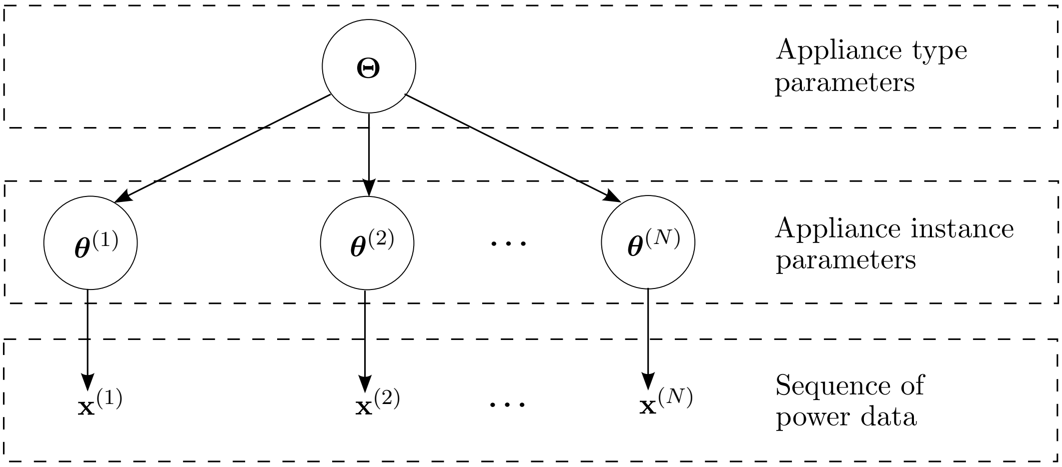


Figure 4.1: Hierarchical model of an appliance type

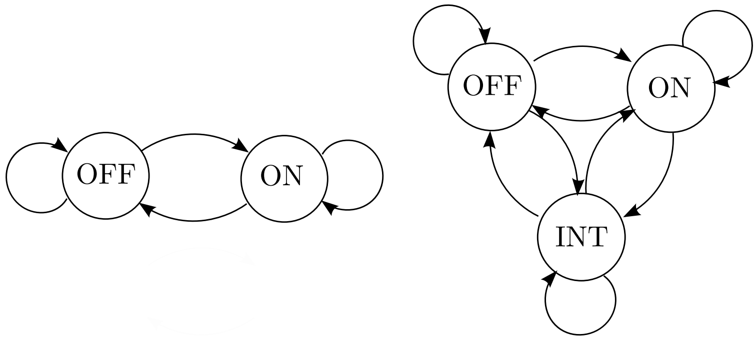
of the refrigerator to provide some intuition into the model choices and role of various parameters.

We adopt a hierarchical approach to model multiple appliances of the same type, as shown by Figure 4.1. In this model, we represent an appliance type as a distribution from which appliance instances are drawn. As such, the appliance type represents any behaviour which is common to all instances of that type, while an appliance instance also represents behaviour which is specific to that single instance and its usage. Furthermore, instead of observing the appliance instance parameters directly, we observe sequences of power data drawn from each appliance instance. More formally, the aim is to infer the parameters of an appliance type, $\boldsymbol{\Theta}$, from sequences of power data, $\mathbf{x}^{(n)} = \{x_1^{(n)}, \dots, x_T^{(n)}\}$, generated by individual appliance instances described by parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(N)}\}$, where n is one of N appliance instance indices.

In order to learn the appliance type parameters, in the following section we estimate the parameters of each appliance instance from a sequence of power readings. We then describe a method for generalising over these parameters in Section 4.3.

4.2 Appliance Instance Parameter Estimation using Hidden Markov Models

We adopt a HMM representation of appliances, as described in Section 2.3.1. HMMs present a natural choice for modelling appliances for the following three reasons. First, such models are well studied in the domain of probabilistic graphical modelling, and as a result many algorithms exist to solve the problems of parameter learning and hidden state inference. Second, HMMs provide a suitable trade-off between the representation



(a) 2 state model (b) 3 state model

Figure 4.2: Appliance state models

of the physical structure of appliances (e.g. finite set of operational states) and also representing the various patterns of usage of the appliance by the household's occupants. Third, a number of extensions of the basic HMM exist which allow more complex types of behaviour to be modelled (e.g. the periodic behaviour of refrigerators).

In the HMM, the value of each discrete variable, z_t, of the Markov chain corresponds to one of K states (e.g. on, off), while each continuous variable, y_t, can be either zero or any positive real number (e.g. 100.5 W), since appliances only consume energy. Figure 4.2 shows a 2 state model (representing an appliance that is either on or off) and a 3 state model (representing an appliance with an on, off and intermediate state), in which nodes represent states and edges represent transitions between states. For the sake of clarity, we omit the appliance instance index $^{(n)}$ throughout Section 4.2.

We use a Gaussian function to model the distribution over each state's power demand, since the distribution has previously been shown to provide a good fit of appliance power demand (Kim et al., 2011). Although an appliance's power demand is strictly positive, we found that the Gaussian distribution's support for negative power demands is negligible for most appliances. However, it is worth noting that other distributions could also be used if a strictly positive (e.g. gamma distribution) or a multi-modal (e.g. a mixture of Gaussians) distribution were required. In the case of the refrigerator, the *off* state emission distribution will likely be a very high precision distribution centered around 0 W, while the *on* state distribution will be a slightly lower precision distribution centered around approximately 100 W. Both distributions are expected to be of relatively high precision since this precision parameter represents only small fluctuations in the appliance's power demand around its expected values.

We adopt a Bayesian approach to learn the parameters of the HMM in which prior distributions are placed over the model parameters, as shown by Figure 4.3. A Bayesian approach is required in this scenario, since it ensures that the states learned for one instance of an appliance type correspond to the same states learned from a different instance of the same appliance type. For example, it ensures that the *spin* state of

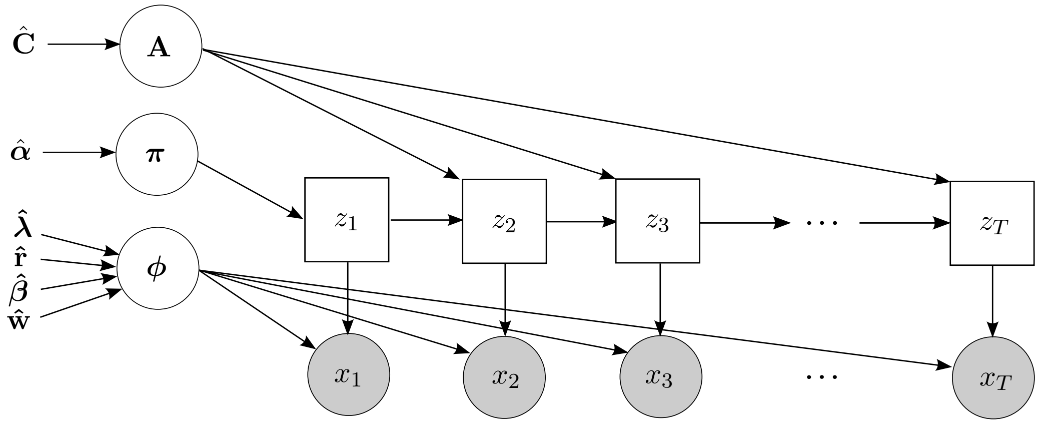


Figure 4.3: Bayesian hidden Markov model

washing machine A corresponds to the *spin* state of washing machine B. By placing conjugate priors over the model parameters, we ensure that both the priors and posteriors belong to the same family of distributions. We now describe the prior and posterior distributions over the model parameters, which for the sake of clarity, we use a hat to denote the hyperparameters of the prior distributions (e.g. $\hat{\boldsymbol{\alpha}}$), and a tilde to denote the parameters of the posterior distributions (e.g. $\tilde{\boldsymbol{\alpha}}$).

The initial probabilities follow a categorical distribution, for which the conjugate prior is the Dirichlet distribution:

$$\boldsymbol{\pi} \sim \mathrm{Dir}(K, \hat{\boldsymbol{\alpha}}) \tag{4.1}$$

where Dir is the Dirichlet distribution parameterised by the number of categories, K, and the concentrations parameters, $\hat{\boldsymbol{\alpha}}$. We denote the parameters of the posterior distribution as $\tilde{\boldsymbol{\alpha}}$. In the case of the refrigerator, we have little a priori information regarding the initial distribution, and so a uniform prior distribution is used.

Similarly, each row, i, of the transition matrix also follows a categorical distribution:

$$\mathbf{A}_i \sim \mathrm{Dir}(K, \hat{\mathbf{C}}_i) \tag{4.2}$$

where Dir is the Dirichlet distribution parameterised by the number of categories, K, and a vector of concentrations parameters, $\hat{\mathbf{C}}_i$. We denote the posterior parameters as $\tilde{\mathbf{C}}_i$. In the case of the refrigerator, this parameter is easily learned from sub-metered training data, and so a uniform prior is also a sufficient distribution.

Finally, the emission variables are Gaussian distributed, for which a conjugate prior is the Gaussian-gamma distribution (Murphy, 2007):

$$\mu_k \sim \mathcal{N}(\hat{\lambda}_k, \hat{r}_k) \tag{4.3}$$

$$\tau_k \sim \mathrm{Gamma}(\hat{\beta}_k, \hat{w}_k) \tag{4.4}$$

where $\mathcal{N}$ is the Gaussian distribution parameterised by mean, $\hat{\lambda}_k$, and precision, $\hat{r}_k$, and Gamma is the gamma distribution parameterised by shape, $\hat{\beta}_k$, and scale, $\hat{w}_k$. We denote the respective parameters of each posterior distribution as $\tilde{\lambda}_k$ and $\tilde{r}_k$, and $\tilde{\beta}_k$ and $\tilde{w}_k$.

It is crucial to incorporate domain knowledge via these hyperparameters to ensure the posterior states correspond between different appliance instances. In the case of the refrigerator, $\hat{\lambda}_{\mathit{off}}$ would be 0 W and $\hat{\lambda}_{on}$ would be 100 W since these represent the expected value of each state's mean power. In addition, $\hat{r}_{\mathit{off}}$ and $\hat{r}_{on}$ represent the precision in the mean values between different appliance instances, and therefore $\hat{r}_{\mathit{off}}$ would be large since all refrigerators consume close to 0 W when they are *off*, while $\hat{r}_{on}$ would be relatively low since the mean *on* power of different refrigerator instances varies between approximately 50 W and 150 W. Since the precision parameter, $\boldsymbol{\tau}$, varies greatly for different states and appliance instances, the hyperparameters $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{w}}$ are used to provide fairly broad prior distributions.

We use this Bayesian approach to parameter estimation in HMMs to individually learn the parameters, $\boldsymbol{\theta}^{(n)}$, of each appliance instance, n, from sequences of their power data, $\mathbf{x}^{(n)}$. Since there is no analytical solution to parameter estimation in HMMs, we performed inference using variational message passing (Winn and Bishop, 2006), as introduced in Section 2.3.7.2. Variational message passing was used since it provides an efficient and deterministic method of Bayesian parameter estimation for which convergence is guaranteed (Minka et al., 2012). We implemented the model as described in this section and performed inference using the Infer.NET framework (Minka et al., 2012). The Infer.NET framework was used because it is the most flexible toolkit for Bayesian inference in probabilistic graphical models. As such, it allows arbitrary Bayesian networks to be defined (in this case a HMM) and general purpose inference algorithms to be used to infer posterior distributions over the unknown variables. The Infer.NET framework is particularly preferable in this scenario since it supports fully Bayesian inference, unlike similar tookits which only support maximum likelihood parameter estimation, e.g. Bayes Net Toolbox for Matlab (Murphy, 2001).

In the following section, we describe how these parameters are combined to form a model of the appliance type which will generalise to previously unseen instances of this appliance type.

4.3 Generalising Over Multiple Appliance Instances

We now describe a method by which the parameters learned in Section 4.2 can be combined to form a model that represents the whole appliance type, and therefore generalises to previously unseen instances of that appliance type. Our method consists of fitting distributions to samples drawn from the posterior distributions over appliance instance parameters. As a result, this method averages over our uncertainty around the appliance instance parameters. We introduce the notation:

$$\mathbf{\Theta} = \{\mathbf{\Theta}^{\alpha}, \mathbf{\Theta}^{C}, \mathbf{\Theta}^{\lambda}, \mathbf{\Theta}^{r}, \mathbf{\Theta}^{\beta}, \mathbf{\Theta}^{w}\} \tag{4.5}$$

to represent the parameters of the general model of an appliance type as defined in the following paragraphs. In the case of the refrigerator, $\mathbf{\Theta}$ represents a distribution over all possible refrigerator instances. Crucially, this general model allows the probability to be calculated that an unknown appliance instance belongs to the refrigerator appliance type.

Samples drawn from the posterior distributions over the initial probabilities and transition matrix are in the form of multinomial distributions, for which the Dirichlet distribution is the conjugate prior. Therefore, we generalise by fitting Dirichlet distributions to the samples using:

$$\mathbf{\Theta}^{\alpha} = \arg\max_{\boldsymbol{\alpha}} \mathrm{Dir}(\pi_{1:M}^{(1:N)} | K, \boldsymbol{\alpha}) \tag{4.6}$$

$$\mathbf{\Theta}_z^{C} = \arg\max_{\mathbf{C}_z} \mathrm{Dir}(A_{1:M}^{(1:N)} | K, \mathbf{C}_z) \tag{4.7}$$

where $\pi_{1:M}^{(1:N)} \sim \mathrm{Dir}(K, \tilde{\boldsymbol{\alpha}})$ and $A_{1:M}^{(1:N)} \sim \mathrm{Dir}(K, \tilde{\mathbf{C}})$ represent sets of $1, \ldots, M$ samples, each of which is drawn from the initial and transition posterior distributions for appliance instances $1, \ldots, N$. We use the Fastfit MATLAB toolbox to estimate the parameters of the Dirichlet distributions, which provides a simple and efficient method for parameter estimation through the generalised Newton method (Minka, 2002a,b).

We also fit a Gaussian distribution to samples drawn from the posterior distribution over the emission mean parameters:

$$\mathbf{\Theta}_z^{\lambda}, \mathbf{\Theta}_z^{r} = \arg\max_{\lambda, r} \mathcal{N}(\mu_{1:M}^{(1:N)} | \lambda, r) \tag{4.8}$$

where $\mu_{1:M}^{(1:N)} \sim \mathcal{N}(\tilde{\lambda}, \tilde{r})$ represents sets of $1, \ldots, M$ samples drawn from the posterior distribution over $1, \ldots, N$ appliance instances' mean power.

Similarly, we fit gamma distributions to samples drawn from the posterior distributions over each state's precision. This results in a distribution which generalises over each posterior distribution of a given state's precision. However, this approach is prone to severe over-fitting when a gamma distribution is fitted to the precisions of the *off* state. In this case, the posterior distributions of the *off* state's precision are often highly peaked and centred around similar values, since they generally only represent the measurement noise around 0 W. However, it is possible for the power demand to be sampled during a transition between the *off* and *on* states for any appliance type. This results in a sample which would receive near zero probability given the tight estimates of each state's precision. In fact, the probability is likely to be beyond the numerical precision

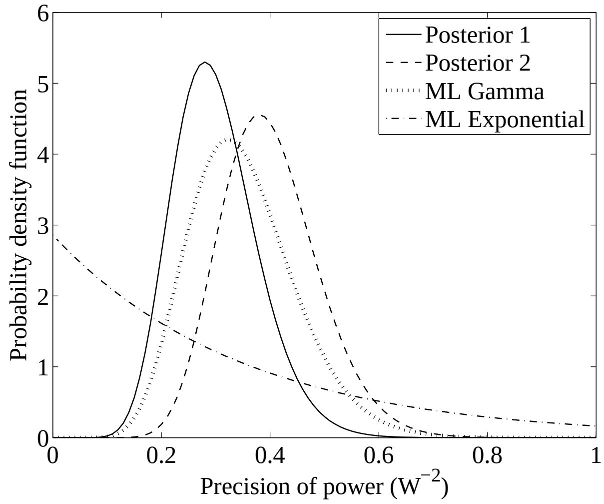


Figure 4.4: Generalising over appliance state precision

of a double precision floating point number and therefore cause the inference to fail. To prevent this, we constrain the gamma distribution for the *off* state ($k = 1$) which is fitted to samples drawn from the posterior distributions such that it follows an exponential distribution, by holding the shape parameter fixed at 1:

$$\mathbf{\Theta}_z^{\beta}, \mathbf{\Theta}_z^{w} = \begin{cases} \arg\max_{w} \text{Gamma}(\tau_{1:M}^{(1:N)} | 1, w\}) & k = 1 \\ \arg\max_{\beta,w} \text{Gamma}(\tau_{1:M}^{(1:N)} | \beta, w\}) & k > 1 \end{cases} \tag{4.9}$$

where $\tau_{1:M}^{(1:N)} \sim \text{Gamma}(\tilde{\lambda}, \tilde{r})$ represents sets of $1, \ldots, M$ samples drawn from the posterior distribution over $1, \ldots, N$ appliance instances' precisions. Figure 4.4 shows two examples of posterior distributions of the precision of an *off* state as learned from two appliance instances. It can be seen that fitting a gamma distribution to the samples drawn from posterior distributions would result in a tight distribution which would assign an extremely low probability to any measurement of the power demand sampled during a transition between states. Figure 4.4 also shows an exponential distribution fitted to the samples drawn from the appliance posteriors. It is clear from the long tail shape of this distribution that it will have non-zero support for data points sampled during a transition between two states.

We use the approach described in this section to build models of an appliance type that will generalise to previously unseen instances of that appliance type. We now go on to describe an empirical evaluation of this approach using the Tracebase data set.

Appliance type	Number of instances	Average signatures per instance
Refrigerator	11	19
Kettle	9	14
Microwave	8	8
Washing machine	9	6
Dishwasher	8	19

Table 4.1: Breakdown of signatures in Tracebase repository, including 3 additional dishwasher instances.

4.4 Empirical Evaluation of Model Generalisation using Tracebase Data Set

We evaluated the benefit of building generalisable models of appliance types using the Tracebase data set (Reinhardt et al., 2012). This data set is particularly useful for such an evaluation since it contains data from many instances of appliances of the same type. The data set consists of samples of appliances' power demands at roughly one second intervals. We extracted between 2 and 60 signatures (durations when the appliance was in use) depending on data availability for each appliance instance in the Tracebase data set. We selected the following 5 common appliance types: refrigerator, kettle, microwave, washing machine and dishwasher. We also extended the Tracebase data set with data collected from 3 additional dishwasher instances, such that at least 8 instances were available for each appliance type. Table 4.1 shows a breakdown of the signatures extracted from the Tracebase data set.

We modelled the refrigerator, kettle and microwave using the 2 state model shown in Figure 4.2 (a) and we modelled the washing machine and dishwasher using a 3 state model as shown in Figure 4.2 (b). We selected the number of states based upon the minimum number of electrical components for each appliance type. For example, kettles consist of a single heating element which can either be on or off, and therefore a model with 2 states (*on*, *off*) was appropriate. In contrast, a washing machine typically has a water heater and drum motor, and therefore a model with 3 states (*heater*, *spin*, *off*) was appropriate. It is important to note that, although different washing machine cycles are available, the cycles consist of the operation of the same components in different orders. As a result, a 3 state model can represent a range of cycles with different temperatures and durations. Furthermore, it was important to restrict the number of states to ensure the learned states correspond between different appliance instances. The hyperparameters for each appliance type used are given in Appendix B.

We use hold-one-out cross validation to determine how well a given appliance model generalises to a previously unseen appliance instance. Hold-one-out cross validation was selected since it penalises appliance models which have been over-fitted to the training appliance instances and therefore to not generalise to new appliance instances, while

Approach	Description
GT	General appliance model as learned from the Tracebase data set without any parameter tuning.
NT	Specific appliance model as learned from a single appliance instance other than the test appliance.
ST	Specific appliance model as learned from the test appliance instance.

Table 4.2: Summary of approaches compared using the Tracebase data set.

also favouring appliance models which are specific enough to accurately represent the behaviour of the specific appliance type. This process requires the construction of a generalisable appliance model using between 2 and 7 training appliance instances, which we show to be sufficient to build a general model of each appliance type. We then test these general models against a single appliance instance that was excluded from the training set. Therefore, a single fold of the set of appliance instances corresponds to an ordered list of 7 training appliances and one test appliance. We refold the set of appliance instances 50 times, and for each fold we evaluate how well the appliance model constructed using between 2 and 7 training appliance instances generalises to the test instance.

We compare our approach which builds general models of appliance types (GT) to two bounding benchmarks. The first benchmark uses training data from a single appliance instance from the training set (NT). This represents a lower bound, in which no effort is made to generalise over multiple appliance instances. The second benchmark uses training data from the test appliance (ST). This represents an upper bound, in which the test appliance is not regarded as previously unseen. Since ST and NT are dependent only on the fold of the set of appliance instances, and not on the size of the training set, both ST and NT can only be evaluated once for each fold of data. We present the mean log-odds for GT, NT and ST over the 50 folds. These approaches are summarised in Table 4.2.

We use the model likelihood as a metric for evaluating how well an appliance model explains the test data, averaged over each fold of the data set. The model likelihood metric was selected since it represents the extent to which an appliance explains a data sequence through a single value, rather than other metrics which examine each of the model parameters individually. This metric represents the likelihood of the test data, $\mathbf{x}$, given a general appliance model, $\mathbf{\Theta}$, with both the appliance states $\mathbf{z}$ and the appliance instance parameters $\boldsymbol{\theta}$ integrated out, as given by:

$$p(\mathbf{x}|\mathbf{\Theta}) = \iint p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{\Theta}) \, \mathrm{d}\mathbf{z} \, \mathrm{d}\boldsymbol{\theta} \tag{4.10}$$

where $p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ is calculated using Equation 2.4. Since this likelihood decreases towards zero as the length of the input data sequence increases, we compare the log-odds rather

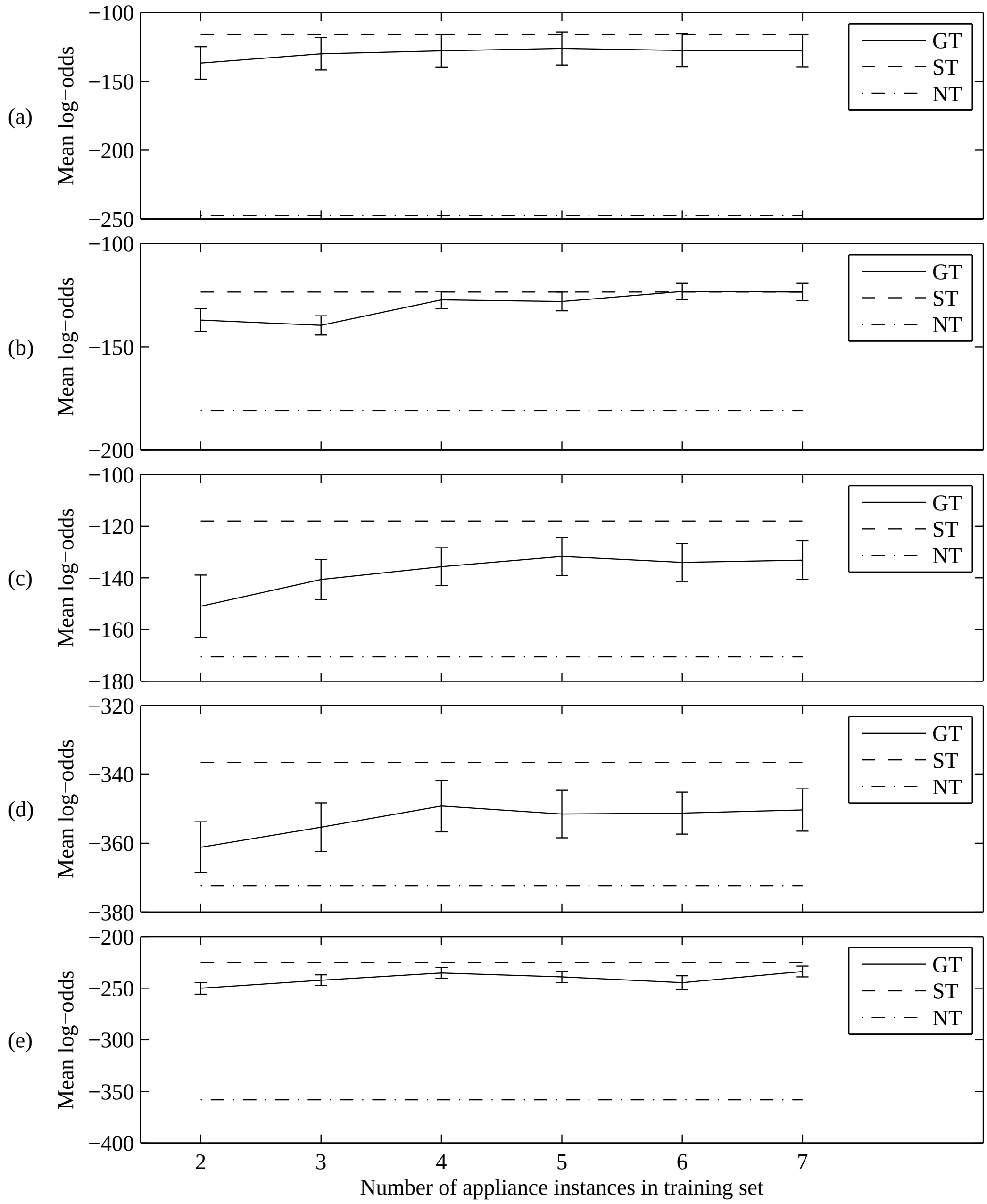


Figure 4.5: Mean cross validation model log-odds for increasing training set sizes. Legend: GT - generalised training, ST - sub-metered training, NT - non-generalised training. Subplots: (a) Kettle, (b) Refrigerator, (c) Microwave, (d) Washing machine, (e) Dishwasher. Error bars represent standard error in the mean.

than the probability. Log-odds, or the logit function, has the advantage that it maps a probability, p, in the range $[0, 1]$ to the domain of real numbers, and therefore avoids problems of numerical precision. This function is defined by:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \tag{4.11}$$

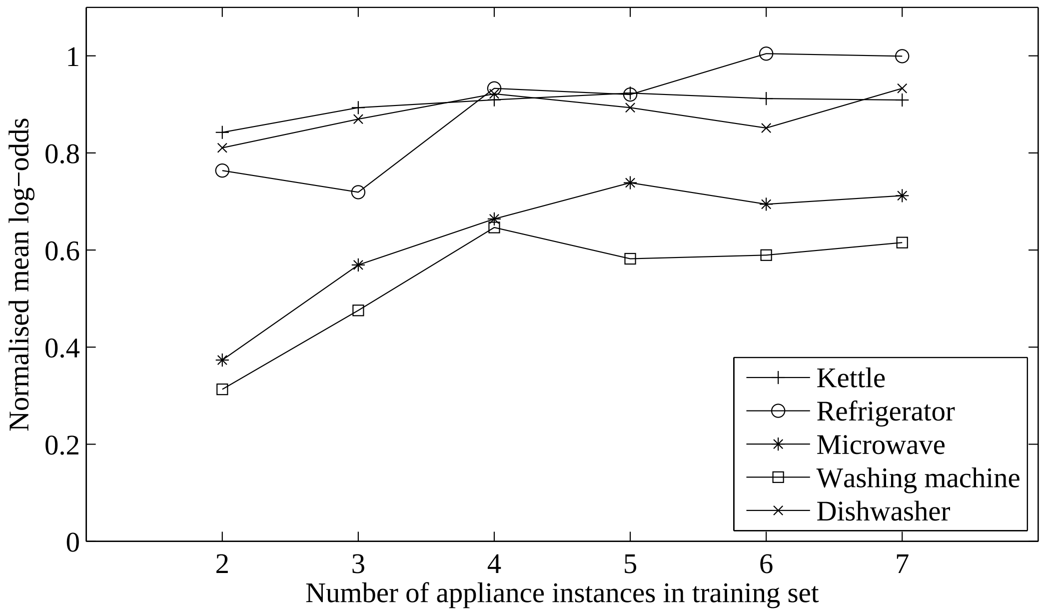


Figure 4.6: Normalised cross validation model log-odds for increasing training set sizes.

Figure 4.5 shows the cross validation model log-odds for 5 common household appliances for training set sizes of between 2 and 7 appliance instances. These are compared with the two benchmarks described above, representing approaches where sub-metered data is available from the test appliance, and where data is only available from a single appliance from the training set. The error bars represent the standard error in the mean. A clear trend common to all appliance types is that the model log-odds increases towards an asymptote as the number of appliance instances in the training set increases. This indicates that the majority of the appliance type's behaviour can be described by a general model learned from a relatively small number of appliance instances. As such, we argue that it is not necessary to build an exhaustive database of all appliance instances as other work has discussed (Lam, 2007; Lai et al., 2012), and instead we propose the use of a database of distributions over possible appliance behaviour.

In addition, all averages lie above the lower bounding benchmark, reflecting the intuition that an approach is always preferable if it generalises over multiple appliances rather than uses data from a single instance. Furthermore, all averages lie below the upper bounding benchmark, reflecting that no general model provides a better explanation of sub-metered data than a model learned from that sub-metered data.

Figure 4.6 shows the normalised cross validation average log-odds for the same 5 appliances. The appliance averages were normalised to lie in the range $[0, 1]$, such that 0 represents the accuracy of the model trained with a single non-test appliance and 1 represents the accuracy of the model trained with the test appliance. This figure enables interesting comparisons between appliance types. First, it can be seen that some appliance types converge towards their asymptote more rapidly than others. This trend is most obvious when comparing the kettle to the washing machine, since the kettle shows relatively little improvement beyond the set of 2 instances due to its single heating component, while the the washing machine shows a much greater improvement between the sets of 2 and 4 appliance instances as a result of its additional electrical components. This indicates that fewer training examples are required for appliances with fewer electrical components before the optimal general model is achieved. Furthermore, it can be seen that some appliances converge to an asymptote that is closer to the benchmark which uses sub-metered training data (normalised log-odds = 1). This trend is most obvious when comparing the refrigerator to the washing machine, since the refrigerator converges to an asymptote very close to the benchmark which uses sub-metered training data, while the washing machine converges to an asymptote noticeably lower than the corresponding benchmark. This is caused by different degrees of variance within an appliance type, for example, there is less variance within the kettle appliance type than the washing machine appliance type. We now summarise the findings of this chapter in the following section.

4.5 Summary

In this chapter, we first proposed a hierarchical structure which links appliance types, appliance instances and the sequences of power data generated by such appliances. We then went on to describe a method by which appliance instance parameters could be learned from sequences of power data using Bayesian inference, which crucially ensured that the learned states correspond between different appliance instances of the same type. Next, we introduced a method which is able to learn a single model of an appliance type which will generalise to previously unseen appliance instances of that type, using only the learned appliance instance models. Last, we provided an empirical evaluation of our approach through cross validation using the Tracebase data set, and showed that general appliance models can be learned from 2–6 appliance instances. We also showed that such general appliance models consistently outperform models learned from a single appliance instance, and furthermore we showed that in some cases the general appliance models can perform comparably to models learned from sub-metered data from the test appliance instance.

Now, having introduced a Bayesian method for inferring the behaviour of appliance instances given a HMM representation, and proposed a method for generalising over the multiple appliance instances, in the following chapter we propose a novel method by which these general appliance models can be tuned to the appliance instances in a new household using only aggregate data.

Chapter 5

Tuning General Models using Aggregate Data

As identified in the previous chapter, some appliance behaviour is unique to a particular household and therefore cannot be captured by the general model of the appliance type. Such behaviour can be due to the unique characteristics of the appliance instances present in a household (e.g. a freezer with a defrost cycle), and also due to their pattern of usage by the household's occupants (e.g. a microwave often used on low power). Therefore, in this chapter we propose a novel method for learning such behaviour that is unique to a single household, which requires only general appliance models and household aggregate data.

More formally, this learning process directly corresponds to tuning the parameters of an appliance type's general model, $\mathbf{\Theta}$, to the specific appliance instance n in a household, $\boldsymbol{\theta}^{(n)}$, given only the household's aggregate data $\mathbf{x}$. Our approach differs from the training approach used by Kim et al. (2011), in which appliances are detected using a FHMM but are also required to be manually labelled. Similarly, Kolter and Jaakkola (2012) proposed a training approach in which individual appliance signatures are extracted from aggregated data, but again each signature was also required to be manual labelled with an appliance name. Last, Johnson and Willsky (2013) used a Bayesian FHMM to update appliance parameters without manual intervention, although they only evaluated the approach on households containing at most 5 appliance types.

In this chapter, we first propose a method which is able to identify and extract periods during which only one appliance is changing state from the aggregate load using only the general appliance model (Section 5.1). We then describe how these extracted signatures can be used to tune general appliance models to represent the specific appliance instances in a single household (Section 5.2). Next, we provide an empirical evaluation of our approach using the REDD data set, and benchmark our approach against the state of

Appliance type	**Window length (minutes)**
Refrigerator	200
Microwave	10
Washing machine	60
Dishwasher	120

Table 5.1: Window length for various appliance types.

the art which uses a FHMM to tune the general appliance model (Section 5.3). Finally, Section 5.4 summarises the findings of this chapter.

5.1 Extracting Appliance Signatures from an Aggregate Load

As discussed above, our proposed approach requires periods during which a single given appliance is operating to be extracted from an aggregate load. This is achieved by calculating the likelihood that a period of aggregate data was generated by a single appliance instance drawn from a given general appliance model. However, it is important to note that our approach aims to extract periods during which only the appliance of interest is changing state, and that other appliances might be drawing a constant power during this period. Therefore, in our approach, the base-load is first subtracted from the aggregate load before calculating the likelihood:

$$\bar{\mathbf{x}}_{i:j} = \mathbf{x}_{i:j} - \min(\mathbf{x}_{i:j}) \tag{5.1}$$

where $\mathbf{x}_{i:j}$ is a window of aggregate data $x_i, \ldots, x_j$, and $\bar{\mathbf{x}}_{i:j}$ is the same window after the base-load has been subtracted. This ensures that the distributions over the mean power demand for each state correspond between different signatures. The approach considers windows of aggregate data, for which the size of the window is determined by the maximum signature length encountered in the training data used in Chapter 4, as shown by Table 5.1. Longer window lengths can be used for appliances for which multiple sequential signatures can be extracted (e.g. refrigerator), while shorter window lengths should be assigned to appliances which are likely to be used once for a short period of time (e.g. microwave). We calculate the likelihood that a period of aggregate data was generated by a single appliance instance drawn from a given general appliance model as follows:

$$accept(\bar{\mathbf{x}}_{i:j}) = \begin{cases} true & \text{if } p(\bar{\mathbf{x}}_{i:j}|\boldsymbol{\Theta}) > D \\ false & \text{otherwise.} \end{cases} \tag{5.2}$$

where $\bar{\mathbf{x}}_{i:j}$ is a window of aggregate data after the base-load has been subtracted, $p(\bar{\mathbf{x}}_{i:j}|\boldsymbol{\Theta})$ is the likelihood of that window of data given the general appliance model

(a)

(b)

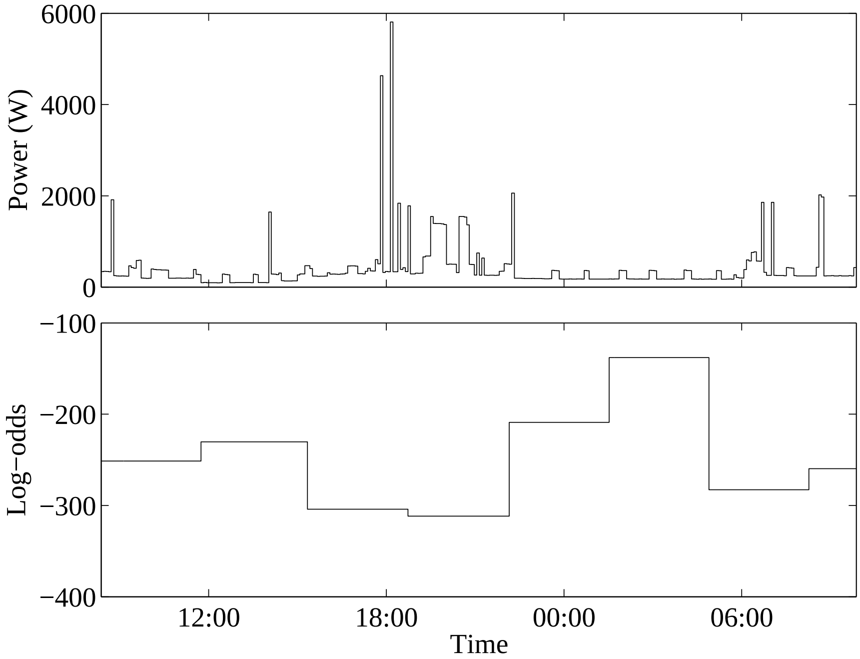


Figure 5.1: (a) Household power demand over 24 hour period. (b) Log-odds of window of aggregate power being generated by only refrigerator.

$\mathbf{\Theta}$ as in Equation 4.10, and D is an appliance specific likelihood threshold. This threshold is set such that the model will accept windows of data which can be explained by a set of appliance parameters drawn from the given appliance type's general model, and reject any windows of data generated by other appliance types or combinations of appliances. Therefore, this process effectively identifies windows of aggregate data during which only an appliance matching its general model changes state. It is worth noting that the data likelihood $p(\bar{\mathbf{x}}_{i:j}|\mathbf{\Theta})$ is inherently dependent upon the variance within an appliance type, since it decreases as the variance of the general appliance model increases. Therefore, we use the sub-metered data from Chapter 4 to calculate D for each appliance type as the minimum of $p(\mathbf{x}^{(n)}|\mathbf{\Theta})$ for each appliance instance n, and as a result this threshold generalises to unseen households.

Figure 5.1 gives an illustrative example of how appliance signatures can be extracted from an aggregate load in the case of the refrigerator. The figure shows the power demand of a household over a 24 hour period, and also the log-odds that each 4 hour window of data was generated by only the refrigerator. For most windows of data, it is clear that step changes in the aggregate power demand were generated by a combination of the refrigerator and a number of other appliances, and therefore received a low log-odds score. However, between 02:00 and 05:00 only the refrigerator contributed to changes in the aggregate power, and as a result the window receives a high log-odds score. Therefore, this period can be extracted from the aggregate load and used as an appliance signature with which the refrigerator general model can be tuned. We found that a step size equal to the window length to be sufficient to extract signatures for each of the modelled appliances. However, in households where aggregate data is more limited or where overlapping appliance usage is more common, we would expect that a smaller step size would allow a greater number of signatures to be extracted. Having introduced a method by which appliance signatures can be extracted from aggregate data, in the next section we describe how such signatures can be used to tune general appliance models.

5.2 Tuning General Appliance Models using Extracted Signatures

Once the signatures of a single appliance instance have been extracted from aggregate data, the aim is to tune the general model to include the behaviour of the appliance instance which is unique to the previously unseen household. Given that both the general model for this appliance type, $\boldsymbol{\Theta}$, and signatures sampled from the specific appliance instance are available, $\bar{\mathbf{x}}_{i:j}$, Bayesian integration (Ghahramani, 2001) provides a natural approach to infer the posterior distribution over such appliance instance parameters with the state sequence marginalised out:

$$\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \int p(\bar{\mathbf{x}}_{i:j}, \mathbf{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\Theta}) \, \mathrm{d}\mathbf{z} \tag{5.3}$$

In this setting, Bayesian updating provides a desirable trade-off between parameter tuning and avoiding model over-fitting. For example, when only a small number of appliance signatures are extracted from the aggregate load, the parameters are prevented from becoming over-fitted to one or two signatures. However, when many signatures are extracted from the aggregate load, the parameters are tuned to represent the repeatable behaviour of the appliance instance specific to that household. Since there is no analytical solution to this integral, we again use variational message passing implemented using Infer.NET for the same reasons as in Section 4.2.

Figure 5.2 illustrates the outcome of the tuning process using the microwave's *on* state as an example. It can be seen that the prior distribution, as learned during the generalisation method described in Chapter 4, shows a broad distribution over the mean power of all microwaves. In contrast, the posterior distribution, as tuned using the method described in this section, shows a more precise distribution over the mean of this specific microwave instance. However, it should be noted that the mean power of the *on* state, μ_k, is just one of the set of appliance model parameters, $\boldsymbol{\theta}$, and therefore it is not expected that appliances will be uniquely distinguishable using only this parameter. Having now introduced a method by which general appliance models can be tuned using only household aggregate data, in the next section we provide an empirical evaluation of our approach against the state of the art.

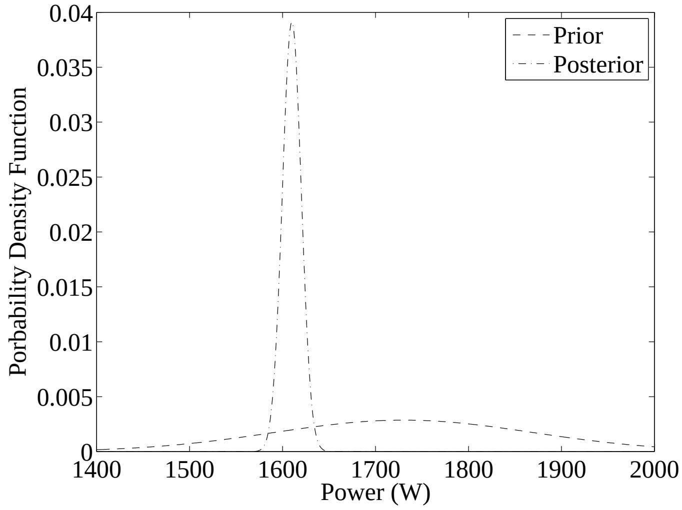


Figure 5.2: Probability density functions of microwave *on* state mean power prior and posterior distributions.

5.3 Empirical Evaluation of Model Tuning using REDD Data Set

We evaluated the benefit of tuning general appliance models using the Reference Energy Disaggregation Data set (REDD) (Kolter and Johnson, 2011). This data set was chosen as it is an open data set collected specifically for evaluating NIALM methods, and contains both household aggregate and circuit-level power demand measurements. Since many circuits contain only one appliance, these circuits represent the ground truth power demand for those appliances. As a result, we were able to evaluate how well a given appliance model explains each appliance's actual power demand. However, of the appliances investigated in Chapter 4, the kettle is not connected to an individual circuit in the REDD data set, and therefore could not be evaluated in this section. Furthermore, due to differences between American and European appliances, it was necessary to artificially increase the mean hyperparameter of the mean power distribution of the *on* state, $\hat{\lambda}_{on}$, of the general model for the microwave and washing machine. However, all other general model parameters were exactly as learned from the Tracebase data set.

We compare the approach described in this section (AT) to three benchmarks. The first benchmark (GT) uses the general appliance model as learned empirically in Chapter 4 without any model tuning. This variant represents the model fit of the general appliance models. The second benchmark (FT) uses standard Bayesian inference via Gibbs sampling over a FHMM when supplied with aggregate data and general appliance priors. This represents the state of the art for unsupervised learning in NIALM (Johnson and Willsky, 2013), and was implemented using `pyhsmm`.[1] This library was a natural choice as it provides an efficient Gibbs sampler for approximate inference in FHMMs. The

[1] `https://github.com/mattjj/pyhsmm`

Approach	Description
GT	General appliance model as learned from the Tracebase data set without any parameter tuning.
FT	General appliance model as learned from the Tracebase data set tuned via a FHMM.
AT	General appliance model as learned from the Tracebase data set tuned using signatures extracted from aggregate data.
ST	General appliance model as learned from the Tracebase data set tuned using signatures extracted from sub-metered data.

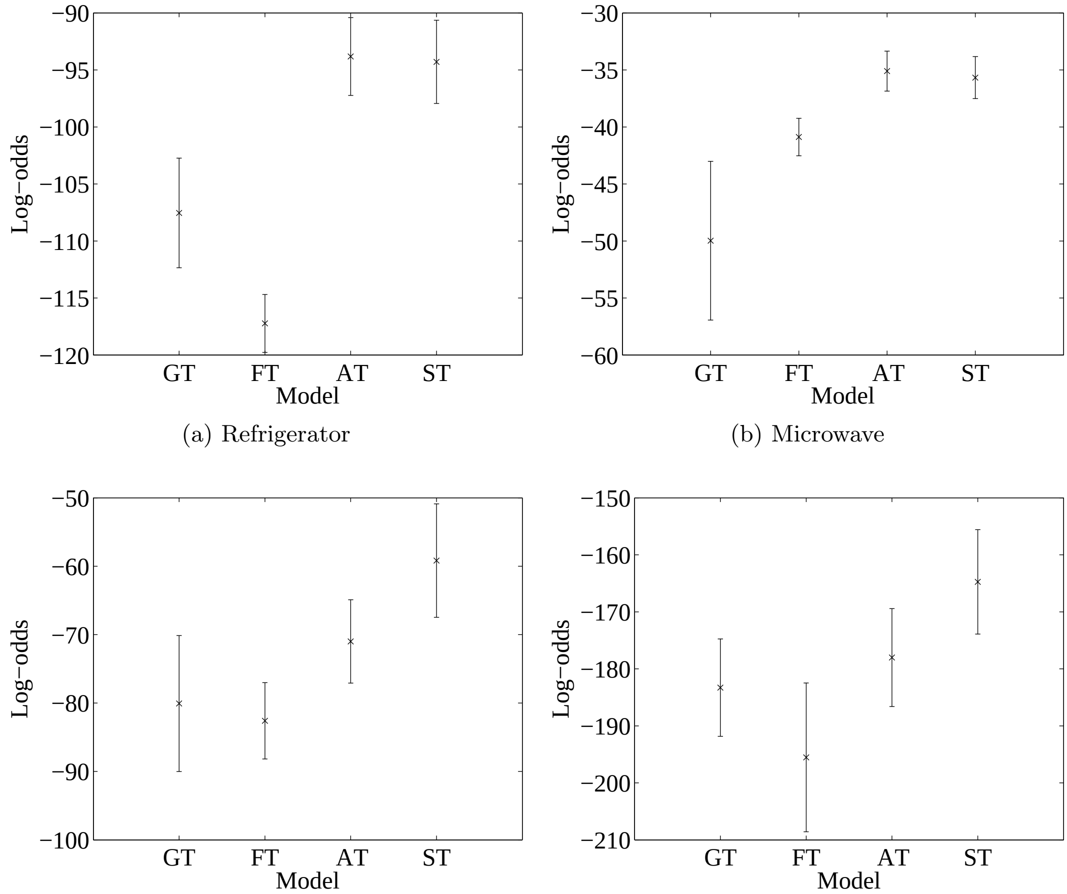
Table 5.2: Summary of approaches compared using REDD data set.

third benchmark (ST) tunes the general models using sub-metered data, through the approach described in Section 5.2. This approach represents the model fit in the ideal case where sub-metered data is available for model tuning. These four approaches are summarised in Table 5.2.

As in the previous section, we evaluate the extent to which an appliance model explains the appliance's power demand using the logit function applied to the model likelihood, given by Equation 4.10 and Equation 4.11.

Figure 5.3 shows the model log-odds for 4 common household appliances, each of which compares the model fit of our proposed approach against the three described benchmarks. The error bars represent the standard error in the mean. A clear trend is visible in that the model tuned using signatures extracted from aggregate data (AT) always outperforms the untuned general model (GT). In fact, in two cases it even performs comparably to the model tuned using sub-metered data (ST). This indicates that unsupervised model tuning using aggregate data is a practical alternative to the more intrusive method of supervised training through sub-metered data. Furthermore, it can be seen that the model tuned using signatures extracted from aggregate data (AT) consistently outperforms the current state of the art (FT) which uses a FHMM to tune appliance parameters. This is due to the FHMM method being unable to distinguish between appliances given only general appliance priors and aggregate data, and as a result learns appliance posteriors that consist of combinations of difference appliances.

It is also interesting to compare the benefit of model tuning shown in Figure 5.3 between appliances. The refrigerator and microwave show the most consistent increase in model log-odds as represented by the distinct error bars. This can be attributed to the many clean signatures that could be extracted by the AT method with which the general model could be tuned. This is in contrast to the washing machine and dishwasher, for which fewer, noisier signatures were extracted. As such, there is a smaller increase in the log-odds of the model tuned using aggregate data relative to the model tuned using sub-metered data. This indicates that model tuning will be less effective for appliances often used simultaneously with other appliances. It is also interesting to compare the performance of FT between different appliance types. For the microwave, the FT method



(a) Refrigerator (b) Microwave

(c) Washing machine (d) Dishwasher

Figure 5.3: Mean model log-odds for different training methods. Legend: GT - general model, FT - general model tuned using FHMM, AT - general model tuned using extracted signatures, ST - general model tuned using sub-metered data. Error bars represent standard error in the mean.

is able to improve the general appliance model using only aggregate data, while for the refrigerator, washing machine and dishwasher the FT tuning method actually produces an inferior appliance model.

We now investigate the benefit of tuning each individual parameter of the appliance models using the Kullback-Leibler (KL) divergence. The KL-divergence was chosen as it allows the difference between individual distributions to be compared, in contrast to the model likelihood which was used to compare complete appliance models in Figure 5.3. Table 5.3 shows the KL divergence between the models tuned using sub-metered data (ST) and the three approximations (GT, FT, AT) as $D_{\mathrm{KL}}(\mathrm{ST}||\mathrm{GT})$, $D_{\mathrm{KL}}(\mathrm{ST}||\mathrm{FT})$ and $D_{\mathrm{KL}}(\mathrm{ST}||\mathrm{AT})$ respectively. This table allows the information lost to be compared when each approach is used to approximate the distributions learned from sub-metered data. It can be seen that $D_{\mathrm{KL}}(\mathrm{ST}||\mathrm{FT})$ is systematically greater than $D_{\mathrm{KL}}(\mathrm{ST}||\mathrm{AT})$ for both the transition and emission distributions across all appliances. These high divergence values further highlight that model tuning via signature extraction is preferable

Appliance	Measure	Initial	Transition	Emission
Refrigerator	$D_{\mathrm{KL}}(\mathrm{ST}\|\|\mathrm{GT})$	0.183	0.164	3.735
	$D_{\mathrm{KL}}(\mathrm{ST}\|\|\mathrm{FT})$	0.369	2.613	26.289
	$D_{\mathrm{KL}}(\mathrm{ST}\|\|\mathrm{AT})$	0.348	0.525	3.376
Microwave	$D_{\mathrm{KL}}(\mathrm{ST}\|\|\mathrm{GT})$	0.072	0.107	1.944
	$D_{\mathrm{KL}}(\mathrm{ST}\|\|\mathrm{FT})$	0.017	1.441	238.544
	$D_{\mathrm{KL}}(\mathrm{ST}\|\|\mathrm{AT})$	1.630	0.469	0.963
Washing machine	$D_{\mathrm{KL}}(\mathrm{ST}\|\|\mathrm{GT})$	0.185	0.057	2.784
	$D_{\mathrm{KL}}(\mathrm{ST}\|\|\mathrm{FT})$	0.008	4.451	5.401
	$D_{\mathrm{KL}}(\mathrm{ST}\|\|\mathrm{AT})$	0.674	0.137	3.599
Dishwasher	$D_{\mathrm{KL}}(\mathrm{ST}\|\|\mathrm{GT})$	0.178	0.356	7.209
	$D_{\mathrm{KL}}(\mathrm{ST}\|\|\mathrm{FT})$	0.599	2.809	9.189
	$D_{\mathrm{KL}}(\mathrm{ST}\|\|\mathrm{AT})$	0.875	0.990	5.079

Table 5.3: KL divergence between the model tuned using sub-metered data (ST) and the three approximations of this model (GT, FT and AT).

to the current state of the art which uses FHMMs. It is also interesting to note that although $D_{\mathrm{KL}}(\mathrm{ST}||\mathrm{GT})$ is often slightly less than $D_{\mathrm{KL}}(\mathrm{ST}||\mathrm{AT})$ for the transition matrix, $D_{\mathrm{KL}}(\mathrm{ST}||\mathrm{GT})$ is almost always much greater than $D_{\mathrm{KL}}(\mathrm{ST}||\mathrm{AT})$ for the emission distribution. This indicates that the tuning process is more important for the emission distributions than the transition matrices. However, it is worth noting that the emission distribution of $D_{\mathrm{KL}}(\mathrm{ST}||\mathrm{FT})$ for the microwave received a high divergence score but also explained the actual appliance data with a reasonable likelihood in Figure 5.3, which indicates that ST is not the only model that can explain actual appliance data with a reasonable likelihood. The next section summarises the findings of this chapter.

5.4 Summary

In this chapter, we began by proposing a method which is able to extract individual appliance signatures from an aggregate load using only general appliance models of appliance types. We then showed how these signatures could be used to tune a general model of an appliance type to the specific appliance instances in a single household. Next, we gave an empirical evaluation of our proposed tuning method using the REDD data set. We showed that models tuned using our proposed approach outperform the general models of appliance types, and can even perform comparably to models tuned using sub-metered data from the test appliance instance. Furthermore, we showed that our tuning approach outperforms the state of the art, which uses standard Bayesian inference in a FHMM to update the appliance parameters. Finally, we compared the general models, the models tuned using our proposed approach, and the models tuned using the FHMM, to the models as learned from sub-metered appliance data. This comparison showed that the tuning process was most important when updating the appliances' emission distributions.

Having introduced a method by which general appliance models could be tuned to the specific appliances in a household given only aggregate data, we now go on to describe a large-scale application of the methods introduced in both Chapter 4 and this chapter.

Chapter 6

Case Study Application to Cold Appliances

This chapter applies the two novel methodologies introduced in Chapter 4 and Chapter 5 to the problem of energy consumption estimation of fridges and freezers using a large scale data set. We begin by motivating the study of a single appliance type at a large scale (Section 6.1), followed by a justification for the use of the Colden Common data set (Section 6.2). We then evaluate the accuracy by which the approach is able to estimate the energy consumption of combined fridge freezers (Section 6.3). However, many households contain a separate fridge and freezer, for which we propose an extension to our model and provide a description of its evaluation (Section 6.4). We next provide a discussion of the generality of the approach described in this chapter to appliances other than cold appliances (Section 6.5). Finally, we summarise the chapter (Section 6.6).

6.1 Motivation

Similar to all related work, the previous chapters have used public data sets containing sub-metered appliance data to evaluate the accuracy of our approach on a number of different appliance types. Such data sets are essential in order to compare the accuracy of different energy disaggregation algorithms. However, sub-metered data sets are intrinsically expensive to collect due to the number of required sensors. Furthermore, existing sub-metered data sets cover a maximum of 10 households, and are often biased towards smaller households for which instrumentation is cheaper and simpler. As a result, existing sub-metered appliance data sets do not sufficiently represent the range of households in the UK. To account for this shortcoming, we evaluate our approach using the Colden Common data set which includes aggregate data from 117 households, as described in Section 6.2.

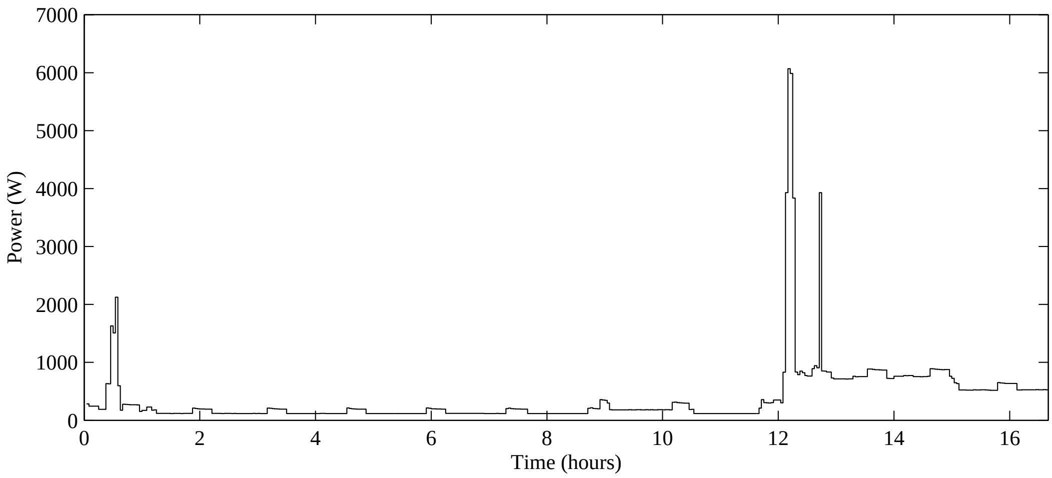
When applying disaggregation algorithms to large amounts of real data, model extensions are often required for each appliance in order to ensure that the disaggregation process is robust to a wide range of households. For this reason, we have chosen to focus upon only cold appliances; combined fridge freezers, individual fridges and individual freezers. This appliance type was selected since it is both a major energy consumer and at least one cold appliance is present in almost all households. Furthermore, it has been estimated that households can save an average of 310 kWh per year by replacing such cold appliances with new energy efficient appliances (Zimmermann et al., 2012). We provide a detailed description of the extensions that were required for the robust modelling of cold appliances in Section 6.3.

6.2 Colden Common Data Set

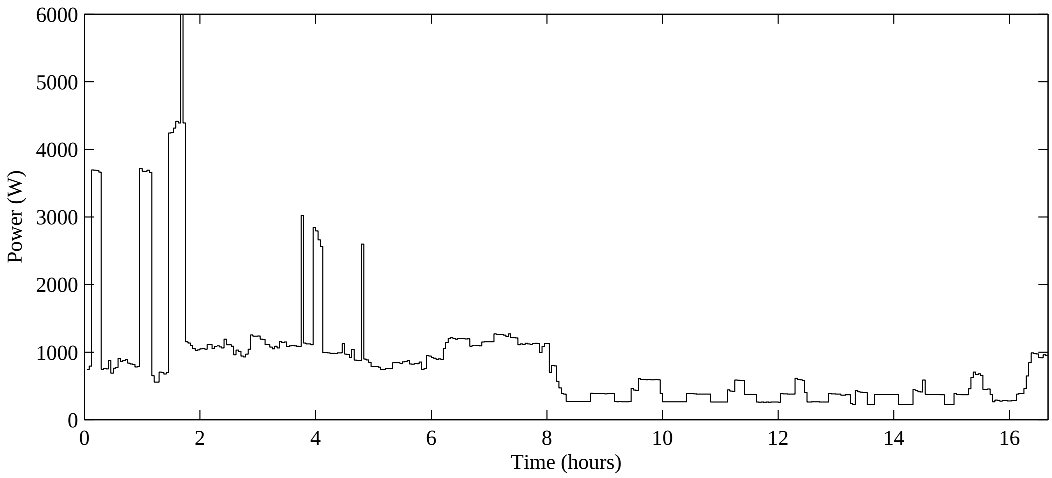
In this chapter, we use the Colden Common data set to evaluate the robustness of our approaches described in Chapter 4 and Chapter 5. The Colden Common data set consists of power data collected from 117 households in the village of Colden Common, Hampshire, UK. Current clamps were used to measure each household's aggregate power demand at 1 second intervals, which we down-sample to 10 second intervals to mimic the reporting rate of UK smart meters (Department of Energy & Climate Change, 2013a).

Appliance sub-meters were not used to collect this data set, and therefore power data for individual appliances is not available. As a result, it was necessary to manually label the power demand of the cold appliances in this data set. In many cases, the fridge and freezer loads can be trivially identified by manual inspection. However, in other households it was not possible due to measurement noise or interference from other appliances. Of the 117 households, we identified 38 households as having a combined fridge freezer, 32 households as having a separate fridge and freezer, and 47 households for which it was not possible to manually identify a fridge or freezer load. Figure 6.1 shows examples of these three cases respectively. Figure 6.1 (a) shows a household in which the power demand of a combined fridge freezer is clearly visible during the hours 2-8. Figure 6.1 (b) shows a household in which the power demands of a separate fridge and freezer are visible during the hours 8-16. Last, Figure 6.1 (c) shows a household in which no fridge or freezer power demands are visible as a result of measurement noise and interference from other appliances. The windows identified in Figure 6.1 (a) and (b) are exactly the periods which our proposed approach exploits in order to infer the energy consumption of fridges and freezers using only a general appliance model and a household's aggregate load.

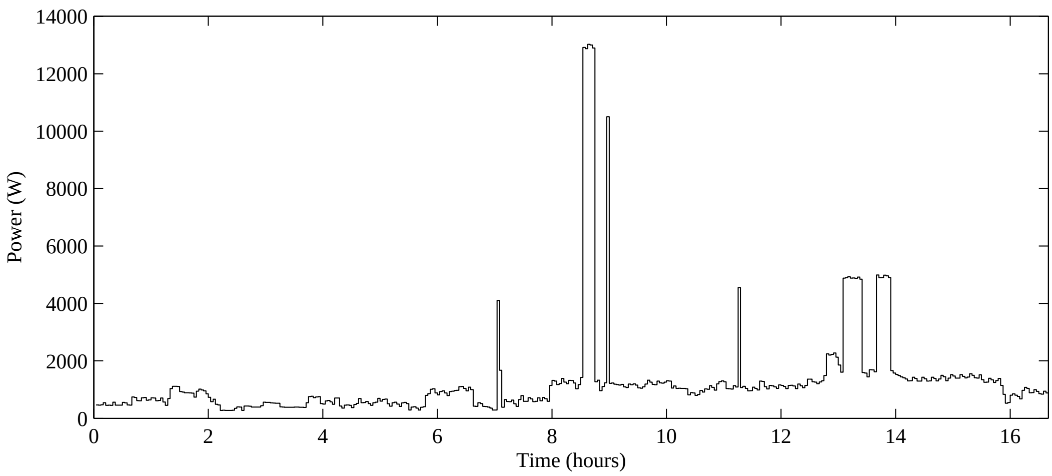
Figure 6.2 shows the annual energy consumption of the combined fridge freezers in the Colden Common households as labelled manually. The figure also shows a dashed line which represents the 217 kWh annual energy consumption of a £429 energy-efficient



(a) House with a combined fridge freezer



(b) House with a separate fridge and freezer



(c) House with large appliance and measurement noise

Figure 6.1: Examples of aggregate power demand of three households.

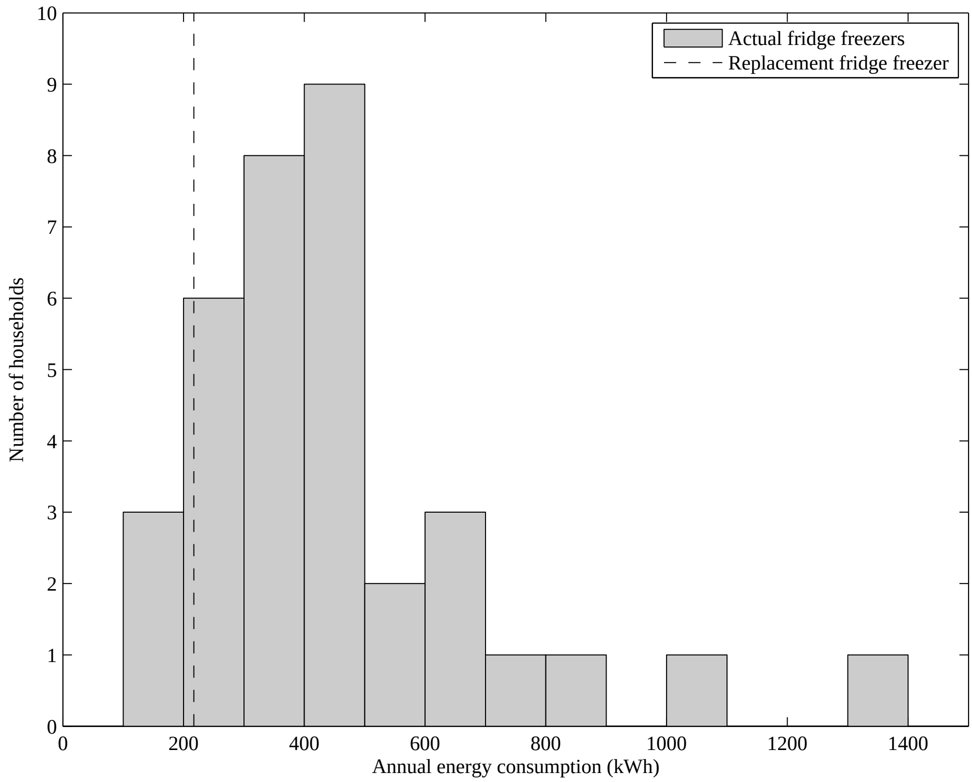


Figure 6.2: Annual energy consumption of combined fridge freezers.

fridge freezer.[1] It can be seen that the vast majority (32 out of the 35 fridge freezers) consume more energy per year than the replacement, and therefore would save energy by replacing their fridge freezer. These annual energy savings are exactly the feedback which we aim to quantify by automatically disaggregating smart meter data.

Figure 6.3 shows the total cold appliance annual energy consumption for the households containing two cold appliances. The figure also shows a dashed line which represents the annual energy consumption of the same £429 energy-efficient combined fridge freezer. It can be seen that almost all households' cold appliances (28 out of the 29 households) consume more energy per year than the replacement, and therefore would save energy by replacing their two cold appliances with a combined fridge freezer. Again, these annual energy savings are exactly the feedback which we aim to quantify by automatically disaggregating smart meter data.

In the following section we describe how the energy efficiency of combined fridge freezers can be inferred (Section 6.3). This is followed by a description of how the energy efficiency of separate fridges and freezers can be inferred (Section 6.4).

[1] `http://www.appliancecity.co.uk/liebherr/fridges-and-freezers/cup3221/product-16287/`

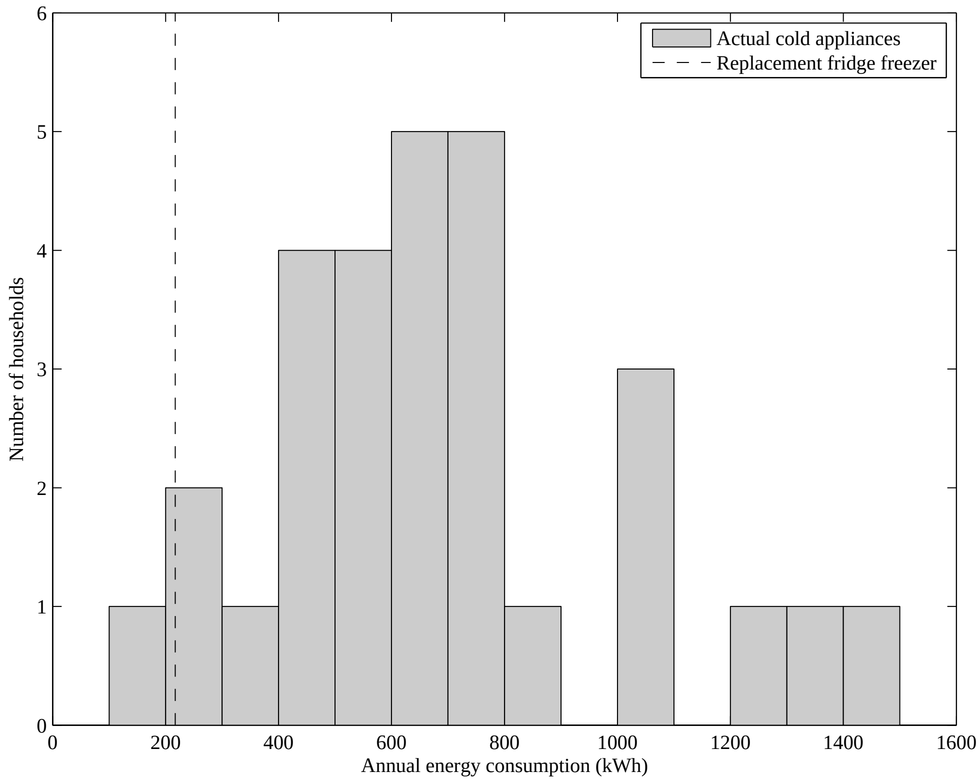


Figure 6.3: Total annual energy consumption of separate fridges and freezers.

6.3 Combined Fridge Freezer Disaggregation Evaluation

According to a recent survey, 62% of UK households contain a combined fridge freezer (Zimmermann et al., 2012). In this section we apply the methodology introduced in Chapter 4 and Chapter 5 to the disaggregation of combined fridge freezers. First, we evaluate the accuracy by which households with a combined fridge freezer can be automatically identified (Section 6.3.1). We then evaluate the accuracy by which the energy consumption of these appliances can be estimated (Section 6.3.2). Last, we show how the estimated appliance energy consumption can be used to provide actionable energy saving feedback (Section 6.3.3).

6.3.1 Detection of Combined Fridge Freezer Households

In order to determine which households contain a combined fridge freezer, we first generated a general fridge freezer appliance model using the approach described in Chapter 4 and the Tracebase data set. We then used the approach described in Chapter 5 to determine the likelihood that any window of aggregate data was generated by only a combined fridge freezer. If this likelihood is greater than a given threshold for any window of data, then that household is classified as containing a fridge freezer and no other cold appliances. This detection therefore relies on the general appliance model parameters, $\mathbf{\Theta}$, and the likelihood threshold parameter, D. This section investigates the detection accuracy for various settings of these parameters.

We evaluate the detection accuracy of our approach using the receiver operating characteristic (ROC), which illustrates the trade-off between an approach's true positive rate (TPR) and false positive rate (FPR), which are defined as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6.1}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{6.2}$$

where:

- TP - The number of combined fridge freezer households classified correctly.
- FP - The number of households without a fridge freezer classified incorrectly.
- FN - The number of combined fridge freezer households classified incorrectly.
- TN - The number of households without a fridge freezer classified correctly.

Since all existing disaggregation algorithms assume knowledge of which appliances are present in a household, we were unable to benchmark our methodology against other approaches. We therefore evaluated the detection accuracy of our approach while varying the appliance threshold, D, as described in Section 5.1. After a manual inspection of the incorrectly classified households using a general appliance model built using the Tracebase data set, we observed a trend in which combined fridge freezers with a higher power demand were being systematically misclassified. This can be attributed to the fact that only refrigerator data was available in the Tracebase data set, and therefore the general model is not representative of the range of combined fridge freezer power demands. To control for this limitation, we created a second appliance model with the same parameters as were learned from the Tracebase data set except with the mean power of the *on* state increased from 100 W to 150 W.

Figure 6.4 shows a ROC curve of the trade-off between the TPR and FPR of the two appliance models. Each curve shows how the detection accuracy of a given appliance model changes as the likelihood threshold is varied. A TPR of 1 and a FPR of 0 represents perfect detection accuracy, while the dashed line along TPR = FPR represents the detection accuracy of a random classifier. The figure compares two different appliance models and one benchmark: the solid line represents the model learned empirically from the Tracebase data set, the dashed and dotted line represents the same model but with the increased *on* state mean power, and the dashed line represents the detection rate of the random classifier benchmark. The figure shows a common trend that the TPR for both appliance models increases rapidly while the FPR is low, while there is little or

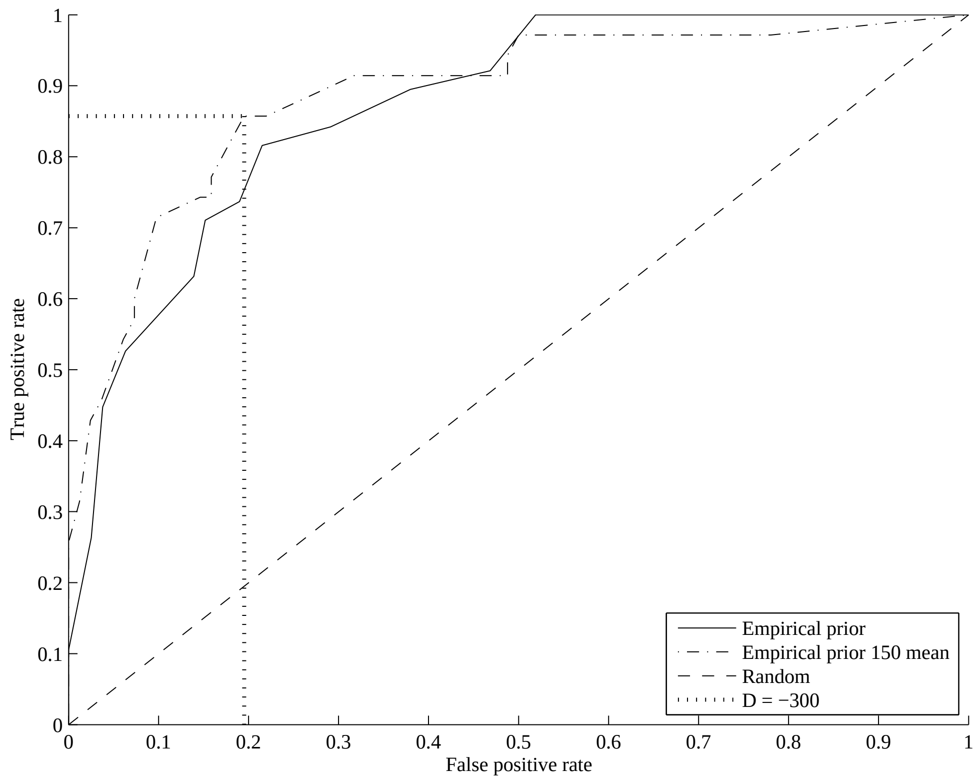


Figure 6.4: ROC curve showing detection accuracy of combined fridge freezer.

no increase in the TPR for increases in high FPR. A particularly favourable trade-off is visible at $D = -300$, which correctly detects more than 85% of households while also producing less than 20% false positives.

Furthermore, it can be seen that the two curves intersect at approximately FPR = 0.5, TPR = 0.9, and therefore the appliance model build from the Tracebase data is preferable in situations where a high FPR is acceptable, while the appliance model with an increased mean power is preferable in situations where only a low FPR is acceptable. It is likely that a recommendation system would favour a conservative method in which a low FPR is required to prevent a household's occupants from being discouraged by obviously incorrect feedback. Therefore, we believe the appliance model with an increased mean power to be preferable in most realistic scenarios. This matches the intuition that a model representing combined fridge freezers would outperform a model representing only refrigerators. Having described a method by which households containing a combined fridge freezer can be identified, we now describe how the energy efficiency of such appliances can be inferred.

6.3.2 Inference of Energy Consumption

Having determined which households contain a combined fridge freezer, the next step is to estimate each appliance's annual energy consumption. This is performed using the

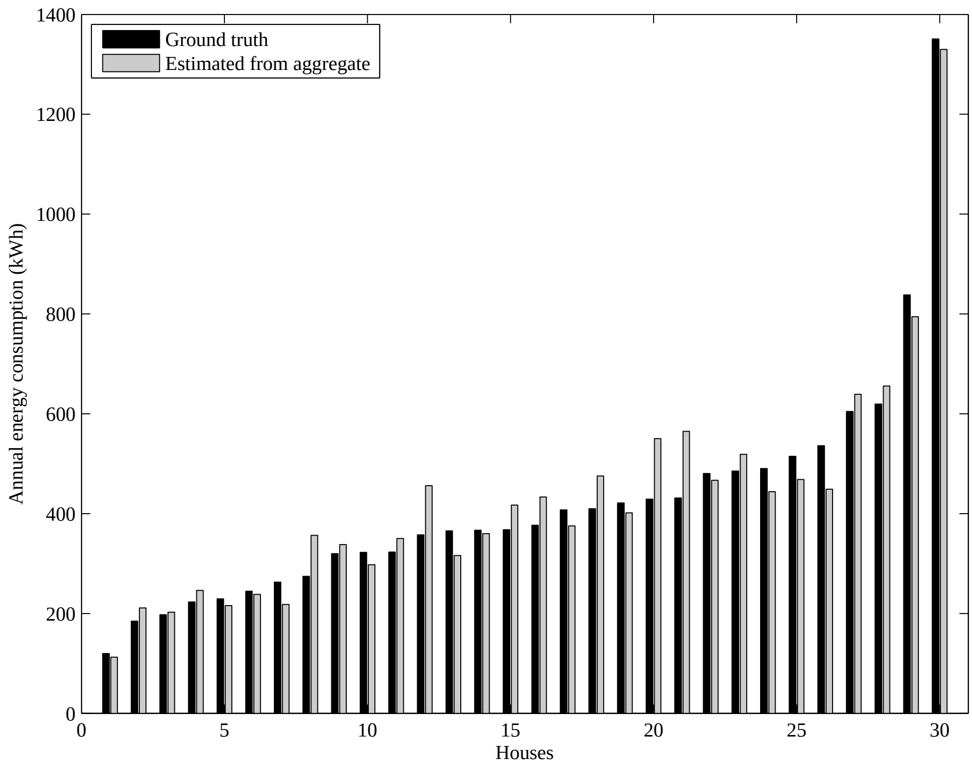


Figure 6.5: Ground truth and estimated annual energy consumption of combined fridge freezers, ordered by ascending ground truth consumption.

general appliance model learned from the Tracebase data set as described in Chapter 4 with the increased mean as described in Section 6.3.1. Furthermore, we tune the general appliance model using the method described in Chapter 5, with a log likelihood threshold of $D = -300$.

Figure 6.5 shows a bar graph of the actual and an estimate of the annual energy consumption of combined fridge freezers for the 30 households which were correctly detected as containing a combined fridge freezer. It can be seen that for most households, including those with high energy consuming fridge freezers, the difference between the actual and estimated energy consumption is small (mean = 42.4 kWh), relative to the actual energy consumption. This indicates that it is possible to provide highly accurate feedback regarding combined fridge freezer replacement for households containing only one cold appliance.

Figure 6.6 shows a bar graph of the error for each household normalised by the ground truth energy consumption, where positive values represent overestimates of the energy consumption and negative values represent underestimates of the energy consumption. It can be seen that almost all households achieve a normalised error of less than approximately 0.3, indicating that most households' fridge freezer energy consumption is estimated to within 30% of the appliances' actual energy consumption. It can also be seen that there is a slight bias towards overestimating the energy consumption of the

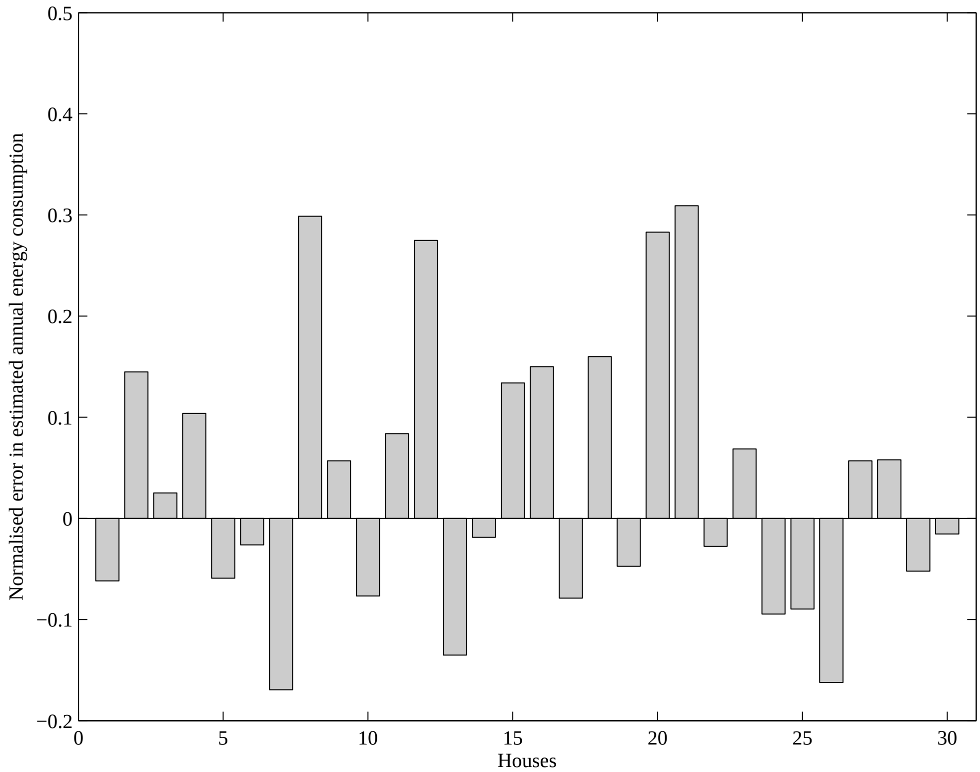


Figure 6.6: Normalised error in the estimate of annual fridge freezer energy consumption, ordered by ascending ground truth consumption. Positive values represent overestimates, negative values represent underestimates.

fridge freezers (mean = 17.4 kWh). This is likely a result of appliances other than the combined fridge freezer contributing to the power demand in the extracted signatures, since appliances only consume energy as opposed to generating energy. This effect could be controlled for by subtracting the mean error from each estimate, and as a result decreasing the overall error. Finally, there does not appear to be any systematic bias for households containing high energy consuming fridge freezers compared to those with low energy consumptions.

Figure 6.7 shows a histogram of the absolute error for each household. It can be seen that 27 of the 30 households have an error of less than 100 kWh per year. Summing the 217 kWh annual energy consumption of the replacement combined fridge freezer introduced in Section 6.2 and the 100 kWh error produces a lower bound of 317 kWh per year, above which we can say with 90% confidence that the replacement of the combined fridge freezer will result in energy savings.

Furthermore, the energy savings for the combined fridge freezers of more than 317 kWh annual consumption can be quantified as the difference between their estimated energy consumption and that of the replacement appliance. The uncertainty around these energy savings can even be estimated using the mean absolute error of 42.4 kWh. Having evaluated the accuracy by which the energy consumption of combined fridge freezers

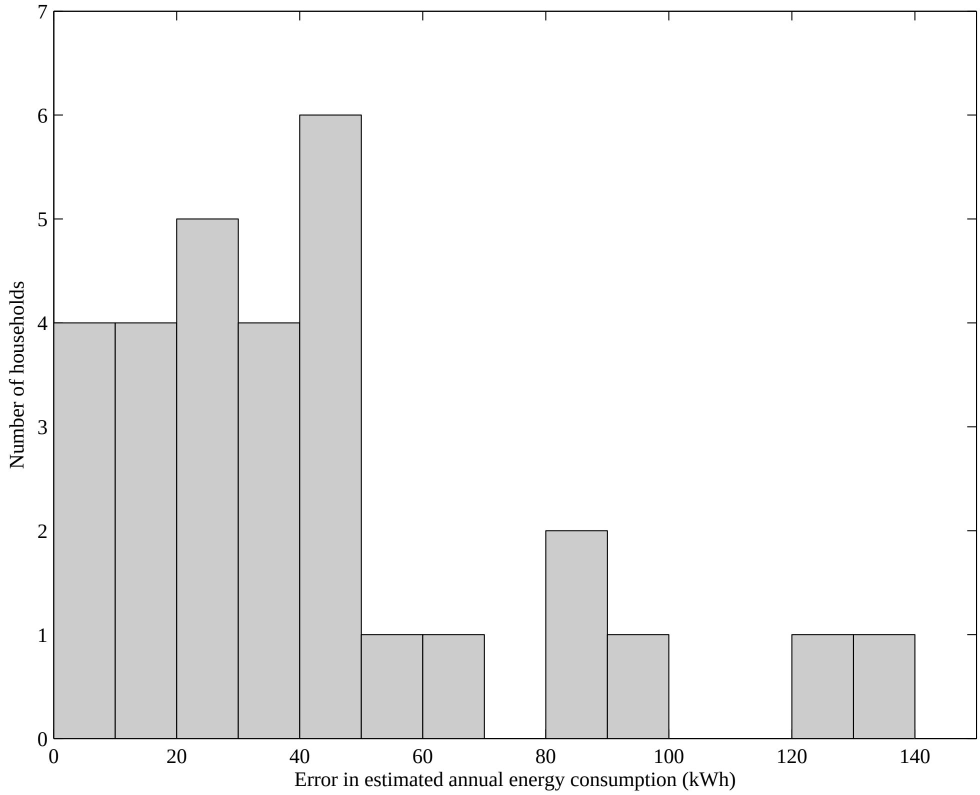


Figure 6.7: Absolute error in the estimate of annual fridge freezer energy consumption.

can be inferred, we now discuss the potential feedback that could be provided to the household's occupants.

6.3.3 Example of Potential Feedback

To illustrate the feedback that could be provided for an actual household, we consider house 5 as an example. House 5 was selected because it contains a C grade combined fridge freezer of median energy consumption for which our system would recommend the appliance's replacement. Our approach estimates the combined fridge freezer's annual energy consumption as 519 $\pm$ 42 kWh. The annual energy savings for replacing this appliance with a new combined fridge freezer can then be quantified as 302 $\pm$ 42 kWh, assuming a replacement which consumes 217 kWh per year.[2] Furthermore, these figures can be converted into financial savings, which for house 5 equates to £45.30 $\pm$ £7.20 per year, assuming a rate of £0.15/kWh. Finally, the time until the annual savings have offset the cost of the new combined fridge freezer can be estimated as 9.5 years, assuming a replacement cost of £429.[2] These figures provide the household occupants with the required information in order to make an informed decision of whether to replace their combined fridge freezer.

[2] `http://www.appliancecity.co.uk/liebherr/fridges-and-freezers/cup3221/product-16287/`

Having shown how the energy efficiency of combined fridge freezers can be inferred in this section, the following section extends our approach to cover households with separate refrigerators and freezers.

6.4 Separate Fridge and Freezer Disaggregation Evaluation

In this section we apply the methodology introduced in Chapter 4 and Chapter 5 to households with separate fridges and freezers, which account for 45% of UK homes (Zimmermann et al., 2012). First, we describe an extension to our appliance model to allow the detection of mutliple appliances using a FHMM (Section 6.4.1). We then evaluate the accuracy by which households with separate fridges and freezers can be automatically identified (Section 6.4.2). Next, we propose a second extension to exploit the quasiperiodic behaviour required to disaggregate two appliances with similar power demands using a FHSMM (Section 6.4.3), for which we subsequently evaluate the accuracy of appliance energy consumption estimation (Section 6.4.4). Last, we show how the estimated appliance energy consumption can be used to provide actionable energy saving feedback (Section 6.4.5).

6.4.1 Extension to Factorial Hidden Markov Model

In households with a combined fridge freezer, the fridge and freezer share a single compressor, and therefore only one signal appears in the household aggregate power demand. However, for households with a separate fridge and freezer, the power demands of the appliances are completely independent and therefore the sum of two out-of-phase signals will appear in the aggregate power demand. As a result, it is not be possible to apply the tuning approach described in Chapter 5 given the current model structure.

To address this shortcoming, we adopt a FHMM structure (Ghahramani and Jordan, 1997) instead of the basic HMM structure. A FHMM consists of multiple Markov chains, and an observation sequence which represents the combination of each Markov chain's emission. Specifically, our model consists of two Markov chains; one for the fridge and one for the freezer, while the observation sequence represents the sum of each appliance's power demand. The following section describes how we apply the FHMM to detect households with separate fridges and freezers.

6.4.2 Detection of Separate Fridge Freezer Households

We now use the FHMM to identify households with separate fridges and freezers through a similar approach as was used for combined fridge freezers in Section 6.3.1. To do so, we consider the same two appliance models, $\mathbf{\Theta}$, as learned from the Tracebase data set in Chapter 4; the empirical general model, and also the extension of this model with an increased *on* state mean. This allows the probability to be evaluated that windows of aggregate data were generated by only two cold appliances. Again, we compare the likelihood of each window against a threshold, D. This section investigates the detection accuracy of both appliance models for various settings of this likelihood threshold.

Figure 6.8 shows a ROC curve of the detection accuracy of our approach using the appliance models learned from the Tracebase data set. It is interesting to note that, unlike the gradual curve of the combined fridge freezer detection, this ROC curve appears to consist of three distinct sections. At values of low FPR, the TPR increases sharply until approximately FPR = 0.1, TPR = 0.5, where the curves run parallel to the TPR = FPR diagonal until approximately FPR = 0.4, TPR = 0.9, where the TPR increases gradually until FPR = 1, TPR = 1. This indicates that the most preferable threshold values lie between the two change points of FPR = 0.1, TPR = 0.5, and FPR = 0.4, TPR = 0.9.

It is also interesting to compare the detection accuracy of the two appliance models. Unlike the ROC curve for combined fridge freezers, neither approach appears to be preferable for a range of likelihood thresholds. This is likely due to the range of cold appliances present in the households (fridges, freezers and combined fridge freezers), and as a result, increasing the mean *on* state power results in no overall improvement in performance. In this following section, we describe a further extension to the FHMM to allow separate fridges and freezers to be disaggregated.

6.4.3 Extension to Factorial Hidden Semi-Markov Model

In households in which a minimum of one appliance was found to be operating in all windows of data, it can be inferred with high confidence that this appliance is a combined fridge freezer. This is due to the fact almost all households have both refrigeration and freezer facilities. However, in households in which a minimum of two appliances are detected across all windows of data, it is possible that either both appliances are cold appliances, or one appliance is a combined fridge freezer and the other is an appliance other than a cold appliance (e.g. lighting). This occurs because HMMs do not represent the periodic nature of certain signals. It is therefore necessary to determine whether both appliances are in fact cold appliances before useful feedback can be provided.

We use the presence of quasiperiodic behaviour to determine whether both appliances are in fact a cold appliance. Quasiperiodicity is defined as behaviour that recurs at a repeatable interval with some random component. As a result, the periodicity is not prefectly constant and may vary slightly. To represent this behaviour, we extend our graphical model of appliances from HMMs to HSMMs (Yu, 2010). In a HSMM, the number of sequential emissions from each state is modelled explicitly, which we model

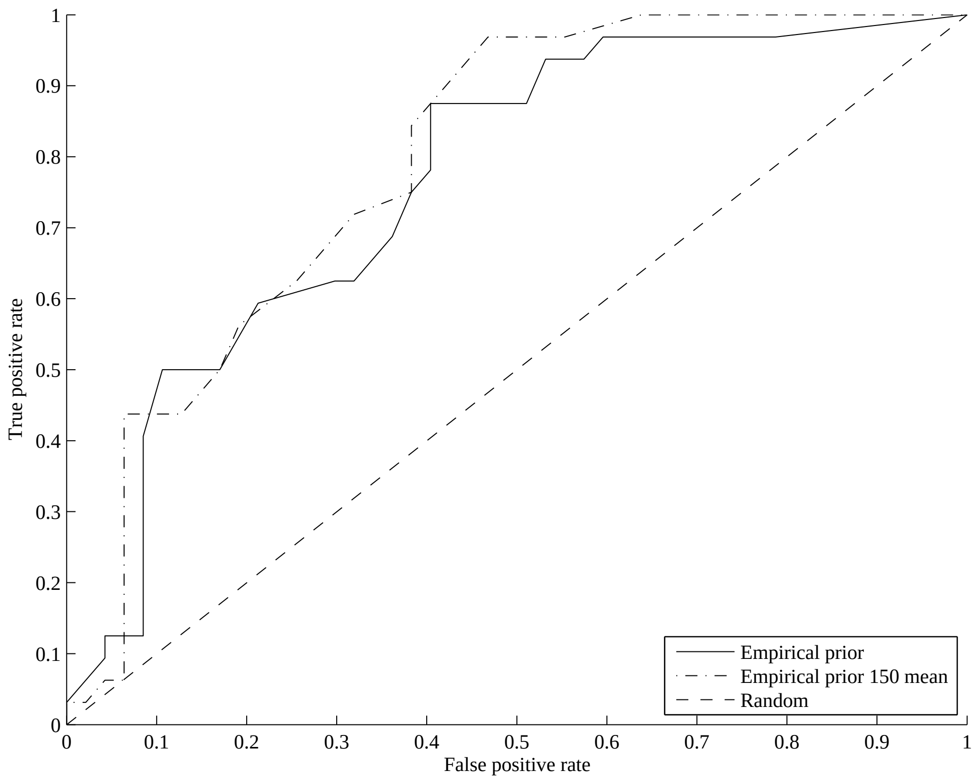


Figure 6.8: ROC curve showing detection accuracy of individual fridges and freezers.

using a Poisson distribution, rather than in a HMM in which the states are assumed to evolve according to a geometric distribution. In the context of appliance modelling, HSMMs provide a more representative model of quasiperiodic appliances, such as fridges, freezers, and combined fridge freezers. As with HMMs, HSMMs can be combined to form a factorial hidden semi-Markov model (FHSMM), allowing inference to be performed over multiple quasi-periodic components. The following section applies the FHSMM to infer the efficiency of separate fridges and freezers.

6.4.4 Inference of Energy Consumption

We determine whether both appliances are cold appliances by inferring the posterior distribution of each appliance's model parameters. This is performed using the general appliance model learned from the Tracebase data set as described in Chapter 4 as the prior distribution over the model parameters. However, to prevent both chains from matching each cold appliance equally, we use asymmetric priors such that the first chain's prior contains increased support for higher frequency periodicity, while the second chain's prior contains support for lower frequency periodicity. This encourages the two chains to converge towards a different cold appliance.

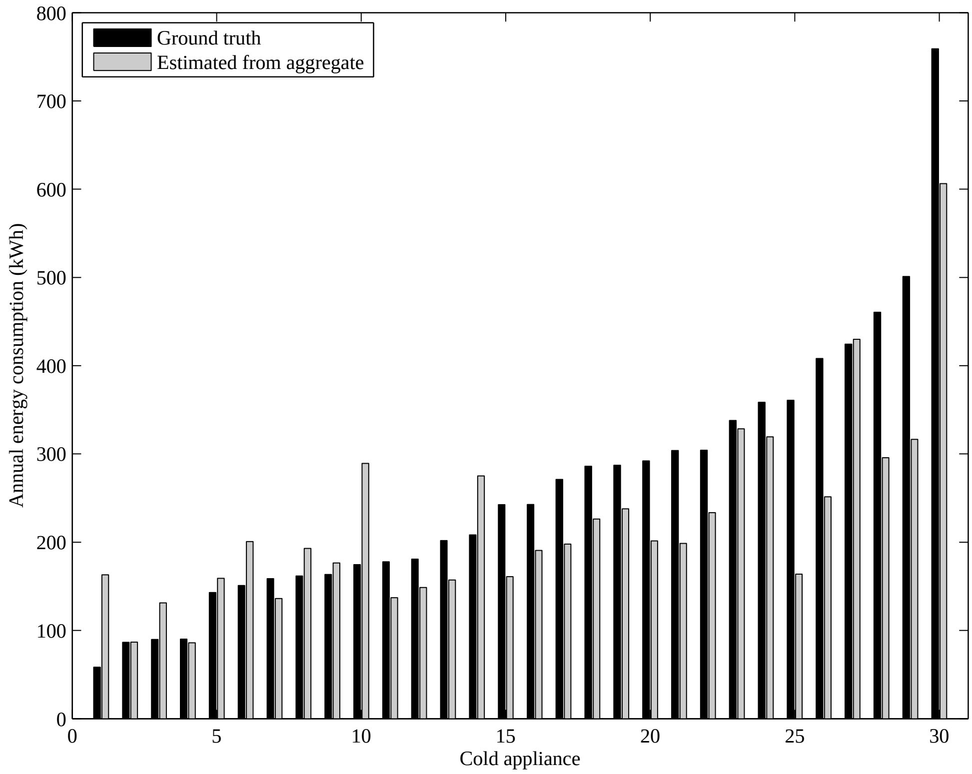


Figure 6.9: Ground truth and estimated annual energy consumption of cold appliances for households with two cold appliances, ordered by ascending ground truth consumption.

We now evaluate the accuracy by which our approach estimates the energy consumption of the cold appliances in two ways. First, we evaluate the accuracy of each individual cold appliance within the households containing two cold appliances. Although this information does not directly enable particular feedback, the learned appliance parameters would be required by any disaggregation algorithm. Second, we evaluate the accuracy by which the total energy consumption of each household's cold appliances can be estimated. This information directly enables feedback to be provided to a household's occupants regarding the potential energy savings as a result of replacing their cold appliances with a single combined fridge freezer.

Figure 6.9 shows a bar graph of the actual and estimated annual energy consumption of each cold appliance for the households containing two cold appliances. It can be seen that for most households, the difference between the actual energy consumption and the estimated energy is small (mean = 69.2 kWh). However, the greatest errors tend to occur in households which contain the highest energy consuming cold appliances. This is likely due to the inference process over-allocating energy to one cold appliance while under-allocating energy to the other cold appliance.

Figure 6.10 shows a bar graph of the error for each cold appliance normalised by the ground truth energy consumption, where positive values represent overestimates of the

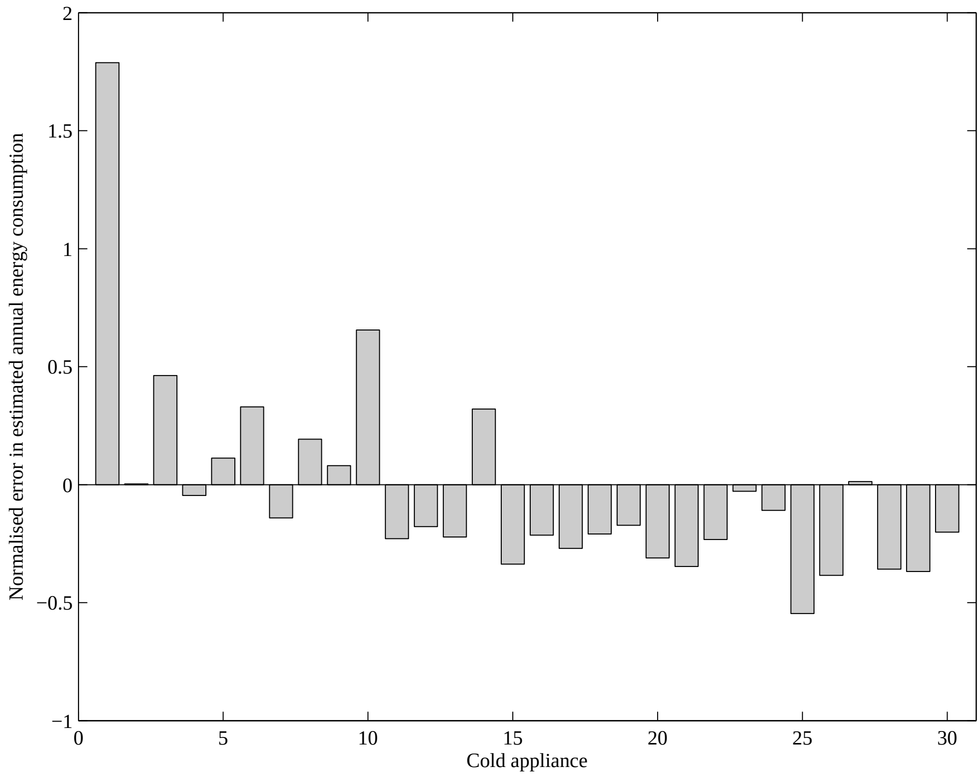


Figure 6.10: Normalised error in the estimate of annual cold appliance energy consumption for households with two cold appliances, ordered by ascending ground truth consumption. Positive values represent overestimates, negative values represent underestimates.

energy consumption and negative values represent underestimates of the energy consumption. It can be seen that most households achieve a normalised error of less than approximately 0.5, indicating that most households' cold appliance energy consumption is estimated to within 50% of the appliances' actual energy consumption.

However, the first cold appliance shown by Figure 6.10 appears to be an outlier to the general trend of low error, in that its energy consumption is estimated at approximately 1.8 times its actual consumption. Although a clean signature containing both cold appliances' signatures was successfully extracted for this household, the FHSMM was unable to disaggregate the two signals. This is likely due to the choice of the Poisson distribution to model each states' duration. Since a Poisson distribution is defined by a single parameter which represents both its mean and variance, the resulting distribution is not restrictive enough to discriminate between the two appliances, whose power demand never overlaps. As a result, each measured power demand is evenly assigned to both appliances. This problem could be avoided by using a distribution defined by multiple parameters to represent appliance state durations, such as the Gaussian distribution, therefore allowing distributions to be described with a large mean but small variance. However, a larger data set would first be required to determine whether this is either a unique or a common scenario.

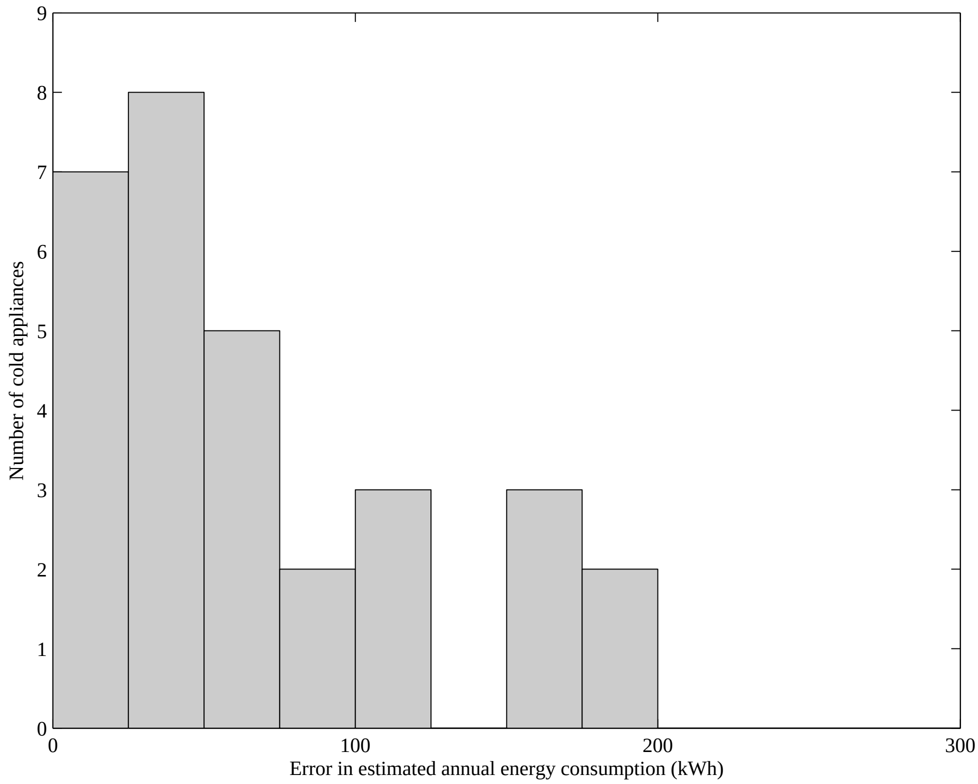


Figure 6.11: Absolute error in the estimate of annual cold appliance energy consumption for households with two cold appliances.

Furthermore, in contrast to the combined fridge freezers shown in Figure 6.6, a systematic bias is visible as the energy consumption of the fridges and freezers increases. It appears that the inference is more likely to overestimate the energy consumption of the appliances when the energy consumption is low, and also to underestimate the energy consumption when the appliance's actual energy consumption is high. This is a result of too few signatures being extracted from the aggregate load, and as a result the general appliance prior has a relatively high influence in comparison to the observed data used to tune the model.

Figure 6.11 shows a histogram of the absolute error for each cold appliance. It is interesting to compare this histogram of the error for cold appliances in households with two cold appliances, to Figure 6.7 which shows the error for combined fridge freezers for households with a single cold appliance. Intuitively, the error is greater for households which contain multiple cold appliances, and therefore household energy disaggregation will be less accurate than those for households containing a single cold appliance.

Having analysed the accuracy by which the energy consumption of individual cold appliances can be estimated, we now evaluate the accuracy by which the total energy consumption of the cold appliances can be estimated in order to draw conclusions regarding the feedback that could be provided to the households' occupants. Figure 6.12 shows a bar graph of the actual and estimated total annual energy consumption of cold

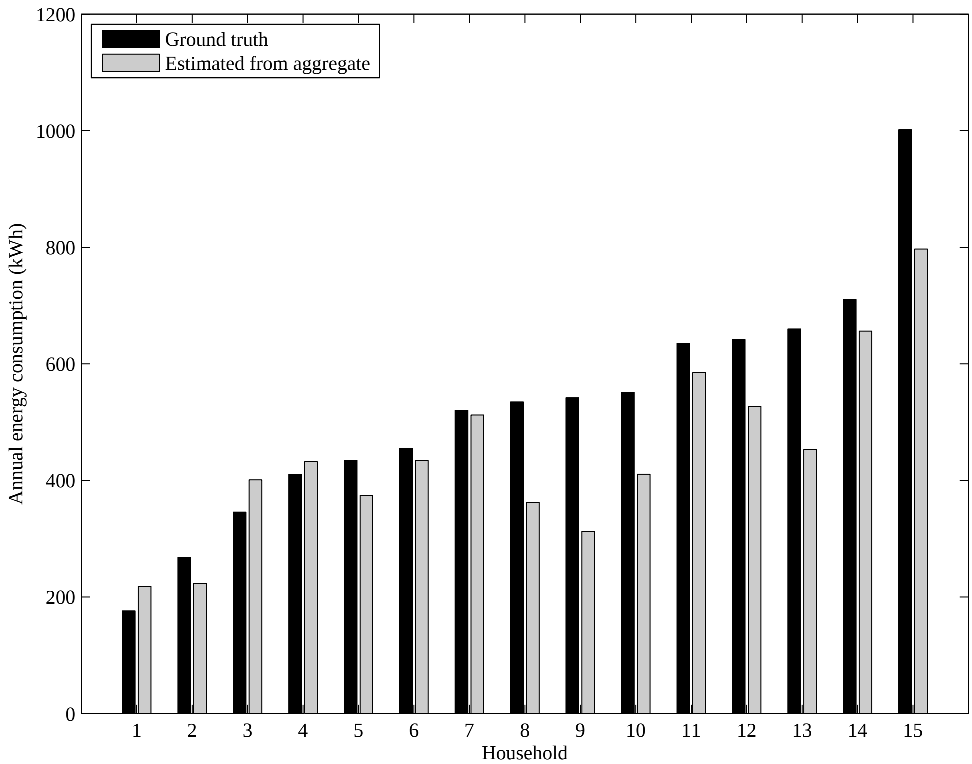


Figure 6.12: Ground truth and estimated total annual energy consumption of cold appliances for households with two cold appliances, ordered by ascending ground truth consumption.

appliances for each household containing two cold appliances. As in Figure 6.5, the greatest errors tend to occur in households which contain the highest energy consuming cold appliances. Furthermore, it is intuitive that the average error of total energy consumption estimation (mean = 95.1 kWh) is greater than for individual cold appliances, due to the greater amount of energy that is consumed. Therefore, there will be greater uncertainty regarding the energy savings of potential feedback.

Figure 6.13 shows a bar graph of the error of the total energy consumption of the cold appliances normalised by the ground truth energy consumption, where positive values represent overestimates of the energy consumption and negative values represent underestimates of the energy consumption. It is interesting to note that, although Figure 6.12 showed the absolute total error to be greater than for individual cold appliances, Figure 6.13 shows that the normalised total error is less than for individual cold appliances. This is a consequence of individual appliance errors cancelling out, since an overestimate for the first cold appliance and an underestimate for the second cold appliance in the same household will result in a smaller total error.

Figure 6.14 shows a histogram of the absolute total error for each household containing two cold appliances. It can be seen that 12 of the 15 households have an error of less than 200 kWh per year. Summing the 217 kWh annual energy consumption of a new

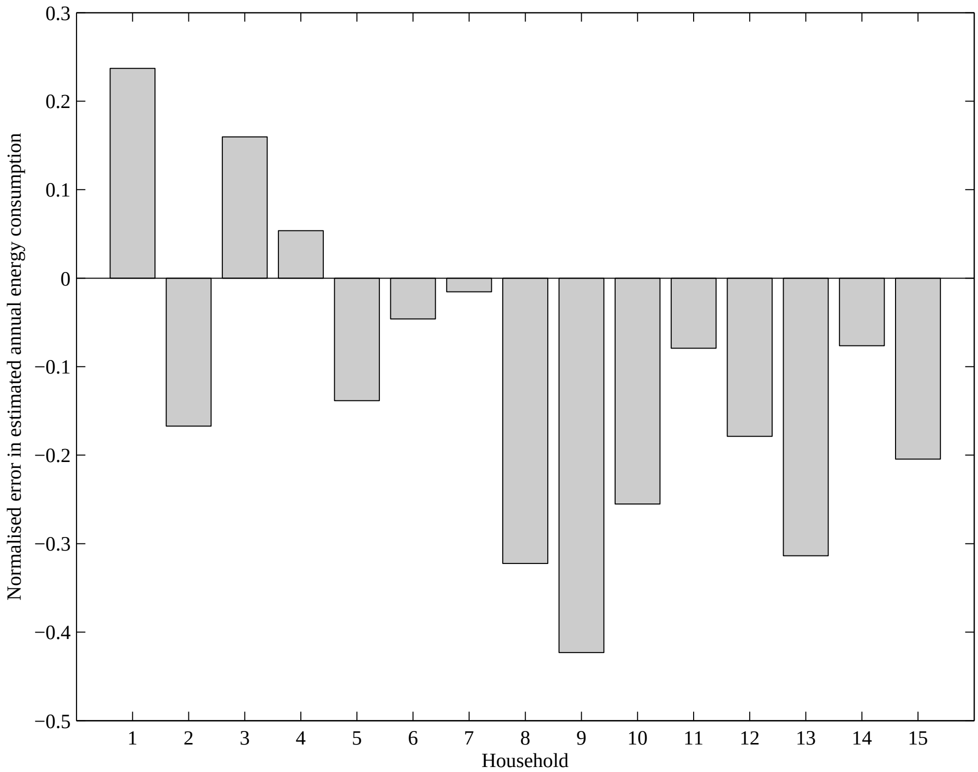


Figure 6.13: Normalised error in the estimate of total annual cold appliance energy consumption for households with two cold appliances, ordered by ascending ground truth consumption. Positive values represent overestimates, negative values represent underestimates.

combined fridge freezer[3] and the 200 kWh error produces a lower bound of 417 kWh per year, above which we can say with 80% confidence that the replacement of the household's cold appliances with a combined fridge freezer will result in energy savings.

Furthermore, the energy savings for the fridge freezers of more than 417 kWh annual consumption can be quantified as the difference between their estimated energy consumption and that of the replacement combined fridge freezer. The uncertainty around these energy savings can even be estimated using the mean absolute error of 95.1 kWh. Having evaluated the accuracy by which the energy efficiency of separate fridges and freezers can be inferred, the next section provides examples of the potential feedback that could be provided to a household's occupants.

6.4.5 Example of Potential Feedback

To illustrate the feedback that could be provided for an actual household, we consider house 108 as an example. House 108 was selected because it contains two cold appliances of median total energy consumption for which our system would recommend the

[3] `http://www.appliancecity.co.uk/liebherr/fridges-and-freezers/cup3221/product-16287/`

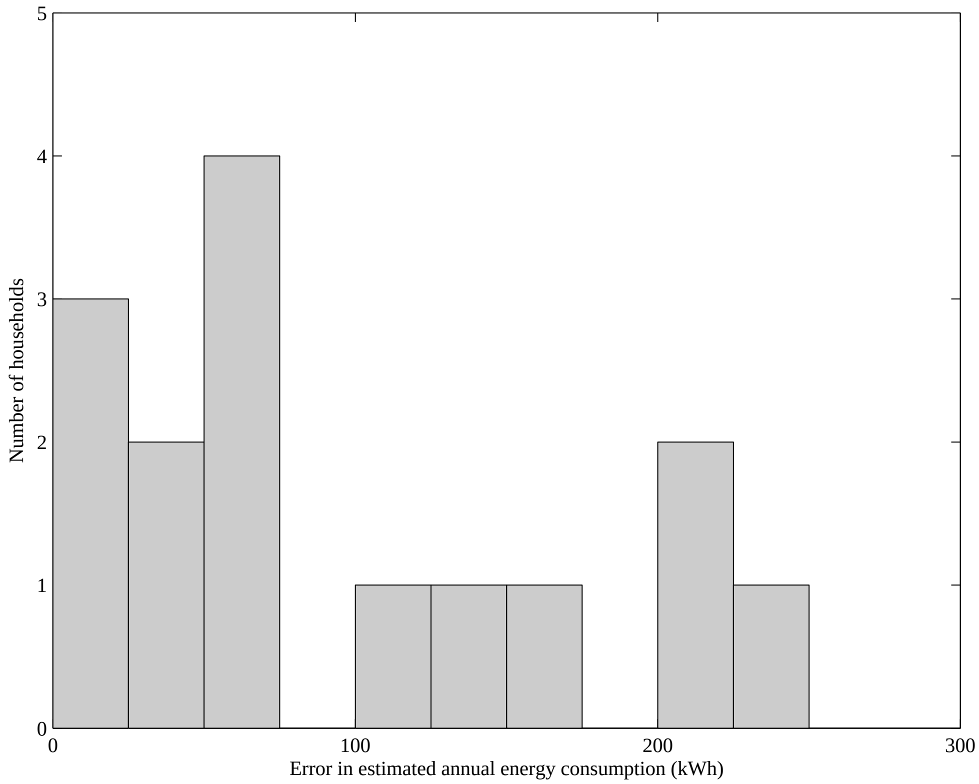


Figure 6.14: Absolute error in the estimate of total annual cold appliance energy consumption for households with two cold appliances.

appliance's replacement. Our approach estimates the total cold appliance annual energy consumption as 542 ± 95 kWh. The annual energy savings for replacing this appliance with a new combined fridge freezer can then be quantified as 325 ± 95 kWh, assuming a replacement which consumes 217 kWh per year.[4] Furthermore, these figures can be converted into financial savings, which for house 108 equates to £48.75 ± £14.25 per year, assuming a rate of £0.15/kWh. Finally, the time until the annual savings have offset the cost of the fridge freezer can be estimated as 8.8 years, assuming a replacement cost of £429.[4] These figures provide the household occupants with the required information in order to make an informed decision of whether to replace their cold appliances with a combined fridge freezer.

6.5 Generality to Appliances other than Cold Appliances

This chapter has focused on a case study application to cold appliances in order to examine the performance of the proposed approaches in a large number of households. The category of cold appliances was chosen because such appliances exhibit a cyclic pattern throughout the whole day, and as such the signature can be easily extracted during the overnight period. However, it is also worth considering whether the same approaches

[4] `http://www.appliancecity.co.uk/liebherr/fridges-and-freezers/cup3221/product-16287/`

can be applied to appliances other than cold appliances. In Chapter 5 we provided 4 examples of appliance types (one of which was a cold appliance) from a single household which benefit from model tuning using only aggregate data. This indicates that the approach presented in Chapter 5 does generalise to appliances other than cold appliances, although it does not test such generality over a large number of households, which still remains an open research question. However, the initial results from Chapter 5 are intuitive in that model tuning appears to be more accurate for appliances for which clean signatures can be extracted from the aggregate load. For example clean signatures of over 3 hours in duration can be extracted during the overnight period for cold appliances, and similarly clean signatures of 10 minutes in duration can be extracted for the microwave. In contrast, the 1 hour washing machine signatures and 2 hour dishwasher signatures contain patterns generated by other appliances (e.g. refrigerator cycles), and as a result the model tuning is less effective. As such, the performance of model tuning will be dependent upon the quality of appliance signatures that can be extracted from the aggregate load.

It is also worth considering whether the feedback methods used in this chapter generalise to appliances other than cold appliances. In this chapter, the feedback was generated based upon the energy efficiency of the appliance instance, and as such it was not necessary to disaggregate all usages of the appliance, and instead the extraction of the overnight periods was sufficient to generate the desired feedback. However, such a feedback mechanism will not be relevant to all appliance types. For instance, although the microwave model could be tuned accurately using the approach presented in Chapter 5, feedback based on the energy consumption of its individual usage would not be sufficient. This is because the context of the microwave's usage is unknown (what food is being cooked) and therefore the appliance's energy efficiency cannot be estimated. Instead, an alternative form of feedback would be more relevant, such as comparing the energy used to defrost frozen food using the microwave to letting the food defrost naturally, in the case that the defrost setting was detected.

6.6 Summary

In this chapter, we have presented a case study application of the approaches presented in Chapter 4 and Chapter 5 to the disaggregation of cold appliances. We have shown that by using these methods both specific appliance models and appliance energy efficiency information can be estimated using only smart meter data. Such specific appliance models could then be used by an existing disaggregation approach (e.g. Kolter and Jaakkola, 2012). In addition, these specific appliance models can also be used to provide actionable feedback regarding the energy efficiency of a specific appliance, as we have shown in this chapter.

Furthermore, we have shown that the methods presented in this work are robust enough to generalise to cold appliances in previously unseen households, and therefore the presented approach can automatically scale to large numbers of households. This addresses a core requirement of energy disaggregation research (Requirement 4) which has so far been neglected by academic research. The following chapter summarises the contributions of thesis and gives directions for future work.

Chapter 7

Conclusions, Limitations and Future Work

This thesis has described an approach to train non-intrusive load monitoring systems for use with household smart meter data. We now summarise the contributions of this work, highlight its limitations and give directions for future work.

7.1 Conclusions

We first defined the problem of non-intrusive load monitoring in Chapter 1. We identified four key requirements that must be fulfilled in order to realise a realistic solution to this problem. The requirements stated that the solution must be able to disaggregate low granularity smart meter data into individual appliances. However, most importantly the solution must not require training data to be collected from each household in which disaggregation will be performed. This requirement is crucial since it allows the approach to scale with the recent national deployments of smart meters.

We then provided a background of existing work in the field of non-intrusive load monitoring in Chapter 2. We showed that solutions which involve the installation of hardware in addition to existing smart meters are too expensive for large scale deployments. We then highlighted existing work based on temporal graphical models and their application to non-event based monitoring, and discussed their potential for robust energy disaggregation. However, each existing approach requires either sub-metered data to be collected from all appliances in each household, or requires a manual labelling process in which a domain expert is required to assign an appliance name to each identified appliance. As a result, they cannot be applied automatically to new households.

Next, we described 8 data sets released since 2011 which provide the potential to evaluate disaggregation systems in Chapter 3. We highlighted the following three data sets as particularly relevant. First, the REDD data set was identified as containing both sub-metered and household aggregate data from a range of different households, and can therefore be used to accurately evaluate the extent to which an appliance model represents the power data as measured from a real appliance. Second, the Tracebase data set was shown to contain data from a range of appliance instances of the same type, and can therefore be used to extract appliance information which generalises to new appliance instances of the same type. We concluded that all existing data sets only contain data collected from 10 or fewer households, and therefore do not represent the variety of households in which disaggregation methods will be required to perform. To address this shortcoming, we introduced a third data set, the Colden Common data set, which contains data collected from 117 households, and is therefore far more representative of the range of UK households than any existing data sets.

Chapter 4 represents the first major contribution of this thesis, in which we propose a method which is able to learn the characteristics of various appliances which generalise to new households. Our approach first learns HMM parameters for multiple instances of the same appliance type. Next, general distributions are fitted to samples drawn from each appliance instance model. As such, the learned general models represent distributions over the range of appliance instances that exist. Through our cross validation experiments using the Tracebase data set, we have shown that such general models outperform models learned from a single appliance instance, and furthermore that accurate general models can be learned from only 2–6 appliance instances.

Chapter 5 represents the second major contribution of this thesis, in which we show how general appliance models can be tuned to the specific appliance instances in a single household using only aggregate data. Our approach uses the general appliance models to extract periods of aggregate data during which only a single appliance is operating. These periods are subsequently used to refine the general appliance models. Through our experiments using the REDD data set, we have shown that such appliance models tuned using only aggregate data better represent the appliances in a household than the general appliance models. Furthermore, we have shown that the tuned appliance models can even perform comparably to when appliance models are learned using sub-metered appliance data from the test household. Finally, we showed that our proposed approach outperforms the state of the art which uses a FHMM to tune appliance models.

Chapter 6 represents the final major contribution of this thesis, in which we describe a large scale application of the approaches presented in Chapter 4 and Chapter 5 to 117 households in the Colden Common data set. We show that our approach is able to detect whether households contain a combined fridge freezer or a separate fridge and freezer, and furthermore we show that our approach is able to learn accurate appliance models in each situation using only a combination of the general appliance models and household aggregate data. Although these tuned appliance models could potentially be used for disaggregation, we show that they can also be used to infer the energy efficiency of the specific appliance instances. We show that the inferred energy efficiency is accurate enough to advise the majority of household occupants with 90% confidence that the replacement of their combined fridge freezer will result in energy savings.

When taken together, the contributions of this thesis represent an advancement to the state of the art in the domain of NIALM. We believe our combination of principled appliance modelling techniques along with a set of reasonable assumptions regarding the available data from each household has resulted in a robust solution which is suitable for deployment on a large scale. However, we recognise that our approach will not perform perfectly for all appliances and households, and such limitations are discussed in the following section.

7.2 Limitations

Although a number of scenarios in which the contributions of this thesis have been successfully applied have been highlighted in the previous section, it is also worth analysing the scenarios where such approaches will be less successful. This section highlights three such scenarios, which will be used to motivate the future work described in the following section.

The most important scenario in which the contributions of this thesis are likely to be limited is a household containing appliances which are poorly modelled by HMMs. Examples of such appliances include those with a continuously variable power demand (e.g. dimmer light) and those with states that are revisited many times in a particular order (e.g. a washing machine with 3 identical cycles). However, although these appliances might be poorly modelled using HMMs, they will only interrupt the modelling or disaggregation of other appliances during their use. Another type of appliance which will be poorly modelled captured by HMMs are those with a high variance between instances of the same type (e.g. collections of different numbers of light bulbs controlled using a single switch). As such, there might be considerable overlap between different appliance types, preventing the automatic labelling of an identified appliance instance.

Another potential limitation is that of households in which no clean appliance signatures can be easily extracted from the aggregate load. An example of such a household might be one in which multiple appliances of different types are continuously changing state throughout the day (e.g. multiple air conditioners or electric heating, multiple cold appliances or automated lighting systems). However, such appliances are likely controlled by an automated system, and therefore might exhibit a repeating structure that can be exploited in a similar way to the proposed method for dealing with households with separate refrigerators and freezers (Section 6.4). Crucially, more complex appliance modelling techniques can be applied to shorter data sequences (e.g. a few hours rather than days), with the aim of explaining the data using less appliance types.

A third potential limitation of the work presented in this thesis is that of appliances for which energy efficiency feedback would not be relevant. For example, the energy consumption of a kettle over a single use would not be particularly indicative of its efficiency, and is instead more closely related to the amount of water which had been boiled. As such, other feedback mechanisms could be used, such as those comparing the energy consumption of different appliances which accomplish the same task (e.g. boiling water using a kettle compared to an electric stove). We now discuss potential future work that could address these limitations.

7.3 Future Work

Future work will focus on a large-scale deployment of the technology presented in this work integrated with AgentSwitch; an agent-based platform designed to help household occupants manage their electricity consumption (Ramchurn et al., 2013). We aim to use the general models as constructed from the Tracebase data set, in combination with household aggregate data, to provide intuitive and actionable energy saving advice to household occupants. The accuracy of the inferred energy efficiency will be evaluated using limited individual appliance sub-meters, and the operating energy efficiency of appliances will be compared with that quoted by the appliance manufacturer. Furthermore, we will also use such appliance sub-meters to measure whether the energy saving advice has resulted in energy and financial savings.

In such a deployment, it might be necessary to construct more extensive general appliance models as larger sub-metered data sets become available. However, the use of longer power sequences will increase the time required to build such general models, and therefore more efficient inference algorithms would be required. One possibility would be to exploit the structure of HMMs through a structured variational inference algorithm. Such an approach could iterate between exact inference over HMM states using the Viterbi algorithm (Viterbi, 1967), and a variational approximation of the HMM parameters.

Furthermore, as gas and water smart meters begin to be deployed alongside electricity smart meters, an interesting research problem is emerging regarding the parallel disaggregation of all three utilities. Recent work has shown that methods similar to electricity disaggregation can be applied individually to gas disaggregation (Cohn et al., 2010) and water disaggregation (Dong et al., 2013). However, since some appliances consume two utilities (e.g. washing machine requires both electricity and water), information derived from one utility could also be used to disaggregate another utility. Therefore, new techniques are required which combine the disaggregation of multiple utilities while allowing mutual information to be shared between each utility.

In most cases, it is trivial to determine the number of states for an appliance type given some examples of power data. However, this might not be the case for more unusual appliances, and as a result an automated approach would be required to determine the number of states. We believe the infinite hidden Markov model (iHMM) (Beal et al., 2001) provides a natural representation of appliances in which the number of states is unbounded, and is free to grow as more data is observed. However, new methods will be required to generalise over these models, since an iHMM will likely learn a different set of states when applied to multiple instances of the same appliance type, and therefore the identification of corresponding states between multiple appliance models becomes a complex problem.

As discussed in this thesis, many extensions to HMMs exist, each of which provide a theoretical advantage over the basic HMM for certain appliances. For example, in Section 6.4.3 we stated that FHMMs were unable to disaggregate two cold appliance loads differentiated only by their periodicity, and also showed that FHSMMs were able to disaggregate such loads. However, each extension to the model increases the complexity of the inference process, and as a result more time will be required to reach good solutions to the inference problem. Therefore, it is an open question for future work whether the increase in disaggregation accuracy offsets the additional complexity of the inference process for such model extensions.

Another interesting challenge for extending this work would be to apply our proposed training methods to appliance models other than those based on HMMs. We have shown HMMs to be a good model for appliances with a discrete set of states (e.g. refrigerator or dishwasher), however HMMs are likely to fail for appliances with a continuously variable power demand (e.g. plasma television or dimmer light). In such cases, different graphical models, such as linear dynamical systems (Kalman, 1963), might represent such continuously variable appliances more appropriately. We believe that the approach proposed in this work, which constructs general appliance models and tunes such models using aggregate data, is general and therefore will be applicable to new graphical models.

Appendix A

Appliance Study

Table A shows approximate values for common appliances' power demands, usage per day and energy consumption per day. The approximate power demands were taken from MacKay (2008). Estimates of daily usage were then used to calculate the expected energy consumption of each appliance per day. The appliances were ordered for consistency with Table 1.1.

Appliance Name	Power demand (W)	Time per day (hours)	Energy per day (kWh)
Clothes drier	2500	0.8	2
Electric hob	3300	0.5	1.65
Dishwasher	2500	0.6	1.5
Electric oven	3000	0.5	1.5
Washing machine	2500	0.4	1
Kettle	3000	0.3	0.9
Incandescent light bulbs	60	8	0.48
Fridge	20	24	0.48
Microwave	1400	0.3	0.42
Wireless router	10	24	0.24
Set top box	10	24	0.24
Television	100	2	0.2
Games console	170	1	0.17
CFL light bulbs	20	8	0.16
Vacuum cleaner	1600	0.1	0.16
Toaster	1200	0.1	0.12
Computer	100	1	0.1
Phone charger	5	12	0.06
Alarm clock	2	24	0.048
LCD monitor	40	1	0.04
Laptop	25	1	0.025
Stereo	10	1	0.01
DVD player	10	1	0.01

Table A.1: Household appliance power demand, usage duration and energy consumption.

Appendix B

Prior Distributions of Appliance Model Parameters

This section provides the hyperparameter values used for the experiments in Chapter 4. We used uninformative uniform priors for both Dirichlet distributions over the initial multinomial distribution and transition matrix. We also used rough hyperparameters for the Gaussian-gamma distribution over the emission function, as stated in Table B.1.

Appliance	**Hyperparameter**	**State**		
		1	**2**	**3**
Kettle	$\hat{\lambda}$	0	2000	-
	$\hat{r}$	10^{-2}	10^{-4}	-
	$\hat{\beta}$	0.2285	4	-
	$\hat{w}$	0.0088	0.01	-
Refrigerator	$\hat{\lambda}$	0	100	-
	$\hat{r}$	10^{-2}	10^{-5}	-
	$\hat{\beta}$	0.2285	4	-
	$\hat{w}$	0.0088	0.01	-
Microwave	$\hat{\lambda}$	0	1350	-
	$\hat{r}$	10^{-2}	10^{-5}	-
	$\hat{\beta}$	0.2285	4	-
	$\hat{w}$	0.0088	0.01	-
Washing machine	$\hat{\lambda}$	0	150	1350
	$\hat{r}$	10^{-2}	10^{-3}	10^{-5}
	$\hat{\beta}$	0.2285	4	4
	$\hat{w}$	0.0088	0.01	0.01
Dishwasher	$\hat{\lambda}$	0	75	1350
	$\hat{r}$	10^{-2}	10^{-3}	10^{-5}
	$\hat{\beta}$	0.2285	4	4
	$\hat{w}$	0.0088	0.01	0.01

Table B.1: Hyperparameters of emission function.

Appendix C

Data Set Examples

This section provides example fragments from the publicly available data sets described in Chapter 3.

C.1 Reference Energy Disaggregation Data Set

Figure C.1 shows a fragment of data from the REDD data set. The fragment represents the first 10 readings of the first circuit channel (channel_3: oven) from house_1. The two columns represent a unix timestamp when the recording was taken and a decimal number corresponding to the instantaneous power demand of that circuit. It can be seen from the data fragment that power readings were recorded every 3 or 4 seconds, and also that the oven drew 0 W throughout the fragment as it was not in use.

```
303132933 0.00
1303132936 0.00
1303132940 0.00
1303132943 0.00
1303132946 0.00
1303132950 0.00
1303132953 0.00
1303132957 0.00
1303132960 0.00
1303132964 0.00
```

Figure C.1: REDD fragment

C.2 Building-Level Fully Labeled Electricity Disaggregation Data Set

Figure C.2 shows a fragment of data from the BLUED data set. The fragment represents 10 consecutive labels of appliance operation from the single household in the data set. The three columns represent a human readable timestamp, a code corresponding to the appliance event label, and the aggregate phase to which the appliance event is associated. Each row represents a single label of an appliance event, and as such the labels are associated with different appliances and occur at a random frequency. It should be noted that the data set also contains household aggregate data in a similar format to the REDD data set.

```
10/20/2011 14:40:49.357,111,A
10/20/2011 15:27:47.707,111,A
10/20/2011 15:30:39.240,127,A
10/20/2011 15:30:39.240,207,A
10/20/2011 15:30:58.907,127,A
10/20/2011 15:30:58.924,207,A
10/20/2011 15:42:34.723,111,A
10/20/2011 15:45:54.590,204,B
10/20/2011 15:47:31.223,204,B
10/20/2011 16:09:00.424,204,B
```

Figure C.2: BLUED fragment

C.3 UMASS Smart* Home Data Set

Figure C.3 shows a fragment of data from the Smart* data set. The fragment represents the first 10 readings from various appliances for house A of the data set. The columns represent the circuit name, circuit number, unix timestamp, real power and apparent power. Each row represents a single reading of a different circuit, and as such consecutive rows often contain the same timestamp. It should be noted that the data set also contains plug level data in a similar format.

C.4 Tracebase Repository

Figure C.4 shows a fragment of data from the Tracebase data set. The fragment represents 10 readings from coffee maker 735E9D. The columns represent a human readable timestamp, 1 second average power, and 8 second average power. Each row represents a single reading for this appliance, and the reading interval can be seen to be between 3 and 7 seconds.

```
Grid,1,1341047862,1004.972,1053.855
Dryer,2,1341047862,1.88,2.854
OfficeOutlets,3,1341047862,10.582,15.326
LivingRoomOutlets,4,1341047862,70.698,92.852
GuestBathOutlets,5,1341047862,0.594,1.553
MasterBathOutlets,6,1341047862,2.022,3.132
LivingRoomPatioLights,7,1341047862,4.562,9.187
GuestBathHallLights,8,1341047862,5.643,13.856
DiningRoomOutlets,9,1341047862,77.542,80.515
OutsideOutlets,10,1341047862,0.582,1.283
```

Figure C.3: Smart* fragment

```
15/05/2012 11:25:49;4;2
15/05/2012 11:25:56;4;2
15/05/2012 11:26:00;4;87
15/05/2012 11:26:03;689;175
15/05/2012 11:26:09;85;147
15/05/2012 11:26:13;94;81
15/05/2012 11:26:18;1275;691
15/05/2012 11:26:22;1139;1134
15/05/2012 11:26:26;113;951
15/05/2012 11:26:31;1184;1011
```

Figure C.4: Tracebase fragment

C.5 Individual Household Electric Power Consumption Data Set

Figure C.5 shows a fragment of data from the Individual Household Electric Power Consumption data set. The fragment represents the first 10 readings each circuit level meter. The columns represent a human readable date and timestamp, household aggregate active power, household aggregate reactive power, voltage, global intensity, and the power demand of three individual circuits. It should be noted that no further sub-metered data is available beyond these three circuits.

```
16/12/2006;17:24:00;4.216;0.418;234.840;18.400;0.000;1.000;17.000
16/12/2006;17:25:00;5.360;0.436;233.630;23.000;0.000;1.000;16.000
16/12/2006;17:26:00;5.374;0.498;233.290;23.000;0.000;2.000;17.000
16/12/2006;17:27:00;5.388;0.502;233.740;23.000;0.000;1.000;17.000
16/12/2006;17:28:00;3.666;0.528;235.680;15.800;0.000;1.000;17.000
16/12/2006;17:29:00;3.520;0.522;235.020;15.000;0.000;2.000;17.000
16/12/2006;17:30:00;3.702;0.520;235.090;15.800;0.000;1.000;17.000
16/12/2006;17:31:00;3.700;0.520;235.220;15.800;0.000;1.000;17.000
16/12/2006;17:32:00;3.668;0.510;233.990;15.800;0.000;1.000;17.000
16/12/2006;17:33:00;3.662;0.510;233.860;15.800;0.000;2.000;16.000
```

Figure C.5: IHEPCDS fragment

C.6 Household Electricity Use Study

Figure C.6 shows a fragment of data from the Household Electricity Use Study. The fragment represents the first 10 readings of a single appliance. The columns represent an interval ID (corresponding to a part of the day), a household ID, an appliance ID, a human readable date stamp, the energy consumed in tenths of a Watt-hour, and a human readable time stamp. The rows correspond to individual readings of the same appliance, and it can be seen that the readings are recorded at two minute intervals.

```
1,202116,0,2010-07-28,0,00:00:00
1,202116,0,2010-07-28,0,00:02:00
1,202116,0,2010-07-28,0,00:04:00
1,202116,0,2010-07-28,0,00:06:00
1,202116,0,2010-07-28,0,00:08:00
1,202116,0,2010-07-28,0,00:10:00
1,202116,0,2010-07-28,0,00:12:00
1,202116,0,2010-07-28,0,00:14:00
1,202116,0,2010-07-28,0,00:16:00
1,202116,0,2010-07-28,0,00:18:00
```

Figure C.6: HES fragment

C.7 Pecan Street Research Institute Sample Data Set

Figure C.7 shows a fragment of data from the Pecan Street Sample Data Set. The fragment represents the first 10 readings for each circuit in Home 01. The columns represent a human readable timestamp, the power demand of the household, the power supplied by the grid, the active power demand of air conditioner 1, the reactive power demand of air conditioner 1 and the active power demand of the dining room sockets. It should be noted that many more columns are available in the data set, and only a fragment of the data is reported here for illustrative purposes. The rows correspond to individual readings across all circuits, and it can be seen that these occur at 1 minute intervals.

```
03/09/2012 00:00  0.347283333 0.347283333 0.001983333 0.00785 0.001616667
03/09/2012 00:01  0.34665 0.34665 0.002 0.008116667 0.001616667
03/09/2012 00:02  1.064433333 1.064433333 0.659316667 0.854666667 0.001616667
03/09/2012 00:03  3.70055 3.70055 2.8687  3.0957  0.001666667
03/09/2012 00:04  3.7279  3.7279  2.895516667 3.122766667 0.00145
03/09/2012 00:05  3.725683333 3.725683333 2.894366667 3.1209  0.00165
03/09/2012 00:06  3.731516667 3.731516667 2.900566667 3.126116667 0.0016
03/09/2012 00:07  3.721 3.721 2.888816667 3.116316667 0.001666667
03/09/2012 00:08  3.722133333 3.722133333 2.891083333 3.12045 0.001683333
03/09/2012 00:09  3.715383333 3.715383333 2.883833333 3.114416667 0.0017
```

Figure C.7: Pecan Street fragment

C.8 The Almanac of Minutely Power Data Set

Figure C.8 shows a fragment of data from the AMPds Data Set. The fragment represents the first 10 readings for the North Bedroom in the single household in the data set. The columns represent a unix timestamp, voltage, current, frequency, displacement power factor, apparent power factor, real power, real energy, reactive power, reactive energy, apparent power and apparent energy. The rows correspond to individual readings of this circuit, and it can be seen that these occur regularly at 1 minute intervals.

```
1333263600,116.2,0.0,60.0,1.0,0.1,0,0,0,0,9,0
1333263660,116.2,0.0,60.0,1.0,0.05,0,0,0,0,9,0
1333263720,116.1,0.0,60.0,1.0,0.05,0,0,0,0,9,0
1333263780,116.9,0.0,60.0,1.0,0.1,0,0,0,0,9,1
1333263840,117.2,0.0,60.0,1.0,0.05,0,0,0,0,9,0
1333263900,116.6,0.0,60.0,1.0,0.05,0,0,0,0,9,0
1333263960,116.4,0.0,60.0,1.0,0.05,0,0,0,0,9,0
1333264020,116.4,0.0,60.0,1.0,0.05,0,0,0,0,9,0
1333264080,116.7,0.0,60.0,1.0,0.05,0,0,0,0,9,0
1333264140,116.2,0.0,60.0,1.0,0.05,0,0,0,0,9,1
```

Figure C.8: AMPds fragment

References

Kyle Anderson, Adrian Ocneanu, Diego Benitez, Derrick Carlson, Anthony Rowe, and Mario Berges. BLUED: A Fully Labeled Public Dataset for Event-Based Non-Intrusive Load Monitoring Research. In *Proceedings of the 2nd Workshop on Data Mining Applications in Sustainability (SustKDD)*, Beijing, China, 2012.

K Carrie Armel, Abhay Gupta, Gireesh Shrimali, and Adrian Albert. Is disaggregation the holy grail of energy efficiency? The case of electricity. *Energy Policy*, 52:213–234, 2013.

Sean Barker, Aditya Mishra, David Irwin, Emmanuel Cecchet, Prashant Shenoy, and Jeannie Albrecht. Smart *: An Open Data Set and Tools for Enabling Research in Sustainable Homes. In *Proceedings of the 2nd Workshop on Data Mining Applications in Sustainability (SustKDD)*, Beijing, China, 2012.

Nipun Batra, Jack Kelly, Oliver Parson, Haimonti Dutta, William Knottenbelt, Alex Rogers, Amarjeet Singh, and Mani Srivastava. NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring. In *Fifth International Conference on Future Energy Systems (ACM e-Energy)*, Cambridge, UK, 2014.

Matthew J Beal, Zoubin Ghahramani, and Carl E Rasmussen. The Infinite Hidden Markov Model. In *Proceedings of Neural Information Processing Systems*, Vancouver, BC, Canada, 2001.

M.J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, University College London, 2003.

Yoshua Bengio and Paolo Frasconi. An Input Output HMM Architecture. In *Proceedings of Neural Information Processing Systems*, pages 427–434, Denver, CO, USA, 1994.

Mario Berges, Ethan Goldman, H Scott Matthews, and Lucio Soibelman. Training Load Monitoring Algorithms on Highly Sub-Metered Home Electricity Consumption Data. *Tsinghua Science & Technology*, 13(Supplement 1):406–411, 2008.

Mario Berges, Ethan Goldman, H. Scott Matthews, Lucio Soibelman, and Kyle Anderson. User-Centered Non-Intrusive Electricity Load Monitoring for Residential Buildings. *Journal of Computing in Civil Engineering*, 25:471–480, 2011.

Mario E. Berges, Ethan Goldman, H. Scott Matthews, and Lucio Soibelman. Enhancing Electricity Audits in Residential Buildings with Nonintrusive Load Monitoring. *Journal of Industrial Ecology*, 14(5):844–858, 2010.

George Casella and Edward I George. Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174, 1992.

Hsueh-Hsien Chang, Ching-Lung Lin, and Jin-Kwei Lee. Load identification in nonintrusive load monitoring using steady-state and turn-on transient energy algorithms. In *International Conference on Computer Supported Cooperative Work in Design*, pages 27–32, Savannah, GA, USA, 2010.

M.-Y. Chen, A Kundu, and S N Srihari. Variable duration hidden Markov model and morphological segmentation for handwritten word recognition. *IEEE Transactions on Image Processing*, 4(12):1675–1688, 1995.

Gabe Cohn, Sidhant Gupta, Jon Froehlich, Eric Larson, and Shwetak Patel. GasSense: Appliance-Level, Single-Point Sensing of Gas Activity in the Home. In Patrik Floren, Antonio Krger, and Mirjana Spasojevic, editors, *Pervasive Computing*, volume 6030 of *Lecture Notes in Computer Science*, pages 265–282. Springer Berlin / Heidelberg, 2010.

Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

Department of Energy & Climate Change. Climate Change Act. Technical report, UK, 2008.

Department of Energy & Climate Change. A consultation on smart metering for electricity and gas. Technical report, UK, 2009.

Department of Energy & Climate Change. Energy Consumption in the United Kingdom. Technical report, UK, 2010.

Department of Energy & Climate Change. Smart Metering Equipment Technical Specifications. Technical report, UK, 2012a.

Department of Energy & Climate Change. Smart Metering Implementation Programme: First Annual Progress Report on the Roll-out of Smart Meters. 2012b.

Department of Energy & Climate Change. Smart Metering Equipment Technical Specifications Version 2. Technical report, UK, 2013a.

Department of Energy & Climate Change. Smart Metering Implementation Programme: Second Annual Report on the Roll-out of Smart Meters. 2013b.

Hail Dong, Bingsheng Wang, and Chang-Tien Lu. Deep Sparse Coding based Recursive Disaggregation Model for Water Conservation. In *International Joint Conference on Artificial Intelligence*, Beijing, China, 2013.

Karen Ehrhardt-Martinez, Kat A. Donnelly, and John A. Skip Laitner. Advanced Metering Initiatives and Residential Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities. Technical report, Research Report E105, American Council for an Energy-Efficient Economy, 2010.

Energy Saving Trust. Powering the Nation. Technical report, UK, 2012.

Rebecca Ford. *Reducing Domestic Energy Consumption Through Behaviour Modification.* PhD thesis, University of Oxford, 2009.

J Froehlich, E Larson, S Gupta, G Cohn, M Reynolds, and S Patel. Disaggregated End-Use Energy Sensing for the Smart Grid. *Pervasive Computing, IEEE*, 10(1):28–39, 2011.

Jon Froehlich. Promoting Energy Efficient Behaviors in the Home through Feedback : The Role of Human-Computer Interaction. Technical report, 2009.

Z. Ghahramani and M. I. Jordan. Factorial Hidden Markov Models. *Machine Learning*, 29:245–273, 1997.

Zoubin Ghahramani. An Introduction to Hidden Markov models and Bayesian Networks. In *Journal of Pattern Recognition and Artificial Intelligence*, volume 15, pages 9–42. 2001.

Hugo Gonçalves, Adrian Ocneanu, and Mario Bergés. Unsupervised disaggregation of appliances using aggregated consumption data. In *ACM Special Interest Group on Knowledge Discovery and Data Mining, workshop on Data Mining Applications in Sustainability*, San Diego, CA, 2011.

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*, pages 139–148, New York, NY, USA, 2010.

G. W. Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(12): 1870–1891, 1992.

G W Hart and A T Bouloutas. Correcting dependent errors in sequences generated by finite-state processes. *Information Theory, IEEE Transactions on*, 39(4):1249–1260, 1993.

International Energy Agency. Electricity/Heat in World in 2008, 2008.

Matthew J Johnson and Alan S Willsky. Bayesian Nonparametric Hidden Semi-Markov Models. *Journal of Machine Learning Research*, 14:673–701, 2013.

Rudolf Emil Kalman. Mathematical description of linear dynamical systems. *Journal of the Society for Industrial & Applied Mathematics, Series A: Control*, 1(2):152–192, 1963.

H. Kim, M. Marwah, M. F. Arlitt, G. Lyon, and J. Han. Unsupervised Disaggregation of Low Frequency Power Measurements. In *Proceedings of the 11th SIAM International Conference on Data Mining*, pages 747–758, Mesa, AZ, USA, 2011.

Younghun Kim, Thomas Schmid, Zainul M Charbiwala, and Mani B Srivastava. ViridiScope: design and implementation of a fine grained power monitoring system for homes. In *Proceedings of the 11th ACM International Conference on Ubiquitous Computing*, pages 245–254, New York, NY, USA, 2009.

J. Z. Kolter, S. Batra, and A. Y. Ng. Energy Disaggregation via Discriminative Sparse Coding. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems*, pages 1153–1161, Vancouver, BC, Canada, 2010.

J. Z. Kolter and T. Jaakkola. Approximate Inference in Additive Factorial HMMs with Application to Energy Disaggregation. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pages 1472–1482, La Palma, Canary Islands, 2012.

J. Z. Kolter and M. J. Johnson. REDD: A Public Data Set for Energy Disaggregation Research. In *ACM Special Interest Group on Knowledge Discovery and Data Mining, workshop on Data Mining Applications in Sustainability*, San Diego, CA, USA, 2011.

Po-hsiang Lai, Mark Trayer, Sudhir Ramakrishna, and Ying Li. Database Establishment for Machine Learning in NILM. In *1st International Workshop on Non-Intrusive Load Monitoring*, Pittsburgh, PA, USA, 2012.

Hoi Yee Lam. *Voltage-Current Trajectory: A 2-Dimensional Approach to Understand Electrical Load Signatures.* PhD thesis, The University of Hong Kong, 2007.

C Laughman, K Lee, R Cox, S Shaw, S Leeb, L Norford, and P Armstrong. Power signature analysis. *IEEE Power and Energy Magazine*, 1(2):56–63, 2003.

Te-Won Lee, M S Lewicki, M Girolami, and T J Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *Signal Processing Letters, IEEE*, 6(4):87–90, April 1999. ISSN 1070-9908.

Mikhail A. Lisovich and Stephen B. Wicker. Privacy Concerns in Upcoming Residential and Commercial Demand-Response Systems. *IEEE Proceedings on Power Systems*, 1 (1):1–10, 2008.

David J C MacKay. *Sustainable Energy - without the hot air.* UIT Cambridge, UK, 2008.

Stephen Makonin, Fred Popowich, Lyn Bartram, Bob Gill, and Ivan V Bajic. AMPds: A Public Dataset for Load Disaggregation and Eco-Feedback Research. In *IEEE Electrical Power and Energy Conference*, pages 1–6, Halifax, NS, Canada, 2013.

A Marchiori, D Hakkarinen, Qi Han, and L Earle. Circuit-Level Load Monitoring for Household Energy Management. *IEEE Pervasive Computing*, 10(1):40–48, 2011.

Richard McWilliam and Alan Purvis. Potential Uses of Embedded RFID for Appliance Identification. *International Appliance Manufacturing*, 10:100–107, 2006.

T Minka, J M Winn, J P Guiver, and D A Knowles. Infer.NET 2.5. Technical report, Microsoft Research Cambridge, 2012.

Thomas P Minka. Beyond Newton's method. Technical report, 2002a.

Thomas P Minka. Estimating a Dirichlet distribution. Technical report, 2002b.

Kevin P Murphy. The Bayes Net Toolbox for MATLAB, 2001.

Kevin P Murphy. Hidden semi-Markov models (HSMMs). Technical report, MIT, 2002.

Kevin P Murphy. Conjugate Bayesian analysis of the Gaussian distribution. Technical report, 2007.

Oliver Parson, Siddhartha Ghosh, Mark Weal, and Alex Rogers. Using Hidden Markov Models for Iterative Non-intrusive Appliance Monitoring. In *Neural Information Processing Systems, Workshop on Machine Learning for Sustainability*, Sierra Nevada, Spain, 2011.

Oliver Parson, Siddhartha Ghosh, Mark Weal, and Alex Rogers. Non-intrusive load monitoring using prior models of general appliance types. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 356–362, Toronto, ON, Canada, 2012.

Oliver Parson, Siddhartha Ghosh, Mark Weal, and Alex Rogers. An Unsupervised Training Method for Non-intrusive Appliance Load Monitoring. *Artificial Intelligence (under review)*, 2013.

S N Patel, T Robertson, J A Kientz, M S Reynolds, and G D Abowd. At the flick of a switch: Detecting and classifying unique electrical events on the residential power line. In *Proceedings of the 9th International Conference on Ubiquitous Computing*, pages 271–288, Innsbruck, Austria, 2007.

L R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Sarvapali Ramchurn, Michael Osborne, Oliver Parson, Talal Rahwan, Sasan Maleki, Steve Reece, Trung Dong Huynh, Muddasser Alam, Joel Fischer, Tom Rodden, Luc Moreau, and Sephen Roberts. AgentSwitch: Towards smart electricity tariff selection. In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems*, Saint Paul, MN, USA, 2013.

P Ramesh and J G Wilpon. Modeling state durations in hidden Markov models for automatic speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 381 –384, San Francisco, CA, USA, 1992.

Andreas Reinhardt, Paul Bauman, Daniel Burgstahler, Matthias Hollick, Hristo Chonov, Marc Werner, and Ralf Steinmetz. On the Accuracy of Appliance Identification Based on Distributed Load Metering Data. In *Proceedings of the 2nd IFIP Conference on Sustainable Internet and ICT for Sustainability*, pages 1–9, Pisa, Italy, 2012.

Matthieu Sanquer, Florent Chatelain, Mabrouka El-guedri, and Nadine Martin. Hierarchical Bayesian Learning for Electrical Transient Classification. In *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, BC, Canada, 2013.

A Schoofs, A Guerrieri, D T Delaney, G O'Hare, and A G Ruzzelli. ANNOT: Automated Electricity Data Annotation Using Wireless Sensor Networks. In *Proceedings of the 7th Annual IEEE Communications Society Conference on Sensor Mesh and Ad Hoc Communications and Networks*, Boston, MA, USA, 2010.

S R Shaw, S B Leeb, L K Norford, and R W Cox. Nonintrusive Load Monitoring and Diagnostics in Power Systems. *Instrumentation and Measurement, IEEE Transactions on*, 57(7):1445–1454, 2008.

F Sultanem. Using appliance signatures for monitoring residential loads at meter panel level. *IEEE Transactions on Power Delivery*, 6(4):1380–1385, 1991.

K H Tiedemann. Using conditional demand analysis to estimate residential energy use and energy savings. *European Council for an Energy Efficient Economy Summer Study*, Panel 6:1279–1283, 2007.

Tsuyoshi Ueno, Ryo Inada, Osamu Saeki, and Kiichiro Tsuji. Effectiveness of an energy-consumption information system for residential buildings. *Applied Energy*, 83(8):868–883, 2006.

Tuomas Virtanen. Speech Recognition Using Factorial Hidden Markov Models for Separation in the Feature Space. In *9th International Conference on Spoken Language Processing*, Pittsburgh, PA, USA, 2006.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

John Winn and Christopher M Bishop. Variational message passing. *Journal of Machine Learning Research*, 6(1):661–694, 2006.

Yu Yi-xin, Li Peng, and Zhao Chun-liu. Non-intrusive method for on-line power load decomposition. In *Proceedings of the China International Conference on Electricity Distribution*, Guangzhou, China, 2008.

S Yu. Hidden semi-Markov models. *Artificial Intelligence*, 174(2):215–243, 2010.

A A Zaidi, F Kupzog, T Zia, and P Palensky. Load recognition for automated demand response in microgrids. In *Proceedings of 36th Annual IEEE Conference on Industrial Electronics Society*, pages 2442–2447, Glendale, AZ, USA, 2010.

M Zeifman and K Roth. Nonintrusive appliance load monitoring: Review and outlook. *IEEE Transactions on Consumer Electronics*, 57(1):76–84, 2011.

Jean-Paul Zimmermann, Matt Evans, Jonathan Griggs, Nicola King, Les Harding, Penelope Roberts, and Chris Evans. Household Electricity Survey A study of domestic electrical product usage. Technical report, Intertek, 2012.