Why less can be more: A Bayesian Framework for Heuristics

Paula Parpart

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Experimental Psychology

University College London

October 9, 2017

I, Paula Parpart, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

When making decisions under uncertainty, one common view is that people rely on simple heuristics that deliberately ignore information. One of the greatest puzzles in cognitive science concerns why heuristics can sometimes outperform full-information models, such as linear regression, which make full use of the available information. In this thesis, I will contribute the novel idea that heuristics can be thought of as embodying extreme Bayesian priors. Thereby, an explanation for less-is-more is that the heuristics' relative simplicity and inflexibility amounts to a strong inductive bias, that is suitable for some learning and decision problems. I will formalize this idea by introducing Bayesian models within which heuristics are an extreme case along a continuum of model flexibility defined by the strength and nature of the prior. Crucially, the Bayesian models include heuristics at one of the Bayesian prior strength and classic full-information models at the other end of the Bayesian prior. This allows for a comparative test between the intermediate models along the continuum and the extremes of heuristics and full regression model. Indeed, I will show that intermediate models perform best across simulations, suggesting that down-weighting information is preferable to entirely ignoring it. These results refute an absolute version of less-is-more, demonstrating that heuristics will usually be outperformed by a model that takes into account the full information but weighs it appropriately.

Thereby, the thesis provides a novel explanation for less-is-more: Heuristics work well because they embody a Bayesian prior that approximates the optimal prior. While the main contribution is formal, the final Chapter will explore whether less is more at the psychological level, and finds that people do not use heuristics, but rely on the full information instead. A consistent perspective will emerge throughout the whole thesis, which is that *less is not more*.

Acknowledgements

I would like to acknowledge and thank all the people who made this thesis possible. The research would not have been possible without the continuing support of UCL’s warm and inspiring research environment and my collaborators and colleagues, the peers gained from the MSc Cognitive Decision Science course, and the people I worked with at the ABC research group during my time at the Max Planck Institute for Human Development in Berlin.

Thanks to my boyfriend Neil Bramley for being there the whole part of the journey, for his love and putting up with me while always being responsive to questions when needed, for providing personal support but also academically inspiring research ideas. I want to thank David Lagnado, Maarten Speekenbrink, Adam Haaris, Takao Noguchi and Anne Hsu for making UCL a great place to work and for supporting me throughout my PhD and time at UCL, and the inspiring exchange of research ideas. Thanks to Eric Schulz for being a great friend, amazing statistician and collaborator. I want to thank my supervisor Bradley Love for the many great meetings on grand ideas, for allowing me to run studies in the lab, and the many invaluable life lessons learned. Thanks to my MSc supervisor Anne Hsu for teaching me Bayesian modelling. Thanks to the people at the ABC research group in Berlin, particularly my former supervisor Edward Cokely for inspiring me to start a research career and moving to London, to Thorsten Pachur, Julian Marewski, Leal Schooler, Mirta Galesic, Gerd Gigerenzer, Bjoern Meder, Jonathan Nelson, Bettina van Helversen, and all those people who supported and motivated me from very early on to do research. Thanks to my collaborator Matt Jones for help with math derivations.

A big thanks to Prof Philip Treleaven, Yonita Carter and the CDT for Financial Computing & Analytics for funding me with an EPSRC grant, and giving me the opportunity to grow beyond a cognitive scientist. Without the support of the Financial Computing Centre I would not have been able to do an MRes and had the chance to take courses in machine learning, or discover my passion for business and startups. Thanks to the Master students of the Msc Cognitive and Decision Sciences as well as the undergraduate students who I have had the pleasure to teach in the last year as Teaching Fellow on the CoDes course, and who have made the experience at UCL more rewarding and interesting, despite the double workload. Thanks for giving me the opportunity to lecture and grow as an academic, and the positive feedback on research supervision.

I also wish to express gratitude to my lab mates, colleagues and friends at UCL and around London who have transformed the experience: Thanks to Saoirse Connor Desai for being a great friend and listener, to Chloe Bukata and Will Crichton for hosting many great game nights, dinners and enduring long philosophical discussions, to Laura Weis, Kasia Borowska, Sonia IcedTea, Danny Onions, Anna Hessenface, Christiane Ahlheim, Lara Kirfel, Moritz Haller, Peter Riefer, Tobias Gerstenberg, Christos Bechlivanidis, Hannah Tickle, Leo Cohen, George Deane, John Draper, Sam Solomon, Neele Engelmann, Tom Hardwicke, Nigel Harvey, Pia Horbacki, Manos Konstantinidis, Jens Koed Madsen, Katie Parker, Simone Malejka, Laura de Moliere, Phil Newell, Keith O’Brien, Toby Pilditch, David Shanks, Sarah Smith, Paul Troop, Sebastian Bobadilla, Amna Ali, Sergio Barbosa, Kurt Braunlich, and anyone I might have forgotten. A special thanks also goes to my sister Anne Parpart and my granddad Gerhard Dietrich.

Finally I wish to thank my parents, Jens and Margret, for their continuous love and the ”safety net” they provide at all times.

Contents

List of Figures

List of Tables

The Chapters in this thesis are based on the following articles:

Chapter 4 and Chapter 5

Parpart, P., Jones, M. & Love, B.C. (under review). Heuristics as Bayesian inference under extreme priors. Proceedings of the National Academy of Sciences (PNAS)

Chapter 6

Parpart, P., Schulz, E., Speekenbrink, M. & Love, B. (2015). Active learning as a means to distinguish among prominent decision strategies. Proceedings of the 37th Annual Conference of the Cognitive Science Society.

Parpart, P., Schulz, E. & Speekenbrink, M. (submitted). Model-based active learning. Psychonomic Bulletin & Review

Chapter 1

Introduction

> *For every complex problem there is an answer that is clear, simple, and wrong.*
>
> \- H. L. MENCKEN

Making decisions under uncertainty is central to everyday life. Every time we buy a cup of coffee, decide what turn to take at the next intersection when driving, or where to submit a paper, there are potentially thousands of cues that could play into the decision, but we do not usually have time nor cognitive resources to use them all. Hence, when making decisions under uncertainty, such as choosing which apartment to rent, one common view is that people rely on *heuristic* algorithms, which deliberately ignore parts of the available information. For example, instead of considering all available information sources such as proximity to work, proximity to schools, crime rates, neighbourhood sport facilities or market trends, the *Take-The-Best* heuristic (Gigerenzer & Goldstein, 1996) would just rely on the first most important cue that is able to discriminates among the flats, and ignore all other cues. For example, if the most important cue was the proximity to work, the Take-The-Best heuristic would decide for the flat that is closer to work. Or else, the *tallying* heuristic would simply tally which flat is better on each cue and choose the flat that has more cues in its favour. Heuristics are often regarded as plausible decision strategies because they do not make full use of the input data and rely on a set of simple rules. In contrast, *full-information* decision models make full use of the available information, taking into account things like covariance among information sources and differential weighting. One of the main unanswered questions is still why heuristics can sometimes outperform full-information models, such as linear regression. These paradoxical findings are called *less-is-more* effects in the literature, and have been repeatedly demonstrated in both artificial and real-world prediction tasks (Chater, Oaksford, Nakisa, & Redington, 2003; Gigerenzer & Brighton, 2009; Gigerenzer, Todd, & Group, 1999).

The existence of less-is-more represents one of the most central debates in the history of judgement and decision making, such as indicated by the debate between the heuristics-and-biases program (Tversky & Kahneman, 1974), and the fast-and-frugal heuristics program (Gigerenzer et al., 1999). The state-of-the-art explanation for less-is-more effects in the field relies on the statistical bias-variance concept (which will be explained in detail in the next Chapter), and proposes that heuristics can excel as a result of lower overfitting rates due to smaller number of parameters (Gigerenzer & Brighton, 2009), i.e., overfitting happens when models are too sensitive to variability in the samples and capture noise, which hurts at generalizing to new data. However, the bias-variance concept alone does not provide a formal computational model which is able to make testable predictions about when and why heuristics or full-information algorithms will perform best. Furthermore, less-is-more effects are very volatile to environmental conditions, and even sometimes reversible as will be seen in Chapter 3. Secondly, the bias-variance concept lacks any formal link between full-information models and heuristics.

In this thesis, I will contribute the idea that heuristics can be thought of as embodying extreme Bayesian priors. Thereby, an explanation for less-is-more is that the heuristics' relative simplicity and inflexibility amounts to a strong inductive bias, that is suitable for some learning and decision problems. In the main body of the thesis I will formalize this idea by introducing Bayesian models wherein heuristics are an extreme case along a continuum of model flexibility defined by the strength of the prior. Crucially, the Bayesian models include heuristics at one extreme end of the Bayesian prior's strength and classic full-information models (such as linear regression) at the other end of the Bayesian prior. This is achieved with regularization methods that are conceptually related to ridge regression from machine learning (Hoerl & Kennard, 1970). Our models' regularization methods contain a penalty term that adjusts model flexibility with a single parameter. Importantly, this penalty term is equivalent to a Bayesian prior on the weights, and the parametric variation of the penalty parameter corresponds to a parametric variation of the Bayesian prior's strength. In the limit of an extremely strong prior, the Bayesian models are equivalent to the heuristics. Thus, this framework provides a formal characterization of the link between traditional statistical models (OLS) and heuristics, which are usually only contrasted. The formal continuum allows for a comparative test including the intermediate prior settings which lie along the continuum between the extremes of heuristics (which entirely ignore some information) and the full regression model (which differentially weights and includes all cues). A crucial difference between the heuristics (in the limit) and the intermediate prior settings along the continuum is that the intermediate models are fully sensitive to the input data such as covariance or cue weight magnitudes. Indeed, I will show that intermediate parametrisations perform best across all simulations in this thesis, suggesting that down-weighting information is preferable to entirely ignoring it. These results refute an absolute version of less-is-more, demonstrating that heuristics will usually be outperformed by an intermediate model that takes into account the full information but weighs it appropriately. Thereby, a novel explanation for less-is-more emerges, suggesting heuristics may perform well because they approximate intermediate models with the optimal prior. This suggests the optimal setting for many familiar situations is often close to the heuristic end of the continuum.

In parallel to the insights into statistical less-is-more effects, the thesis will suggest a new perspective on how Bayesian models relate to heuristics. In contrast to the opposing relationship between Bayesian models and heuristics in cognitive science and behavioural economics - where probabilistic inference models and heuristic models are seen as competitors and pitted against each other (Katsikopoulos, Schooler, & Hertwig, 2010; Martignon & Hoffrage, 2002; Tversky & Kahneman, 1974) - the thesis shows that heuristics are part of Bayesian inference (for an extremely strong prior). Since the nineties, scientists have argued that the impressive less-is-more findings were in dire need of a rational explanation (Chater et al., 2003; Gigerenzer & Goldstein, 1996) - however instead, the fast-and-frugal heuristics approach relied on *ecological rationality* to explain their findings, and is largely yet to address where this fits into a broader framework of rational decision making. Instead, this thesis provides a Bayesian explanation for less-is-more, suggesting a unification for Bayesian inference models and heuristics. The probabilistic formalization puts heuristics on the same playing field as other full-information models.

Over the course of the thesis, a consistent perspective will emerge. Each Chapter will contribute to the perspective that *less is not more*. Importantly, the main contribution of the thesis is formal, relying on a series of derivations and computational studies. However, the final Chapter will use behavioural experimentation combined with modelling to explore what the information-gathering behaviour of people can tell us about their use and representation of information for decision-making. The question in the final Chapter will be: Do people use a fast-and-frugal heuristic (Take-The-Best) or a full-information model (logistic regression)? In that way, the psychological Chapter will look at a different kind of less-is-more effect, i.e., whether people fully and systematically ignore information in the input data as proposed by the fast-and-frugal heuristics (Gigerenzer & Brighton, 2009) (i.e., a descriptive *psychological* less-is-more effect). While the Bayesian inference Chapters 4 and 5 come to the conclusion that less is not more in the sense that heuristics can always be outperformed with a model that uses all information (i.e., refuting *absolute* less-is-more), the psychological Chapter 6 concludes that people do not fully ignore presented information but are much more adaptive to the full information presented, along the lines of full-information models. In that way, all Chapters conclude on a similar note regarding less-is-more, however on different levels of analysis. Chapter 3 will clarify the differences among less-is-more definitions.

The structure of the thesis is as follows: In Chapter 2, I will introduce heuristics, and contrast probabilistic approaches with heuristic approaches to cognition. I will develop the two most prominent approaches to heuristics in decision making, i.e., the *heuristics-and-biases* account (Tversky & Kahneman, 1974), and the *fast-and-frugal heuristics* account (Gigerenzer et al., 1999), and provide definitions and terminologies. Most importantly, less-is-more effects will be introduced and the most common explanation in the literature will be given. I will point out its short-comings that the thesis will advance.

Chapter 3 will critically analyse previous less-is-more effects in the literature. I will identify three factors in the statistical method and the statistical environment that lead to less-is-more, however I find that these same factors can also be used to make them disappear. In 4 computational studies, I will show that when these factors are reversed, often the less-is-more effects disappear. The Chapter will highlight the volatility of less-is-more and the limitations of the bias-variance concept as an explanation. The chapter will end by highlighting the need for a formal model that can account for *why* less is more.

Chapter 4 will formalize the idea that heuristics represent extreme Bayesian priors and provide a novel explanation for less-is-more. I will develop the first computational Bayesian inference model for the tallying heuristic (Dawes, 1979). In this model, parametric variation of a prior's strength generates a continuum of models, with a variant of linear regression at one extreme and the tallying heuristic at the other extreme. Although the Bayesian model can mimic tallying perfectly, a crucial difference is that the Bayesian account regulates weights, but never discards any information. In a computational study of real-world prediction tasks, I will show that novel intermediate models usually perform best across a wide range of real-world environments. A novel interpretation of why heuristics work will be discussed: Heuristics may excel because they approximate the intermediate models, which have the optimal prior for the environment. Finally, this Chapter will attempt to derive the TTB heuristic as an extreme Bayesian prior from a different kind of regularization method in machine learning, i.e., *lasso regression* (Ripley, 2007). Lastly, psychological implications of the model findings will be discussed.

Chapter 5 will develop the second Bayesian account that formally relates both the tallying and Take-The-Best heuristic to ordinary linear regression, by relying on a prior that focuses on sensitivity to covariation among predictors (Rieskamp & Dieckmann, 2012). This relates to the fact that linear regression is fully sensitive to covariance whereas heuristics completely ignore it. Parametric variation of the covariance prior's strength results in a continuum that contains, as limiting cases, both heuristics (TTB and tallying heuristic) as well as (ordinary least-squares) linear regression. In a computational study of real-world environments, I will show that, along the continuum, the best-performing models for the real-world datasets tested are novel intermediate models that do not entirely ignore covariance among predictors, but that nonetheless down-weight this information via the influence of their priors. Again, Chapter 5 will conclude that less is not more. The theoretical and psychological implications will be discussed, as well as the model's limitations.

Finally, Chapter 6 will look at people's representations and decision making processes, asking the question: To what extent do people fully ignore information (i.e., such as the weights by the TTB heuristic) or try to incorporate it into their decision? It does so by looking at people's behaviour in an active learning task as a window on their decision models. Thereby I propose a new model selection method for psychological models based on active learning. The active learning method is based on the assumption that an agent's information gathering behaviour reflects how they represent and go on to use that information in decisions. Chapter 6 will contrast computational active learning algorithms for two model classes that differ in how they represent the decision process and value the information: One is based on a heuristic (Take-The-Best), and the other one is based on a full-information model (logistic regression). In one active learning experiment with both a learning and a test phase, I will ask the question whether people's active learning behaviour is better described by the goal of learning the cue weights, or the goal of identifying a cue rank order among cues for usage with Take-the-best. Interestingly, I find that both while gathering information and while using it to make decisions, people are more consistent with the full-information representation. These findings are consistent with the hypotheses arising from Chapter 5 and Chapter 4, proposing that potentially people's psychological processing may represent multi-attribute decision problems in a much more ’weighted additive’ way than the heuristics literature would have us believe. Interestingly, I also find that people who appeared to use a Take-The-Best heuristic when making decisions often behaved as if they were learning cue weights while gathering information. Thereby, the final Chapter concludes on the same note as previous Chapters - less is not more - however, on a psychological processing level. This final Chapter brings together learning theories and decision making theories.

In Chapter 7, I bring all these chapters together and construct a cohesive picture of less-is-more. I will propose that a revised understanding of heuristics may be required, and discuss the implications for other disciplines such as neuroscience, psychology, behavioural economics, computer science and machine learning.

The following describes the core contribution of the thesis in an umbrella statement for the entire document: *Less-is-more is observed for comparing simple and complex models (e.g., Take-The-Best and regression), but less-is-more is not true in that one can always do better by including all information rather than throwing it out. That is, one can always do better by including the information and down-weighting it instead. This is established in a Bayesian framework. Heuristics work not because they throw out information, but because they embody a prior that approximates the optimal prior. Although this statement does not address psychological processing, it is relevant to directing research on what people may actually do.*

Chapter 2

Heuristics and Less-is-More

”Less is more.”
- ROBERT BROWNING

”Less is more.”
-LUDWIG MIES VAN DER ROHE

2.1 A brief history of heuristics

Heuristics are simple decision algorithms that deliberately ignore information. The origin of the word *heuristic* is Greek and means ”serving to find out or discover”. Mathematicians such as George Polya saw heuristics as separate from analytical methods. Heuristics were used to find a proof, whereas analysis was used for checking a proof (Groner, Groner, & Bischof, 1983). Einstein included the term heuristic in the title of his Nobel prize-winning paper from 1905 on quantum physics, indicating that the heuristic view he presented was incomplete but highly useful (Holton, 1988, pp. 360 361). Around the same time, Gestalt psychologists including Max Wertheimer saw heuristics as methods for looking around and guiding search for information.

In the field of artificial intelligence (AI), the notion of heuristics as indispensable search strategies emerged. For example, Pearl's book titled *Heuristics: Intelligent search strategies for computer problem solving* (Pearl, 1984) discussed many AI problems that involve such large search spaces that heuristics are required to reduce them to a manageable size given the system's limited resources of time and space (Groner et al., 1983). Similarly, in computer science, until today heuristics are used to solve NP-complete, i.e., computationally intractable, problems by finding approximate solutions when classic methods such as logic and probability theory fail to find exact solutions. Yet, this does not suggest heuristics are similar to other approximate solutions such as Gibbs sampling or particle filters (i.e., Monte Carlo algorithms that approximate a probability distribution by sampling a set of samples from that distribution) (Doucet, De Freitas, & Gordon, 2001; Geman & Geman, 1984; Gilks, Richardson, & Spiegelhalter, 1996). In comparison to these approximate methods, heuristics are assumed to be much simpler, taking into account limited processing resources. Recent work in cognitive science suggests a heuristic-like strategy is more akin to a single-particle particle filter (Sanborn, Griffiths, & Navarro, 2010) relying on fewer samples. Furthermore, it is important to note that despite some viewing heuristics as approximate solutions, heuristics are not necessarily suboptimal or incorrect, and one can also find situations where more complex models, such as those using more samples, produce inferior solutions to the heuristics.

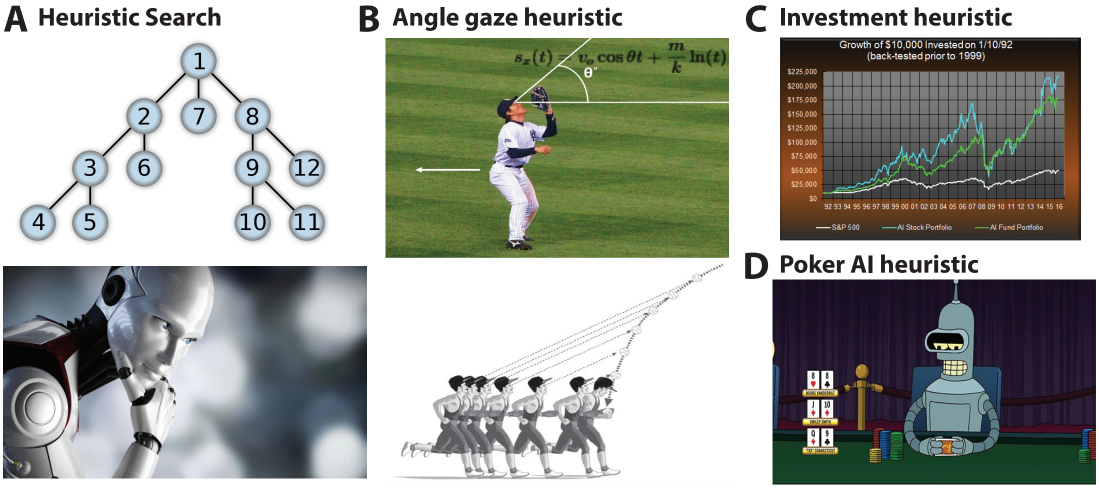


Figure 2.1: **Image**: Areas where heuristics are applied. (**A**): Heuristic tree search: Heuristics are a common type of metric in AI that estimate how far away from the goal state a particular state is, without guaranteeing to be perfect. (**B**): The angle gaze heuristic is used for catching a ball in the air, based solely on the optical angle between the ball and the catcher. (**C**): The simple 1/N rule allocates financial resources equally and could outperform portfolio optimization models. (**D**): AI poker agents rely on the same simple heuristic strategies as expert human poker players, and could not improve upon these heuristics.

Heuristics play an integral role in psychology and particularly decision making. Consider the following example: Imagine you were playing baseball and a friend prompts you to catch a ball that is already high up in the air (Fig. 2.1*B*). You might immediately start running towards the ball to make sure you arrive in the right spot for a catch. How are people able to solve this complex problem? Underlying it are several complex differential equations to predict the trajectory of the ball, as well as a multitude of additional physical variables such as wind resistance. Psychologists suggest people rely on a simple heuristic instead (Gigerenzer, 2007). McLead and Dienes (1996) let baseball fielders catch balls projected towards them from a blowing machine at 45 m distance. They noticed that the fielders were running back and forth at a speed that kept the optical angle between the ball and themselves constant. Paying attention to only this one piece of information (i.e., the constant angle) results in a very simple algorithm which does not tell the fielders where or when the ball will land, however, it ensures that they run through the place where the ball drops at the precise moment that the ball arrives there. This simple strategy also automatically results in interception of the ball irrespective of the effect of wind resistance on the trajectory. McLeod and Dienes (1996) called this simple algorithm the *gaze heuristic*.

The above research presents a classic example where scientists proposed that, instead of solving a set of differential equations, people rely on a set of simple heuristics which follow simple rules and which only pay attention to few pieces of information while deliberately ignoring the rest (Gigerenzer, 2007). Hence, in psychology, a common view is that people use heuristics because our cognitive capacities do not allow for complex strategies that take into account all possible variates and optimal probabilistic computations, as people usually do not have the time, knowledge, or capacity (Simon, 1990). In contrast, a heuristic ignoring data makes the calculation easier and thus may be more compatible with inherent cognitive limitations (Gigerenzer et al., 1999; Tversky & Kahneman, 1974). The appeal of heuristics as theories of human cognition stems from their simplicity, being easy to grasp and easy to model, as they do not have any parameters that need to be optimized.

One influential school of thought on heuristics interprets people's use of heuristics as suboptimal and flawed. The *heuristics and biases* program identified multiple situations in which people's reliance on a heuristic lead to reasoning *biases* - that is, systematic violations of the probabilistic axioms and a set of logical axioms (Tversky & Kahneman, 1974). For example, they found that people ignore the base rates in making probability judgements (Tversky & Kahneman, 1983). Tversky and Kahneman (1974) interpreted these deviations as irrational behaviour and consequentially heuristics became predominantly associated with suboptimal cognitive algorithms in psychology, behavioural economics and related fields since the 1970s. In contrast, the *bounded rationality* approach looked at how people reason when the conditions for rationality as postulated by neoclassical economics were not met (Simon, 1990; Simon et al., 1989). Instead, heuristics were defined as cognitive satisficing mechanisms that allow people to make reasonably accurate predictions despite the limitations of the human mind, while taking into account the structure of the environment. The second main influential approach to heuristics that developed out of this perspective is the *fast and frugal heuristics* program (Gigerenzer et al., 1999), which posits *ecological rationality* instead. In this ecological rationality framework, the success of cognitive algorithms is always judged in conjunction with the environments in which they succeed. Rather than focus on human failings, this program catalogs various cases in which humans excel by using simple heuristics. Yet, the most important contribution of this program were statistical *less-is-more* effects (Czerlinski et al., 1999; Dawes, 1979; Gigerenzer & Brighton, 2009), whereby heuristics are able to outperform complex full-information models in real-world prediction tasks and simulations. For example, the famous city size task involved a heuristic matching multiple linear regression in performance at predicting which of two German cities has the larger population size based on a set of cues (Gigerenzer & Goldstein, 1996). This finding became very influential in the scientific community as it was not expected that a simpler algorithm could match or outperform a more complex one. However, despite this approach receiving great popularity, it also received criticism for not being compatible with rational analysis and other forms of rationality such as Bayesian rationality (Anderson, 1990; Chater et al., 2003). For example, the city size task was taken as evidence by the fast-and-frugal proponents that classical rational norms such as probability theory could be replaced with ecological rationality. However, instead of the fast-and-frugal approach entirely dispensing with rational explanation of the less-is-more effect, a search for a rational explanation could have been launched. Yet, this unification of rational models and heuristics never happened, and it appears that the ecological rationality program could still be developed further with Bayesian rationality.

Understanding why less can be more still represents an unsolved question in cognitive science: How can a simple heuristic that relies on less complex calculations and less information outperform a more complex model that takes into account the full information?

Examples of these *less-is-more* effects in the real world are given next. For example, considering the immense uncertainty in financial markets, how do people decide where to invest their financial resources? Bernatzi and Thaler (2001) noticed that some individuals, when deciding how to allocate financial resources among N options, relied on a simple rule of 1/N. The 1/N rule (Benartzi & Thaler, 2001) allocates financial resources equally across all alternatives (sometimes called naive diversification heuristic (Benartzi & Thaler, 2001)). Interestingly, when the 1/N rule was compared to 14 optimizing models in predicting stock performance across multiple investment problems, including the Nobel Prizewinning Markowitz's mean-variance portfolio model, the 1/N rule's prediction performance was surprisingly high compared to all 14 optimization models (DeMiguel, Garlappi, & Uppal, 2009). All strategies had to iteratively make predictions for the next month's stock performance. While the optimization models were trained on data of 10 years of stock data to estimate the models' parameters, the 1/N rule does not learn anything from the data as it does not have any free parameters. The authors found that, surprisingly none of the optimization models were consistently better than the 1/N rule in terms of Sharpe ratio, certainty-equivalent return, or turnover. They concluded that in order for the classic Markowitz's mean-variance model to beat the simple 1/N heuristic, it would need access to data of around 6000 months for a portfolio with 50 assets. These results suggest that the heuristic could reach an equivalent performance level to more complex models with much less data and shorter time. The question that evolves from this research is *why was less more?*

Consider another example. Recent research shows remarkable results in the area of poker AI. Bowling, Burch, Johanson, and Tammelin (2015) recently solved heads-up limit hold'em poker which is the simplest form of poker, however it is still part of a family of games that exhibit imperfect information where players do not have full knowledge of past events. A novel insight is that both skilled poker players and optimal AI agents are in agreement on a set of very simple opening heuristics. Research already established that skilled poker players use strategic heuristics (Newall, 2011, 2013), but interestingly, so do unboundedly optimal AI poker agents (Bowling et al., 2015) and they seem to not be able to improve upon these simple heuristics. It is remarkable that an AI poker agent, capable of maximizing a strategy of arbitrary complexity, could not improve upon a heuristic strategy (Bowling et al., 2015). Again, the question that this research raises is *Why could a simple heuristic not be improved upon*?

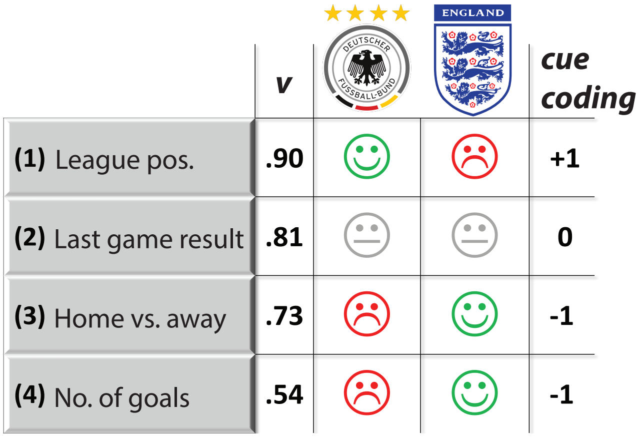
In sum, why less is sometimes more still represents one of the greatest puzzles in judgement and decision making. The fast-and-frugal heuristics program provided one possible explanation based on the statistical *bias-variance* concept, which will be introduced below. While this approach is invaluable for our understanding, it is only limited as it does not formalize the problem: When and why do heuristics succeed? Furthermore, less-is-more effects can be very volatile (Chapter 3) and can easily be made disappear with changing conditions in the environment, emphasizing yet again the need for a formal model.

2.2 Heuristic versus Probabilistic Approaches to Cognition

This thesis will develop a formal explanation for why less is more. The approach taken to achieve this goal is a Bayesian approach to heuristics. Hence, the current section will firstly outline why an integration between Bayesian approaches and heuristics is needed, and crucially, why it has not happened yet.

Homo economicus is a rational decision-making creature with the ability to fully maximize its utility. Unfortunately, Homo economicus sightings are as rare and difficult to verify as Bigfoot sightings. In fact, early proponents of the probabilistic approach to cognition assumed humans act in line with Homo economicus, i.e., assuming *unbounded* rationality Gigerenzer et al. (1999). In contrast to the probabilistic approach, heuristics are commonly viewed as more psychologically plausible than models that take into account all available information (Czerlinski et al., 1999; Simon, 1990). To illustrate, consider choosing which of two opponents, England versus Germany, will win a football match (Fig. 2.2). The available information, usually termed *cues*, might include the official league position, the result of the last game, whether the match is home or away, and which team has scored more goals during a recent competition. One popular heuristic is to first order the cues by their cue validity v (i.e., predictive value), then to proceed from the most valid to least valid until a cue is found that favors one team over the other (Gigerenzer & Goldstein, 1996). What makes this decision heuristic, known as *Take-The-Best*, frugal is that it terminates at the first discriminative cue, discarding all remaining information. In the example (Fig. 2.2), only the first cue, league position, would be used by Take-The-Best to predict Germany as the winner.

Why were heuristics and probabilistic models never integrated? The unbounded rationality view proposed that the human mind relies on probability calculus and is equipped with unlimited reasoning capacity. For example, early economic models such as rational choice theory (Friedman, 1953; Scott, 2000) and expected



	v	★★★★ DEUTSCHER FUSSBALL-BUND	ENGLAND	*cue coding*
(1) League pos.	.90	☺	☹	+1
(2) Last game result	.81	😐	😐	0
(3) Home vs. away	.73	☹	☺	-1
(4) No. of goals	.54	☹	☺	-1

Figure 2.2: Illustrative binary prediction task where a heuristic can be used. Predicting whether Team Germany or England will win is based on four cues: league position, last game result, home vs. away match, and recent goal scoring. Cue validities (v) reflect the relative frequency with which each cue makes correct inferences across many team comparisons. Smiley and frowning faces indicate which team is superior on each cue, whereas a grey face indicates the two teams are equal on that cue. A cue is coded +1 when it favors the team on the left (Germany), -1 when it favors the team on the right (England), and 0 when the teams are equal along that cue.

utility theory (Von Neumann & Morgenstern, 1944,1947,1953,2007) portrayed humans as always acting rationally with complete knowledge, out of self-interest and with the desire to maximize wealth. Despite acknowledging that Homo economicus assumes unrealistic mental abilities, proponents of the unbounded rationality argued that people act *as if* they were unboundedly rational. Where did this confidence come from? To understand their positioning, one needs to take into account the *probabilistic revolution* (Gigerenzer & Murray, 1987). After two millennia following Aristotle, who saw logic as the theory for ideal human reasoning and inference, probability theory emerged in the mid-17th century and replaced logical certainty with a more modest theory of rationality, acknowledging the fundamental uncertainty of human nature (Daston, 1980). Probability theory became the new calculus of uncertainty (Laplace 1814/1951) and some even saw probability as equivalent to human thought, as exemplified in the famous treatise *An Investigation of The Laws of Thought* (Boole, 1854). This probabilistic revolution has shaped and continues to influence our understanding of the mind across cognitive science, economics, and animal behaviour until today. In psychology, in the 1940s, inferential statistics started dominating as the method to draw inferences from data to hypotheses (Gigerenzer & Murray, 1987). The second probabilistic revolution in psychology happened around 1955 on the level of theory construction, and together these developments resulted in an increasing understanding of cognitive processes as statistical inference (Gigerenzer & Murray, 1987). The impact of the probabilistic revolution on nowadays cognitive science can be seen in the number of Bayesian frameworks that exist for modelling various higher-level and lower-level cognitive phenomena. For example, memory systems, perception, motor control and vision systems have been shown to be consistent with optimal Bayesian inference, as they accurately reflect the statistics of the world (e.g., Anderson (1990); Anderson and Schooler (1991); Jacobs (2002); Kersten, Mamassian, and Yuille (2004); Körding and Wolpert (2004); Ma, Beck, Latham, and Pouget (2006); Weiss, Simoncelli, and Adelson (2002); Yuille and Kersten (2006)). Higher-level cognitive phenomena that have been successfully modelled with Bayesian inference include everything from learning (Tenenbaum, 1999), human reasoning under uncertainty (Oaksford & Chater, 1994), categorisation (Tenenbaum & Griffiths, 2001), counterfactual inference and causal representation (Pearl, 2000; Sloman & Lagnado, 2005), causal learning and theory change (Bramley, Dayan, Griffiths, & Lagnado, 2017), to language acquisition (Hsu & Chater, 2010).

However, importantly, many Bayesian inference models of higher-level cognition are not proposed as psychological process models, but rather as computational-level theories (Jones & Love, 2011). Cognitive scientists make a clear distinction between Bayesian inference models at the computational or functional level of analysis, the algorithmic level of analysis (corresponding to psychological processes, e.g., heuristic decision making processes), the implementational level of analysis (corresponding to the information integration at the neuronal level (Marr, 1982b) (see Marr's levels in Table 2.1). In that way, many probabilistic models were formulated as descriptive rational models of cognition rather than rational process models. For example, Oaksford and Chater (2007) introduced a rich general probabilistic framework for inference which sees probability theory as a normative

Level	Definition
Computational	What problem is the brain solving? What is the input and output? What is the structure of the environment?
Algorithmic	What processes does the mind execute to produce the solution? What algorithms are computed?
Implementational	Hardware: How are those algorithms implemented in the brain?

Table 2.1: Marr's levels of analysis still continue to influence cognitive science (Marr, 1982b).

theory of how people should reason about uncertainty, and as a descriptive theory of how good reasoners actually reason, but makes no assumptions about how cognitive processes solve the probabilistic problems they face. In fact, Oaksfoard and Chater consider explicit probabilistic calculation at the processing level to be highly unlikely given how hopelessly poor people are at explicit mathematical reasoning about probability. On the other hand, there is a multitude of evidence at the neuronal level suggesting the brain represents probability distributions and performs probabilistic inference, i.e., at Marr's implementational level (Ma et al., 2006; Pouget, Beck, Ma, & Latham, 2013).

In contrast, heuristics were formulated as algorithmic process models (Table 2.1), proposed as possible algorithms the mind relies on to arrive at a solution (Gigerenzer & Brighton, 2009). Instead of integrating these heuristic process models with probabilistic approaches at the computational level, and trying to understand how heuristic processes may give rise to optimal Bayesian inference at the computational level, research on heuristics and rational probabilistic models of cognition diverged into two different research paradigms with little communication between them. For example, proponents of *fast-and-frugal* heuristics focused on the successes of their heuristics at capturing people's behaviour, or their success in statistical competitions (Gigerenzer et al., 1999). At the same time, they strongly argued against any probabilistic implications for the processing level, arguing that the computational-level approaches to cognition have the problem of apparent intractability of rational probabilistic calculation (Brighton & Gigerenzer, 2008, p. 189), (Chater, Tenenbaum, & Yuille, 2006, p. 293). Despite many of the computational-level theories not being meant to capture algorithmic processes, proponents of the heuristics paradigm expected that this would make them useless as models of human cognition, and the only way this could be overcome would be by formulating a computational-level theory that can also take into account process level models (such as heuristics), assessing their compatibility (Brighton & Gigerenzer, 2008). While many have proposed this as a solution, no such framework was provided. Only in recent years, some computational level theorists, mostly coming from the rational probabilistic paradigm, have attempted to bridge the gap with the algorithmic processing level (Griffiths, Lieder, & Goodman, 2015; Lieder, Griffiths, Huys, & Goodman, 2017; Sanborn et al., 2010). Nevertheless, none of these recent approaches have attempted to extent the probabilistic inference models to understand why heuristics work.

Despite the success of simple heuristics at competing with more complex models, such as in the famous city size task and other real-world environments (Czerlinski et al., 1999; Gigerenzer & Goldstein, 1996), proponents of the heuristics believed a rational explanation was no longer necessary. Their line of reasoning was that because a simple heuristic is able to match performance with a more traditionally rational strategy, this was enough evidence to dispense with any rational explanation based on probability theory. However, just because a simpler model could outperform a traditionally rational model does not mean that refuting a rational explanation of these impressive results is the way forward, and instead a rational explanation should be found (Chater et al., 2003). Yet, this integration of rational inference models and heuristics did not happen. Lastly, the two most prominent accounts of heuristics in psychology both in fact regard heuristics as incompatible with Bayesian inference for different reasons which are outline below (next Section).

In sum, an integration between heuristic and probabilistic approaches has been difficult in the past, and has not happened yet for the historic reasons outlined. Despite the success of Bayesian inference models at capturing human behaviour both on the neuronal and the computational level, the heuristic proponents refused to assess whether an integration with computational Bayesian inference models would be possible. However, I argue that real progress can only be made by integrating the the Bayesian inference models with heuristics, which will advance our understanding of heuristics.

2.3 Prominent Accounts of Heuristics

I will now introduce the two most prominent approaches to heuristics, before discussing less-is-more effects and explanations thereof below. The two most prominent heuristic accounts are Kahneman and Tversky's *heuristics-and-biases* program (Tversky & Kahneman, 1974), and the *fast-and-frugal heuristics* program by Gigerenzer and the ABC research group. I will evaluate both heuristic programs with regard to central aspects of the thesis: 1) To what extent has the program contributed to our understanding of why heuristics work? 2) To what extent has the heuristic program aided an integration of probabilistic rational models and heuristics?

2.3.1 The Heuristics-and-Biases Program

The heuristics-and-biases program emphasizes heuristics' deviations from Bayesian rationality, interpreting their suboptimal performance as a consequence of their computational efficiency (Kahneman, 2003; Tversky & Kahneman, 1974). That is, it assumes that heuristics are subject to the *accuracy-effort* trade-off, which posits that heuristics require lower cognitive effort, but necessarily pay with lower accuracy levels. Note that this means in the heuristics-and-biases account less-is-more cannot exist. Instead, the heuristics-and-biases account focuses on heuristics as *biases*. Importantly, the reason for classifying heuristics as irrational lies in the rational norms applied as normative standards, derived from Bayesian rationality and traditional economic theories. Hence, Kahneman and Tversky interpreted any deviance in people's behaviour from the axioms of probability theory and logic as irrational behaviour. For example, consider the *representativeness* heuristic which results in the conjunction fallacy (Tversky & Kahneman, 1983). When asked to judge the probability of two events (”*Linda is a bank teller and is active in the feminist movement.*” and ”*Linda is a bank teller.*”) after seeing a personality description of Linda that matches a feminist profile very well, people give a higher probability judgement to the conjunctive event than to the single event of Linda being a bank teller. This clearly violates probability theory as the probability of two events occuring in conjunction is always less than or equal to the probability of either one occurring alone. Tversky and Kahneman (1983) explained this bias with people relying on a representativeness heuristic, which judges the probability of an event by how representative and similar it is in essential characteristics to its parent population.

To what extent has the heuristics-and-biases program contributed to our understanding of why heuristics work? The heuristics-and-biases program did not claim that heuristics work well (although leaving the option for good and bad depending on situation), and less-is-more effects do not exist. Instead, every heuristic in this program is tied to a reasoning bias that violates the laws of probability theory. Biases are assumed to be the result of people’s insensitivity to prior probability of outcomes for example (Tversky & Kahneman, 1974). However, no formal explanation for why and when heuristics perform well was provided.

To what extent has the heuristics-and-biases program aided an integration of probabilistic rational models and heuristics? The heuristics-and-biases program has not aided an integration of probabilistic models with heuristics, as it focused on their divide instead, moving them further away from each other. Furthermore, due to the lack of formalization of these heuristics (such as representativeness), it becomes more difficult to place them into a rational probabilistic framework, or make any meaningful predictions about when and why heuristics perform well. For example, one main criticism of the heuristics-and-biases account has been that heuristics are vague labels rather than theories and explanations that make testable predictions (Gigerenzer & Goldstein, 1996; Gigerenzer, Hertwig, Hoffrage, & Sedlmeier, 2008). In sum, the heuristics-and-biases account sees the relationship between heuristics and Bayesian models is one where heuristics are regarded as biased approximations to the optimal benchmark of Bayesian inference (Table 2.2).

	Heuristics-and-Biases Account	**Fast-and-Frugal Heuristics Account**
Compatibility?	Heuristics are conceived as deviating from rational Bayesian inference, because they often do not conform to probability theory.	Heuristics and rational Bayesian approaches are seen as opponents and are often pitted against each other in modelling competitions (Katsikopoulos et al., 2010; Martignon & Hoffrage, 2002).
Rationality of Heuristics?	Heuristics are irrational.	A heuristic is *ecologically rational* to the degree that it is adapted to the structure of the environment (Gigerenzer et al., 1999).
Rational Norms?	Axioms of Logic & Probability Theory.	Ecological Rationality (no more logic and probability theory)
Relationship between heuristics and Bayesian models?	Heuristics are considered as biased approximations to optimal Bayesian inference (Kahneman & Tversky, 1972; Tversky & Kahneman, 1974)	Heuristics are considered psychologically plausible algorithms, while Bayesian inference models are considered too computationally heavy. The models are perceived as mutually exclusive.

Table 2.2: Compatibility with Bayesian models. The relationship between the two most prominent heuristic programs and optimal Bayesian inference.

2.3.2 The Fast-and-Frugal Heuristics Program

In contrast, two decades later, Gigerenzer and the ABC research group introduced a novel account wherein heuristics are defined as computational models with a set of rules, that specify precise steps of information gathering and processing involved in generating a decision (Gigerenzer et al., 1999). In contrast to the heuristics-and-biases account, the fast-and-frugal heuristics account does not rely on Bayesian rationality as normative standards any longer, but instead introduced a different form of rationality, *ecological rationality*, which emphasizes how well heuristics are adapted to the structure of the environment (Gigerenzer et al., 1999). The authors proposed that people rely in an *adaptive toolbox*, containing a collection of fast-and-frugal heuristics, from which they select heuristics in an adaptive manner depending on the surrounding task environment (Payne, Bettman, & Johnson, 1993; Todd & Gigerenzer, 2000) (A list of heuristics in the adaptive toolbox is in Appendix A). Rather than focusing on human failings and biases, this program catalogues various cases in which humans excel by using simple heuristics in everyday decisions. The fast-and-frugal program does not assume an *accuracy-effort* trade-off any longer, as heuristics are shown to excel despite using less information and computation (Czerlinski et al., 1999). This is shown with impressive less-is-more effects, which the fast-and-frugal approach became famous for, where a heuristic can sometimes outperform a more complex model that uses more information. Less-is-more effects and the fast-and-frugal's approach to why heuristics work are discussed in more detail below in Section 2.4.

To what extent has the fast-and-frugal heuristics program aided an integration of probabilistic rational models and heuristics? In fact, ecological rationality entirely dispenses with probability theory and logic as normative standards. Instead, Gigerenzer and colleagues argue that human behaviour should never be measured against probability theory, and instead posits that the fit between heuristic strategy and environment should determine the rationality of a strategy. The ecological rationality approach originates in the *bounded rationality* approach (Simon, 1990). Simon rejected the notion of human rationality as *optimization under constraints*, which still underlies Kahneman and Tversky's work and most work in behavioural economics and psychology, suggesting if people behaved optimal they would be consistent with the axioms of probability, however often fail to do so due to capacity constraints. Instead, according to the ecological rationality approach, people behave optimal when they rely on the strategy which matches the environment they are in. That is, this programs's answer to the question of when and why a heuristic performs well is *if the structure of the heuristic matches that of the environment*

(Gigerenzer et al., 1999). The main reason that Bayesian rationality was rejected by the fast-and-frugal heuristics account is that Bayesian probabilistic integration was assumed to be computationally too heavy for the human mind, addressing the compatibility only on the psychological processing level. Furthermore, an interpretation of optimal Bayesian inference models as opponents to heuristics in statistical simulations did not aid an integration (Martignon & Laskey, 1999). In sum, the fast-and-frugal account saw the relationship between heuristics and Bayesian models as mutually exclusive models (Table 2.2). Nevertheless, despite the seeming incompatibilities between the ecological and Bayesian rationality approach, I conjecture that these two accounts are in fact more compatible than they first appear, and the idea of a strategy being optimal when it matches the structure of the environment need not be opposing the idea of strategy as optimal when it matches probabilistic inference. As will be seen in this thesis, the Bayesian frameworks developed here show that both interpretations are compatible.

2.3.3 Heuristic Definitions and Terminology

This section will introduce important definitions and terminology from the fast-and-frugal heuristics program which are central to the Bayesian frameworks developed in Chapters 4 and 5. These include formal definitions of the two most prominent fast-and-frugal heuristics, i.e., the *tallying* and *Take-The-Best* heuristic (TTB), and *full-information* models on the other hand. Also, a formal definition of heuristic cue validities will be given.

First, I will define the tallying and TTB heuristic. With respect to the binary football prediction task (Fig. 2.3), the tallying heuristic would tally which team is better on each cue and chooses the team that has more cues in its favor. In the scenario depicted in Fig. 2.3*A*, this algorithm would choose England. That is, the tallying heuristic simply tallies the valences of all cues, ignoring any possible differences in weight magnitudes (Czerlinski et al., 1999; Dawes, 1979). Notice, that in contrast, the TTB heuristic ignores cues instead, e.g., in the football example TTB would have chosen Germany as the winner instead, as the highest ranked cue was able to discriminate among alternatives. Hence, TTB relies on cue validity (i.e., v in

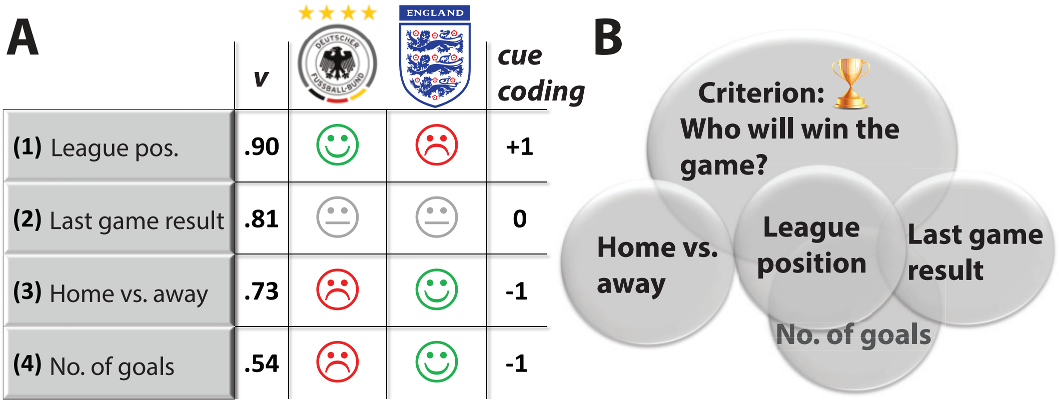


Figure 2.3: Illustrative example of a binary prediction task. (*A*) Predicting whether Team Germany or England will win is based on four cues: league position, last game result, home vs. away match, and recent goal scoring. Cue validities (v) reflect the relative frequency with which each cue makes correct inferences across many team comparisons (Equation 2.1). Smiley and frowning faces indicate which team is superior on each cue, whereas a grey face indicates the two teams are equal on that cue. A cue is coded +1 when it favors the team on the left (Germany), -1 when it favors the team on the right (England), and 0 when the teams are equal along that cue. (*B*) Irrespective of cue validity, cues can co-vary (illustrated by overlap) with the criterion variable but also with each other. The heuristics considered here ignore this covariance among cues.

Fig. 2.3*A*) to create a ranking order of cues and to be able to sequentially search through cues (Gigerenzer & Goldstein, 1996). The tallying heuristic also relies on cue validities, however only for extracting the cue valences. Both the tallying and the TTB heuristic can also be defined in terms of a search rule, stopping rule, and decision rule (Gigerenzer et al., 1999) as laid out in Box 2.3.2 and 2.3.3: 1. Search rules specify in what direction the search extends in the search space. 2. Stopping rules specify when the search is stopped. 3. Decision rules specify how the final decision is reached.

In contrast to the above heuristic algorithms, a *full-information* model such as a full regression model would make greater use of the available information: cue weight magnitudes, predictive values, and covariation among the cues. For example, league position and number of goals scored are highly correlated (Fig. 2.3*B*). Although such covariances naturally arise and can be meaningful, the cue validities used by the tallying and TTB heuristics completely ignore them (Martignon & Hoffrage, 1999). Instead, cue validities assess only the probability with which a single cue can identify the correct alternative (e.g., which team won the football match in Fig. 2.3), derived as the proportion of correct inferences made by each cue alone across the set of binary object comparisons (Martignon & Hoffrage, 1999). Thus, cue validities reflect how predictive each cue is in isolation of other cues. When two cues co-vary highly, they essentially provide the same information, but heuristics ignore this redundancy and treat the related cues as independent information sources. In contrast, regression weights as estimated by multiple regression would always naturally consider covariation among cues as part of the parameter estimation (Box 2.3.4). Cue validity is formally defined as:

Box 2.3.1: Cue validity

$$v = \frac{R}{R+W}, \tag{2.1}$$

R = number of correct predictions

W = number of incorrect predictions, and it follows that $0 \leq v \leq 1$.

The learner is usually assumed to learn cue validities from past experiences (Gigerenzer & Goldstein, 1996; Gigerenzer et al., 1999; Martignon & Hoffrage, 1999). Hence, in applying heuristics to datasets, the cue validities are often learned from the training data.

Lastly, one other feature of heuristics to consider that will be relevant for Chapter 6, is compensatoriness. The tallying heuristic is a *compensatory* strategy, whereas TTB is a *non-compensatory* strategy. Compensatory strategies have the property that a cue can be compensated for by combinations of subsequent cues and tallying is a typical example thereof: It integrates all available cues (however it equally weighs them) and later cues can compensate for earlier cues (Fig. 2.4*B*). In contrast, the non-compensatory TTB heuristic ignores most cues to make decisions, as the most powerful cue C_k can outweigh any combination of the subsequent cues $C_{k+1},\ldots,C_{k+n}$ (Gigerenzer & Goldstein, 1999) (Fig. 2.4*A*). Not surprisingly, both heuristics perform better in environments that match their weighting structure, i.e., tallying performs best in compensatory environments, while TTB performs best in non-compensatory environments (Martignon & Hoffrage, 1999, 2002). Note that multiple regression is also a compensatory strategy, as weaker cues can be compensated for by stronger cues.

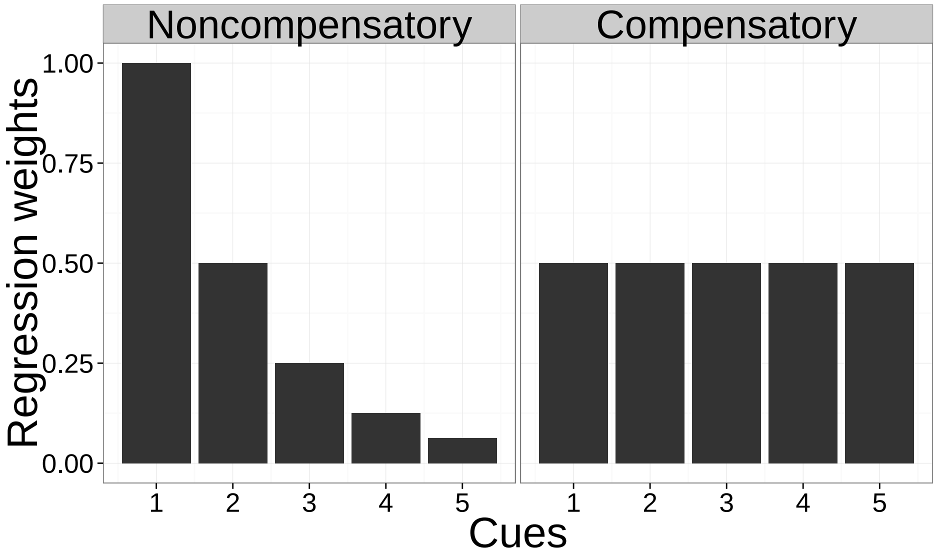


Figure 2.4: Non-compensatory and compensatory environmental structures with five cues. **A**: A perfectly non-compensatory environment has cue weights of 1, 1/2, 1/4, 1/8 and 1/16. In this environment, TTB is as accurate as any linear weighted combination of cues. **B**: A compensatory environmental structure where cue weights are all 0.5. In this environment, the tallying heuristic is as accurate as any linear weighted combination. Taken from Martignon and Hoffrage (2002).

Box 2.3.2: Take-The-Best Heuristic (TTB)

Mechanism:

1. Search through cues in order of their (absolute) validity.

2. Stop on finding the first cue that discriminates between the alternatives.

3. Infer from this cue that the alternative with the higher cue value has the higher criterion value.

The TTB heuristic uses sequential search, meaning when a cue does not discriminate between alternatives (e.g., a value of 0 in Fig. 2.3), the search moves onto the next valid cue until a cue is found that discriminates. The TTB heuristic ignores all remaining cues to make a decision, so sometimes it will rely only on a single cue, if the highest ranked cue discriminates (Gigerenzer & Goldstein, 1996; Martignon & Hoffrage, 1999).

Take-The-Best is a called a *noncompensatory* strategy as the more powerful cue C_k can outweigh any combination of the subsequent cues $C_{k+1}, \dots, C_{k+n}$, i.e., no combination of subsequent cues can compensate for the weight of the more valid cue (as defined by the cue rank order).

Box 2.3.3: Tallying Heuristic

Mechanism:

1. Search through cues in any order.

2. Stop search after m out of a total of M cues (with $1 < m \leq M$). If the number of positive cues is the same for both alternatives, search for another cue. If no more cues are found, guess.

3. Decision rule: Decide for the alternative that is favoured by more cues.

The tallying heuristic entirely ignores cue weight magnitudes and weighs all cues equally (unit-weight) (Dawes, 1979; Gigerenzer et al., 1999).

As cue weights, tallying relies only on the cue directionalities. In one definition, the cue directionalities are assumed to be known in advance (Dawes, 1979), and in another definition they are learned from the data (Gigerenzer et al., 1999).

Tallying is a *compensatory* strategy because a cue can be compensated for by combinations of subsequent cues, i.e., the negative and positive cue valences tradeoff.

Box 2.3.4: Full-information models

Full-information models make full use of *all* available information in the input data such as cue weight magnitudes, predictive values, and covariation among the cues. These algorithms embody principles of classical rationality:

- Complete search - they use all cue values
- Complete integration - they combine all cues into a single value.
- Optimal weighting - they optimally weight the cues

Examples of full-information models:

- Multiple linear regression / Logistic Regression
- Optimal Bayesian models: integrate observed data with prior information.
- Exemplar models for categorization and memory such as

    Nearest-neighbour classifier (Cover & Hart, 1967)

    Nosofsky's generalized context model (Nosofsky, 1990)

- Prospect Theory
- Neural networks
- Regularized regressions (e.g., ridge or lasso regression)

2.4 Less-Is-More Effects

Yet, the most central phenomenon to heuristics that this thesis will address are *less-is-more* effects. These effects were originally defined by the fast-and-frugal program (Gigerenzer & Brighton, 2009), wherein heuristic algorithms surprisingly outperformed full-information models in real-world prediction tasks.

In the 1970s, Dawes made the surprising discovery that a unit-weight strategy, i.e., the tallying heuristic can predict as accurate as and sometimes even better than multiple linear regression (Dawes, 1979; Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975). How could using less information, i.e., disregarding any differential weighting, perform better than relying on more information? These results came as a surprise to the scientific community. When Robin Dawes presented the results at professional conferences, distinguished attendees told him that they were ”impossible”. However the results were correct and spoke for themselves. There was a small difference in how Dawes' applied tallying and multiple regression to the data. According to Dawes' original definition of tallying, people already know the cue directions in advance from past experience and these do not need to be learned from data (Dawes, 1979). In contrast, in 1999, Gigerenzer and the ABC research group conducted a more comprehensive test in 20 real-world environments (Czerlinski et al., 1999), in which both tallying and TTB, and multiple regression were cross-validated and learned weights from the data. That is, all algorithms were trained on 50% of the dataset and made predictions for the other 50% of the dataset. The real-world datasets ranged across various domains from predicting high school dropout rates, to predicting mammals' average sleep time, to predicting the attractiveness ratings of famous women and men (Czerlinski et al., 1999). The task was to predict the outcome of binary comparisons, for instance, estimating which of two Chicago high schools has a higher dropout rate, based on cues such as writing score and proportion of Hispanic students.

Results showed that, averaged across all datasets, multiple regression outperformed the heuristics at fitting, however the TTB heuristic and the tallying heuristic achieved higher predictive accuracy than multiple regression at generalization (Fig. 2.5). At fitting, multiple regression performed best with 77% correct inferences (i.e., fitting parameters to data which is already known), but then dropped in performance to 68% at predicting novel data which has not been encountered before (i.e., the test sample). In contrast, the heuristics' relative drop in performance was not as steep dropping only 4 percentage points on average. Particularly the TTB heuristic,

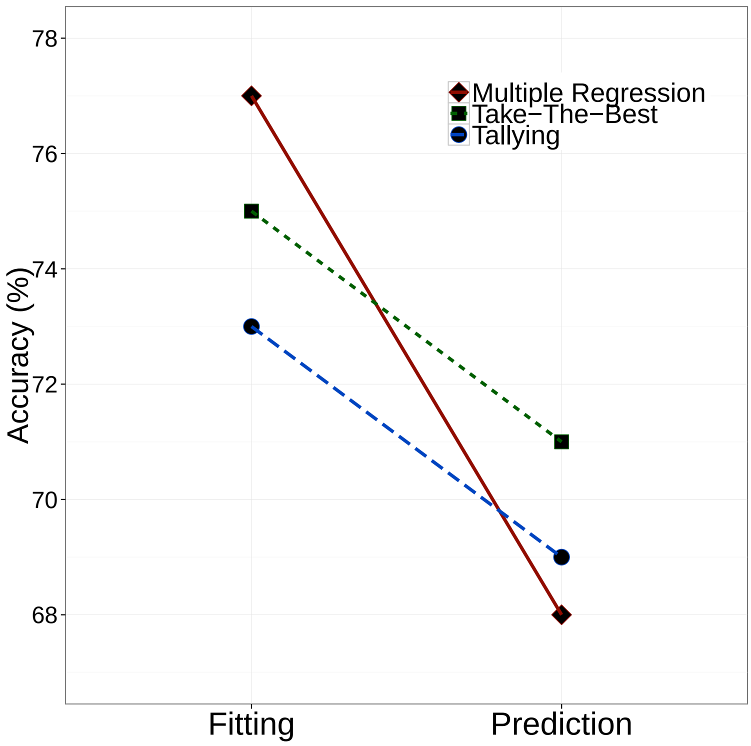


Figure 2.5: Famous less-is-more finding by the fast-and-frugal heuristics program across 20 real-world environments (Czerlinski et al., 1999). Both tallying and TTB predict more accurately than multiple regression during prediction (generalization), despite using less information. Note that multiple regression excels in data fitting, but performs poorly at prediction during cross-validation. The cross-validation method split the datasets in half, i.e., the training sample contained 50% and the test set contained the other 50% of the dataset. Figure is a replica of the graph in the original publication by Czerlinski et al. (1999).

which often relies only on one good reason, excelled at prediction, outperforming both tallying and multiple regression.

These findings had a huge impact, as hitherto, the wide-spread assumption was that more information is always better (Gigerenzer & Brighton, 2009). These findings also undermined the widespread accuracy-effort trade-off assumption, because heuristics were able to exhibit higher accuracy without loosing to higher computational effort. Three years before, Gigerenzer and Goldstein (1996) had stirred up the scientific community with the famous city size task (i.e., predicting which of two German cities has the larger population size), which for the first time showed that a TTB heuristic could match the performance of a multiple regression model.

The above shows a classical example of a less-is-more effect. However, how is less-is-more defined? The term less-is-more is used to label a number of different effects and different people use it to mean different things. Hence, it is particularly important to clarify what definition one refers to. In their original less-is-more discussion, the ABC research group says:

> *Less-is-more effects: More information or computation can decrease accuracy; therefore, minds rely on simple heuristics in order to be more accurate than strategies that use more information and time.* (Gigerenzer & Brighton, 2009, p. 110)

Note that this definition already contains two definitions of less-is-more in one, the classical relative definition of less-is-more and a psychological one. The *relative* definition of less-is-more is the most prevalent in the literature. It is for example described by Gigerenzer and Brighton (2009) in reference to the findings in Fig. 2.5: *Heuristics can lead to more accurate inferences than strategies that use more information and computation. Thus, the accuracy-effort trade-off does not generally hold; there are situations where one attains higher accuracy with less effort.* At a different point in the article, the authors later clarify that *more information* refers to things such as cues, weights, or dependencies. Hence, we define the first, relative less-is-more effect as:

Box 2.4.1: *Relative* Less-is-more

Simple heuristics, that ignore information, can lead to more accurate inferences than strategies that use more information.

However, importantly, the fast-and-frugal program went beyond in their definition of less-is-more surpassing the more trivial relative definition of less-is-more. In the same article, they argued that there is a point where throwing out information actually leads to better performance, and where including it would be detrimental.

This is evident in multiple occasions such as,

Note that the term less-is-more does not mean that the less information one uses, the better the performance. Rather, it refers to the existence of a point at which more information or computation becomes detrimental, independent of costs. (Gigerenzer & Brighton, 2009, p. 111),

and similarly:

A less-is-more effect, however, means that minds would not gain anything from relying on complex strategies, even if direct costs and opportunity costs were zero. (Gigerenzer & Brighton, 2009, p. 111)

Hence, we define the second less-is-more effect as:

Box 2.4.2: ***Absolute*** **Less-is-more**

There is a point where more information becomes detrimental and less information (as used by heuristics) leads to higher accuracy.

Hence, this absolute definition implies that there is actually a point where the simpler model cannot be improved upon with strategies that rely on more information. Note that neither the first nor the second definition of less-is-more have said anything about psychology yet. However, ABC's understanding as indicated by the first quote above also clearly state the assumption that the mind relies on simple heuristics - in fact, all statistical less-is-more findings in Gigerenzer and Brighton (2009) (such as Fig. 2.5) are directly used as implying that people rely on heuristics. This kind of less-is-more effect could be called a descriptive psychological effect, as it expects people to rely on heuristics rather than full-information models, as indicated by behavioural data:

Box 2.4.3: *Psychological* Less-is-more

1) Descriptive: People rely on simple heuristics rather than full-information models.
2) Capacity: There is a point where information gets too much, and processing less information results in better performance.

However, there is also another psychological definition: If one thinks of people as systems with capacity limitations, one may expect that less is more for humans, such that when information gets too much, processing less information becomes advantageous (e.g., by closing one's eyes, or turning to a simpler strategy under high cognitive load (Hoffmann, von Helversen, & Rieskamp, 2013)). However, even if less is more for humans at the capacity level, it does not invalidate any of the other less-is-more effects.

The fact that the central less-is-more definition given by the fast-and-frugal heuristics is a relative claim is evident in some studies showing that heuristics are not always performing best, but can also be outperformed by other full-information models. For example, Martignon and Hoffrage (Martignon & Laskey, 1999) compared heuristics to two Bayesian models across the same 20 datasets as above (Fig. 2.5). The authors compared the TTB heuristic against a naive Bayes classifier which assumes that cues are conditionally independent, and a Bayesian network which assumes that cues are interdependent. On average, the Bayesian network performed four percentage points better than TTB at fitting, and three percentage points better at generalization. While the predictive accuracy of TTB was 71%, the predictive accuracy of the Bayesian network was 74% and that of naive Bayes 73%.

In sum, the above clarifies that there are many definition of less-is-more. By giving the precise definitions, it will be possible to clarify what version of less-is-more each Chapter refers to. While the thesis will address most less-is-more definitions, the main advance will be the Bayesian inference frameworks in Chapter 4 and 5 tackling the *absolute* less-is-more effect.

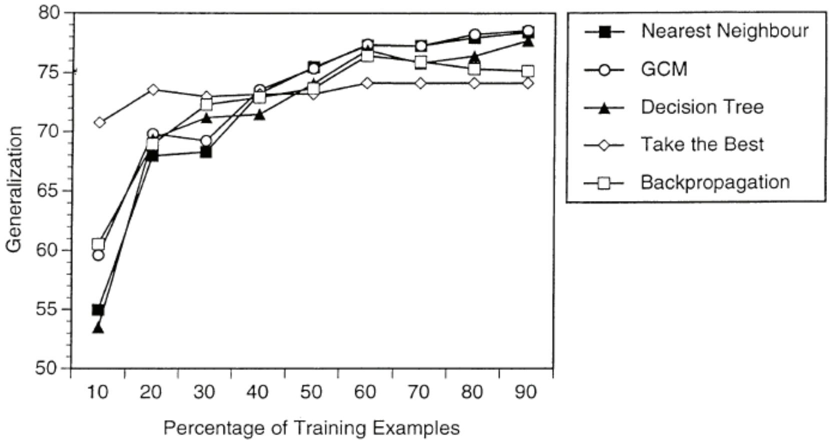


Figure 2.6: Results by Chater et al (2003) in the city size task. Generalization performance of the TTB heuristic in comparison to a Nearest Neighbour classifier, the Generalized Context Model (Nosofsky, 1986), a C4.5 decision tree model (Quinlan, 1993), and a neural network using the backpropagation algorithm (Rumelhart et al., 1986). The ordinate represents the percentage of correct inferences made in predicting the outcome of city comparisons during cross-validation. The training sample size was varied from 10% to 90% of all city comparisons.

2.4.1 When is less more?

Soon after the initial less-is-more demonstration (Fig. 2.5), other cognitive scientists critical of less-is-more effects (*relative* less-is-more claim), put it to a more difficult test. For example, sceptical about the robustness of less-is-more, Chater et al. (2003) used the city size task to test TTB against more powerful machine learning models, such as a nonlinear strategy, i.e., a three-layer feedforward connectionist neural network trained using the backpropagation algorithm (Rumelhart, McClelland, Group, et al., 1986). These multi-layer neural networks are very robust with respect to domain and data structures and are used across psychological and applied research (Russell & Norvig, 2002). The authors also tested two exemplar-based models, a nearest-neighbour classifier (Cover & Hart, 1967) and Nosofsky's generalized context model (Nosofsky, 1986, 1990), a popular model for inductive categorization inferences. The third type of algorithm was the C4.5 decision tree model (Quinlan, 1993), a standard classification learning algorithm in machine learning. Results (Fig. 2.6) showed an impressive performance for the TTB heuristic again compared to the complex machine learning models.

However, crucially, TTB outperformed other algorithms when training data was small, i.e., up to 40% of all pairwise comparisons, however when the sample size was larger, models that utilized more information performed better. This study represents one of the many important studies identifying a factor of *when* less is more. Chater et al. (2003) not only showed that less is not more with larger training sample sizes (i.e., relative less-is-more), but also that simple heuristics seem to have a competitive advantage with smaller training samples. Other studies replicate this finding. Brighton (2006) extended the study by Chater et al. (2003) into a larger study with 25 environments based on widely available regression problems, relying on the same machine learning models. In the paper, Brighton (2006) finds that for half of the environments, TTB clearly outperforms the competitors, and for the other half, TTB performs less well, however it still performs relatively well for smaller training sample sizes (e.g., up to 20 objects). What's more, Katsikopoulos et al. (2010) showed that both TTB and tallying have a performance advantage over more complex models with very small, minute-size training samples. In comparing the predictive accuracy of the heuristics against that of naive Bayes across 19 of the original datasets by (Czerlinski et al., 1999), they find that with training samples of only 2 objects, tallying had the highest predictive accuracy and TTB was more accurate than naive Bayes; and for 3-10 objects, TTB had the highest accuracy, with naive Bayes being more accurate than tallying. However, with training sizes of 50% and larger, the model performances reverse again, and the full-information models outperform the heuristics. In conclusion, a key insight from these less-is-more studies is that heuristics have a competitive advantage over more complex models when observations are sparse. Other factors play a role in *when* less is more, which will be identified and critically evaluated in Chapter 3.

I now turn to *why* less can be more, the topic central to the thesis. The next section will introduce the existing explanation of less-is-more based on *bias-variance*. I will focus on what it can provide and where its limitations lie.

2.4.2 Current Explanations of Less-is-more

The less-is-more effects pose a paradox: How can heuristics, which ignore information, outperform full-information models? Gigerenzer and Brighton (2009) offered an explanation based on the statistical concept of the *bias-variance* tradeoff. This thesis will extent the frequentist bias-variance idea (Gigerenzer & Brighton, 2009) into a Bayesian framework that can formally link heuristics and full-information models.

From a statistical perspective, every model, including heuristics, has an inductive *bias*, which makes it best-suited to certain learning problems (Gigerenzer & Brighton, 2009). A model's bias and the training data are responsible for what a model learns. Subsequently, a model can apply what it has learned from past experiences (i.e., the training data) to novel test cases in *cross-validation*, the core tool for evaluating the performance of learning models in machine learning and psychology (Kohavi, 1995). From a psychological standpoint, a model's cross-validation performance can be understood as its ability to generalize from past experience to guide future behaviour. Thus we are typically interested in the models' *generalization performance* (J. Friedman, Hastie, & Tibshirani, 2001). In addition to differing in bias, models can also differ in how sensitive they are to sampling variability in the training data, which is reflected in the variance of the model's parameters after training (i.e., variance in their estimates over different training samples). The inductive bias and the parameters' variance jointly determine how well a model classifies novel test cases, as can be seen in the equation for prediction error:

$$\text{Total error} = \text{bias}^2 + \text{variance} + \text{noise}. \tag{2.2}$$

Higher flexibility can in fact hurt a model's performance because it makes the model more sensitive to the idiosyncrasies of the training sample. This phenomenon, commonly referred to as *overfitting*, is characterized by high performance on experienced cases from the training sample but poor performance on novel test items. Overfitted models have high goodness of fit but low generalization performance (Pitt & Myung, 2002), see Fig. 2.7*A*. For example, the right-most model in the bottom panel of Fig. 2.7*A* attempts to capture the pattern underlying the data with a higher-degree polynomial model. While it fits the data points better than the

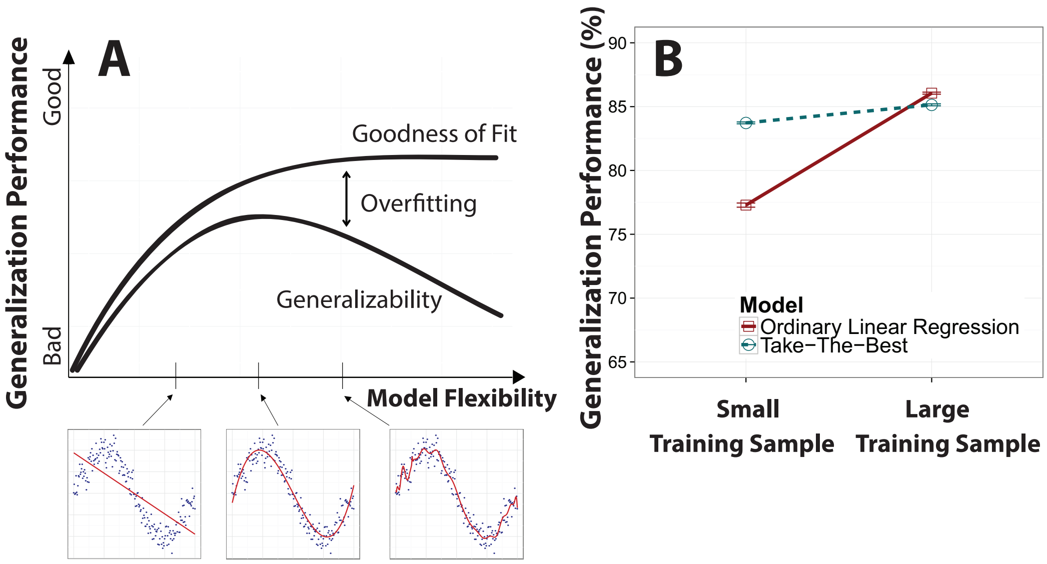


Figure 2.7: The concept of overfitting. *(A)* More flexible models can fit the training sample better (goodness of fit), accounting for most of the variability. However, these models can fare poorly in generalization tasks that test on novel samples (generalizability) Pitt and Myung (2002). *(B)* Our re-analysis of a dataset Czerlinski et al. (1999) used to evaluate heuristics (predicting house prices) finds that TTB outperforms ordinary linear regression at generalization when the training sample is small (20 training cases). However, the pattern reverses when the training sample is enlarged (100 training cases). Error bars represent ± SEM. Details are in the Appendix C.

lower-degree polynomials in the middle and left-most graph, the model's flexibility also results in larger prediction errors due to suffering from increased variance, i.e., resulting in overall lower generalizability.

Bias and variance tend to trade off with one another such that models with low bias suffer from high variance and vice versa (Geman, Bienenstock, & Doursat, 1992), which implies that more flexible (i.e., less biased) models will overfit small training samples and can be bested by simpler (i.e., more biased) models that overfit less, such as heuristics. Hence, this explains why sometimes heuristics, which have a larger bias but low variance, can be more successful than more complex (i.e., less biased, high variance) models. However, as the size of the training sample increases, more complex models should fare better (Chater et al., 2003). Indeed, in a reanalysis of a dataset used to evaluate heuristics (Czerlinski et al., 1999), we find that the advantage for the heuristic over linear regression disappears when training sample size is increased (Fig. 2.7*B*). This reversal with training sample size was also found by Chater et al. (2003) above (Fig. 2.6) and by Davis-Stober, Dana, and Budescu (2010); Einhorn and Hogarth (1975); Katsikopoulos et al. (2010).

In conclusion, the bias-variance tradeoff can explain why sometimes simpler, more biased heuristics can generalize better than more complex, less biased, models. The same bias-variance concept can also account for why, when the amount of data is increased, the disadvantage of more complex models fades (i.e., the problem of overfitting is overcome with more data). Thus, the bias-variance explanation is very useful by giving existing findings, such as the reversal of less-is-more above, a statistical interpretation which had been previously lacking (Brighton & Gigerenzer, 2008). However, what the bias-variance concept is still lacking is a formal computational model which can make testable predictions about when and why heuristics or full-information models will outperform. As could be seen above, frequentist less-is-more effects are often volatile, i.e., the optimal model may be the full-information model in one instance and the heuristic in the next, and it is not always clear why that is the case. Chapter 3 will show several of these reversals depending on environmental conditions. However, in contrast, in a Bayesian account (to the extent that the model is correctly specified), the optimal model is not expected to change with these environmental conditions (such as training sample size), as the optimal Bayesian model automatically weighs prior against data always. In addition, the bias-variance account also does not formally link those models that take into account all or most information when making decisions (e.g., full regression models) to those that focus on very little of the available information (i.e., heuristics). By not attempting to understand the formal relationship between heuristics and full-information models, a full understanding of why less is sometimes more is impossible. Why this is the case will become clear in the Bayesian frameworks developed in Chapters 4 and 5.

To refer back to the definitions of less-is-more in the preceding Section: While bias-variance provides a post-doc statistical explanation of the first less-is-more effect (*relative* less-is-more), it cannot provide insight into the second definition (*absolute* less-is-more), nor provide a formal explanation for relative less-is-more. In contrast, the Bayesian account developed in this thesis (Chapters 4 and 5) aims to provide formal explanations for both the relative and the absolute less-is-more effects.

2.5 Summary

In this chapter, I introduced heuristics, and contrasted heuristics with probabilistic approaches to cognition. I first set out the background for why an integration between probabilistic approaches and heuristics approaches has not happened yet in the past. Next I introduced the most prominent approaches to heuristic, i.e., the *heuristics-and-biases* account, and the *fast-and-frugal heuristics* account, and evaluated both accounts with respect to their explanations of why heuristics work, and their positioning on an integration between heuristic and probabilistic approaches. I then defined two fast-and-frugal heuristics more carefully, tallying and TTB, and gave definitions and terminologies relevant to the thesis. Finally, I introduced less-is-more effects providing the different definitions that exist in the literature. I gave the most common explanation of less-is-more effects based on bias-variance, and discussed its short-comings, which the thesis will advance.

Chapter 3

When (and why) is less more?

Things aren't always what they seem.

-SHAKESPEARE, Macbeth Act II

This Chapter will look at less-is-more effects and investigate what factors lead to less-is-more. That is, this Chapter addresses the relative less-is-more effect (Box 2.4.1) and will identify several environmental and statistical factors that influenced less-is-more effects, which include 1) the type of data the algorithms are trained on, 2) training sample sizes and 3) the sampling method used for cross-validation. Interestingly, I find that these same factors can also be used to make the less-is-more effects disappear. In 4 computational studies, I show that when each of these factors is reversed, a previous less-is-more effect can be reversed. This demonstrates the volatility of relative less-is-more effects and the necessity of a formal framework for heuristics.

The city size task (Gigerenzer & Goldstein, 1996) represents one of the landmarks of less-is-more effects in the history of decision making research. I will show that in the original city size task by Gigerenzer and Goldstein (1996), these effects stem from the fact that the models were not trained on the same dependent variable, and hence the model comparison was not done properly. Furthermore, I will show the effect of using the appropriate method in training both models on the same data, and will apply both Gigerenzer and Goldstein's and my approach to the 20 datasets in Czerlinski et al. (1999) historically used for evaluating heuristics. When the models are trained correctly, the average less-is-more effect disappears across all 20 datasets. In another computational study, I will show that some of the individual datasets still exhibit less-is-more findings provided a small enough training sample is used, in line with previous literature (Chater et al., 2003; Katsikopoulos et al., 2010). However, when the training sample size is increased, the less-is-more effects disappear. Lastly, I will look at the effect of cross-validation method on the relative performance of algorithms, and show that the initial less-is-more effect in the original 20 datasets disappears when training cases are sampled by comparisons instead of objects. This study demonstrates that heuristics and regression models may benefit from different sampling techniques.

Crucially, at the end of each computational study I will assess whether it is clear why less was more in the first place. Especially given that each less-is-more effect was shown to be reversible, in the Discussion I will critically ask what bias-variance can provide as an explanation and what it can not provide. I will come to the overall conclusion that bias-variance is insufficient as an explanation as it cannot fully account for why *relative* less-is-more (Box 2.4.1) was true in these instances in the first place, except for the trivial fact that the heuristics bet on lower overfitting than complex models. It also appears as if the conditions in some studies by the fast-and-frugal heuristics program were particularly advantageous for heuristics in terms of an overfitting advantage. What was missing from these past studies is a formal explanation for relative less-is-more. Moreover, the frequentist bias-variance explanation does not assess whether there is a point where more information becomes detrimental and less information (as used by heuristics) leads to higher accuracy, i.e., the *absolute* less-is-more effect (Box 2.4.2).

3.1 What environmental factors aid heuristics?

Before analysing what led to less-is-more effects in past studies, an obvious question to ask is: What other factors are known to help produce less-is-more effects, i.e., *when* is less more? Various environmental aspects have been identified such as the degree of compensatoriness (Section 2.3.3)(Martignon & Hoffrage, 1999). Both

TTB and tallying perform best in environments that match their non-compensatory and compensatory nature. Other factors that have been identified are small training sample sizes (Chater et al., 2003; Davis-Stober et al., 2010; Einhorn & Hogarth, 1975; Katsikopoulos et al., 2010) (Section 2.4), moderate to high redundancy among cues in the environment for aiding the TTB heuristic, as well as moderate to high variability in cue weights (Brighton, 2006; Dieckmann & Rieskamp, 2007; Gigerenzer & Brighton, 2009; Hogarth & Karelaia, 2007; Martignon & Hoffrage, 2002) for aiding the TTB heuristic. In contrast, tallying seems to do well with lower variability in cue validities and with low redundancy among cues (Hogarth & Karelaia, 2005, 2006). For example, Dieckmann and Rieskamp (2007, 2012) showed that in a simulation study TTB benefited over logistic regression from environments with high covariance levels among cues and higher cue validity dispersion (i.e., more non-comepensatory). However, it was not clear from this study which of these two aspects was more important for TTB's success. The relationship with covariance in the environment makes intuitive sense, as relying only on the most important piece of information is sensible when covariance levels are high, while integrating all information (such as tallying) is more beneficial when covariance levels are low. Crucially, in much of this research, despite advancing our understanding of *when* less is more, it does not clarify *why* less was more even in those instances such as under high covariance or high cue dispersion. While there is evidence that heuristics can excel under these conditions more than others, there exists no formal theory explaining why a heuristic should be able to perform better than the more complex model. For example, do heuristics usually excel due to their ignorance of information, as suggested by the fast-and-frugal program?

3.2 Less is not always more (1): Discovery of an asymmetry in the original city size task

The city size task (Gigerenzer & Goldstein, 1996) was the first time that the TTB heuristic was defined and the first time a relative less-is-more effect was demonstrated by the fast-and-frugal heuristics program. Since the nineties, the city size task has become a popular dataset which has been used many times to demonstrate less-is-more effects (e.g., Chater et al. (2003); Gigerenzer and Brighton (2009); Goldstein and Gigerenzer (2002); Katsikopoulos et al. (2010)). In contrast to the principles of Bayesian rationality, the TTB heuristic was proposed as a satisficing algorithm (Simon, 1990) that purposefully violates the fundamental tenets of rational full-information models: It neither looks up nor integrates all information (Box 2.3.2 and 2.3.4). However, Gigerenzer and Goldstein (1996) claimed that despite it being fast and frugal, it would not show a significant loss of inferential accuracy. A second novelty that Gigerenzer and Goldstein introduced was comparing the performance of a satisficing algorithm against what they considered a classically ”rational” algorithm, i.e., multiple regression, in a real-world environment. The argument was that a real test of the validity of cognitive algorithms would have to be in a complex natural environment resembling the problems our cognitive system faces. They chose the city size task:

City Size Task

The task is to make a choice between two alternatives on a quantitative dimension (Gigerenzer & Goldstein, 1996):

Which city has a larger population? (a) Hamburg (b) Cologne.

The city size task consists of many of the above binary comparisons. In total, the dataset includes all German cities with more than 100,000 inhabitants (83 cities after German reunification). The dataset has the estimated population size of each of these cities as the dependent variable (Appendix D), and cities are described on a set of 9 binary cues, which are visible in Table 3.1. Each cue has an associated cue validity, indicative of how well a cue predicts correctly with respect to the city comparisons (i.e., *which city has the larger population?*). The full environment with all 9 x 83 cue values is provided in Appendix D. All pairwise comparisons between the 83 cities are created, and the task for both the TTB and multiple regression algorithm is make predictions with respect to which city had the larger population on each comparison.

What did Gigerenzer and Goldstein (1996) find? The authors found that TTB was able to match multiple regression in accuracy in making binary city size predictions. The heuristic also outperformed multiple regression in terms of inference speed due to being more frugal, i.e., relying on fewer cues than multiple regression. Hence, TTB was declared the overall winner of the competition taking into account both predictive accuracy and frugality. These results were interpreted as revolutionary because an algorithm that purposefully violates the fundamental tenets of classical rationality (i.e., does not use complete search and complete integration, Box 2.3.4), was able to match or outperform a classically ”rational” algorithm (full-information algorithm). Hence, Gigerenzer and Goldstein (1996) declared the result ”*an existence proof that cognitive mechanisms capable of successful performance in the real world do not need to satisfy the classical norms of rational inference.* (Gigerenzer & Goldstein, 1996, p. 23). This paper went on to spawn hundreds if not thousands of follow up papers with the perspective that heuristics can outperform rational models (e.g., see Gigerenzer and Gaissmaier (2011) for a review of a decade).

This was the first time a less-is-more effect of this type was demonstrated. However, the critical question one needs to ask is: *Why* was the heuristic able to match performance with multiple regression? To answer this question, we need to take a closer look at the methods used in the original simulation.

3.2.1 Computational methods in the original city size task

To train and test TTB and multiple regression, all possible pairwise comparisons between the 83 German cities were created, resulting in 3,403 total comparisons. Both models were fit to these 3,403 comparisons but also made predictions for the exhaustive set of 3,403 city pairs. That is, the algorithms were trained and tested on the same data, essentially measuring within-sample fitting performance.

At training, both algorithms estimate weights from the training sample. The TTB heuristic estimates cue validities for each of the 9 cues (Table 3.1), assessing the frequency with which a single cue can identify the correct alternative (e.g., which city had the larger population size). Cue validity is derived as the proportion of correct inferences made by each cue alone across the set of binary object comparisons (Equation 2.1) (Martignon & Hoffrage, 1999). That is, the calculation of cue validities relies on both the binary outcome variable (i.e., which city actually had the higher population size on each comparison) and the cue differences vectors from each city comparison (i.e., an example for a cue difference vector for a single comparison is in the cue coding column of Fig. 2.3). Importantly, the actual city populations in Appendix D do not enter the equation for heuristics' parameter estimation.

Cue	Ecological cue validity
National capital (Is the city the national capital?)	1.00
Exposition site (Was the city once an exposition site?)	.91
Soccer team (Does the city have a team in the major league?)	.87
Intercity train (Is the city on the Intercity line?)	.78
State capital (Is the city a state capital?)	.77
License plate (Is the abbreviation only one letter long?)	.75
University (Is the city home to a university?)	.71
Industrial belt (Is the city in the industrial belt?)	.56
East Germany (Was the city formerly in East Germany?)	.51

Table 3.1: Cue validities in the city size task.

On top of the regular training process, Gigerenzer and Goldstein also simulated people with partial knowledge about the cities to imitate inference from memory. Limited knowledge took two forms: One was limited recognition of objects (cities), and the other was limited knowledge about the cue values of recognised objects. To model limited recognition knowledge, they simulated people who recognized between 0 and 83 German cities, with the recognition rate proportional to the city size. To model limited knowledge of cue values, they simulated 6 classes of people, who knew 0%, 10%, 20%, 50%, 75%, or 100% of the cue values associated with the objects they recognized. Combining the two sources of limited knowledge resulted in 6 x 84 types of people, each having different degrees and kinds of limited knowledge. Within each type, they created 500 simulated individuals. That is, instead of performing cross-validation splits on the full dataset of 3,403 comparisons, the authors trained and tested the models with each of the 500 x 6 x 84 simulated profiles on the full dataset.

Multiple linear regression was trained to estimate weights in a different way to TTB: Concretely, instead of training on the cue differences vectors and binary outcomes like TTB, multiple regression estimated its weights by accessing the actual population sizes of the German cities (e.g., such as Berlin having 3,433,695 citizens and Hamburg having 1,652,363 citizens from the Appendix D). Specifically, weights were learned from the original dataset containing 83 cities (in the Appendix D) and the the nine cues were regressed onto the continuous population variable. That means the weights were not derived from the city comparisons and the binary dependent variable, such as in the case of heuristics. The multiple regression weights were then used to predict continuous population sizes of all cities in the test set comparisons, before thresholding them into binary predictions. In sum, the multiple regression and TTB were trained on different dependent variables. It is explicitly stated in the original article that "*Unlike any of the other algorithms, regression had access to the actual city populations (even for those cities not recognized by the hypothetical person) in the calculations of the weights.*" (Gigerenzer & Goldstein, 1996, p. 15/16). In addition, multiple regression was made to include an additional simulated cue: whether the city was recognized. Hence while multiple regression trained on 10 cues, the TTB heuristic trained on 9 cues. Presumably, and as stated in their manuscript, the authors did not consider the asymmetries in the procedure for training TTB and multiple regression incorrect, as they assumed they were strengthening multiple regression by giving it more information than any of the simpler algorithms [1].

Results of the competition are in Table 3.2 for different levels of cue knowledge, averaged across all levels of recognition knowledge. The bar graph in Fig. 3.1 also visualizes results for 100% cue knowledge. Both the Table and the Figure show there were almost no differences in predictive accuracies of TTB and multiple regression. Despite these results showing only marginal differences, they were interpreted as a less-is-more effect because, for the first time, a more frugal algorithm

[1] While Gigerenzer and Goldstein (1996) applied an asymmetric way of training and testing models in the city size task, later papers by the fast-and-frugal heuristics program often trained regression on the same binary dependent variable as the heuristics such as in Czerlinski et al. (1999)

was able to match accuracy of a more complex algorithm (Gigerenzer & Goldstein, 1996).

	Percentage of cue values known				
Algorithm	10	20	50	75	100
Take The Best	.621	.635	.663	.678	.691
Multiple Regression	.625	.635	.657	.674	.694

Table 3.2: Results of the competition in the city size task: Average proportion of correct inferences by cue knowledge levels, averaged across all levels of recognition knowledge. Table is adapted from (Gigerenzer & Goldstein, 1996).

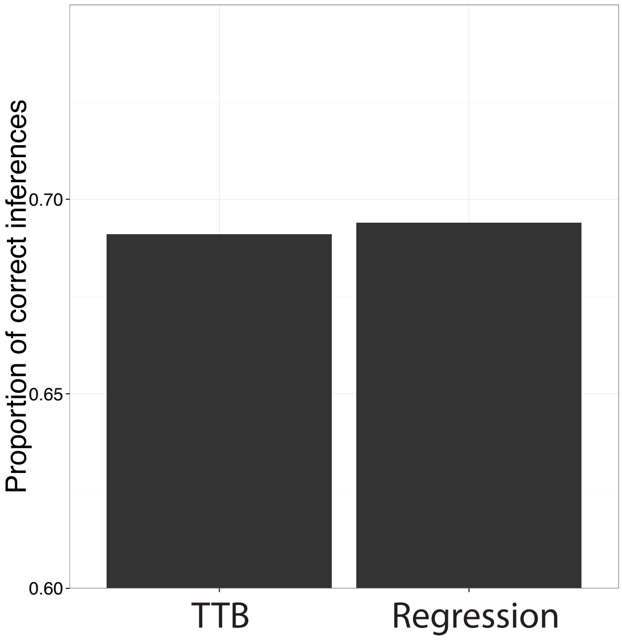


Figure 3.1: City Size Task: Average proportion of correct inferences for the TTB heuristic and multiple regression for cue level knowledge of 100%. Note that this represents an unfair model comparison because it is comparing TTB to the inappropriate regression which trains on population data, while TTB trains on the binary outcome data (adopted from Gigerenzer and Goldstein (1996)). Standard errors were not reported in Gigerenzer and Goldstein (1996).

3.2.2 Summary

In sum, the heuristic and regression model were not trained on equal data in the original city size demonstration. This means the model comparison in Gigerenzer and Goldstein (1996) was not appropriate, as two models should not be compared on the grounds of training on different data, as well as different number of cues. A true test of less-is-more is only impossible when no additional asymmetries are added.

Hence, it is possible that less was more due to the fact that linear regression had additional access to population sizes, resulting in higher overfitting. Regression represents the more flexible model compared to TTB and suffers from higher variance than TTB (which has a larger bias), and its sensitivity to sampling variability can hurt its performance (Section 2.4.2). It is possible that learning the idiosyncrasies of the population sizes in the city size task may have have not helped but harmed regression's performance at making binary predictions for city pairs. If this was the case, one would expect that the less-is-more effect should disappear when both models are trained on the same data. The next computational study will test this hypothesis by comparing both TTB and multiple regression while training on the same binary dependent variable.

3.3 Computational Study 1

This computational study re-models the original city size simulation, with the small change that both the heuristic and multiple regression are trained on the same dependent variable. The goal is to test whether a less-is-more effect (i.e., matching performance of TTB and regression, Fig. 3.1) in the city size task disappears when multiple regression learns weights from the same binary dependent as the TTB heuristic. This comparison does not introduce any additional asymmetries, i.e., the number of cues is kept constant between models and no additional factors are introduced such as varying cue knowledge or recognition of the cities. Instead, 100% cue knowledge is assumed for all models. While not introducing limited recognition or cue knowledge means the comparison of our simulation and the original findings in Fig. 3.1 will not be perfectly possible, it is not the primary goal of this study to replicate all statistical asymmetries of Gigerenzer and Goldstein (1996) — instead, the goal is to test whether there *is* a less-is-more effect in the city size task when models are compared appropriately and no other factors are introduced. TTB will be compared to both the linear regression model that trains on the same binary dependent variable (referred to as *appropriate regression*), and the linear regression model which additionally trains on population size (referred to as *inappropriate regression*). Note that linear regression is used here instead of logistic regression merely to be consistent with the previous work (Czerlinski et al., 1999; Gigerenzer & Goldstein, 1996).

3.3.1 Methods

As in Gigerenzer and Goldstein (1996), all possible pairwise comparisons between the 83 German cities were created, resulting in 3,403 possible comparisons. The criterion variable encodes which of two cities has the higher population size (coded as +1 when the first city is larger and -1 when the second city is larger, i.e., the order coding is consistent and cities are randomly selected). As cue values, cue (trinary) differences vectors resulting from the comparisons were used (coded as 1, -1 or 0 such as cue differences values in Fig. 2.3). All algorithms trained on the set of all comparisons and also made predictions for the exhaustive set of city pairs. The simulation was repeated 100 times and the variation between simulation iterations lies only in the way the comparisons are sampled, i.e., one can sample the comparison between a first and a second city either as city1 - city2 or city2 - city1, where the order results in different differences vectors. The order of sampling was randomly determined for each comparison on each simulation run. During training, the TTB heuristic learned weights by computing cue validities on the differences data and binary criterion variable, and appropriate regression also learned weights from the same differences data and binary criterion variable. In contrast, the inappropriate regression, to replicate the method by Gigerenzer and Goldstein (1996) from above, first learns weights by regressing the population size onto all 83 cities' nine cues (data in Appendix D), which are then used to make continuous population size predictions for each city of the exhaustive set of city comparisons. These continuous predictions are then thresholded to binary predictions for the test comparisons.

Statistical Parameters in the Simulation

Number of objects	83
Number of pairwise comparisons	$N = 3403$
Number of cues	$m = 9$
Class variable	Binary, ±1 (Which city has the larger population size?)
Absolute correlation between cues (averaged over cue pairs)	mean = 0.19
Training sample size	100% of all pairwise comparisons
Test sample size	100% of all pairwise comparisons (equivalent to training sample)
Number of repetitions	100

Table 3.3: Statistical parameters in the simulation on the city size task (Gigerenzer & Goldstein, 1996) as presented in Fig. 3.2.

3.3.2 Results

Fig. 3.2 shows that the appropriate linear regression models outperformed TTB, suggesting there is no less-is-more effect in the city size task, when fitting performance is assessed, and both models are trained equally. The performance of the appropriate regression was significantly greater than TTB ($t(99) = 55.84$, $p < 0.001$), and so was the performance of the inappropriate regression ($t(99) = 29.38$, $p < 0.001$). The performance of the inappropriate regression was no longer matched to the performance of TTB (Fig. 3.1, however study results cannot be directly compared as noted). Interestingly, the inappropriate regression which trains on different data was able to outperform TTB in the current study, suggesting it does not overfit so much that it looses in performance to the TTB heuristic in this dataset. However, the best performance is still achieved by the appropriate regression which trains on the same binary outcome variable as TTB. Appropriate regression outperformed the inappropriate regression ($t(99) = 24.06$, $p < 0.001$) suggesting inappropriate regression overfit more than the appropriate regression.

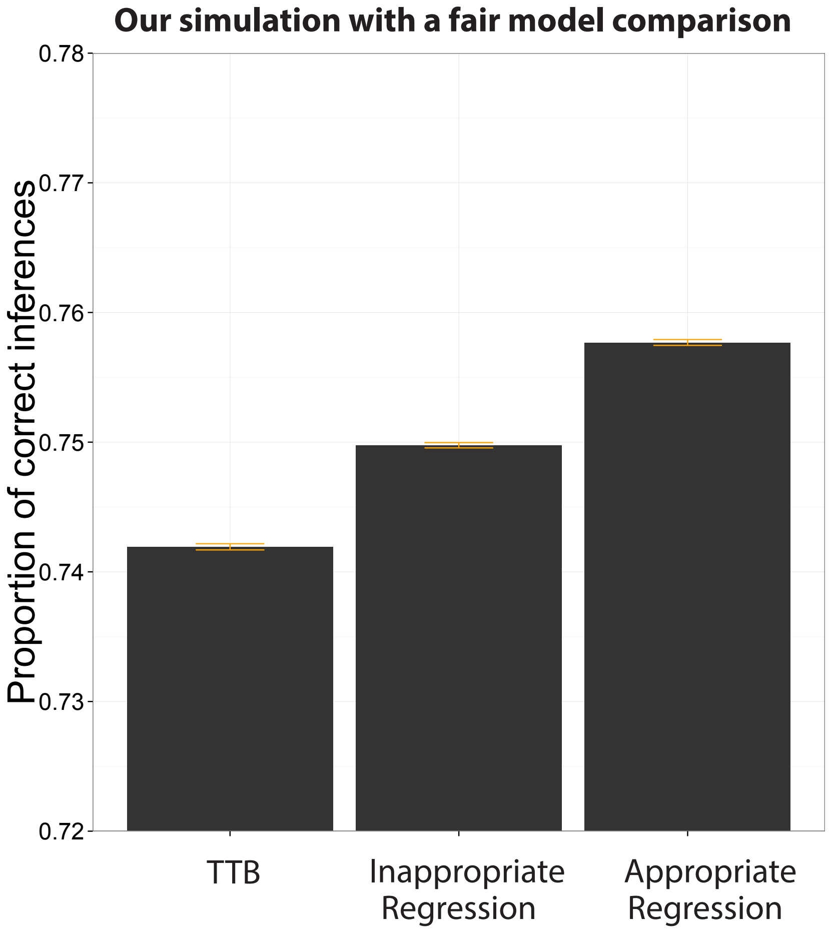


Figure 3.2: City Size Task: Fair model comparison comparing TTB to the appropriate regression model which trains on the same binary outcome data as TTB, as well as the inappropriate regression model for comparison. The ordinate axis represents the average proportion of correct inferences made in the test set. Error bars represent ± SEM across simulation runs.

3.3.3 Discussion

In sum, results show that less is not more in the city size task, when models are trained equally. The success of TTB in the original city size task may stem from inappropriate regression overfitting due to training on different data (or due to accessing additional recognition knowledge). However, in contrast, this study performed a clean model comparison between TTB and appropriate regression keeping the training method constant, and finds no less-is-more effect. Overall, the finding demonstrates how volatile the relative less-is-more effect was in the hallmark dataset for less-is-more effects.

3.3.3.1 Extension

To further understand the effect of training on different data, the next computational study will assess the size of the effect in the original 20 environments frequently used to assess less-is-more effects (Czerlinski et al., 1999) (Fig. 2.5). The goal is to investigate whether a less-is-more effect exists across the original 20 datasets when inappropriate regression is used, and whether an initial less-is-more effect can be reversed with appropriate regression. Thereby, the study will be able to shed more light on why less was more when applying the approach by Gigerenzer and Goldstein (1996). It will help understand whether the less-is-more effect was due to the inappropriate regression model overfitting, in line with the frequentist bias-variance idea. Furthermore, it is not clear how representative the findings in the city size task were, as it could also be that the city size task represents a special case with respect to the effect, which can only be assessed by looking at a more diverse set of datasets. Another extension is that computational study 1 only looked at fitting performance, as the models trained and tested on essentially the same data. Instead, the next computational study will use cross-validation and compare the model's generalization performance.

3.4 Computational Study 2

Computational study 2 will assess the effect of training on different data (Gigerenzer & Goldstein, 1996) in the original 20 real-world datasets (Czerlinski et al., 1999). These popular real-world datasets range across various domains from psychology to biology, health to environmental science. Tasks range from predicting house prices to predicting mammals sleep time to predicting the attractiveness of famous men and women (Czerlinski et al., 1999) (listed in Appendix B). The city size task is among them.

3.4.1 Methods

The summary statistics of the 20 datasets are in Table 3.4 and a description of each dataset is in Appendix B. The task is to make binary predictions for object comparisons as in the city size task, such as predicting ”Which house has the higher sales price?” or ”Which fish has the larger number of eggs?”. The dependent variable in each of the datasets is continuous (e.g., the amount of eggs in fish as indicator of fish fertility, or the house prices in dollars). As in the city size task, the cue values are binary, after being binarized at their median from originally continuous data by Czerlinski et al. (1999). The number of cues ranged from 3 (fish fertility) to 18 (high school dropout). For each dataset, all possible pairwise comparisons of the objects were created to generate the cue differences vectors and the binary criterion variable encoding which of two objects is superior on a comparison (e.g., which of two fish has more eggs?). Both the TTB and the regression models were cross-validated on each dataset by splitting the total set of pairwise comparisons randomly into training and test sets. The size of the training set was 50% of all comparisons, and the test set represented the complementary set of comparisons always. The rationale for choosing 50% as training sample size was to reflect the original cross-validation procedure by Czerlinski et al. (1999). For each training set size, the cross-validation split into training and test set was repeated 1000 times and performance was averaged across all.

As in Computational Study 1, the appropriate regression and TTB learn weights from the cue differences vectors and the binary dependent variable. In contrast, the inappropriate regression derives weights by regressing the continuous dependent variable onto the binary cue values. To give an example, in the house dataset, the selling prices of 22 houses are predicted based on current property taxes, number of bathrooms, number of bedrooms, lot size, total living space etc.. The inappropriate regression learns weights by regression the actual house prices onto all cues for those objects (houses) that are part of the training set. That also means, while the appropriate regression’s predictions are binary, the predictions of the inappropriate regression are initially continuous and are thresholded into binary predictions for the comparisons (i.e., Did house 1 or house 5 have a higher sales price?).

Statistical parameters in the 20 datasets

Number of objects	11 to 395
Number of pairwise comparisons	$N = 55$ to $N = 77815$
Number of cues	$m = 3$ to $m = 18$
Class variable	Binary, ±1 (e.g., Which house has the higher actual sales price? Which professor has the higher salary? Which child is more obese? Which Arctic charr fish has the larger number of eggs? Which city has the larger population size?)

Absolute correlation between cues (averaged over cue pairs)	range = 0.12 to 0.63,
	mean = 0.31,
	median = 0.28, sd = 0.14

Training sample size	50% of all pairwise comparisons
Test sample size	N - training sample (complementary set of comparisons)
Number of cross-validation repetitions	1000

Table 3.4: Statistical parameters in the 20 datasets by Czerlinski et al. (1999) as presented in Fig. 3.3. A full description of the 20 datasets is in Appendix B.

3.4.2 Results

Fig. 3.3 demonstrates the models' generalization performance averaged across the 20 datasets (Czerlinski et al., 1999). As can be seen, TTB outperforms inappropriate regression which trains on the population size data ($t(19) = 1.96,\ p < 0.05$). However, a cross-over in performance can be observed when regression trains on the same data as TTB, i.e., appropriate regression. The performance of the appropriate regression was significantly greater than inappropriate regression, with a mean change in performance of 5.5% ($t(19) = 3.19,\ p < 0.01, (d = 1.46)$ indicating a large effect size when averaged across all 20 datasets (Cohen, 1988)). Appropriate regression also outperformed the TTB heuristic ($t(19) = 2.54,\ p < 0.05$). Table 3.5 displays results for each of the 20 datasets.

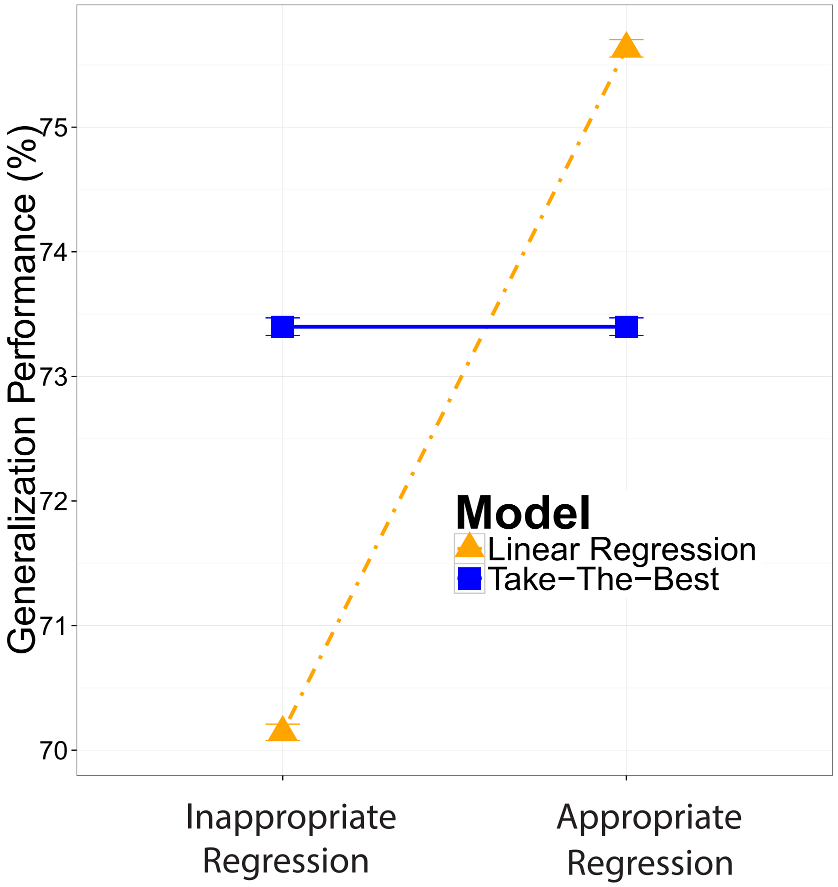


Figure 3.3: Generalization performance across all 20 environments (Czerlinski et al., 1999). Superiority of the TTB heuristic over linear regression disappears when regression is trained appropriately, i.e., on the same data as the TTB heuristic. Inappropriate Regression: Regression trains on different dependent variable than TTB, i.e., the continuous criterion (Gigerenzer & Goldstein, 1996). Appropriate Regression: Regression trains on the same binary dependent variable as TTB. Training sample size was 50% of object comparisons in each dataset. Error bars represent ± SEM.

	Environment	*N*	LR_inapp	LR_app	TTB	Δ (LR_app - LR_inapp)	Cohen's *d* (Δ)	SD_inapp	SD_app
1	City Size	3403	75.12	75.35	73.98	0.23	0.29	0.74	0.84
2	Professors' Salaries	1275	78.96	82.46	79.93	3.50	3.03	1.12	1.20
3	Body Fat	23653	46.14	61.47	58.79	15.33	46.41	0.32	0.34
4	Car Accidents	666	74.24	78.51	70.54	4.28	2.45	1.77	1.73
5	Cloud Rainfall	276	71.09	67.81	64.39	-3.28	-1.16	2.64	3.02
6	High School Dropouts	1596	68.52	73.22	64.37	4.70	3.97	1.17	1.20
7	Obesity	1035	62.94	73.45	74.12	10.51	7.40	1.42	1.42
8	Fuel Consumption	1128	51.07	79.61	76.94	28.54	20.89	1.46	1.27
9	Biodiversity	325	79.68	80.56	82.07	0.89	0.35	2.31	2.77
10	Homelessness	1225	58.59	65.03	63.85	6.44	4.54	1.40	1.44
11	House Prices	231	80.84	86.25	85.07	5.41	2.18	2.45	2.52
12	Land Rent	1653	79.85	81.06	79.56	1.21	1.19	1.00	1.03
13	Mammals' Sleep	595	76.39	79.45	77.82	3.06	1.68	1.75	1.89
14	Mortality Rates	190	67.28	77.69	75.73	10.41	3.03	3.26	3.61
15	Oxidants in L.A.	136	67.61	81.95	68.39	14.33	3.72	3.99	3.72
16	Oxygen	91	81.40	74.21	77.33	-7.19	-1.65	4.07	4.61
17	Ozone in S.F.	55	72.76	76.78	79.07	4.02	0.63	5.92	6.83
18	Attractiveness Men	496	66.70	72.74	72.62	6.04	2.70	2.21	2.26
19	Attractiveness Women	435	68.33	69.74	70.19	1.41	0.59	2.24	2.53
20	Fish Fertility	77815	75.35	75.30	73.19	-0.05	-0.31	0.16	0.17

Table 3.5: Mean performance per dataset: LR_inapp = inappropriate regression, LR_app = appropriate regression. Mean performance (% correct) of inappropriate, appropriate linear regression, and TTB. The difference in performance between inappropriate and appropriate regression is shown with Δ, and Cohen's *d* shows effect size of the change in performance. Training sample size is 50% of all pairwise comparisons (Table 3.4).

Individual datasets differed in the size of the effect of training on different data (Table 3.5). As can be seen, the change in performance (i.e., the difference of appropriate regression - inappropriate regression as indicated by Δ) was positive for 17 out of 20 datasets, and negative only for 3 datasets. These positive differences ranged from 0.23% (city size) to 28.54% (fuel consumption). The respective effect sizes were large for most datasets (Cohen, 1988) with Cohen's d as large as 46.42 for the bodyfat dataset. It can also be seen that the city size had the smallest effect size out of all datasets (Cohen's d of 0.29). Interestingly, for many of the datasets, the TTB heuristic performed better than the inappropriate regression but worse than the appropriate regression, being located in between the two regression models. In conclusion, the size of the effect of training on different data was large for most datasets, and the city size task had the smallest effect out of all datasets. Results demonstrate that the effect of training on different data is greater overall when assessing generalization performance, i.e., out-of-sample performance, as opposed to fitting performance as in the city size task above (Fig. 3.2). Many of the datasets exhibited an initial relative less-is-more effect under the inappropriate regression, which disappeared when regression trained on the same data as TTB.

3.4.3 Discussion

Fig. 3.3 shows that across all 20 datasets by Czerlinski et al. (1999), an initial relative less-is-more effect with the inappropriate regression could be eliminated when regression trained on the same dependent variable as TTB. These findings suggest that the cause of TTB's success was the inappropriate regression overfitting. Hence, despite having access to more information (i.e., the continuous measures in each dataset), this in fact hurt the more complex model (i.e., due to fitting additional noise) in comparison to the simpler TTB heuristic which deliberately ignores information (e.g., cue weights). The current results lend further support to the hypothesis that in the initial city size task, the TTB had an unfair advantage over the regression model due to the way the model comparison was set up (Gigerenzer & Goldstein, 1996). Yet, while the bias-variance concept is useful for understanding less-is-more due to overfitting, it does not address whether less was more in an *absolute* sense in this study, i.e., could TTB have been improved upon when less was more in Fig. 3.3 (see Discussion)?

Next, I will show that in the same 20 datasets, using a smaller training sample often gives rise to less-is-more effects. As introduced in Chapter 2, the size of the training sample can have a big influence on less-is-more effects (Chater et al., 2003; Katsikopoulos et al., 2010).

3.5 Computational Study 3

The goal of this computational study is to demonstrate how the relative performance of TTB and regression changes as a function of training sample size.

3.5.1 Methods

The datasets and simulation methods were the same as in Computational Study 2. However, this time we only considered a linear regression model training on the same binary outcome data as TTB, consistent with Czerlinski et al. (1999). After creating all possible comparisons, the datasets were randomly split into training and test samples 1000 times. Concretely, the size of the training sample was varied between small training samples, i.e., 10 and 20 pairwise comparisons, and larger training samples, i.e., 50% of all pairwise comparisons. 50% of all pairwise comparisons corresponds to $M = 2907$ ($SD = 8857$) comparisons on average across datasets. The test set represented the complementary set of comparisons always. All parameters are also listed in Table 3.6. Hence, the models are compared on their ability to generalize predictions to out-of-sample comparisons they have not seen before. In line with previous research, the hypothesis was that TTB would be more likely to outperform regression with small training sample sizes. However, we expected that the with larger training sample size, the effect could be reversed.

Statistical Parameters in the Simulation

Number of objects	11 to 395
Number of pairwise comparisons	$N = 55$ to $N = 77815$
Number of cues	$m = 3$ to $m = 18$
Class variable (e.g., which house had the higher actual sales price?)	Binary, ±1
Absolute correlation between cues (averaged over cue pairs)	range = 0.12 to 0.63, mean = 0.31, median = 0.28, sd = 0.14
Training sample size	10, 20 & 50% of all pairwise comparisons
Test sample size	$N-10$, $N-20$, $N-50\%$
Number of cross-validation repetitions	1000

Table 3.6: Statistical parameters in simulation as presented in Fig. 3.4. A full description of the 20 datasets is in Appendix B.

3.5.2 Results

Fig. 3.4 shows results for each of the 20 datasets as a function of training sample size. It can be seen that for 18 of the 20 datasets, when training samples were small (10 or 20 training cases), the TTB heuristic outperformed multiple regression ($t(19) = 4.28,\ p < 0.001$ for 10 and $t(19) = 3.84,\ p = 0.001$ for 20 training cases). However, a reversal of performances could be observed when training sample size was enlarged (50% of the training comparisons) for 15 out of the 20 datasets ($t(19) = 2.74,\ p < 0.05$). For example, in the House Prices, Mortality, Professor Salaries, Car Accidents, or High School Dropouts datasets, with 10 and 20 training cases, the TTB heuristic was superior to regression.

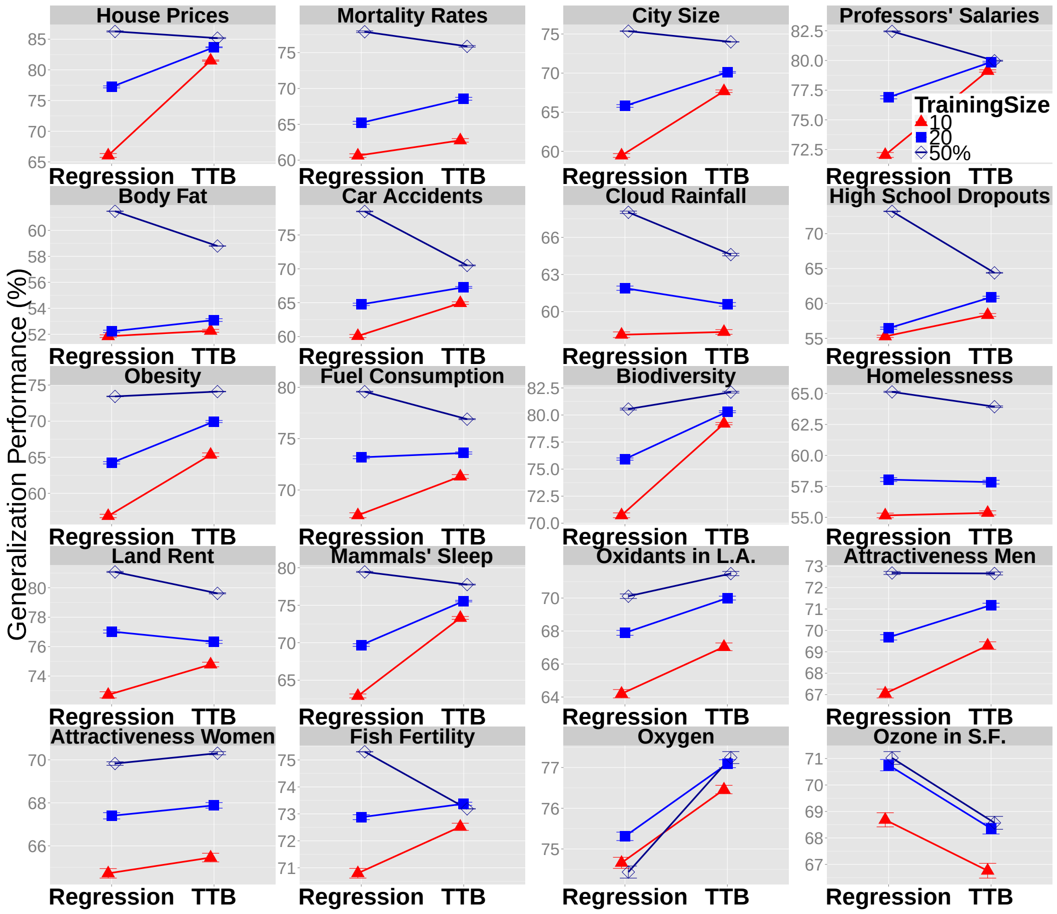


Figure 3.4: Generalization performance of the TTB heuristic and linear regression by training sample size in the 20 heuristic datasets (Czerlinski et al., 1999). The ordinate represents predictive accuracy in predicting test comparisons. With smaller training sample size, i.e., 10 or 20 training cases, the TTB heuristic performed better than the linear regression model in 18 out of the 20 datasets. However, this performance effect was typically reversed when training sample size is large, i.e., 50% of all object comparisons. The average performance of the regression model at 50% training reflects the performance of the appropriate regression model in Fig. 3.3. Error bars represent ± SEM across simulation runs.

3.5.3 Results: As a function of all training sizes

Next, I will demonstrate performance of linear regression and TTB in the 20 datasets (Czerlinski et al., 1999) as a function of a larger range of training sample sizes, ranging from 10% to 100% of all pairwise comparisons (Fig. 3.5). The simulation methods used here are otherwise unchanged from above. As can be seen on the far left of the abscissa (Fig. 3.5), data points for the small training sample sizes of

10 and 20 training comparisons are also included from above (Fig. 3.4). Fig. 3.5 demonstrates the full extent of the superiority of the more complex linear regression model over the simple heuristic: For the vast range of training sample sizes, the predictive accuracy of linear regression (training on binary outcome) exceeds that of TTB. In comparison, only for the small training samples of 10 and 20 training cases, the TTB heuristic outperforms regression, displaying a less-is-more effect.

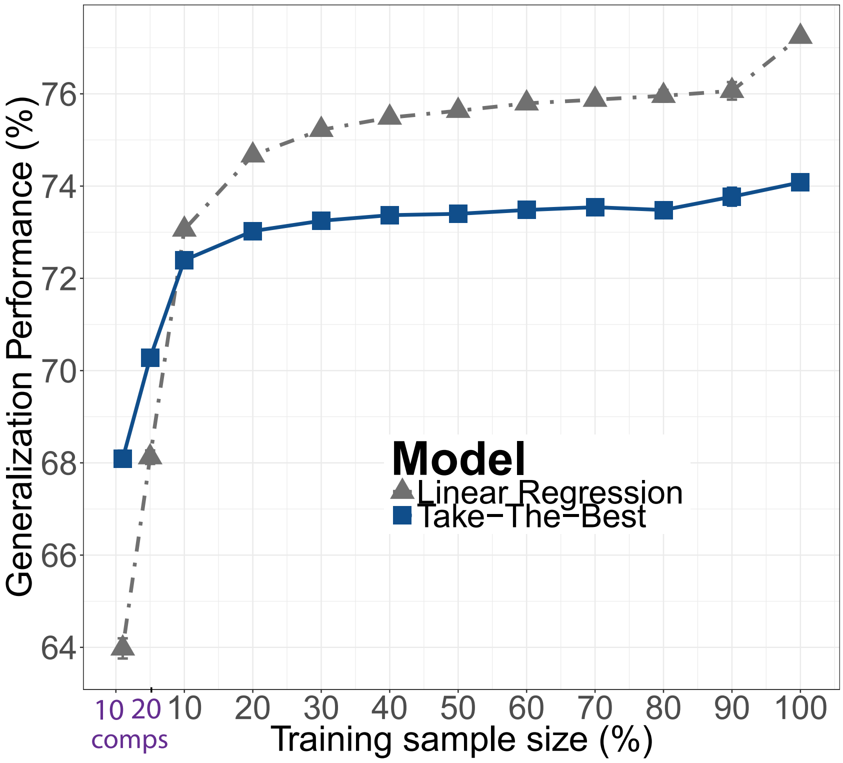


Figure 3.5: Generalization performance of the TTB heuristic and linear regression as a function of training sample sizes across all 20 datasets (Czerlinski et al., 1999). The ordinate represents the predictive accuracy at test during cross-validation. Training sample sizes were varied between 10% to 100% of all pairwise object comparisons, as well as 10 and 20 training comparisons (far left). The two far left data points for 10 and 20 training comparisons map onto results in Fig. 3.4 above. Linear Regression outperformed TTB for almost all training sample sizes. Only with very small training sizes (10 or 20 training comparisons), TTB had a performance advantage. Error bars represent ± SEM.

3.5.4 Discussion

Results in this study showed that less-is-more effects often occur with small training samples and can often be made disappear when the amount of data is increased, leading to a reversal of performances. This finding is in line with previous research insights that simpler algorithms can sometimes be more robust at generalizing to new data than more flexible algorithms (e.g., regression models), when data is sparse (Brighton, 2006; Chater et al., 2003; Katsikopoulos et al., 2010) (as discussed in Section 2.4.2). According to bias-variance, this makes sense: As bias and variance trade off with one another, more flexible models (i.e., less biased and high variance) overfit small training samples and can be beat by simpler (i.e., more biased) models that overfit less, such as heuristics. Yet, as the size of the training sample increases the complex model fares better again usually (Chater et al., 2003). Bias-variance is a useful concept to understand results in terms of overfitting again, yet again, we do not know whether less was more in an absolute sense, i.e., if there could have been a third model that outperforms both regression and TTB (more in Discussion).

An obvious question arising at this point is, how did the fast-and-frugal heuristics program rely on these 20 datasets to demonstrate less-is-more effects, if Fig. 3.5 barely displays any less-is-more effects? How did Czerlinski et al. (1999) analyse the data in Fig. 2.5 where TTB was supposedly able to outperform the appropriate linear regression with 50% training sample size? In short, it needs to be investigated where the discrepancy to our work (Fig. 3.5) stems from. At a closer inspection of the original methods in Czerlinski et al. (1999), it becomes clear that the difference lies in how the training cases were sampled. While the ABC research group sampled objects rather than comparisons among objects, the computational studies in this Chapter sampled comparisons between objects. The final study of this chapter will explain and demonstrate the impact of the sampling method on the relative performance of models.

3.6 Less is not always more (2): Sampling objects versus sampling comparisons

In all computational studies so far, training samples were defined by sampling a subset of all possible comparisons (i.e., object pairs) in a dataset. However, in some other past work including that of Czerlinski et al. (1999) training samples were defined by sampling a subset of the objects first and then training on all pairs within the sampled subset. Both methods can be equally found in the literature and utilized by the fast-and-frugal heuristics program, e.g., sampling objects (Brighton, 2006; Czerlinski et al., 1999) and sampling comparisons (Chater et al., 2003; Rieskamp & Dieckmann, 2012). A systematic study of the impact of the sampling methods on relative model performance has not been done before. To determine whether the classic less-is-more finding would be dependent on this sampling decision, I systematically compared both sampling methods in the original 20 datasets (Czerlinski et al., 1999).

3.7 Computational Study 4

This study compares simulations on the original 20 datasets by (Czerlinski et al., 1999) with both sampling methods for a training sample size of 50% of all comparisons or objects, respectively. When cross-validation is performed such that objects are sampled for both the training and testing sample, this means that that predictive accuracy refers to comparisons between novel, unseen objects. In contrast, when comparisons are sampled, some comparisons in the testing sample will contain objects that have already been encountered in the training sample. For running the simulation with sampling by objects, the method used in (Czerlinski et al., 1999) was applied: For each of the 20 datasets, objects were split in half, assigning 50% of objects to the training sample and 50% of objects to the test sample. Then all possible comparisons between objects were created for both the training and the test sample. Except for the sampling methods, all other factors were kept constant between simulations.

3.7.1 Results

Fig 3.6 shows both sampling methods side-by-side. As expected, the overall generalization of models was lower with sampling by objects. The pattern of results also suggests a reversal of performances with different sampling methods: While sampling by comparisons results in linear regression outperforming TTB ($t(19) = 2.54,\ p < 0.05$) by 2%, sampling by objects results in TTB outperforming regression ($t(19) = 2.14,\ p < 0.05$) with a mean difference of 1.8%. This suggests that sampling by objects gave the heuristic a small advantage over ordinary linear regression.

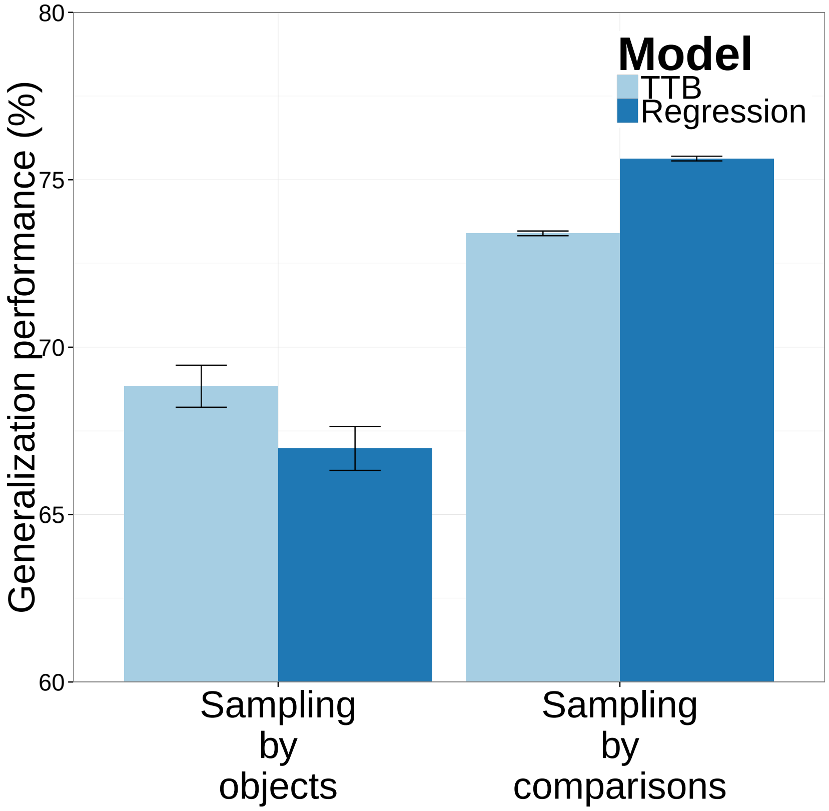


Figure 3.6: Generalization performance of TTB and Linear Regression by cross-validation sampling methods, i.e., sampling by objects and sampling by comparisons, across all 20 datasets by (Czerlinski et al., 1999). Training sample size was 50% of the training cases. Sampling by objects gives the TTB heuristic a small advantage, while sampling by comparisons gives regression a small advantage. Error bars represent ± SEM per dataset.

In conclusion, the cross-validation sampling method impacts the existence of a less-is-more effect (Fig. 3.5), such that changing the sampling method could make the less-is-more effect disappear in the original 20 datasets (Czerlinski et al., 1999).

3.8 Discussion

This Chapter explored several existing demonstrations of less-is-more effects and showed that they are reversible by varying factors in the procedure for training and testing the models. These factors included training algorithms on the same dependent measure, varying training sample size, and whether models were trained on samples of comparisons, or on sub-samples of the objects themselves.

The fact that less-is-more effects were so easily reversible strengthens the need to understand why less was more and why it could be reversed. Next I will evaluate what aspects of less-is-more in the computational studies could be accounted for by the frequentist bias-variance concept. Linking back to the findings in computational study 2, where a reversal of less-is-more could be observed with training both TTB and regression on the same dependent measure, the bias-variance concept can account for the fact that the inappropriate regression overfit due to relying on continuous dependent data which hurts at generalization. However, can it tell us anything about whether an absolute less-is-more effect applied in Fig. 3.3? Next, in computational study 3 a reversal of less-is-more was observed with an increase from small to large training sample size. Why was less more under small training sample size (Fig. 3.4)? The bias-variance concept would suggest that less was more because heuristics are relatively more robust when there are fewer observations: With very little data, the more complex algorithms are more likely to overfit than the simpler model. Simpler heuristics bet on lower overfitting (large bias, low variance) which becomes particularly advantageous compared to complex models when data is sparse. But does this tell us whether TTB outperformed due to the heuristic dropping information such as cues, cue weights and cue covariance (simplicity), or because of its large inductive bias compared to regression? Note that these two factors are confounded in the frequentist bias-variance explanation. Only a model that allows for a continuum of different degrees of bias, and controls both simplicity and bias, can answer this question. In a Bayesian framework, it becomes possible to control for bias and simplicity with a Bayesian prior, as will be presented in the following Chapters. Secondly, another aspect that is not clear from looking at Fig. 3.4, is whether there could not be another, third model that performs better than either TTB and multiple regression. That is, was there really a point where more information was detrimental and less information (as used by heuristics) resulted in higher accuracy, suggesting TTB was the optimal model? It is not clear whether the simpler model could not have been improved upon by including more information (e.g., cues, weights, covariance) with a different strategy. This can only be answered with a formal continuum among models, which will be given in the following Bayesian Chapters. In sum, while the bias-variance gives a powerful statistical interpretation of the relative reversal of less-is-more with an increase in training data, it does not account for a possible *absolute* less-is-more effect (Box 2.4.2).

With respect to computational study 4's findings, the question that follows is why was less more with sampling by objects, but not with sampling comparisons? An obvious explanation is that when sampling subsets of objects, models have to generalize to subsets of comparisons among novel objects only. This means multiple regression overfits with less experienced data. When sampling comparisons, the more complex model has an advantage relatively, as the amount of experienced data increases. Specifically, this bias-variance explanation is based on the same explanation for why less can be more with small training sample sizes. For now I point out that the less-is-more effect that is induced by sampling objects can also be reproduced by selecting fewer training comparisons to train on in the sampling-by-comparisons approach. Hence, the choice of training method does not matter that much, as long as the differences are clearly pointed out and tested. We could find no compelling reason to prefer either method for the studies here, as both are equally present in the literature. However, there may be a small psychological argument for modelling cognitive strategies with sampling-by-comparison (Pachur and Olsson (2012) showed that humans learn better by comparison than direct object learning, and are able to better generalize to new un-encountered objects as well as providing more accurate continuous estimates after learning from comparisons). In the Chapters to follow, we will compare both methods and acknowledge the differences, such as the fact that heuristics seem to benefit slightly from sampling by objects, while full-information models benefit slightly from sampling by comparisons.

In conclusion, bias-variance is insufficient as an explanation for less-is-more as it cannot fully account for why *relative* less-is-more (Box 2.4.1) was true in many of these instance, except for the trivial fact that the heuristics bet on lower overfitting, and it cannot provide insight into the second definition of less-is-more (*absolute* less-is-more, Box 2.4.2). Importantly, bias-variance does not provide a formal computational model which is able to make testable predictions about when and why heuristics or full-information algorithms perform best. Note that in contrast, in a Bayesian framework, if the model is correctly specified, less-is-more is not volatile anymore, as the optimal model is expected to necessarily be the same regardless of training sample size, because the optimal model optimally weighs prior against likelihood. Furthermore, to fully understand why relative less-is-more can occur, we argue that one needs to understand the formal link between full-information models and heuristics, and the continuum of decision strategies between them. The frequentist bias-variance explanation fails to provide such a formal continuum. These short-comings of existing less-is-more explanations will be addressed in the Bayesian framework for heuristics developed in the next Chapters 4 and 5.

3.9 Summary

In this chapter, I identified three important factors that can induce and reverse less-is-more effects, i.e., training algorithms on the same data, training sample size, and the sampling method in cross-validation. I found that in the original city size task, the regression model was not trained on the same dependent variable as the heuristic, and when both models were trained on the same outcome data, the less-is-more effect disappeared. The full size of this effect was then observed in modelling all original datasets used to evaluate heuristics (Czerlinski et al., 1999). Next, this chapter demonstrated the impact of training sample size on less-is-more effects in computational study 3. I showed that an initial less-is-more effects under small training sample sizes could often be reversed with larger training sample sizes in the original 20 datasets (Czerlinski et al., 1999). Finally, this chapter found that the sampling method for training models can reverse less-is-more effects. This insight could explain the contrast between findings in the computational studies here and past results established by the fast-and-frugal heuristics researchers. All less-is-more findings were evaluated with respect of the explanatory power of bias-variance, concluding that bias-variance only has limited explanatory power and insight into why less was more.

Chapter 4

Heuristics as Bayesian Inference - The half-ridge model

Golden Middle.
Aurea mediocritas.
- HORAZ, 65 BC - 8 BC

The golden middle.
- GERMAN IDIOM suggesting an ”ideal middle way between two extremes”

4.1 A Bayesian explanation for why less is more

Less is more, because a simpler model can outperform a more complex model. This is the definition of the *relative* less-is-more effect. On various occasions less-is-more is observed for comparing simple and complex models (e.g., Take-The-Best and regression). For example, the ABC research group is famous for their less-is-more presentations, showing that a TTB and a tallying heuristic could beat the more complex multiple regression at generalization performance (Fig. 2.5). Or else, Chater et al. (2003) and Katsikopoulos et al. (2010) established that heuristics perform well under small training samples sizes compared to complex machine learning models such as neural networks, decision trees, nearest neighbour or naive Bayes classifiers. In addition, our simulation in Chapter 3 demonstrated that TTB was able to outperform regression in nearly all of the original 20 datasets when training samples were small (Fig. 3.4). All of this evidence seems to culminate in the fact that less can be more. However, importantly these studies all target the *relative* less-is-more definition (Gigerenzer & Brighton, 2009), which states that simple heuristics can lead to more accurate inferences than strategies that use more information (Box 2.4.1), e.g., such as exemplified in a definition by Gigerenzer and Brighton (2009) as: *Less-is-more effects: More information or computation can decrease accuracy; therefore, minds rely on simple heuristics in order to be more accurate than strategies that use more information and time.* (Gigerenzer & Brighton, 2009, p. 110).

In response to these statements, other scientists such as Chater et al. (2003) made the point that the effect could actually be reversed by increasing training sample size. However, importantly, this reversal does not really disprove a relative less-is-more effect in any way by showing that under other circumstances heuristics are no longer superior to the complex model. That is because the relative less-is-more effect merely states that heuristics *can* lead to higher accuracy, meaning the relative less-is-more definition allows for the effect to go both ways (and therefore also difficult to falsify). This is quite different from an *absolute* less-is-more effect where such a reversal would not be possible. That is, crucially, none of the studies discussed in the thesis so far were able to address whether less is more in the sense that one could not improve upon the heuristic theoretically by incorporating more information. The *absolute* less-is-more effect states that there is a point where more information becomes detrimental and less information (as used by heuristics) leads to higher accuracy (Box 2.4.2) (Gigerenzer & Brighton, 2009). The fast-and-frugal heuristics program advocated absolute less-is-more, as evident in various places such as the following quote: *Note that the term less-is-more does not mean that the less information one uses, the better the performance. Rather, it refers to the existence of a point at which more information or computation becomes detrimental, independent of costs.* (Gigerenzer & Brighton, 2009, p. 111).

This Chapter will investigate whether absolute less-is-more holds. Interestingly, we find that less is not more in the sense that one could always perform equally well or better than a heuristic by incorporating more information sources, however appropriately down-weighted. This is established in a Bayesian framework. Two separate Bayesian frameworks are developed in the thesis, a Bayesian inference model for the tallying heuristic (this Chapter), and a Bayesian model for both the TTB and the tallying heuristic (next Chapter 5).

Thereby, the thesis will contribute the novel idea that heuristics embody strong Bayesian priors. In the Bayesian frameworks, a formal continuum between simple heuristics and full-information models is created based on a single Bayesian prior. Crucially, heuristics represent an extreme case on the Bayesian continuum of decision strategies, corresponding to entirely ignoring information rather than including all available information (and down-weighting) such as strategies along the continuum do. In the Bayesian model for tallying, the parametric variation of the prior's strength generates a continuum of model flexibility, with a variant of linear regression at one extreme (most flexible and least biased) and the tallying heuristic (least flexible and most biased) at the other extreme. Although the Bayesian model can mimic tallying in the limit perfectly, a crucial difference is that the Bayesian account always regulates weights, but never discards any information. This continuum of models allows for a comparative analysis: I discover that intermediate models, which do not throw out information, perform best across all simulations. Indeed, this suggests that down-weighting but using all of the information is preferable to entirely ignoring it. These results refute the absolute version of less-is-more claims whereby entirely discarding some information sources (as heuristics do) can be optimal.

Resulting from this Bayesian framework is a novel explanation for less-is-more: An explanation for the success of heuristics is that their relative simplicity and inflexibility amounts to a strong inductive bias, which is close to the optimal prior for many learning and decision problems. That is, we find evidence across both Bayesian models that heuristics work not because they throw out information, but because they embody a prior that approximates the optimal prior. The heuristics approximate the intermediate models which are actually optimal and rely on the full information but down-weight the information via the influence of their priors. Thereby, this work extends the frequentist bias-variance idea (Gigerenzer & Brighton, 2009) into a Bayesian explanation for why less is more. The Bayesian explanation moves beyond the bias-variance by providing a formal link between heuristics and full-information models. As a consequence, the Bayesian explanation can address both the relative and the absolute less-is-more phenomena, which was previously impossible in the frequentist bias-variance interpretation. Furthermore, the continuum of decision strategies allows the Bayesian framework to control for simplicity and bias in heuristics, as previously it was not clear in the frequentist bias-variance explanation whether heuristics succeed due their large bias or simplicity (ignorance of information). In the Bayesian account, it becomes possible to disentangle the two factors as potential explanations. Another advance of the Bayesian model derives from the fact that the optimal model should not be as volatile as it was in the frequentist case. While Chapter 3 demonstrated that the optimal model could be easily reversed in relative less-is-more effects, in the Bayesian framework the optimum model is expected to be in the same place as to the extent that the model is correctly specified.

Notice that the above less-is-more definitions have not addressed psychological processing. The computational Bayesian frameworks' core contribution is formal and not psychological, however they have strong implications for directing research into human cognition and decision making. Both the Bayesian model in this Chapter and the next investigate the relative and the absolute less-is-more effects, but are not able to address either a descriptive or capacity-based psychological less-is-more effect (Box 2.4.3). However, implications of the formal framework for psychology will be discussed at the end of each Chapter. In contrast, Chapter 6 will directly explore people's use of decision strategies with the goal of better understanding people's representations.

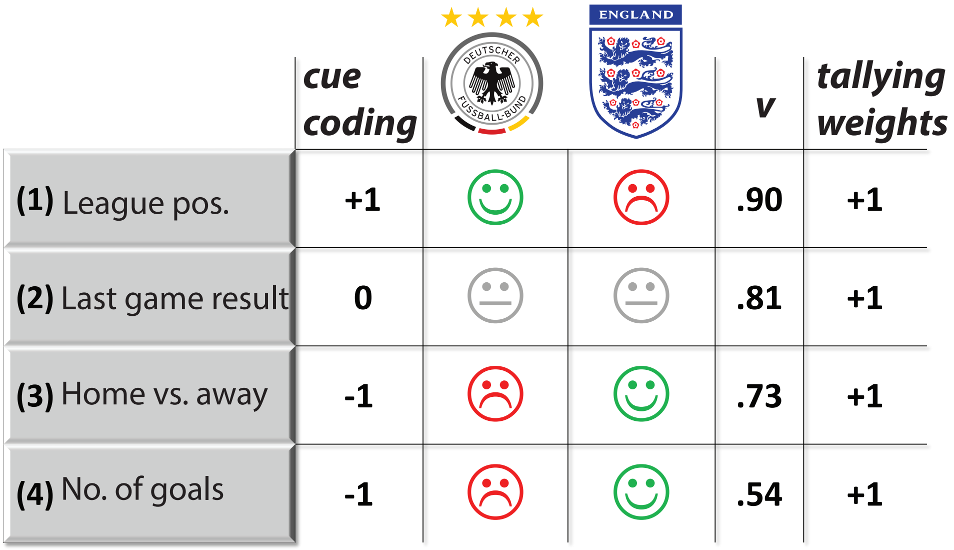
This Chapter is structured as follows: First, I will briefly review how the tallying heuristic works. Next I will demonstrate how the tallying heuristic can be mathematically derived as an extreme Bayesian prior in what we call the Bayesian *half-*

ridge model, which is conceptually related to ridge regression, a regularized regression method from machine learning. The generalization performance of the Bayesian half-ridge model will be assessed in a simulation study on the original 20 real-world environments used to evaluate heuristics (Czerlinski et al., 1999). Finally, Chapter 4 will attempt to derive the TTB heuristic as an extreme case of lasso regression, a different kind of regularized regression from machine learning. The theoretical and psychological implications of this work will be discussed. The half-ridge model and the derivations in this Chapter were developed in collaboration with Matt Jones (University of Colorado, Boulder) and Brad Love (UCL).

4.2 Linking tallying and regression through Bayesian inference

The tallying heuristic ignores information in the input data by dropping cue weight magnitudes. Thus, in contrast to the frugal TTB heuristic, it includes all available cues but weighs them equally. For example, in Fig. 4.1, tallying would simply add up the positive and negative cues values for each alternative, weighted by their respective cue directionality. In the example, all cue directionalities are positive (column ”tallying weights”) and hence team England would be predicted to be the winner in this football prediction task (A full definition of the tallying heuristic is in Box 2.3.3).

This Chapter will contribute the idea that the tallying heuristic represents a strong Bayesian prior. In this approach, the heuristic succeeds because of its relative inflexibility which is equivalent to a strong inductive bias, suitable for various decision problems. This approach will be formalized by proposing a Bayesian model wherein the tallying heuristic is an extreme case along a continuum of model flexibility defined by the strength of the prior. The Bayesian inference model for tallying generates a continuum of models, with a variant of linear regression at one extreme (most flexible and least biased) and the tallying heuristic (least flexible and most biased) at the other extreme.



	cue coding	★★★★ DEUTSCHER FUSSBALL-BUND	ENGLAND	v	tallying weights
(1) League pos.	+1	☺	☹	.90	+1
(2) Last game result	0	😐	😐	.81	+1
(3) Home vs. away	-1	☹	☺	.73	+1
(4) No. of goals	-1	☹	☺	.54	+1

Figure 4.1: Binary prediction task. Predicting whether Team Germany or England will win is based on four cues: league position, last game result, home vs. away match, and recent goal scoring. Smiley and frowning faces indicate which team is superior on each cue, whereas a grey face indicates the two teams are equal on that cue. A cue is coded +1 when it favors the team on the left (Germany), -1 when it favors the team on the right (England), and 0 when the teams are equal along that cue. Heuristic cue validities (v) reflect the relative frequency with which each cue makes correct inferences across many team comparisons (Equation 2.1). Tallying ignores the cue-outcome magnitudes and instead relies on cue directionalities only, as indicated in the *tallying weights* column.

4.3 Tallying as a limiting case of ridge regression

The Bayesian model we develop for tallying is conceptually related to ridge regression (Hoerl & Kennard, 1970), a successful regularized regression approach in machine learning. Ridge regression extends ordinary linear regression by incorporating a penalty term that adjusts model flexibility to improve weight estimates and avoid overfitting (Fig. 2.7). The types of tasks we model in this chapter are binary comparisons of the type in Fig. 4.1, where each input represents comparisons between two alternatives on a set of cues, and the output represents which alternative has the greater value on some outcome variable. Consider a training set of input-output pairs $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ with $\mathbf{x}_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$. An example is Fig. 4.1, where the explanatory variables ($\mathbf{x}$) encode which soccer team is superior on each cue, and the outcome variable (y) indicates which team won each comparison (match). The aim in any linear regression problem is to estimate the weights, i.e., a vector of regression coefficients $\mathbf{w} = [w_1, ..., w_m]^T$, such that prediction error between y and $\mathbf{Xw}$ is minimized. The weights estimated by ridge regression are defined by

$$\hat{\mathbf{w}}_{\text{ridge}} = \arg\min_{\mathbf{w}} \left\{ \underbrace{||\mathbf{y} - \mathbf{Xw}||^2}_{\text{Goodness-of-Fit}} + \underbrace{\theta ||\mathbf{w}||^2}_{\text{Penalty Term}} \right\}, \tag{4.1}$$

where the penalty parameter θ is nonnegative. $||.||^2$ denotes the square of the Euclidean norm, $\mathbf{y} = [y_1, ..., y_n]^T \in \mathbb{R}^n$ is the outcome variable defined over all n binary comparisons in the training sample, and $\mathbf{X}$ is an $n \times m$ matrix with one column for each of the m predictor variables x_j. When the penalty parameter equals zero, ridge regression is concerned only with goodness of fit (i.e., minimizing squared error on the training set). For this special case, ridge regression is equivalent to ordinary linear regression, which is highly sensitive to sampling variability in the training set. As the penalty parameter increases, the pressure to shrink the weights increases, reducing them to zero as $\theta \to \infty$. Thus larger values of θ lead to stronger inductive bias, which can reduce overfitting by reducing sensitivity to noise in the training sample. However, the optimal setting of θ will always depend on the environment from which the weights, cues, and outcomes were sampled.

Importantly, the ridge penalty term is mathematically equivalent to a Gaussian Bayesian prior on the weights, where θ is inversely proportional to the prior variance η^2 of each w_i, that is, $\theta = \sigma^2/\eta^2$ (where σ^2 is the error variance in y, also assumed to be Gaussian). In the Bayesian interpretation, the strength of the prior is thus reflected by $1/\eta^2$, growing stronger as $\eta^2 \to 0$. This prior distribution is combined with current observations (i.e., the training sample) to form a posterior distribution (also Gaussian) over the weights. Like ordinary linear regression, ridge regression provides a point estimate for the weights, equal to the mean (and also the mode) of the full Bayesian posterior distribution (Marquaridt, 1970; Ripley, 2007). The conceptual relationships among ridge regression, ordinary linear regression, and the Bayesian model are illustrated on the left-hand side of Fig. 4.2.

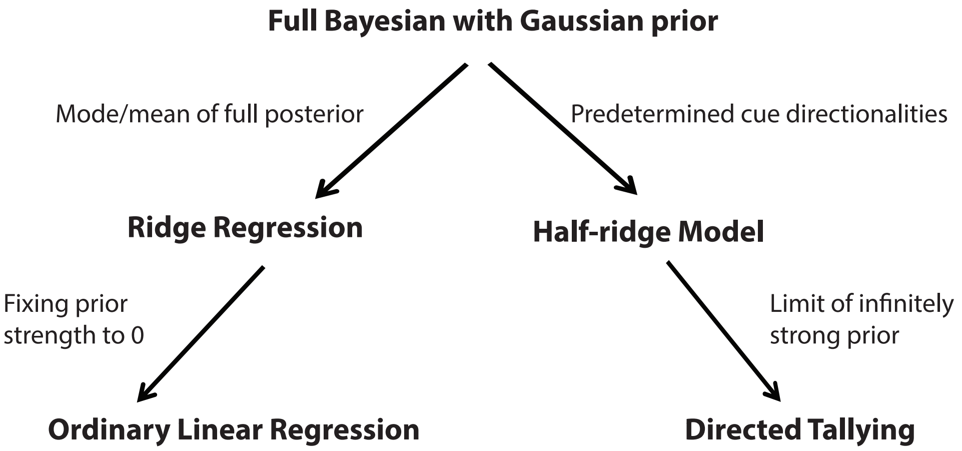

Full Bayesian with Gaussian prior
Mode/mean of full posterior
Predetermined cue directionalities
Ridge Regression
Half-ridge Model
Fixing prior strength to 0
Limit of infinitely strong prior
Ordinary Linear Regression
Directed Tallying

Figure 4.2: Formal relationships among full Bayesian regression, ridge regression, ordinary least-squares linear regression, the Bayesian half-ridge model, and the directed tallying heuristic. The lower-right arrow represents the main contribution of this Chapter: The directed tallying heuristic (explained below) is a limiting case of Bayesian inference (here, the half-ridge model) with an infinitely strong prior.

4.3.1 Half-ridge model and tallying

Our Bayesian derivation of the tallying heuristic extends ridge regression by assuming the directionalities of the cues (i.e., the signs of the true weights) are known in advance. For example, being higher in the league standings will, if anything, make a team more likely (not less) to win a given match. This assumption is concordant with how the tallying heuristic was originally proposed in the literature (Dawes, 1979) (Box 2.3.3). We refer to this definition of the tallying heuristic as *directed tallying* in order to differentiate it from the version of the tallying heuristic that learns cue directionalities from the training data (Czerlinski et al., 1999). Thus, we define the prior for each weight as half-Gaussian, truncated at zero (right-hand side in Fig. 4.2), and we refer to this Bayesian model as the *half-ridge* model. Formally, the joint prior is defined by

$$\begin{aligned} \mathbf{w} &\sim \mathscr{N}(0, \mathbf{\Sigma})_{|\mathbf{w} \in \mathscr{O}} \\ \mathbf{\Sigma} &= \eta^2 I, \end{aligned} \tag{4.2}$$

where $\mathbf{\Sigma}$ is the covariance matrix among the weights (prior to truncation) and η^2 determines the variance for each weight. The restriction notation, $|\mathbf{w} \in \mathscr{O}$, indicates we truncate the distribution to one orthant $\mathscr{O} \subset \mathbb{R}^m$, defined by the predetermined directionalities of the cues, and renormalize. For example, if the cues were assumed all to have positive (or null) effects on the outcome, then $\mathscr{O}$ would equal $[\mathbf{w} \in \mathbb{R}^m | \forall i, w_i \geq 0]$. Linear regression with an untruncated Gaussian prior (L2 regularization) yields a Gaussian posterior for the weights, having mean

$$\left(\mathbf{X}^T\mathbf{X} + \sigma^2\mathbf{\Sigma}^{-1}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{4.3}$$

and variance

$$\sigma^2\left(\mathbf{X}^T\mathbf{X} + \sigma^2\mathbf{\Sigma}^{-1}\right)^{-1}. \tag{4.4}$$

The posterior for the half-ridge model inherits the same truncation from the prior above (Eq. 4.2) and is otherwise unchanged except for renormalization:

$$\begin{aligned}\mathbf{w}|\mathbf{X},\mathbf{y} \sim \mathscr{N}\Big(&\left(\mathbf{X}^T\mathbf{X} + \sigma^2\mathbf{\Sigma}^{-1}\right)^{-1}\mathbf{X}^T\mathbf{y}, \\ &\sigma^2\left(\mathbf{X}^T\mathbf{X} + \sigma^2\mathbf{\Sigma}^{-1}\right)^{-1}\Big)_{|\mathbf{w}\in\mathscr{O}}.\end{aligned} \tag{4.5}$$

The important question is what happens to this posterior as the prior becomes arbitrarily strong, that is, as $\eta \to 0$. To understand how the posterior behaves as the prior becomes arbitrarily strong, we can rescale the weights by $1/\eta$ and substitute Eq. 4.5 to rewrite the posterior as

$$\begin{aligned}\frac{\mathbf{w}}{\eta}\bigg|\mathbf{X},\mathbf{y} \sim \mathscr{N}\Big(&\eta\left(\eta^2\mathbf{X}^T\mathbf{X} + \sigma^2 I\right)^{-1}\mathbf{X}^T\mathbf{y}, \\ &\left(\frac{\eta^2}{\sigma^2}\mathbf{X}^T\mathbf{X} + I\right)^{-1}\Big)_{|\mathbf{w}\in\mathscr{O}}.\end{aligned} \tag{4.6}$$

Rescaling the weights has no impact in a binary comparison task, so we can work with the distribution of $\mathbf{w}/\eta$ in place of that of $\mathbf{w}$. The convenience of this rescaling is that the resulting distribution obeys a simple convergence:

$$\frac{\mathbf{w}}{\eta} \xrightarrow{d} \mathcal{N}(0,I)_{|\mathbf{w}\in\mathcal{O}} \text{ as } \eta \to 0 \tag{4.7}$$

conditional on $\mathbf{X}$ and $\mathbf{y}$ (where $\xrightarrow{d}$ indicates convergence in distribution). Consequently, the rescaled weights all converge to the same value in the limit, namely

$$\lim_{\eta\to 0} \mathbb{E}\left[\frac{w_i}{\eta}\middle|\mathbf{X},\mathbf{y}\right] = \pm\sqrt{\frac{2}{\pi}}, \tag{4.8}$$

with signs determined by each cue's assumed directionality. In particular, for any two weights j and k, their ratio converges to unity:

$$\lim_{\eta\to 0} \frac{\mathbb{E}\left[w_j|\mathbf{X},\mathbf{y}\right]}{\mathbb{E}\left[w_k|\mathbf{X},\mathbf{y}\right]} = 1. \tag{4.9}$$

Hence, just as with increasing the penalty parameter in regular ridge regression, strengthening the prior in the half-ridge model shrinks the weights toward zero (Equation 4.7). However, the ratios of the weights — that is, the relative inferred strengths of the cuesall converge to unity (Equation 4.9). Therefore, the optimal decision-making strategy under the Bayesian half-ridge model converges to a simple summation of the predictors – that is, a tallying strategy. To understand this result intuitively, refer to Eq. 4.3 and Eq. 4.4 and note that the posterior mean and posterior variance both scale with the priors covariance matrix $\mathbf{\Sigma}$ as $\mathbf{\Sigma}$ approaches 0 (i.e., as the precision of the prior approaches ∞). Thus, the mean of each weight goes to 0 faster than its standard deviation, or in other words the coefficient of variation goes to 0. That fact is not consequential when the directions of the cues are unknown, but it is significant when the cue directions are known. In the latter case, the signs of $\mathbf{w}$ become the most important information the learner has.

Note that, under this limit, the model becomes completely invariant to the training data. In particular, it ignores how strongly each cue is associated with the outcome in the training set (i.e., magnitudes of cue validities) as weights reduce to zero (Equation 4.7). At the other extreme, as the prior becomes extremely weak, i.e.,

$\eta \to \infty$ and $1/\eta^2 \to 0$, the Bayesian half-ridge model converges to a full regression model akin to ordinary linear regression in that it differentially weighs cues, e.g., more predictive cues receive higher weights than less predictive cues, however, this will be less and less the case as the prior's strength approximates the directed tallying heuristic. In conclusion, the half-ridge model demonstrates how the directed tallying heuristic can be derived as an extreme case of a Bayesian prior, due to assumptions about the distributions of the weights in the environment.

4.4 Computational Study 5: Heuristics vs. Intermediate Models

From a Bayesian perspective, the model that fares best on a given decision task should be the one with a prior most closely matching the data's generative process. In many decision environments, cues differ in their predictiveness, but these differences are not arbitrarily large (i.e., the cue weights are not drawn uniformly from all real numbers). An advantage of the Bayesian half-ridge framework is that it specifies a continuum of models between the extremes of linear regression and the directed tallying heuristic. Thus, we expect that for many environments, the best-performing model should lie somewhere between these two extremes. We also expect that the best-performing model should not change with different training set sizes (cf. Fig. 3.4), because - unlike the frequentist phenomenon of bias-variance tradeoff - in a Bayesian setting, the optimal model is guaranteed to find the optimal tradeoff between prior and likelihood, for any sample size. We simulated the Bayesian half-ridge model on the original 20 datasets that have been used to compare heuristic and regression approaches (Czerlinski et al., 1999).

4.4.1 Methods

The goal of this simulation was to assess the generalization performance of the Bayesian half-ridge model as a function of its prior's strength, and as a function of training sample size. A full list of the 20 datasets is in Appendix B, and the key parameters in the simulation are listed in Table 4.1 below. For each dataset, we created all possible pairwise comparisons of the objects to generate the cue differences vectors and the binary criterion variable, encoding which of two objects is superior on a comparison (e.g., which of two houses has a higher sales price?). The Bayesian half-ridge model was cross-validated on each dataset by splitting the total set of pairwise comparisons randomly into training and test sets. The size of the training sample was varied between 10, 20 comparisons (small) and 115 comparisons (large), and the test set represented the complementary set of comparisons. As two of the datasets, Oxygen and Ozone, only have 91 and 55 object pairs in total respectively, the large training sample size of 115 was excluded for those datasets. For each training sample size, the cross-validation split into training and test set was repeated 1000 times and performance was averaged across all. Model weights were derived by fitting the posterior distribution in Eq. 4.5 to the training sample. The posterior samples were drawn from a truncated multivariate Normal distribution, and the truncation in Eq. 4.5 (the choice of orthant $\mathscr{O}$) depended on the actual cue directions in the full dataset, following the assumption that the cue directions are known in advance. We derived mean posterior weights under different values of the Bayesian prior's strength ($1/\eta^2$), as reported in Table 4.1. Next, the mean posterior weights were used to make predictions for the novel test sets. To compute predictive accuracy, we compared the model predictions to the actual outcomes on the criterion variable in each dataset.

Statistical Parameters in the Simulation

Number of objects	11 to 395
Number of pairwise comparisons	$N = 55$ to $N = 77815$
Number of cues	$m = 3$ to $m = 18$
Class variable (e.g., which house had the higher actual sales price?)	Binary, ±1
Absolute correlation between cues (averaged over cue pairs)	range = 0.12 to 0.63, mean = 0.31, median = 0.28, sd = 0.14
Training sample size	10, 20, 115
Test sample size	$N-10$, $N-20$, $N-115$
Number of cross-validation repetitions	1000
Error variance	$\sigma_\varepsilon^2 = 1$
Strength of prior	$1/\eta^2$ = [1000000, 100000, 1000, 700, 330.08, 156.81, 74.50, 35.39, 16.81, 7.99, 3.80, 1.80,0.86, 0.41, 0.19, 0.09, 0.03, 0.01, 0.001, 0.0001, 0.00001]

Table 4.1: Parameters in the simulation of the 20 datasets as presented in Fig. 4.3. A full description of the 20 datasets is in Appendix B.

4.4.2 Results

Our main prediction held across simulations. With small and large training sample sizes, the performance peak could be found for an intermediate model, i.e., with medium-strength prior (Fig. 4.3). Note that an approximately infinitely strong prior on the far right of each graph (small values of η) corresponds to the directed tallying heuristic, and a prior strength of zero (in the limit of $\eta \to \infty$) corresponds to the regression model. Interestingly, we find that the regression model outperforms the directed tallying heuristic in all cases. This difference to past less-is-more results arises because cue directions are not learned in these simulations with the directed tallying heuristic, and therefore there is no opportunity for the more flexible regression model to misestimate the cue directions from the data. However, we replicate previous less-is-more results (Czerlinski et al., 1999) when comparing models that estimate cue directions from the training set (next Chapter 5). The key finding reported in Fig. 4.3 is that intermediate half-ridge models outperformed tallying in all 20 datasets, independent of training sample size. This suggests that ignoring

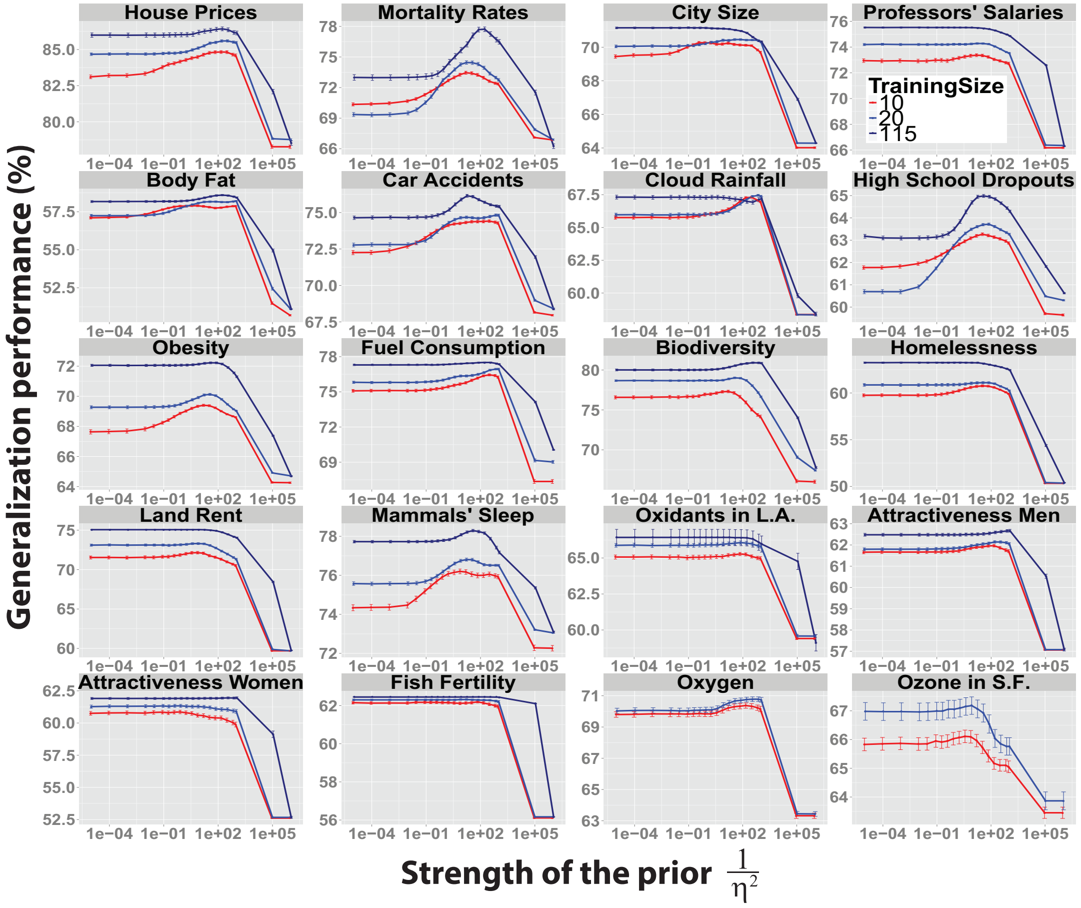


Figure 4.3: Generalization performance of the Bayesian half-ridge model by training sample size and as a function of the strength of the prior for 20 datasets for which heuristics have been previously evaluated (Czerlinski et al., 1999). The abscissa represents the strength of the prior on a logarithmic scale, and the ordinate represents the predictive accuracy of the model on test comparisons. Note that an approximately infinitely strong prior on the far right of each graph (small values of η) corresponds to the directed tallying heuristic. Intermediate models, i.e., with a medium-strength prior, performed best in all datasets regardless of training sample size. Error bars represent ± SEM. Because the Oxygen and Ozone data sets contain less than 115 object pairs in total, the 115 cases sample is not included (see Methods for details).

information was never the best solution in these simulations. Instead, the best performing model used all the information in the training data, but down-weighting it via the influence of the prior.

Note that in contrast to the performance reversal with training sample size observed in the frequentist case (Fig. 3.4), in this case the most flexible model is no longer guaranteed to perform best when training sample size is large, as intermediate priors performed best regardless of training sample size. To assess whether the accuracy maximizing prior did not change as a function of the sample size, we performed a statistical test. We regressed the maximum prior strength onto datasets and training sample size and find that the slope for training sample size is not significant (i.e., beta = -0.009, p = 0.977), suggesting the best performing model stays approximately in the same location across training sample sizes. For 7 out of the 20 datasets, the difference in peak prior strength between training sample sizes was exactly zero, meaning the location of the accuracy maximizing prior did not change.

4.5 TTB as a limiting case of lasso regression?

Given that ridge regression (L2 regularization) yields tallying, one might wonder whether a strong prior of a different functional form might yield the TTB heuristic. In particular, *lasso regression* (L1 regularization) (Ripley, 2007) is known to produce sparsity in cue selection (i.e., many weights are estimated as zero), and thus might be expected to yield TTB in the limit. Just as ridge regression's penalty term can be interpreted as a Gaussian Bayesian prior on the weights, lasso's penalty term can be interpreted as a Laplacian Bayesian prior on the weights.

However, instead, we find that lasso regression also converges to tallying in the limit when the cue directionalities are known a priori (mathematical derivation is in Appendix E). Thus, L1-regularized truncated regression converges to tallying, just like in the ridge regression case (L2 regularization). This result highlights the robustness of tallying arising as a limiting case of Bayesian inference under a variety of different priors and norms (L1 norm and L2 norm). Given this formal result, we motivate the TTB heuristic as an extreme Bayesian prior with a different approach in the next Chapter via the COR model, which is able to capture the sequential nature of the TTB heuristic.

4.6 Discussion

We find that a tallying heuristic and a full-information model can be linked through Bayesian inference, in which tallying, that deliberately ignores information, is equivalent to the limit of an infinitely strong prior. We showed that by relying on a prior that biases all weights toward zero as in ridge regression, we could derive the tallying heuristic in the limit. This work also demonstrates limitations of the absolute less-is-more assumptions: models with intermediate priors, which use all available information in the training data, always performed better than heuristics, and full regression models.

A central message of this work is that ignoring information is never more. This stands in stark contrast to the (absolute) less-is-more claims (Gigerenzer & Brighton, 2009; Gigerenzer et al., 1999; Tsetsos et al., 2016) which argue that when heuristics outperform, they cannot be improved upon with a strategy that relies on more information. Our findings of intermediate models always outperforming heuristics shows that including the information while down-weighting is optimal. This argument will become even clearer in the next Bayesian framework (Chapter 5), where less-is-more effects can be observed but intermediate models still perform best throughout. Heuristics may work well in practice because they correspond to an infinitely strong prior that is oblivious to the training data, but they will usually be outperformed by a prior of finite strength that leaves room for learning from experience. That is, the claim that one can do better with heuristics by throwing out information entirely (corresponding to a strong prior, e.g., $1/\eta^2 = \infty$ in the half-ridge model) rather than using it, is false. That is because the optimal solution always uses all of the information (e.g., a finite value of $1/\eta^2$), but it combines that information with the appropriate prior. In contrast, no amount of data can overcome the heuristics' inductive biases (as can be seen in Fig. 4.3). The tallying heuristic is defined to entirely ignore relative cue weight magnitudes, unlike the intermediate half-ridge models.

Another insight arising from this and the next Chapter is that heuristics do not work well because of their simplicity but due to their bias. This represents a strong contrast to the less-is-more claims (Gigerenzer & Brighton, 2009; Gigerenzer et al., 1999; Tsetsos et al., 2016) which argue that heuristics work well because of their ignorance of information. However, the evidence in this Chapter and the next suggest differently: Heuristics may work well because they embody a prior that approximates the optimal prior. The variation of bias along the continuum of the Bayesian prior's strength means our findings tell us that models which are highly biased (similar to heuristics) but not as simple as heuristics (no ignorance of information) are optimal. The Bayesian conceptualisation allows a de-confounding of simplicity and inductive bias in heuristics, as only the extreme end case (i.e., prior strength of $1/\eta^2 = \infty$) on the Bayesian continuum is simple (due to its complete ignorance of information information). The Bayesian approach offers the novel explanation that heuristics do well because they approximate a strong inductive bias.

4.6.1 Psychological Implications

While the core intended contribution is formal, this work has implications for psychology. In the psychological literature, heuristics have been repeatedly pitted against full-information algorithms (Chater et al., 2003; Czerlinski et al., 1999; Katsikopoulos et al., 2010) that differentially weight the available information. The current work indicates that the best-performing model will usually lie between the extremes of ordinary linear regression and fast-and-frugal heuristics, i.e., at a prior of intermediate strength. Between these extremes lie a host of models with different sensitivity to cue-outcome correlations in the environment. One question for future research is whether heuristics give an accurate characterization of psychological processing, or whether actual psychological processing is more akin to these more complex intermediate models. On the one hand, it could be that implementing the intermediate models is computationally intractable, and thus the brain uses heuristics because they efficiently approximate these more optimal models. This case would coincide with the view from the heuristics-and-biases tradition of heuristics as a tradeoff of accuracy for efficiency. On the other hand, it could be that the brain has tractable means for implementing the intermediate models (i.e., for using all available information but down-weighting it appropriately). This case would be congruent with the view from ecological rationality where the brains inferential mechanisms are adapted to the statistical structure of the environment. However, this possibility would also suggest a reinterpretation of the empirical evidence used to support heuristics is necessary: heuristics might fit behavioral data well in many instances because they closely mimic a more sophisticated strategy used by the mind. This would indicate future research should identify these more sophisticated intermediate strategies and investigate whether they are better at predicting people's behaviour than traditional heuristics. We believe that this endeavour will present a natural next step for extending this work in the future.

Although speculative at this point, some recent research also points towards evidence for strategies more akin to the intermediate models at the psychological level. Bergert and Nosofsky (2007) showed that people's pattern of behaviour matches a probability mixture version of TTB rather than the strong version of TTB or a full-information strategy (RAT). The authors hypothesized that the strong forms of TTB and RAT (a weighted additive model similar to linear regression, Lee and Cummins (2004)) would be psychologically implausible, and formulated generalizations of these models which allowed for subjective cue weighting, probabilistic orders of cue inspection, and noisy decision making. Under conditions in which the standard TTB and RAT strategies yielded equivalent decisions, reaction times and the estimated cue weights suggest that most participants adopted a generalized TTB strategy, in which cue weights deviate from the ones in the traditional heuristic. These results suggest participant's pattern of cue weighting challenges the strong versions of both RAT and TTB and provides evidence for mixture models. There is also various evidence showing people do not adhere to the strong from of TTB which entirely ignores information, but instead rely on the full information with individual weighting structures (Newell & Shanks, 2003; Newell, Weston, & Shanks, 2003; van Ravenzwaaij, Moore, Lee, & Newell, 2014) (discussed further in Chapter 5). However, it needs to be stressed that at this point these ideas are speculative and it needs to be established empirically whether people choose to rely on the full information and down-weighting it, rather than ignoring it.

Chapter 6 will in fact look at whether people rely on heuristics or a full-information model in a decision making and active learning task, which will shed some light onto what information people choose to include and drop during learning and decision making.

4.6.2 Discovery of new heuristics?

The general framework for creating a model continua between prominent strategies may apply more broadly than is presented in this thesis. Concretely, the framework's logic may itself be a good discovery heuristic for identifying novel heuristics that work well, i.e., for different environmental structures and Bayesian priors. The same approach could yield alternative heuristics that have value compared to the established ones, by creating Bayesian prior continua between full-information models and heuristics, and taking a prior to extreme values (i.e., infinite prior strength and zero prior strength). However, crucially the choice of prior needs to be informed by the structure of the environment and the strategy. One possible example may be the idea that a Gaussian process regression can result in a Nearest Neighbour heuristic with a strong prior (Gramacy & Lee, 2008). When taking a Gaussian Kernel's lengthscale parameter to extremes (i.e., zero and approximating infinity), on the one end the predictions of the Gaussian kernel would converge onto a strategy that only relies on nearby observations such as in the nearest neighbour classification. On the other end of the extreme (when the lengthscale parameter approximates infinity), its results would become equivalent to those of linear regression. Hence, potentially other interesting new heuristics that people might use could be uncovered in the limit of strong Bayesian priors. Nevertheless, this is an idea that yet needs to explored in the future.

4.6.3 Integration of Probabilistic and Heuristic Approaches

By taking a Bayesian inference perspective as a lens on heuristics, we could arrive new insights about heuristics that were invisible before, e.g., highlighting the strategies along the prior's continuum which are usually ignored. However, by providing a Bayesian explanation for heuristics, we also offer an integration of probabilistic approaches with heuristics. For decades, probabilistic approaches to cognition and heuristic approaches evolved in parallel due to historic developments in the field (e.g., due to little integration between computational-level theories and algorithmic-level theories, and due to both the heuristics-and-biases model and the fast-and-frugal model perceiving heuristics as incompatible with Bayesian inference). The vast amount of previous literature pitted heuristics against rational Bayesian models in modelling competitions, e.g., (Katsikopoulos et al., 2010; Martignon & Hoffrage, 2002), and the underlying implicit assumption was that heuristics are necessarily competitors to Bayesian inference models. In contrast, we show that heuristics are equivalent to Bayesian inference for extremely strong priors. Thereby the current research helps to move closer the different levels of analysis in cognitive science, in particular the computational level and the algorithmic or process level (Marr, 1982a).

There have been other more recent approaches looking at the compatibility between psychologically plausible processes and probabilistic models of cognition (Brown & Steyvers, 2009; Daw & Courville, 2008; Griffiths et al., 2015; Jones & Love, 2011; Lee & Cummins, 2004; Lieder et al., 2017; Sanborn et al., 2010; Scheibehenne, Rieskamp, & Wagenmakers, 2013; van Ravenzwaaij et al., 2014). While these investigations are interlinked with our own, most of that work has focused on finding algorithms that approximate Bayesian models, whereas we have taken the opposite approach. The relationship between our work and some of these recent formal approaches for heuristics will be discussed in the General Discussion.

4.6.4 Possible Extensions of the Model and Simulations

In the current half-ridge simulations, training sets were defined by sampling a subset of all possible comparisons (i.e., object pairs). As discussed in Chapter 3, in some past work training sets were defined by sampling a subset of objects instead and then training on all pairs within the sampled subset (Czerlinski et al., 1999). For our purposes, we could find no compelling reason to prefer either method. However, as we have seen, this sampling method can make a difference in the relative performance of heuristics and regression models (Fig. 3.6). To determine whether our results reported here would be dependent on this sampling decision, we compared both sampling methods. In short, the qualitative pattern of results is not dependent on the sampling method. When sampling objects rather than comparisons, we varied the training sample size between sampling 5, 7 and 16 objects, which correspond to 10, 21 and 120 possible comparisons for the training sets, respectively.

We chose these training sample sizes to roughly match the training sample sizes used for the half-ridge simulations when sampling comparisons (i.e., 10, 20 and 115 training cases in Fig. 3). The qualitative pattern of results between the directed tallying heuristic and regression was the same. Performance of all models was lower overall by a few percent in accuracy when sampling objects, confirming the results from Chapter 3. Additionally, models with weaker priors (i.e., closer to regression) showed a slightly larger drop in performance under object sampling than did models with stronger priors (i.e., closer to the heuristics). Thus, sampling objects gives the directed tallying heuristic a small advantage over ordinary regression for the training sample sizes considered here. However, this performance advantage was never enough to reverse the performance patterns in Fig. 4.3, and the location of the performance peak was in approximately the same location under both sampling methods.

Typically in past work, continuous cue values were discretized at the median in the original 20 datasets to form binary variables with the cue value indicating which object in the pair was superior along that dimension (e.g., Chater et al. (2003); Czerlinski et al. (1999); Katsikopoulos et al. (2010)). The dependent variables were also binary arising from the comparisons and detail which object has the higher criterion value on each comparison. As such, logistic regression rather than least squares regression would be more appropriate for binary outcome variables. However, we chose least squares regression to be consistent with past work on the 20 heuristic datasets (Czerlinski et al., 1999) in order to be able to replicated findings, and build a continuum between these models traditionally used in the heuristic literature. To confirm that the choice of link function was not critical, we re-analysed the data using logistic regression. We found no significant difference between the two approaches on the datasets considered. In particular, we considered the 20 datasets across all training sample sizes (10, 20 and 115) to compare both regression approaches. Ordinary least squares regression had an average generalization performance of 0.687 with sd = 0.08 (where 1 corresponds to 100% correct out-of-sample inferences and 0 corresponds to 0% correct inferences), across all training sample sizes (10, 20 and 115 training cases) and across all 20 environments. Logistic regressio's generalization performance was 0.682 with sd = 0.08, across all training sample sizes (10, 20 and 115 training cases) and across all 20 environments as used in Fig. 4.3. The mean difference in performance scores between least squares and logistic regression, across all training sizes and datasets, was 0.0049 with sd = 0.008. A possible future model extension could be to build the model continua within a (Bayesian) logistic regression framework rather than the linear one used here. We believe this would be a useful avenue for further research, as part of a general program to build on the theoretical ideas introduced here to develop new, more powerful decision algorithms and to further link heuristics to other modeling approaches. For now, we note that the linear models used here support the main conclusions just as well although they are not ideally tailored to the task being analysed (i.e., where criterion values are binarised).

4.6.4.1 Robustness of the intermediate peak?

The performance peak was found for intermediate prior strength in all 20 datasets. As predicted, these findings are in contrast to the frequentist case (e.g., Fig. 3.4, Fig. 3.5), where the training sample size results in a cross-over of performances. We performed a statistical sign test to ensure that the performance peak was not moving left or right with a different training sample sizes, which revealed that best performing prior strength stayed approximately constant across training sample sizes (Section 4.4.2). Despite these findings, we emphasize that the optimal prior is expected to stay in the same location only to the extent that the model is correctly specified. The fact that the performance peaks in our simulations were not significantly affected by training sample size suggests that the degree of model misspecification was not sufficient to reveal itself in this way. In contrast, the Bayesian model developed in the next Chapter is more clearly misspecified, as will be seen. However, whether the models are misspecified or not does not impact the mathematical fact that the Bayesian model converge to the heuristics under strong priors, or the empirical fact that intermediate models outperformed the heuristic on real datasets.

4.6.5 Practical Implications

Knowing when heuristics can be superior (though not optimal) to more costly flexible strategies has useful implications for various real-world domains where quick actions are required. For example, soldiers at a checkpoint must quickly decide whether an approaching car contains terrorists or civilians (Keller & Katsikopoulos, 2016) and recent research suggests a simple heuristics can aid quick decision making in these scenarios. Likewise, doctors need to decide whether to assign a patient to a coronary care unit or a regular nursing bed (Marewski, Gaissmaier, & Gigerenzer, 2010). The ability of the Bayesian approach to down-weight and up-weight information (e.g., in estimating regression weights) means the model places both costly flexible strategies and heuristic strategies on equal footing which allows for an analysis of when one model will outperform the other and quantify by how much. In that way, the Bayesian approach can be applied to real-world data from different situations and through predictive modelling the performance of different strategies can be determined a priori. These kinds of analyses could lead to developments of new decision tools that are helpful in medical contexts, legal contexts, or military contexts to name a few. Our formulation places the heuristics within the same framework as other Bayesian models with the same prior, varying only in strength. The optimal prior strength (i.e., the best model within the continuum) will vary from one domain to another, but other than this choice of free parameter, the half-ridge model can be straightforwardly applied in any settings where the heuristic can (in fact, the former models are not limited to binary comparison tasks, whereas the heuristics are). Should a practicing researcher abandon heuristics such as the tallying heuristic for the more complex intermediate framework? Whether the intermediate models are enough better to merit use than the heuristics will depend on the context and the situation a researcher tries to model. For example, researchers in machine learning are always interested in gaining higher accuracy with new regularization methods to avoid overfitting in generalizing to new data, and hence an intermediate method that is able to improve upon existing simpler algorithms would be regarded as essential. Or else, in applied contexts, in professions where accuracy is central as errors can have devastating consequences, such as the astronautical engineering profession (where astronauts cannot make a mistake in replacing machine parts on the international space station for example), the method that results in higher accuracy would always be preferred over the method that is quicker but less accurate (such as heuristics). To give another example, in the medical domain (where a patient's life may be at stake), a machine learning method that recognizes a cancer with higher accuracy but requires more computation should be preferred to a quicker and less accurate method. However, in other contexts, when there is no time for more computation and integrating the full information, a heuristic that drops most information may be more useful under time pressure, such as when making decisions at war in the example above.

4.6.6 Extending the half-ridge model to an encompassing Bayesian framework

In this Chapter, we showed that the tallying heuristic could be derived as an extreme Bayesian prior by placing a truncated Gaussian prior on the weights. This resulted in a Bayesian continuum that is oblivious to the training data in the limit. We also attempted to derive the TTB heuristic as the limit of an extreme Bayesian Laplacian prior, i.e., lasso regression. The reasoning was that *lasso regression* (Ripley, 2007) is known to produce sparsity in cue selection. However, instead, lasso regression also converged to tallying in the limit when the cue directionalities are known a priori. Instead, in the following Chapter, we take a different approach to construct our second Bayesian model which unifies both TTB and tallying with ordinary linear regression. This Bayesian inference model (COR) is based on the key observation that, unlike linear regression, both TTB and tallying rely on isolated cue-outcome relationships (i.e., cue validity) that disregard covariance information among cues. The COR model will extend the current Bayesian half-ridge model with an inference model that is not oblivious to the training data, but instead learns cue directions and cue weights from the training data.

Chapter 5

Heuristics as Bayesian inference - The COR model

All models are wrong, but some are useful.

-GEORGE E. P. BOX

5.1 A Bayesian explanation for why less is more

Given the formal result of Chapter 4 that lasso regression also yields tallying, we take a different approach to unify both TTB and tallying with ordinary linear regression. One key observation is that, unlike linear regression, both TTB and tallying rely on isolated cue-outcome relationships (i.e., cue validity, Fig. 5.1*B*) that disregard covariance information among cues. We use this insight to construct our second Bayesian model, which contains a prior that at one extreme suppresses all information about cue covariance but retains information about cue validity. We refer to this model as Covariance Orthogonalizing Regularization (COR), because our regularization method essentially behaves so as to make the cues appear orthogonal to each other. The strength of the prior yields a continuum of models defined by sensitivity to covariation among cues, which smoothly vary in their mean posterior weight estimates from those of ordinary linear regression to weights that are linear transforms of the heuristics cue validities (derivations below).

This Chapter extends the Bayesian half-ridge model with a second Bayesian model that learns cue directions and cue weights from the training samples, thereby being able to exhibit less-is-more effects (in contrast to the previous Chapter), i.e., tallying or TTB outperforming linear regression. Yet, crucially, we find again that even in less-is-more situations, the strategies that perform best along the Bayesian continuum are intermediate models. Hence, the goal of this Chapter is identical to the previous Chapter. We show that while relative less-is-more holds, the Bayesian framework provides a different message for *absolute* less-is-more (Box 2.4.2): We find heuristics can in fact be improved upon with strategies that rely on all information but appropriately down-weight it. While no other previous study on heuristics has addressed the absolute less-is-more meaning in the past (Gigerenzer & Brighton, 2009), the current Bayesian framework allows for a continuum between heuristics and regression which makes this possible.

As with Chapter 4, this Chapter contributes to a novel Bayesian explanation of why heuristics work. We do so by formulating heuristics as the limiting cases of strong Bayesian priors. In the COR model, the prior varies sensitivity to covariation among predictors (Rieskamp & Dieckmann, 2012). Crucially, the heuristics represent an extreme case on this Bayesian prior, corresponding to entirely ignoring covariance rather than including it and down-weighting it. Parametric variation of the covariance prior's strength results in a continuum that contains, as limiting cases, both heuristics (TTB and tallying heuristic) as well as (ordinary least-squares) linear regression. Although the Bayesian model contains tallying and TTB as special cases, a crucial difference is that along the continuum the Bayesian account regulates covariance estimation, but never entirely dispenses with it. Similar to the half-ridge model, we investigate the performance of the Bayesian COR model in real-world environments and replicate previous less-is-more effects. Importantly, we find that the best-performing models for the real-world problems tested were intermediate models that do not entirely ignore covariance among predictors but that nonetheless down-weight this information via the influence of their priors. Hence, the findings in this Chapter add to the Bayesian explanation proposed in the preceding Chapter: Heuristics may excel because they have a strong inductive bias which approximates the optimal prior for many environments. Again, it needs to be noted that while the framework in this Chapter has interesting implications for psychological processes, it does not aim to model less-is-more effects at the psychological level, and primarily provides a formal computational account. Implications for psychology will be discussed.

This Chapter is structured as follows: First, I discuss why covariance is a sensible factor to rely on for building a Bayesian model continuum between heuristics and regression. Next, I introduce the Bayesian learning model based on covariance sensitivity. A first computational study will assess the convergence of the Bayesian model to the heuristics in the limit of a strong prior. A second computational study will assess the generalization performance of the Bayesian model in the classic 20 real-world environments (Czerlinski et al., 1999). Finally, the theoretical and psychological implications, and limitations of the model are discussed. All mathematical derivations for the COR model in this Chapter were derived in collaboration with my collaborator Matt Jones (University of Colorado, Boulder) and my supervisor Brad Love (UCL).

5.2 Linking heuristics and regression through a covariance prior

Why do we focus on covariance as a prior for linking heuristics and full-information models? As explained at the beginning (Section 2.3.3), a central difference between heuristics that rely on cue validity and full-information models, such as linear regression, is that regression coefficients naturally take into account covariance as part of the parameter estimation, however cue validity naturally ignores any covariance among cues.

For example, in Fig. 5.1*B*, league position and number of goals scored are highly correlated. Although such covariances naturally arise and can be meaningful, the cue validities used by the tallying and TTB heuristics completely ignore them (Martignon & Hoffrage, 1999). Instead, cue validities assess only the probability with which a single cue can identify the correct alternative (e.g., which team won the

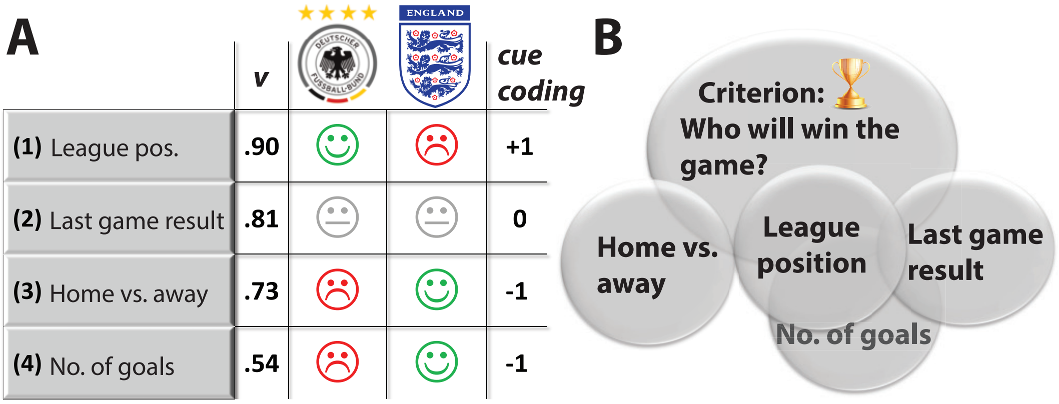

A
v
cue coding
(1) League pos. .90 ☺ ☹ +1
(2) Last game result .81 😐 😐 0
(3) Home vs. away .73 ☹ ☺ -1
(4) No. of goals .54 ☹ ☺ -1
B
Criterion: Who will win the game?
Home vs. away
League position
Last game result
No. of goals

Figure 5.1: Illustrative example of a binary prediction task. (***A***) Predicting whether Team Germany or England will win is based on four cues: league position, last game result, home vs. away match, and recent goal scoring. Cue validities (v) reflect the relative frequency with which each cue makes correct inferences across many team comparisons (see definition in this section below). Smiley and frowning faces indicate which team is superior on each cue, whereas a grey face indicates the two teams are equal on that cue. A cue is coded +1 when it favors the team on the left (Germany), -1 when it favors the team on the right (England), and 0 when the teams are equal along that cue. (***B***) Irrespective of cue validity, cues can co-vary (illustrated by overlap) with the criterion variable but also with each other. The heuristics considered here ignore this covariance among cues.

football match), derived as the proportion of correct inferences made by each cue alone across the set of binary object comparisons (Equation 2.1). When two cues co-vary highly, they essentially provide the same information, but heuristics ignore this redundancy and treat them as independent information sources. An intuitive example may be having two of copies of the same newspaper as input, where the second newspaper does not really provide news anymore.

However, this simple difference with respect to covariance estimation between full-information strategies and heuristics may play an important role for their differential behaviour and success. For example, some recent evidence indicates that heuristics can generalize well compared to complex models under conditions of high co-variation among cues (Dieckmann & Rieskamp, 2007; Rieskamp & Dieckmann, 2012). One explanation for these findings is the heuristics complete ignorance of co-variation among cues may function as a strong form of bias that can result in better generalization to novel, unseen samples. Although the relationship between the success of heuristics and external covariance in the environment is complex, the fact that heuristics deliberately ignore covariance in their cue weight estimates appears to be an important factor that differentiates heuristics from full-information models. Evidence supporting this notion was developed by Gigerenzer and Brighton (2009), where the authors compared regular TTB to a version of TTB that orders cues by conditional validity, thereby taking into account covariance among cues. They pitted both versions of TTB against classical full-information models that do take into account covariance (Fig. 5.2). The authors claimed that the conditional TTB's cue weight estimation is closer to the more resource-intensive machine learning models, i.e., such as the C4.5 and CART model. Fig. 5.2 shows that, interestingly, while TTB outperforms all other full-information models for training sizes up to 50 objects, when the search rule of TTB is altered to conditional validity, its performance drops to the level of the other full-information models (bottom left panel). From these findings, the authors concluded that perhaps what helps the TTB heuristic is a deliberate disregard of cue covariance which can be adaptive in some environments. One possibility is that this disregard of covariance results in lower overfitting compared to models that estimate covariance among cues. Gigerenzer and Brighton (2009) say that the ignorance may be ”an adaptive processing policy when observations are sparse”, i.e., with smaller training sample sizes. This is a valuable insight, however, what they did not investigate is whether the heuristic was the optimal model when training sample size was small, or whether it could not have been improved upon by a model that does take into account the full information (covariance) (i.e., absolute less-is-more claim). In this particular instance, we do not know why ignoring covariance in the weight estimates was advantageous for the TTB heuristic. Nevertheless, these results highlight the important role that ignorance of covariance in cue validity estimates plays for heuristics. We make use of this fact that heuristics and regression are differently sensitive to covariance to create our Bayesian model continuum.

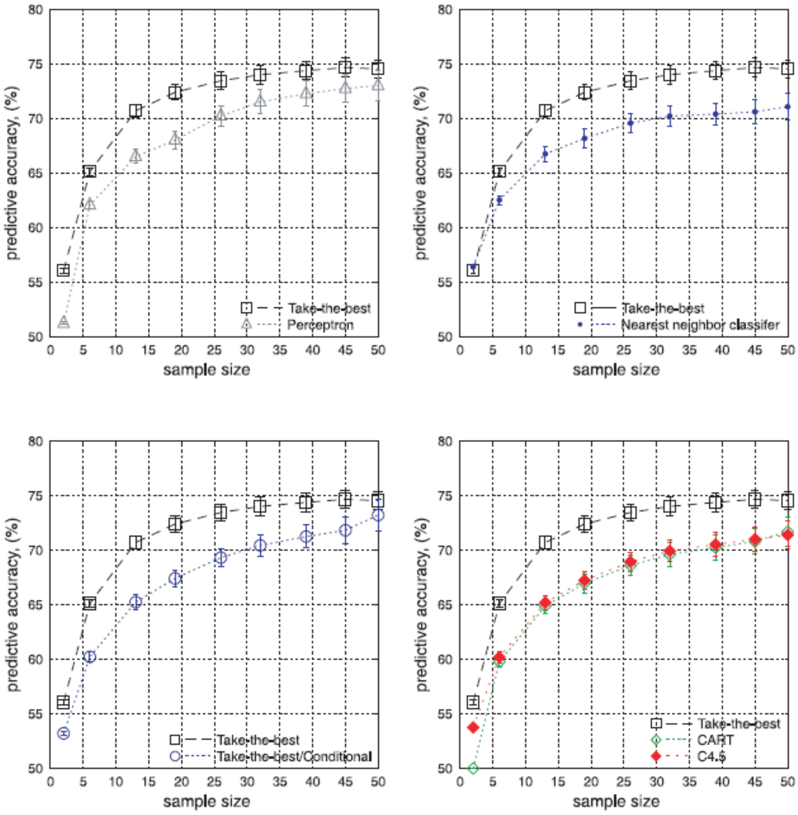


Figure 5.2: In the city size task, generalization performance of the TTB heuristic was superior to five full-information models as a function of the number of objects included in the training sample. TTB was compared to: a linear perceptron, i.e., similar to logistic regression (top left), the nearest neighbor classifier (top right), two tree induction algorithms, C4.5 and CART (classification and regression trees) (bottom right), a variant of TTB that orders cues by conditional validity (bottom left). Error bars are standard errors of means. Results are taken from Gigerenzer and Brighton (2009).

5.3 COR: A Bayesian framework relating regression and heuristics

We use the this insights to construct our second Bayesian model, which contains a prior that at one extreme suppresses all information about cue covariance but retains information about cue validity. The strength of the prior yields a continuum of models (Fig. 5.4) defined by sensitivity to covariation among cues, which smoothly vary in their mean posterior weight estimates from those of ordinary linear regression to weights that are linear transforms of the heuristics' cue validities. That is, importantly, cue validities, which ignore cue covariance, are essentially like single regression coefficients, that result from regressing the outcome variable onto a single cue alone (A proof of the linear relationship between cue validity and single-predictor regression weights is in the Chapter's Appendix 5.8). The mathematical derivations for the posterior mean of COR are in the Chapter's Appendix 5.9, however the essential model properties are explained next.

The COR framework also relies on a regularized regression method such as ridge regression in the previous Chapter (Equation 4.1). But, in contrast to ridge regression, we express the regression problem in multivariate terms by multiplexing the outcome m times (the number of predictors), which allows the model to capture the sequential nature of TTB. As shown in Fig. 5.3, every copy of the output receives input from every cue, and thus the weights can be represented as an $m \times m$ weight matrix $\mathbf{W}$. The model architecture of this matrix is illustrated in Fig. 5.3, were the solid arrows represent the diagonal weights (direct weights) and the dashed arrows represent the off-diagonal weights (cross-weights). Unlike in ridge regression, where the Gaussian prior shrinks all model weights toward zero, only the cross-weights (i.e., the off-diagonal elements) are penalized. Penalizing only the cross-weights has the effect that the strength of the prior $(1/\eta^2)$ modulates the model's sensitivity to covariation among cues, leading to the continuum in Fig. 5.4. In the limiting case, when the precision of the prior $1/\eta^2$ approaches infinity as $\eta \rightarrow 0$,

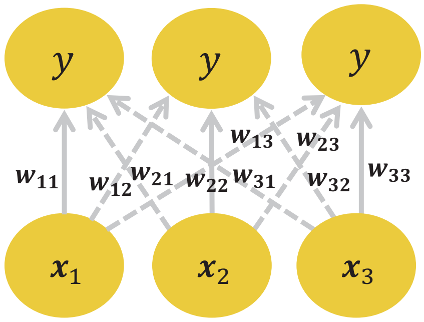


Figure 5.3: COR model architecture with $m = 3$ cues as presented in Fig. 5.4. All $\mathbf{y}$ variables are replicas of one another and contain the same outcome information. Dashed arrows are called *cross-weights*, and solid arrows are called *direct weights*. Weight indices refer to the weight matrix $\mathbf{W}$.

the cross-weights reduce to zero, visualized in the rightmost panel of Fig. 5.4 (see mathematical derivations in Appendix 5.9). In this limit, the posterior estimates for the direct weights are equivalent to cue validities as used by the heuristics, i.e., neglecting covariance information, up to a linear transformation (Equation 5.9).

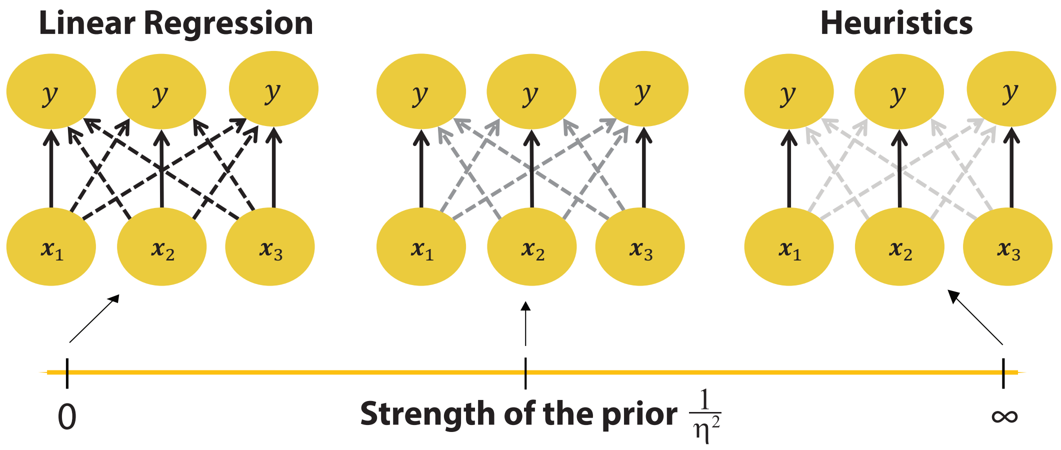


Figure 5.4: The prior of the COR model influences the posterior solution (i.e., the mean of the posterior on w) such that the model can encompass linear regression and the heuristics as extreme cases. In this example, there are $m = 3$ cues, i.e., $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}$ are vectors representing the explanatory variables containing information on all binary comparisons, where one vector entry x_{i1}, x_{i2}, x_{i3} pertains to the ith binary comparison. As in Fig. 5.1, cues can take values of $[-1, +1, 0]$, depending on whether the left or right option has a greater value on each cue, and likewise the criterion y_i can take values of +1 or -1, depending on the outcome of the comparison (e.g., which team won the match). In order to establish a continuum of covariation sensitivity, the criterion variable is multiplexed as many times as there are cues, i.e., m times. The result is a multivariate regression problem with a dependent matrix Y of m columns of identical criterion variables. We refer to the dashed arrows as *cross-weights*, and the solid arrows as *direct weights*, corresponding respectively to the off-diagonal and diagonal entries of the weight matrix $\mathbf{W}$. In an ordinary linear regression model with three predictors $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}$, the predictors' weights are determined by taking into account their covariances. In contrast, a model structure without any of the cross-weights would revert to three simple regressions with exactly one predictor each (either $\mathbf{x_1}$, $\mathbf{x_2}$, or $\mathbf{x_3}$). Therefore, when $1/\eta^2 = 0$, in analogy to ridge regression, the prior does not penalize the cross-weights, and the set of mean posterior weights to each copy of the criterion variable are equal to those of the ordinary linear regression solution (leftmost network). At the other extreme, when $1/\eta^2 \to \infty$, the cross-weights are shrunk to zero, and the knowledge captured in the direct weights becomes equivalent to that embodied by cue validities in heuristics that ignore covariation information (rightmost network). Between these two extreme values of $1/\eta^2$ lie models that are sensitive to covariation to varying degrees (middle network).

At the other extreme, when $1/\eta^2 = 0$, none of the cross-weights are penalized, and all weights in the model architecture are unchanged (leftmost network in Fig. 5.4). When $1/\eta^2 = 0$, every copy of $\mathbf{y}$ has the same posterior for its set of weights, and the mean (and mode) of this posterior is equal to the ordinary linear regression solution. That is, the set of weights are identical for all copies of $\mathbf{y}$, i.e., which is like repeating ordinary linear regression m times and averaging across these. In particular, the covariance information is reflected in the posterior weights as it is in the ordinary regression solution.

The model weights are paired with a decision rule to classify a test item such as $x_i = [x_{i1}, x_{i2}, x_{i3}]$, with three cues such as in Fig. 5.4. The vector x_i is multiplied by the mean posterior weight matrix $\mathbf{W}^*$ to generate an output vector $\hat{y}_i = [\hat{y}_{i1}, \hat{y}_{i2}, \hat{y}_{i3}]$:

$$\hat{y}_i = x_i W^*. \tag{5.1}$$

Note that using the posterior mean is equivalent to integrating over the full posterior distribution, due to the linearity of the prediction. The *TTB decision rule* is then applied to the resulting $\hat{y}_i$ as

$$z_i = sgn\left(y_{i,j_i^*}\right) \text{ where } j_i^* = \arg\max_j \left|y_{ij}\right| \tag{5.2}$$

and

$$\text{choice}_i = \begin{cases} +1\ (left), & \text{if } z_i = 1 \\ -1\ (right), & \text{if } z_i = -1 \\ 0\ (guess), & \text{if } z_i = 0. \end{cases} \tag{5.3}$$

The TTB decision rule selects the maximum absolute output (Equation 5.2) and takes the valence of that output as its choice (Equation 5.3). Notice that the decision rule is applied over a weighted combination across all cues (Equation 5.1) that depends on the posterior weight matrix. When $1/\eta^2 = \infty$ (and the cross-weights are thus zero), the decision rule exhibits the exact sequential nature of the TTB heuristic, because then each output $\hat{y}_{ij}$ in Equation 5.1 will equal the value of each cue $\mathbf{x}_j$

times its cue validity. In other words, the cues will be perfectly ordered according to their cue validity, e.g., where the largest output will correspond to the most valid cue that is not indifferent, and so on. Importantly, those cases where the cue is indifferent between alternatives, the weighted combination output is $\hat{y}_{ij} = 0$ resulting in the TTB rule to move onto the next most valid cue. Thus, when the TTB decision rule is adopted, the COR model converges to the TTB heuristic as $1/\eta^2 \to \infty$. This convergence is shown in Computational Study 6 (Fig. 5.5) in an artificial binary prediction task similar to Fig. 5.1, demonstrating that the COR model (with TTB decision rule) and the TTB heuristic reach perfect agreement in their predictions as the prior becomes strong enough.

Notably, the tallying heuristic can also be derived from the COR model, in its undirected version that uses cue validities in the training data to infer cue directionalities. The *tallying decision rule* is defined by

$$z_i = \sum_j sgn\left(y_{ij}\right). \tag{5.4}$$

and

$$\text{choice}_i = \begin{cases} +1\ (left), & \text{if } z_i = 1 \\ -1\ (right), & \text{if } z_i = -1 \\ 0\ (guess), & \text{if } z_i = 0. \end{cases} \tag{5.5}$$

The tallying decision rule chooses the alternative with a majority of outputs in its favor (conveyed by their valences), irrespective of the magnitudes of the outputs. The choice is determined by Equation 5.5), as in the TTB decision rule. When the tallying decision rule is adopted by the COR model, the model converges to the tallying heuristic in the limit as $1/\eta^2 \to \infty$ (Fig. 5.6, computational study 6).

Lastly, in the limit of $1/\eta^2 = 0$, either decision rule will yield decisions equivalent to ordinary linear regression as outlined above. That is because, in the limit of $1/\eta^2 = 0$, the posterior weights are all equal to the ordinary linear regression weights, and the outputs $\hat{y}_i$ produced according to Equation 5.1 are all equal to the ordinary linear regression prediction. Applying either a TTB or tallying decision rule to these equivalent outputs $\hat{y}_i$ (Equation 5.1) will yield a choice equal to the valence of that prediction.

Thus, we have demonstrated how ordinary linear regression and both TTB and the tallying heuristic can be derived as extreme cases of a single Bayesian prior defined by covariance expectation. Importantly, the element varying across the continuum is the priors strength (as reflected in the posterior mean $\mathbf{W}^*$, Equation 5.18), and the prior is responsible for recovering the heuristics in the limit, rather than decision rules or the choice of regularization type. The COR model converges to ordinary regression as the strength of the prior goes to zero regardless of the decision rule, and these model properties hold under other forms of regularization as well we find (e.g., lasso regularization). That is, when the regularization in the posterior for COR is replaced with an L1 regularization, the model properties in the limiting cases still hold, emphasizing the robustness of the COR model.

Next, we demonstrate convergence of the COR model with heuristics and ordinary linear regression in a simulation study (computational study 6). Then, as with the half-ridge model in Chapter 4, we investigate how the COR model performs in real-world environments (computational study 7).

5.4 Computational Study 6: Convergence of Bayesian COR model and heuristics

In this computational study, our goal is to demonstrate that the COR model converges to the heuristics (i.e., tallying and TTB) as the prior grows arbitrarily strong (in the limit of $1/\eta^2 \rightarrow \infty$) and to ordinary linear regression as the prior's strength approximates zero (in the limit of $1/\eta^2 = 0$).

5.4.1 Methods

For the purpose of this simulation, we created artificial binary prediction tasks of the type in Fig. 5.1. Each artificial dataset was created as follows: We generated cue values for $m = 3$ cues on 20 objects by uniformly sampling cue values of 0 or 1. Cue values refer to the smiley and frowning faces in the illustrative example of Fig. 5.1. An object refers to a single football team (e.g., Team England). We then created all possible pairwise comparisons between the 20 objects which results in 190 possible comparisons. For each pair, we computed the cue difference vector by subtracting the cue values of the second object from the first object. For example, in Fig. 5.1, the cue coding column contains the cue differences values for comparing Team England to Team Germany, which can take values of 1, -1 and 0. The object comparisons represent a matrix of cue difference vectors with one row for each object pair. Next, we sampled $m = 3$ weights from an exponential distribution with rate parameter equal to 2 as generating weights. Finally, we calculated a criterion variable by relying on the cue differences matrix, the generating weights, and additional Gaussian noise. The criterion variable contains the outcome for each object comparison, indicating which team won the comparison. In total, we simulated 1000 datasets in this way and results are averaged across them. All models were fit to the artificial datasets, and subsequently made predictions for a novel test set to assess convergence. The test set was constructed according to a complete sampling approach where each possible combination of cue differences occurs once. For three cues with possible cue difference values of $\{-1, +1, 0\}$, there are 27 possible cue combinations. However, as one test item has all zeros as cue values, it was eliminated for not providing any additional information, and hence, the test matrix contained 26 test comparisons. Each test pair corresponds to a novel pairwise comparison, e.g., between two football teams. The predictions on the test set are used to assess agreement among different models.

The COR model, TTB and tallying heuristic, and ordinary linear regression were trained on each of the 1000 datasets to derive the weights for prediction. The COR model weights were derived by fitting the exact posterior mean in Equation 5.18 to the data. As we were interested in the change of the posterior weight matrix as a function of the strength of the prior, we derived a different posterior estimate for each value of the strength of the prior. The range of the prior strengths tested is listed in Table 5.1. Next, the mean posterior weight matrix was used to make predictions with respect to the test comparisons via matrix multiplication. If the cue differences for the test set are represented by a matrix $\mathbf{M}$ containing $m = 3$ columns and 26 rows, and the mean posterior weight matrix $\mathbf{W}^*$ is a 3×3 square matrix, then by matrix multiplication, the output is also a matrix $\mathbf{Y}$ with dimensions 26×3,

$$\mathbf{Y} = \mathbf{M}\mathbf{W}^*. \tag{5.6}$$

Due to the multivariate representation of the problem, the output matrix $\mathbf{Y}$ contains the continuous predictions of the Bayesian model with respect to the three copies of $\mathbf{y}$ (Section 5.9). The output matrix is then paired with either a TTB or a tallying decision rule according to Equation 5.2 and 5.4 to generate outcome predictions for the test comparisons. Convergence results are presented next.

Statistical Parameters in the Simulation

Number of objects	20
Number of pairwise comparisons	$N = 190$
Number of cues	$m = 3$
Class variable	Binary, ± 1
Absolute correlation between cues averaged over cue pairs	0.26
Generating weights	randomly sampled from an exponential distribution with rate parameter equal to 2
Training Sample Size	190
Test Sample Size	26
Number of cross-validation repetitions	1000
Error variance	$\sigma_\varepsilon^2 = 1$
Strength of prior	$1/\eta^2 =$ [1E+06, 8E+05, 5E+05, 1E+05, 10000, 1000, 700, 600, 500, 400, 330.08, 200, 156.81, 74.50, 35.39, 16.81, 7.99, 3.80, 1.80, 0.86, 0.41, 0.19, 0.09, 0]

Table 5.1: Parameters in the simulation on artificial dataset as presented in Fig. 5.5 and Fig. 5.6.

5.4.2 Results

Fig. 5.5 demonstrates that the COR model with the TTB decision rule and the TTB heuristic reach perfect agreement in their predictions as the prior becomes strong enough. At the other extreme, the COR model reaches perfect agreement with ordinary linear regression when the prior becomes weak enough (approaching zero). Parallel results hold for the tallying decision rule (Fig. 5.6). The COR model with the tallying decision rule and the tallying heuristic converge on perfect agreement as the prior becomes arbitrarily strong, and when the prior strength approaches zero, COR converges to ordinary linear regression. These convergence findings were also verified by estimating the posterior mean of the COR model through Markov chain Monte Carlo (MCMC) to sample from the true posterior probability distribution over the weight matrix $\mathbf{W}^*$.

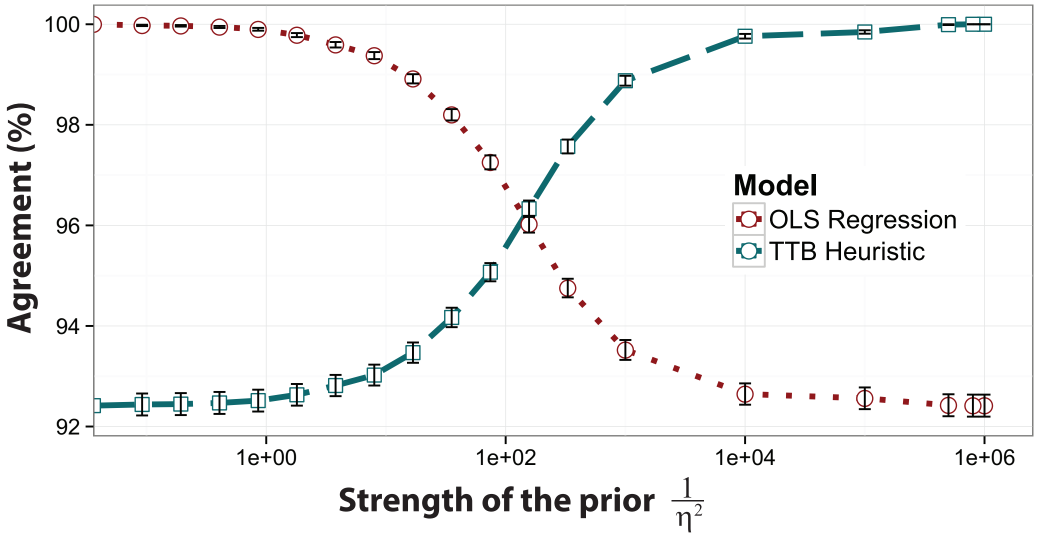


Figure 5.5: Agreement between the COR model (with TTB decision rule) and the TTB heuristic, as well as ordinary linear regression, as a function of the strength of the prior. As expected, agreement (i.e., proportion of equal predictions on test items) between the Bayesian COR model and TTB heuristic increased with a stronger prior, reaching an asymptote of perfect agreement as $1/\eta^2 \rightarrow \infty$. The opposite pattern held for ordinary linear regression, with agreement being perfect at $1/\eta^2 = 0$ and declining as the prior strength increases. The ordinate indicates the percentage agreement on test item choices in a simulated binary decision task (see Methods). Results are averaged across 1000 simulated datasets. Error bars represent ± SEM.

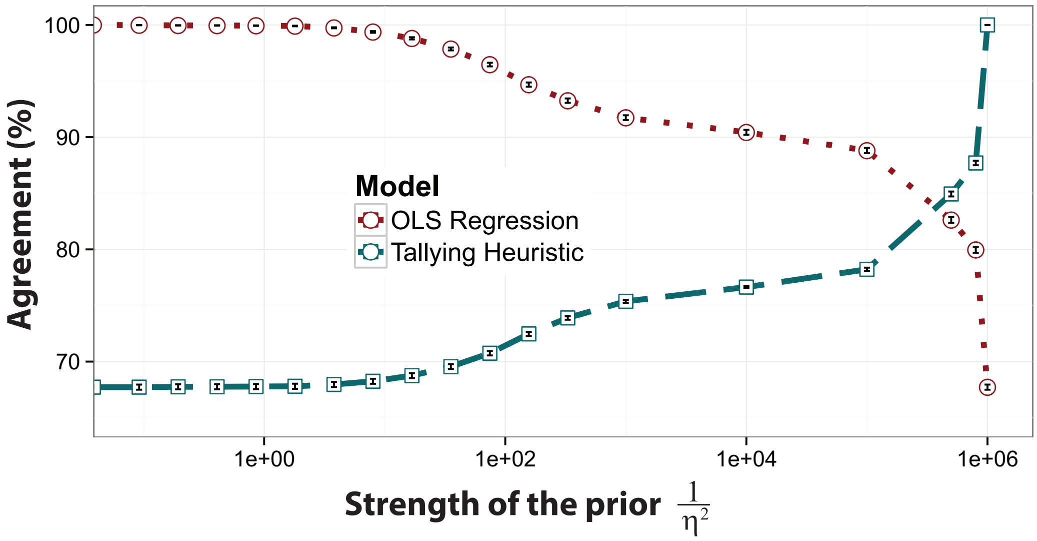


Figure 5.6: Agreement between the COR model (with tallying decision rule) and the tallying heuristic, as well as ordinary linear regression, as a function of the prior strength. As expected, agreement (i.e., proportion of equal predictions on test items) between the Bayesian COR model and the tallying heuristic increased with a stronger prior, reaching an asymptote of perfect agreement as $1/\eta^2 \to \infty$. The opposite pattern held for ordinary linear regression, with agreement being perfect at $1/\eta^2 = 0$ and declining as the prior strength increases. The ordinate reflects the percentage agreement on test item choices in simulated binary decision task (see Methods). The datasets used for this simulation are equivalent to the ones displayed in Fig. 5.5. Error bars represent ± SEM.

5.5 Computational Study 7: Performance in real-world datasets

In parallel to the half-ridge simulations, we investigate how the COR model performs in the 20 original real-world environments that have been frequently used to demonstrate less-is-more effects (Czerlinski et al., 1999). We ask: In data sets where heuristics can outperform ordinary linear regression, can intermediate Bayesian models still outperform both?

One exciting aspect of our Bayesian COR approach is that it specifies a continuum of models between the extremes of linear regression and the heuristics. For many environments, the best- performing model should lie somewhere between these two extremes of covariance expectation. That is because, from a Bayesian perspective, the model that fares best on a given decision task should be the one with a prior most closely matching the data's generative process. Decision environments may vary in the amount of cue covariance, however these covariances between cues are not arbitrarily large (i.e., the covariance levels are not drawn uniformly from all real numbers). Thus, we expected that for many environments, the best-performing model should lie somewhere between these two extremes of covariance estimation.

5.5.1 Methods

The parameters in the 20 datasets (Czerlinski et al., 1999) were equivalent to those in Chapter 4 and are listed below in Table 5.2. We created all possible comparisons between the objects in each dataset, resulting in a binary criterion variable encoding which of two objects is superior on each comparison. The COR model was cross-validated on each dataset by splitting the total number of pairwise comparisons into training and test sample. The size of the training sample was varied between 10, 20 and 115 comparisons, and the test set represented the complementary set of comparisons. We repeated this cross-validation process 1000 times and performances were averaged across all. As in the half-ridge simluations, two of the datasets, Oxygen and Ozone, only have 91 and 55 object pairs in total respectively, so the large training sample size of 115 was excluded for those datasets. COR model weights were derived by fitting the exact Bayesian posterior (Equation 5.18) to the training data. As the posterior weight matrix depends on the strength of the Bayesian prior (i.e., $1/\eta^2$ in the matrix of penalties Λ in Equation 5.18 and 5.19), we derived posterior weight matrices for different values of the Bayesian prior's strength (see Table 5.2). Next, we used the mean posterior weight matrix to make predictions with respect to the test set via matrix multiplication (Equation 5.6). Both decision rules are then applied to the resulting output matrix to generate predictions for the test comparisons according to Equation 5.2 and 5.4. We also validated all model weights and results with Markov chain Monte Carlo (MCMC) sampling directly from the Bayesian posterior over weight matrices.

Statistical Parameters in the Simulation

Number of objects	11 to 395
Number of pairwise comparisons	$N = 55$ to $N = 77815$
Number of cues	$m = 3$ to $m = 18$
Class variable (e.g., which house had the higher actual sales price?)	Binary, ± 1
Absolute correlation between cues (averaged over cue pairs)	range = 0.12 to 0.63, mean = 0.31, median = 0.28, sd = 0.14
Training sample size	10, 20, 115
Test sample size	$N-10, N-20, N-115$
Number of cross-validation repetitions	1000
Error variance	$\sigma_\varepsilon^2 = 1$
Strength of prior	$1/\eta^2$ = [1000000, 100000, 1000, 700, 330.08, 156.81, 74.50, 35.39, 16.81, 7.99, 3.80, 1.80,0.86, 0.41, 0.19, 0.09, 0.03, 0.01, 0.001, 0.0001, 0.00001]

Table 5.2: Parameters in the simulation of the 20 datasets as presented in Fig.5.7 and Fig. 5.8. A full list and descriptions of the 20 datasets are in Appendix B.

5.5.2 Results

Fig. 5.7 shows generalization performance for the COR model with the tallying decision rule, and Fig. 5.8 shows results for the TTB decision rule. Importantly, as with the half-ridge model, we find that CORs performance peaks for intermediate priors in all 20 datasets.

Note that an approximately infinitely strong prior on the far right of each graph (small values of η) in Fig. 5.7 corresponds to the tallying heuristic which learns cue directions from the data, and a prior strength of zero (in the limit of $\eta \to \infty$) corresponds to ordinary linear regression. As expected, we find that the results differ from those in the half-ridge model which assumed that tallying already knows cue directions in advance. Allowing the tallying heuristic to learn from data (Fig. 5.7) results in better performance relative to ordinary linear regression (cf. Fig. 4.3). We find that, in 11 out of 20 datasets, a less-is-more effect could be observed where the heuristic model, e.g., the tallying heuristic ($\eta \to 0$) outperformed ordinary linear regression (prior strength of zero), with 10 and 20 training cases. However, importantly, in all of those less-is-more cases, the models with intermediate prior strength still outperformed both models.

Similarly, Fig. 5.8 displays results of the COR model with a TTB decision rule. An infinitely strong prior (as $\eta \to 0$) on the far right corresponds to the TTB heuristic which learns cue rank orders from the data. The TTB heuristic ($\eta \to 0$) outperformed ordinary linear regression (prior strength of zero), in 18 out of the 20 heuristic datasets with 10 and 20 training cases. However, the performance peak could still be found in the middle, i.e., for medium-strength priors. The performance peak was approximately in the same place across training sample sizes, to the degree that the model is correctly specified, however overall the optimal performance in the COR model was less stable than in the half-ridge simulation. This is due to the COR model being more clearly mis-specified (discussed below in Section 5.9 and the Discussion).

The key finding reported in Fig. 5.8 and Fig. 5.7 is that intermediate COR models outperformed tallying and TTB in all 20 datasets, independent of training sample size. This suggests that that less was not more in these datasets as the heuristics were outperformed by a prior of finite strength that learns covariance from the training data but nonetheless down-weights that information. Interestingly, these findings contrast with the frequentist findings of Fig. 3.4 in Chapter 3: While the frequentist case suggested that less-is-more effects are typically reversed when training sample sizes are increased (i.e., regression model performs best), this relative reversal of less-is-more effects often still happens in the COR model (Fig. 5.7). However, it is no longer true that the most flexible model performs best with large training sample sizes, as the intermediate strategies perform best independent of training sample size.

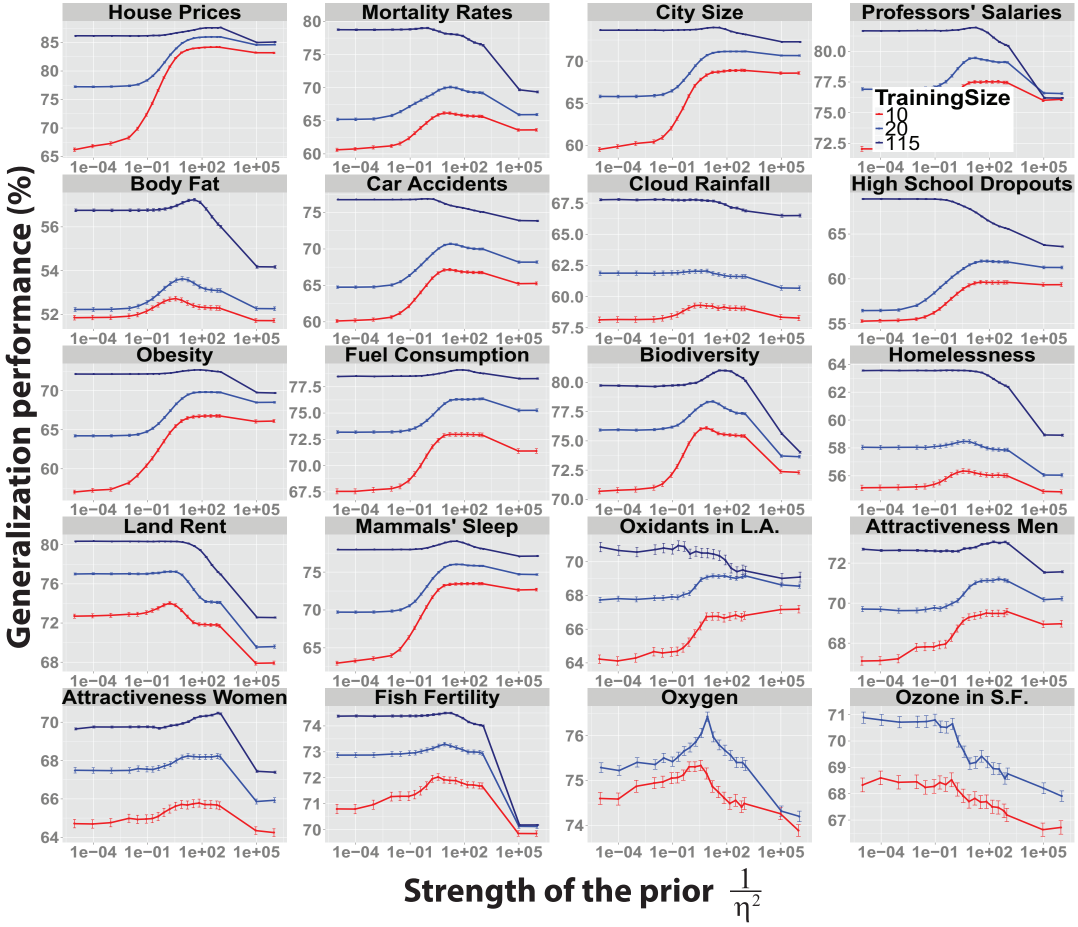


Figure 5.7: Generalization performance of the Bayesian COR model with the tallying decision rule by training sample size in all 20 datasets that heuristics have been extensively tested on Czerlinski et al. (1999). The abscissa represents an increasing prior strength from left to right, and the ordinate represents the predictive accuracy of the model. Note that an approx. infinitely strong prior (e.g., $1/\eta^2$ = 1e+06) corresponds to the tallying heuristic that learns cue directions from the data. A prior strength of zero ($1/\eta^2 = 0$) corresponds to ordinary linear regression. In 11 out of the 20 datasets, a less-is-more effect can be observed, where the tallying heuristic outperformed ordinary linear regression, with 10 and 20 training cases. For example, in the City Size, Car Accidents, and Mammals datasets, the tallying heuristic outperformed ordinary linear regression for training samples sizes of 10 or 20 training cases. However, the optimal performance could be found in the middle, i.e., for medium-strength priors. In other datasets, such as Homelessness, Fish Fertility, and Womens Attractiveness, ordinary linear regression outperformed tallying. However, the optimal performance for all datasets was found for intermediate COR models, i.e., for medium-strength priors. Error bars represent ± SEM.

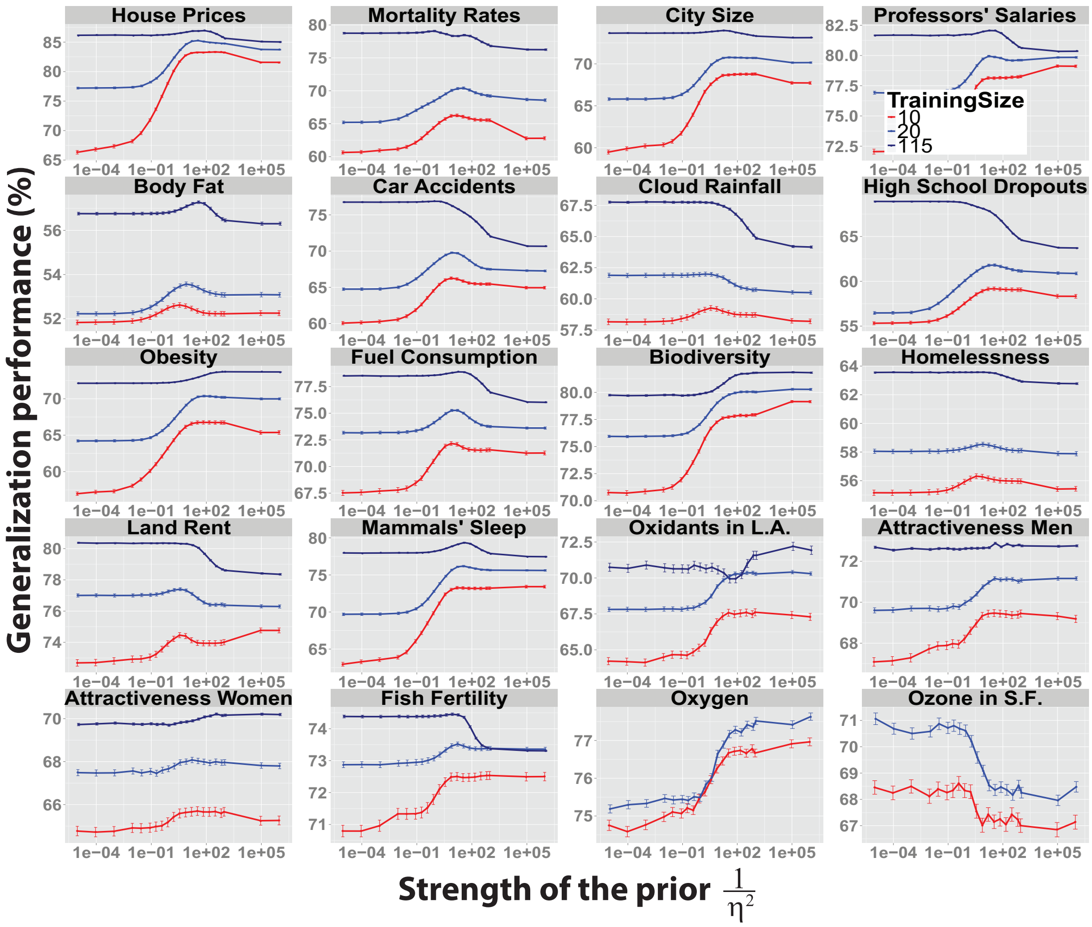


Figure 5.8: Generalization performance of the Bayesian COR model with the TTB decision rule by training sample size in all 20 datasets that heuristics have been extensively tested on Czerlinski et al. (1999). The abscissa represents an increasing prior strength from left to right, and the ordinate represents the predictive accuracy of the model. Note that an approx. infinitely strong prior (e.g., $1/\eta^2$ = 1e+06) corresponds to the TTB heuristic, and a prior strength of zero ($1/\eta^2$ = 0) corresponds to ordinary linear regression. In 18 out of the 20 datasets, a less-is-more effect can be observed, where the TTB heuristic outperformed ordinary linear regression, with 10 and 20 training cases. For example, in the House Prices, Mortality, City Size, and Professor Salaries datasets, the TTB heuristic outperformed ordinary linear regression for training samples sizes of 10 or 20, but the optimal performance could be found in the middle, i.e., for medium-strength priors. In other datasets, such as the Cloud Rainfall or the Ozone levels dataset, ordinary linear regression outperformed the TTB heuristic, but the optimal performance can still be found in the intermediate COR models, i.e., for medium-strength priors. Error bars represent ± SEM.

As with the half-ridge simulations, the COR simulations reported here defined training sets by directly sampling pairs of objects (i.e., comparisons). We compared this approach to one of sampling objects (and training on all pairs in the sampled subset), to determine whether our results would be dependent on this sampling decision. In short, the qualitative pattern of results is not dependent on the sampling method. When sampling objects rather than comparisons, we varied the training sample size between sampling 5, 7 and 16 objects, which correspond to 10, 21 and 120 possible comparisons, respectively. We chose these training sample sizes to approximate the training sample sizes used for the COR simulations when sampling comparisons (i.e., 10, 20 and 115 training cases in Figs. 5.7 and 5.8). For both the tallying and the TTB decision rule, the pattern is almost the same under both sampling methods. Performance of all models is lower overall by a few percent in accuracy when sampling objects, which makes sense as the models do not encounter test objects in the training set first. Additionally, models with weaker priors (i.e., closer to ordinary regression) showed a larger drop in performance under object sampling (especially for smaller training sizes) than did models with stronger priors (i.e., closer to the heuristics). Thus, sampling objects gives the heuristics a small advantage over ordinary regression for the training sample sizes considered here. However, the number of less-is-more effects (i.e., datasets in which heuristics outperform ordinary regression) is the same and they occur in the same environments for both sampling methods. Also, the location of the performance peak is the same (with some small error) under both sampling methods for both the TTB and tallying decision rules.

5.6 Discussion

We find that both tallying and the TTB heuristic can be formally linked to full regression models through a Bayesian prior on covariance among cues. We show that heuristics correspond to an extreme Bayesian prior that deliberately ignores information (i.e., covariance among cues). Interestingly, less-is-more is observed for comparing simple and complex models (e.g., Take-The-Best and regression), but less-is-more is not true in that one can always do better by including all information and down-weighting it rather than throwing it out. That is, we find that intermediate models which are sensitive to the information in the training data, always performed better than heuristics.

The COR model extends the Bayesian half-ridge model with a Bayesian inference model that learns cue directions and cue weights from the training samples, thereby exhibiting less-is-more effects (e.g., in 18 out of 20 datasets in Fig 5.8). However, typically the best performance is achieved by intermediate models of intermediate prior strength along the Bayesian continuum. It can be observed that the location of the best-performing model was more volatile in the COR model demonstrations than in the half-ridge model (Fig. 4.3). This is because the COR model is more clearly misspecified (see discussion below). Similarly to the preceding Chapter, a central message of this Chapter is that ignoring information is never more. Even for those cases where less-is-more effects could be found with either the TTB or the tallying heuristic, intermediate models that down-weight covariance sensitivity rather than entirely ignoring it did best (in contrast to the *absolute* less-is-more claim (Box 2.4.2) (Gigerenzer & Brighton, 2009; Gigerenzer et al., 1999; Tsetsos et al., 2016)). The Chapter contributes to the thesis' Bayesian explanation for why (relative) less can be more: Heuristics may excel not because of their ignorance of information, but because they embody a prior that approximates the optimal prior. Furthermore, the de-confounding in the Bayesian prior's continuum again provides evidence that heuristics perform well due to their large bias and not their simplicity (i.e., dropping cues or dropping weights). This is demonstrated in the performance comparison between the intermediate models (peak performance) and the heuristics which only represent the limiting cases of the Bayesian prior. The COR model also provides a model for unifying TTB (and tallying) with regression models on a single dimension, depending on the decision rule applied.

On top of these formal developments, we have gained new insights into the role that covariance estimation plays in heuristics vs. full-information models. Heuristics may sometimes work well in practice as they correspond to an infinitely strong prior that is completely oblivious to covariance in the training data, but they will usually be outperformed by a prior of finite strength that leaves room for learning covariance from experience. We expect that natural environments rarely correspond to the extreme prior of a heuristic, assuming complete independence among cues, however, potentially more often the environments match the intermediate prior's strength of some covariance among cues. That is, we believe that most real-world environments probably exhibit some level of covariance between predictors in the environment, as indicated by the average covariance levels in the 20 original datasets for example (Table 5.2).

5.6.1 Psychological Implications

As stated in the half-ridge Chapter, the current work is formal and does not attempt to provide a model of human performance. Nevertheless, it has implications for psychological processes. In previous psychological studies, full-information models that are fully sensitive to covariance in the environmental are usually only contrasted with heuristics that ignore this covariance information (e.g., Chatper 3). However, this Chapter provides evidence that the best-performing model uses a prior that down-weights covariance sensitivity instead of entirely ignoring it or being completely sensitive to it (i.e., a prior of zero strength in the COR model).

To the extent that people are tuned to the structure of the environment, intermediate solutions might reflect the functioning of psychological processes. However, tractable algorithms for these intermediate strategies have yet to be formulated. On the other hand, it could be that the intermediate strategies are computationally intractable. In that case, these intractable algorithms might embody the extant heuristics, meaning that heuristics represent the correct psychological models and are successful because they efficiently approximate the optimal solution (e.g., COR with a finite prior strength). Alternatively, people might be doing something much more sophisticated and sensitive to the data, along the lines of faithful implementation of COR. As suggested in Chapter 4, under this theory, the empirical evidence taken to support fast and frugal heuristics is mistaken: heuristics fit data well only because they closely mimic the much more sophisticated strategy the brain is carrying out.

Some empirical evidence that could potentially indicate intermediate strategies at the psychological processing will be discussed, despite being merely speculative. If people were relying on the TTB heuristic, search should stop after identifying the most valid cue that discriminates among options. Hence, the TTB heuristic predicts people rely on few cues and entirely ignore remaining cues (i.e., an infinitely strong prior in the COR model with a TTB decision rule). However, a recent study by van Ravenzwaaij et al. (2014) indicates that people integrate the full information relying on *all* cues. van Ravenzwaaij et al. (2014) used a hierarchical Bayesian model ('stop and search model') to model people's searching and stopping behaviour in looking up cues in the German city size task. In one of the experiments, participants were free to determine the number of cues they wish to examine. The author's "stop and search model" was no better than TTB at predicting the specific cue at which participants would stop their search. However, when it came to how many cues people take into account overall, the stop and search model predicted that participants would include more cues than predicted by TTB. In fact, the authors found that on at least half of the trials, participants relied on all of the available cues as predicted by a the full-information model strategy. There are a multitude of studies showing similar evidence - when participants are presented with multiple cues on the screen, they seem to make use of all the information, going against the predictions of the TTB heuristic (Newell & Shanks, 2003; Newell et al., 2003). These results are relevant for our work, as they suggest that people might be using all of the available information but down-weighting it, rather than entirely dropping cues such as suggested by extreme TTB. Future research could potentially try to apply the current Bayesian COR framework to data from studies such as in van Ravenzwaaij et al. (2014), to uncover whether intermediate weighting strategies better explain people's inference from cues to prediction. However this is speculative at this point and it needs to yet be investigated first whether, when people rely on the full information, they down-weight this information.

5.6.2 Limitations and Extensions of the COR model

Decision Rules

A potential limitation of the COR model are the decision rules which are applied to COR's posterior weight matrix. Yet, as noted above, importantly, the element varying across the continuum is the priors strength (as reflected in the posterior mean $\mathbf{W}^*$, Equation 5.18), and the prior is responsible for recovering the heuristics in the limit, rather than decision rules or the choice of regularization type (i.e., both L2 and L1 regularization result in the heuristics in the limit). Nevertheless, we point out the decision rules as a limitation of the model architecture. In response to this potential criticism, we developed the half-ridge model which supports the same conclusions but does not require downstream decision rules. The half-ridge model's decision rule is the Bayes-optimal one for the task: It integrates predictions over the posterior distribution of weight estimates, which is equivalent to using the mean of the posterior, by linearity of the mapping from weights to predictions.

Crucially, our argument is more general and goes beyond the model-specific details of the COR model. Based on the different Bayesian frameworks developed in this thesis, and the observation that different choices of regularization scheme (corresponding to Gaussian vs. Laplacian priors) lead to the same heuristics in the limit, we conjecture that heuristics can arise as limiting cases of many different Bayesian models that assume different generative processes. We show that, for a variety of formalisms, extreme priors lead to heuristic-like models.

Misspecification

As mentioned in Section 5.9, the COR model is more clearly misspecified than the half-ridge model. We emphasize that, to a certain extent all models in this thesis are misspecified (like all models are when it comes to real-world environments). The fact that the performance peaks in the half-ridge simulations (Fig. 4.3) were not significantly affected by training sample size suggests that the degree of model misspecification was not sufficient to reveal itself in this way. Importantly, in contrast to the half-ridge model, the COR model is misspecified because of the multiplexing of the criterion variable and a model architecture that is artificially multivariate despite the original prediction problem being univariate. The higher degree of misspecification is responsible for the optimal performance being less stable than in the half-ridge model simulations (Fig. 5.7 and 5.8). To avoid confusion on this topic and stick to the central point, we focus the main results of the Bayesian framework on the Bayesian half-ridge model and rely on COR to demonstrate how TTB can be accommodated within this larger theoretical framework. Crucially, whether the models are misspecified or not does not impact the mathematical fact that the Bayesian models converge to the heuristics under strong priors, or the empirical fact that intermediate models outperform the heuristics on real datasets.

Logistic Regression

As in the half-ridge Chapter, we relied on least squares regression to be consistent with past work on the 20 heuristic datasets (e.g., Czerlinski et al. (1999)) and to replicate less-is-more effects with our Bayesian framework. However, as the originally continuous criterion variables were discretized to form binary dependent variables, a logistic regression would be more appropriate for the binary outcomes. To confirm that the choice of link function was not critical, we re-analysed the data using logistic regression as presented in the preceding half-ridge Chapter. We found no significant difference between the two regression approaches on the datasets considered (See Section 4.6.4 for the statistics). A future extension could be to build the COR model continua within a (Bayesian) logistic regression framework rather than the linear one used here.

Why were the particular heuristics chosen?

We chose the two particular fast-and-frugal heuristics, i.e., tallying and TTB, for our Bayesian frameworks as they are among the most well-known fast-and-frugal heuristics, and because they are intuitive and arise in a number of contexts. Both heuristics have been repeatedly contrasted with rational full-information linear regression approaches (Czerlinski et al., 1999; Gigerenzer & Goldstein, 1996; Katsikopoulos et al., 2010), which makes them very suitable for consideration as part of a Bayesian inference model for our purpose. However, including other heuristics will provide a great extension of the current Bayesian program to better understand heuristics and their relationship to full-information models and Bayesian inference models. Fortunately, analyses with the initial two heuristics proved tractable, providing an opportunity to argue that less is not more when less involves ignoring (rather than down-weighting) information. To justify the current focus on the fast-and-frugal heuristics as a sub-problem for the Bayesian inference frameworks, refer to Chapter 2, where I explained that heuristics in the heuristics-and-biases account are not formalized and therefore more difficult to be included in any formal framework, as they do not make falsifiable predictions and can be explained with multiple theories (Gigerenzer & Goldstein, 1996) (Nevertheless, see recent work by Lieder et al. (2017) for an example of a rational process model that tries to capture the anchoring-and-adjustment heuristic).

5.6.3 A Novel Regularization Account

Our regularization approach (COR) is related to ridge regression, and although ridge regression can reduce overfitting, many problems researchers face across fields such as genetics, machine learning, neuroscience, or finance may be better addressed by CORs regularization-by-co-variation approach. This may especially helpful when datasets have high levels of redundancy. Applying COR allows a continuum of covariance sensitivity where the optimal solution can lie anywhere along that continuum. The Bayesian framework also reiterates the importance of applying fundamental machine learning concepts to psychological findings (Gigerenzer & Brighton, 2009). In doing so, we provide a formal understanding of why heuristics can outperform full-information models.

5.7 Summary

This Chapter provided a Bayesian model formally unifying both tallying and TTB with regression based on a prior of covariance sensitivity. First I discussed the existing literature on the role of covariance in heuristics and why it may be an important factors that differentiates heuristics from full-information models. Next, I introduced the Bayesian learning model based on covariance sensitivity. A first computational study demonstrated convergence of the Bayesian model with either heuristic in the limit with a strong prior. A second computational study assessed the generalization performance of the COR model in the classic 20 real-world environments (Czerlinski et al., 1999) and found again that intermediate models performed best independent of dataset. The Chapter concluded on the same note as the Bayesian half-ridge Chapter 4: We find that while relative less-is-more is possible, absolute less-is-more is not. Finally, the limitations of the COR model were discussed such as its misspefication, as well as the psychological implications. Next, I will explore whether less-is-more at the psychological level, i.e., as none of the previous Chapters have investigated what people do. I will lay out how the findings in the psychological Chapter are connected to the current Bayesian framework, and will come to conclude on a consistent note regarding less-is-more.

5.8 Chapter Appendix: Cue validities as transformations of single regression weights

These derivations will show that heuristic cue validities are direct linear transformations of single-predictor regression weights.

Cue validities are defined for binary decision tasks, wherein two objects (e.g., two soccer teams, Fig. 5.1) are compared on several cues and the inference is made about which object has the higher criterion value (i.e., which team will win the match). The criterion variable encodes the actual outcomes (e.g., which teams actually win the soccer matches), and can be coded as -1 and +1 as in Figure 5.1. Cue validities, v, reflect the probability with which single cues can identify the correct alternative, and can be derived as the proportion of correct inferences made by each cue across a set binary comparisons Martignon and Hoffrage (1999):

$$v = \frac{R}{R+W} \tag{5.7}$$

where R = number of correct predictions, W = number of incorrect predictions, and consequently, $0 \leq v \leq 1$.

For example, Table 5.3 portrays a binary decision environment where five object comparisons are made on the basis of three cues. Note that the computation of cue validities ignores those cases where a cue predicts indifference between objects. A fundamental difference between cue validities and the regression weights derived by linear regression is that cue validities completely ignore covariance among cues. In contrast, regression weights as estimated by a multiple linear regression model always consider the covariation among cues, as seen in the expression for the parameter estimate,

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{5.8}$$

where $\mathbf{X}^T\mathbf{X}$ captures the covariances. In an ordinary linear regression analysis with multiple cues, the covariance among cues has a direct influence on the regression weights $\hat{\mathbf{w}}$. If the regression weights were instead derived by regressing the criterion variable on each cue alone, i.e., eliminating all other cues from the model (single-predictor regression analysis), the weight magnitudes, valences as well rank order of weights would change (see for example Appendix B, Fig. **??**). It can be shown that cue validities are a linear transformation of single-predictor regression weights (Martignon & Hoffrage, 1999), according to the following relationship:

$$\hat{w} = 2v - 1. \tag{5.9}$$

This relation holds because, when there is a single predictor ($\mathbf{x}$), the $\mathbf{X}^T\mathbf{X}$ term in Eq. 5.8 is equal to the number of cases where the predictor makes a prediction ($\mathbf{x} = \pm 1$), with cases where the predictor is indifferent ($\mathbf{x} = 0$) excluded. This can be seen from the computation in Table 5.3. That is,

$$\mathbf{X}^T\mathbf{X} = R + W. \tag{5.10}$$

At the same time, $\mathbf{X}^T\mathbf{Y}$ counts up all cases where a cue predicts the criterion (i.e., $x_i = x_i$) and subtracts those cases where the cue makes the opposite prediction (i.e., $x_i = -x_i$), while ignoring indifferent cases of $x_i = 0$ (see Table 5.3). Thus

$$\mathbf{X}^T\mathbf{Y} = R - W. \tag{5.11}$$

Therefore, the single-predictor regression coefficient estimate $\hat{w}$ can be reformulated as

$$\begin{aligned} \hat{w} &= \frac{R-W}{R+W} \\ &= \frac{2R}{R+W} - \frac{R+W}{R+W} \\ &= 2v - 1. \end{aligned} \tag{5.12}$$

Note also that the expression $\frac{R-W}{R+W}$ in the first line of Eq. 5.12 represents the Goodman-Kruskal rank correlation as suggested by Martignon and Hoffrage

Comparison	Cue $\mathbf{x}_1$	Cue $\mathbf{x}_2$	Cue $\mathbf{x}_3$	$\mathbf{y}$	r_1	w_1	$\mathbf{x}_1{}^T\mathbf{y}$	$\mathbf{x}_1{}^T\mathbf{x}_1$
1	-1	-1	0	-1	1	0	1	1
2	1	-1	1	-1	0	1	-1	1
3	0	-1	0	1	0	0	0	0
4	1	1	1	1	1	0	1	1
5	1	1	-1	1	1	0	1	1
	$v_1 = \frac{3}{4}$	$v_2 = \frac{4}{5}$	$v_3 = \frac{1}{3}$		$R_1 = 3$	$W_1 = 1$	$R_1 - W_1 = 2$	$R_1 + W_1 = 4$

Table 5.3: Computation of cue validities: A binary prediction task where five object comparisons are made on the basis of three cues. The cue columns represent cue difference values, $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ respectively, and are coded in the same way as the coding column in Fig. 1 in the main text. The criterion variable $\mathbf{y}$ contains the outcome of each comparison. r_1 and w_1 indicate whether cue $\mathbf{x}_1$ predicted the outcome correctly or incorrectly (r = right, w = wrong) on each comparison, and R_1 and W_1 are the sums across all comparisons, $R_1 = \sum r_1$ and $W_1 = \sum w_1$. Then, the cue validity for cue $\mathbf{x}_1$is computed as $v_1 = \frac{R_1}{W_1 + R_1}$. The validities for $\mathbf{x}_2$ and $\mathbf{x}_3$ are defined similarly.

(1999). The linear relationship in Eq. 5.9 reveals that cue validities are a positive linear rescaling of single-predictor regression weights. Therefore they yield the same predictions in binary comparisons when used in the TTB heuristic for example, due to returning the same cue rank orders. Note that this will no longer be the case when regression weights are computed in the presence of multiple cues.

5.9 Chapter Appendix: Posterior Mean for COR

The COR model approach differs from standard regularized regression in that the prior modulates sensitivity for covariation among cues. This is achieved by expressing the regression problem in multivariate terms, by replicating the criterion variable $\mathbf{y}$ as many times as there are cues (i.e., m times). Due to this multiplexing, the model architecture implements m regression problems at once, meaning the criterion variable $\mathbf{y}$ is regressed onto all cues m times (Fig. 5.3). The weights constitute an $m \times m$ matrix $\mathbf{W}$, with each column, $\mathbf{W}_{\cdot j}$, representing the weights for the

jth copy of the outcome, $\mathbf{y}_j$:

$$\mathbf{W} = \begin{bmatrix} w_{11} & \cdots & w_{1m} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mm} \end{bmatrix}. \tag{5.13}$$

As in standard regression, the likelihood for each $\mathbf{y}_j$ is given by a Gaussian with error variance σ^2:

$$p(\mathbf{y}_j|\mathbf{X},\mathbf{W}) \propto \exp\left(-\frac{(\mathbf{X}\mathbf{W}_{\cdot j}-\mathbf{y}_j)^T(\mathbf{X}\mathbf{W}_{\cdot j}-\mathbf{y}_j)}{2\sigma^2}\right) \tag{5.14}$$

where $\mathbf{X}$ is the matrix that contains the cue data and is indexed by trials and cues (i.e., $n \times m$).

In contrast to ridge regression, where all weights are penalized equally, in the COR model only the off-diagonal elements of the weight matrix $\mathbf{W}$ are penalized, while the diagonal weights are left unpenalized. This is implemented by assuming an improper uniform prior on all $\mathbf{W}_{ii}$ ($1 \leq i \leq m$) and a prior of $\mathcal{N}\left(0,\eta^2\right)$ for all $\mathbf{W}_{ij}$ ($i \neq j$). The joint distribution on $\mathbf{W}$ treats all weights as independent. The model architecture is illustrated in Fig. 5.3, were the solid arrows represent the diagonal weights (direct weights) and the dashed arrows represent the off-diagonal weights (cross-weights). Penalizing only the cross-weights has the effect that the strength of the prior ($1/\eta^2$) modulates the model's sensitivity to covariation among cues. When $1/\eta^2 = 0$ (uniform prior on all weights), the posterior for the weights $\mathbf{W}_{\cdot j}$ is identical for all $\mathbf{y}_j$, with mean (and mode) equal to the ordinary least squares linear regression solution. As $1/\eta^2 \rightarrow \infty$, the estimated cross-weights converge to zero, while the direct weights stay un-penalized. Thus, in the limit, the direct weight w_{jj} is the only non-zero weight in each column $\mathbf{W}_{\cdot j}$. This means that each cue effectively has its own isolated regression (i.e., as if only direct weights were present in Fig. 5.3, with no cross-weights). These single-predictor regression weights are linear transforms of the cue validities as used by the heuristics (see proof in Eq. 5.12). Therefore, in the limit, when the COR model weights are paired with a decision rule

(Eq. 5.2 or 5.4), the model's behaviour converges to that of the respective heuristic. To derive the posterior distribution for COR's weight matrix, we observe first that the weights for the different copies of $\mathbf{y}$ are decoupled. More precisely, the prior, likelihood, and hence posterior all factor into separate functions, one for each set of weights $\mathbf{W}_{\cdot j}$. Therefore we can derive the posterior separately for each set. The prior for each set of weights is given by

$$p(\mathbf{W}_{\cdot j}) \propto \exp\left(-\frac{1}{2}\mathbf{W}_{\cdot j}^T \Sigma \mathbf{W}_{\cdot j}\right) \tag{5.15}$$

where Σ is the precision matrix, defined by $\Sigma_{jj} = 0$, $\Sigma_{ii} = \frac{1}{\eta^2}$ for $i \neq j$, and $\Sigma_{ik} = 0$ for $i \neq k$. Combining this with the likelihood in Eq. 5.14 yields the posterior:

$$p(\mathbf{W}_{\cdot j}|\mathbf{X},\mathbf{y}) \propto \exp\left(-\frac{1}{2}\mathbf{W}_{\cdot j}^T\left(\Sigma + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)\mathbf{W}_{\cdot j} + \frac{1}{\sigma^2}\mathbf{W}_{\cdot j}^T\mathbf{X}^T\mathbf{y}\right) \tag{5.16}$$

$$\propto \exp\Bigg(-\frac{1}{2\sigma^2}\left(\mathbf{W}_{\cdot j} - \left(\Lambda + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right)^T$$
$$\left(\Lambda + \mathbf{X}^T\mathbf{X}\right)\left(\mathbf{W}_{\cdot j} - \left(\Lambda + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right)\Bigg). \tag{5.17}$$

That is, the posterior for $\mathbf{W}_{\cdot j}$ is a multivariate Gaussian with mean at

$$\left(\Lambda + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{5.18}$$

and covariance matrix equal to

$$\sigma^2\left(\Lambda + \mathbf{X}^T\mathbf{X}\right)^{-1}. \tag{5.19}$$

The matrix Λ is interpretable as a matrix of penalties on the components of $\mathbf{W}_{\cdot j}$, with $\Lambda_{jj} = 0$, $\Lambda_{ik} = 0$ for $i \neq k$, and $\Lambda_{ii} = \frac{\sigma^2}{\eta^2}$ for $i \neq j$.

The posterior weight matrix depends on the strength of the prior, as expressed by $1/\eta^2$ in Λ (Equation 5.18), resulting in different posterior weight estimates with different η.

It needs to be noted that, in contrast to the half-ridge model, the COR model is misspecified, because of the multiplexing of the criterion variable. The resulting model architecture is artificially multivariate despite the original prediction problem being univariate. Nevertheless, the COR model opens up new insights into the role of cue covariance in establishing a continuum between heuristics that rely on cue validity and full-information models. Penalizing only the cross-weights in the COR model architecture results in a regularization of covariance sensitivity in the model, with a continuum ranging from ordinary linear regression (fully sensitive to the covariance structure among cues) to heuristics that rely on cue validities (insensitive to any cue covariance).

Chapter 6

Do people learn weights or ranks? Psychological evidence for less-is-more

"Man is not what he thinks he is, he is what he hides."
ANDRÉ MALRAUX

None of the previous Chapters addressed psychological processing yet. We developed a formal computational framework for why less can be more (or appear to be more). In contrast, this Chapter explores to what extent less is more for humans, i.e., to what extent people choose to ignore information via simple heuristics, or include the full information. Referring back to the original definitions of less-is-more in Chapter 2, importantly, we address the *descriptive psychological* definition (Box 2.4.3) which asks: To what extent do people rely on simple heuristics rather than full-information models? We do not address the *capacity psychological* definition (Box 2.4.3) which states that when information gets too much, people may process less information as it results in better performance. Hence, the less-is-more effect at the capacity level may still hold, while it does not invalidate the other descriptive psychological less-is-more effect. We are interested in whether people usually rely on simple heuristics or full-information models. While the work in previous Chapters looked at statistical less-is-more effects (relative and absolute), this Chapter looks at a different kind of psychological less-is-more phenomena (see definitions in Section 2.4).

This Chapter compares people's use of the TTB heuristic and logistic regression by comparing the requisite representations for TTB heuristic and logistic regression against people's active information gathering behaviour using information theory. The goal is to better understand what representations people choose to learn in order to understand what information people have at their disposal for their decision strategies: Do people use the full information and weight each piece of information (regression)? Or do they use only the cue rank orders and ignore cues altogether (TTB)?

We find evidence that people in fact use all available information and learn cue weights rather than cue rank orders. That is, people do not exactly follow the extant fast-and-frugal heuristics in our experiment which throw out information entirely, but are more sensitive to learning the precise weights in the training data. This is consistent with the Bayesian framework's predictions in previous Chapters, proposing that people may be following a more sophisticated strategy which is sensitive to the full information, even while matching a heuristic strategy. We establish this with an active learning paradigm, which focuses on people's active information-gathering behaviour. The reason we use an active learning paradigm to understand people's decision strategies is that it can move beyond traditional passive methods in decision making experiments: While typically decision making experiments place participants into a passive, controlled experiment where stimuli have already been carefully selected by the experimenter and participants make forced choices, an active learning paradigm allows the experimenter to look at what stimuli people choose to learn in the first place, and subsequently what they rely on for decision making. That way, we are able to track both people's active learning process but also their passive decision making process in a test phase. Overall, the Chapter finds support for the same conclusion as previous Chapters, on a psychological level. Less is not more in our experiment (descriptive less-is-more). Furthermore, the method of using information gathering behaviour to separate behaviourally hard-to-distinguish decision models is novel and we argue that it could become a promising new model selection method for psychology.

Next, we will motivate our active learning paradigm as a new method to look at people's decision making by first introducing limitations of existing evidence for people' use of heuristics compared to full-information models. We will conclude that the existing evidence is sparse, inconclusive, or sometimes appears artificially induced. That is because the methods currently in use for discriminating among model classes are too homogeneous. What is missing is a way of moving beyond the typical passive behavioural model fits, by finding a way of looking at people's decision making process earlier on, i.e., when gathering information relevant for the decision. Then we will introduce active learning as a method in psychology. Subsequently, the two active learning algorithms will be introduced: an active learning algorithm for both Take-The-Best and Logistic Regression, which assume that people learn to establish cue rank orders or cue weights, respectively. Next, an active learning experiment designed to distinguish between these models is introduced. By letting both models and humans actively learn, we can compare their active queries in the active learning task. The experiment also looks at people's behaviour in dependence of the compensatoriness (Fig.2.4) in the environment, assessing whether TTB is be a better psychological model under non-compenstory environments, while a full-information model may be more appropriate in compensatory environments (Martignon & Hoffrage, 1999). Finally, we will draw a link between our findings and the Bayesian framework in this thesis.

6.1 Traditional model testing approaches

While the fast-and-frugal heuristics approach gained great popularity based on showing statistical less-is-more effects as presented above in (Brighton, 2006; Chater et al., 2003; Czerlinski et al., 1999), the empirical evidence for any specific use of heuristics remained relatively sparse. One of the core questions in this debate concerns the way in which people look up and integrate information, and whether this behaviour conforms to the search and stopping rules as predicted by the heuristics or full-information models (Box 2.3.2, 2.3.4, 2.3.3). While some studies find that people's search behaviour and response times conform to the TTB heuristic at least sometimes, and in an adaptive manner (e.g., Bergert & Nosofsky, 2007; Bröder & Gaissmaier, 2007; Dieckmann & Rieskamp, 2007; Rieskamp & Dieckmann, 2012), other studies find that people are better fit by TTB only in a third of the cases while adopting other full-information strategies instead (Bröder, 2000; Glöckner & Betsch, 2008; Newell & Shanks, 2003; Newell et al., 2003).

One common method to disentangle people's use of noncompensatory heuristics (TTB) versus compensatory full-information models has been probing whether participants look up additional information (Newell et al., 2003). However, this is only a limited approach to the problem at hand since it is only ever possible to check for $k+1$ look-ups given C_k cues presented so far and it has been argued that people look up additional information but do not use it (Marewski & Mehlhorn, 2011). If people were using TTB, their information search of cues should stop in accordance with its stopping rule. In contrast, if they were using a compensatory strategy, information search should continue beyond that. Interestingly, as outlined above, a large body of studies found the opposite to TTB - people were integrating the full information available (Bröder, 2000; Glöckner & Betsch, 2008; Newell & Shanks, 2003; Newell et al., 2003). Hence, fast-and-frugal heuristics proponents tested people's use of TTB when acquiring additional information was costly (Newell & Shanks, 2003; Newell et al., 2003). For example, Dieckmann and Rieskamp (2007, 2012) first showed in computer simulations that in environments with high redundancy among cues, TTB can be as accurate as naive Bayes, and then experimentally tested whether people's choice of strategies conforms to this pattern, such that in high-redundancy environments, TTB would better predict participants' judgments, whereas in low-redundancy environments, full-information strategies would predict best. In the learning phase, participants rarely stopped their information search after finding a first discriminating cue as predicted by TTB: Stopping consistent with TTB was observed in 26% of decisions in the low-redundancy condition and 23% in the high. This pattern changed dramatically when information search costs were introduced in the test phase. In the low-redundancy environments, people stopped search in accordance with TTB for 44% of decisions, but in the high-redundancy environments it became the predominant strategy, with 77% of all decisions. These findings suggest the prevalence of TTB use can be increased by increasing cue information cost. Nevertheless, how surprising is this finding and how much does it tell us about people's actual use of heuristics? In order to elicit higher prevalence of TTB use, the above studies introduced an external information cost (Dieckmann & Rieskamp, 2007; Newell & Shanks, 2003). This suggests that the noncompensatory search behaviour in line with TTB may have been partly artificially induced, as it is sensible to use less information when it is costly, i.e., people did what any sensible model would do. It would present much stronger evidence if people readily used TTB without the additional search cost in the experiment.

What other model testing approaches have been used? The model dispute over heuristics and full-information models has been about their psychological plausibility. A repeated argument has been that non-compensatory strategies are simpler and require less computational capacity and are therefore more plausible (Todd & Gigerenzer, 2000). Yet, the most common method of pitting full-information models and heuristics against each other have been statistical simulations, showing that one outperforms the other (i.e., *relative* less-is-more, Box 2.4.1). For example, as demonstrated in previous Chapters, sometimes heuristics outperform full-information models (Czerlinski et al., 1999; Gigerenzer & Goldstein, 1996; Katsikopoulos et al., 2010). In contrast, other studies show that there is no strong reason to prefer TTB over other cognitive models as it does not perform noticeably better (Chater et al., 2003; Schulz, Speekenbrink, & Shanks, 2014). While statistical simulations are appropriate for understanding relative and absolute less-is-more effects, it is not sufficient to derive conclusions about people's psychological representation. We argue that the psychological and statistical less-is-more effects need to be distinguished such as by definitions in Chapter 2. Unfortunately, in the literature, evidence for the relative less-is-more effect seems to often get used as evidence for the psychological less-is-more effect, as evident in Gigerenzer and Brighton (2009)'s following statement:

Less-is-more effects: More information or computation can decrease accuracy; therefore, minds rely on simple heuristics in order to be more accurate than strategies that use more information and time. (Gigerenzer & Brighton, 2009, p. 110)

However, we argue, that just because one class of models can beat another with better predictions, it does not follow that this class is necessarily a better psychological representation of what people actually do. We also conjecture that it is not enough to assume people use heuristics because of their simplicity and accuracy. Instead, psychological processing has to be investigated independently with appropriate methods that can elicit people's representations. Although whole research paradigms are dedicated to solving the question about whether people rely on simpler heuristics or full-information mechanisms, different methodologies currently in use to answer this question are scarce and homogeneous. Especially given the current replication crisis discussion (Maxwell, Lau, & Howard, 2015; Stroebe & Strack, 2014), psychology is in dire need of novel model selection methods.

We propose *active learning* as a novel method to solve the dilemma of discriminating among full-information and heuristic strategies as psychologically plausible decision models. What most previous decision making studies have in common is that they study peoples decision making in static, passive and highly controlled experiments. Yet, in order to answer the crucial question about what information people hold in memory and how they look up knowledge when making decisions, we believe one has to look at an earlier stage in the process –at the stage of learning the relevant information in the first place. We argue that stronger evidence for peoples use of either TTB or a full-information strategy comes from the way people actively acquire information, i.e. cue weights or cue orders, in the respective environments. If a cognitive agent has evolutionarily developed to prefer a certain class of models as her means to learn a cognitive representation in a particular environment, then the way she sequentially selects information should (at least partially) reflect this representation. For example, if an agent has come to apply TTB, then -intuitively- she should try to gather information about the cue rank orders that will reduce her uncertainty maximally, and so forth. Using this way of re-creating the structure of a cognitive mechanism, it becomes possible to set up active learning algorithms for many different cognitive models over time. We firstly introduce active learning as a method in psychology in general and then our novel active learning algorithms for decision models of TTB and logistic regression (a full-information model). We suggest that the active learning method can be used as a general new method to perform model selection in psychology. Thereby, we rely on the active learning method to answer the question: Do people use heuristics or full regression models, i.e., is less more at the psychological (descriptive) level?

6.2 Active Learning: Do people learn with respect to cue weights or cue orders?

The main idea behind psychological theories of active learning is to describe a learning agent as optimally designing experiments (Chaloner & Larntz, 1989). That is, given that one wishes to find the true hypothesis out of many potential explanations as fast as possible, an agent assigns prior probabilities to each hypothesis according to some objective criterion such as available frequency data or according to the subjectively judged plausibility of each hypothesis. Each possible outcome of each possible experiment can thus be considered in a ”preposterior analysis” (Raiffa & Schlaifer, 1961) assessing the ways in which each possible experimental outcome could modify beliefs about the hypothesis. Optimal experimental design (OED) relies on increasing *information gain* and maximizing an informational utility, which is typically a measure of how much the beliefs about the hypotheses have changed and the uncertainty has been reduced. A common measure of *uncertainty reduction* is entropy reduction (Shannon, 1948). Shannon entropy expresses the prior uncertainty about a hypothesis, while the reduction in entropy refers to the reduction in uncertainty about the hypothesis after seeing some evidence. There has been a great deal of interest in both normative and descriptive questions surrounding human information acquisition. In a probabilistic framework, many OED models have been used to model human behavior on cognitive tasks such as feature learning (Griffiths

& Austerweil, 2009), reward-specific information search (Meder & Nelson, 2012), and to assess the trade-off between exploration and exploitation (Knox, Otto, Stone, & Love, 2011). Oaksford and Chater (1994) were among the first researchers to define participants' information query behaviour as active information selection. In a series of experiments they showed that the way people select cards in the Wason card selection task is in line with optimal experimental design, thereby re-redefining what was thought of as irrational behaviour into a sensible strategy to test hypotheses. Markant and Gureckis (2013) investigated if it is better to select or receive information for testing hypothesis about categories and showed that participants actively selecting categories tend to search in high information regions more along the category boundaries, resulting in a faster learning curve and lower classification error after fewer learning trials (Markant & Gureckis, 2013).

In this current study, we want to assess to what extent different active learning models based on uncertainty reduction can match participants' behaviour in an active learning experiment, but with the goal of distinguishing among decision models. Thereby we use the notion of efficient information gathering to assess what people are trying to learn about Bramley et al. (2017). As there are no active learning counterparts to these decision models yet, we developed two entropy-minimizing learning algorithms, one for a cue-ranking strategy and one for a cue-weighing strategy. Next, we compare the models' a priori search queries to the queries made by participants in an active learning task with pairwise comparisons. By letting people freely choose among pairwise comparisons during learning, we can investigate whether people pick information such that they learn about cue orders, or instead learn cue weights directly as proposed by the active logistic regression strategy.

6.3 Active learning algorithms

Both active learning algorithms essentially rely on a one-step ahead greedy uncertainty minimization of the uncertainty over possible input queries. Input queries in the active learning context are all those observations that an active agent could make next, e.g., choosing from different pairwise comparisons (such as cities in the city size task). Greedy algorithms always choose as the next observation that query which currently promises to reduce the uncertainty about the learning model maximally. In that way, the active algorithms' one-step ahead queries can be compared with the queries made by participants.

6.3.1 TTB

The TTB heuristic assumes that people look up cues sequentially in the order of their cue importance, and stop search as soon as a cue favours one option over the other (Box 2.3.2). Hence, the active TTB algorithm learns with the goal of establishing the cue rank orders and not the precise weights. We implemented a Bayesian version of TTB that estimates a distribution over cue weights via Metropolis-Hastings sampling and then generates multiple realizations of the heuristic given its current posterior's cue rank orders. These multiple realizations of TTB are called proposal TTBs and differ in their cue rank orders. The proposal TTBs can then be used to create multiple realizations of model predictions over the input queries (i.e., possible next pairwise comparisons), and the predictive variances for input queries can be assessed as the amount of disagreement over all proposal TTBs. Higher disagreement means that the different proposal TTBs generated a higher predictive variance for an input query. Queries with higher uncertainty lead to higher disagreement and are therefore (in the long run) expected to reduce uncertainty most in a one step ahead greedy search. The active TTB algorithm therefore chooses that query as its next observation where the uncertainty is maximal, as this is the query where uncertainty reduction is expected to be largest. In other words, the active TTB algorithm makes choices that are expected to learn most about the underlying cue rank orders.

6.3.2 Logistic Regression

Logistic regression is set up as the competing full-information model to the TTB model. In contrast to the heuristic which relies on the cue rank order and ignores cues, logistic regression optimally weighs each cue and integrates all of them (Box 2.3.4). The active logistic algorithm was set up to learn with the goal of establishing the cue weight magnitudes.

We implemented a Bayesian version of logistic regression based on a random walk Metropolis algorithm. We use Gibbs sampling to draw posterior MCMC-samples of the regression weights. These multiple realizations of posterior weights result in multiple realizations of the model, i.e., it creates proposal logistic regressions estimates. These proposal models can then be used to generate model predictions over all input queries. As in the active TTB model, the predictive variance for each query can finally be summarized by the disagreement among proposal models. Thereby we built a logistic regression analogue to the active TTB that works in the same way. The active logistic model hence chooses that input query next which has the largest predictive variance, i.e., uncertainty, and is expected to reduce uncertainty maximally. Instead of trying to drive down the uncertainty with respect to cue rank orders, the logistic algorithm tries to drive down uncertainty with respect to the regression weights.

6.4 Degrees of Compensatoriness

We are interested in the performance of the two proposed active learning models in environments with different ”compensatoriness” (Martignon & Hoffrage, 1999) (Fig. 2.4). Note that a non-compensatory environment can be defined as a logistic regression environment in which the β weights are exponentially decreasing. In order to create different degrees of ”compensatoriness”, we make use of a mathematical trick that allows us to rely on a single parameter to smoothly vary from compensatory to non-compensatory environments through a ”stick breaking process”. The generation would be of a set of 4 weights $\beta_{k=1}^{4}$ through:

$$\beta_k' \sim \text{Beta}(1, \theta) \tag{6.1}$$

$$\text{Define } \{\beta_k'\}_{k=1}^{4} \text{ as:} \tag{6.2}$$

$$\beta_k = \beta_k' \prod_{i=1}^{k-1} (1 - \beta_i') \tag{6.3}$$

As the expectation of the Beta-distribution is defined as $\frac{\alpha}{\alpha+\theta}$, a perfect TTB environment corresponds to setting θ to 1 or greater as this would lead to a perfectly non-compensatory weight structure. Given the strict boarder of $\theta = 1$ that separates compensatory from non-compensatory strategies, we will use $\theta = [\sim 0, 0.5, 1, 2, \sim \infty]$ for all the upcoming scenarios as this generates degrees of compensatoriness starting from uniform weights ($\theta \sim \infty$) all the way to an environment where only one cue matters ($\theta \sim 0$). Fig. 6.1 shows the weighting structures that result from simulating different levels of compensatoriness with four cues by increasing θ.

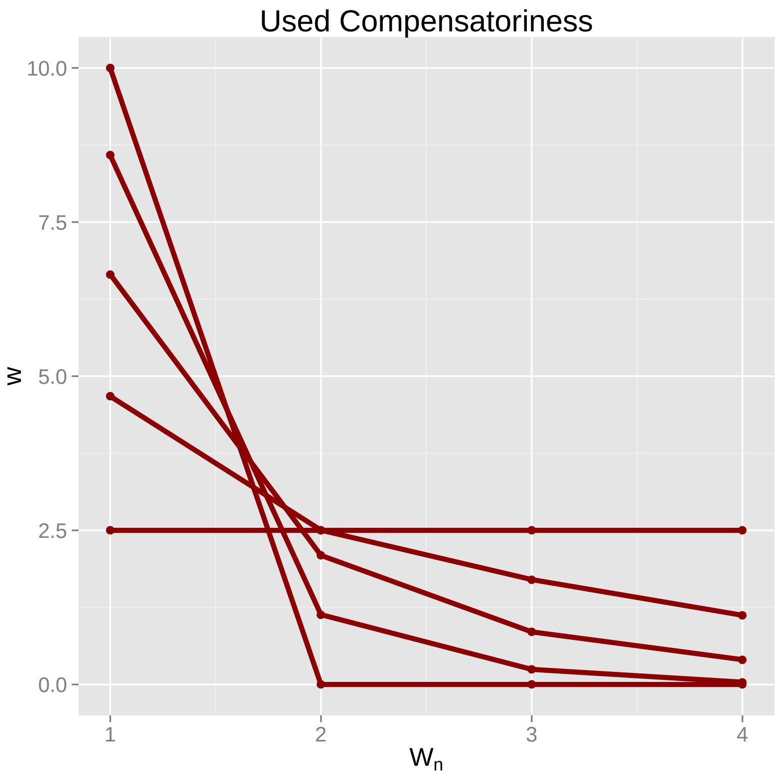


Figure 6.1: Compensatoriness for five different levels of θ. The x-axis represents four different cues and the y-axis displays the cue weight magnitudes. These five levels of compensatoriness were used as five conditions in the Experiment below.

The heuristic literature predicts people's choice of decision model is adaptive to compensatoriness and we sought to see whether this is also the case for active learning models (Martignon & Hoffrage, 1999).

6.5 Experiment

The experiment was designed to find out whether people are more likely to follow a rank-based or a weight-based active learning algorithm. The outcome has implications for either decision mechanism as plausible decision strategies, i.e., either people might learn the exact weights and integrate the full information (logistic regression), or people might only learn the cue rank orders instead (TTB). We hypothesized that the active model that best describes people's information querying behaviour would map onto the most likely cognitive decision model. This constitutes the basic assumption of our approach. We also wanted to investigate whether people are sensitive to the structure of the environment (the degree of compensatoriness) in their active queries, such that in non-compensatory environments participants would be better matched by an active TTB algorithm, while in compensatory environments they would be better matched by an active logistic regression algorithm. We assigned people randomly to one of the five above-mentioned compensatoriness conditions (Fig 6.1).

6.5.1 Participants

Two hundred and sixty-four (N = 264) participants (M = 35.4 years) were recruited via Amazon Mechanical Turk to take part in the ”Alien Olympics” study. Participants were paid $0.50 for participation plus an additional bonus between $0 and $0.5 depending on their performance.

6.5.2 Procedure and Stimuli

The experiment was divided into a learning phase and a test phase. The learning phase consisted of participants actively choosing Alien pairs to fight against each other, with a binary outcome. On each learning trial, participants had to choose a pair of Aliens to compete with the goal of learning about the Alien's strengths and features. Aliens varied on four different cues, which are displayed in Fig. 6.2. All cues were designed to be helpful in fights, e.g., wings enabled an Alien to fly which helps in attacking enemies, while camouflage is useful for hiding from enemies, and antennas give surrounding vision. The cues were explained to participants at the start and they were told that the different characteristics might not all be of equal importance for an Alien's strength in a fight.

Importantly, we emphasized that people should pick their Aliens wisely by selecting informative comparisons out of the presented Aliens, as the goal was to learn how the different cues influenced an Aliens chances to win. Participants were informed that they would need this feature knowledge later in the experiment for an assessment task (test phase). As there are four cues, we generated all possible cue combinations which results in 16 different Alien types. On a given trial, participants were presented with four random Aliens on the screen such as in the screenshot in Fig. 6.3. They had to choose only one out of the six possible resulting Alien comparisons. After selecting a pair, participants received feedback about which Alien had won the competition. They were also told that sometimes a weaker Alien could win against a stronger competitor as in any sport, which reflects the probabilistic generation of the outcomes. The underlying weights of the four cues that people could

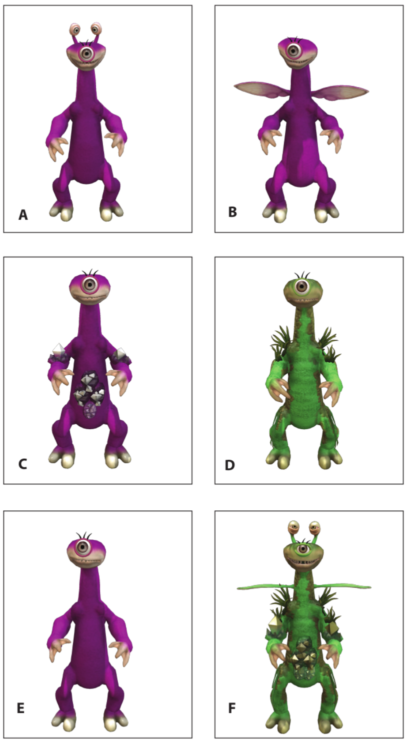
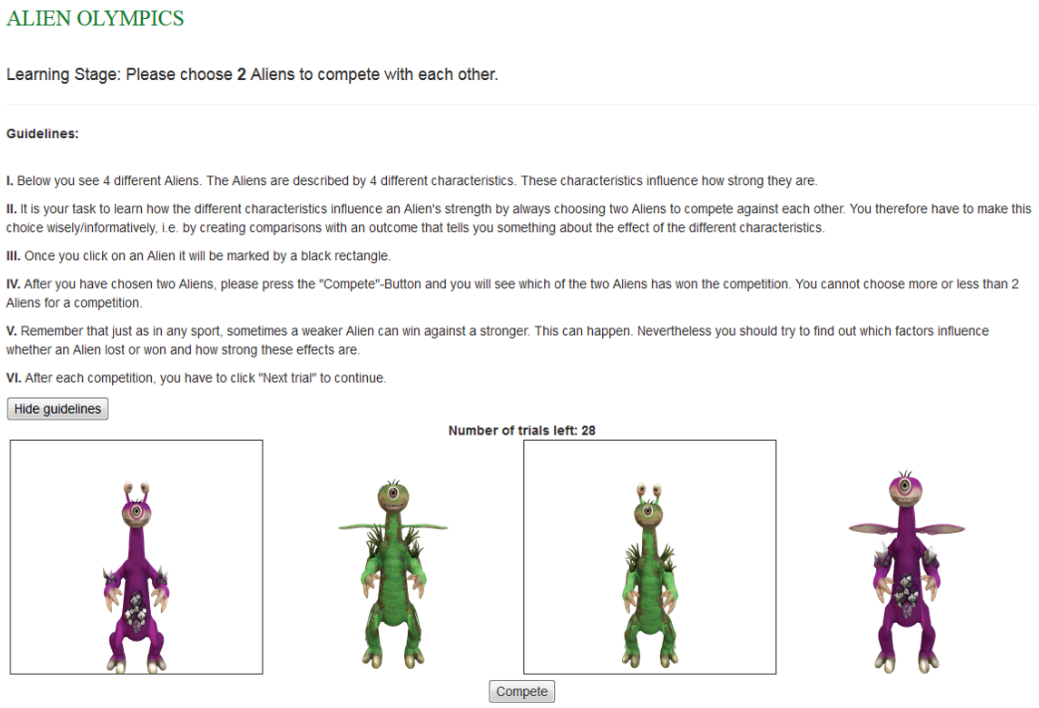


Figure 6.2: Aliens varied on 4 different cues (A-D): Antennae, Wings, Diamonds, and Camouflage. E: Alien without cues, F: Alien with all cues.

learn depended on the compensatoriness condition a participant was in (Fig. 6.1). The actual outcomes observed in feedback were generated by using the weights from the respective compensatoriness conditions (standardized to always add up to 10) and applying logistic regression in order to determine an Alien's strength, i.e., likelihood of winning against another Alien. The learning phase consisted of 30 trials overall.


ALIEN OLYMPICS

Learning Stage: Please choose **2** Aliens to compete with each other.

Guidelines:

I. Below you see 4 different Aliens. The Aliens are described by 4 different characteristics. These characteristics influence how strong they are.

II. It is your task to learn how the different characteristics influence an Alien's strength by always choosing two Aliens to compete against each other. You therefore have to make this choice wisely/informatively, i.e. by creating comparisons with an outcome that tells you something about the effect of the different characteristics.

III. Once you click on an Alien it will be marked by a black rectangle.

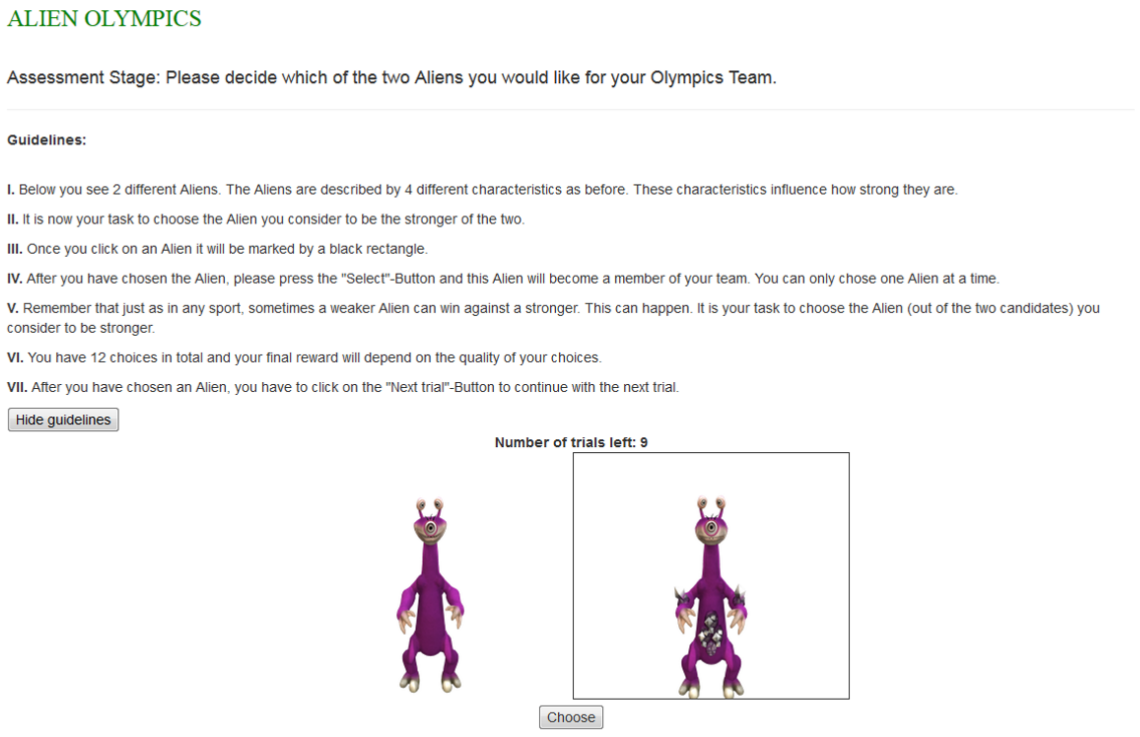
IV. After you have chosen two Aliens, please press the "Compete"-Button and you will see which of the two Aliens has won the competition. You cannot choose more or less than 2 Aliens for a competition.

V. Remember that just as in any sport, sometimes a weaker Alien can win against a stronger. This can happen. Nevertheless you should try to find out which factors influence whether an Alien lost or won and how strong these effects are.

VI. After each competition, you have to click "Next trial" to continue.

Figure 6.3: Training: An example screenshot of a learning trial in the Experiment.

The test phase was designed to assess what people had learned and was structured as follows: On each test trial, participants were presented with only 2 different Aliens that were again randomly drawn from the Alien database. We told participants that these Aliens were the candidates for their Olympic Team, and it was their task to choose the Alien they considered to be stronger based on what they had learned about the characteristics in the learning phase. The test phase consisted of 10 trials forcing participants to make binary choices. Participants were reminded that a bonus payment would depend on their performance in this test phase.



ALIEN OLYMPICS

Assessment Stage: Please decide which of the two Aliens you would like for your Olympics Team.

Guidelines:

I. Below you see 2 different Aliens. The Aliens are described by 4 different characteristics as before. These characteristics influence how strong they are.

II. It is now your task to choose the Alien you consider to be the stronger of the two.

III. Once you click on an Alien it will be marked by a black rectangle.

IV. After you have chosen the Alien, please press the "Select"-Button and this Alien will become a member of your team. You can only chose one Alien at a time.

V. Remember that just as in any sport, sometimes a weaker Alien can win against a stronger. This can happen. It is your task to choose the Alien (out of the two candidates) you consider to be stronger.

VI. You have 12 choices in total and your final reward will depend on the quality of your choices.

VII. After you have chosen an Alien, you have to click on the "Next trial"-Button to continue with the next trial.

Figure 6.4: Test: An example screenshot of a test trial in the Experiment.

6.5.3 Results

First, we present test phase results, and then the active learning results. For purpose of clarity and differentiation, model results from the test phase will sometimes be referred to as "passive" while model results from the learning phase will be referred to as "active".

6.5.4 Passive Model Fits at Test

Participants' performance at identifying the stronger Aliens during the test stage was highly above chance, $t(263) = 27.44,\ p < 0.001$. The average percentage of correct choices made was 74% with a range of [30%, 97%]. Performance varied as a function of the compensatoriness condition that participants were in. Fig. 6.5 represents the average performance score at test as a function of compensatoriness: As the environmental structure got more non-compensatory (i.e., more weight on just a few cues), the average performance dropped. This intuitively makes sense as there is less information to be learned when one cue dominates all others, which

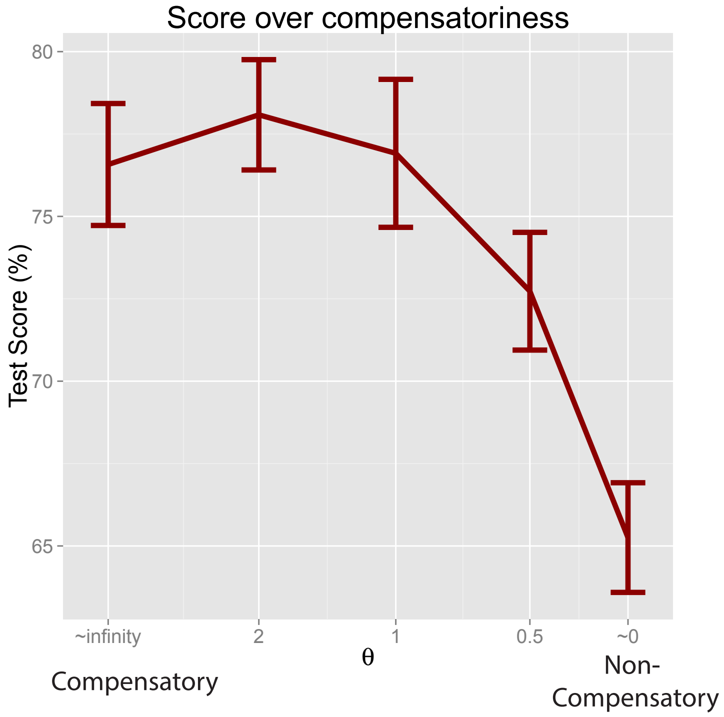


Figure 6.5: Average test performance by compensatoriness conditions. Y-axis represents the percent of correct choices that participants made across the 10 test trials. Error bars represent ± SEM.

makes draws among Aliens more likely and informative comparisons less likely. However, peak performance was observed for an environment not entirely compensatory ($\theta \sim \infty$), but slightly compensatory ($\theta = 2$).

Next, we demonstrate the correspondence of the TTB heuristic and Logistic Regression at and participants' choices at test. That is, we are not looking at any active learning models yet (below), but are assessing the passive behavioural model fits at test first. That is, we assessed the how well the two decision making models, regular TTB (Box 2.3.2) and logistic regression (Box 2.3.4) would perform at predicting people's choices at test correctly. This is essentially assessing the models' predictive accuracy in cross-validation, where the training sets are represented by a different training set for each participant - that is because each participant had a different training experience depending on what queries they selected on learning trials (and the attached feedback). Hence, we let both the TTB and logistic regression learn in the same training environments as the participants, by creating as many simulated participant profiles as there were participants. During learning, both models extract cue weights, e.g., regression weights or cue validities and cue rank orders, and the established weights are then used to make predictions for the novel alien pairs encountered at test. That is, again both models made predictions for each unique participant profile (test sets differ between participants, as Aliens are randomly selected). We were interested in the overlap of the models' test predictions and participants actual choices at test. This model fit assesses how well the TTB heuristic and a full-information model, i.e., logistic regression, capture what people's decision making strategy is at test, based on what they have learned in training.

Fig. 6.6 presents the performances of the TTB model and logistic model in comparison with a model that predicts at random. It can be seen that both logistic regression and TTB were better than the random model at predicting people's behaviour at test. However, logistic regression was a lot better than the TTB heuristic at capturing people's choices. These initial results suggest that people's decision strategy was more in line with a logistic regression strategy. Hence, this speaks for the psychological plausability of the full-information model rather than the heuristic as a psychological model for decision making. Importantly, this is where psychology experiments usually stop. That is, most psychological studies do not go beyond comparing models in a passive decision making experiments, such as the test phase in this experiment. That is, the model selection is often completed with the passive model fit - based on how well the models capture people's behaviour as measured by a criterion (e.g., predictive accuracy (%), R-squared, AIC or BIC). However, these results, although insightful, still represent passive model fits as they model participants' choices made in environments where the stimuli were already chosen by the experimenter in a highly controlled manner, but were not actively chosen. Instead, we focus our attention on the active learning results.

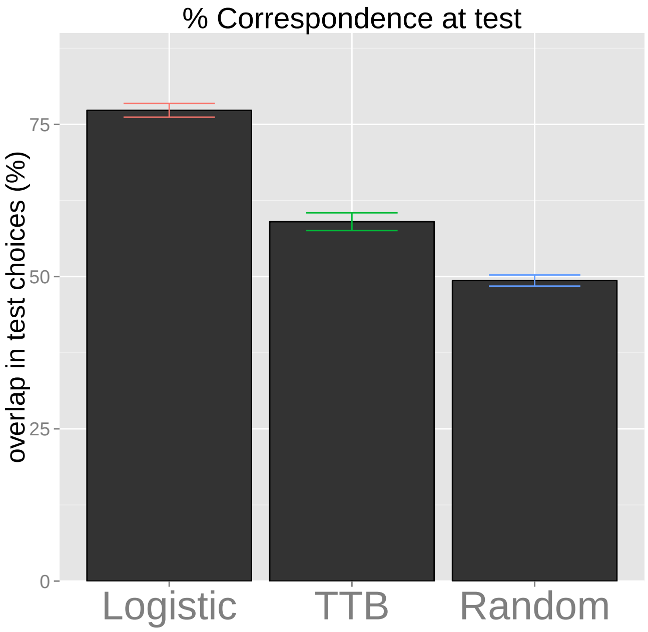


Figure 6.6: Correspondence between participants and (passive) decision models at test. Regular TTB, logistic regression, and a random model are trained on people's observed alien queries and then make predictions for what binary choices people will make at test (across 10 test trials). The y-axis refers to the accuracy at capturing people's choices at test. Results are averaged across participants and compensatoriness conditions. Error bars represent ± SEM.

6.5.5 Active Learning Patterns

We categorized all possible pairwise comparisons people could create into the 8 subtypes that can be seen on the x-axis of Fig. 6.7. For example, a query of +000 signifies a comparison of two Aliens with 3 equal cues (0 for draw), but where one Alien had one more cue than the other Alien (+ for advantage). A +-00 query compares two Aliens that are matched on two cues but differ on two other cues (+ for advantage, - for disadvantage), e.g., one Alien may have the additional cue 'Wings' and the other may have the cue 'Camouflage', while both have Diamonds and Antennaes. This query would hence test whether the cue 'Wings' is more important than 'Camouflage' for the outcome.

An important test is to see how people's frequencies of choosing queries differ from what is expected under the base rates of the study. That is, each query has a different base rate probability of occuring on any given learning trial, i.e., some Alien comparisons are more likely than others due to the nature of how the 16 different Alien types were generated (see above). We compared the absolute frequencies with which participants *chose* each type of query across the whole experiment (i.e., 8 subtypes in Fig. 6.7) with the probabilities of each query *occuring* at any trial for any participant (as measured by the relative frequency of query occurence across the full experiment). A Chi-square test of goodness of fit reveals that the observed distribution of frequencies was significantly different from the expected frequency distribution under the base rate probabilities, $\chi^2 = 460.23$, $df = 7$, $p < 2.2e - 16$. Hence, it can be concluded that people's behaviour in the learning task was non-random and followed a deliberate pattern. Interestingly, the Chi-square test also

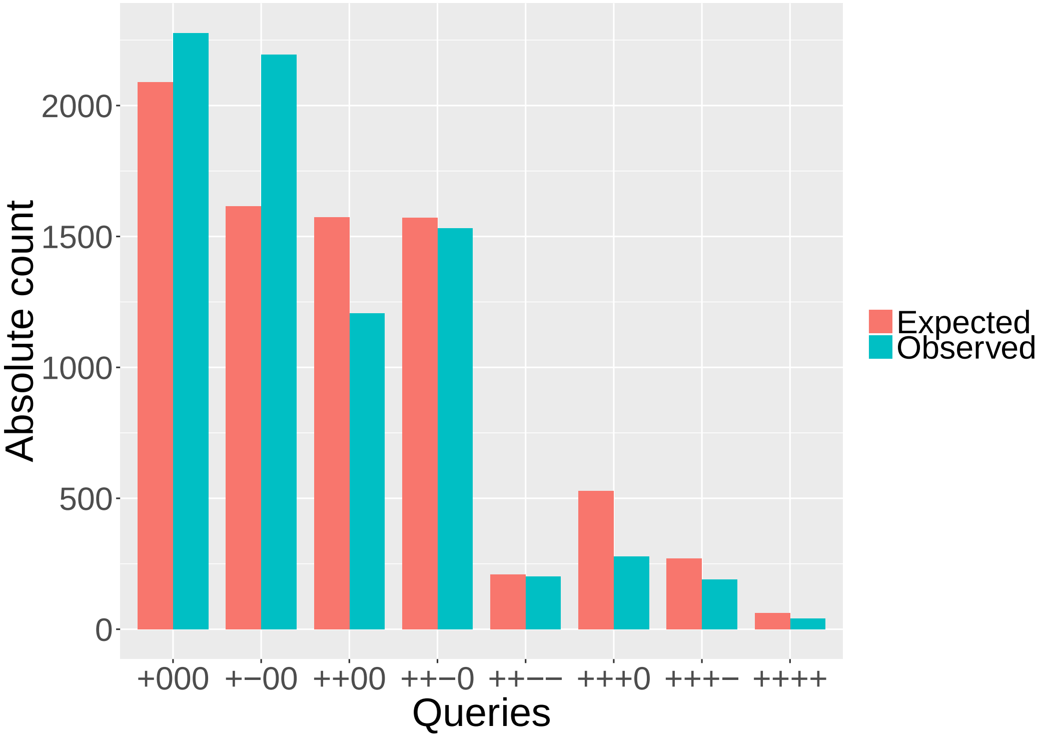


Figure 6.7: The 8 subtypes of active learning queries that participants could make. Y-axis represents the frequency of choosing the query across the 30 learning trials across 264 participants. Coding is as follows: '+' = Alien has a cue that the other alien does not have (advantage); '-' = Alien lacks a cue that the other alien has (disadvantage); '0' = Both aliens have the cue, or both aliens do not have the cue (draw). The figure plots the observed frequency of choosing a particular query against the expected frequency of choosing a particular query under the base rate (probability of occuring naturally) in the experiment.

provides a vector of the estimated frequency counts under the base rate probabilities. Henceforth, it is possible to compare the observed frequency counts of participants with the counts expected under the null hypothesis of the base rate, which is plotted in Fig. 6.7. From these results we can see that the biggest difference between expected and observed count was that participants were querying the +000 and the +-00 comparisons much more. People rarely chose comparisons where it is unclear what feature was responsible for an outcome, e.g., a comparison of an Alien with 3 or 4 more cues than its competitor (+++0, ++++). Instead, the most common comparisons were simpler and controlled, such as the +000 query, assessing whether a cue improves the outcome. This query suggests participants were testing whether a cue has a positive or negative effect on an Alien's strength which is a very sensible query given that this was not clear. For example, the instructions to participants explained the properties of each cue (e.g., ”Camouflage means an Alien can hide from enemies”), however it was never explicitly stated that all cue valences were positive.

Participants queried the +-00 a lot more than expected under the base rate. Specifically the +-00 is a controlled test which tests for the relative effect of one cue in comparison to another. For example, such as comparing two Aliens differing on 2 cues, e.g., ’Wings’ and ’Camouflage’, and assessing the relative importance of Wings and Camouflage for the outcome. This behaviour may suggest people were potentially trying to learn the weights of cues (when performing this query multiple times over time), aligned with the active logistic model, or it may suggest people were learning a cue rank order (i.e., Which cue is more important for the outcome?), aligned with the TTB active model. In general, it can be said that all queries, i.e., even queries such as +++0, exhibit some information about cue weights, while for TTB the best queries to learn about cue order seem to be queries such as +-00, ++-0, and +++-.

Finally, we assessed strategy change across trials. That is, as the learning trials progress from the 1st to 30th learning trial, people may prefer different kind of queries. Fig. 6.8 plots query frequencies as a function of learning trial. It stands out that the top two lines (query +000 and +-00) cross over, where the +-00 query takes over the +000 query during the first half of trials and specifically in the middle trials. This is interesting as the base rate frequencies would predict that the +-00 query is always below the +000 query (Fig. 6.7). However, it appears that people are systematically choosing +-00 more often up to a certain point, while later in the experiment, e.g., from around trial 21, the simpler +000 query starts to dominate again. In fact, the slope for the +000 query was significant, $\beta = 0.53 \pm 0.16,\ p < 0.01$. It can be assumed that since the +000 query is a relatively uninformative query, once the cue valence has been established by the learner, the +000 query may reflect a preference to perform confirmatory, i.e., positive tests (Markant & Gureckis, 2012). Recent research indicates that people have a tendency to fall back onto confirmatory testing towards later trials in learning studies (Bramley et al., 2017). Other research also suggests people may use a combination of discriminatory and confirmatory testing, e.g., in causal intervention studies (Coenen, Rehder, & Gureckis, 2015). It is also noteworthy that participants overall preferred the more complex query ++-0 to the ++00 query (middle green lines) despite the expected base rates being equal (Fig. 6.7). The ++-0 query is a more non-compensatory query comparing an Alien with two more cues to an Alien with one other cue, e.g., assessing how much better the cue 'Antennae' is compared to 'Wings' and 'Diamonds' combined. So far we have looked at people's active query patterns, however we have not yet looked at what kind of active queries the active learning algorithms would predict (assuming people minimize uncertainty) and how these compare to people's active queries. Next we look at how well the active learning models could capture people's sequential information search behaviour.

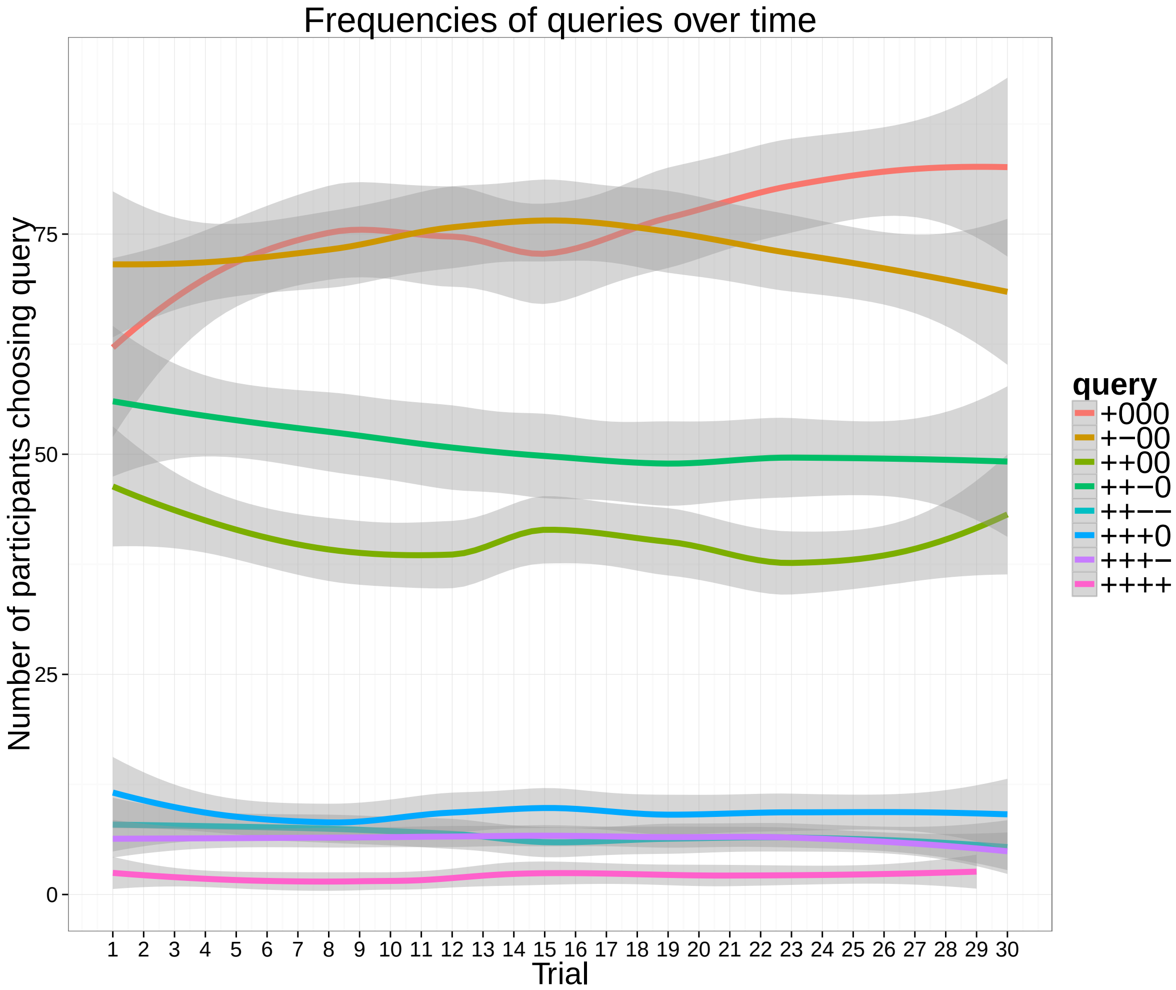


Figure 6.8: The 8 subtypes of active learning queries that participants made as a function of time, i.e., learning trials progressing from 1 to 30 (x-axis). The y-axis represents the number of participants out of 264 participants in total that chose the query at each trial. The data points for each graph were smoothed to a line graph with a smoothing function (geomsmooth() in R), where the boundaries represent the 95% confidence interval.

6.5.6 Active Model Fits at Training

Finally, we compared queries selected by the two learning algorithms against queries chosen by participants. We let both the TTB and logistic regression algorithms learn in the same compensatory and non-compensatory environments as the participants, by creating as many simulated participant profiles as there were participants in each compensatoriness condition. Then, we let the models learn over time. That is, the algorithms made one-step ahead predictions for each learning trail from

1 to 30. Both active algorithms relied on uncertainty sampling, always choosing that query next which maximally reduces uncertainty about the underlying model (see above, 6.3.2 and 6.3.1). This method allows us to compare the participants' queries to those made by the active algorithms and assessing correspondence with a correlation or regression.

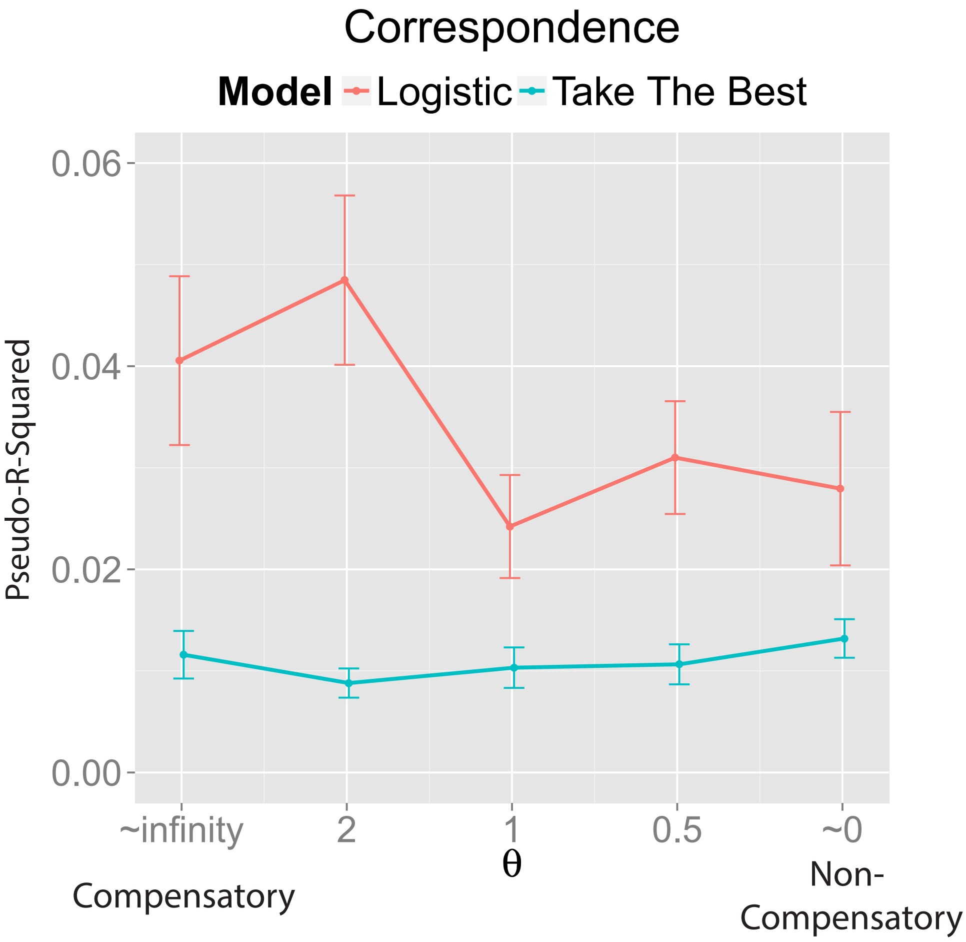


Figure 6.9: Correspondence between the active laerning algorithms and participants' active queries, as a function of the compensatoriness condition. Results are established from simulating active participants with each active model in a step-by-step fashion predicting participants' next query. The active logistic algorithm was better at capturing people's active queries regardless of compensatoriness condition. Error bars represent ± SEM.

Fig. 6.9 shows the match between the active algorithms and participants' queries, as measured by Pseudo-R-squared, as a function of compensatoriness conditions. Firstly, it needs to be noted that results in the active learning part were noisy. Results demonstrate that the active logistic algorithm captured people's queries better in all environments. Interestingly, no clear relationship between the compensatoriness among cues and the best fit active models could be seen, e.g., it was not the case that the active TTB model better fit people's queries in non-compensatory conditions and the active logistic model better fit in compensatory conditions. Instead, people seem to be learning the weights regardless of compensatoriness.

These results are very interesting as they go against much of the heuristic literature claiming people are not able to learn weights due to capacity limitations and instead rely on heuristics which are less computational demanding (Gigerenzer et al., 1999). Particularly in highly non-compensatory environments which present the perfect environments for the TTB heuristic and where it is assumed to be relied on more extensively (Martignon & Hoffrage, 1999, 2002; Rieskamp & Dieckmann, 2012), it is surprising that people were still behaving as if they are learning cue weights rather than cue rank orders.

These results are confirmed with the AIC measures below. Fig. 6.10*b* compares the mean AIC of the active logistic and the active TTB model to that of a random model across all conditions. The mean AIC of the logistic regression model was lower than the mean AIC of the active TTB model, suggesting participants were actively querying in a manner more consistent with the cue-weight based model. It is also interesting to compare the AIC results to those from the passive model fits at test from above (Fig. 6.10*a*). Fig. 6.10*a* and Fig. 6.10*b* demonstrate that across the experiment, in both learning and test phase, people were better fit by an active logistic regression model. Taken together the evidence from both passive and active part of the experiment is stronger and more insightful than would have been by looking at passive model fits. However, we go a step further and look at the AIC model fits of the active learning models on an individual level rather than group level, in order to see how many people were best fit by an active logistic and an active TTB model.

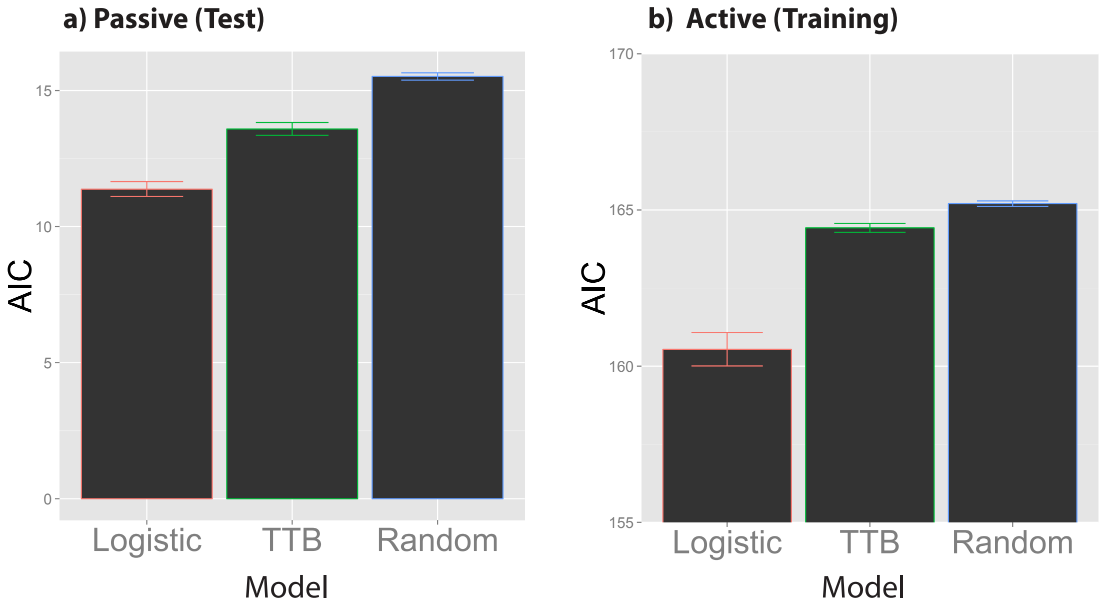


Figure 6.10: AIC model fits for the logistic, TTB and random model in predicting peoples' active queries (Training) **(b)** and passive choices at test (Test) **(a)**. Results are averaged across all 264 participants. For both the passive and active part, the logistic regression model had the lowest AIC and performs best at describing what people do. The AIC magnitudes between **a)** and **b)** cannot be compared, as they correspond to capturing entirely different data (e.g., the active data corresponds to capturing behaviour during training, and the passive data corresponds to choices at test.)

6.5.7 Participants best fit by Active Logistic Model vs. Active TTB Model

We look at the number of participants best classified by an active TTB model versus active logistic model (as indicated by a lower AIC) in relation to those best fit by either model in the passive test phase. Table 6.1 shows a 2-by-2 contingency table displaying frequencies for both learning and test phase. The "active" columns refer to the learning phase and the "passive" rows refer to the test phase. We would expect that the diagonal entries of this 2-by-2 matrix should be higher than the non-diagonal entries for indicating internal consistency among strategies between learning and test. Results show that most people, that were best fit by an active logistic model (LOG-active) compared to an active TTB model (TTB-active), were also best fit by a logistic decision model (LOG-passive) at test (i.e., 107 participants). This is good evidence for the internal consistency among an individual's use

Passive: No. best fit by AIC	**Active:** No. best fit by AIC TTB-active	 LOG-active	 Total by Passive
TTB-passive	47	67	114
LOG-passive	43	107	150
Total by Active	90	174	264

Table 6.1: Number of participants best classified by an active TTB model versus active logistic model (as indicated by a lower AIC) at training (columns), in relation to those best fit by a passive TTB model and passive logistic model at test (rows).

of strategies in our experiment. While there was a clear majority of people better fit by an active logistic model compared to the active TTB model in the active phase (174 vs. 90), the total number best fit by TTB-passive and LOG-passive at test is much more balanced (114 vs. 150).

Most striking is the top right cell size (67), which are people that appear to be best fit by TTB-passive at test, but are in fact better fit by the LOG-active at learning. Hence, this indicates these were people that were in fact actively learning the cue weight magnitudes, however looked like they were applying a TTB rule at test in making binary choices. This finding is particularly interesting as it may indicate people often know more, and are more sensitive to the information than can be found from a passive model fit of a heuristic decision model to people's empirical data in forced choice tasks. This has interesting implications for the less-is-more phenomena, and the Bayesian heuristic frameworks, which will be discussed below. Overall, it can be concluded that the active logistic learning algorithm which minimizes uncertainty with respect to cue weights was a better description of how people learn in our experiment.

6.6 Discussion

Results demonstrated that people were best fit by a logistic regression model in both the learning and test phase in our experiment, instead of a TTB heuristic. The active logistic algorithm was better at capturing people's active queries regardless of compensatoriness condition. This suggests people learned to establish cue weights precisely, rather than cue rank orders. These findings lends support to the full-information model (logistic regression) as a more psychologically plausible mechanism in these environments. Referring back to the original questions of less-is-more at the beginning, importantly, we show that the *descriptive psychological* definition (Box 2.4.3) does not hold in our experiment: People relied on the full information rather than the heuristic, and less was not more.

Even in highly non-compensatory environments, people seemed to learn cue weights rather than cue rank orders, lending even more support to the cue-weight strategy as the underlying psychological model. This goes against what the fast-and-frugal heuristics program propose (Martignon & Hoffrage, 1999; Rieskamp & Otto, 2006), and our results may be different, because we rely on active learning thereby getting a very different window on the decision making process which is usually invisible with passive model fitting to empirical heuristic data.

Overall, people seemed to have a good idea of the actual weights. Specifically, what stands out is that some people even appeared to be using TTB from the empirical data in the forced-choice test trials (of the type in Fig. 6.4), however had actually learned to cue weights at learning (Table 6.1). This suggests people just appeared to be using the TTB heuristic as the TTB rule fits the empirical data well, but this does not reflect what information people hold in their memory and use, and it may be that they were actually relying on the full weighting structure and much more sensitive to the full information. This maps onto the theoretical predictions made by the Bayesian half-ridge and COR model. Interestingly, there are two possibilities to explain these findings: Either people know the weights but decide not to use them, and rely on a TTB rank order instead. Evidence for this theory comes in fact from Marewski and Mehlhorn (2011) who argued that people sometimes know information but decide to ignore it on purpose. The other alternative is that people know the weights and actually use the weights, but the weights map onto the cue order of the non-compensatory TTB, such that both the logistic regression model and the TTB heuristic make the same prediction and are indistinguishable. This is particularly likely when environments have non-compensatory cue structure since a non-compensatory environment can be defined as a logistic regression environment in which the β weights are exponentially decreasing. Hence, in these environments the logistic regression strategy that learns its weights from the environment would map onto the TTB heuristic which relies only on the highest ranked cue. A post-doc analysis of the 2-by-2 contingency table (Table 6.1) in different compensatoriness conditions seems to indicate this is roughly the case: In more non-compensatory conditions, the top right cell sizes are larger than in more compensatory conditions. Although speculative, a possibility is that even in some past heuristic studies, the evidence for the heuristics as psychological models may have been mistaken as people might have known the weights but applied them with a non-compensatory decision rule. Especially in tests with more non-compensatory environments, people might have appeared to be using TTB because it is the optimal weighting strategy (and matches logistic regression), however in reality had a much more nuanced idea of the differential weights. An even stronger argument to make would be that potentially people can only use a heuristic if they know the cue weights , i.e., to extract the cue rank order in TTB, people may first learn the actual weight magnitudes. Or, in order to rely on tallying, one needs to extract the signs of these weights. However note that it does not work the other way around, one cannot use a regression model based on the cue rank order. Future research needs to approach these previous experiments with an active learning approach to compare passive model fits with active model fits similar to our study.

Furthermore, the type of active queries that people chose give a more complete picture. The queries that participants chose significantly more often than predicted by the base rate in the experiment (Fig. 6.7) were controlled comparison queries (+-00), that assess the relative importance of one cue over another. With regard of the above findings, the queries can be re-interpreted in a new light given that most people were reducing uncertainty with respect to cue weights rather than cue ranks: The +-00 query may be a sensible test assessing not only the relative order among cues (i.e., Is Camouflage better than Wings?) but also how strong the outcome relationship is with every cue (i.e., how likely is it that a Camouflage Alien outperforms an alien with Wings?). As the experiment was designed stochastically, this is a good query for learning about the differential weights.

Our findings demonstrate that purely passive model fits are limited in their ability to distinguish between common decision models, even in a prediction-based test (Fig. 6.6). While the passive model fits at test came to the same conclusion that people are best fit by a logistic regression model compared to a TTB heuristic, the full picture only becomes clear when looking at the active learning results. The active learning results represent more processing-based evidence than the usual predictive accuracy findings or descriptive model fits.

6.6.1 What do these results suggest for Bayesian heuristic models?

People relied on the full-information, confirming the Chapter 4 and Chapter 5's hypotheses that people take into account the full information (but potentially down-weighting). To answer the question *Why is less more?* this Chapter's answer might be that it sometimes appears that way, when empirical data fits a heuristic well, but people may actually rely on the full weighting structure such as strategies along the continuum in COR and half-ridge would suggest.

Hence, the findings of this Chapter are much aligned with the propositions of the COR and half-ridge models. Both the half-ridge and the COR model speculated that potentially the optimal intermediate strategies (located between traditional regression and heuristics) may represent people's psychological processing, however more research needs to be done. Crucially, the intermediate models that performed best in real-world environments did not throw away any cues or cue weights, and instead relied on the full information but down-weighted these weights (regularized weights). The COR and half-ridge frameworks also suggested that, if indeed people's psychological processing was more aligned with the intermediate strategies rather than the extreme end cases of the existing heuristics (e.g., which deliberately ignore information), this would suggest the empirical evidence taken to support heuristics is typically misinterpreted: people just appear to be using the heuristics as they fit the data well, however their minds are actually carrying out a much more sophisticated strategy. It cannot be concluded from the current experiment whether people were indeed down-weighting weights or covariance and relied on a strategy that is located between logistic regression and the TTB heuristic, as we did not test this. However, an interesting extension of the current thesis would be to incorporate a half-ridge active learning model and a COR active learning model to predict people's active queries, moving beyond the standard logistic regression. The reason that this was not done yet is merely the fact that the active learning Chapter was done prior to developing the COR and half-ridge models. We believe this will present an exciting new avenue for future research.

A potential limitation of the experiment might have been that the base rate probabilities of the different type of queries (e.g., Fig. 6.7) were not equal. A future experiment should try to hold them constant. However, doing that would also mean a lower external validity of the experiment, as having more of the ++++, +++-, +++0 or ++– queries in the learning phase would mean it becomes much more unlikely that people are able to learn as much, because these queries are less controlled and the source of an effect is less clear. It is unlikely that people rely on these more uninformative queries to reduce uncertainty about a model as evidence suggests people are quite sensible in their choice of active queries (Bramley et al., 2017).

6.6.2 Model-based active learning

We give our novel active learning approach to discriminate psychological model classes the name of *model-based active learning* as its goal is to learn about possible underlying psychological models. The current experiment demonstrates that model-based active learning experiments can be used to distinguish among prominent decision strategies. Our results revealed a more informative picture than the traditional passive model fitting procedures. Model-based active learning is based on the assumption that a cognitive agent actively queries information in the environment, in order to minimize uncertainty about the cognitive model they utilize in that particular environment.

We see this experiment only as a first step towards a general model-based active learning methodology, since the possible applications of this approach are vast.

From moral decision making to function learning, from evidential reasoning to categorization learning and decision making tasks, for many researched domains it is possible to formulate active versions of existing algorithms by simply taking a given model, re-define it to generate uncertainties over its predictions, and then to test it in a model-based active learning task. For example, one can imagine doing the same as above for discriminating among exemplar and prototype models to try to find new insights on a long-standing debate about how people learn about categories. We therefore believe that the methods introduced here can be beneficial to many psychological domains in which researchers have argued about the validity of different models. However, there are also still many short-comings such as testing the underlying assumption that people do in fact learn about the models defined as well as possible by minimizing uncertainty, and are not doing something else instead which we have not uncovered yet, i.e., this is an issue in any active learning research (Coenen, Nelson, & Gureckis, 2017). Furthermore, the approach currently does not test the possibility that models might be developed on the go while active learning is happening. Or else, a participant might consider multiple models to represent a given environment and only later makes a decision on which one represents the data best. Despite these shortcomings we believe that they do not make this approach less exciting. Given that many psychological findings currently seem to be hard to replicate (Maxwell et al., 2015), we have introduced a novel and hopefully exciting way to compare different psychological models, which can be added to psychology's methodological tool kit.

Chapter 7

General Discussion

7.1 Less-Is-Not-More

The initial introduction in Chapter 2 cast heuristics as simple algorithms that can often perform surprisingly well, however often with no clear explanation of why. I gave examples of where heuristics excel in the real world. For example, in financial forecasting a simple 1/N rule was able to outperform a much more complex model in predicting stock performance over time from limited data. Or else, an AI poker agent relied on a set of simple opening heuristics similar to humans, but could not find a better strategy to improve upon the heuristics. The question that was asked at the end was why was less more?

Throughout the thesis a new perspective on heuristics and less-is-more effects emerged, which is that *less is not more*. We learned from the Bayesian formalization of heuristics that heuristics may often appear to perform best, but when the full range of decision strategies is taken into account with a Bayesian prior, one can always find a model that performs better by including all information rather than throwing it out as the heuristics do. Chapter 4 and Chapter 5 found that the best strategy relies on all the information but down-weights it instead via the influence of an appropriate prior. That is, while less-is-more is observed for comparing simple and complex models (e.g., *relative* less-is-more, such as comparing a heuristic with regression), the thesis suggests that the *absolute* less-is-more, wherein heuristics could be optimal, is never true. That is because, consistent with our Bayesian principles, our empirical evidence suggests the intermediate models, that combine the input data with the prior, outperform heuristics. In this view, ignoring data does not improve performance, but, instead, heuristics may often appear successful relative to full regression models, because heuristics approximate the intermediate models that are actually optimal. Thus, the Bayesian explanation for less-is-more is that heuristics work because they embody a prior that approximates the optimal prior.

While the Bayesian framework did not explicitly investigate its implications at the psychological level, the final psychological Chapter explored to what extent people represent and use information for decision making in an active learning experiment. In that way, the psychological Chapter looked at a different kind of less-is-more effect, i.e., whether people fully and systematically ignore information in the input data as proposed by the fast-and-frugal heuristics (i.e., a descriptive *psychological* less-is-more effect). Interestingly, the psychological Chapter 6 came to the conclusion that people do not fully ignore presented information but are much more adaptive to the full information presented, along the lines of full-information models. In that way, the conclusions reached in the psychological Chapter were very similar to the Bayesian Chapters regarding less-is-more despite being at a different level of analysis: less was not more in that people integrated information rather than ignoring it. Furthermore, the empirical evidence suggested that while people sometimes appear to use a simple heuristic during decision making, they had often behaved as if learning the exact cue weights in a learning phase. This suggests that sometimes the empirical evidence taken to support heuristics may be mislead in psychology, which echoes the conclusions of the Bayesian frameworks in Chapters 4 and 5 implying that potentially heuristics may fit empirical data well because heuristics are close to the more sophisticated full-information strategy that people may actually be using. Despite these interesting parallels between the formal approach in this thesis and the psychological study, it would be pre-emptive to draw conclusions with respect to people's use of intermediate strategies, as this needs to be validated in a separate set of studies first.

7.1.1 Implications for work in psychology

As outline at the end of Chapter 4, one possible interpretation of the intermediate models performing best in the Bayesian frameworks is that the brain actually has tractable means for implementing these intermediate models (i.e., for using all available information but down-weighting it appropriately). The other alternative is that the intermediate models are intractable and the brain uses heuristics. In the former case, the hypothesis that people implement intermediate strategies would have to be validated. Hence, future work in psychology should attempt to see whether an intermediate strategy with a weighting structure deviating from that of the classical heuristics and regression models better captures people's behaviour in behavioural decision making tasks.

One idea would be to re-analyse empirical data from existing multiple-cue integration tasks that assessed people's use of TTB as opposed to a weighted additive strategy. For example, the studies by Newell et al. (2003) and Newell and Shanks (2003) asked people to make forced choices between two shares based on a set of cues indicative of the shares' performance. Each cue had a cost attached to it for ”buying” the additional information. Newell et al. (2003) argued that the best measure distinguishing between fast-and-frugal behaviour and weighted-additive compensatory behaviour would be the acquisition of unnecessary information (after discovering a discriminating cue). Hence, the authors sought to see whether people's cue look-up behaviour would stop in line with TTB predictions (i.e., where no more cues are looked up once the highest weight cue discriminates). Interestingly, while people's search behaviour at the group level was consistent with the cue rank orders of TTB, few people behaved completely consistent with TTB in terms of the search, stopping and decision rule. Two thirds of participants violated at least one of TTB's rules and a considerable amount used a weighted-additive strategy. The majority of people did not exactly use a perfect stopping rule according to TTB, but a different frugal stopping rule that took into account more cues than predicted by TTB, even when it was cognitively demanding and financially disadvantageous. Many of these people also violated TTB's decision rule (deciding based on the highest weighted cue only) and indicated they weighted further cues to make a decision. Yet, these participants' strategy also did not appear to be perfectly in line with weighted-additive strategies according to the authors, as participants did not buy all cues available on each trial (Newell et al., 2003). These findings represent an interesting opportunity for research with our Bayesian frameworks and assessing the validity of intermediate models, as the evidence indicates people were behaving close to the heuristic (i.e., cue ranking behaviour) however weighted all pieces of information to different degrees. This leaves scope for assessing whether people's weighting structure might have been not perfectly aligned with TTB nor linear regression, but more in line with an intermediate weighting structure that regularizes weights. The first step in such a study would be establishing the prior for the intermediate model in the specific task environment, which could be done by modelling the task with the Bayesian models. Once the prior strength is established, the intermediate model's suggested posterior cue weights could formulate a new intermediate strategy located between TTB and weighted-additive, and its cue weighting structure could be compared to people's cue buying behaviour, and the intermediate model's predictions with respect to forced choices compared to people's choices. This represents just one potential avenue for researching the psychologically plausability of intermediate models.

Another test going beyond standard forced-choice tasks could be to analyse people's information-gathering behaviour with an active learning paradigm similar to Chapter 6, however adding a third active learning model that learns with respect ot intermediate model weights. By defining an active learning algorithm not only for the extreme heuristics and regression model but the intermediate models, it could be tested whether people's active learning and passive decision making behaviour is more in line with an intermediate model strategy. If the research into intermediate models suggested that people's behaviour is more in line with these strategies than the extant decision models, these should be added to psychology's repertoire of decision making models. Hence, the current framework can be used as a tool to identify new psychological decision strategies.

Lastly, another potential way in which our work could be extended in psychology is by applying the general framework to discover novel heuristics, as briefly outlined in Chapter 4' Discussion. The logic behind our framework, i.e, identifying the correct prior for a particular heuristic, and formalizing the Bayesian prior as drawing a continuum between a prominent strategy such as linear regression at one end, and the heuristic at the other extreme end, may be in itself a good "discovery heuristic" for identifying strategies for different environmental structures and priors.

7.2 A Revised Understanding of Heuristics?

This work potentially attaches a new meaning to heuristics. What we call simple heuristics in decision sciences may in fact correspond to strategies that hide a lot complexity and are more sensitive to learning the full information than we assumed. This idea echoes previous ideas by Dougherty, Franco-Watkins, and Thomas (2008) and Juslin and Persson (2002) arguing that heuristics hide a lot of the complexity in the computations for ordering the cues. If the interpretation of heuristics as intermediate strategies was correct, a new terminology for these intermediate strategies would be required to distinguish them from the extant heuristics in the literature. Note that the ideas in this section are based on the assumption that heuristics only match empirical evidence because they closely mimic a more sophisticated strategy, and hence this section needs to be judged with caution, as it is speculative.

Heuristics are thought to work as they rely on simple decision rules (such as counting the positive and negative evidence in tallying), which are believed to be robust and dont overfit as much. However, if the assumptions in this section are correct, what appeared to be the advantage of a simple decision rule that ignores information may in fact have been the advantage of a large inductive bias in combination with using the full information. This would also suggest what we previously called a heuristic may in fact be the more complex strategy compared to the standard full-information models such as linear regression. In that case, the strategy underlying the heuristics would be a smart strategy which is well adapted to the statistical structure of the environment by using the right amount of inductive bias. Although speculative, a question arising would be: How does the brain know what the correct prior is for a given environment? One idea is that it might potentially rely on something similar to an ”overfitting regularizer” (e.g., by putting a hyperprior on the prior as in a hierarchical Bayesian model), that learns what inductive biases are appropriate for different environments, and learns to adjust the information integration rules to be biased as much as needed for the noisy, sparse environments that we navigate in. If this was correct, the brain might have a mechanism for optimizing biases for different environments, and is well adapted to the statistical structure of the environment.

Another possibility that goes against the assumptions in this section may be that instead of using intermediate strategies, people may be sensitive to all the information in the input data, but deliberately decide not to use it all (Marewski & Mehlhorn, 2011), by applying a simple heuristic decision rule. However, this would also suggest that the decision rules as used by TTB for example (deciding based on highest ranked cue) work due to the more complex computations used to establish the cue rank orders (Dougherty et al., 2008). Hence, in this interpretation, despite the mind being able to learn more nuanced information (e.g., such as optimal weights), it may still just use a simple cut-off rule (along the lines of fast-and-frugal heuristics) for decision-making. However, some data (Newell & Shanks, 2003; Newell et al., 2003; van Ravenzwaaij et al., 2014) goes against this idea as people did not follow the frugal stopping and decision rules. In conclusion, this section outlined some interesting implications following from the current research’s interpretations, but importantly it needs to be emphasized that these were speculative ideas based on assumptions that need to yet be validated.

7.3 Bayesian Models and Heuristics Revisited

Linking back to the original less-is-more findings in the classic 20 datasets (Czerlinski et al., 1999) (Chapter 2), and in the city size task (Chapter 3), many of these less-is-more effects could not be explained with a rational probabilistic model since the nineties.

The fast-and-frugal heuristic program interpreted the success of heuristics in the city size task (Gigerenzer & Goldstein, 1996) as evidence that Bayesian rational norms could be entirely replaced with ecological rationality, because a much simpler model was able to trump what they called a "rational" model (multiple regression). Instead, the fast-and-frugal account (and the heuristics-and-biases account) perceived heuristics and Bayesian models as opponents (Table 2.2) and as a result never assessed to what extent heuristic fit into the broader framework of rational decision making. In contrast, I argued that the ecological rationality approach and the Bayesian rationality approach are likely to be compatible and not mutually exclusive. Hence, in reaction to the original less-is-more findings in the city size task, I believe that a rational explanation was missing for *why* the heuristic succeeded. Hence, my interpretation is very much aligned with Nick Chater's interpretation of the city size task from 2003: "*I suggest that, analogously, Gigerenzer and Goldstein's impressive demonstration of the success of Take-the-Best should lead to a search for a rational analysis of why it succeeds, rather than the conclusion that rational explanation is dispensable.*" (Chater et al., 2003, p. 72). Chater et al. (2003) convincingly argued that, while TTB provides "*an outstanding example of how a fast and frugal algorithm can succeed in the real world, and exemplifies that environmental success does not require that the cognitive system engages in rational calculation using probability or statistical theory.*" (p.72), this still does not suggest that descriptive rational theories are not useful, and more importantly, incompatible with the success of TTB. The thesis demonstrated that heuristics are in fact compatible with Bayesian inference. In contrast to the widespread incompatibility assumptions, heuristics are part of Bayesian inference for extreme Bayesian priors. Thereby, we provide a rational explanation for heuristics that was missing since the nineties (Chater et al., 2003) and resolve the tension between the ecological rationality approach and the Bayesian rationality approach.

What does the Bayesian explanation add beyond the frequentist explanation of less-is-more? Our Bayesian characterization of heuristics does not offer an alternative explanation of (absolute) less-is-more effects. Instead, it refutes the absolute less-is-more claim. The Bayesian characterization opened up ways of defining intermediate models, by taking the infinitely strong priors that yield heuristics and weakening them to priors of finite strength. The result (for both half-ridge and COR) was a continuum of models that use all of the available information, except for the special case of heuristics (i.e., infinitely strong priors). The evidence across Bayesian models, priors and regularization techniques suggested that the intermediate models that combine the input data with the prior outperform those extreme models. What we learn from taking a Bayesian perspective are all of the aspects mentioned in this paragraph and the three sections above (e.g., Revised Understanding of Heuristics, Implications for Psychology and Less-is-Not-More), which follow naturally from the Bayesian framing, in a way that would not be possible under the frequentist bias-variance approach. The Bayesian framework furthermore establishes a formal characterization of the link between traditional statistical models (OLS) and heuristics.

7.4 Implications for work in neuroscience, behavioural economics, computer science, machine learning and other fields

The Bayesian framework developed here can potentially be applied to less-is-more effects in other fields in order to assess whether a solution that entirely throws out information could not be improved upon with an intermediate model. I will give one example of recent work on a neural encoding model which identified a less-is-more effect, that was brought to our attention by a reviewer. Tsetsos et al. (2016) found that when choosing among multiple alternatives (e.g., among 3 holiday destinations based on cues) people make intransitive choices (irrational according to traditional logical norms, (Von Neumann & Morgenstern, 1944,1947,1953,2007)), however these intransitive choices paradoxically improved accuracy when decision formation was corrupted by internal neural noise. The authors provide evidence that people accumulate evidence over time (i.e., cues) using a ”selective integration” policy which deliberately discards information about alternatives (e.g., holiday destination) with lower momentary value. Tsetsos et al. (2016) claim that the selective integration model (called SINT) displays a less-is-more effect on the level of neural encoding. Tsetsos et al. make the normative claim that the SINT strategy is optimal in the context of an accumulator model that has late noise, i.e., random diffusion in the value of the accumulator that is separate from the values of the inputs themselves. The SINT model is shown to improve model performance over a model that uses untransformed inputs (analogous to a relative statistical less-is-more effects with heuristics). However, as in the current research, there are an array of intermediate models that use all the information in the inputs and that would outperform the SINT model, which were not considered. Hence, although the problem that Tsetsos et al. investigate is substantially different from ours (i.e., theirs is a question of encoding, not of decision making), we believe there are some interesting parallels. Looking to the future, there may be interesting new avenues for integrating this kind of research with our formal frameworks by investigating neural encoding strategies for models that down-weight information sources to different degrees without entirely discarding them.

Another area where our work may be useful to other fields is in those contexts where regularization methods are required. Our discoveries are made possible by a novel extension and application of core ideas in machine learning (i.e., ridge regression) to psychological theory. The new half-ridge and COR regularization methods may be useful to researchers in genetics, neuroscience, machine learning, finance and anywhere where noisy datasets and overfitting models are an issue. Furthermore, the COR model may be especially useful (in comparison to ridge regression) in environments where covariance is high, as it allows for a continuum of covariance sensitivity. The Bayesian framework may also potentially be interesting to those who work in data-intensive environments where time can be at a premium (e.g., traders, financial forecasters, doctors, Big Data architects), because the Bayesian frameworks allows for a performance comparison of models of different complexity and simplicity in advance (i.e., how well does a heuristic perform compared to the optimal model in this environment?).

7.5 Relationship to other rational approaches to heuristics

While there are various approaches looking at the compatibility between psychologically plausible processes and probabilistic models of cognition (Daw & Courville, 2008; Griffiths et al., 2015; Jones & Love, 2011; Lee & Cummins, 2004; Marr, 1982a; Sanborn et al., 2010; Scheibehenne et al., 2013; van Ravenzwaaij et al., 2014), which are in line with our approach trying to further bridge the gap between algorithmic-level approaches and computational-level approaches, none of these approaches tried to directly place heuristics within a Bayesian inference framework.

Yet recently, researchers at the University of Berkeley developed a rational process model for a different kind of heuristics, i.e., from the heuristics-and-biases program (Lieder et al., 2017). Lieder et al. (2017) propose a sampling approach to derive the anchoring-and-adjustment heuristic (Tversky & Kahneman, 1974). The authors explain the anchoring bias with a sampling approach where the assumption is that the brain solves numerical estimation tasks by engaging in a sampling process similar to Markov Chain Monte Carlo sampling (MCMC) (Gilks et al., 1996). According to their model, the anchoring bias is the result of suboptimal sampling due to limited cognitive resources, an approach they call ”resource-rational” which is supposedly compatible with bounded rationality. What their work and ours have in common is that both try to explain a heuristic (e.g., the anchoring-and-adjustment heuristic or fast-and-frugal heuristics) with rational inference, however the model by Lieder et al. (2017) relies on the assumption of limited cognitive resources to derive the heuristic rather than putting the heuristic on equal footing with other rational inferences strategies and formalizing the mathematical relationship. This presents a fundamentally different approach. Furthermore, while their work tries to find a resource-rational algorithm that approximates optimal inference, we take the opposite approach and define optimal Bayesian inference models that approximate the heuristics. In addition, a crucial difference is the work by Lieder et al. (2017) do not explain less-is-more effects and focus on explaining cognitive biases and suboptimal behaviour, while our goal is to explain why sometimes less appears to be more. Despite the different approaches and goals, the work is related and we welcome other work looking to further integrate research at the computational level of analysis with the process level (Marr, 1982a).

7.6 Reconciling irrational and adaptive notions of heuristics?

Referring back to the introductory discussion on different views on heuristics - one being the heuristics-and-biases viewpoint and the other being the ecological rationality viewpoint, our Bayesian extension of heuristics may help a reconciliation between the adaptive and irrational notion of heuristics. For decades, the two most prominent heuristic programs have not been able to come to an agreement on the nature of heuristics and rationality. While the heuristics and biases program (Kahneman, 2003; Tversky & Kahneman, 1974) focused on heuristics deviations from Bayesian rationality as a sign of errors, the fast and frugal program (Gigerenzer et al., 1999) emphasized heuristics simplicity and more psychologically plausible computations in comparison to Bayesian models. We show that the two accounts are not incompatible by making use of parts of both theories. Specifically, we relied on Bayesian inference as rational norms to identify the optimal model (given particular prior and data), echoing Tversky and Kahneman's approach. However, crucially, we do not measure the performance of the heuristic against the optimal Bayesian model - instead - the heuristic represents a particular prior setting (prior strength) on the same Bayesian prior's continuum as the optimal model and any model along the continuum can theoretically be optimal (however this will rarely be the heuristics, as they only corresponds to limiting cases which are outperformed by intermediate priors). At the same time, heuristics are successful when the structure of the environment (environmental prior) matches that of the strategy (strategy's prior strength). This echoes ecological rationality and the interpretation shows a heuristic can be ecologically rational while being compatible with probabilistic inference.

Resulting from this theoretical reconciliation are interesting implications for human rationality. In this framework, different versions of human rationality coincide. Concretely, instead of judging normative rationality (most often referring to models as optimal when accuracy is highest, i.e., economic rationality as mentioned in Tsetsos et al. (2016)), ecological rationality (a model is judged optimal when the strategy matches the environment), and Bayesian rationality (a strategy is optimal when it has the appropriate prior for the environment), as separate and irreconcilable, in the current framework, all forms of rationality overlap and apply at once. Under the assumption that the model is correctly specified, heuristics are rational when they match the structure environment (ecological rationality), which is when they are expected to perform optimal (i.e., a normative rationality), which is when the strategy's prior (i.e., inductive bias) matches the prior of the environment (Bayesian rationality). Thus, just because heuristics excel does not need to mean rational norms have to be abandoned. Heuristics can in fact match all three forms of rationality.

7.7 Conclusions

In conclusion, this thesis developed a novel Bayesian account for heuristics explaining why less can be more. Less-is-more effects have represented one of the single greatest puzzles in the decision making literature for the past decades. While existing explanations based on bias-variance concepts only have limited explanatory power in capturing why heuristics can sometimes excel, the Bayesian framework provided moved beyond these limitations by not only addressing a relative less-is-more effect, that occurs in comparing comparing simple and complex models, but also an absolute less-is-more effect that addresses whether heuristics can be the optimal solution in less-is-more findings.

The Bayesian framework developed in this thesis addressed less-is-more from a novel angle by proposing that heuristics can be thought of as embodying extreme Bayesian priors. Thereby, an explanation for less-is-more is that the heuristics' relative simplicity and inflexibility amounts to a strong inductive bias, that is suitable for many learning and decision problems. Chapters 4 and 5 formalized this idea with two Bayesian models wherein heuristics are an extreme case along a continuum of model flexibility defined by the strength of the prior. Importantly, both Bayesian models included heuristics at one extreme end of the Bayesian prior's strength and a full-information models (regression models) at the other end of the Bayesian prior. The central finding was that models that pertain an intermediately strong prior performed best across all real-world simulations in this thesis, suggesting that down-weighting information is preferable to entirely ignoring it. Chapters 4 and 5 concluded that while a relative less-is-more effect is possible, absolute less-is-more is not, as heuristics will usually be outperformed by an intermediate model that takes into account the full information but weighs it appropriately. Chapter 6 explored whether less is more at the psychological processing level. An active learning experiment was used as a window on people's decision models assuming that people's information-gathering behaviour reflects how they represent and go on to use the information in decisions. Chapter 6 came to the same conclusion, that less is not more, however on the psychological processing level. Chapter 7 drew all Chapters together and discussed implications for future work in psychology and other disciplines, proposing that the potential intermediate models should be identified assessing their psychologically plausability. Chapter 7 also suggested that our understanding of heuristics may need to be revised based on the findings in this thesis.

In sum, the story of heuristics is more complex than it appears at first glance, and only through interdisciplinary novel approaches applying fundamental machine learning concepts to psychological theory is it possible to gain deeper insights and provide a formal understanding of heuristics by placing them in a common probabilistic inference framework.

Appendix A

List of Fast-and-Frugal Heuristics

Heuristic	**Definition**	**Less-is-more effect**
Recognition heuristic (Goldstein & Gigerenzer, 2002)	If one of two alternatives is recognized, infer that it has the higher value on the criterion.	Less-is-more effect with recognizing fewer objects (Goldstein & Gigerenzer, 2002), and with systematic forgetting (Schooler & Hertwig, 2005)
Fluency heuristic (Jacoby & Dallas, 1981)	If both alternatives are recognized but one is recognized fast, infer that it has the higher value on that criterion	Less-is-more effect; systematic forgetting can be beneficial (Schooler & Hertwig, 2005)
Take-The-Best (Gigerenzer & Goldstein, 1996)	see definition in Box 2.3.2	Can predict more accurately than multiple regression (Czerlinski et al., 1999) and other machine learning models (Brighton, 2006)

Tallying (unit-weight linear model) (Dawes, 1979)	To estimate a criterion, do not esatimte weights but simply count the number of positive cues, see definition in Box 2.3.3	Can predict equally or more accurately than multiple regression (Czerlinski et al., 1999)
Satisficing (Simon, 1955; Todd & Miller, 1999)	Search through alternatives and choose the first one that exceeds your aspiration level.	Aspiration levels can lead to significantly better choices than chance, even if they are arbitrary (e.g., the secretary problem, see Gilbert & Mosteller, 1966)
$1/N$, equality heuristic (DeMiguel et al., 2009)	Allocate resources equally to each of N alternatives.	Can outperform optimal asset allocation portfolios. (DeMiguel et al., 2009)
Default heuristic (Johnson & Goldstein, 2003; Pichert & Katsikopoulos, 2008)	If there is a default, do nothing.	Explains why mass mailing has little effect on organ donor registration; predicts behavior when trait and preference theories fail.
Tit-for-tat (Axelrod, 1984)	Cooperate first and then imitate your partners last behavior	Can lead to a higher payoff than optimization (backward induction).
Imitate the majority (Boyd & Richerson, 2005)	Consider the majority of people in your peer group and imitate their behavior	A driving force in bonding, group identification, and moral behavior.

Imitate the successful (Boyd & Richerson, 2005)	Consider the most successful person and imitate his or her behavior	A driving force in cultural evolution.

Table A.1: Description of 10 heuristics in the adaptive toolbox, taken from Gigerenzer and Brighton (2009).

Appendix B

A Description of the 20 Environments

Domain	**Environment**
Psychology	*Attractiveness of men*: Predict average attractiveness of 32 famous men based on the subjects' average likeability ratings of each man, the percentage of subjects who recognized the man's name (subjects saw only the name, no photos), and whether the man was American. (Based on data from a study by Henss, 1996, using 115 male and 131 female Germans, aged 17-66 years) *Attractiveness of women*: Predict average attractiveness of 30 famous women based on the subjects' average likeability ratings of each woman, the percentage of subjects who recognized the woman's name (subjects saw only the name, no photos), and whether the woman was American. (Based on data from a study by Henss, 1996, using 115 male and 131 female Germans, aged 17-66 years)

Sociology	*High school dropout rates*: Predict dropout rate of the 57 Chicago public high schools, given the percentage of low-income students, percentage of nonwhite students, average SAT scores, etc. (Based on Morton, 1995, and Rodkin, 1995) *Homelessness*: Predict the rate of homelessness in 50 US cities given the average temperature, unemployment rate, percentage if inhabitants with incomes below the poverty line, the vacancy rate, whether the city has rent control, and the percentage of public housing. (From Tucker, 1987)
Demography	*Mortality*: Predict the mortality rate in 20 US cities given the average January temperature, pollution level, the percentage of nonwhites, etc. (Based on McDonald & Schwing, 1973; reported in StatLib) *City population*: Predict populations of the 83 German cities with at least 100,000 inhabitants based on whether each city has a soccer team, university, intercity train line, exposition site, etc. (From Fischer Welt Almanach, 1993.)
Economics	*House Price*: Predict the selling price of 22 houses in Erie, PA, based on current property taxes, number of bathrooms, number of bedrooms, lot size, total living space, garage space, age of house, etc. (Based on Narula & Wellington, 1977; reported in Weisberg, 1985)

	Land rent: Predict the rent per acre paid in 58 countries in Minnesota (in 1977 for agricultural land planted in alfalfa) based on the average rent for all tillable land, density of dairy cows, proportion of pasture land, and whether liming is required to grow alfalfa. (Alfalfa is often fed to dairy cows.) (Data provided by Douglas Tiffany; reported in Weisberg, 1985) *Professors' salaries*: Predict the salaries of 51 professors at a midwestern college given gender, rank, number of years in current rank, the highest degree earned, and number of years since highest degree earned (Reported in Weisberg, 1985)
Transportation	*Car Accidents*: Predict the accident rate per million vehicle miles for 37 segments of highway, using the segment's length, average traffic count, percentage of truck volume, speed limit, number of lanes, lane width, shoulder width, number of intersections, etc. for Minnesota in 1973. (Based on an unpublished master's thesis in civil engineering by Carl Hoffstedt; reported in Weisberg, 1985.) *Fuel consumption*: Predict the average motor fuel consumption per person for each of the 48 contiguous United States using the population of the state, number of licensed drivers, fuel tax, per capita income, miles of primary highways etc. (Based on data collected by Christopher Bingham for the American almanac for 1974, except fuel consumption, which was given in the 1974 World Almanac, reported in Weisberg, 1985)

Health	*Obesity at age 18*: Predict fatness at age 18 of 46 children based on body measurements from age 2 to age 18. The body measurements included higher, weight, leg circumference, and strength. (Based on the longitudinal monitoring of the Berkeley Guidance Study, Tuddenham & Snyder, 1954; reported in Weisberg, 1985) *Body fat*: Predict percentage of body fat determined by underwater weighing (a more accurate measure of body fat) using various body circumference measurements (which are more convenient measures than underwater weighing) for 218 men. (Data supplied by A. Garth Fisher form the study of Penrose et al., 1985; reported in StatLib)
Biology	*Fish fertility*: Predict the number of eggs in 395 female Arctic charr based on each fish's weight, its age, and the average weight of its eggs.(Data courtesy of Christian Gillet, 1996) *Mammal's sleep*: Predict the average amount of time 35 species of mammals sleep, based on brain weight, body weight, life span, gestation time, and predation and danger indices (From Allison & Cicchetti 1976; reported in StatLib) *Cow manure*: Predict the amount of oxygen absorbed by dairy wastes given the biological oxygen demand, chemical oxygen demand, total Kjedahl nitrogen, total solids, and total volatile solids for 14 trials (Moore, 1975; reported in Weisberg, 1985)

Environmental Science	*Biodiversity*: Predict the number of species on 26 Galapagos islands, given their area, elevation, distance to the nearest island, area of the nearest island, distance from the coast, etc. (Based on Johnson & Raven, 1973; reported in Weisberg, 1985)
	Rainfall from cloud seeding: Predict the amount of rainfall on 24 days in Coral Gables, FL, given the types of clouds, the percentage of cloud cover, whether the clouds were seeded, number of days since the first day of the experiment, etc. (From Woodley et al., 1977; reported in Weisberg, 1985)
	Oxidant in Los Angeles: Predict the amount of oxidant in Los Angeles for 17 days given each day's wind speed, temperature, humidity, and insolation (a measure of the amount of sunlight). (Data provided by the Los Angeles Pollution Control District; reported in Rice, 1995.)
	Ozone in San Francisco: Predict the amount of ozone in San Francisco on 11 occasions based on the year, average winter precipitation for the past two years, and ozone level in San Jose, at the southern end of the Bay. (From Sandberg et al., 1978; reported in Weisberg, 1985)

Table B.1: Description of the 20 datasets from Czerlinski et al. (1999). The datasets were retrieved from the public repository at http://www-abc.mpib-berlin.mpg.de/sim/Heuristica/environments/. The descriptions of the datasets above are taken from Czerlinski et al. (1999). In each dataset, the cues were either binary or dichotomized at the median, and the task is always to predict which of two objects has the higher criterion value.

Appendix C

Reanalysis of a heuristic dataset (Fig. 2.7*B*)

Linear regression and the TTB heuristic were both fit to one of the original 20 datasets reported by the ABC Research Group Czerlinski et al. (1999). In these original simulations (Czerlinski et al. (1999)), the continuous values were transformed to binary values of 0 and 1 by median split. The criterion variable of the dataset analyzed in Fig. 2.7*B* encodes which of two houses has a higher actual sales price. There are 10 cues, which include things like the number of bedrooms, number of fireplaces, number of garage spaces, living space, current taxes, and the age of the house. We created all 231 possible pairwise comparisons of the original 22 houses. Both the linear regression model and TTB were cross-validated on the dataset by splitting the total number of pairwise comparisons randomly into training and test sets. The size of the training set was 20 comparisons ($\sim$9% of all comparisons) or 100 comparisons ($\sim$43% of all comparisons), and the test set was always the complementary set of comparisons. For each training set size, the cross-validation split into training and test sets was repeated 1000 times and performance of each model was averaged across these replications. Fig. 2.7*B* in the main text demonstrates the generalization performance, i.e., the out-of-sample performance, of both multiple linear regression and TTB as a function of the training set size (small or large). Error bars in Fig. 2.7*B* represent the variation in performance across all cross-validation splits, expressed as standard errors of the mean.

Statistical Parameters in the Simulation

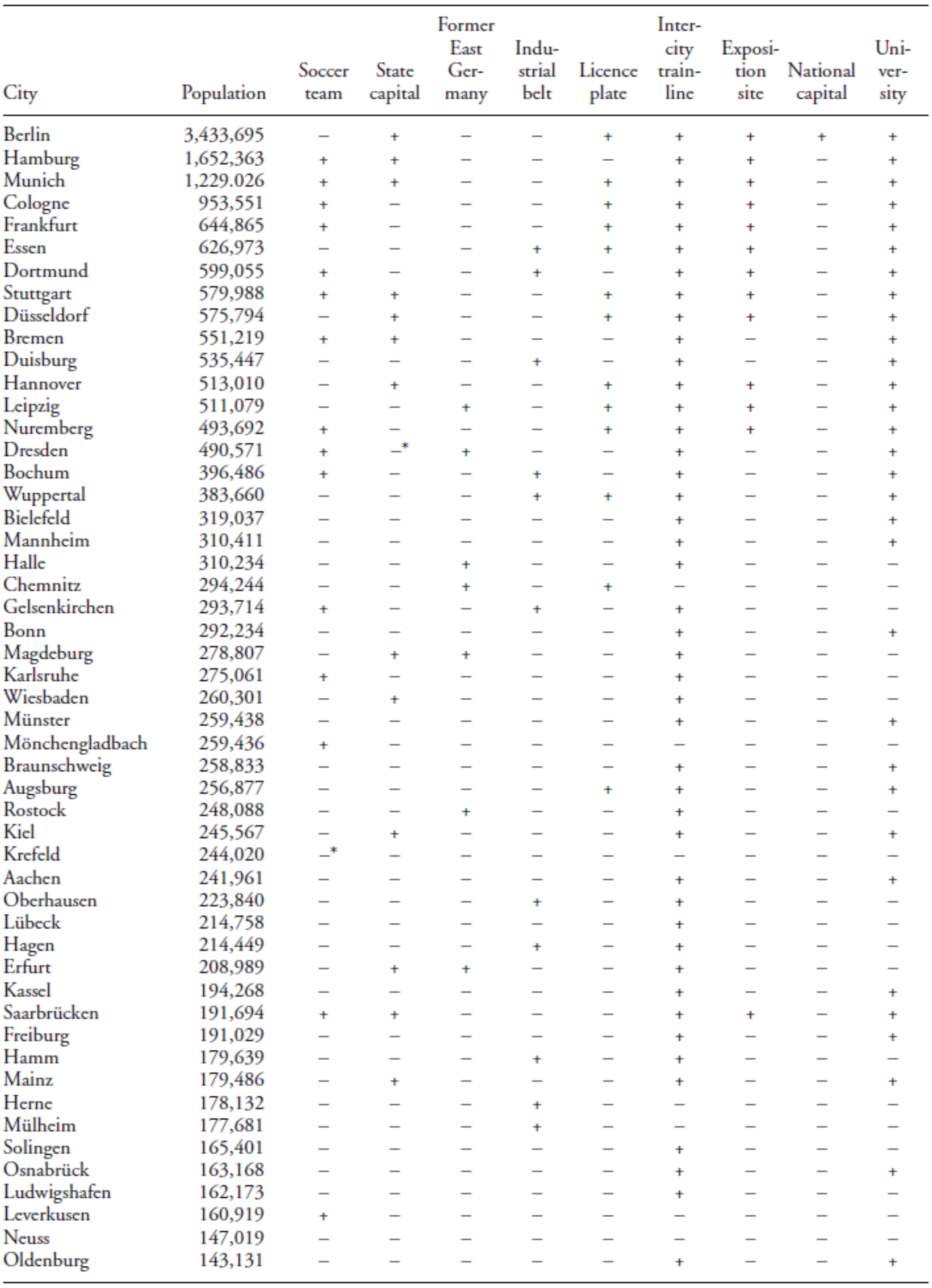
Number of objects	22
Number of pairwise comparisons	$N = 231$
Number of cues	$m = 10$
Class variable (which house had the higher actual sales price)	Binary, ± 1
Absolute correlation between cues averaged over cue pairs	0.35
Sample cue validities	[1.00, 0.99, 0.94, 0.88, 0.83, 0.76, 0.73, 0.73, 0.72, 0.31]
Small training sample size	20 ($\sim$ 9% of all pairwise comparisons)
Large training sample size	100 ($\sim$ 43% of all pairwise comparisons)
Test sample size	$N - 20$, $N - 100$
Number of cross-validation repetitions	1000

Table C.1: Parameters in the house dataset as presented in Fig. 2.7*B*.

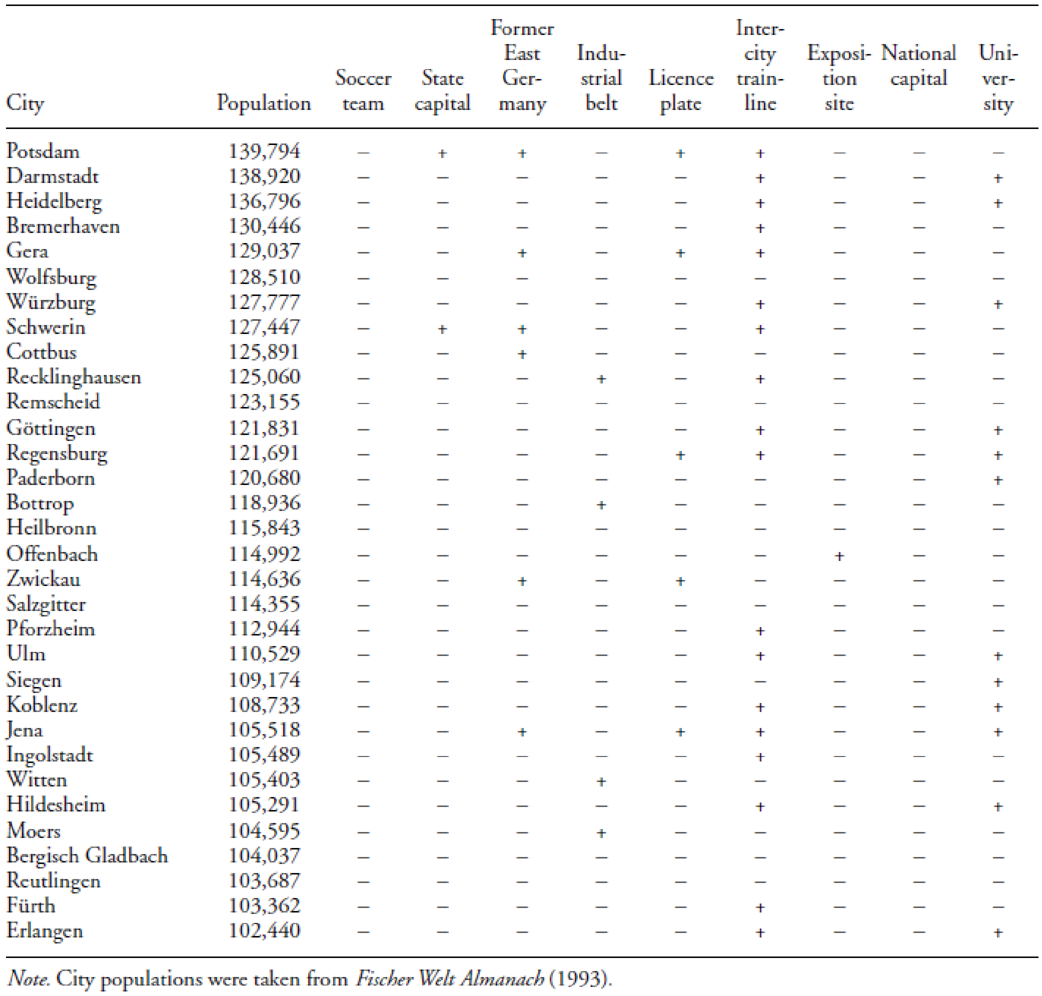
Appendix D

City Size Task - Dataset

D.1 The Environment



City	Population	Soccer team	State capital	Former East Germany	Industrial belt	Licence plate	Intercity trainline	Exposition site	National capital	University
Berlin	3,433,695	–	+	–	–	+	+	+	+	+
Hamburg	1,652,363	+	+	–	–	–	+	+	–	+
Munich	1,229.026	+	+	–	–	+	+	+	–	+
Cologne	953,551	+	–	–	–	+	+	+	–	+
Frankfurt	644,865	+	–	–	–	+	+	+	–	+
Essen	626,973	–	–	–	+	+	+	+	–	+
Dortmund	599,055	+	–	–	+	–	+	+	–	+
Stuttgart	579,988	+	+	–	–	+	+	+	–	+
Düsseldorf	575,794	–	+	–	–	+	+	+	–	+
Bremen	551,219	+	+	–	–	–	+	–	–	+
Duisburg	535,447	–	–	–	+	–	+	–	–	+
Hannover	513,010	–	+	–	–	+	+	+	–	+
Leipzig	511,079	–	–	+	–	+	+	+	–	+
Nuremberg	493,692	+	–	–	–	+	+	+	–	+
Dresden	490,571	+	–*	+	–	–	+	–	–	+
Bochum	396,486	+	–	–	+	–	+	–	–	+
Wuppertal	383,660	–	–	–	+	+	+	–	–	+
Bielefeld	319,037	–	–	–	–	–	+	–	–	+
Mannheim	310,411	–	–	–	–	–	+	–	–	+
Halle	310,234	–	–	+	–	–	+	–	–	–
Chemnitz	294,244	–	–	+	–	+	–	–	–	–
Gelsenkirchen	293,714	+	–	–	+	–	+	–	–	–
Bonn	292,234	–	–	–	–	–	+	–	–	+
Magdeburg	278,807	–	+	+	–	–	+	–	–	–
Karlsruhe	275,061	+	–	–	–	–	+	–	–	–
Wiesbaden	260,301	–	+	–	–	–	+	–	–	–
Münster	259,438	–	–	–	–	–	+	–	–	+
Mönchengladbach	259,436	+	–	–	–	–	–	–	–	–
Braunschweig	258,833	–	–	–	–	–	+	–	–	+
Augsburg	256,877	–	–	–	–	+	+	–	–	+
Rostock	248,088	–	–	+	–	–	+	–	–	–
Kiel	245,567	–	+	–	–	–	+	–	–	+
Krefeld	244,020	–*	–	–	–	–	–	–	–	–
Aachen	241,961	–	–	–	–	–	+	–	–	+
Oberhausen	223,840	–	–	–	+	–	+	–	–	–
Lübeck	214,758	–	–	–	–	–	+	–	–	–
Hagen	214,449	–	–	–	+	–	+	–	–	–
Erfurt	208,989	–	+	+	–	–	+	–	–	–
Kassel	194,268	–	–	–	–	–	+	–	–	+
Saarbrücken	191,694	+	+	–	–	–	+	+	–	+
Freiburg	191,029	–	–	–	–	–	+	–	–	+
Hamm	179,639	–	–	–	+	–	+	–	–	–
Mainz	179,486	–	+	–	–	–	+	–	–	+
Herne	178,132	–	–	–	+	–	–	–	–	–
Mülheim	177,681	–	–	–	+	–	–	–	–	–
Solingen	165,401	–	–	–	–	–	+	–	–	–
Osnabrück	163,168	–	–	–	–	–	+	–	–	+
Ludwigshafen	162,173	–	–	–	–	–	+	–	–	–
Leverkusen	160,919	+	–	–	–	–	–	–	–	–
Neuss	147,019	–	–	–	–	–	–	–	–	–
Oldenburg	143,131	–	–	–	–	–	+	–	–	+



City	Population	Soccer team	State capital	Former East Ger- many	Indu- strial belt	Licence plate	Inter- city train- line	Exposi- tion site	National capital	Uni- ver- sity
Potsdam	139,794	–	+	+	–	+	+	–	–	–
Darmstadt	138,920	–	–	–	–	–	+	–	–	+
Heidelberg	136,796	–	–	–	–	–	+	–	–	+
Bremerhaven	130,446	–	–	–	–	–	+	–	–	–
Gera	129,037	–	–	+	–	+	+	–	–	–
Wolfsburg	128,510	–	–	–	–	–	–	–	–	–
Würzburg	127,777	–	–	–	–	–	+	–	–	+
Schwerin	127,447	–	+	+	–	–	+	–	–	–
Cottbus	125,891	–	–	+	–	–	–	–	–	–
Recklinghausen	125,060	–	–	–	+	–	+	–	–	–
Remscheid	123,155	–	–	–	–	–	–	–	–	–
Göttingen	121,831	–	–	–	–	–	+	–	–	+
Regensburg	121,691	–	–	–	–	+	+	–	–	+
Paderborn	120,680	–	–	–	–	–	–	–	–	+
Bottrop	118,936	–	–	–	+	–	–	–	–	–
Heilbronn	115,843	–	–	–	–	–	–	–	–	–
Offenbach	114,992	–	–	–	–	–	–	+	–	–
Zwickau	114,636	–	–	+	–	+	–	–	–	–
Salzgitter	114,355	–	–	–	–	–	–	–	–	–
Pforzheim	112,944	–	–	–	–	–	+	–	–	–
Ulm	110,529	–	–	–	–	–	+	–	–	+
Siegen	109,174	–	–	–	–	–	–	–	–	+
Koblenz	108,733	–	–	–	–	–	+	–	–	+
Jena	105,518	–	–	+	–	+	+	–	–	+
Ingolstadt	105,489	–	–	–	–	–	+	–	–	–
Witten	105,403	–	–	–	+	–	–	–	–	–
Hildesheim	105,291	–	–	–	–	–	+	–	–	+
Moers	104,595	–	–	–	+	–	–	–	–	–
Bergisch Gladbach	104,037	–	–	–	–	–	–	–	–	–
Reutlingen	103,687	–	–	–	–	–	–	–	–	–
Fürth	103,362	–	–	–	–	–	+	–	–	–
Erlangen	102,440	–	–	–	–	–	+	–	–	+

Note. City populations were taken from *Fischer Welt Almanach* (1993).

Figure D.1: The original city size dataset taken from Gigerenzer and Goldstein (1996). The dataset contains cue values on all 9 binary cues and the population size of all 83 cities.

Appendix E

Mathematical Derivations: TTB as a limiting case of lasso regression?

E.1 L1 regularization

These mathematical derivations attempt to derive the TTB heuristic as a limiting case of L1 regularization. The derivations are developed by my co-author Dr. Matt Jones (University of Colorado, Boulder) and printed here with his approval.

E.2 Posterior Mean for Lasso

To derive the posterior, we start with the special case of a single cue, **X** (matrix notation). The prior for the single weight, w, is Laplacian such that its components are independent and identically distributed:

$$p(w) = \tfrac{\lambda}{2} e^{-\lambda |w|}. \tag{E.1}$$

where the prior's strength is given by λ.

As in the half-ridge model, we assume that the cue directionalities (i.e., the signs of the true weights) are known in advance. The posterior for the single weight, w, is

$$
\begin{aligned}
p\left(w|\mathbf{X},\mathbf{y}\right) &\propto e^{-\lambda|w|-\frac{1}{2\sigma^2}\left((w\mathbf{X}-\mathbf{y})^T(w\mathbf{X}-\mathbf{y})\right)} \\
&\propto \begin{cases} e^{-\frac{1}{2\sigma^2}\left(\mathbf{X}^T\mathbf{X}w^2-2\left(\mathbf{X}^T\mathbf{y}+\lambda\sigma^2\right)w\right)} & w<0 \\ e^{-\frac{1}{2\sigma^2}\left(\mathbf{X}^T\mathbf{X}w^2-2\left(\mathbf{X}^T\mathbf{y}-\lambda\sigma^2\right)w\right)} & w>0 \end{cases} \\
&= \begin{cases} e^{\frac{\left(\mathbf{X}^T\mathbf{y}+\lambda\sigma^2\right)^2}{2\sigma^2\mathbf{X}^T\mathbf{X}}}e^{-\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2}\left(w-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\right)^2} & w<0 \\ e^{\frac{\left(\mathbf{X}^T\mathbf{y}-\lambda\sigma^2\right)^2}{2\sigma^2\mathbf{X}^T\mathbf{X}}}e^{-\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2}\left(w-\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\right)^2} & w>0 \end{cases} \\
&\propto f\left(w;\mathbf{X},\mathbf{y},\lambda\right) = \begin{cases} e^{\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}e^{-\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2}\left(w-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\right)^2} & w<0 \\ e^{-\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}e^{-\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2}\left(w-\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\right)^2} & w>0. \end{cases}
\end{aligned}
$$

The function $f\left(w;\mathbf{X},\mathbf{y},\lambda\right)$ is a pair of truncated Gaussians that meet at 0, with the (common) normalization constant omitted. To determine the mean of the normalized distribution, we use the following relations:

$$
\begin{aligned}
\int_{w>0} e^{-a(w-b)^2}\mathrm{d}w &= \sqrt{\frac{\pi}{a}}\Phi\left(b\sqrt{2a}\right) \\
\int_{w<0} e^{-a(w-b)^2}\mathrm{d}w &= \sqrt{\frac{\pi}{a}}\Phi\left(-b\sqrt{2a}\right) \\
\int_{w>0} we^{-a(w-b)^2}\mathrm{d}w &= \left(-\tfrac{1}{2a}e^{-a(w-b)^2}\right)\Big|_{w=0}^{w=\infty} + \int_{w>0} be^{-a(w-b)^2}\mathrm{d}w \\
&= \tfrac{1}{2a}e^{-ab^2} + b\sqrt{\frac{\pi}{a}}\Phi\left(b\sqrt{2a}\right) \\
\int_{w<0} we^{-a(w-b)^2}\mathrm{d}w &= \left(-\tfrac{1}{2a}e^{-a(w-b)^2}\right)\Big|_{w=-\infty}^{w=0} + \int_{w<0} be^{-a(w-b)^2}\mathrm{d}w \\
&= -\tfrac{1}{2a}e^{-ab^2} + b\sqrt{\frac{\pi}{a}}\Phi\left(-b\sqrt{2a}\right). \qquad \text{(E.2)}
\end{aligned}
$$

From these relations, we have

$$\begin{aligned}\int f(w;\mathbf{X},\mathbf{y},\lambda)\,\mathrm{d}w =&\sqrt{\frac{2\pi\sigma^2}{\mathbf{X}^T\mathbf{X}}}e^{\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)+\\ &\sqrt{\frac{2\pi\sigma^2}{\mathbf{X}^T\mathbf{X}}}e^{-\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)\\ =&\sqrt{\frac{2\pi\sigma^2}{\mathbf{X}^T\mathbf{X}}}e^{\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\left[\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)+e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)\right]\end{aligned}$$

and

$$\begin{aligned}\int wf(w;\mathbf{X},\mathbf{y},\lambda)\,\mathrm{d}w =&e^{\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\left(-\frac{\sigma^2}{\mathbf{X}^T\mathbf{X}}e^{-\frac{\left(\mathbf{X}^T\mathbf{y}+\lambda\sigma^2\right)^2}{2\sigma^2\mathbf{X}^T\mathbf{X}}}+\right.\\ &\left.\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\sqrt{\frac{2\pi\sigma^2}{\mathbf{X}^T\mathbf{X}}}\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)\right)\\ &+e^{-\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\left(\frac{\sigma^2}{\mathbf{X}^T\mathbf{X}}e^{-\frac{\left(\mathbf{X}^T\mathbf{y}-\lambda\sigma^2\right)^2}{2\sigma^2\mathbf{X}^T\mathbf{X}}}+\right.\\ &\left.\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\sqrt{\frac{2\pi\sigma^2}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)\right)\\ =&\sqrt{\frac{2\pi\sigma^2}{\mathbf{X}^T\mathbf{X}}}e^{\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\left[\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)+\right.\\ &\left.\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)\right].\end{aligned}$$

The posterior mean of w is then given by

$$\begin{aligned}\frac{\int wf(w;\mathbf{X},\mathbf{y},\lambda)\,\mathrm{d}w}{\int f(w;\mathbf{X},\mathbf{y},\lambda)\,\mathrm{d}w} &= \frac{\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)+\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)+e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\\ &= \frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}+\frac{\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\frac{\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)-e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)+e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}.\end{aligned}$$

This expression is difficult to evaluate directly. We develop solution methods in the next two subsections.

E.2.1 A. Simplified Case

For simplification, assume $\mathbf{X}^T\mathbf{X} = 1$, $\mathbf{X}^T\mathbf{y} = y$, $\sigma = 1$. The posterior becomes

$$p(w|y) \propto g(w;y,\lambda) = \begin{cases} e^{\lambda y} e^{-\frac{1}{2}(w-y-\lambda)^2} & w < 0 \\ e^{-\lambda y} e^{-\frac{1}{2}(w-y+\lambda)^2} & w > 0. \end{cases}$$

To evaluate the mean of this posterior (and its asymptotic behavior with large λ), we decompose the mean into three components: the posterior probability of $w > 0$, and the conditional means of w given $w > 0$ and given $w < 0$.

The posterior mean conditioned on $w > 0$ can be calculated using Equation E.2 as follows:

$$\begin{aligned} E[w|w>0,y] &= \frac{\int_0^\infty wg(w;y,\lambda)\,\mathrm{d}w}{\int_0^\infty g(w;y,\lambda)\,\mathrm{d}w} \\ &= \frac{\int_0^\infty we^{-\frac{1}{2}(w-y+\lambda)^2}\mathrm{d}w}{\int_0^\infty e^{-\frac{1}{2}(w-y+\lambda)^2}\mathrm{d}w} \\ &= \frac{e^{-\frac{(y-\lambda)^2}{2}} + (y-\lambda)\sqrt{2\pi}\Phi(y-\lambda)}{\sqrt{2\pi}\Phi(y-\lambda)} \\ &= y-\lambda+\frac{\phi(y-\lambda)}{\Phi(y-\lambda)}. \end{aligned}$$

To evaluate how this expression behaves in the limit $\lambda \to \infty$, observe that (by repeated application of l'Hospital's rule)

$$\begin{aligned} \lim_{\lambda\to\infty} \lambda\left(y-\lambda+\frac{\phi(y-\lambda)}{\Phi(y-\lambda)}\right) &= \lim_{\lambda\to\infty} \frac{\left(y\lambda-\lambda^2\right)\Phi(y-\lambda)+\lambda\phi(y-\lambda)}{\Phi(y-\lambda)} \\ &= \lim_{\lambda\to\infty} \frac{(y-2\lambda)\Phi(y-\lambda)+\phi(y-\lambda)}{-\phi(y-\lambda)} \\ &= \lim_{\lambda\to\infty} \frac{-2\Phi(y-\lambda)+\lambda\phi(y-\lambda)}{-(y-\lambda)\phi(y-\lambda)} \\ &= 1. \end{aligned}$$

Therefore the posterior mean, conditioned on a positive weight, is asymptotically inversely proportional to the prior's strength, with no dependence on the cue-outcome correlation (i.e., on y):

$$\lim_{\lambda\to\infty} \lambda E\left[w|w>0,y\right] = 1.$$

We will also need to know the component of order λ^{-2}, which is obtained as follows:

$$\begin{aligned}
\lim_{\lambda\to\infty} \lambda\left(\lambda E\left[w|w>0,y\right]-1\right) &= \lim_{\lambda\to\infty}\left(\lambda^2 y-\lambda^3+\lambda^2\frac{\phi\left(y-\lambda\right)}{\Phi\left(y-\lambda\right)}-\lambda\right)\\
&= \lim_{\lambda\to\infty}\left(\frac{\left(\lambda^2 y-\lambda^3-\lambda\right)\Phi\left(y-\lambda\right)+\lambda^2\phi\left(y-\lambda\right)}{\Phi\left(y-\lambda\right)}\right)\\
&= \lim_{\lambda\to\infty}\left(\frac{\left(2\lambda y-3\lambda^2-1\right)\Phi\left(y-\lambda\right)+3\lambda\phi\left(y-\lambda\right)}{-\phi\left(y-\lambda\right)}\right)\\
&= \lim_{\lambda\to\infty}\left(\frac{\left(2y-6\lambda\right)\Phi\left(y-\lambda\right)+\left(\lambda y+4\right)\phi\left(y-\lambda\right)}{-\left(y-\lambda\right)\phi\left(y-\lambda\right)}\right)\\
&= y.
\end{aligned}$$

The posterior mean conditioned on a negative weight can be calculated as:

$$\begin{aligned}
E\left[w|\mathbf{X},\mathbf{y},w<0\right] &= \frac{\int_{-\infty}^{0} wg\left(w;y,\lambda\right)\mathrm{d}w}{\int_{-\infty}^{0} g\left(w;y,\lambda\right)\mathrm{d}w}\\
&= \frac{\int_{-\infty}^{0} we^{-\frac{1}{2}\left(w-y-\lambda\right)^2}\mathrm{d}w}{\int_{-\infty}^{0} e^{-\frac{1}{2}\left(w-y-\lambda\right)^2}\mathrm{d}w}\\
&= \frac{-e^{-\frac{\left(y+\lambda\right)^2}{2}}+\left(y+\lambda\right)\sqrt{2\pi}\Phi\left(-y-\lambda\right)}{\sqrt{2\pi}\Phi\left(-y-\lambda\right)}\\
&= y+\lambda-\frac{\phi\left(y+\lambda\right)}{\Phi\left(-y-\lambda\right)}.
\end{aligned}$$

This expression behaves similarly to the one above for a positive weight:

$$\lim_{\lambda\to\infty}\lambda\left(y+\lambda-\frac{\phi\left(y+\lambda\right)}{\Phi\left(-y-\lambda\right)}\right) = -1.$$

Thus once again, the conditional posterior mean is asymptotically inversely proportional to the prior's strength, with no dependence on the cue-outcome correlation:

$$\lim_{\lambda\to\infty} \lambda E\left[w|\mathbf{X},\mathbf{y},w<0\right] = -1.$$

Also paralleling the result for $w>0$, the component of order λ^{-2} is given by

$$\begin{aligned}
\lim_{\lambda\to\infty} \lambda\left(\lambda E\left[w|\mathbf{X},\mathbf{y},w<0\right]+1\right) &= \lim_{\lambda\to\infty}\left(y\lambda^2+\lambda^3-\lambda^2\frac{\phi\left(y+\lambda\right)}{\Phi\left(-y-\lambda\right)}+\lambda\right)\\
&= \lim_{\lambda\to\infty}\left(\frac{\left(y\lambda^2+\lambda^3+\lambda\right)\Phi\left(-y-\lambda\right)-\lambda^2\phi\left(y+\lambda\right)}{\Phi\left(-y-\lambda\right)}\right)\\
&= \lim_{\lambda\to\infty}\left(\frac{\left(2y\lambda+3\lambda^2+1\right)\Phi\left(-y-\lambda\right)-3\lambda\phi\left(y+\lambda\right)}{-\phi\left(y+\lambda\right)}\right)\\
&= \lim_{\lambda\to\infty}\left(\frac{\left(2y+6\lambda\right)\Phi\left(-y-\lambda\right)+\left(-4+y\lambda\right)\phi\left(y+\lambda\right)}{\left(y+\lambda\right)\phi\left(y+\lambda\right)}\right)\\
&= y.
\end{aligned}$$

Finally, the odds of a positive weight can be calculated as

$$\begin{aligned}
\frac{\Pr\left[w>0|y\right]}{\Pr\left[w<0|y\right]} &= \frac{\int_0^\infty g\left(w;y,\lambda\right)\mathrm{d}w}{\int_{-\infty}^0 g\left(w;y,\lambda\right)\mathrm{d}w}\\
&= \frac{e^{-2\lambda y}\Phi\left(y-\lambda\right)}{\Phi\left(-y-\lambda\right)}.
\end{aligned}$$

Before evaluating this quantity, note that

$$\begin{aligned}
\lim_{\lambda\to\infty}\frac{e^{-2\lambda y}\Phi\left(y-\lambda\right)}{\phi\left(-y-\lambda\right)} &= \lim_{\lambda\to\infty}\int_{-\infty}^{y-\lambda} e^{-\frac{1}{2}z^2+\frac{1}{2}(y+\lambda)^2-2\lambda y}\mathrm{d}z\\
&= \lim_{\lambda\to\infty}\int_{-\infty}^{0} e^{-\frac{1}{2}(u+y-\lambda)^2+\frac{1}{2}(y-\lambda)^2}\mathrm{d}u\\
&= \lim_{\lambda\to\infty}\int_{-\infty}^{0} e^{-\frac{1}{2}u^2+u(\lambda-y)}\mathrm{d}u\\
&\leq \lim_{\lambda\to\infty}\int_{-\infty}^{0} e^{u(\lambda-y)}\mathrm{d}u\\
&= \lim_{\lambda\to\infty}\frac{1}{\lambda-y}\\
&= 0.
\end{aligned}$$

Evidently, the original limit in question is nonnegative, so we have

$$\lim_{\lambda\to\infty}\frac{e^{-2\lambda y}\Phi(y-\lambda)}{\phi(-y-\lambda)}=0.$$

Now we can show that the odds of a positive weight converge to 1:

$$\begin{aligned}\lim_{\lambda\to\infty}\frac{e^{-2\lambda y}\Phi(y-\lambda)}{\Phi(-y-\lambda)} &= \lim_{\lambda\to\infty}\frac{-2ye^{-2\lambda y}\Phi(y-\lambda)-e^{-2\lambda y}\phi(y-\lambda)}{-\phi(-y-\lambda)}\\ &= e^{-2\lambda y-\frac{(y-\lambda)^2}{2}+\frac{(y+\lambda)^2}{2}}+2y\lim_{\lambda\to\infty}\frac{e^{-2\lambda y}\Phi(y-\lambda)}{\phi(-y-\lambda)}\\ &= e^{-2\lambda y+\lambda y+\lambda y}+0\\ &= 1,\end{aligned}$$

which implies the probability of a positive weight converges to $\frac{1}{2}$. To see how it converges, consider

$$\begin{aligned}&\lim_{\lambda\to\infty}\lambda\left(\frac{e^{-2\lambda y}\Phi(y-\lambda)}{\Phi(-y-\lambda)}-1\right)\\ &=\lim_{\lambda\to\infty}\frac{\lambda e^{-2\lambda y}\Phi(y-\lambda)-\lambda\Phi(-y-\lambda)}{\Phi(-y-\lambda)}\\ &=\lim_{\lambda\to\infty}\frac{(1-2\lambda y)e^{-2\lambda y}\Phi(y-\lambda)-\lambda e^{-2\lambda y}\phi(y-\lambda)-\Phi(-y-\lambda)+\lambda\phi(-y-\lambda)}{-\phi(-y-\lambda)}\\ &=\lim_{\lambda\to\infty}\frac{2\lambda ye^{-2\lambda y}\Phi(y-\lambda)}{\phi(-y-\lambda)}-\lim_{\lambda\to\infty}\frac{e^{-2\lambda y}\Phi(y-\lambda)}{\phi(-y-\lambda)}+\lim_{\lambda\to\infty}\frac{\Phi(-y-\lambda)}{\phi(-y-\lambda)}\\ &=\lim_{\lambda\to\infty}\frac{2y(1-2\lambda y)e^{-2\lambda y}\Phi(y-\lambda)-2\lambda ye^{-2\lambda y}\phi(y-\lambda)}{(-y-\lambda)\phi(-y-\lambda)}-0+\lim_{\lambda\to\infty}\frac{-\phi(-y-\lambda)}{(-y-\lambda)\phi(-y-\lambda)}\\ &=\lim_{\lambda\to\infty}\frac{2y(1-2\lambda y)}{-y-\lambda}\frac{e^{-2\lambda y}\Phi(y-\lambda)}{\phi(-y-\lambda)}+\lim_{\lambda\to\infty}\frac{-2\lambda y\phi(-y-\lambda)}{(-y-\lambda)\phi(-y-\lambda)}+0\\ &=0+\lim_{\lambda\to\infty}\frac{-2\lambda y}{(-y-\lambda)}\\ &=2y.\end{aligned}$$

This relation implies the probability of a positive weight converges to $\frac{1}{2}$ as

$$\begin{aligned}
\lim_{\lambda\to\infty} \lambda\left(\Pr[w>0|y]-\tfrac{1}{2}\right) &= \lim_{\lambda\to\infty} \tfrac{1}{2}\lambda\left(\Pr[w>0|y]-\Pr[w<0|y]\right)\\
&= \lim_{\lambda\to\infty} \frac{\frac{1}{2}\lambda\left(\Pr[w>0|y]-\Pr[w<0|y]\right)}{2\Pr[w<0|y]}\\
&= \tfrac{1}{4}\lim_{\lambda\to\infty}\lambda\left(\frac{\Pr[w>0|y]}{\Pr[w<0|y]}-1\right)\\
&= \frac{y}{2}.
\end{aligned}$$

Combining the above results, we have an asymptotic expression for the posterior mean:

$$\begin{aligned}
\lim_{\lambda\to\infty}\lambda^2 E[w|\mathbf{X},\mathbf{y}] =& \lim_{\lambda\to\infty}\lambda^2\Big(\Pr[w<0|\mathbf{X},\mathbf{y}]E[w|\mathbf{X},\mathbf{y},w<0]+\\
&\Pr[w>0|\mathbf{X},\mathbf{y}]E[w|\mathbf{X},\mathbf{y},w>0]\Big)\\
=&\lim_{\lambda\to\infty}\lambda\left(\Pr[w<0|\mathbf{X},\mathbf{y}]-\tfrac{1}{2}\right)\lambda E[w|\mathbf{X},\mathbf{y},w<0]+\\
&\lambda\left(\Pr[w>0|\mathbf{X},\mathbf{y}]-\tfrac{1}{2}\right)\lambda E[w|\mathbf{X},\mathbf{y},w>0]+\\
&\tfrac{1}{2}\lambda\left(\lambda E[w|\mathbf{X},\mathbf{y},w<0]+1\right))+\tfrac{1}{2}\lambda\left(\lambda E[w|\mathbf{X},\mathbf{y},w>0]-1\right)\\
=&-\frac{y}{2}(-1)+\frac{y}{2}\cdot 1+\frac{y}{2}+\frac{y}{2}\\
=&2y.
\end{aligned}$$

E.2.2 B. General Case

We now solve the general case, with posterior given by

$$p(w|\mathbf{X},\mathbf{y}) \;\propto\; f(w;\mathbf{X},\mathbf{y},\lambda)=\begin{cases} e^{\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}e^{-\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2}\left(w-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\right)^2} & w<0\\ e^{-\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}e^{-\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2}\left(w-\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\right)^2} & w>0.\end{cases}$$

The posterior mean conditioned on $w>0$ can be calculated using Equation E.2 as follows:

$$
\begin{aligned}
E\left[w|w>0,\mathbf{X},\mathbf{y}\right] &= \frac{\int_0^\infty w e^{-\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2}\left(w-\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\right)^2}\mathrm{d}w}{\int_0^\infty e^{-\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2}\left(w-\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\right)^2}\mathrm{d}w} \\
&= \frac{\frac{\sigma^2}{\mathbf{X}^T\mathbf{X}}e^{-\frac{\left(\mathbf{X}^T\mathbf{y}-\lambda\sigma^2\right)^2}{2\sigma^2\mathbf{X}^T\mathbf{X}}}+\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\sqrt{\frac{2\pi\sigma^2}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\sqrt{\frac{2\pi\sigma^2}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)} \\
&= \frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}+\frac{\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}\frac{\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}.
\end{aligned}
$$

This expression converges to 0 and is proportional to λ^{-1} with constant of proportionality equal to

$$
\begin{aligned}
&\lim_{\lambda\to\infty}\lambda\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}+\frac{\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}\frac{\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\right) \\
&=\lim_{\lambda\to\infty}\frac{\left(\mathbf{X}^T\mathbf{y}\lambda-\lambda^2\sigma^2\right)\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)+\lambda\sigma\sqrt{\mathbf{X}^T\mathbf{X}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\mathbf{X}^T\mathbf{X}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)} \\
&=\lim_{\lambda\to\infty}\frac{\left(\mathbf{X}^T\mathbf{y}-2\lambda\sigma^2\right)\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)+\sigma\sqrt{\mathbf{X}^T\mathbf{X}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{-\sigma\sqrt{\mathbf{X}^T\mathbf{X}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)} \\
&=\lim_{\lambda\to\infty}\frac{-2\sigma^2\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)+\frac{\lambda\sigma^3}{\sqrt{\mathbf{X}^T\mathbf{X}}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{-\frac{\sigma\left(\mathbf{X}^T\mathbf{y}-\lambda\sigma^2\right)}{\sqrt{\mathbf{X}^T\mathbf{X}}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)} \\
&=1.
\end{aligned}
$$

The component of order λ^{-2} is given by

$$\begin{aligned}
&\lim_{\lambda\to\infty} \lambda\left(\lambda\left(E\left[w|w>0,\mathbf{X},\mathbf{y}\right]\right)-1\right)\\
=&\lim_{\lambda\to\infty}\left(\lambda^2\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}+\lambda^2\frac{\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}\frac{\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}-\lambda\right)\\
=&\lim_{\lambda\to\infty}\left(\frac{\left(\mathbf{X}^T\mathbf{y}\lambda^2-\lambda^3\sigma^2-\mathbf{X}^T\mathbf{X}\lambda\right)\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)+\lambda^2\sigma\sqrt{\mathbf{X}^T\mathbf{X}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\mathbf{X}^T\mathbf{X}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\right)\\
=&\lim_{\lambda\to\infty}\left(\frac{\left(2\mathbf{X}^T\mathbf{y}\lambda-3\lambda^2\sigma^2-\mathbf{X}^T\mathbf{X}\right)\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)+3\lambda\sigma\sqrt{\mathbf{X}^T\mathbf{X}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{-\sigma\sqrt{\mathbf{X}^T\mathbf{X}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\right)\\
=&\lim_{\lambda\to\infty}\left(\frac{\left(2\mathbf{X}^T\mathbf{y}-6\lambda\sigma^2\right)\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)+\frac{\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}\left(\mathbf{X}^T\mathbf{y}\lambda+4\mathbf{X}^T\mathbf{X}\right)\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{-\frac{\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}\left(\mathbf{X}^T\mathbf{y}-\lambda\sigma^2\right)\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\right)\\
=&\frac{\mathbf{X}^T\mathbf{y}}{\sigma^2}.
\end{aligned}$$

For $w<0$, the posterior conditional mean is

$$\begin{aligned}
E\left[w|w<0,\mathbf{X},\mathbf{y}\right] &= \frac{\int_0^\infty we^{-\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2}\left(w-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\right)^2}\mathrm{d}w}{\int_0^\infty e^{-\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2}\left(w-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\right)^2}\mathrm{d}w}\\
&= \frac{-\frac{\sigma^2}{\mathbf{X}^T\mathbf{X}}e^{-\frac{\left(\mathbf{X}^T\mathbf{y}+\lambda\sigma^2\right)^2}{2\sigma^2\mathbf{X}^T\mathbf{X}}}+\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}\sqrt{\frac{2\pi\sigma^2}{\mathbf{X}^T\mathbf{X}}}\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\sqrt{\frac{2\pi\sigma^2}{\mathbf{X}^T\mathbf{X}}}\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\\
&= \frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}-\frac{\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}\frac{\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}.
\end{aligned}$$

This expression converges to 0 and is proportional to λ^{-1} with constant of proportionality equal to

$$\begin{aligned}
&\lim_{\lambda\to\infty}\lambda\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}-\frac{\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}\frac{\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\right)\\
=&\lim_{\lambda\to\infty}\frac{\left(\mathbf{X}^T\mathbf{y}\lambda+\lambda^2\sigma^2\right)\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)-\lambda\sigma\sqrt{\mathbf{X}^T\mathbf{X}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\mathbf{X}^T\mathbf{X}\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\\
=&\lim_{\lambda\to\infty}\frac{\left(\mathbf{X}^T\mathbf{y}+2\lambda\sigma^2\right)\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)-\sigma\sqrt{\mathbf{X}^T\mathbf{X}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{-\sigma\sqrt{\mathbf{X}^T\mathbf{X}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\\
=&\lim_{\lambda\to\infty}\frac{2\sigma^2\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)+\frac{-\lambda\sigma^3}{\sqrt{\mathbf{X}^T\mathbf{X}}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\frac{\sigma\left(\mathbf{X}^T\mathbf{y}+\lambda\sigma^2\right)}{\sqrt{\mathbf{X}^T\mathbf{X}}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\\
=&-1.
\end{aligned}$$

The component of order λ^{-2} is given by

$$\begin{aligned}
&\lim_{\lambda\to\infty}\lambda\left(\lambda\left(E\left[w|w<0,\mathbf{X},\mathbf{y}\right]\right)+1\right)\\
=&\lim_{\lambda\to\infty}\left(\lambda^2\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\mathbf{X}^T\mathbf{X}}-\lambda^2\frac{\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}\frac{\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}+\lambda\right)\\
=&\lim_{\lambda\to\infty}\left(\frac{\left(\mathbf{X}^T\mathbf{y}\lambda^2+\lambda^3\sigma^2+\mathbf{X}^T\mathbf{X}\lambda\right)\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)-\lambda^2\sigma\sqrt{\mathbf{X}^T\mathbf{X}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\mathbf{X}^T\mathbf{X}\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\right)\\
=&\lim_{\lambda\to\infty}\left(\frac{\left(2\mathbf{X}^T\mathbf{y}\lambda+3\lambda^2\sigma^2+\mathbf{X}^T\mathbf{X}\right)\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)-3\lambda\sigma\sqrt{\mathbf{X}^T\mathbf{X}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{-\sigma\sqrt{\mathbf{X}^T\mathbf{X}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\right)\\
=&\lim_{\lambda\to\infty}\left(\frac{\left(2\mathbf{X}^T\mathbf{y}+6\lambda\sigma^2\right)\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)+\frac{\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}\left(\mathbf{X}^T\mathbf{y}\lambda-4\mathbf{X}^T\mathbf{X}\right)\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\frac{\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}\left(\mathbf{X}^T\mathbf{y}+\lambda\sigma^2\right)\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\right)\\
=&\frac{\mathbf{X}^T\mathbf{y}}{\sigma^2}.
\end{aligned}$$

Finally, the odds of a positive weight can be calculated as

$$\begin{aligned}\frac{\Pr\left[w>0|\mathbf{X},\mathbf{y}\right]}{\Pr\left[w<0|\mathbf{X},\mathbf{y}\right]} &= \frac{\int_0^\infty f\left(w;\mathbf{X},\mathbf{y},\lambda\right)\mathrm{d}w}{\int_{-\infty}^0 f\left(w;\mathbf{X},\mathbf{y},\lambda\right)\mathrm{d}w}\\ &= \frac{e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}.\end{aligned}$$

Asymptotically, the odds approach 1:

$$\begin{aligned}&\lim_{\lambda\to\infty}\frac{e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\\ &=\lim_{\lambda\to\infty}\frac{-\frac{2\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)-\frac{\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{-\frac{\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}\phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\\ &=e^{\frac{\left(\mathbf{X}^T\mathbf{y}+\lambda\sigma^2\right)^2-\left(\mathbf{X}^T\mathbf{y}-\lambda\sigma^2\right)^2}{2\sigma^2\mathbf{X}^T\mathbf{X}}-\frac{2\lambda\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}+\frac{2\mathbf{X}^T\mathbf{y}}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\lim_{\lambda\to\infty}\frac{e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\\ &=1,\end{aligned}$$

with the component of order $\lambda^{=1}$ given by

$$\begin{aligned}
&\lim_{\lambda\to\infty}\lambda\left(\frac{e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}-1\right)\\
=&\lim_{\lambda\to\infty}\frac{\lambda e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)-\lambda\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\\
=&\lim_{\lambda\to\infty}\frac{\left(1-\frac{2\lambda\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}\right)e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)-\frac{\lambda\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{-\frac{\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}+\\
&\frac{-\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)+\frac{\lambda\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}\phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{-\frac{\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\\
=&\lim_{\lambda\to\infty}\frac{2\lambda\mathbf{X}^T\mathbf{y}e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}-\lim_{\lambda\to\infty}\frac{\sqrt{\mathbf{X}^T\mathbf{X}}e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\sigma\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}+\\
&\lim_{\lambda\to\infty}\frac{\sqrt{\mathbf{X}^T\mathbf{X}}\Phi\left(-\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\sigma\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\\
=&\lim_{\lambda\to\infty}\frac{2\mathbf{X}^T\mathbf{y}\left(1-\frac{2\lambda\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}\right)e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)-\frac{2\lambda\sigma\mathbf{X}^T\mathbf{y}}{\sqrt{\mathbf{X}^T\mathbf{X}}}e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{-\frac{\sigma}{\sqrt{\mathbf{X}^T\mathbf{X}}}\left(\mathbf{X}^T\mathbf{y}+\lambda\sigma^2\right)\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}-0+0\\
=&\lim_{\lambda\to\infty}\frac{2\mathbf{X}^T\mathbf{y}\left(\mathbf{X}^T\mathbf{X}-2\lambda\mathbf{X}^T\mathbf{y}\right)}{-\sigma\sqrt{\mathbf{X}^T\mathbf{X}}\left(\mathbf{X}^T\mathbf{y}+\lambda\sigma^2\right)}\frac{e^{-2\lambda\frac{\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{X}}}\Phi\left(\frac{\mathbf{X}^T\mathbf{y}-\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}+\lim_{\lambda\to\infty}\frac{2\lambda\mathbf{X}^T\mathbf{y}\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}{\left(\mathbf{X}^T\mathbf{y}+\lambda\sigma^2\right)\phi\left(\frac{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}{\sigma\sqrt{\mathbf{X}^T\mathbf{X}}}\right)}\\
=&0+\lim_{\lambda\to\infty}\frac{2\lambda\mathbf{X}^T\mathbf{y}}{\mathbf{X}^T\mathbf{y}+\lambda\sigma^2}\\
=&\frac{2\mathbf{X}^T\mathbf{y}}{\sigma^2}.
\end{aligned}$$

Therefore the posterior probability of a positive weight obeys

$$\begin{aligned}
\lim_{\lambda\to\infty} \lambda \left(\Pr[w>0|\mathbf{X},\mathbf{y}] - \tfrac{1}{2}\right) &= \lim_{\lambda\to\infty} \tfrac{1}{2}\lambda \left(\Pr[w>0|\mathbf{X},\mathbf{y}] - \Pr[w<0|\mathbf{X},\mathbf{y}]\right) \\
&= \lim_{\lambda\to\infty} \frac{\tfrac{1}{2}\lambda \left(\Pr[w>0|\mathbf{X},\mathbf{y}] - \Pr[w<0|\mathbf{X},\mathbf{y}]\right)}{2\Pr[w<0|\mathbf{X},\mathbf{y}]} \\
&= \tfrac{1}{4}\lim_{\lambda\to\infty} \lambda \left(\frac{\Pr[w>0|\mathbf{X},\mathbf{y}]}{\Pr[w<0|\mathbf{X},\mathbf{y}]} - 1\right) \\
&= \frac{\mathbf{X}^T\mathbf{y}}{2\sigma^2}.
\end{aligned}$$

Combining all the above results, we have an asymptotic expression for the posterior mean:

$$\begin{aligned}
&\lim_{\lambda\to\infty} \lambda^2 E[w|\mathbf{X},\mathbf{y}] \\
&= \lim_{\lambda\to\infty} \lambda^2 \left(\Pr[w<0|\mathbf{X},\mathbf{y}] E[w|w<0,\mathbf{X},\mathbf{y}] + \Pr[w>0|\mathbf{X},\mathbf{y}] E[w|w>0,\mathbf{X},\mathbf{y}]\right) \\
&= \lim_{\lambda\to\infty} \lambda \left(\Pr[w<0|\mathbf{X},\mathbf{y}] - \tfrac{1}{2}\right) \lambda E[w|w<0,\mathbf{X},\mathbf{y}] + \\
&\quad \lambda \left(\Pr[w>0|\mathbf{X},\mathbf{y}] - \tfrac{1}{2}\right) \lambda E[w|w>0,\mathbf{X},\mathbf{y}] + \\
&\quad \tfrac{1}{2}\lambda \left(\lambda E[w|w<0,\mathbf{X},\mathbf{y}] + 1)\right) + \tfrac{1}{2}\lambda \left(\lambda E[w|w>0,\mathbf{X},\mathbf{y}] - 1\right) \\
&= -\frac{\mathbf{X}^T\mathbf{y}}{2\sigma^2}(-1) + \frac{\mathbf{X}^T\mathbf{y}}{2\sigma^2}\cdot 1 + \frac{\mathbf{X}^T\mathbf{y}}{2\sigma^2} + \frac{\mathbf{X}^T\mathbf{y}}{2\sigma^2} \\
&= \frac{2\mathbf{X}^T\mathbf{y}}{\sigma^2}.
\end{aligned}$$

E.2.3 C. Conclusion

If the directionalities of the weights are taken as known, then all weights are asymptotically identical (and independent of the data). Thus L1-penalized truncated regression converges to tallying, just like in the ridge regression case (L2 regularization). If the directionalities of the weights are unknown, then L1-penalized regression converges to a scheme that uses all the predictors, with relative weights determined by their respective inner products with the data. The one caveat to these results is that they ignore collinearity between predictors, but we conjecture this factor becomes irrelevant for large penalty parameters (as is true in the L2 case). When ignoring collinearity, the derivations of the posterior show that all cue weights become equal in the limit.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological science*, *2*(6), 396–408.

Benartzi, S., & Thaler, R. H. (2001). Naive diversification strategies in defined contribution saving plans. *American economic review*, 79–98.

Bergert, F. B., & Nosofsky, R. M. (2007). A response-time approach to comparing generalized rational and take-the-best models of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 107.

Boole, G. (1854). *An investigation of the laws of thought on which are founded the mathematical theories of logic and probabilities by george boole*. Walton and Maberly.

Bowling, M., Burch, N., Johanson, M., & Tammelin, O. (2015). Heads-up limit holdem poker is solved. *Science*, *347*(6218), 145–149.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing neuraths ship: Approximate algorithms for online causal learning. *Psychological review*, *124*(3), 301.

Brighton, H. (2006). Robust inference with simple cognitive models. In *Aaai spring symposium: Between a rock and a hard place: Cognitive science principles meet ai-hard problems* (pp. 17–22).

Brighton, H., & Gigerenzer, G. (2008). Bayesian brains and cognitive mechanisms: Harmony or dissonance. *The probabilistic mind: Prospects for Bayesian cognitive science, ed. N. Chater & M. Oaksford*, 189–208.

Bröder, A. (2000). Assessing the empirical validity of the" take-the-best" heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1332.

Bröder, A., & Gaissmaier, W. (2007). Sequential processing of cues in memory-based multiattribute decisions. *Psychonomic Bulletin & Review*, *14*(5), 895–900.

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes [Journal Article]. *Cogn Psychol*, *58*(1), 49-67.

Chaloner, K., & Larntz, K. (1989). Optimal bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, *21*(2), 191–208.

Chater, N., Oaksford, M., Nakisa, R., & Redington, M. (2003). Fast, frugal, and rational: How rational norms explain behavior [Journal Article]. *Organizational behavior and human decision processes*, *90*(1), 63-86.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: where next? *Trends in Cognitive Sciences*, *10*(7), 292–293.

Coenen, A., Nelson, J. D., & Gureckis, T. (2017). Asking the right questions about human inquiry.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive psychology*, *79*, 102–133.

Cohen, J. (1988). Statistical power analyses for the social sciences. *Hillsdale, NJ, Lawrence Erlbauni Associates*.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21–27.

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? [Book Section]. In G. Gigerenzer, P. Todd, & A. R. Group (Eds.), *Simple heuristics that make us smart* (p. 97118). New York: Oxford University Press.

Daston, L. J. (1980). Probabilistic expectation and rationality in classical probability theory. *Historia Mathematica*, *7*(3), 234–260.

Davis-Stober, C. P., Dana, J., & Budescu, D. V. (2010). A constrained linear estimator for multiple regression. *Psychometrika*, *75*(3), 521–541.

Daw, N., & Courville, A. (2008). *The pigeon as particle filter.* (Vol. 20) [Conference Paper]. MIT Press.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making [Journal Article]. *American psychologist*, *34*(7), 571.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making [Journal Article]. *Psychological bulletin*, *81*(2), 95.

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies*, *22*(5), 1915–1953.

Dieckmann, A., & Rieskamp, J. (2007). The influence of information redundancy on probabilistic inferences [Journal Article]. *Mem Cognit*, *35*(7), 1801-13.

Doucet, A., De Freitas, N., & Gordon, N. (2001). An introduction to sequential monte carlo methods. In *Sequential monte carlo methods in practice* (pp. 3–14). Springer.

Dougherty, M. R., Franco-Watkins, A. M., & Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental models and the fast and frugal heuristics. *Psychological review*, *115*(1), 199.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making [Journal Article]. *Organizational Behavior and Human Performance*, *13*(2), 171-192.

Friedman. (1953). *Essays in positive economics*. University of Chicago Press.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics Springer, Berlin.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma [Journal Article]. *Neural computation*, *4*(1), 1-58.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*(6), 721–741.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. Penguin.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: why biased minds make better inferences [Journal Article]. *Top Cogn Sci*, *1*(1), 107-43.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making [Journal Article]. *Annu Rev Psychol*, *62*, 451-82.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality [Journal Article]. *Psychol Rev*, *103*(4), 650-69.

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: take the best and its relatives. *Simple Heuristics that Make Us Smart. Oxford University Press, New York*, 75–95.

Gigerenzer, G., Hertwig, R., Hoffrage, U., & Sedlmeier, P. (2008). Cognitive illusions reconsidered. *Handbook of experimental economics results*, *1*, 1018–1034.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Psychology Press.

Gigerenzer, G., Todd, P. M., & Group, A. R. (1999). *Simple heuristics that make us smart* [Book]. Oxford University Press.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing markov chain monte carlo. *Markov chain Monte Carlo in practice*, *1*, 19.

Glöckner, A., & Betsch, T. (2008). Multiple-reason decision making based on automatic processing. *Journal of experimental psychology: Learning, memory, and cognition*, *34*(5), 1055.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: the recognition heuristic [Journal Article]. *Psychol Rev*, *109*(1), 75-90.

Gramacy, R. B., & Lee, H. K. (2008). Gaussian processes and limiting linear models. *Computational Statistics & Data Analysis*, *53*(1), 123–136.

Griffiths, T. L., & Austerweil, J. L. (2009). Analyzing human feature learning as nonparametric bayesian inference. In *Advances in neural information processing systems* (pp. 97–104).

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic [Journal Article]. *Topics in cognitive science*, *7*(2), 217-229.

Groner, M., Groner, R., & Bischof, W. F. (1983). Approaches to heuristics: A historical review. *Methods of heuristics*, 1–18.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems [Journal Article]. *Technometrics*, *12*(1), 55-67.

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2013). Deliberations blindsight: How cognitive load can improve judgments. *Psychological science*, *24*(6), 869–879.

Hogarth, R. M., & Karelaia, N. (2005). Ignoring information in binary choice with continuous variables: When is less more? *Journal of Mathematical Psychology*, *49*(2), 115–124.

Hogarth, R. M., & Karelaia, N. (2006). take-the-best and other simple strategies: Why and when they work well with binary cues. *Theory and Decision*, *61*(3), 205–249.

Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of judgment: matching rules and environments. *Psychological review*, *114*(3), 733.

Holton, G. J. (1988). *Thematic origins of scientific thought: Kepler to einstein*. Harvard University Press.

Hsu, A. S., & Chater, N. (2010). The logical problem of language acquisition: A probabilistic perspective. *Cognitive science*, *34*(6), 972–1016.

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in cognitive sciences*, *6*(8), 345–350.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, *110*(3), 306.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition [Journal Article]. *Behav Brain Sci*, *34*(4), 169-88; disuccsion 188-231.

Juslin, P., & Persson, M. (2002). Probabilities from exemplars (probex): A lazy algorithm for probabilistic inference from generic knowledge. *Cognitive science*, *26*(5), 563–607.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality [Journal Article]. *Am Psychol*, *58*(9), 697-720.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. In *The concept of probability in psychological experiments* (pp. 25–48). Springer.

Katsikopoulos, K. V., Schooler, L. J., & Hertwig, R. (2010). The robust beauty of ordinary information [Journal Article]. *Psychological Review*, *117*(4), 1259.

Keller, N., & Katsikopoulos, K. V. (2016). On the role of psychological heuristics in operational research; and a demonstration in military stability operations [Journal Article]. *European Journal of Operational Research*, *249*(3), 1063-1073.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as bayesian inference. *Annu. Rev. Psychol.*, *55*, 271–304.

Knox, W. B., Otto, A. R., Stone, P., & Love, B. C. (2011). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in psychology*, *2*.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection [Conference Proceedings]. In *Ijcai* (Vol. 14, p. 1137-1145).

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247.

Lee, M. D., & Cummins, T. D. (2004). Evidence accumulation in decision making: unifying the ”take the best” and the ”rational” models [Journal Article]. *Psychon Bull Rev*, *11*(2), 343-52.

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2017). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, 1–28.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature neuroscience*, *9*(11), 1432–1438.

Marewski, J. N., Gaissmaier, W., & Gigerenzer, G. (2010). Good judgments do not require complex cognition [Journal Article]. *Cognitive processing*, *11*(2), 103-121.

Marewski, J. N., & Mehlhorn, K. (2011). Using the act-r architecture to specify 39 quantitative process models of decision making. *Judgment and Decision Making*, *6*(6), 439–519.

Markant, & Gureckis, T. M. (2012). One piece at a time: Learning complex rules through self-directed sampling. In *Proceedings of the 34th annual conference of the cognitive science society. austin, tx: Cognitive science society.*

Markant, & Gureckis, T. M. (2013). Is it better to select or to receive? learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94.

Marquaridt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation [Journal Article]. *Technometrics*, *12*(3), 591-612.

Marr, D. (1982a). *Vision: A computational investigation into the human representation and processing of visual information* [Aggregated Database]. Freeman.

Marr, D. (1982b). Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY*, *2*, 4–2.

Martignon, L., & Hoffrage, U. (1999). Why does one-reason decision making work. a case study in ecological rationality. [Book Section]. In *Simple heuristics that make us smart* (p. 119-140). New York: Oxford University Press.

Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, *52*(1), 29–71.

Martignon, L., & Laskey, K. B. (1999). Bayesian benchmarks for fast and frugal heuristics.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? what does failure to replicate really mean? *American Psychologist*, *70*(6), 487.

McLeod, P., & Dienes, Z. (1996). Do fielders know where to go to catch the ball or only how to get there? *Journal of Experimental Psychology: Human Perception and Performance*, *22*(3), 531.

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*(2), 119–148.

Newall, P. (2011). *The intelligent poker player.* Two Plus Two Publishing.

Newall, P. (2013). *Further limit hold em: Exploring the model poker game.* Two Plus Two Publishing.

Newell, B. R., & Shanks, D. R. (2003). Take the best or look at the rest? factors influencing" one-reason" decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 53.

Newell, B. R., Weston, N. J., & Shanks, D. R. (2003). Empirical tests of a fast-and-frugal heuristic: Not everyone takes-the-best. *Organizational Behavior and Human Decision Processes*, *91*(1), 82–96.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, *34*(4), 393–418.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, *65*(2), 207–240.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker.* Cambridge University Press.

Pearl, J. (1984). Heuristics: intelligent search strategies for computer problem solving.

Pearl, J. (2000). *Causality: models, reasoning, and inference.* Cambridge University Press Cambridge, UK:.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad [Journal Article]. *Trends Cogn Sci*, *6*(10), 421-425.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature neuroscience*, *16*(9), 1170–1178.

Quinlan, J. R. (1993). C4. 5: Programming for machine learning. *Morgan Kauffmann*, *38*.

Raiffa, H., & Schlaifer, R. (1961). Applied statistical decision theory. *Division of Research, Harvard Business School, Boston, MA*.

Rieskamp, J., & Dieckmann, A. (2012). Redundancy: Environment structure that simple heuristics can exploit [Book Section]. In *Ecological rationality: Intelligence in the world* (p. 187-215). New York: Oxford University Press.

Rieskamp, J., & Otto, P. E. (2006). Ssl: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(2), 207.

Ripley, B. D. (2007). *Pattern recognition and neural networks* [Book]. Cambridge university press.

Rumelhart, D. E., McClelland, J. L., Group, P. R., et al. (1986). Parallel distributed processing, vol. 1&2. *Cambridge, MA: The MIT Press*.

Russell, S. J., & Norvig, P. (2002). Artificial intelligence: a modern approach (international edition).

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning [Journal Article]. *Psychological review*, *117*(4), 1144.

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A bayesian hierarchical approach [Journal Article]. *Psychological review*, *120*(1), 39.

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological review*, *112*(3), 610.

Schulz, E., Speekenbrink, M., & Shanks, D. R. (2014). Predict choice: A comparison of 21 mathematical models..

Scott, J. (2000). Rational choice theory. *Understanding contemporary society: Theories of the present*, *129*.

Shannon. (1948). Ba mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379423.

Simon, H. A. (1990). Invariants of human behavior. *Annual review of psychology*, *41*(1), 1–20.

Simon, H. A., et al. (1989). The scientist as problem solver. *Complex information processing: The impact of Herbert A. Simon*, 375–398.

Sloman, S. A., & Lagnado, D. (2005). The problem of induction. *The Cambridge handbook of thinking and reasoning*, 95–116.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59–71.

Tenenbaum, J. B. (1999). Rules and similarity in concept learning. In *Nips* (pp. 59–65).

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and brain sciences*, *24*(04), 629–640.

Todd, P. M., & Gigerenzer, G. (2000). Précis of simple heuristics that make us smart. *Behavioral and brain sciences*, *23*(05), 727–741.

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences*, *113*(11), 31023107.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases [Journal Article]. *Science*, *185*(4157), 1124-31.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, *90*(4), 293.

van Ravenzwaaij, D., Moore, C. P., Lee, M. D., & Newell, B. R. (2014). A hierarchical bayesian modeling approach to searching and stopping in multi-attribute judgment [Journal Article]. *Cogn Sci*, *38*(7), 1384-405.

Von Neumann, J., & Morgenstern, O. (1944,1947,1953,2007). *Theory of games and economic behavior*. Princeton university press.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature neuroscience*, *5*(6), 598–604.

Yuille, A., & Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, *10*(7), 301–308.