UNIVERSITY OF AMSTERDAM

Faculty of Science
Master Artificial Intelligence
Track Intelligent Systems

Behavior Analysis of Elderly using Topic Models

Master Thesis of

Kristin Rieping

Supervisors:
Prof. Dr. ir. B.J.A. Kröse
Dr. G. Englebienne

Amsterdam, August 2013

Abstract

In this thesis behavior patterns of people are analyzed with Topic models. Two novel variations of the Latent Dirichlet Allocation (LDA) model are presented. The models give the opportunity to detect patterns in low-dimensional sensor data in an unsupervised manner. LDA-Gaussian, the first variation of the model, is a combination of a Gaussian Mixture Model and the LDA model. Here the multinomial distribution of the topics, that is normally used in the LDA model, is replaced by a set of Gaussian Distributions. In this way similar looking sensor data is automatically grouped together and captured in the same topic. LDA-Poisson, the second variation of the model, takes a set of Poisson Distribution for the topic descriptions. This model is more suited to handle counts of stochastic events. The parameters of both models are determined with an EM-algorithm. Both models are applied to real sensor data, which is gathered in the homes of elderly people. It is shown that meaningful topics can be found and that a semantic description of these topics can be given.

Contents

List of Figures

Chapter 1

Introduction

The life expectancy of people is assumed to rise continuously in the following years [13]. As a consequence the percentage of elderly increases. Elderly people often need more health care but studies show that they prefer to live at home [4]. The manpower to take care of elderly people is not always available. That is why monitoring the health condition of people in their home environment becomes more and more important. In this way slowly emerging declines in the health condition of people can be detected and appropriate health care can be granted if it is necessary before critical health conditions are reached.

New systems give the possibility to monitor elderly from the distance or even automatically [15]. Different systems give the possibility to detect accidents or monitor the health condition on a long time basis. Some systems still need the action of a person involving into the system. For example the inhabitant has to push a button if an accident occurs [10]. Other systems use cameras to monitor the elderly [11, 12]. But these methods are privacy-sensitive and often not adopted by the elderly.

Less intrusive methods use simple sensors like motion sensors or pressure mats that are placed in the homes of people [16, 7]. Extracting valuable information out of this data is however difficult and that is why activity recognition is often done to make the data more easily to interpret. Different techniques attempt to extract 'Activities of daily living' (ADL's). Changes in the ADL's can be a sign for declines of peoples health [16]. Some researchers depend on annotated data to find ADL's in sensor data [16, 8, 18]. But the task of labeling the data is time consuming and also intrusive, if the labeling is done by a different person. The knowledge of being observed by a sensor, for example a camera, might change the behavior patterns of a person and in this way the data is not accurate. Annotation that is done by the subject himself also might not by accurate, because the person has to interrupt his daily behavior to write down the annotation of the activities.

A way to automatically find behavior pattern in sensor data is by applying a topic model to the data. Topic models are initially designed for classifying text documents and they are able to find abstract topics, such as 'politics', 'sports', 'finances' etc. But a number of researchers have applied the topic model 'Latent Dirichlet Allocation' (LDA) to different kind of sensor data [6, 5]. The found topics are comparable to ADL's and can give a good representation of peoples daily behavior. There is however a mayor difference between textual and sensor data. In the LDA model presented by Blei [1] the documents are represented by a Bag-of-words (BOW) model. A BOW model is an unordered representation of words, that does not take the grammar or the order of the words into account. To make use of the LDA model on sensor data one has to create artificial words. But sensor data is mostly time dependent and that is why in most approaches a time value is added to the artificial words. There are numerous ways ways to create words from sensor data. Some researchers create artificial words by simply adding a time-value to the the sensor data and directly apply the BOW model to these 'words' [6, 3]. This approach requires a large dictionary of words to find behavior patterns in the data, to be able to capture all the variations in artificial words. For this reason other researchers first cluster the data and then apply the LDA model [9, 2]. In this way the size of the dictionary is smaller and less data is required. But finding the correct clusters is however difficult.

In this thesis the two new topic models 'LDA-Gaussian' and 'LDA-Poisson' are developed. These models are able to capture similar observations/words into the same topic automatically. The clustering and the LDA model are combined into one model and except for the construction of the artificial words, no further pre-processing is necessary . In the original LDA model (Blei) the topics are described by a multinomial distribution over the words of the vocabulary. In the 'LDA-Gaussian' model this distribution is replaced with a Gaussian distribution, so that similar words are caught in the same topic. 'LDA-Poisson' is a variation on this model, where the underlying distribution is Poisson. In this way event based sensor data can be modeled. The parameters of both models are found with an EM-procedure, which uses the likelihood of the model to converge to the optimal model parameters. The models are applied on real sensor data, that is obtained from the houses of solitary living elderly. All sensors are binary and are experienced as non-intrusive by the inhabitants. These elderly people live in a care home and get health care on a regular basis. Still they are able to live on their own.

In the next chapter of this thesis an overview of related approaches are given. In chapter 3 the data that is used is described in more detail and after that representation of the features is given in 4. This chapter is followed by chapter 5 which introduces the 'LDA-Gaussian' and 'LDA-Poisson' models. Chapter 6 contains the different experiments that are performed on the available data. In chapter 7 the conclusions of this thesis is presented and some suggestions for future work are given.

Chapter 2

Related Work

The goal of monitoring the health of people is to detect accidents or even more important to prevent accidents and critical health conditions. Changes in the daily behavior patterns can be a sign of changes in the health of people. This can be both mental or physical declines. There are different ways to monitor the health condition of people. Cameras or microphones can be very useful to monitor peoples behavior [12, 19], but these sensors are invading the privacy of people and often not accepted as sensors in peoples homes.

Simple binary sensors such us motion sensors, contact switches or pressure mats are preferable for health monitoring in home environments. These sensors are low in cost and easy to install. Moreover, they are also experienced as non-intrusive and not disturbing by the inhabitants. Numerous researchers implemented different approaches to apply activity recognition on data generated by these kind of sensors. These activities and especially changes in these activities, often referred to as ADL's, can then give valuable information on peoples health [14]. Tapia et al. [16] uses a naive Bayes classifier to find activities in annotated, sensor data. They show that it is possible to find activities in ubiquitous, simple sensor data, that was obtained in real-life environments. In the work of Kasteren et al. [17] two approaches for recognizing activities in sensor data are compared. The Hidden Markov Model and the Conditional Random Field are both applied to annotated, real-life sensor data. They also vary between different kind of sensor readings and show that this can improve the results for recognizing activities with their approaches. Wilson et al. [18] implemented a Particle Filter to find activities in simulated as well as real-life data. They are able to distinguish the actions between multiple people in the environment. Hong et al. [8] uses ontologies to describe daily activities. They use an evidential network to describe activities in a hierarchical way.

All of the previous approaches used annotated data. Generating this labeled data is however difficult, time consuming and the determined labels are not always accurate. For this reason supervised methods are preferred above supervised methods in this field. Various authors applied LDA to different kind of data. This topic model is able to find abstract descriptions of activities in data automatically.

Chikhaoui et al. [5] uses the topic model LDA in combination with sequential pattern mining to find activities in various datasets. The sequential pattern are used as words, which are the input for the LDA model. They test their method on varied annotated data sets. The topics that are found describe activities and the accuracy is measured by comparing the topics with annotation labels. In their work the focus lays on detecting activities and not so much on the global idea of detecting behavior patterns as it is done in this work. The sequential patterns might be interesting to investigate for further research as a variation on the feature representation that is given in chapter 4.

Huyhn et al. [9] and Casale et al. [2] both try to discover daily routines from sensor data.

Acceleration sensors that are attached to the human body generate a continuous stream of data. The data is sampled with different time intervals and in this ways artificial words are created. The dictionary, which contains all unique words, is quite big, because small variations in the sensor data generates different words. Therefore the authors cluster the data on forehand with the k-means algorithm. In this way the size of the dictionary can be reduced. The choice of k however is of big influence on the outcome of the LDA model. That is why in this thesis a different approach is used to handle the big amount of variations in the artificial words.

Farrahi et al. [6] applies LDA to location data gained from cell information of mobile phones. A lot of data is available and artificial words are created by combining the location of a person at three consecutive time steps and adding the time value. In this way every time of a day is described with a word. This approach of dividing a day into words is also adopted in this thesis. In Farrahi's work only 512 different words are possible, but about 2800 days of 68 different people are available. This ratio of vocabulary size and available data makes their approach a successful way to find latent topics in location data, without the need of clustering the data on forehand. Their way of describing the features is a good option to capture transitions in locations and could also be applied too this work.

In the work of Castanedo et al. [3] they also apply the LDA model on sensor data without pre-clustering the dictionary. For their work much data was available, which was obtained in an office environment. The words are however represented differently than it is done in the work of Farrahi et al. Not every time of a day is divided into words, but only time periods that contain sensor activations build the artificial words. They indicate that it can become difficult to give a good interpretation of the topics, that are found. So it is questionable if leaving out time periods without sensor activations in the feature representation leads to a better result.

Chapter 3

Data

In this chapter a description of the houses and the inhabitants is given. The different kind of sensors are explained and an impression of the received data is given.

3.1 Homes and persons

In the homes of five different people sensors are installed. The floor-plan of these homes is for all residents the same and is shown in figure 3.1. There might be small differences of the locations of the sensors due to the personal arrangements of peoples personal belongings. The persons that live in the homes are people that need healthcare on a regular basis, they are further able to live on their own. The amount of data collected differs for the different houses, but there is at least 63 days of data available for every house. An overview of the different houses is given in table 3.1

Table 3.1

HouseNr	1	2	3	4	5
# of days	142	98	89	63	73
mean number of events per day	488	523	668	565	427
age	84	77	79	93	95
gender	female	female	female	male	female

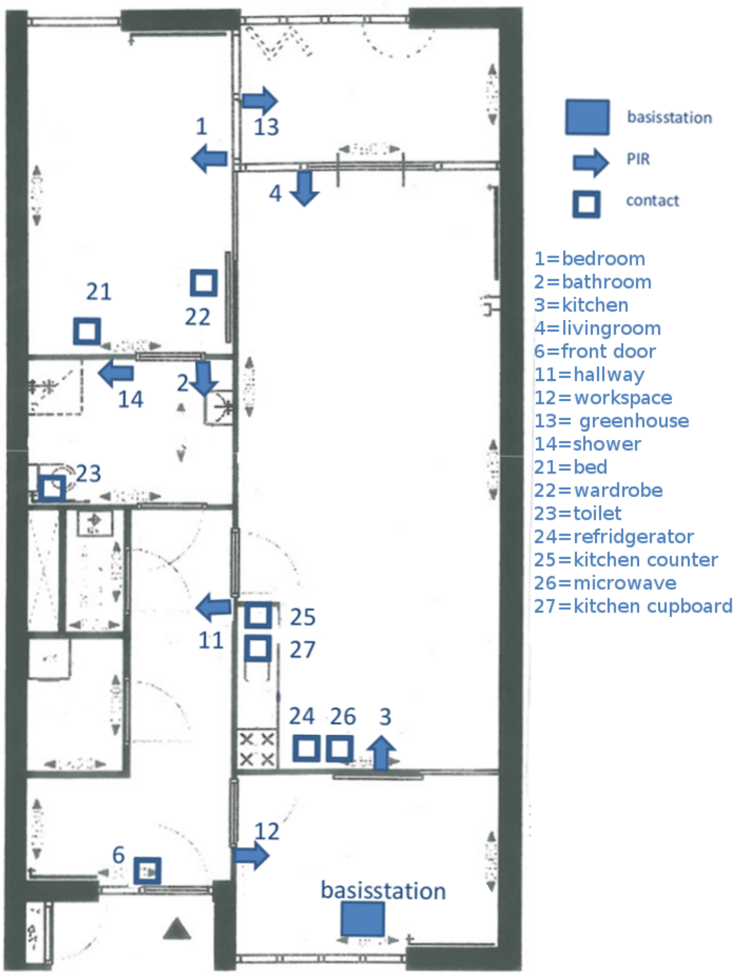


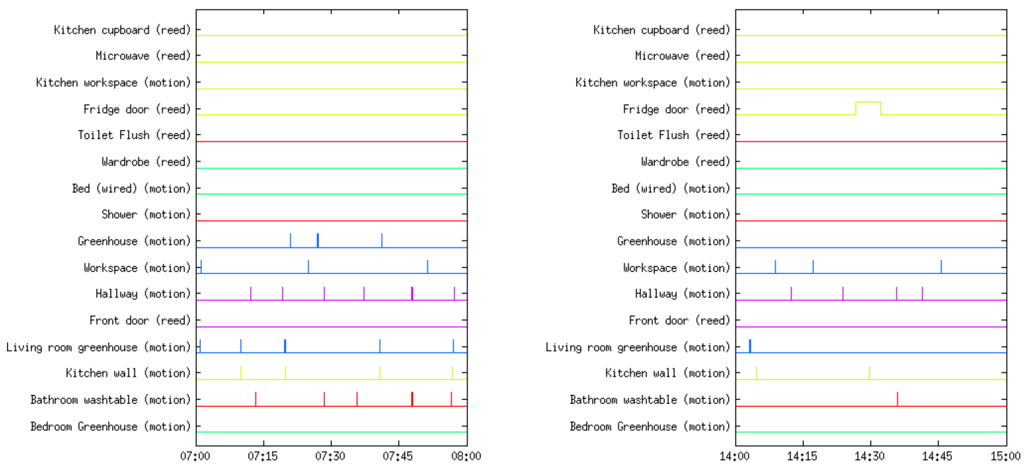
Figure 3.1: Floorplan of the houses with sensor descriptions

3.2 Sensors

There are different types of sensors installed in the homes. The contact switches are mostly installed at doors and cupboards. They get the value 'one' if a door is opened and the value 'zero' if the door is closed again. The motion-sensor (PIR) are placed at different places in the homes, mostly against the walls. They have a range of 5 meters. If a motion occurs in the region the sensor sends an impulse value, which means that the value becomes 'one' and immediately 'zero' again . After that the sensor is set to mute for about 3 minutes, which means that in this time there is no motion captured. In this way constantly firing of the sensor will be avoided. The sensor system is active 24 hours and 7 days a week. However failure can occur due to network problems, sensor failing or other unexpected problems.

3.3 Received Data

In figure 3.2 the data stream of two different hours of one day is shown. The data belongs to one person. Several sensors that are located in the same room are manually grouped together in a 'field'. In the following we use the 'fields' {'kitchen', 'living room', 'bathroom', 'bedroom', 'hallway'}. They are marked in the figure with different colors.



a) b)

Figure 3.2: Sensor Data for two different hours at a day of one person. The fields 'kitchen','living room','bathroom','bedroom','hallway' are marked with the colors 'yellow', 'blue', 'red', 'green', 'purple' respectively.

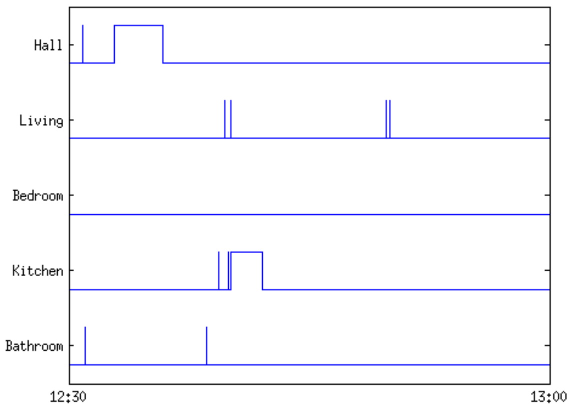
In the figure one can see the different types of data that are generated by the different sensor types. The fridge sensor is a reed sensor which has the value '1' for a longer period of time, when the door is opened for a while. The motion sensors on the other hand only give a impulse value as mentioned before. Some sensors are not triggered at all in the time intervals that are shown. In the next chapter the feature representation will be described.

Chapter 4

Features

The data that is obtained from the sensors generates a continuous data stream for every sensor. A good feature representation of the data is required so that the LDA model can be applied. First the sensors are divided into five fields and all the data that is generated from one field is combined into one data stream. In this way the data is reduced to five dimensions. The continuous data stream cannot be used as input for the LDA model. Therefore, like it is done in the work of Farrahi [6], the data of one day is divided into time-slices of length l. For a chosen length of the time-slices l in minutes, the number of time-slices for one day can then be calculated with $n = 1440/l$ (there are 1440 minutes per day) So for example if the length of the time-slices is $l = 30$ min. then the number of time-slices on one day is $n = 48$. A day starts and ends at 3 a.m. in the morning. In this way the chance to cut between activities is reduced. It still can occur that a person goes to bed late or that he needs to visit the toilet. For now this fact is left out in the part of modeling.

For every field the number of activations is counted in each time-slice. An sensor activation is defined as the change of the signal from zero to one. The duration of an active signal is not taken into account. In this way a door that accidentally is left open will not generate a high value, which will otherwise disturb the data. Every field then forms a dimension of the observations o_n. An observation can be seen as an artificial word. In figure 4.1 an example on how the data is translated into a vector representation is given for one time-slice.



$$\vec{o}_n = \begin{bmatrix} o_{Hall} \\ o_{Living} \\ o_{Bedroom} \\ o_{Kitchen} \\ o_{Bathroom} \\ o_{time} \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 0 \\ 3 \\ 2 \\ t \end{bmatrix}$$

Figure 4.1: Vector representation of the data. The data of the sensors is shown in the left image. It is translated in the vector shown on the right-hand side.

The last dimension of the observation $\vec{o}_n$ represents the time value of a time-slice. There are

Table 4.1

HouseNr	1	2	3	4	5
# of days	142	98	89	63	73
observations	6816	4704	4272	3024	3504
unique observations	3017	2110	2288	1269	1490

two different ways how the time dimension is added to the observations. In the fine-grain representation the time value becomes the number of the time-slice. In figure 4.1 the 30 minutes time interval $[12:30, 13:00]$ is the 20th time-slice on a day that starts at 3 a.m. In this way the observation will become $\vec{o}_n = \{2, 4, 0, 3, 2, 20\}$. In the coarse-grain representation, which is also used in the work Farrahi et al. [6] and Castanedo et al. [3], the 24 hours of a day are divided into the five time intervals $\{3am - 8am, 8am - 1pm, 1pm - 18pm, 18pm - 23pm, 23pm - 3am\}$. In this way the observation of figure 4.1 will become $o_n = \{2, 4, 0, 3, 2, 2\}$.

The data sets that are used in this thesis contain a minimum of 68 and a maximum of 142 days of data. With the chosen feature representation there are a lot of possible observations, which are not contained in the data sets. In the data the maximum value of 28 is observed in one field, when the number of time-slices per day is $n = 48$. This is an extreme value and occurs not that often in the data. If the maximum value for each field is assumed to be 15 and the time-dimension is fine grain, there are approximately 36 million ($15^5 * 48$) possible observations that can be made. As a comparison: In the work of Farrahi [6] only 512 ($4^3 * 8$) different observations are possible and 2856 days of data for 68 people is available.
In table 4.1 an overview is given how many unique words are actual observed for the different houses and how many words are totally observed ($48 * \#$ of days). One can see that the number of unique observations scales with the number of days available for each house. This also shows that in the data a lot of possible observations are not included. In figure 4.2 the occurrence for each unique word is shown on the y-axis for the data of house number one. The words that occur the most have zero values for the first five dimensions and different values for the sixth dimension (time). Most other unique words only occur once. Therefore applying the LDA model to this simple BOW representation is not feasible.

In figure 4.3 for each dimension of the observations a histogram is created of the activation values in the set of unique observations. For the first five dimension, which corresponds to the sensor values, small values are more probable to appear in the vocabulary. The found distribution is comparable to a Poisson-distribution. The sixth dimension, which corresponds to the time, shows that there are more unique observations that have a time value between 15 and 43. This makes sense because there is more activity during daytime which leads to more variate observations in this time period.

The given feature representation, with the two variation of the time dimensions, fine grain and coarse grain, are used in the following chapters. There are much more possibilities to describe the features, which is further explained in the future work (see chapter 7).

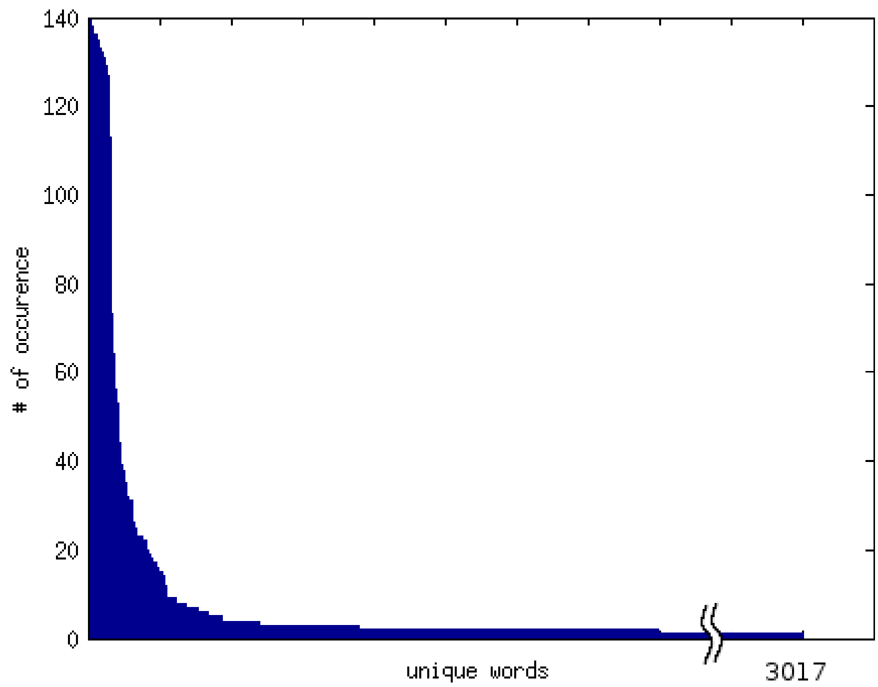


Figure 4.2: Occurrence of unique observation for house number one.

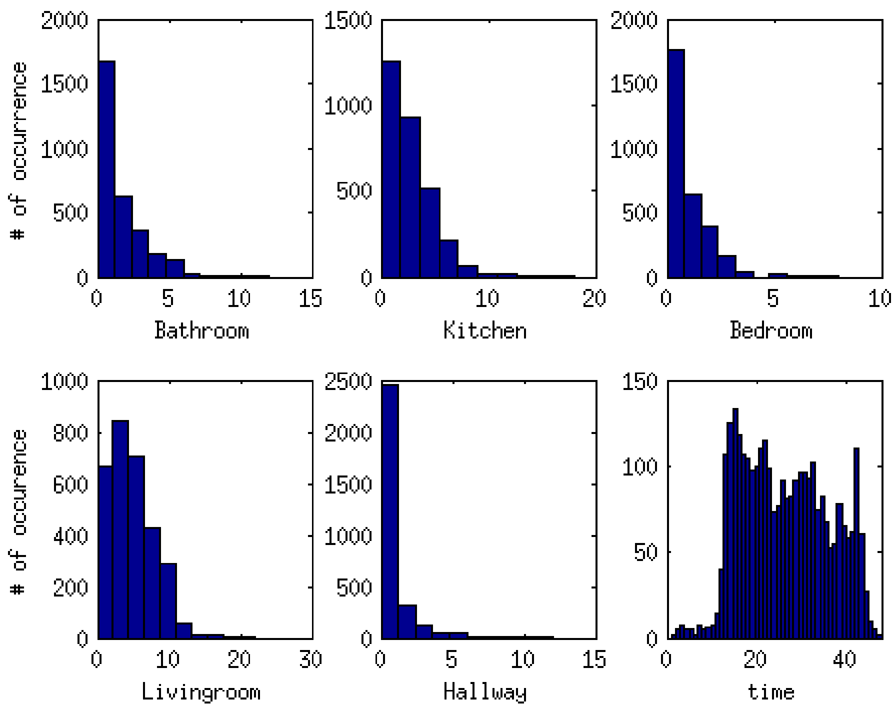


Figure 4.3: Occurrence of observation per field/dimension in the set of unique words for house number one.

Chapter 5

Topic Models

In the first section of this chapter the general idea of topic models is introduced. This section is followed by a description of how this kind of models can be applied the sensor data. After that the extension of the LDA model 'LDA-Gaussian' is given, including the EM-algorithm that is used to determine the model parameters. This chapter is finalized with the description of the LDA-Poisson model.

5.1 Introduction to Topic Models

Topic models are often used in the field of document classification. Given a set of documents (corpus) it is assumed that every document is a mixture of topics. For example a news article may belongs for some percentage, let us say 30 %, to the topic 'Economy' and for 70 % to the topic 'Politics'. Another document of the same Corpus may belong to the topic 'Economy' with 50 %, 'Politics' with 30 % and 'Finance' with 20 %. There might be a lot of different topics and the topics can have different level of details.
The topics are defined by several words that can occur in the documents. The topic 'Economy' may be defined by the list of words {'trade', 'industry', 'GDP'}. Other topics have different lists of words that describe them. The list might be longer or shorter and the words in the list will depend on the corpus that is used to generate the topics. It might also be the case that one word belongs to multiple topics. Eventually we can find the topic distribution of a document according to the words that are included in this document.
In the topic model 'Latent Dirichlet Allocation' (LDA) it is assumed that a Corpus can be made out of a generative process. The parameters that generate the corpus are than used to describe the model of the corpus. The generative process which builds a corpus is as follows:
For every document that is generated in the Corpus

1. Choose the number of words in the document from $N \sim Poisson(\xi)$.
2. Choose a topic distribution $\theta \sim Dir(\alpha)$ for the document.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$.
   (b) Choose a set of words w_n from the set of all words V from $p(o_n|z_n;\beta)$, a multinomial probability conditioned on the topic z_n. Where β is the distribution over words given a topic.

The model is also shown in figure 5.1. The parameters α and β define a corpus.

To determine the model parameters an EM-algorithm can be used. How this algorithm can be applied is extensively described in [1].
In the next section it is described how this model can be used with the sensor data, that is gained in the different houses.

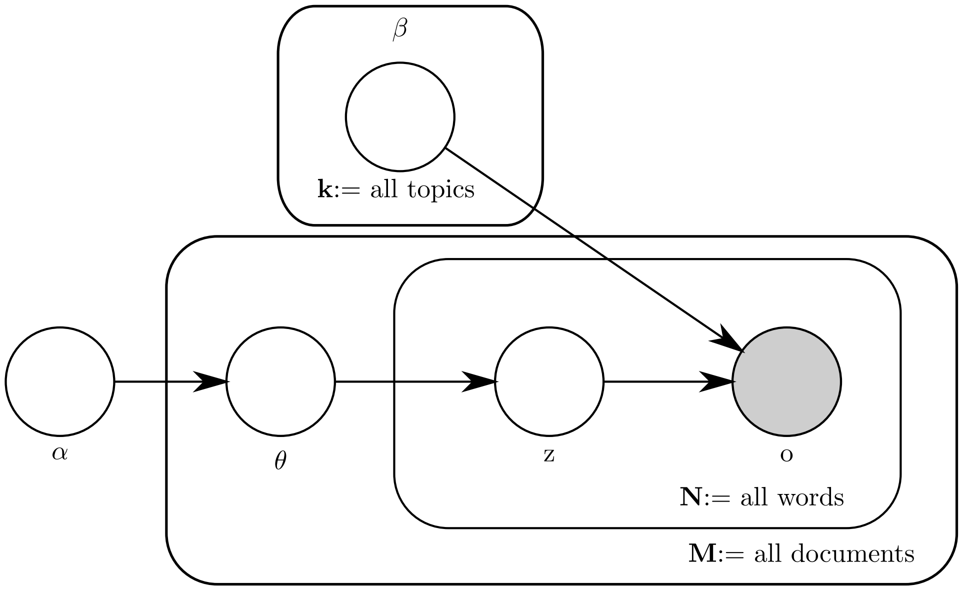


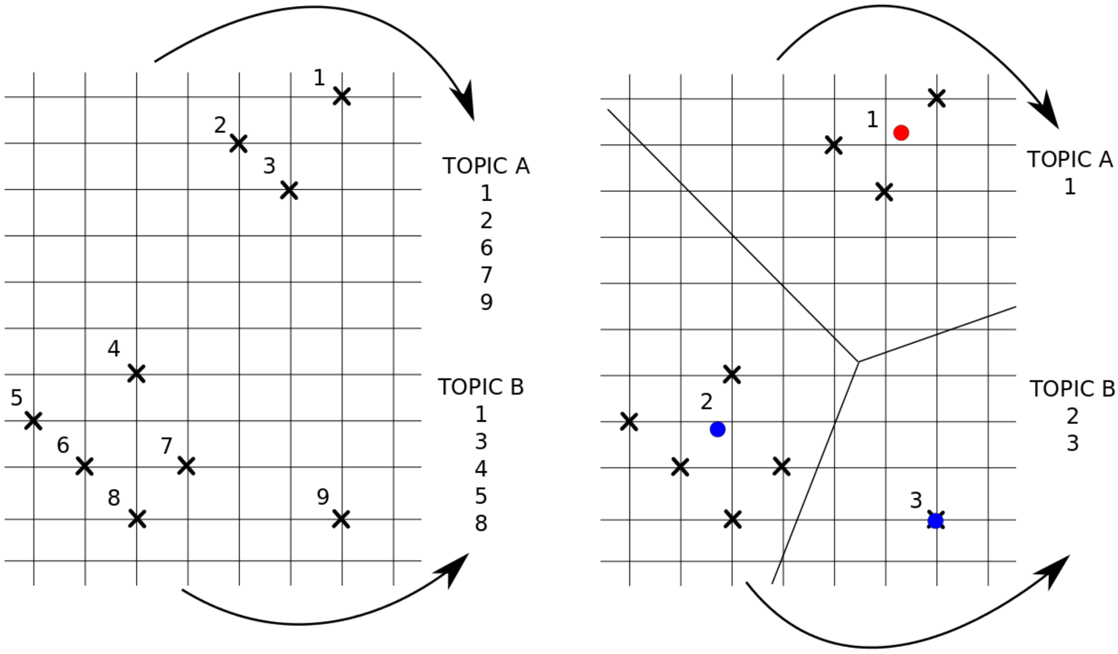
Figure 5.1: Graphical representation of the LDA model

5.2 Topic models with Sensor Data

In order to employ the topic model with the sensor data, first the different levels of description are introduced and related to the terms of document classification.

- **Dataset/Corpus**: One dataset C describes the sensor data that is obtained in the home of a single person. So for every person there is a separate dataset. This set of data can be compared with a Corpus in document classification.
- **Day/Document**: Every dataset is divided in days, which is a natural choice because daily routines are a good indication for behavior patterns. A day can be compared with a document in a Corpus.
- **Observations/Words**: Finally every day is built of a set of observations, which are already introduced in chapter 4. Observations can be roughly compared with words in document classification. A different name for this observations is 'Artificial words'.

There are some differences between the data that is used for topic detection in text documents and in data obtained from binary sensors. The main difference is that words that look similar to each other, like "illusion" and "allusion", may belong to a totally different topic in the document classification. But in the case of sensor data, two observation that are similar to each other, are more likely to refer to the same topic. So for example if the topic "preparing breakfast" has the observation $o_n = 2, 4, 0, 3, 2, 2$ a similar observation like $o_n = 2, 4, 0, 4, 2, 2$, where only the fourth value of the vector is changed, may also refer to the same topic. With a BOW model, where every unique observation forms a new dimension, this similarity cannot directly be captured. The only way to capture this relationship is by observing the similar observations with another observation. Say observation A and B are similar to each other. So if observation A is seen with observation C and also B is seen with C, then it might be possible that the LDA algorithm assigns the same topic for A and B. A lot of data is necessary to find these kind of relations. Another difference is that in the topic model LDA, it is assumed that the order of the words in the text is not important. However with sensor data the time when an observation is made is of big influence. By adding the time value to the observation we can overcome this problem. The fine grained time representation (see chapter 4) will contain more information about the sequential order of the observations than



a) b)

Figure 5.2: a) The Bag-of-words model with topic assignment. b) Feature space with k-means and topic assignment.

the coarse grain representation.

If the Bag-of-Words model will be applied to the data, a dictionary of all unique observations must be made. All this observations then can be seen as a different dimension which are independent of each other. In figure 5.2 a) an example is shown how a topic model can assign topics to the different observations. It is not necessary that two observations lay close to each other if they are assigned to the same topics. So the spatial relationship between the observation is not taken into account. Observations can be assigned to multiple topics.

To make sure that similar observations will be grouped together and appear in the same topic, the k-means algorithm can be applied to find clusters in the data. The id's of the centroids of the clusters then function as the 'artificial words'. The size of the dictionary of the BOW model is in this way reduced. Applying LDA on this reduced representation can lead to a better topic description. This is shown in figure 5.2 b).

The outcome of the LDA model does strongly depend on the outcome of the apriori used cluster algorithm. Choosing the correct number of clusters for the k-means algorithm can be difficult, as it is also pointed out in the work of Casale [2]. And still it can occur that the clusters do not give a good representation of the data. All clusters have the same influence on the LDA model, which is not always desirable. A Gaussian Mixture model (GMM) can give a better description of the clusters and give a degradation of the topics. Every Gaussian distribution can be assigned to a topic with the LDA model. This is shown in figure 5.3 a) with the red and blue Gaussian distributions. The red distribution is assigned to one topic and the two blue distributions are assigned to another topic.

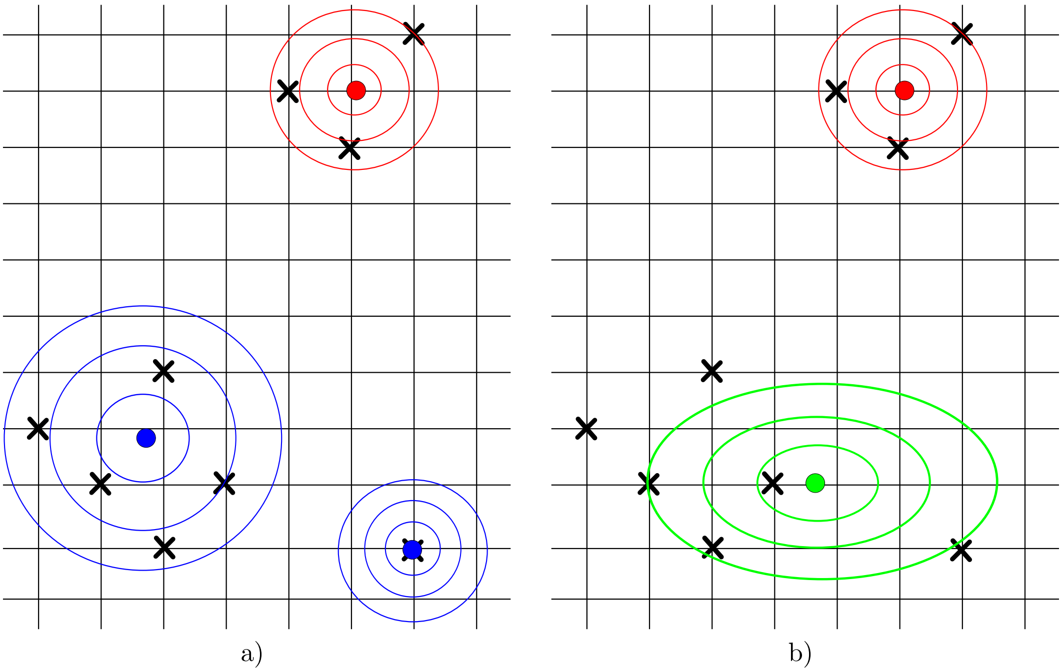


Figure 5.3: a) GMM and LDA. b) GMM and LDA combined in one model.

Combining the clustering and the Topic model directly together can lead to a more global representation of the topics. This is indicated with the green Gaussian distribution in figure 5.3 b).

5.3 LDA-Gaussian Model

In this section the LDA-Gaussian model is described. First the model is explained, which is followed by the description of the variational inference. This inference step is necessary to calculate the model parameters. In this subsection also the two steps of the EM-algorithm are explained.

5.3.1 Model Description

For the model it is assumed that every day in a dataset can be represented as random mixture of latent topics, where every topic can be described as a distribution over observations. The following generative process is assumed for every day m in the data set C:

1. The number of observations on a day m is fixed with size N (for every day the same size).
2. A day has a distribution over the topics given with $\theta \sim Dir(\alpha)$.
3. For each of the N observations on a day o_n:
   (a) Estimate the topic $z_n \sim Multinomial(\theta)$.
   (b) An observation o_n is gained from $p(o_n|z_n, \boldsymbol{\mu}, \boldsymbol{\sigma})$, which is a probability that can be drawn from a set of Gaussian distributions. This probability is conditioned on the topic z_n and the Gaussian Parameters $\vec{\mu_{z_n}}$ and $\vec{\sigma_{z_n}}$ of length d that belong to the estimated topic z_n.

In this model the number of topics k is assumed to be known and fixed and with it the size of the topic variable z. The probability for the observations is parametrized with two matrices $\boldsymbol{\mu}$ and

$\boldsymbol{\sigma}$, both of size $D \times k$, where D is the number of dimensions in an observation and k the number of topics. They present the mean and standard deviation respectively and for every topic i and every dimension d there is a set of parameters, which describes a Gaussian distribution. Every value of a dimension for an observation o_{ndi} can then be drawn from a Gaussian Distribution $\mathcal{N}(\mu_{di}, \sigma_{di})$. With the feature representation given in chapter 4 the size of the dimension is fixed with $D = 6$. α represents the Dirichlet parameter and is vector of length k.

In figure 5.4 the graphical representation of the model is shown. It differs from the basic LDA model (figure 5.1) in the description of the topic distribution β, which is here replaced with μ and σ.

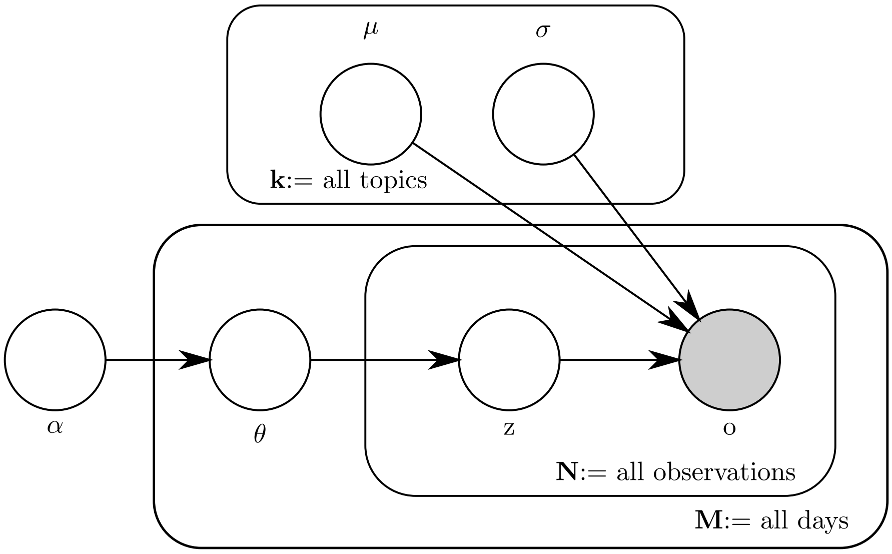


Figure 5.4: Graphical representation of the LDA-Gaussian model

Assuming that the generative process described above can describe the available data properly, the optimal parameters need to be found. Therefore the probability for the dataset C needs to be maximized, with respect to the parameters α, μ and σ. This probability looks like this

$$p(C|\alpha, \mu, \sigma) = \prod_{m=1}^{M} p(\mathbf{o}_m|\alpha, \mu, \sigma) \tag{5.1}$$

where M is the number of days within the dataset.

For one day the joint distribution of a topic mixture θ, a set of N topics $\mathbf{z}$ and a set of N observations $\mathbf{o}$ is:

$$p(\theta, \mathbf{z}, \mathbf{o}|\alpha, \mu, \sigma) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(\vec{o}_n|z_n, \mu, \sigma) \tag{5.2}$$

. Integrating over θ and summing over z the marginal distribution of one document is obtained with

$$p(\mathbf{o}|\alpha, \mu, \sigma) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(\vec{o}_n|z_n, \mu, \sigma) \right) d\theta \tag{5.3}$$

This probability can then be substituted into the equation 5.1.

In the next section it is described how the optimal parameters for the model given a data set can be found.

5.3.2 Variational Inference

The marginal distribution, which is given in the previous section, can be written in terms of the parameter α, μ and σ as

$$p(\mathbf{o}_m|\alpha,\mu,\sigma) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}\right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{d=1}^{D} \theta_i \mathcal{N}(o_{nd};\mu_{id},\sigma_{id})\right) d\theta \tag{5.4}$$

Due to the coupling between θ and the Gaussian parameters μ and σ this probability is intractable to compute.

That is why a convexity-based variational algorithm is applied, so that the log-likelihood of a given dataset can be approximated. An approximation of the model is given with

$$q(\theta, z|\gamma,\phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n). \tag{5.5}$$

In this model γ represents the Dirichlet parameter and ϕ are the multinominal parameter which can be viewed as the probability $p(z_i|o_n)$ and is given as a $k \times N$-matrix for every day m. The graphical representation of the model is shown in figure 5.5.

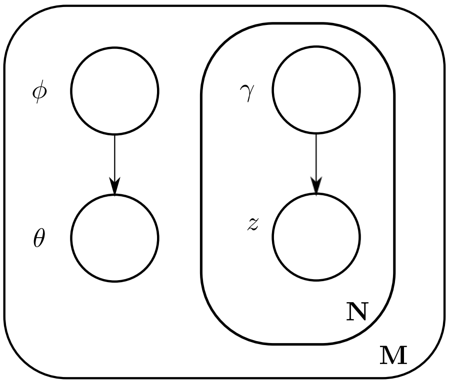


Figure 5.5: Approximation of the model.

Given the variational distribution for the approximate model the lower bound of the log-likelihood can be estimated with the Jensen inequality

$$\begin{aligned} L(\gamma;\phi;\alpha;\mu;\sigma) =& E_q[\log p(\theta|\alpha)] + E_q[\log p(\mathbf{z}|\theta)] + E_q[\log p(\mathbf{o}|\mathbf{z},\mu,\sigma)] \\ &- E_q[\log p(\theta)] - E_q[\log q(\mathbf{z})] \end{aligned} \tag{5.6}$$

In terms of the model parameters and the variational parameters this becomes

$$\begin{aligned} L(\gamma;\phi;\alpha;\mu;\sigma) =& \log\Gamma(\sum_{j=1}^{k}\alpha_j) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) + \sum_{i=1}^{k}(\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)) \\ &+ \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_i)) \\ &+ \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{d=1}^{D}\phi_{ni}\log(\mathcal{N}(o_{nd};\mu_{id},\sigma_{id})) \\ &- \log\Gamma(\sum_{j=1}^{k}\gamma_j) + \sum_{i=1}^{k}\log\Gamma(\gamma_i) - \sum_{i=1}^{k}(\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)) \\ &- \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\log\phi_{ni} \end{aligned} \tag{5.7}$$

Notice that in the implementation the Gaussian distribution is integrated over a time interval of $[o_{nd} - 0.5, o_{nd} + 0.5]$, so that the log-likelihood becomes a probability value. In this way numerical errors in the log-likelihood are smoothed.

An EM-process is able to maximize this lower bound on the log-likelihood. The two steps are:

1. **E-step:** For each day m, optimize the variational parameters $\{\gamma_m*, \phi_m*\}$
2. **M-step:** Maximize the resulting lower bound on the log-likelihood with respect to the model parameters α, μ and σ.

Now a more detailed description on both of these steps is given:

E-step In the E-step of the algorithm the variational parameters ϕ and γ are optimized. To get the update function for ϕ all terms of the lower bound of the log-likelihood in equation (5.7) are gathered, that contain the variable ϕ. Take $y_i = \sum_{d=1}^{D} \mathcal{N}(o_{nd}; \mu_{id}, \sigma_{id})$ and add the constraint $\sum_{i=1}^{k} \phi_{ni} = 1$. This results in the following formula

$$L_{[\phi_{ni}]} = \phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)) + \phi_{ni} \log(y_i) + \lambda_n(\sum_{j=1}^{k} \phi_{ni} - 1) \tag{5.8}$$

From this equation the derivative can be taken and set it to zero. This leads to the first update function

$$\phi_{ni} \propto \log(y_i) \exp(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)) \tag{5.9}$$

A similar approach is handled for γ. Again all terms of equation 5.7 that contain this variable is selected and the derivative is set to zero. This leads to the second update equation

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{5.10}$$

M-Step In the M-step the parameters of the Gaussian distribution μ and σ are estimated with the weighted arithmetic mean calculated over all observation in a dataset given the parameter ϕ, which is gained in the previously e-step. This leads to the update formulas

$$\mu_{di} = \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} o_{dn} \phi_{ni}}{\sum_{m=1}^{M} \sum_{n=1}^{N} \phi_{ni}} \tag{5.11}$$

and

$$\sigma_{di} = \sqrt{\frac{\sum_{m=1}^{M} \sum_{n=1}^{N} o_{dn}^2 \phi_{ni}}{\sum_{m=1}^{M} \sum_{n=1}^{N} \phi_{ni}} - \mu_{di}^2} \tag{5.12}$$

To calculate the parameter α again all terms of the likelihood that contains the variable α are selected. The derivative in the Hessian form is

$$\frac{\partial L}{\partial \alpha_i \alpha_j} = m(i,j) M \Psi'(\alpha_i) - \Psi'(\sum_{j=1}^{k} \alpha_j) \tag{5.13}$$

On this equation the Newton-Rhapson method can be used to calculate the optimal α.

5.4 LDA-Poisson Model

The feature description in chapter 4 generates positive integers. Small values are more likely to occur in the observations and that is why a Poisson distribution is a better choice for the description of the topics. In figure 5.6 the graphical representation of LDA-Poisson model is shown.

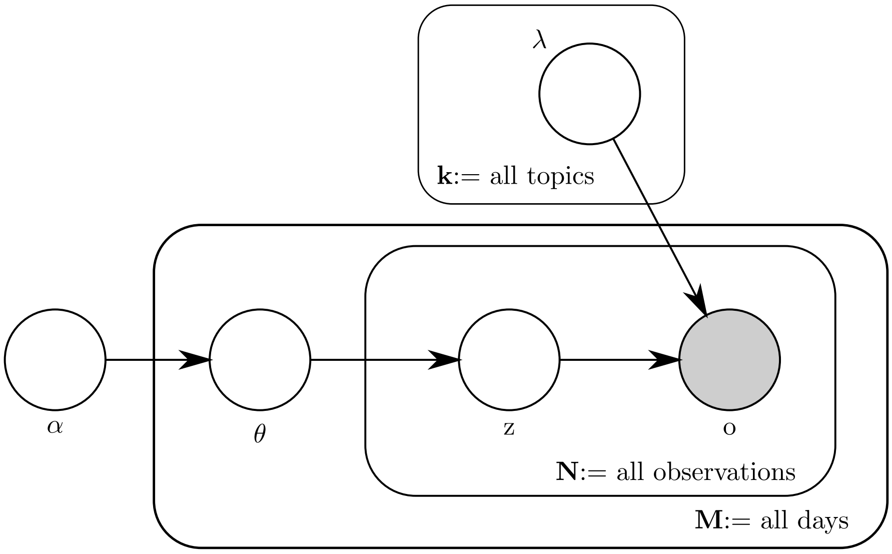


Figure 5.6: Graphical representation of the LDA-Poisson model.

The variational Inference and the EM-procedure will be the same as described before, except for the Gaussian distribution that is exchanged with the Poisson distribution. The parameter *lambda* which describes the Poisson distribution is calculated in the M-step with

$$\lambda_{di} = \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} o_{dn} \phi_{ni}}{\sum_{m=1}^{M} \sum_{n=1}^{N} \phi_{ni}}. \tag{5.14}$$

Chapter 6

Experiments

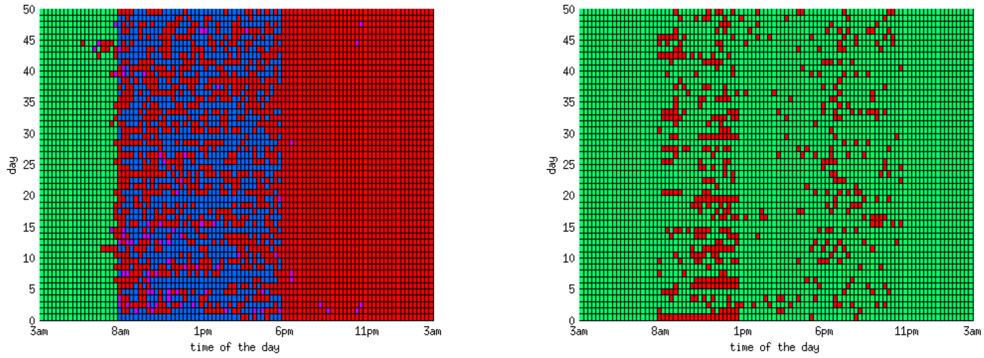
In this chapter the experiments and the results are presented. In the section of the qualitative results visualizations of the topics are given for different models. It is shown that the topics are meaningful and that semantic descriptions can be given to some of these topics. In the section of the quantitative results the two developed models are compared with each other and the performance of the model for different number of topics and different number of time-slices are shown.

6.1 Qualitative Results

6.1.1 Comparison of the different models

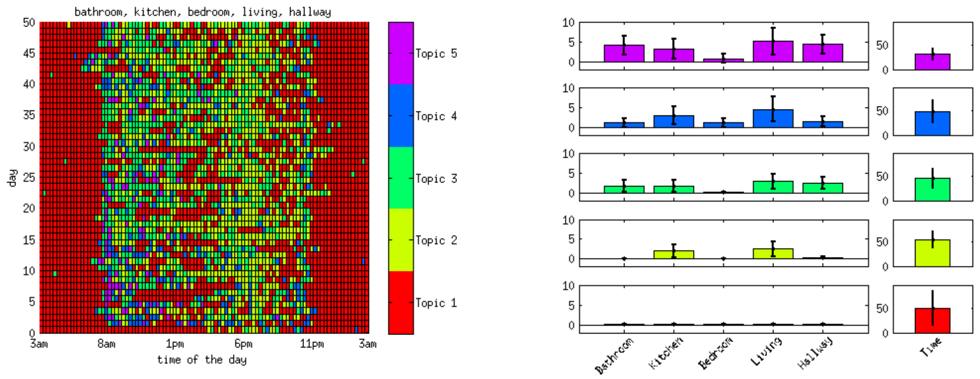
In this section four different models are compared with each other. The first two models are the LDA model with a Bag-of-word representation, where the dictionary one the one hand is pre-clustered with the k-means algorithm(LDA with BOW + k-means, figure 6.1 a)) and on the other hand not clustered at all (LDA with BOW, figure 6.1 b)). The third and fourth model are LDA-Gaussian and LDA-Poisson (figure 6.2 and 6.3).

In the following figures the number of time-slices on a day is $N = 96$. The time dimension is a coarse-grain representation, which means that there are five different values for the time. The number of topics that are initialized are $k = 5$. Figures 6.1 a) + b), 6.2 a) and 6.3 a) are a representation of the topic distribution on different days. There are in total 50 days represented on the y-axis in each figure. On the x-axis the time-slices are shown. The different colors represent the different topics that are assigned to the time-slices.



a) b)

Figure 6.1: a) LDA model applied to BOW representation with k-means clustering. b) LDA applied to data with simple BOW representation of the data.

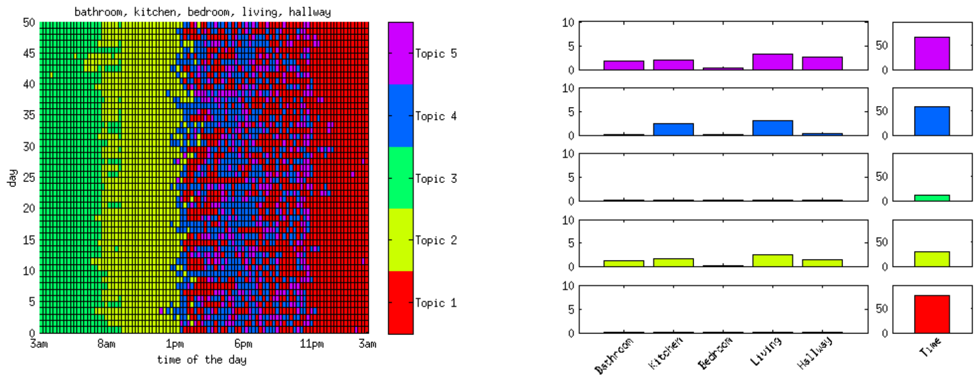


a) b)

Figure 6.2: Topic distribution per day and the Topic visualization for LDA-Gaussian

The LDA model with a BOW-representation is still able to find different topics in the data, although the BOW representation without the clustering only distinguishes between two topics (see figure 6.1 b)). If the data is clustered beforehand LDA is able to find four different topics in the data (see figure 6.1 a)). The number of clusters for the k-means algorithms is set to $V = 6$. This was the best value to find the most topics in the data. Before the data was clustered, all dimensions where normalized.

Figure 6.2 shows the outcome of the LDA-Gaussian model. In the figure on the right-hand side (6.2 b)) the 6 dimensions of the observations are shown on the x-axis. Every topic is marked with a different color and corresponds to the topics in the left image. On the y-axis the mean value μ of every Gaussian distribution in every dimension is indicated with the height of the bar-chart. The vertical black line represents the standard deviation σ of every distribution. The red topic (topic 1) contains a high σ-value for the time-dimension and captures all time-slice where the value is zero for the sensor values. LDA-Gaussian is able to find more notable topics and leads to more meaningful results than it can found with LDA for a BOW representation. For example the daily structure of a person can be distinguished more easily than it can be done in figure 6.1 a) or b).



a) b)

Figure 6.3: Topic distribution per day and the Topic visualization for LDA-Poisson

In figure 6.3 the outcome of the LDA-Poisson model is shown. The topics that are found are again shown on the right-hand side (6.3 b)). The height of the bars represent the λ-value of the Poisson-distributions. In the figure 6.3 a) five different topics are found. These topics depend however a lot on the time. This is caused by the fact that the Poisson distribution is not a good way to model the time dimension, because small values will have a higher probability in this distribution. Nevertheless LDA-Poisson still performs better, than LDA on the BOW representation.

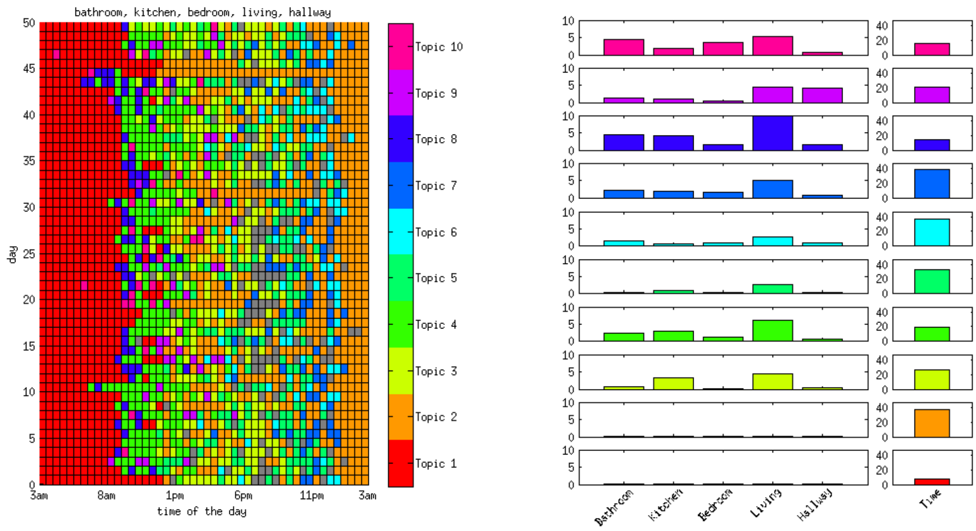
6.1.2 Semantic topic description

In the following figures the topic distribution for different houses are shown. A semantic description of some topics is given to illustrate the usefulness of the models. The number of time-slices is set to $N = 48$. In every figure 50 days are shown. In the left images a) the topic distribution of the days are shown and in the left images b) the topic description for every dimension is given with respectively the same color. For more results see the appendix A and B.

In figure 6.6 the outcome for the LDA-Gaussian model is shown. The time dimension has a fine grained representation. The model is initialized with $k = 20$ topics and the 10 most important topics are visualized with a different color, the remaining topics have the color gray in the left image. The first topic (red) is an easy topic to find. In all time-slices that are marked with red no activations of the sensors are captured. The time dimension has a big standard deviation, because this topic can occur all day long. In figure 6.4, which is outcome of the LDA-Poisson model with same values, thi topic is divided in two different topics (red and orange). These topics only distinguish between the time value. Comparing the two outcomes (figure 6.6) and 6.4) one can see that topics of the LDA-Gaussian model are much easier to interpret. For example the ninth topic (purple) can be seen as the topic 'preparing for bed'. This topic cannot be found in the outcome of the LDA-Poisson model. The topic 'going to toilet', which is the third topic (yellow) in the LDA-Gaussian model is also not shown in the outcome of the LDA-Poisson model.

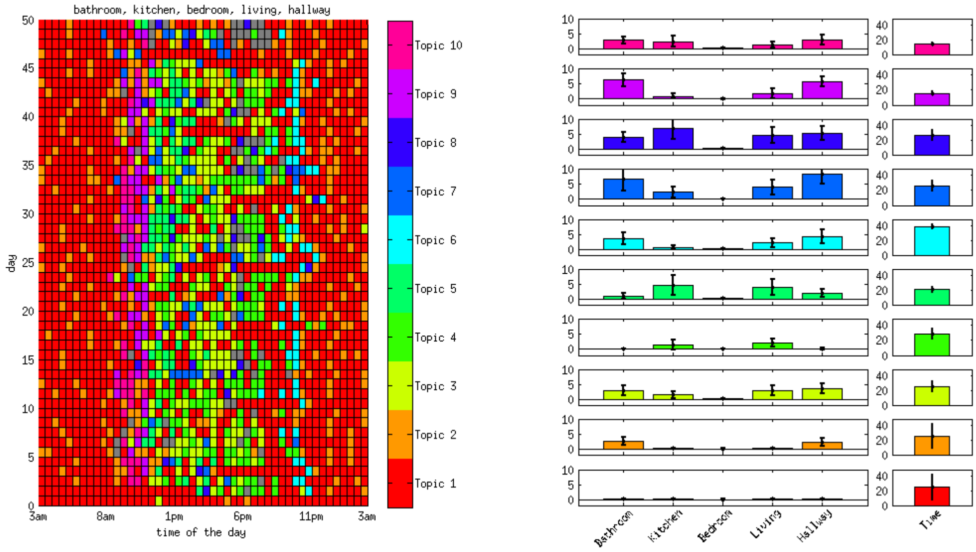
In figure 6.5 the 'going to toilet' is also present. This figure shows the outcome for House number 5 with the LDA-Gaussian model. Here two topics (purple and pink) can be found, which indicates the 'morning preparations'. Every week the person seems to leaving the house in the afternoon and comes back around 11 pm. Comparing house number 1 and 5 with each other one can see that the person that is living in house number 5 goes to bed a little bit earlier than the person living in house number 1.

More topics can be appointed and for some person it appears easier than for other person. The main point however is that indeed a daily or even weekly structure can be found with distributions of the topics during a day.



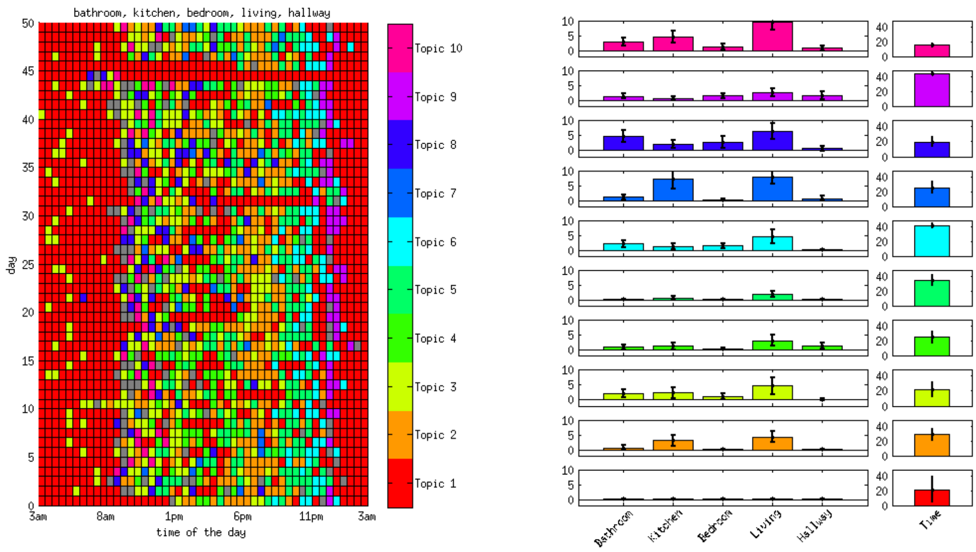
a) b)

Figure 6.4: House number 1, 20 topics, fine grain time, LDA-Poisson



a) b)

Figure 6.5: House number 5, 20 topics, fine grain time, LDA-Gaussian



a) b)

Figure 6.6: House number 1, 20 topics, fine grain time, LDA-Gaussian

6.1.3 Quantitative Results

The log-likelihood indicates on how well the model with the estimated parameters fits the data. This can easily overfit and new data can not easily modeled with the found parameters. Therefor it is necessary to also find a high log-likelihood on a hold-out-set. The perplexity gives a good indication on how and can be calculated with

$$\text{perplexity}(D_{HOS}) = \exp\left\{-\frac{\sum_{m=1}^{M}\log p(\mathbf{o}_d)}{M * N}\right\} \tag{6.1}$$

A low perplexity value indicates a good result on the hold-out-set.

In the next experiments 10% percent of the data is used as a hold-out set. The model is trained on the remaining 90% with different initialization values and the perplexity is calculated on the hold-out-set with the estimated parameters. In the following it is shown what the effect is on different number of topics and different numbers of time-slices. Also the two models LDA-Gaussian and LDA-Poisson are compared with each other.

Different Sets of Data In figure 6.7 the perplexity is shown for the 5 different data-sets gained from the 5 houses. Every run is performed ten times and the mean over these runs is taken for different number of topics k. Every run is initialized with 5 random days. The data sets vary in length, but as one can see in the figure, the amount of data is not necessary of influence how well the LDA-Gaussian model can be trained. Some data sets are much more stable than others and this is probably due to the way how regular peoples behavior is.

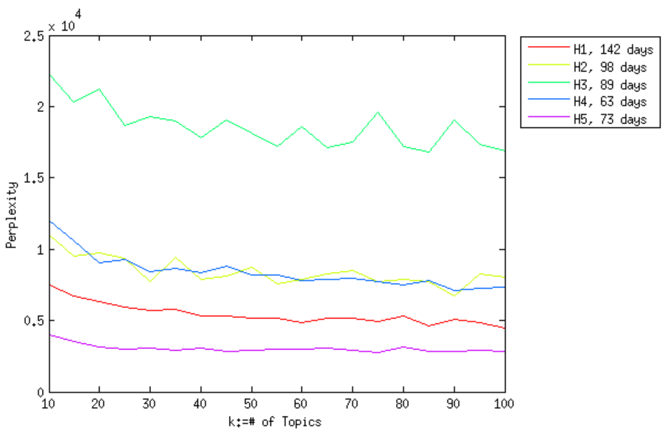


Figure 6.7: Perplexity for different number of topics for 5 different House

Comparison LDA-Gaussian and LDA-Poisson The LDA-Poisson model is not capable to model the time dimension properly. To be able to compare both models directly with each other the time dimension is left out in the following experiments. The perplexity is calculated for the hold-out set with different amount of time-slices and the mean of 10 runs for different number of time-slices is shown in figure 6.8. The perplexity gets better if the number of time-slices increases, but at one point the there will be no improvement. If the length of the time-slices becomes so small that only one activation is captured in an observation, the model perplexity becomes indeed very high, because a topic only contains one value. The topics are then however not very informative.

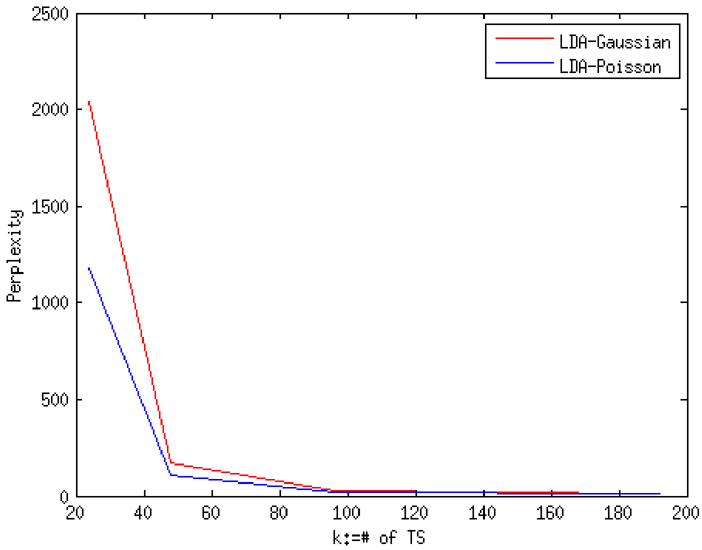


Figure 6.8: Perplexity for LDA-Gaussian and LDA-Poisson with different amount of time-slices

In figure 6.9 the perplexity for both models is shown with different number of topics. The mean of 10 runs is shown and also the standard deviation is indicated with the vertical errorbar in the figure. The standard deviation increases for a higher number of topics, which is reasonable, because more topics increases the chance of more variation. If the number of topic increases the perplexity first decreases and then increases. If the number of topics is to high the model does overfit on the training data and the perplexity indeed becomes worse.

To creating the figures the data of house number 2 is chosen. For other data sets the outcomes looks similar and are therefore not presented. Both experiments show that the LDA-Poisson model outperforms the LDA-Gaussian model.

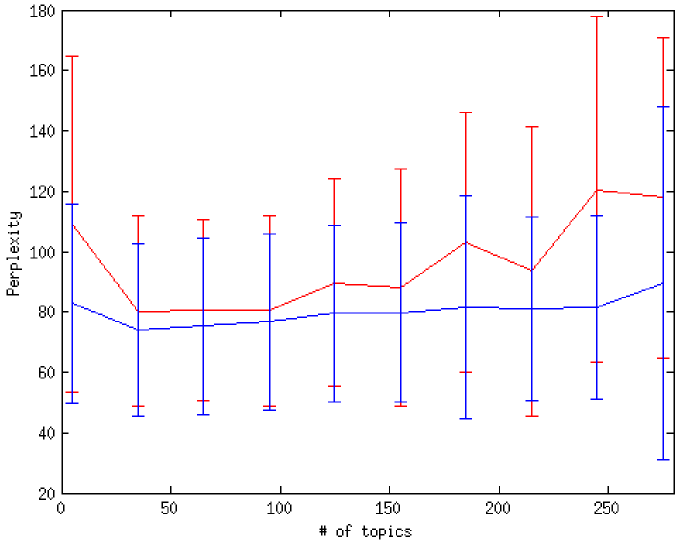


Figure 6.9: Perplexity for LDA-Gaussian and LDA-Poisson with different amount of topics

Chapter 7

Conclusions

In this thesis two novel models, LDA-Gaussian and LDA-Poisson, are presented. The models are both successfully applied to real-life, binary sensor data, which is gathered from five different houses of elderly people. It is shown that the models can find meaningful topics in relative little data.

The topics and how these topics are distributed over a day can give a good insight on peoples daily behavior. This information can be used to monitor peoples health condition on a long term basis. Changes in the patterns can be a sign of decline in the health of people. These declines can be detect manually or even automatically.

The experiments show that, when time dimension is not taken into account, the LDA-Poisson model performs better than LDA-Gaussian model. This is an expected outcome, because without the time dimension, the feature representation of all dimension is event based. The Poisson distribution is not appropriate to model the time dimension, because this distribution is not symmetric. The Gaussian distribution is a better fit, but time is periodical and topics that are found during the night might be separated because of the linear representation of the time. The von Mises distribution, which is a continuous probability distribution on a circle, is a potential better choice.

In future work it might be a good idea to combine different distributions for the different dimensions in the observations. Event based dimensions can be modeled with the Poisson distribution and time can be modeled with the von Mises distribution. Another idea might be to investigate the LDA-Gaussian model in more detail. In the implementation the covariance is modeled with a diagonal matrix. A fully trained covariance matrix is able to capture relationships between the different dimensions and could lead to better topic descriptions. To be able to find a correct covariance matrix a lot of data needs to be available and this is however contradictory to one main benefits of the new topic models: the fact that it can find topics in relative little data.

In this thesis a simple feature representation was chosen. In the experiments only a variation in size of non-overlapping time-slices was considered. Small time-slices did lead to a higher likelihood on a hold-out-set, but the topics that were found were not informative enough. In addition, smaller slices leaded to a higher complexity of the model. This is not desirable, because of the high computational costs. There are many other representation that could be used to find good features. For example time-slice can be overlapping each other or a word can be a combination of the values from sequential time-slices. In this way the transitions are taken into account. Other possibilities are: Taking the activation length of a sensor into account or looking at every sensor separately than rather put them together in one field. In fact there are infinite ways to combine the data into features and the best representation is probably not yet found. With a more complex representation the computational costs will raise and even more important the features are more likely to overfit the data.

Finding the best feature representation is in fact a big challenge. The proposed models give the opportunity to investigate feature representation. For example a Genetic algorithm can be used to investigate the feature space. The likelihood with a combination of a penalty for high complexity can be a good fitness function to automatically find the best feature representation.

Bibliography

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. 1, 13

[2] Pierluigi Casale, Oriol Pujol, Petia Radeva, and Jordi Vitrià. A first approach to activity recognition using topic models. In *Proceedings of the 2009 conference on Artificial Intelligence Research and Development: Proceedings of the 12th International Conference of the Catalan Association for Artificial Intelligence*, pages 74–82, Amsterdam, The Netherlands, The Netherlands, 2009. IOS Press. 2, 3, 15

[3] Federico Castanedo, Diego López de Ipiña, Hamid K. Aghajan, and Richard Kleihorst. Learning routines over long-term sensor data using topic models. *Expert Systems*, pages n/a–n/a, 2013. 2, 4, 10

[4] F Cavallo, G van den Broek, and C Wehrmann. *AALIANCE Ambient Assisted Living Roadmap*. Ambient Intelligence and Smart Environments. IOS Press, Amsterdam, 2010. 1

[5] Belkacem Chikhaoui, Shengrui Wang, and Hélène Pigot. Adr-splda: Activity discovery and recognition by combining sequential patterns and latent dirichlet allocation. *Pervasive and Mobile Computing*, 8(6):845–862, 2012. 1, 3

[6] Katayoun Farrahi and Daniel Gatica-Perez. Daily routine classification from mobile phone data. In *Proceedings of the 5th international workshop on Machine Learning for Multimodal Interaction*, MLMI '08, pages 173–184, Berlin, Heidelberg, 2008. Springer-Verlag. 1, 2, 4, 9, 10

[7] Tao Gu, Zhanqing Wu, Xianping Tao, H.K. Pung, and Jian Lu. epsicar: An emerging patterns based approach to sequential, interleaved and concurrent activity recognition. In *Pervasive Computing and Communications, 2009. PerCom 2009. IEEE International Conference on*, pages 1–9, 2009. 1

[8] Xin Hong, Chris Nugent, Maurice Mulvenna, Sally McClean, Bryan Scotney, and Steven Devlin. Evidential fusion of sensor data for activity recognition in smart homes. *Pervasive and Mobile Computing*, 5(3):236 – 252, 2009. ¡ce:title¿Pervasive Health and Wellness Management¡/ce:title¿. 1, 3

[9] Tâm Huynh, Mario Fritz, and Bernt Schiele. Discovery of activity patterns using topic models. In *Proceedings of the 10th international conference on Ubiquitous computing*, UbiComp '08, pages 10–19, New York, NY, USA, 2008. ACM. 2, 3

[10] Ohbyung Kwon, Jae Moon Shim, and Geunchan Lim. Single activity sensor-based ensemble analysis for health monitoring of solitary elderly people. *Expert Systems with Applications*, 39(5):5774 – 5783, 2012. 1

[11] Muhammad Mubashir, Ling Shao, and Luke Seed. A survey on fall detection: Principles and approaches. *Neurocomputing*, 100(0):144 – 152, 2013. ¡ce:title¿Special issue: Behaviours in video¡/ce:title¿. 1

[12] Norihiro Nagai, Yusaku Fujii, Hiroshi Ueda, Koichi Maru, and Shigeru Kumakura. Pc-based monitoring system for efficient use of day care centers for elderly people. *Procedia - Social and Behavioral Sciences*, 2(1):204 – 208, 2010. ¡ce:title¿The 1st International Conference on Security Camera Network, Privacy Protection and Community Safety 2009¡/ce:title¿. 1, 3

[13] United Nations. 1

[14] Wendy A. Rogers, Beth Meyer, Neff Walker, and Arthur D. Fisk. Functional limitations to daily living tasks in the aged: A focus group analysis. *Human Factors*, (1):111–125. 3

[15] Toshiyo Tamura, Tatsuo Togawa, Mitsuhiro Ogawa, and Mikiko Yoda. Fully automated health monitoring system in the home. *Medical Engineering & Physics*, 20(8):573 – 579, 1998. 1

[16] Emmanuel Munguia Tapia, Stephen S. Intille, and Kent Larson. Activity recognition in the home using simple and ubiquitous sensors. In *In Pervasive*, pages 158–175, 2004. 1, 3

[17] Tim van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben Kröse. Accurate activity recognition in a home setting. In *Proceedings of the 10th international conference on Ubiquitous computing*, UbiComp '08, pages 1–9, New York, NY, USA, 2008. ACM. 3

[18] D. H. Wilson and C. Atkeson. Simultaneous tracking and activity recognition (star) using many anonymous, binary sensors. In *Proceedings of the Third international conference on Pervasive Computing*, PERVASIVE'05, pages 62–79, Berlin, Heidelberg, 2005. Springer-Verlag. 1, 3

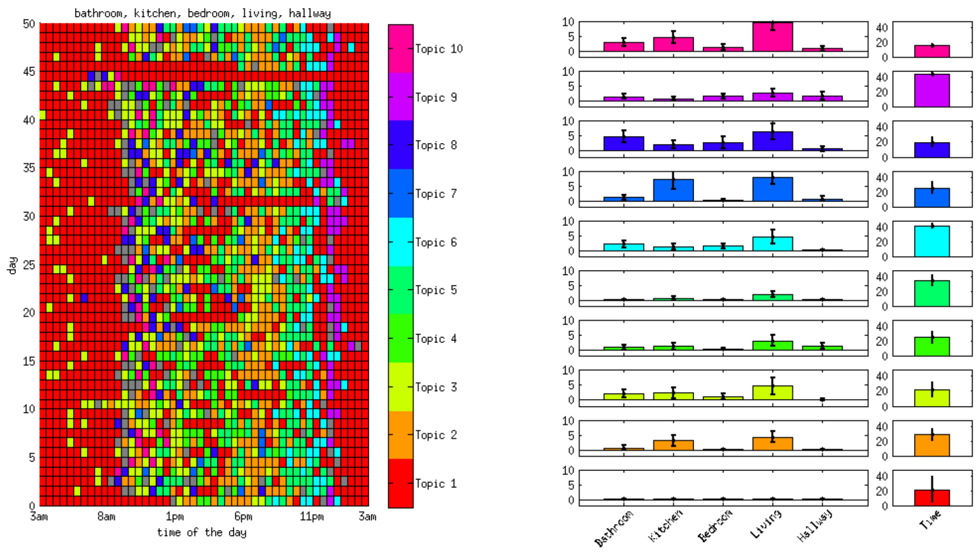
[19] Huadong Wu. *Sensor Fusion for Context-Aware Computing Using Dempster-Shafer Theory.* PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, December 2003. 3

Appendix A

LDA-Gaussian

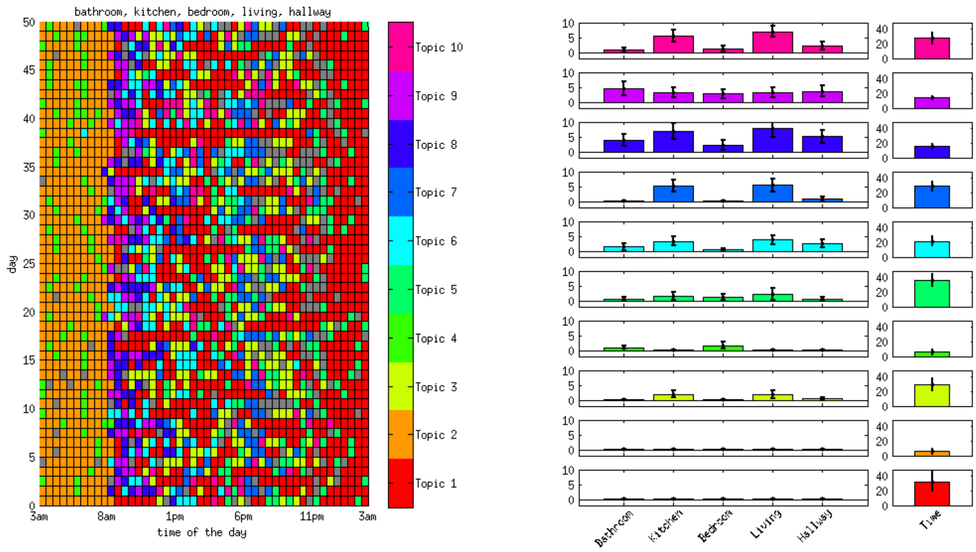
In this part of the appendix the results for all houses are shown for the LDA-Gaussian model. The number of time-slices is $n = 48$ and the number of topics is $k = 20$. The time dimension has a fine grain representation.

LDA-Gaussian Here the outcome of the LDA-Gaussian model is shown.



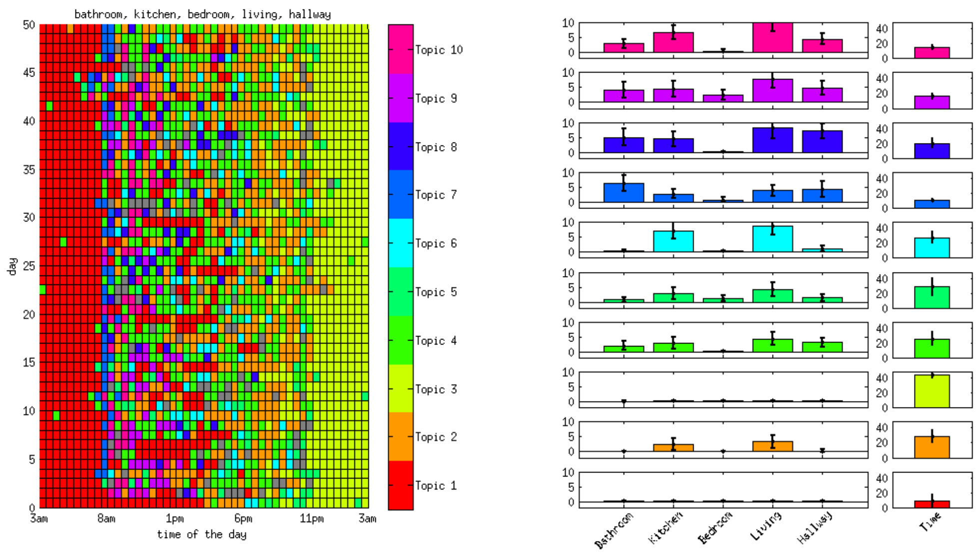
a) b)

Figure A.1: House number 1, 20 topics, fine grain time, LDA-Gaussian



a) b)

Figure A.2: House number 2, 20 topics, fine grain time, LDA-Gaussian



a) b)

Figure A.3: House number 3, 20 topics, fine grain time, LDA-Gaussian

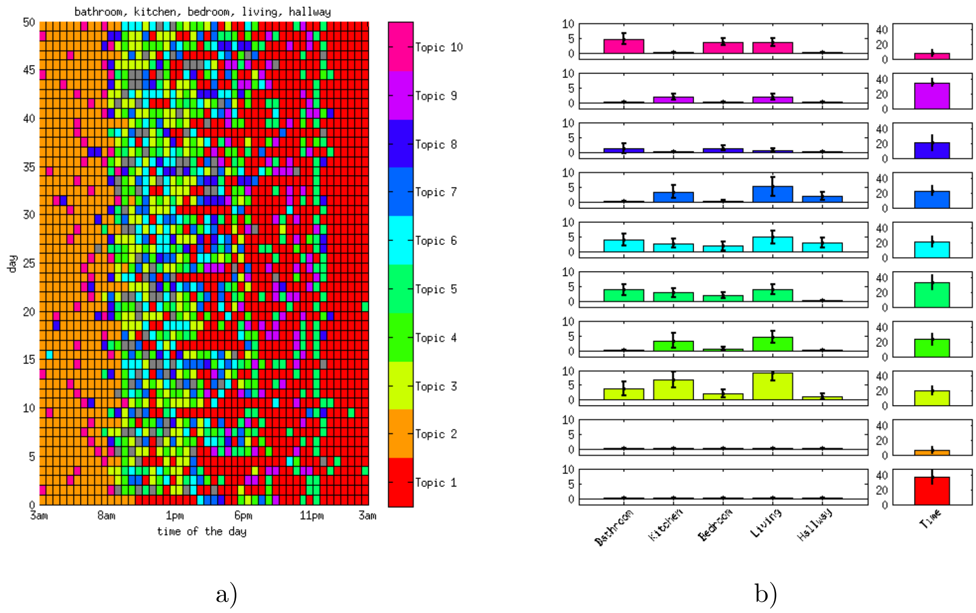


Figure A.4: House number 4, 20 topics, fine grain time, LDA-Gaussian

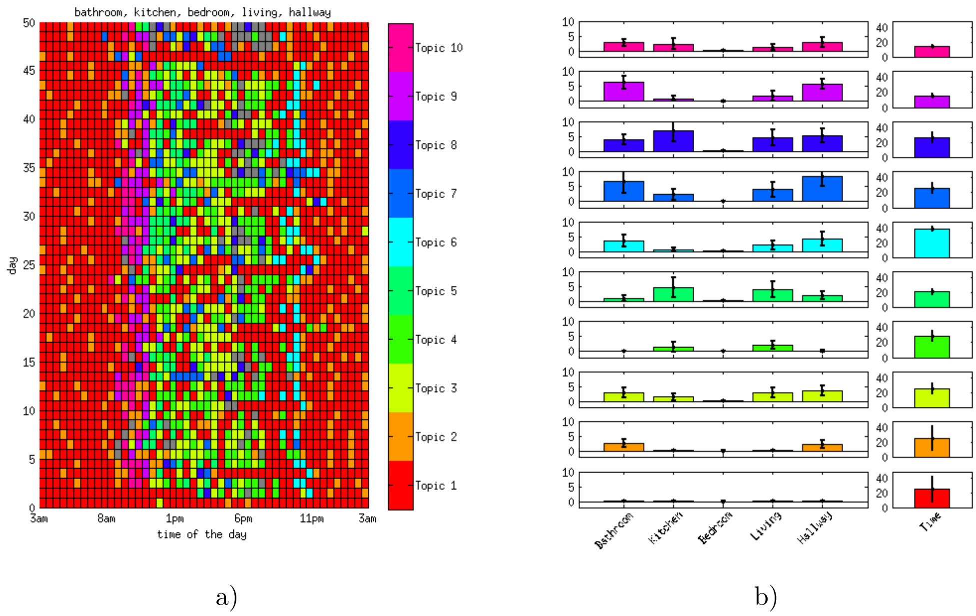
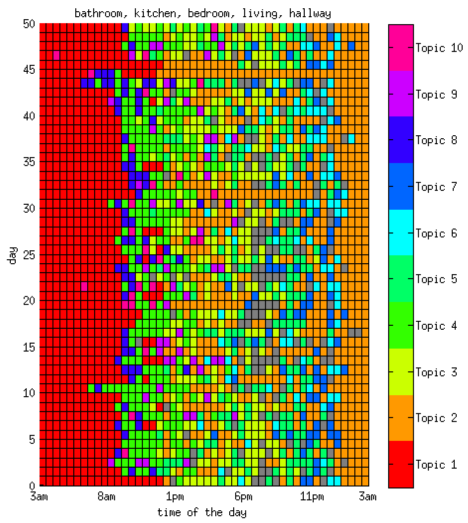


Figure A.5: House number 5, 20 topics, fine grain time, LDA-Gaussian

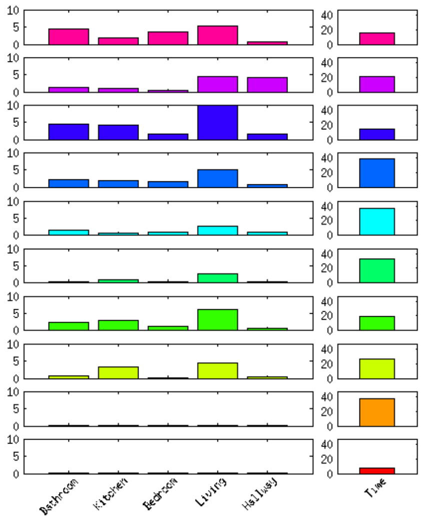
Appendix B

LDA-Poisson

Here the results for all houses are shown for the LDA-Poisson model. The number of time-slices is $n = 48$ and the number of topics is $k = 20$. The time dimension has a fine grain representation.



a)



b)

Figure B.1: House number 1, 20 topics, fine grain time, LDA-Poisson

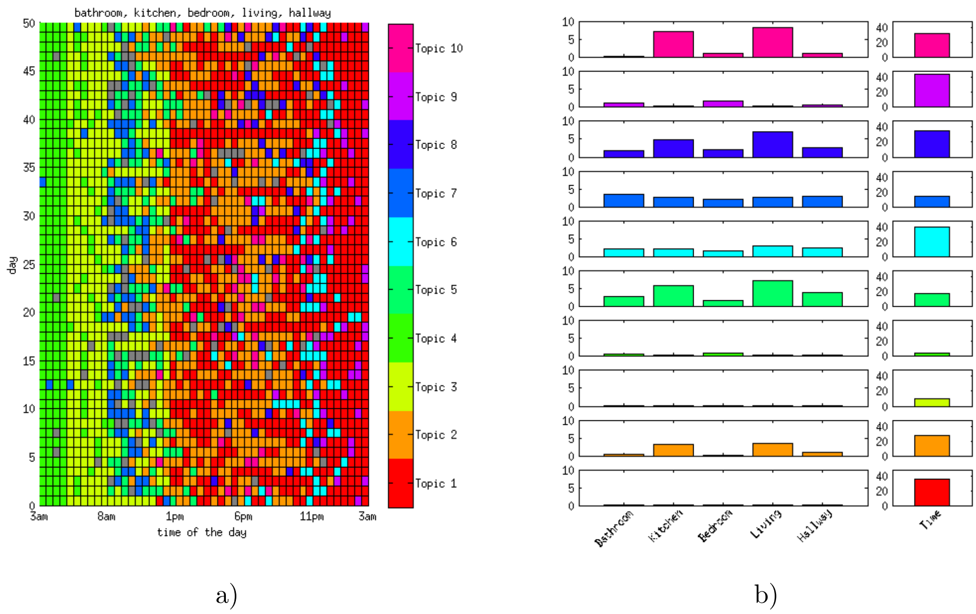


Figure B.2: House number 2, 20 topics, fine grain time, LDA-Poisson

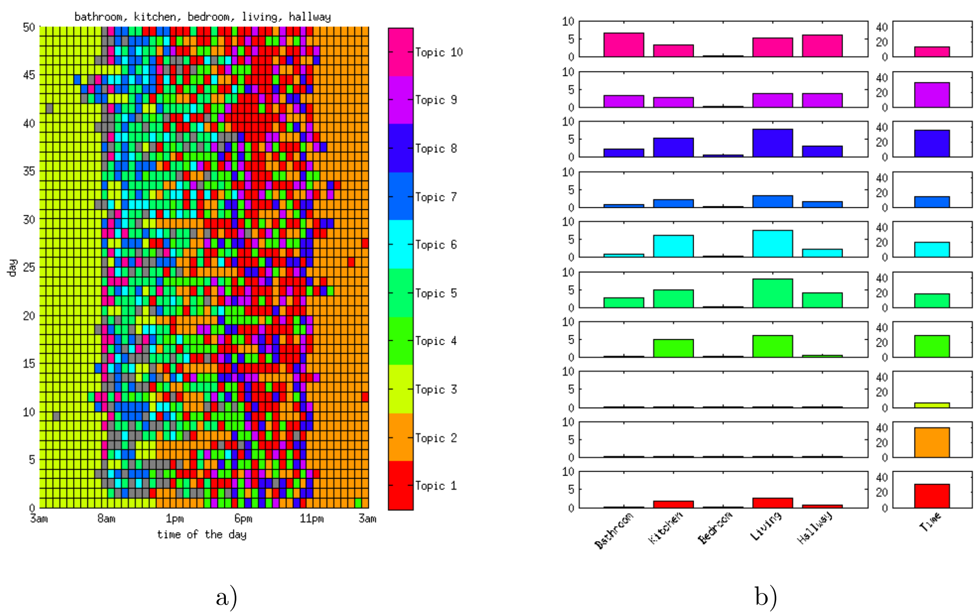
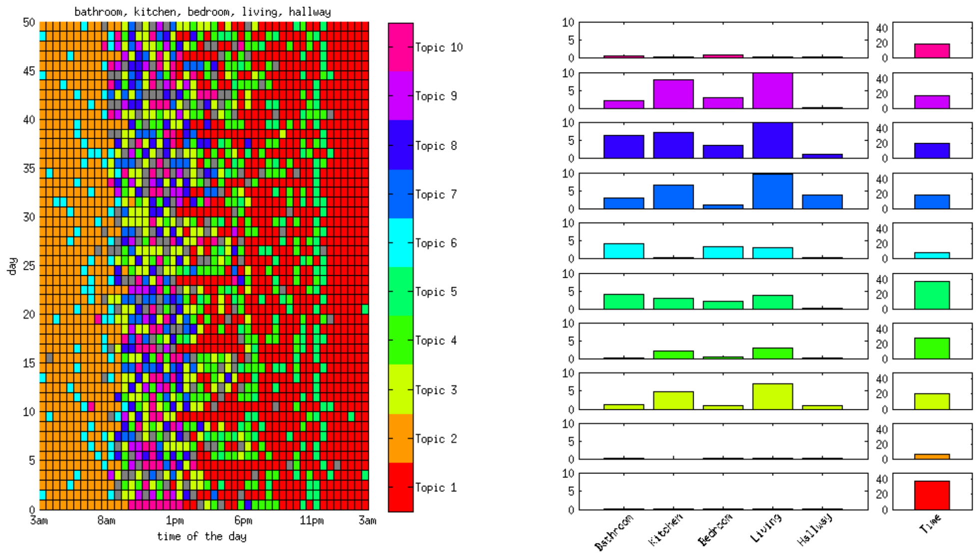
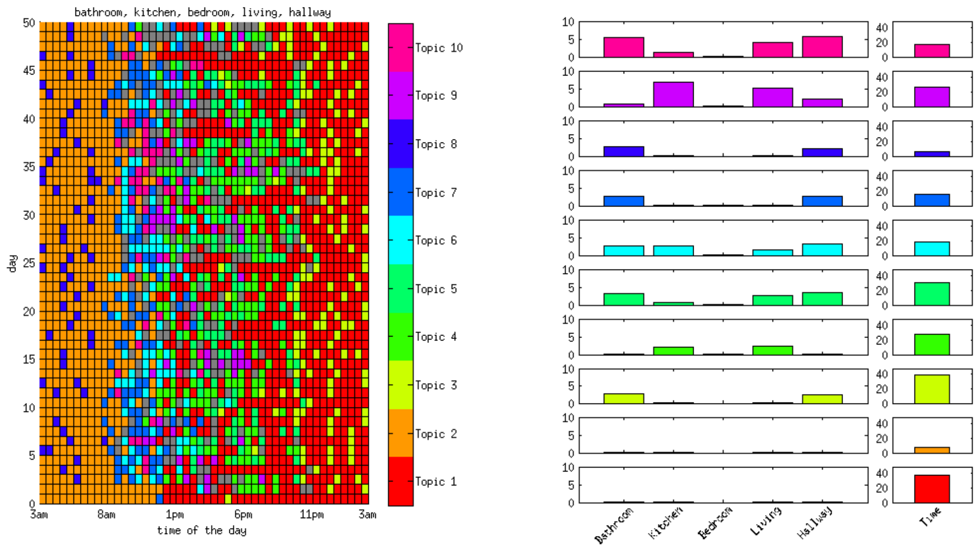


Figure B.3: House number 3, 20 topics, fine grain time, LDA-Poisson



a) b)

Figure B.4: House number 4, 20 topics, fine grain time, LDA-Poisson



a) b)

Figure B.5: House number 5, 20 topics, fine grain time, LDA-Poisson