Learning the Unknown
A Computational Approach to Incrementally Developing a Hypothesis Space with Gaussian Mixture Models

Master's Thesis in Artificial Intelligence by

Erwin de Wolff
s4244907

Supervised by

Johan Kwisthout[1]

[1] Donders Institute for Brain, Cognition and Behaviour

Department of Psychology and Artificial Intelligence
Radboud University Nijmegen
June 2, 2020

Abstract

We present a new computational model that captures a key behaviour found in humans we wish to replicate, namely the ignoring of singular anomalous stimuli. Our model is built around incremental learning in Gaussian mixture models combined with an explicit 'unknown' hypothesis with posterior reasoning. We show through simulations that our model indeed captures the behaviour of interest as opposed to IGMM, the model it was based on. We also show that our model improves the performance of IGMM, providing theoretical evidence that ignoring anomalies is a superior method of learning as opposed to fitting all the data.

1 Introduction

How do infants build a complex model of the world? This question is at the heart of developmental cognitive science. How do we go from a confusing stream of sensory input to a well-structured model of beliefs that we use as adults? The problems any cognitive agent needs to overcome are truly difficult. They need to learn how to segment the visual input into distinct shapes, and attribute meaning to them. They need to learn to segment sound into distinct groups. A further challenge comes in the form of language learning. Before any sense of semantics and grammar can be learned, the agent needs to have solved what the distinct sounds, or phonemes, of the language are, as well as how different sounds are used together to form words.

What all these problems share is that the agent needs to assign meaning to sensory input without the availability of a broader belief system to explain them with. This means that they have to develop a belief system from the ground up based on regularities in these observations (for more on these types of problems, see Quine (1974)). Finding out what principles underlie this process of development is very much an open question, and a fascinating topic of research. In this thesis we will examine the early development of a hypothesis space through computational modelling. The computational model we will be using is a Gaussian mixture model, that needs to develop a set of hypotheses (or clusters) based on 'fitting' incoming data. We describe this more at the end of this section.

Computational models allow us to think more clearly about theories about cognition and the brain. Because verbal theories use language it is often unavoidable to create misunderstandings and confusion, however minor, between different expert groups. Because formal models require to be explicitly defined in all regards, the ambiguity that is inherent of language is resolved, leading to a better shared understanding of the theory. During the process of formalising a theory, researchers often find gaps in existing theory, areas where the original researchers made problematic assumptions or simply did not see that an additional step was needed. The process of filling these gaps can lead to new developments in theories about cognition.

The use of computational modelling for cognitive research is not new. Computational models that aim to capture aspects of cognition are easy to find, be it word segmentation (Perruchet & Vinter 1998), contradictory beliefs (Thagard 1989), decision making (Gigerenzer & Todd 1999) and many more. More recently, a common way to think about cognition is as (approximate) Bayesian inference (L Griffiths et al. 2008, Lee 2011). This idea has since been used to explain how infants learn words (Xu & Tenenbaum 2007), perform action-understanding (Baker et al. 2006), do visual processing (Kok et al. 2013, Lee & Mumford 2003) and even to explain cognition as a whole (Clark 2013, Friston 2010).

Many of these cognitive theories make use of machine learning models that originated in data-science.[1] This is not surprising, as data-science has been a very rich and active field in recent years. In turn, cognitive science has also inspired the development of new learning models (Lotter et al. 2016), further showcasing how these two fields can be used to inspire one another.

Researchers should, however, take care that the data-scientific learning models are not blindly taken as biologically plausible. Algorithms in data-science are designed to solve a problem as best they can, usually by learning to 'fit' the data they were given, rather than explain (human) cognition. They are not, by default, explanatory models. Because of the different demands that go into designing the models, the models often behave differently, subtly or overtly, from humans. Success in solving a problem should not be equated to success in modelling how humans solve the problem.

This research look at one of these discrepancies between what is 'best' in terms of solving a problem versus what is most 'human'. The central feature of interest is the property of humans to ignore unexpected stimuli that appear infrequently, which we will refer to as (singular) anomalous observations. Current methods of learning in Gaussian mixture models do not account for this feature of cognition, as we will show in the Previous Work section.

Romberg & Saffran (2013) performed an experiment where infants had to make predictive saccades to a stimulus which would appear on either the left or right side of the screen. When the stimuli were always presented on the left side of the screen, a single stimulus on the right side did not influence the infants' looking behaviour much. This suggests, the authors claimed, that "*...a single novel instance did not strongly shift infants' expectations given their highly consistent global context.*"(Romberg & Saffran 2013). In other words, infants did not seem willing to create a new hypothesis upon making a single anomalous observation.

In another saccadic-planning study by Kayhan et al. (2019), infants ignored unexpected stimuli that did not add to their knowledge structure. In this experiment, the colour of an unexpected stimulus carried information about the location of future stimuli. When the unexpected stimulus had the wrong colour, infants would not change their predicting saccades. In other words, they did not change or add new hypotheses based on 'noisy' observations. When the unexpected stimulus did have the right colour, there was a change

[1] Data-science is also called Artificial Intelligence, but this thesis chose the former to make clear that the models are not specifically created for any AI-inspired goal, but rather a data-fitting goal.

in their predictions. Although it could be argued that the infants learned the effect of each colour on future locations of stimuli, a less complex solution is that the infants simply ignored single outliers, but were responsive to unexpected, persistent patterns. In other words, although they did develop a hypothesis for what happened after the right colour cue, they had no hypothesis about what happened after the wrong colour cue.

So, in order to better capture actual human behaviour, models that learn a hypothesis space should be less sensitive to outlying observations than might be expected. That is, unless the agent is faced with a consistent series of unlikely observations, no new hypothesis should be formed. In this thesis, we propose a computational model that can replicate the behaviour of ignoring anomalies. To specify the problem as best we can, we will outline the explicit assumptions under which we operate.

1. New hypotheses are generated through made observations only, not through 'teaching' or using labels. As such, we are excluding direct teaching and 'changes of mind' that might occur from conclusions drawn in another part of a belief network. These methods of learning are certainly important and present in human learning, but are not addressed in the scope of this thesis.

2. Observations can only be made at the perceptive level. Beliefs at the higher level can only ever be inferred in a bottom-up fashion. Note that these higher levels can still influence the agent in a top-down fashion as well. A simple example of this is that a human cannot 'see' a cat. Rather, the sensory input of the eye is processed bottom-up to allow us to conclude that there is a cat. However, knowing that I am at my friend's house who owns a cat (top-down information), even a glimpse of a furry texture could allow me to conclude that I see a cat.

3. Agents only have access to a single observations at each learning step. That is, the agent can not look back at observations past or look ahead at observations to come. From this limited input, the agent needs to develop its belief model incrementally. This assumption is in fact more strict than would be the case for humans. Of course humans do not have perfect, eternal memory, as would be needed for an agent to have full access to the entirety of all past observations. They do have, however, a short-term memory of more than one previous observation. In this thesis, we chose not to model the impact of short-term memory, and instead focus on the agent processing the current observation only.

4. New hypotheses are automatically linked to the rest of the belief model in a clearly defined manner. In other words, there is no distinction made between discovery of a new concept and causal learning. Whenever a new hypothesis is discovered, it is assumed to be independent of all other hypotheses about the world.

The computational model the we will be using is a Gaussian Mixture Model (GMM). GMMs are a popular framework for clustering algorithms. Clustering, also known as unsupervised learning, is the name for any learning problem where the label or 'class' of an observation is not available to the agent, which matches our assumption 2. GMMs cluster data by learning the parameters of a set of Gaussian functions. Each Gaussian is its own cluster, which we will refer to as a hypothesis.

Because of assumption 3, the agent will need to learn the Gaussian functions through one observation at a time. This is known as incremental or online learning. There have been several ways of learning GMMs incrementally proposed in previous research (see Previous Work section). We will base our research on one of these existing approaches, the IGMM model by Engel & Heinen (2010).

Current state-of-the-art approaches to incrementally learning GMM's do not capture the selective ignoring of singular anomalous observations. The goal of this thesis will be to propose a new model we call PRIGMM that can replicate this feature of cognition. We do so by explicitly representing an 'unknown' hypothesis in the agent's model. The agent uses this hypothesis to see whether its belief model needs expanding through Bayesian hypothesis testing (details in Methods). The central question we will answer is:

> Does the addition of the explicit 'unknown' hypothesis to the hypothesis space allow agents to ignore uninformative anomalous observations?

We want to make clear that the goal is not related to best 'fit' a set of data. The model we are proposing is an explanatory model. Furthermore, this research is exploratory in nature. There are undoubtedly many ways to model ignorance of anomalous observations. Rather, we hope that this thesis can be used as a stepping stone for future research, be that in development psychology, cognitive science or data-science.

If the answer to the main research question is yes, we will also look at a two more aspects of the model. The first of these investigates the effect of time on the ability to learn a new hypothesis.

> How does the amount of previously acquired observations influence the chance that an anomalous observation is ignored?

The expectation is that the more observations an agent has already made, the less likely it will be for it to learn a new hypothesis. In other words, a certain rigidity of beliefs develops.

The second has to do with the representational strength of our approach. A danger of any approach to learning that is designed to inhibit the response to certain stimuli is that it does not manage to learn all the relevant hypotheses that are present in the data, because of said inhibition. To see whether this is also the case in this thesis, we will also answer the following question.

> Does the agent still learn all relevant hypotheses if it ignores anomalous observations?

To answer this last question, we compared the Bayesian-information criterion (BIC for short, see Methods section for the definition) between the learned hypotheses and the ‘true’ generating process. Success is mostly dependent on a small difference in the BIC values between the true model and our learned model, which translates to the agent having learned all relevant hypotheses.

The remainder of this thesis is set up as follows. In Previous work, we review the basics of learning the parameters of a Gausian mixture model, as well a review newer approaches to learning incrementally. Then in Methods we explain our the methodology of how we created our proposed model, both at the conceptual and formal level, and describe a set of simulation tests to test our research questions. Lastly, we present the results and interpretation, and end with the conclusion and an overview of future directions of research.

2 Previous Work

The field of unsupervised learning or clustering has had much research devoted to it. Clustering describes the general process of grouping data together into ’clusters’ through some metric of similarity. This metric can be distance between points, similarity in shape etc. In literature, multiple models exist to learn this grouping of data, including hierarchical models (Johnson 1967, Reddy & Vinzamuri 2018), centroid models (Arora et al. 2016, Taillard 2003) and density models (Ester et al. 1996, Schubert et al. 2017).

This thesis will looks at yet another model of clustering called distribution models, where the clustering of a data set is learned through a linear combination of so called mixture components. One reason to use distribution models to cluster is because they are generative models, which means that the models are, in principle, capable of generating data that is similar to the data they used to learn. As a model of cognition, this property is very desirable, as it allows agents to make predictions about future observations (i.e. data).

As mentioned in the introduction, the mixture components used in this thesis are (multivariate) Gaussian distributions. Distribution models that use Gaussian functions are called Gaussian Mixture Models (GMMs). A GMM can be written as the triple of vectors.

$$M = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

where $\boldsymbol{\pi}$ is a discrete probability distribution over the Gaussians, $\boldsymbol{\mu}$ is the vector containing the means for each Gaussian, and $\boldsymbol{\Sigma}$ is the vector containing the covariance matrix for each Gaussian.

Because these models are inherently probabilistic, they can be easily used to assess the likelihood of a given set of observations X. For any mixture model, the likelihood of this set is given by:

$$P(X|\boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_i \sum_j \pi_j L(x_i|\theta_j) \tag{2}$$

where $\boldsymbol{\theta}$ are the parameters of the mixture components. In the case of a Gaussian mixture model, this translates to:

$$P(X|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_i \sum_j \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j) \tag{3}$$

When a full dataset is readily available (unlike in this thesis, see assumption 3), one popular method of learning the hypothesis space is through an algorithm called Expectation-maximization (EM) (Dempster et al. 1977). EM uses a iterative two-step approach to find the maximum likelihood estimate of the parameters for each cluster. Initially, each mean and covariance matrix is chosen randomly.[2] Then, each observation (or data point) is assigned to all clusters, weighted based on their relative likelihood. This means that each observation is assigned to multiple clusters, a method called 'soft' clustering. This is opposed to 'hard' clustering, where each observation is given to a single cluster exclusively. This soft assignment of each observation to the clusters is called the *expectation* step.

The next step is the *maximization* step, where the new mean and covariance matrix for each cluster k are determined. For the mean this update is given by:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} w_{k_n} o_n \tag{4}$$

where o_n is the n-th observation in the dataset, w_{k_n} is the probability that observation n belongs to cluster k, and $N_k = \sum_{n=1}^{N} w_{k_n}$. The value w_{k_n} can represent either the relative likelihood or the posterior probability, where for the latter a prior probability is assigned to each cluster.

The new covariance matrix of cluster k is determined using:

$$\Sigma_{k_{i,j}} = \frac{1}{N_k} \sum_{n=1}^{N} w_{k_n} (\mu_i - o_{n_i})(\mu_j - o_{n_j}) \tag{5}$$

Here, Σ_k has dimensions equal to $N \times N$.

After the maximization step, we can repeat the cycle of expectation and maximization until the clusters converge on some parameter values. Then, these are the maximum likelihood estimates of the parameters for each cluster.[3]

We note again that the EM-algorithm requires constant access to the entire set of observations in order to estimate the most likely parameters for each cluster. This contradicts our assumption 3 (see Introduction). As such, EM can not be used in its basic form to learn

[2] A common choice is to only choose random means, and have each covariance matrix be given by $\Sigma = \lambda I$ for some scalar λ.

[3] Another method of approximating parameters would be to determine a distribution of the parameters of each Gaussian instead of a fixed-point estimate. These models are not explored in this thesis.

incrementally. We will now discuss alternative clustering methods that can do so, as found in literature. The papers mentioned are not exhaustive, but do provide a good overview of the approaches found in data-science to learn GMM's incrementally in recent years. This thesis did not look at Bayesian non-parametrics, which is a related field of data-science that aims to find a distribution over hypothesis parameters, rather than fixed point estimates as in the following research.

2.1 Incremental Learning

Song & Wang (2005) published an approach where the EM algorithm was used on smaller sets of data, so-called batches. In their approach, an agent would have some model $g^N(x)$ existing, which is a mixture model learned from all previously seen observations, as well as a small set of new observations $x_{N+1}...x_{N+M}$. These new observations are clustered using EM, where the number of hypotheses (i.e. components) is determined by the Bayesian Information Criterion. Now, the two different hypotheses sets are combined. During this combination step, all new hypotheses are compared to their most similar previously existing hypothesis. If they are similar enough, they are merged. Else, they are added to the model separately.

Kristan et al. (2008) published an approach where a GMM was incrementally learned through a method of bandwidth estimation and compression. After each new observation, a new hypothesis is added to the model through an approximation function that attempts to find the best parameters. In addition, the model has a set number of hypotheses that are allowed. If there are more hypotheses present in the model, the model is compressed until the model no longer exceeds the threshold. Their compression algorithm uses a concept of ‘unlearning‘, which defines how a model can deal with so-called ‘negative examples‘.

Declercq & Piater (2008) proposed a different version of updating the hypothesis space after each new observation. Their approach emphasized a trade-off between the number of hypotheses and how well they represent previously made observations. To do so, they created a hierarchical system, where each hypothesis at the highest level is itself a mixture of lower hypotheses. Their design controlled the trade-off by combining very specific hypotheses at the lowest level into ‘non-overfitting‘ hypotheses at the higher level.

Unlike the previously mentioned approaches, where each observation was in principle its own hypothesis unless merged with an already existing hypothesis, Engel & Heinen (2010) slowly built up the number of hypotheses. In their approach, called IGMM, each new observation x was checked for novelty. This was done by comparing the density height at the observation x with the density height at the mean, for each Gaussian in the mixture model. If the ratio of the former divided by the latter was below some (pre-defined) parameter τ, the observation x was novel for that hypothesis. If x was novel for all hypotheses, a new hypothesis was added to the model. If no new hypothesis were created, each hypothesis was updated with x based on their relative likelihood.

All these approaches either create a new hypothesis for each observation (possibly to be merged later), or use the likelihood to determine whether a new hypothesis is warranted. They have no explicit design to encourage outliers to be ignored. To represent singular points with its own hypothesis is a case of 'overfitting', where the models lose generality in favour of fitting the training data better. While there are checks in place in some of these papers to avoid overfitting, the particular problem of deciding whether an outlier is just an outlier or the first of a new hypothesis that should be represented is not a main concern. Suppose, for example, that an outlier is observed. If the agent decides to create a new hypothesis, it may have made the mistake of overfitting. However, if it holds of on making a new hypothesis, and instead incorporates it into an existing one, new outliers of that same group might also be merged with that same hypothesis. This would be an example of underfitting, failing to be specific enough to disambiguate the data. What we want, therefore, is a way to allow the agent to 'change their mind' about whether an outlier was just that or not. The previous work mentioned here does not have the tools to allow agents to do so. We will now look at our proposed model, PRIGMM, that aims to capture this behaviour.

3 Methods

We will now describe our proposed model. This model, we call PRIGMM (short for Posterior-Reasoning Incremental Gaussian Mixture Model), is based in the IGMM model by Engel & Heinen (2010). In that model, each hypothesis has its own distinct novelty value, which is relative to the mode of that hypothesis. The reason we chose IGMM as the basis for our own model is because the other approaches all assume that each observation forms its own hypothesis unless it fits well in a previously established hypothesis. This is a poor basis for any model that can conditionally ignore anomalous observations, which is the behaviour observed in developmental psychology we aim to capture here. IGMM, on the other hand, gradually expands its hypothesis space if a certain criterion is met. By changing the nature of this criterion, we can allow the agent to ignore anomalies in a cognitively plausible manner.

We will first rewrite IGMM in such a way that the criterion can be explained as Bayesian hypothesis testing within pairs of hypotheses. Next, we suggest a new variation of the model, where an agent explicitly represents an 'unknown' hypothesis in its hypothesis space.

3.1 PRIGMM

We start by noting that the IGMM model can be rewritten to allow for direct comparison between hypothesis pairs. To start, we are going to rewrite the definition of the IGMM to make the next steps. First, we change each hypothesis h in the IGMM model to a pair of hypotheses (h_k, h_u), representing 'known' and 'unknown' respectively (see figures 1 and 2). We let the likelihood $P(x|h_k)$ be described by the original Gaussian function of h, and let the likelihood of $P(x|h_u)$ be represented by $\tau \times Mo(h_k)$, where τ is a fixed hyperparameter

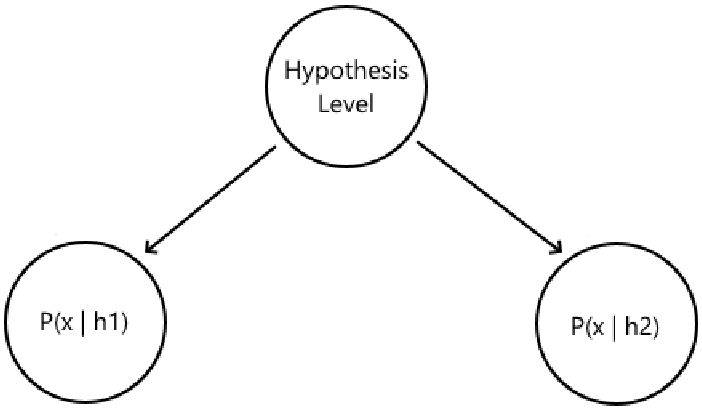

Figure 1: *The original design as in Engel & Heinen (2010). Here, the hypothesis space consists of two hypotheses, which are (multivariate) Gaussians.*

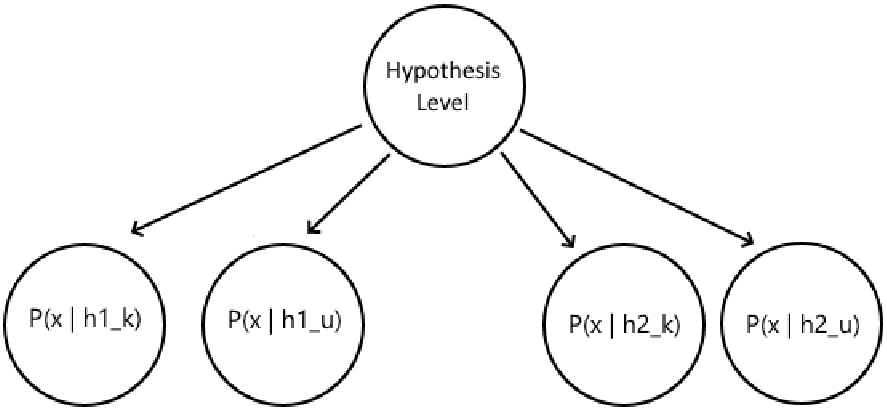

Figure 2: *The Rewritten IGMM model, where each hypothesis is turned into a pair of hypotheses, h_k and h_u. This model is functionally equivalent to that in figure 1, given equal prior weights to both hypotheses in a pair.*

and $Mo(h_k)$ is the mode of the density function $P(x|h_k)$. This likelihood value for h_u is equivalent to the value Engel & Heinen (2010) used as a novelty criterion. Because of this, any time the likelihood of h_u is bigger than that of h_k, the observation would be novel in the IGMM model as well. As such, we can change perspective from reasoning about criterion rules to paired-hypothesis testing, without any change in behaviour between models.

For the next step, we break up the hypothesis pairs, and merge all instances of h_u into a single 'unknown' hypothesis h_u. See figure 3 for a visualization of this step. This allows the agent to start thinking about hypothesis testing in general rather than in pairs. Either one of the known hypotheses is most likely, or the unknown hypothesis is. If the unknown hypothesis h_u is most likely, the agent will add a new hypothesis to the model.

Previously we defined the likelihood of each h_u by using the mode of its paired hypothesis

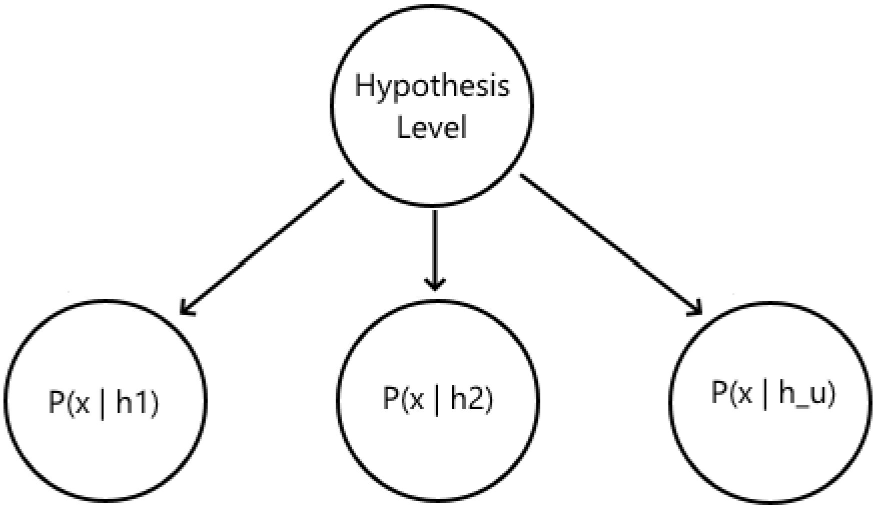

Figure 3: *The variation proposed in this thesis. As opposed to the rewritten version depicted in figure 2, this model has a single hypothesis h_u representing the unknown. The prior over this hypothesis is subject to Bayesian updating exactly like the other hypotheses in the model.*

h_k. To allow a similar metric for this new, general h_u, we look at the expected mode instead, which is defined as:

$$Mo_{expected}(H) = \frac{\sum_{h=1}^{H_k} P(h)Mo(h)}{\sum_{h=1}^{H_k} P(h)} \tag{6}$$

where H_k is the hypothesis subset of H containing all hypotheses except h_u, and $Mo(h)$ is once again the mode of h.

We can now take the same fixed parameter τ and multiply this with the expected mode, to get the likelihood for our unknown hypothesis:

$$P(x|h_u) = \tau \times Mo_{expected}(H) \tag{7}$$

As this likelihood function is a measure of the modes of all known hypotheses, is can be quite off for any single hypothesis. However, because the prior probability of each hypothesis is taken into account, this weighted sum of modes should provide a reasonable approximation of the overall mode of the landscape.

So far, the model has only relied on the highest likelihood to determine whether a new hypothesis should be added to the hypothesis space. The problem with this approach is that any time this likelihood is highest, a new hypothesis is formed. To counter this, we make our final change to the model. Instead of adding a new hypothesis to the model whenever the likelihood of h_u is highest, we demand that the posterior probability of h_u is highest. To find the posterior probability, we make use of Bayes' rule, commonly written as:

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} \tag{8}$$

where H is the hypothesis and E is the evidence. In our case, H consists of all our known hypotheses h_k along with the unknown hypotheses h_k, and E is the observation x.

Now, the posterior $P(H|x)$ can be calculated per equation 8. Each observation now has two possible outcomes. Either a new hypothesis is added to the model, or the parameters of all existing hypotheses are updated. If no new hypothesis is added, the parameters of the model ($\boldsymbol{\pi}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) are updated. The recursive equations for this are taken from Engel & Heinen (2010) with adjusted notation.

$$\boldsymbol{\alpha}^{t+1} = \boldsymbol{\alpha}^{t} + P(\boldsymbol{\pi}^{t}|x) \tag{9}$$

$$\pi_i^{t+1} = \frac{\alpha_i^{t+1}}{\sum_j \alpha_j^{t+1}} \tag{10}$$

$$\mu_i^{t+1} = \mu_i^{t} + \frac{P(i|x)}{\alpha_i^{t+1}}(x - \mu_i^{t}) \tag{11}$$

$$\Sigma_i^{t+1} = \Sigma_i^{t} - (\mu_i^{t+1} - \mu_i^{t})(\mu_i^{t+1} - \mu_i^{t})^T + \frac{P(i|x)}{\alpha_i^{t+1}}((x - \mu_i^{t+1})(x - \mu_i^{t+1})^T - \Sigma_i^{t}) \tag{12}$$

where the superscript indicated the number of observations, and the subscript indicates the ith component from the model. Note that the weight of h_u is updated just like all other hypotheses. This means that the posterior probability of the unknown hypothesis can become bigger and bigger over multiple observations.

In the case that the posterior probability is highest for h_u, a new hypothesis is added to the model instead. In that case, the equations above are not used. Instead, a new component h_{i+1} is added to the model according to:

$$\alpha_{i+1} = \alpha_{h_u} \tag{13}$$

$$\mu_{i+1} = x \tag{14}$$

$$\Sigma_{i+1} = \Sigma^0 \tag{15}$$

where α_{h_u} is the original weight of the unknown hypothesis, x is the observation that led to the new hypothesis, and Σ^0 is the default prior for the covariance matrix, which should be set by the modeller. When a new hypothesis is added to the model, it replaces the unknown hypothesis. That is, the new hypothesis h_{i+1} is given the weight of h_u, denoted by α_u. Then, a new h_u is added with initial $\alpha_u = 1$.

At the beginning of the learning task, the agent starts with an empty model. The first observation x it makes becomes the first known hypothesis with mean $\mu = x$ and $\Sigma = \Sigma^0$. It then also adds h_u. Both parameters in $\boldsymbol{\alpha}^1$ start at 1.

By making these changes the agent is able to ignore single anomalous observations for which the likelihood of h_u is highest if the prior over h_u is small. In other words, if most observations are well explained by the known hypotheses, a single outlier may be ignored safely without much risk. This does however, depend on the scale of the outlier. It could happen that the likelihood for h_u is so much bigger than for any other hypothesis that the posterior probability of h_u is highest despite a small prior. Taken together, this means that outliers can be ignored by the agent unless they are too surprising.

Conversely, if many subsequent observations are well, but not best, explained by h_u, the agent could conclude that a new hypothesis is best. By repeated observations in the tail-end of known hypotheses, the agent becomes more and more convinced that a better model can explain the data, even though no single observation was best explained by ignorance.

Lastly, it should be noted that because h_u starts with a small prior after a new hypothesis is added, an agent will generally be less eager to form new hypotheses after more observations have been made. This is because as the overall number of observations n grows, the ratio of $\frac{1}{n}$ becomes smaller and smaller. This can be thought of as a sort of 'rigidity of beliefs' that slowly develops.

3.2 Simulation

To see whether the proposed variation could correctly ignore anomalous stimuli, we ran a series of computer simulations. Two models were tested, the IGMM model and the PRIGMM model. We included the IGMM model to see whether ignored outliers are ignored as a result of our modifications, or whether they were also not treated as 'novel' by the IGMM model. In addition, the IGMM model provided a good baseline to compare the overall quality of learning with, as relates to our third research question.

We created three similar sets of observations in a 2-dimensional grid. This data set was created by pseudo-random draws from five different multivariate Gaussians, which can be seen in figure 4. The purple group in the top right was the so called 'outlier group'. We tested different versions of the data set where this outlier group was empty, contained 1 outlier, 3 outliers or 10 outliers. During learning, the agent always saw the outliers in immediate succession.

In addition to the number of outliers, we also varied the time of their occurrence. This was done to test how the property that a larger amount of previous observations lowers the prior of the 'unknown' hypothesis would affect the ability to ignore outliers. As such, we also tested how the agent responded to the outliers group after 9 different moments. These moments were when the agent had already learned 10%, 20%, 30% ... up to 100% of the other observations.

We ran each configuration of the simulation a 100 times with randomized draws from the Gaussians that generated the data set. We recorded what percentage of these 100 simulations the agent made a new hypothesis to explain the outlier group. In addition, we recorded the average Bayesian information criterion (BIC) for four three models: two fixed models, IGMM and PRIGMM. The BIC is a criterion for model-selection formulated by Schwarz et al. (1978). It provides a trade-off between the 'fit' of a model on data, and the amount of parameters

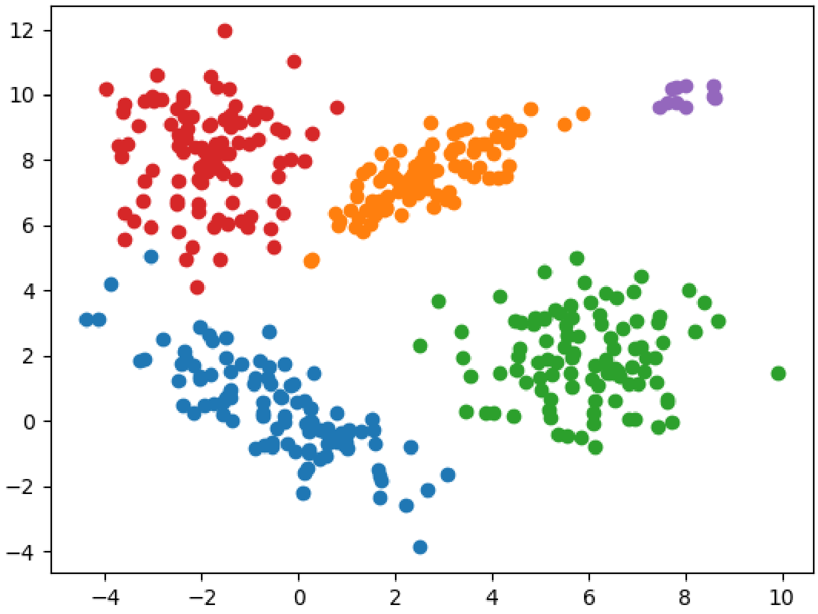

Figure 4: *The two-dimensional dataset, colour-coded to show the group each data point belongs to. The outlier group is denoted in purple in the top right corner of the image. Whether the agent detects this group should depend on the number of consecutive outliers, as well as the amount of previously made observations.*

is requires to do so, and is given by:

$$BIC = k \ln(n) - 2 \ln(\hat{L}) \tag{16}$$

where k is the number of parameters in the model, n is the total number of observations made, and $\hat{L}$ is the best likelihood estimate. In our case, $\hat{L}$ is the likelihood from the final model learned by the agent, and k is the number of hypotheses.

We expect that the IGMM model will always make a new hypothesis for the outlier cluster, regardless of the amount of outliers (with the obvious exception for zero outliers). By contrast, we expect that our approach will be able to ignore outliers. We hypothesize is that there is a linear relation between the number of previously made observations and the number of outliers required to form a new hypothesis.

3.2.1 Implementation & Parameters

We used Python 3.7 to implement both approaches, using the numpy, scipy and matplotlib packages. For the IGMM model, we used their suggested parameters $\tau = 0.01$ and $\Sigma^0 = 7.5I$, where I is the identity matrix. These parameters were based on some preliminary empirical testing.

To generate the data, four main hypotheses were used along with a single outlier hypothesis, with the following parameters:

Hypothesis	**Mean**	**Covariance**
Main 1	[0, 0]	[[2, -1.5], [-1.5, 2]]
Main 2	[2.5, 7.5]	[[1, 0.75], [0.75, 1]]
Main 3	[6, 2]	[[2,0], [0, 2]]
Main 4	[-2, 8]	[[1, 0], [0, 2]]
Outlier	[8, 10]	[[0.1, 0], [0, 0.1]]

Table 1: *Parameters used to generate the training data shown to the agent. For the testing data to find BIC score, only the four main hypotheses were sampled.*

4 Results

We will now summarize the results of the simulations. All the data shown in this section is based on the means of the data across all simulations. The standard deviations can be found in the tables in the appendix. We chose to exclude the standard deviations from the graphs because of their size: including them severely reduced to quality of the graphs.

4.1 Ignoring Outliers

The agent using PRIGMM successfully managed to ignore anomalous observations in the simulations. This led to a notable difference in the probability that the outlier group would cause a new hypothesis between IGMM and PRIGMM. This difference can be seen in figure 5.

Both IGMM and PRIGMM had an increased chance of ignoring outliers after more observations had been made by the agent. However, the PRIGMM model had a much higher chance of doing so. In addition, the curve for PRIGMM is sharper, suggesting a separate effect of the amount of observations made before outliers on the chance to ignore those outliers. Lastly it should be noted that, across the simulations, the single outlier was less often ignored than the three outliers, which goes against expectations.

The fact that the IGMM model did not always create a new hypothesis due to the outlier group means that these outliers were not always novel to each hypotheses at that time. In other words, at least one hypothesis was close enough to the outlier that it was not longer an outlier. We can expect that a similar part of the PRIGMM chance to ignore outliers came from the same circumstances. However, the much increased chance of ignoring the

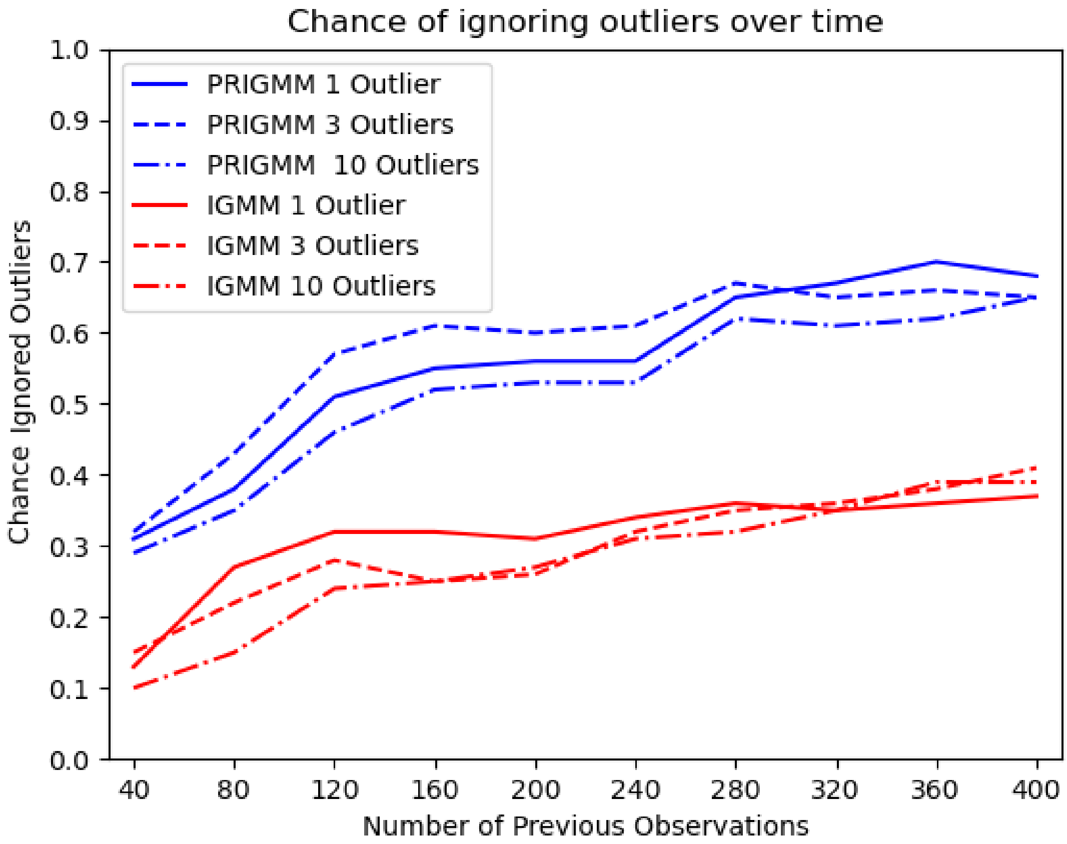

Figure 5: *The chance that the outliers group did not create a new hypothesis given the amount of previously made observations. The upper three lines in blue show the results for the PRIGMM model, and the lower three lines in red those of the IGMM model.*

outlier group suggests an additional effect of the posterior reasoning as opposed to likelihood reasoning.

Because the PRIGMM model ignored outliers more often, it also ended up with fewer hypotheses than the IGMM model, as can be seen in figure 6. PRIGMM learned an average of 3 hypotheses with no outliers to 3.7 with 10 outliers, whereas IGMM learned an average of 6 hypotheses with no outliers to 6.7 with 10 outliers.

4.2 Quality of Learning

In addition to the main two models IGMM and PRIGMM, two fixed models were tested. These models were predefined and acted as a 'gold standard' to compare the learning models with. The first fixed model, the Fixed Base model, consisted of four hypotheses. These were the exact hypotheses used to generate the data for the training set, excluding the outlier

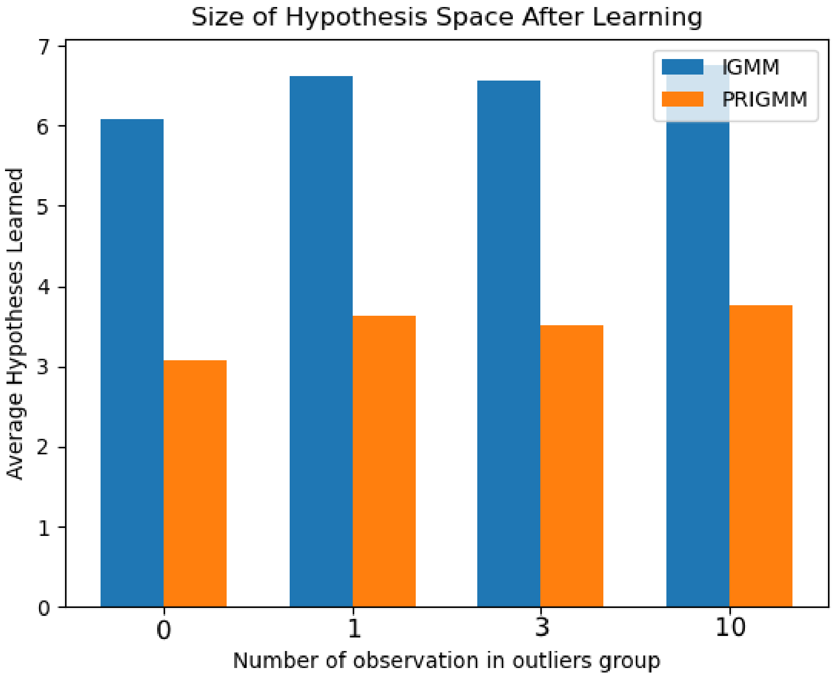

Figure 6: *The average amount of hypothesis after the training phase. The IGMM model is shown in blue on the left, and the PRIGMM in orange on the right.*

hypothesis. The Fixed Outlier model also had these four hypotheses, but also had the outlier hypothesis as an additional hypothesis.

Because these fixed models could not acquire any new hypotheses through learning, the time at which the outliers were shown was irrelevant for their BIC score. Thus, the BIC was recorded once for each condition of 0, 1, 3 and 10 outliers, each averaged over 100 trials. The results can be seen in figure 7.

The first obvious observation is that including the outlier group severely worsened the BIC score compared to the Fixed Base model. This is most probably because the very small covariance values in the outlier hypothesis. This hypothesis would effectively give a likelihood of zero to all data points not directly in the outlier group, which would drastically reduce the likelihood on the whole set of observations. This is severely punished by BIC.

The second thing to note is that the BIC score for the PRIGMM was lowest for all sizes of the outlier group, even lower than the Fixed Base, which should act as golden standard. This could be explained by the nature of BIC: it is a trade-off between the likelihood and the number of parameters used to find it. A lower BIC suggest an advantageous trade-off between these two in terms of the score. A possible example of how PRIGMM did this is shown later in this section.

To see what the effect was of the amount of observations made by the agent before the

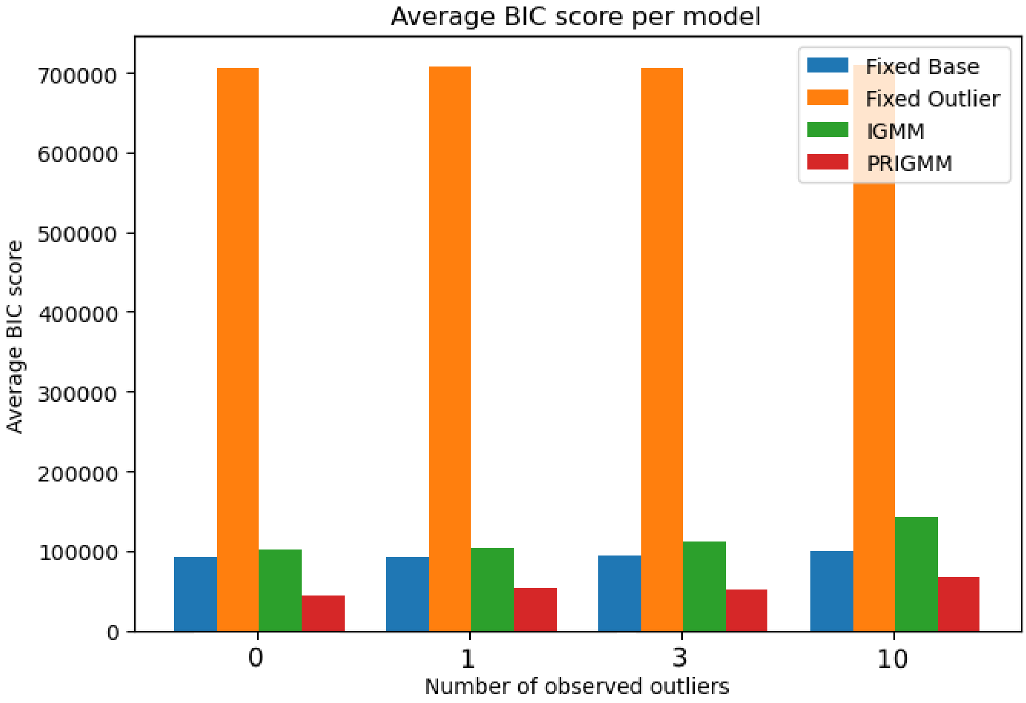

Figure 7: *Average BIC per tested model. The fixed base model uses a static model consisting of the four main hypotheses. The fixed outlier adds the outlier group as a known hypothesis as well. For the IGMM and PRIGMM models, the BIC scores were averaged across all simulations. Results per group from left to right are: fixed base, fixed outlier, IGMM and PRIGMM.*

outlier group was shown, we plotted the BIC score against the this variable. This can be seen in figure 8. Exactly when the outliers were seen did not seem to affect the BIC score for either model if the outlier group consisted of 1 or 3 outliers. When there were 10 outliers in the outliers group however, the BIC score improved as more observations were made by the agent before confronted with those outliers.

4.3 Example of Learned Hypothesis Space

To see what sort of hypotheses were learned by the agents, we plotted examples of the hypothesis spaces learned by IGMM and PRIGMM. We applied the learned model to the data set shown in figure 4. Using the learned hypotheses, each point in that new set was 'classified' as belonging to a certain hypothesis, which we all gave a unique colour. The results can be seen in figure 9 for the IGMM model and figures 10 and 11 for the PRIGMM model. The figures have been colour-coded such that uncertain observations are a mix of the colours of the possible clusters. In addition, a black cross has been placed at the mean of each Gaussian, so signify the 'prototypical' point for that hypothesis. As can be seen, the

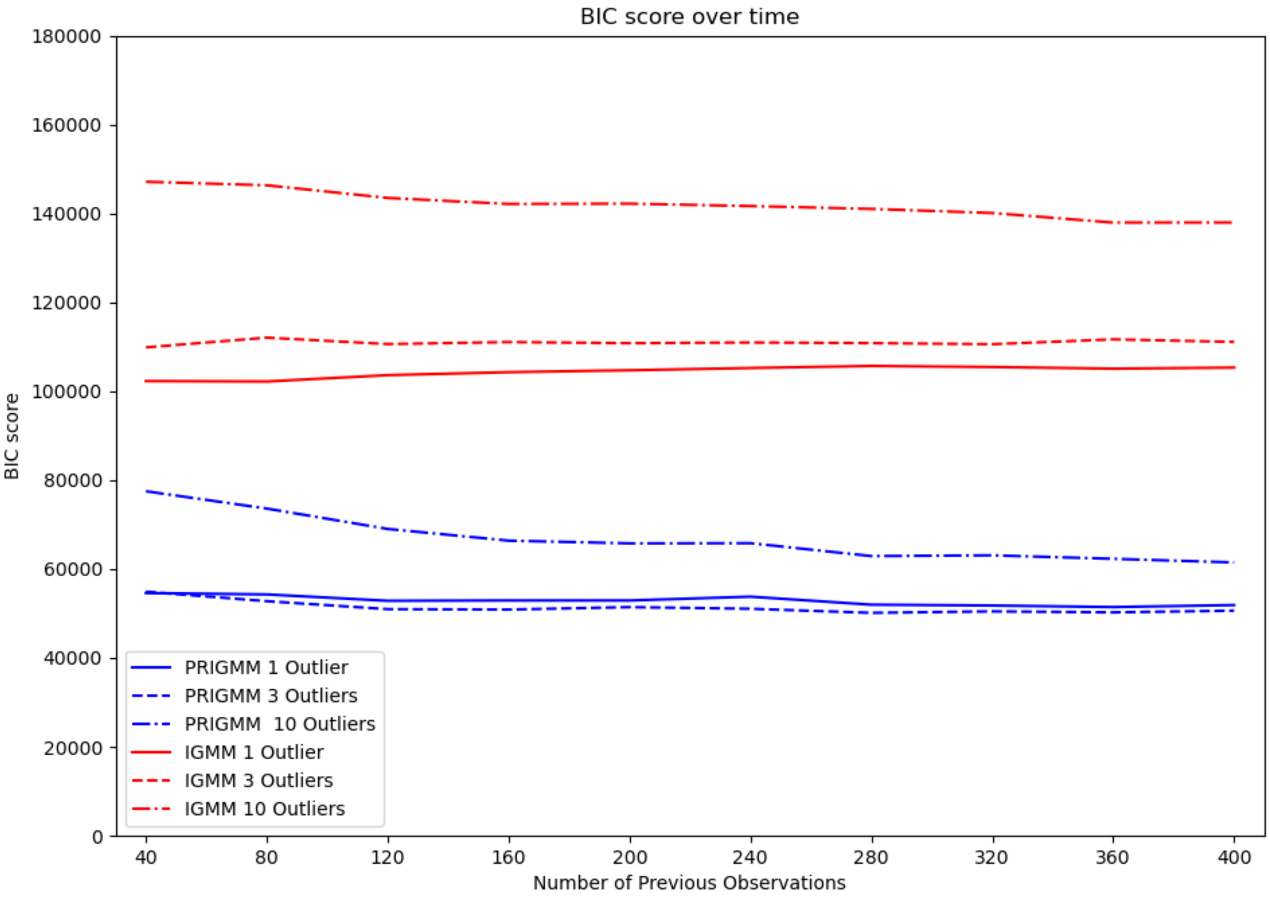

Figure 8: *Average BIC relative to the number of observations made before the outlier group was shown. The upper three lines in red show the results for the IGMM model, and the lower three lines in blue those of the PRIGMM model.*

IGMM model resulted in 7 hypotheses as opposed to the 5 and 3 of the PRIGMM model. This is a typical result throughout the simulations. Figure 11 also shows how the BIC score can become very low for this set of data.

5 Discussion

The PRIGMM model successfully managed to ignore outliers, which led to much smaller hypothesis spaces that those found in the IGMM model. As such, our proposed model was successful in capturing this aspect of human cognition. It should be noted that the IGMM model also ignored the outlier group on occasion. The only explanation for this is that the

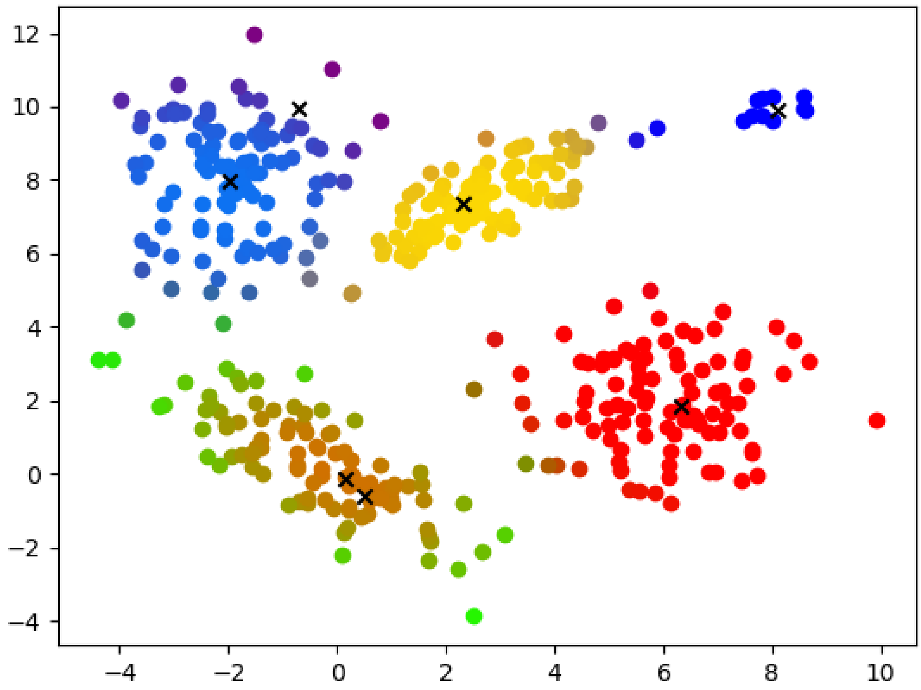

Figure 9: *Example hypothesis set learned from the data show in figure 4, using IGMM. This image shows the classification of each observation after the agent learned on a separate set of observations. In this example, 3 outliers were shown to the agent after 200 out of 400 observations during the learning phase. The black crosses denote the centre of each learned hypothesis. The colour denotes the probabilistic membership of an observation to each hypothesis. The more resemblance the colour of the observation has with the colour of the hypotheses, the higher the probability it belongs to that hypothesis according to the agent. The BIC of this simulation was 133531.*

outliers were not novel for all hypotheses in the model at that time. We can expect that a similar percentage of the ignored outliers in the PRIGMM model were similarly ignored because of a likelihood that was too low. As such, the exact influence of using the posterior is unknown. We can note that effect is a quite large by looking at the difference between the two models.

The chance that the outlier group was ignored seemed higher for 3 outliers than for a single outlier. This is unlikely to be true overall: if one outlier is enough to form a new hypothesis, three consecutive definitely should. Our explanation for this phenomenon is simply a difference caused by the pseudo-random ordering of the data. No strong differences were found between 1 and 3 outliers for any measurement in the simulations. As such, there is no reason to assume any real difference between the agent seeing 1 or 3 consecutive outliers, which is not what we expected.

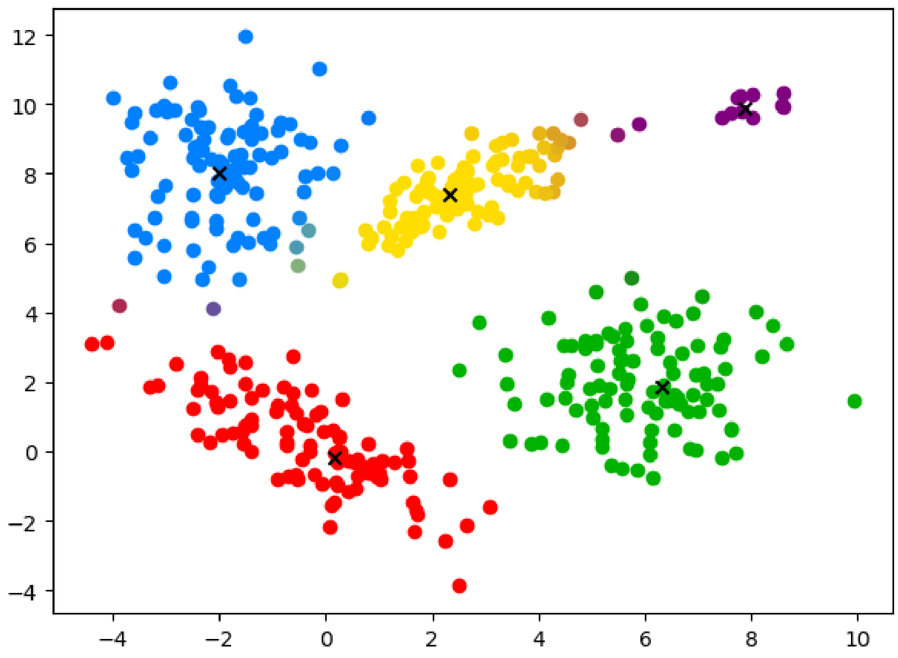

Figure 10: *Example hypothesis set learned from the data set shown in figure 4 using PRIGMM. The details are the same as in figure 10. The BIC of this simulation was 103570.*

From figure 5, we can also conclude that there is a moderate effect of the amount of previously made observations on the chance that the outlier group will lead to a new hypothesis. Although the effect seems strong, we must keep in mind that it is probably that a percentage similar to the IGMM model is due to the outliers not being 'surprising enough'. When we subtract the chance to ignore outliers of the IGMM model from the PRIGMM model, the effect becomes moderate. This is opposed to our expectations, which were that this effect would be more strongly present. The presence of the moderate effect property did not, seem to translate into a higher or lower BIC score. This suggests that the outlier group had a moderate effect on the BIC score as a whole, as ignoring or accommodating for it did not seem to influence the score by much. A follow-up design would have investigated how the chance of a new hypothesis is influenced by the amount of previous observations during the very early observations (5, 10, 15 etc.). This mainly because of their more direct use as a baseline expectation for empirical experiments in psychology, that rarely, if ever, show as many cues as in our simulations. In addition, the chance that outliers cause a new hypothesis is likely to decrease the most quickly during the early observations.

The fact that the PRIGMM model outperformed the model that generated our data is quite surprising. There are two possible explanations for this. The first is that this highlights a flaw in the BIC as a criterion. It is known that BIC harshly punishes the amount of parameters used for a model. It could be that this is an example of a situation where such

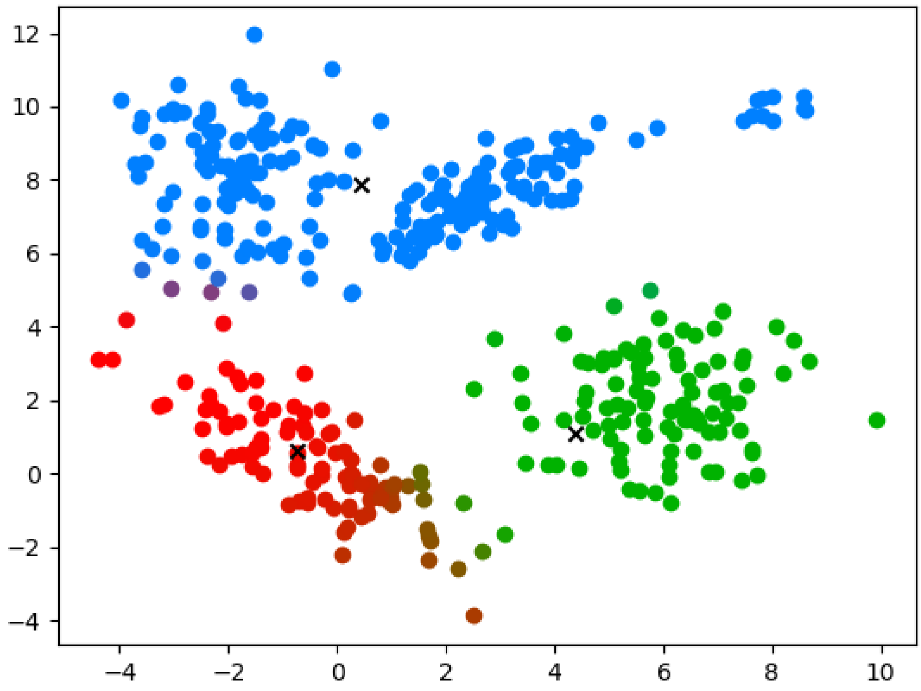

Figure 11: *Example hypothesis set learned from the data set shown in figure 4 using PRIGMM. The details are the same as in figure 10. The BIC of this simulation was 39917.*

a punishment is not warranted. The alternative explanation is that our test set was simply better represented as three main hypotheses as opposed to our four. In that case the loss in likelihood would very small compared to adding a fourth hypothesis. If that is the case, our model found a superior representation of the data. Whether the low BIC score holds up in a variety of 'environments' would be a natural follow-up question.

Either way, the PRIGMM model seem to overfit far less than the IGMM model. This provides us with a theoretical argument why humans might ignore anomalous observations. Doing so creates a sparser, yet arguably more predictive hypothesis space. As such, the ignoring of certain stimuli can be explained as a method used by the agent to keep the complexity of their predictive model low. Empirical studies where humans need to perform learning tasks could shed light on whether ignoring outliers is a persistent feature of human behaviour, or whether it arises circumstantially. In any case, our theoretical results provide new opportunities to learn more about these aspects of (human) cognition.

5.1 Future Work

In the current model, the 'unknown hypothesis' does not contain any information. While multiple observations may be needed together to cause a new hypothesis to be formed, only the last will help shape the contents of that hypothesis. A natural extension for the model would be to incorporate the observations into a ‘pre-hypothesis’ that has some decaying memory of what was most unknown recently. A chief concern here would be that outliers from different hypotheses would be joined together in the unknown hypothesis, which would need to be kept separate in the final belief model.

A next step is to look at ways of expanding the model. This has two natural extensions, both of which are valuable improvements. The first is to allow the agent to take the previous state(s) of the world into account. That is, instead of just making a prediction about the observation conditioned on nothing, allow the agent to make this prediction dependent on the previous observation(s). It should be noted that such a model would only provide extra use if the world has some stability. This stability was notable absent in our simulations, as observations could be drawn from any main hypothesis at any time.

The second way to expand the model is to add hierarchical layers on top of the current hypothesis layer. An initial ‘easy’ path would be to look at models with a tree structure of several layers deep, which keeps the independence assumption from this thesis. Following that, subsequent work would need to focus much more on how relations between concept are formed, be that between or within ‘layers’ of a belief network.

6 Conclusion

We proposed a new model, PRIGMM, that learns a hypothesis space represented by a Gaussian mixture model using an ‘unknown’ hypothesis. We showed through simulations, that this model successfully replicates the main aspect of human cognition of interest: the ignoring of anomalous observations. In addition, we showed that the hypothesis space developed by an agent using this model has a very low BIC score, suggesting that it learns a sparse yet accurate representation of the world. This provides theoretical evidence for the idea that humans choose to ignore anomalous observations because it is beneficial to the quality of their belief models. Lastly, we showed that, contrary to what we predicted, the amount of observations made by an agent before they were confronted with a series of outliers has only a moderate effect on the chance that a new hypothesis is formed or not. Still it held that the more observations were seen by the agent, the higher the chance that no new hypothesis was formed. This effect was similar for all tested outlier groups.

References

Arora, P., Varshney, S. et al. (2016), 'Analysis of k-means and k-medoids algorithm for big data', *Procedia Computer Science* **78**, 507–512.

Baker, C., Saxe, R. & Tenenbaum, J. B. (2006), Bayesian models of human action understanding, *in* 'Advances in neural information processing systems', pp. 99–106.

Clark, A. (2013), 'Whatever next? predictive brains, situated agents, and the future of cognitive science', *Behavioral and brain sciences* **36**(3), 181–204.

Declercq, A. & Piater, J. H. (2008), Online learning of gaussian mixture models-a two-level approach., *in* 'VISAPP (1)', pp. 605–611.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 1–22.

Engel, P. M. & Heinen, M. R. (2010), Incremental learning of multivariate gaussian mixture models, *in* 'Brazilian Symposium on Artificial Intelligence', Springer, pp. 82–91.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996), A density-based algorithm for discovering clusters in large spatial databases with noise., *in* 'Kdd', Vol. 96, pp. 226–231.

Friston, K. (2010), 'The free-energy principle: a unified brain theory?', *Nature reviews neuroscience* **11**(2), 127.

Gigerenzer, G. & Todd, P. M. (1999), Fast and frugal heuristics: The adaptive toolbox, *in* 'Simple heuristics that make us smart', Oxford University Press, pp. 3–34.

Johnson, S. C. (1967), 'Hierarchical clustering schemes', *Psychometrika* **32**(3), 241–254.

Kayhan, E., Hunnius, S., O'Reilly, J. & Bekkering, H. (2019), 'Infants differentially update their internal models of a dynamic environment', *Cognition* **186**, 139–146.

Kok, P., Brouwer, G. J., van Gerven, M. A. & de Lange, F. P. (2013), 'Prior expectations bias sensory representations in visual cortex', *Journal of Neuroscience* **33**(41), 16275–16284.

Kristan, M., Skocaj, D. & Leonardis, A. (2008), Incremental learning with gaussian mixture models, *in* 'Computer Vision Winter Workshop', pp. 25–32.

L Griffiths, T., Kemp, C. & B Tenenbaum, J. (2008), 'Bayesian models of cognition'.

Lee, M. D. (2011), 'How cognitive modeling can benefit from hierarchical bayesian models', *Journal of Mathematical Psychology* **55**(1), 1–7.

Lee, T. S. & Mumford, D. (2003), 'Hierarchical bayesian inference in the visual cortex', *JOSA A* **20**(7), 1434–1448.

Lotter, W., Kreiman, G. & Cox, D. (2016), 'Deep predictive coding networks for video prediction and unsupervised learning', *arXiv preprint arXiv:1605.08104* .

Perruchet, P. & Vinter, A. (1998), 'Parser: A model for word segmentation', *Journal of memory and language* **39**(2), 246–263.

Quine, W. V. (1974), 'The roots of reference'.

Reddy, C. K. & Vinzamuri, B. (2018), A survey of partitional and hierarchical clustering algorithms, *in* 'Data Clustering', Chapman and Hall/CRC, pp. 87–110.

Romberg, A. R. & Saffran, J. R. (2013), 'Expectancy learning from probabilistic input by infants', *Frontiers in psychology* **3**, 610.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. (2017), 'Dbscan revisited, revisited: why and how you should (still) use dbscan', *ACM Transactions on Database Systems (TODS)* **42**(3), 1–21.

Schwarz, G. et al. (1978), 'Estimating the dimension of a model', *The annals of statistics* **6**(2), 461–464.

Song, M. & Wang, H. (2005), Highly efficient incremental estimation of gaussian mixture models for online data stream clustering, *in* 'Intelligent Computing: Theory and Applications III', Vol. 5803, International Society for Optics and Photonics, pp. 174–183.

Taillard, É. D. (2003), 'Heuristic methods for large centroid clustering problems', *Journal of heuristics* **9**(1), 51–73.

Thagard, P. (1989), 'Explanatory coherence', *Behavioral and brain sciences* **12**(3), 435–467.

Xu, F. & Tenenbaum, J. B. (2007), 'Word learning as bayesian inference.', *Psychological review* **114**(2), 245.

Appendix

		Number of Outliers			
Observations	**Model**	0	1	3	10
40	IGMM	**101332 ± 31407**	102288 ± 22569	**109866±29156**	147128 ± 34073
	PRIGMM	**44770 ± 22924**	54590 ± 22142	54895 ± 24509	77497 ± 30429
80	IGMM	-	**102195 ± 22152**	112056 ± 30749	146331 ± 33420
	PRIGMM	-	54289 ± 22767	52765 ± 25391	73602 ± 31461
120	IGMM	-	103617 ± 23135	110608 ± 31230	143480 ± 33663
	PRIGMM	-	52854 ± 23354	50971 ± 25055	69017 ± 33140
160	IGMM	-	104290 ± 23068	111063 ± 31133	142133 ± 34639
	PRIGMM	-	52931 ± 23757	50878 ± 25091	66388 ± 32239
200	IGMM	-	104717 ± 22642	110808 ± 29332	142202 ± 36622
	PRIGMM	-	52928 ± 24192	51425 ± 25693	65772 ± 32638
240	IGMM	-	105223 ± 22942	110968 ± 30154	141660 ± 37630
	PRIGMM	-	53779 ± 24608	51056 ± 25957	65812 ± 32185
280	IGMM	-	105688 ± 23348	110825 ± 30262	141018 ± 37461
	PRIGMM	-	51981 ± 23974	**50139±25721**	62914 ± 32396
320	IGMM	-	105445 ± 22890	110578 ± 30201	140094 ± 37357
	PRIGMM	-	51792 ± 24021	50471 ± 25806	63083 ± 32251
360	IGMM	-	105082 ± 22540	111670 ± 31333	**137944±37769**
	PRIGMM	-	**51446±24161**	50244 ± 25881	62299 ± 31842
400	IGMM	-	105317 ± 22560	111102 ± 30515	137961 ± 38055
	PRIGMM	-	51891 ± 24310	50636 ± 26216	**61480 ± 31995**

Table 2: *BIC score for the simulations as recorded on a separate test set of 400 'normal' observations. All agents trained on 400 observations in total, excluding the outlier group. The left column shows after how many observations the outlier group was shown to the agent during training. For the 0 outliers group, only the first row shows the result, as all results are identical. Each column highlights the best BIC score for both models.*

	Number of Outliers			
Model	0	1	3	10
Fixed-Base	92036 ± 2179	92687 ± 1956	94461 ± 2077	99876 ± 2022
Fixed-Outlier	706637 ± 2179	708453 ± 14024	706007 ± 15108	710359 ± 2022
IGMM	101332 ± 31407	104386 ± 22785	110954 ± 30407	141995 ± 36069
PRIGMM	44770 ± 22924	52848 ± 23729	51348 ± 25532	66786 ± 32058

Table 3: *BIC score for both fixed models (with and without inclusion of the outlier group). Scores are indicated as the mean plus or minus the standard deviation.*