Computational Modelling of Visual Search

By Vilius Narbutas

A thesis submitted to the
University of Birmingham
For the degree of
DOCTOR OF PHILOSOPHY

School of Psychology
College of Life and Environmental Sciences
University of Birmingham
September 2017

Abstract

Visual search traditionally has two main competing theories of parallel and serial search and this architectural issue has not been solved to this day. The latest developments in the field have suggested a possibility that response time distributions may aid in differentiating the two competing theories. For this purpose we have used the best available serial model Competitive Guided Search and two biologically-plausible parallel models inspired by the theory of biased competition. The parallel models adopted a winner-take-all mechanism from Selective Attention for Identification Model as base model that was extended to form a novel model for explaining response time distributions. These models are analytically intractable, therefore we adopted a more accurate kernel density estimator for representing unknown probability density function. Introduced robustness properties to the fitness method and developed a more efficient algorithm for finding the parameter solutions. Then these methods were applied for comparison of the respective models and concluded that winner-takes-all model poorly generalises to response time distributions. The results were followed by introducing a novel Asymmetrical Dynamic Neural Network model that managed to explain distributional changes better than Competitive Guided Search model.

Acknowledgements

I must thank Dr Dietmar Heinke for being an incredible supervisor, for always listening and engaging, for tackling issues together and being enthusiastic about every little new idea, great or foolish. It was truly a fun journey that I secretly wished to last just a little bit longer.

I also have to thank EPRC for funding most of my PhD that has been a true saviour over the years.

Finally, I have to thank my wife Mai for always being there, lovely and cheerful as usual. For listening through my ridiculous out space ideas. We always know there will be something that was not considered but acknowledging these random thoughts nonetheless.

Table of Contents

List of Figures

Chapter 1:

General Introduction

1.1 Response Time Distributions

Many psychological experiments produce response time (RT) distributions as a manipulated variable. Historically, this crucial information has been used to study and raise hypothesis of underlying processes that have generated these distributions (Luce, 1986; Hohle, 1965). This time from stimulus appearance to the response covers a number of underlying psychological processes within brain such as encoding, identification, decision making and motor execution times. Understanding these processes requires identifying how every process contributes to overall execution time by manipulating experimental designs. The importance of RT-distributions has been appreciated but application was limited due to violations of normality, violations of independence per trial and present contaminations.

It is well established that often RT distributions are not normally distributed and contain skewness as well as kurtosis, thus logarithmic transformations frequently are not applicable. Moreover, these features can cover different attributes of the underlying processes within brain such as working memory, attentional lapses and other psychological processes (see Matzke & Wagemakers, 2009 for review). It has been shown on several occasions that different features of the distribution vary differently across different conditions. For instance, in visual search experiments the set size increase changes mean and standard deviation values more or less linearly. However, other features such as skewness and kurtosis do not change linearly and the average of RTs cannot capture these changes (Palmer et al., 2011).

It is generally believed that many trials are required to appropriately approximate unknown RT-distribution (van Zandt, 2000), which may not be feasible for the given task. Repeatedly performing the same/similar task is subject to learning which may not be desirable. Consider a visual search example, the purpose is to identify whether a search is perform in parallel or serially in conjunction search task. However, it has been shown that participants learn to quickly discount a set of distractors within conjunction search by focusing on a single feature such as colour. Some larger scale visual search experiments confirmed that such strategy was used by participants (Wolfe, Palmer & Horowitz, 2010; Moran et al., 2013). In statistical literature, this is also known as violation of independence as the future outcome depends on previous trials. Thus, it is reasonable to assume that a true RT-distribution is different at the beginning of the experiment when compared to the distribution at the end of the experiment. In this thesis, this assumption is also taken but datasets are kept smaller and a method with greater accuracy is developed instead.

Contaminants are the most problematic issue as it indicates processes that are irrelevant to the task such as attentional lapses (Ratcliff & Tuerlinckx, 2002) and introduces biases to the results. The contaminants consist of fast responses that are too fast for the given task. Slow contaminants that generally reside within a tail and can be an indicators of attentional lapses. The issue is with hidden contaminants which are within the RT-distribution itself. It can be difficult to differentiate a genuinely prolonged task execution from attentional lapses. The distribution of these hidden contaminants are not known and a number of assumptions have to be made when modelling them. We propose relying on robust statistical methods which reduce the effect of these contaminants when a proposed model properly defines the studied phenomena (For more information refer to 3rd Chapter).

1.2 Visual Search

Many visual tasks in our daily life would be impossible without our capability to find visual objects of interest. This is a key feature aiding us in processing an enormous amount of visual information available in our environment. This search procedure is also regularly used for day-to-day tasks such as seeking for ingredients in the shopping list or cooking a meal. Researchers study this phenomena in a more controlled settings and visual search experiments are one of the most commonly used experimental paradigms for studying underlying cognitive processes within brain (see Eckstein, 2011 for a review). For example, determining the presence of predefined target among distractors is a standard visual search task. The time taken from the stimuli onset and a decision is known as response or reaction time (RT) and repeated measurements produce an RT-distribution. This variable is studied to determine the difficulty of visual search experiment in relation to the number of the distractors presented to the participant. A standard approach averages these measurements for the given number of distractors and the slope of the reaction time from few to many distractors is determined. This slope indicates the difficulty of the task and has been pivotal in the development of visual search theories.

The underlying cognitive processes have been elusive thus far and two main competing theories have emerged to explain this information processing phenomenon. One theory is inspired by the base knowledge of neuronal systems in the brain which proposes the parallel processing system (e.g., Palmer, 1995; Ward & McClelland, 1989). This theory roots from the realisation that humans are extremely quick at discounting irrelevant information within the enormous quantity present in our immediate environment. However, this theory has been challenged by the observations that some searches are quite slow and various objects have to be identified in a serial manner before a goal target is found (e.g. Sternberg, 1966). These observations have firmly positioned the visual search theory into two completely different architectural beliefs of serial vs parallel because parallel theory could reproduce the effects attributed to serial processing (Townsend, 1976).

Treisman and Gelade (1980) in their groundbreaking study proposed a Feature Integration Theory that introduced a two-stage model. This theory suggested that there is a feature map that is processed in parallel and in case a relevant map triggers more than one location, a serial process is begun to reject the items that were not discounted in the first phase. This enabled to explain a flat search slopes (aka, pop-out search) commonly attributed to parallel theory as well as steep slopes. However, the discussion did not end with this new model and alternative single-stage proposals emerged. In 1989, Duncan and Humphreys have suggested Biased Competition theory that assumes a parallel interference between competing items within visual field. This interference is measured by the similarity of the items and results in the overall slowdown of the processing time. The biggest bottleneck for models assuming serial component was to identify when a parallel process ends and a serial process begins because the observed slopes were not limited to pop-out outcomes or visibly serial identification of the items. Instead, there seemed to be no obvious switch and the slopes have a gradual increase from flat to very steep. As a result, Wolfe (1994) coined a term efficient vs inefficient search which has been widely accepted throughout visual cognition community.

Throughout the research, central tendencies remained as the main experimental measurement even though these central tendencies failed to determine the correct architecture. The issue resides within the capability of the parallel models to explain slower processing with each additional distractors intuitively attributed to serial processing (e.g., Townsend & Ashby, 1984; Townsend & Nozawa, 1995). Overall, this mimicry of the opposing theories failed to resolve an initial architectural issue (Townsend & Werger, 2004). Due to this unfalsifiability of the two architectures, the question has been largely ignored up until recently. Palmer et al. (2010) following the findings that different aspects of the response time distribution refer to different cognitive processes looked how different properties of the response time distributions vary in visual search paradigm. This study found quantifiable properties of the response time distributions changed differently and not always linearly as mean RTs approach assumes. A follow-up study by Wolfe, Palmer and Horowitz (2010) have shown that this falsifies the best available model inspired by serial architecture (Wolfe, 1994; Chun & Wolfe, 1996). This failure was solved by Moran et al. (2013, see Chapter 5 for more details) in their proposed Competitive Guided Search model. To this day there is no successful parallel model that could also explain RT distributions with a growing number of distractors.

Whether RT-distributions could aid in solving this unresolved issue is an open scientific question. These two theories posit two completely different predictions on how the RT-distributions are formed as the number of distractors is increased. Serial theory states that a resulting RT-distribution is a convolution of multiple identification processes, i.e. every selected item is identified in a serial manner. Each additional item indicates an increase in required number of serial scans. On the other hand, a parallel theory predicts that a change in the shape of the RT-distribution is due to other cognitive limitations such as limited attentional capacity or a direct competition between the presented stimuli. Parallel theory is stationed within knowledge that decision is attained via diffusion process (Ratcliffe, 1978) of accumulating evidence for as well as against presence of the target (aka. accumulator). The issue is, this diffusion process is a single two-choice accumulator of evidence while multiple distractors are present in visual search. Moreover, it would not be difficult to fit RT-distributions using the diffusion process because is has been proven that this model can fit any two-choice distribution (Jones & Dzhafarov, 2014). This indicates that only a relationship between diffusion model parameters and setsize effect would have to be established to form a model, in other words, there is a strong suggestion that a 2-stage unfalsifiable parallel model already exists. Thus, one of the purposes of the thesis is to design a single-stage parallel model that can account for distributional changes present in visual search experiments without changing any of its variables.

At this point it is worth stressing that the thesis focuses on the dichotomy of parallel vs. serial rather than on how visual features, such as colour, orientation, etc. influence visual search. In other words, such perceptual features in this framework turn into free model parameters rather than explicit computations.

1.3 Models

1.3.1 Serial Model

The most influential serial model is Guided search (e.g., Wolfe et al., 1989; Wolfe, 1994, 2007; Chun & Wolfe, 1996) which introduced a two stage model. The first stage computes an activation map (Koch & Ullman, 1995). This activation map is assumed to be a combination of bottom-up saliencies that define how distinctive the items are among themselves and top-down influence that indicates interest in the particular item(s) (Itti & Koch, 2000). This activation map guides selection for the second stage which identifies each chosen item. More active items will have a higher probability to be selected in compared to other items within visual display. When all items have an equivalent activation, the search slope becomes completely probabilistic and produces very steep slopes. However, this activation map is the main workaround for serial model to explain highly efficient searches with flat slopes because highly active items can have a probability to be selected equivalent to one.

A recent realisation of potential importance of RT-distributions for visual search (Palmer et al., 2011) and the failure to fit RT-distributions using Guided Search model (Wolfe, Palmer & Horowitz, 2010) has sparked a speculation that RT-distributions may falsify some models. The biggest failure of the model emerged due to RT-distributions of target absent condition being very similar to the target present condition. This weakness was tackled by Moran et al. (2013) in their novel Competitive Guided Search (CGS, for more detailed introduction see chapter 5) model which extended Guided Search model by introducing a better termination criteria. This proved to be crucial change for explaining RT-distribution in absent trials. Namely, it was disproven that people perform exhaustive search nor the half of the items within display and this quitting criteria has introduced a viable solution for termination (Townsend & Wenger, 2004).

1.3.2 Parallel Models

1.3.2.1 Race Models

Unlike a single model representing the serial class of models, parallel models divide into a couple of classes. The simplest parallel model, race model, assumes that the individual items compete in non-interfering fashion with each other for determining the response. Importantly the items (racers) don't interfere with each other in this race. Linear Ballistic Accumulator (LBA: Brown & Heathcote, 2008) is one of the best and simplest multi-choice decision making race model that has become a gold standard for studying cognitive decision making processes (Forstmann et al., 2008; 2010; Ho, Brown, & Serences, 2009) as well as a toy model for testing cognitive modelling methodology (Turner et al., 2013; Turner & Sederberg, 2014; Jones, & Dzhafarov, 2014). Different versions of the race model were also applied for stroop effect by introducing independent accumulators for colour and word (Eidels, 2012) and for visual search using signal detection theory (Verghese, 2001).

Overall, none of these models were adapted or fitted to response time distributions. Therefore, Moran et al., (2016) have developed and fitted a parallel race model to RT-distributions using quantile maximum likelihood (QML) method and compared it to CGS model. This model made a number of assumptions common in parallel modelling research. They adopted a race model with drift rates adapting to the number nodes within display which is a common practice for modelling n-choice decision making (Usher, Olami, & McClelland, 2002; Brown & Heathcote, 2008). Additionally, they adopted a traditional self-terminating process within present conditions meaning a decision is made right after the first identification but not an exhaustive search for absent trials (Townsend & Wenger, 2004). Instead they introduced an additional diffuser that accumulates after each rejection for absent trial as in CGS model (Moran et al. 2013). Another very important assumption was within enforcing the input to be the same for all diffusers. No input bias is assumed, they motivate this by stating that modifying the initial accumulation level is sufficient for the bias.

They found that CGS model outperformed this parallel model despite it having more parameters. There were a number emerging differences and the main issue was within balancing error rates with RT-distributions. When the model would match the distributions, it would fail with error rates and when it would match error rates, it would fail with the distributions. The issues with independent accumulators are as outlined by the authors within statistical chance of different accumulator reaching a decision threshold. While this work only considered single race model, it does illustrate the possible underlying issues with such models.

1.3.2.2 Competitive Models

A class of parallel models that was not considered by Moran et al. (2016) is based on the Biased Competition Theory which probably is the most influential parallel search theory (e.g., Desimone & Duncan, 1995; Duncan, Humphreys, & Ward, 1997). In contrast to the race model, this theory suggests that models are competing for activation in an inhibitory fashion. The nodes accumulate information under constant interference from other nodes until a winner is declared. The biases in this model are introduced in an identical approach to activation maps used for CGS model by combining bottom-up and top-down influence (e.g., expectations about targets, Anderson, Heinke & Humphreys, 2010; or short-term memory, Woodgate, Strauss, Humphreys & Heinke, 2015). Competitive models explain various cognitive processes such as decision making (Wang, 2002; Bogacz et al., 2007) or selective attention (Mordkoff & Yantis, 1991). Finally, studies using the Selective Attention for Identification Model (SAIM; Mavritsaki, Heinke, Humphreys & Deco, 2006; Heinke & Humphreys, 2003; Heinke & Backhaus, 2011) have shown that search efficiency is another area that competitive models may be able to explain.

It is worth noting that this theory has been overlooked by authors testing whether RT-distributions are sufficient to distinguish serial vs parallel problem, thus this thesis attempts defining a competitive model that is capable to explain RT-distributions. It is biologically plausible class of models but it raises a multitude of applied modelling challenges due to being analytically intractable (this issue will be covered in greater detail within next section) therefore a few chapters will be dedicated to addressing these underlying issues.

1.4 Modelling RT-distributions

1.4.1 Statistical Modelling

A viable option to model RT-distributions is to use parametric functions as models such as Weibull, ex-Gauss or ex-Wald (see Burbeck & Luce, 1982; Schwarz, 2001; Logan, 1992; respectively) to characterise such distributions, though these are limited to one distribution. Though such functions are not perfect, they can explain individual distributions relatively well and can provide a valuable insight into how distributions change themselves (Palmer et al., 2011).

In a number of cases a non-parametric function as a model is more suitable. Such models, unlike parametric functions, do not assume the shape of the distribution and the data defines it instead. These methods choose to approximate either probability density functions (PDF) or cumulative density functions (CDF). CDFs are most commonly approximated functions using quantiles (Heathcote, Brown & Mewhort, 2002). The main advantage of CDFs is that it allows for constructing a super participant. PDFs were considered as an alternative but it has not been used much other than for considerations even though its effectiveness has been verified (van Zandt, 2000). The choice is justifiable due to higher mathematical complexity of the approximations with little to no gain in accuracy of the approximation. It also leads to higher computational costs for parameter estimation and is generally perceived a that a large number of samples is required. Due to these reasons PDF approximation has been overlooked and its potential has not been fully explored.

The most commonly used statistical models for approximating unknown probability functions are non-parametric, although, parametric models have been and still is a viable choice when computational resources are an issue (for example, Ratcliffe, 1978). The choice of the statistical model mainly depends on the theoretical model used and available computational resources as some methods are more accurate than other methods. This thesis aims to determine if RT-distributions can aid in separating parallel from serial search thus we opted for emphasising the accuracy over efficiency. As discussed in Statistical Models section, approximation of the CDFs is the most commonly used method. Though, PDF approximations have reemerged in recent years thanks to the work by Turner, Sederberg (2014) showcasing that the simplest PDF approximation via kernel density estimation (KDE, Silverman, 1986) outperforms the most commonly used CDF approximation in posterior approximation.

1.4.2 Theoretical Modelling

There is a whole class of models that make architectural assumptions about the cognitive processes (see Teodorescu & Usher, 2013; for a review). These models hypothesise how systems generating response time distributions work and can be targeted by carefully designed experiments. In general, the processes generating these distributions are produced by stochastic systems where some information accumulation with random noise occurs before the actual response. Two simplest models of such systems are drift diffusion (DDM) and Ornstein-Uhlenbeck (OU) (Ratcliff, 1978; Busemeyer & Townsend, 1993). These models are relatively simple and have known probability functions that determines the probability of various model outcomes. However, even the simplest psychologically plausible extensions make these probability functions analytically intractable. For example, an extended DDM for two alternative choice decision making introduces additional variabilities (e.g. between trials) which renders it intractable (Smith & Ratcliff, 2009). The same issue arises with an OU extension for multi-choice decision making (Usher & McClelland, 2001) thus no analytical function is available for either model.

1.4.3 Intractable models

Analytically, intractable models create a wealth of issues because parameter estimation of the given model becomes biased for the data. Generally, parameter estimation is performed via finding the maximum likelihood value (see robust likelihood chapter for more in depth discussion). The likelihood value indicates how well the model presents the data given a certain parameter setting. Hence the maximum likelihood determines the parameter values of the model that represent the data most accurately. In other words, maximum likelihood is the mode within a likelihood function which describes how well the model represents data as a function of parameter values. If this likelihood function is known analytically, it is easily computed and the maximum is trivially found. Analytically intractable models, by extension, do not have a known likelihood function. These models are solved using numerical methods by simulating a number of times until a reasonable approximation of their function can be determined. In essence it is an identical process as attempting to model RT-distributions because these models approximate these distributions via simulations, therefore, statistical models are used as an intermediary. These intermediary approximations can be used for approximating likelihoods of the data, e.g. response time data.

The approximation of the PDF is not the only issue concerning modelling analytically intractable models. Unavailability of the likelihood function makes methods that rely on gradient such as Nelder-Mead algorithm (Singer & Nelder, 2009) to perform poorly or fail completely. In such cases intelligent sampling methods are desirable because grid sampling is too costly even with small number of dimensions. Sampling methods use some strategy with randomisation to inform the sampler where a good solution is likely to reside. A good sampler reduces the overall optimisation costs by reducing the number of required samples (Doucet & Johansen, 2008). These methods use MCMC algorithms which are also known as Bayesian approximators. In fact, an optimal sampler produces the samples that are distributed equivalently to the posterior distribution. In other words, a sampler minimises the deviation between the distribution of produced samples and true posterior distribution. These approximators either rely on a good approximation of the likelihood values or introduce some acceptance threshold. Algorithms that use acceptance threshold as the sampling guidance are also known as likelihood-free Bayesian methods or Approximate Bayesian Computational (ABC) methods. These methods rely on arbitrary error threshold which does not guarantee the correctness of the approximator (Beaumont, 2010). In our work we found that these methods perform very poorly when maximum likelihood is undefined (Grazian & Robert, 2015) and proposed an alternative algorithm for approximating analytically intractable models that predict RT-distributions (see 4th Chapter for more information).

1.5 Outline

The structure of this research is as follows. We divided it into two parts since the entire research is dealing with two larger scientific questions. In the first part we will introduce the new methods for broader scientific usage in analysing intractable models as well as their associated unknown probability density function (PDF). In the three chapters within this part we will introduce the most common methods and present improvements and alternatives for these methods.

The main focus of the thesis addresses an unsolved problem appropriately approximating the reaction time (RT) distribution of humans as well as analytically intractable models. Characterising these models is a crucial factor in opening a whole class of models for speculation and falsification such as various accumulator models (e.g., Smith & Ratcliff, 2009; Usher & McClelland, 2001). All these models predict RT-distributions which cannot be expressed in a analytical PDF. Thus, researchers have generated a wealth of literature on various approaches to approximate PDFs and in the first chapter we will propose a new method as an improvement compared to most other methods.

The second chapter addresses an issue that RT data have the tendency to be contaminated by processes that were not generated by the cognitive processes of interest. Moreover, data cleaning tends to be biased and some poor samples cannot be identified within distributions because they are hidden within an actual RT distribution which can lead to poor inferences about the cognitive processes of interest (Miller, 1991; Ulrich & Miller, 1994). Some authors started to model the contaminants as part of the models (Ratcliff & Tuerlinckx, 2002; Ratcliff & McKoon, 2008; Wagenmakers et al., 2008) but we argue against such practice and suggest instead to adopt robust methods by introducing a robust likelihood method for recovering true parameters without requiring to identify contaminants within data.

The third chapter deals with arising issue from the use of the described two methods within the first two methodological chapters. This new issue emerges when parameter estimation is performed and most of the currently available methods fail to find the best solutions within a reasonable time. Moreover, these methods negatively impact posterior approximation algorithms therefore we have developed a novel non-parametric posterior approximation algorithm for analytically intractable models when the key two methods are adapted.

The second part will apply a combination of these methods to compare a serial CGS model with parallel SAIM-WTA and ADyNeN models using the data generated from visual search experiments (feature search, conjunction search and spatial configuration search). The fifth chapter will compare SAIM-WTA model with CGS using some of the outlined methods demonstrating their applicability as well as how that translates to a practical example. It demonstrates that basic SAIM-WTA model does worse than CGS in representing RT-distributions combined with accuracy. The sixth chapter extends the findings in the fifth chapter and will compare ADyNeN model that overall has a capacity to explain RT-distributions slightly better than CGS model. We will follow the findings of the two studies with broader overview of implications into a long-standing question of serial vs parallel search and suggest that ADyNeN model has a capacity to explain a wide range of visual search processes as well as possible natural extensions to explain additional processes.

Part 1:

Methodology

Chapter 2:

Kernel Density Estimators

Abstract

For decades now, reaction time (RT) analysis has been a main outcome of many experimental paradigms encompassing psychology in humans as well as animals. However, while importance has been appreciated the usage has been limited due to its inherently skewed properties, violation of independence assumption and vulnerability to contaminations. Additionally, the associated processes and by extension the models are too complex to solve analytically. There have been a number of attempts to address these issues but the preference for usage of mean RTs and other less accurate methods has remained strong. An alternative approach of using kernel density estimator (KDE) has often been overlooked, possibly due to its higher computational demands to compute multiple distributional approximations. However, modern computers have reduced this issue substantially. In this research we present one of the latest publicly available KDE methods, apply it to modelling of RT-distributions and show its performance in relation to a better known traditional KDE method. We found that it outperforms traditional KDE method yet retains fewer distributional approximations in order to accurately represent underlying RT distributions. Moreover, we show that as few as 100 trials produce a relatively reasonable approximation of target RT distribution.

2.1 Introduction

Response time is a vital experimental measurement for studying and understanding cognitive processes. A set of these measurements form an RT distribution which has eluded cognitive modellers to this day and the most common modelling approach is to simplify the data by finding a central tendency of the distribution. This approach assumes that moments of the unknown PDF do not change as the task is modified or that these changes do not affect core findings of the study. However, the assumption of the moments generally is not true and visual search is a prime example where the shape of the distribution changes as the number of distractors is changed (Wofle, Palmer & Horowitz, 2010). Moreover, a second assumption makes a very gross generalisation about experimental conditions and may lead to statistical evidence when there is none if precaution analysis is not performed (demonstrating it is beyond the scope of the thesis, though, see general discussion chapter for further look at this particular issue).

Various summary statistics have been proposed to accomplish this; Ratcliff (1978) used intermediate model ex-Gauss while Heathcote, Brown and Mewhort (2002) suggested quantile maximum likelihood method (QML). For instance, QML divides cumulative functions into percentiles (usually 5 quantiles but can contain more for higher accuracy given sufficiently large data is present) containing the same number of data points. Quantiles work reasonably well, they are fast and robust to noise within data, thus a viable option when a single decent solution is desirable. It has been used in a wide range of RT distributional analysis (Pleskac & Busemeyer, 2010; Schmiedek et al., 2007; Vandekerckhove & Tuerlinckx, 2008) and while reasonable solutions will be produced, these will be subject to biases which can be evidently shown when posterior distributions are approximated (Turner & Sederberg, 2014).

Indeed, a more accurate representation of RT distributions would produce better predictions for the best solutions and a more stringent comparison of the models. For this purpose kernel density estimate (KDE) methods are attractive as they can model RT distributions without requiring summary statistic. These methods combine multiple smaller probability functions into one common function. Such method is flexible enough to approximate any desirable PDF. KDEs are a natural extension to histogram methods by having probability functions rather than bars; this extension has a property of naturally smoothing the approximation of the underlying distribution. Surprisingly, it is relatively uncommon to use KDEs to represent RT probability functions and even when it is used, the simplest KDE where every single datapoint have their own kernel (Silverman, 1986) is used (Turner, van Zandt & Brown, 2011; Turner & Sederberg, 2014; Miletic et al., 2017; Turner, Sederberg, & McClelland, 2016), we will refer to this implementation as traditional KDE. The usage of such methods has received little attention in psychological literature mainly due to perception that a lot of trials are required to approximate an underlying PDF (van Zandt, 2002).

However, KDE methods have progressed extensively from commonly used traditional KDEs and distributions can be approximated with relatively few kernels as well as fairly high accuracy. Recently we have presented the usage of relatively advanced online kernel density estimate (oKDE) method in practice (Narbutas, Lin, Kristan & Heinke, 2017) designed by Kristan, Leonardis and Skočaj (2011). While this implementation is designed for online distributional approximation, we were only using it for offline approximations. There we confirmed earlier results by Van Zandt (2000) that relatively few trials (100) are enough to represent response time (RT) distributions to some degree of accuracy. We also speculated that this method is better than a traditional KDE; however we did not perform any direct comparison. To address this shortcoming we carried out simulation studies that specifically compared these two methods.

2.2 Description

2.2.1 Traditional Kernel Density Estimate

Kernel density estimate is expressed as a mixture of weighted probability functions

$$f(X, h) = n^{-1} h^{-1} \sum_{i=1}^{n} K((X - \mu_i)/h) ,$$

with h corresponding to a smoothing parameter (commonly termed as bandwidth), n stands for the number of kernels while μ is their centre and K a kernel function itself. A kernel function represents an expected form of uncertainty surrounding the point this function represents. Most commonly used functions are Gaussian and Epanechnikov (we used the former) though other distributions such as Poisson can be utilised as well. The smoothing parameter is determined by Silverman's rule of thumb: $h_n = 0.9/N^{0.2} min(\sigma, (q_3 - q_1)/1.349)$; where σ is standard deviation and q_3, q_1 are first and third quartiles of the given dataset containing N points. 0.9 is a scalar value that can be adjusted as long as it remains below 1. It should be set to the value that best describes data which minimises the difference between target KDE distribution and an unknown probability function. A general suggestion is to start with 0.9 and change until the most reasonable value is found (van Zandt, 2002). The number of kernels as well as their locations can be chosen rather arbitrarily or centred on all available datapoints. The most accurate choice would correspond to data but can be computationally very inefficient with large datasets as it grows linearly with each data point, a reasonably small but not too small number of kernels could be the best compromise, e.g. van Zandt (2000) has chosen 60 equally spaced points within the range of the given data but this is far from optimal choice as it will be demonstrated later.

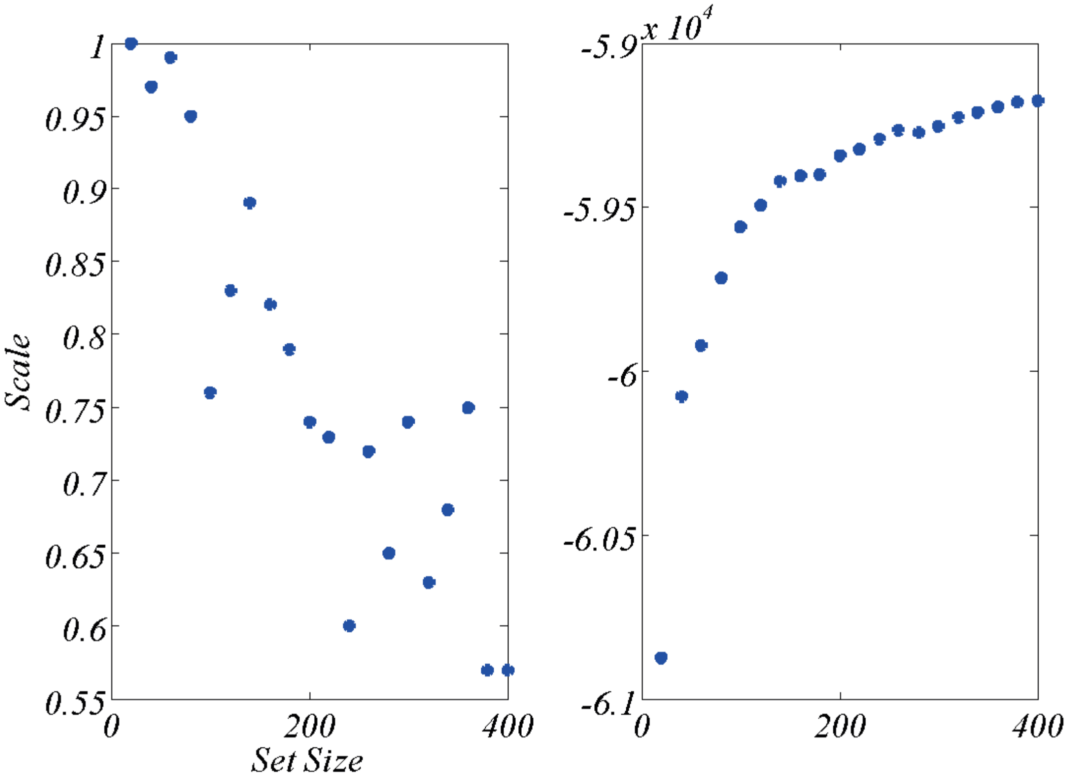


Figure 2.1: An illustration of the approximately of the best solutions for traditional KDE. The figure on the left show that scale parameter has a decreasing best smoothing parameter values but this decrease is sharper for smaller KDE. The figure on the right show log-likelihood values illustrating an increase in accuracy of the underlying distribution as the set size grows.

2.2.2 Online Kernel Density Estimate

To illustrate the progress of KDE methods we employed oKDE (Kristan, Leonardis, & Skočaj, 2011). This method has been shown to find a good balance between the number of components and accuracy of the target distribution in complex multivariate space. It is Gaussian mixture distribution and, unlike traditional KDE, does not rely on equivalent components for each point but adjusts them by clustering these points. As a result, it retains relatively compact mixture distribution that places components to characterise uneven sparsity within data in a more efficient way. Each component has its own weight corresponding to the relative number of points falling within it and covariance matrix. This can be expressed as a convolution of kernels: $\sum_{i=1}^{n} w_i \phi(X - \mu_i, h + \sigma_i)$.

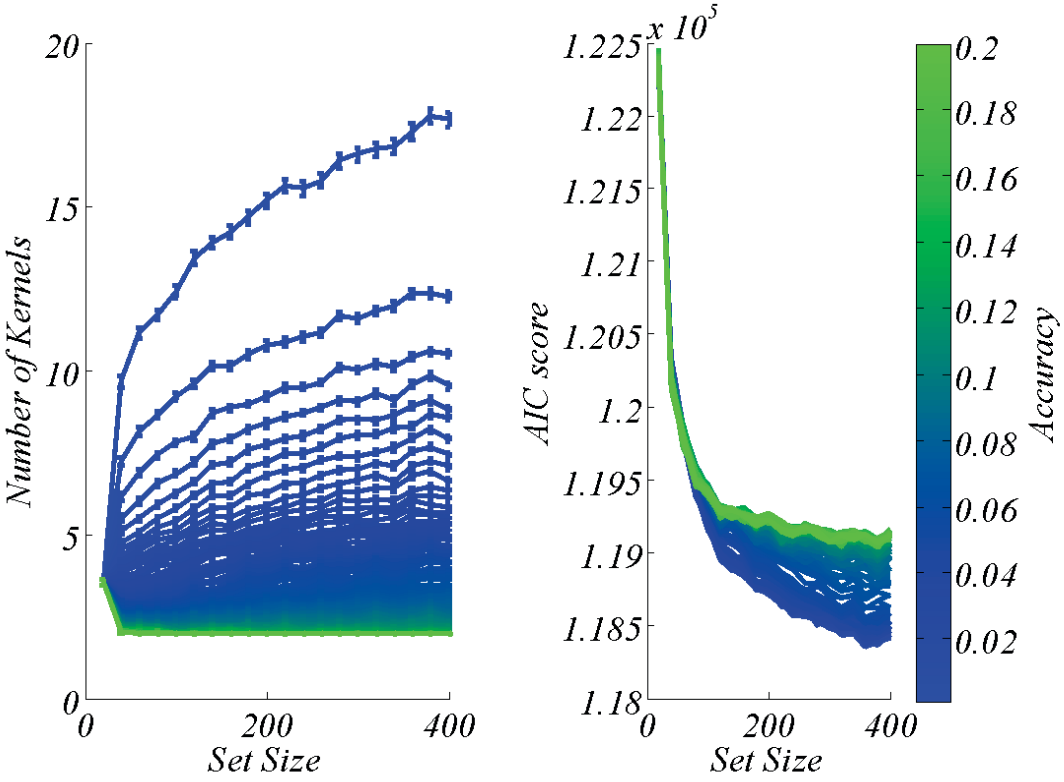


Figure 2.2: An illustration of results as permissible error is manipulated for oKDE. An illustration on the left shows how many kernels are required as the permissible error is modified. Lower accuracy leads to fewer components that are used to represent original data (with minimum being at 2). The figure on the

right demonstrates how AIC changes as accuracy is changed. The best solutions are concentrated at relatively low accuracy levels.

Most importantly, oKDE approximates an optimal bandwidth. Since the original distribution is unknown, it can be approximated by minimising an *asymptotic mean integrated squared error* (AMISE, Wand & Jones, 1995). Once the approximate optimal smoothing parameter is known, the algorithm clusters kernels within KDE in order to find a mixture distribution that approximates an original distribution with fewer components $m: \sum_{i=1}^{n} w_i \phi(X - \mu_i, \sigma_i) = \sum_{j=1}^{m} \hat{w}_j \phi(X - \hat{\mu}, \hat{\sigma}_j)$ where $m < n$. This smaller KDE function induces an error within approximation of original distribution which is expressed using Hellinger distance metric that shows how similar two distributions are. The entire clustering is performed using hierarchical top-down clustering (Goldberger & Roweis, 2005) in order to minimise the number of components but retain acceptable level of disparity between original sample and an approximated KDE. This is done by making sure that each component is below this threshold. Each component representing its subset data worse than this threshold is divided into two new components. These new components are centred by clustering the corresponding subset data using k-means algorithm. This reduces oKDE set-up variables to permissible level of error within clusters. For comparison, traditional KDE requires to determine the scale of the bandwidth and locations as well as the number of kernels for more efficient representation of the data. Clustering reduces these requirements with an introduction of some bias, as we will show, a good choice for permissible error leads to a better generalisation over unknown PDF.

2.3 Materials and methods

To compare the two methods we used Weibull parametric function with the same parameters values as van Zandt (2000) used. These parameters were set to 163.15 for scale, 1.5 for shape and 652.72 for shift which produces a distribution that arguably resembles a possible response time distribution. These values were kept constant throughout and were used to generate hypothetical datasets of sizes between 20 and 400 with 20 being an increment. The largest set size of 400 was chosen to represent a realistic size of the RT datasets as gathering larger datasets can be cumbersome and is widely unpopular. We also do not expect any significant difference between the two methods at larger datasets in terms of accuracy. From each data set 100 samples were taken and which were used to the determine a KDE. We also generated one large dataset containing 10000 data points. This large dataset was used to evaluate each KDE which was created from the smaller datasets. This evaluation used the likelihood principle i.e., the likelihood that this large dataset were generated by the given distributions. According the likelihood principle, a more accurate representation of true distribution would produce higher likelihood values over points that were generated by this distribution.

2.4 Identifying good settings for KDE methods

2.4.1 Optimal Bandwidth for traditional KDE

The traditional KDE has components centred on datapoints and has a common bandwidth for all kernels (Silverman, 1986). Identifying the best common bandwidth provides the most accurate PDF approximation that can be used for comparing two KDE methods. This section of the article focuses on finding the approximate best solutions for the scale parameter for bandwidth computation as the sample size is changed. The variable had values set between 0.01 and 1, the intermediate step was equivalent to 0.01 (100 parameter settings in total). Since there were no changes in the KDE other than the scale itself, log-likelihoods are sufficient for identifying the best performing parameter values. These settings were found by identifying the highest average log-likelihood values at each set size. Fig. 1b? illustrates the best average setting across increasing dataset size and a decrease in the best performing scaling values can be observed (Fig. 1a). Log-likelihoods illustrate the increasing average accuracy of the approximate distribution and an availability of more data would improve accuracy.

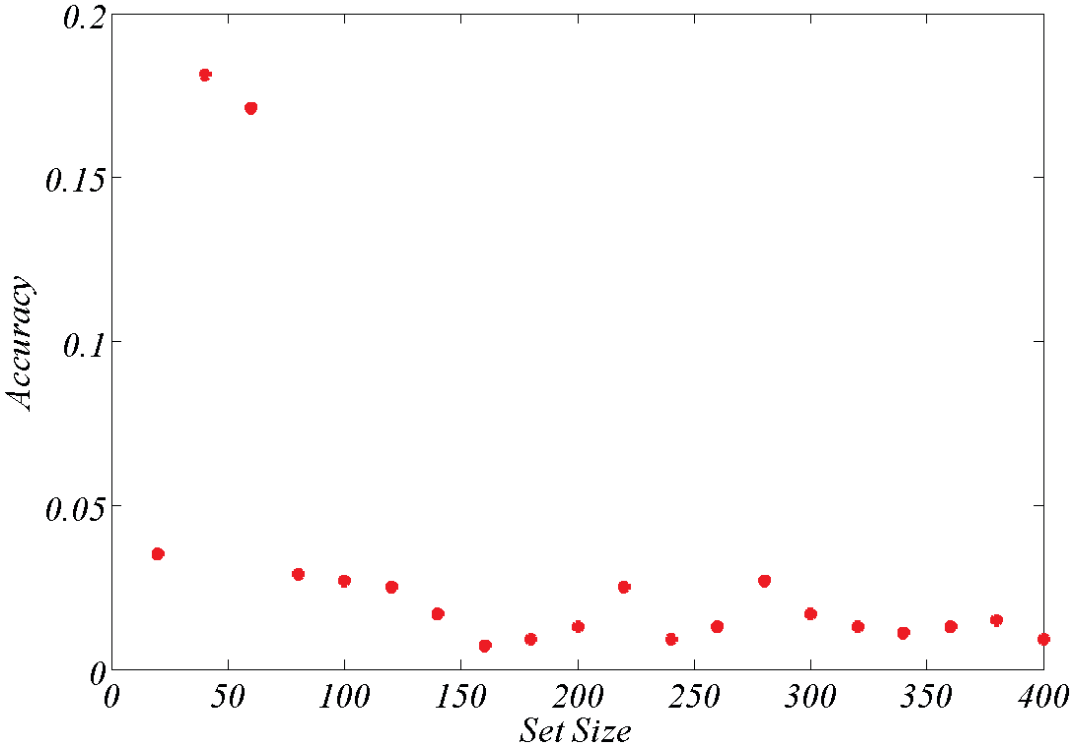


Figure 2.3: This figure show the optimal levels of accuracy for different set sizes. The best solutions have a small negative slope as the number of items is increased but the level is relatively consistent.

2.4.2 Optimal oKDE accuracy

The previous results established the best performing values for smoothing parameter as a function of set size. The oKDE method approximates the optimal bandwidth and no identification is required by the user. Instead, it has an accuracy measure as threshold for data clustering. We repeated the experimental design as used for smoothing parameter earlier and found the optimal permissible error level for the same set size settings. This accuracy level is expressed as Hellinger distance which we varied from 0.002 to 0.2 with 0.001 being an increment step. There were no changes in the procedure of generating simulated PDFs as described in general methods. Since no alternative method formulation was compared, we focused on AIC scores in order to identify the best settings. AIC was used instead of log-likelihoods since it takes into account the number of variables available. Due to clustering, the number of components varies depending on accuracy as well as the data size.

The results showed that the number of components varied from 2 components at the minimum accuracy but has not exceeded more than 20 at the maximum accuracy. Though, in general it required around 6 kernels to represent response time distribution (Fig 3a). This may change when a differently shaped distribution is being used, but a similar number of optimal set of components should be expected. Possibly, as sample size grows and the number of kernels required to represent data settles, the only further improvement will be within accuracy of kernel placements in space. Fig 3b demonstrates how this number of components impacts the AIC score with the best performing values concentrated at relatively low values. The best accuracy solutions for different set sizes are illustrated in Fig 4, evidently, the best solutions reside around 0.02 permissible error. Low sample sizes show more variance with two values being relatively large (though they may simply be outliers) while the other values showed a consistent slight dip in the level of permissible error as the sample size is increased.

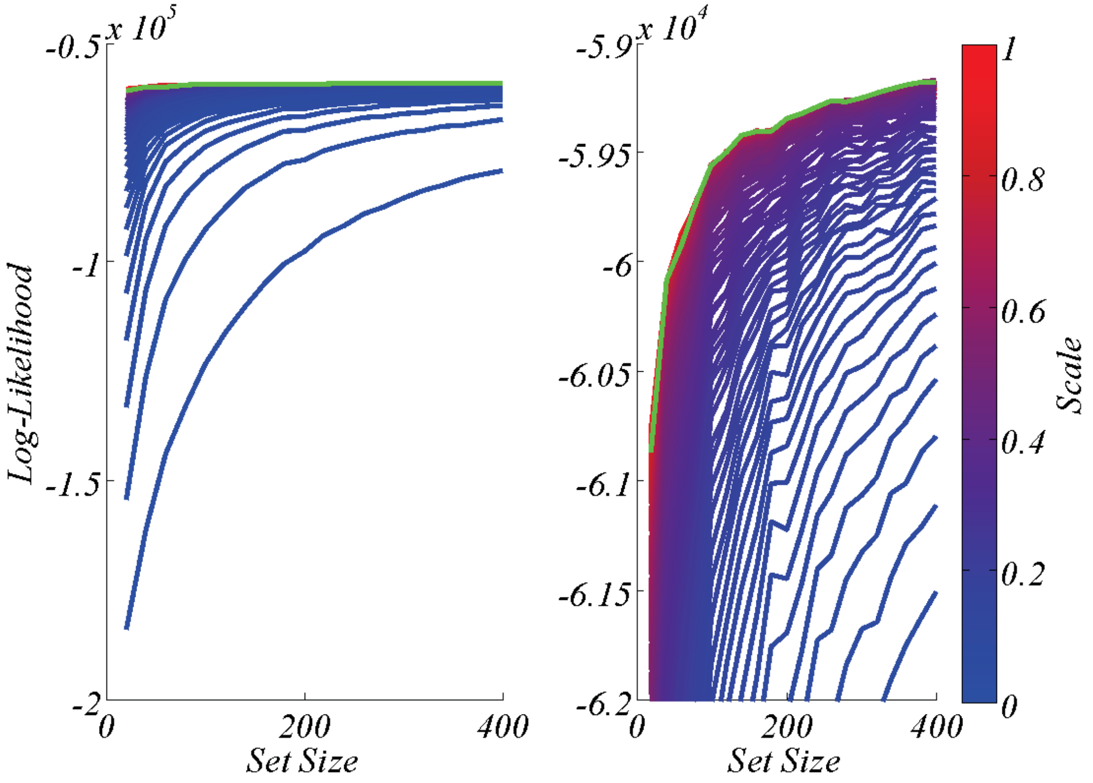


Figure 2.4: Comparison between oKDE and traditional KDE with different smoothing scales when the locations and the number of kernels correspond to data. The figure on the left show log-likelihood values with green curve representing oKDE. The figure on the right show zoomed-in perspective. Traditional KDE performs equivalently under relatively high bandwidth values.

2.5 Comparison between two KDE methods

2.5.1 Smoothing setting

In order to illustrate the working of the two KDE methods and identify where the main weaknesses lie we have isolated individual parameters of the two mixture models by profiling these parameters. We started with the traditional KDE and used the best oKDE accuracy across increasing set size as identified previously. Fig. 5 shows that most of the scaling values had a worse likelihood value compared to oKDE. In order to further the comparison, we performed Wilcoxon signed-rank test across different set sizes and parameter values (Fig 6). In this graph all values between -2 and 2 indicate non-significant difference while values outside this boundary show significantly better fits in favour of oKDE when values are positive and in favour of traditional KDE when values are negative. The decrease in the best smoothing parameter value was observed with larger number of data points as observed previously. Importantly, regardless of the number of kernels available to approximate data, traditional KDE performed significantly better only in very small number of samples (e.g. 60). This may be due to clustering which results in a higher variation between samples due to generalisation that clustering provides, though it may also be due to the best accuracy value being an outlier. A similar characteristic can be observed at the area between 250 and 300. It has a brighter colour, but accuracy is slightly above a general curve (Fig. 6). Note, that regardless of the sample size obtained for data, traditional mixture model can only perform equivalently to oKDE since matching kernels to data reduces the comparison to the accuracy of bandwidth. The approximate of the optimal bandwidth has not produced significantly better results and the only noteworthy difference is much fewer number of components required to equivalently represent target distribution. Therefore, carefully selected locations of kernels as well as their variances are more important than the values set for bandwidth or the number of kernels themselves.

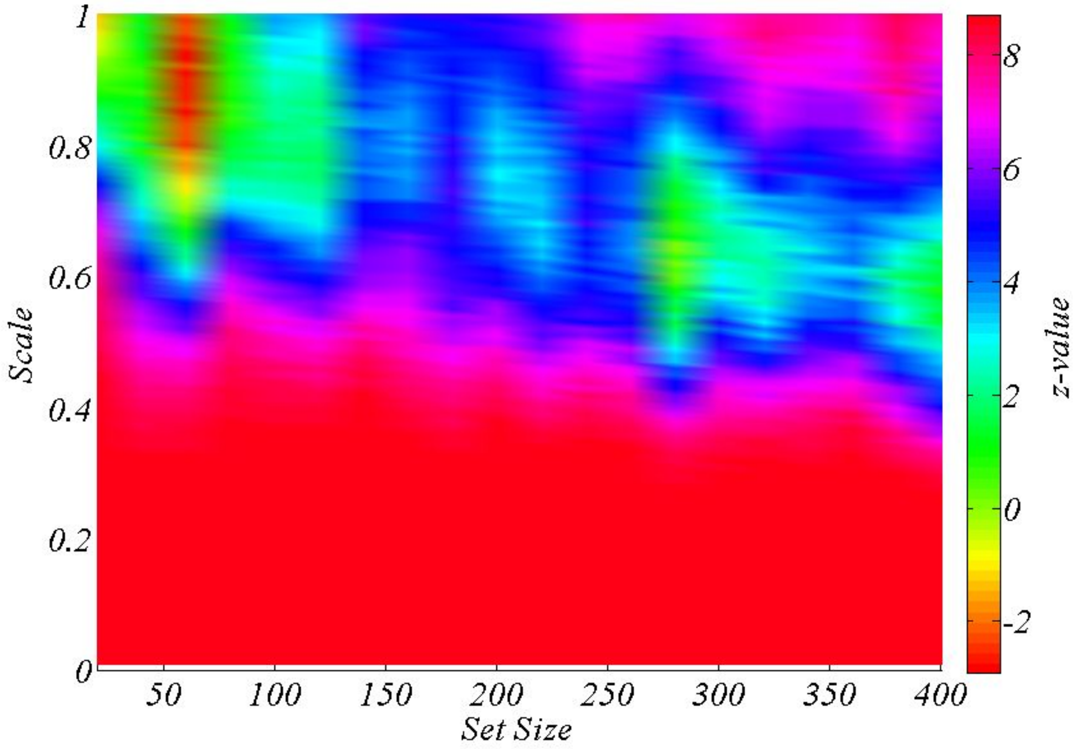


Figure 2.5: This figure illustrates how z-values between two KDE methods change as a scale of bandwidth and set size changes when the number of kernels and their locations correspond to the data. Z-values were produced by Wilcoxon signed-rank test as a function of set size and smoothing scale. Positive values favour oKDE while negative values favour traditional KDE. Colours corresponding to values between -2 and 2 indicate non-significant areas. Overall, the results favour oKDE but there are bandwidth values that produce non-significant differences between two methods.

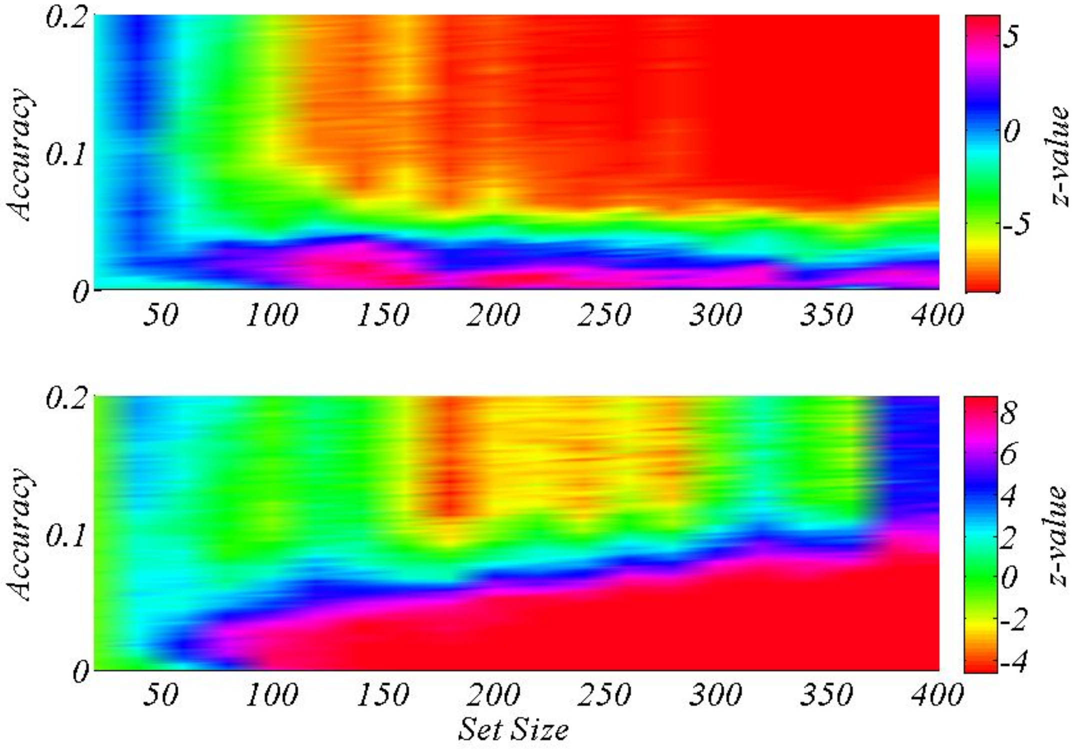


Figure 2.6: *This figure illustrates how z-values between two KDE methods as permissible error and set size changes. Z-values were produced by Wilcoxon signed-rank test as a function of set size and compression accuracy in terms of log-likelihood and AIC score. Positive values favour oKDE while negative values favour traditional KDE. Colours corresponding to values between -2 and 2 indicate non-significant areas. The top figure show log-likelihood comparisons and indicate that there's no significant difference when a small sample sizes are available but from around 80 samples there are permissible levels of error that produce significantly better performance favouring oKDE. A bottom figure illustrates AIC scores and how the change in the number of components used to approximate distribution impact which mixture model is favoured. Since the number of components used for oKDE increases logarithmically while traditional KDE has a linear increase, the area favouring oKDE expands as the set size grows.*

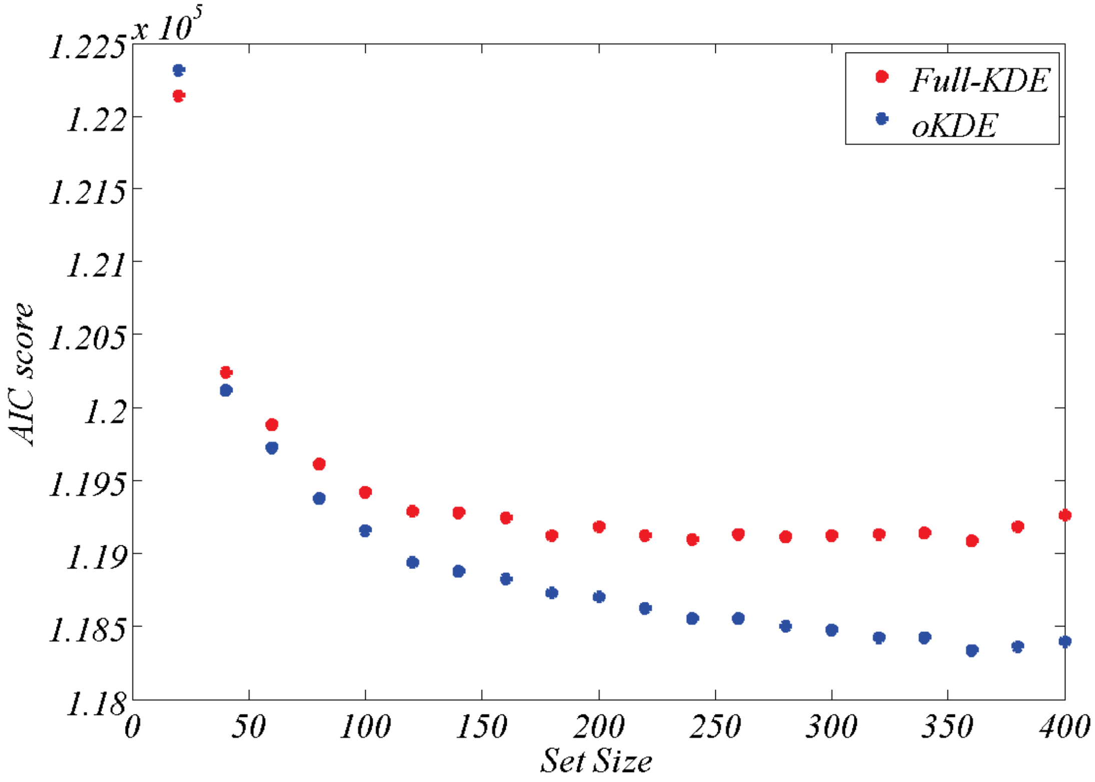


Figure 2.7: This figure displays AIC score under the best performing settings of each method. 50 and more samples consistently favour oKDE over traditional KDE.

2.5.2 Accuracy

We followed the same procedure for accuracy parameter to repeat likelihood variation analysis as in previous study. The results in Fig. 7 show hardly any difference between two methods and permissible error when a low number of samples are being used but that changed with larger set sizes in favour of oKDE. Only relatively low levels of permissible error led to significantly better performance when oKDE was used. In general, a significantly better performance can be observed from around 80 samples. However, a confidence area where significance can be observed gets smaller as set size gets larger and larger. This further suggests our earlier claim that two methods would hardly differ at large set sizes. Notice that oddities observed in previous study disappears strengthening our claim that these patches were likely to be outlier outcomes.

	Task	**Participant**	**Number of items**	**Condition**
1	Feature	3	18	Absent
2	Feature	6	6	Absent
3	Spatial	5	3	Present
4	Conjunction	3	6	Present
5	Spatial	2	18	Present
6	Conjunction	8	12	Absent

Table 2.1: Datasets used to compare two mixture models.

We believe that clustering of data within oKDE generalises about PDF functions better which leads to a better performance overall at smaller sample sizes. This clustering that creates larger components with larger weights and bandwidths better captures missing points within data. However, such benefit should drop as sample sizes increase. Then, this generalisation is all about efficiency in having fewer components. We looked how the number of components impacts the support on KDEs using Akaike information criteria score (AIC) and compared these scores with traditional mixture model. The number of components corresponds to the degrees of freedom. The oKDE had settings that outperformed traditional traditional KDE in majority of cases (Fig.

8) as AIC was smaller for oKDE in these cases. On the other hand, oKDE varied in its performance permissible error parameter and only relatively high accuracy has consistently outperformed traditional mixture model at larger set sizes. Notice the traditional KDE has saturated in performance at around 200 components and any further improvement only compensates for an increase in the number of components (Fig. 9). As the number of samples increase, the number of kernels matching data no longer is a viable choice as log-likelihood improvement becomes increasingly small.

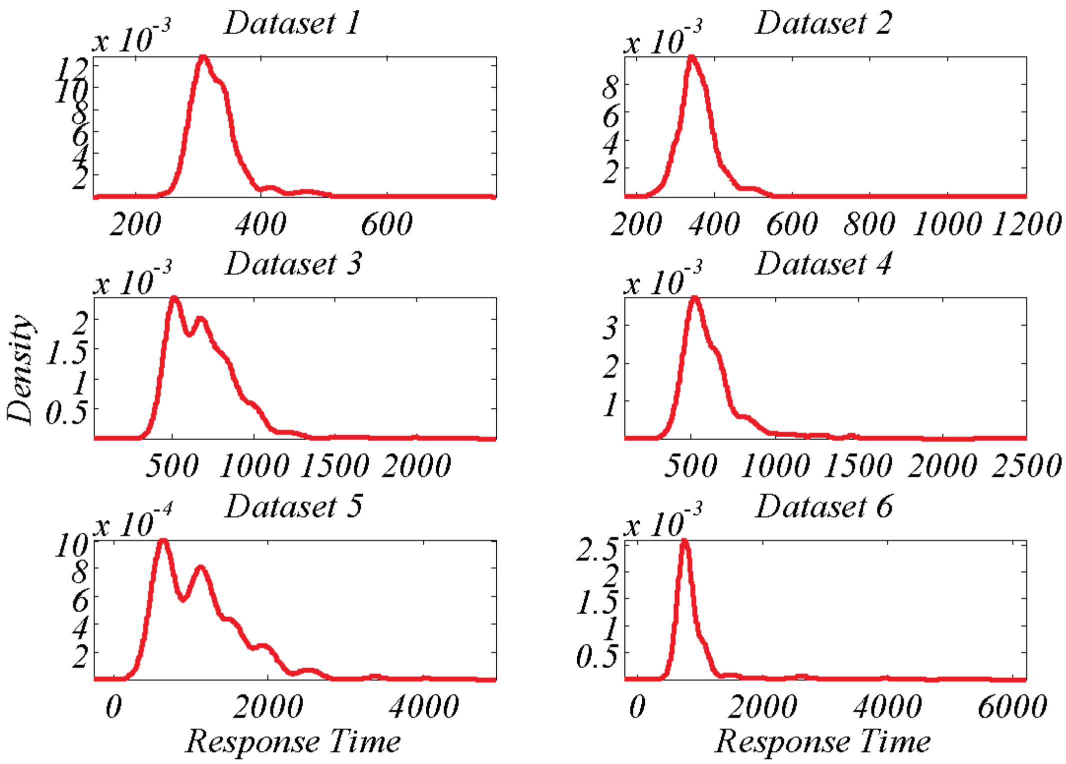


Figure 2.8: Six distributions that were pseudorandomply (random participants but variation in set sizes and tasks was enforced) selected to be fitted.

2.6 Real Data Example

To affirm our findings of the two methods we have used real human data taken from publicly available dataset (Wolfe, Palmer & Horowitz, 2010). This dataset has data taken from several participants performing visual search task with varying number of distractors. There were tasks that varied in difficulty, with colour being the easiest search task, a combination of colour and orientation being the average and 2s among 5s being the hardest task. 6 datasets were pseudorandomly (random participants but enforcing variation in display size and present vs absent conditions so the choices would represent a wide range of distributions) chosen, 2 from each task and random participants, 2 from each item set size, out of 6 datasets 3 were from target present and 3 from target absent conditions (see Table 1 and Fig 10). These datasets varied substantially, some were very fast (a, b) but one had a few very unlikely data points in the tail (b). There also were wider differently skewed distributions (c, d) and very wide more ones (e, f).

Each dataset we bootstrapped 100 times generating new datasets that had the same set sizes as previously of 20 to 400 and 20 was an increment step. This time generated PDFs were tested by original dataset utilising likelihoods as previously. As expected, in this experiment oKDE outperformed traditional KDE at smaller set sizes as well. However, the likelihoods of two methods converge at larger set sizes (Fig 11) and no significance can be observed there some datasets. Fig 12 shows all 6 examples with AIC score and none of the examples favour traditional KDE and Wilcoxon signed-rank test confirms this observation across all setups retaining significant values. Notice that AIC score of traditional KDE approaches the minimum roughly at 150 components A few datasets show an increase in AIC score suggesting that there are excessive degrees of freedom (number of components) which provide improvement in fitness that does not compensate for additional components required to represent data. Such bend does not exist in the case of oKDE which is expected as the number of components for each dataset settles where the number of components varied from 3 to 8 (see Fig 13) and is favoured by AIC.

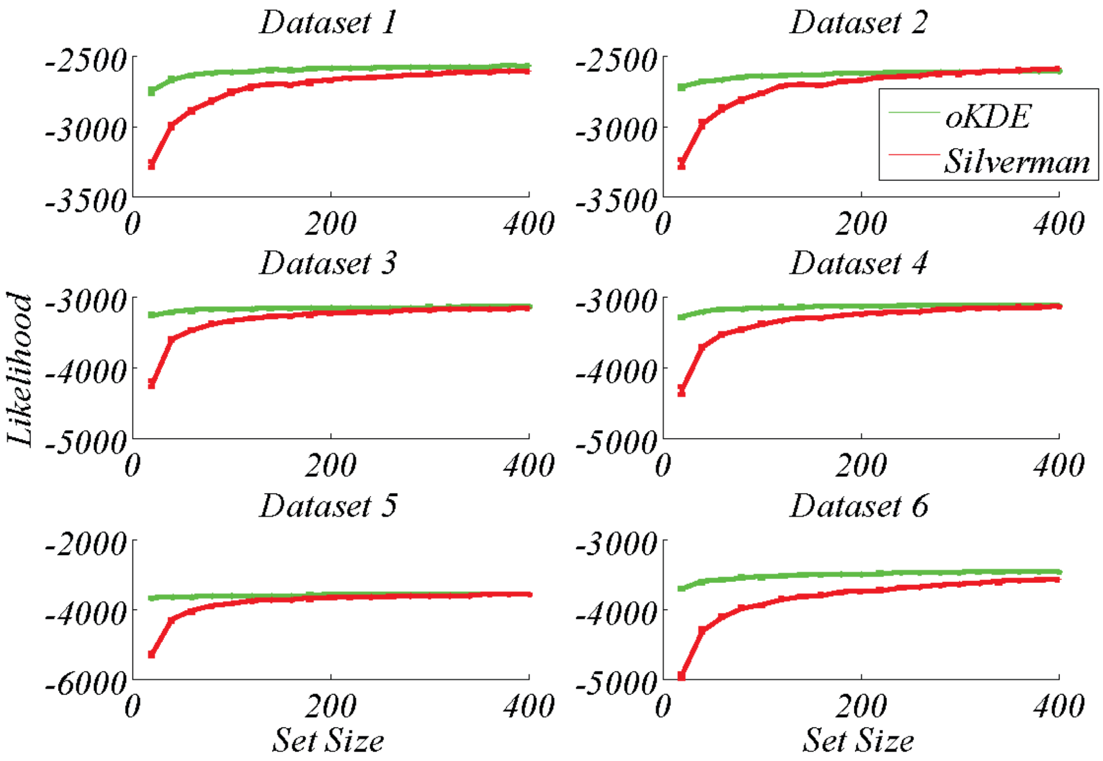


Figure 2.9: Likelihood fits over 6 distributions where green curve is produced by oKDE and red curve by traditional KDE. Traditional mixture model under a couple of distributions show similar performance when many components are available to approximate target distribution.

2.7 Discussion

In this section we introduced a novel online kernel density estimator (oKDE: Kristan, Leonardis, & Skočaj, 2011) for non-parametric description of unknown probability density functions (PDF). We propose this method as one of the best available methods to describe response time (RT) distributions and designed a series of experiments to illustrate its performance in relation to a more commonly used Silverman's KDE implementation (Silverman, 1989). Numerical assessment was performed on both methods to identify the best respective parameter configurations for the two methods and evaluate their efficiency in describing RT data in relation to the size of the dataset. We believe that this introduction of the oKDE method for approximating RT-distributions will contribute to the available methods for describing this essential dataset.

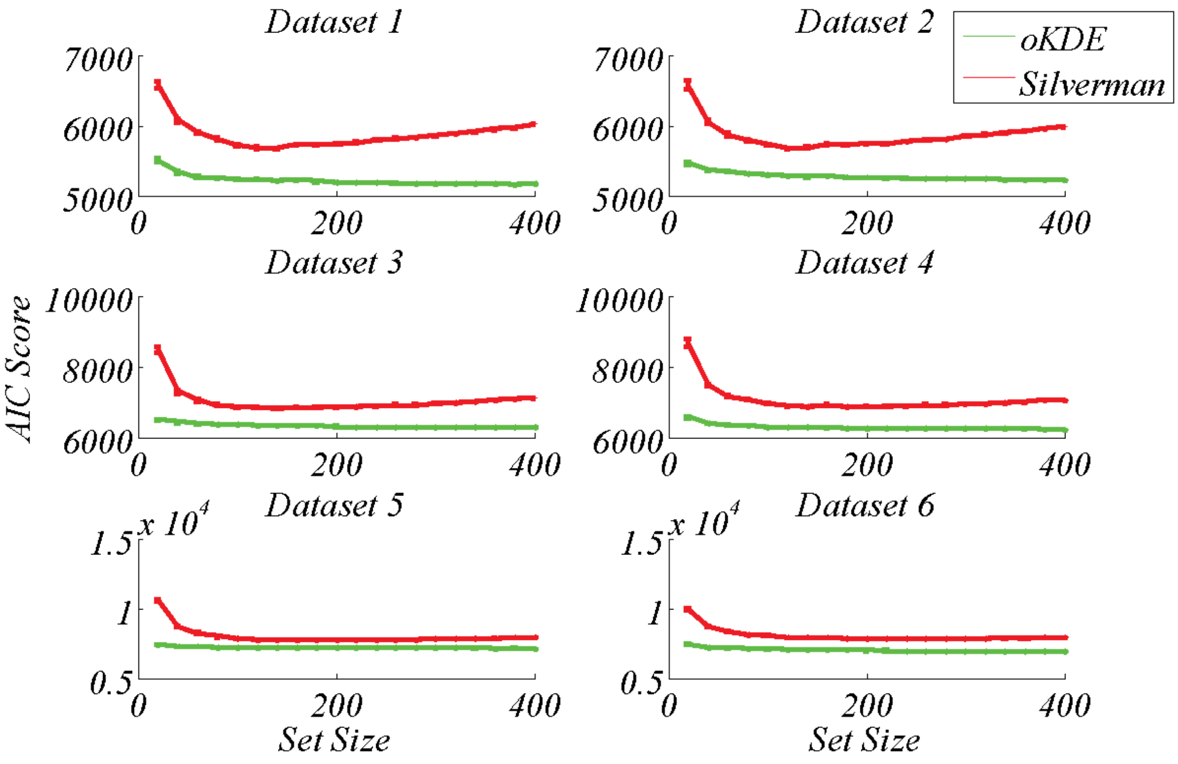


Figure 2.10: AIC score of 6 distributions where green is oKDE while red is traditional KDE. In all cases oKDE shows a better performance.

Earlier we have mentioned that psychologists rarely exploit KDE methods and in those rare occasions mainly traditional KDE is being used while there is a number of alternative and better methods. In this study we showed how oKDE performs in relation to traditional KDE in various circumstances. For the purpose of comparison, firstly we established the best model descriptions for both methods. Traditional KDE requires setting the scale parameter of the bandwidth to properly smooth the pdf approximation which tends to be higher with smaller datasets. oKDE on the other hand approximates the optimal bandwidth thus it does not require identical assessment. However, it has accuracy parameter which determines how much precision is valued over the compactness of the KDE. It clusters components based on this accuracy threshold and the KDE of the collection of larger components is formed to represent approximation of underlying PDF. We found that the optimal accuracy drops with the growing size of the dataset, though that change was very small. A general guide for a good starting choice would be to set it to 0.02 and decrease or increase accordingly.

Using the identified best parameter configurations for both KDE methods we performed numerical comparisons by manipulating one configuration at a time, constraining the other method at its best set-up for give dataset size. Overall findings show that oKDE performs significantly better than Silverman's KDE from ~50 datapoints. This superior performance drops with the growing dataset but it also increases the number of components that are used to represent traditional KDE. When the number of components used is considered, we find that oKDE is substantially more compact and retains a similar or better approximation that traditional KDE. This is a vital feature because matching the number of components with size of the dataset leads to substantial computational costs to evaluate other data against the same KDE. This indicates that a right choice for the number of kernels is an essential feature of efficient KDE method. However, a more important factor in producing a superior outcome relates to the locations components are placed as well as flexible bandwidths across components allowing to adjust for local sparsity. These features make significant difference in the performance of the two methods and precisely due to them; oKDE can use fewer kernels and still perform better than traditional KDE on a number of settings.

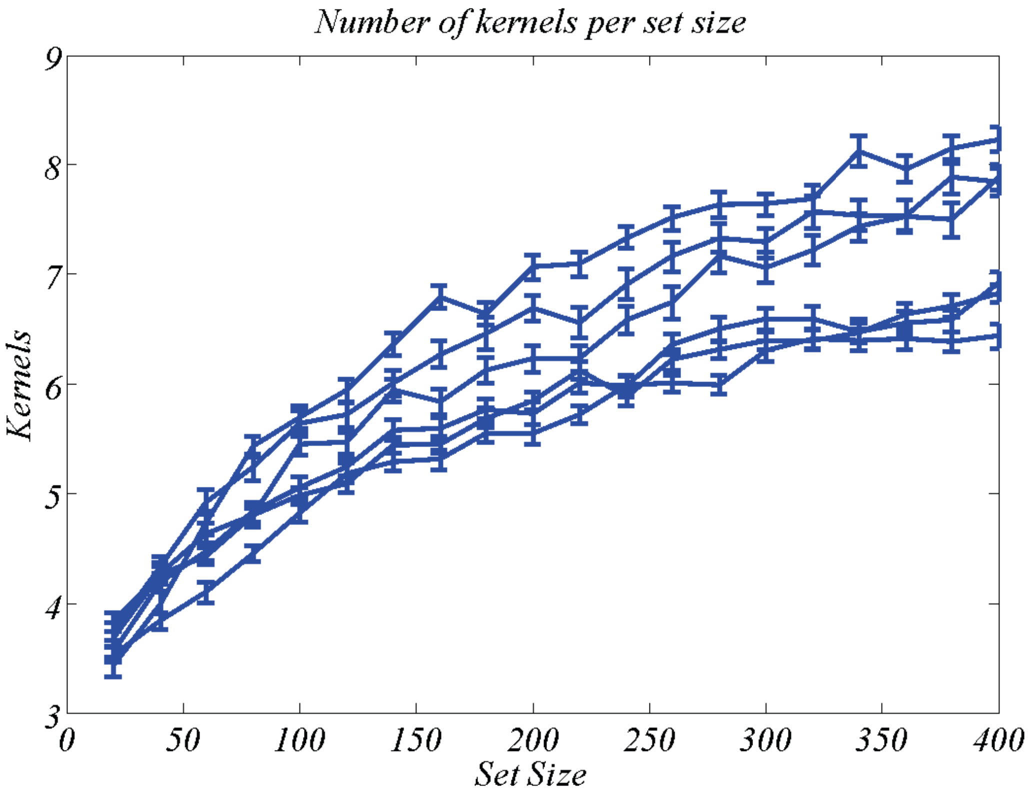


Figure 2.11: The number of kernels as function of set size. In general, 5-8 well placed components is enough to properly represent respond time distributions.

Additional very important observation is in the fact that oKDE improves approximations for smaller datasets. A common perception for requirement of many trials to perform distributional analysis is a general factor in ruling out analysis of RT-distributions. We observed that 100+ datapoints may be satisfactory to produce a good PDF approximation of the RT-distribution. However, this number varies depending on the task participants perform as well as how individual participants perform themselves. The oKDE method could describe better the data within the tails of RT-distributions. It could be argued that QML (Heathcote, Brown, & Mewhort, 2002) method is a more suitable than KDE because it divides cumulative function into equally likely bars. As a result, it deals better with increasing sparsity within tails. In other words, when tails are wide but unlikely, a common bandwidth provides a poor approximation of the probabilities at the tail and an extension for varying minimal bandwidth depending on which part of RT-distributions is being approximated would further improve their approximations. There are some KDE implementations that attempt to generalise data when the variance of the data changes (Wand, Jones, 1995; Cwik & Koronacki, 1998; Vincent & Bengio, 2003) which could be incorporated within oKDE. An alternative solution would be to replace Gaussian or Epanechnikov kernels with skewed kernels such as Poison (Wang et al., 1996; Byers & Shenton, 1999). Skewed kernels could be better suited option to represent RT because both distributions can represent skewness properties. Most likely such kernel choice would further reduce the number of kernels required to represent RT-distributions improving efficiency of the evaluations and comparisons.

KDE is a non-parametric method that approximate unknown PDF functions. This generally implies that an approximated distribution will not be normally distributed and general statistical methods assuming normality would fail. However, an approximation of PDF equips researchers with a possibility to compare datasets that have various distributional shapes. Distance measures such as KL-divergence or Hellinger distance become available for approximation. However, currently KDE application is still limited for RT analysis because researchers are often interested in the comparison of groups and no aggregated KDE method is available. One especially attractive feature of QML is its ability to generate a super participant. However, a possible solution may rest within approximation of distance metrics such Hellinger because Hellinger comparison is only available when a common denominator is found (Kristan, Leonardis, & Skočaj, 2011). This common denominator may also be a solution for aggregating RT distribution from different participants. Further mathematical research is needed to clarify and develop a method for determining a superparticipant for more practical use of KDE methods.

To sum up, we have introduced a novel KDE method that approximates RT-distributions in order to aid psychological analysis of cognitive processes. We provide a good illustrative guide for choosing appropriate settings for the configuration of the two methods and show that this oKDE is a better choice than traditionally used KDE method. However, there is further room for improvement of approximations for the RT-distributions that would enhance the description of the data sparsity at the tails as well as the right choice of the data size provided the knowledge of expected RT-distribution of the experimental task. Finally, an addition of appropriate statistical tests for comparison of different RT data approximations would increase the attractiveness of the method and development of KDE aggregation for between population analysis would be welcome as well. Moreover, while we have limited this study to RT-distributions, it has much wider research applicability and we will show two additional applications for KDE methods in the next two chapters.

Chapter 3:

Robust Likelihood

Abstract

An introduction of KDE approximation has additional properties that can be employed for statistical robustness. This chapter proposes a novel approach to deal with data that is misrepresented by the approximation of probability density functions (PDF). It is well established fact that RT-distributions are contaminated and that this negatively affect parameter estimation of the models. Various approaches showed that general practices such as cutoff can aid in finding the correct solutions but frequently fails in properly identifying these outliers. Moreover, RT-distributions are known to contain hidden contaminants that are within tail but do not correctly represent a processing time of the given task. The usage of known robust methods is still limited possibly due to limited development in statistical inference, especially in Bayesian inference. However, there is a growing interest in these methods though mainly outside of psychology field. In this research we aim to introduce robust statistics as a good alternative to traditional maximum likelihood estimation (MLE) approach. We present a novel robust likelihood method that behaves equivalently to traditional MLE method when the model is well defined but contains robustness properties similar to robust methods when that is not the case (i.e. presence of contaminants).

3.1 Introduction

In previous section we have proposed the use of one of the best KDE methods for approximating analytically intractable models. However, KDE simply provides with a more accurate unknown model description than a number of methods are generally used. Ratcliff (1978) showed that using moments fails to accurately capture skewness and kurtosis and suggested using ex-Gauss as intermediate simpler model. Heathcote, Brown & Mewhort (2002) and van Zandt (2000) further extended the literature by introducing quantile (QML) and cumulative density estimation (CML) methods. All these methods use some goodness-of-fit function for estimating the best fitting parameters. Arguably, MLE is the most commonly used goodness-of-fit measurement for fitting models to RT-distributions which was used for ex-Gauss parameter estimation (Hockley, 1984; Ratcliff, 1978) as well as parameter estimation using quantile maximisation (Moran et al., 2013). An alternative commonly used method is Least Squares Estimation (LSE: Usher & McClelland, 2001; Bogacz & Cohen, 2004; Tsetsos, Usher, & McClelland, 2011) which was shown to be the most efficient method for models with intractable PDFs (van Zandt, 2000; Heathcote, Brown & Mewhort, 2002).

Another issue some of these methods attempt to address is robustness. It is a very intriguing property that could potentially provide the solution for response time distribution modelling because RT data which is often plagued with outliers and general contaminations, whether it appears as fast, slow or hidden within main distribution (Miller, 1991; Ulrich & Miller, 1994). Some contaminants such as fast responses can be easily recognised as long as it does not overlap with the RT distribution, but slow and especially hidden are trickier or simply impossible to distinguish. These contaminants can badly affect parameter estimation with bias being a likely outcome (van Zandt, Colonius, & Proctor, 2000). QML also was designed with robustness properties in mind showing better robustness properties than CML.

The most successful approach produced a workaround by modelling the contamination itself and employed MLE over that model (Ratcliff & Tuerlinckx, 2002; Ratcliff & McKoon, 2008; Wagenmakers et al., 2008). However, extending models by adding another component to the resulting PDF creates a number of theoretical issues on modelling RT-distributions. Firstly, it assumes how contaminants affect RT-distributions and research in this area appears to be limited at best. Ratcliff and Tuerlinckx (2002) proposed using uniform noise for hidden contaminants but there's not theoretical backing over this suggestion. Moreover, it ignores the fact that the model may in fact be able to accommodate contaminants within itself by modifying the parameters accordingly. In these circumstances expecting parameter recover is a wrong initial expectation because the model is too flexible. Robust methods are a better choice for recovering parameters for cognitive models of RT-distributions because they will identify a possibility that a model is too flexible. If the resulting outcomes are equivalent to the outcomes produced by standard parameter estimation approaches, then it is flexible enough to account for these assumed contaminations. In fact, this is a potential extension to the assessment of the model complexity in terms of its flexibility since models can contain the equivalent number of parameters but one use more flexible functions.

Another purpose for employing more robust goodness-of-fit is due to our introduced KDE method. Basically, many of psychological processes are modelled using stochastic models and these often have no known likelihood function as such problems quickly become mathematically intractable. In this modelling enterprise the model's PDF needs to be created via Monte Carlo sampling which sometimes leads to a misrepresentation of the model's PDF particularly for unlikely simulation outcomes (i.e., unlikely data points). Unfortunately, response time distributions associated with these psychological processes are skewed while our used KDE method was the Gaussian mixture model. This means that sporadic RTs within tail have lower density and simulations may fail to produce any response at the low end of the RT distribution completely. In other words, the KDE constructed from such a sampling error assigns small probabilities to these data points; in fact even smaller than implied by their presence in the dataset. But other sampling runs with similar parameter settings may assign reasonable probabilities to these data points. Such variations in sampling can lead to large problems in finding optimal parameters during the parameter estimation.

The following chapter will introduce the methods we used to assess our robust likelihood method and how they compare to it. We designed multiple experiments emulating fast and slow but hidden contaminants. However, we do not introduce additional noise component and perform no data truncation (Ratcliff & Tuerlinckx, 2002). We use these contaminated distributions for showing how our method behaves in relation to the likelihood as well as Hellinger distance.

3.2 Robust Divergence Functions

It is relatively unknown in psychological literature that likelihoods and square distance belong to a broader class of goodness-of-fit power functions for finding the best solutions (Cressie & Read, 1984). These functions are also known as divergence functions that provide with comparison measurement between two density functions. The likelihood equivalent divergence functions is commonly known as Kullback-Leibler (KL) divergence. In general, various power divergence measurements have a trade-off between robustness and efficiency where efficiency means that less data is needed to recover proper parameters. Two most common goodness-of-fit, likelihoods and square distances, methods in Psychology are both efficient methods (Lindsay, 1994) and only these evaluation criterias were generally compared (van Zandt, Colonius, & Proctor, 2000; Ratcliff & Tuerlinckx, 2002). The observations found in their work is not surprising since the least squares implementation is represented by Pearson's χ^2 which is more efficient than likelihood equivalent KL-divergence. Therefore, it recovers parameters better when contaminants are removed or incorporated within a model. In fact, it is well understood how these methods differ in their asymptotic properties with MLE and χ^2 being the best suited when the model is a good representative of the data.

However, other very useful power divergence measures such as Hellinger distance, to the best of our knowledge, were never used in psychology, even though it is inherently robust to the contaminants present in the data. This metric is a special case of divergence measures because it is symmetric, i.e. a reverse comparison between two density functions produce identical value. The most important feature Hellinger distance possess is an inherent robustness to the contaminations when the model is poorly specified. It's limited usage may in part be explained due to Hellinger distance requiring known density functions It does not have a convenient likelihood form as KL-divergence does, which implies that PDF approximation for the data is required. Overall, robust methods such as Hellinger are influenced substantially less by contaminants than commonly used methods within psychology. Bogacz and Cohen (2004) did observed that LSE methods can be adapted to be more robust by modifying normalising factor. In fact, a lesser known Neyman's χ^2 is robust to the outliers which has a different normalisation factor than Pearson's χ^2. However, modifying this normalisation factors trades efficiency in exchange of robustness.

Generally, the Hellinger distance is rarely used outside statistical sciences despite its inherent robustness properties. Methods were also developed, albeit delayed compared to MLE, for statistical testing, for example, there is a likelihood ratio equivalent test designed for Hellinger distance (Simpson, 1989). Hellinger distance does require for a known PDF functions though more recently a number of methods were developed for Hellinger approximation (Lu, Hui, & Lee, 2003; Karlis & Xekalaki, 1998; 2001; Kristan, Leonardis, & Skočaj, 2011). However, its usage was still hampered by increased interest in Bayesian methods in the 90s due to growing available computer resources. Regardless, more recently there has also been growing interest in robust Bayesian inference by replacing likelihood function with various power divergence measures such as Hellinger distance itself (Hooker & Vidyashankar, 2014; Liu et al., 2014). Unfortunately, to the best of our knowledge there's still no method available for model selection that would take into account model complexity when only the best solution is known. Finally, Hellinger distance compared to efficient method behaves poorly with low number of samples (Basu, & Lindsay, 1994; Lu, Hui, & Lee, 2003; Patra, Mandal, & Basu, 2008) which is a common issue in psychology but the same issue has been in RT-distribution analysis thus we expect the sample sizes for RT experiments to grow.

All in all, we opted to introducing a likelihood modification that would behave equivalently to the likelihood method when the model is well-defined but contain similar robustness properties as Hellinger distance. For this purpose we used Hellinger distance as performance evaluation of the introduced modification to the likelihood function. Hellinger distance is expressed as follows:

$$HD(D, M(\theta)) = 2\int(\sqrt{(D)} - \sqrt{(M(\theta))})^2;$$

Where D represents an approximated PDF of the process that produced data and $M(\theta)$ is a PDF of the model's output given θ parameters.

3.3 Maximum Likelihood Estimation

Suppose we have data set $X = \{x_1, x_2, ..., x_{n-1}, x_n\}$ and a model $M(\theta)$ where θ is a parameter set of the model M. Then the likelihood is computed as follows

$$L(X|\theta) = \prod_{i=1}^{n} M(x_i|\theta).$$

$L(X|\theta)$ is also known as likelihood function and maximising it is finding the best solution. Likelihood functions can be categorised into three groups based on the structure of the data.

The simplest form of data contains only discrete values. This generally corresponds to a frequency of possible outcomes. Let's denote a set of possible outcomes as $S = \{s_1, s_2, ..., s_{m-1}, s_m\}$, then all outcomes in this set will have some probability value that corresponds to a frequency of outcome s_j in the data set. This could be a number of correct and incorrect responses by a participant. On the other hand, our task could be any two alternative choice task which then provides us with data on correct/incorrect responses for both choices. This could be expanded with even more possible choices.

Such data would require a corresponding model to explain the observed differences and predict possible different outcomes if more data was gathered. For data sets that have only two possible outcomes we would construct some binomial distribution to represent our observed data. For more outcomes we would further extend to a multinomial distribution. For a relative example assume that our chosen model can produce all m outcomes contained in the set S. Also, our data contains at least one example for each outcome in S with a varying frequency. Furthermore, our model has a known probability mass function that gives some probability value for each possible outcome $P(S) = \{P(s_1), P(s_2), ..., P(s_{m-1}), P(s_m)\}$. Now we can use these probabilities to construct mass function directly into an equation of likelihood function $L(X|\theta) = \prod_{i=1}^{n} P(x_i = s_j) = \prod_{j=1}^{m} P(s_j)^{P(X=s_j)}$. Continuous data in that regard does not differ much from discrete data. It simply is an extension from a fixed number of outcomes to an infinite number of outcomes and probability for the categories is replaced with probability for the density forming a probability density function.

The biggest different from the description provided above is when data is mixed. A slightly more straightforward example of mixed data could be a response time data generated from two alternative choice tasks. In this case we have two continuous distributions X and Y containing response time data. For illustration let's return to discrete example, each possible outcome has only one associated value. If we were to express this in a probability distribution, we would use dirac delta distribution and rewrite our equation as follows

$$L(X|\theta) = \prod_{j=1}^{m} P(s_j)^{P(X=s_j)} = \prod_{j=1}^{m} P(s_j) \prod_{i=1}^{X=s_j} \delta_{s_j}(x_i) ,$$

this means that in a continuous space this outcome has only one value with a probability of one and the rest has a probability of zero, thus these values did not occur. Dirac delta function has a variance of 0, but as you increase this variance the amount of values $X \in s_j$ increases and the probability drops. In other words, we can overwrite dirac delta function with any probability density function $f(x)$. Therefore, a maximum likelihood for response time distributions X and Y is

$$L(X, Y|\theta) = P(X) \prod_{i=1} f_x(x_i) * P(Y) \prod_{i=1} f_y(y_i)$$

and for any number of data distributions $L(S|\theta) = \prod_{j=1}^{m} P(s_j) \prod_{i=1}^{X=s_j} f_{s_j}(x_i)$.

3.4 Robust Maximum Likelihood

We improved the robustness of estimating the likelihood function at unlikely data points by introducing a dataset-defined threshold for the model's pdf. Note that thresholding very small pdf-values is common practice to avoid numerical issues (i.e., underflow) in the calculation of likelihood values. Consider a discrete example of two-choice decision making task with very low error rates such as one error. A model unable to produce any errors would not even be considered due to multiplicative properties of likelihood. However, if the model does not have a known likelihood function and some stochastic process of the model has outcomes that have very low chance of occurring, you would regularly fail to acknowledge the best solution. Furthermore, if a considered model differs from data only by failing to produce that single error, it is not a bad solution and does not imply that it can't produce it. Importantly, a probability of that single point is $1/n$, where n is the number of datapoints. This means that it cannot have a probability between zero and one. A viable solution to this numerical issue could be half of the minimum value: 1/(2*n*). The same reasoning can be applied to continuous data. In order to determine this threshold, we fitted a KDE to the dataset.

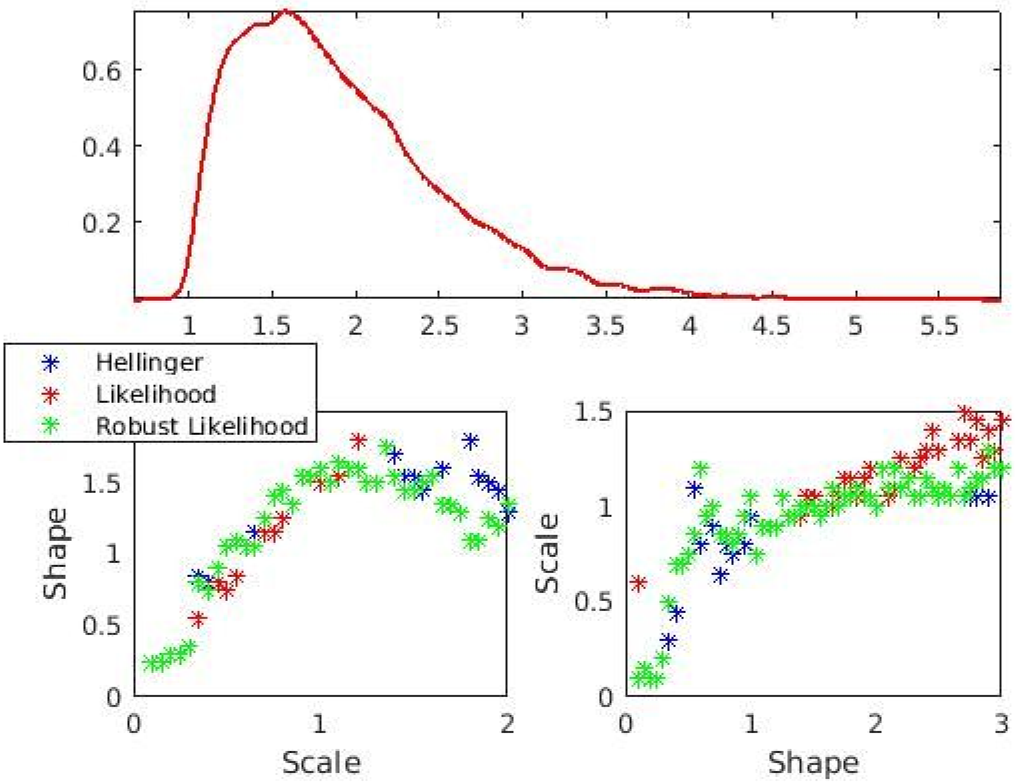


Figure 3.1: This figure show Weibull dataset when no noise was introduced and how different goodness-of-fit measures behave. Top graph illustrates the dataset while the bottom two graphs show profiled parameter values with blue stars representing Hellinger, red stars showing likelihood and green stars show our proposed dynamic likelihood. Bottom left figure show the profiled shape parameter when the scale is being changed and right figure show the profiled scale when shape is being changed. When shape is being profiled, the robust likelihood behaves similarly to both methods while in terms of scale, it behaves more similarly to Hellinger rather than likelihood.

As stated in the KDE chapter, the oKDE method initially assumes a kernel for each data point (similar to the traditional KDE-approach) and determines the smallest bandwidth for this initial KDE using an optimality criteria. This bandwidth defines a lower bound for a probability of a data point to occur. Hence, any PDF constructed for a model (KDE) should produce at least this probability for each datapoint. In other words, if the model is correct, a failure to produce reasonable probabilities for unlikely data points has to be due to the sampling error. Thresholding the model's KDE in this situation corrects this error, which means that this bandwidth forms a reasonable threshold for the model's PDF (KDE). We used Gaussian function which has the mean equivalent to data point and some bandwidth h, then a probability for non-represented points would be expressed as follows: $1/(2n)\phi(x, h)$. To finalise, an ideal solution for a given continuous dataset is data itself thus a KDE constructed from data would be identical to the KDE constructed from simulated data. Therefore, we can construct a KDE for data prior simulations, find a value for non-represented points and re-express our original likelihood function to:

$$L(\theta) = \prod_{i=1}^{n} max\ (M(\theta), \phi(0, h)/2).$$

3.5 Materials and Methods

In order to demonstrate how a dynamic thresholding of the likelihood impacts the parameter space we have conducted a series of numerical experiments. All experiments used shifted Weibull as a baseline model with scale, shape and shift being set to 1, 1.5 and 1 respectively. Using these parameters, a hypothetical large dataset containing 10000 points was generated. We followed with constructing an equally distributed parameter grid of scale and shape for evaluation while shift was retained constant throughout the study. This grid had scale varied from 0.1 to 2 and shape varied from 0.1 to 3 with a common intermediate step of 0.05, this formed a 39x59 parameter set matrix. Each parameter combination had 1000 simulated datapoints generated to construct an approximate pdf of the model. All pdf approximations were performed using oKDE with a maximum error from data being set to 0.02. For comparison purposes we used likelihood with very low numerical threshold to protect against unlikely numerical underflow ($t = 10^{-200}$) as well as proposed dynamic threshold. Both methods were compared in relationship to the Hellinger distance as a reference to how robust methods behave.

We constructed 4 case scenarios of possible contaminations present within given dataset. The first is a ground truth example to illustrate the differences in behaviour when there is no contamination other than asymptotic deviation produced 10000 random samples. Fig ? show the distribution produced by these samples and the parameter performance using the 3 outlined criterias. The top figure show the hypothetical data distribution used and the bottom two graphs show how the parameters interact. Since the goodness-of-fit measures cannot be compared directly, we looked how they affect the parameter landscape at each parameter dimension. We chose to use profiling rather than marginalising due to the same issue of comparison. Profiles are defined as maximum value across dimensions of interest: $f(X|\theta_{\theta-\theta^*}) = max(f(X|\theta^*)$; where θ^* is the dimensions we want to profile through and $\theta_{\theta-\theta^*}$ are remaining dimensions. The results show that robust likelihood behaves more similarly to the Hellinger distance rather than likelihood. A more rapid drop of the shape parameter when larger scale values are considered (left figure) can be observed for robust likelihood compared to the Hellinger. However, the main difference emerges in the right figure where the rate of change for scale parameter as shape is being increased is much higher for likelihood than the other two evaluation measures. Robust likelihood has almost identical behaviour compared to the Hellinger.

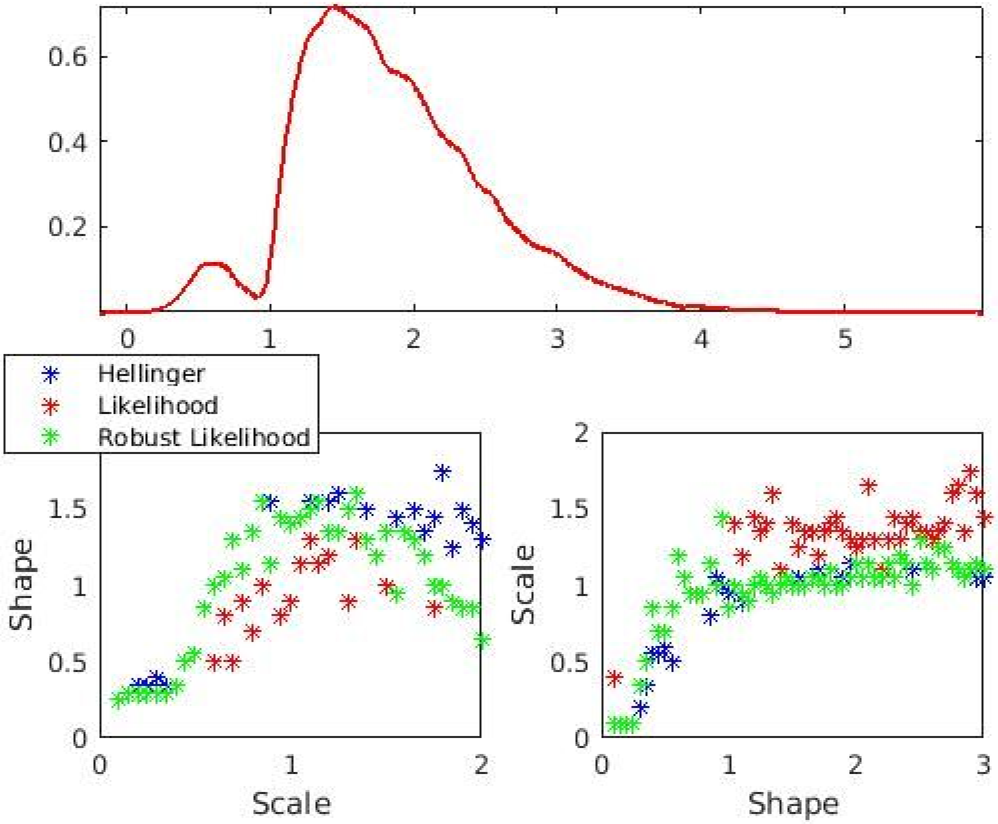


Figure 3.2: This figure show Weibull dataset when responses that are too fast are recorded and how different goodness-of-fit measures behave. Top graph illustrates the dataset while the bottom two graphs show profiled parameter values with blue stars representing Hellinger, red stars showing likelihood and green stars show our proposed dynamic likelihood. Bottom left figure show the profiled shape parameter when the scale is being changed and right figure show the profiled scale when shape is being changed. In both cases, the robust likelihood behaves more similarly to Hellinger rather than likelihood.

We considered 2 different contaminations separately and in combination. Firstly, we generated a hump of fast responses by replacing 0.05% of the datapoints with normally distributed noise which had a mean of 0.6 and standard deviation of 0.15. It is easily distinguishable noise and simple data cleaning could remove these points but it is a good standard example for robustness study. Our next contaminants introduce slow error which visibly extend tail but difficult to determine as outliers. One contaminant noise followed uniformly distributed noise as modelled by Ratcliff's & Tuerlinckx's (2002). It was an additive noise to the RT data and the frequency was assumed to be 0.05% of the time while the range was between 0 and 5. Such noise within RT data makes it difficult to determine how much of that tail is caused by unrelated interfering processes within brain and how much is a genuine processing time.

3.6 Contamination studies

Firstly, we will take a look at three different observable contaminations separately and then show how it impacts results when there's a combination of noise. Fig ? show the considered dataset with contamination by fast responses and the bottom two figures show parameter relationships in the same way when no noise was considered. The parameter relationships clearly show that likelihood deviates from the other two curves for both parameters. Moreover, it does not stay in the 1 and 1.5 intersection for scale and shape respectively while the other two evaluation criterias stay within this intersection. Robust likelihood behaved very similarly to the Hellinger except for the same area of bottom left figure as in no noise case and this difference appears to increase showing that methods while behaving similarly, they do behave differently away from the true solution.

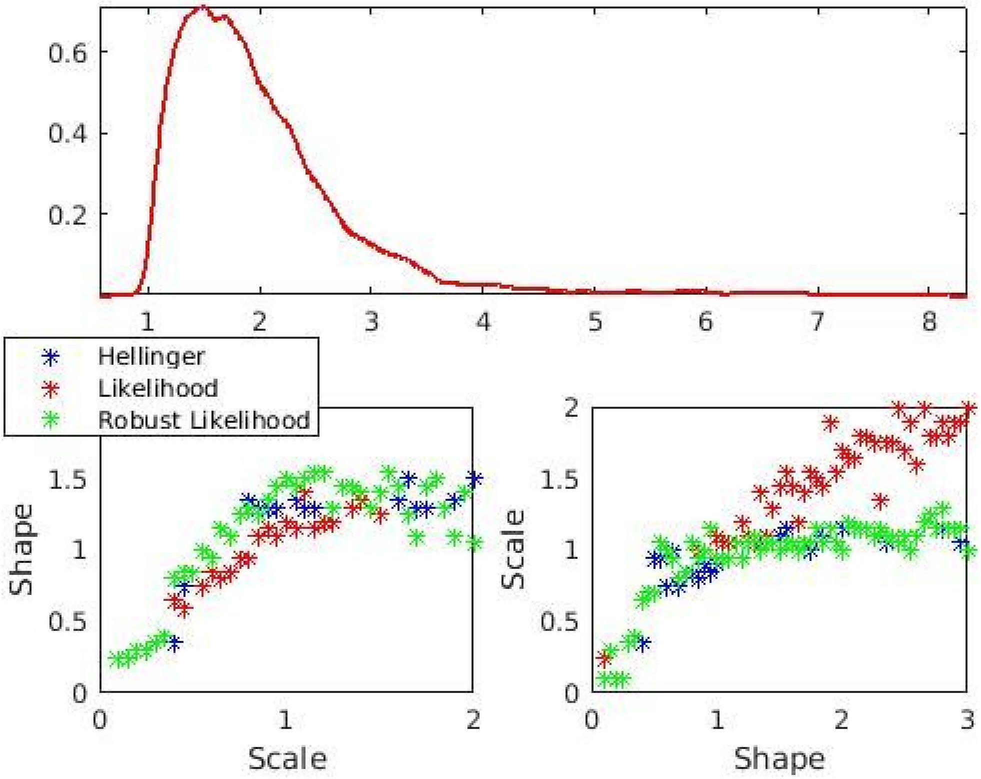


Figure 3.3: This figure show Weibull dataset when additive flat noise is added to a 5% of the dataset and how different goodness-of-fit measures behave. Top graph illustrates the dataset while the bottom two graphs show profiled parameter values with blue stars representing Hellinger, red stars showing likelihood and green stars show our proposed dynamic likelihood. Bottom left figure show the profiled shape parameter when the scale is being changed and right figure show the profiled scale when shape is being changed. In both cases, the robust likelihood behaves more similarly to Hellinger rather than likelihood.

Our next considered example introduced a identical model for contamination as proposed by Ratcliff & Tuerlinckx (2002) which adds uniform noise to small subset of the RT data. In this case the noise simply extends the length of the tail and the cutoff is not so straightforward. Fig ? show this example with distribution having a noticeably longer tail. This inclusion has changed likelihoods behaviour which still shows a preference for lower scale values but the effect is smaller. However, it has clearly change the behaviour around shape parameter because it shows a clear tendency to increase value linearly with the increase in scale. The likelihood also misses 1 and 1.5 interception as it did in the case of the fast contaminants. Importantly, the behaviour of dynamically thresholded likelihood and Hellinger distance was almost identical with little observable difference between their behaviour. Both measures did not deviate from the proper interception illustrating that dynamic likelihood possess desirable robustness properties.

Our final example contained fast responses as well as flat slow contaminants, this adds to 10% contamination which could be considered as bad data. Fig ? illustrates this scenario and results repeat an obvious advantage for Hellinger distance as well as dynamic likelihood. The typical likelihood exhibited the behaviour that is a mix two behaviour observed in previous examples. A more scarce shape parameter distribution tending to have values below 1 while correct solutions should have been around 1.5. The scale tended to go higher with increasing shape parameter but the slope is less pronounced compared to slow contaminant example but more pronounced compared to fast contaminant example. Dynamic likelihood very behaved similarly to Hellinger disparity measure with little observable difference.

Fig ? show how dynamic likelihood behave under different contaminants. There are no noticeable differences in terms of scale parameter as shape is changed but there is observable difference when shape is considered. When contaminants are being introduced, the preferred shape parameter values appear to dip below 1.5 with the highest dip being introduced under the largest 10% contamination (both contaminants). It is not a surprising result since the possibility to remove outliers via parameter fitting and robust methods depends on the model itself. If there are model specifications that can perfectly describe data (even if outliers are present), robust methods will find this model specification thus it is not surprising that there may be a Weibull model specification which somewhat describes data with outliers. However, it may also be the result caused by the difference between non-contaminant parts of the data. The results are not sufficient to declare a systematic bias and asymptotic properties of the proposed method should be studied to properly identify its sensitivity to biases. Regardless, the proposed method behaved very similarly to a Hellinger method which has known asymptotic properties thus it could be expected to have them very similar. These results show that the proposed method is a viable alternative to Hellinger distance for robust statistical inference and is a good solution for finding the best parameter settings for RT data despite the presence of contaminants.

3.7 Discussion

In this section we introduced a modification to the approximate likelihood values for contaminated data. The general idea of the suggested method emerges from common practice to apply some small probability values to non-represented data points to prevent numerical underflow. We replaced this constant thresholding mechanism with dynamic thresholding that depends on the data itself. This modification introduces robustness properties to the likelihood function which is a desirable feature for model fitting on response time (RT) distributions. We propose this method as a viable option for performing model fitting in various contexts and designed a few experiments designed to outline its behaviour. Numerical comparisons were performed on most commonly used likelihood method, a known robust Hellinger distance and our suggested robust likelihood method to find how the robust likelihood behaves in relation to these two methods. We believe that this likelihood modification for recovering model parameters modelling RT-distributions in the presence of various contaminants is a potential new method for identifying the best fitting parameters as well as applying it to Bayesian inference.

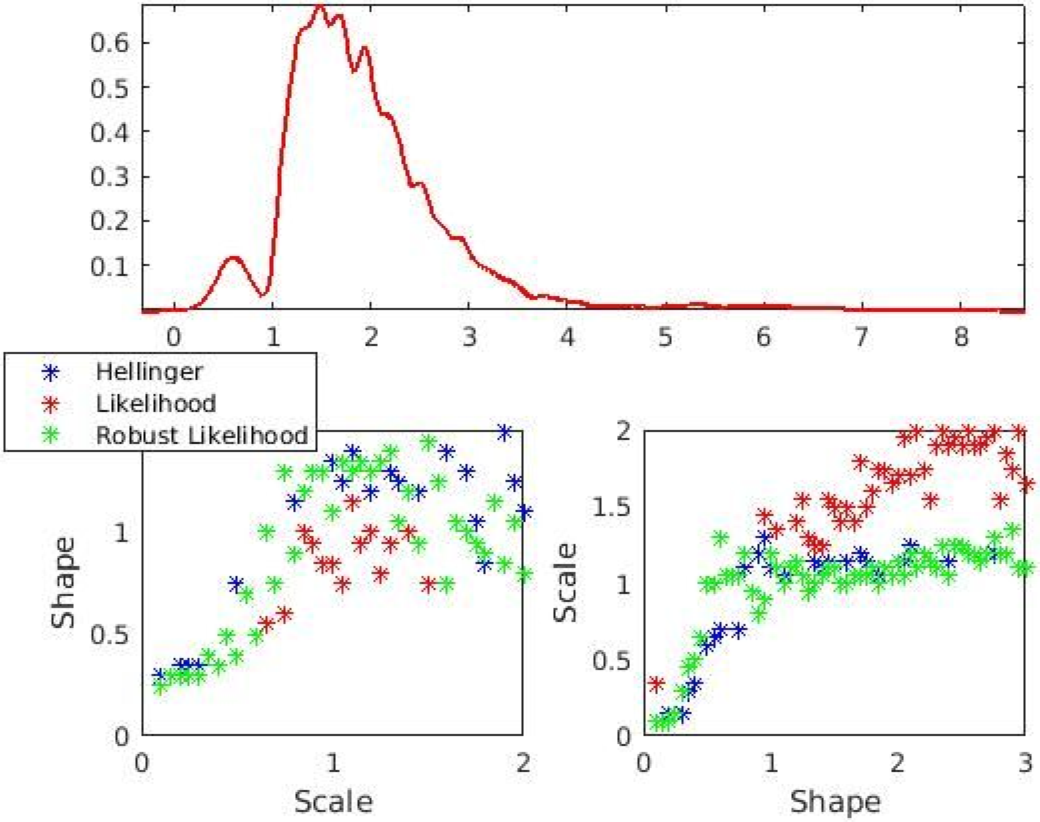


Figure 3.4: This figure show Weibull dataset when additive flat noise is added to a 5% of the dataset as well as 5% of the dataset is replaced with fast contaminants and how different goodness-of-fit measures behave under these contaminants. Top graph illustrates the dataset while the bottom two graphs show profiled parameter values with blue stars representing Hellinger, red stars showing likelihood and green stars show our proposed dynamic likelihood. Bottom left figure show the profiled shape parameter when the scale is being changed and right figure show the profiled scale when shape is being changed. In both cases, the robust likelihood behaves more similarly to Hellinger rather than likelihood.

To illustrate robust likelihood properties we have constructed a hypothetical dataset using a Weibull function. The parameters for Weibull function were set to the best parameter settings of human RT distribution in order to create a realistic as well as controlled settings. We followed the example by contaminating this dataset with typical RT contaminants of fast responses as well as slow responses (Ratcliff & Tuerlinckx, 2002). Overall, there were four contamination settings with none, fast, slow and both contaminations that were considered. We used these 4 datasets to assess how well these metrics recover original parameters by observing their general behaviour as parameters are changed. The results suggest that robust likelihood behaves almost identically to the Hellinger distance in most scenarios suggesting equivalent robustness properties in most of the situations. Both metrics show little responsiveness to the presence of the outliers while original likelihood displayed deviation from original behaviour.

No noise example result of the robust likelihood was disappointing when compared to the other two methods because the results showed equivalent behaviour to Hellinger distance. This result suggests that this modified likelihood also possess lower efficiency than traditional likelihood method. This means that larger samples are required to be able to recover original parameters compared to MLE or LSE methods (Lindsay, 1994). However, we used thresholding of 50% over the minimum probability but it most likely is not the best option. Further studies on specifying a proper thresholding criteria could improve suggested issue of efficiency since no thresholding is a true likelihood method. Moreover, our adopted KDE method is more efficient than Silverman's KDE which possibly already improves efficiency of the robust methods in general.

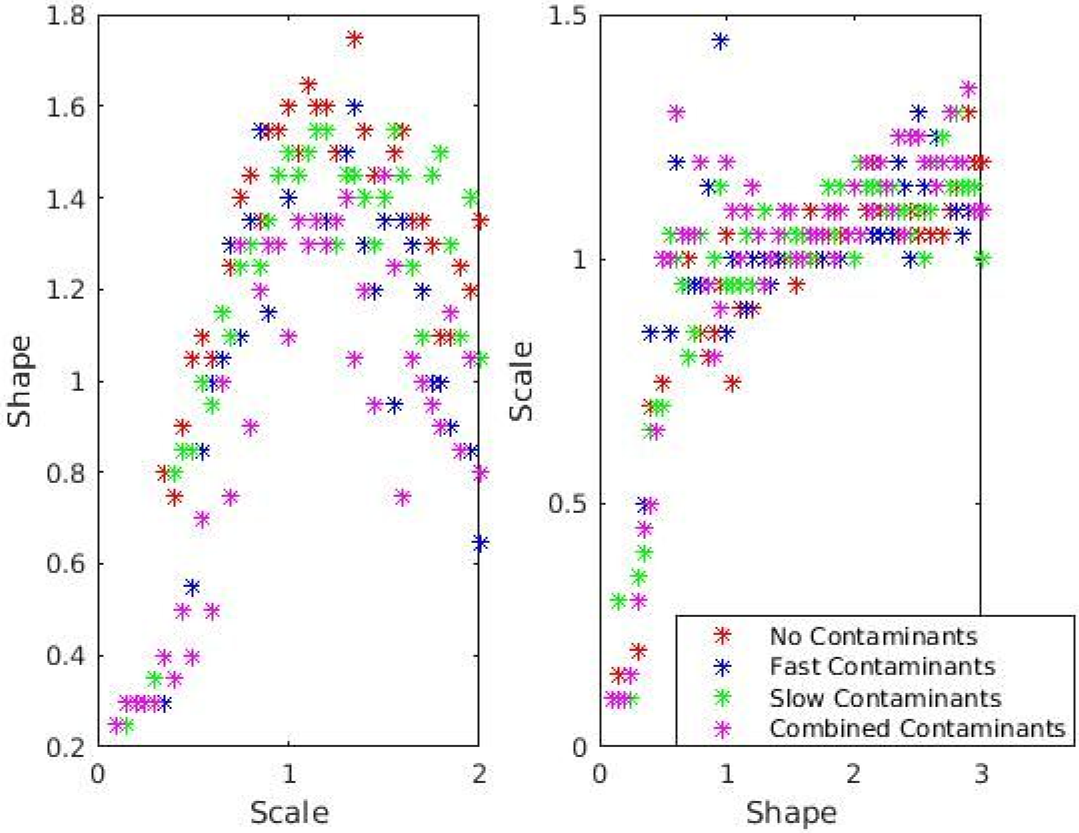


Figure 3.5: This figure show how dynamic likelihood behaved under different contaminants. Red stars show the situation with no contamination, blue with fast contaminants, green with slow and purple shows a combination. The scale parameter does not show and difference in behaviour (right figure) and shape suggests a small deviation from the situation with no noise.

In conclusion, we have modified a likelihood function by introducing robustness properties informed by the data itself and showed that this simple modification introduces robustness properties that are on part to the Hellinger distance which is known for its robustness properties. We demonstrated that both methods recover true or close to true parameters when commonly observed contaminants are included within RT-distributions and no additional complication of the model is required (Ulrich & Miller, 1994). While robustness is a nice extension for cognitive modelling, there’s further work required to identify optimal implementation of the robust likelihood and how its introduction would impact Bayesian inference.

Chapter 4:

Approximate Bayes Sampling with Excessively High Likelihood Errors.

Abstract

Every approximation produces some error from true solution which introduces some uncertainty variable over the true solution. Simple undefined models with its unknown PDF are easily dealt by adding some normally distributed error function which enables Bayesian Inference over the model. Such uncertainty simply expands confidence interval of the posterior distribution estimates without much additional impact. However, the KDE approximation over the model is a mixture approximation which is known to have undefined true maximum likelihood. A simple addition of the normal process to the variables does not aid in estimating model parameters. This is a mild issue for the best solution estimation but confidence intervals are difficult to define, especially for computationally expensive models. A standard method for posterior approximation uses acceptance/rejection criteria of Metropolis-Hastings mechanism which is a core mechanism in Markov Chain Monte Carlo methods. However, this mechanism fails with undefined maximum likelihood and this chapter illustrates this issue as well as designs a new algorithm that deals with the issue of undefined maximum likelihood.

4.1 Introduction

Earlier we have shown that reaction time distributions can be represented using kernel density estimator methods (see KDE section). The same method can be used to characterise models with analytically intractable PDFs. A capability to describe models that do not have a known probability density function expands assessable models by extending solvable complexity of the models. This ability is particularly useful if the models are based on neurobiologically plausible mechanisms which go beyond the currently very popular decision models (Usher & McClelland, 2001; Wang, 2002; Shadlen & Newsome, 2001). It is important to show that these accumulator models can produce equivalent RT-distributions to probabilistic models. However, currently available methods are limited and further work in designing accurate and efficient methods that are able to approximate these models is required.

A classical approach to optimise the models is by estimating maximum likelihood (MLE) via likelihood function. Analytically intractable models do not have a known PDF thus quantile approximations are usually used (Maritz & Jarrett, 1978; Heathcote, Brown &, Mewhort, 2002; Heathcote, Brown, & Cousineau, 2004). These quantiles discretize probability functions into bins which are used as an alternative to the unknown PDF. Once approximation of the PDF is constructed, MLE methods can be used to identify the best fitting parameters. An alternative approach exploits Bayesian methods which in the past several years have been growing in popularity in terms of usage as well as applicability and psychology is no exception (Turner & Van Zandt, 2012; Vincent, 2015). Classical Bayes is expressed as follows:

$$P(\theta|X) = L(X|\theta)P(\theta)/\int(L(X|\theta)P(\theta)),$$

where left-side of equation is a probability distribution of the parameters (expressed as θ) given data X referred as posterior, $P(\theta)$ is a prior, $L(X|\theta)$ is a likelihood function and an integral on the right-side of the equation is also known as marginal which makes posterior a proper probability distribution. This is a direct extension from MLE methods that provide distributions for the parameters of the model. This method is superior to MLE as it reuses available knowledge expressed within prior to construct probabilities for parameter settings given the new data. It also generalises better for model selection as an entire parameter space is used for comparing which model described data better. However, since it still uses likelihood functions, it inherits all the issues present in MLE methods for analytically intractable models. Moreover, MLE approximation methods such as QML based approach for approximation of RT-distributions are biased in Bayesian settings and an alternative KDE methods are better suited (Turner & Sederberg, 2014) which, as we will show later, create other issues in posterior approximation for analytically intractable models.

A number of sampling algorithms have been developed to approximate the posterior distribution when likelihood functions are not available. Most of the algorithms are designed for models that do not have a known likelihood function but likelihoods themselves are easily evaluated. Well known examples involve using generic functions to describe model's behaviour (Ratcliff, 1978) or using some stable summary statistics such as QML (Heathcote, Brown, & Mewhort, 2002). More notable Bayesian approximation algorithms follow Metropolis-Hastings MCMC probabilistic acceptance/rejection (Hastings, 1970). Other MC algorithms such as Importance Sampling exploit the knowledge of sampled space by resampling available samples (Doucet & Johansen, 2008). However, all these methods fail when likelihood values are not easily retrieved.

Contrary to other sampling methods, Approximate Bayesian Computation (ABC) algorithms assume likelihood functions are not available and replaces them with two alternative methods. One method introduces tolerance function to filter out poor parameter samples (Sisson, Fan & Tanaka, 2007; Beaumont, Cornuet, Marin & Robert, 2009) and another method weights these samples by mapping them via some kernel function (Beaumont et al. 2002; Wilkinson, 2008, 2013). Both methods extend to posterior approximation algorithms though their accuracy is not fully understood while the algorithms tend to suffer from the curse of high-dimensionality (Beaumont, 2010).

In this section we propose a new Importance Sampling algorithm based on clustering space as means to remove accept/reject notion to deal with the curse of high-dimensionality. We will also extend it by integrating tolerance threshold and propose a new method to determine this threshold by approximating the most likely model approximation at the best parameter solution. We will start this section with a brief introduction to a couple of main algorithms available for Bayes approximation. We chose DE-MCMC algorithm to represent a class of algorithms that use MH acceptance criteria because it is a main component that hinders this class of algorithms. This class includes other algorithms such as Hamiltonian Monte Carlo (Neal, 2011) but will suffer from the same efficiency issues that are created due to undefined maximum likelihood, thus will not be looked at. However, DE-MCMC is one of the least costly methods computationally wise. Another class of algorithms originates from resampling mechanism used in Importance Sampling class of algorithms. We use this particular mechanism to develop a new algorithm based on parameter space clustering. It will follow with working examples when likelihood values are computable but not the likelihood function. Afterwards we will consider the situation when likelihood values are not available and propose adjustments suitable for such situations.

4.2 Differential Evolution - Markov Chain Monte Carlo (DE-MCMC)

Recently a new algorithm based on differential evolution (DE: Storn & Price, 1997) for analytically intractable and linearly correlated models has been proposed (DE-MCMC: Turner, Sederberg, 2012; Turner et al., 2013). This algorithm has been successfully applied to a few well known analytically intractable models such as Leaky Competing Accumulator model (LCA: Miletic et al., 2017) and Feed-Forward Inhibition model (FFI: Turner, Sederberg, & McClelland, 2016). It combines a couple of major algorithms to deal with issues of posterior approximation when such models are being used.

Firstly, it incorporates a traditional Metropolis-Hastings probabilistic acceptance/rejection procedure of the algorithm (Hastings, 1970). It is a classic sampling algorithm for approximating posterior distribution which has been used in psychology as a standard method to approximate posterior distribution. This algorithm works by employing Markov Chain Monte Carlo (MCMC) property that accepting or rejecting proposal given a previous state of the chain would converge to the posterior distribution. The probability to accept or reject is defined by likelihood ratio between between a new proposal and previous state which is expressed as follows: $\alpha = min\,(1, \frac{\pi(\theta_n)K(\theta_n)}{\pi(\theta_{n-1})K(\theta_{n-1})})$, where $K(\theta_n|\theta_{n-1})$ is a transition kernel and $\pi(\theta)$ is fitness value of the given state (parameter settings). The algorithm uses a single chain which is perturbed to generate a new proposal governed by the transition kernel which is positioned at the centre of the current state of the chain. Usually this transition kernel is normally distributed with some unknown covariance matrix. This matrix is one of the free parameters that are used to determine how far away from the current state the new proposal should be generated. This can be tricky when analytically intractable models are being used because parameter relationships are not always easily determined. Current state corresponds to the last parameter settings proposal that was accepted using the outlined ratio criteria, in other words, the chain is a sequence of accepted parameter settings that form a posterior distribution (Hastings, 1970).

DE-MCMC algorithm introduces an evolutionary approach to generate proposals within posterior distribution by expanding MCMC chain into multiple chains jumping within posterior distribution at the same time. This forms a population which then can be used in a standard evolutionary approach by using its two core features of mutation and crossover to find the solutions. Mutation works by randomly perturbing certain individuals within a population while crossover combines members of the population to generate a new population. DE-MCMC uses differential evolution algorithm which performs crossover by exploiting parameter relationships. In other words, it adapts to local parameter space by adapting their variances as well as correlations. This feature removes the requirement for transition kernel which otherwise has to be provided by the researcher. These parameter settings transitions are computed by mixing existing parameter solutions within current populations. It is computed as follows:

$$\theta_{n,k_0} = \theta_{n-1,k_0} + \gamma_1(\theta_{n-1,k_1} - \theta_{n-1,k_2}) + \gamma_2(\theta_b - \theta_{n-1,k_0}) + \varepsilon;$$

where γ indicates a weight contributing to the shift from original location, k indicates some chain, $n, n-1$ current and previous states while θ_b are the best known solutions thus far. The equation also has error component ε which introduces random perturbation to the new proposals.

A second additive component yields a shift towards the best known solution. This second component violates equality condition of the parent and child particles and is suited for finding the best solution. Setting $\gamma_2 = 0$ satisfies stationary conditions and will converge to the posterior distribution.

The last two features of the algorithm are mutation and migration. Mutation feature is used to keep exploring the parameter space by introducing changes to the parameter settings. The migration on the other hand works by dividing population into groups. Then, a few poorly performing parameter settings are exchanged between between groups to mix them. This migration prevents groups getting stuck in their local minima since a bad member of one group may be a good member to other groups. Overall, an introduction of evolution within algorithm makes it substantially more efficient than traditional MCMC algorithms because it can adapt to local space in terms of correlations as well as variance of the parameters.

4.3 Importance Sampling

An alternative class of algorithms for posterior sampling is Importance Sampling algorithms (IS; Geweke, 1989). IS is based on the idea that random sampling can be improved substantially by generating proposals from dense areas of solutions. IS algorithms work by constructing the Importance distribution and sampling from it rather than from prior distribution. Importance distribution provides sampling probabilities based on whether certain areas are undersampled or oversampled in order to guide future samples. Consider a target posterior distribution

$$P(\theta|X) = M(X|\theta)p(\theta)/(\int M(X|\theta)p(\theta)d\theta),$$

where M indicates some model and X a given dataset. The importance distribution is defined by introducing the weight distribution $W(\theta)$ as follows: $I(\theta) = W(\theta)P(\theta|X)$. Weight distribution itself is defined as a ratio between likelihood function and the posterior distribution: $W(\theta) = M(\theta)/P(\theta|X)$. An ideal solution would have this weight at one for all parameter settings as this is an indication of convergence between importance and posterior distributions. Basic IS algorithm would use this weight distribution to resample previous samples after n iterations by introducing copies of the undersampled samples to guide sampling algorithm (Doucet & Johansen, 2008).

This algorithm is easily extended to work iteratively by treating a posterior distribution also as a prior or sampling distribution and every consecutive draw immediately updates a sampling distribution, such algorithms are also called Sequential Importance Sampling (SIS) algorithms and belong to a class of Sequential Monte Carlo sampling (SMC) algorithms (Sisson, Fan & Tanaka, 2007; Doucet & Johansen, 2008; Beaumont et al., 2009). In this section we will introduce a new SIS algorithm which can be applied for computationally expensive models. Our SIS algorithm introduces clustering and additional dimension to the posterior distribution. The clustering helps to better approximate the expected likelihood values in specific areas and replace resampling by producing duplicates with sampling from local areas. The additional dimension represents the distribution of approximated likelihood values. This is motivated by the fact that fitness outcomes have a distribution of their own and parameter values are correlated with these fitness values. This modifications to the original algorithm substantially improves efficiency compared to MH-style algorithms as all defined samples are retained.

4.4 Sequential Importance Sampling by Clustering (SISC)

4.4.1 The Algorithm

```
$Initialise\ \theta_{1:n}\ by\ sampling\ from\ a\ prior$
$L_{1:n} \leftarrow M\ (X|\theta_{1:n})$
$K \leftarrow Cluster\ (\theta_{1:n})$
$for\ 1 \leq k \leq K$
    $W_k \leftarrow Weight(L_{L \in K_k}|K_k);$
$end$
$for\ n < i \leq m$
    $\theta_i \leftarrow Generate(\theta_{1:i-1}, W)$
    $L_i \leftarrow M\ (X|\theta_i)$
    $K \leftarrow Cluster(\theta_i)$
    $for\ 1 \leq k \leq K$
        $W_k \leftarrow Weight(W_{k,}L_{i \in K_k}|K_k);$
    $end$
$end$
```

Fig. 2.3.1: This figure shows a pseudocode for proposed algorithm.

Our Sequential Importance Sampling by Clustering (SISC) algorithm replaces the concept of samples with the concept of weighted areas which are represented by kernels. Fig 2.3.1 outlines a general description of the algorithm where θ represents parameter settings, L is a likelihood or any other fitness measure, K indicates the clusters within posterior distribution and W are resampling weights. We propose to replace the notion of chains with clusters of samples within parameter space that are dynamically modified to accommodate knowledge about parameter space. We cluster the drawn samples to represent the posterior distribution (lines 3 and 10). This clustering can be performed in multiple ways and it will be discussed later. Once the clusters are formed, the weights for the clusters are computed (lines 4-6 and 11-13). These values are normalised to form the selection of individual clusters probabilities and is a crucial feature of the algorithm. The main body of lines 7-13 iteratively generate new proposals that are used to update the knowledge about the parameter space. It updates clusters as well as fitness values to guide future samples and with sufficient number of iterations will converge to the posterior distribution.

4.4.2 Importance Sampling by Clustering

The simplest posterior distribution is normally distributed and has linearly or near-linearly correlated parameters. Algorithms such as DE-MCMC can also approximate skewed posterior distributions but its performance would deteriorate with increasingly complex posterior distributions. Generally, complex posterior distributions are represented using mixture models (**Cite**) and can accurately approximate posterior distributions when the flexibility of the mixture models are somewhat constrained, i.e. centres are allowed to vary but not variances (Grazian & Robert, 2015). However, we suggest that compression via clustering of the mixture model can reduce degrees of freedom and ultimately enable to approximate a proper unknown posterior distribution using a mixture model. In order to approximate unknown posterior distribution we incorporated a number of features within online kernel density estimator (oKDE: Kristan, Leonardis, and Skočaj, 2011) to our proposed SIS algorithm.

Firstly, we propose treating parameter samples as kernels which enables sampling from undersampled local space rather than producing copies. Consider a parameter sample distribution Θ with its corresponding weight distribution W. During resampling phase, the basic IS produces a new distribution Θ^* which has parameter values equally weighted. The primary goal is to obtain a sample distribution that is equally weighted. The same can be achieved when parameter samples are treated as kernels and new parameter distribution Θ^* has $W\Theta$ new proposals drawn in the neighbourhood of previous samples. This improves efficiency in terms of the variation of various samples but sampling can be further improved by expanding this local neighbourhood by compressing local kernels into one. These compressed kernels forms a clustered local sampling space. Each cluster is represented using a weighted Gaussian distribution which is a convolution of the points within i-th cluster: $w_i N(\mu_i, \sigma_i)$, where w_i indicates the weight, μ_i is a mean of the points within the i-th cluster and σ_i is the covariance of the points within that cluster. Then the posterior approximation can be re-expressed as a mixture of Gaussian clusters:

$$P(\theta|X) = \sum w_i N(\theta|\mu_i, \sigma_i).$$

As a result, importance distribution is re-expressed as follows:

$$I(\mu) = W(\mu) / \sum w_i N(\mu|\mu_i \sigma_i);$$

where μ represents the centres of the clusters and σ their standard deviations.

The vital difference emerges in determining the weight distribution because it is a ratio between likelihood and posterior. Working with clusters requires to assign some likelihood value to these clusters that would appropriately define their fitness. This can be obtained by merging likelihood values of all points within that cluster which is equivalent to computing the expected likelihood value: $L_n(X|\mu) = n^{-1} \sum L(X|\theta)$; where L is a likelihood of the data. Then the weight distribution for importance sampling algorithm is expressed into:

$$W(\mu) = n_{\mu}{}^{-1} \sum L(X|\mu)/ \sum w_i N(\mu_i \sigma_i).$$

4.4.3 Sequential Updating

An important feature of the algorithm is its extension to sequential updating of the sampling distribution because iterative sampling, albeit more costly, is much more efficient since it can react to very good new samples. It is especially useful in early stages of sampling as some new samples can change the direction of posterior sampling. The new proposals are drawn from this weight distribution which are either integrated into the old clusters or begin to form a new cluster. Parameters that are being integrated into old clusters are simply added to the old integral $L_n(X|\mu) = (1 - n^{-1})L(X|\Theta)_{n-1} + n^{-1}L(X|\theta^*)$, where μ is a centre of the cluster, n indicating the number samples represented by the component, Θ being all the previous samples and θ^* corresponding to the new proposal. When clusters are merged, the new likelihood integral is computed in the same manner, except for the weights representing the weights of the clusters. We select the area for generating a new proposal with the probability equivalent to normalised weight distribution.

A new random sample is drawn from local Gaussian distribution (see Local Sampling Distribution section) which is evaluated against data and added to the posterior approximation with the contributing weight $w_{\theta^*} = max(1, L(X|\theta^*)P(\theta^*|X)/(L(X|\theta_b)P(\theta_b|X)))$; where θ_b is the best known solution. This further increases efficiency of the system because the contribution is increased or reduced depending on whether it is oversampled or undersampled proposal. It also guards from contributions produced by oddly shaped likelihood functions, i.e. multi-modal function.

Importantly, we do not compress posterior approximation iteratively because it is computationally a relatively expensive feature. Instead, we use the a maximum effective number of clusters as a threshold for clustering to be performed as defined in oKDE (Kristan, Leonardis, & Skočaj, 2011). During this phase, the algorithm assesses whether the present combination of clusters can represent the parameter samples thus far within desirable accuracy. If it can represent, a smaller set of clusters is formed to construct a new posterior distribution approximation. When that is not the case, the maximum effective number of clusters is expanded by some scalar c and bad clusters are divided to form a new posterior approximation with a larger number of components. We had this scalar set to 1.5 since Kristan, Leonardis, and Skočaj (2011) showed it to be sufficient.

4.4.4 Local Sampling Distribution

Using our new weight distribution we could perform sampling directly through mixture model. However, it is prone to degeneration because it does not explore space further away from already known samples. We propose extending this local neighbourhood by expanding chosen component from the mixture model. All current parameter samples can be evaluated against separate weighted components. This provides us with a weighted vector $wp(\mu, \sigma|\theta)$, where weights indicate the probability of the component within mixture model of the posterior, we can construct a new Gaussian kernel that combines these weighted components using this computed weighted vector. Then the new local neighbourhood kernel is defined as follows:

$$K_i = N(\mu_i, \sum w_i p\ (\mu_i, \sigma_i|\theta)\sigma_i / \sum w_i p\ (\mu_i, \sigma_i|\theta))\ ;$$

where K_i is i-th new sampling kernel centered at the same value as the i-th cluster but expanded to the neighbourhood with the weighted probability of that centre in other kernels (a measure of overlap).

4.4.5 Importance Sampling by Clustering with Forgetting

We also introduced a burn-in sampling mode for finding the good initial posterior when initial population does not represent prior distribution. It simply weighs new arriving samples higher than older samples retaining the overall weight of the sample distribution set to some constant value c. We set this weight to $c = wn$, where n corresponds to the number of dimensions being approximated and w is a scaling factor which we always set to 100. Such constant weight was relatively efficient and consistent in converging to the posterior distribution, though we expect that there are better ways to converge to good initial posterior distribution.

However, in early sampling a weighted sequential updating of the posterior approximation is unstable because the probability of the new proposals is likely to be close to zero in comparison to the best known solution. In this circumstance we suggest setting this contribution weight to one for all new samples until a good initial approximation is produced. Reverting to proper weight contribution during posterior approximation will gradually phase any accidental poor samples.

4.5 Materials and Methods

To illustrate the working of the algorithm we created multiple tests in order to compare it to DE-MCMC algorithm. For this purpose we used an analytically tractable 3 parameter Weibull model. This model has scale, shape and shift as parameters. To assign realistic parameter values this model was fitted to human dataset (information about dataset) (Wolfe ?). The best solution to this dataset is approximately 755, 1.24 and 426. Throughout the simulations, the hypothetical dataset (500 samples) was generated from these parameters which was used to compute the likelihood under provided model, this same dataset was used in all examples. We used this hypothetical dataset to reduce error arising from incorrect model choice, in this case the difference between Weibull and human participant. Throughout the study we used non-informative bounded uniform priors with the best solutions being roughly centred within the range of the uniform bounds (see Table ?).

	Min/Max Bounds
Scale	0:1500
Shape	0:3
Shift	0:900

Table 4.1: This table shows the prior range used for each parameter setting of the Weibull parametric function.

This study was divided into two parts. First part assumed that a kernel method used is the Weibull function itself. This reduces the flexibility associated with kernel methods since there are fewer degrees of freedom. It also enforces the best solution to correspond to the actual best solution and limits the possible variations within approximated likelihood values. In the second part we will examine the issues arising when a pdf approximator with unlimited degrees of freedom is used and how it impacts SISC and DE-MCMC. In this case we used mixture model as pdf approximator as described in KDE section. In both cases each proposed parameter setting for the model would be used to generate 100 samples. We chose 100 samples since it illustrates potential limitations for simulating the model.

The methods differ in their underlying processes and strengths as well as weaknesses. To adapt to these differences the running process differed for the algorithms, the requirements also differed within separate studies. In the first section, during the burn-in period DE-MCMC was run to produce 25000 proposals from 25 chains and their initial parameter sets drawn from the priors while SISC algorithm produced 2000 proposals with 100 initial parameter settings drawn from the identical prior. For posterior approximation we produced 50000 samples for DE-MCMC and 10000 samples for SISC algorithm. Additionally, SISC algorithm used oKDE for approximating unknown posterior distribution. The accuracy of oKDE was set to default value at 0.02 which is not the optimal option but this level of the accuracy did produce satisfactory approximations of the posterior distributions.

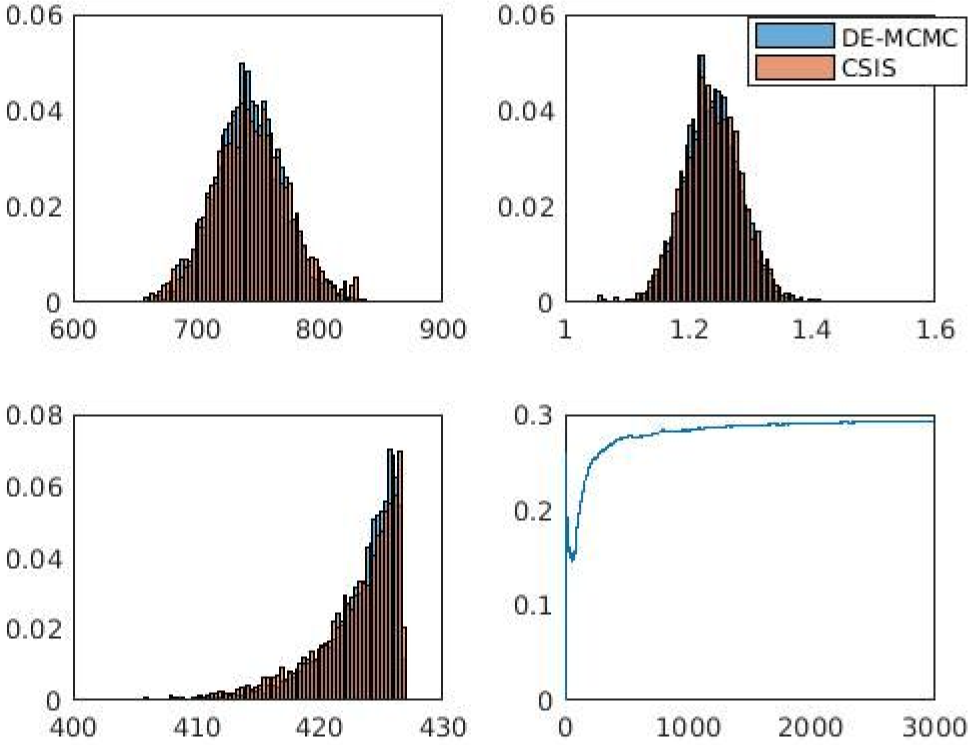


Fig 4.1: This figure illustrates the posterior approximations produced by DE-MCMC (blue) and SISC (red) algorithms. Each graph shows marginalised dimensions and approximates produced similar distributions. The fourth graph shows the acceptance rate of DE-MCMC algorithm

4.6 Noise-free posterior approximation

We begin our algorithm assessment in perfect conditions when the model is analytically tractable and likelihood values are easily evaluated. We perform posterior approximation using DE-MCMC and SISC algorithms on the shifted Weibull model. All parameters were free to vary and had the best solutions and non-informative priors as described in general section. There were no changes in the settings of the algorithms as well.

Fig 4.1 shows the fitted results indicating that both methods fitted the underlying posterior distributions similarly well. There’s very little noticeable difference for SISC even though there was a compounded error introduced by the accuracy of KDE. It also shows the approximations are close despite the limited number of samples for SISC which is an important feature when lengthy MCMC sampling is too costly to perform. The fourth illustration of the figure show that DE-MCMC attained closed to 30% of proposals being accepted. This a very efficient acceptance that we will use as reference in later sections because this efficiency is the highest possible. Any drop in efficiency will show how uncertainty over likelihood values impacts it.

4.7 Model Error

The main advantage of the algorithm emerges when likelihood values cannot be accurately measured as this is not always the case and more complex models require simulations to approximate this value. The fewer simulations are performed, the less accurate a resulting approximation of the model is. Depending on the model, many simulations can be infeasible. This induces error on posterior results which can take numerous forms and these will negatively impact posterior approximation algorithms. Under such conditions, a likelihood approximation can be expressed as follows: $L(D|\theta, \varepsilon) \approx M(D|\theta) + \varepsilon;$ where ε is a random variable characterised by some probability distribution and M is a likelihood function of the model. This error term could take a range of distributions and the likelihood value error generally changes as the parameter settings are modified. In general, this can be divided into two situations, when likelihood function is unknown and likelihood values are strictly bounded at the maximum and when they are not.

Firstly, we will consider a situation that has likelihood function is unknown but its values are strictly bounded at the maximum. This is an interesting situation where the kernel estimator will introduce incorrect parameter setting by some error probability. As a result, this may produce the maximum likelihood value but will not produce higher likelihood values than the true maximum. Such situation emerges when the estimator of the model is limited to a set of simple functions which can range from linear regressors to simple distributions. This also known as a parametric likelihood approximation. For example, this situation would arise when a complex model is considered and its output is mapped to some simpler form of the model. It is a nice approach to make sense of complex problems which was initially introduced by Ratcliff (1978) where ex-Gauss model was a simplifying parametric PDF for the proposed diffusion model. This approach was widely used (Smith & Mewhort, 1998; Hockley, 1984; Leth-Steenson, Elbaz, & Douglas, 2000) until a less biased QML method was proposed by Heathcote, Brown and Mewhort (2002). More recently a new approach has emerged that creates hierarchical models to explain complex models where lower layers use intractable models and these are mapped to simpler tractable models to form one large model (Liu, Shum & Zhang, 2001; Lee & Mumford, 2003; Sanborn, 2017).

The main argument against using simpler model approximators is their bias. An alternative approach would approximate model’s PDF shape using non-parametric methods. The most common methods use quantiles as statistical description of such models (Maritz & Jarrett, 1978; Heathcote, Brown & Mewhort, 2002; Heathcote, Brown & Cousineau, 2004). Quantiles create a limited number of bins of data which later are fitted to simulated RT-distributions. This approach has little variation in likelihoods values and any standard MLE method would work. However, it is heavily biased posteriors are approximated. An alternative approach uses KDE methods where multiple simple pdf functions are used to approximate unknown distributions (van Zandt, 2002; Turner, van Zandt, & Brown, 2011; Turner & Sederberg, 2014). This is a challenging situation where Bayesian inference may not be possible due to ill-conditioning, multi-modality, non-linearity as well as unbounded likelihood values (Grazian & Robert, 2015). We will focus on unboundedness of error which makes sampling methods highly unstable. In this section we will propose an extension to the SISC algorithm to deal with unbounded likelihood errors by introducing the concept of overfitting to the model evaluation.

Overfitting is a rather artificial situation created by the introduction of robustness in previous chapter. In robust likelihood chapter we have emphasised that it is a known fact that RT-distributions are generally contaminated (Ulrich & Miller, 1994). We suggested that contamination removal is not necessary because there are robust methods that can find optimal solutions despite their presence in the data. However, using a robust method introduces a new issue for approximating posterior distribution. Identifying contamination using a robust goodness-of-fit means that the weight over certain values within data has to be shifted to other parts of the data (Mandal, Basu, & Pardo, 2010). The most interesting contaminants are hidden contaminants which imply that the tail is weighted more than it should (Ratcliff & Tuerlinckx, 2002). However, this weight depends on the model itself which is being approximated by the sampling because not analytical solution is available. This means that there is a probability that an approximated model PDF will fit data perfectly and show excessively good likelihood values. We coined this emerging situations as overfitting of the data.

An overfitting issue will cause issues for posterior approximators that rely on likelihood ratios when datasets are relatively large. In this situation a likelihood ratio between a good PDF approximation and overfitting PDF approximation will be very close to zero. Moreover, the probability of overfitting can be relatively likely (5%). Once this outcome is generated, most of the future proposals will be rejected when DE-MCMC algorithm is being used. On the other hand, SISC algorithm will fail completely because it will only generate proposals from the cluster that has this overfitting value. Most of the new proposals will be within its cluster and occasional proposals outside the cluster will be unlikely to generate overfitting PDF approximation. Depending on the PDF approximation, the scale of overfitting can range from mild to severe thus likelihood values are almost unbounded and most of the posterior approximation algorithms will struggle with such situations.

A way to deal with overfitting is to treat it as a probability which can be linked to a method called tolerance threshold and is a method to approximate unknown posterior distribution (Sisson, Fan & Tanaka, 2007; Beaumont et al., 2009). This method was introduced for posterior approximation when summary statistics are used while our proposed method suggests approximating likelihood values and determining the most likely maximum likelihood value as threshold point. We argue, that placing this threshold at this point indicates where the overfitting begins provides the most efficient posterior estimator. To determine the most likely likelihood value when the best parameters are used central tendency methods can be adopted since the generated values form a distribution as well. Out of the three central tendency methods (mean, median and mode) mean is the least suitable since this likelihood value distribution can take skewed forms as well. However, we also think that mode is a better choice than a median because it indicates the most probable likelihood value which is the best guess for the overfitting threshold. Throughout further work we assume that the mode of the likelihood value distribution at the maximum indicates this threshold level.

4.8 Bounded Likelihood Values

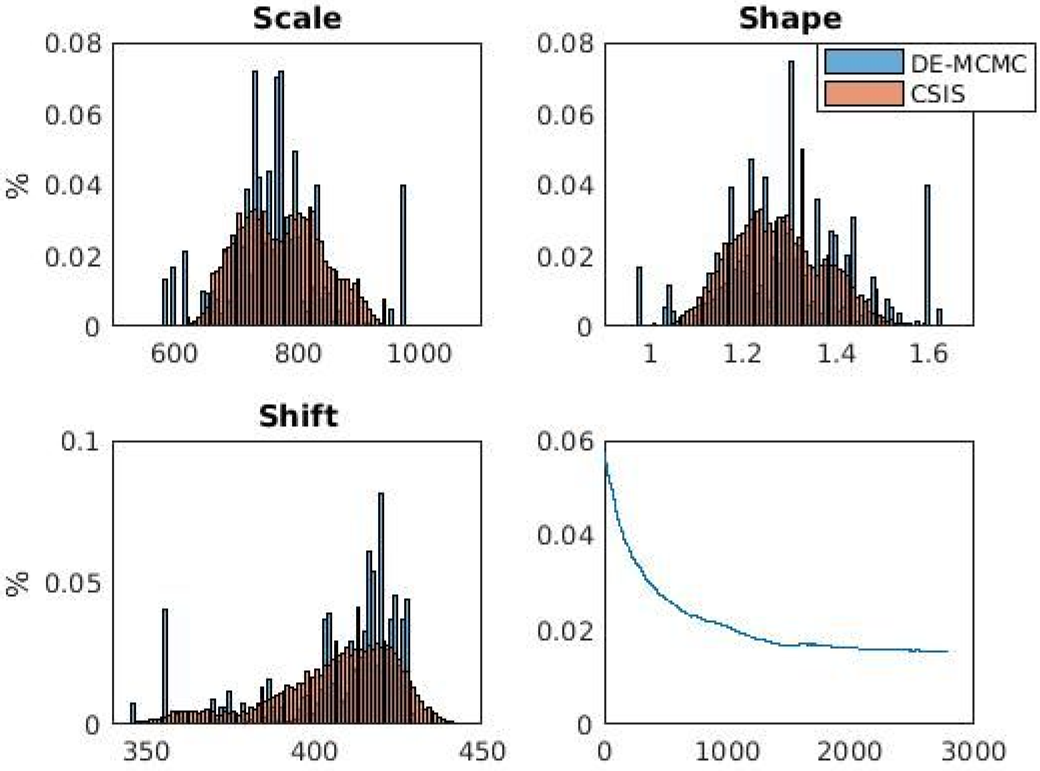


Fig 4.2: This figure illustrates the posterior approximations produced by DE-MCMC with 3000 generations (blue) and SISC (red) algorithms when the likelihood values are bounded at the maximum. The distributional approximations poorly distinguishable approximation produced by DE-MCMC thus comparison is difficult. Fourth graph shows the acceptance rate of DE-MCMC algorithm.

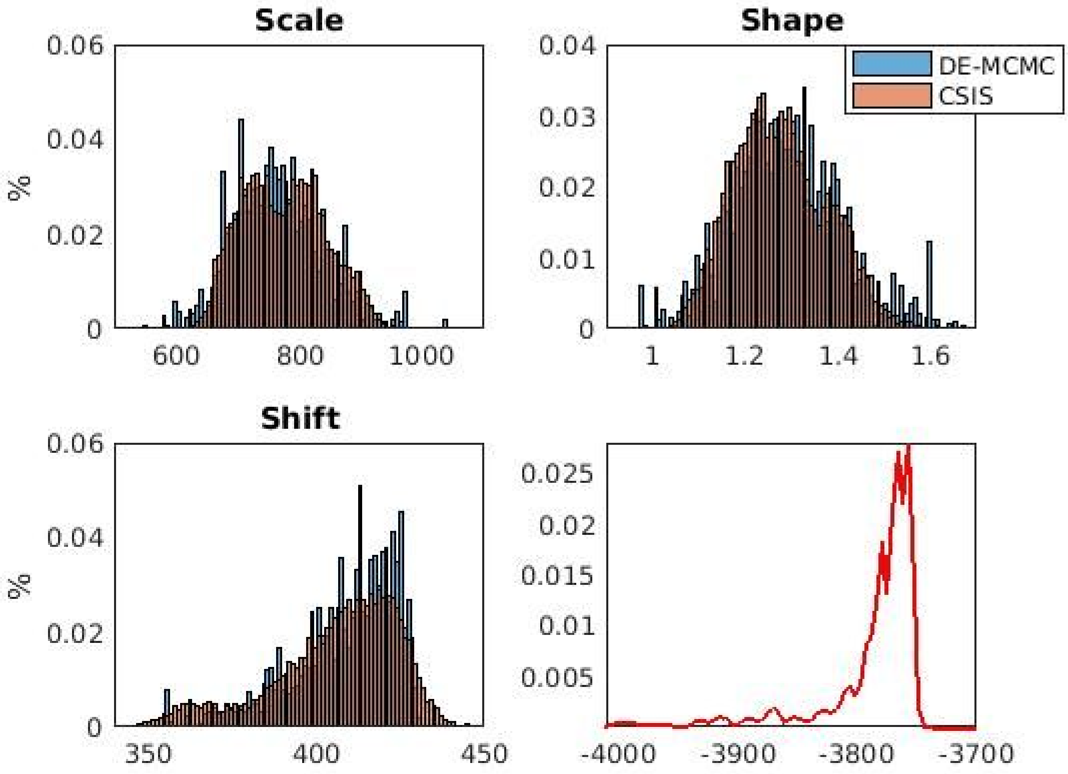


Fig 4.3: This figure illustrates the posterior approximations produced by DE-MCMC with 10000 generations (blue) and SISC (red) algorithms when the likelihood values are bounded at the maximum. The distributional approximations are close except for shift parameter where CSIS algorithm produced more values at the right-tail. Fourth graph shows the distributions of likelihood values produced when posterior was approximated.

4.8.1 Materials and Methods

We will generate this situation by using the approach used in RT modelling before QML methods were introduced. We consider a parametric simplifying model as our kernel approximator. To illustrate this, we picked a 'perfect' kernel model which is identical to the model being approximated. Such specification removes the possibility of unbounded likelihoods which is the case with mixture models. Instead, the error will be within MLE estimation of parameters for simulated samples.

As stated in general methods we assumed that unknown model is Weibull function and it was simulated to generate 100 samples for every parameter proposal. Using 100 samples is not sufficient to estimate unbiased maximum likelihood for Weibull parameters since it is asymptotically biased at shift parameter due to hard probability bound at the left side of the parameter range (Fig 4.1c;). However, we accept this bias as it is an illustrative example for posterior approximation and these samples were used to determine the Weibull parameters that best describe simulated distribution. Note that when Weibull function is used as a kernel, it has multi-modality as well as negatively skewed likelihood value distribution. Multi-modality is caused by shift parameter and would be easily observed if the posterior was only approximated for this parameter. We focused on all 3 parameters since multi-modal examples are simply equivalent to discrete probability of whether a sample distribution will give zero or non-zero probabilities and the probability of the two outcomes which is not an interesting example. Restraining on a full model replaces obvious multi-modality with very long tail instead. Since, Weibull is only 3 parameter distribution, we used a sufficiently fast fminsearch to determine the parameters for kernel representation.

Additionally, at this stage we make one change to SISC algorithm. New random samples are drawn from local Gaussian distribution as described in non-extended algorithm version. However, due to high variation in likelihood values produced by the new samples the weighted contribution is completely omitted. Instead, all new samples receive a weight of one.

Finally, we increased the number of generations for DE-MCMC from 3000 to 10000 to improve posterior approximation since 3000 will not be enough to produce a decent approximation.

4.8.2 Results

Fig. 3 shows the results of posterior approximation when DE-MCMC had 3000 generations as in previous example. The result shows a poor approximation of the posterior distribution when DE-MCMC is used and the comparison has little value. The Fig. 3d demonstrate the acceptance rate under such conditions which is below 2%. The result is not surprising given such a low efficiency.

We expanded the number of samples produced by DE-MCMC with additional 7000 generations which produced better posterior approximations (Fig 4). Evidently, the posterior approximations are no longer identically looking for two algorithms as in no-noise example. DE-MCMC result shows less smooth distribution which is the case due to an increased chance for chains getting stuck with very good likelihoods. The distributions of likelihood values shows a large variation in the observed outcomes and most of these outcomes are rejected when likelihood ratio approach is used. However, more generations would smooth posterior distribution further until it would converge to a nice approximation.

On the other hand, SISC algorithm is as smooth as it was in no-noise example though some built-up modes are visible. More sampling and higher accuracy would remove this effect. However, a more interesting outcome is with a shift parameter, it illustrates the disadvantage by the algorithm produced when all proposals are accepted with an equal weight and poor clustering accuracy. This an artificial outcomes caused by encouraging some additional exploration by expanding local sampling area. A way to reduce this issue would be to increase the accuracy of clustering, though working with more clusters makes algorithm slower. Irrespective of these limitations, the satisfactory posterior approximation was obtained with substantially smaller number of samples.

4.9 Approximate Likelihood Values

4.9.1 Sequential Importance Sampling by Filtered Clustering (SISFC)

We extended the SISC algorithm by introducing an additional dimension to the mixture model. This is motivated by the argument made earlier in the section on the likelihood error that fitness outcomes for each parameter setting have a distribution of their own. In order to deal with unbounded distributions of likelihood value error, this method relies on kernel filtered likelihood values to deal with high uncertainty about the accuracy of the result. The convolution of this high uncertainty and actual likelihood values completely covers an entire likelihood value function and is an important observation to solve this problem. Given that error in likelihood value can take a distributional form from Normal to highly skewed as well as have one or more modes. We used Gaussian mixture model as our complex kernel filter and integrated this new dimension into SISC algorithm by updating it after each new proposal to improve the knowledge about likelihood value distribution.

Across an entire parameter space, poor fits are more common than overfitting fits and these will be the most likely thus to keep accepting all proposals, some directional sampling is required. In order to introduce direction when likelihood values are not reliable, we introduced a dynamic overfitting threshold as described in previous section but adapt it to clustering based sampling.

Instead of identifying it at the best solutions which is difficult to find under these circumstances, we propose sampling based on the probability of producing overfitting points as predicted by the most probable likelihood value within the best fitting cluster. This value is retrieved from the suggested additional dimension of likelihood value distribution. Our weight distribution for all clusters is re-expressed as the proportion of likelihood values exceeding the threshold within each cluster relative to the probability of the cluster which is a probability in current state of the posterior: $W(k) = P(L(X|\theta_k \in k)/P(\theta_{1:n}|X)$; where k is a cluster of parameter settings. Due to the introduction of the likelihood dimension, each cluster have their weighted mixture distribution of the likelihood values, thus

$$P(L(X|\theta_k \in k) = 1 - F(L_\varepsilon|k);$$

where F is a cumulative function of all the likelihood values within cluster and L_ε is tolerance threshold.

4.9.2 Materials and Methods

In order to illustrate the example, we used identical settings as in bounded maximum likelihood value example except for kernel approximator being non-parametric. We used a Gaussian mixture model generated by oKDE with accuracy of 0.002. DE-MCMC algorithm in this scenario struggles to accept new proposals even more and produces rates of <1% in this simple example. To improve its performance we introduced overfitting threshold to DE-MCMC for a fair comparison. Our overfitting threshold was determined during a burn-in period of SISFC model and corresponds to the approximated point when the model begins producing overfitting proposals.

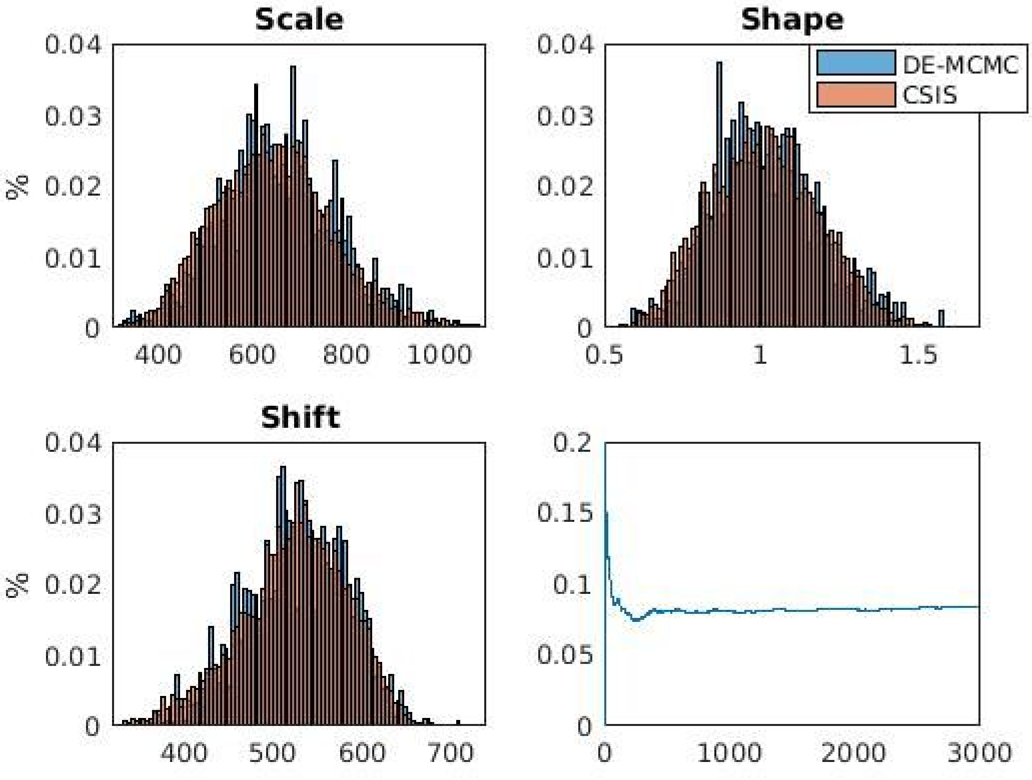


Fig 4.4: This figure illustrates the posterior approximations produced by DE-MCMC (blue) and SISFC (red) algorithms when the likelihood values are unbounded at the maximum. The distributional approximations are closely matched for both algorithms. The fourth graph shows the acceptance rate of DE-MCMC algorithm

4.9.3 Results

Fig 2.3.5 show the results of the two algorithms when a non-parametric pdf approximation was used. Both algorithms performed very similarly and produced similar posterior approximations.

Our introduction of the overfitting threshold has greatly improved efficiency of the DE-MCMC algorithm since acceptance rate has jumped to 8% which is much higher than acceptance rate of <2% observed in the bounded likelihood example. Moreover, such acceptance did not require running more generations for a proper comparison of the two algorithms. However, this is still the main advantage for SISFC algorithm since the number of proposals required to approximate the posterior distribution remained much lower. The algorithm used the same number samples as in previous cases to arrive to the posterior approximation. We found that approximating posterior distribution of Weibull function was sufficient to perform 10000 parameter evaluations while DE-MCMC used 50000 samples. However, SISFC had all samples added to posterior distribution approximation while DE-MCMC had roughly 6000 contributing samples despite evaluating 50000 samples in total. This suggests that SISFC algorithm is a better choice as long as additional computation time per each new proposal for using KDEs for posterior approximation is not 12x higher.

Moreover, unlike Weibull function, Gaussian function is not strictly bounded in any direction, thus there is no multi-modality in the results compared to the example when Weibull is used as a model approximator. As a result, our proposed algorithm does not have issues observed in bounded maximum likelihood example due to accepting all proposals with the same weight. However, the asymptotic bias of the shift parameter has expanded to the opposite direction compared to the same example because Gaussian PDF does not have equivalent hard-bound as Weibull PDF. Therefore, the drift in the median of the simulated distribution does not produce zero probability values which tend to dominate when Weibull is used.

4.10 Discussion

In this section we introduced a novel sequential importance sampling algorithm for approximating unknown posterior distributions. We present this algorithm in two implementations with minor modifications based on the situation with likelihood values. Sequential Importance Sampling by Clustering (SISC) is proposed for usage when likelihood values are known or contain a small error and likelihood function is unknown. The second implementation Sequential Importance Sampling by Filtered Clustering (SISFC) deals with the situation when likelihood function is unknown as well as likelihood values are unbounded at the maxima. We assessed these two modification in their respective applications by designing numerical experiments for scenarios with known likelihood function, with unknown likelihood function but likelihood values are strictly bounded and when both are analytically intractable. This new algorithm was compared with the performance of one of the best available posterior approximation algorithm Differential Evolution Markov Chain Monte Carlo (DE-MCMC) in these three scenarios. The results show that our proposed algorithm can be a default algorithmic choice in situations when each evaluation of the model is too costly for lengthy MCMC sampling posterior approximation approach.

In order to assess a proposed algorithm we have constructed hypothetical datasets equivalent to robust likelihood study with no contamination. The comparison of the algorithms is in their ability to handle unknown likelihood functions. Initially, we produced the situations for SISC algorithm. The first experiment illustrates the circumstance when likelihood function is known and demonstrated the performance of both algorithms. We found that both algorithms had little issue approximating the posterior distribution and there was little observable difference. In the next experiment we have considered the situation when likelihood function is not known but likelihood values are strictly bounded at the maximum likelihood because a more complex model is replaced with simpler model (Ratcliff, 1978). The performance of the two methods remained satisfactory though SISC struggled with a sharp drop in probability values at shift parameter. This is not surprising since accepting all values also accepts values outside the bound. It could have been improved by improving the accuracy and the number of the clusters but such decision would require more samples. This example illustrated a situation where accepting all proposals can lead to poor approximation of the posterior distribution. Regardless, there was no increase in the number of required samples for SISC algorithm and with tailed distributions the outcome would be much more satisfactory while DE-MCMC efficiency has dropped.

Afterwards we assessed SISFC algorithm in comparison to DE-MCMC. In this case the likelihood values are no longer bounded to maximum and possess very unlikely extremely good fits instead. In such situations approximate Bayesian computation (ABC) methods are used to construct a likelihood-free posterior distribution. We chose to introduce likelihood approximation and use these to estimate the 'true' maximum likelihood value which then is used as a tolerance threshold when likelihood approximations are being used. It is a novel idea to treat likelihood values as informative outcomes that define a proper tolerance threshold as a point of the best PDF approximation and better likelihood values are overfits. For a fair comparison, we used this threshold for both algorithms when model PDFs are estimated using KDE methods. Our constructed example combined with overfitting threshold has increased the efficiency of DE-MCMC sampler. However, the most striking result was in SISFC approximating posterior distribution with the same number of samples from posterior distribution while DE-MCMC had its efficiency drop.

These demonstrated examples used a single RT-distribution produced by Weibull model. Researchers are increasingly interested in mixed datasets with multiple RT-distributions (Pleskac & Busemeyer, 2010; Ratcliff & Starns 2009) and a combination of these distributions scales the uncertainty of the overall likelihood values since each RT-distribution approximation is unbounded. In these circumstances the DE-MCMC algorithm acceptance without very tolerant threshold is abysmal. Moreover, such models are explaining relatively complex cognitive processes and can contain a dozen of parameters to be estimated. These parameters in general have complex non-linear relationships that can reduce acceptance beyond reasonable levels due to proposals being frequently generated outside the dense space of the posterior distributions. DE-MCMC algorithm was applied in a couple of such models (Miletic et al., 2017; Turner, Sederberg, & McClelland, 2016) but this may be a feasible limit. In this research we have developed a novel model for visual search tasks which fits multiple distributions as well as have costly approximations for each parameter proposal (see Part II, Chapter 2 for more details) and posterior approximation using DE-MCMC is beyond reasonable scope under current implementations.

Our proposed SISC algorithm introduces an idea of approximating posterior distribution using a compressed mixture model by constructing a smaller clustered mixture model. This is a novel idea and to the best of our knowledge has not been implemented before. The algorithm clusters sampled parameters to form a compact posterior distribution and guide future samples. Such implementation improves efficiency of the proposal generation and constructs complex representation of parameter space. However, current examples were quite limited in their parameter space and more complex examples should be generated to test how it works. However, this algorithm can approximate multi-modal posterior distributions with different probabilities of each mode. DE-MCMC is capable to capture different modes (Turner & Sederberg, 2012) but modelling different relative probabilities for each mode could be tricky. Moreover, the algorithm may be able to approximate non-linear relationships further improving sampling beyond DE-MCMC algorithm's capabilities which is well suited for linear or near-linear relationships.

SISFC algorithm introduced an idea of identifying overfitting model predictions and use this as a boundary for posterior approximation. This idea is similar to tolerance threshold used when summary statistics are considered (Sisson, Fan & Tanaka, 2007; Beaumont et al., 2009) except that it is directly inferred from the likelihood outcomes to enforce the boundedness to the likelihood values. Working with overfitting threshold assumes that better results are accidental and non-representing of an actual model behaviour. As a result, the probability to produce better solutions than this threshold is assumed to be properly describing posterior distribution. At the moment we cannot prove or illustrate examples when this assumption would lead to incorrect approximations and a comparison to Beaumont et al. (2009) might be a good starting point for confirmation of the proposed method. However, a small drift away from optimal threshold expands the posterior distribution approximation and getting it poorly introduces a bias within approximation. Identifying it is subject to the accuracy of the likelihood value distribution at the best solution which has an approximation error thus a bias from a true posterior distribution is inevitable.

An alternative proposal for approximating posterior distribution is via a mapping function (Wilkinson, 2013; Turner & Sederberg, 2012). This function replaces the threshold with some probability function that maps deviation from the true model. This mapping function could follow the likelihood value distribution itself but normality assumption at the expense of efficiency for the sake of simplicity and accuracy can be taken as well. It is important to be cautious in using maximum likelihood value error distribution as a function since simply assigning the function corresponding to that distribution can produce highly misleading posterior. Recall an example with bounded likelihood error, it had a multi-modal situation which could also be the case for the best parameter values. Under such circumstances, a large probability would be assigned to non-performing likelihood values such as zero. This mapping method was designed for summary statistics as well with the probability of the deviation from the best solution being a replacement for fitness value. However, in terms of unbounded likelihoods that would indicate a very unlikely value and we ruled this out by designing overfitting threshold. Regardless, we do believe that it can improve posterior approximation by reducing reliance on identifying overfitting threshold to some degree. However, further research is required to identify the right probability function.

To sum up, we have developed a novel algorithm for posterior approximations. Showed that it is well-suited for parameter approximations when analytically intractable models are used when a concept of overfitting is introduced. We compared this algorithm to a recently developed DE-MCMC algorithm (Turner, Sederberg, 2012) and showed that it is more efficient at approximating posterior distributions even when the considered model is relatively simple. These features are highly attractive for future cognitive model development and analysis.

General Discussion

1 Summary

In the three chapters of this part of the research we have discussed various methods available for modelling cognitive processes. We focussed on modelling analytically intractable models which involves constructing approximation of the unknown probability density function (PDF), designing appropriate goodness-of-fit measure, designing efficient parameter estimation algorithm and dealing with excessive uncertainty around likelihood values. All these features are an important contribution to the cognitive modelling on their own but can be combined to contribute to our capabilities to do Bayesian analysis of complex models. It is a challenging problem that is being studied in a wide range of fields and new methods are constantly being developed. Currently, the Bayes inference methodology for such models is still quite limited and most of the methods are mainly available for multivariate distributions that are constrained by a single mode and linear or near-linear relationships.

In the first chapter we introduce an improved method to approximate response time (RT) distributions using a compact online kernel density estimator (oKDE) designed by Kristan, Leonardis, and Skočaj (2011). This novel KDE method outperforms usually used (Turner, van Zandt & Brown, 2011; Turner & Sederberg, 2014; Miletic et al., 2017; Turner, Sederberg, & McClelland, 2016) Silverman's KDE (Silverman, 1989) when lower RT datasets are being used and is more compact for larger datasets. As a results, its usage is more efficient in practice.

The second chapter dealt with the contamination of RT-distributions which is a commonly known issue (Miller, 1991; Ulrich & Miller, 1994). We suggested to exploit the usage of KDE to introduce robustness properties to the likelihood function and showed that it has similar properties as Hellinger distance which is known for its robust properties (Lindsay, 1994). The results were very promising and a wider adoption is possible to replace current practice of modelling contamination itself (Ratcliff & Tuerlinckx, 2002; Ratcliff & McKoon, 2008; Wagenmakers et al., 2008).

The third chapter dealt with implications produced by the introduction of the methods in the first two chapters. The use of these methods creates unbounded likelihood values making most of the algorithms highly inefficient. In this chapter we proposed a new sequential importance sampling (SIS) algorithm as well as overfitting concept to improve efficiency of the sampling algorithms. We show that these introductions significantly reduce the number of required samples to approximate unknown posterior distribution compared to recently developed Differential Evolution-Markov Chain Monte Carlo (DE-MCMC: Turner et al., 2013).

2 Future Directions

The three covered chapters dealt with issues of representing response time distributions, contamination and Bayesian approximation when these methods are used. Though Bayes is a very powerful method it is often limited for analytically tractable models and extensions to difficult problems usually require to make a number of underlying assumptions to be able to perform Bayes inference. However, these assumptions can be perilous for models that are not well understood. When model is not tractable, computing marginal requires approximation of the unavailable likelihood function which in itself is a challenging and complicated issue. Furthermore, providing a good prior is often overlooked issue. This complication is one of the main reasons the modelling was focused on the best solutions informed by local parameter curvature rather than Bayes methods (Jeffries, 1946; Ly et al., 2017). In fact, defining proper non-informative priors for ill-defined models is impossible and relying on them may lead to inconsistent posterior approximations. It can be shown that Jeffrie's prior is improper (does not integrate to 1) for ill-defined models and posterior under such circumstances also is not a proper probability distribution. Furthermore, a non-informative prior is often preferred for model comparisons and an intuitive choice of a uniform prior is not a proper prior since it also does not integrate to 1 and resulting posterior is not a pdf. However, improper posterior can be used as an approximation of the proper posterior. The main issue within uniform prior is that it still does not imply being non-informative prior because such thing as a non-informative prior does not exist. By choosing uniform prior we make an assumption that parameter space is equally distributed which may not, and generally is not, the case. A correct non-informative prior has the minimum possible influence on data and such a prior has to be provided to perform correct Bayesian inference (Liu et al., 2014). In general, there are two methods to arrive to non-informative priors: Jeffrie's and reference prior (Jeffries, 1946; Bernardo, 1979; respectively) and neither is analytically tractable if the model isn't. This substantially limits the possibility of Bayesian inference when complex models are being used. MacLachlan and Peel (2000) argued that non-informative priors should not be used for such problems since arising posteriors from improper priors are meaningless. However, simply opting to mildly informative prior may not be a proper choice as well. Recently, Nalisnick and Smyth (2017) proposed learning prior distributions when it cannot be determined analytically and show that their method converges to a proper Jeffrie's prior. They also showed a model example for which scientists often apply normal distribution as a prior and found that a proper non-informative prior is completely different from usually applied Normally distributed prior. In psychology, Turner and Sederberg (2014) suggested the usage of approximate methods for understanding analytically intractable models such as LCA (Usher & McClelland, 2001) and used mildly informative normally distributed priors to make observations about the model but most likely it suffers from the same issues outlined above.

Our model is analytically intractable as well and its' parameter behaviour is not well understood. In order to be able to perform Bayesian inference, non-identifiable parameters have to be identified and removed and only then we would be able to approximate a non-informative proper prior for our model, therefore we focused on finding the maximum likelihood (ML) rather than approximate posterior distributions for inference purposes. However, posterior approximation methods still are vital for identifying these ill-defined parameters. The most common approach to identify these parameter dimensions is by using local sensitivity around the best solution (Sun & Hahn, 2006). Such approach ignores how the model behaves globally to reduce the complexity of the system, however, this approach fails to appreciate that these parameters may have different global sensitivities and reparameterisation would lead to removal of other (if any) parameters. As a result, Bayesian inferences under these circumstances are expected to differ. An alternative for parameter reduction uses profiling, this is a special case of Bayes which uses local MLEs rather than marginal likelihoods as in Bayes. Ronald Fisher and others (Ly et al., 2017) propagated and focused on frequentist approach rather than Bayes because it did not suffer from the issues with poor priors. When the model is poorly specified, we know that defining proper priors is intractable, therefore profiling is an equivalent better choice to picking the right model as was

MLE in 50s (Raue et al., 2009). It is important to note that removing parameters based on their curvature can be misleading (?) showed an example where such choice completely changes the prediction the model entails. However, parameters with very low curvature may be impossible to approximate thus either additional data has to be gathered or removed from consideration.

3 Conclusions

Our research was limited to analytically intractable modelling for visual search tasks but the methods are transferable to other fields of the research because it introduces a few methods that arguably are better approaches than many currently used methods as we illustrated with commonly used approaches in cognitive psychology. We showed that the approximations of the RT-distributions can be improved and the removal of the contaminants is not necessary when a good goodness-of-fit is chosen. Moreover, we dealt with the issues produced by these methods by introducing a novel algorithm informed posterior approximation. There’s still a lot of research to be done on these methods to identify their additional weaknesses or strengths but the results shown here are promising for the further development of cognitive modelling in psychology.

Part 2:
Visual Search

Chapter 5:
Selective Attention for Identification Model - Winner-Take-All (SAIM-WTA) model

Abstract

For 50 years or so, visual search experiments have been used to examine how humans find behaviourally relevant objects in complex visual scenes. For the same length of time there has been a dispute over whether this search is performed in a serial or parallel fashion. In this paper, we approach this dispute by numerically fitting a serial search model and a parallel search model to RT-distributions from three visual search experiments (feature search, conjunction search and spatial configuration search). In order to do so, we used a free-likelihood method based on a novel kernel density estimator (KDE).

The serial search model was the Competitive guided search (CGS) model by Moran et al. (2013). We were able to replicate CGS's ability to model RT-distributions from visual search experiments and demonstrated that CGS generalizes well to new data. The parallel model was based on the biased-competition theory and utilized a very simple biologically-plausible winner take all (WTA)-mechanism from Heinke and Humphreys's (2003) Selective Attention for Identification model (SAIM). With this mechanism, SAIM has been able to explain a broad range of attentional phenomena but it was not specifically designed to model RT-distributions in visual search. Nevertheless, the WTA was able to reproduce these distributions.

However, a direct comparison of the two models suggested that the serial CGS is slightly better equipped to explain the RT-distributions than the WTA-mechanism. The CGS's success was mainly down the usage of the Wald-distribution which was specifically designed to model visual

search. Future WTA versions will have to find a biologically plausible mechanism to reproduce such a RT-distribution. Finally, both models suffered from a failure to generalise across all display sizes. From these comparisons we developed suggestions for improving the models and motivated empirical studies to devise a stronger test for the two types of searches.

5.1 Introduction

Its most recent instalment, VS-SAIM, was also able to simulate visual search experiments (Heinke & Backhaus, 2011). However, in this study SAIM was evaluated in a qualitative way. This research aims to rectify this shortcoming by taking SAIM's core mechanism, a Winner-Take-All network; and evaluating it quantitatively with RT-distributions taken from Lin et al.'s (2015) data. It is also worth noting that Bundesen's (1990) biased-competition model (TVA) was fitted quantitatively albeit based on mean RTs. Of course, any parallel search model is likely to represent a simplification of how visual search is implemented and a serial component is certainly part of a fuller picture. For one thing, eye movements are an obvious candidate for a serial component (see Hulleman & Olivers, 2016; on the importance of eye movements). It is also conceivable that some serial 'internal rejecting' after a parallel search may take place (see SEarch via Recursive Rejection SERR; Humphreys & Müller, 1993; for an example). It is also clear that such serial components are particularly important in target-absent trials (even though perhaps not in terms of eye movements as discussed in Wolfe, 2007). Hence, we will focus in this research on present trials. Nevertheless, previous studies with parallel-only models such as SAIM imply that such a parallel approach may go a long way in explaining visual search, particularly for target-present trials. Here, we will follow this up and utilize RT-distributions to further evaluate this approach. Moreover, the direct comparison with a serial approach will give us a good insight into what may be missing in the parallel-only approach. In fact, CGS represents a particularly strong challenge as CGS is developed especially for modelling visual search while SAIM was developed to account for a broad range of attentional effects. Another argument for fitting a simple parallel model comes from the methodology issue posed by utilizing RT-distributions.

5.2 Visual search experiment

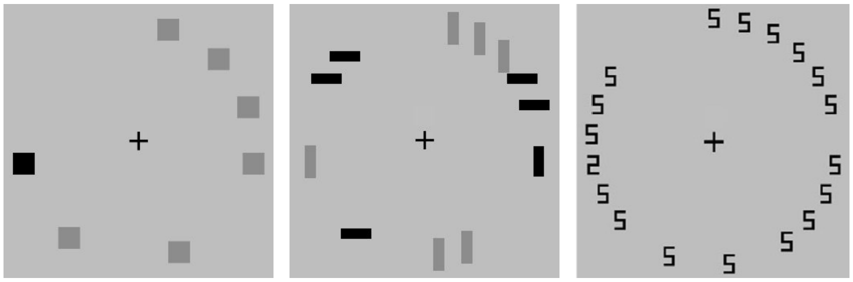


Fig. 5.1: Schematic representation of the tasks. From left to right: feature search, conjunction search and spatial configuration search.

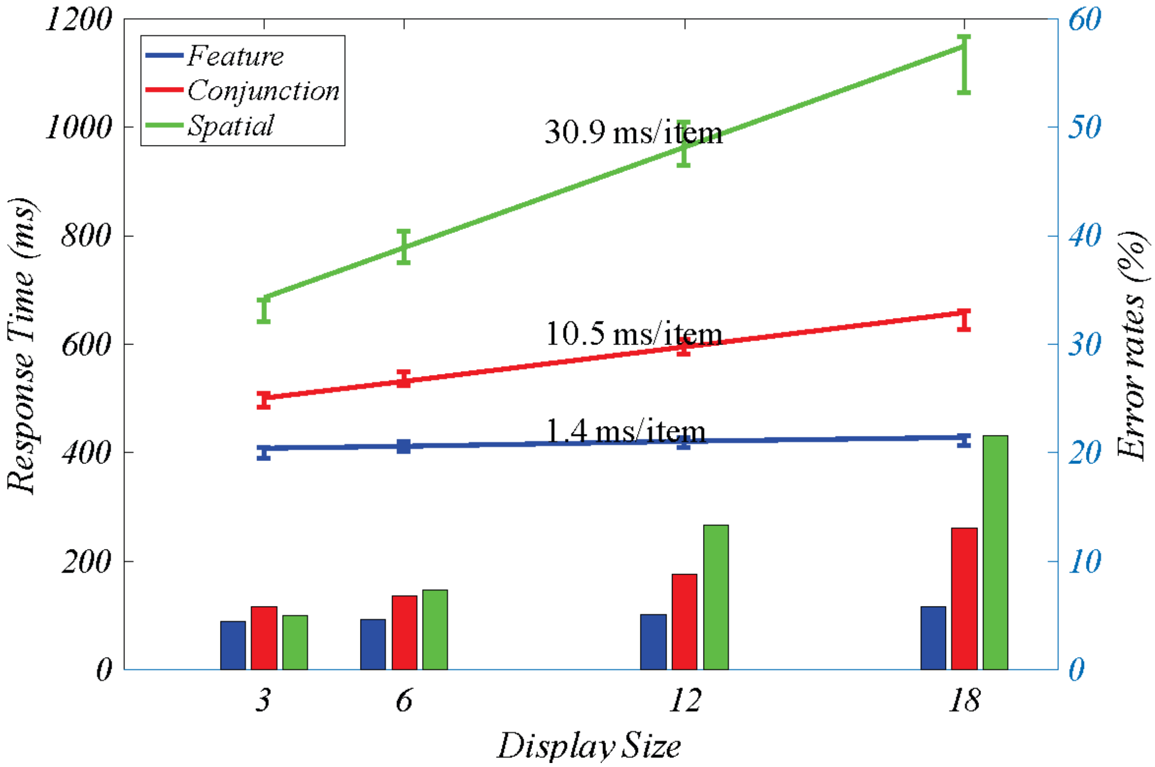


Fig. 5.2: Search functions of the three tasks (feature search, conjunction search and spatial configuration search) used in this study. The error bars indicate the standard error.

The data used for this research was collected as part of Lin et al.'s (2015) experiments. Details of the design can be found in Lin et al. (2015). The search displays were arranged in a circular layout (see Fig. 1; for an illustration) in which items can be placed in 25 locations. The display size was either 3, 6, 12 or 18 items. Each condition comprised 100 trials. Three different search experiments were conducted: feature search, conjunction search and spatial configuration search. In the feature search task, the target was a dark square while the distractors were gray squares. In the conjunction search, participants looked for a vertical dark bar amongst two types of distractors, vertical gray bars and horizontal dark bars. The spatial configuration task used two items, digit 2 (target) and digit 5 (distractor). Each search task was completed by 20 participants though one participant has been removed from feature and conjunction search tasks due to high error rates. Figure 2 shows the resulting search functions for the present condition. In addition, Lin et al. (2015) found that for feature search and conjunction search the RT-distribution's skewness increased with increasing display size. For spatial configuration search the relationship between skewness and display size was more complex. The skewness first increased over the smaller display sizes (3, 6, 12) but then decreased from 12 to 18. We will return to this finding in the discussion section.

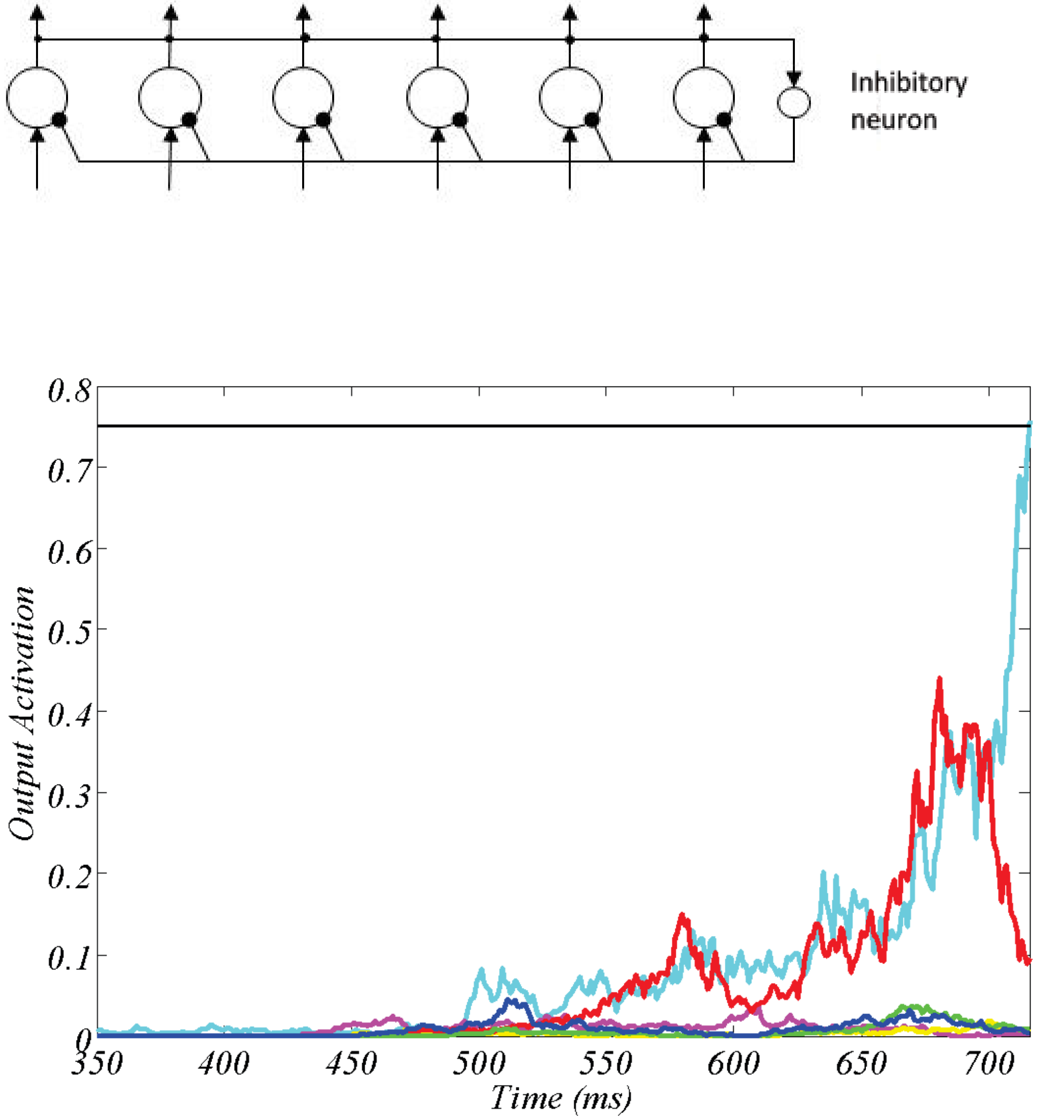


Fig. 5.3: The graph at the top shows SAIM-WTA's architecture. All nodes compete via global inhibitory neuron. The time course at the bottom shows an example of SAIM-WTA's output activation. The line colours correspond to the network's nodes. The dotted line indicates the threshold for this particular simulation. The simulation result came from the spatial configuration search with 6 search items with the parameters from the 8th participant (see appendix I).

The results indicate that the data provides a good basis for testing the models on a range of search task difficulties similar to Wolfe et al.'s (2010) data. However, there is an interesting difference between their data and our data. In our study, participants made roughly twice as many as errors as in their experiments. This difference can be explained by the fact that in Wolfe et al.'s (2010) experiments participants completed all three tasks with 500 trials in each condition. In contrast, in our study participants completed only 100 trials per condition and not all tasks. In other words, in Wolfe et al.'s (2010) study participants were highly practiced compared to our study. In fact, when we re-analysed Wolfe et al.'s (2010) data and included only the first 100 trials of each condition the error rates were similar to our error rate. Hence our dataset poses the interesting challenge to CGS whether CGS will also be able to model less practiced participants.

This dataset was chosen so CGS model would be tested on alternative dataset and challenged by parallel models. This dataset had unlimited duration which allowed the participants to choose their preference on speed or accuracy. As a consequence, this dataset allows eye movements that are not measured and by extension serial processes. The main criticism of such experiments is its inability to discriminate between serial vs parallel theories because parallel models can explain flat as well as steep RT slopes (Townsend & Ashby, 1983; Townsend & Wenger, 2004). An alternative experimental paradigm uses short display duration to prevent eye movements, thus rule out the possibility of serial search (e.g. Palmer, 1995, Palmer, Verghese, & Pavel, 2000; Verghese, 2001). Since parallel models can mimic serial data but not the other way around, long display duration study is more valuable if it can falsify parallel theory under some conditions.

This study aims to assess whether this is also the case for entire RT-distributions as well as improve available parallel models in case it is viable.

5.3 Computational models

5.3.1 SAIM’s Winner-Take-All (SAIM-WTA)-model

The biased competition model is based on the SAIM’s WTA-mechanism (Heinke & Humphreys, 2003; Heinke & Backhaus, 2011; Zhao, Humphreys & Heinke, 2011). This WTA-mechanism uses a single layer of ‘neurons’ which are connected by a lateral inhibition (see Figure 3a; for an illustration). If the correct parameters are chosen, the neuron with the highest input is activated while all other neurons are shut down (see Figure 3b for an exemplar simulation result). In other words, all neurons compete with each other and the neuron with the largest input wins the competition. The mathematical description of the model is the following:

$$dx_i =- \tfrac{dt}{\tau}x_i + \tfrac{dt}{\tau}(w(\sum_{j=1}^{n} y_j) + I_i) + \sqrt{\tfrac{dt}{\tau}}\xi_i \, ;$$

$$y_j = f(x_j) = \frac{1}{1+e^{-m(x_j-s)}} \, ; \qquad \xi = N(0,\sigma) \, ;$$

f(x) is a sigmoid function with parameters slope (m) and intercept (s). τ is the accumulation rate of input activation and w is the strength of the lateral inhibition. ξ is Gaussian noise with the variance (σ). I_i is the input to the *i*th neuron. y_i is the output activation of *i*th neuron. x_i is the internal activation of the *i*th neuron. These equations are based on a mathematical description of neurophysiological processes using a spiking-rate neuron model. The Gaussian noise takes into account the randomness of neural processes. The sigmoid-function models the non-linear relationship between cell activation and output spiking rate. The differential equation models the leaky accumulation behaviour of synapses. The summation term realizes the lateral inhibition within the layer (inhibitory neuron).

To adapt the model to modelling visual search data, we made several simple assumptions. Each 'neuron' is assumed to correspond to an item location in the search display. If a location is empty the input is set to zero. The neuron for the target location is set to one while the distractor neurons are set to a saliency value. To model the response time, we introduced a decision boundary and computed the time it takes for a neuron to pass this threshold. If it is a distractor neuron the response is recorded as 'target absent' whereas if it is target neuron the response is 'target present'.

It is worth noting that SAIM-WTA is similar to the Leaky Competing Accumulation (LCA) model (Usher and McClelland, 2001). However, to the best of our knowledge, LCA has never been applied to visual search. Moreover and similar to LCA, SAIM-WTA stands in the tradition of the Parallel Distributed Processing (PDP) framework (Rumelhart &McCleland, 1986) in that it draws on principles of neural information processing in order to understand phenomena at the behavioural level (see also Mavritsaki, Heinke, Allen, Deco & Humphreys, 2011 for a discussion of linking the neural level with the behavioural level through means of computational models).

In addition, some mechanisms and conceptualisations are also similar to stochastic drift diffusion models (e.g., Busemeyer & Diederich, 2010; Ratclif, 1978). These models assume that perceptual decision making is based on an accumulation of perceptual information. Once, this accumulation has reached a certain level (i.e., threshold) a decision is made (i,e, a response is generated). The time it takes for the accumulation to reach the threshold is interpreted as response time. SAIM-WTA can be framed in terms of these drift diffusion models in that

SAIM-WTA's model accumulates information about the search items (i.e., identifies them) and once this information has reached a certain level the model / participant initiates a corresponding response. However and different from drift diffusion models, the accumulators interfere with each other.

SAIM-WTA has seven free parameters. In explorations of the parameter space prior to work presented here, we found several regions where it was possible to achieve a similar quality of fit. Subsequently, we focused on a region where it was possible to reduce the number of free parameters to the smallest possible number (i.e., three) while still obtaining the best fits for all participants (see appendix I for the values of the fixed parameters). In addition, the remaining free parameters (accumulation rate, decision boundary and distractor saliency) allowed us to ask interesting theoretical questions about the factors which influence visual search performance. Given the biased-competition theory's assumptions, we expect that the distractor saliency (target-distractor similarity) increases with task difficulty, but it is not clear if the distractor saliency is sufficient to explain the differences between the tasks or is there also a difference in terms of the difficulty of identifying items (as expressed by accumulation rate)? In fact, computational models such as SAIM (and CGS) suggest the involvement of a separate object identification stage.

5.3.2 Competitive guided search (CGS)

Moran et al.'s (2013) CGS implements a serial search based on Wolfe's (2007) two stage architecture. The guidance through the saliency map is implemented through a probabilistic selection where target item has the probability to be selected:

$$p_{target} = \frac{w_{target}}{n-1+w_{target}}.$$

and the distractor items:

$$p_{distractor} = \frac{n-1}{n-1+w_{target}}.$$

w_{target} is the saliency of the target relative to the distractor. If the target saliency is smaller than one there is no guidance. n is the number of items currently available for selection and is decremented after each search step. Hence, CGS assumes that once an item is identified it is not revisited again. Prior to each search step, CGS decides whether to continue with the search, or to quit the search and decide that the target is absent. The probability to quit is calculated in the following way:

$$p_{quit} = \frac{w_{quit}}{w_{quit}+n-1+w_{target}}$$

Again n is decremented after each search. Hence, the probability to quit increases with each search step. This effect is increased further by modifying w_{quit} at each search step

$$w_{quit,\, new} = w_{quit,\, old} + \Delta w_{quit}$$

Note that the value of $w_{quit,}$ is zero at the beginning of the search.

At each search step, CGS assumes that an item is identified as to whether it is a target or a distractor. This identification process is modelled as a drift diffusion process which is also used to describe the behaviour of SAIM-WTA's nodes. However, instead of simulating the drift diffusion process Moran et al. (2013) used the Wald-distribution to represent the distribution of the identification time (i.e., passing of threshold):

$$p_{wald}(\alpha, v, \sigma) = \sqrt{\frac{(\frac{\alpha}{\sigma})^2}{2\pi x t^3}}\, e^{\frac{-(\frac{\alpha}{\sigma})^2 (t-\frac{\alpha}{v})^2}{2(\frac{\alpha}{v})^2 t}}$$

with the three parameters identification drift rate (υ), identification threshold (θ) and noise level (σ). The noise level was fixed at 0.1 throughout the studies. The total response time is the sum of all identification times from all search steps. Moran et al. (2013) also chose this distribution as Palmer et al. (2011) found this distribution to be the best distribution to describe Wolfe et al.'s (2010) data.

Note that mathematically the probability distribution of the sum of independent random variables can be determined by convolving the probability distributions of the individual random variables. Hence, CGS's total response time can be described as multiple convolutions of a Wald-distribution where the number of convolutions depends on the number of search steps. One consequence of this convolution is that CGS's RT-distribution is more and more skewed the more search steps take place. Hence, CGS should produce an increase in skewness with increasing display size (depending on the search task). This relationship should enable CGS to model RT-distributions from visual search tasks.

CGS also assumes that at the response execution stage an erroneous response can occur due to a motor error with a certain probability (m). Since the identification stage is assumed to be perfect, misses of targets can only occur through motor errors. It is also worth noting that motor errors can "correct" misses as it is possible that search terminates without finding the target but due to an error the model still reports "target present". Finally, a residual time accounts for the duration of processes which are outside the actual search process such as encoding of items, post-decisional processes, response planning and execution. The residual time is assumed to be distributed as a shifted exponential distribution with non-decision shift time (T_{min}) and non-decision drift time (γ) as parameters.

5.3.4 Discussion

Apart from implementing two different types of searches, the models relate differently to the neural substrate. SAIM-WTA aims to be ‘biologically-plausible’ while CGC is less rooted in neural processes, even though the identification stage has a similar link (drift diffusion model) to neural processes as SAIM-WTA. Also both models have in common that they assume item identification plays a critical role in visual search.

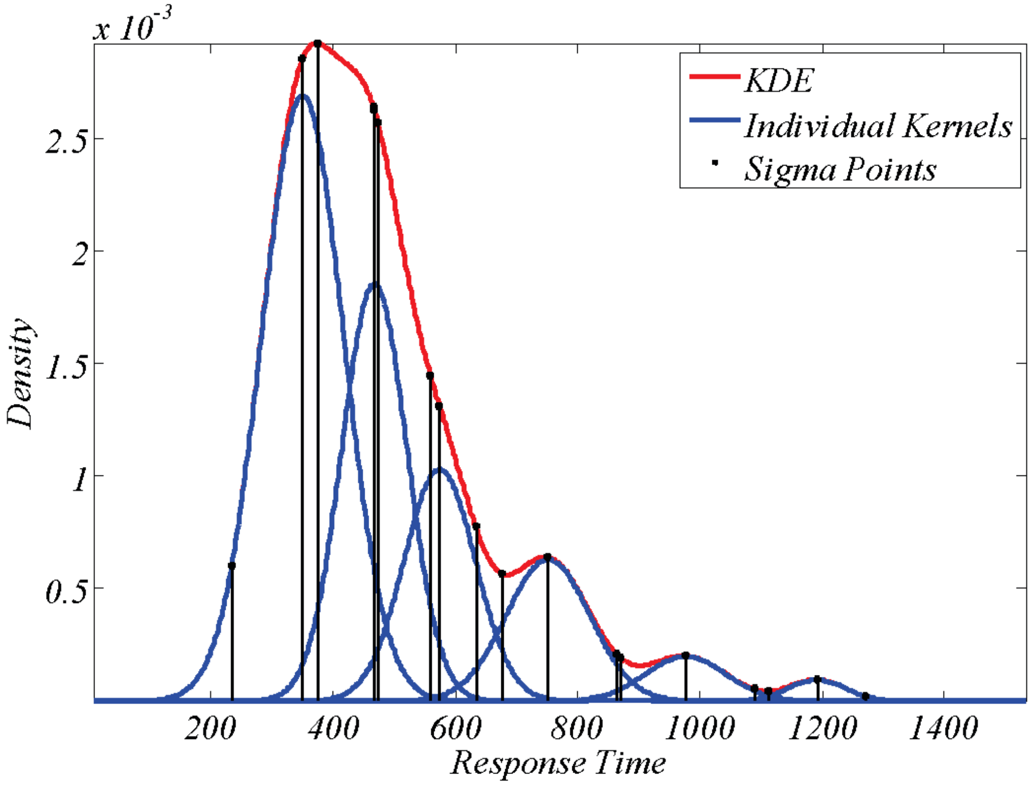


Fig. 5.4: An illustrative example of how oKDE decomposes a distribution in several kernels with varying variance.

The selection process from the saliency map is seen by CGS's authors as an approximation of a competition process (hence *competitive* guided search). However, the approximation does not involve interference between items as SAIM-WTA implements. Hence, the selection process is probably better understood as a randomized selection process which is modulated by item saliency.

However, for the purpose of this research these differences and commonalities are less important. More important is the fact that SAIM-WTA has fewer free parameters than CGS. SAIM-WTA absorbs CGS's stages (such as identification stage, encoding stage, etc.) into the competition process. Hence, a numerical comparison between the models can look at whether this more parsimonious model is more successful than a more complex model.

5.4 Materials and Methods

Both models and all the methods used were implemented using Matlab. Both models were fitted to each participant separately. The resulting quality of fit and parameter settings were averaged in order to represent the population level.

The best fitting parameter settings were determined using the maximum likelihood estimation (MLE). MLE allows us to base model fit on RT-distributions together with accuracy. In order to employ MLE, traditionally it is necessary for models to possess an analytic probability density function (PDF). However, models such as SAIM-WTA or GCS don't possess such PDFs. Recent developments in model fitting, often termed approximate Bayesian computation (ABC) or 'likelihood-free methods' (see Beaumont, 2010; for a review) solve this issue by approximating the model's PDFs. Here, we utilized a likelihood-free method based on a KDE-approach which estimates the model's PDF for a given parameter setting using Monte

Carlo sampling (see also Turner & Sederberg, 2014). We utilised the online kernel density estimator (oKDE) introduced by Kristan, Leonardis and Skočaj (2011). We used DE-MCMC algorithm for parameter estimation for both models (Turner, Sederberg, Brown & Steyvers, 2013, see chapter 4 for description).

5.4.1 Likelihood-function

The fit of the model with the data was evaluated with the likelihood principle using a pdf for mixed data (Turner and Sederberg, 2014):

$$L(\theta) = \prod_{j=1}^{2} \prod_{i=1}^{X=s_j} P(s_j) M_j(x_i\,|\theta)$$

where P(s_1) is the probability of correct response and P(s_2) is the probability of incorrect response. M_j(x_i|θ) denotes the model's probability density function (pdf) for any observation x_i, the parameters θ and response type. As stated earlier, here the model's pdf is represented by a kernel density estimate (KDE). However, this likelihood function is not fully suitable for our modelling approach, as we don't consider the response times for incorrect responses. Therefore, Turner and Sederberg's (2014) equation turns into:

$$L(\theta) = (1 - P(X_c))^{n_i} \prod_{j=1}^{n_c} P(X_c) M(x_j|\theta)$$

Where n_i indicates the number of incorrect responses and n_c is the number of correct responses. P(X_c) is the probability of correct responses in the model.

5.4.2 Comparison with Moran et al.'s (2013) methods

To fit parameters, Moran et al. (2013) used the popular algorithm by Nelder and Mead (1965) which is implemented in MatLab's fminsearch. However, it is also well-known that this method is very sensitive to the choice of the starting points of the parameter Estimation. Our DE-MCMC method reduces this problem by using a population of starting points[1]. Moreover, to estimate the RT-distributions Moran et al. (2013) employed the commonly used Quantile Maximal Probability (QMP) method by Heathcote, Brown, and Mewhort (2002). However, Turner and Sederberg (2014) showed that this method can lead to misleading results.

Thus, given the differences between our approach and Moran et al. (2013), attempting to replicate Moran et al.'s (2013) parameter settings is unlikely to be successful. However, to demonstrate that our approach is more reliable than Moran et al.'s (2013) approach, we fitted CGS to Wolfe et al.'s (2013) data twice, using different starting points. First, Moran et al.'s (2013) parameter settings from the individual participants were used as starting points for the parameter estimation. Even though these starting points are unlikely to be the best fits given the differences in methods, they should at least be close to very good solutions which DE-MCMC would be able to find. Second, we used our parameter settings established by fitting Yishin et al.'s (2015) data. Interestingly, the quality of fit for Moran et al.'s (2010) parameter settings as starting points was not as good as for our parameter settings as starting points. Hence, we

[1] This sensitivity to the starting point of a search is due to the fact that complex models like the ones used here have many local solutions. These local solutions are the best solutions in particular areas of the parameter space but it is not clear whether a particular local solution is the overall best solution (global solution). Most, if not all, methods for parameter estimation find (get trapped in) local solutions and cannot guarantee that this is the global solution. Amongst other factors the starting point of the search is critical for which local solution is found. Broadly speaking, search algorithms tend to find local solutions near the starting point.

conclude that our parameter settings generalize better across different datasets while Moran et al.'s (2015) settings seem very specific to their chosen starting point of the search.

5.4.3 Removal of outlier parameter settings

It turned out that some participants' parameter settings were extreme. We therefore applied an outlier elimination method, the median absolute deviation (MAD; Leys, Ley, Klein, Bernard, & Licata, 2013), to each parameter in each task. As criterion for an outlier we used five standard deviations. A participant was identified as outlier if at least one parameter value was considered to be an outlier. This participant was removed from the further analysis.

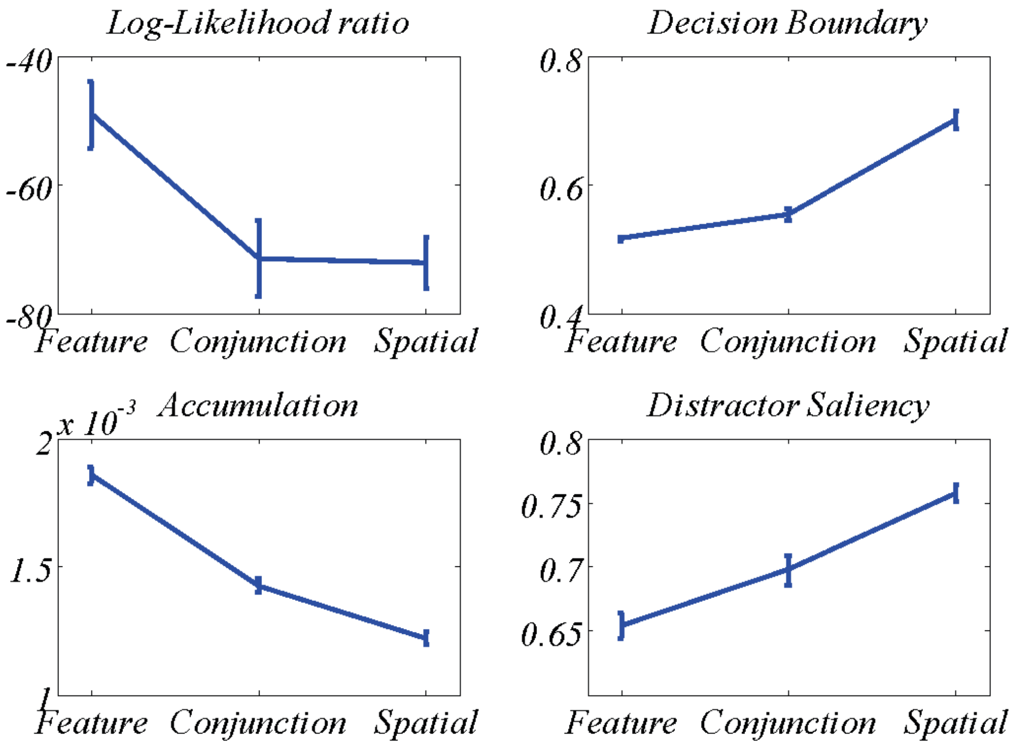


Fig. 5.5: Results of fitting SAIM-WTA. The top-left graph shows the mean log-likelihood ratios (quality of fit) for the different tasks. The remaining graphs show the mean parameters. The error bars indicate the standard error.

5.5 Results and discussion: SAIM-WTA

5.5.1 Results

We fitted SAIM-WTA with three free parameters (distractor saliency, decision boundary, accumulation rate) to Lin et al.'s (2015) 58 datasets from three visual search experiments, feature search, conjunction search and spatial configuration search. Hence we obtain 58 parameter settings (see appendix II for values), 19 parameter settings (participants) for feature search, 19 settings (participants) for conjunction search and 20 settings (parameters) for spatial configuration. Eyeballing the parameter settings we noticed that there were a few settings which could be considered as outliers. Our outlier detection procedure led to the removal of 2 participants from feature search, 3 settings from conjunction search and no participant from spatial configuration search.

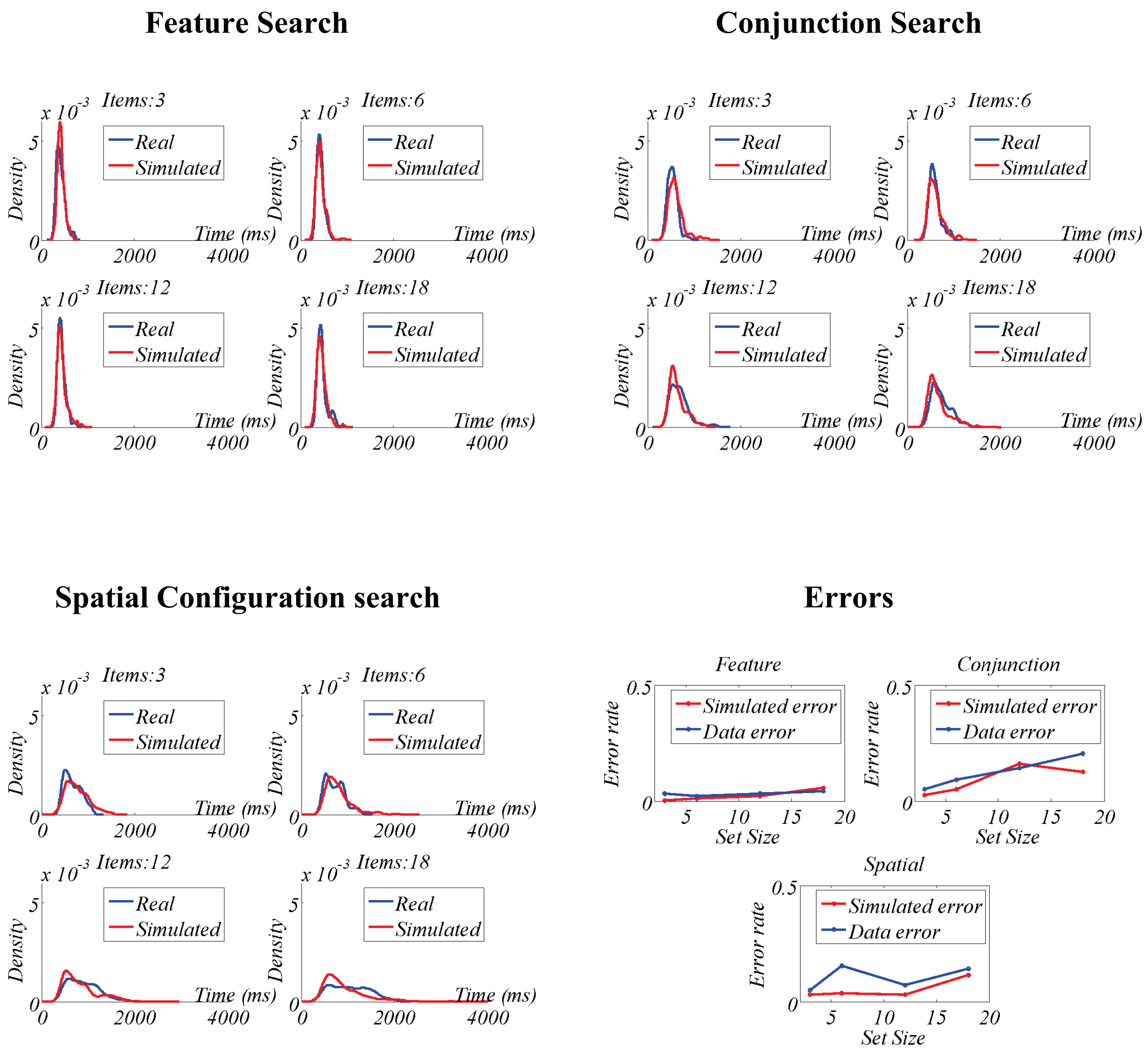


Fig. 5.6: SAIM-WTA: KDE-based illustration of RT-distributions and response errors. Note that these graphs show the RT distributions for three participants.

To assess the overall fit for each participant we calculated the log likelihood ratios, log likelihood value from the model divided by the log likelihood value of the KDE's dataset (see Fig. 5 for the results). We compared ratios from the different tasks with the Wilcoxon rank-sum test and found a significant decline of ratios between feature search and conjunction search ($z = 2.36$; $p = 0.018$); and feature search and spatial configuration ($z = 2.85$; $p = 0.004$). There was no significant difference between conjunction search and spatial configuration search

$(z = 0.14; p = 0.886)$. To illustrate the quality of fit (likelihood ratio), Fig. 6 shows the outcome from three participants. Note that the choice of these participants was made randomly by MatLab to avoid an author bias. The likelihood ratio was -46.32 for feature search; -61.10 for conjunction search and -71.98 for spatial configuration search. The graphs indicate that SAIM-WTA was able to produce an increased skewness with increased display size. This increase broadly matched the increase of skewness in the data, but not to the same degree. The failure to match skewness is particularly pertinent in spatial configuration search for display size 18. This effect is illustrated further in Fig. 10 where the likelihood ratio declined with increasing display size. Nevertheless, it is important to note that the only source of this effect is the increase of number of distractors in the input of the model since all parameters are kept constant. In other words, the competition between items due to lateral inhibition is able to explain the skewness found in the visual search.

Figure 5 shows how the three free parameters (distractor saliency, decision boundary, accumulation rate) changed across the three tasks. The parameters were entered into a Wilcoxon rank-sum test. For accumulation rate there was significant difference between feature search and conjunction search (z=4.845, p<0.001) as well as between conjunction search and spatial configuration search (z=4.056, p<0.001). For distractor saliency the difference between feature search and conjunction search was significant (z=-2.684, p=0.007) and so was the comparison between conjunction search and spatial configuration (z=-3.659, p<0.001). Finally for decision boundary both comparisons were also significant (z=-3.368, p<0.001; z=-4.619, p<0.001). Hence all comparisons were significant. As expected, the distractor saliency increased with increasing task difficulty. However and interestingly, the other two parameters were also related to task difficulty. Accumulation rate decreased with increasing task difficulty, i.e., framed in terms of drift diffusion models, discussed earlier, the accumulation of perceptual information about items was more and more difficult. Finally it is worth noting that the level of decision boundary increased with task difficulty. A closer inspection of this finding showed that it is a by-product of the increase of competition leading to higher activations of distractors requiring higher decision boundaries to avoid response errors.

5.5.2 Discussion

On the whole, SAIM-WTA exhibited a reasonable fit with three search tasks. In particular, it was able to model the increasing skewness with increase in display size. However, the quality of the fit decreased with task difficulty. In particular, SAIM-WTA was not able to match the increase in skewness with increased display size.

The three free parameters (distractor saliency, accumulation rate, decision boundary) were different for the three tasks. Distractor saliency increased across tasks as predicted by all major theories on visual search. Importantly, SAIM-WTA identifies distractor saliency as an important source for the increase in skewness across display sizes. The results also identify an increase of accumulation rate with task difficulty. As discussed in the introduction, accumulation can be linked with object identification. In other words, SAIM-WTA also indicates that not only does selection of target get more difficult across the three tasks, but also object identification gets harder. This is consistent with other theories on visual search such as biased-competition theory, and also with computational models such Moran et al.’s model and SAIM. However, both models consider object identification as a separate processing stage. Moreover, this finding offers an interesting explanation for findings that responses in pop-out searches can be speeded up even further by additional manipulations such as priming (e.g., Maljkovic, Nakayama, 2000; Woodgate et al., 2015). SAIM-WTA suggests that the speed-ups are due to improved target identification. Fitting SAIM-WTA to data from priming experiments should support this prediction.

Finally, the increase of the decision boundary makes an interesting prediction for an application of the Speed-Accuracy-Tradeoff (SAT)-procedure in visual search experiments. In a SAT-procedure participants' response time is controlled by requiring them to respond within a set time window (e.g., Wickelgren, 1977; Zhao, Heinke, Ivanoff, Klein & Humphreys, 2011). For early time windows where participants have to sacrifice accuracy for speeded responses, SAIM-WTA should detect lower decision boundaries, but distractor saliency and accumulation rate remain unchanged.

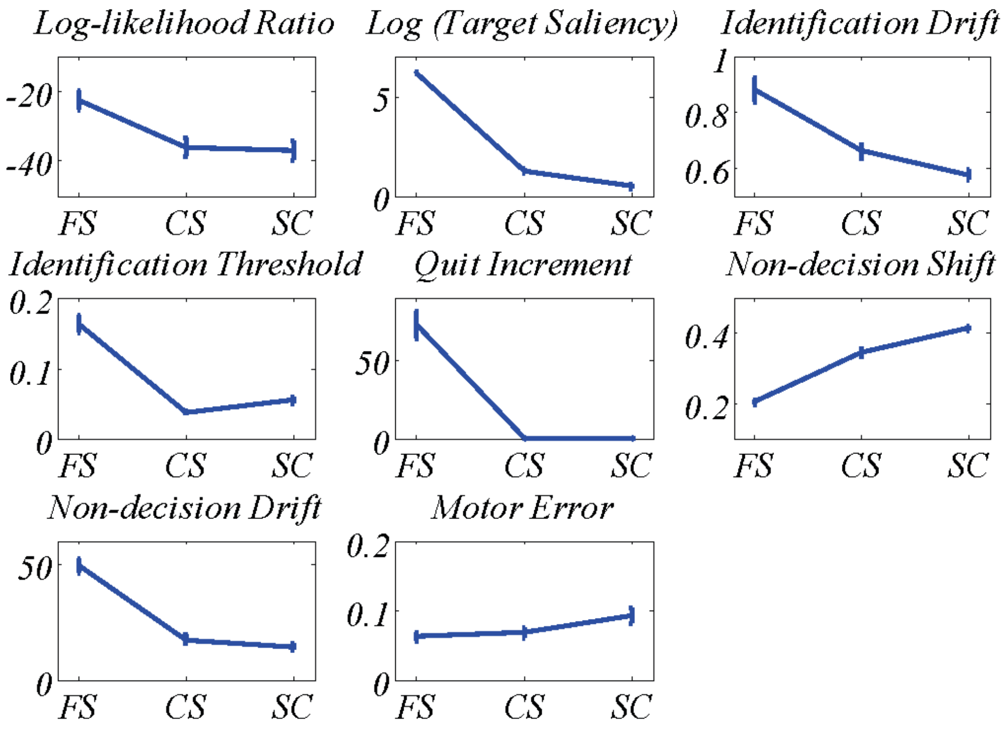


Fig. 5.7: Results of fitting CGS: The top-left graph shows the mean log-likelihood ratios (quality of fit) for the different tasks. The remaining graphs show the means of CGS's seven parameters for the three tasks (Feature Search = FS; Conjunction Search = CS; Spatial configuration search = SC). Note, for the purpose of a better illustration the target saliency parameter was scaled logarithmically. The results replicate Moran et al.'s (2013) findings. The error bars indicate the standard error.

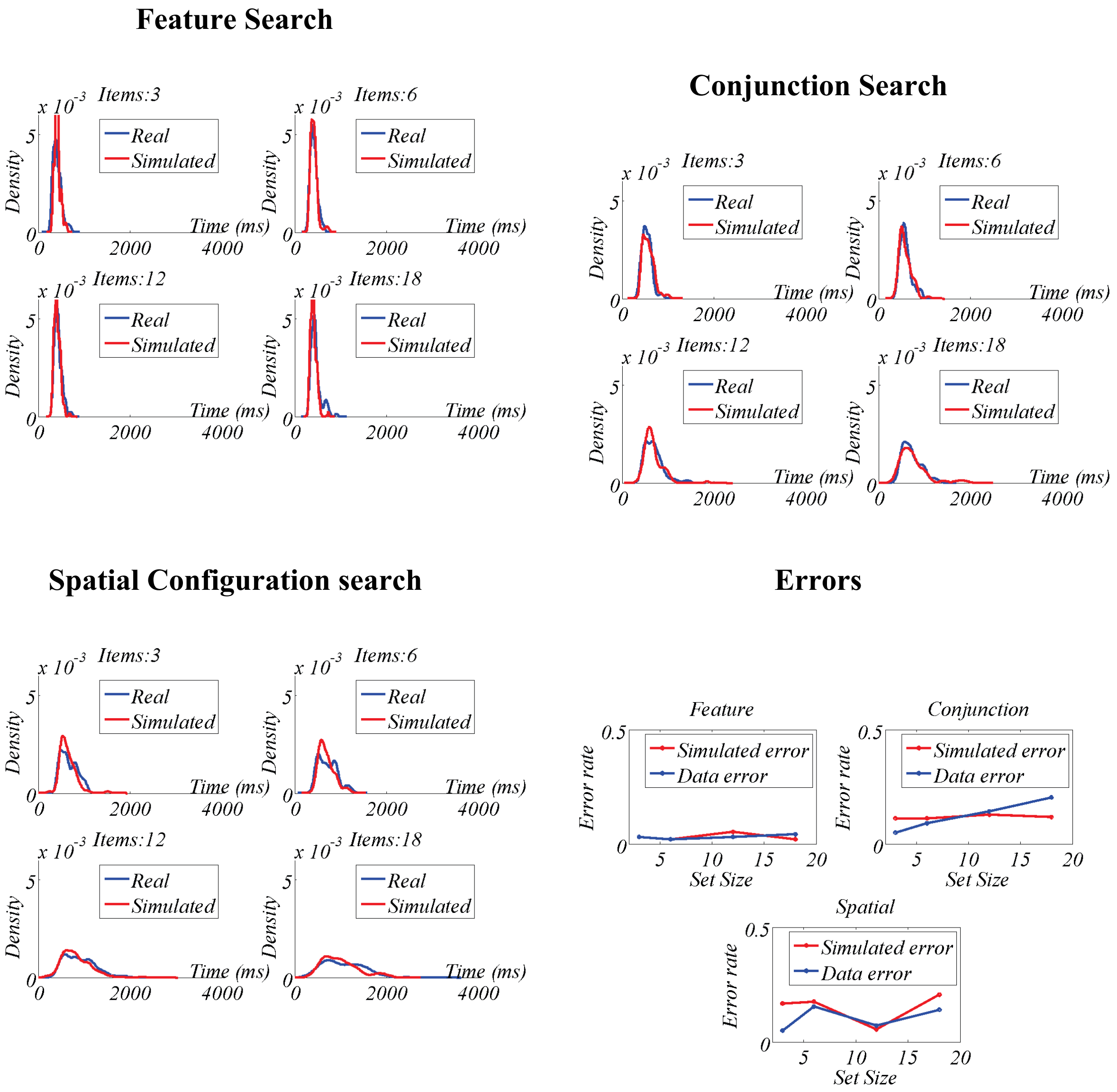


Fig. 5.8: CGS: KDE-based illustration of RT-distributions and response errors. Note that these graphs show the RT distributions for three participants.

5.6 Results and discussion: Competitive Guided Search (CGS)

We fitted CGS with seven free parameters (Target Saliency, Identification Drift, Identification Threshold, Quit weight increment, Non-decision time shift, Non-decision time drift and Motor error) to Lin et al.'s (2015) 58 datasets from three visual search experiments, feature search, conjunction search and spatial configuration search. Hence we obtained 58 parameter settings (see appendix II for values), 19 parameter settings (participants) for feature search, 19 settings (participants) for conjunction search and 20 settings (parameters) for spatial configuration. Our outlier removal procedure detected no outlier for feature search, two outliers for conjunction search and two outliers for spatial configuration. It is also worth noting that the parameter estimation revealed that there are good fits for conjunction search and feature search where the saliency values are implausibly high. This is not very surprising as fast target searches can be executed with arbitrarily high saliency value. To solve this problem, we first fitted spatial configuration and used these resulting parameter values as starting point for fitting the other searches. This way the best fits produced saliency values which were relatively small (see footnote 1 for an more explanations).

To assess the overall fit for each participant, we calculated the log likelihood ratios, i.e., log likelihood value from the model divided by the log likelihood value of the KDE's dataset. We compared ratios from the different tasks with the Wilcoxon rank-sum test and a significant difference was found between feature search and conjunction search ($z = 2.92; p = 0.004$); and feature search and spatial configuration search ($z = 3.11; p = 0.002$). And there was no significant relation between conjunction search and spatial configuration search ($z = 0.25$; $p = 0.81$). Fig. 7 shows that the quality of fit declined with task difficulty. To illustrate the quality of fit, Fig. 8 shows the outcome from three participants. Note that the choice of these participants was made randomly by MatLab to avoid an author bias. The likelihood ratio was –18.44 for feature search; -38.82 for conjunction search and -44.10 for spatial configuration search. The graphs illustrate that CGS's distributions nicely overlap with the RT-distributions from the respective tasks. Hence, we were able to qualitatively replicate Moran et al.'s (2013) results.

Interestingly, we were also able to replicate the qualitative relationship of parameter values with the search tasks (see Fig. 7; in appendix II and Moran et al., (2013); appendix C). The parameters accounting for encoding, post-decisional process, etc. showed longer delay (non-decision shift) and more variance (non-decision drift) with increasing task difficulty. Also the motor error increased with task difficulty. The identification drift showed slower accumulation rate with increasing task difficulty. The identification threshold decreased with task difficulty (albeit counterintuitively). The likelihood to stop scanning the search display (w_{quit}) increased less the more difficult the task was. Finally, the guidance (target saliency) was smaller the harder the task was. Interestingly and similar to Moran et al.'s (2013) findings, there was still residual guidance in the spatial configuration search task. In fact, there was no significant difference between our guidance parameters and Moran et al.'s (2013) parameters (see appendix II; Table II.1 for a comparison using Wilcoxon rank-sum test). This finding questions Moran et al.'s (2013) explanation for their result. They stipulated that guidance may be possible due to the fact that participants were highly practiced in Wolfe et al.'s (2010) dataset. However, in our experiment participants were not practiced and CGS still suggests that there is guidance involved. Note that due to the difference in numerical methods, as discussed earlier, this problem needs to be interpreted with caution and further studies are required, e.g. a direct comparison between fits with Wolfe et al.'s (2010) dataset and our dataset using our method.

However, how practice effects are explained with these models goes beyond the topic of the present research.

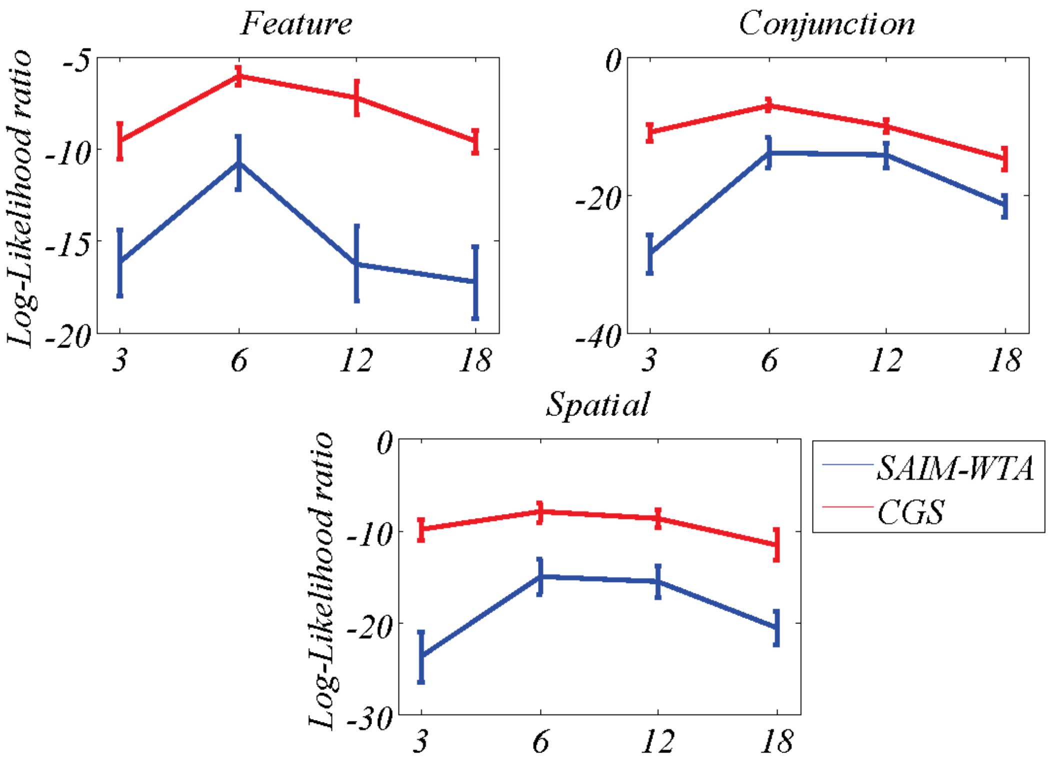


Fig. 5.9: Comparison of mean log likelihood ratios from SAIM-WTA and CGS for the three tasks (feature search, conjunction search and spatial configuration). The graphs indicate the contributions from the different display size to the overall log likelihood ratios. The error bars indicate the standard error without within-participant variance (Cousineau, 2005).The graphs demonstrate that CGS was better at explaining the data than SAIM-WTA. However, they also show that the performance of both models is best at display size 6 and worse at all other display sizes (see main text for more discussion).

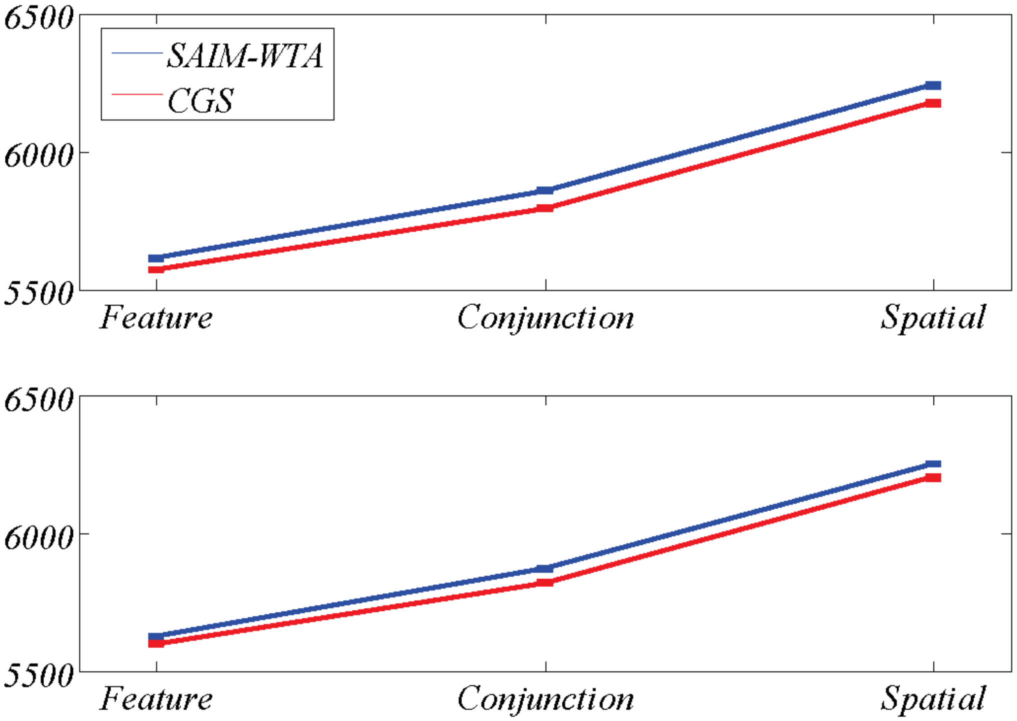


Fig. 5.10: Comparison of BIC scores and AIC scores from SAIM-WTA and CGS for the three tasks (feature search, conjunction search and spatial configuration search). The error bars indicate the standard error without within-participant variance (Cousineau, 2005). BIC and AIC penalize the quality with the model complexity as measured with the number of parameters. The graphs indicate that CGS performed better than SAIM-WTA despite SAIM-WTA being the simpler model.

5.7 Comparison of SAIM-WTA and CGS

Fig. 9 compares SAIM-WTA and CGS in terms of log-likelihood ratio. Overall, it demonstrates that CGS explained the visual search data better than SAIM-WTA. Fig. 11 also breaks down to the results in terms of how well the two models explained the data for the different display sizes. The figure gives us an insight into why SAIM-WTA is worse than CGS and highlights a problem with both models. The graphs show that irrespective of display size SAIM-WTA performed worse than CGS. Hence, SAIM-WTA failed to replicate the RT-distributions with the same precision as CGS. However, both models showed their best fit with display size 6 and then a decline in quality for larger and smaller display sizes. In other words, their ability to explain the influence of display size and their effect on RT-distributions does not generalize to all display sizes equally well. Hence, both models need to be improved in this respect.

z-value, p-value	**AIC**	**BIC**
Feature	**3.62, <0.001**	**2.29, 0.022**
Conjunction	**3.82, <0.001**	**3.58, <0.001**
Spatial	**3.55, <0.001**	**2.95, 0.003**

Table 5.1: Comparison of BIC scores and AIC scores from SAIM-WTA and CGS for the three tasks (feature search, conjunction search and spatial configuration search) using Wilcoxon sign-rank test. All comparisons were significant. For all comparisons CGS showed better results than SAIM-WTA.

The log-likelihood ratio does not take into account the complexity of models. This difference of complexity (as measured in terms of number of parameters) is particularly marked between CGS and SAIM-WTA. Normally, model complexity is included in a model comparison by using the Akaike information criterion (AIC) and/or the Bayesian information criterion (BIC). Both criteria penalize the quality of fit by the number of parameters needed to achieve this quality whereby BIC penalizes the quality more than AIC. Also note that the smaller the AIC/BIC scores is the better the model. Figure 10 shows the results for AIC and BIC. Since the results for the two models are quite close we also entered the AIC values into a Wilcoxon sign-rank test (see Table 1; for results). Again, the results show that apart from for conjunction search, CGS performs better than SAIM-WTA. However, the difference in the range of 2 and 3.82 for AIC and the range of 2.29 and 3.58 for BIC is not very large.

5.8 General Discussion

This research aimed to contribute to the long-standing dispute on parallel versus serial search. In order to do so, we numerically fitted two computational models to RT-distributions from three visual search experiments (feature search, conjunction search and spatial configuration search). The two computational models, Competitive Guided Search (CGS; Moran et al., 2013) and SAIM-WTA (e.g., Heinke & Backhaus, 2011) implement a serial and parallel search respectively. The comparison of the two models' success to explain the RT-distributions is expected to advance our understanding of visual search in humans. It also allows us to demonstrate how RT-distributions can contribute to this enterprise.

The results with SAIM-WTA showed for the first time that a biased-competition model is able to reproduce RT-distributions from visual experiments in particular the increased skewness linked with increased display size. However, a direct comparison between SAIM-WTA and CGS revealed that CGS fits better to RT-distributions than SAIM-WTA. This is the case even if the evaluation takes into account that CGS is a more complex model than SAIM-WTA. In other words, GCS's Wald-distribution modelling the item identification at search step constitutes a better description of search behaviour than the RT distribution generated by SAIM-WTA's competition process. Also the addition of identification times by the way of serial search scans (mathematically the multiple convolutions of Wald-distributions) represents fairly well the increase of mean RTs and their increased skewness. However, it is notable that both models don't generalize well across different display sizes as the quality of fit for both models decreases with increasing display size. It is also worth pointing out that we fitted CGS and SAIM-WTA to

Wolfe et al.'s (2010) data. As this data set does not contain as many participants as Yishin et al.'s (2015) dataset, the results were not as statistically conclusive as the ones presented here. Also some parameter settings were statistically different than ones presented here. These differences are certainly due to methodological differences, some of which we discussed in the section about Yishin et al.'s (2015) dataset. Since these differences go beyond the scope of this research we have not included reports of those results. However and importantly, the overall findings presented here, that CGS is slightly better than SAIM-WTA in explaining visual search experiments, was replicated. In other words, we can be quite confident that our results are valid for these three visual experiments irrespective of the methodological details, such as display geometry, practice effects, etc.

So what are the lessons from this model comparison for SAIM-WTA and CGS? It shows that SAIM-WTA produced excellent fits, but CGS produced slightly better fits.This is probably due to the fact that Moran et al. (2013) chose to model the identification of each item with a Wald-distribution which in turn was motivated by Palmer et al.'s (2010) finding that the Wald-distribution is the best distribution compared to other skewed distributions such as ex-Gaus or Weibull. Hence, it will be important for the progress of SAIM-WTA to find a way to produce more Wald-like distributions. A possible solution is to add an identification stage. Such an identification stage would lead to more skewed RT-distributions possibly enabling SAIM-WTA to produce better fits for larger display sizes. Finally, it is also worth noting that SAIM-WTA was not specifically designed to model visual search and instead aimed to capture a broad range of experimental evidence typically associated with visual selective attention. Hence, matching CGS's performance or even surpassing it was always a difficult goal to achieve.

In addition, SAIM-WTA (and CGS: see below) will have to improve on how the display size influences its response times. It is interesting to note that on the whole the quality of fit decreases increased with increasing display size. This may point to a possible cause of this problem. Of course, with increased display density the spatial proximity between items increases. Hence, it is conceivable that perceptual grouping (e.g. Wertheimer, 1923) plays an increasing role in higher display sizes. Hence, a sensible extension of SAIM-WTA may be to integrate a grouping mechanism into the competition process. For instance, at present the inhibitory connections are homogenous independent of an item's position. An extension of SAIM-WTA may modulate these weights depending on the distance between items. A corollary of this line of argument is that perceptual similarities between items may also play a role in visual search. Of course, this is not a new idea and there is already evidence for this, in particular from the seminal paper by Duncan and Humphreys (1998) (see also Müller-Plath & Elsner, 2007 for a systematic variation of spatial proximity and item similarity). In any case, this extension of SAIM-WTA will have to be tested with a series of studies manipulating grouping in visual search possibly along the lines of Müller-Plath and Elsner's (2007) work.

Obviously, an integration of a grouping mechanism into CGS's saliency map along similar lines is also possible, and this modification may lead to the desired effect of improving the fit with higher display sizes. However, such a modification would not improve CGS's serial process as such. Instead, a simple modification of CGS consistent with its serial tenet could be to let the parameters of the identification stage depend on at which point in the serial search scan the items are identified. This additional mechanism could slow down identification or make RT-distributions more skewed the later an item is selected, possibly improving CGS's performance for larger display sizes. This modification can be seen as some sort of inhibition effect on the identification stage (object-based inhibition, e.g., Egly, Driver, & Rafal, 1994; Heinke & Humphreys, 2003; Study 5). However, it is highly questionable whether this new mechanism can successfully explain the potential influence of perceptual grouping as discussed earlier since it does not consider spatial proximity of items or item similarity. On a more general note, it is worth pointing out that it is difficult to imagine how CGS's slow serial identification process can model perceptual grouping in a plausible way. (E.g., serially scanning through some items and making them as a group if they are the same would be far too slow). Hence, these difficulties of CGS with perceptual grouping suggest that the series of visual experiments manipulating grouping, as suggested earlier, can produce data which allow for a stronger comparison between serial and parallel models where even the parallel approach may win the competition.

In conclusion, we have demonstrated that it is possible to constrain computational models of visual search with RT-distributions. We were also able to replicate findings from a serial model of visual search (CGS: Moran et al., 2013). In addition we successfully fitted a parallel (biased competition) model of visual search (SAIM-WTA, e.g., Heinke & Backhaus, 2011) to RT-distributions for the first time. When the two models were compared the serial model was able to explain better RT-distributions from three visual search tasks. However, both models exhibited deficits in how they dealt with different display size. From the discussion of possible mechanisms to iron out this problem we inferred that a series of visual experiments manipulating perceptual grouping should lead to a stronger test for the models.

Chapter 6:
Asymmetrical Dynamic Neural Network Model (ADyNeN)

Abstract

In previous chapter we attempted differentiating serials from parallel process by numerically fitting the corresponding model representations to RT-distributions and their accuracys to visual search data representing very easy to very hard experimental tasks. The serial search model was the Competitive Guided Search (CGS) model by Moran et al. (2013) and winner takes all (WTA) mechanism from Selective Attention for Identification (Heinke and Humphreys (2003)) and we found that CGS model outperformed SAIM-WTA model. The given study was limited to the trials when the target was present because of SAIM-WTA was not designed to explain target 'absent' trials. In this study we present a novel asymmetrical dynamic neural network (ADyNeN) proposing asymmetrical inhibition as means to flip WTA between target 'present' or 'absent' trials. A comparison between CGS and the new model indicates that a new ADyNeN model performs slightly better than CGS model. Moreover, a proposed model naturally aligns with biased competition theory and is predicted to explain a wider range of phenomena observed in visual search experiments. The future work should directly challenge the model under different constraints different.

6.1 Introduction

SAIM-WTA model was fitted only to ‘present trials, the ‘absent’ trials were omitted from the modelling fitting. This model was able to perform well on either `present' or `absent' trials but not both. The issue was that the feature search task would have to have a high difference between target saliency and distractor saliency for `present' trials in order to produce fast response. In contrast, 'absent' trials would have to be as fast as 'present' trials. Setting low saliency for distractors means a very slow `absent' response instead. Setting high saliency means that there is a lot of competition between target and distractors, therefore a slow response in `present' trials. In order for SAIM-WTA to fit both conditions, at least one parameter would have to be changed depending on whether the target is present or absent. This is not a desirable feature; therefore we opted to structural change of the model.

In this chapter we introduce a novel Asymmetrical Dynamic Neural Network (ADyNeN) model inspired by winner-takes-all mechanism described in SAIM-WTA chapter. This model suggest that to determine the absence of the target in a competitive parallel system within homogenous distractors there has to be a substantial top-down effort to break down similarity between distractors and this can be modelled by modifying inhibitory weights between items. Firstly, we will present the description of the model and perform equivalent studies to the SAIM-WTA chapter which will also include absent trials.

6.2 Asymmetrically Dynamic Neural Network (ADyNeN) model

In our work we introduce a modified SAIM-WTA model that is capable to explain present as well as absent visual search RT-distributions without changing parameters. Typically, an input is treated as a saliency value which represents the bottom-up strength of the visual stimuli and similarity to the identity of the target somewhat contributes to it (Mavritsaki, Heinke, Humphreys & Deco, 2006; Heinke & Humphreys, 2003; Heinke & Backhaus, 2011). We propose to use bottom-up saliency meaning (Itti, & Koch, 2000) that two items belonging to the same category such as colour (for the sake of simplicity we assume there are no saliency differences or these differences are negligible between certain colours) are similarly salient whether they are targets or distractors. We motivate our choice by the fact that neuronal populations corresponding to the identity of the target fire more frequently than the populations corresponding to the distractors because top-down modulation can suppress as well as enhance the firing rate of the neuronal populations depending on whether these items are relevant to the task or not (Treue & Trujillo, 1999).

SAIM-WTA model was modified by replacing a global inhibitory node with mutual connection for mutual inhibition. Mathematically, this is equivalent to original model since all inhibitions add-up to the same value. Next, the model had its inhibitory connections relaxed so the competition between nodes resemble how similar distractors are to the identity of the target. As a result, we have introduced separate inhibitory weights for inhibition from the target and from distractors. The inhibitory connections between distractors depended on the similarity of the distractors to the target. In addition, we set the inputs to 1 for all items within display since the input is equivalently strong for all items. Fig ? illustrates the proposed architecture of the model. Each node retains the direct identity to the item being displayed within the particular location, none if the location is empty. The lines show mutual inhibition with end dots indicating the strength of the inhibition. Larger end dots represent greater inhibition between particular nodes of the model.

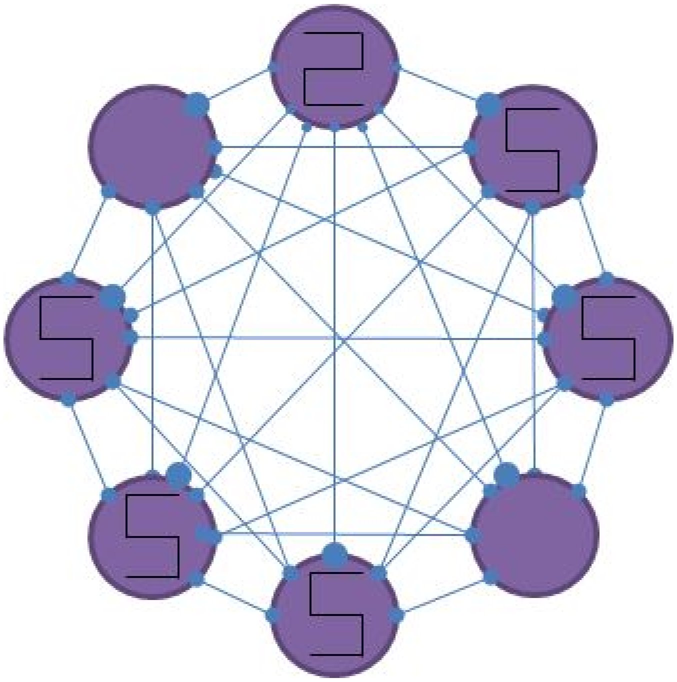


Fig 6.1: This graph illustrates ADyNeN model's architecture. Each node demonstrates a corresponding locations of possible stimuli within. The given example shows "2 vs 5" task where "2" is a target and "5" is distractor. All nodes compete by mutual inhibition. Inhibitory weights between distractors is symmetric but connections between targets and distractors is asymmetric. The size of the inhibitory dot indicates the inhibitory strength towards the particular node.

Our novel dynamic asymmetrical neural network model's (ADyNeN) mathematical equation is expressed as follows:

$$dx_i =- \tfrac{dt}{\tau} x_i + \tfrac{dt}{\tau}(- \sum_{j=i;i \neq j,t}^{n_d} w_{ij} f(x_j) - w_{ti} f(x_t) + I) + \sqrt{\tfrac{dt}{\tau}} \xi_i;$$

Where I is an input which is either 1 or 0 depending on whether there is some item or not. The difference from previous equation is within weights w as now it depends on which node is being inhibited. The number of distractors is indicated by n_d and t corresponds to the target node. This model has two conditions, target is present and target is absent. A simpler form represents a situation when the target is absent where $n_d = n$. When this is the case, it is a SAIM-WTA model as indicated earlier and the only difference is a constant input of 1 which does not change the behaviour of the model in 'absent' condition. The observed delay in the predicted RT outcomes is due to the competitive interference between the distractors and other combinations of the parameters. However, trials with the target present will have a SAIM-WTA submodel among n_d nodes and an additional inhibitor from the target. This weight will generally be higher than the inhibitory strengths coming from the common distractors. An important node is a target itself as it receives a common inhibitory input from all distractors. Note that the inhibitory weights from distractors to target and distractors to distractors are not the same. These are treated as separate parameters and allowed to vary freely for different datasets. The establishment of these asymmetrical weights enables a target to have a biased competitive advantage for the node representing the target. $f(x)$ is a sigmoid function (see SAIM-WTA chapter) that has slope and shift parameters which modulate the sensitivity towards the incoming input and the suppression of baseline neuronal activity. This function models non-linear relationship between output activation and firing rate. The populations' sensitivity to the identity of the input can be modelled by varying the parameters of this function. We argue, that this sensitivity is a more sensible parameter to explain saliency than the biased inputs as commonly used in practice. Nodes representing items from different categories such as shape and colour would have different sensitivity to the stimulus. This predicts attentional grab by more salient distractor present within display.

ADyNeN model is a self-terminating model where the decision of target presence is made when some item passes the threshold. If the target passes the decision threshold, then the models determines that a target is present. When non-target passes the threshold, it makes a decision that target is absent. Items perceived as potential targets accumulate information and evidence. The noise within this system indicates the uncertainty around the identity of the item corresponding to the node. Such a decision making indicates a pure parallel process where the winning of the distractor is treated as a proof that a target does not exist within display. It also assumes that no serial process exists. Additionally, the model introduces a motor error to explain the errors within absent trials. Finally, in 'absent' trials there is no target node which could trigger an incorrect response, therefore we assume that people are perfect at identifying the absence but make a mistake in their motor response. This motor error may also work in favour of the correct choice during the present trials, but these cannot be determined, therefore we treat them like correct responses. If the decision is incorrect, it is measured as an error.

6.3 Materials and Methods

In the following section, we will discuss the methods we used to find the best-fitting parameters. KDE-approach was identical to the study used for SAIM-WTA comparison with CGS model. We adapted the same approximate likelihood method for dataset containing present as well as absent trials. The main differences emerge in parameter estimation compared to SAIM-WTA study, hence we included a brief explanation for the algorithms and why these were chosen.

6.3.1 Parameter Estimation

The methods used to find the optimal parameters differed from the methods in previous SAIM-WTA chapter. We still used DE-MCMC algorithm (Turner, Sederberg, Brown & Steyvers, 2013) but only for establishing initial parameter population to be used for our suggested Sequential Importance Sampling by Filtered Clustering (described in Methodology chapter) algorithm. Early in parameter estimation using DE-MCMC algorithm the acceptance rate is satisfactory and across all participants will converge to a number of highly dense areas. Since a good non-informative prior was not available, SISFC algorithm had a higher risk in settling to a wider good area while DE-MCMC explores space better as long as the best chain value is not stuck in unlikely local maximum likelihood. Overall, each participant had 20 initial parameter settings drawn from the prior distribution (Table III.1) and had a greedy implementation of DE-MCMC run for 50 generations.

We used a modified SISFC algorithm for the main parameter estimation. This algorithms is much more efficient than DE-MCMC in searching local area when uncertainty is excessively large and the used model is computationally expensive. However, to further improve efficiency and the convergence speed of the searching we introduced a stricter dynamic tolerance threshold. We used individual mean log-likelihoods of each cluster as a performance measure rather than proportion of values being above the threshold. Such choice will filter out more clusters from producing new proposals which will lead to faster convergence overall. We also used the median expectation of overall likelihood value distribution as filter instead of the proposed mode though median should produce similar results as compared to mode-based approach when unbounded maximum likelihood is present. Our weights to choose clusters were computed the same way as in SISFC algorithm, except the earlier defined mean log-likelihoods were used to determine whether an entire cluster is above threshold or not and at what percentile it falls. This new percentile was used as a fitness value to compute a weights using in methodology defined equation.

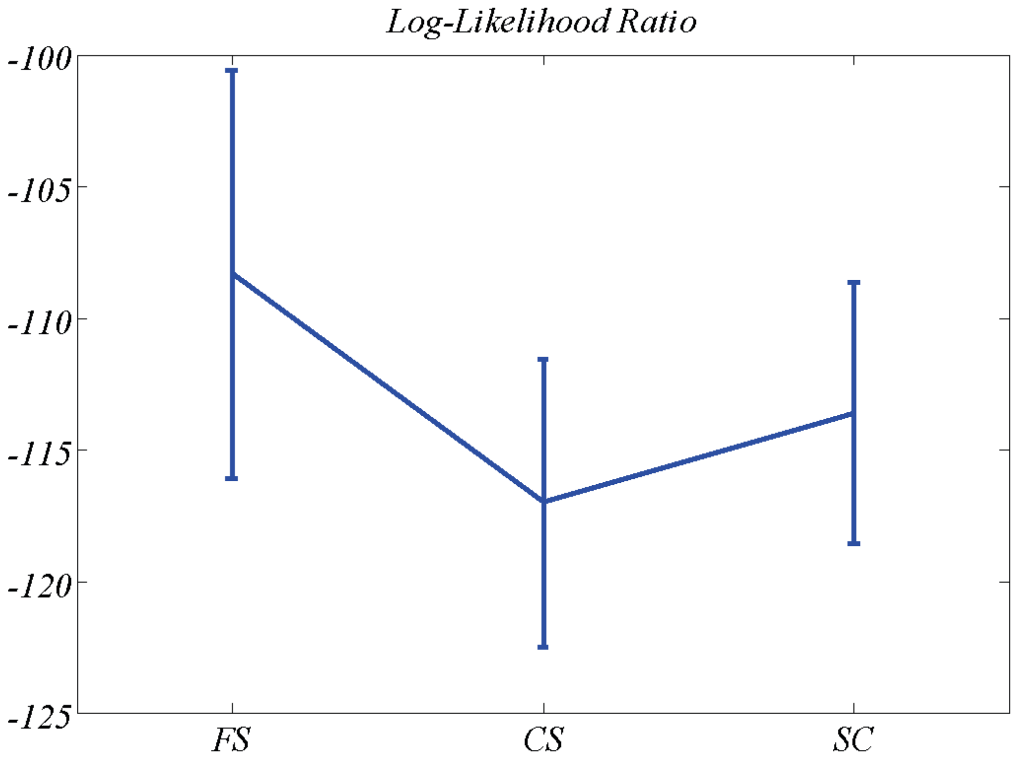


Fig. 6.2: Results of fitting ADyNeN. This graph shows the mean log-likelihood ratios (quality of fit) for the different tasks. The error bars indicate the standard error.

We initialised all participants using the best solutions generated by DE-MCMC, 58 initial common parameter settings for all datasets. We performed fitting for all participants and tasks concurrently, each participant had a new parameter sample generated at at time. If this new proposal had the best fitting likelihood value for other participants, we used it to update the knowledge about that participant by adding it to the posterior approximation. We fitted each participant for at least 1000 iterations though some tasks and participants required additional fitting processes. CGS model had the same issue with saliency parameter as in SAIM-WTA study and conjunction results failed to produce consistently good solutions when all participants and tasks were fitted concurrently. Firstly, we carried out 3000 iterations for all participants and tasks together. To deal with outlined issue we used the best solutions for spatial configuration dataset and performed additional parameter estimation for conjunction dataset together spatial configuration task for another 2000 iterations. ADyNeN model did not have issues with the same tasks, 1000 iterations were sufficient to find very good performing solutions for both tasks. However, feature task struggled to find the best solutions when fitted concurrently with other tasks. To help with this task we performed another 6000 iterations for all participants from feature task.

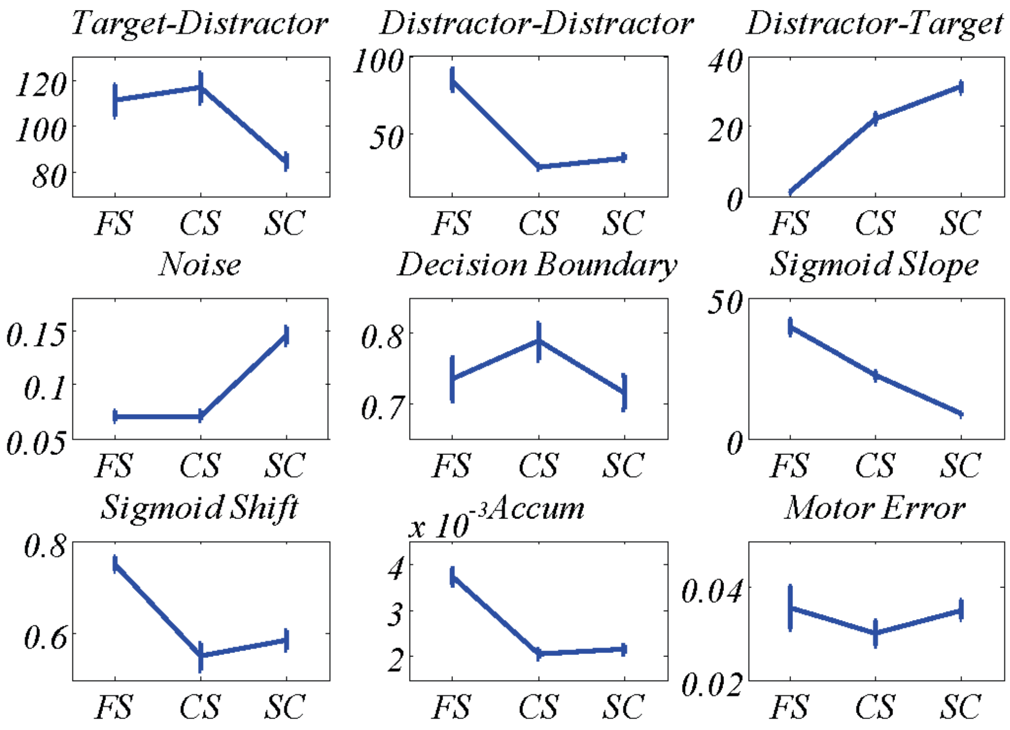


Fig. 6.3: Results of fitting ADyNeN. The graphs show the mean parameters. The error bars indicate the standard error. Accum - Accumulation.

6.3.2 Likelihood-function

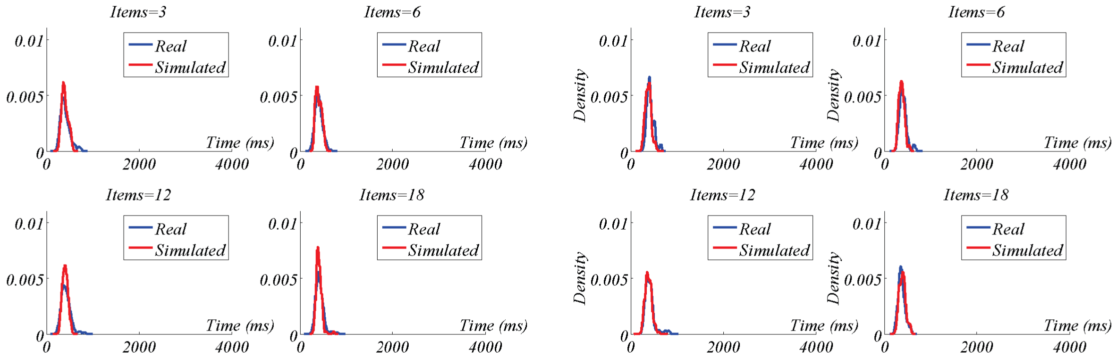
The fitness of the model over data was evaluated using likelihoods which were determined using approximate pdf of the model. The likelihood equation used was for mixed data (Turner and Sederberg, 2014) as in SAIM-WTA study with one key difference. The equation also incorporated absent trials which expanded overall multiplied components in likelihood function by two. The probability of target being present or absent were equal thus we introduced a scalar of 0.5 for both condition to form a proper probability function.

6.4 Results and discussion: ADyNeN

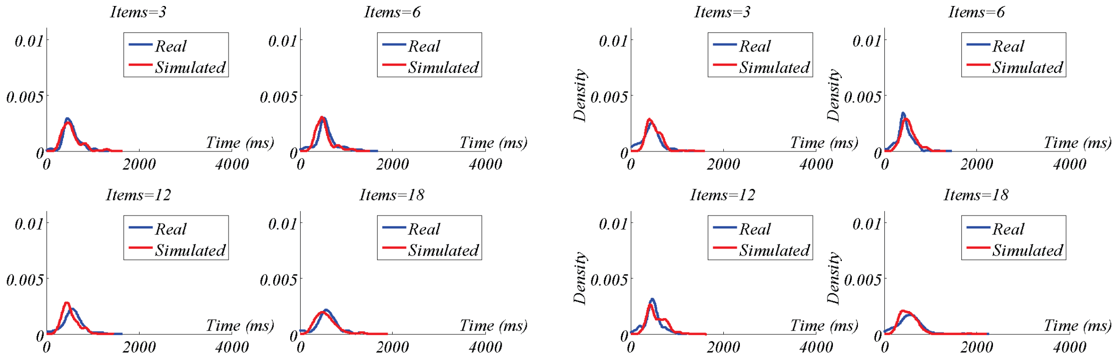
6.4.1 Results

We fitted ADyNeN with nine free parameters (target-to-distractor, distractor-to-distractor, distractor-to-target, noise, decision boundary, sigmoid slope, sigmoid shift, accumulation, motor error) to Lin et al.'s (2015) 58 datasets from three visual search experiments, feature search, conjunction search and spatial configuration search. Hence we obtain 58 parameter settings (see appendix III for values), 19 parameter settings (participants) for feature search, 19 settings (participants) for conjunction search and 20 settings (parameters) for spatial configuration. Our outlier detection procedure led to the removal of 3 participants from feature search, 2 settings from conjunction search and 3 participants from spatial configuration search.

Feature Search



Conjunction Search



Spatial Configuration Search

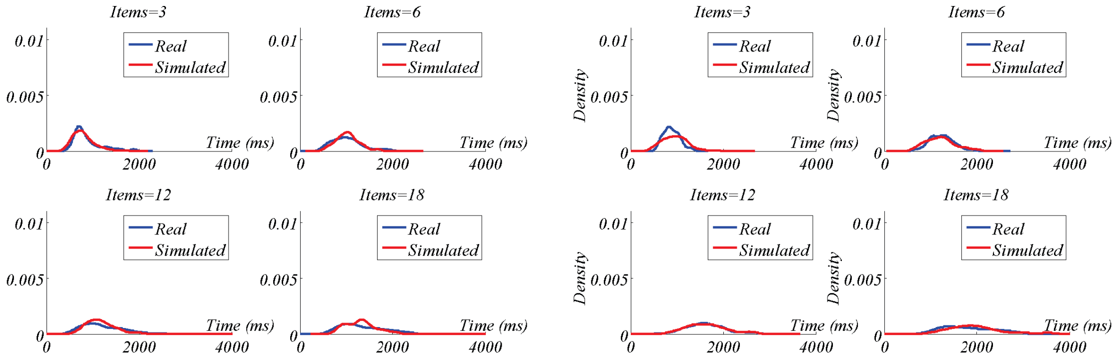


Fig 6.4: ADyNeN: KDE-based illustration of RT-distributions for one participant from each task. Figures on the right are from tasks with target present and the left figures are from tasks with target absent from display.

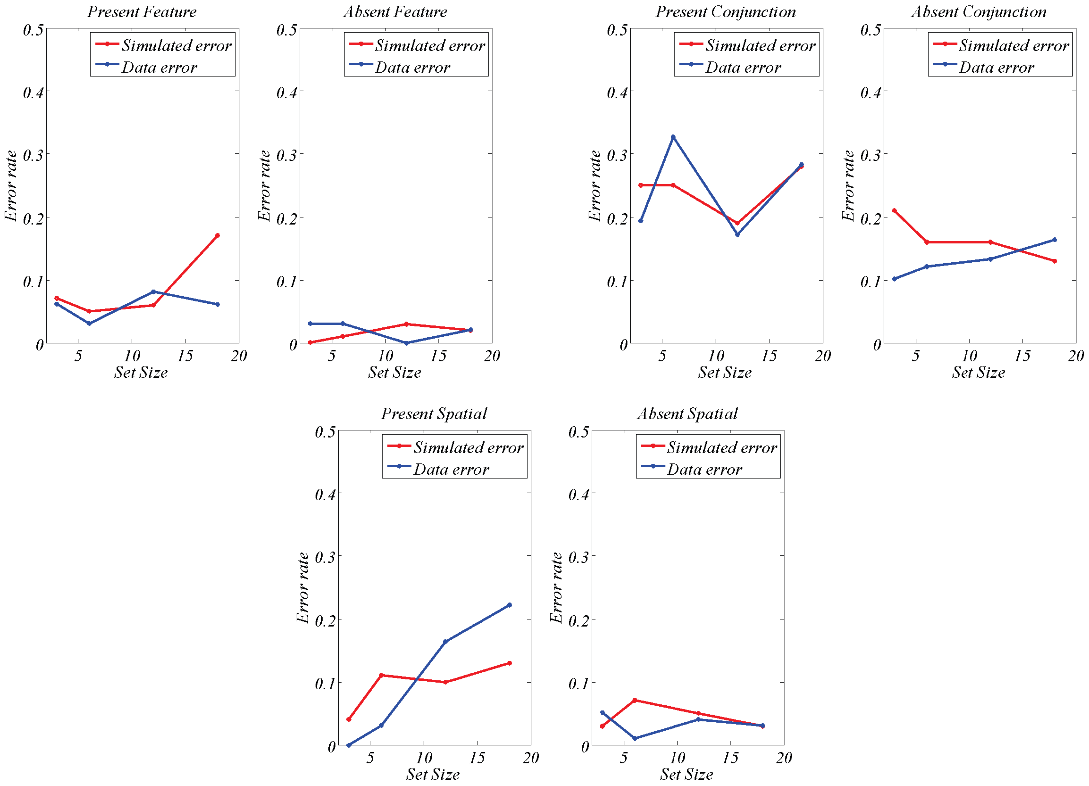


Fig 6.5: ADyNeN: An illustration of error rates produced by participants and the simulated model for each task. All error rates are from the same participants as shown by RT-distributions in Fig ?

To assess the overall fit for each participant we calculated the log likelihood ratios, log likelihood value from the model divided by the log likelihood value of the KDE's dataset (see Fig. ? for the results). We compared ratios from the different tasks with the Wilcoxon rank-sum test and found a significant decline of ratios between feature search and conjunction search ($z = 2.04$; $p = 0.042$); but not for feature and spatial configuration searches ($z = 1.49$;

$p = 0.135$). There also was no significant difference between conjunction search and spatial configuration search ($z = -0.62; p = 0.535$). Fig. ? shows the outcome from three participants. Note that the choice of these participants was made randomly by MatLab to avoid an author bias. The likelihood ratio was -95,16 for feature search; -102.11 for conjunction search and -99.36 for spatial configuration search. The graphs indicate that ADyNeN was able to explain distributional changes with the increased display size. The changes broadly matched the distributional changes in the data. There was little indication of potential shortcomings without broader comparison with other models which indicates a great success of the model to account for the effects of display size on the resulting RT-distributions. The model had constant parameter settings for all datasets for the given participant and the only modifier was the number of competing accumulators which matched the display size. The broad results suggest that model parameterisation is sufficient to explain distributional changes with additional distractors.

Figure 6.3 shows how the nine free parameters (target-to-distractor, distractor-to-distractor, distractor-to-target, noise, decision boundary, sigmoid slope, sigmoid shift, accumulation, motor error) changed across the three tasks. The parameters were entered into a Wilcoxon rank-sum test (Appendix ?; Table ?). The parameters varied differently from task to task showing complex relationships. A crucial asymmetrical weights novelty of the model over its predecessor SAIM-WTA show varying interactions between tasks. The weights of target-to-distractor and distractor-to-target are the only two parameters that make the difference between target present or absent conditions. Fig ? show that these two parameters are not linearly related with increasing task difficulty. Distractor-to-target show a relatively linear increase in the parameter values with increasing difficulty of the task. However, target-to-distractor illustrate a bound of the inhibition coming from a item of interest within visual display. Feature and conjunction tasks have this inhibition parameter set to ~110 while spatial configuration best fitting settings fall to ~80. This suggests that inhibition of distractors by the presence of the target is limited and does not differentiate between feature or conjunction tasks. However, distractors have a much lesser influence to the target during feature search. In fact, this value ~1 and often less. This means that target in feature task accumulates almost unhindered by the presence of the distractors and the distributional variance is more due to non-competition processes in the brain. This is not the case for conjunction and spatial tasks as both show an increase of the parameter value with values for spatial configuration being the highest. The last weight interaction is between distractors. This weight show no difference between conjunction and spatial configuration tasks but a big difference for feature task. It appears the system prefers an active suppression of various irrelevant competing items within display to make a rapid decision about the presence or absence of the target. Decision boundary showed little difference between tasks with conjunction task indicating slightly higher decision threshold. This is not a surprising result since the task did not urge participants in any way, thus speed accuracy tradeoff is more of the individual preference rather than emerging property due to difficulty of the tasks. The noise show no difference between feature with conjunction tasks but a comparingly large difference compared to spatial configuration task. The noise modulates the confidence in the identity of the item and the result for spatial configuration show much greater fluctuation in the belief of the identity of the item over time hurting overall decision making. However, it is interesting that there was no difference between feature and conjunction tasks. Motor error had relatively low values at around 3% across the tasks and the tasks had no influence on the tendency to make a motor error. The sigmoid modulates firing rate of the neuronal population and it had two unique parameters of slope and shift. The slope indicates the sensitivity over incoming evidence while shift is an overall suppression of the population or the suppression of allowed baseline activity. The slope showed a linear decrease in sensitivity to the incoming information with the task difficulty while activity suppression was greater for feature task than other two tasks which had little difference between them. These parameters show that the system favoured quick responsiveness to the available information for feature task and reduced baseline activity to prevent unnecessary interference. This was not the case for other two tasks for which baseline activity was very similar but a greater sensitivity was for conjunction search task rather than structural configuration task. Accumulation parameter showed a small difference between conjunction and spatial configuration tasks where accumulation had a slightly higher value than conjunction task. On the other hand, feature task had much higher accumulation values than these two tasks. This parameter indicates a general information gathering which is much greater for feature search as expected. It is interesting, that this is not the case for the other two tasks indicating an equivalent information gathering. Our closer inspection showed that sigmoid shift and accumulation had significant partial correlations for all 3 tasks indicating a strong relationship between parameters (Table ?). Higher sigmoid shift delays neuronal firing while higher accumulation speeds up information gathering thus observed differences between tasks may be completely explained by the relationship of these two parameters. There were no other consistently significant relationships between parameters.

6.4.2 Discussion

On the whole, ADyNeN exhibited great fits for the given three search tasks. It was able to explain the distributional changes with increase in display size and different visual search tasks. There is little evidence for the change in quality of the fits in relation task difficulty suggesting that a parallel competitive model is sufficient to explain datasets that are traditionally deemed to illustrate serial processes. Though, a comparison to the best solutions of the serial model may aid in identifying potential weaknesses which are not observable by the resulting fits alone.

The free nine parameters (target-to-distractor, distractor-to-distractor, distractor-to-target, noise, decision boundary, sigmoid slope, sigmoid shift, accumulation, motor error) varied differently for the three tasks. Unlike general modelling practice, input was the same for all items, instead it was replaced with dynamical weight system as a driving force of visual search. The inhibition is affected by the similarity of the target and distractors which can be explained by biased competition theory as connections between neuronal populations are biased in favour of the target. Distractor-to-target weight can be perceived as a level of interference in finding the target and pop-out feature search has little to no interference by the distractors as is in most of the experiments. Another main task predictor was sigmoid slope which decreased with the task difficulty. This parameter could be thought as a replacement of a general saliency term as it determines the sensitivity of the neuronal population. We think this could be a more appealing interpretation of the saliency since colours used in feature tasks are salient regardless of whether they are targets or distractors. The noise parameter of the ADyNeN model suggests that identification of structural items has greater degree of uncertainty and people are more likely to make a wrong identification. Though this a common observation in speed-accuracy-tradeoff of visual experiments. This also explains why decision boundary is not affected by SAT between feature and spatial configuration tasks. On the other hand, the noise is relatively the same for feature and conjunction tasks while decision boundary is slightly elevated for conjunction task. These parameters have related behaviour and it is common practice to constrain noise to a constant value but Donkin, Brown and Heathcote (2009) discourage this idea as noise have some properties that help explain RT distributions. Indeed, decision boundary is an urgency to make a decision parameter while noise indicates the fluctuations in the evidence for the identity of the item. The last parameter is motor error and it shows no difference between different tasks as generally expected.

There were limited persistent interactions between parameters though accumulation with shift to response time remained significant across all tasks. It is not surprising since these parameters have very similar roles in the equation as both affect how fast the information is gathered. Accumulation directly modifies this gathering while shift modifies the baseline activity of the neuronal population which results in similar overall influence on the behaviour. However, this relationship does not have a known realisation via equations that would allow replacing one variable with another. The observed linear relationship does not have solutions on a line, thus parameter confidence intervals for each participant would have to be identified in order to determine the possibility of removal of one parameter. A future study using Bayesian inference may help to confirm an actual linear relationship which would enable the model simplification.

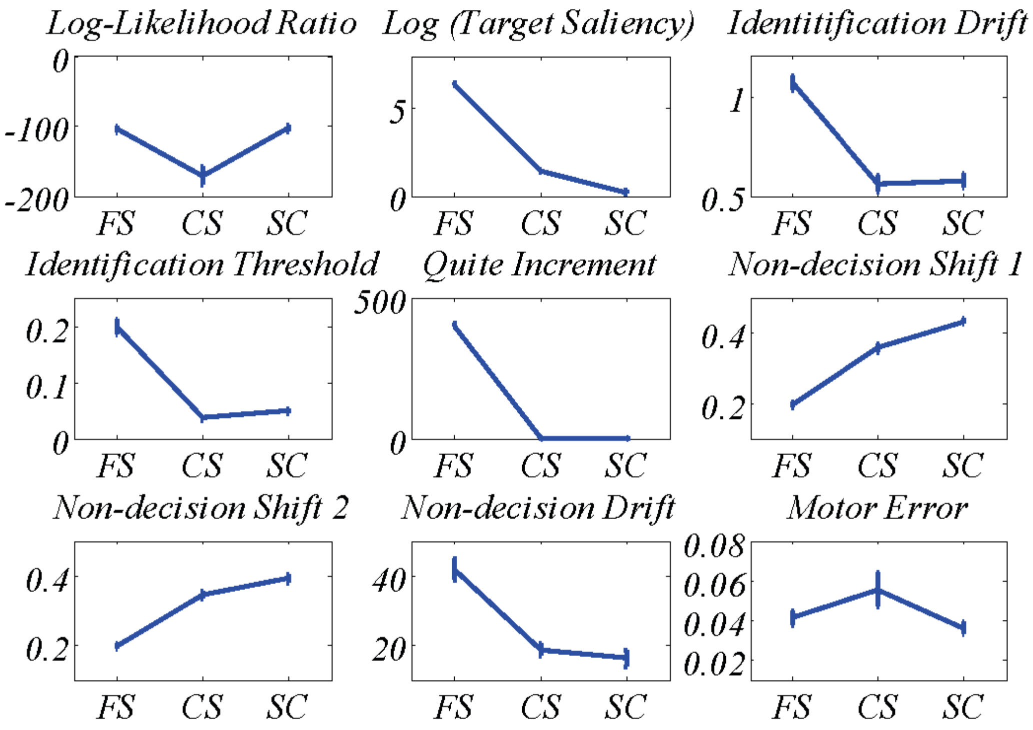


Fig. 6.6: Results of fitting CGS: The top-left graph shows the mean log-likelihood ratios (quality of fit) for the different tasks. The remaining graphs show the means of CGS's eight parameters for the three tasks (Feature Search = FS; Conjunction Search = CS; Spatial configuration search = SC). Note, for the purpose of a better illustration the target saliency parameter was scaled logarithmically. The results replicate Moran et al.'s (2013) findings. The error bars indicate the standard error.

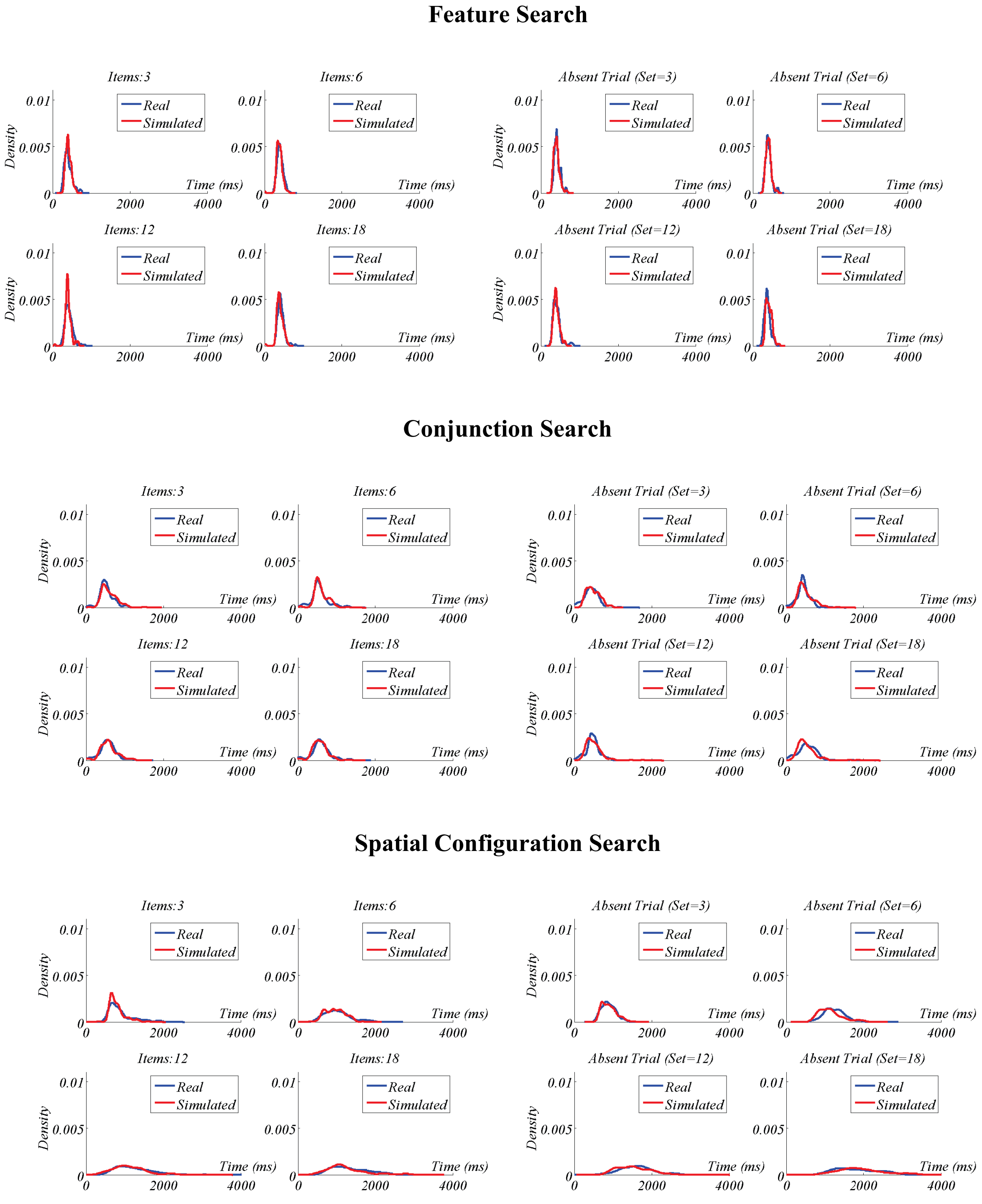


Fig 6.7: CGS: KDE-based illustration of RT-distributions for one participant from each task. Figures on the right are from tasks with target present and the left figures are from tasks with target absent from display.

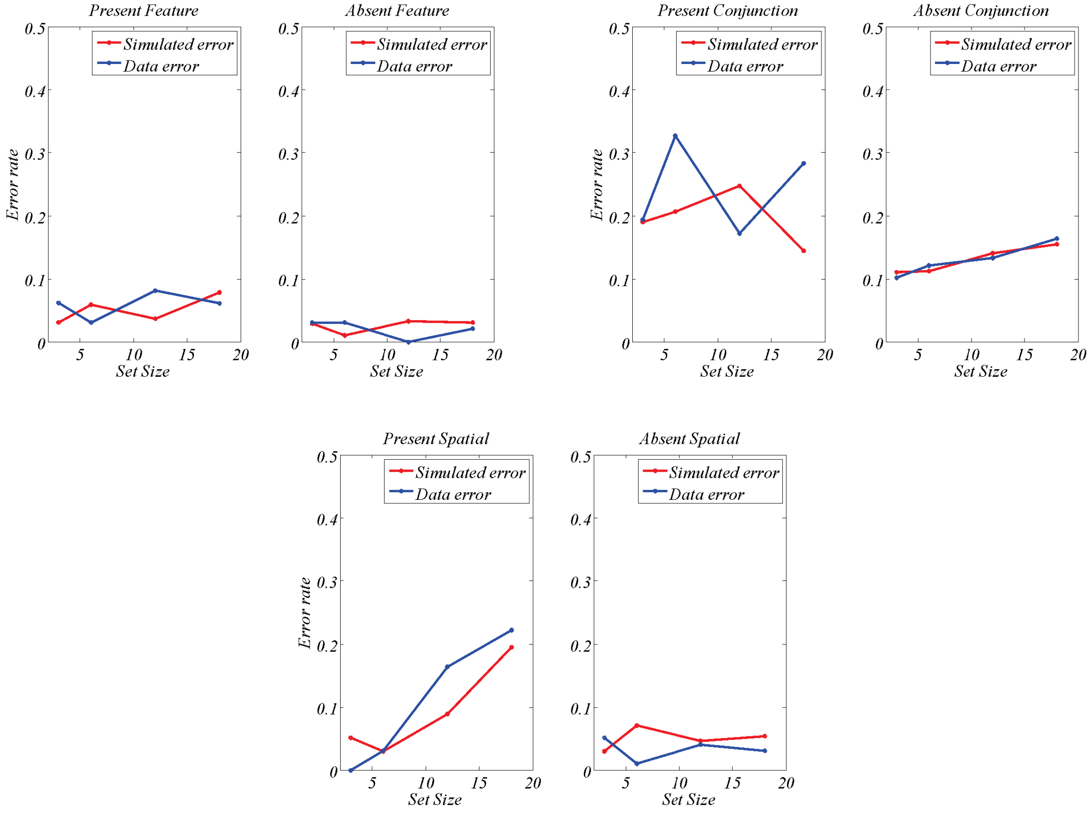


Fig 6.8: CGS: An illustration of error rates produced by participants and the simulated model for each task. All error rates are from the same participants as shown by RT-distribution in Fig?

6.5 Results and discussion: Competitive Guided Search (CGS)

We fitted CGS with eight free parameters (Target Saliency, Identification Drift, Identification Threshold, Quit weight increment, separate values for Non-decision time shift for present and absent trials, Non-decision time drift and Motor error) to Lin et al.'s (2015) 58 datasets from three visual search experiments, feature search, conjunction search and spatial configuration search. Hence we obtained 58 parameter settings (see appendix II for values), 19 parameter settings (participants) for feature search, 19 settings (participants) for conjunction search and 20 settings (parameters) for spatial configuration. Our outlier removal procedure detected three outliers for feature search, four outliers for conjunction search and 5 outliers for spatial configuration. It is also worth noting that the parameter estimation revealed that there are good fits for conjunction search and feature search where the saliency values are implausibly high. The number of outliers was higher than the study containing only trials from present condition. Some participants showed general shift in parameters and firstly using fitted spatial configuration parameter settings as starting point for fitting the other searches has not aided since some participants in spatial search showed outlier tendences.

To assess the overall fit for each participant, we calculated the log likelihood ratios, i.e., log likelihood value from the model divided by the log likelihood value of the KDE's dataset. Fig. 7 shows that the quality of fit declined for conjunction task but not the other two tasks. We compared ratios from the different tasks with the Wilcoxon rank-sum test and a significant difference was found between feature search and conjunction search ($z = 4.17; p =< 0.001$); but not feature search and spatial configuration search ($z = 0.15; p = 0.877$). And there was a significant relationship between conjunction search and spatial configuration search tasks ($z =- 4.06; p =< 0.001$). To illustrate the quality of fit, Fig. 8 shows the outcome from three participants. Note that the choice of these participants was made randomly by MatLab to avoid an author bias. The likelihood ratio was –18.44 for feature search; -38.82 for conjunction search and -44.10 for spatial configuration search. The graphs illustrate that CGS's distributions nicely overlap with the RT-distributions from the respective tasks. Hence, we were able to qualitatively replicate Moran et al.'s (2013) results.

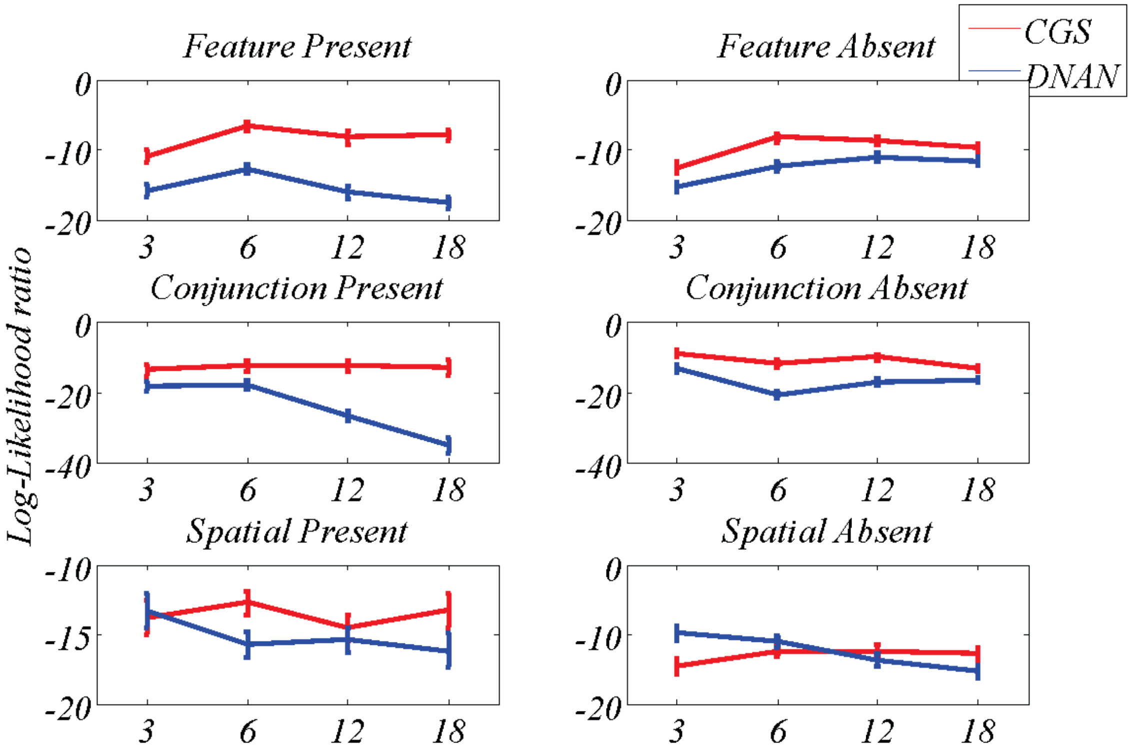


Fig. 6.9: Comparison of mean log likelihood ratios from ADyNeN and CGS for the three tasks (feature search, conjunction search and spatial configuration). The graphs indicate the contributions from the different display size to the overall log likelihood ratios. The error bars indicate the standard error without within-participant variance (Cousineau, 2005).The graphs demonstrate that ADyNeN was better at explaining the data than CGS.

Overall, the parameter relationships matched relationships observed in fitting study for trials from present condition. The non-decision shift (for both conditions) and non-decision drift parameters suggest that non-decision processes such as encoding or motor execution predict that harder tasks have more variation and require more time to perform these processes. The same parameter predictions also were for identification drift showing slower accumulation rates with more difficult tasks while the identification threshold showed an increased preference for lower settings with task difficulty. Motor error had little change with conjunction being slightly elevated in motor error than other tasks. The parameters also predict that the difficulty of the task reduces the probability to quit searching after identifying the item. Finally, the the saliency of the target has dropped with increasing task difficulty as expected though guidance for spatial configuration task remained above 1 reproducing findings Moran et al.'s (2013) and our findings (see SAIM-WTA section).

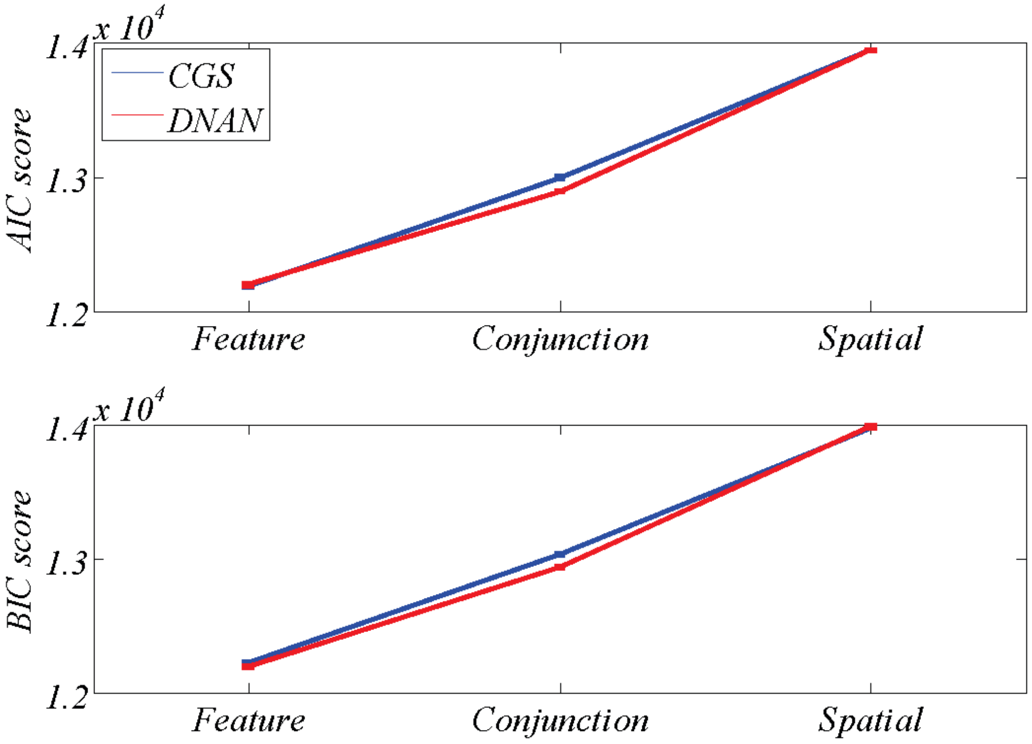


Fig. 6.10: Comparison of BIC scores and AIC scores from ADyNeN and CGS models for the three tasks (feature search, conjunction search and spatial configuration search). The error bars indicate the standard error without within-participant variance (Cousineau, 2005). BIC and AIC penalize the quality

with the model complexity as measured with the number of parameters. The graphs indicate that ADyNeN performed equivalently with CGS in feature and spatial tasks but better for conjunction task.

6.6 Comparison of ADyNeN and CGS

Fig. 9 compares ADyNeN and CGS models in terms of log-likelihood ratios across different tasks and conditions. Additionally, it shows how the model adapts to different display sizes without manipulating parameter settings. Overall, ADyNeN model performs better than CGS model but the difference in performance appears to drop with the difficulty of the tasks. Spatial configuration has the smallest overall difference. Interestingly, the CGS model in the absent trials show smaller difference in performance than present trials and disappears in spatial configuration task. This may be due to numerical reasons as additional parameter in present condition (target saliency) negatively impacts fitting process for this condition. However, this does suggest that there is a trade-off in whether present or absent conditions are represented better. Such situation would indicate that simply adding separate non-decision time for absent trials is not sufficient to explain the differences between tasks. Additionally, CGS model show a tendency for a drop in fit across display sizes but this was mainly the case for present trials which could be explained by the same numerical issue since this tendency is not consistent in absent condition. However, conjunction task shows the deepest slip which may be the case caused by the constrained introduced to identification drift and threshold (Moran et al. 2013). We found that the model performed better in conjunction task when this biologically plausible constraint was omitted which matches their findings as well. Contrary to the CGS model, ADyNeN show similar performance on both conditions but the most surprising result was in its capability to fit different display sizes similarly way. Only the fits of small set sizes seem to be slightly lower for feature and spatial configuration tasks. This is a counter-intuitive result because parallel models are expected to perform well in these conditions. Overall, CGS predicted RT distributional changes with a lower precision than ADyNeN model. However, the weaknesses emerge in different areas, CGS should further be investigated in its ability to account for both conditions since this was a main selling point (Moran et al., 2013) and ADyNeN model requires to address the issue of small set sizes.

z-value, p-value	**AIC**	**BIC**
Feature	0.36,0.717	**-2.05, 0.040**
Conjunction	**-3.70, <0.001**	**-3.62, <0.001**
Spatial	0.45,0.654	0.67, 0.502

Table 6.1: Comparison of BIC scores and AIC scores from ADyNeN and CGS for the three tasks (feature search, conjunction search and spatial configuration search) using Wilcoxon sign-rank test. Only conjunction comparisons were significant in favour of ADyNeN. BIC score also was different for feature task in favour of ADyNeN. The remaining comparisons were not significant.

Considering that ADyNeN model has nine parameters while CGS eight we needed to asses the performance by taking account the complexity of the models. We used two standard model comparison methods for comparisons when maximum likelihood is known: Akaike information criterion (AIC) and the Bayesian information criterion (BIC). This difference of complexity (as measured in terms of number of parameters) is particularly marked between CGS and ADyNeN. Fig 6.10 show the resulting scores with AIC being represented by upper graph while BIC by the lower graph. Both scores show very similar results for feature and spatial configuration tasks but not conjunction task. The scores clearly indicate a better performance by ADyNeN model. We performed Wilcoxon sign-rank test on these scores (see Table ?; for results) and most of the observation made in the graph match significance test results. Both scores show a significantly better penalised fits for ADyNeN model in conjunction search task and both scores show non significantly better penalised fits for CGS model. However, the scores show significantly better penalised fits for feature search favouring ADyNeN model when BIC score is considered but there is no significance when AIC is considered. Overall, the results show a slight preference of ADyNeN model over CGS despite CGS having one parameter less.

6.7 SAIM-WTA in absent trials

ADyNeN model is an extension of SAIM-WTA with an introduction of asymmetrical inhibitory weights, i.e. target inhibits more than distractors. A key observation is that absent trials are modelled using SAIM-WTA model and the success is significantly better than in the results of the initial study. This setup has seven free parameters (mutual inhibition weights, noise, decision boundary, sigmoid slope with shift, accumulation and motor error) in contrast with the three used for present trials study. Saliency value is constant for all nodes and are set to one since all inputs are the same. We can assess its performance in isolation of present trials since there are no significant differences between the fits of present and absent conditions suggesting near optimal solutions (feature: $z = 1.14; p = 0.255$, conjunction: $z =- 1.17; p = 0.243$ and spatial: $z =- 0.34; p = 0.735$). This assumption is quite reliable because fitting is biased towards the absent trials since it has fewer dimensions (equivalent to observed issue for CGS). We did not compare these results to the CGS model because it had issues balancing two conditions thus performance is likely to be below its capability when a single condition is considered. Such comparison would only be possible if CGS was fitted to absent trials in isolation of present condition.

6.8 General Discussion

6.8.1 Summary

This section aimed to introduce a novel parallel model for explaining distributional changes with an increasing display size. In order to do so, we repeated a study in section about SAIM-WTA with some practical changes. We numerically fitted two computational models to RT-distributions from three visual search experiments (feature search, conjunction search and spatial configuration search) but also included absent trials which were deliberately left out in SAIM-WTA study. A serial Competitive Guided Search model (CGS; Moran et al, 2013) was fitted again in order to accommodate it to absent trials. A proposed new Asymmetrical Dynamic Neural Network (ADyNeN) model representing a parallel search was fitted to compare their performances. The introduction of this competitive parallel model extends the knowledge about visual search and suggests novel perspectives to think about this unresolved issue.

Our study introduced a first parallel model that can explain visual search RT-distributions in present as well as absent conditions without changing any variables. It managed to successfully reproduce changes observed within distributions with additional items within display. Our direct comparisons between ADyNeN and CGS models illustrates that ADyNeN explains RT-distributions better than CGS but this difference become insignificant when model complexity is considered. However, there is significance in conjunction search task which may be due to parameter constraint imposed by Moran et al. (2013) to make model more plausible.

We confirmed their findings that this constraint negatively impacts overall fits in this task. This finding suggests that there is a focus on a single feature rathen two features in conjunction search tasks even substantially smaller number of trials compared to Wolfe et al. (2010) dataset. Such strategy effectively reduces the number of distractors that impair visual search. Overall, ADyNeN model is slightly better than CGS model in explaining distribution changes of these visual search experiments.

Crucially, ADyNeN model adapts better to the presence of absent trials and does not require any parameter changes to accommodate the lack of target. It explained the two conditions equally well, while CGS model introduced a different shift for trials without target to better explain the distributional differences between two conditions. This parameter models residual time which is not built explicitly for ADyNeN model. ADyNeN model explains this distributional difference between two conditions by the presence, or the lack, of inhibition from the target. The absence of the target increases competitive processes between distractor items because overall neuronal activity becomes higher. This leads to slower or faster decision making depending on how similar the distractors are to the target. Therefore, ADyNeN does not predict a difference in residual time compared to CGS model.

6.8.2 Perceptual Grouping

An interesting weakness of ADyNeN model is its performance with few items within display. This is unexpected finding because it is a parallel model which is expected to perform well when there’s few items within display. A plausible explanation for the result is in the change of the processes required to search for the items because the average distance between items drops.

Lower average distance between similar items encourages perceptual grouping which disappears when only few items exist in the display. In this case the model would fit better the other three display sizes because of its inability to fit items that do not group or due to numerical reasons where larger display sizes together contribute more to overall fit. Interestingly, this was not observed with conjunction search task which suggests that people struggle to group items in such task unlike the other two tasks. Importantly, these findings confirm predictions about perceptual grouping in Biased Competition theory (Duncan, & Humphreys, 1989). Notice that distractor-to-distractor inhibition has the lowest value of the three tasks which is an indication of network rejection as a group. Additionally, the low item set effect is slightly lower for spatial configuration task when compared to feature task. In order to test these predictions the experiments would have to be designed that would modulate target-distractor and distractor-distractor similarities as well as control for distance between distractors. Probably, the most suitable experiment would reproduce grouping results for RT-distributions by Müller-Plath and Elsner (2007).

6.8.3 Similarity vs Sensitivity

Aside qualitative performance of ADyNeN model, it also introduces a few cognitive predictions that possibly could be tackled to push our understanding of underpinning processes within brain. We separate a traditional understanding of saliency into two processes contributing to the observations attributed to saliency (Itti & Koch, 2000). We believe that these two processes are a similarity between the components and a sensitivity to the stimuli itself. This means that the brain is equivalently sensitive to two items of the same feature but these items may be interfering with each other due to their similarity. For example, there may be a greater sensitivity to the colour than orientation. This distinction suggests that it is possible to find a common perceptual similarity between target and distractor for the two features. However, the colour would produce a response at a different pace. This is a problematic comparison because equivalent similarity would have to be established for completely different stimuli in order to test this prediction. However, hypothetically such relationship introduces the possibility to model heterogeneity but produce effects via sensitivity to the stimulus, such involuntary attentional grab.

6.8.4 Limitations

Overall, the findings indicate that ADyNeN model describes RT-distributions with additional competing components better than a serial identification process convoled Wald-distribution. Palmer et al. (2011) showed that ex-Wald (convolution of Wald and exponential function) characterises RT-distributions the best in comparison to other parametric functions. Results by Palmer et al. (2011) suggest potential structural issues as the cause because they show that ex-Wald parameters change depending on whether it is absent or present condition as well as on the number of distractors. CGS model assumes that probabilistic termination combined with convolved Wald functions is sufficient to explain the parameter changes. For the model to work well, these parameter changes should match the changes observed in Palmer et al. (2011) study. The same study shows that exponential function is influenced by the display size which is assumed to be constant by Moran et al. (2013). Future studies should explore whether CGS reproduces Wald distribution parameter changes by combining a serial process with dynamic termination. In case the model cannot reproduce similar results to ex-Wald distribution, it would raise questions on validity of serial architecture during long display visual search tasks as envisioned by the authors.

Chapter 7:
General discussion

7.1 Summary

In our research we have attempted to expand the understanding about visual search by designing new models and developing novel methods to improve their assessment. Our goal was to design a parallel model capable of explaining distributional changes with a increasing number of distractors without changing the parameter values. We considered a winner-takes-all (WTA) mechanism used in Selective Attention for Identification Model (SAIM:) adapted for visual search task as a base model. Upon investigating its performance we developed a novel Asymmetrical Dynamic Neural Network (ADyNeN) model which relaxed symmetry of mutually inhibitory nodes within original architecture. We compared both models to Competitive Guided Search (CGS; Moran et al.) model designed for explaining serial search strategies evidenced by eye movement studies and identified as the best model to describe visual search data. We hope that these model developments will contribute to an already existing wide literature on the issue of serial vs parallel search and a general visual search understanding.

In order to fit the models to the RT-distributions, we employed a novel non-parametric kernel density estimator (KDE) method. As a result, we successfully employed a robust maximum likelihood estimator (see Robust Likelihood Approximation section in Methodology part) to identify the best fitting parameters. We used the differential evolution Markov chain Monte Carlo (DE-MCMC) method to find the best fits for the SAIM-WTA chapter study and for identifying initial parameter populations for ADyNeN chapter study. For further fitting in the second ADyNeN study we used a novel Sequential Importance Sampling by Filtered Clusters (SISFC) algorithm which showed superior efficiency compared to DE-MCMC. This section contributed to the claims made in Methodology chapter that these methods are capable to find solutions for analytically intractable computational models that fit RT-distributions produced by visual search experiments. Moreover, this research demonstrated that these methods can successfully fit computational models to RT-distributions from the visual experiments. We also showed that the KDE is an excellent method to represent RT distributions.

7.2 Methodology

7.2.1 Approximating RT-distributions

The thesis has harnessed some of the most advanced methods in order to determine whether RT-distributions in respect to accuracy can sufficiently constrain serial or parallel theory and falsify either. An important introduction in the methodology was approximation of the unknown PDF using KDE. This approach has not been used much but has resurfaced in recent years (Turner et al...) as means to quantify models that do not have an analytical likelihood function. The KDE introduced in the thesis (Kristan et al. 2011) is more powerful than traditional KDE (Silverman, 1986) because it shows greater accuracy with fewer components. In general, this thesis shows that there is a more accurate solution with relatively few components (e.g. 8) than produced by traditional KDE. This indicates, by extension to be a more accurate solution than other approximation methods (Turner & Sederberg, 2014; van Zandt, 2001). The main benefit is achieved via clustering of the simulated RTs, which is a costly process. Though, the cost may not be acceptable for larger scale approximation such as Bayesian. We found that a single KDE approximation generally costs ~1s which adds to ~13h for 50000 parameter evaluations.

Note that the benefit of the clustering is an important observation because quantile methods cluster data by discretizing cumulative function. Intuitively, it suggests that QML should be a better method, contrary to the findings by Turner and Sederberg (2014). Note, that the number of components required to approximate well KDE depended on the size of the data. Larger data is better approximated with more components. QML method has become a gold standard method that uses five bins irrespective of the data size. Turner and Sederberg (2014) did not question this established notion in their study and used identical settings. Increasing the number of percentile bins should have produced equivalent to superior performance. Therefore, this was an unfair comparison from the very beginning. It is expected for future studies to confirm this observation which would firmly establish quantile-based method as the most accurate non-parametric density approximator due to its lower computational costs. Unlike proposed KDE, QML does not require hierarchical clustering and sorting is the most costly function within the method. The oKDE does not require sorting but performs multiple complex mathematical operations that add to greater overall costs.

7.2.2 Robust Parameter Evaluation

This thesis has proposed a robust likelihood method as an alternative approach to dealing with contaminants within RT-distributions. Currently, there exists a few approaches to handling this issue. The most common approach is to use QML (Heathcote, Brown & Mewhort, 2002) which averages data, though using mixture model by explicitly modelling contamination has proven to be the most successful approach thus far (Ratcliff & Tuerlinckx, 2002; Ratcliff & McKoon, 2008; Wagenmakers et al., 2008). The thesis did not compare these methods explicitly but outlined issues on these methods within chapter enable additional verification studies.

The most significant limitation of the introduced method is establishing proper weighting between fitting RT-distributions and accuracy. Accuracy is a discrete value while RTs are continuous values which can impact the relative importance during parameter estimation process. Frequently, matching accuracy via simulations is undervalued though can be overvalued as well. Currently there is no evaluations of the merits of chosen weighting of fitting importance between the two components. However, overall results show that this lack of proper weighing did not cause major issues to the parameter estimation process. In fact, our weighing approach is arguably better than the method proposed by Turner & Sederberg (2014). They suggested using defective cumulative functions which works only in the best case scenario. Generally, likelihood values for PDF distribution can vary from very small to very large depending on the width of the distribution, thus the weighting becomes determined by the width of the PDF distribution which is not a desirable property. An alternative QML-based approach treats error rates as additional bin (Moran et al., 2016). It gives identical weighting to this new bin than larger percentiles irrespective of the error rate which violates the expectation that probabilities for correct and incorrect components sum to one. In conclusion, our robust likelihood reduces the effect of underrepresented RT points boosting the overall weighting towards the accuracy at matching error rates within data. However, some additional future mathematical work is required to establish proper weighting between correct and incorrect decisions.

7.2.3 Efficient Parameter Estimation

There was an algorithm developed as part of the thesis for more efficient Bayesian approximation. This algorithm was successfully employed for parameter estimation. The purpose of the algorithm was to find the means to identify solutions when no maximum likelihood exists. The lack of maximum likelihood creates a numerical problem where virtually any algorithm fails to approximate the posterior distribution in a reasonable amount of time. Under these circumstances this novel algorithm has substantially reduced the parameter estimation costs for the given models. However, it remained insufficient for approximating posterior distribution of ADyNeN model because a single parameter set evaluation take ~8s, coupled with KDE evaluation it amounts to ~9s. Performing 50000 evaluations that are common in Bayesian settings would take 5 days per participant.

Moreover, the algorithm lacks correctness analysis via mathematical proofs as all experiments are performed via simulations which does not establish a proof of correctness. Though results are very promising, some slight adjustments are expected before a true posterior distributions can be found. At the moment, this algorithm is within the Approximate Bayes Computation category since it lacks formal mathematical relationship between analytical and approximated posterior distributions. Regardless, the algorithm is good for finding good parameter estimates when no maximum likelihood exists which has been shown via cognitive modelling applications.

7.3 Discussion and Future Directions

7.3.1 SAIM-WTA vs ADyNeN

Our follow-up revisit of SAIM-WTA model in absent conditions (6th chapter) showed strikingly better fits than produced for present trials only (5th chapter). There's a couple of possible causes for this inconsistency. Firstly, it may be the case that our observation that SAIM-WTA model does not explain RT-distributions well with increasing number of distractors when present trials are considered is correct. On the other hand, absent conditions had more free parameters than SAIM-WTA used for present trials. We may have overconstrained the model reducing its flexibility to explain distributional differences. This contradicts an earlier statement that the model performed similarly with additional parameters but the relationships between parameters also suggest overparameterisation for ADyNeN model. Thus, the best performing model is simpler than ADyNeN but more complex than SAIM-WTA.

7.3.1 Race model overview

Moran et al. (2016) had attempted fitting a parallel race model to the response time distributions. The model they have considered is architecturally different from our proposed SAIM-WTA and ADyNeN models and the results of our and their studies possibly solve some architectural issues outlined for designing parallel models of visual search task (Townsend & Wenger, 2014). It relatively difficult to disentangle the main causes of the failure of their parallel race model because one poor decision masks other good decisions. Firstly, for naming convenience we will refer to Moran et al’s. (2016) model simply as RT-race model so there would not be confusion with general race models.

The following views are deduced using the knowledge about success of our models on Lin et al. (2015) dataset and architecture of RT-race model. We have established that SAIM-WTA is a smaller version of ADyNeN model (see 6th chapter) thus we will focus on establishing the connection between SAIM-WTA and RT-race models. RT-race model does not have competitive interference as SAIM-WTA is the largest difference. However, RT-race model is a limited capacity model which enables to factor the increased interference observed as display size effects. Another difference is in the number of decision boundaries. RT-race model treats each node as individual diffuser that accumulates information about the identity of the corresponding item. SAIM-WTA model assumes there are no identification errors, thus a single decision boundary is used.

For future analysis consider a simplified RT-race model that makes identical assumption that there is no identification error. This can be accomplished by introducing a separate drift rate as equivalent to different inputs for distractors in SAIM-WTA. This is a common simplifying model assumption for modelling n-diffusers (Usher & McClelland, 2001). However, the observed increase in errors for absent trials would no longer exist. By assuming it is always a motor error, the observed increase in errors would no longer be present (Moran et al., 2016). Note that authors made an identical assumption for CGS model (Moran et al., 2013). Their model assumes that identification follows a diffusion process governed by Wald distribution which also is a simplifying diffusion process that was enforced on RT-race model (Schwarz, 2001). If this assumption was relaxed and replaced with identical misidentification error probability as RT-race model has, it would fail at fitting absent trials as well. The performance of RT-race model would have been substantially better under equivalent assumption of perfect identification.

Perfect identification does not prove equivalent performance. However, this RT-race model version can be further simplified. SAIM-WTA model terminates after the the first identification while RT-race model has a separate node for termination. For the sake of comparison RT-race model could have also terminated after the first identification. With this simplification the two models differ in the way interference is modelled. SAIM-WTA has mutual inhibition while RT-race model has a limited capacity, both processes slow down accumulation though not equivalently. However, upon closer look at quantile by Moran et al. (2016), these simplifications suggest RT-race model to be capable to account for RT-distributions equivalently well. Though, this has to be confirmed with a future study.

7.3.2 Competition Implications

The crucial difference between competitive and race models is within inhibitory connections between the components. A race model does not have inhibitory connection but frequently possess processing capacity limitations which slow down processing when more items are introduced by replacing the drift rate with the function dependent on the number of items (Ward & McClelland, 1989). Competitive model has a similar component in mutually inhibitory connections. These inhibitory connections limit the overall activity within network to some maximum activity which slows down accumulation with additional items. However, it differs in respect that race model does not limit an overall activity within the network and simply slows down the overall accumulation while competitive architecture has a clear bound on the average activity and any gain for some nodes have to come at the expense of other nodes. It is difficult to tell to what degree the two can mimic the behaviour but it may be possible to introduce a capacity function such as Hick's law (Usher & McClelland, 2001) that would capture similar overall accumulation slow-down as for competitive model. However, a parallel race model would not be able to exhibit suppression of all items but one though may be capable of producing identical RT-distributions as that one item that passed the threshold.

7.3.3 Competitive vs Race parallel models

Provided our observations that leveling out playing field for parallel race model with CGS model should yield equivalently good results, a future study comparing parallel race model with competitive race model is required. However, there are underlying assumptions that influence overall preference for either model. Firstly, a race model is analytically tractable model that is easy to evaluate and parameters are easily found (Moran et al., 2016). Moreover, there is no issue of absent trials being too slow during feature search because of the competitive nature of the model, an observation that has resulted into ADyNeN model. On the other hand, ADyNeN has relaxed a mutual inhibition parameter that was constrained for SAIM-WTA. In fact SAIM-WTA can be reduced to the race model by setting this inhibitory part to 0. This creates two possible solutions that can explain RT-distributions, ADyNeN model which introduces asymmetrical interactions between competing items within visual display or a parallel race model which changes its drift rate and decision boundary depending on the number of parameters within visual field (Moran et al., 2016). It may seem that a race model would have too many free parameter, however, we expect that introducing the same identification assumption would reduce the number of required parameters (Usher, Olami, & McClelland, 2002) because Moran et al. (2016) made a genuine effort to fit RT-distributions by freeing up as many variables as possible. However, this has to be confirmed in the future studies.

So where does ADyNeN stand if the observations are confirmed? Firstly, ADyNeN is an extension of SAIM-WTA which can be reduced to the race model. It is architecturally more complex and has a potential to adapt and explain a wider range of cognitive phenomena with observed race conditions being one side of the coin. Unlike the race model, ADyNeN can capture complex interactions between items, their relatedness and possibly group items if they are related enough. Moreover, the relations can be expanded to the spatial interactions of the items with distant but similar items being less likely to be grouped. On the other hand, it is not tractable analytically and reduction to parallel race model may be preferred when such model is applicable.

7.3.4. RT-distributions

The main theme of the thesis addressed the question whether RT-distributions produced via standard visual search task could constrain the serial or parallel theory to falsify either. This was suggested by Wolfe, Palmer and Horowitz (2010) in his work showcasing the failure of the best serial model at the time in fitting RT-distribution. As a result, Moran et al. (2013) introduced an extension to the model that equipped it with such capability and failed to develop a parallel model Moran et al. (2016) that has such capacity. This raised two questions, did they falsify the theory positing a single-stage parallel architecture or there is an architecture that has been overlooked and is there a more stringent evaluation to showcase the weaknesses within serial theory. This thesis answers both questions positively but not the main theme.

The findings within thesis indicate that RT-distributions are a weak constraint at distinguishing the two theories. This indication was already present within their new and somewhat overlooked 'x-score transform' method, which briefly appeared in the literature (Wolfe, Palmer, & Horowitz, 2010; Moran et al., 2013), though, never got published. This transform resulted in a remarkably similar distributions of the three visual search tasks, especially when averaged across participants. This finding somewhat negatively impacts the significance of the message that RT-distributions could be crucial for visual search because mean and variance are linear which is removed once transform is performed. This finding strongly suggests that an inclusion of the whole RT-distribution in parallel vs serial dichotomy is unlikely to aid in distinguishing the two tasks and an addition of variance may have been sufficient. Such transformation prior to proceeding the complex KDE analysis would have shown if simulations of the models could differentiate between different tasks. Note, that the transform was performed on quantiles which are known to be less accurate. Regardless, an improved transformation may prove that mean RTs are insufficient for this particular task, unless the data is very large and analytical solutions are available.

A crucial development within the thesis is an introduction of the novel Asymmetrical Dynamic Neural Network model which represents the biologically plausible class of parallel models consisting of the single-stage. It is well-established that serial models require two-stages for the models to explain fast searches that are parallel accepted to be parallel (Treisman & Gelade, 1979). This has not been proven for parallel model assumptions and ADyNeN model introduces another single-stage model that is capable at explaining RT-distributions without requiring the change in the parameters as it would be required by employing diffusion model. The model on average outperformed the serial counterpart but it remains to see if it would perform as good or even better as individual diffusion models for each set size and task difficulty. Such study is required to finalise the correctness of the model for setsize effects.

The dispute resides within the presence of the seriality as a serial identification process (Wolfe et al., 1989; Wolfe, 1994, 2007; Chun & Wolfe, 1996; Moran et al., 2013). It would be difficult to identify the transition from parallel to serial as evidence relying solely on RT-distributions or linear changes of these distributions (Wolfe, Palmer & Horowitz, 2010). All limited capacity models predict that it can be exceeded and a serial process is likely to fill the gap but this limitation proved to be illusive. It is possible that it would not be possible to falsify a single-stage model unless RT-distribution consistently (repeated studies produced the same outcome) has more than one mode in the resulting distribution. A standard visual search task may not be sufficient to consistently produce such outcome and alternative measurements, such as eye movements, may be required.

We have sufficient evidence to believe that serial processes exist which generally are predicted by the eye movements (Deubel & Schneider, 1996; Hulleman & Olivers, 2016). The WTA mechanism is commonly used to model saccades (Parkhurst, Law, & Niebur 2002; Itti & Koch, 2001) since selective attention precedes these eye movements (Peterson, Kramer, & Irwin, 2004). However, overall the studies are inconclusive. While the first saccade takes time to process, the following saccades frequently are within a very short period compared to the initial computation of saccade. Moreover, the second saccade is frequently towards an actual target which is in the neighbourhood of the initial eye movement (Findley, Brown & Gilchrist, 2000). These findings show that most of the eye movement cannot be treated as an indication of the serial process as only an eye gaze to completely different visual processing area can be considered as close to serial. In fact, these eye movement findings suggest ADyNeN’s extension for explaining RT-distributions combined with eye movements produced by parallel processes. While current ADyNeN version does not consider the effects of distance between items, such introduction via modulation of the inhibition is viable. Items within groups could be rejected in rapid succession. The modified model would encourage accumulating until all the items are either above decision boundary or completely suppressed. Such mechanism enables a natural non-exhaustive parallel search. However, absence of eye movements indicate attention but attention can be shifted without saccades. Some studies show similar performance despite the presence/absence of saccades (Wolfe, 1998), thus an addition of eye movement would not introduce sufficient constraint to ADyNeN model.

7.3.5 Speed/Accuracy Trade-off

An important experimental paradigm focuses on speed/accuracy trade-off (SAT: Dosher, Han & Lu, 2004, 2010). It emphasis that RTs do not provide all the information about the underlying cognitive processes. This paradigm looks how RTs and accuracy change as various levels of speed are encouraged. A general average accuracy for the given speed is measured which illustrates flattening error rate for longer display sizes. The RT-distributions also change differently as a function of display size at various levels of speed being emphasised. These studies show that unlimited capacity parallel processes often are sufficient to explain SAT effects. However, for increasingly longer viewing durations distinguishing two architectures becomes illusive. Lin et al. (2015) data represent the later category and the findings within the thesis is not unexpected for the proponents of the paradigm (Townsend & Ashby, 1984). Neither ADyNeN nor CGS were tested in this paradigm and it remains an open question whether either theory could explain SAT data. However, ADyNeN model is better positioned since it can be reduced to the race model.

7.4 Conclusions

To sum up, we have used some of the most advanced numerical methods to compare competing parallel and serial theories and their representative models in visual search paradigm. Specifically we looked at response time distributions and whether models can account for distributional changes with an increasing number of distractors. We used Competitive Guided Search (CGS; Moran et al., 2013) model which is the latest prominent installment of the serial search Guided model (Chun & Wolfe; 1996). We used a winner-take-all mechanism from Selective Attention for Identification Model (SAIM; Mavritsaki, Heinke, Humphreys & Deco, 2006; Heinke & Humphreys, 2003; Heinke & Backhaus, 2011) as a base model which we compared to CGS model and found that on average it performed worse than SAIM-WTA model. We followed these findings with developing an extension of SAIM-WTA by relaxing inhibitory connection between competing nodes to introduce bias depending on the identity of the item the node represents. A new Asymmetrical Dynamic Neural Network (ADyNeN) model has successfully explained the distributional changes observed within different tasks and on average outperformed CGS model. It is slightly more complex model than CGS but future research could identify relationships between parameters that can be simplified to reduce the complexity.

References

Anderson, G. M., Heinke, D., & Humphreys, G. W. (2010) Featural guidance in conjunction search: The contrast between orientation and color. *Journal of Experimental Psychology: Human Perception and Performance*, 36, 5, 1108-1127. http://dx.doi.org/10.1037/a0017179

Balota, D. A., & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry: The power of response time distributional analyses. Current Directions in Psychological Science, 20, 160–166, doi:10.1177/0963721411408885.

Basu, A., & Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, *46*(4), 683-705.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 113-147.

Bogacz, R., & Cohen, J. D. (2004). Parameterization of connectionist models. *Behavior Research Methods*, *36*(4), 732-741.

Bogacz, R., Usher, M., Zhang, J., & McClelland, J. L. (2007). Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362, 1655–1670. http://doi.org/10.1098/rstb.2007.2059

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. Cognitive psychology, 57(3):153-178.

Bruce, N. D. B., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. Journal of Vision, 9, 1–24, http://www.journalofvision.org/content/9/3/5, doi:10.1167/9.3.5.

Bundesen, C. (1990). A theory of visual attention. Psychological Review, 97(4), 523-547. http://dx.doi.org/10.1037/0033-295X.97.4.523

Burbeck, S. L., & Luce, R. D. (1982). Evidence from auditory simplereaction times for both change and level detectors. Perception & Psy-chophysics, 32(2), 117–133.

Busemeyer, J. R. & Diederich, A. (2010) Cognitive modelling. Sage.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson’s method. Tutorial in Quantitative Methods for Psychology, 1(1), 4–45.

Byers, R. H., & Shenton, L. R. (1999). Sister chromatid exchange data fit with a mixture of Poisson distributions. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *427*(2), 157-162.

Chun, M. M., & Wolfe, J. M. (1996). Just say no: How are visual searches terminated when there is no target present?. *Cognitive psychology*, *30*(1), 39-78.

Cressie, N., & Read, T. R. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 440-464.

Cwik, J. & Koronacki, (1998) A combined adaptive-mixtures/plug-in estimator of multivariate probability densities, Computational Statistics and Data Analysis 26, 199–218.

Desimone, R. & Duncan, J. (1995) Neural mechanisms of selective visual attention. Annual Review Neuroscience, 18, 193–222.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision research*, *36*(12), 1827-1837.

Dosher, B. A., Han, S., & Lu, Z. L. (2004). Parallel processing in visual search asymmetry. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(1), 3.

Dosher, B. A., Han, S., & Lu, Z. L. (2010). Information-limited parallel processing in difficult heterogeneous covert visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1128.

Doucet, A., & Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, *12*(656-704), 3.

Duncan, J., Humphreys, G. W., & Ward, R. (1997). Competitive brain activity in visual attention. Current Opinion in Neurobiology, 7, 255–261.

Duncan, J., & Humphreys, G. W. (1989) Visual search and stimulus similarity. *Psychological Review*, 96, 433–458.

Duncan, J., & Humphreys, G. W. (1992). Beyond the search surface: Visual search and attentional engagement. Journal of Experimental Psychology: Human Perception and Performance, 18(2), 578 –588.

Eckstein, M. P. (2011) Visual search: A retrospective. *Journal of Vision*, 11(5):14. doi: 10.1167/11.5.14.

Eidels, A. (2012). Independent race of colour and word can predict the Stroop effect. *Australian Journal of Psychology*, *64*(4), 189-198.

Egeth, H., Jonides, J., & Wall, S. (1972). Parallel processing of multielement displays. Cognitive Psychology, 3, 674–698, doi:10.1016/ 0010-0285(72)90026-6.

Egly, R., Driver, J., & Rafal, R. D. (1994) Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. Journal of Experimental Psychology: General, 123 (2), 161–177.

Findlay, J. M., Brown, V., & Gilchrist, I. D. (2001). Saccade target selection in visual search: The effect of information from the previous fixation. *Vision research*, *41*(1), 87-95.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., Von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E. J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*, *105*(45), 17538-17542.

Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S., Wagenmakers, E. J., ... & Turner, R. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *Proceedings of the National Academy of Sciences*, *107*(36), 15916-15920.

Grazian, C., & Robert, C. P. (2015). Jeffreys' priors for mixture estimation. In *Bayesian statistics from methods to models and applications* (pp. 37-48). Springer, Cham.

Heathcote, A., Brown, S., & Mewhort, D. J. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic bulletin & review*, *9*(2), 394-401.

Ho, T. C., Brown, S., & Serences, J. T. (2009). Domain general mechanisms of perceptual decision making in human cortex. *Journal of Neuroscience*, *29*(27), 8675-8687.

Hohle, R. H. (1965). Inferred components of reaction times as functions of foreperiod duration. Journal of Experimental Psychology, 4, 382–386.

Hulleman, J. & Olivers, C. N. L. (2016).The impending demise of the item in visual search. Behaverial Brain Science doi:10.1017/S0140525X15002794

Humphreys, G. W., & Müller, H. J. (1993). SEarch via Recursive Rejection (SERR): A connectionist model of visual search. *Cognitive Psychology*, 25, 43–110.

Itti, L., Koch, C. & Niebur, E. (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11), 1254-1259. doi: 10.1109/34.730558

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, *40*(10), 1489-1506.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature reviews. Neuroscience*, *2*(3), 194.

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological review*, *121*(1), 1.

Karlis, D., & Xekalaki, E. (1998). Minimum Hellinger distance estimation for Poisson mixtures. *Computational Statistics & Data Analysis*, *29*(1), 81-103.

Karlis, D., & Xekalaki, E. (2001). Robust inference for finite Poisson mixtures. *Journal of Statistical Planning and Inference*, *93*(1), 93-115.

Kristan, M., Leonardis A., & Skočaj, D. (2011) Multivariate online kernel density estimation with Gaussian kernels. Pattern Recognition 44(10):2630-2642.

Knuth, G. (1998). The Art of Computer Programming, Seminumerical Algorithms, Vol. 2, Addition Wesley. *Reading, Massachusetts*.

Kwak, H., Dagenbach, D., & Egeth, H. (1991). Further evidence for a time-independent shift of the focus of attention. Perception and Psychophysics, 49(5), 473-480.

Letham, B., Letham, P. A., Rudin, C., & Browne, E. P. (2016). Prediction uncertainty and optimal experimental design for learning dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *26*(6), 063110.

Leys, Ley, Klein, Bernard, & Licata (2013) Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median, *Journal of Experimental Social Psychology*, 49, 4, pp. 764-766

Lin, Y., Heinke, D., & Humphreys, G. W. (2015) Modeling visual search using three-parameter probability functions in a hierarchical Bayesian framework. Att., Perc., & Psych., 77, 3, 985-1010.

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 883-914.

Lu, Z., Hui, Y. V., & Lee, A. H. (2003). Minimum Hellinger distance estimation for finite mixtures of Poisson regression models and its applications. *Biometrics*, *59*(4), 1016-1026.

Luce, R. D. (1986). Response times: Their role in inferring elementary mental organization. New York: Oxford University Press.

Mandal, A., Basu, A., & Pardo, L. (2010). Minimum disparity inference and the empty cell penalty: Asymptotic results. *Sankhya A-Mathematical Statistics and Probability*, *72*(2), 376-406.

Matzke, D., & Wagenmakers, E. J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic bulletin & review*, *16*(5), 798-817.

Mavritsaki, E., Heinke, D., Allen H., Deco, G., & Humphreys, G. W. (2011) Bridging the gap between physiology and behavior: Evidence from the sSoTS model of human visual attention. *Psychological Review*, 118(1), 3-41

Michael, J. R., Schucany, W. R., & Haas, R. W. (1976). Generating random variates using transformations with multiple roots. *The American Statistician*, *30*(2), 88-90.

Miletić, S., Turner, B. M., Forstmann, B. U., & van Maanen, L. (2017). Parameter recovery for the Leaky Competing Accumulator model. *Journal of Mathematical Psychology*, *76*, 25-50.

Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. *The quarterly journal of experimental psychology*, *43*(4), 907-912.

Müller-Plath, G. & Elsner, K. (2007) Space-based and object-based capacity limitations in visual search, *Visual Cognition*, 15:5, 599-634, DOI:10.1080/13506280600845572

Moran, R., Zehetleitner, M., Mueller, H. J., & Usher, M. (2013) Competitive guided search: Meeting the challenge of benchmark RT distributions. Journal of Vision, 13(8): 1-31.

Moran, R., Zehetleitner, M., Liesefeld, H. R., Müller, H. J., & Usher, M. (2016). Serial vs. parallel models of attention in visual search: accounting for benchmark RT-distributions. *Psychonomic bulletin & review*, *23*(5), 1300-1315.

Mordkoff, J. T., & Yantis, S. (1991). An interactive race model of divided attention. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(2), 520.

Narbutas, V., Lin, Y. S., Kristan, M., & Heinke, D. (2017). Serial versus parallel search: A model comparison approach based on reaction time distributions. *Visual Cognition*, 1-20.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization Computer Journal. 7: 308–313. doi:10.1093/comjnl/7.4.308

Palmer, J. (1995). Attention in visual search: Distinguishing four causes of a set-size effect. *Current directions in psychological science*, *4*(4), 118-123.

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision research*, *40*(10-12), 1227-1268.

Palmer, E.M., Horowitz, T. S., Torralba, A., &Wolfe, J.M. (2011).What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, 37, 58–71.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision research*, *42*(1), 107-123.

Patra, R. K., Mandal, A., & Basu, A. (2008). Minimum Hellinger distance estimation with inlier modification. *Sankhyā: The Indian Journal of Statistics, Series B (2008-)*, 310-322.

Peterson, M. S., Kramer, A. F., & Irwin, D. E. (2004). Covert shifts of attention precede involuntary eye movements. *Attention, Perception, & Psychophysics*, *66*(3), 398-405.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological review*, *117*(3), 864.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, *85*(2), 59.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347-356.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic bulletin & review*, *9*(3), 438-481.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873-922.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological review*, *116*(1), 59.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., & Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, *25*(15), 1923-1929.

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195-223.

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H. M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*(3), 414.

Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. Behavioral Research Methods, Instruments & Computers, 33(4), 457– 469.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). CRC press.

Singer, S., & Nelder, J. (2009). Nelder-mead algorithm. *Scholarpedia*, *4*(7), 2928.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*(3736), 652-654.

Storn, R., & Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization, 11(4):341-359.

Sun, C., & Hahn, J. (2006). Parameter reduction for stable dynamical systems based on Hankel singular values and sensitivity analysis. *Chemical engineering science*, *61*(16), 5393-5403.

Teodorescu, A. R., & Usher, M. (2013). Disentangling decision models: From independence to competition. *Psychological review*, *120*(1), 1.

Townsend, J. T. (1976). Serial and within-stage independent parallel model equivalence on the minimum completion time. *Journal of Mathematical Psychology*, *14*(3), 219-238.

Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. Psychological Bulletin, 96(2), 394–401.

Townsend, J. T., & Nozawa, G. (1995). Spatiotemporal properties of elementary perception: An investigation of parallel, serial, and co-active theories. Journal of Mathematical Psychology, 39, 321–359, doi:10.1006/jmps.1995.1033.

Townsend, J. T., & Wenger, M. J. (2004). The serial-parallel dilemma: A case study in a linkage of theory and method. *Psychonomic Bulletin & Review*, *11*(3), 391-418.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. Cognitive Psychology, 12, 97–136.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. Psychological Review, 95, 15–48.

Treue, S., & Trujillo, J. C. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*(6736), 575.

Tsetsos, K., Usher, M., & McClelland, J. L. (2011). Testing multi-alternative decision models with non-stationary evidence. *Frontiers in neuroscience*, *5*.

Turner, B. M., & Sederberg, P. B. (2012). Approximate Bayesian computation with differential evolution. *Journal of Mathematical Psychology*, *56*(5), 375-385.

Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, *56*(2), 69-85.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological methods*, *18*(3), 368.

Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. Psychonomic Bulletin & Review, 21(2):227-250.

Turner, B. M., Van Zandt, T., & Brown, S. (2011). A dynamic stimulus-driven model of signal detection. *Psychological review*, *118*(4), 583.

Turner, B. M., Sederberg, P. B., & McClelland, J. L. (2016). Bayesian analysis of simulation-based models. *Journal of Mathematical Psychology*, *72*, 191-199.

Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, *123*(1), 34.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. Psychological Review, 108(3):550-592.

Usher, M., Olami, Z., & McClelland, J. L. (2002). Hick's law in a stochastic race model with speed–accuracy tradeoff. *Journal of Mathematical Psychology*, *46*(6), 704-715.

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review*, *7*(2), 208-256.

Van Zandt, T. (2000). How to fit a response time distribution. Psychonomic Bulletin and Review, 7, 424-465.

Zandt, T. V. (2002). Analysis of response time distributions. *Stevens' handbook of experimental psychology*.

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods*, *40*(1), 61-72.

Verghese, P. (2001). Visual search and attention: A signal detection theory approach. *Neuron*, *31*(4), 523-535.

Vincent, P., Bengio, Y., (2003) Manifold Parzen windows, Adv. Neural Inf. Process. Syst. 849–856.

Vincent, B. T. (2015). A tutorial on Bayesian models of perception. *Journal of Mathematical Psychology*, *66*, 103-114.

Wagenmakers, E. J., van der Maas, H. L., Dolan, C. V., & Grasman, R. P. (2008). EZ does it! Extensions of the EZ-diffusion model. *Psychonomic Bulletin & Review*, *15*(6), 1229-1235.

M.P. Wand, M.C. Jones, Kernel Smoothing, Chapman & Hall/CRC, 1995.

Wang, P., Puterman, M. L., Cockburn, I., & Le, N. (1996). Mixed Poisson regression models with covariate dependent rates. *Biometrics*, 381-400.

Ward, R., & McClelland, J. L. (1989). Conjunctive search for one and two identical targets. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(4), 664.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt II [Principles of perceptual organization]. Psychologische Forschung, 4 , 301- 350

Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychologica*, 41, 67–85.

Woodgate P. J., Strauss S., Sami S. A., Heinke D. (2015) Motor cortex guides selection of predictable movement targets. *Behavioural brain research*, 287, 238-246. doi:10.1016/j.bbr.2015.03.057.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided Search: An alternative to the Feature Integration model for visual search. Journal of Experimental Psychology: Human Perception and Performance, 15, 419–433.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2), 202-238.

Wolfe, J. M. (2007). Guided search 4.0. In W. D. Gray (Ed.), Integrated models of cognitive systems (cognitive models and architectures) (pp. 99–120). Oxford, UK: Integrated Models of Cognitive Systems.

Wolfe, J. M. (1998). Visual search. In H. Pashler (Ed.), Attention (Vol. 1, pp. 13–73). London, UK:University College London Press.

Wolfe, J. M., Palmer, E. M., & Horowitz, T. S. (2010) Reaction time distributions constrain models of visual search. Vision research, 50(14), 1304-1311.

Wolfe, J. M., Palmer, E. M., & Horowitz, T. S. (2010). Reaction time distributions constrain models of visual search. *Vision research*, *50*(14), 1304-1311.

Zhao, Y., Humphreys, G. W., Heinke, D. (2012) A Biased-Competition Approach to Spatial Cuing: Combining Empirical Studies and Computational Modelling. Visual Cognition, 20(2), 170-210.

Zhao, Y., Heinke, D., Ivanoff, J., Klein, M. K. & Humphreys, G. W. (2011) Two components in IOR: Evidence for response bias and perceptual processing delays using the SAT methodology. Attention Perception & Psychophysics, 73(7), 2143-2159.

Appendix I: SAIM-WTA

Constant parameters	Noise	Sigmoid slope	Sigmoid shift	Inhibitory weight
	0.1	50	0.5	-1

The gamma-parameter in DE-MCMC was set to $2.38/\sqrt{2d}$ where *d* refers to the number of free parameters. A constant random noise was set to all parameter proposals of 0.001.

Feature Search Participant	Decision Boundary	Accumulation	Distractor Saliency	Outlier
1	0.5128	0.0019	0.6642	No
2	0.5214	0.0018	0.6537	No
3	0.5020	0.0017	0.6196	No
4	0.5067	0.0018	0.6381	No
5	0,5276	0.0016	0.0635	Yes
6	0.5011	0.0018	0.6672	No
7	0.5035	0.0019	0.6849	No
8	0.5079	0.0019	0.7145	No
9	0.5190	0.0020	0.6636	No
10	0.5297	0.0020	0.6413	No
11	0.5249	0.0016	0.5686	No
12	0.5221	0.0018	0.6756	No

13	0.5068	0.0020	0.6695	No
14	0.5184	0.0019	0.2815	Yes
15	0.5503	0.0018	0.6467	No
16	0.5047	0.0020	0.6612	No
17	0.5412	0.0017	0.6884	No
18	0.5058	0.0021	0.5642	No
19	0.5075	0.0019	0.6968	No

Conjunction Search Participant	Decision Boundary	Accumulation rate	Distractor Saliency	Outlier
1	0.5762	0.0015	0.7192	No
2	0.5709	0.0017	0.7067	No
3	0.5056	0.0015	0.6181	No
4	0.5503	0.0014	0.7780	No
5	0.7373	0.0012	0.6394	Yes
6	0.5228	0.0013	0.6968	No
7	0.5226	0.0014	0.8034	No
8	0.5927	0.0013	0.7208	No
9	0.5177	0.0015	0.6557	No
10	0.8288	0.0043	0.8395	Yes
11	0.5326	0.0013	0.6682	No

12	0.5696	0.0014	0.6703	No
13	0.5260	0.0015	0.6937	No
14	0.5445	0.0016	0.5276	Yes
15	0.6008	0.0014	0.6723	No
16	0.5666	0.0014	0.6761	No
17	0.5499	0.0012	0.6795	No
18	0.5120	0.0015	0.6671	No
19	0.6416	0.0014	0.7351	No

Spatial Configuration Search Participant	Decision Boundary	Accumulation rate	Distractor Saliency	Outlier
1	0.7613	0.0013	0.7352	No
2	0.7843	0.0014	0.7419	No
3	0.6618	0.0013	0.7079	No
4	0.7788	0.0012	0.7797	No
5	0.5340	0.0012	0.7579	No
6	0.7609	0.0011	0.7613	No
7	0.7340	0.0012	0.7292	No
8	0.7389	0.0013	0.7465	No
9	0.7639	0.0011	0.7918	No
10	0.7017	0.0012	0.7877	No

11	0.6516	0.0014	0.7788	No
12	0.7391	0.0010	0.7551	No
13	0.7196	0.0011	0.7632	No
14	0.7221	0.0011	0.7393	No
15	0.6336	0.0011	0.8028	No
16	0.6639	0.0012	0.7321	No
17	0.6368	0.0013	0.7043	No
18	0.7149	0.0014	0.7519	No
19	0.6088	0.0012	0.8059	No
20	0.7118	0.0014	0.7651	No

Appendix II: Competitive Guided Search (CGS)

In order to fit CGS we followed the same procedure as in SAIM-WTA apart from the fact that DE-MCMC was initialized with Moran et al.'s (2013) parameters rather than randomly.

Feature Search Participant	w_{target}	υ	θ	Δw_{quit}	T_{min}	γ	m	Outlier
1	509.73	0.9012	0.1936	87.46	0.1750	59.309	0.0679	No
2	601.31	0.9948	0.2353	92.20	0.1632	53.809	0.0671	No
3	427.05	1.3736	0.2190	138.58	0.2381	45.042	0.0586	No
4	414.69	0.8981	0.1497	51.52	0.2202	41.927	0.0627	No
5	703.09	0.9442	0.2183	10.79	0.2367	66.394	0.0030	No
6	519.11	0.8869	0.1715	83.47	0.1964	64.822	0.0687	No
7	470.43	1.0780	0.1787	57.55	0.2049	38.983	0.1109	No
8	603.56	0.7128	0.1240	77.35	0.2003	53.991	0.0978	No
9	541.55	0.9225	0.1748	78.00	0.1883	76.430	0.0726	No
10	522.10	0.7007	0.1310	123.69	0.1514	28.384	0.0871	No
11	374.24	1.1795	0.2349	75.31	0.2422	38.389	0.0187	No

12	313.05	0.7203	0.1452	53.64	0.1854	24.450	0.0690	No
13	566.21	0.6860	0.0867	95.44	0.2251	49.145	0.0855	No
14	486.48	0.7606	0.1221	81.15	0.2149	34.367	0.0306	No
15	383.20	0.6314	0.0997	26.61	0.2311	42.630	0.0433	No
16	539.41	0.8274	0.0964	116.18	0.2310	63.428	0.0852	No
17	323.05	0.7722	0.2064	34.15	0.1694	47.416	0.0596	No
18	651.63	0.7536	0.0746	1.12	0.2386	46.928	0.0399	No
19	482.04	1.0000	0.2288	101.74	0.1539	60.497	0.0657	No
Mean	496.42	0.8813	0.1627	72.93	0.2035	49.281	0.0628	

Conjunction Search Participant	w_{target}	δ	θ	Δw_{quit}	T_{min}	γ	m	Outlier
1	2.6567	0.7215	0.0385	0.0296	0.3235	22.987	0.0670	No
2	3.9355	0.7221	0.0385	0.0062	0.3076	21.287	0.1006	No
3	5.3075	0.4667	0.0243	0.0236	0.3804	38.041	0.0606	No
4	3.8513	0.8845	0.0538	0.0886	0.3268	8.6027	0.2885	Yes

5	1.4917	0.5548	0.0287	0.0047	0.4121	7.3934	0.0148	No
6	5.6134	0.6043	0.0414	0.0093	0.3641	15.702	0.0720	No
7	6.2059	0.8420	0.0540	0.0387	0.3214	13.922	0.2933	Yes
8	3.3363	0.6934	0.0436	0.0094	0.3609	10.769	0.1021	No
9	3.3293	0.6954	0.0367	0.0323	0.3248	19.404	0.0334	No
10	0.9580	0.8742	0.0439	0.0590	0.2298	12.375	0.1097	No
11	4.0060	0.5743	0.0337	0.0028	0.3974	11.466	0.0452	No
12	4.9320	0.6380	0.0422	0.0078	0.3572	14.878	0.0547	No
13	4.4002	0.6305	0.0319	0.0242	0.3430	21.556	0.1323	No
14	3.2589	0.6728	0.0362	0.0048	0.3081	24.146	0.0208	No
15	3.9908	0.5219	0.0323	0.0011	0.3907	20.714	0.0419	No
16	4.9043	0.6137	0.0369	0.0024	0.3598	17.587	0.0550	No
17	4.5750	0.6768	0.0459	0.0222	0.3855	12.596	0.0588	No
18	4.3901	0.7781	0.0400	0.0236	0.3163	18.593	0.0940	No
19	2.6121	0.8141	0.0519	0.0128	0.3034	7.3051	0.1020	No
Mean	3.8818	0.6831	0.0397	0.0212	0.3428	16.807	0.0919	

Spatial Configuration Search Participant	w_{target}	υ	θ	Δw_{quit}	T_{min}	γ	m	Outlier
1	0.7808	0.6625	0.0535	0.0002	0.4152	13.623	0.0143	No
2	0.8427	0.4798	0.0383	0.0016	0.3729	13.779	0.0567	No
3	2.4182	0.5085	0.0385	0.0119	0.4176	17.661	0.0993	No
4	0.5615	0.6705	0.0627	0.0021	0.4086	15.636	0.0821	No
5	2.9063	0.4395	0.0232	0.0132	0.4108	13.377	0.1779	No
6	1.3379	0.5003	0.0644	0.0122	0.4648	17.407	0.0370	No
7	0.8685	0.5045	0.0382	0.0002	0.4322	31.391	0.0369	Yes
8	1.9042	0.6199	0.0474	0.0120	0.3735	7.285	0.1433	No
9	1.1208	0.6009	0.0767	0.0279	0.4161	9.546	0.1216	No
10	2.0736	0.6429	0.0778	0.0436	0.4443	17.560	0.1167	No
11	2.6494	0.6500	0.0492	0.0319	0.3630	9.308	0.1728	No
12	1.5317	0.4810	0.0501	0.0211	0.4816	3.984	0.1087	No
13	1.8925	0.6139	0.0776	0.0288	0.4324	7.596	0.0944	No

14	1.3086	0.5378	0.0489	0.0109	0.4229	9.263	0.0551	No
15	3.2230	0.6341	0.0829	0.0324	0.3912	15.537	0.1257	No
16	2.2699	0.4245	0.0418	0.0304	0.4431	14.322	0.0563	No
17	2.3946	0.6768	0.0406	0.0047	0.3850	14.435	0.0559	No
18	0.8305	0.6198	0.0395	0.0040	0.4052	26.181	0.0837	No
19	4.0731	0.7604	0.0932	0.0554	0.3875	18.036	0.3250	Yes
20	2.2644	0.5925	0.0691	0.0225	0.3915	27.515	0.0630	No
Mean	1.8438	0.5659	0.0537	0.0176	0.4142	13.932	0.0917	

z-value, p-value	w_{target}	υ	θ	Δw_{quit}	T_{min}	γ	m
Feature	-0.84, 0.403	**3.15, 0.002**	**3.00, 0.003**	**-3.84, <0.001**	-1.13, 0.258	**-4.18, <0.001**	**4.18, <0.001**
Conjunction	-1.35, 0.176	**3.74, <0.001**	**3.38, <0.001**	**-2.59, 0.010**	-0.89, 0.371	0.85, 0.396	**4.02, <0.001**
Spatial	0.87, 0.383	**4.22, <0.001**	**3.32, <0.001**	-0.35, 0.724	0.92, 0.358	**-4.22, <0.001**	**4.22, <0.001**

Table II.1 Comparison between Moran et al.'s (2013) parameters and our parameters using Wilcoxon sum-rank test. Results in bold font indicate a significant difference.

Appendix III: Asymmetrical Dynamic Neural Network (ADyNeN)

A list of ADyNeN parameters:

1. Target to Distractor weight - TD;
2. Distractor to Distractor weight - DD;
3. Distractor to Target weight - DT;
4. Noise - ζ;
5. Decision Boundary - thr;
6. Sigmoid Slope - s;
7. Sigmoid Shift - d;
8. Accumulation - t;
9. Motor error - m;

TD	DD	DT	ζ	thr	s	d	t	m
0,200	0,150	0,100	0,0.4	0.5,1	0,150	0,0.2	0,0.005	0,0.5

Table III.1: The parameter bounds for ADyNeN model.

Outliers:

1. Feature: 4th, 7th, 19th
2. Conjunction: 4th, 10th
3. Spatial: 2nd, 5th, 8th,

Feature Search	TD	DD	DT	ζ	thr	s	d	t	m
1	124.3903	79.9497	1.069	0.0895	0.5624	35.4516	0.6608	0.0027	0.0212
2	118.3452	64.5135	0.7137	0.1018	0.6666	27.8293	0.6573	0.0027	0.0417
3	112.3337	75.9271	1.9872	0.0465	0.8845	49.311	0.8055	0.0044	0.0402
4	86.4986	0.379	0.2005	0.092	0.6226	37.012	0.6796	0.0029	0.0237
5	82.0935	135.284	0.1323	0.0579	0.6296	46.1509	0.7864	0.0034	0.0043
6	114.1549	58.7303	2.2657	0.0578	0.8662	40.6812	0.7966	0.0044	0.037
7	150.4281	79.0322	4.5986	0.0618	0.8539	38.5298	0.7647	0.0042	0.0616
8	69.6937	27.4476	2.1202	0.0709	0.617	33.6018	0.7405	0.0035	0.0465
9	137.2734	64.2373	3.1389	0.0557	0.8337	42.0765	0.7637	0.0041	0.0254
10	63.4552	84.1684	0.3601	0.0936	0.7135	29.0801	0.8159	0.0047	0.0609
11	144.3466	136.101	0.0288	0.0648	0.7215	50.6529	0.7647	0.0032	0.0214
12	94.3155	103.7935	1.1335	0.0923	0.6408	27.5563	0.7356	0.0033	0.0592
13	96.9523	77.7047	2.5708	0.0569	0.926	32.6229	0.77	0.0048	0.058
14	107.0716	85.1993	0.5197	0.0661	0.5536	36.5859	0.7274	0.0033	0.0091
15	126.5389	75.5599	0.452	0.0673	0.7572	36.3168	0.7997	0.0042	0.0343
16	159.5913	81.0231	0.7746	0.0526	0.8737	58.2801	0.7371	0.004	0.0463
17	83.715	128.0563	1.2521	0.1019	0.6539	26.442	0.8437	0.0041	0.053
18	142.0826	82.7039	0.3158	0.0574	0.8473	60.8336	0.6	0.0028	0.009
19	78.5121	0.3114	0.1514	0.0906	0.6216	36.0105	0.6741	0.0029	0.0248
Mean	111.0221	85.024	1.177	0.070	0.734	39.592	0.7503	0.0037	0.0354

Conjunction Search	*TD*	*DD*	*DT*	*ζ*	*thr*	*s*	*d*	*t*	*m*
1	124.3903	79.9497	1.069	0.0895	0.5624	35.4516	0.6608	0.0027	0.0212
2	118.3452	64.5135	0.7137	0.1018	0.6666	27.8293	0.6573	0.0027	0.0417
3	112.3337	75.9271	1.9872	0.0465	0.8845	49.311	0.8055	0.0044	0.0402
4	86.4986	0.379	0.2005	0.092	0.6226	37.012	0.6796	0.0029	0.0237
5	82.0935	135.284	0.1323	0.0579	0.6296	46.1509	0.7864	0.0034	0.0043
6	114.1549	58.7303	2.2657	0.0578	0.8662	40.6812	0.7966	0.0044	0.037
7	150.4281	79.0322	4.5986	0.0618	0.8539	38.5298	0.7647	0.0042	0.0616
8	69.6937	27.4476	2.1202	0.0709	0.617	33.6018	0.7405	0.0035	0.0465
9	137.2734	64.2373	3.1389	0.0557	0.8337	42.0765	0.7637	0.0041	0.0254
10	63.4552	84.1684	0.3601	0.0936	0.7135	29.0801	0.8159	0.0047	0.0609
11	144.3466	136.101	0.0288	0.0648	0.7215	50.6529	0.7647	0.0032	0.0214
12	94.3155	103.7935	1.1335	0.0923	0.6408	27.5563	0.7356	0.0033	0.0592
13	96.9523	77.7047	2.5708	0.0569	0.926	32.6229	0.77	0.0048	0.058
14	107.0716	85.1993	0.5197	0.0661	0.5536	36.5859	0.7274	0.0033	0.0091
15	126.5389	75.5599	0.452	0.0673	0.7572	36.3168	0.7997	0.0042	0.0343
16	159.5913	81.0231	0.7746	0.0526	0.8737	58.2801	0.7371	0.004	0.0463
17	83.715	128.0563	1.2521	0.1019	0.6539	26.442	0.8437	0.0041	0.053
18	142.0826	82.7039	0.3158	0.0574	0.8473	60.8336	0.6	0.0028	0.009
19	78.5121	0.3114	0.1514	0.0906	0.6216	36.0105	0.6741	0.0029	0.0248
Mean	116.4638	28.2972	22.193	0.0710	0.788	22.450	0.5499	0.0020	0.0299

Spatial Configuration Search	TD	DD	DT	ζ	thr	s	d	t	m
1	65.4658	24.6815	23.2552	0.1553	0.5616	6.4034	0.5961	0.0018	0.0265
2	89.2456	41.389	37.0931	0.168	0.7325	6.8741	0.5292	0.0019	0.0639
3	66.9934	52.2721	43.158	0.1125	0.8534	9.7997	0.4118	0.0021	0.0303
4	71.7787	43.3261	38.8835	0.1919	0.7584	5.1527	0.5873	0.0026	0.0353
5	189.579	23.9195	20.9947	0.0896	0.5815	21.9964	0.7442	0.0025	0.0315
6	83.4007	41.6787	37.6695	0.2081	0.8529	5.4695	0.4529	0.002	0.0327
7	83.4981	32.8276	31.2112	0.1342	0.6934	7.9076	0.5567	0.0019	0.0211
8	78.1335	29.6979	29.405	0.1016	0.6883	8.683	0.8179	0.0038	0.055
9	84.1313	29.2898	28.8115	0.155	0.6606	7.187	0.6184	0.0019	0.0333
10	70.3223	27.0412	26.5685	0.102	0.7035	10.0902	0.7228	0.0028	0.0484
11	107.828	24.9001	21.8516	0.1312	0.5539	14.0163	0.5666	0.0015	0.0252
12	83.0408	46.8087	41.3161	0.2025	0.8622	5.5766	0.4649	0.0021	0.0421
13	77.6819	32.421	29.8364	0.1689	0.6969	6.9992	0.5781	0.0018	0.0315
14	74.8653	30.0233	27.4714	0.1679	0.6444	6.1196	0.6336	0.0021	0.0394
15	96.7771	36.6325	32.0488	0.1319	0.7813	9.9764	0.699	0.0027	0.0562
16	100.656	38.1026	31.7875	0.1344	0.7611	9.2423	0.5926	0.0019	0.0295
17	104.948	28.8846	25.4629	0.0915	0.6572	13.547	0.6412	0.0023	0.0297
18	67.596	44.3735	40.8627	0.0934	0.8424	8.7495	0.5549	0.0031	0.0444
19	97.8113	22.471	22.0493	0.1305	0.5957	12.7919	0.6627	0.0019	0.0377
20	104.523	29.2569	30.1452	0.1566	0.6706	8.3339	0.6169	0.0022	0.0339
Mean	84.7835	34.4112	31.3170	0.1451	0.71468	8.668	0.5856	0.0021	0.0351

Appendix IV: Competitive Guided Search (CGS) 2nd Study

Outliers:

1. Feature: 18th and 19th
2. Conjunction: 1st, 4th, 7th
3. Spatial: 5th, 11th, 15th, 19th, 20th

Feature Search	w_{target}	υ	θ	Δw_{quit}	$T_{present}$	T_{absent}	γ	m
1	597.9089	1.0603	0.2094	409.3828	0.1704	0.1696	31.8507	0.0379
2	630.8515	1.1033	0.2764	429.6111	0.1352	0.1428	41.9189	0.0449
3	1020.412	0.86	0.1304	399.054	0.2606	0.2484	56.1228	0.0698
4	366.9662	0.9234	0.1184	325.5649	0.241	0.1617	28.4238	0.0148
5	347.0001	1.2604	0.2744	405.3118	0.2296	0.2286	39.2696	0.0085
6	600.5712	1.0433	0.1739	413.7916	0.1965	0.2086	25.7711	0.0462
7	991.651	0.8733	0.1471	391.3397	0.218	0.2256	63.7069	0.0619
8	606.7712	1.0765	0.1856	409.3845	0.1924	0.1911	30.7785	0.0457
9	600.054	1.1118	0.1572	415.1492	0.2084	0.2176	25.835	0.0336
10	584.4948	1.1036	0.227	406.913	0.1435	0.1446	39.0397	0.0415
11	318.4201	1.3898	0.3014	468.0363	0.2249	0.207	39.801	0.023
12	598.2879	1.1749	0.2665	409.486	0.1495	0.1672	41.3642	0.0539
13	1054.131	1.147	0.2352	428.6241	0.1692	0.2088	65.1582	0.0629

14	841.2155	1.0925	0.1719	431.0299	0.2119	0.2438	55.4448	0.0184
15	600.1325	1.0462	0.1825	409.8216	0.1992	0.2004	28.6878	0.0463
16	831.305	1.1305	0.1333	373.8617	0.2204	0.217	48.6937	0.0471
17	704.8592	0.772	0.1933	341.0633	0.1685	0.171	49.0647	0.0429
18	369.7622	0.5622	0.029	231.244	0.2737	0.263	39.8363	0.0194
19	393.1467	0.9783	0.1337	311.3181	0.2303	0.1836	39.8196	0.0434
Mean	664.4136	1.0687	0.1990	403.9662	0.1964	0.1972	41.8194	0.04113

Conjunction Search	w_{target}	υ	θ	Δw_{quit}	$T_{present}$	T_{absent}	γ	m
1	52.5353	1.1506	0.3696	56.3225	0.1614	0.2167	71.0159	0.0192
2	5.2457	0.7163	0.0483	0.8014	0.2967	0.3019	21.6476	0.1149
3	3.8363	0.489	0.0263	1.7863	0.384	0.3687	15.844	0.0539
4	65.007	0.4914	0.1613	125.533	0.1829	0.0533	12.2833	0.1558
5	1.7965	0.5189	0.0357	0.0504	0.4083	0.3508	12.8873	0.0019
6	4.2975	0.6242	0.0461	1.2103	0.3621	0.3643	18.5246	0.0248
7	23.8112	2.34	0.8519	265.8012	0.1257	0.1151	70.7832	0.0219
8	3.8356	0.5054	0.0346	0.3282	0.378	0.3483	18.7242	0.0575
9	4.1196	0.694	0.052	0.6805	0.3321	0.3346	24.3372	0.0415
10	2.9134	0.6236	0.0455	0.6693	0.2378	0.2774	11.1925	0.1066
11	4.5345	0.3237	0.0164	0.3247	0.4145	0.4364	12.482	0.0614
12	5.2258	0.4464	0.0258	0.5925	0.3786	0.3252	11.1809	0.0402
13	6.7002	0.8888	0.0452	0.9433	0.3543	0.3731	31.2117	0.0868
14	4.7236	0.6456	0.044	1.7929	0.3126	0.3578	19.8347	0.0019
15	4.4064	0.3743	0.0212	0.4451	0.4023	0.3504	16.3568	0.0418
16	4.4924	0.6649	0.0488	1.2718	0.3546	0.3565	17.0843	0.0406
17	4.5279	0.2466	0.0144	0.4823	0.453	0.3906	15.2474	0.0893
18	6.4921	0.5757	0.0358	1.1877	0.3308	0.3218	36.4661	0.0394
19	2.869	0.686	0.0495	0.4627	0.3532	0.2627	13.3171	0.0864
Mean	4.3760	0.5639	0.0368	0.8143	0.3595	0.3450	18.5211	0.0555

Spatial Configuration Search	w_{target}	υ	θ	Δw_{quit}	$T_{present}$	T_{absent}	γ	m
1	0.9297	0.6967	0.0713	0.0024	0.4028	0.3788	19.9092	0.013
2	2.1928	0.8838	0.1042	0.0441	0.4097	0.4074	40.5937	0.0417
3	1.4702	0.5267	0.0293	0.0195	0.4125	0.336	16.9954	0.0437
4	0.9402	0.5367	0.0517	0.0096	0.4602	0.4159	21.5886	0.036
5	55.5617	1.2818	0.5128	43.1337	0.2195	0.1561	77.7109	0.0607
6	1.0644	0.5017	0.0569	0.0246	0.4519	0.4469	9.0538	0.011
7	1.2943	0.4488	0.0283	0.0026	0.4337	0.4046	10.1367	0.0352
8	1.2943	0.4488	0.0283	0.0026	0.4337	0.4046	10.1367	0.0352
9	1.0255	0.5555	0.0613	0.0317	0.444	0.3324	9.0722	0.0561
10	0.4439	0.6108	0.0424	0.0162	0.4558	0.3802	28.5303	0.0583
11	3.1182	0.5596	0.0464	0.4081	0.3757	0.3178	10.2446	0.0528
12	0.9459	0.5037	0.0406	0.0149	0.4953	0.4749	5.5664	0.0378
13	1.7779	0.5445	0.0487	0.0688	0.4696	0.4264	7.2598	0.0451
14	1.3536	0.4296	0.046	0.0173	0.4242	0.3983	12.0196	0.0369
15	14.5046	0.9167	0.5055	11.9088	0.1725	0.198	62.2988	0.0458
16	1.2943	0.4488	0.0283	0.0026	0.4337	0.4046	10.1367	0.0352
17	2.471	0.8632	0.0545	0.1096	0.3858	0.3057	23.3617	0.0277
18	1.0814	0.6347	0.0407	0.0269	0.4053	0.3405	19.2556	0.0289
19	10.8925	0.7959	0.31	17.2083	0.2329	0.2375	58.0916	0.0136
20	8.3673	1.1836	0.388	3.4779	0.2392	0.2972	66.7059	0.023
Mean	1.3052	0.5755	0.0488	0.0262	0.4345	0.3904	16.2410	0.0361