Modelling Human Uncertainty in Predictive Data Mining

Development of a Neuro-Stochastic Model to Describe Unreliable User Feedback and its Impact on User-Adaptive Information Systems

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie (Dr. phil.)
durch die Philosophische Fakultät der
Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Kevin Jasberg
aus Essen

Betreuer: Dr. Dr. Sergej Sizov

Düsseldorf, Juli 2020

I dedicate this dissertation to my grandfather
Bruno Gwiasdowski *in memoriam.*

He was a great and generous person
to whom I owe a lot.

Acknowledgement

I would like to take this opportunity to express my gratitude to those people who have accompanied me during the stages of my research and the process of creating this thesis.

In particular, I would like to thank Sergei Sizov for his excellent support. As my supervisor, he taught me everything a scientist needs to succeed, especially how to produce interesting research questions beyond the mainstream and how to publish them successfully. He always did his utmost to provide me with the best possible support. He made me think beyond my borders. Thanks for the patience and effort!

At this point, I would also like to thank Wiebke Petersen. She supported me both academically and emotionally as my co-supervisor. Her feedback was crucial for the success of this dissertation and contributed significantly to its quality. Moreover, she provided me with follow-up financing after the closure of my department. Without her help, the realisation of this work would not have been possible. Thank you very much!

Furthermore, I would like to thank Laura Kallmeyer for funding an important conference trip. This allowed me to present a large part of my research to the scientific community and to receive important feedback. In this sense, I also wish to thank Markus Werning for the expert discussions and for the opportunity to present my research in his colloquium to receive valuable feedback. Last but not least, I would like to thank all my friends who have always supported and encouraged me.

Thank You!

Modelling Human Uncertainty in Predictive Data Mining

Development of a Neuro-Stochastic Model to Describe Unreliable User Feedback and its Impact on User-Adaptive Information Systems

submitted by
Kevin Jasberg

Abstract

The present dissertation was developed within the scope of "Systems with Empathy for the Human Nature", i.e. systems that explain individual human behaviour along with its peculiarities rather than reflecting aggregated group data. This thesis, in particular, focuses on the phenomenon of unreliable user feedback (human uncertainty) in the context of recommendation and personalisation.

At the beginning, the concept of measurement uncertainty is used to render human uncertainty measurable and to analyse its impact on the comparative assessment of predictive systems. The findings reveal a difficulty to distinguish between systems regarding a given accuracy metric. Furthermore, human uncertainty is shown to induce an offset on such metrics, which limits the detection of improvements. This furnishes the need for a mathematical model of unreliable decision-making, feasible test procedures for significant system distinction, and a user model to plausibly explain the present phenomenon. To this end, concepts of statistics will be combined with those of metrology and theoretical neuroscience. Using human uncertainty as an example, it is illustrated how systems with empathy for the human nature can be designed.

The knowledge gained in this dissertation comprises a technical and an epistemological component. On the one hand, a specific characteristic of human decision-making is investigated and its origin is discussed against the background of a possible model of cognition. On the other hand, a mathematical framework is developed to analyse and implement this phenomenon for future systems of predictive data mining. This possibly paves the way for a new perspective within a currently prominent research direction.

Contents

List of Figures

List of Tables

1 | Introduction

The purpose of this chapter is to introduce the topic of this thesis, to motivate its research questions and to demonstrate its relevance for related fields of research. Some parts of this chapter are mainly based on my work Jasberg and Sizov (2019). In particular, Sec. 1.2, 1.3 and 1.4 have been published almost verbatim there. The motivating example has also been published in my contributions Jasberg and Sizov (2017b), Jasberg and Sizov (2017c), and Jasberg and Sizov (2018a). However, all these sections underwent small content-related modifications such as the underlying storyline or the addition of another research goal (i.e. research goal D).

1.1 Preliminary Settings and Terminology

This thesis' contribution is related to the field of predictive data mining which is a sub-field of applied computer science. Predictive data mining can be defined as the "search for very strong patterns in big data that can generalize to accurate future decisions." (Weiss and Indurkhya, 1998, p. 1). In other words, predictive data mining is about searching the right data within records that is suitable for predicting future events by generalising from the past. Such algorithms are employed in a multitude of technologies nowadays, e.g. fraud detection, (online) marketing, healthcare outcomes, and investment analysis (cf. Weiss and Indurkhya, 1998, p. 7).

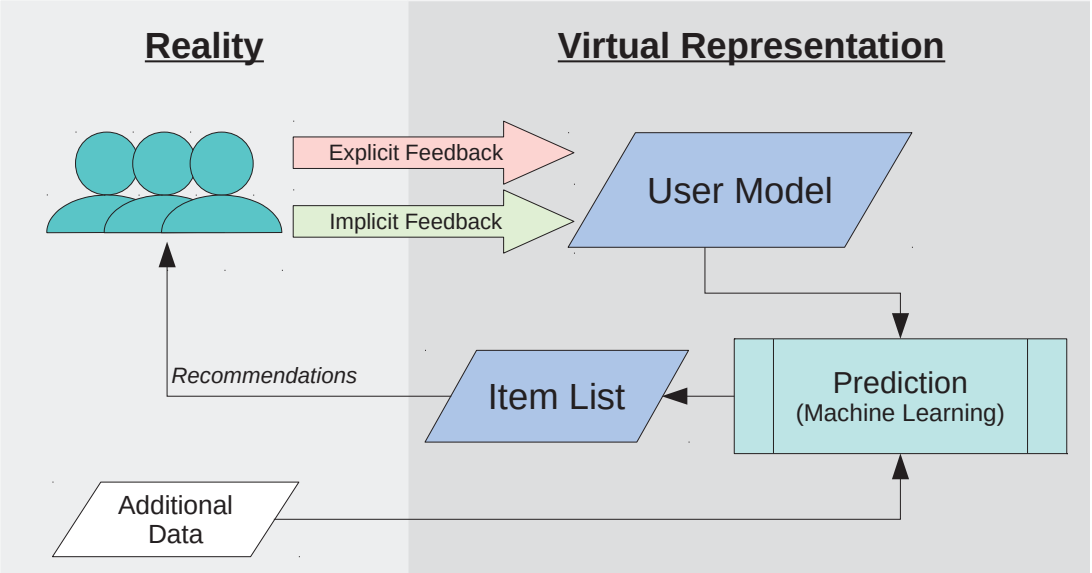


Figure 1.1: Schematic operation principle of recommender systems

A prominent exponent of predictive data mining is the branch of so-called recommender systems. These are "software tools and techniques providing suggestions for items to be of use to a user" (Ricci et al., 2010, p. 1). The wide range of possible applications for these systems comprise entertainment, content personalisation, e-commerce and service which can be regarded as the most common use cases (cf. Ricci et al., 2010, p. 14). Utilisation in entertainment thereby comprises recommendations for films or music while content personalisation comprises recommendations for documents, web pages, news, and e-mail filters (cf. Ricci et al., 2010, p. 14). E-commerce makes use of automated recommendations for purchasable products such as electronic devices, books, clothes and many more while service recommendations are often related to travel services, the suggestion of appropriate consultant experts or matchmaking services (cf. Ricci et al., 2010, p. 14).

The basic operation principle of a recommender system is illustrated in Fig. 1.1. Each recommender system is dependent on user interactions for its proper working and its adaptation to individual preferences and behaviour. Such user interactions are commonly divided into either explicit feedback or implicit feedback (cf. Ricci et al., 2010, p. 2). Whilst explicit feedback is entered directly by a user himself, implicit feedback is inferred from interactions with an online interface (cf. Ricci et al., 2010, p. 2). Explicit feedback is usually collected by requiring a user's personal opinion using

a scale (e.g. numeric or binary) in an online questionnaire (cf. Ricci et al., 2010, p. 9). Typical examples of implicit feedback are user actions such as displaying additional information on items, adding items to a wish list, or entering search queries (cf. Ricci et al., 2010, pp. 9–10). All this information is then stored into the so-called user model. Therefore, a user model can be defined as the entirety of stored user data that encodes a user's preferences and needs (cf. Ricci et al., 2010, pp. 8–9). Quite often it may just be "a simple list containing the ratings provided by the user for some items" (Ricci et al., 2010, p. 8). This user model can be augmented by additional data such as inferred "mood, weather as well as other people's votes" (Yang et al., 2012, p. 1) to rule out biases. Jannach et al. point out that "although the existence of a user model is central to every recommender system, the way in which this information is acquired and exploited depends on the particular recommendation technique" (Jannach et al., 2010, p. 2). In other words, most recommendations arise from intricate algorithms of (modern) machine learning that are capable of learning a person's preferences based on manifold database entries concerning past user behaviour. These recognised patterns are then transferred to new and unseen items to obtain an estimate of a user's degree of preference in advance. These techniques usually provide for each user an individual list of items that are sorted by the inferred item preference. The top n of this list is recommended to a user whose reaction then provides additional feedback to refine the underlying user model and hence to improve further recommendations.

According to the automated evaluation of contemporary research contributions conducted by Enríquez et al. (2019), current efforts of system improvement strongly focus solely on the optimisation of machine learning techniques. In doing so, the "majority of the published empirical evaluations of recommender systems [...] has focused on the evaluation of a recommender system's accuracy" (Herlocker et al., 2004, p. 19), i.e. the degree of matching between prediction and real user feedback. The rise of accuracy-driven evaluations has also led to criticism which is comprehensively described in Ch. 2. In parallel, a smaller branch of research focused on the user model's data quality. Koren and Sill describe that "different users tend to have different internal scales" (Koren and Sill, 2011, p. 117) and therefore, the validity of so-gathered information is questionable. The same holds true for implicit feedback as well since "mapping the user actions into a numerical scale [...] would be somewhat arbitrary" (Koren and Sill, 2011, p. 117). However, the validity concerns regarding user models and related problems (described in more detail in Ch. 2) had no discernible effect on the scientific mainstream

as the results by Enríquez et al. (2019) demonstrated. This thesis is supposed to describe further investigations related to the data quality of user models, particularly on reliability (instead of validity or objectiveness) and its impact on the frequently pursued accuracy optimisation. All these investigations will be carried out in the light of recommender systems, but the results of this thesis are likely to apply for the entire field of predictive data mining because explicit (and implicit) human-generated data is more or less ubiquitous for other use cases as well. For the rest of this thesis, the common terminology of recommender systems research will be used as introduced, for example, in Ricci et al. (2010); Jannach et al. (2010); Herlocker et al. (2004).

1.2 Motivation and Choice of Topic

The quality of data is an important topic for empirical sciences and initial research has already been done for predictive data mining as well. In contrast to the research of Koren and Sill (2011) elaborating on questions of validity, this thesis is dedicated to reliability concerns. Reliability in terms of recommender systems means that user feedback is not constant and therefore not absolutely credible but subject to a certain degree of uncertainty. This section is supposed to motivate the relevance of this phenomenon for the field of recommender systems. For example, the question arises whether a deviation from predicted feedback is due to inadequate system operation or due to lacking reliability, meaning that the system is actually working quite well. Having this question in mind, one can start to question existing efforts to determine values of accuracy and rankings that are built upon these comparative evaluations.

As a motivating example, it is referred to an initial pilot study conducted by Sizov (2017b) to investigate unreliable user feedback. As a motivating example, an initial pilot study will be presented which was conducted by Sizov (2017b) to investigate unreliable user feedback. In this study, 110 participants were shown 220 photos of well-known places and famous attractions (e.g. the Leaning Tower of Pisa) which were supposed to be rated on the usual 5-star scale. However, one photo was not shown once but five times in total and has thus been rated repeatedly. The main result of Sizov (2017b) is that a virtual community of users can be described in total (globally), but the same system is then again unable to describe individual user behaviour (locally):

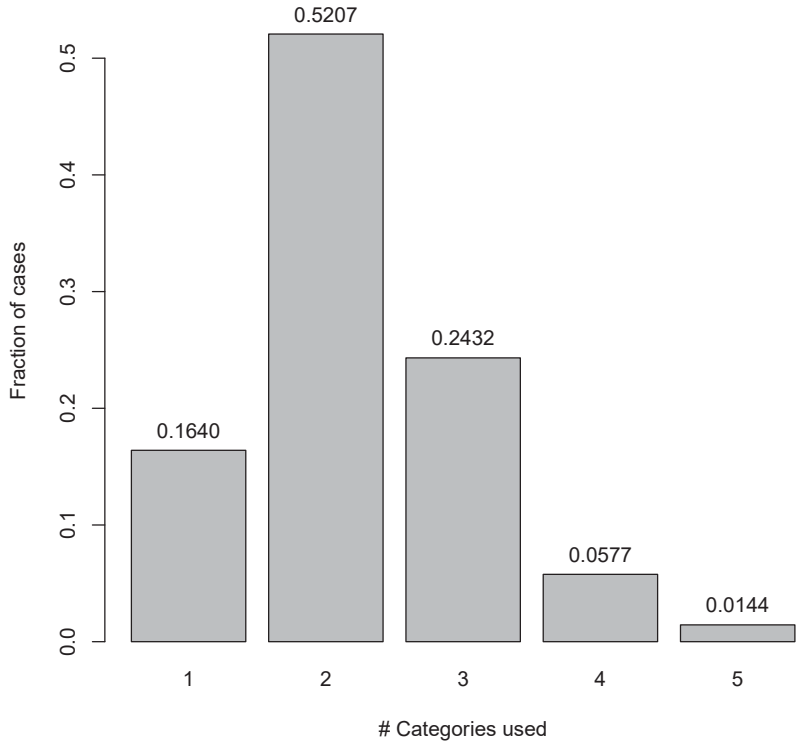


Figure 1.2: Pilot study from Sizov (2017b): Repeated ratings of photos with $N = 550$ participants reveal the existence and extent of unreliable feedback for the same picture.

> "In other words, the considerable fraction of users exhibits some (unfitting) behaviour that contradicts the [global] model. [...] Consequently, it would not be correct to claim that such a model provides a reasonable 'explanation' of the individual user behaviour in the population." (Sizov, 2017b, p. 870)

This result is substantial for two reasons: (1) It implicitly questions accuracy-driven evaluations in which prediction quality is usually measured for an entire community at once and (2) it demands a new perspective of system design in which users are to be understood individually (i.e. local explanation) rather than within a virtual community (i.e. global explanation). These results implicitly support the research of McNee et al. who find that those "recommendations that are most accurate according to the standard metrics are sometimes not the recommendations that are most useful to

users" (McNee et al., 2006, p. 1097). Therefore, further research about data quality in user models – and reliability in particular – is mandatory as it challenges contemporary methods of research and might lead to proposals for future system design to ensure that recommender systems will fulfil their actual purpose even better than before.

In contrast to former contributions related to this phenomenon (cf. Ch. 2), Sizov (2017b) was the first to describe and illustrate the lack of reliability explicitly and in a discrete way. From Fig. 1.2 it can be seen that only 16% of all participants had used the same response category to rate the photography in all five rating repetitions. This basically means that only four of twenty-five users give constant ratings. Just over 50%, on the other hand, utilised two different response categories for five ratings, so they changed their mind once. Moreover, 25% - a quarter - of all users have even used three categories and thus changed their opinion twice. Against this background, the argument that users are sufficiently understood when a system is highly accurate may need to be reconsidered. For example, let the predictor for a particular user-item pair be $\pi = 2$ stars given on the usual 5-star scale. Let this user rate the corresponding item five times, assigning the ratings $\{4, 2, 2, 4, 4\}$ stars. Such a scenario of utilising two response categories is very likely according to Fig. 1.2. Can the predictor $\pi = 2$ stars be considered to be false, or can it be considered to be correct, or is it $3/5 = 60\%$ false and $2/5 = 40\%$ correct? More importantly, how to quantify the difference between this predictor and the real rating? And finally, how to calculate the overall prediction quality of a whole system under these conditions?

To elaborate on these questions, a separate experiment was planned and conducted. In this experiment, 67 users were required to rate video trailers on a 5-star scale. Five of these trailers have been presented five times in total with certain temporal gaps and other rating tasks in between. A comprehensive description and detailed analysis can be found in Ch. 3. The first idea for possible effects of lacking reliability on the global prediction accuracy is obtained by a simple simulative analysis. To this end, three sample recommender systems are defined by determining their predictors:

$$\text{RS 1} \quad \pi^1_{u,i} := \quad \text{mean of all ratings for the u-i-pair} \tag{1.1}$$

$$\text{RS 2} \quad \pi^2_{u,i} := \quad \text{3 const.} \tag{1.2}$$

$$\text{RS 3} \quad \pi^3_{u,i} := \quad \text{first given rating for the u-i-pair} \tag{1.3}$$

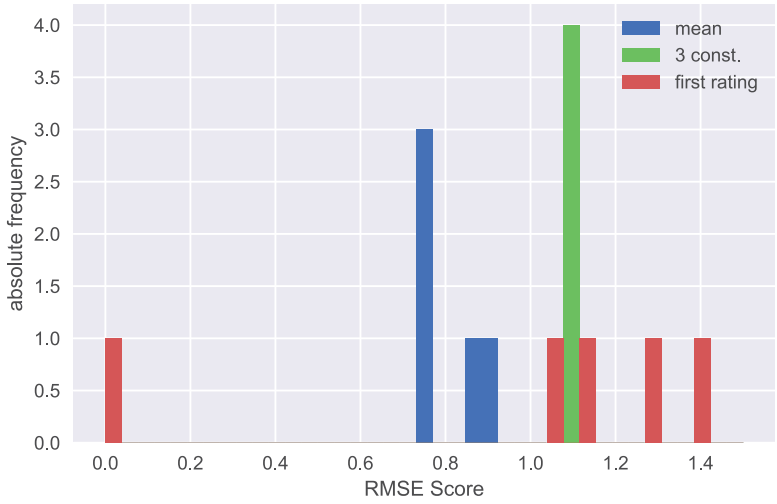
The rationale behind these definitions is as follows: The first recommender system is a statistical one and accounts for each rating trial. Moreover, it will be revealed in

forthcoming chapters that this is also the best recommender possible when a special accuracy metric is considered. The second recommender system does not account for any given rating at all and provides a constant prediction to each user-item pair instead. The choice for constantly predicting three stars (instead of any other possible star rating) is motivated by the work of Sizov (2017a). The author proves $\pi = 3$ to be a good choice for a constant recommendation as it "shows better performance than at least one User-User comparison" (Sizov, 2017a, p. 892), i.e. a constant three-star prediction outperforms a setting in which users serve as recommenders for themselves. The third recommender system is an example of such a "User-User comparison" as mentioned by Sizov (2017a). The predictor is defined as the first user rating to a specific item since this reflects the current reality of rating scenarios, i.e. a user rates each item only once and without further re-evaluations. For each of these recommender systems $k = 1, 2, 3$ all $N = 335$ user-item pairs are used to compute the root mean squared error

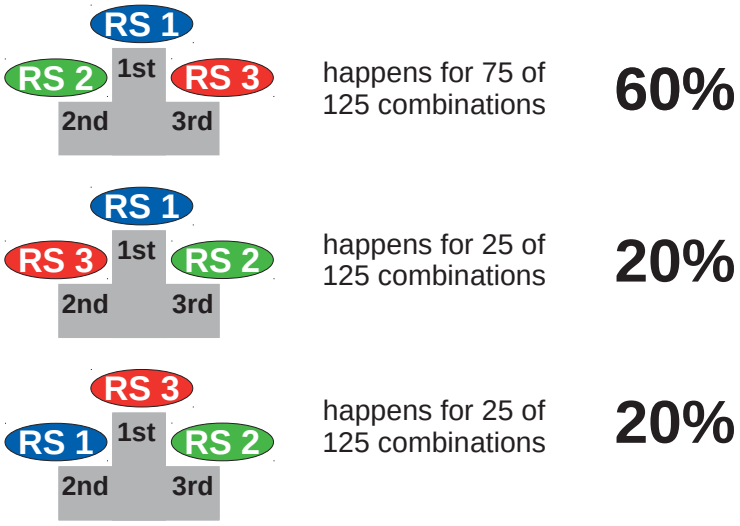
$$\mathrm{RMSE}(k, t) := \sqrt{\frac{1}{N} \sum_{u,i} \left(r^t_{u,i} - \pi^k_{u,i} \right)^2} \tag{1.4}$$

as a commonly used accuracy metric for each rating trial $t = 1, \ldots, 5$ where $r^t_{u,i}$ denotes the t-th rating from user u given to item i. This results in five possible RMSE outcomes for each recommender system. A histogram of these scores is shown in Fig. 1.3a. It can be recognised that those RMSE scores scatter significantly which is most evident for RS 3 since it can produce both, the best or even the worst prediction accuracy. Having five scores for three recommender systems, one yields $5^3 = 125$ possible combinations to form a ranking of these systems. In Fig. 1.3b the factual rankings are shown together with their absolute and relative frequencies. At this point, it is clear that the uncertainty of explicit user feedback propagates when computing accuracy metrics. Accordingly, taking the lack of reliability into account, there is no longer an absolutely valid ranking, but each possible ranking is subject to a particular probability for being drawn in a single trial. This inevitably changes the perspective one should have on explicit user feedback and its interpretation. As seen in this first example, reliability concerns - which are always inherent in explicit user feedback - tackle the statistical soundness of system evaluation. This alone motivates the effort of a better understanding of this propagation process.

In the following sections, other issues will also be highlighted that need to be considered when analysing unreliable user feedback, e.g. biasing effects on accuracy



(a) histogram of RMSE scores for each system and each repetition trial



(b) possible rankings based upon all RMSE score combinations

Figure 1.3: Impact of human uncertainty on the reliability of prediction accuracy

metrics as well as natural offsets which are limiting the detection of prediction quality for well-working systems. All these consequences could potentially support misjudgements in comparative evaluations. These can hardly be solved with big data, meaning the collection of even more explicit user feedback, since new data will also come along with missing reliability. In this light, what are the essential implications for real use cases?

1. Someone could opt for a supposedly better system, whose superiority is just due to uncertainty that comes along with reliability issues.

2. Someone is investing financial and human resources in the further improvement of a system, although a statistically sound detection of improvements is not possible anymore.

3. The magnitude of improvement is misjudged because the uncertainty has a different bias for small metric scores than for larger scores. On this basis, competitions in which a reference system must be outperformed by a certain threshold may possibly need further consideration.

4. Problems are ignored due to the assumption that soliciting explicit user feedback to obtain big data would solve these issues, which is only partially true.

These phenomena justify a sound investigation of lacking reliability in explicit user feedback and motivate a well-developed theory of comparative assessment when reliability concerns are involved. For the rest of this dissertation, the phenomenon of lacking reliability of user feedback will be denoted as **human uncertainty**. This term reflects, on the one hand, the concept that is employed to operationalise reliability, namely the concept of measurement uncertainty as it is introduced by JCGM (2008a) and thoroughly described in Ch. 2. On the other hand, it also describes the presumed human origin of this phenomenon and hence reflects that all findings are likely to apply whenever explicit user feedback is considered within the field of predictive data mining.

1.3 Research Objectives

The essence of this thesis is to add a new dimension to given structures in the field of predictive data mining and to provide indications that digital footprints (i.e. the information that human beings inevitably provide by using modern technology) should

not always be considered as absolutely credible. Consequently, current tendencies in predictive data mining, i.e. aiming for the highest accuracy and collecting more and more explicit user feedback (big data) can only lead to real innovations when it comes along with a well-considered analysis of its inherent uncertainty. For this purpose, subgoals and research questions (RQ) are defined which will be worked off gradually in this thesis:

Subgoal A: Revealing that uncertainty actually exists in a concrete rating scenario with items that are closer to reality than photos.

RQ A1: How can uncertainty in explicit user feedback be modelled?

RQ A2: How can uncertainty information be effectively measured?

Subgoal B: Demonstrating what happens if uncertainty is not considered, especially when determining the prediction accuracy of multiple systems.

RQ B1: What could a model for the propagation of uncertain user feedback look like when computing accuracy metrics?

RQ B2: What characterises the propagation of uncertainty, i.e. are there any amplifying or weakening effects or confounding variables?

RQ B3: What impact does uncertainty propagation have on the comparative evaluation of prediction accuracies?

Subgoal C: Convincingly introducing methodological basics and analysis approaches for dealing with uncertainty in order to turn this phenomenon to good account.

RQ C1: What are current solutions and what are they able to enhance?

RQ C2: How can human uncertainty be used to create recommendation benefits?

By pursuing the previously mentioned subgoals, it became apparent that human uncertainty is indeed an additional dimension of system design and evaluation that should not be neglected. For this reason, a thorough knowledge of human uncertainty is certainly relevant to evaluate these impacts beforehand. In other words, it is most imperative to predict human uncertainty for future events and a multitude of circumstantial dependencies (e.g. fatigue, stress, mood, etc.). Predicting human uncertainty for each user-item pair might support further (strategic) decisions, e.g. to sort a list of possible

recommendations by the best predictor along with the least uncertainty, to only present recommendations whose anticipated uncertainty does not exceed a certain threshold, or to explicitly present these items of high uncertainty in order to evoke surprise in some users. The multitude of additional opportunities constitutes the need for a new user model that is capable of representing human uncertainty as well as passing the relevant information on to common machine learning techniques. Moreover, each event of a random process as well as the extent of scattering itself has an informative value. In theory, this information might also serve to improve recommendation rather than just raising doubts. For this purpose, the decision was made to employ a cognitive model from the field of theoretical neuroscience. The rationale behind this decision is that human uncertainty is likely to originate from a cognitive process and can, therefore, best be described by a corresponding theory. Such a theory also holds the advantage that it might provide additional information about the human cognition process which can be used for recommendation. Moreover, a neuroscientific user model could be expanded more easily by many influencing factors such as fatigue and stress. Neurological concepts in conventionalised form have already proven to be fruitful in the past, e.g. (recurrent) artificial neural networks, deep learning (cf. Savage, 2019, p. 16) or hierarchical temporal memory (HTM) models (cf. Ahmad et al., 2017, p. 1).

During the research of subgoal A, certain doubts arose as to whether the applied measurement procedures indeed measure an existing phenomenon or simply produce it themselves. In other words, the validity of both measurement procedures introduced in Ch. 3 is not fully evident. On the one hand, the repeated rating of items confronts the user with a continuously changing environment as the presentation of preceding items is always different. This could result in different situational biases distorting an otherwise constant user rating. The second measurement approach, on the other hand, requires a user's belief about the adequateness of each possible user rating. This could represent a leading question for it suggests that different weightings of simultaneous response categories are mandatory. At first, this does not seem to be a problem after all. Even if human uncertainty was induced by exposing a user to contextual changes, one has still to clarify which of these contexts is the right one for a proper evaluation of recommendation quality. The uncertainty about this individual context and its corresponding bias then again justifies the utilisation of a probability density. This perspective implies that the entire feedback distribution can be reinterpreted in a way that different situations lead to different biases and that the evaluation of recommendation quality is carried

out across any of these possible contexts. The problem about this reasoning is the underlying assumption that user feedback consists of a true and constant opinion plus a constant but context-dependent bias. This assumption entails that by capturing a user's context, one can straighten out the corresponding bias and obtain a true and constant opinion. Considering any other context than this would be obsolete because the true opinion is already known. One simply has to find the dependencies between context and related bias. This issue is currently tackled by a variety of research as described in more detail in Yang et al. (2012).

The essential core of this dissertation thesis is, however, that user feedback has nothing like a true value but is (to a certain degree) random by nature. Although there is currently not enough evidence to reject one of these opposing hypotheses in favour of the other, the neurological user model of Ch. 6 provides several indications to support the probabilistic assumption instead of a true constant value plus bias. Therefore, the development of a neurological user model not only results in novel ideas of system design but also serves to provide initial indications for uncertainty to be a natural human characteristic. This in return supports the hypothesis of human uncertainty to be present in most databases containing implicit or explicit user feedback. In the light of all these points, it appears necessary to define another subgoal:

Subgoal D: Discovering a neurological user model to predict uncertainty and to substantiate the possibility for a human-inherent origin.

> **RQ D1:** What does a possible cognition model look like that naturally explains human uncertainty?
>
> **RQ D2:** To what extent can such a model be considered as biologically plausible?
>
> **RQ D3:** What are the benefits of this user model in particular and what are the benefits of this novel paradigm in general?

In brief, subgoal D provides additional support to the previous subgoals as it demonstrates that human uncertainty is very likely to be existent in any database containing implicit or explicit user feedback. Furthermore, this subgoal can be used to derive proposals on how systems may be designed in the near future so that they explain individual user behaviour even better than before.

1.4 Outline of this Thesis

The goals of this thesis are closely interwoven with its structure and a supplementary illustration can be found in Fig. 1.4. In the last sections of this introduction, the interdisciplinary nature of this thesis will be described along with its applied methodology and possible contributions to certain fields of study.

The related work is actually split into two distinct parts in order to improve readability. The first part in Ch. 2 is concerned with measuring and modelling uncertainty and how it should be handled. The second part in Ch. 6.2 focuses on (computational) neuroscience and its potentials to support a new perspective of system design. Both parts will be preceded to those chapters introducing the relevant research conducted in this dissertation project.

The scientific and technical main core of this work refers to subgoal A to C whilst each subgoal has its chapter dedicated. In Ch. 3, the subgoal A will be addressed and evidence will be given to prove the existence of uncertainty in a concrete rating scenario with realistic items. To this end, a stochastic model of human uncertainty will be defined which is based on the theory of measurement error from metrology and physics. This model is chosen because it already constitutes a well-established conjunction of lacking reliability and the uncertainty about a quantity of interest which is operationalised and represented by probability distributions. Moreover, this model provides some explicit hints to possible approaches for the measurement of uncertainty which will be further elaborated through a comparative analysis.

After substantiating the existence of human uncertainty by using these measurement techniques, the impact of uncertain data on prediction accuracy will be elaborated in Ch. 4. For this purpose, an additional concept from physics and metrology will be employed which has been developed to analyse the propagation of uncertain data through a computational model. This concept will be applied to a common prediction accuracy metric in order to derive viable estimations and to investigate the dependence on confounding variables. After reporting on potential issues induced by human uncertainty, possible solutions will be discussed in Ch. 5. Here, new proposals will also be made to obtain benefits from uncertain user feedback.

After all these elaborations, it makes sense to examine the origin of human uncertainty more closely and to take a closer look at the potential cognitive process in order to propose a user model for uncertainty representation. This research, as it is presented

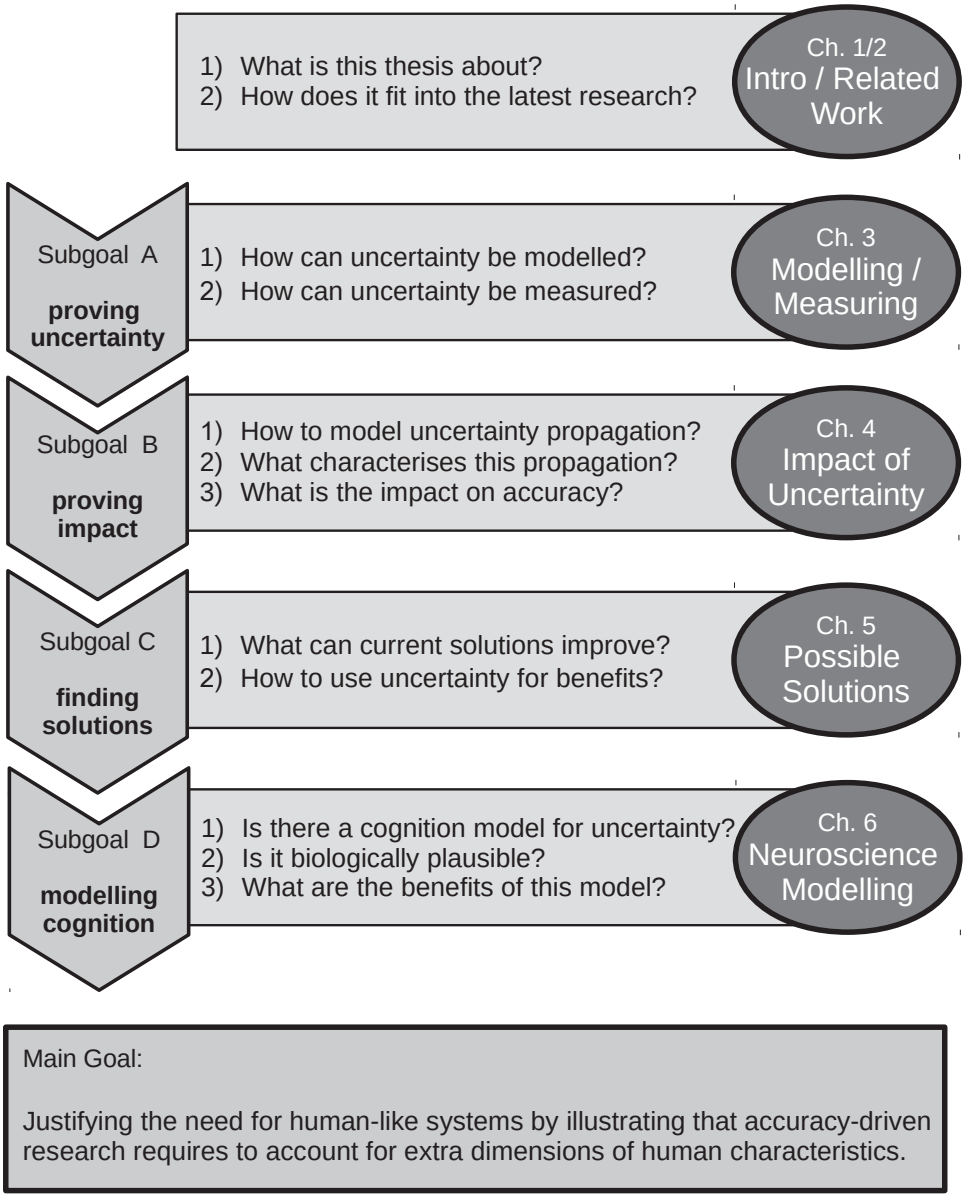


Figure 1.4: Graphical representation of the organisation of this thesis

in Ch. 6, should not be understood as a second technical core of this thesis but rather as (1) a proposal for new system design concepts and (2) a very first initial research substantiating previous findings. For this reason, the associated analyses have not been divided into distinct chapters in order to symbolise its subordinate position within the big picture. This initial research idea involves another related work in which the utilised model of behavioural variability is motivated. Afterwards, the basics and the operating principle of this model will be clarified and potential connections to uncertain user feedback will be drawn. The technical implementation of this model and the realisation of its learning phase (model fitting) will constitute a large part of this chapter. Finally, it will be examined whether this cognition model (fitted on real human data) leads to implications which are in concordance to biological or medical literature. The conclusion of this chapter is formed by an analysis of a possible connection between this cognitive model and the field of predictive data mining.

In Ch. 7 the results of this thesis will be recapitulated and all subgoals will be brought together in a concluding discussion. Although this thesis focuses solely on human uncertainty as one example of human peculiarities, the results will serve to draw conclusions for the interpretation of user data in general. In particular, conclusions are drawn with respect to the demand of Sizov (2017b) for systems that explain and describe users as individuals with all their human characteristics. What such systems might look like and how they are integrated into standard concepts of machine learning can be deduced from the neurological analysis. The final chapter of this thesis will elaborate on the scientific impact of this dissertation. Finally, additional and novel ideas will be raised to describe further research.

1.5 Fields of Study, Methods, and Contributions

Despite its interdisciplinary nature, this thesis is supposed to be understood mainly as a contribution to information science. The interdisciplinarity, in particular, is constituted by the transfer of methods from other fields of research that are not naturally inherent to information science. Those are in particular:

applied computer science: The example of recommender systems was borrowed from this field of research as a representative for predictive data mining. This also applies to the current methodology for the (comparative) assessment of such

systems via accuracy metrics as revealed by Enríquez et al. (2019) and thoroughly described by Herlocker et al. (2004). Many evaluations within this area rely on user experiments. Especially, the measurement of human uncertainty via repeated ratings (Ch. 3) has been carried out before by Amatriain et al. (2009b). To fully account for human uncertainty and to make this information available to recommender systems, a new user model is designed in Ch. 6. User model design is also part of recommender system research as previously described. In this thesis, suitable user models will be assessed in Ch. 6 in terms of A/B testing as it has been advocated in the industry for years (cf. Kohavi and Thomke, 2017). For this purpose, various machine learning techniques, which are currently studied in applied computer science (cf. Enríquez et al., 2019, p. 12), will be fitted with these user models and their performance will then be evaluated.

mathematics and statistics: Mathematics (and statistics in particular) serves as an ancillary science. From this wide field, descriptive methods, as well as inductive methods, are employed. The last ones include hypothesis tests (i.e. KS-test, Welch's t-test, Levene's test) to compare differently obtained distributions (Ch. 3). Those distributions rely on discrete measurement data and sufficient statistics for distributions from the parametric family are hence estimated using the maximum likelihood method. The impact of human uncertainty on metrics of recommendation accuracy will be addressed in Ch. 4 by using convolutions of probability density functions. Respective approximations will be yielded using the method of Taylor expansion as described in Ku (1966).

metrology and physics: Metrology is the science of accurate measurement whose methods are often applied in physics (and engineering as well). According to the Joint Committee for Guides in Metrology, there are basically two types of evaluation (cf. JCGM, 2008a, p. 7) which are thoroughly introduced in Ch. 2. Both types will be used in Ch. 3 to develop instruments of measuring human uncertainty in an online rating scenario. One key element of this dissertation is the representation of lacking response reliability (human uncertainty) through probability densities. This key idea also stems from metrology concepts (cf. JCGM, 2008a, p. 6). The explicit choice of distributions is guided by the maximum entropy principle as proposed by metrology (cf. JCGM, 2008c, pp. 18–20). For the case considered in this thesis, this principle points to the utilisation of Gaussians as

they serve to represent the discrete drawings whilst assuming the least unknown additional information. Using Gaussians also allows to simplify the convolutions regarded in Ch. 4 by using the method of Gaussian Error Propagation (cf. Ku, 1966, pp. 265–267). To further simplify mathematical derivations of Ch. 4 and to enable investigations in the absence of analytical solutions respectively (Ch. 6), Monte-Carlo simulations will be employed. These simulations are also proposed by the Joint Committee for Guides in Metrology (cf. JCGM, 2008c, pp. 27–32).

computational neuroscience: The core of computational neuroscience can be described as "the application of mathematics and systems theory to the analysis of neural systems and, reversely, the application of neural procedures to the solution of technical problems" (Mallot, 2013, p. v). The first part is done in Ch. 6 by connecting the theory of probabilistic population codes from the field of neuroscience to the phenomenon of human uncertainty from the field of recommender systems. The theory of probabilistic population codes is suitable for this kind of connection as it assumes behaviour to be represented by probability densities just as it is discovered for user feedback in Ch. 3. This connection is additionally made plausible by an analysis of the biological implications of such a fusion. The second part, i.e. the application of neural[1] procedures to the representation of human uncertainty is done by mapping user behaviour to neuronal[2] features that in return constitute a user model for recommendation purposes.

Findings revealed by applying these above-mentioned methods will be interpreted mainly in the light of information science. Possible conclusions for other fields of study are only briefly mentioned within the main body of this thesis, but they are described in detail in terms of further research (Ch. 7). This dissertation makes a contribution to the field of information science by

1. revealing insights about the limited credibility of knowledge that is gained about people on the basis of their interactions with digital systems,

2. evaluating current developments in the field of predictive data mining on a meta-level, especially the predominant role of accuracy-driven research and the subordinated role of user models and their representation of human characteristics,

[1] The term 'neural' refers to characteristics of the (central) nervous system in general.
[2] The term 'neuronal' refers to characteristics of a particular a neuron.

3. proposing a new perspective of future system design with the goal to further sensitise predictive data mining to human beings, i.e. to better explain individual user behaviour and to stronger account for human characteristics.

The third point suggests that this work is part of a larger series of research which is following a special credo: Systems of predictive data mining should seek to further understand human beings through their individual psyche, the neurological foundations of behaviour, emotional states, and life circumstances. The rationale behind this claim is given by the fact that "appropriate interpretation of collective feedback requires the development of suitable models that summarize and 'explain' observations" (Sizov, 2017b, p. 869). It is exactly this 'explaining' which, according to Sizov (2017b), has moved away from truly understanding human beings and therefore needs to be redefined. This criticism is expressed qua

> "Model components and parameters are often interpreted as an 'explanation' of observations. From this perspective, users form groups with homogeneous behaviour, individuals of each group are characterized by the corresponding distribution [...]. Unfortunately, it is technically possible to construct models that are simple, fit well with summary (macro-level) aggregated data, but do not appropriately fit the individual (micro-level) behaviour of individual users." (Sizov, 2017b, p. 869)

This criticism alludes to the results of Enríquez et al. (2019) which show that contemporary research is accuracy-driven as well as mainly technical, meaning that the majority of contributions address the issue of solely improving machine learning techniques. In doing so, the explanation of individual users becomes nothing more than finding optimal weights of a machine learning model which is tuned to community behaviour – the individual moves into the background and can no longer be explained. This conclusion has been substantiated by hypothesis tests revealing that a "considerable fraction of users exhibits some (unfitting) behaviour that contradicts the [tuned] model" (Sizov, 2017b, p. 870). In other words: Current efforts in predictive data mining to mainly optimising machine learning techniques in an accuracy-driven fashion may possibly not possess enough explicative power to understand individual human beings. One possible solution might be the following:

> "Our work was motivated by the intention to make statistical models more 'human-like', i.e. better describing individual human behaviour [...].

> Experimental evaluations have shown that our methodology allows for constructing user behaviour explanations that go beyond established mixture models." (Sizov, 2017b, pp. 875–876)

This demand for human-like systems has motivated the focus of this dissertation. This thesis shows how such systems can be designed, namely by reinterpreting the knowledge about users against human characteristics (here: missing reliability of user feedback) and by transferring neural theories of cognition and mind into the user model itself. However, the majority of publications within the field of predictive data mining and recommendation is presenting evidence that their systems work well for (very) large amounts of users. In the course of this thesis, it is exactly this label of "well-working" that will be reviewed and questioned against the phenomenon of human uncertainty. In this sense, this debate is not simply about two scientific streams with different definitions of 'explaining users'. Rather, possible shortcomings and chances for misinterpretation will be revealed for the accuracy-driven recommender research insofar individual human behaviour is not considered appropriately. Of course, human beings are much more multifaceted and it can not be claimed that the mere consideration of human uncertainty leads to absolutely human systems, nor that it makes the call for more human-like systems obsolete. Rather, this dissertation can be seen as a starting point for a new paradigm amid established methods.

2 | Related Work

The purpose of this chapter is to introduce the different branches of research which are related to this thesis. In doing so, a preliminary description of research fundamentals will be given together with the respective terminology. This will lay the foundation to describe the current state-of-the-art within each research branch which will finally allow for the exact positioning of this thesis' contribution, i.e. commonalities, differences and amendments. This chapter has not yet been published in its current form.

2.1 Predictive Data Mining and Recommendation

This thesis is mainly related to the field of predictive data mining, a sub-field of applied computer science. Weiss and Indurkhya define (predictive) data mining as follows:

> "Data Mining is the search for valuable information in large volumes of data. It is a cooperative effort of humans and computers. Humans design databases, describe problems and set goals. Computers sift through data, looking for patterns that match these goals. Predictive data mining is a search for very strong patterns in big data that can generalize to accurate future decisions." (Weiss and Indurkhya, 1998, p. 1)

In easy terms, data mining is the process of searching the right data to solve a predefined problem or to fulfil a predefined goal. According to Weiss and Indurkhya, those problems

and goals respectively can be addressed by either knowledge discovery (decision-support) or prediction (decision-making) (cf. Weiss and Indurkhya, 1998, p. 7). It is the latter type of data mining that aims to gather the right information to predict future outcomes and is referred to as predictive data mining. Typical applications of predictive data mining are fraud detection, (online) marketing, healthcare outcomes, and investment analysis (cf. Weiss and Indurkhya, 1998, p. 7). To fulfil its goal, algorithms of predictive data mining aim to solve related problems "by looking at past experience with known answers, and then projecting to new cases" (Weiss and Indurkhya, 1998, pp. 7–8).

One special application of predictive data mining is the development of so-called recommender systems. Ricci et al. define these systems as follows:

> "Recommender Systems [...] are software tools and techniques providing suggestions for items to be of use to a user [...]. The suggestions relate to various decision-making processes, such as what items to buy, what music to listen to, or what online news to read." (Ricci et al., 2010, p. 1)

Examples of possible items that are usually recommended to users are news, web pages, books, CDs, movies, electronic devices, but also insurance policies, financial investments, travels and jobs (cf. Ricci et al., 2010, p. 8). In general, such systems perform data mining across data of past human behaviour that is reflecting personal preferences. According to Jannach et al., the choice of data as well as its representation is denoted as the underlying user model (cf. Jannach et al., 2010, p. 1).

The author additionally points out that "although the existence of a user model is central to every recommender system, the way in which this information is acquired and exploited depends on the particular recommendation technique" (Jannach et al., 2010, p. 2). In other words, there is a variety of recommender systems that do not only differ by the algorithm itself but also by the underlying user model, i.e. the data collection on which the system performs its learning task. Such a user model can be "a simple list containing the ratings provided by the user for some items" (Ricci et al., 2010, p. 8) and/or "sociodemographic attributes such as age, gender, profession, and education" (Ricci et al., 2010, p. 8). Another degree of freedom that has already been mentioned is the method of augmenting the data collection according to the chosen user model:

> "[Recommender systems] collect from users their preferences, which are either explicitly expressed, e.g., as ratings for products, or are inferred

> by interpreting user actions. For instance, a [recommender system] may consider the navigation to a particular product page as an implicit sign of preference for the items shown on that page." (Ricci et al., 2010, p. 2)

According to this description, explicit feedback about an item is defined as being directly entered by the user himself. This is typically done by transferring the personal opinion onto a scale utilised in an online questionnaire (cf. Ricci et al., 2010, p. 9). Two popular scales are the numerical scale (e.g. '1' to '5' stars as used by Amazon or the Google Playstore) and the binary scale (e.g. 'like' or 'dislike' as used on Tinder). Implicit user feedback consists of preference information that is inferred by a user's action, e.g. displaying additional information to an item, adding items to a wish list, or entering search queries (cf. Ricci et al., 2010, pp. 9–10).

As mentioned before, recommender systems do also differ by the employed algorithm. Qualitatively, there are a few ideas of algorithm design. Two design concepts that can be found in standard textbooks are the content-based filtering and the collaborative filtering respectively. Both methods can be described as follows:

> "At its core, content-based recommendation is based on the availability of (manually created or automatically extracted) item descriptions and a profile that assigns importance to these characteristics. If we think [...] of [a] bookstore [...], the possible characteristics of books might include the genre, the specific topic, or the author." (Jannach et al., 2010, p. 4)

> "The basic idea of [collaborative filtering] is that if users shared the same interests in the past [...] they will also have similar tastes in the future, so, if, for example, user A and user B have a purchase history that overlaps strongly and user A has recently bought a book that B has not yet seen, the basic rationale is to propose this book also to B." (Jannach et al., 2010, pp. 2–3)

According to Ricci et al., collaborative filtering can be seen to be the most popular concept for recommender system design (cf. Ricci et al., 2010, p. 12). The research about recommender systems is yet still prospering due to their relevance for the economy, e.g. increasing sales number, selling more diverse items, increasing user satisfaction and user fidelity (cf. Ricci et al., 2010, p. 5). In order to describe the current state-of-the-art,

Enríquez et al. analysed 1195 papers resulting in 80 primary studies. The main results in terms of technical novelty are

> "that the most studied technique in recommendation systems is recommendation with the use of collaborative filters, closely followed by those that use content-based filters. Only 14 used hybrid recommendation systems, whereas 31 used collaborative filtering and 29 used content-based methods." (Enríquez et al., 2019, p. 14)

Furthermore, "the most researched [innovations] correspond to naive Bayes, SVM vectors, and neuronal [sic!] networks, representing almost 55% of the techniques used for this purpose" (Enríquez et al., 2019, p. 14). This study thus indicates that contemporary research is focusing on traditional algorithm designs and seeks to further improve techniques from the 1960s. The authors further reveal:

> "Although many of the proposals present a validation, few of them use real data sources instead of synthetic ones (artificially generated rather than generated by real-world events) to carry out their experiments. In this sense, a lack of technology transfer of these proposals to real case studies has been detected." (Enríquez et al., 2019, p. 14)

> "[Current research projects] respond to the design and implementation phase but are far from offering solutions in earlier stages such as requirements and analysis. This makes it very difficult to find efficient and effective solutions that support real business needs from an early stage." (Enríquez et al., 2019, p. 14)

> "Finally, we can accomplish that even having executed this rigorous study, there is still a big difficulty in deciding about which algorithm is better than another depending on the context in which it is used." (Enríquez et al., 2019, p. 15)

In short terms, actual efforts strongly focus on technical considerations, optimising well-known algorithms within artificial environments. On the other side, the current state-of-the-art is lacking an explicit elaboration for real-world applicability, which the authors acknowledge as a potential for future work.

This potential is covered by this thesis for its contribution questions possible issues related to the utilisation of real human data. In particular, special attention is given to data quality and its impact on predictive algorithms and related systems. Questioning the quality of data is indeed a topic of interest and not entirely new within this branch of research. For instance, although numerical scales are frequently used to collect explicit feedback, the validity of so-gathered information is questionable:

> "When user feedback is related to absolute numbers, taking the scores as numerical may not reflect the user intentions well. Different users tend to have different internal scales. For example, taking star ratings as numeric will put the same distance between '3 stars' and both '4 stars' and '2 stars'. However, one user can take '3 stars' as similar to '4 stars', while another user strongly relates '3 stars' to low quality, being similar to '1-2 stars'." (Koren and Sill, 2011, p. 117)

But also for implicit feedback, such validity questions can be raised since user actions are quite often internally encoded by numerical scales as well:

> "For example, a user can search and browse a product page, which is a weak indication of interest in the product. A stronger indication would be bookmarking the product or adding it to a 'wish list'. An even stronger indication would be entering the product to the 'shopping cart' or bidding on the product. The strongest indication would be actually purchasing the product. [...] Yet, mapping the user actions into a numerical scale would not be natural or trivial. Any decision to map the actions into a numerical scale, e.g. coding 'search and browse' as a 1, bookmarking or wish-listing as 2, etc., would be somewhat arbitrary." (Koren and Sill, 2011, p. 117)

It can be seen that the debate about the effects of real human data has been held in 2011 and that it had no major effect on the current state-of-the-art according to Enríquez et al. (2019). It can thus be assumed that such issues might not be perceived as relevant by a certain majority or that such ideas do not propagate rigorously enough to demonstrate a long life period. Since recommender systems are intended to be applied in real-world scenarios, they are supposed to deal with realistic data created by human beings. For this reason, the question of the effects of human-generated data and especially its quality criteria is quite relevant.

From predictive data mining, the example of recommender systems is adopted along with the comparative assessment through accuracy metrics. However, this thesis will not focus on the techniques of machine learning themselves but on the underlying user model and its data quality. While Koren and Sill (2011) focus on the validity, i.e. whether the data indeed reflects the user's true intention, it is the additional quality criterion of reliability that is in the scope of this present thesis. For this purpose, the first part of this thesis (Ch. 3 to 5) is dedicated to measuring reliability and discussing its impact on the comparative assessment of recommender systems. The second part of this thesis (Ch. 6) is dedicated to the development of a new user model that operationalises reliability and serves to deduce proposals for future system design. This will be exemplified by using the above-mentioned collaborative filtering approaches as this is the most common technique for recommendation (cf. Ricci et al., 2010, p. 12). The results of this thesis are exemplified with recommender systems but also serve as indications to question similar issues for the broader field of predictive data mining.

2.2 Validation Methods in Predictive Data Mining

As described above, research on predictive data mining – and recommender systems in particular – mainly relies on the optimisation of existing algorithms of machine learning. A comprehensive introduction of existing approaches is given in Bishop (2006) as well as in Manning et al. (2008) and Kubat (2015). A comprehensive guide to system validation and comparative assessment in the field of predictive data mining can be found in Herlocker et al. (2004) and Bobadilla et al. (2013). According to Herlocker et al., the task of identifying the best predictive algorithm is rather difficult and a variety of related metrics have been developed for this purpose (cf. Herlocker et al., 2004, p. 6). By 2004, however, the "majority of the published empirical evaluations of recommender systems [...] has focused on the evaluation of a recommender system's accuracy" (Herlocker et al., 2004, p. 19). This practise has been maintained to this day so that accuracy metrics are still considered as "the most commonly used metrics by recommender systems" (Rawat and Dwivedi, 2019, p. 17). This reported supremacy is also in line with other recent publications (cf. Heinrich et al., 2019, p. 2). In general,

> "an accuracy metric empirically measures how close a [...] predicted ranking of items for a user differs from the user's true ranking of preference.

> Accuracy measures may also measure how well a system can predict an exact rating value for a specific item." (Herlocker et al., 2004, p. 19)

Typical measures of accuracy, according to Herlocker et al., are precision and recall, the mean average error (MAE), the mean squared error (MSE) and the root mean squared error (RMSE) (cf. Herlocker et al., 2004, pp. 20–23). Especially the RMSE plays a prominent role after it was used in the Netflix Prize in 2009 (cf. Netflix Inc., nd). It will therefore be employed as the accuracy metric for the considerations in Ch. 4.

Although accuracy has become the most frequently used attribute for recommendation quality, it only evaluates a small fraction of applicable perspectives. Based on the above-given definition, accuracy does not measure other important attributes like diversity, i.e. to what extent a user gets offered songs from other artists rather than from one and the same. Accuracy does also not measure a user's surprise about a novel and unknown item, nor does it measure a user's satisfaction (which might be present even if a rating does not match its model-based prediction). This is the reason for some sort of scepticism that has built against the exclusive use of accuracy metrics. For example, McNee et al. (2006) claims that

> "the recommendations that are most accurate according to the standard metrics are sometimes not the recommendations that are most useful to users." (McNee et al., 2006, p. 1097)

The author's main criticism about accuracy-driven evaluation is the range of consequences of not considering all the above-mentioned attributes that accuracy does not account for. Neglection of diversity can, for example, lead to blindness towards so-called similarity holes, i.e. "only giving exceptionally similar recommendations (e.g. once a user rated one Star Trek movie she would only receive recommendations for more Star Trek movies)" (McNee et al., 2006, p. 1098). To tackle this issue, Ziegler et al. (2005) introduced an intra-list similarity metric that is sensitive to topic diversification. Users were given a list of recommendations which could be altered in the extent of item diversity. The results demonstrated, that the users preferred lists with increased diversity despite their worse accuracy compared to less diverse lists (cf. Ziegler et al., 2005, pp. 28–30). Another often neglected attribute is serendipity, i.e. "the experience of receiving an unexpected and fortuitous item recommendation" (McNee et al., 2006, p. 1099). The authors argue that users often prefer "recommendations that are for items they would not have thought of themselves" (McNee et al., 2006, p. 1099). The

last attribute that has already been mentioned is user satisfaction. It has been shown that user satisfaction does not always correlate with high recommender accuracy (cf. Ziegler et al., 2005, pp. 28–30) and that explaining user satisfaction cannot be reduced to accuracy considerations (cf. Knijnenburg et al., 2012, p. 443). This is so far only the criticism related to what accuracy metrics do not account for. In addition, the informative value of accuracy itself is questionable:

> "How large a difference does there have to be in the value of a metric for a statistically significant difference to exist? Complete answers to these questions have not yet been substantially addressed in the published literature." (Herlocker et al., 2004, p. 19)

Here, statistical significance indicates the authors' belief in reliability issues.

With this thesis, the argumentation presented above is extended by the additional aspect of human response reliability. In particular, we follow Herlocker's idea and thoroughly investigate the impact of lacking reliability on accuracy metrics in Ch. 4 and provide instruments for distinguishing systems with statistical significance. The contribution of this thesis thus provides further evidence against the isolated use of accuracy metrics. However, while other authors investigate shortcomings of neglecting further dimensions of validation (i.e. similarity holes, serendipity, satisfaction), this thesis questions the credibility of accuracy itself due to lacking reliability of user-generated data. Although the specific methodology introduced in this thesis is exemplified by using the RMSE, the main results can easily be adapted for alternative assessment metrics without substantial loss of generality, insofar they require for (uncertain) human input. This has been demonstrated by Zhang et al. (2018), who applied the reliability analysis of Sec. 4.1 to the accuracy metric MAE and obtained similar results.

2.3 Human Uncertainty in Recommender Systems

The first reported discovery of reliability issues for user feedback was done by Hill et al. (1995) through an experiment to evaluate the power of collaborative filtering. A total of 291 users participated through an automated e-mail interface and each of them was given a list with 500 films to rate on a numeric 1-to-10-scale (Hill et al., 1995, pp. 197,

199). In order to estimate the limitations of the recommendation results, the authors considered data reliability:

> "Six weeks after they initially [participated], 100 early users were asked to re-rate exactly the same list of movie titles as they had rated the first time. 22 volunteers replied with a second set of ratings." (Hill et al., 1995, p. 199)

The authors found that the Pearson correlation between these two rating trials was 0.83 (cf. Hill et al., 1995, p. 199) and concluded that

> "since a person's rating is noisy (i.e., has a random component in addition to their more underlying true feeling about the movie), it will never be possible to predict their rating perfectly." (Hill et al., 1995, p. 200)

This supports the hypothesis from above that also the reliability of data constituting the underlying user model has an impact on a recommender system. On this foundation, Xavier Amatriain initiated research in this direction and coined the term "natural noise" (Amatriain et al., 2009b, p. 173). This term comprises the same phenomenon of unreliable user ratings as already introduced, but it further reflects the concept that user ratings consist of a true feeling plus an additional random component (i.e. superimposed noise). The authors operationalised the phenomenon by using a psychometric test-retest-reliability in the form of a noise-to-signal-ratio (cf. Amatriain et al., 2009a, p. 249). The experimental setup scheduled three distinct rating trials in which 118 users had to provide ratings for 100 movie titles from the Netflix Prize on a 1-to-5-star scale using a web interface (cf. Amatriain et al., 2009a, pp. 249–250). The second rating trial has been carried out 24 hours after the first trial and the third rating trial has been carried out 15 days after the second trial (cf. Amatriain et al., 2009a, pp. 249–250). The noise-to-signal ratio has been found to be 0.924 (cf. Amatriain et al., 2009a, p. 252) and the author confirmed that "any value over 0.9 is usually considered 'good' in classical test theory" (Amatriain et al., 2009a, p. 252). This is in concordance with the observation of Hill et al. (1995) and reveals that existent deviations are relatively small. These small deviations nonetheless hold great potential to negatively affect recommendation: Using a sample recommender system, "the calculated RMSE between different trials ranged between 0.557 and 0.8156 [sic!]" (Amatriain et al., 2009a, p. 257), depending on the specific rating trial. The authors interpreted the lower bound as the empirical

minimum for this specific recommender system which is then impaired by natural noise (cf. Amatriain et al., 2009a, p. 257). To solve this issue, Amatriain et al. (2009b) introduced a de-noising algorithm which replaces repeated ratings with large deviation by artificial ratings with smaller deviation. The authors achieved an improvement above 14% compared to the original RMSE (cf. Amatriain et al., 2009b, p. 180). Another strategy of coping with user noise is proposed by Koren and Sill (2011). The authors associate model-based predictions with artificial noise so that they better resemble noisy feedback. This strategy has been implemented in a recommender system that performed better than its reference on standard data records. A more detailed analysis of these proposed solutions can be found in Ch. 5.

One remarkable interpretation of Amatriain et al. is that the lowest value gained from considering the RMSE for each rating trial constitutes an empirical lower bound for the respective recommender system (cf. Amatriain et al., 2009a, p. 253). The idea for an absolute limit of prediction accuracy was also initially discussed by Hill et al. (1995) but has not been in the main scope of the respective paper. The assumption of such a limitation has later been mentioned once again:

> "Though the new algorithms often appear to do better than the older algorithms they are compared to, we find that when each algorithm is tuned to its optimum, they all produce similar measures of quality. We – and others – have speculated that we may be reaching some 'magic barrier' where natural variability may prevent us from getting much more accurate." (Herlocker et al., 2004, p. 6)

The essence of this magic barrier seems to be that the comparative evaluation of best-performing recommender systems becomes obsolete when reaching a certain value of accuracy, i.e. there is only an equivalence class of optimal systems left where no sensible rankings can be performed. Said et al. assumed that every improvement beyond this barrier might indicate over-fitting rather than better performance (cf. Said et al., 2012, p. 238). This reflects the importance of being able to estimate this magic barrier. A theoretical framework for this purpose has been introduced by Said et al. (2012) using an empirical risk minimisation principle. In doing so, Said et al. found that the magic barrier – in the case of the RMSE – constitutes as the square root of the average variance in user responses (cf. Said et al., 2012, p. 243). The peculiarity of this work is that it was the first time that a theoretical framework had been developed to model natural

noise (i.e. unreliable user responses) in recommender systems whilst other authors like Amatriain et al. (2009a) kept on working phenomenologically.

In further progress, the authors exploited this magic barrier to enhance recommendation in terms of accuracy (Said and Bellogín, 2018). Their idea is to compute a user's coherency within an attribute space (e.g. a film genre). For example, a coherent user would rate all horror films equally whilst an incoherent user would demonstrate more fluctuations throughout this genre. In the course of this, Said and Bellogín (2018) conducted an online experiment with 308 users providing ratings to 2329 items in two distinct rating trials. It was found

> "that user coherence is correlated with the magic barrier; we exploit this correlation to discriminate between easy users (those with a lower magic barrier) and difficult ones (those with a higher magic barrier)." (Said and Bellogín, 2018, p. 97)

> "[that] this experiment confirms that it is possible [...] to build different training (and test) models in such a way that the error decreases for the easy users, i.e., to increase the accuracy of the recommender system." (Said and Bellogín, 2018, p. 117)

Indeed, the authors were able to show that the RMSE can be improved by 10 to 40%, depending on the number of difficult users in the training set (cf. Said and Bellogín, 2018, p. 121). This result points to a practical advantage as it allows for "cheaper recommendation cycles (in terms of computational effort, time, and parameter tuning) for easier users" (Said and Bellogín, 2018, pp. 120–121), which might be a large fraction within usual data records according to the reliability measurements from above.

In this thesis, we also address the missing reliability of user responses and its impact on recommender systems. We also address the measurement of reliability as well as its numerical representation by appropriate metrics. One of the major differences between this thesis and preceding research is the perspective of doing so. Hill et al. described the underlying phenomenon by a quantity of psychometrics and test theory (cf. Hill et al., 1995, p. 199) while Amatriain et al. used a noise-to-signal ratio from electrical communication engineering (cf. Amatriain et al., 2009a, p. 249). This thesis will describe yet another access to this phenomenon which is based on the theory

of measurement uncertainty as proposed by metrology and physics. This concept is thoroughly introduced in the next section. Its essence is to regard measurable quantities such as user feedback as a distribution which comprises a central tendency together with its uncertainty. One advantage of this approach is the potential to investigate reliability issues for each user-item pair rather than the aggregation of hundreds of uncertain ratings as it was done in Hill et al. (1995) and Amatriain et al. (2009a). This in-depth information might be key to a deeper understanding of this phenomenon.

To separate this novel perspective terminologically from the other perspectives (i.e. psychometrics and electrical communication theory) it will henceforth be referred to as human uncertainty. This term reflects both, the application of the concept of measurement uncertainty as well as the assumption about a human origin of this phenomenon. The phenomenon itself – namely the missing reliability in user feedback and the assumption of an inherent random component – is exactly the same throughout all of these three different terminologies. Simply put, the terminological variation does not refer to the phenomenon itself but to the perspective that is applied to it.

The second difference to preceding research is the measurement approach. Hill et al. (1995) collected two ratings per item with a temporal gap of six weeks while Amatriain et al. (2009a) collected three ratings with a temporal gap of 24 hours for the second and additional 15 days for the third rating trial. Although the measurement described in Ch. 3 also relies on repeated ratings, the temporal gap between each trial is significantly shorter while the number of trials is higher. The rationale behind this course of action is to reduce the chance of substantial changes in a user's external conditions (time of day, weather, etc.) or a user's internal states (emotions such as joy or anger, fatigue, stress, health). Otherwise one could always argue that uncertainty is just another manifestation of environmental biases. In this light, it has been found that "ratings are always irrational, because they may be affected by many unpredictable factors like mood, weather and other people's votes" (Yang et al., 2012, p. 1). Detecting human uncertainty in a setting of constant internal and external conditions would prove that user feedback indeed contains a random component rather than just different (constant) biases corresponding to environmental changes. This in turn would support an important hypothesis, namely that missing reliability is indeed a human characteristic which is always present whenever a particular rating is done. A pointer to this hypothesis is given by a second measurement procedure (cf. Ch. 3) which collects all information about a feedback distribution in one single rating trial only.

The third difference can be found within the analysis of accuracy exemplified by the RMSE. Amatriain et al. further examined possible impacts of missing reliability on recommender accuracy and found that

> "although the reliability of the survey as an instrument and the stability of user opinions are high, inconsistencies negatively impact the quality of the predictions that would be given by a [recommender system]." (Amatriain et al., 2009a, p. 257)

The authors substantiate this claim by the fact that the RMSE values changed for each of the three rating trials. This conclusion is based on the assumption that different ratings within the learning set influence the underlying machine learning algorithm. This in return would lead to better or worse recommendation quality. From a metrologist point of view, the scattering of RMSE values do not necessarily indicate an impact on the machine learning technique but rather indicate the propagation of uncertain user ratings while computing the RMSE. The explicit difference can be understood by considering the comparison function within the RMSE formula, i.e. $\Delta_\nu := r_\nu - \pi_\nu$ where $r_\nu \in \mathbb{R}$ is the rating for user-item pair $\nu = (u, i)$ and $\pi_\nu \in \mathbb{R}$ is the model-based prediction. In this notation, ν is a multi-index that aggregates the user id u and item id i for the purpose of abbreviation. The conclusion of uncertainty affecting model-based predictions would imply that π_ν is a random variable with realisations in $\mathbb{R}$. However, the formation of this random variable and its related density function on the basis of uncertain user feedback is still unknown. By contrast, metrology naturally regards the measured input quantity as being uncertain. Hence, r_ν (instead of π_ν) is considered to be a random variable whose density can be obtained by measurement. This also leads to an uncertainty propagation with regard to Δ_ν and the RMSE. This propagation is elaborated in Ch. 4. The legitimation of the metrologist perspective is given by our initial experiment in which recommender systems have been built by simply defining constant predictors. Accordingly, those predictors have not been generated by machine learning on the basis of different rating trials. Nonetheless, the computation of RMSE scores for each rating trial resulted in scattering as revealed by Amatriain et al. (2009a). In simple terms, the main difference is that preceding research assumed a negative impact on recommendation quality itself while the metrologist perspective just assumes an uncertainty with which recommendation quality can be detected. Certainly, both perspectives have their specific raison d'être and this thesis does not aim to falsify the

other point of view. This thesis rather aims to introduce this novel perspective and to discover a phenomenon (along with its impact) in a different light than before.

Using the metrologist approach, the uncertainty propagation for the RMSE can explicitly be derived in a closed mathematical formulation as to see in Sec. 4.1. The RMSE hence takes the form of a probability density itself and has to be seen as a generalisation of those findings reported by Amatriain et al. (2009a). This finally allows to answer an important question concerning the comparative assessment given two different recommender systems:

> "How large a difference does there have to be in the value of a metric for a statistically significant difference to exist? Complete answers to these questions have not yet been substantially addressed in the published literature." (Herlocker et al., 2004, p. 19)

When both metrics are represented by probability densities, these might overlap for small differences between their respective location parameters. As a result, each ranking in terms of order relation is subject to a certain degree of error that can be expressed by probability. This error probability of comparative assessments is derived and exemplified by the revaluation of the Netflix Price competition in Sec. 4.3.

This metrologic concept is also capable of deriving the magic barrier exactly as it is introduced in Said et al. (2012). Moreover, its derivation in Sec. 4.4 proves to be a generalisation of the preceding work: Said et al. (2012) introduced the magic barrier as a sharp boundary while it is demonstrated in Sec. 4.4 that the magic barrier must also be represented by a probability density. This basically means that the magic barrier might be lower or higher than initially presumed and that its exceeding can only be expressed in terms of probability.

2.4 Uncertainty Concepts in Metrology

The Bureau International des Poids et Mesures (engl. International Bureau of Weights and Measures) defines metrology as "the science of measurement, embracing both experimental and theoretical determinations at any level of uncertainty in any field of science, and technology" (BIPM, 2004). Traditionally, the concepts of metrology are applied in physics, engineering, astronomy, and other experimental sciences. However, they might apply for predictive data mining a well. The collection of user feedback

can certainly be understood as a "process of experimentally obtaining one or more [...] values that can reasonably be attributed to a quantity" (JCGM, 2008b, p. 16) which is exactly the BIPM definition of measurement for which these particular concepts have been designed for. Moreover, the credibility issues addressed by metrology have already been found to be present in underlying user models as well:

measurement accuracy according to the BIPM denotes the "closeness of agreement between a measured quantity value and a true quantity value of a measurand" (JCGM, 2008b, p. 21). This quality criterion matches the discovery of so-called biases as introduced above and neatly summarised by Yang et al. (2012). Such a bias, in the light of recommender systems, denotes a systematic shift of a user's true opinion triggered by a particular environmental factor, e.g. weather bias, social bias and many more.

measurement precision according to the BIPM denotes the "closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions" (JCGM, 2008b, p. 22). This quality criterion matches the discovery of missing reliability of user responses (also called natural noise or human uncertainty) as initially discovered by Hill et al. (1995) and described in the previous section.

This correspondence between the descriptions of data quality in two different research areas indicates the potential for a straightforward transfer of concepts from one area to the other. One concept that can be adapted is the operationalisation of uncertainty which, in general, can be described as a "non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used" (JCGM, 2008b, p. 25). In particular, the BIPM operationalises (measurement) uncertainty as follows:

> "Uncertainty of measurement comprises, in general, many components. Some of these components may be evaluated from the statistical distribution of the results of series of measurements and can be characterized by experimental standard deviations [also known as type-A evaluation]. The other components, which also can be characterized by standard deviations, are evaluated from assumed probability distributions based on experience or other information [also known as type-B evaluation]." (JCGM, 2008a, p. 2)

The representation of human uncertainty by standard deviations for each user-item pair is a difference to the research approaches used by Hill et al. (1995) and Amatriain et al. (2009a). Moreover, the metrologist perspective does not consider a true value plus uncertainty separately but holistically combined in a probability density (cf. JCGM, 2008a, p. 6). This reflects the idea that the true value remains unknown and can only be located within intervals to a certain degree of probability. For a measurand, there are two different ways of gathering the necessary information to construct a probability density. The type-A evaluation is based on the frequentist definition of probability, i.e. that the probability of an event equals the relative frequency of its occurrence for an infinite number of observations (cf. Schurz, 2015, p. 3). The repeated rating of films as applied in Hill et al. (1995) and Amatriain et al. (2009a), for example, can be considered as type-A evaluation. In contrast to this, the type-B evaluation relies on the Bayesian definition of probability, i.e. probability is the personal degree of belief for an event to occur (cf. Schurz, 2015, p. 3). Such evaluation might be needed to account for different types of measuring devices (or to develop novel instruments of collecting uncertainty information as in our case).

An important concern within the field of metrology is the so-called propagation of uncertainty. This becomes relevant as many quantities can not explicitly be measured and must be calculated on the basis of other quantities. In this light, JCGM (2008c) defines propagation as a

> "method used to determine the probability distribution for an output quantity from the probability distributions assigned to the input quantities on which the output quantity depends." (JCGM, 2008c, p. 5)

From an era in which computational capacity was still very limited, a method has been established to approximate the uncertainty propagation. This method assumes that all uncertain quantities are independent and normally distributed while the output's standard deviation is then calculated by using a Taylor-series where terms of higher order are omitted. This well-known approach is also called the Gaussian Error Propagation. Comprehensive introductions and derivations can be found in Ku (1966) and Taylor (1997). While being easy to compute, this approach has the limitation of not being able to account for type-B evaluation completely: Measured quantities are nowadays assigned an arbitrary probability density that is not necessarily a Gaussian. Quantities calculated therefrom are then assigned a distribution by means of a convolution of all

their argument densities. This model is comprehensively described in JCGM (2008a). Moreover, a computational extension via Monte-Carlo simulations can be found in JCGM (2008c). Monte-Carlo simulation in this sense is regarded as any "method for the propagation of distributions by performing random sampling from probability distributions" (JCGM, 2008c, p. 5).

In this thesis, the missing reliability of user responses is considered as the uncertainty of a user rating (measurand). The metrologist approach is adopted and this uncertainty is operationalised as the standard deviation of a respective probability density. To some extent, this model is similar to those already used by Said et al. (2012) with the difference that these authors considered ratings and standard deviation separately whilst the whole probability density is employed in this thesis. In order to obtain the uncertainty information related to user feedback, both evaluations (type-A and type-B) are used. The type-A evaluation dictates a repeated rating task as it has already been used by other authors before (cf. previous section). In contrast, the type-B evaluation is the foundation for a novel measuring instrument introduced in Ch. 3. Its essence is that users are required to provide their personal belief about the adequacy of each possible rating option on a Likert scale. The advantage of this novel instrument over the type-A based rating is that the entire probability density is surveyed in just one single rating trial. This information is then used to construct feedback distributions for each user-item pair. According to JCGM (2008c) and the maximum entropy principle that is introduced there, the parametric model of choice is a Gaussian (cf. JCGM, 2008c, pp. 18–20). The rationale behind this data model is its minimisation of hidden assumptions given available data. However, by using the Monte-Carlo simulation, one is not restricted to Gaussians as a variety of parametric and even non-parametric distributions can be used as well.

In this thesis there is also an answer given to the Herlockers question, i.e. how large does a potential accuracy improvement has to be in order to be statistically significant (cf. Herlocker et al., 2004, p. 19). In doing so, the uncertainty propagation is performed for the RMSE in Ch. 4 and by analysing the overlapping of two distributions, an error probability for the most likely ranking order will be derived. This forms the basis for a hypothesis test providing information about the statistical significance of improvements. The uncertainty propagation will be carried out using both methods, the Monte-Carlo simulation as well as the Gaussian Error Propagation. The Monte-Carlo simulation

guarantees freedom of choice for the distribution models while the Gaussian Error Propagation provides fast solutions in big data scenarios.

2.5 Uncertainty Concepts in Computational Neuroscience

For two particular reasons, there is a need to seek for biological theories that might explain the present phenomenon: (1) Having found user feedback to be distributed does not necessarily lead to the conclusion that decision-making yields a random component. Another possibility might be that prior information such as film trailers seen before (user history) leads to different constant biases. This implies that user feedback is indeed reliable and any kind of uncertainty can be resolved by simply knowing the user's surroundings together with its corresponding bias. (2) Since nature often provides solutions for technical problems, e.g. the idea of neural networks, it might be fruitful to develop novel approaches for predictive data mining that rely on a biological basis. A biologically deduced user model may represent uncertainty in recommender systems so that uncertainty can be predicted as well. The following paragraph is a brief overview of related biological contributions. A comprehensive discussion in the light of model selection is given in Sec. 6.2.

The assumption of underlying distributions for sensory perception (and decision-making) is the essence of the Bayesian brain hypothesis which is thoroughly described in Doya et al. (2007). The utilisation of probability densities is supported by experiments of cue integration: In Knill and Pouget (2004), auditory and visual stimuli have simultaneously been presented to test subjects. Given these stimuli, the subjects had to estimate the direction of an (imaginary) event that evoked both stimuli. The certainty of such stimuli has been altered by adjusting the volume (auditory certainty) and the contrast (visual certainty). The results show that the subjects' estimations were close to model-based estimations that originated from applying the Bayes rule to corresponding stimuli distributions. The authors concluded that human beings develop approximation strategies of Bayes-optimal integration and must hence possess an internal (neuronal) representation of probability densities as well (cf. Knill and Pouget, 2004, pp. 712–713). A possible explanation for the occurrence of internal probability distributions is given by Faisal et al. (2008). The authors describe random mechanisms that lead to so-called neuronal noise and they make this noise responsible for trial-to-trial variability which is considered "a prominent feature of behaviour" (Faisal et al., 2008,

p. 292). This explanation is picked up by Ma et al. (2006) and is extended to the theory of probabilistic population codes. This theory describes the genesis and representation of internal distributions which might be fruitful for the development of a user model.

In this thesis, the theory of probabilistic population codes is transformed into a user model that can be combined with a variety of machine learning techniques. In doing so, the theory of population codes is fully adopted as described by Ma et al. (2006). Missing specifications are discussed for those degrees of freedom that influence the neuronal activity and for possible aggregation functions that map neuronal activity onto probability distributions. Both specifications are determined through computer simulation. A novelty for this theory is its application to decision-making for it has so far only been studied in the light of sensory perception.

3 | Modelling and Measuring Uncertainty

The purpose of this chapter is to introduce the metrological model of uncertainty, to deduce two measurement approaches for this phenomenon, and to comparatively analyse their specific properties. These measurement approaches will be used for conducting an online user experiment whose data set will be relevant for the further analyses of this thesis. This chapter is mainly based on my work Jasberg and Sizov (2019) and was published almost verbatim there. Furthermore, the results of this chapter and especially the figures along with their interpretations were also used for Jasberg and Sizov (2017c) and Jasberg and Sizov (2018c). However, all these sections underwent a linguistic revision.

3.1 Uncertainty Models and Measurement Approaches

The concept of uncertainty is understood disparately in different scientific disciplines. In terms of human-computer interaction, human uncertainty is considered as a lack of reliability caused by the fact that people cannot reproduce their decisions during repeated interactions (cf. Amatriain et al., 2009a). In this sense, human uncertainty could constitute a characteristic feature of the cognitive process, which significantly influences the outcome and thus leads to condition-dependent and temporal instability.

This assumption is going to be evaluated in forthcoming chapters. In natural science and metrology, uncertainty is understood as a motivated doubt about the validity of a result. Uncertainty can therefore be seen as an additional dimension assigned to the measurement result, which denotes the scattering of values for a measured quantity (cf. JCGM, 2008a, p. 2). Both perspectives can be combined and thus the lack of reliability can be described by metrological metrics of scattering. The most important features of this modelling can be summed up briefly:

> "Thus a Type A standard uncertainty is obtained from a probability density function [...] derived from an observed frequency distribution [...], while a Type B standard uncertainty is obtained from an assumed probability density function based on the degree of belief that an event will occur [...]." (JCGM, 2008a, p. 7)

Uncertain quantities are therefore modelled as random variables X following a specific probability density f_X, whereby the distribution's shape and its width, in particular, is determined by uncertainty. This model fits well with the results of the RETRAIN study (as described later) and examples of assigned distributions can be seen in the histograms of repeated ratings as depicted in Fig. 3.6. From theoretical considerations, normal distributions prove to be the models of choice in this case: On the one hand, this is supported by theories of neuroscience in which the propagation of prior densities is carried out in a recurrent neural network. In this case, one necessarily yields a normal distribution as the posterior density due to the central limit theorem and the law of large numbers (cf. Ma et al., 2006, p. 1433). But also in metrology and physics, the normal distribution is proposed to be the best model for a given central tendency and scattering. In this case, the normal distribution maximises the information entropy (maximum entropy principle) (cf. JCGM, 2008c, p. 20), i.e. this distribution is the one that makes the least additional (unknown) assumptions. In this contribution as well, the normal distribution will be utilised. However, computations are not limited to this type and other distributions may be appropriate as well. At this point, the question arises of how to obtain the necessary information about the specific parameters of the underlying distribution. Metrology distinguishes between two different measurement methods based on different definitions of probability:

The Frequentist Approach: This way of deriving a user's feedback distribution is based on the frequentist definition of probability, i.e. the probability of an event

to occur is equal to its relative frequency for infinite trials. This procedure thus requires a repeated rating of the same products (re-rating) and has already proved itself in Amatriain et al. (2009b). Re-rating produces a rating tensor $(R_{u,i,t})$ where the coordinates (u, i, t) encode the rating that has been given to item i by user u in the t-th trial. From this record one can derive a unique rating distribution for a fixed user-item pair $\nu := (u, i)$ by considering tensor-slices in trial-dimension $R_{\nu,\bullet} := (R_{\nu,t})_{t=1,\dots,5} = (R_{\nu,1}, \dots, R_{\nu,5})$ and computing the maximum likelihood parameters for the chosen data model.

The Bayesian Approach: An alternative approach to accessing human uncertainty is based on the Bayesian definition of probability, i.e. the probability of an event to occur is the degree of one's personal confidence in this occurrence. Under this assumption, one can obtain a rating-distribution directly from requiring a user's personal belief that a particular rating score appears to be adequate for the corresponding item. This procedure is denoted as pdf-rating (**p**robability **d**ensity **f**unction). In mathematical terms: Having a 5-star scale $S = \{1, \dots, 5\}$, a user associates to each possible rating $s \in S$ a degree of personal confidence or belief c_s about the appropriateness of s concerning the item to be rated. The personal confidence is provided by a second scale $C = \{1, \dots, 5\}$. Hence, a pdf-rating

$$\mathfrak{p}_\nu = \{(1, c_1), (2, c_2), (3, c_3), (4, c_4), (5, c_5)\} \tag{3.1}$$

is given by a family of two-dimensional vectors in $S \times C$ where the values for ones personal belief are considered as specific weights for each of the associated ratings. In order to retrieve the feedback distribution, this rating is converted into its frequentist equivalent by use of the transformation

$$\tau \colon \mathfrak{p}_\nu \mapsto (\underbrace{1, \dots, 1}_{c_1\text{-times}}, \underbrace{2, \dots, 2}_{c_2\text{-times}}, \dots, \underbrace{5, \dots, 5}_{c_5\text{-times}}) \tag{3.2}$$

since the absolute histogram of this frequentist translation will exactly reproduce the data entered by the user, i.e. this transformation does not systematically bias the intention of a user's response. Subsequently, a maximum likelihood estimation is performed on $\tau(\mathfrak{p}_\nu)$ to find the optimal parameters for a chosen data model.

3.2 User Study: Repeated Trailer Rating (RETRAIN)

Since there are only a few datasets in which human uncertainty was collected, the RETRAIN (**Re**liability **Tra**iler Rat**in**g) study was conducted as an online experiment in which $N = 67$ participants had watched theatrical trailers of popular films and provided ratings according to the frequentist and the Bayesian approach. As far as it can be ascertained, this is the very first time that uncertainty in user responses has been gathered by Bayesian approaches. The goals of this experiment are

1. to prove unreliable user feedback in an explicit rating scenario,
2. to explore the viability of survey methods to measure this uncertainty,
3. to find indicative hints for the origin of this uncertainty, and
4. to gather uncertainty data to use for further explorations.

Goal (1) has already been "proven" by Hill et al. (1995), but the main criticism of the corresponding study is that the individual interviews have always been conducted with a temporal gap of several days in between. In this time, the situational contexts along with the cognitive and emotional states are very likely to change substantially. This challenges the credibility of the results. Moreover, human uncertainty would just emerge as the manifestation of a conglomerate of diverse situation biases and thus, this data record does not allow for unprejudiced answers to explore other possibilities for its origin. Complete control of the situational context would otherwise require a laboratory experiment, but this can often target only a certain group (mostly university students looking for experimental credits) rather than gathering a representative group with a realistic composition of people. Also, users who almost always do their ratings at home in front of their computer are certainly brought into a new and unfamiliar environment, which causes distorting effects such as an increase in concentration, which would normally not be present. For this reason, the decision was made against a laboratory experiment in favour of an online survey instead.

For the conducted RETRAIN study, the repetition trials had to be done in a short time sequence to minimise the probability of a change of situational contexts. This method, in turn, allows the criticism that the previous trailers could affect the rating of each subsequent trailer, so there might be a distorting Markov process. Unfortunately,

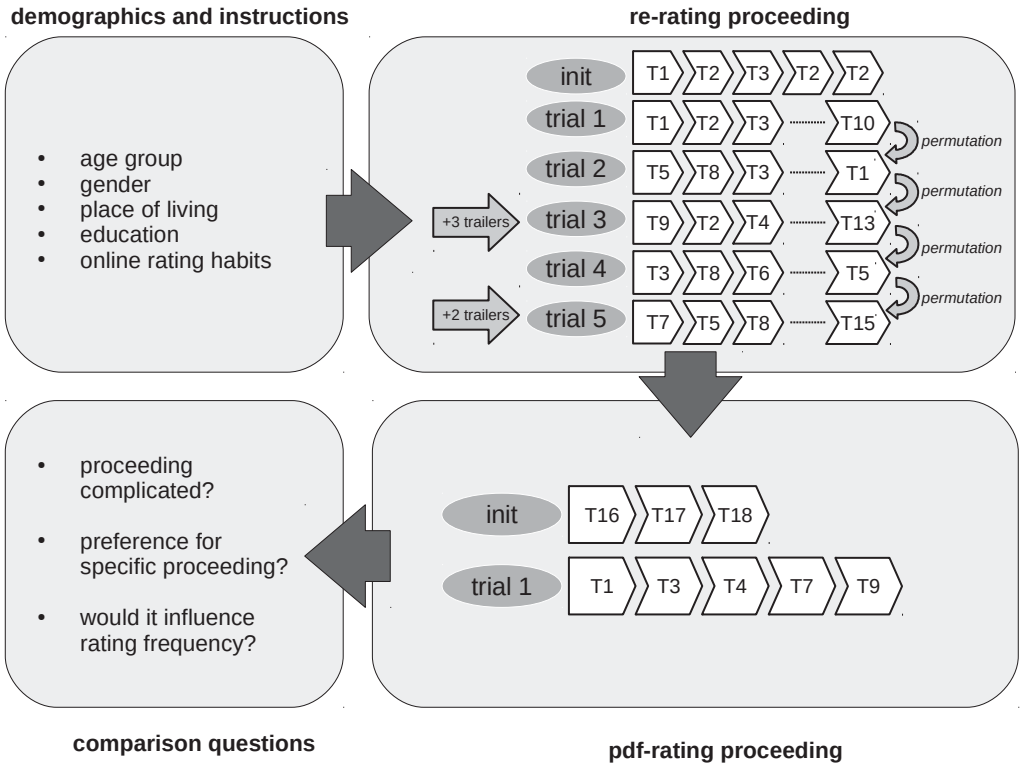


Figure 3.1: Conduction of the RETRAIN study

this can not be excluded, but there have been several rating trials in which the order of all trailers was randomised, i.e. each trailer always had a different "history" preventing one trailer to permanently have a different (i.e. larger or smaller) bias than another trailer. At this point, the origin of missing response reliability can not be clarified without a doubt. On the one hand, strong changes in the emotional and cognitive states can be excluded on average, since these variables usually do not change significantly within one hour (average duration of the experiment) as long as no secondary activities have been performed. On the other hand, even the randomised history of each trailer can be understood as a (small) contextual change which may be important. However, an argument will be introduced later that favours a natural and context-independent uncertainty rather than a genesis through contextual change. The RETRAIN study was set up with Unipark's[1] survey engine whilst the participants were engaged from the

[1] `http://www.unipark.com/de/` (last accessed on Mar 10, 2020)

crowdsourcing platform Clickworker[2]. All participants had to complete four phases:

1. demographics and instructions
2. re-rating proceeding
3. pdf-rating proceeding
4. comparison questions

The corresponding storyboard of this experiment can be found in Fig. 3.1. The different phases can be described as follows:

Demographic Data and Instructions: The pre-defined goals of this experiment require a generality which can only be assumed if the examined user group is as representative as possible for a specific cultural domain. This is verified by considering demographic information. Afterwards, the participants are instructed for the upcoming re-rating tasks using screen messages that ensure equal conditions in each experimental run. The participants are also informed that it might come to repetitions for technical reasons. This prevents manual termination ahead of schedule due to the assumption of technical errors. However, the true meaning of those repetitions remain unclear to the participants.

Re-Rating Proceeding: During this phase, the participants watch several theatrical trailers of popular films and television shows and after each trailer, a rating is required on a 5-star scale (as used on Amazon, etc.). The re-rating phase starts with an initiation run in which four very short trailers are shown and rated. One of these introductory trailers is shown twice to prepare the participants for an emerging redundancy so that confusion doesn't arise in further progress. These ratings are not recorded and won't affect latter evaluations. After the initiation phase, ten trailers are shown consecutively and must be seen completely ere a rating can be given to continue (trial 1). Immediately at the end of the first trial, all ten trailers are shuffled and presented once again (trial 2), i.e. there is no noticeable break when switching to the next trial. In the second trial, ratings can be submitted after 20 seconds. This ensures a shortening of runtime and acts as a prevention of rapid loss of interest when watching the same trailer repeatedly. This intention is also supported by adding five additional trailers (during the third and fifth trial) which are not to be repeated, just to keep things new and interesting. From the third trial, the trailers can be evaluated immediately, i.e.

[2] `https://www.clickworker.de/` (last accessed on Mar 10, 2020)

Trailer ID	Title	Genre (IMDb)	Duration (mm:ss)
I1	Xena Intro	adventure	00:59
I2	Terminator 2	science fiction	01:15
I3	Der König der Löwen	family	01:25
T1	**Star Wars 8**	**fantasy**	**02:16**
T2	Fack Ju Göthe 2	comedy	02:44
T3	**James Bond 007 - Spectre**	**action**	**02:40**
T4	**Minions**	**animation**	**02:44**
T5	Fifty Shades of Grey	romance	02:27
T6	The Walking Dead (Season 5)	horror	01:25
T7	**Big Bang Theory (Season 8)**	**comedy**	**01:18**
T8	Suits (Season 1)	comedy	01:55
T9	**Arrow (Season 1)**	**crime**	**01:01**
T10	Shannarah Chronicles (Season 1)	fantasy	01:11
D1	Resident Evil	horror	01:54
D2	Avengers: Age of Ultron	action	02:12
D3	I, Frankenstein	horror	02:37
D4	Jurassic World	action	02:32
D5	Die Tribute von Panem - Mockingjay	action	01:43

Table 3.1: Trailer information. Trailers in boldface have been recorded.

there is no blocking time. This procedure continues until all five rating trials are completed. The submitted ratings have been recorded for five of ten trailers so that the remaining ones acted as distractors, triggering the misinformation effect, i.e. memory is becoming less accurate due to interference from post-event information. Further trailer information can be found in Tab. 3.1.

pdf-Rating Proceeding: This phase starts with another instruction (screen messages) telling the participants that they are about to face a new method for providing user ratings. At this point, no additional information was given to examine if the new design is intuitive and self-explanatory. The entire rating for an item is entered by five sliders. A screenshot of the utilised rating interface can be seen in Fig. 3.2.

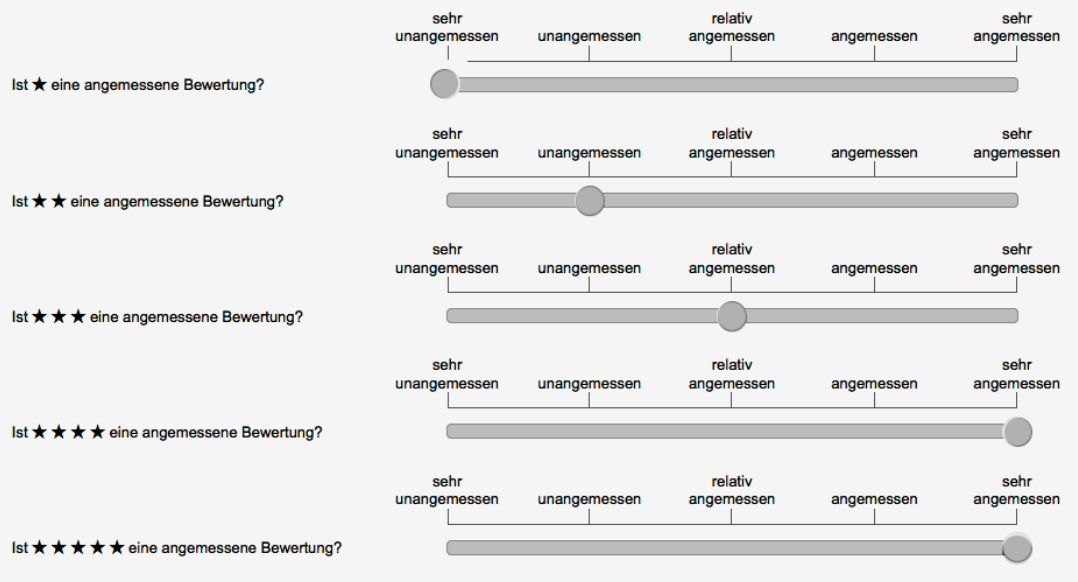
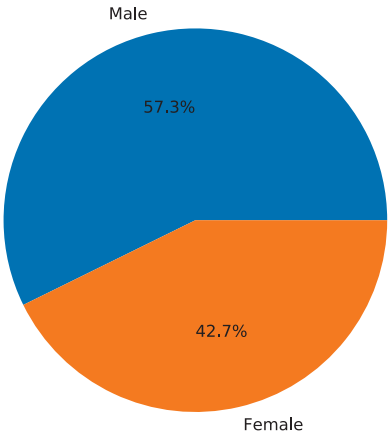


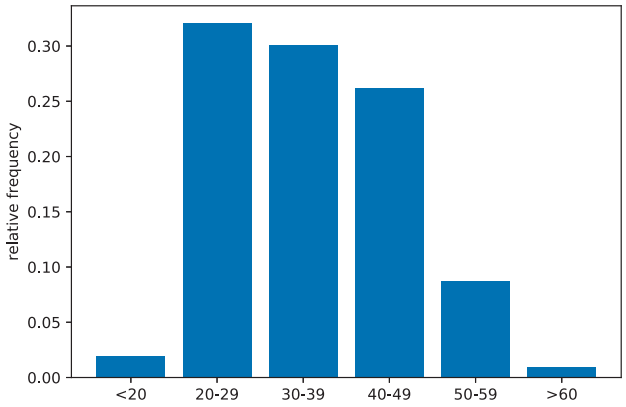
Figure 3.2: Implementation of the pdf-rating in the RETRAIN study. By entering the belief about the appropriateness of each rating, a probability density can be constructed.

Each of those sliders is used to set an answer to the question "Is a s-star-rating appropriate for this item?" The answers do vary from "not appropriate at all" to "very appropriate" and are internally encoded by integers 1 to 5. The initial run consists of one single pdf-rating for training purposes (understanding the new procedure) which is not recorded. Next, the participants have to rate those five trailers that were recorded throughout the re-rating proceeding by using this previously described new technique.

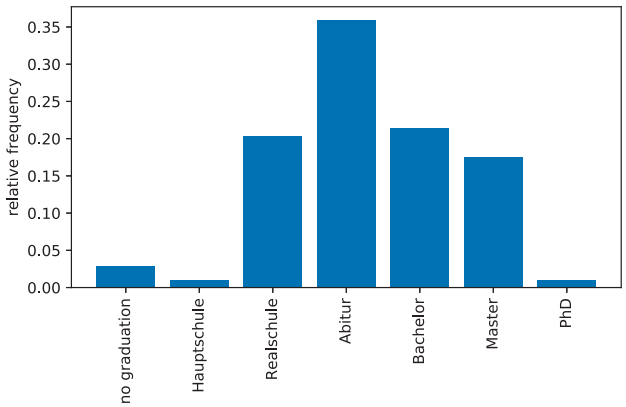
Altogether $N = 67$ people from Germany, Austria and Switzerland participated in this experiment. This group can be parted into 57% males and 43% females as to see in Fig. 3.3a. Although there is a slight gender imbalance, this is not significant considering the sample size. Therefore, statements based on the RETRAIN study are not to be regarded as gender-biased. The age distribution is depicted in Fig. 3.3b. There are almost equal fractions of the most relevant age groups to be spotted, so there is no age bias for further evaluations of this study. The education of participants is shown in Fig. 3.3c. A bias toward higher education can be recognised. In 2017, the fraction of those without graduation was 4.0%, those with the graduation of "Realschule" was 23.1% and those with "Abitur" was 31.9% (cf. Statistisches Bundesamt, 2019). Accordingly,



(a) gender

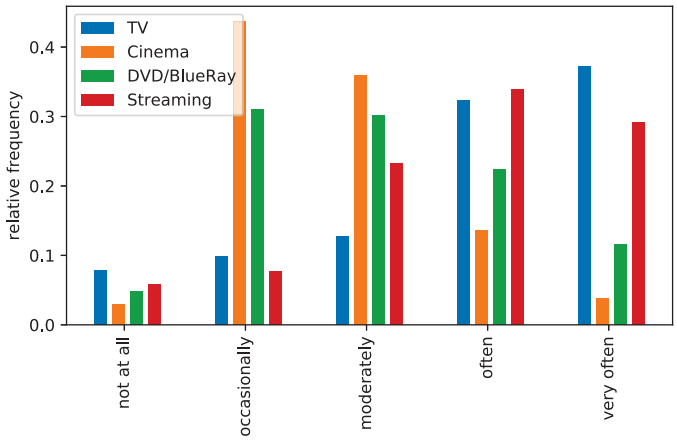


(b) age groups

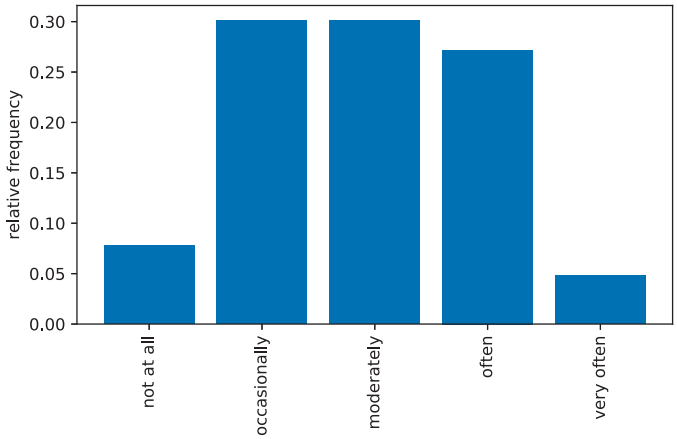


(c) education

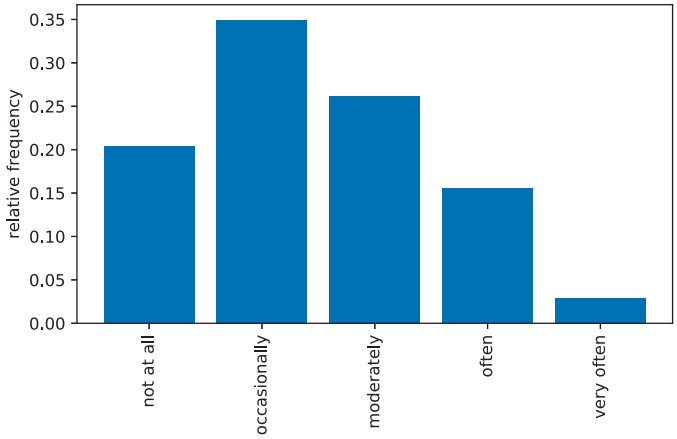
Figure 3.3: Demographic data of the RETRAIN participants



(a) using media



(b) rating items



(c) rating films etc

Figure 3.4: Rating experience/know-how of the RETRAIN participants

the medium educational levels in the RETRAIN study are in very good agreement with those of the German population. However, in 2017, 30.4% had a "Hauptschule" diploma (but only 3% in the study), while bachelor and master degrees were at 2.2% and 1.4% (but each over 20% in the study) (cf. Statistisches Bundesamt, 2019). So, the RETRAIN study contains significantly less lower educational levels and significantly more participants with higher education. This may be explained by the fact that many students might use crowdsourcing platforms to earn money and finance their studies. The habits of using media as well as rating products and especially films are depicted in Fig. 3.4. All participants are sufficiently media affine and their rating frequency habits range from "occasionally" to "often" in (almost) uniform distribution. According to this analysis, it can be assumed to have gathered representative cross-sectional data throughout the German-speaking population from Germany, Austria and Switzerland.

3.3 Existence of Uncertainty in Feedback Scenarios

The first goal of the RETRAIN study is to prove the existence of unreliable user feedback in an explicit rating scenario.The obtained data set comprises 335 user-item pairs and 1675 individual ratings for the re-rating proceeding. From all user ratings, only 35% manifested a consistent response behaviour whilst 50% changed their rating once and 15% changed their rating twice or more. A detailed breakdown can be found in Fig. 3.5. The overall proportion of reliability is depicted by the outer ring: 65% of all users do not provide consistent and reliable feedback in case of repetitions. Figure 3.6 depicts exemplary relative histograms of repeated ratings given by the same users to the same items. With these results, the criticism about the study of Hill et al. (1995) can be ruled out that changes of opinion only take place over long periods. Instead, even on short time intervals, one's opinion can vary significantly so that a single measurement is only an uncertain momentum of the user's actual opinion. These results are in accordance with the study of Amatriain et al. (2009b) and prove the existence of human uncertainty in feedback scenarios. In contrast to this study, yet another measurement approach for human uncertainty has been tested in the RETRAIN study in order to provide an alternative and more practical method. At this point, the question arises as to the equivalence of their results as well as possible advantages and disadvantages.

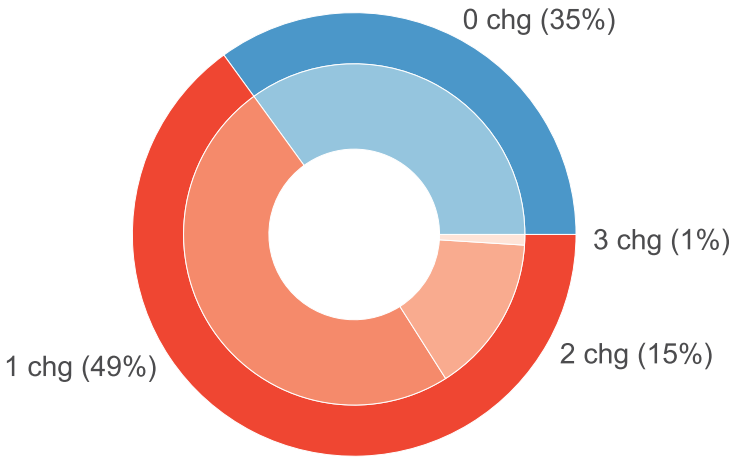


Figure 3.5: Change of response behaviour in five consecutive rating trials aggregated from all user-item pairs. This analysis indicates the existence of human uncertainty.

3.4 Systematic Comparison of Measurement Approaches

The second goal of the RETRAIN study is to explore the viability of survey methods to measure human uncertainty. It can be shown that both measurement methods do not lead to major differences between the resulting feedback distributions. To this end, different attributes of the corresponding feedback distributions (obtained from both approaches) are compared by using the following tests:

KS-test: With this test it can be determined if the user feedback distributions are the same for both measurement methods, i.e. the Kolmogorov-Smirnov test (cf. Krapp and Nebel, 2011, pp. 90–91) checks for the agreement of two probability distributions in general. This is done by comparing the empirical (discrete) distribution functions F_X and F_Y of the random variables X and Y via $d := \sup_{t \in \mathbb{R}} |F_X(t) - F_Y(t)|$. The equality of distributions is rejected at a significance level of α if d exceeds the critical value K_α, which can be taken from tables.

Welch's t-test: Even if the general equality of two distributions has to be rejected, they may nevertheless possess the same expectation that could be assigned to a user for future predictions. Therefore, Welch's t-test (cf. Janczyk and Pfister, 2013,

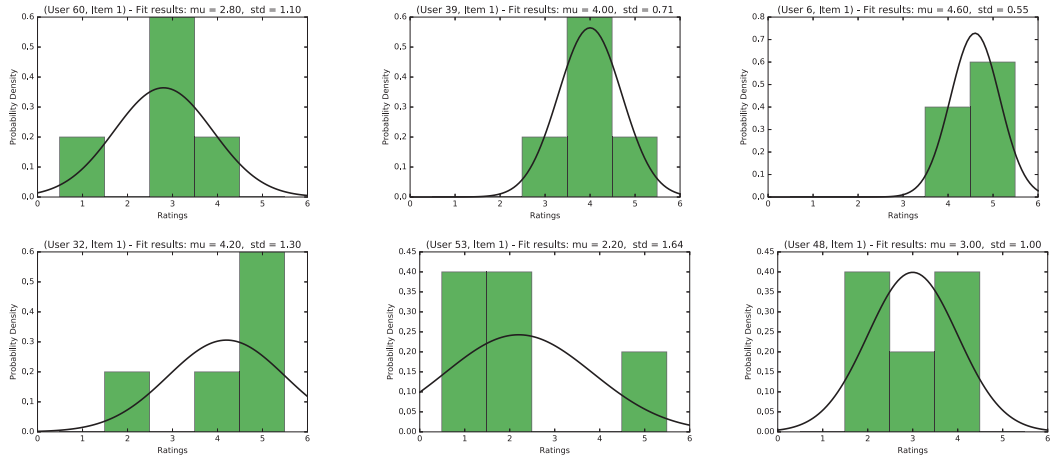


Figure 3.6: Visualisation of human uncertainty in the RETRAIN study: Exemplary histograms of the ratings given by users in five consecutive trials for the same item.

p. 53) is performed to compare the expected values. Welch's t-test is an adaptation of Student's t-test, however more reliable and suitable when the two samples have not necessarily equal variances. On the basis of two underlying random variables X and Y, Welch's t-test defines a statistic by $\mathrm{t} := (\overline{\mathrm{x}} - \overline{\mathrm{y}})/(\mathrm{s}_{\mathrm{X}}^2/\mathrm{n}_{\mathrm{X}} + \mathrm{s}_{\mathrm{Y}}^2/\mathrm{n}_{\mathrm{Y}})^{1/2}$, where $\overline{x}, \overline{y}$ are the sample means, s_X, s_Y the sample variances and n_X, n_Y the sample sizes, respectively. Again, the equality of means is rejected at a significance level of α if t exceeds the critical value (quantile of the t-distribution) $\mathrm{t}_{(\alpha,\mathrm{m})}$ with m degrees of freedom, which can be taken from tables.

Levene's Test: For the case of mismatching distributions and mismatching expected values, those distributions may still have the same variance that can be used for further analyses. Thus, Levene's test (cf. Ramachandran and Tsokos, 2009, pp. 722–723) is used to investigate homoscedasticity, which is suitable for random variables that are not necessarily normally distributed. Given the variables X and Y with sample sizes n_X and n_Y, the mean absolute deviations $a_X := \sum_{j=1}^{n_X} |x_j - \bar{x}|$ and a_Y respectively are computed along with their mean value $\bar{a} := (a_X + a_Y)/2$. The respective test statistic then constitutes as

$$\mathrm{L} := (\mathrm{n_X} + \mathrm{n_Y} - 2) \cdot \frac{\mathrm{n_X}(\mathrm{a_X} - \bar{\mathrm{a}})^2 + \mathrm{n_Y}(\mathrm{a_Y} - \bar{\mathrm{a}})^2}{(\sum_{\mathrm{j}=1}^{\mathrm{n_X}} |\mathrm{x_j} - \mathrm{a_X}|)^2 + (\sum_{\mathrm{j}=1}^{\mathrm{n_Y}} |\mathrm{y_j} - \mathrm{a_Y}|)^2}.$$

Homoscedasticity is rejected with a significance level of α when L exceeds the quantile $\mathrm{F}_{(\alpha,1,\mathrm{n_X}+\mathrm{n_Y})}$ of the F-distribution with 1 and $n_X + n_Y$ degrees of freedom.

	KS-test		Welch's t-test		Levene's test	
	n. rejected	rejected	n. rejected	rejected	n. rejected	rejected
Item 1	59 (1.00)	0 (0.00)	52 (0.88)	7 (0.12)	52 (0.88)	7 (0.12)
Item 2	39 (0.98)	1 (0.02)	34 (0.85)	6 (0.15)	33 (0.82)	7 (0.17)
Item 3	31 (0.94)	2 (0.06)	23 (0.70)	10 (0.30)	26 (0.79)	7 (0.21)
Item 4	45 (1.00)	0 (0.00)	38 (0.84)	7 (0.16)	37 (0.82)	8 (0.18)
Item 5	33 (0.97)	1 (0.03)	28 (0.82)	6 (0.18)	26 (0.76)	8 (0.24)

Table 3.2: Hypothesis testing for the feedback distributions (absolute counts first, fractions in brackets). The absolute count varies as the number of ratings with non-vanishing variance changes for each item.

All results of hypothesis testing presented hereinafter has been performed by comparing the respective p-values with a significance level of $\alpha = 0.05$.

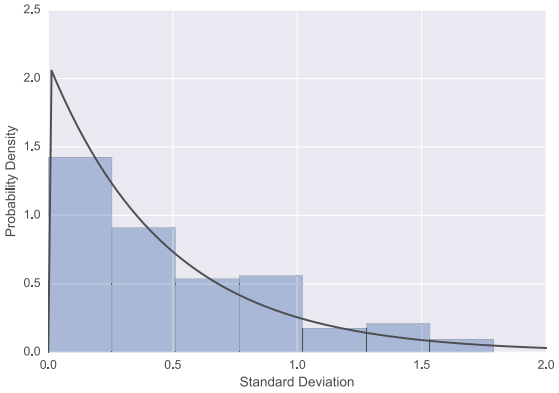
The test results are recapped in Tab. 3.2. For the equality of distributions (KS-test), it can be noticed that only very small fractions of distribution-pairs can be deemed to be significantly different. Vice versa, both measurement approaches produce feedback distributions that do not differ significantly in almost any case. The comparison of corresponding expectations (Welch's t-test) reveals that these do not differ significantly from one another in 83% (on average) of all cases. Similarly, a deviation from homoscedasticity (Levene's test) was only significant in 82% on average. It is noticeable that the KS-test does not detect significant differences between two distributions, although Welch's t-test and Levene's test detect different moments of these. This can be explained by the robustness and the power of these tests or respectively by how the test statistic is computed. The test statistic of the KS-test is linear in both arguments, whilst the other tests take into account squared deviations. This ensures that small deviations (less than one) are attenuated and larger deviations (greater than one) are amplified. In the end, this means that the applied tests for the moments of a particular distribution are much more sensitive and attest significant inequalities even for the smallest deviations. Having this in mind, rejection rates of only 20% are still quite good. Considering all the tests together, it can be argued that both methods, the re-rating and the pdf-rating, lead to feedback distributions that closely approximate each other.

At this point, some hints have likewise been found for the third goal of the RETRAIN experiment, which was to narrow down possible origins of unreliable user feedback. Earlier, the argument has been introduced that the varying "history" of a trailer, i.e.

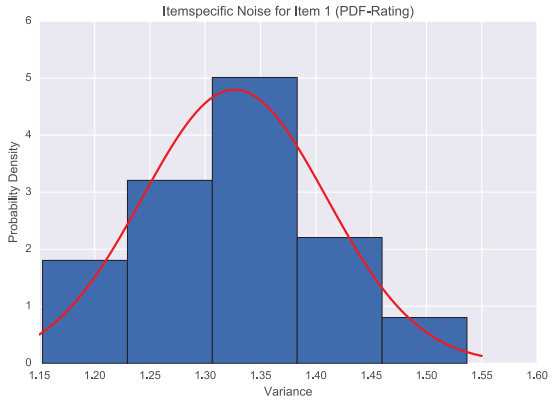
the variation of trailers that have been shown before, can be considered as a change in the situational context of a user and that this change of context might lead to differences in the current rating. However, for the pdf-rating, there is no such thing as a varying "history" during the measurement process since all uncertainty data is collected at once. The equality of resulting distributions, therefore, suggests that the "history" of a trailer has no impact on the uncertainty. The unreliable feedback must thus possess another origin. So, the alternative explanation that uncertainty is inherent to the human's natural cognition process becomes very attractive. Accordingly, the term "human uncertainty" is well chosen because it expresses this property very well.

In this contribution, human uncertainty is represented by the standard deviation or variance of the individual feedback distributions. In contrast to the feedback distributions that have been discussed above, the uncertainty distribution itself differs strongly when changing the measurement approach. For the re-rating, we yield a power-law-distribution as to see in Fig. 3.7a, i.e. many people are quite certain whereas larger uncertainty only manifests for a few people. It can be assumed that this is an artefact of the conventional rating instrument in which customers are forced to choose precisely one element of a discrete rating scale whilst other rating possibilities or even weightings cannot be considered at all. In contrast to this, the pdf-rating provides normally distributed uncertainties as to see in Fig. 3.7b and 3.7c, which are often found when considering human characteristics. The pdf-rating allows for a simultaneous weighting of different rating options and can be deemed to be more suitable for uncertain decision-making. A remarkable property of the resulting distributions of the pdf-rating when aggregated for each item is the common mean value of 1.3 stars. This can already be observed visually in those examples of Fig. 3.7. In particular, Welch's t-test indicates that only 10% of all distributions possess a significantly different expectation. The fact that human uncertainty appears to be an equally strong property in every human being (with some natural fluctuations) is hence another indication for a cognitive origin. As already indicated in Fig. 3.6, the user feedback can be modelled with normal distributions. Further use of the KS-test reveals that the deviations between the empirical densities and associated normal distributions are not significant in any case. In forthcoming sections, the normal distribution will provide significant advantages when modelling the uncertainty propagation in scenarios of comparative recommender system assessment.

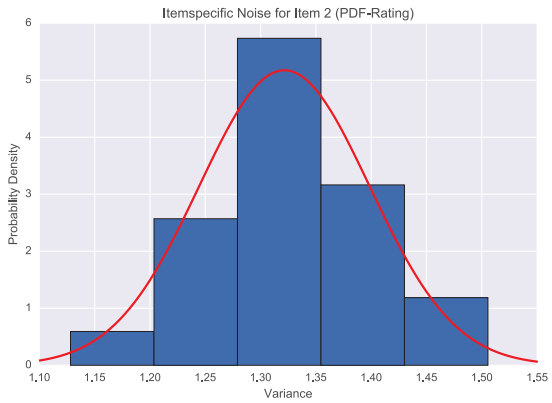
It should be noted that approximated statistics of a distribution model can only be located within confidence intervals since they are computed on samples rather than



(a) re-ratings



(b) pdf-ratings for item 1



(c) pdf-ratings for item 2

Figure 3.7: Distributions of feedback variances representing human uncertainty

on the whole population. This means that due to the limited amount of information gathered by both approaches, the statistics themselves are subject to some kind of measurement uncertainty as well. In the following, it will be investigated whether the choice of a particular measuring method influences this measurement uncertainty. For the assumption of normality, the confidence interval for the parameter μ_ν is given as

$$\mu_\nu \in \left[\bar{x}_\nu - \mathrm{t}_{(\mathrm{n}-1,1-\frac{\alpha}{2})} \frac{\mathrm{s}_\nu}{\sqrt{\mathrm{n}}} \; ; \; \bar{\mathrm{x}}_\nu + \mathrm{t}_{(\mathrm{n}-1,1-\frac{\alpha}{2})} \frac{\mathrm{s}_\nu}{\sqrt{\mathrm{n}}} \right] \tag{3.3}$$

where $\bar{x}$ and s are the point estimates for the mean and Bessel-corrected standard deviation and $\mathrm{t}_{(\mathrm{p,k})}$ represents the p-quantile of the t-distribution with k degrees of freedom (cf. Henze, 2013, p. 341). Additionally, the confidence interval of σ_ν for normality assumption is given as

$$\sigma_\nu \in \left[s_\nu \sqrt{(n-1)/\chi^2_{(1-\frac{\alpha}{2},n-1)}} \; ; \; s_\nu \sqrt{(n-1)/\chi^2_{(\frac{\alpha}{2},n-1)}} \right] \tag{3.4}$$

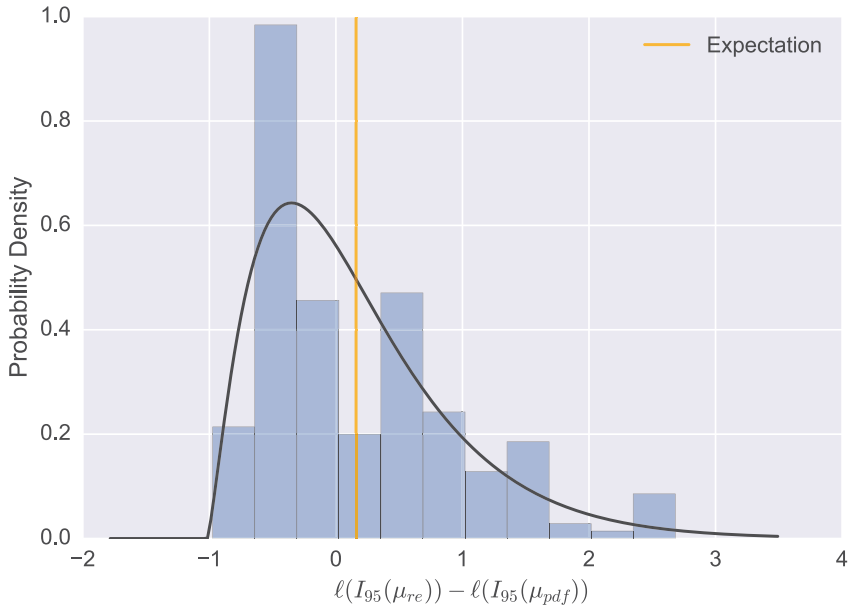
where $\chi^2_{(q,m)}$ is the q-quantile of the χ^2-distribution with m degrees of freedom (cf. Roxy and Devore, 2011, p. 295). To evaluate the measurement uncertainty for the feedback distribution's parameters, one can rely on the length of the 95% confidence intervals for the mean and variance. The shorter this interval, the smaller the measurement uncertainty, i.e. the more precise can a particular parameter be determined. The superiority of a specific measurement approach can then be expressed by the auxiliary variable

$$\Delta_\mu := \ell(I_{95}(\mu_{re})) - \ell(I_{95}(\mu_{pdf})), \tag{3.5}$$

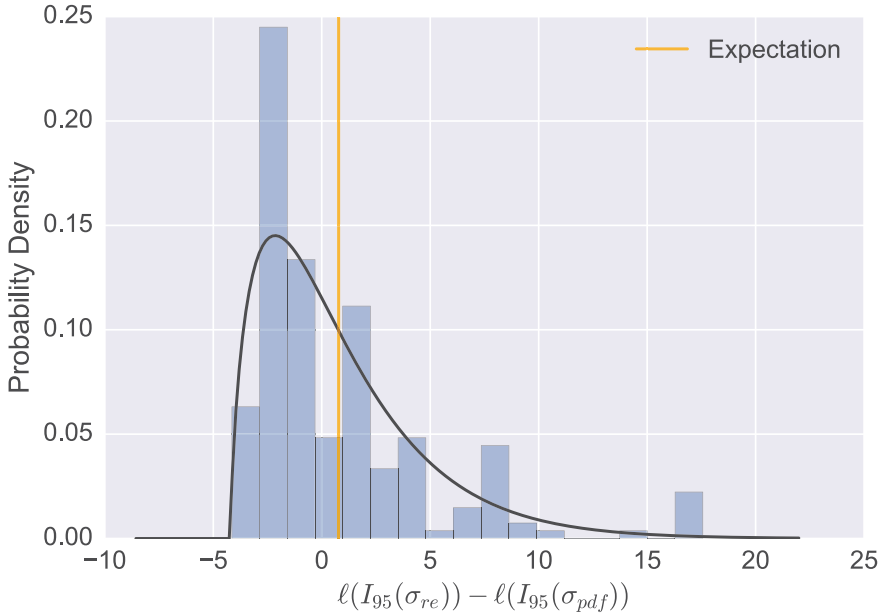
where $I_{95}(\mu)$ is computed according to Eq. 3.3. If $\Delta_\mu > 0$, then the length $\ell(I_{95}(\mu_{re}))$ of the re-rating-interval is greater than the length $\ell(I_{95}(\mu_{pdf}))$ of the pdf-rating-interval, i.e. the pdf-rating appears to be more precise in locating the mean value. The analysis of the standard deviation is done analogously by using Eq. 3.4.

Figure 3.8 depicts the distribution of these length differences. It can be seen that the mass-ratio of improvements and deteriorations is very balanced. At the same time, it can be seen that the strength of these deteriorations is small in comparison to the strength of improvements. The expectations show that on average, the pdf-rating will produce a slight increase in overall precision. This may be explained by the fact that the pdf-rating allows for options that cannot be captured at all with the re-rating.

Another suitable approach to compare the measurement precision is to compute the percentile distributions by re-sampling which turned out to be normal distributions



(a) measurement uncertainty of the mean value



(b) measurement uncertainty of the variance

Figure 3.8: Analysis of the feedback distributions' precision depending on the applied method of gathering uncertainty

again. Accordingly, the standard deviations of these percentile distributions naturally become representative for the inherent measurement precision. Therefore, let q be a particular percentile of a feedback distribution. A percentile distribution is then obtained qua re-sampling with respect to the uncertain feedback distribution parameters. Let $\sigma_{re}(q)$ and $\sigma_{pdf}(q)$ be the standard deviation of those percentile distributions when re-sampling from the re-rating proceeding or the pdf-rating proceeding, respectively. The auxiliary quantity

$$\delta_q := \sigma_{re}(q) - \sigma_{pdf}(q) \tag{3.6}$$

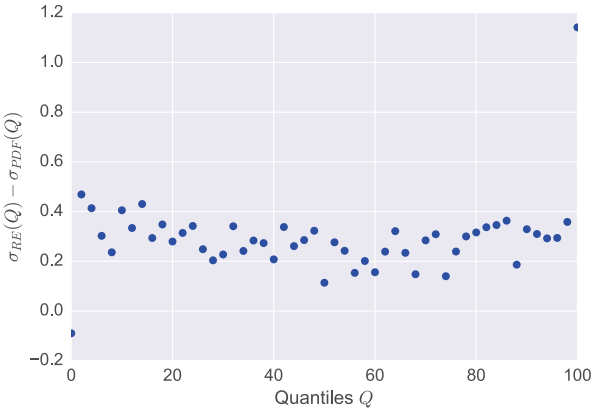
is positive for $\sigma_{pdf}(q) < \sigma_{re}(q)$ indicating superiority of the pdf-rating. When computing scatter plots for δ_q against q for each user-item pair, there are three repetitive archetypes to be spotted, which are monotonic behaviour (**homogeneity**), at least two clusters (**clustered**), and high dispersion with no visible relationship (**irregularity**). Representatives of these archetypes can be seen in Fig. 3.9. When interpreting these archetypes, it is important to remember that the auxiliary variable is a measure of how much precision is gained by choosing between the re-rating and the pdf-rating proceeding. This provides a hint as to which approach applies best to a current rating and allows for conclusions about the respective user:

- Homogeneous users either show no significant precision effect (constant line) or a functional relationship so that the "uncertainty by action" as measured by the re-rating can be converted into "uncertainty by cognition" as measured by the pdf-rating and vice versa. Cognition and action are closely linked for these users, i.e. they make their decisions very thoughtfully and possibly not based on feelings.

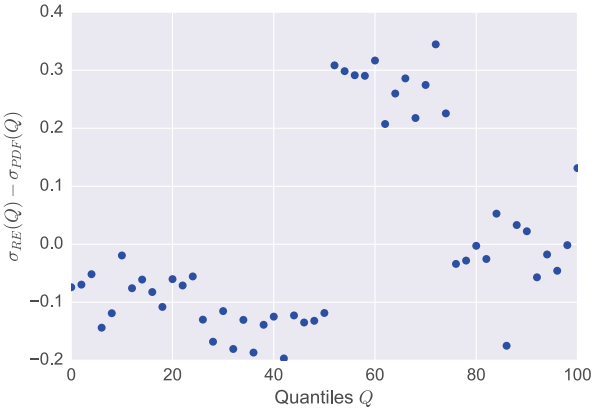
- For the cluster archetype, a functional relationship is to be seen which is interrupted at different points in which uncertainty by cognition seems to have more precision. There is no functional relationship between cognition and action in these cases so that they can be understood as the manifestation of a "gut feeling".

- The irregular archetype does not show any relationship between action and cognition. Probably, those users have not rated seriously and just clicked through the online survey.

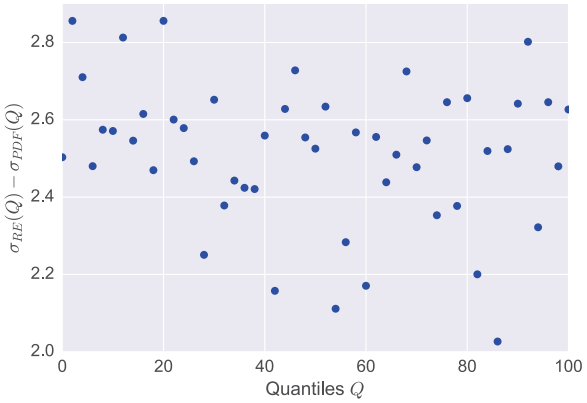
The fractions of these groups are similar to those of human uncertainty (cf. Fig. 3.5): Half of the users utilise their gut feeling when giving feedback, while only about a third of the users turn out to be a reliable source of information.



(a) homogeneous (28% of all users)



(b) clustered (45% of all users)



(c) irregular (27% of all users)

Figure 3.9: Examples of archetypes showing where cognition is more precise than action

3.5 Measurement Applicability and User Satisfaction

The second goal of this study was to explore the viability of survey methods for gathering human uncertainty information. To this end, both methods have been compared in a small follow-up questionnaire. As a result, two-thirds find the re-rating proceeding easier whilst a quarter does not see any differences in the difficulty of the two approaches. This can be explained by the fact that the re-rating simply repeats a long-known and standardised method. Despite the repetition, the basic principle has long been internalised. In the case of the pdf-rating, however, the user has to adapt to a completely new questioning approach which seems to be more cumbersome and therefore more complicated. This can certainly be improved by granting users more time to get used to this new approach. On the other hand, when questioning the participants about their very own motivation to continue giving product ratings, both methods perform equally, i.e. there is no preference for one particular method. So, if it is up to the users, both methods are equally applicable, even if the pdf-rating is perceived as more difficult at first sight. From a technical point of view, the collection of a re-rating is much more demanding: If the evaluation is repeated to rapidly, i.e. with too short temporal gaps or with too few distractors in between, this proceeding will not result into a valid uncertainty measurement due to memory effects. Then again, if the time intervals are too long, incisive changes of one's situational context may occur so that a possible bias does not allow a valid measurement as well. Applying constant laboratory conditions to a group of real users (to guarantee measurement validity) is certainly difficult. By contrast, the pdf-rating is much more flexible to use and far less demanding in its control.

3.6 Chapter Summary

In this section, two methods are presented to gather information about human uncertainty from explicit feedback. The re-rating proceeding measures by employing a user's action, i.e. when repeating the same feedback task after the memory of the first time has subsided. On the contrary, the pdf-rating measures by employing a user's perception, that is, by asking for the personal confidence for the correctness of each possible rating. Both methods lead to feedback distributions that do not differ significantly from each other. However, the distribution of uncertainty for the entire

population is considerably different. This may be explained by the fact that for the re-rating proceeding only discrete values are required each time, while the pdf-rating allows for considerably more gradings and relations between possible ratings. If people are indeed making decisions based on internal distributions, the pdf-rating constitutes a much more natural input method. In terms of measurement precision, the pdf-rating is also slightly superior to the re-rating. Both approaches can be easily adapted for all forms of explicit user feedback. Either one repeats the questioning of users multiple times or requires the personal confidence or belief for all response options at the same time. In a follow-up survey to the RETRAIN study, the participants stated that they have no special preference for any of these measurement approaches despite the assumed simplicity of the re-rating. This fact is especially important if one considers how to collect more explicit user feedback and motivate its input in the future. The collected data set from this experiment will be essential for any of the further analyses, i.e. all further explorations are either based directly on this data set or arise from simulations that transfer this data set to other situations.

4 | Impact of Human Uncertainty

The purpose of this chapter is to elaborate on subgoal B, i.e. to demonstrate and analyse limitations for comparative assessments of recommender systems in terms of accuracy metrics. This chapter is mainly based on my work Jasberg and Sizov (2019) and has been published almost verbatim there. In addition, various parts of this chapter have been formerly published in corresponding articles as well: The Monte-Carlo analyses for the RMSE density along with the worst/best-case consideration have been published Jasberg and Sizov (2017a). The ranking error probability as well as the sensitivity analysis for the error probability and RMSE density has been published in Jasberg and Sizov (2018a). The magic barrier estimation has been published in Jasberg and Sizov (2017b). A summary has been published in Jasberg and Sizov (2018c), along with the computation of ranking errors for the Netflix Prize. However, this chapter underwent a linguistic revision for this dissertation.

4.1 Modelling Uncertainty Propagation

It has already been demonstrated in Ch. 1 that the presence of human uncertainty in explicit user feedback leads to some kind of uncertainty in metric scores that are calculated therefrom. This section is dedicated to describe possible models of uncertainty propagation (research question B1). An arbitrary metric Z can be seen as a composed

quantity, i.e. it cannot be measured directly but has to be computed on the basis of measurable quantities such as user feedback. Mathematically speaking, let $(X_k)_{k=1,\dots,n}$ be an arbitrary family of random variables (e.g. user feedback) and $g \in C^\infty(\mathbb{R}^n)$ a smooth function that is not necessarily linear. Then $Z = g(X_1, \dots, X_n)$ is denoted a composed quantity (e.g. an accuracy metric) and becomes a random variable itself. The probability density f_Z of Z emerges as a convolution of all densities $(f_{X_k})_{k=1,\dots,n}$ with respect to the mapping g (cf. JCGM, 2008c, p. 8). This reasoning can be understood heuristically: For each draw, there is a variety of possibilities for an outcome x_k of a random variable X_k. Having one outcome $x_1, \dots, x_n$ for each random variable, one can be compute a single outcome $z = g(x_1, \dots, x_n)$ of the composed quantity Z. Accounting for all the possibilities for $x_1, \dots, x_n$ (e.g. when repeating draws infinitely) will then result in a variety of calculated outcomes z. The normed relative histogram of all z-values is a representation of the probability density f_Z (according to the frequentist definition of probability). For a better understanding, this explanation is additionally illustrated in Fig. 4.1.

Determining propagation via Monte-Carlo simulation. A typical example of such composed quantities are accuracy metrics, e.g. the mean average error (MAE), the mean squared error (MSE) and the root mean squared error (RMSE) among many others. The derivation of a metric's probability density will be exemplified by using the prominent accuracy metric

$$\text{RMSE} := \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{\nu}(\mathfrak{F}_\nu - \pi_\nu)^2}, \tag{4.1}$$

where $\mathfrak{F}_\nu \sim \mathcal{N}(\mu_\nu, \sigma_\nu^2)$ is the feedback distribution for a user-item pair $\nu := (u, i)$, π_ν is the corresponding prediction of an arbitrary recommender, and N is the number of user-item pairs. For each of the $\mathfrak{F}_\nu$ a sample $\mathcal{S}_m(\mathfrak{F}_\nu) := \{\mathfrak{f}_\nu^1, \dots, \mathfrak{f}_\nu^m\}$ is computed with m pseudo-random numbers (Monte-Carlo trials) that are drawn from the underlying feedback distribution. With these samples, one can compute a sample for the RMSE via

$$\mathcal{S}_m(\text{RMSE}) = \left\{ z_j = \sqrt{\frac{1}{N}\sum_{\nu}(\mathfrak{f}_\nu^j - \pi_\nu)^2} \;\middle|\; j = 1, \dots, m \right\}. \tag{4.2}$$

Post hoc illustration of this sample by a normalised relative histogram leads to a representation of the RMSE's density, which can then be assigned a distribution model with statistics derived from a maximum likelihood estimation. Due to randomness,

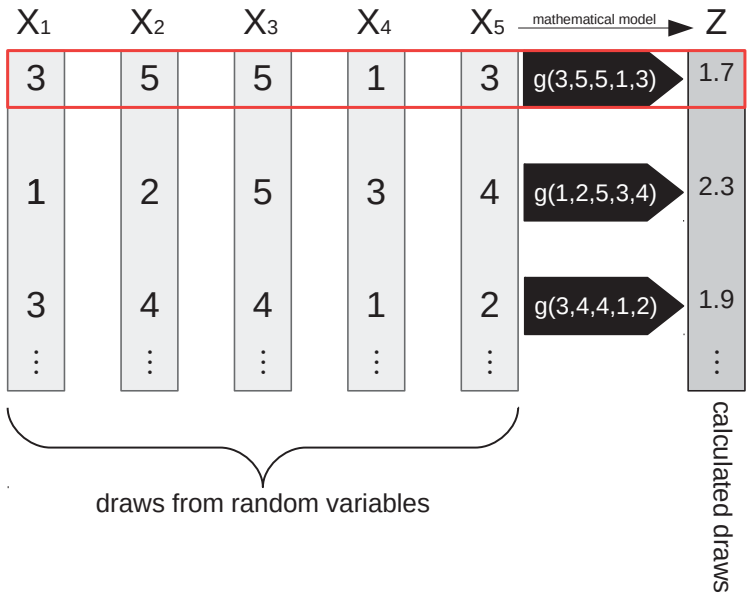


Figure 4.1: Illustration of a Monte-Carlo convolution of probability densities. Let $X_1, \ldots, X_5$ be arbitrary random variables. For each variable, single outcomes (columns) are drawn and for each complete set of draws (rows), a quantity can be computed through a mathematical function g. All these calculated values can be understood as the representation of a novel random variable $Z = g(X_1, \ldots, X_5)$. For infinite repetitions, the normed relative histogram of all calculated representations converges into the probability density of Z.

these Monte-Carlo simulations may fluctuate slightly for computation repetitions, but this effect diminishes for a high number of trials. In the present analyses, stable results have been reached by setting $m = 10^6$.

In Sec. 3.4, it has been shown that measurement approaches for human uncertainty only suffice to locate the feedback distribution's statistics within confidence intervals. Thus, the question arises how this measurement uncertainty affects our results. To investigate this issue it is crucial to have sample RMSE distributions as well as sample recommender systems. In the following, there are six recommender systems to be used which are defined by their predictors via

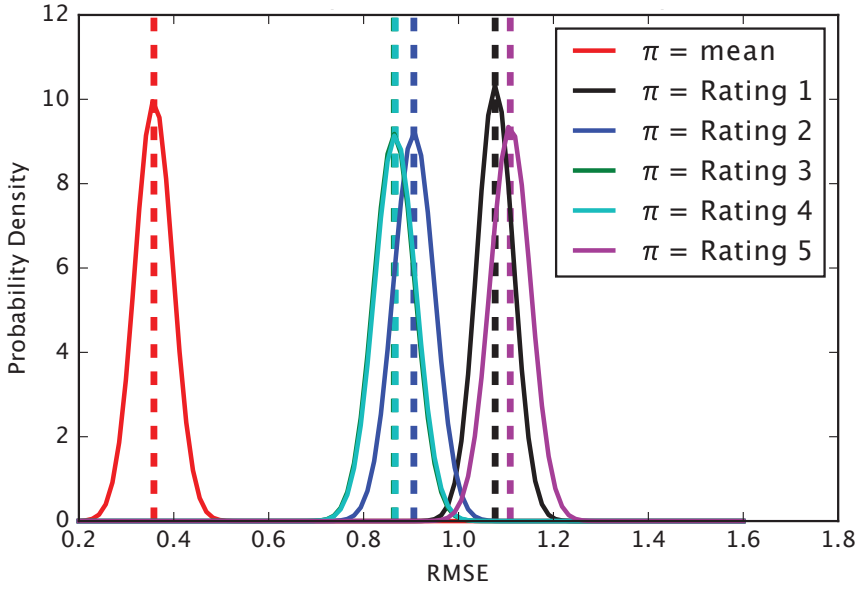
Recommender R1	$\pi :=$ mean of ratings from user to item
Recommender R2	$\pi :=$ 1st ratings from user to item
Recommender R3	$\pi :=$ 2nd ratings from user to item

Recommender R4	$\pi :=$ 3rd ratings from user to item
Recommender R5	$\pi :=$ 4th ratings from user to item
Recommender R6	$\pi :=$ 5th ratings from user to item.

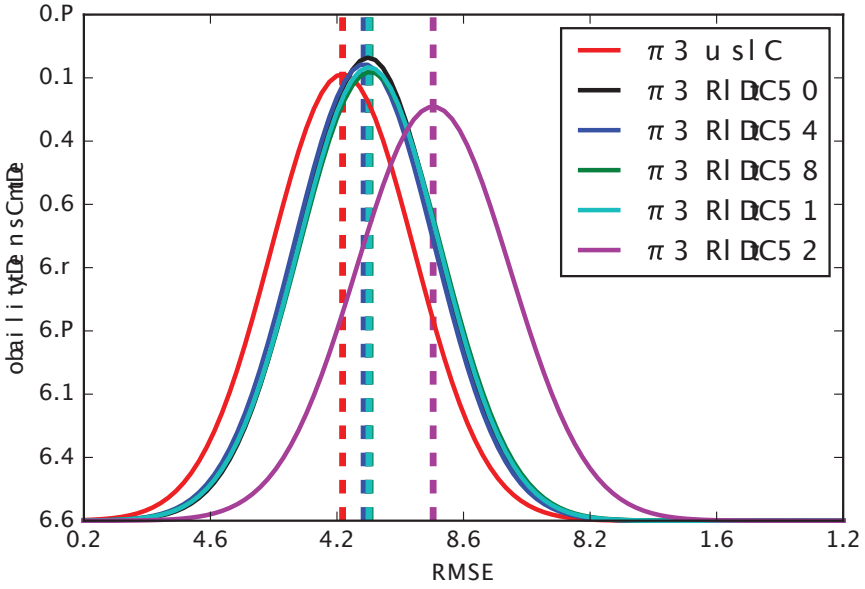
For each of the recommender systems, samples $\mathcal{S}_m(\mathrm{RMSE}(\mathrm{R}\,k))$ are computed on the basis of the RETRAIN data record. To explore the impact of measurement uncertainty, one simply computes the borderline cases for the RMSE distributions by assigning the parameters μ_ν and σ_ν of each feedback distribution as the lower limits of their corresponding confidence interval and the upper limits, respectively.

Figure 4.2 visualises the impact of measurement uncertainty for the re-rating proceeding (the pdf-rating produces similar results). It follows that the resulting distributions of the RMSE are ambiguous. A good distinction can be recognised for three groups of RMSEs in the minimum case. Therefore, a ranking of these groups will be possible without having large probabilities of error. However, this clear distinction is no longer possible for the maximum case. In this case, one cannot build a ranking order since all recommender systems are more or less the same in terms of this specific accuracy metric. The true distributions of those RMSEs can vary between these two limits but remain unknown on the basis of the information collected. In short, with only five re-ratings or one pdf-rating, it is not possible to get high-quality uncertainty information. This deficiency is not grounded within this new probabilistic perspective itself. In reality, one has to distinguish between two different types of uncertainty: On the one hand, there is the human uncertainty (leading from feedback scores to distributions) which is in the main focus of this thesis. But on the other hand, there is also a kind of measurement error which is denoted as measurement uncertainty. The variability for the RMSE distributions in Fig. 4.2 is completely explained by the impact of this measurement uncertainty and the small amount of information that can be measured.

But how much information is needed to reduce the ambiguity of the RMSE? To answer this question it is necessary to gradually increase the number of re-ratings (or to allow for finer graduation of the confidence scale when performing the pdf-rating). Both will result in larger data sizes and decreasing widths of corresponding confidence intervals. A simulation is used to estimate the number of re-ratings that are necessary to bring both borderline cases closer together. As a measure of this convergence, the intersection area of the minimum and maximum density is computed for each recommender system as explained in Fig. 4.3. High intersection areas will indicate that



(a) best case



(b) worst case

Figure 4.2: Borderline cases for the RMSE distributions emerging from the ambiguity of feedback distributions

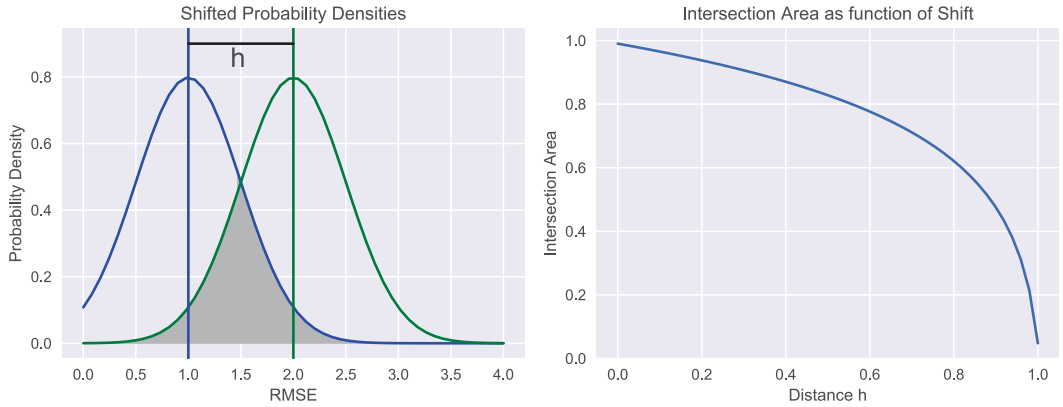


Figure 4.3: Measure of convergence for two probability densities that differ by a shift of their location parameter. When this shift (distance h) decreases, the intersection area of both densities converges to one.

both densities are close together. As can be seen from Fig. 4.4, between 1000 and 2000 re-ratings are required to achieve a convergence of both borderline cases by more than 90%. This means that users in a real rating scenario would have to re-evaluate the same item at least 1000 times in order to locate the RMSE distribution accurately. For the pdf-rating, this amount of information is equivalent to having a discrete confidence scale $C = \{0, 1, \ldots, 200\}$. Therefore, both solutions are virtually not feasible. No user would re-evaluate the same item for a thousand times, nor would a user be able to cope with the size of an ordinal scale providing 200 different options. The development of novel feedback procedures with higher informative value is still at an infancy stage and remains as an interesting direction of research for the future.

Although the statistical simulation of convolutions produces excellent results while also being easy to realise, run-time problems arise for big data. This is demonstrated by computing the RMSE for different data sizes on a bullx B510 computing node having a 2.7 GHz Intel E5-2697v2 (Ivy Bridge EP) along with 55 GB DDR3 1866MHz (59.8 GB/s). A constant predictor $\pi_\nu := 3$ will be used for each user-item pair as a possible recommender system. The runtimes for different data sizes are depicted in Fig. 4.5. For example, with $N = 80\,000$ user-item pairs, the simulation already takes up to an hour of runtime. To compute the RMSE on the Netflix test record ($N = 2.8 \cdot 10^6$), one would need about 30 hours. For this reason, it is necessary to find an analytical solution or at least an approximation of the resulting metric distribution in a closed

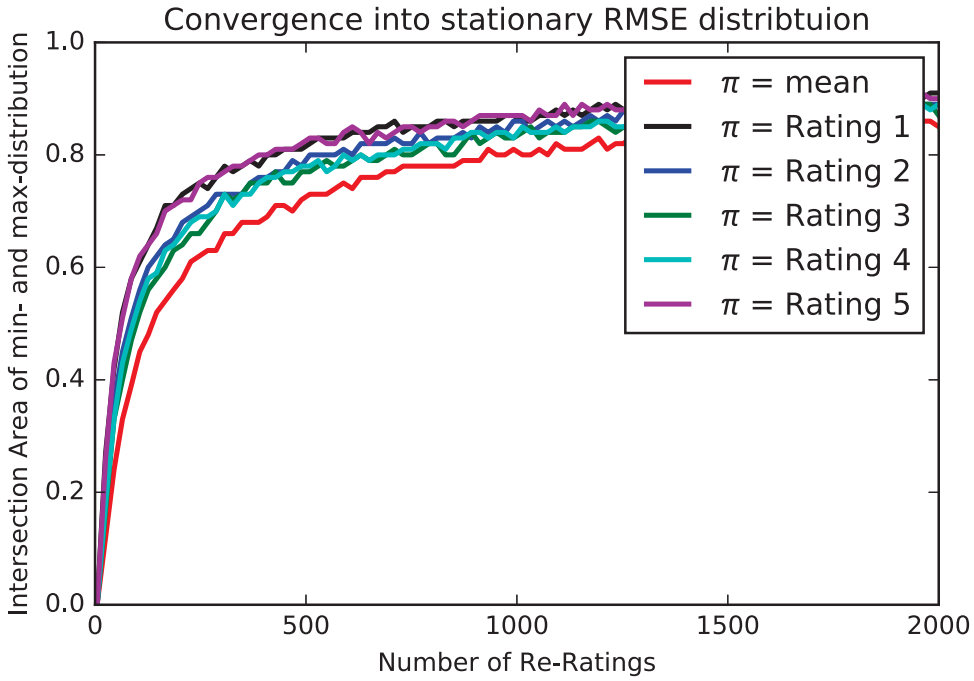


Figure 4.4: Intersection of both borderline cases of the RMSE distribution. The higher this intersetction, the closer are both borderline distributions together.

form. In the following, such estimation will be derived with which the same results can be obtained, however, in just a fraction of the simulation time. The runtime results of this approximation are also depicted in Fig. 4.5 and shows that the RMSE can be computed in less than 200 milliseconds on the Netflix test record.

Determining propagation via analytic derivation. The idea of deriving the RMSE's density analytically is to calculate its moments step by step for each particular transformation that is used to process the user data. This effort becomes straightforward if the mapping g of the composed quantity $Z = g(X_1, \ldots, X_n)$ is linear in each of its arguments since the (pseudo-)linearities

$$\mathbb{E}[aX \pm bY] = a\mathbb{E}[X] \pm b\mathbb{E}[Y] \tag{4.3}$$

$$\mathbb{V}[aX \pm bY \pm c] = a^2\mathbb{V}[X] + b^2\mathbb{V}[X] \tag{4.4}$$

for the mean and variance can be used, where $a, b, c \in \mathbb{R}$ are scalars and X, Y are arbitrary independent random variables. By means of the central limit theorem, $\mathbb{E}[g(X_1, \ldots, X_n)]$

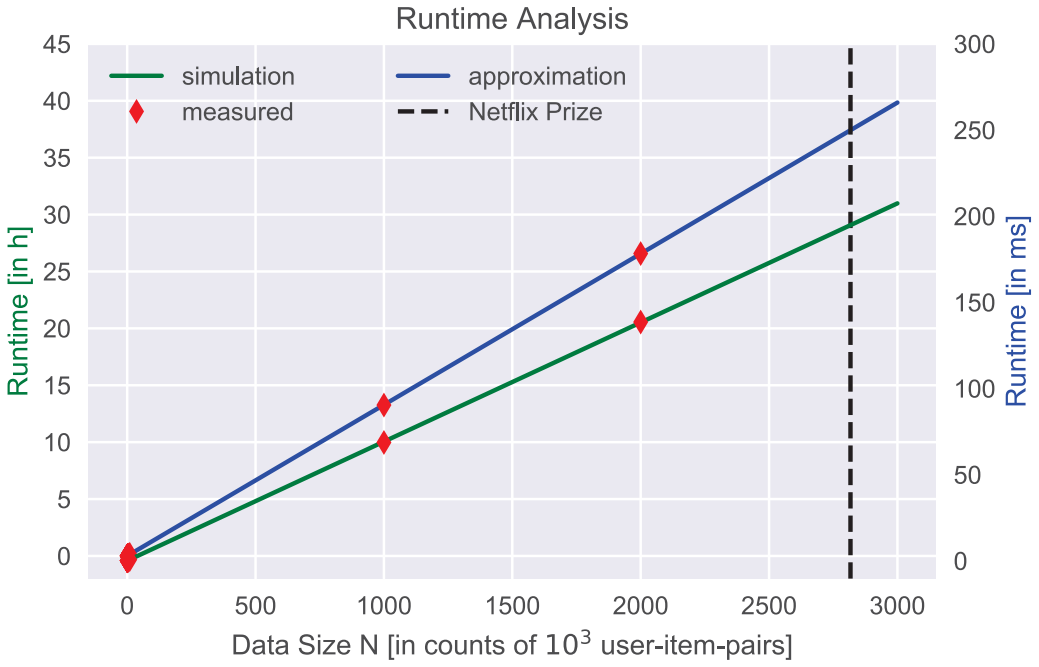


Figure 4.5: Runtime of computing the probability density of the RMSE using Monte-Carlo simulation (green) as well as approximates (blue) on a bullx B510 computing node

and $\mathbb{V}[g(X_1, \ldots, X_n)]$ become sufficient statistics for Z when large n are considered.

Considering the measured feedback distributions $\mathfrak{F}_\nu$, the expected value and variance will first be derived for the accuracy metric

$$\text{MSE} := \tfrac{1}{N}\textstyle\sum_\nu (\mathfrak{F}_\nu - \pi_\nu)^2. \tag{4.5}$$

Subsequently. the square root will be considered to obtain the corresponding statistics for the RMSE. By using Gaussians as the underlying data model, each user feedback $\mathfrak{F}_\nu \sim \mathcal{N}(\mu_\nu, \sigma_\nu)$ can be written as $\mathfrak{F}_\nu = \sigma_\nu \mathcal{I} + \mu_\nu$ where $\mathcal{I} \sim \mathcal{N}(0, 1)$.
Then, $Y_\nu := (\mathfrak{F}_\nu - \pi_\nu)^2$ receives the parameters

$$\begin{aligned} \mathbb{E}[Y_\nu] &= \mathbb{E}[(\sigma_\nu \mathcal{I} + \Delta_\nu)^2] = \mathbb{E}[\sigma_\nu^2 \mathcal{I}^2 + 2\mathcal{I}\sigma_\nu \Delta_\nu + \Delta_\nu^2] \\ &= \sigma_\nu^2 \mathbb{E}[\mathcal{I}^2] + 2\sigma_\nu \Delta_\nu \mathbb{E}[\mathcal{I}] + \mathbb{E}[\Delta_\nu^2] \\ &= \sigma_\nu^2 + \Delta_\nu^2 \end{aligned} \tag{4.6}$$

$$\begin{aligned}\mathbb{V}[Y_\nu] &= \mathbb{V}[(\sigma_\nu\mathcal{I}+\Delta_\nu)^2] = \mathbb{V}[\sigma_\nu^2\mathcal{I}^2+2\mathcal{I}\sigma_\nu\Delta_\nu+\Delta_\nu^2] \\ &= \sigma_\nu^4\left(\mathbb{E}[\mathcal{I}^4]-\mathbb{E}[\mathcal{I}^2]^2\right)+4\sigma_\nu^2\Delta_\nu^2 \\ &= 2\sigma_\nu^4+4\sigma_\nu^2\Delta_\nu^2 \end{aligned} \tag{4.7}$$

where $\Delta_\nu := \mu_\nu - \pi_\nu$ represents the local prediction quality of an arbitrary recommender system. In a second step, the random variable $\mathrm{MSE} = \frac{1}{N}\sum_\nu Y_\nu$ is a sum of χ^2-densities and turns into a Gaussian for a large number of ratings. For this Gaussian, one yields

$$\mathbb{E}[\mathrm{MSE}] = \tfrac{1}{N}\textstyle\sum_\nu \mathbb{E}[Y_\nu] = \tfrac{1}{N}\sum_\nu \sigma_\nu^2+\Delta_\nu^2 \tag{4.8}$$

$$\mathbb{V}[\mathrm{MSE}] = \tfrac{1}{N^2}\textstyle\sum_\nu \mathbb{V}[Y_\nu] = \tfrac{2}{N^2}\sum_\nu \sigma_\nu^4+2\sigma_\nu^2\Delta_\nu^2. \tag{4.9}$$

For some composed quantities like the RMSE, this approach will not work properly due to non-linearity, e.g. as for the root function. However, the expectation and the variance can be obtained in integral-form due to the identity $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(t)f_X(t)\,\mathrm{d}t$ for arbitrary functions g. For the example of the RMSE, one yields

$$\mathbb{E}[\sqrt{\mathrm{MSE}}] = \int_{-\infty}^{\infty}\sqrt{t}\cdot f_{\mathrm{MSE}}(t)\,\mathrm{d}t \tag{4.10}$$

$$\begin{aligned}\mathbb{V}[\sqrt{\mathrm{MSE}}] &= \mathbb{E}[\sqrt{\mathrm{MSE}}^2]-\mathbb{E}[\sqrt{\mathrm{MSE}}]^2 \\ &= \mathbb{E}[\mathrm{MSE}]-\left(\int_{-\infty}^{\infty}\sqrt{t}\cdot f_{\mathrm{MSE}}(t)\,\mathrm{d}t\right)^2\end{aligned} \tag{4.11}$$

where f_{MSE} is the probability density function of the MSE - which is a Gaussian with parameters given by Eq. 4.8 and 4.9. Unfortunately, these integrals can not be solved analytically and must be approximated using numerical methods. The density function of the RMSE can be derived by its cumulative distribution

$$\begin{aligned}F_{\mathrm{RMSE}}(t) &= P(\sqrt{|\,\mathrm{MSE}\,|}\le t) = P(-t^2\le \mathrm{MSE}\le t^2) \\ &= F_{\mathrm{MSE}}(t^2)-F_{\mathrm{MSE}}(-t^2) \\ &= 2\cdot F_{\mathrm{MSE}}(t^2)-1\end{aligned} \tag{4.12}$$

and the probability density function can thus be written as

$$f_{\mathrm{RMSE}}(t) = \frac{d}{dt}F_{\mathrm{RMSE}}(t) = 2\cdot\frac{d}{dt}F_{\mathrm{MSE}}(t^2) = 4t\cdot f_{\mathrm{MSE}}(t^2). \tag{4.13}$$

This method is exact and fast but may also be too complicated for an easy utilisation in individually designed recommender systems. To increase simplicity, it would be advantageous to have neat approximations in a closed form.

Determining propagation via approximation. Approximating a metric's density can be done by the so-called Gaussian Error Propagation, which is a very common approach in physics and metrology (cf. Ku, 1966). The core of this estimation is to expand $g \in C^\infty(\mathbb{R})$ into its Taylor series

$$g(X) = \sum_{k=0}^{\infty} \frac{g^{(k)}(\mu)}{k!} (X-\mu)^k \tag{4.14}$$

where $g^{(k)}(\mu)$ denotes the k-th derivative of g evaluated at the expectation of X. Due to the linearity of the expectation, it follows that

$$\begin{aligned} \mathbb{E}[g(X)] &= \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{g^{(k)}(\mu)}{k!} (X-\mu)^k\right] = \sum_{k=0}^{\infty} \frac{g^{(k)}(\mu)}{k!} \mathbb{E}\left[(X-\mu)^k\right] \\ &= \sum_{k=0}^{\infty} \frac{g^{(k)}(\mu)}{k!} m_k \end{aligned} \tag{4.15}$$

where m_k is the k-th central moment. For the variance and its quasi-linearity one yields

$$\begin{aligned} \mathbb{V}[g(X)] &= \mathbb{V}\left[\sum_{k=0}^{\infty} \frac{f^{(k)}(\mu)}{k!} (X-\mu)^k\right] = \sum_{k=0}^{\infty} \left(\frac{f^{(k)}(\mu)}{k!}\right)^2 \mathbb{V}\left[(X-\mu)^k\right] \\ &= \sum_{k=0}^{\infty} \left(\frac{f^{(k)}(\mu)}{k!}\right)^2 (m_{2k} - m_k^2) \end{aligned} \tag{4.16}$$

where the last line has been simplified by using the common identity $\mathbb{V}[(X-\mu)^k] = \mathbb{E}[(X-\mu)^{2k}] - \mathbb{E}[(X-\mu)^k]^2 = m_{2k} - {m_k}^2$. The usual approximation is now to omit terms of higher orders, like

$$\mathbb{E}[g(X)] = g(\mu) + g'(\mu) \cdot m_1 + \ldots \approx g(\mu) \tag{4.17}$$

$$\mathbb{V}[g(X)] = g'(\mu)^2 m_2 + g''(\mu)^2 (m_4 - m_2^2)/4 + \ldots \approx g'(\mu)^2 m_2, \tag{4.18}$$

where $m_2 = \sigma^2$ when using Gaussians. This approach can now be applied to the accuracy metric $\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$ by setting

$$g(X) = \sqrt{X} \quad \text{and} \quad g'(X) = \tfrac{1}{2\sqrt{X}}. \tag{4.19}$$

By inserting the MSE's mean and variance from Eq. 4.8 and 4.9, one yields the following estimations for the RMSE's statistics

$$\mathbb{E}[\mathrm{RMSE}] = \mathbb{E}[g(\mathrm{MSE})] \approx g(\mu) = \sqrt{\tfrac{1}{N} \textstyle\sum_\nu \sigma_\nu^2 + \Delta_\nu^2} \tag{4.20}$$

$$\mathbb{V}[\mathrm{RMSE}] = \mathbb{V}[g(\mathrm{MSE})] \approx g'(\mu)^2 \sigma^2 = \tfrac{\sum_\nu \sigma_\nu^4 + 2\sigma_\nu^2 \Delta_\nu^2}{2N \cdot \sum_\nu \sigma_\nu^2 + \Delta_\nu^2}. \tag{4.21}$$

Since this is an approximation, its quality has to be investigated, i.e. the degree of matching the true distribution obtained by simulation. In doing so, simulated means and variances are compared with calculated ones in a regression analysis. For this, $N \in \{50, 100, 150, 200, 500, 1000\}$ means μ are sampled uniformly from the interval $[1, 5]$ and N variances σ^2 are sampled uniformly from $[\sigma^2_{min}, \sigma^2_{max}]$. The variance boundaries result from the assumption of five repeated ratings (as happened in the RETRAIN study) with the commonly used 5-star scale. Under these conditions, the variance yields the limitations

$$\sigma^2_{min} = \mathrm{Var}(\{1, 1, 1, 1, 2\}) = 0.16 \quad \text{and} \quad \sigma^2_{max} = \mathrm{Var}(\{1, 1, 1, 5, 5\}) = 3.86. \tag{4.22}$$

For each pair (μ, σ^2) a sample of random numbers is drawn from $\mathcal{N}(\mu, \sigma^2)$ to perform the convolution via Monte-Carlo simulation (Eq. 4.2) and via approximation (Eq. 4.20 and 4.21). For many repetitions, a lot of simulated means/variances can be plotted against the approximated counterparts using linear regression. A perfect match between simulation and approximation would lead to the regression $y = 1 \cdot x + 0$ with a coefficient of determination $\rho^2 = 1$. The results

$$\mathrm{Sim}(\mathbb{E}) = 0.999 \cdot \mathrm{Apr}(\mathbb{E}) - 0.003 \qquad (\rho^2 = 0.99) \tag{4.23}$$

$$\mathrm{Sim}(\mathbb{V}) = 0.981 \cdot \mathrm{Apr}(\mathbb{V}) + 0.000 \qquad (\rho^2 = 1.00) \tag{4.24}$$

show that this condition is almost fully achieved and hence these approximations can be considered as appropriate. However, not only the mean and the variance are of great importance, but rather the entire probability density itself. While the simulated distribution arises naturally from convolution, it is predetermined by assumptions for the approximation approach. Therefore, it is necessary to evaluate the degree of matching for both distributions. For each sample, discrete probability distributions P_{sim} and P_{apr} are computed and analysed through the Jensen-Shannon-Divergence (JSD) as it is introduced in Lee (2000) where it is described as a "useful measure of the distance [i.e. similarity in this context] between distributions" (Lee, 2000, p. 2). It is defined as

$$\mathrm{JSD}(P_{sim}, P_{apr}) := \frac{1}{2}\,\mathrm{KL}(P_{sim}, M) + \frac{1}{2}\,\mathrm{KL}(P_{apr}, M) \tag{4.25}$$

where $\mathrm{KL}(P, Q) := \sum_i P(i)\,\log(P(i)/Q(i))$ denotes the commonly used Kullback-Leibler-Divergence and $M = \frac{1}{2}(P_{sim} + P_{apr})$. When using the binary logarithm for the Kullback-Leibler-Divergence, the JSD yields the boundaries $0 \leq \mathrm{JSD} \leq 1$ (cf. Lin, 1991,

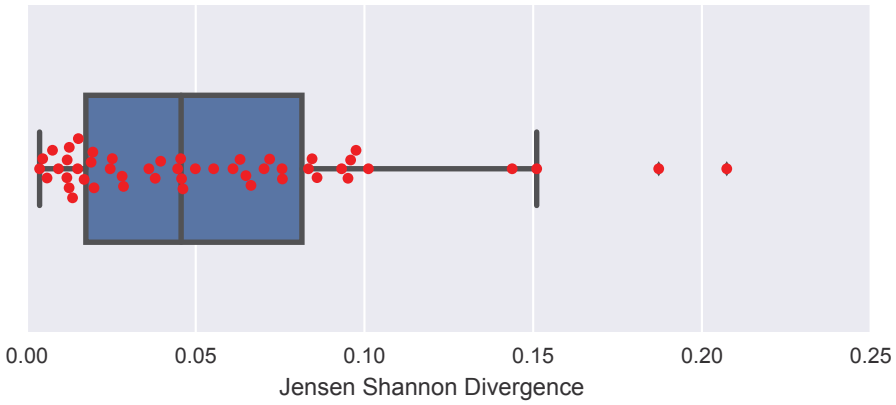


Figure 4.6: Jensen-Shannon-Divergence for comparing the simulated distribution with a predetermined Gaussian. Scores close to zero indicate a perfect matching.

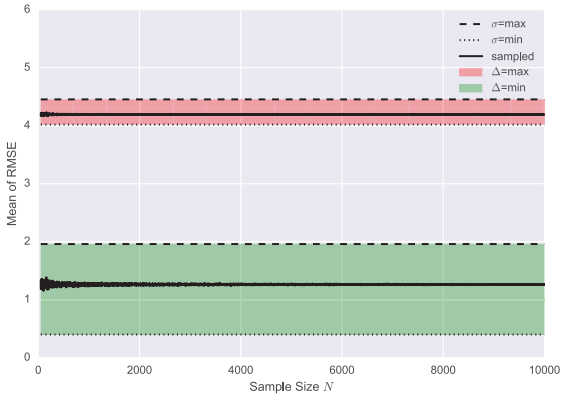
pp. 147–148). This is also the variant used in this thesis. The outcomes for the JSD are shown in Fig. 4.6. It can be observed that the mid-range of all outcomes is located between 0.01 and 0.08 confirming a high similarity of the simulated distribution and the assumed Gaussian respectively. There are, however, some outliers which only occur for $N = 50$ ratings. This can be explained by the fact that the RMSE contains the sum of squared normal distributions, which is χ^2-distributed, but quickly converges to the normal distribution for large N (cf. Walck, 1996, p. 39). Thus, the more ratings are considered, the more adequate becomes a Gaussian as the assumed density.

4.2 Properties of Uncertainty Propagation

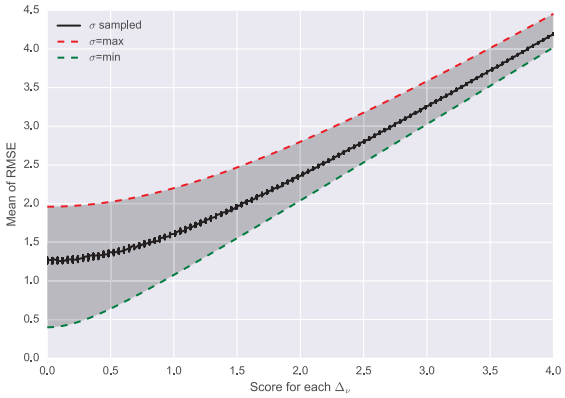
For the exploration of propagation properties (research question B2), the individual dependencies between explicit user feedback and the RMSE density are considered by employing a sensitivity analysis, i.e. it is determined how the distribution parameters respond to the variation of its arguments. Therefore, one argument is varied within reasonable boundaries while fixing all the other arguments at the same time. These boundaries depend on the utilised scale as well as the measuring approach for human uncertainty. Again, a 5-star scale is assumed as well as five repetition trials. Thus, the deviations between feedback and prediction yield the boundaries $0 < \Delta_\nu < 5$ whereas the possible non-vanishing variances range between $0.16 < \sigma_\nu^2 < 3.86$.

Impact on the Metric's Mean. Figure 4.7 depicts the outcomes for the expectation μ of the RMSE in correspondence to the number N of user-item pairs, the average deviation Δ_ν as well as the average human uncertainty σ_ν^2. Subfigure 4.7a shows that the mean of the RMSE is not affected by N (straight line with vanishing slope). It can also be recognised that the impact of human uncertainty is much higher for small deviations than for large ones (width of coloured bands). For small deviations (green/lower band) in particular, uncertainty may shift the RMSE's location from 0.5 to 2.0 (+300%) whereas a shift can only increase values from 4.0 up to 4.5 (+13%) for large deviations (red/upper band). This can be explained by the fact that although both quantities are equally represented in Eq. 4.20, the magnitude of differences is much higher for the deviations Δ_ν than for the human uncertainty σ_ν^2. Thus, for large deviations, the additional contribution of human uncertainty is lower. However, this implies serious problems, because the better a recommender system becomes (lower deviations), the more impact is given to human uncertainty – and this uncertainty is unlikely to be improved if its origin lies within the cognitive process. Subfigure 4.7b shows the reaction of the RMSE's mean on the variation of the average deviation between prediction and user feedback. The curve clearly shows that there is a functional dependency with asymptote $f(x) = x$. Here, the width of the grey band is an indicator of the influence of human uncertainty, which fades for large deviations. Again, a shift of the expected value can be identified, but for a much finer gradation than in the figure before. Subfigure 4.7c depicts the dependency on the average human uncertainty. The corresponding curve looks different from the curve obtained for the average deviation, although both quantities contribute equally to the mean of the RMSE. This indeed demonstrates that the magnitude of human uncertainty is much more limited than that of the deviations. Certainly, this curve looks like a zoom into the beginning part of the graph in Fig. 4.7b.

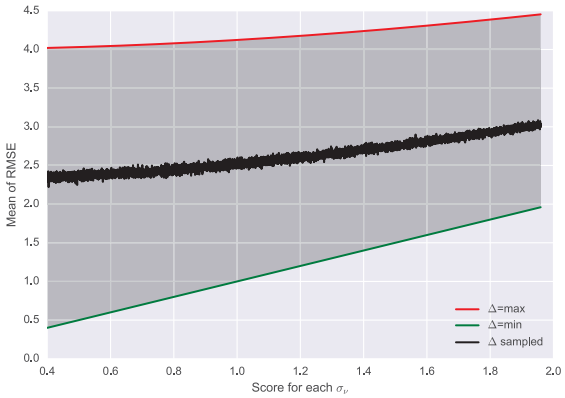
Briefly restated: In general, the mean of the employed quality metric is mainly determined by the deviations themselves. So far, this is a very good sign for the research of recommender systems, because this proves the validity of the RMSE, i.e. it indeed measures what it is supposed to – however with notable uncertainty. This uncertainty receives the influence of up to 300% for well-operating systems, i.e. those systems with small deviations between predicted and real user behaviour. Technically spoken: the better a system becomes, the more impact is given to random fluctuations such as comprised by human uncertainty. This is certainly not optimal for the current research seeking accuracy optimisation and has so far only scarcely been considered in the latest



(a) μ with respect to N



(b) μ with respect to Δ_ν

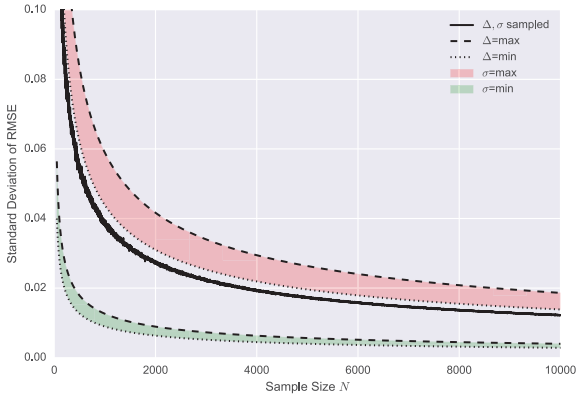


(c) μ with respect to σ_ν^2

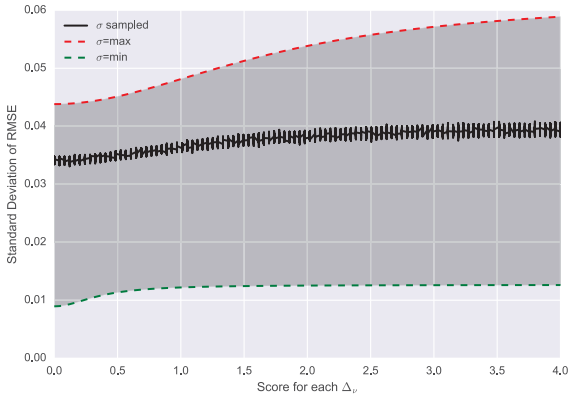
Figure 4.7: Sensitivity analysis for the expectation μ of the RMSE distribution

reports (cf. Enríquez et al., 2019). It is also striking that the RMSE, as small as it becomes, does never vanish. This proves that the mere existence of human uncertainty induces an offset, i.e. an RMSE score that cannot be fallen below. The existence of such a barrier has already been predicted in Herlocker et al. (2004) and is denoted as the magic barrier. For the RMSE in particular, this barrier has been theoretically estimated in Said et al. (2012).

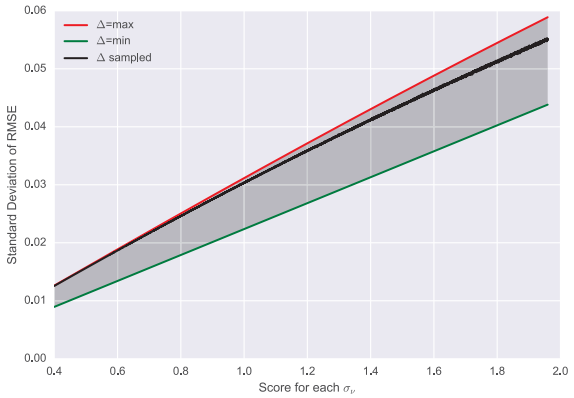
Impact on the Metric's Variance. Figure 4.8 depicts the sensitivity analysis for the RMSE's variance σ^2 in correspondence with variation of N, Δ_ν and σ_ν^2. In Subfigure 4.8a, the big impact of human uncertainty can be recognised (a large range between the coloured bands). Although the deviations between prediction and action also have an impact on the metric's variance, it is relatively weak and is dependent on human uncertainty itself. That is, the impact of deviations is poor (width of the green/lower band) for a small uncertainty, but it can be amplified by large uncertainties (width of the red/upper band). The most striking dependency of the metric's variance is the dependence on the number N of user-item pairs. On the one hand, it is surprising that the precision of a particular metric gains from adding more data with additional uncertainty. On the other hand, it is known from Eq. 4.21 that the variance scales with $1/(2N)$. This means that one yields a gain in precision for larger data sets very quickly. This also means that the increase in precision for even larger data sets rapidly fades. The decrease in the variance (i.e. the gain in precision) up to $N = 3\,000$ is tremendous (-133%). Thereafter, a further increase of data no longer leads to such a remarkable precision gain anymore. This finding may hold consequences on the economics of smaller studies (e.g. testing of new interfaces) since there is a point from which on additional participants will cost money but will not bring much benefit. This fast convergence means that one still has to deal with the impact of human uncertainty for big data. Subfigure 4.8b depicts the influence of the deviations. There is only a weak dependency to be spotted (borders are approximately straight lines with vanishing slope). For example, the magnitude of these deviations can increase the variance of the RMSE from 0.045 to 0.06 (+33%, difference of red/upper curve representing high human uncertainty) or from 0.01 to 0.012 (+20%, difference of green/lower curve representing low human uncertainty). In contrast, the human uncertainty itself may impact the variance much stronger (+300%, width of grey band). Subfigure 4.8c demonstrates the impact of human uncertainty. Linear growth of the variance can be seen in dependency



(a) σ with respect to N



(b) σ with respect to Δ_ν



(c) σ with respect to σ_ν^2

Figure 4.8: Sensitivity analysis for the variance σ^2 of the RMSE distribution

of human uncertainty and is amplified for large deviations (slope of the red/upper line compared to the green/lower line). At the same time, human uncertainty can increase the variance of the RMSE tremendously (difference in height of the coloured/outer lines) in comparison to the deviations (width of grey band).

In summary, the variance of the RMSE is affected by deviations (small impact), by human uncertainty (large impact) as well as by the number of user-item pairs (enormous impact). This might be a very good sign for accuracy-driven research: Since human uncertainty is unlikely to be improved, one possible way of dealing with its impact is simply to use big data. However, this way of thinking works only within certain limits, as the precision gain itself quickly decreases with an increasing amount of additional data. Here one has to find the golden mean between the necessary precision and monetary expense induced by more data. It will be demonstrated in later sections that even for large data records such as the Netflix Prize, there is yet still a considerable variance that corresponds to different RMSE distributions. The consequence of this non-vanishing variance is that it induces a probability of error whenever a ranking of systems is built with respect to a particular metric. This has to be investigated more closely.

4.3 Misjudgements in Comparative Evaluations

Research question B3 is about the impact of uncertainty on the comparative assessment of recommender systems when using accuracy metrics. It has already been reported in Sec. 1.2, that system rankings can not be considered as absolutely reliable anymore. This phenomenon will now be described mathematically. Considering metrics Z_1 and Z_2 as random variables, their distributions may possess a significant intersection. This means that each ranking order $Z_1 < Z_2$ and $Z_2 < Z_1$ is possible, however with different probabilities of occurrence. Consequently, no matter what ranking is finally chosen, there is always a probability of error for this decision. Now that the RMSE distribution can be determined on the fundament of uncertain user feedback, possible intersections of two RMSE's can be discussed along with induced ranking errors.

For a proper mathematical description of this error probability, let $Z_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $Z_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ be two random variables representing arbitrary metric results for two different recommender systems. The assumption of normality is justified by previous considerations, i.e. the density of a composed quantity quickly converges into a Gaussian for a large number of user-item pairs. Additionally, an auxiliary variable is

defined as follows:

$$W := (Z_1 - Z_2) \sim \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right) \tag{4.26}$$

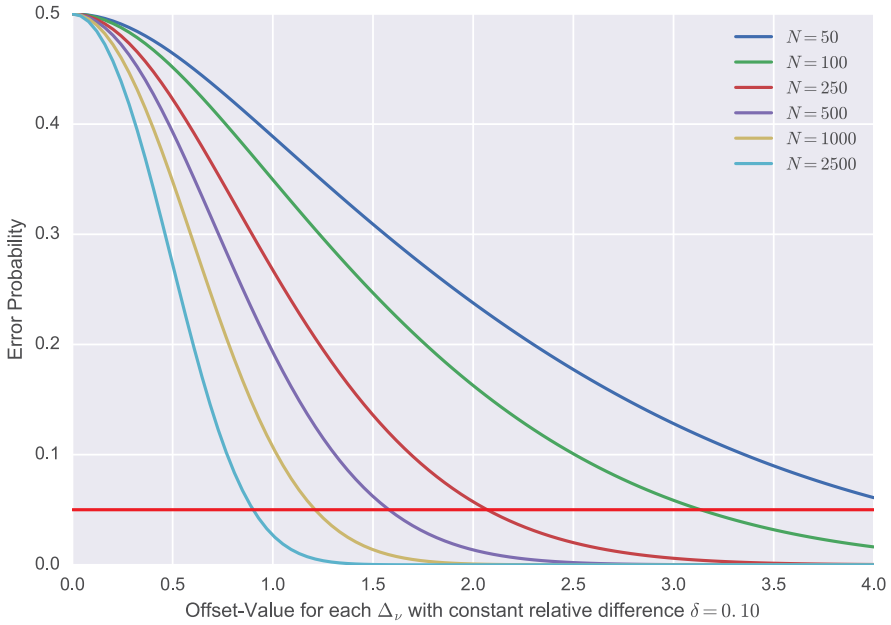
The most intuitive way to build a ranking of two distributions is to compare their expected values. If $\mu_1 < \mu_2$, then approach 1 can be considered to be better than approach 2. Due to the non-vanishing variance, this decision may be subject to an error which occurs with a probability of

$$P(Z_1 \geq Z_2) = P(W \geq 0) = 1 - F_W(0), \tag{4.27}$$

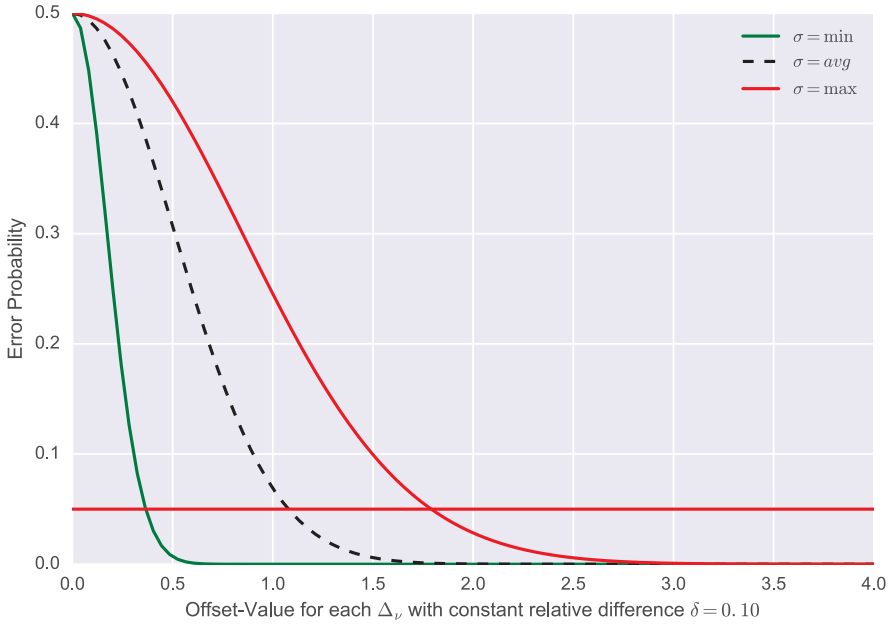
where F_W is the cumulative distribution function of W. Since W is normally distributed, it can be represented by a transformation of the standard-normal distribution and therefore F_W can be expressed in terms of the standard-normal cumulative distribution function Φ. This finally leads to

$$P(Z_1 \geq Z_2) = \Phi\left(\frac{\mu_1 - \mu_2}{(\sigma_1^2 + \sigma_2^2)^{1/2}}\right). \tag{4.28}$$

This error probability naturally accounts for the means of both metric distributions (impacted by the recommender accuracy and human uncertainty) as well as their variances (impacted by the amount of data). For an investigation of these quantities' impact, two recommender systems are defined by determining the average local prediction quality Δ in such a way that one system's mean is constantly 10% better than the mean of the other system. This constant difference has been chosen following the Netflix Prize where improvements had to be at least 10% according to a reference. Other choices for this constant difference will mainly produce equivalent results. Subfigure 4.9a depicts the error probability for two RMSE distributions concerning the average deviation of prediction and rating together with the impact of the amount N of data. Here, $P = 0.05$ has been chosen as the borderline of distinguishability (in accordance with the significance levels for hypothesis testing). Astonishingly, the error curves are no constant lines, meaning that the distinguishability is different for two well-performing recommender systems than for poor-performing ones, even though the difference of the RMSE's mean remains the same. Moreover, it can be observed that for $N = 50$ no system can be distinguished from another without making an error in less than 5% of all cases (blue/upper line). For $N = 100$, for example, only poor systems ($\Delta > 3.2$) can be distinguished with an error probability of less than 5%. Well-performing systems



(a) family of error probability curves parametrised by N



(b) family of error probability curves parametrised by σ_ν

Figure 4.9: Sensitivity analysis for the error probability

($\Delta < 1$), on the other hand, can only be sufficiently distinguished with at least 2 500 user-item pairs. It is shown, that the influence of the data size has got as much influence on the distinguishability (i.e. error probability) as on the metric's variance itself. Moreover, the convergence of the variance for $N \to \infty$ can be seen, represented by the fading distance of the curves relative to one another for increasing size of data.

Subfigure 4.9b depicts the error probability for two RMSE distributions concerning the average deviation together with the impact of human uncertainty for a fixed data size of $N = 1\,000$. It turns out that human uncertainty can significantly shift the borderline of sufficient distinguishability. For a low uncertainty (green/lower curve), extremely well-performing systems ($\Delta < 0.5$) can be brought into a ranking with only a low probability of error. For high uncertainty (red/upper curve), only medium-quality systems ($\Delta \approx 2$) can be distinguished through low-error rankings.

In conclusion, the impact of human uncertainty on the distinguishability is remarkable but also gives the impression of not being as striking as the impact of the data size. Even with big data, one cannot completely get rid of this problem as the gain of distinguishability is fading for additional data. Another surprising fact is that two systems with a relative accuracy difference of 10% can be distinguished and put into a ranking order only if these are low-quality systems. On the other hand, such systems cannot be distinguished anymore if they are of high quality, although the relative difference is still 10%. This indicates that the better recommender systems become, the more additional improvement is required to recognise a superior system as such with statistical evidence.

Up to this point, these ranking errors have only been considered within a statistical theory or in a small experiment with simple recommender systems. However, the interesting question is whether this phenomenon can be found in more sophisticated prediction tasks and whether the use of big data minimises the chance of ranking errors. Therefore, this probabilistic framework is employed to discuss possible ranking errors for one of the largest recommender system competitions in recent years, namely the Netflix Prize (cf. Netflix Inc., nd). At this point, it appears to be challenging that Netflix did not collect any information about human uncertainty. However, for the size of Netflix's test record ($N = 2.8 \cdot 10^6$), this is not a problem at all since the RMSE's variance scales with $1/(2N)$ which is illustrated in Figure 4.8a. It can be seen that the specific magnitude of the metric's uncertainty is much more dependent on the size of

the data set rather than on the feedback uncertainty itself (compare heights of the grey area). Indeed, human uncertainty has been estimated for the Netflix Prize in three different ways and always produced the same result:

1. ML-fitting of human uncertainty based on the RETRAIN study provided an uncertainty distribution from which random draws were made to be associated with each rating of the Netflix record (see Fig. 3.7a).

2. Human uncertainty was randomly sampled from different distributions (e.g. uniform, triangular, normal) and associated with each rating of the Netflix record.

3. Having a 5-star scale, human uncertainty yields certain limitations. Associating the minimum or maximum uncertainty to each Netflix rating produces an interval in which the RMSE's variance is located.

With Eq. 4.20, each RMSE *score* in the Netflix leaderboard can then associated with the average prediction quality Δ that will produce the predefined mean $\mu = score$. Subsequently, σ^2 can be computed using Eq. 4.21 and Δ. This approach will transform each RMSE *score* into a random quantity $Z \sim \mathcal{N}(\mu, \sigma^2)$. For these distributions, one can estimate the error probabilities for pairwise rankings using Eq. 4.28.

Table 4.1 lists the error probabilities for the paired rankings within the leaderboard of the Netflix Prize. The most likely error probabilities can be found in Subtable 4.1a. For example, the decision that the third-placed algorithm is better than the fourth-placed algorithm is afflicted with an error probability of 25%, i.e. these positions would swap in one of four repetitions. In the same way, position four to six hold very high probabilities of error, so that a permutation of these positions is likely to occur in the light of human uncertainty. An analysis of the positions nine to twelve appears to be much more remarkable since their errors are approaching the maximum value, i.e. the entry into the Top 10 (providing glory and honour) is dependent solely on chance and not on model-based prediction quality. These estimations are so far only based on point estimated for the feedback variances. However, it has been explained in Sec. 3.4 that these parameters can only be located within confidence intervals. Accordingly, the error computation could also be done by using the upper bound of the interval in Eq. 3.4 to obtain the worst case of possible ranking errors. These scores can be seen in Subtable 4.1b. For this worst case (which is not very likely but still possible), substantially higher error probabilities can be observed. The decision for recommender

	$R_{1/2}$	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}
$R_{1/2}$	.50	.04	.01	.00	.00	.00	.00	.00	.00	.00	.00
R_3		.50	**.24**	.14	.08	.01	.00	.00	.00	.00	.00
R_4			.50	**.36**	**.24**	.06	.00	.00	.00	.00	.00
R_5				.50	**.36**	.12	.01	.00	.00	.00	.00
R_6					.50	**.20**	.02	.00	.00	.00	.00
R_7						.50	.10	.01	.00	.00	.00
R_8							.50	.12	.10	.10	.08
R_9								.50	**.45**	**.45**	**.41**
R_{10}									.50	**.50**	**.45**
R_{11}										.50	**.45**
R_{12}											.50

(a) most likely case of error probabilities for pairwise rankings

	$R_{1/2}$	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}
$R_{1/2}$	.50	.14	.07	.04	.03	.01	.00	.00	.00	.00	.00
R_3		.50	**.34**	**.26**	**.20**	.09	.02	.00	.00	.00	.00
R_4			.50	**.42**	**.34**	.18	.04	.01	.01	.01	.01
R_5				.50	**.42**	**.24**	.07	.01	.01	.01	.01
R_6					.50	**.31**	.10	.02	.02	.02	.02
R_7						.50	**.22**	.07	.06	.06	.05
R_8							.50	**.24**	**.22**	**.22**	**.20**
R_9								.50	**.47**	**.47**	**.44**
R_{10}									.50	**.50**	**.47**
R_{11}										.50	**.47**
R_{12}											.50

(b) worst case of error probabilities for pairwise rankings

Table 4.1: Estimated error probabilities for pairwise rankings within the leaderboard of the Netflix Prize when human uncertainty from the RETRAIN record is assumed

"1/2" (both systems obtained the same RMSE score) to be better than recommender "3" is subject to a chance of error of about 15%, i.e. this ranking is incorrect in 3 out of 20 cases. This prominent example demonstrates that the supposed verification of systems, even under the protection of big data, can sometimes be deceptive and needs to be rethought against the background of human uncertainty.

4.4 Limitations of System Improvements

A noteworthy property of error probabilities is their functional dependency on the prediction quality: Two recommender systems with a constant quality difference of 10% can be distinguished in terms of their RMSEs, insofar they are poor systems (see Fig. 4.9). The better two systems operate on a data set, the more indistinguishable they become while maintaining a constant difference in quality. This theoretical property certainly has an impact on the reality of evaluating systems. Considering the Netflix Prize, already Koren stated

> "None of the 3 400 teams actively involved in the Netflix Prize competition could reach, as of 20 months into the competition, lower RMSE levels, despite the big incentive of winning a $1M Grand Prize. Thus, the range of attainable RMSEs is seemingly compressed." (Koren, 2008, p. 432)

and has thus recognised the manifestation for natural limitations of prediction accuracy. If this property (i.e. the better the overall performance, the more indistinguishable) is transferred to the optimisation process of a single system (distinguishability of an improved product to its predecessor version), then such a high accuracy may be achieved that the remaining potential of further improvements is just smaller than the required quality difference for a sound detection. In other words, from a certain quality of systems, further improvements can no longer be identified without significant doubt. Hence, there is only one equivalence class of excellent systems. This specific limit is commonly known in the literature as the magic barrier (cf. Said et al., 2012).

This magic barrier can be described as the minimum of a metric which results for an optimal recommender system given human uncertainty. For the RMSE in particular, the magic barrier results from setting $\Delta_\nu := \mu_\nu - \pi_\nu = 0$ for each user-item pair, i.e. the difference between prediction and rating is zero on average. Hence, the expected

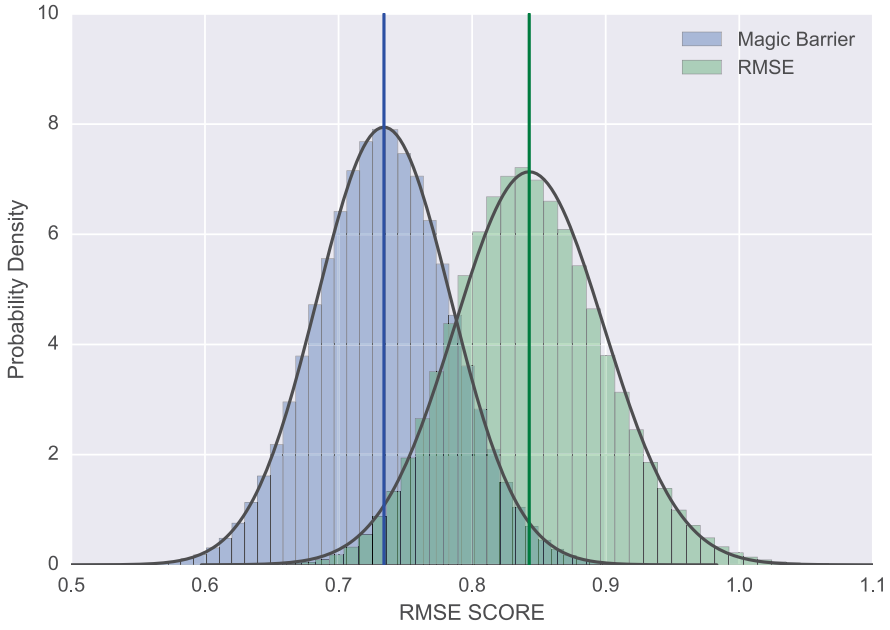


Figure 4.10: Example of an RMSE distribution possibly interfering with the magic barrier. The chance of inference is $P(\mathcal{MB} > \text{RMSE}) \approx 0.33$.

value in Eq. 4.20 becomes

$$\mathbb{E}[\mathcal{MB}] \approx \sqrt{\frac{1}{N}\sum_{\nu}\sigma_{\nu}^{2}}, \tag{4.29}$$

showing that the magic barrier can be located near the expected value of human uncertainty which is consistent with the assumption of Hill et al. that a recommender system can never predict more accurately than the variance of the user responses considered (cf. Hill et al., 1995, p. 200). Thus, the magic barrier demonstrates the uncertainty bias for optimal systems and can thus be understood as a background noise on a particular metric. Consequently, all rankings on the basis of smaller scores are completely random and the associated systems become completely indistinguishable. However, it is not a good idea to understand this limit as a sharply localised value since human uncertainty additionally induces a variance

$$\mathbb{V}[\mathcal{MB}] \approx \frac{1}{2\,N}\,\frac{\sum_{\nu}\sigma_{\nu}^{4}}{\sum_{\nu}\sigma_{\nu}^{2}} \tag{4.30}$$

according to Eq. 4.21. This means that systems can be interfered by human uncertainty

even if the expectation of a metric's distribution has not yet fallen below the magic barrier. Vice versa, systems with expectation below this specific limit need not necessarily be interfered by human uncertainty. Figure 4.10 illustrates the interference of the magic barrier with an example recommender system. Although the RMSE mean is still above the mean of the magic barrier, there is a significant probability that the RMSE outcome is already affected. This probability $P(\mathcal{MB} > \text{RMSE})$ of interference is simply the error probability from Eq. 4.28 with $\mu_1 = \mathbb{E}[\mathcal{MB}]$ and $\sigma_1^2 = \mathbb{V}[\mathcal{MB}]$. At this point, it can be seen once more that a dichotomous decision criterion (better or worse) is not adequate for a probabilistic understanding of human behaviour and that all possibilities must be considered along with their probabilities (i.e. how likely is it that a system can still be improved and what risk of error is still acceptable?).

As a realistic application, the magic barrier can be calculated for the Netflix Prize. By estimating the user variances as already done above, the magic barrier can be determined to $\mathcal{MB} \sim \mathcal{N}(0.6687, 0.0007)$. Using the contest winner as a reference, the interference probability vanishes, i.e. it can be assumed that the magic barrier has not yet been reached. To be more precise, there is still a potential for about 20% of improvement when taking the winner as a reference. The existence of this barrier is particularly relevant when monetary decisions are made regarding the optimisation of recommender systems, e.g. employee's efforts or financial resources are invested in an alleged optimisation process but the results are purely random which remains unnoticed. For Netflix, the magic barrier may soon be reached within the next few years. From this point, further costly improvements in terms of optimising the RMSE would simply make no sense anymore. With this spirit, a differentiated analysis of human uncertainty according to metrologic models appears as an essential component of future comparative studies on recommender systems when user feedback is considered.

An extension of this magic barrier is given when the predictor is assumed to be a random variable as well. This is the case, e.g. for supervised learning where human beings carry out classification which impacts the prediction of future user behaviour. In this case, an optimal system would assign the same distribution to the predictor that will be provided by subsequent user feedback, i.e. a human being is always the best predictor for himself. Let therefore be $\mathfrak{F}_\nu \sim \mathcal{N}(\mu_\nu, \sigma_\nu^2)$ and $\pi_\nu \sim \mathcal{N}(\mu_\nu, \sigma_\nu^2)$ respectively. The difference is hence given as

$$Y_\nu := \mathfrak{F}_\nu - \pi_\nu \sim \mathcal{N}(0, 2\sigma_\nu^2) \tag{4.31}$$

and can be written as $Y_\nu := \sqrt{2}\sigma_\nu \mathcal{I}$ where $\mathcal{I} \sim \mathcal{N}(0,1)$. With this substitution, it follows that

$$\mathbb{E}[Y_\nu^2] = \mathbb{E}[(\sqrt{2}\sigma_\nu \mathcal{I})^2] = \mathbb{E}[2\sigma_\nu^2 \mathcal{I}^2] = 2\sigma_\nu^2 \mathbb{E}[\mathcal{I}^2] = 2\sigma_\nu^2 \tag{4.32}$$

$$\mathbb{V}[Y_\nu^2] = \mathbb{V}[(\sqrt{2}\sigma_\nu \mathcal{I})^2] = \mathbb{V}[2\sigma_\nu^2 \mathcal{I}^2] = 4\sigma_\nu^4 \mathbb{V}[\mathcal{I}^2] = 8\sigma_\nu^4 \tag{4.33}$$

and the MSE's statistics are hence given by

$$\mathbb{E}[MSE] = \frac{1}{N}\sum_\nu \mathbb{E}[Y_\nu^2] = \frac{2}{N}\sum_\nu \sigma_\nu^2 \tag{4.34}$$

$$\mathbb{V}[MSE] = \frac{1}{N^2}\sum_\nu \mathbb{V}[Y_\nu^2] = \frac{8}{N^2}\sum_\nu \sigma_\nu^4 \tag{4.35}$$

Finally, when using the Gaussian Error Propagation, the RMSE for optimal prediction (which will be denoted as the human barrier in this thesis) is given by

$$\mathcal{HB} \sim \mathcal{N}\left(\sqrt{2}\cdot\sqrt{\frac{1}{N}\sum_\nu \sigma_\nu^2}\ ,\ \frac{1}{N}\cdot\frac{\sum_\nu \sigma_\nu^4}{\sum_\nu \sigma_\nu^2}\right). \tag{4.36}$$

It can be seen the expectation of the human barrier is $\sqrt{2}$ times larger than the expectation of the magic barrier whereas the variance of the human barrier is twice as much as the variance of the magic barrier. This means that for supervised learning, i.e. when the predictor is also subject to human uncertainty, the limitation of statistically sound distinction of two systems is much more restrictive. In other words, the possibilities for further improvements are much more limited. Another scenario in which the predictor is also a random variable is given when the recommender system uses human uncertainty to extend the predictor score. This case will be discussed in more detail in Ch. 5.

4.5 Chapter Summary

Based on the latest research in metrology, the uncertainty of quantities propagates with respect to a specific mathematical model when composed quantities are computed. In a probabilistic sense, the composed quantity is distributed by a probability density which emerges as a convolution of all arguments' densities. Typical approaches to determine resulting distributions are analytical derivations, Monte-Carlo simulations as well as the Gaussian Error Propagation. Transferred to the comparative assessment of personalisation systems, the results of well-established accuracy metrics turn out to be

distributions rather than single scores. It can be assumed that it is not uncommon for two such distributions to have an intersection, resulting in a probability of error when creating a ranking. This error can be thought like this: Although a ranking according to the expected values may imply a system A to be better than system B, it does (more or less frequently) occur that system B even outperforms system A when considering only single draws from underlying distributions. The frequency for this ranking inversion can be seen as a probability of error that is associated with each ranking. When this error probability is too high, there is no sufficient evidence for any possible ranking order and both systems become undistinguishable by means of a ranking. These limitations can be thought of as uncertainty-induced barriers. In conclusion, the presented results reveal that human uncertainty has a great impact on the comparative assessment of recommender systems. They hence justify an even more differentiated consideration of user data. Future comparative evaluations of recommender systems may hence require to account for human uncertainty. Possible strategies for doing so are discussed in the next chapter.

5 | Possible Solutions

The purpose of this chapter is to present a brief overview of existing strategies for dealing with human uncertainty (subgoal C). The eponymous chapter published in Jasberg and Sizov (2019) provides the basis for this, although it has received an extensive revision for this dissertation. In addition to its contents, single sentences have also been taken verbatim from this work. Moreover, the concept of the so-called sRMSE was first presented in Jasberg and Sizov (2017a) and was then published in its analytical form in Jasberg and Sizov (2018a) from where it was copied verbatim.

5.1 Statistically Sound Improvement Detection

One possible solution to improve the handling of uncertainty is to examine its concrete impact for a given scenario of comparative assessment in reality. Whether two systems are eventually distinguishable or not deserves to be determined by hypothesis testing due to the statistical nature of the phenomenon.

Such a test has already been illustrated in the last chapter when considering the Netflix Prize. At this point, the above-mentioned test is to be explained in general and an applicable short form is to be derived: Let z_1 and z_2 be two realisations from Z_1 and Z_2 representing the results of two distinct systems' accuracy metrics. Both random variables are then determined by $Z_k \sim \mathcal{N}(z_k, \sigma^2)$ where σ is obtained from uncertainty propagation based on the score itself and the size of the data set. For

smaller improvements of a system where distinguishability matters, the dependence of σ^2 on the score itself can be neglected. A rough estimate can then be taken from Figure 4.8a and has been proven to be reliable for big data scenarios in the previous chapter. Moreover, it can be assumed that the standard deviation is equal for each Z_k as long as all systems operate on the same data set. The relationship $Z_1 < Z_2$ can then be assumed to hold with a significance level of α if the opposite case occurs with a probability $P(Z_1 \geq Z_2) \leq \alpha$ (type I error). By rewriting $z_1 = z_2 + h$ with $h \geq 0$ and using the identity from Eq. 4.28, it follows that

$$P(Z_1 \geq Z_2) = \Phi\left(\frac{z_1 - z_2}{\sqrt{2}\sigma}\right) = \Phi\left(\frac{-h}{\sqrt{2}\sigma}\right) \leq \alpha \quad \Leftrightarrow \quad h > -\sqrt{2}\sigma\,\Phi^{-1}(\alpha). \tag{5.1}$$

This means that both systems are distinguishable with a significance level of α if the difference of both metric scores is at least $-\sqrt{2}\sigma\Phi^{-1}(\alpha)$ where Φ^{-1} is the inverse cumulative distribution function for the standard-normal distribution. Considering the common case of $\alpha = 0.05$, the difference has to be $h > 2.33\,\sigma$ for a significant detection.

Another approach is to shift the magic barrier along the x-axis of metric scores and test whether it is possible to cover both metric results z_1 and z_2 within the 95% confidence interval I_{95}. Heuristically explained, two metric scores can not be distinguished by means of the relation $Z_1 < Z_2$ if there exists a single solution (i.e. a probability density) which can explain both outcomes (i.e. draws) with sufficient significance. In other words: For $\alpha = 0.05$, both scores z_1 and $z_2 = z_1 + h$ must not exceed the I_{95} interval which has the length $\ell(I_{95}) \approx 4\sigma$. This implies that both scores must differ by at least $h > 4\sigma$ to avoid being explained by a single model. For an arbitrary level of significance α, one has to find $a > 0$ so that $I_{1-\alpha} = [\mu - a\sigma\,;\,\mu + a\sigma]$. Since the shifted magic barrier X is assumed to be normally distributed, it follows that

$$\begin{aligned} P(\mu - a\sigma \leq X \leq \mu + a\sigma) &= F_X(\mu + a\sigma) - F_X(\mu - a\sigma) \\ &= \Phi(a) - \Phi(-a) \\ &= \operatorname{erf}(a/\sqrt{2}) \end{aligned} \tag{5.2}$$

Accordingly, the $I_{1-\alpha}$ interval covers the probability mass $1 - \alpha = \operatorname{erf}(a/\sqrt{2})$ which implies $a = \sqrt{2}\,\operatorname{erf}^{-1}(1-\alpha)$. For a statistically sound distinguishability, both scores z_1 and $z_2 = z_1 + h$ must hence differ by $h > \ell(I_{1-\alpha}) = 2a\sigma = 2\sqrt{2}\,\sigma\,\operatorname{erf}^{-1}(1-\alpha)$.

It is noteworthy that both tests exhibit a different sensitivity which is due to their structural design. The first approach works by considering the intersection of two

densities and the second approach tries to find a translation of one density so that both scores are within a certain confidence interval. For the special case of $\alpha = 0.05$, the first approach can detect even smaller differences than the second approach. In particular, differences between $2.33\sigma < h < 4\sigma$ can be detected by the first approach but not by the second approach. On the other hand, the second approach avoids the assumption that given metric scores represent the mean values of two distinct distributions. Overall, both approaches involve the challenge of estimating the standard deviation of the metric distribution which is computationally cumbersome if human uncertainty has not been measured (cf. Netflix Prize considerations). Accordingly, these test methods can only represent a mere estimate of potential distinguishability in the absence of uncertainty information. They should not be misunderstood as a substitute for the need for a proper uncertainty measurement for the individual use case.

5.2 Improvement by Subsequent Uncertainty Reduction

Another approach to improve the negative aspects of human uncertainty is to artificially reduce it through additional processing steps.

Pre-processing steps. A prominent example of de-noising algorithms has been introduced in Amatriain et al. (2009b) and has already been mentioned in Ch. 2. The test setup included three different rating trials, in which 118 users had to rate 100 film titles on a scale of 1 to 5 stars using a web interface. Human uncertainty was thereby gathered via re-rating: The second rating trial was conducted 24 hours after the first trial and the third rating trial was conducted 15 days after the second trial (cf. Amatriain et al., 2009a, pp. 249–250). It has been found that "the calculated RMSE between different trials ranged between 0.557 and 0.8156 [sic!]" (Amatriain et al., 2009a, p. 257). To tackle this issue, Amatriain et al. introduced a de-noising algorithm which recursively replaces repeated ratings with large scattering (i.e. above an arbitrary threshold) by artificial ratings with smaller scattering (cf. Amatriain et al., 2009b, p. 175). The authors achieved an improvement above 14% compared to the original RMSE (cf. Amatriain et al., 2009b, p. 180). Heuristically, human uncertainty is artificially limited by manually replacing real data with smaller (but fictive) deviations. Of course, this reduces the uncertainty of the RMSE and thus the corresponding bias of its mean, leading to better RMSE scores. The main issue with this approach is

that data is subsequently manipulated (without a real basis) until the desired result is achieved. If an uncertain quantity is measured and then the uncertainty is replaced by a smaller one, the true extent of uncertainty remains the same while not being properly detectable any longer. In other words, one just considers other data than it exists in reality. For this reason, it can be assumed that the stated improvement is only virtual and does not reflect reality.

Partially omitting data. Another idea is to modify existing accuracy metrics to make them sensitive to human uncertainty. In other words, each time a rating is compared with a model-based prediction, it is necessary to examine whether the observed deviations are significant or whether they are simply due to human uncertainty. This involves dividing the set of all deviations into two subsets: One subset contains all deviations around a predictor π_ν which can be considered as being caused by human uncertainty. The other subset contains all deviations whose extent cannot be explained by this uncertainty and which therefore appear to be induced by the prediction model itself. In this case, it seems practicable to calculate the quality metric by considering only those deviations that are related to the algorithm and not to human uncertainty.

This idea will be exemplified using the RMSE: Following the argumentation above, it seems appropriate to employ statistical hypothesis testing to decide whether a realisation $\mathfrak{f}_\nu$ of a feedback distribution $\mathfrak{F}_\nu$ is equal to a model-based prediction π_ν or not. In mathematical notation, one has to test $H_0 \colon x_\nu = \pi_\nu$ vs. $H_1 \colon x_\nu \neq \pi_\nu$ at a given significance level of α. For known density functions, the region of rejection can be constructed as the complement $I_{1-\alpha}^{\complement}$ of $I_{1-\alpha} = [\pi_\nu - a \,;\, \pi_\nu + a]$ where $a > 0$ is chosen to satisfy

$$\int_{\pi_\nu - a}^{\pi_\nu + a} f_{\mathfrak{F}_\nu}(t)\,\mathrm{d}t = 1 - \alpha. \tag{5.3}$$

The assumption of normality for each user feedback $\mathfrak{F}_\nu$ allows further simplification according to Eq. 5.2 from which follows that $a = \sqrt{2}\,\mathrm{erf}^{-1}(1-\alpha)$. While the traditional RMSE density constitutes as a convolution of all $f_{\mathfrak{F}_\nu}(t)$ through a mathematical model, the density of the modified RMSE emerges as a convolution of restrictions

$$f_{\mathfrak{F}_\nu}\big|_{I_{1-\alpha}^{\complement}}(t) := \mathbb{I}_{I_{1-\alpha}^{\complement}}(t) \cdot f_{\mathfrak{F}_\nu}(t) \tag{5.4}$$

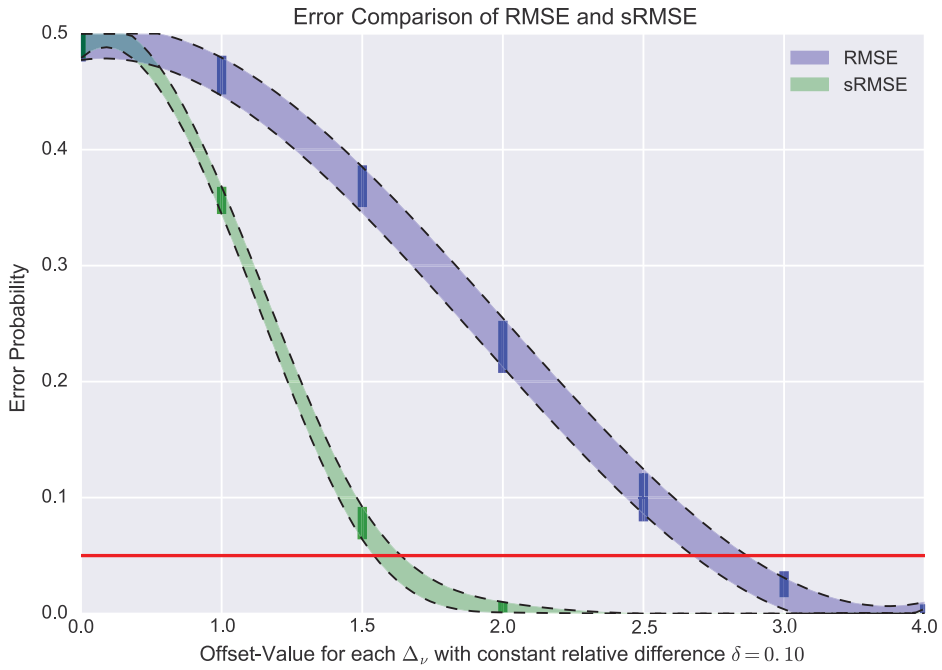


Figure 5.1: Comparison of error probabilities computed with the RMSE and sRMSE

where $\mathbb{I}$ represents the indicator function

$$\mathbb{I}_A(t) := \begin{cases} 1 & \text{if } t \in A \\ 0 & \text{else} \end{cases} \tag{5.5}$$

Since being similar to the traditional RMSE, the modified metric will be referred to as the significant RMSE (sRMSE). The sRMSE guarantees a comparison between different systems with much lower probability of error. This is achieved by excluding the stabilising centre of all feedback distributions. More precisely, because the RMSE amplifies the remaining extremes by its quadratic term (cf. Eq. 4.1), the resulting distributions differ rapidly with increasing false predictions.

Using this algorithm along with a confidence level of $\alpha = 0.05$, the respective error probabilities have been computed for two fictitious systems with an arbitrary average prediction quality Δ along with a constant accuracy difference of 10%. This is exactly the same simulation setup as previously employed to generate Fig. 4.9. The results of this new simulation are depicted in Fig. 5.1. It is obvious that the error curve

drops significantly faster for the sRMSE than it does for the RMSE and thus falls below the critical limit of 5% much earlier. This means that by using the new metric sRMSE, better systems can still be distinguished significantly whereas an evaluation using the RMSE already produces indistinguishability. The discrimination between deviations that can be explained by human uncertainty and those that can not be explained, respectively, may be understood as some kind of sensitisation of a metric to this phenomenon. This may indicate that human-like metrics can be regarded as fruitful for future system evaluation. However, this sensitivity is based on computing accuracy metrics with those 5% of deviations that are large enough. In other words, this approach denies 95% of the available data which may not always be in accordance with real user behaviour. Another shortcoming is that all problems of distinguishability are only diminished but still existent.

5.3 Uncertainty as Information Source

The methods from the last section essentially considered human uncertainty as undesirable and modelled it with the objective of elimination. However, the possibilities of extracting additional information have only insufficiently been considered so far. In this section, new ideas are presented that aim to benefit from human uncertainty.

Clustering Uncertainty. Sizov notes that systems nowadays describe the behaviour of an entire community well but cannot explain individual behaviour at the same time (cf. Sizov, 2017b, p. 869). Against this background, the author presents an approach with the aim "to collect additional information about individual users [i.e. uncertainty information in this case], and to include gained knowledge into adjusted mixture models" (Sizov, 2017b, p. 876). Such a mixture model can be defined as

$$f(x) = \sum_{k=1}^{K} p_k \cdot f_k(x, \theta_k) \tag{5.6}$$

where K is the number of components in the mixture, f_k represents the individual group density parametrised by θ_k, and the coefficients $0 \leq p_k \leq 1$ can be considered as component probabilities satisfying $\sum_{k=1}^{K} p_k = 1$ (cf. Sizov, 2017b, p. 871). The idea behind using mixture models is that for each user behaviour that cannot be explained by a global distribution (i.e. a global model in the terminology of Sizov), another distribution can be added. This will theoretically lead to a family of different distributions, which

in sum can describe any individual user behaviour. This argumentation is supported by Bishop who explains:

> "By using a sufficient number of Gaussians [as group density functions], and by adjusting their means and covariances as well as the coefficients in the linear combination, almost any continuous density can be approximated to arbitrary accuracy." (Bishop, 2006, p. 111)

In principle, the problem of user explanation can be solved by adding another component to this mixture model for each user action, which then again strongly increases the model complexity. To find an optimal trade-off between complexity and fitting quality, so-called information criteria are usually employed (cf. Bishop, 2006, pp. 32–33). Following Sizov, the commonly employed "information criteria do not account for micro level of individual user behaviour and thus may under-estimate the necessary number of components in favor of a less complex solution" (Sizov, 2017b, p. 873). The author's solution is to 'humanise' the information criteria (as a function of K) by multiplying them by the relative frequency of observations that would not be explained by a K-component model. This is exemplified for the Deviance Information Criterion (DIC) by using the transformation

$$\mathrm{HDIC}(K) := \mathrm{DIC}(K) \cdot h_{reject}(K) \tag{5.7}$$

in order to obtain the Human Deviance Information Criterion (cf. Sizov, 2017b, p. 873). At this point, $h_{reject}(K)$ represents "the fraction of individual per-user observation [...] that falsify the global model in the sense of common hypothesis testing procedures (such as KS-test)" (Sizov, 2017b, p. 873). On a self-collected data set including uncertainty information, Sizov could show that this new model indeed creates a new (corrected) trade-off between complexity and individual user explanation: While the traditional DIC suggests a global model with only one component (which does not explain 38% of all observations), the HDIC points to 4 or 5 components (which reduces the failure of explanation to 3% or 0%, respectively), depending on the chosen significance level of the underlying hypothesis test (cf. Sizov, 2017b, p. 875). Moreover, the identified groups represent interpretable characteristics in terms of rating behaviour and uncertainty:

> "By inspecting the model, we observe a significant fraction of 'uncertain' users with high variance and ratings in the middle of the scale (component 3). At the same time, two further fractions of users tend to assign 'mostly negative' or 'mostly positive' scores with reasonably high variance (components 1 and 5). In addition, the model highlights the presence of small fractions of 'focused' users, with low decision variance and 'below average' vs. 'above average' rating behaviour (components 2 and 4)." (Sizov, 2017b, p. 875)

Briefly summarised, the consideration of uncertainty allowed a comparison of the resulting feedback distributions with a global model. This allowed overcoming the situation that too simple models were constructed which did not reflect the observations of user behaviour. Also, frequently recurring behaviour patterns have been identified which may be used to modify future recommendations. These results indicate that the consideration of human uncertainty not only causes negative effects but is also a useful source of information which can be easily exploited in systems.

Clustering local Barriers: Another idea to make use of human uncertainty was introduced in Said and Bellogín (2018). The essence is to calculate the coherence of a user within an attribute space (e.g. a film genre). For example, a coherent user would rate all horror films equally while an incoherent user would exhibit more variation within this genre. For this purpose, the author conducted an online experiment with 308 users who repeatedly rated 2 329 films in two separate trials (cf. Said and Bellogín, 2018, pp. 104–105). It has been found that the user coherence is correlated with the magic barrier and that this correlation can be used to discriminate between easy users and difficult ones (cf. Said and Bellogín, 2018, p. 97). By systematically composing the learning set and the test set with different proportions of simple and difficult users for a sample recommender system, it was found that it is possible "to build different training (and test) models in such a way that the error decreases for the easy users, i.e., to increase the accuracy of the recommender system" (Said and Bellogín, 2018, p. 117). The authors were able to show that the RMSE could be improved by 10 to 40%, depending on the number of difficult users in the training set (cf. Said and Bellogín, 2018, p. 121). At this point, it has been demonstrated that very good improvements can indeed be achieved by using a more human-like user model.

5.4 Chapter Summary

This chapter has shown that there are currently three basic ways of dealing with human uncertainty and its side effects: The first approach refers to the subsequent reduction of measured uncertainty. However, these methods are questionable as they no longer reflect the existing reality. A consideration of uncertainty is only supposedly carried out. The second approach no longer disguises human uncertainty but accepts its presence and investigates whether it affects a given use case when there is no particular uncertainty information available. Then again, these are only rough estimates and do not replace a proper uncertainty measurement. Accordingly, this is unlikely to be a viable solution for the future and motivates to explore further alternatives. The third approach is about turning uncertainty into a benefit by exploiting the additional information. The ideas presented are very promising and demonstrate that user model tuning offers great potential, both for prediction accuracy as well as for user insights with which predictions can be individually corrected. Nevertheless, these approaches are merely statistical and technical procedures that capture human nature only phenomenologically. This represents just an external perspective, so another solution might be to take the perspective from within. Human uncertainty from an inner perspective, i.e. on the basis of human cognition as represented in theories of psychology and neuroscience, should therefore be examined more closely. Indeed, a few machine learning techniques (e.g. NN, rNN, HTM, etc.) have benefited from neuroscience by adopting some of its foundational ideas. For the case of human uncertainty, a comparison between the results of the internal and the external perspective may perhaps provide new insights about the future design of recommender systems as well. However, this would require an adequate object of comparison which represents the internal perspective. Such an object of comparison will be developed in the next chapter.

6 | A Neuroscience Model of Human Uncertainty

The purpose of this chapter is to introduce an adequate cognitive model to (1) substantiate the assumption of a human-inherent origin of unreliable user feedback and (2) to discover potential benefits for user explanation and recommendation. The very first results of this research have been published in my work Jasberg and Sizov (2018b). From this publication, the description of the probabilistic population codes (Sec. 6.3) was taken almost verbatim. Likewise, a large proportion of verbatim text segments can be found in the description of the user model (Sec. 6.4) and the parameter boundaries (Sec. 6.5) respectively. However, the material has undergone an extensive linguistic revision for this dissertation and has been subject to substantial extensions.

6.1 Modelling Theory and Epistemology

Models are widely used in natural science, especially in physics. Accordingly, these models are in the focus of epistemological discussions in the respective scientific fields. In physics, for example, Kircher et al. define a model as a mental or a material object that is

used as a substitute for an original (cf. Kircher et al., 2009, pp. 732–733). It is by nature a simplification of the original or reality (cf. Kircher et al., 2009, pp. 741–743) and thus enables a mathematical description (cf. Kircher et al., 2009, p. 753). For this reason, the model must match some (but not all) features with the original (cf. Kircher et al., 2009, pp. 741–743). Accordingly, a model is neither right nor wrong but only suitable for a particular purpose or not (cf. Kircher et al., 2009, pp. 736, 754). The cybernetic concept of modelling pushes this definition forward and explicitly discusses the involvement of human beings, i.e. there is a mutual dependency of the real object O to be modelled, the model M itself and the addressee which is called the subject S (cf. Kircher et al., 2009, p. 737). The dependency M-O is characterised by epistemology, scientific methods, and repeated cycles of induction and deduction (cf. Kircher et al., 2009, pp. 739–742). The relationship to the subject S is characterised by group-specific conventions of modelling, the kind of simplification (scientific standards of the community), or by useful representations for explanatory approaches or learning processes (cf. Kircher et al., 2009, pp. 745–758). The subject or rather the addressee (here: the scientific community) along with its needs and ideas of science directly influences the perspective on the real object as well as the act of modelling and thus the model itself (cf. Kircher et al., 2009, pp. 745–758).

For this thesis, human uncertainty is the object to be modelled or, to be more precise, it is the complex cognitive process that leads to this phenomenon. The subject can be either the community of (computational) neuroscience or the community of predictive data mining and recommender systems. Based on the previous research done in this thesis, it is only consistent to decide for the latter community. On the one hand, the scientific standard of modelling is quite simple for this community: According to Weiss and Indurkhya, a possible model is adequate for describing human behaviour if it generates recommendations for items or products that users like, i.e. if this model optimises accuracy metrics (cf. Weiss and Indurkhya, 1998, p. 36). On the other hand, this standard is exactly what has been reasonably questioned in the previous part of this thesis and hence cannot serve as a standard for evaluating a model in the second part of the same thesis. This standard must hence be slightly modified: A cognitive model of human uncertainty should provide a verifiable positive benefit for the community as it is created for the very same. Two obvious benefits of a cognitive uncertainty model would be

R1 the adequacy for predicting the entire feedback distributions, especially human uncertainty in terms of variance and

R2 the simplicity of integration into existing machine learning algorithms and other data science techniques.

These benefits also serve as the first two model requirements. In short, the benefit to the community is the ability to provide a holistic view of human feedback in existing systems without major adjustments. This in return enables to realistically estimate human uncertainty, its genesis dependent on time and content, and its technical processing within information systems. Possible advantages of a holistic prediction of human uncertainty are statistically sound comparisons (cf. Ch.4), in-depth user insights (cf. Sizov, 2017b), better learning strategies (cf. Said and Bellogín, 2018), and new opportunities for system design, e.g. "when picking among several items with the same expected rating, the system can favor the item for which the confidence in the prediction is greatest" (Koren and Sill, 2011, p. 123). Of course, these are only the minimum requirements. Additional requirements could be, e.g. runtime efficiency as well as the efficiency of computational resources. However, this thesis is supposed to focus exclusively on a proof-of-concept and leave further enhancements to further research.

According to these criteria, one does not necessarily need a neurological model. The idea of developing such a model is to obtain information from a possible and human-inherent way of information processing that may eventually lead to system improvements and a better understanding of human beings at the same time. In doing so, it is essential to ensure that this model indeed reflects a 'true' cognition process. This entails two major challenges: First, it must be emphasised once again that this model can only be a simplified mathematical construct and must not be put on a par with reality and the ultimate truth. In this respect, Kircher argues that explanations through a model are not about the 'why' but rather about the possible 'how' (cf. Kircher et al., 2009, p. 754). Second, such a model can virtually not be proven without doubt regarding neuroscience research, since this thesis does not consider the relevant measurement approaches (e.g. electroencephalogram, magnetic resonance imaging, positron emission tomography, etc.) or any anatomical structures of the human brain. But against this background, how can it then be examined whether such a model is indeed close to 'real' human cognition? One possible approach is the implicit validation by comparing the model's implications with recent findings in medical or neuroscientific publications. This

approach is supported by the fact that explanations through models should never be understood locally but have to be considered against the background of other prevalent models, common assumptions, and obtained data (cf. Kircher et al., 2009, p. 754). Therefore, the criteria for model adequacy concerning cognition and decision-making can be summarised by the following requirements:

R3 Model implications are supported by medical and neuroscience research.

R4 There is a reasonable integration of other common neuroscientific models.

R5 The model itself is internally consistent.

In conclusion, it can be said that – in concordance to the main objectives of this chapter – the model adequacy has to be examined explicitly for the community of predictive data mining but also has to include validation in the light of epistemology. The latter ensures that, at least hypothetically, a model is created that indeed considers human uncertainty from an inner perspective. This cognitive model can then be compared to a purely behavioural model, i.e. a model that is founded on mere human observation rather than on assumed human-inherent processes. From this comparison, conclusions may be drawn for the future design of (more) human-like systems.

6.2 Finding Adequate Models

In this section, a neuroscientific model is presented that will probably explain human uncertainty as it has been measured in Ch. 3. Surprisingly, several months of intensive literature research revealed only a single model that seemed up to the task. This should not be taken as a proof that this is the only suitable model. Rather, the continued search for any other suitable model was simply not successful. Therefore, the path of enquiry will be described that finally led to this particular model used henceforth. In doing so, the adequacy of this model will explicitly be emphasised for the case of user response behaviour. This will shed light on its biological plausibility and credibility for naturally reducing human uncertainty to neuronal mechanisms.

A particular difficulty in interdisciplinary literature enquiries is that the phenomenon of human uncertainty – and the term uncertainty in particular – has different meanings within different research areas. At this point, the most important examples for this chapter should be mentioned:

Metrology: "The word 'uncertainty' means doubt, and thus in its broadest sense 'uncertainty of measurement' means doubt about the validity of the result of a measurement. Because of the lack of different words for this general concept of uncertainty and the specific quantities that provide quantitative measures of the concept, for example, the standard deviation, it is necessary to use the word 'uncertainty' in these two different senses. [...] The formal definition of the term uncertainty of measurement [...] is as follows: uncertainty (of measurement) [is a] parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand." (JCGM, 2008a, p. 2)

Decision Theory: "Uncertainty in this context refers to a state in which the probability of the result of selecting an alternative is not known. Decision-making under uncertainty can be sub-classified as follows [...]. The first group is decision-making under ambiguity. This ambiguity refers to a state in which, although the condition and results that will occur are known, the probabilities of the condition and results to occur are unknown" (Takemura, 2014, p. 8). "The second category [...] is decision-making under ignorance when the elements of the set of states or the elements of the set of results are unknown [or not clearly known]" (Takemura, 2014, p. 9).

Neuroscience: Here, two types of uncertainty are distinguished: "Animals are constantly faced with the challenge of interpreting signals from noisy sensors [i.e. uncertainty type 1] and acting in the face of incomplete knowledge about the environment [i.e. uncertainty type 2]. A rigorous approach to handling uncertainty is to characterize and process information using probabilities" (Doya et al., 2007, p. 239). Yu and Dayan propose "that the neuromodulators acetylcholine and norepinephrine play a major role in the brain's implementation of these uncertainty [computations]" (Yu and Dayan, 2005, p. 682). Friston confirms that "the most obvious candidates [...] are classical neuromodulators like dopamine and acetylcholine" (Friston, 2010, p. 132).

Information Theory: "Entropy is a measure of the uncertainty or surprise associated with a stochastic variable, such as a stimulus" (Dayan and Abbott, 2001, p. 28 of Ch. 4). The concept of information theory is frequently used in theoretical neuroscience

and therefore mingles with the vocabulary of this research area. Friston argues that each mammal (including human beings) tries to reduce surprise and that a cognitive "density with low entropy means that, on average, the outcome is relatively predictable" (Friston, 2010, p. 127). The more predictable an outcome is, the less uncertainty it has for the respective mammal.

Given this variety of different definitions and terminology, it is very difficult to gather information about the previously described phenomenon of human uncertainty. Apart from that, the above definitions are the result of a systematic and thorough literature study. The very first enquiry mainly revealed content applying the definition of decision theory which did not seem expedient in the context of the present thesis. The focus was hence changed towards the description of internal distributions within decision-making, as this is an implication of human uncertainty. In doing so, the work of Karl Friston has been found which contributes to the Bayesian brain hypothesis. This hypothesis represents

> "the idea that the brain uses internal probabilistic (generative) models to update posterior beliefs, using sensory information, in an (approximately) Bayes-optimal fashion." (Friston, 2010, p. 129)

The essence of this hypothesis is that the (human) brain uses probability distributions to represent the world internally. The existence of such internal distributions has been postulated earlier in this dissertation. It will be demonstrated later that several indications indeed point to the appropriateness of the Bayesian brain hypothesis. Friston focused on how the brain learns to predict its surrounding world with high accuracy:

> "The underlying idea is that the brain has a model of the world that it tries to optimize using sensory inputs. [...] Central to this hypothesis is a probabilistic model that can generate predictions, against which sensory samples are tested to update beliefs about their causes [i.e. to minimise prediction error]." (Friston, 2010, p. 129)

This assumption is logically sound since the correct interpretation and prediction of the world is crucial to survival. Such a generative model which is able to continuously update prior beliefs about states of the world (i.e. perceived parts of reality or entities of the world) with sensory input is introduced in Friston (2010). The key element of

this generative model is entropy (here: surprise or prediction error) and the information-theoretic concept of free energy optimisation (as an upper bound of entropy).

Although a theory has been found in which internal probability distributions are assumed and which provides information-theoretical concepts for modelling, two problems arise for the application to human uncertainty. The first problem is the fact that this theory describes learning from perception rather than decision-making. However, Friston has provided further hints for transferring this theory to the case of decision-making:

> "The basic idea is that behavior can be cast as inference: in other words, action, and perception are integral parts of the same inferential process and one only makes sense in light of the other." (Friston et al., 2013, p. 2)

This statement clarifies that behaviour as the direct manifestation of preceding decision-making can also be understood in the light of the Bayesian brain hypothesis. Accordingly, it can be assumed at this point that the feedback behaviour of users along with human uncertainty has a Bayesian counterpart within the cognitive process. Despite this statement, Friston's work does not contain any information or specifications that lead to a formal model to build upon. The second problem so far is that internal distributions (even when they are assumed for decision-making) do not necessarily contradict determinism. For example, the proposed Maximum-A-Posteriori approach will always give the same value for a given prior and likelihood. This leads to the conclusion that either the computational process based on constant distributions is unreliable or the representation of such distributions itself must be statistical rather than deterministic. As already stated above, Friston and others ascribe the biological modulation of uncertainty (represented by a density's width) to neurotransmitters such as acetylcholine, noradrenaline (cf. Yu and Dayan, 2005, p. 682), and dopamine (cf. Friston, 2010, p. 132).

Therefore, the next step was to initiate further investigations in neuroscience literature using the keywords "behaviour variability" and "neurotransmitters". This quickly revealed the work of Faisal et al. which describes the genesis and exploitation of noise in the nervous system. It is noteworthy that this contribution implicitly suggests a causality between behavioural variability and neuronal noise:

> "Variability is a prominent feature of behaviour [and] in perception and action [it] is observed even when external conditions, such as the sensory

> input or task goal, are kept as constant as possible. Such variability is also observed at the neuronal level." (Faisal et al., 2008, p. 292)

For the very first time during the literature study, a description was found that is consistent with the observed phenomenon of human uncertainty. By the short consecutiveness of two supposedly distinct phenotypes, Faisal implicitly insinuates a correlation between behavioural variability under constant externals on the one hand and the so-called trial-to-trial variability of nervous cells on the other. Trial-to-trial variability in this context paraphrases "the differences between [neuronal] responses that are observed when the same experiment is repeated in the same specimen (for example, in the same neuron or the same subject)" (Faisal et al., 2008, p. 292). It is indeed remarkable that a repetition of the same task reveals a variability of behaviour and, at the same time, the same trigger in the case of recurrence never leads to the same neuronal responses. This correlation yields a sense of causality when remembering that neuronal responses are ultimately responsible for a particular behaviour. Faisal et al. continue with this chain of causality and discuss the origin of trial-to-trial variability. The culprit is swiftly found to be noise in the nervous system which originates from various mechanisms, i.e.

- motor noise (cf. Faisal et al., 2008, p. 293),
- sensory transduction and the non-deterministic amplification of sensory inputs (sensory noise) (cf. Faisal et al., 2008, p. 293),
- random opening of voltage-gated ion channels located on excitable membranes as well as specific network structures of neurons (cellular noise) (cf. Faisal et al., 2008, p. 293),
- biological constraints which lead to a differential release of neurotransmitters (synaptic noise) (cf. Faisal et al., 2008, p. 293).

Having these observed mammal-inherent properties in mind, it is argued that "small biochemical and electrochemical fluctuations [...] can significantly alter whole-cell responses" (Faisal et al., 2008, p. 294). The rationale behind this is that "when the membrane potential is near the firing threshold, the generation of an AP becomes highly sensitive to noise" (Faisal et al., 2008, p. 294). The abbreviation AP stands for action potential and means a temporal change in a membrane's local electric field which propagates along a neuron's axon and ultimately leads to neuronal responses and information transmission

(cf. Bear et al., 2018, pp. 83–110). After this comprehensive line of argumentation, Faisal converts the previously expressed implicit correlation between behaviour and neuronal characteristics into an explicit conjecture of causality that has long been suspected:

> "Noise is an inescapable consequence of brains operating with molecular components at the nanometer scale, sensors that are sensitive to individual quanta and complex networks of noisy neurons that generate behaviour." (Faisal et al., 2008, p. 300)

Although this could be seen as a disruptive factor in information transmission, Faisal emphasises that there are ample benefits related to neuronal noise. For example, a certain level of noise can be used to detect and transmit weak signals within neural circuits which has been directly demonstrated in human balance control (cf. Faisal et al., 2008, p. 294). Friston even assumes that this noise is deliberately used by the brain to represent probability densities qua neuronal activity adjustment and the formation of specific connection strengths (cf. Friston, 2010, p. 129). This idea is also discussed positively by Faisal:

> "Psychophysical experiments have confirmed that humans use these Bayesian inferences to allow them to cope with noise (and, more generally, with uncertainty) in both perception and action. However, the neural mechanisms that are involved in Bayesian computations are unknown. One idea is that neurons encode probabilities or beliefs about the state of the world and this concept has been incorporated into Bayesian models of neuronal population codes." (Faisal et al., 2008, p. 299)

This backlink to the Bayesian brain hypothesis provides additional references for further investigation, i.e. the keyword 'neuronal population codes'. A search query for these keywords exposed mainly the work of Alexandre Pouget and his probabilistic population codes (PPC). It is a mathematical model that explains the formation of internal probability distributions based on neuronal noise. Unlike Friston's generative model, the PPC approach is not a theory on updating prior probabilities but rather describes the cognitive occurrence of estimates or predictions (of states of the world). The first thing that is striking about this model is that it is capable of unifying all the seemingly different definitions of uncertainty mentioned above. This is made clear by the following example:

> "For instance, imagine hiking in a forest and having to jump over a stream. To decide whether or not to jump, you could compute the width of the stream and compare it to your internal estimate of your jumping ability. If, for example, you can jump 2 m and the stream is 1.9 m wide, then you might choose to jump. The problem with this approach, of course, is that you ignored the uncertainty in the sensory and motor estimates. If you can jump 2 ± 0.4 m and the stream is 1.9 ± 0.5 m wide, jumping over it is very risky – and even life-threatening if it is filled with, say, piranhas." (Ma et al., 2006, p. 1432)

From the perspective of decision theory, jumping or not is a decision under uncertainty. A person does not know the stream width and the own jumping abilities with absolute certainty and hence has to rely on guessing. From the metrology viewpoint, one regards this uncertainty (or guessing) in the form of a probability density over the range of possible outcomes. From the neuroscientific point of view, it is exactly this probability density that is supposed to be represented by neuron populations (i.e. agents). According to Friston, the brain is indeed organised in agency and each of these adaptive agents has to focus on a limited amount of states of the world (cf. Friston, 2010, pp. 2–3). For the example above, a person would employ single agents for the stream width and jumping width, respectively. Each agent would deliberately employ noise to form probability densities that can be evaluated in a Bayes-optimal fashion to finally make a decision under uncertainty. From the perspective of information theory, if one decides to jump, then getting wet would be a surprise (high entropy) and due to the tendency of entropy minimisation one would adjust the prior probability for upcoming decisions (memory). As can be seen from this example, these estimation agents can be deemed as some kind of adhesive between the individual disciplines related to uncertainty.

This reveals the conviction of Ma et al. that probabilistic population codes indeed constitute the neuronal representation of decision-making in a case in which crucial facts are missing and have to be guessed or inferred by internal probability distributions. This has implicitly been elaborated in the first half of this dissertation: People make decisions about personal preferences, but they lack a solid evidence base in the form of standardised and measurable variables they can rely on. So there is uncertainty about this quantity and each person has to use an individual estimate which, in case of repetition, exhibits the existence of an internal probability density. Even the assumption

that user feedback has multiple dimensions (e.g. quality, usability, ergonomics, price, etc.), which are aggregated into a single estimation, is supported and even proposed by this model through agency. Most importantly, those agents along with their ability to form and pass internal probability densities have been purely fictitious until now and are being concretised for the first time by Pouget's PPC model.

The findings of the previous literature research convincingly indicate a neuronal mechanism that could indeed be transferred to the observed case of human uncertainty. As a potential disadvantage, it should be mentioned that this model is demonstrated and proven solely by sensory perception and motor control. This is not because such mechanisms cannot be applied to higher cognition (the opposite is illustrated by the example above) but because decision-making can hardly be measured neurologically and the explicit train of thought cannot be controlled in laboratory settings. Nevertheless, this example implicitly reflects the belief of neurologists that the Bayesian brain hypothesis can be applied to the uncharted field of decision-making and that the brain may generate estimates using the PPC approach.

But how do those probabilistic population codes work? This question will be thoroughly answered in Sec. 6.3 and Sec. 6.4 using mathematical descriptions. Therefore, a more verbalised explication along with an introduction to the most basic ideas will be given at this point. Preceding to every neuronal response, there is a so-called excitation variable or stimulus that causes neuronal activity. For the case of decision-making, this excitation variable is not an external quantity which is recognised by sensory perception but rather an internal quantity, virtually a hidden variable in terms of Bayesian modelling. This hidden variable is also substantially embedded within Friston's generative model (cf. Friston, 2010, p. 128). A crucial point is that this excitation variable or stimulus (i.e. the cause of neuronal responses) can be almost anything:

> "In probabilistic models, the variable s [denoting the stimulus] is referred to as a latent variable (the width of the [river] in the previous example) or [...] a set of latent variables [...]. Note that latent variable is a broad term and need not refer to concrete quantities in the outside world. In motor control, s can be a goal ([e.g.] reaching an object at a particular location), and, in the cognitive domain, it can be relational structures, such as who in our circle of friends gets along with whom." (Pouget et al., 2013, p. 1171)

In other words, this excitation variable or stimulus is a metaphor that is needed to initiate neuronal activity within an artificial system to satisfy causality (i.e. cause and effect). However, there is a biological counterpart to this. When thinking about an estimation to be made or user feedback to be provided, something indeed triggers the necessary thinking process and induces neuronal activity. Finally, it is exactly this activity that represents a probability density over the range of possible activity causes: "Put simply, although agents can never know the causes of their [activations], the causes can be inferred" (Friston et al., 2013, p. 3). To stay in concordance with the scientific literature about the PPC model, this internal excitation variable will henceforth be denoted as the stimulus which is the original terminology of Pouget.

For a given stimulus, a neuron responds with a specific spiking frequency which is the number of action potentials within a defined time interval (cf. Dayan and Abbott, 2001, p. 12 of Ch. 1). A variation of the stimulus (e.g. direction of motion or wind direction) will demonstrably lead to a change of a neuron's response (cf. Dayan and Abbott, 2001, pp. 12–14 of Ch. 1 in addition to pp. 3–4, 12–14 of Ch. 3). It therefore makes sense to understand the neuronal response r as a function of the stimulus attributes s and in fact, it is possible to find such functional dependency $r = f(s)$ in many cases. This mapping f is referred to as the tuning curve (cf. Dayan and Abbott, 2001, pp. 12–15 of Ch. 1). One important characteristic of many tuning curves is the existence of a (local) maximum, i.e. there is a specific stimulus value for which the neuron maximises its spiking frequency (cf. Zemel et al., 1998, p. 405). This specific value is denoted as the preferred stimulus of the corresponding neuron.

The brain makes use of this property in certain situations, for example, to orientate itself in space: The discovery of so-called place cells happened in 1971 by O'Keefe and Dostrovsky during an experiment in which the authors measured spiking activity for a neuron population in a rat's brain while the rodent moved through a cage (cf. O'Keefe and Dostrovsky, 1971). The main result is that for each location within the cage, another neuron maximised its frequency. The discovery of these place cells suggests that neurons can be organised in such a way that their tuning curves fully cover the whole range of possibilities for a stimulus so that the true state of the world can be inferred. Noteworthy is also the "considerable spatial overlap between the fields" (Pouget et al., 2000, p. 125) when tuning curves are organised to cover a range of possible values. This finding is essential for understanding how the brain operates efficiently in the presence of noise: Since neuronal noise can support or suppress stimuli-related action potentials

(cf. Faisal et al., 2008, p. 294), one neuron may respond stronger to a certain location within the maze than the neuron whose preferred value is actually fixed to this location. Put simply, noise can cause the wrong neuron to respond stronger than the correct neuron. This effect is diminished when the whole population is considered, as each neuron represents an uncertain voter (for the true stimulus value). The overlapping of tuning curves ensures that multiple neurons near the true stimulus respond stronger than the remaining ones on average. In this light, it makes sense not to select that particular neuron with a maximum response (whose preferred stimulus may be far from the true stimulus), but to locate the true stimulus within this local group, where many neurons respond strongly, and then use the local maximum instead of the global one. This line of argument is consistent with the acknowledged opinion that population codes indeed serve to overcome issues related to noise:

> "One key property of the population coding strategy is that it is robust [...] because the information is encoded across many cells. [...] Population codes turn out to have [...] computationally desirable properties, such as mechanisms for NOISE [sic!] removal [...]." (Pouget et al., 2000, p. 125)

Of course, the given example is only about spatial orientation so that generalising to other brain tasks seems difficult. However, population coding has additionally been revealed for a plethora of other brain tasks:

> "For instance, in primary visual cortex (V1) and area V4 of the macaque, population codes exist for orientation, color, and spatial frequency. In the hippocampus in rats, a population code exists for the animal's body location. A population code for spatial location in a visual scene or of the body of the organism is also called a topographic map. The cercal system of the cricket has a population code for wind direction. The secondary somatosensory area (S2) in the macaque has population codes for surface roughness, speed, and force. The postsubiculum in rat contains a population code for head direction. Primary motor cortex (M1) in macaque uses populations coding for direction of reach. Even abstract concepts such as number appear to be encoded by population codes in the prefrontal cortex." (Ma and Pouget, 2009, p. 751)

In the light of this evidence, it is justified to believe Pouget who declares that "all neural circuits share similar features and, in neocortex, the detailed circuitry is remarkably

well preserved across areas. It is therefore quite possible that these circuits share common computational principles" (Pouget et al., 2013, p. 1177). In addition, it could be demonstrated that probabilistic population codes greatly simplify a very important task of cognitive systems, namely cue integration which is clearly described by Trommershauser as follows:

> "When an organism estimates a property of the environment so as to make a decision [...], there are typically multiple sources of information (signals or 'cues') that are useful. Information may [...] derive from multiple senses such as visual and haptic information about object size, or visual and auditory cues about the location of a sound. In most cases, the organism can make more accurate estimates of environmental properties or more beneficial decisions by integrating these multiple sources of information." (Trommershauser et al., 2011, p. 5)

From a theoretical Bayesian perspective, cue integration is the process of merging several probability densities for the same stimulus into a single density. Such processing is inevitable under the assumption that the brain works by agency. As stated above, each agent (i.e. neuron population) can only focus on a limited amount of parameters. In order to get a holistic estimation about a stimulus it seems fruitful to combine the different agents' sources of information.

For instance, having two independent agents A_1 and A_2 representing the location probabilities of a sound source (e.g. a car crash) based on distinct sensation (e.g. auditory and visual), these pieces of information can be combined with Bayes' rule $P(s|A_1, A_2) \propto P(A_1|s)P(A_2|s)P(s)$. Given the assumption of a flat prior $P(s)$ (inexperienced observer) and normality of both likelihoods $P(A_i|s)$ with $i = 1, 2$, Pouget shows that the combined posterior is yet a Gaussian with sufficient statistics

$$\mu = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\,\mu_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\,\mu_2 \quad \text{and} \quad \frac{1}{\sigma} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \tag{6.1}$$

(cf. Ma et al., 2006, p. 1433). The assumption of normality is justified by the fact that "neuronal variability is due to the combined effect of a large number of stochastic processes" (Ma and Pouget, 2009, p. 750) and indeed, the above considerations can be verified in a variety of real life experiments (cf. Jacobs, 1999; Knill and Saunders, 2003; Hillis et al., 2004; van Beers et al., 1999; Ernst and Banks, 2002; Battaglia et al., 2003; Alais and Burr, 2004). The interesting point is that the involvement of the

above considerations describes Bayes-optimality throughout sensory perception. This means that the combination of both cues is weighted according to the certainty or reliability of each agents' information. This cue-weighting means that an observer naturally accounts strongest for the most reliable information. Evidence that mammals behave Bayes-optimal contributes in favour of the Bayesian brain hypothesis and it implies that agents are capable of encoding entire probability distributions which must inevitably be preserved during the cognition process. From this follows that "these distributions are collapsed onto estimates only when decisions are needed" (Ma et al., 2006, pp. 1436–1437). This conclusion has also been drawn in the first half of this thesis. The true process of such Bayesian-optimal cue integration is still unknown (cf. Ma et al., 2006, p. 1432). However, Pouget was able to demonstrate that the PPC model is not only able to encode such densities but that it also simplifies cue integration:

> "Specifically, this [Poission-like] variability has a unique property: it allows neurons to represent probability distributions in a format that reduces optimal Bayesian inference to simple linear combinations of neural activities." (Ma et al., 2006, p. 1432)

The mathematical evidence can be found in Ma et al. (2006); Pouget et al. (2013); Beck et al. (2007) and relies on the assumption that the noise distribution is part of the exponential family. Surprisingly, this is exactly what has been found within the nervous system, i.e. the noise indeed follows a Poisson distribution (cf. Moreno-Bote, 2014). Besides this theoretical evidence for PPC involvement, there are also experimental hints to be found: The application of PPC in cue integration has been positively evaluated for robots in artificial environments facing an auditory-visual integration task (cf. Bauer et al., 2015). Moreover, the linear merging of particular neurons within two populations "is consistent with the responses of neurons in areas such as [the] lateral intraparietal cortex" (Pouget et al., 2013, p. 1173). The involvement of PPC in cue integration might be fruitful for user feedback as well. So it may be possible that a single feedback is formed by evaluating different sub-aspects (e.g. ergonomics, visual appearance, usefulness, price-performance ratio, etc.) which are then aggregated. The PPC model provides useful approaches for this aggregation within a Bayesian framework and under consideration of real biological operating principles.

At this point, the facts and the success of the PPC model should be recollected along with an explicit elucidation for its appropriateness when considering uncertain user feedback. The above literature research revealed that many authors assume a connection between neuronal noise and decision-making in an uncertain world. Just as many authors assume that neuronal noise is also responsible for a trial-to-trial variability of behaviour which is exactly what has been observed for user feedback. This is particularly interesting, because there is some doubt about the idea that human uncertainty arises from a different history, i.e. a changing context. At the beginning of this thesis, the alternative hypothesis was presented that this phenomenon could rather originate from a complicated cognitive process. Now the first evidence has been found that this idea is shared by other authors and that a plausible origin of this uncertainty has been identified as noise in the nervous system. In particular, there is indeed a neural basis for modelling behaviour as a probability density and it is assumed that the brain does the very same; i.e. that it works specifically with such densities (Bayesian brain hypothesis). This is demonstrated by various cue integration experiments, all of which provide evidence that the entire density must be fully available for the cognitive process. The theory about a single user rating being only a draw from an underlying feedback distribution was also substantiated by neuroscience since Pouget proposed "that these distributions are collapsed onto estimates only when decisions are needed" (Ma et al., 2006, pp. 1436–1437). So, the considerations from the Bayesian brain hypothesis are in total concordance to those made in the first half of this dissertation, i.e. human uncertainty is likely to have a biological origin and feedback estimates are likely to be drawn from internal distributions.

However, many research projects in the field of the Bayesian brain hypothesis deal with the integration of two densities, the storage and update of prior densities, the reinterpretation of neuroimaging data, and so forth (cf. Doya et al., 2007). At their very essence, all of these research directions are not concerned with the development of probability densities and how they can be used to explicitly obtain single estimates, for instance, a 3-star rating. In this network of contributions on the Bayesian brain hypothesis, only one single theory has been found during the literature study that deals with the emergence and representation of densities along with collapsing densities onto estimates when needed. This is the model of probabilistic population codes (PPC). This model is consistent with the presumed brain agency and, in the light of explicit measurements of neuronal noise, it simplifies agency communication in such a way that

it can be performed by neurons in the cortex. The applicability of such encoding has so far been positively evaluated for

- motor commands and neural prosthetics (cf. Chapin, 2004),
- tactile perception of crickets (cf. Dayan and Abbott, 2001, pp. 12–14),
- auditory perception (cf. Knill and Pouget, 2004),
- visual recognition of orientation, colour, direction of motion, and depth (cf. Pouget et al., 2000, p. 125).

It must be emphasised how remarkably the biological measurements and theoretical derivations fit together with the phenomenon of human uncertainty when the PPC model is applied. Moreover and even more important, the PPC model provides a mathematical framework to describe how such densities arise on the basis of noisy neuronal responses. Such a mathematical description and computability is important for developing novel computer systems for prediction. One potential disadvantage of this model is that it was mainly evaluated in sensory perception and motor control so far. Other authors always assumed an involvement in higher cognition as well and quite recently, evidence has been found for PPC models to explain memory effects:

> "Errors in short-term memory increase with the quantity of information stored [...]. An alternative perspective attributes recall errors to noise in tuned populations of neurons [...]. I show that errors associated with decreasing signal strength in probabilistically spiking neurons reproduce the pattern of failures in human recall under increasing memory load. In particular, deviations from the normal distribution that are characteristic of working memory errors [...] are shown to arise as a natural consequence of decoding populations of tuned neurons. Observers possess fine control over memory representations and prioritize accurate storage of behaviorally relevant information [...]. I show that changing the input drive to neurons encoding a prioritized stimulus biases population activity in a manner that reproduces this empirical tradeoff in memory precision. In a task in which predictive cues indicate stimuli [...], human observers use the cues in an [Bayes-]optimal manner to maximize performance, within the constraints imposed by neural noise." (Bays, 2014, p. 3632)

Evidence for the area of decision-making is still pending. However, considering the previous evidence in favour of this model, there is sufficient substantiation that probabilistic population codes are nevertheless a good choice. Therefore, all further investigations in this chapter assume that the PPC approach can be used for modelling unreliable decision-making. Heuristically, why would the evolution of the human brain give birth to two distinct principles of operation (i.e. one for sensory perception and another for cognition) when there is one single principle to explain them all. A supposedly bigger point of criticism is the choice of stimulus when this model is transferred to decision-making. In sensory perception, the stimulus is clearly defined and neuronal representations of such states of the world can be measured explicitly. For example, it is possible to measure the neuronal representation of space and location in place cells. Considering decision-making, however, it is hardly possible to identify a neuronal representation of something as abstract as a star-rating. In this regard, a brief reminder has to be given that the stimulus is not a physical object of the real world but just a hidden variable, i.e. a metaphor for starting the modelled (artificial) cognitive process. The fact that this hidden variable (i.e. the cause of neuron activation) can be detected during sensory perception does not necessarily mean that there is no hidden variable that triggers a cognitive process in unmeasurable situations. The existence of such a cause of neuronal activation is based on the fact that people are obviously able to make decisions and to provide user feedback – and the related cognition process has to be started somehow. The only lack of clarity is whether the user feedback options s are directly represented or whether there is a transformation τ so that a representation of $\tau(s)$ takes place. For the initial research in forthcoming sections, $\tau \equiv \mathrm{id}_{\mathbb{R}}$ will be used.

It has now been clarified that this model addresses the popular assumption of behavioural variability being caused by neuronal noise. It has been pointed out that this model has been positively evaluated in many areas of cognition. It was also emphasised that this model provides the necessary mathematical description for implementation in predictive systems. What has not yet been clarified is what such an implementation could look like and how it could be evaluated. Since the population responses solely depend on how tuning curves cover the range of possible stimuli values, it seems obvious to explicitly parametrise those tuning curves and to specify the covering approach (e.g. equidistant, adaptive, etc.). For this, Pouget gave initial hints in various publications (cf. Pouget et al., 2000; Ma and Pouget, 2009; Doya et al., 2007). In this way, several parameter vectors can be obtained, each characterising a population's basic configuration. These

vectors naturally span a feature space and each user feedback can then be represented in the form of a neuronal feature vector. This may enable user clustering based on the neuronal disposition of individual users along with many other possible applications in predictive data mining. The evaluation of such a model can be based, for example, on the quality with which it reconstructs the densities measured in the RETRAIN study without violating biologically realistic limits.

6.3 Probabilistic Population Codes

In this section, the modelling of probabilistic population codes is introduced according to Ma et al. (2006). However, minor changes are applied to fill missing model details and to extend this model to the scenario of user feedback. Furthermore, evidence is discussed to make this model more plausible concerning its appropriateness for human decision-making (higher cognition). But first of all, it is important to understand how the human brain processes information and where uncertainty arises during the cognitive process. These mechanisms are then translated into an adequate user model.

The Single Neuron Model

The response of a single neuron to a stimulus is limited to the transmission of electric impulses (spiking) and since each neuron has only got two states of activation, i.e. spiking or not, theories of neural coding assume that information is encoded by the spiking frequency (rate) (cf. Doya et al., 2007, p. 53). The functional relationship between such responses r of a neuron and the attributes $s \in S \subset \mathbb{R}$ of a stimulus is given by the so-called tuning curve $r = f(s)$ (cf. Dayan and Abbott, 2001, pp. 12–15 of Ch. 1).

A fundamental experiment to reveal this functional relationship was conducted by Hubel and Wiesel (1968). They measured the spiking rate of a single neuron from a monkey's visual cortex V1 whilst presenting a rotating black bar. The experimental setting and the data obtained are visualised in a simplified form in Fig. 6.1 following the description by Doya (cf. Doya et al., 2007, pp. 73–74) and Dayan (cf. Dayan and Abbott, 2001, p. 13 of Ch. 1). The boxplots represent the repeated measurement of spiking rates for always the same rotation angle. Two results can be observed: The first result is that there is indeed a functional relationship between the rotation angle of the black bar and the measured spiking rate of a particular neuron. The second result is the

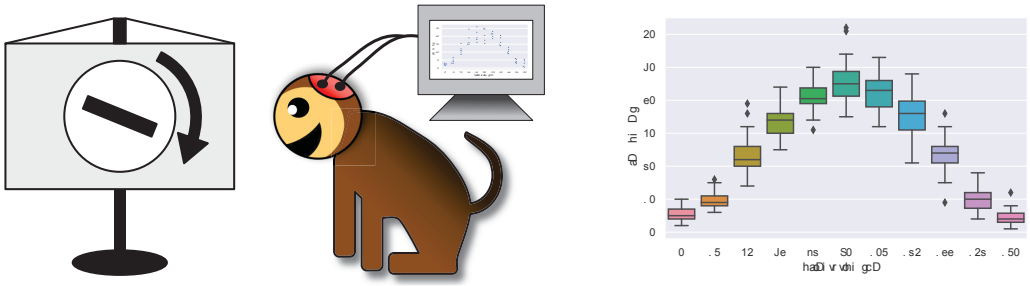


Figure 6.1: Experimental measuring of a tuning curve according to Hubel and Wiesel (1968); Dayan and Abbott (2001); Doya et al. (2007)

unreliability of spiking, i.e. the same stimulus does not necessarily result in the same spiking rate when the measurement is repeated. Even from this single experiment, it can already be concluded that the internal representation of the outer world is not fully reliable and repeated decisions based on these representations might change slightly. This basic idea will indeed constitute the source of human uncertainty according to the model of probabilistic population codes.

To this day, further experiments revealed that each neuron holds an entire set of possible tuning curves for different stimuli and their respective characteristics (cf. Mallot, 2013, p. 113). Besides irregular shapes, tuning curves have frequently been measured to be bell-shaped or sigmoid-shaped, respectively (cf. Dayan and Abbott, 2001, pp. 12–15 of Ch. 1). Moreover, each tuning curve maximises for a particular value $p := \operatorname{argmax} f$ which is denoted as the preferred stimulus. The approach in this thesis follows Doya et al. (2007); Pouget et al. (2000); Ma and Pouget (2009) and confines to bell-shaped tuning curves. Unfortunately, the relevant literature is missing an explicit parametrisation of these curves so that an independent parameter model has to be defined according to the descriptions in Pouget et al. (2000). For $p \in \mathbb{R}$ and $w, g, o \in \mathbb{R}^{>0}$, a bell-shaped tuning curve with preferred value p can be defined as

$$f_p \colon \begin{cases} S \to \mathbb{R} \\ s \mapsto g \cdot h(p, w)(s) + o \end{cases} \tag{6.2}$$

where the shape emerges from the Gaussian density function $h(p, w)$ with sufficient statistics p and w. The parameter w will henceforth be referred to as the tuning curve width. The remaining parameters can be explained heuristically: The tuning curve width naturally induces the maximum spiking rate $\max f_p$ due to the normalisation of

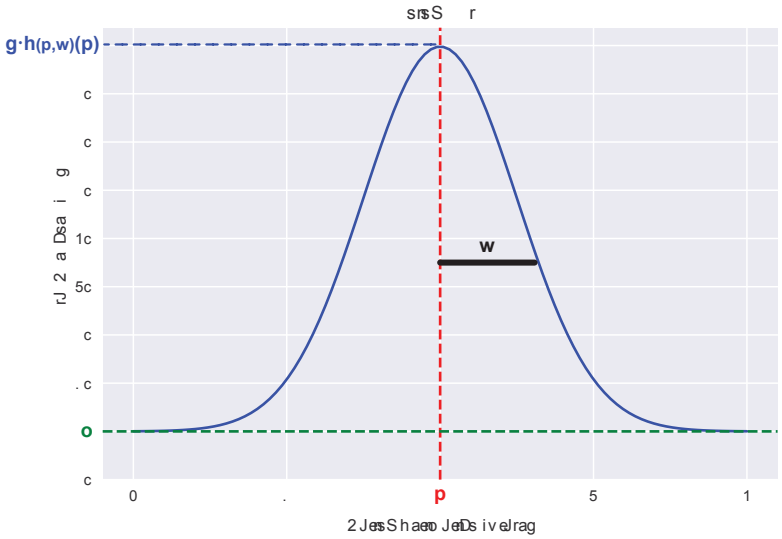


Figure 6.2: Parametrisation of a bell-shaped tuning curve

probability densities. To provide more flexibility, i.e. modelling higher or lower spiking rates for a given width, the Gaussian has to be multiplied with an additional stretching factor g which will be denoted as the frequency gain. The measurement results depicted in Fig. 6.1 suggest yet another component that adds to the frequency, i.e. a spiking offset. This offset becomes obvious when considering f_p at the limits of its domain S (stimulus space). This assumption is consistent with the setting of Pouget (cf. Pouget et al., 2000, p. 126) and Doya (cf. Doya et al., 2007, p. 116). For this reason, a positive constant o is added which will be called the spiking offset. All these parameters and their correspondence to forming the curve of f_p is depicted in Fig. 6.2.

So far, this model has been based upon static and reliable neuronal responses. This is by no means the case when measuring tuning curves in reality because of the perpetual interference with so-called neuronal noise (cf. Faisal et al., 2008). As the results from Fig. 6.1 demonstrate, one will find that spiking rates associated with a specific stimulus form a distribution of possible responses. These must hence be seen as random variables R and it has indeed been found that $R \sim \text{Poi}(\lambda)$ follows a Poisson-like distribution (cf. Moreno-Bote, 2014). Accordingly, a given stimulus value s will produce spiking rates $r \in \mathbb{N}$ each with a different probability determined by

$$P_\lambda(R = r) = \frac{\lambda^r}{r!} \exp(-\lambda) \tag{6.3}$$

with parameter $\lambda \in \mathbb{R}^{>0}$ which represents the expectation and variance of the random variable R. Practically, tuning curves are measured by multiple repetitions carried out for each value s of a stimulus followed by averaging $f_p(s) = \mathbb{E}(R)$ (cf. Dayan and Abbott, 2001, pp. 12–13 of Ch. 1). The Poisson parameter can hence be set to $\lambda = f_p(s)$ and it follows that

$$P(R = r) = \frac{f_p(s)^r}{r!} \exp(-f_p(s)) \quad \text{for} \quad r \in \mathbb{N}. \tag{6.4}$$

Given a measured spiking rate r of a single neuron with well-known tuning curve f, the task is now to guess the stimulus value s that had triggered this response. This problem is easily solved in the absence of neuronal noise just by computing the fibre $\{s\} = f^{-1}(r)$ over r. Please note that the inverse image of f is not necessarily unique since f is not injective for bell-shaped tuning curves. In reality, i.e. when including noise, this recent approach is hardly appropriate. Each s could have triggered this response with a certain probability $P(s|r)$. Thus, the stimulus cannot be precisely determined but it can be inferred, e.g. by opting for a stimulus that maximises this probability, i.e. $s = \text{argmax}_{s \in S} P(s|r)$. This argumentation mathematically exemplifies the essence of the Bayesian brain hypothesis as it has already been introduced in Sec. 6.2 through a detailed comparison of literature. If noise is involved, it is no longer possible to draw reliable conclusions about the cause of a neuronal response. The only way to gain clarity about the cause is to consider all the probabilities of possible causes and assume the most probable of them. For this purpose, the brain would necessarily need to incorporate representations of probability distributions. This makes the human brain indeed a Bayesian inference machine. As described in Sec. 6.2, this Bayesian inference becomes more stable when multiple neurons are utilised as each of them contributes additional information. It is thus reasonable to consider whole populations of neurons.

Probabilistic Population Codes

Up to this point, the phenomenon of neuronal noise has been considered together with its possible role in encoding sensory perception and higher cognition for single neurons. This approach is feasible to get a first understanding of the Bayesian brain hypothesis along with its potentials for modelling unreliable user feedback. However, cognitive tasks are executed by a vast number of interacting neurons rather than by lone individuals (cf. Dayan and Abbott, 2001, p. 6 of Ch. 1).

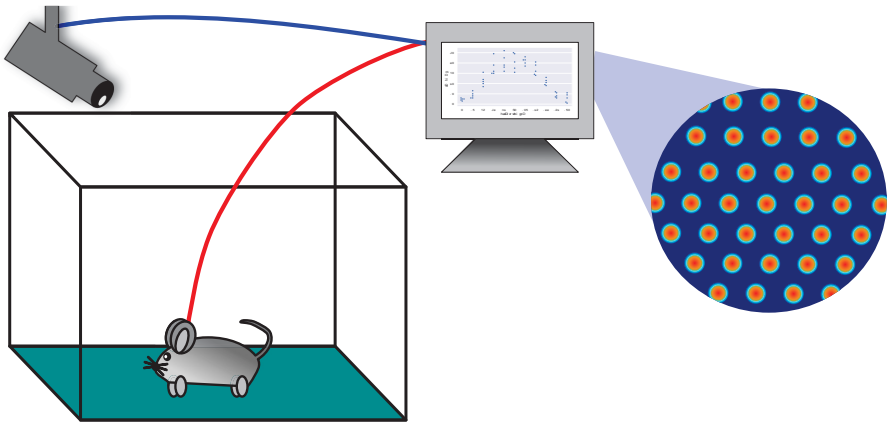


Figure 6.3: Simplified illustration of the discovery of place cells. For each position in space, another neuron maximises its spiking frequency (illustrated by the red dots).

One important finding related to neuron populations – at least for this dissertation – is the discovery of place cells. These have already been outlined in Sec. 6.2) and must be described in more detail at this point. In 1971, O'Keefe and Dostrovsky conducted an experiment in which they measured spiking activity for a population of neurons of a rat's brain (cf. O'Keefe and Dostrovsky, 1971). The rat was brought into a cage where its video recorded trajectory was aggregated with its population spiking behaviour. This experiment is depicted in a simplified form in Fig. 6.3. The main result is that for each location within the cage, another neuron of this population maximised its frequency. This means that all neurons within this population were encoding the entire two-dimensional space. The spiking rate of each neuron encoded a range of possible spatial coordinates for estimating the rat's current location. This principle holds a huge advantage: By representing the entire space with a multitude of neurons, the brain yields much more probabilities $P(s|r_j)$ $(j = 1, \ldots, n)$ about possible locations which can be integrated for a more precise spatial orientation. Transferring this finding to the mechanism of decision-making means that there is a universal set S of all possible decision outcomes for a particular situation and that each neuron within an assigned population is maximising its spiking frequency for another decision so that the entire population is able to cover the whole range of S. Mathematically this is achieved by spreading the preferred values of the neuron's tuning curves equidistantly along S.

To this end, let $S \subset \mathbb{R}$ be a proper subset representing the universal set of all decision outcomes. Let $n \in \mathbb{N}$ be the number of assigned tuning curves with respect to

Eq. 6.2, all having the same parameters $w, g, o \in \mathbb{R}^{>0}$. Let the sequence

$$(p_j)_{j=1,\dots,n} := \inf(S) + j \cdot \frac{\sup(S) - \inf(S)}{n} \tag{6.5}$$

be an equidistant partition of S. Note that S does not necessarily need to be a closed set and hence the supremum and infimum are required to determine its boundaries because a minimum and a maximum does not exist in this case. To provide simplicity to this model, all parameters determining the population size, the shape of all tuning curves, as well as the assumed stimulus (i.e. the unknown decision in the form of a hidden variable) $s \in S$ are summarised in a vector $\xi = (n, g, w, o, s)$ which will be referred to as the cognition vector. Such a cognition vector ξ hence becomes the representation of a whole population together with its cause of activation. Given a fixed ξ, each neuron will respond according to its specific tuning curve and distortion due to neuronal noise, i.e.

$$f_j(s) := f_{p_j}(s) = g \cdot h(p_j, w)(s) + o \tag{6.6}$$

$$R_j \sim \mathrm{Poi}(f_j(s)). \tag{6.7}$$

The realisation r_j of the random variable R_j is denoted as the response of the j-th neuron. To keep in mind that these responses are always dependent on the parameters of the cognition vector, the notation $r_j(\xi)$ will be used henceforth to indicate a realisation of $R_j(\xi)$. The response of the entire population is formed by the response of each neuron and so the n-dimensional random variable

$$\mathcal{R}(\xi) := (R_1(\xi), \dots, R_n(\xi)) \tag{6.8}$$

is denoted as the population response for ξ with realisation $\varrho(\xi) = (r_1(\xi), \dots, r_n(\xi))$.

This theory of the origin of noisy population responses is illustrated in Fig. 6.4. In this example, $\xi = (11, 10, 0.5, 5, 3)$ is used as the cognition vector, i.e. there are $n = 11$ neurons that respond to the assumed stimulus (i.e. the cognition result) of $s = 3$ stars where each tuning curve has the offset $o = 5\,\mathrm{Hz}$, the width $w = 1\,\mathrm{Hz}$, and the gain $g = 7$. In the left-hand subfigure, one can see the tuning curves which are distributed equidistantly over the possible range of a rating scale S with five stars. For $s = 3$ stars, the responses of each neuron can be fetched from its tuning curve. For a better representation of the population response, it has become a standard to plot the individual responses against the corresponding preferred values (cf. Pouget et al., 2000, p. 126) which can be seen in the middle subfigure. These are the theoretical (static)

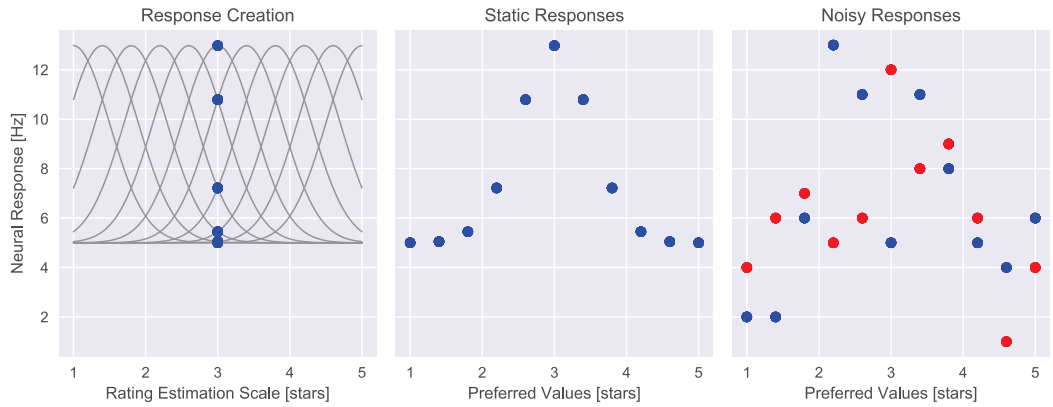


Figure 6.4: Genesis of noisy population responses demonstrating the alteration for each cognition trial (red and blue)

responses without any consideration of neuronal noise. To add this neuronal noise, each static response $r_j^{\text{static}}(\xi)$ is replaced by the draw of a random number from the Poisson distribution with parameter $\lambda = r_j^{\text{static}}(\xi)$. This can be seen in the right-hand subfigure. In addition, the very same sampling has been repeated one more time, i.e. the blue and the red dots in each case represent a noisy population response and it is obvious that these differ not only from the theoretical reference but also from each other. Again, it can be seen that the repetition of the same cognition leads to different neuronal activities, even when numerous neurons are involved. Whilst being computationally easy to infer the most likely cause of activation for a single neuron, this task is far more complicated for an entire population since there are many probability densities to be evaluated simultaneously. This provokes the discussion of multiple alternative strategies of decoding population activities as they are introduced in Ma and Pouget (2009).

Decoder Functions

The main question that arises when observing activities of numerous neurons is: How does the human brain translate this population activity into estimates for a state of the world or a cognition, respectively. Theories assume the utilisation of so-called decoder functions (cf. Ma and Pouget, 2009, pp. 752–754). Mathematically, a decoder function is a mapping

$$\varphi\colon \mathbb{R}^n \to S \tag{6.9}$$

from population activity onto the estimation scale S representing a stimulus or cognition. Several of these decoders are introduced in Ma and Pouget (2009). Based on this publication, a brief overview of the most frequently discussed decoder functions shall be given at this point.

Mode Value Decoder (MVD): Based on the construction of tuning curves, the MVD assumes that it is exactly that neuron with maximum spiking frequency that is most likely to be addressed by a given stimulus or state of the world. The decoder function is thus given as

$$\varphi_{\text{MVD}} \colon \varrho(\xi) \mapsto \underset{p_j \in S}{\operatorname{argmax}} \{r_1(\xi), \ldots, r_n(\xi)\}. \tag{6.10}$$

Figure 6.5 depicts a population response for a stimulus of 3 stars (red line) together with possible estimates (green lines) for this decision. This decoder is very prone to neuronal noise and its estimates are subject to a great ambiguity which, however, diminishes for higher frequencies in neuronal responses.

Weighted Average Decoder (WAD): The WAD accounts for all responses by setting the specific frequency as a weight to the corresponding preferred value and considers its contribution to the total response. Mathematically, the WAD is given by

$$\varphi_{\text{WAD}} \colon \varrho(\xi) \mapsto \frac{\sum_{j=1}^{n} r_j(\xi) \cdot p_j}{\sum_{j=1}^{n} r_j(\xi)}. \tag{6.11}$$

As to see in Fig. 6.5, this decoder function does not produce ambiguous estimates and is very stable against neuronal noise.

Maximum Likelihood Decoder (MLD): For a given population response, the MLD chooses the estimate $\hat{s}$ with a view to maximise the likelihood function:

$$\varphi_{\text{MLD}} \colon \varrho(\xi) \mapsto \underset{s \in S}{\operatorname{argmax}} \; P(\varrho(\xi)|s), \tag{6.12}$$

where the likelihood itself is given by the i.i.d. assumption together with the Poisson probability mass function

$$P(\varrho(\xi)|s) = P(r_1(\xi), \ldots, r_n(\xi)|s) = \prod_{j=1}^{n} \frac{f_{p_j}(s)^{r_j(\xi)}}{r_j(\xi)!} \exp\left(-f_{p_j}(s)\right). \tag{6.13}$$

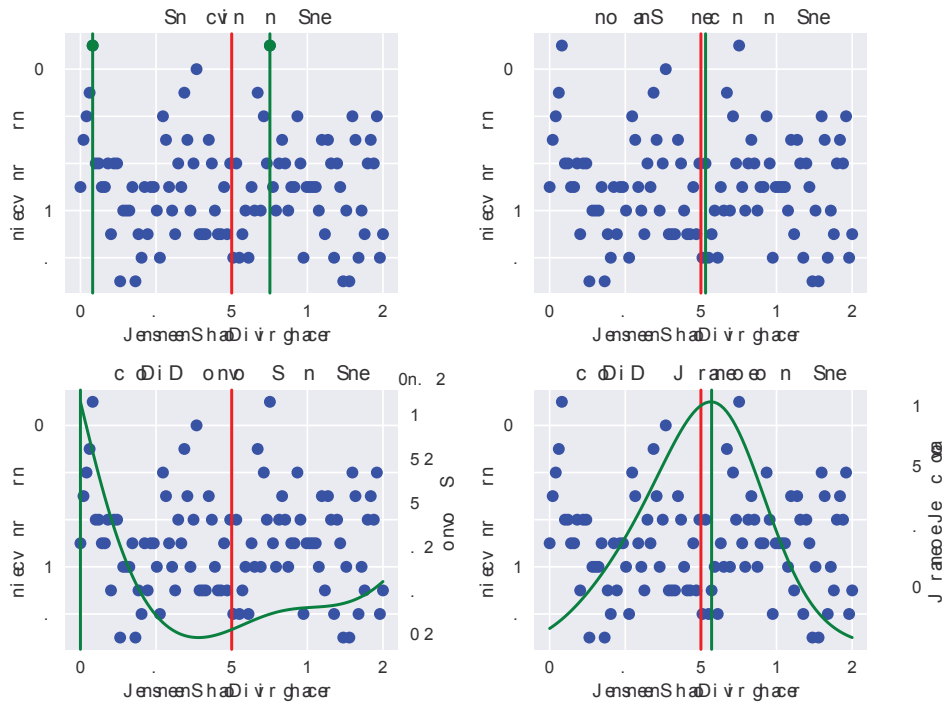


Figure 6.5: Visualisation of decoder functions for a population response constructed with $\xi = (100, 1, 1, 5, 3)$. The red and green lines show the true and the decoded stimulus.

Figure 6.5 depicts the likelihood function (green curve) for the population response together with the MLE estimates (green line). It can be seen that the estimate coincides with the boundaries of S which often occurs for lower frequencies. The MLD is the first decoder that explicitly accounts for neuronal noise through the Poisson probability.

Maximum A Posteriori Decoder (MAD): The likelihood can be transformed into a posterior probability over the stimulus via Bayes' theorem, i.e. $P(s|\varrho(\xi)) \propto P(\varrho(\xi)|s) \cdot P(s)$ where $P(s)$ denotes the prior belief about the stimulus or the state of the world that has been learned through former experiences. The estimate is then chosen to maximise the posterior, i.e.

$$\varphi_{\text{MAD}} \colon \varrho(\xi) \mapsto \underset{s \in S}{\text{argmax}} \; P(s|\varrho(\xi)) \tag{6.14}$$

The MAD is much like the MLD but with less variability since the prior works as a stabiliser. For the example depicted in Fig. 6.5, a Gaussian with $\mu = 3$ and $\sigma^2 = 0.75$ has arbitrarily been chosen as prior belief. The Bayesian brain hypothesis assumes a

prominent role of this decoder function since each population would then naturally represent a probability density over a stimulus or state of the world which integrates memory and which can easily be aggregated with other populations' densities by mere addition (cf. Beck et al., 2007). As discussed above, this was found to be a plausible description of the brain's operating principles for multiple sensory inputs (cf. Knill and Pouget, 2004).

All decoder functions represent plausible strategies for the translation of neuronal activities into real-world estimates. Unfortunately, "the manner in which probability distributions are [represented] in the brain remains unclear, and thus the neural code of uncertainty is unknown" (cf. Walker et al., 2020, p. 122). Because it is still unclear which code is indeed applied, all four decoders will be considered for the upcoming analyses and the most appropriate strategy concerning unreliable user feedback will be determined afterwards.

6.4 Neuroscientific User Model

So far, the basic model has been introduced that allows different population activities to encode one and the same stimulus. For sensory perception, this model can be seen as a noisy translation from outside reality into inside representation. For cognition, however, this model provides a possible translation from a cognitive black box into unreliable but yet measurable representations of decisions and thinking patterns. It is important to understand that this model should not be understood as a subsequent stage of transformation that distorts a reliable cognition (e.g. $s = 3$ stars) with noise. It is rather the case that the cognition itself is given as a noisy neuronal activity through population responses. In this sense, the input of a constant stimulus s is just a mathematical necessity to initiate the artificial coding process and to apply some kind of calibration which allows the resulting estimates to build a distribution around s. Therefore, this cognition model is not supposed to explain the process of choice preference itself but to solely explain the genesis of human uncertainty that comes along with it.

The good fit of this model to the phenomenon of human uncertainty has often been suggested. In particular, when repeating decision-making for a specific feedback task, neuronal noise will lead to different population responses which then lead to different estimates. These model-based estimate distributions can be seen as an equivalence to those feedback distributions measured during the RETRAIN study. This means that for

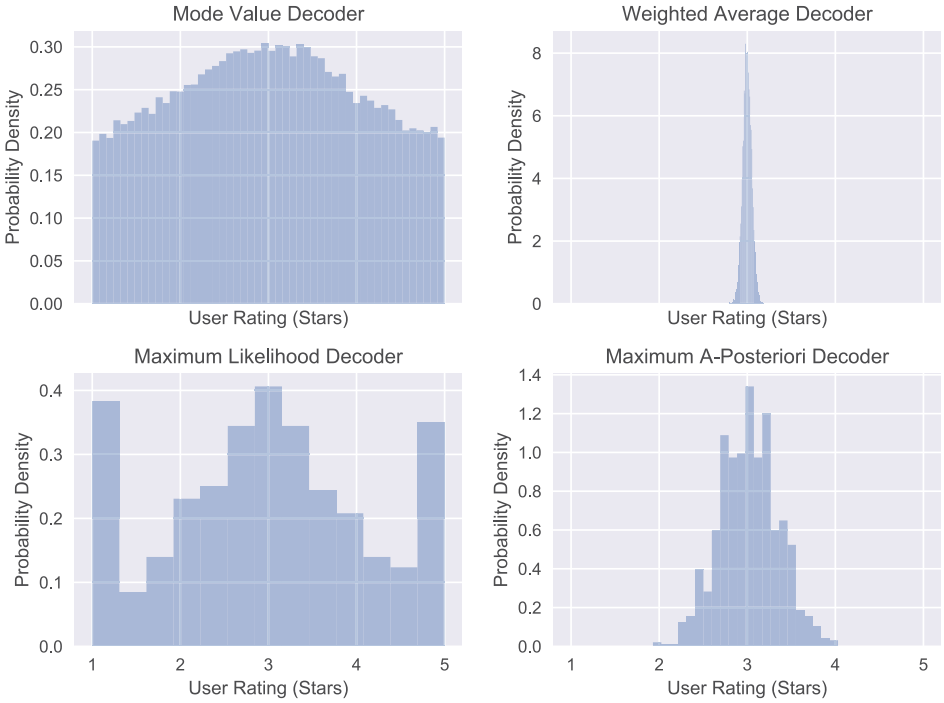


Figure 6.6: Model-based feedback distributions obtained from repeated computation of neuronal activities for $\xi = (100, 1, 1, 5, 3)$ and subsequent decoding

a particular user-item pair $\nu = (u, i)$, an artificial estimation $\widehat{\mathfrak{f}}$ of a single feedback choice can be obtained directly from the realisation of a population response, i.e. $\widehat{\mathfrak{f}} = \varphi(\varrho(\xi))$. Hence, noisy user feedback $\mathfrak{F}_\nu$ can be represented as a random variable given as

$$\mathfrak{F}_\nu = (\varphi \circ \mathcal{R})(\xi). \tag{6.15}$$

This is exemplified in Fig. 6.6. For these illustrations, 1000 estimates were generated with each decoder function based on the fixed cognition vector $\xi = (100, 1, 1, 5, 3)$. For the MVD, the vulnerability for neuronal noise is clearly visible since the corresponding feedback distribution exhibits the largest spread. Even at the boundaries of 1 and 5 stars, there is a high probability of occurrence for estimates generated by this decoder function. The resulting distribution seems only slightly more informative than a uniform distribution. Using the Bayesian definition of probability (which is a measure for one's personal confidence), such user feedback would have been provided by users who are not sure about which rating seems appropriate. For the WAD, its robustness to neuronal noise and the quality of estimation can be noticed. A user who utilises this

decoder function would surely give constant ratings. Conversely, users with larger uncertainties can probably not be modelled with this decoder. The MLD reveals a remarkable property: Due to the small size of the rating scale $S = [1, 5]$, the likelihood's maximum frequently coincides with the scale's boundaries. Therefore, this theoretical set-up might explain users who often only choose between these boundary ratings. At first glance, the MAD provides the most plausible feedback distributions when compared to those measured in the RETRAIN experiment. This seems to strengthen the Bayesian brain hypothesis.

The challenge of this neuroscientific user model is to find a specific cognition vector ξ_ν for each user-item pair $\nu = (u, i)$ along with a decoder function φ, so that the model-based feedback $\widehat{\mathfrak{F}}_\nu$ minimises the difference to the real user feedback $\mathfrak{F}_\nu$ in terms of an arbitrary similarity metric d. Mathematically, this user model is given by $\mathfrak{F}_\nu \equiv (\xi_\nu, \varphi)$ together with the specification

$$\begin{aligned}(\xi_\nu, \varphi) &:= \underset{(\xi,\varphi)}{\operatorname{argmin}}\ d\big(\mathfrak{F}_\nu\,,\,\widehat{\mathfrak{F}}_\nu\big) \\ &= \underset{(\xi,\varphi)}{\operatorname{argmin}}\ d\big(\mathfrak{F}_\nu\,,\,(\varphi\circ\mathcal{R})(\xi_\nu)\big).\end{aligned} \tag{6.16}$$

The advantage of this model is that there is a cognition vector from a five-dimensional vector space associated with each user-item pair. This vector space therefore contains five times more information than the standard case (one variable per user-item pair, i.e. the rating itself) or respectively, 2.5 times more information than a statistical user model (two variables per user-item pair, i.e. mean and variance). At this point, it is reasonable to ask why it is advantageous and expedient to expand the distribution's properties for each user-item pair, that is the mean and variance (i.e. a two-parameter model) into a model with five parameters. In other words, where does the additional information come from? Well, this information comes from the internal processes as they are proposed in neuroscience research. One source of additional information is, for example, the number and shape of tuning curves as well as the Poisson-like neuronal noise which is dependent on a particular response frequency rate. Furthermore, the subsequent data aggregation done by the decoder functions also influences the shape of feedback distributions and thus contributes to a possible information gain. In short: All these components provide enough information to support the extension of a two-dimensional space into five dimensions by incorporating the latest neuroscientific insights into the human brain.

user model	feedback representation	dimensionality
standard	$\mathfrak{F}_\nu \equiv \mathfrak{f}_\nu$	1
repetition	$\mathfrak{F}_\nu \equiv (\mathfrak{f}_\nu^1, \ldots, \mathfrak{f}_\nu^5)$	5
statistical	$\mathfrak{F}_\nu \equiv (\mu_\nu, \sigma_\nu)$	2
neuroscientific	$\mathfrak{F}_\nu \equiv (n_\nu, g_\nu, w_\nu, o_\nu, s_\nu)$	5

Table 6.1: User models for (uncertain) feedback representation

Application

To apply this user model as it is given by Eq. 6.15 and Eq. 6.16, one needs to transform each measured feedback distributions into a particular cognition vector. To be more precise, one needs to associate a unique cognition vector ξ_ν to each user-item pair $\nu = (u, i)$ in such a way that this vector produces a model-based feedback distribution $\widehat{\mathfrak{F}}_\nu$ which comes as close as possible to the real distribution $\mathfrak{F}_\nu$ with respect to a particular decoder function φ. This cognition vector is henceforth the representation of a user's opinion about an item. An overview of all conceivable user models and their respective feedback representation as well as the feature space dimensionality is given in Tab. 6.1.

The best way to determine a cognition vector would certainly be a closed mathematical formalism in terms of a certain transformation $\tau\colon \mathfrak{F}_\nu \mapsto \xi_\nu$ or, alternatively, an efficient approximation. However, Ma and Pouget claim that a purely mathematical approach is not possible for probabilistic population codes (cf. Ma and Pouget, 2009, p. 751). This statement seems reasonable for the following reasons: First of all, the response of each neuron is given as an individual random variable. Thus, convolutions must be calculated for often more than 200 independent and non-identical variables (i.e. neurons), which can hardly be described in a closed form. Then again, most decoder functions are designed to work only with realisations of random variables. When focusing on the random variable rather than on its realisations, the likelihood (needed for the MLD and MAD) would simply become the multi-dimensional joint density $P(\mathcal{R}(\xi)|s) = P(R_1(\xi), \ldots, R_n(\xi)|s)$ whose maximum may be ambiguous modulo marginal distributions. Moreover, the joint density of 200 variables can impossibly be captured in a closed formalism as well. Another approach might be to approximate the feedback distributions by Gaussians and find functional dependencies for their sufficient statistics, that is $\mu = \mu(n, g, w, o, s)$ and $\sigma = \sigma(n, g, w, o, s)$, respectively. This solution

is mainly undermined by model complexity and altering functional dependencies to the neuronal parameters whenever one single parameter is (only slightly) changed.

In the absence of a closed mathematical formalism, one has to rely on large-scale simulations. A simple way is to translate a predefined family of cognition vectors into model-based feedback distributions and compare these to the original user feedback in a brute force manner. This procedure requires the following steps of pre-operation:

- Determining reasonable boundaries for each neuronal variable in the PPC model and generate a finite set of corresponding cognition vectors.
- Choosing the right combination of similarity distance and decoder function that will produce the best fit.

These tasks will be done in the following sections. After "learning" the correct model properties in this preliminary work, they will be applied on the full RETRAIN data record to find the best fitting cognition vectors. To maintain the neuroscientific foundation of this model, it will be examined whether its implications are consistent with the latest biological or medical findings published in the scientific literature.

6.5 Parameter Boundaries

Unfortunately, there was no quantitative information on the parameter boundaries in any published paper about probabilistic population codes. From Pouget et al. (2000); Doya et al. (2007); Ma and Pouget (2009), however, some parameters can be reconstructed from corresponding graphics. In this regard, initial parameter settings are

$$n = 10 \quad ; \quad g = 1 \quad ; \quad w = 1 \quad ; \quad o = 5. \tag{6.17}$$

Starting from this basis, attempts were made to reproduce randomly selected feedback distributions on a single machine with different decoder functions. In doing so, reasonable parameter boundaries have been found that would work optimally with all decoders. These empirically determined parameter ranges are:

$$10 \leq n \leq 250 \quad ; \quad 1 \leq g \leq 100 \tag{6.18}$$

$$0.1 \leq w \leq 2.0 \quad ; \quad 0 \leq o \leq 15 \tag{6.19}$$

For the further analyses, especially for the upcoming user classification task, individual sets N, G, W, O of 100 equidistantly distributed values will be computed for each parameter range determined above. A single cognition vector can hence be regarded as an element $\xi \in N \times G \times W \times O \times \{s\}$ where the stimulus $s := \mathbb{E}(\mathfrak{F}_\nu)$ is passed to the model as the expectation of the corresponding feedback distribution.

6.6 Similarity Metrics

The comparison of distributions through statistical testing is out of question for it only supports dichotomous decisions. Rather, a quantitative measure is needed to indicate the degree of similarity. In the field of machine learning, the Jensen-Shannon-Divergence (based on the Kullback-Leibler-Divergence) is often used for this purpose. This distance is formulated for discrete probability distributions as well as for continuous densities. This reflects the present modelling since it relies on discrete data (and hence needs appropriate analysis), but it also assumes the existence of an underlying continuous density (whose analysis requires different methods). Both formalisms can be covered with this distance. In addition, metrics from the field of psychometry will also be examined. Standard measures for similarity are, for example, Cohen's Kappa (typically used to determine inter-rater reliability) (cf. Döring and Bortz, 2016, pp. 567–568) and Cohen's D (often denoted as effect size) (cf. Döring and Bortz, 2016, pp. 816–819).

Jensen-Shannon-Divergence: A definition of the Jensen-Shannon-Divergence has already been given by Eq. 4.25, but shall now be repeated to improve comprehensibility: Let X be a probability space and let P, Q be discrete probability mass functions on X. The Kullback-Leibler-Divergence is then defined as

$$\mathrm{KL}(P, Q) := \sum_{x \in X} P(x) \cdot \log\left(\frac{P(x)}{Q(x)}\right) \tag{6.20}$$

(cf. Lee, 2000, Ch. 2). The Kullback-Leibler-Divergence has some major disadvantages: By definition, it is not symmetric and has no upper bound. These issues are corrected by the Jensen-Shannon-Divergence (JSD) which is defined via

$$\mathrm{JSD}(P, Q) := \frac{1}{2}\,\mathrm{KL}(P, M) + \frac{1}{2}\,\mathrm{KL}(Q, M) \tag{6.21}$$

where $M = \frac{1}{2}(P + Q)$ (cf. Lee, 2000, Ch. 2). For the continuous case, the JSD is defined analogously. When using the binary logarithm for the Kullback-Leibler-Divergence, the JSD can be shown to be bound by $0 \leq \text{JSD} \leq 1$ (cf. Lin, 1991, pp. 147–148) which is advantageous when evaluating the overall quality of model-based feedback distributions. In this thesis, the JSD is implemented in three different ways:

The pureJSD has no additional assumptions about the RETRAIN data. This means that the probability mass function of the real data is given as the relative histogram of user ratings. To maintain similar discretisation (i.e. the same bins) the model-based feedback has to be rounded to the next integer rating before determining the respective relative frequencies. The JSD is then computed by Eq. 6.21.

The nJSD uses the metrologic user model and works with assumptions of normality for both, the real user feedback as well as the model-based feedback. In doing so, the mean and standard deviation of the corresponding feedback sets are computed and then a Monte-Carlo sample of 10^4 trials is created from the respective normal distributions. In other words, the low number of repeated ratings is completed using the normality assumption to make the JSD outcomes continuous, thus supporting an improved ranking procedure for all model variations. To keep the computation executable, standard deviations of less than 0.007 have to be prohibited, i.e. lower values are fixed to this specific limit. The subsequent discretisation is done according to the original rating scale as with the pureJSD. The JSD is then computed by Eq. 6.21.

The JSD(b) is the generalisation of the nJSD which allows to discretise the normality-completed feedback distributions into an arbitrary amount b of bins rather than breaking it down to the five original bins. This further supports the continuity of the JSD and hence the extinction of ambiguous scores. For the upcoming evaluations, $b = 50$ bins and $b = 200$ bins will be used. The corresponding metrics will be referred to as **JSD50** and **JSD200**, respectively.

These metrics are gradually diverging from discreteness. At first, the feedback distributions are implemented through relative histograms applying the original rating scale $S = \{1, 2, 3, 4, 5\}$ for discretisation as it was also used in the RETRAIN study. The next approach complements the original five ratings into 10^4 ratings under normality assumption while maintaining the same discretisation of S. One step further, this discretisation is re-

placed by a new one where b equidistantly spread bins are chosen within the interval $[1, 5]$. This procedure guarantees that the JSD scores will not cluster into a small number of possible values which would prevent rankings due to ambiguity.

Cohen's Kappa: Cohen's Kappa is often used in psychometry to determine the inter-rater reliability, i.e. whether two (or more) independent observers (raters) do assign similar categories to the same objects (cf. Döring and Bortz, 2016, pp. 567–568). In doing so, the relative matching frequency p_0 of both raters is compared to the probability of a random matching p_c via

$$\kappa = \frac{p_0 - p_c}{1 - p_c}. \tag{6.22}$$

To compare two independent feedback realisations, the real user can be considered as one rater and the cognition model can be considered as the second rater. Each user and each cognition vector will associate an item (object) with specific feedback (category) so that a kappa score can be computed for multiple rating trials. A score of $-1 \leq \kappa < 0$ means that random guessing performs better than using a specific cognition vector and $0 < \kappa \leq 1$ indicates the degree of advantage when using a cognition vector rather than random guessing. When computing Cohen's Kappa, the reference through random guessing can be chosen to be either informed or uninformed. The uninformed guessing is based on the uniform distribution whereas the informed guessing is drawn from the marginal distributions

$$P(X = i) = \sum_{j=1}^{k} P(X = i, Y = j) \tag{6.23}$$

$$P(Y = j) = \sum_{i=1}^{k} P(X = i, Y = j) \tag{6.24}$$

of the joint probability $P(X = i, Y = j)$ that the first rater associates category i while the second rater assigns category j. Here, k denotes the total number of available categories. It would actually be useful to apply an informed guessing procedure since the assignment from either the user or the cognition vector is according to a particular non-uniform distribution. However, uniform guessing is (initially) used for simplicity. The only effect of this simplification might be that the calculated Kappa scores are higher because it is compared against a worse reference. If further results prove this metric to be superior to all the other distances, informative guessing can still be

applied to yield optimised scores. When using uninformed guessing p_c simply reduces to the Laplace probability $p_c = |A|/|B|$ where B is the set of all re-rating vectors $R_{\nu,\bullet} = (R_{\nu,1}, \ldots, R_{\nu,5})$ with $R_{\nu,t} \in \{1, \ldots, 5\}$ for each rating trial t. This set has the cardinality $|B| = 5^5 = 3\,125$. Since the sequence order is irrelevant in terms of the resulting distribution, A must be defined as the set of a particular re-rating along with all its permutations. This set has the cardinality $|A| = 5! = 120$. It follows that $p_c = 120/3125 = 0.0384$.

For the technical implementation there is the re-rating $R_{\nu,\bullet} = (R_{\nu,1}, \ldots, R_{\nu,5})$ on the one side and $m \equiv 0 \bmod 5$ model-based estimations $e = (e_1, \ldots, e_m)$ on the other. The relative frequency p_0 of agreement is then computed by rounding each component of e to an integer and then separating e into $m/5$ chunks of the same size as $R_{\nu,\bullet}$ to check for an agreement modulo permutation. Consequently, this measure stabilises only for large m of model-based estimations. This implementation is still very close to the original case since it works on the untransformed user feedback.

Cohen's D: When using hypothesis testing, any existing effect can be made significant by increasing the sample size. However, significance gives no indication regarding the extent of an existing effect. Therefore, psychologists also consider measures of effect size to determine how large or small a certain effect really is. The most common measure is Cohen's D which considers the difference of two distribution means along with the distribution variances. Cohen's D is defined as

$$D := \frac{\mu_2 - \mu_1}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}} \tag{6.25}$$

(cf. Cohen, 1988, p. 44) and has no upper bound by definition. However, an upper bound can be computed for the present rating scenario since a bounded rating scale is utilised together with five re-ratings. From these conditions it follows that $1 \leq \mu_1 \leq 5$ and $1 \leq \mu_2 \leq 5$ for both distribution means. The Bessel-corrected standard deviation for the user feedback is limited by $0.45 \leq \sigma_1 \leq 2.19$ while the lower bound for the model-based feedback is bounded by zero as some cognition vectors might produce a set of constant ratings. However, since this induces computational issues, the standard deviation is set to $\sigma_2 = 0.07$ for this is the largest variance that still produces constant predictors when rounding corresponding pseudo-random numbers to full integers. Accordingly, the

maximum value for D is given by

$$D_{max} = \max_{(\mu_1,\mu_2,\sigma_1,\sigma_2)} \frac{\mu_2 - \mu_1}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}} \tag{6.26}$$

and the normalised version of Cohen's D can be calculated qua $D_n := D/D_{max}$. This score is based on the assumption of normality for the user feedback and the model-based feedback, respectively.

Monte-Carlo Impact on Classification

One problem of the initial classification algorithm using the pureJSD and Cohen's Kappa has been its unreliability. This was not apparent at first glance since the computation took about several weeks and multiple repetitions have thus been out of question at this early stage. It took months to discover that the first promising results reported in Jasberg and Sizov (2018b) could not be reproduced. The essence of this flaw lies within the Monte-Carlo approach that is utilised to represent probability densities. The utilisation of pseudo-random numbers inevitably results in fluctuations when the same task is repeated. Similarity must hence be seen as a distributed random quantity rather than a single score. This justifies a deeper elaboration of unreliable user classification when Monte-Carlo approaches are involved.

For analysing the effect of pseudo-random numbers on the reliability of user classification, a reference distribution R is defined as the vector of re-ratings

$$R = (1, 1, 5, 5, 5) \qquad \text{with} \qquad m_R = 3.4 \quad \text{and} \quad s_R \approx 2.19 \tag{6.27}$$

together with a family of normally distributed random variables

$$(C_j)_{j=1,\dots,9} \sim \mathcal{N}(\mu_j, \sigma) \qquad \text{with} \qquad \mu_j = 1 + j \cdot \frac{5-1}{9} \tag{6.28}$$

representing distorted copies of R. Here, $\sigma = s_R \approx 2.19$ is chosen as the common standard deviation for the reference and its distorted copies to make the perfect match fully dependent on the mean only. On this basis, the metric score $d(R, C_j)$ is computed a thousand times to get a score density for each C_j so that their intersections can be compared.

The results of this analysis are depicted in Fig. 6.7. The x-axis of this violin plot represents the distribution means μ_j for all distorted copies C_j of R. The violin itself is a kernel density estimation of 1000 scores $d(C_j, R)$ that has been mirrored at the centre line. Consequently, the y-axis shows the range of similarity scores. The red vertical line represents the mean value $m_R = 3.4$ of the reference R. A well-working similarity distance should hence maximise (Cohen's Kappa) or minimise (all other metrics) for C_6 with $\mu_6 = 3.5$. During the initial user classification, only Cohen's Kappa and the pureJSD had been used as a measure for similarity. All the other metrics were added subsequently in an attempt to develop a new measure that is insensitive to the Monte-Carlo uncertainty. The resulting distributions of Cohen's Kappa and the pureJSD have strong overlaps. So if the classification is repeated with otherwise identical input data, different mean values and thus different copies C_j may be identified as the best approximation to R.

Analogous to the error probabilities when selecting the best recommender system (cf. Ch. 4), the same effect is likely to occur here as well. The error probabilities for the case that C_i is a better fit than C_j according to Cohen's Kappa or the pureJSD, respectively, can be seen in Tab. 6.2a and Tab. 6.2b. Considering the probabilities of choosing $\mu_6 = 3.5$ as the best fitting mean, Kappa holds high chances of error in direct comparison with μ_4, μ_5, μ_7, μ_8 and μ_9 which are all in the region of 50%. Accordingly, a classification with this similarity score turns out to be extremely arbitrary. For the pureJSD, there is a slightly different situation. The chance of error that $\mu_6 = 3.5$ fits better than other means is especially high for μ_7, μ_8, and μ_9. A classification in terms of the pureJSD together with a Monte-Carlo approach is therefore too random to be practicable as well.

However, there is a slightly more problematic feature of the pureJSD, namely the strong bias towards larger means. This becomes evident in Tab. 6.2b in which the error probabilities increase monotonically in each row. In Fig. 6.7, this bias can be recognised as monotonically decreasing scores. Moreover, this bias is also dependent on the standard deviation of the measured densities. In Fig. 6.8, for example, the same analysis has been performed but with the standard deviation $\sigma = 1.0$ for each copy C_j rather than the maximum which is considered above. One can observe a different bias towards the middle of the scale.

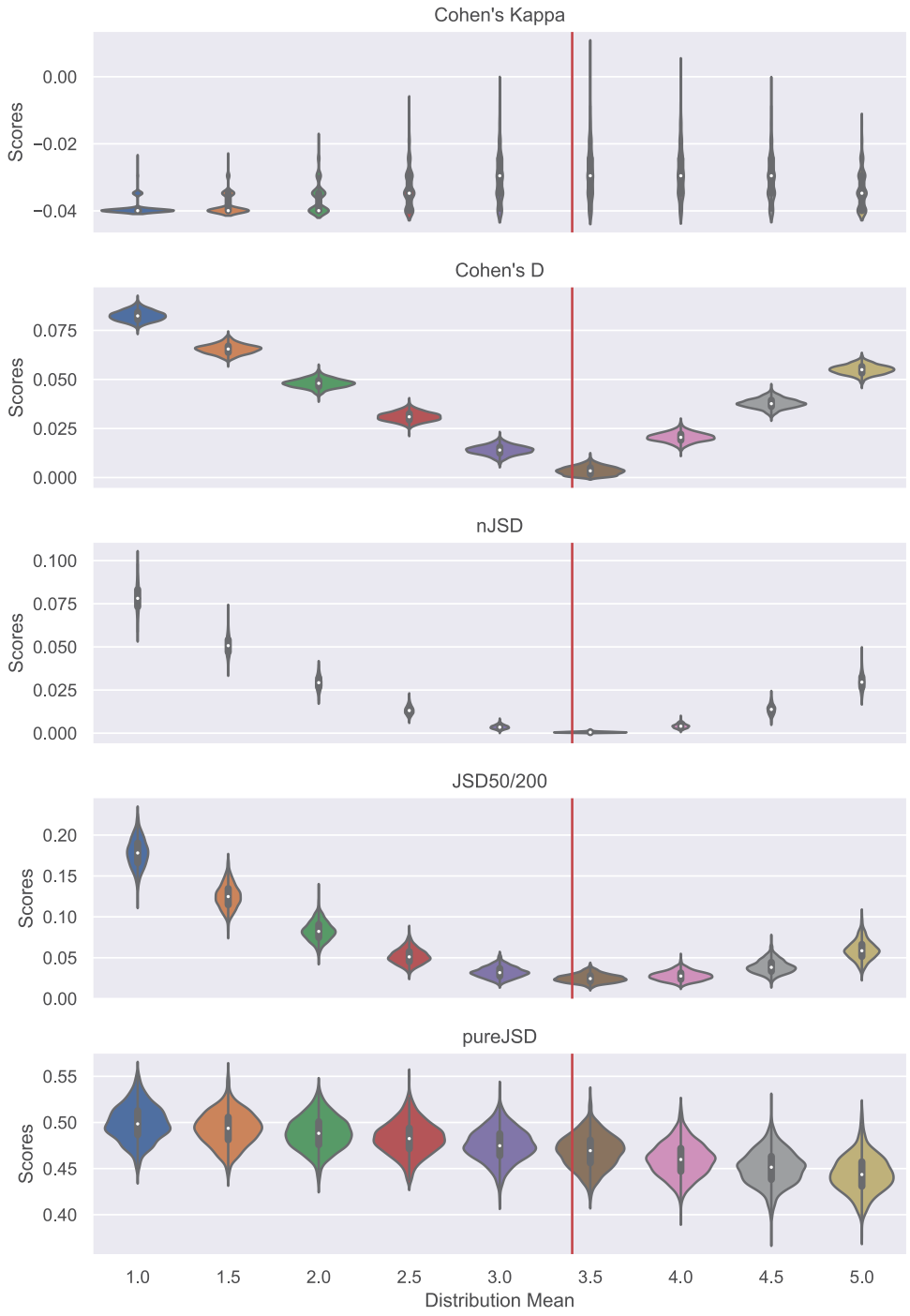


Figure 6.7: Reliability of metric scores: The similarities between the reference distribution and each of its nine copies (with corresponding means on the x-axis) can be regarded as distributions (represented by violin plots). The true mean of the reference is marked as a red vertical line.

μ_i/μ_j	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	–	0.90	0.93	0.96	0.97	0.99	0.98	0.98	0.96
1.5	0.76	–	0.86	0.91	0.95	0.96	0.96	0.95	0.92
2.0	0.55	0.61	–	0.80	0.88	0.91	0.92	0.87	0.81
2.5	0.33	0.39	0.50	–	0.74	0.79	0.81	0.74	0.64
3.0	0.19	0.24	0.35	0.48	–	0.67	0.70	0.61	0.49
3.5	0.15	0.19	0.28	0.40	0.52	–	0.62	0.53	0.41
4.0	0.12	0.16	0.26	0.39	0.52	0.59	–	0.51	0.39
4.5	0.20	0.25	0.35	0.49	0.61	0.67	0.69	–	0.50
5.0	0.33	0.39	0.50	0.63	0.74	0.79	0.80	0.73	–

(a) Cohen's Kappa

μ_i/μ_j	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	–	0.59	0.66	0.74	0.82	0.89	0.94	0.96	0.98
1.5	0.41	–	0.58	0.67	0.75	0.84	0.91	0.93	0.97
2.0	0.34	0.42	–	0.60	0.69	0.80	0.87	0.91	0.96
2.5	0.26	0.33	0.40	–	0.60	0.72	0.82	0.87	0.94
3.0	0.18	0.25	0.31	0.40	–	0.63	0.74	0.8	0.90
3.5	0.11	0.16	0.20	0.28	0.37	–	0.62	0.70	0.83
4.0	0.06	0.09	0.13	0.18	0.26	0.38	–	0.59	0.74
4.5	0.04	0.07	0.09	0.13	0.20	0.30	0.41	–	0.66
5.0	0.02	0.03	0.04	0.06	0.10	0.17	0.26	0.34	–

(b) pureJSD

Table 6.2: Error probabilities for the statement that C_i with expectation μ_i is a better fit than C_j with expectation μ_j

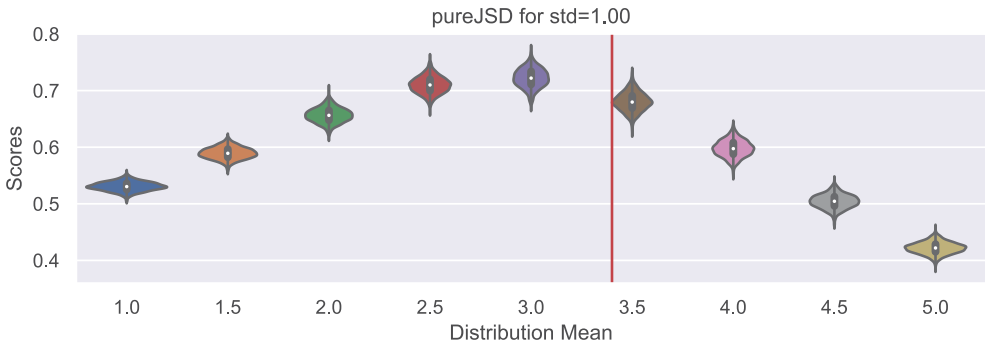


Figure 6.8: pureJSD bias towards the scale's midpoint when all the copies C_j share the common standard deviation of $\sigma = 1.0$

The same analysis shall be repeated for the standard deviation, i.e. distorted copies are defined by

$$(C_j)_{j=1,\dots,9} \sim \mathcal{N}(\mu, \sigma_j) \qquad \text{with} \qquad \sigma_j = 0.07 + j \cdot \frac{1.96 - 0.07}{9} \tag{6.29}$$

with $\mu = m_r = 3.4$ as the common mean in order to make the metric score solely dependent on the standard deviation. The similarity to R is computed for each metric d and each copy C_j a thousand times to yield score distributions. These can be seen in Fig. 6.9. The x-axis represents the standard deviation σ_j and the vertical red line represents the standard deviation of the reference. In this analysis, yet another bias can be recognised, i.e. Cohen's D minimises for small standard deviation and thus chooses the minimum σ_1 along with C_1 as the best fit for R. This means that only the nJSD or the JSD50 (or JSD200 respectively) can be considered as adequate metrics. These metrics only weaken the Monte-Carlo effect but it is still existent and might impair classification reliability, however to a smaller extent. To fully eliminate the effects of Monte-Carlo uncertainty, one has to freeze the so-called random seed. The random numbers used to represent probability densities are, in fact, not random but deterministic. Based on a previously stored start value, an algorithm generates a so-called pseudo-random number. By explicitly setting the seed, the start value is overwritten which serves as the basis for calculating the next pseudo-random number. If the random seed is reset to the same value each time a set of random numbers is computed, the same samples will always be obtained which represent a certain distribution. This in return removes all Monte-Carlo uncertainty. For the upcoming classification, the random seed has been set to 41 856[1]. Now the question arises as to whether this determination does not nullify the randomness based on neuronal noise that should be modelled. This would be the case if the seed is reset after every single random number. In fact, more than 10^4 random numbers will be drawn at once. This will result in distributions that are completely based on simulated neuronal noise and this is exactly what this model is supposed to do: Explain a possible mechanism of translating neuronal noise into feedback uncertainty. Of course, this question can also be extended: Is it realistic that the resulting distributions are constant? It can be assumed that the state of all neurons

[1]The number 41 856 was chosen due to my love for aviation but also as an acknowledgement to Prof Dr Laura Kallmeyer whose generosity supported me to attend an international conference and to present my recent research. The hexadecimal notation of the number 41 856 is A380 which was the aeroplane with which I flew to this conference.

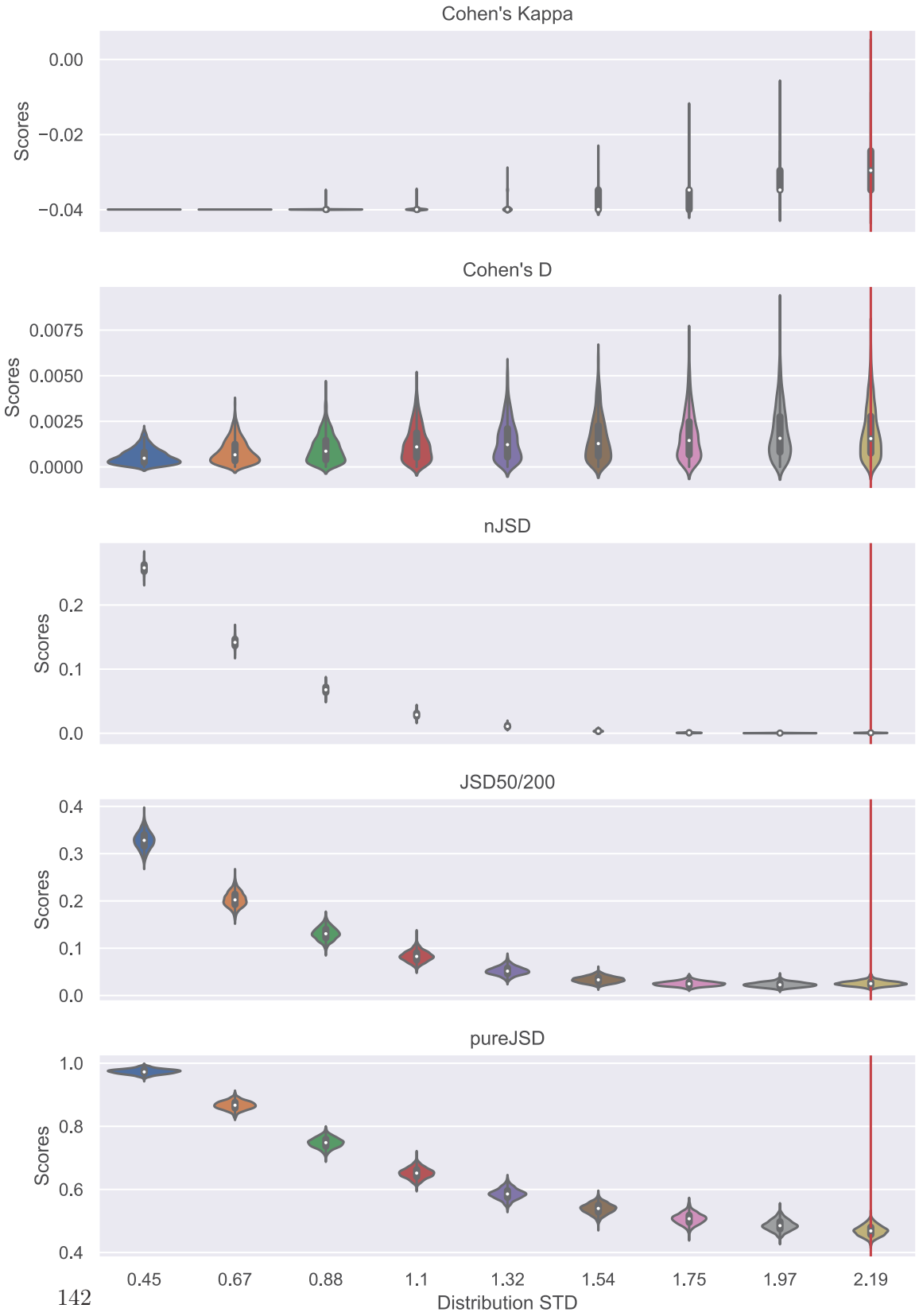


Figure 6.9: Reliability of metric scores

involved in a repeated decision-making is never completely identical. Accordingly, even the distributions must be subject to natural fluctuation. Then again, this fluctuation is reduced by a higher number of Monte-Carlo trials. In consequence, the representation of probability densities through pseudo-random numbers is just a required technical approach. If this technical trick brings additional uncertainty to the table, then it is legitimate to reduce it as long as the genesis of feedback distributions is still ensured.

Although it is already evident that Cohen's D and the pureJSD are no good choices for classification, they will nevertheless be investigated in the upcoming small-case scenario due to comprehensiveness, i.e. to examine their properties without Monte-Carlo uncertainty in a real use case.

6.7 Fitting User Behaviour

In the last two sections, preliminary work was done to enable user classification through computer simulation. In Sec. 6.5, reasonable subsets of $\mathbb{R}$ were identified in which the individual neuronal parameters can be located. In Sec. 6.6, possible metrics were analysed and possible risks and solutions for classification reliability were discussed. On this basis, these metrics can now be used to associate a cognition vector to each feedback distribution in such a way that this vector, in conjunction with a decoder function, optimally reproduces this distribution.

The essential simulations are computationally expensive, especially if they have to be done for every combination of decoder function and metric. This will be clarified by an example: If an equidistant partition with ten values is chosen for each domain N, G, W, O of a neuronal parameter, this will result in 10^4 different combinations of cognition vectors. These must be checked for each of the six metrics and four decoder functions for all 335 user-item pairs. Therefore, $10^4 \cdot 6 \cdot 4 \cdot 335 = 80\,400\,000$ individual computations have to be made. When 10^6 Monte-Carlo trials are used to represent both, the model-based feedback as well as the real user feedback, the computation ultimately involves $1.608 \cdot 10^{14}$ float64 numbers. With 4 bytes per float, the entire data volume that has to be generated sums up to 643 200 terabyte or rather 643.2 petabyte. At this point it must be realised that the necessary simulations can not be calculated in a time-efficient manner, even with intelligent calculation sequencing, buffering of frequently occurring values and reasonable parallel computing.

Decoder	MVD	WAD	MLD	MAD	TOTAL
Runtime [in s]	2.36	0.61	788.99	785.77	1577.73

Table 6.3: Runtime analysis for each decoder function processing ten representative cognition vectors without considering the runtime for computing a metric

Similar calculations can be made concerning the runtime. In doing so, a model-based feedback density is computed for each decoder function processing ten representative cognition vectors. These vectors are in particular

$$\Big\{(n, 20, 1, 5, 3) : n \in \{25, 50, 75, 100, 125, 150, 170, 200, 225, 250\}\Big\} \tag{6.30}$$

and these are representative since the runtime is solely dependent on the population size n which is altered exactly the same way as it will be done in the classification later. The runtime results can be seen in Tab. 6.3. The total runtime of all ten vectors sums up to 26.3 minutes. If this is extrapolated to the 10^4 different combinations of cognition vectors, one yields a runtime of 18.3 days for each of the 335 user-item pairs, resulting in a total of 16.5 years. To get this classification task done in one week (i.e. the maximum runtime allowed on the university's high-performance cluster HILBERT[2]), one would need 874 computing nodes working in parallel which is hard to realise even on HILBERT. Before downscaling the simulation, many attempts were made to realise this classification without loss of information. Some of these attempts involved at least one of the following solution approaches:

Computation Order: The feedback distributions in the form of Monte-Carlo trials should not be recalculated for each metric. Instead, all six metrics are calculated for each feedback distribution at once which reduces costly redundancy.

Caching: Another way to avoid redundancy is to cache frequent values, that is, to keep them in the main memory or to store them into an extra file and retrieve them whenever needed (instead of recalculating). This method is used for the Bayesian decoder functions where values of the Poisson density have to be computed for repetitive arguments and additional parameters.

[2]HILBERT is the high-performance cluster (HPC) maintained by the Centre for Information and Media Technology at the University of Duesseldorf (Germany). Further information is available at `https://wiki.hhu.de/display/HPC/Wissenschaftliches+Hochleistungs-Rechnen+am+ZIM` (last accessed on Jun 21, 2020).

Parallelisation: A highly efficient way of economising runtime is to use a high-performance cluster. For the research described in this chapter, the university's HILBERT HPC has been employed. These computing nodes can be virtually assembled with a variable number of CPU cores and RAM. In addition to the actual parallelisation by computing nodes, this also allows the use of multiprocessing on each node. In the beginning, 200 cores were requested per computing node so that each core only had to compute $10^4 / 200 = 50$ cognition vectors for each user-item pair. However, the high number of cores led to the situation that only one computing node was running at a time. Consequently, the 335 user-item pairs were still calculated sequentially. Additionally, the applied usage restrictions concerning the computing infrastructure entailed certain delays in initialising subsequent computing nodes.

At this time, the cognition vectors were not yet optimally sorted as they are in Eq. 6.30. This is essential, however, as the runtime depends on the number of neurons within a population. If the distribution of cognition vectors across all processes is not optimal, some processes would terminate within minutes while others would take days to finish. Considering the optimal runtimes in Tab. 6.3, the MVD would have finished in 65.88 minutes but took five days in reality. Theoreticaly, the WAD would have needed 17.03 minutes but in fact took one day to compute. Both, the MLD and the MAD, would have taken around 15 days to compute (exceeding runtime). In short, the improvements made so far are not yet sufficient to realise an efficient classification.

Node/Core-Balance: In order to fulfil given technical regulations, the number of cores has been reduced so that several nodes can run simultaneously. The optimal core number was advised to be 20 which resulted in 500 cognition vectors for each core. Unfortunately, none of the initialised nodes finished within the maximum computing time when using this configuration. Further partitioning of the input data for each node (i.e. smaller than a single user-item pair) would have been possible but would have also caused many inconveniences to adequately prepare the data set (i.e. proper partitioning and merging). After a better distribution of computationally expensive cognition vectors (in terms of population size) was achieved, this approach could be successfully implemented.

Common Data Pool: The most effective step was indeed to improve the distribution of computationally expensive cognition vectors. This was initially solved by a common data pool that all processes could access simultaneously. The advantage is that cores do not remain unused after completing the calculation of an allocated set of cognition vectors. The expensive cognition vectors are thus processed by all 20 cores rather than being allocated to only two or three cores which results in exceeding the maximum runtime. This first version of a common data pool with pointers and record locking was very cumbersome to adjust even for small changes of the classification approach. Therefore, in the final version, the data pool was split as usual and allocated to the individual cores, but all cognition vectors have been shuffled before. Shuffling the data will cause expensive vectors to be distributed equally across all cores. This solution is both practical and very simple, and its runtime is similar to the original solution.

Random Number Generation: Fixing the random seed requires a new data structure for repeated neuronal representations. Instead of computing a single estimator from a list of n random numbers and repeating this m times, the new implementation requires to generate $m \cdot n$ random numbers at once and to store them into an $(m \times n)$ - matrix. Each row is then aggregated into an estimate, creating an m-dimensional vector which represents the underlying probability distribution. This eliminates $m - 1$ individual initialisations of the random number generator and reduces runtime considerably.

All these solutions made the computation rather efficient in the end but still resulted in either exceeding the maximum RAM or the maximum runtime. At this point, the decision was made to overcome this obstacle by analysing the metric-decoder interplay on a small but representative data record. After "learning" the optimal model settings, these are then used for a subsequent simulation on the full RETRAIN record.

Small Case

The key to analysing the best system configuration on a small dataset is to make this dataset as representative as possible so that the analysis and the corresponding quality of the configuration are valid even on larger datasets. Since the neuroscientific approach is intended to model human uncertainty, this uncertainty also becomes an indicator of the representative quality. Hence, a small dataset is selected based on the

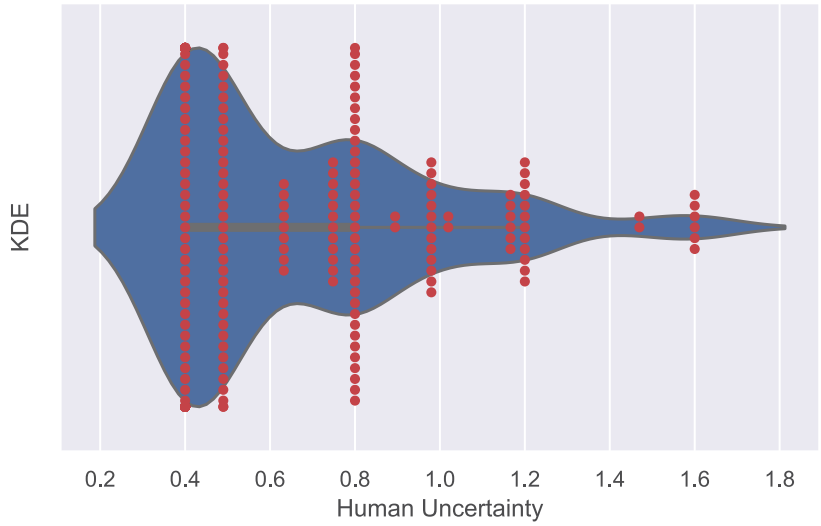


Figure 6.10: Violinplot and swarmplot of non-vanishing human uncertainty within the RETRAIN record

uncertainty (i.e. the standard deviation) of the feedback (distributions) measured in the RETRAIN study. Figure 6.10 depicts the kernel density estimation of the non-vanishing human uncertainty (since the present model is not designed for $\sigma = 0$) in a violin plot. Mirroring the actual KDE along the centre axis makes smaller features of the density's shape more visible. In this plot, four accumulations of human uncertainty can be seen which are in particular

$$\begin{aligned} A_1 &= (0\,,\,0.6] & A_2 &= (0.6\,,\,1.1] \\ A_3 &= (1.1\,,\,1.4] & A_4 &= (1.4\,,\,1.8] \end{aligned} \tag{6.31}$$

where A_3 and A_4 are unified for the subsequent selection procedure due to their small cardinalities. The goal is to select a dataset incorporating each of these accumulations. For each of these accumulations, three feedback distributions are randomly selected and visually compared. This process is to be repeated as long as these distributions differ significantly by eye. The random selection is used to speed up this process by not having to look at each user-item pair individually. For example, A_1 comprises 115 distributions, A_2 still has 72 distributions, and even the cardinality of $A_3 \cup A_4$ is 26. During this process, the set

$$\big\{(26,2),(5,0),(59,3),(58,4),(24,1),(47,0),(27,0),(62,0),(24,3)\big\} \tag{6.32}$$

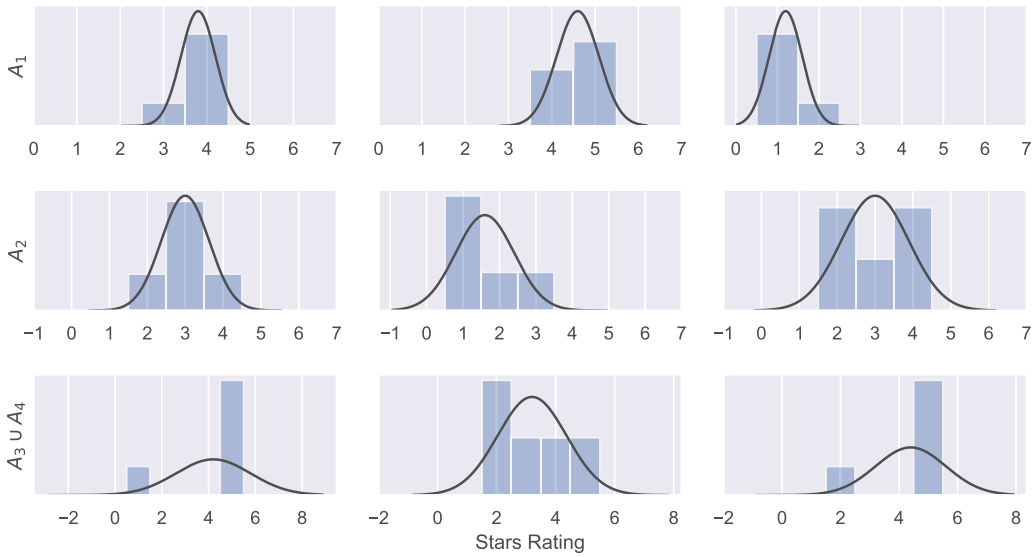


Figure 6.11: Representative feedback distributions constituting a small data set for classification. Each row depicts three feedback distributions from a specific accumulation.

of user-item pairs (u, i) was chosen to represent the totality of possible distributions. Those distributions are also shown in Fig. 6.11 in which the pairwise difference in human uncertainty (standard deviation) can be clearly seen.

It is exactly this small dataset with which the classification is performed for all combinations of decoder function and similarity metric. For the technical implementation, an equidistant discretisation into five numbers is applied for each parameter space from Eq. 6.18 and Eq. 6.19, respectively. Altogether, this provides $5^4 = 635$ different combinations of cognition vectors. For each of these vectors, a separate node with 10 CPUs and 8 GB of RAM was requested on the HILBERT HPC for a maximum runtime of 30 minutes (cf. Fig. 6.12). On each node, one parallel process has been created for each of the nine user-item pairs from the small data record (cf. Fig. 6.13). An individual neuron population was then initialised for each process which provided 10^4 responses according to Eq. 6.8. These population responses were gradually transformed into a model-based distribution by each decoder function and all similarity metrics were computed (cf. Fig. 6.14). By this, each computing node generates nine records (one for each user-item pair) stored on the HDD (cf. Fig. 6.13). All 5 715 files were merged to one single data frame after all nodes had finished their computation (cf. Fig. 6.12).

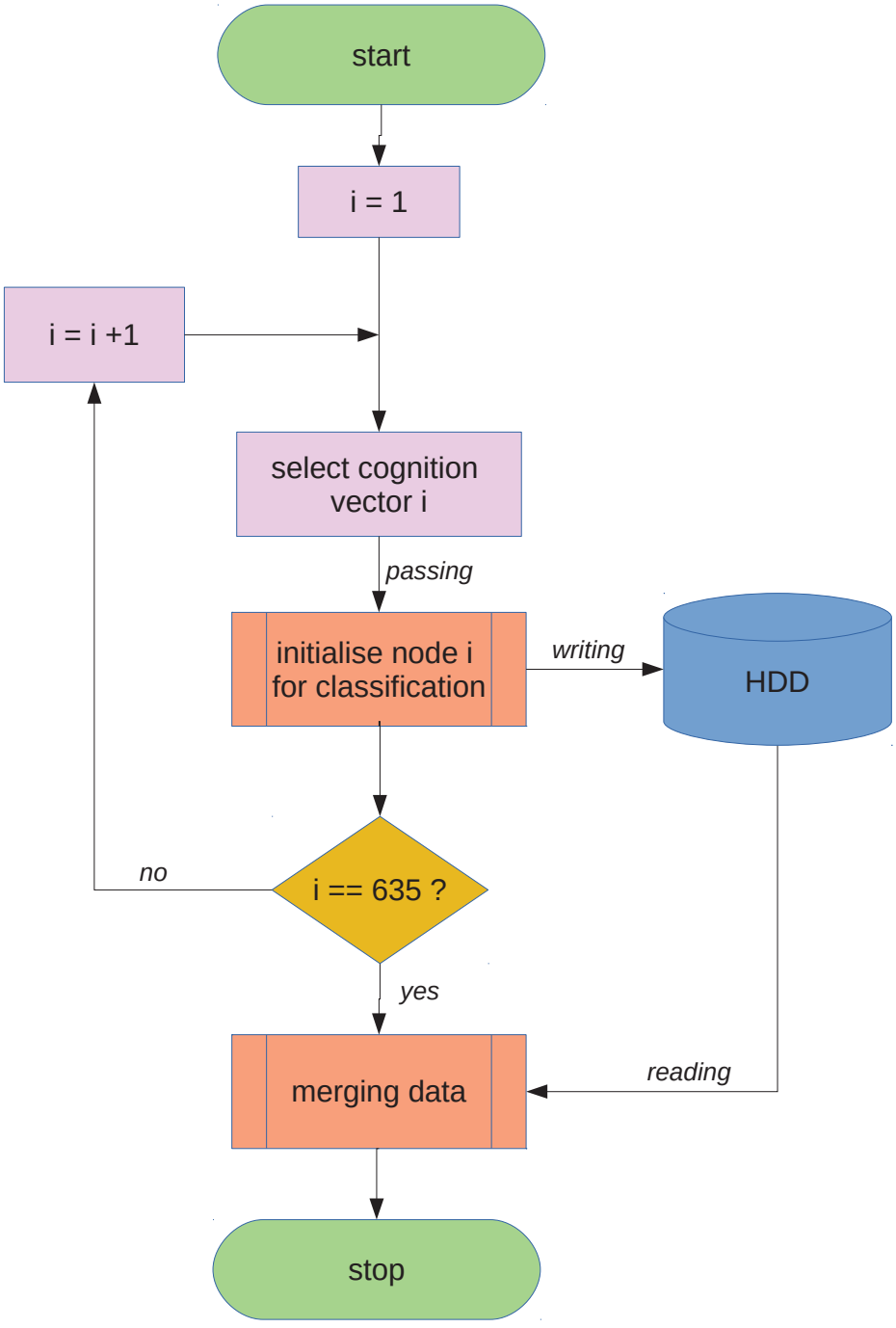


Figure 6.12: Flow chart representing the algorithm of splitting, distributing, and merging the data across the HPC

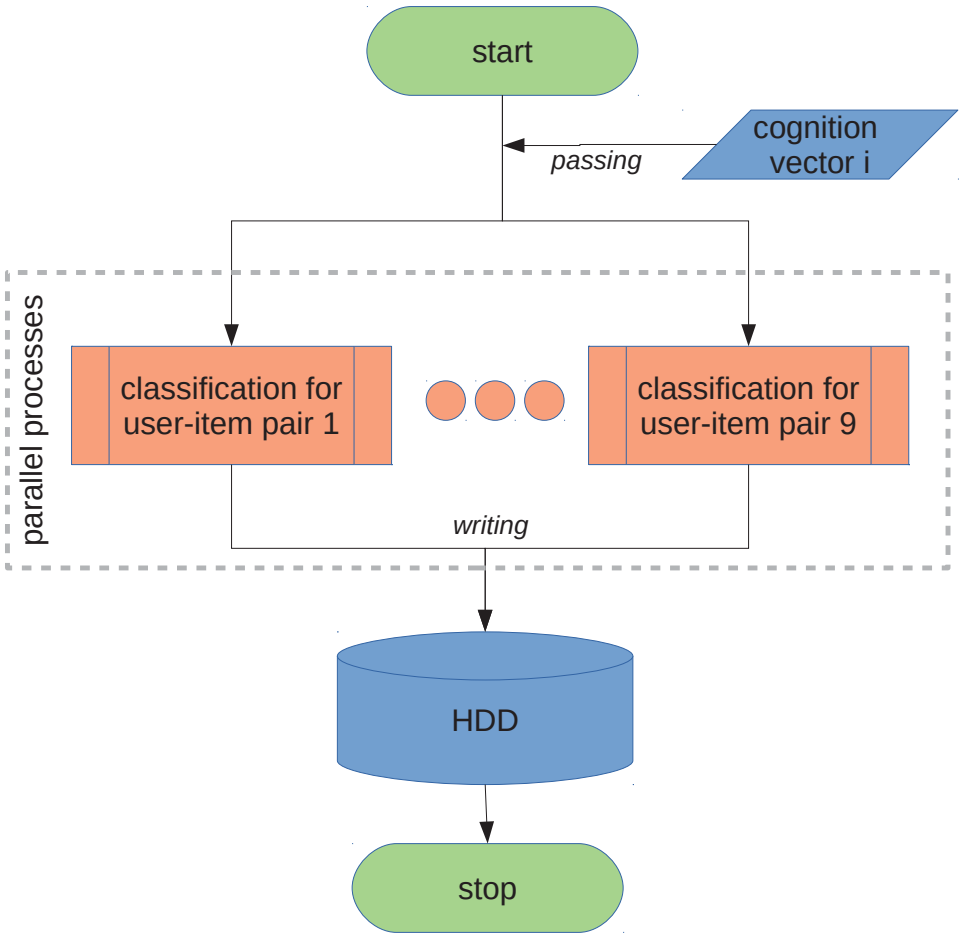


Figure 6.13: Flow chart representing the algorithm of processing and distributing data on a single HPC computing node

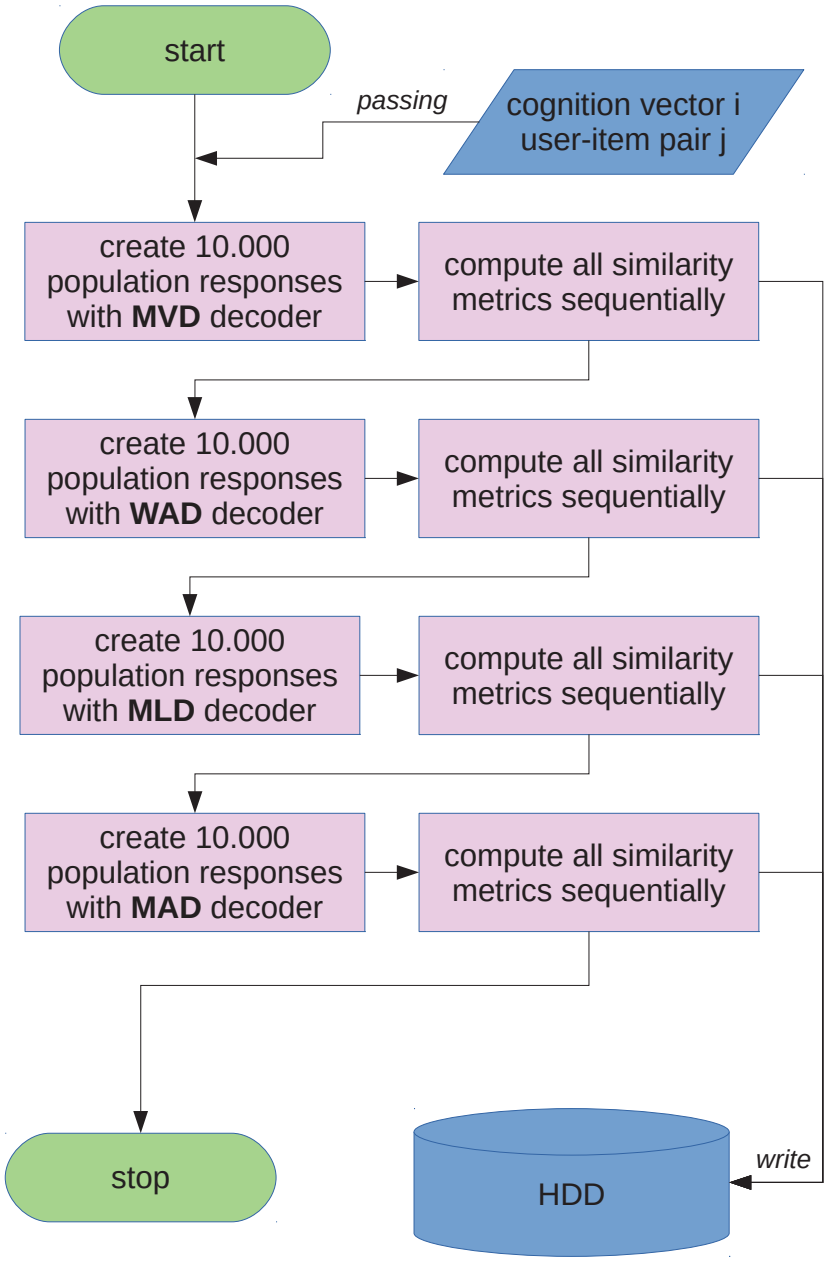


Figure 6.14: Flow chart representing the main classification algorithm performed by each parallel process on each node

The resulting data set comprises 180 cognition vectors, one for each user-item pair, decoder function, and similarity metric. The quality of each metric-decoder combination is analysed with regard to four criteria, namely

- uniqueness of classification results (ambiguity),
- quantifiable similarity, and
- visual similarity.

It will turn out that these indicators – when taken together – are perfectly sufficient to determine the optimal combination of decoder function and similarity metric even with a small data set.

The ambiguity of classification is shown in Tab. 6.4. A brief look into the metrics column reveals that ambiguity occurs mainly with Cohen's Kappa and Cohen's D. These metrics have already been shown to be insufficient for the present case, and this new evidence serves as a supplementary indication that these metrics should not be used for further classifications. In contrast, the metrics JSD50 and nJSD do not produce any ambiguities, leading to advantages for the classification process. Considering the decoder functions, it can be seen that ambiguity is not particularly related to any of these, i.e. all decoder functions only lead to ambiguity when being combined with Cohen's Kappa and Cohen's D. Therefore, no preference or rejection of a special decoder function can be deduced at this point. With a few exceptions, all user-item pairs that are subject to ambiguity belong to the accumulation A_2 or $A_3 \cup A_4$ and thus represent a medium to high human uncertainty. Actually, ambiguity issues would have been expected to occur only with a small uncertainty, since uncertainty is an essential component of this neuronal model. However, little importance is attributed to this result at this point, as it is obvious that ambiguity only occurs with metrics whose appropriateness have already been falsified. Nevertheless, since only a record of $N = 9$ has been considered, it cannot be fully excluded that ambiguities may also occur with other metrics. The full classification addresses this problem with an additional restriction that complies with the brain's least energy principle: The human brain always has to work in an energy-efficient manner and thus it will always choose a cognition strategy that minimises the loss of energy (cf. Niven and Laughlin, 2008, p. 1793). In the case of ambiguity, that is, when several different cognition vectors lead to the same minimum of a metric, this principle will favour the vector $\xi = (n, g, w, o, s)$ inducing all n neurons to spike as sparsely as

user	item	accumulation	decoder	metric	ambiguity count
27	0	$A_3 \cup A_4$	MVD	Kappa	7
58	4	A_2	MVD	CohenD	20
47	0	A_2	MVD	CohenD	20
58	4	A_2	WAD	Kappa	2
24	1	A_2	WAD	Kappa	625
47	0	A_2	WAD	Kappa	625
27	0	$A_3 \cup A_4$	WAD	Kappa	625
62	0	$A_3 \cup A_4$	WAD	Kappa	625
24	3	$A_3 \cup A_4$	WAD	Kappa	625
27	0	$A_3 \cup A_4$	WAD	PureJSD	2
27	0	$A_3 \cup A_4$	MLD	Kappa	2
26	2	A_1	MLD	CohenD	2
58	4	A_2	MLD	CohenD	6
47	0	A_2	MLD	CohenD	6
27	0	$A_3 \cup A_4$	MLD	CohenD	2
24	1	A_2	MAD	Kappa	2
47	0	A_2	MAD	Kappa	3
26	2	A_1	MAD	CohenD	2
58	4	A_2	MAD	CohenD	7
47	0	A_2	MAD	CohenD	7

Table 6.4: Total counts of classification ambiguity within the small data record. Samples with no ambiguity have been omitted.

possible. Consequently, this vector minimises the population energy

$$E \propto n \cdot (g + o). \tag{6.33}$$

The classification results were adjusted for the ambiguities using the least energy principle so that a more sensible assessment of the classification quality is furnished. First, a purely quantitative analysis is performed by focusing on the distributions of similarity scores that result from the associated cognition vectors for each metric-decoder combination. Figure 6.15 depicts these distributions in a boxplot diagram using different groupings for a better comparability. The left boxplot is grouped by similarity metrics and subdivided by decoder functions. It is obvious that Cohen's Kappa performs poorly for all decoder functions and that Cohen's D produces excellent scores. The latter, however, is due to the strong bias, that is, there are a lot of cognition vectors that

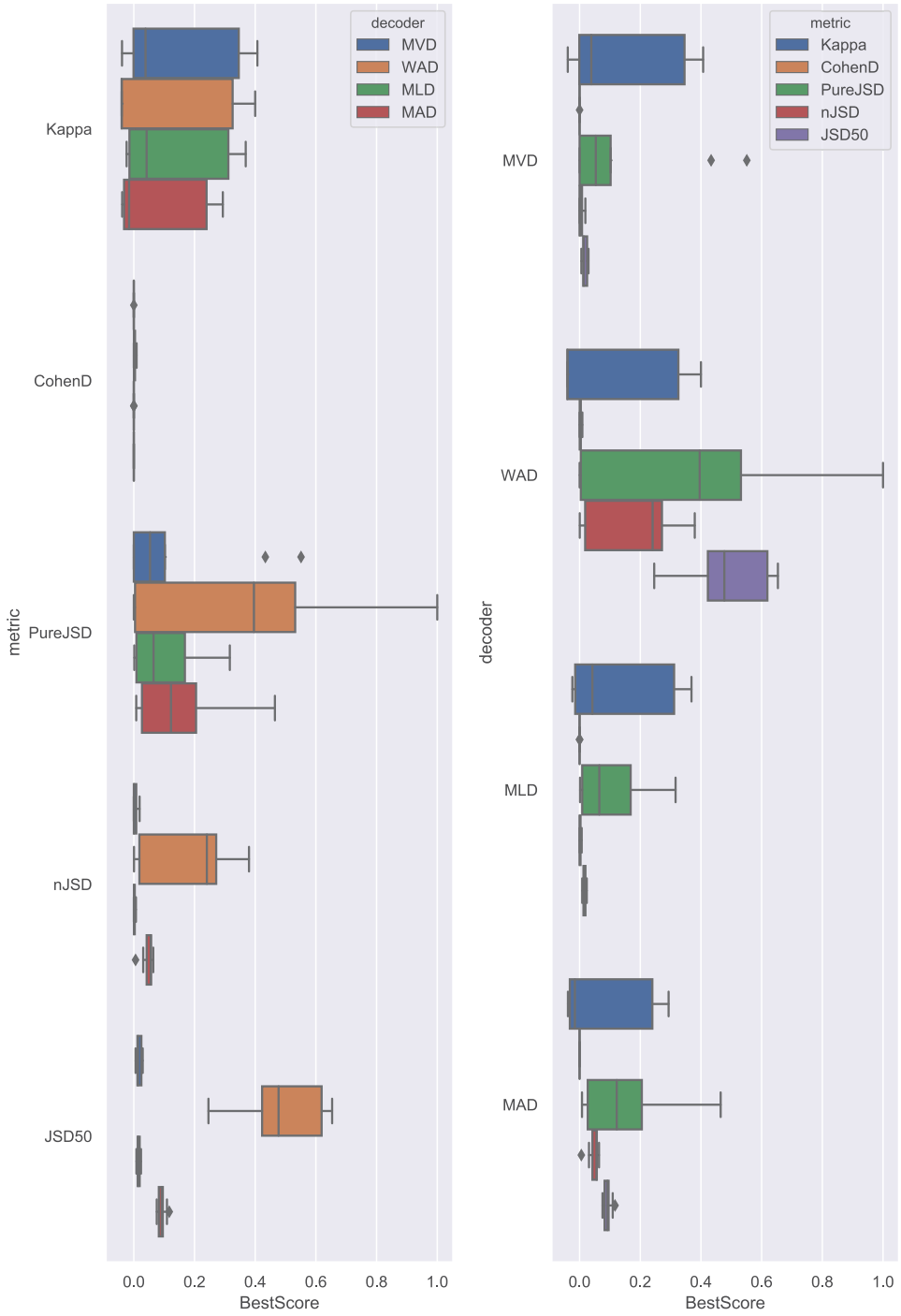


Figure 6.15: Similarity scores for the best fitting cognition vectors. The left boxplot is grouped by similarity metrics and subdivided by decoder functions. The right boxplot is grouped by decoder functions and subdivided by similarity metrics.

will produce too small variances that will be rated with good scores. Moreover, it is striking that the nJSD and JSD50 (no ambiguities for both) produce excellent scores except for the WAD. This provides an additional argument for the potential use of these metrics in the upcoming classification. The right boxplot is grouped by decoder functions and subdivided by similarity metrics. Remarkably, the WAD generally has very bad scores except for Cohen's D. The remaining three decoder functions perform rather well. Descriptively, the MLD seems to be slightly better than the MAD and the MVD. Overall, the combinations of MVD, MLD, and MAD together with the nJSD or JSD50 appear to be the best decoder-metric combinations so far.

It has already been demonstrated in Sec. 6.6 that relying on numerical scores as the only source of information can be very misguiding. Fortunately, a visual check of the model fitting quality is possible for a small dataset. Figures 6.16 to 6.19 depict the RETRAIN feedback distribution (green) for each user-item pair (y-axis) of the small data set along with the model-based feedback distribution (blue) under utilisation of a particular decoder function (separate figures). It can be seen immediately that the WAD (cf. Fig. 6.17) and the MAD (cf. Fig. 6.19) can not be considered as working well. However, it is most likely for the MAD that this is simply due to a wrong prior distribution. Here the real feedback was used as a prior for constructing the posterior distribution which was only a first idea to start with. Fortunately, subsequent research to determine optimal prior distributions is not necessary, since the MLD (cf. Fig. 6.18) – which is a special case of the MAD using an uninformative prior – performs excellently. Within the paradigm of the Bayesian brain and the PPC model, in particular, this fact can be interpreted as the absence of user beliefs or memories, respectively, while completing the study. This interpretation is in concordance with the use of distractors between each repetition, triggering the misinformation effect and thus preventing memory to play a major role within the RETRAIN study. Moreover, this interpretation is supported by the pdf-rating procedure which excluded memory effects while giving distributions that do not significantly differ from the repeated ratings. The superiority of the MLD along with these former findings can thus be seen as a hint for the irrelevance of memory considering this specific cognitive task. In addition to the MLD, the MVD is also a very good candidate (cf. Fig. 6.16) that is even much simpler in its construction. Nevertheless, the MLD is still slightly superior which can be seen for the user-item pairs $(59, 3), (62, 0)$, and $(24, 3)$. If the MLD is selected as the decoder function, then the most appropriate metrics would be the nJSD and the JSD50.

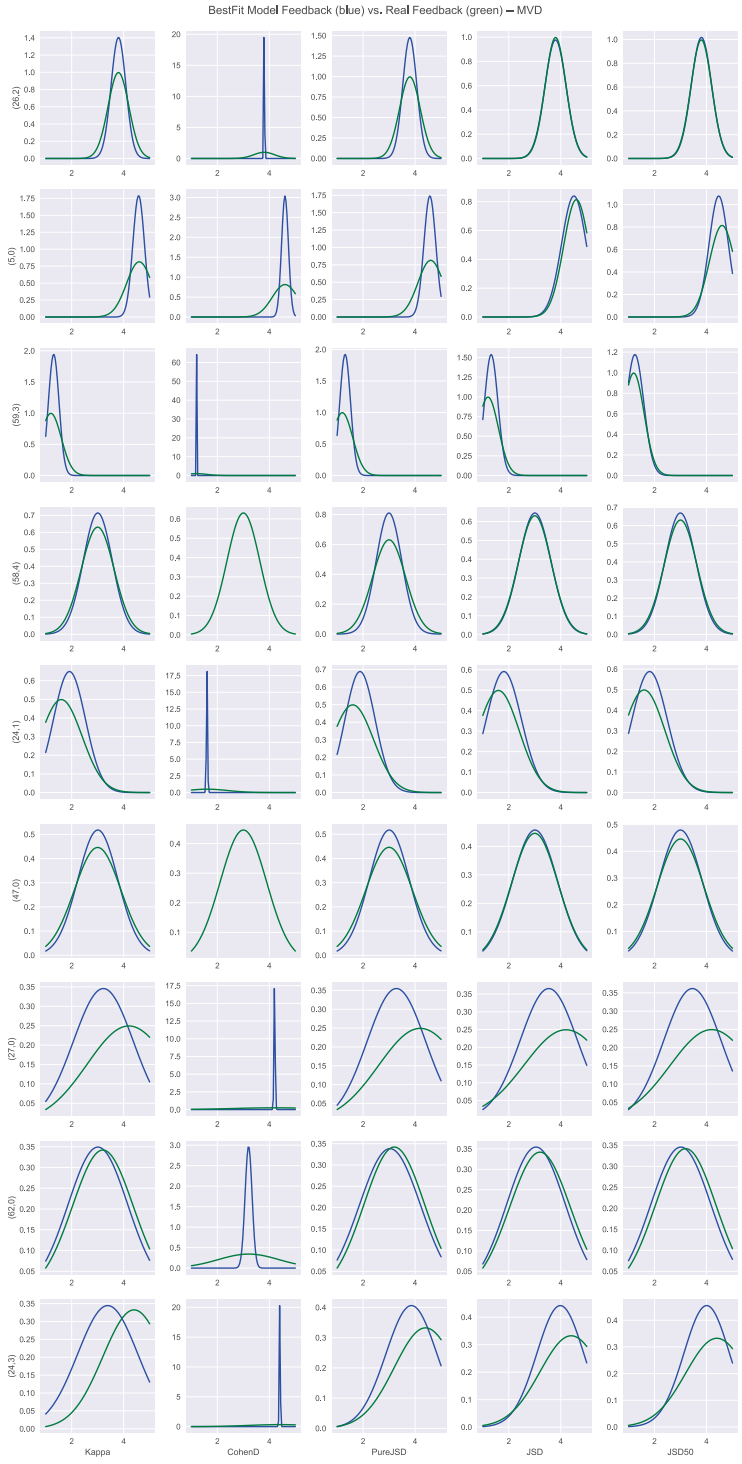


Figure 6.16: Visual fitting quality for the MVD along with all utilised metrics for each user-item pair from the small data set

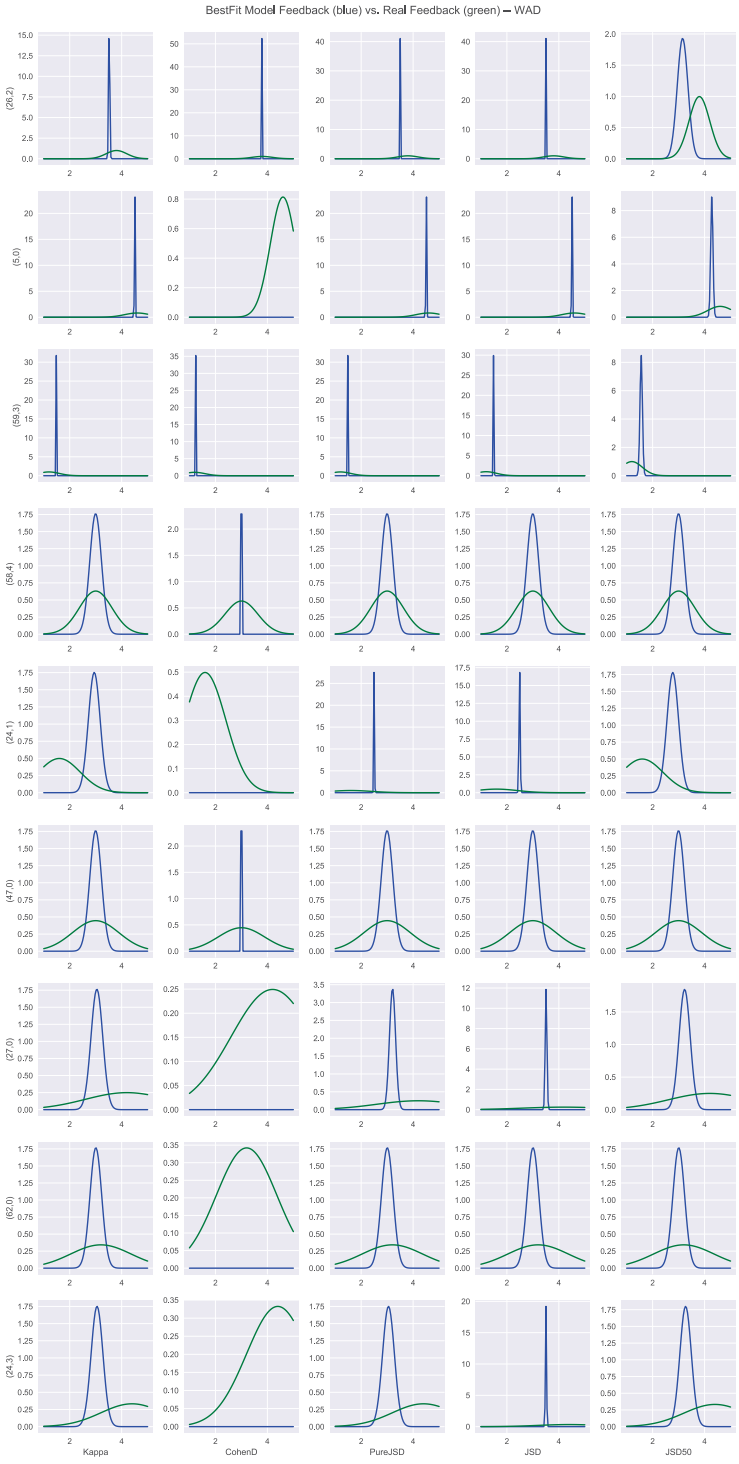


Figure 6.17: Visual fitting quality for the WAD along with all utilised metrics for each user-item pair from the small data set

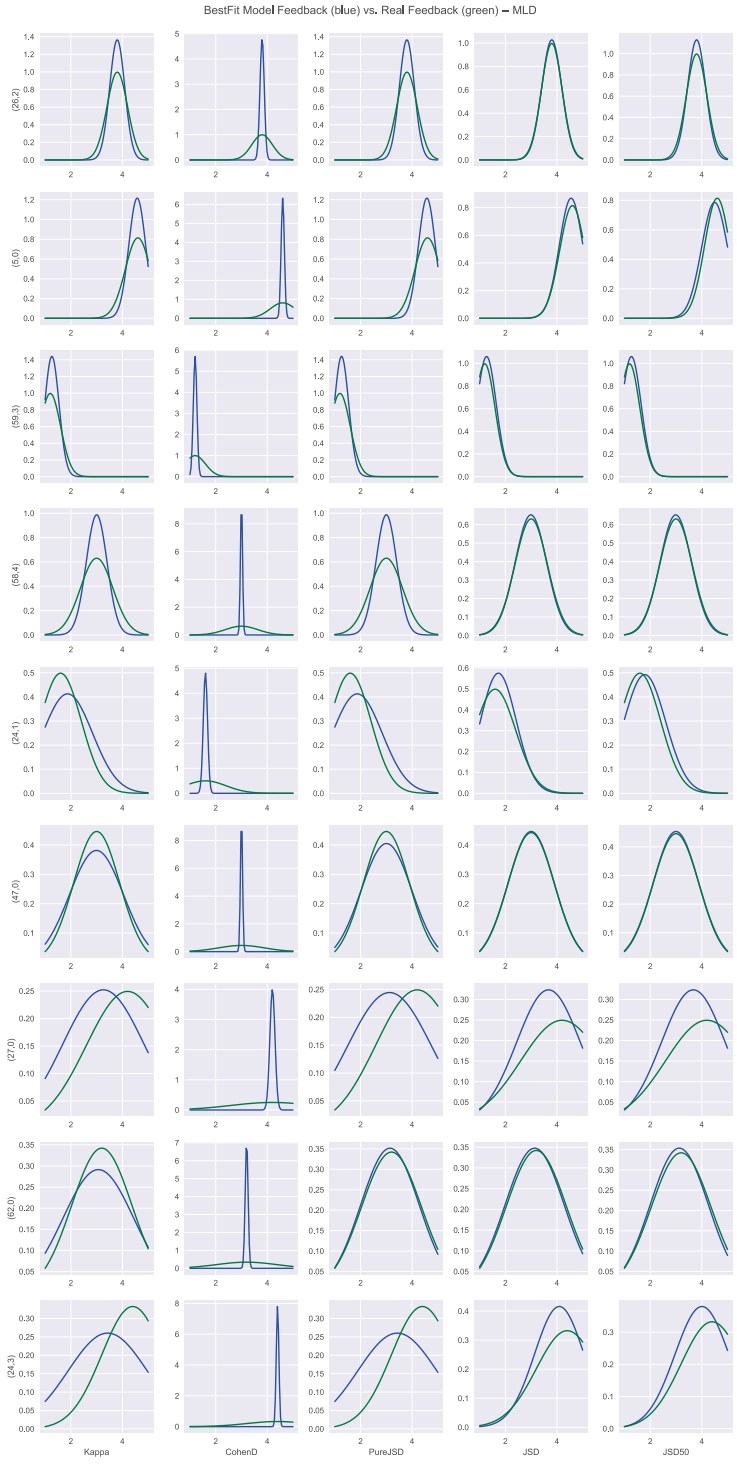


Figure 6.18: Visual fitting quality for the MLD along with all utilised metrics for each user-item pair from the small data set

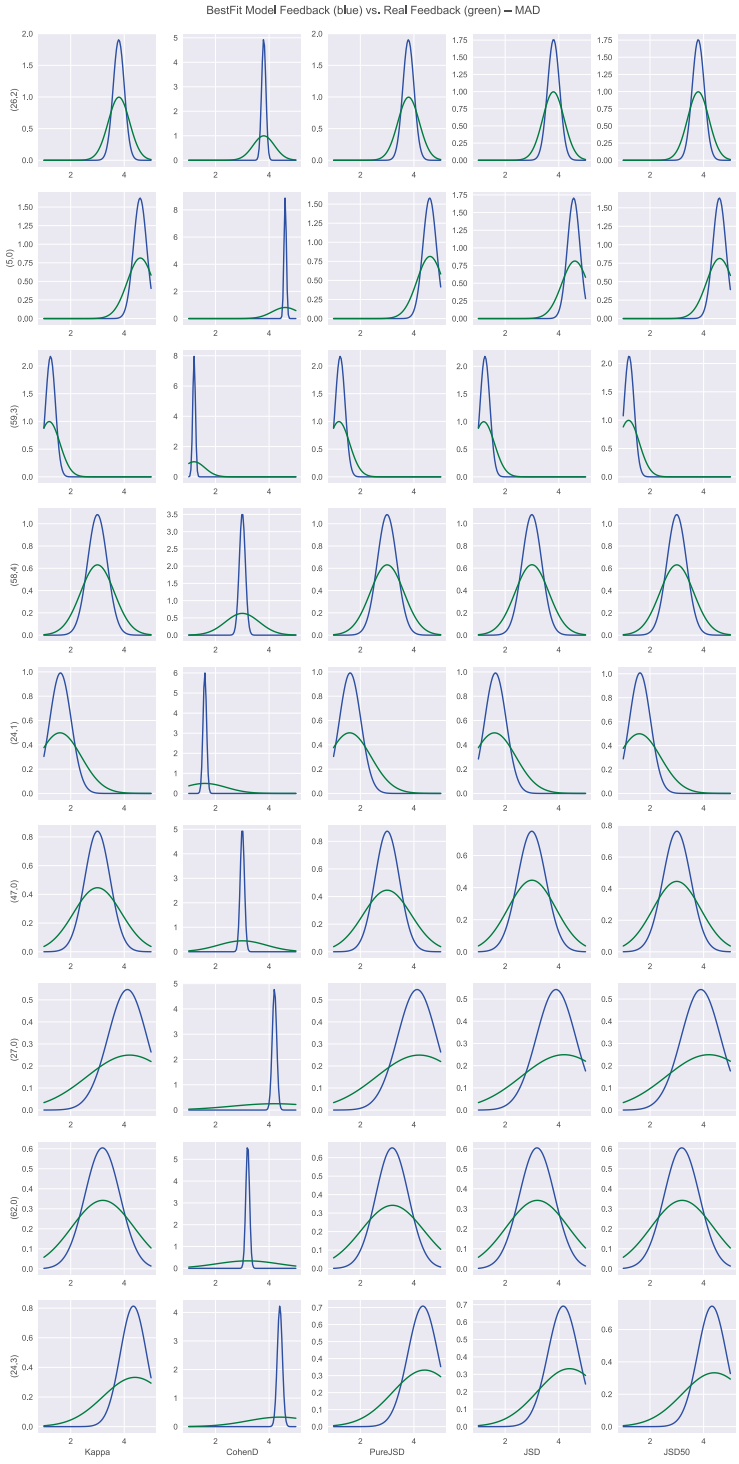


Figure 6.19: Visual fitting quality for the MAD along with all utilised metrics for each user-item pair from the small data set

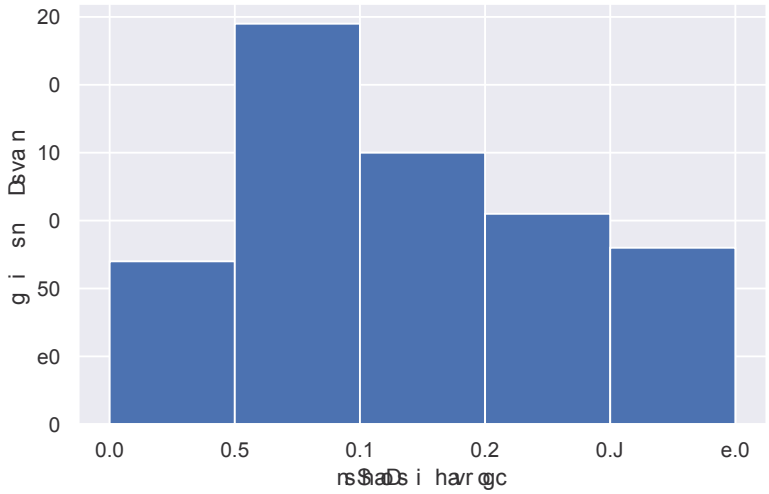


Figure 6.20: Number of cognition vectors with regard to their relative matching when being determined using the re-rating and the pdf-rating

In fact, there are no significant differences between the nJSD and the JSD50 concerning their approximation quality. Nevertheless, a decision is made in favour of the JSD50 since there is a slight descriptive superiority to be spotted for the user-item pairs $(24, 1)$ and $(24, 3)$ in Fig. 6.18.

Digression: Re-Rating vs. PDF-Rating

From Tab. 3.2 it is known that the resulting feedback distributions based on the re-rating and the pdf-rating procedure do not differ significantly in 83% of all cases. Since the neuronal classification is dependent on these distributions, it has to produce almost similar results in terms of cognition vectors. In this short digression, the true impact of the measurement approach on the association of cognition vectors will be examined. For this purpose, the same user-item pairs from the small data set are reused, but now the pdf-rating is chosen as the basis for the respective feedback distribution. Except for this adjustment, the classification is carried out as described above.

In a first analysis, the degree of matching is examined amongst the corresponding cognition vectors, i.e. the relative number of their equal components. For instance, the vectors $(25, 1, 1, 0, 3)$ and $(25, 2, 1, 5, 4)$ have two equal components (the first and the third) and thus the relative matching is $2/5 = 40\%$. Figure 6.20 depicts the number of

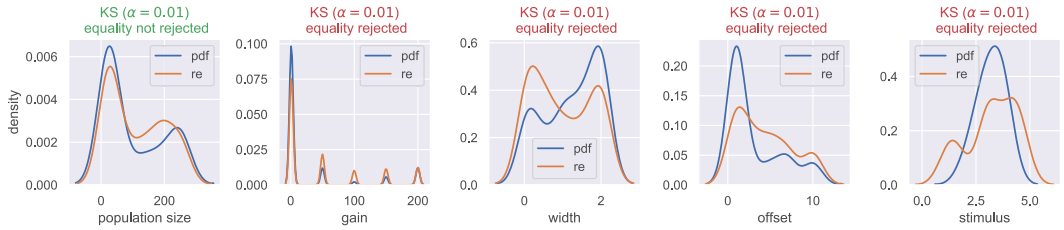


Figure 6.21: Distributions for the neuronal parameters when the cognition vectors are computed on the re-rating set and the pdf-rating set, respectively

cognition vectors for each class of relative matching. Out of 180 cognitive vectors, only about 25 have a match between 80% and 90%. With about 60 vectors, the modal value can be located between 20% to 40% of relative matching. These values indicate that the differences between cognition vectors based on different measurement approaches are substantial. Some of the neuronal parameters match more frequently than others. From all cognition vectors, 42.7% show equality for parameter n, 64.4% for parameter g, 41.6% for parameter w, 37.7% for parameter o, and none of the cognition vectors shows equality for the stimulus s. However, the stimulus s is the only parameter that is computed in advance and passed through the classification approach from outside. Hence, this parameter is not suitable for evaluating the impact of the applied measuring approach on the classification. From Tab. 3.2 it can be seen that stimuli differences are only descriptive in 83% of all cases. If the null hypothesis of equality is assumed to hold for all non-significantly different stimuli (i.e. the maximum of possible equalities), a maximum probability can be computed for the case that the re-rating and pdf-rating both lead to the same cognition vector. This probability is given by

$$P_{\max}(\xi_{re} = \xi_{pdf}) = 0.43 \cdot 0.64 \cdot 0.42 \cdot 0.38 \cdot 0.83 \approx 0.04 \tag{6.34}$$

and therefore, both measurement approaches lead to the same cognition vector with a probability of only 4%. This is also reflected in the parameter distributions depicted in Fig. 6.21 as these differ significantly. A KS-test with $\alpha = 0.01$ confirms that equality can be rejected for all parameter distributions except for those related to population size. However, the same location of maxima and minima suggests that the individual differences might not be very strong.

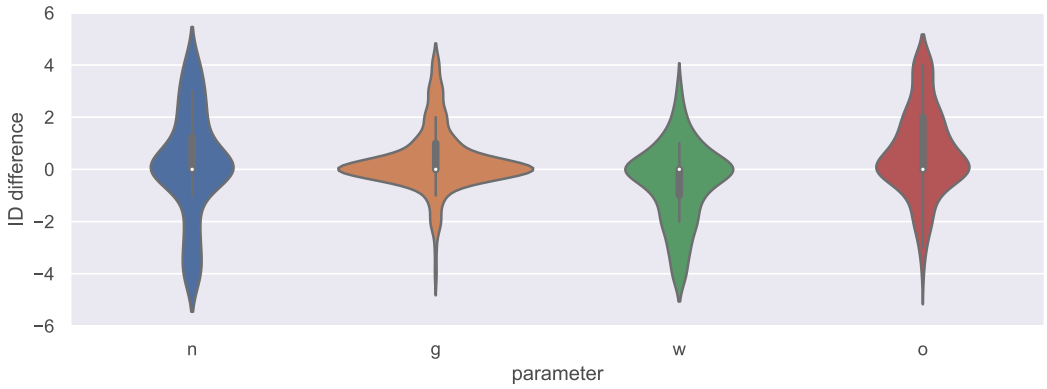


Figure 6.22: Magnitude of parameter differences considering all decoder-metric combinations for each user-item pair from the small data set

Figure 6.22 shows the distributions (violins) of these individual differences per parameter. Since the analyses are performed on roughly discrete scales, it is not practical to use the specific parameter units (e.g. ±200 neurons, ±5 Hz, etc.). Instead, the more general unit of ID difference is used. A value of ±1 means that the next (or the previous) value was chosen within the parameter list. In this particular analysis, the difference of the pdf-rating is compared using the re-rating as a reference, i.e. +1 means that the pdf-rating will choose the next value compared to the re-rating. Overall, it can be seen that large deviations of more than two values (next or previous) occur very seldom and that the magnitude of these differences is rather small. The grey boxes in the violins each represent the middle 50% of the data. This interquartile range often only reaches ±1.

In summary, different methods of measuring a feedback distribution almost certainly lead to different classification results in terms of cognition vectors. However, these differences tend to be rather small in their magnitude. This might be an artefact of the rough discretisation of parameter scales. Perhaps, both parameter values would approach a common value if a much finer discretisation is used. The following full case will hence involve a discretisation which comprises twice as many intermediate values as it was used for the small case. This represents a trade-off between the above arguments and the associated computational complexity. This complexity likewise requires that the following full case is limited exclusively to the re-rating data set.

	ambigue	not ambigue	$\sum$
constant	43	79	122
not constant	0	213	213
$\sum$	43	292	335

Table 6.5: Classification ambiguities of constant and uncertain ratings in a 2×2 contingency table. Both features are statistically dependent ($\alpha = 0.01$).

Full Case

For the full case, the classification is performed as for the small case but with the following changes:

- Solely the MLD and the JSD50 metric are used with 10^3 MC-trials.
- A larger pool of cognition vectors is generated by dividing the parameter spaces into ten instead of just five equidistant steps. This results in $7^4 = 2\,401$ cognition vectors instead of $5^4 = 625$.
- Due to the larger data pool a different parallelisation approach is required for the HILBERT HPC.

If each computing node is set up with seven CPUs, the computation of all 2 401 cognition vectors would require $2\,401/7 = 343$ nodes. Each node is then able to compute seven processes in parallel and each process computes the distribution associated to a single cognition vector along with all corresponding similarities for all 335 user-item pairs. As for the small case, the classification quality for the full case is likewise assessed by considering ambiguity and similarity for all model-based and real feedback distributions. The absolute frequency of ambiguities (i.e. the number of cognition vectors that result in the same minimum) for the JSD50 is shown in Tab. 6.5 along with the occurrence of human uncertainty. It can be suspected that both features are statistically dependent, especially because

$$P(\text{ambigue} \,|\, \text{not constant}) = \frac{0}{213} \neq \frac{43}{335} = P(\text{ambigue}). \tag{6.35}$$

An additional χ^2 contingency test reveals that these features are indeed statistically dependent ($\alpha = 0.01$). This basically means that the occurrence of human uncertainty substantially reduces the probability of ambiguity, i.e. the existence of human uncertainty

maximises the quality of the neuronal classification. At first sight, two reasons seem to suggest that this phenomenon is inherent to the model construction itself. The first reason is that neuronal noise – which is the core element of the PPC model – requires training data in the form of distributions and constant ratings do not meet this requirement. This argument can be refuted since constant ratings do not necessarily contradict the assumption of a distribution. Sizov already assumed that constant ratings are simply the manifestation of distributions with a small variance that cannot be resolved by the applied rating scale (cf. Sizov, 2017b, p. 872). This points to the second reason: There may be multiple cognition vectors leading to a distribution with small variance from which drawings will produce constant ratings. This argument can also be refuted as it implies that constant ratings will always lead to ambiguities, i.e. the same cognition vectors (modulo stimulus) will inevitably produce distributions with equally small variances. However, it can be seen from Tab. 6.5 that 65% of all constant ratings do not produce ambiguities at all. With this line of argument, the present phenomenon can probably not be explained by the mere construction of the PPC model. Therefore, this model is most likely capable of tracing back human uncertainty to neuronal noise with high quality in terms of uniqueness. This provides one indication that the explanatory PPC model is suitable to capture human uncertainty.

For the upcoming analyses, one cognition vector is chosen for each of the ambiguous user-item pairs by minimising the energy score according to Eq. 6.33. The fitting quality can be determined by considering the extent of similarity (JSD50) as well as the visual matching of distributions. Both characteristics are shown in Fig. 6.23. Except for a few outliers, all user-item pairs have a score of less than 0.025 which means a great fit. Even the outliers do not exceed the upper bound of 0.175 which is still acceptable. Figure 6.23 also depicts three exemplary distribution fits to provide an idea of the classification quality. It can be rated as being excellent. These results can be seen as a second indication that the PPC model is capable of capturing the phenomenon of human uncertainty. It was not clear from the beginning that this theory – which had only been reviewed in the light of perception and motor control so far – is also applicable to decision-making. It would have also been possible that the model-based distributions do not match the measured feedback distributions (e.g. shape, mean, variance, etc.) at all. In this sense, the PPC model represents a remarkably accurate mapping between measurable human uncertainty and the neuronal noise of a specific neuron population.

To investigate the plausibility of this model, the intercorrelations of its population

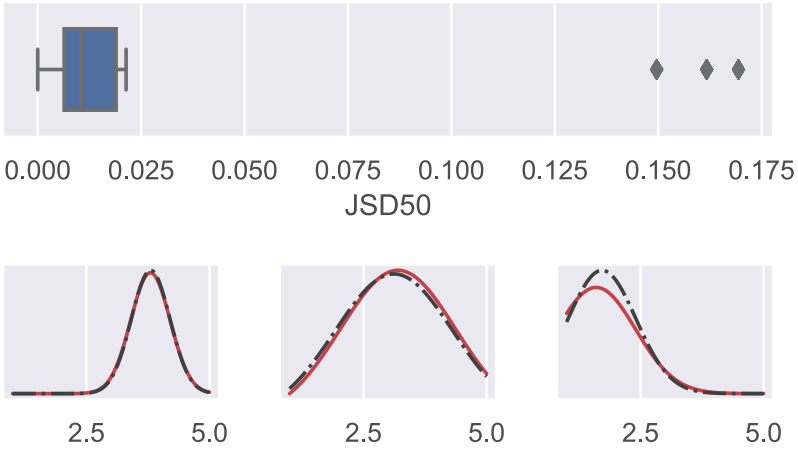


Figure 6.23: Full case fitting quality using the JSD50 and three visual examples of the real feedback distribution (red) against the model-based distribution (black)

parameters are examined as well as the neuronal response frequencies (i.e. model-based biological implications) resulting from the best fitting cognition vectors. The pairwise correlation of each population parameter and the uncertainty of user feedback (standard deviation STD) is depicted by the correlation heatmap in Fig. 6.24. The behavioural variability (STD) mainly correlates with the frequency gain g and the tuning curve width w. The latter is not surprising, since a Gaussian is used to model the tuning curve shape and, therefore, smaller widths automatically lead to higher response frequencies due to normalisation. To further adjust this frequency despite a fixed width, an additional stretching factor (frequency gain) has been introduced. The present correlation can hence be interpreted as follows: The higher the frequency, the smaller the influence of neuronal noise, and the lower the standard deviation for repeated decision-making. Since attention (focus) is correlated with spiking frequency (cf. Dresler, 2011, p. 170), this data can also be interpreted in such a way that greater attention leads to less uncertainty in decision-making.

Further correlations worth mentioning are those for n-w and g-o. The negative correlation between n and w is plausible, because a larger number of neurons potentially increases the resolution when representing a scale (cf. Erdmann et al., 2015, p. 43). To exploit this potential, the subrange of a stimulus space in which each neuron is triggered must be smaller. On the contrary, a larger stimuli-response range induces an

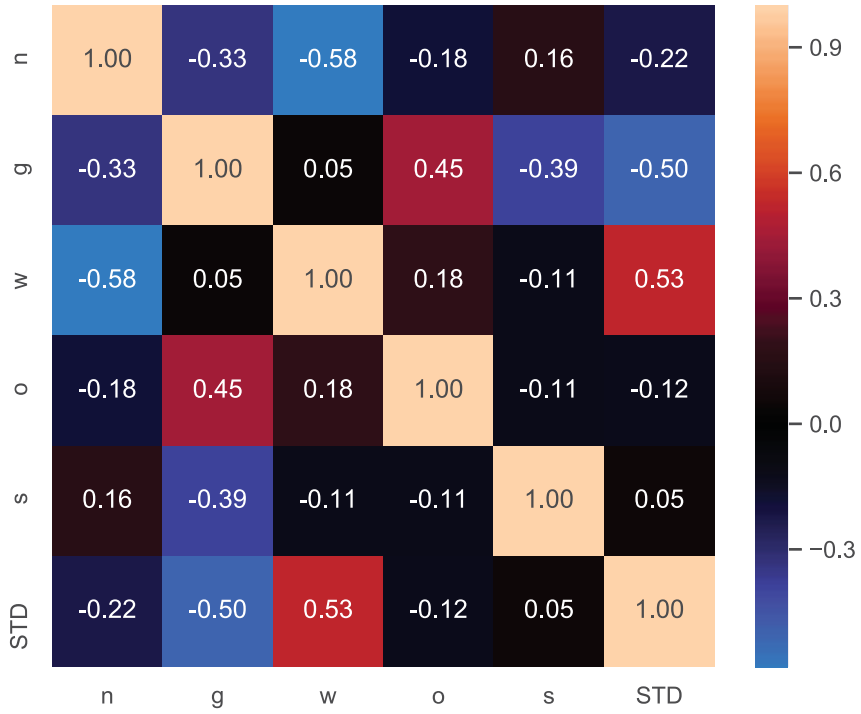


Figure 6.24: Intercorrelations between the neuronal parameters and the human uncertainty (STD) in a heatmap

overlapping of tuning curves and the resulting resolution would again decrease. This would constitute an energy-consuming constellation that delivers the same results as an energy-saving alternative (e.g. fewer neurons with narrow width). At this point, the model therefore automatically operated energy optimisation. This behaviour can not be explained by the artificial energy score correction as this was only done for less than 13% of the data. Considering only uncertain ratings (no energy correction at all), the Pearson-correlation between n and w is $\rho = -0.57$ and shows no significant difference. Accordingly, the PPC model acts – fully on its own – in the same way as it is often postulated in the academic literature, namely to carry out its operations in an energy-efficient and optimising manner (cf. Niven and Laughlin, 2008, p. 1793).

The positive correlation between g and o is a non-obvious property of the model. Both parameters influence the frequency with which a neuron responds to a stimulus. Nevertheless, they have a very different effect on the resulting feedback uncertainty. The offset increases the frequency of each neuron within the population by the same

amount, regardless of the individual dependence on a particular stimulus. This, in turn, increases the uncertainty of a resulting feedback distribution, because neurons whose preferred values deviate substantially from a given stimulus will respond more strongly. In contrast, the gain amplifies neuronal responses to a greater extent if the corresponding preferred values are similar to the given stimulus. An increase in gain thus leads to a decrease in the uncertainty of a feedback distribution. This correlation can hence be regarded as a constant noise-to-signal ratio for decision-making. Since frequency is directly related to the amount of noise according to Eq. 6.7, both parameters also constitute a degree of freedom for influencing the so-called noise correlation – a phenomenon observed in real neuron populations (cf. Averbeck et al., 2006, p. 360). However, the range of this particular degree of freedom is naturally limited by the present correlation, since one parameter cannot be changed independently of the other. It is noteworthy that all these aspects, which are unified in this particular correlation, have indeed been found to essentially influence the quality of neuronal processing. Herrero et al. specify that such processing is improved by enhancing firing rates, reducing firing rate variability and noise correlations (cf. Herrero et al., 2013, p. 729). They argue that "all of these alterations can improve the signal-to-noise ratio when decoding the [population] activity" (cf. Herrero et al., 2013, p. 729). The present correlation reflects these mutual dependence.

To investigate the population response in terms of spiking frequency, the neuronal model is set up with each best fitting cognition vector and the noisy population response is computed by Eq. 6.8 several times. The distribution of spiking rates for all user-item pairs is depicted in Fig. 6.25. One can observe a log-norm distribution between 0 and 70 Hz and an expectation of about 8 Hz (red line). The evoked distribution shape is in line with medical observations:

> "The distribution of firing rates across the population closely resembled a lognormal distribution [...]. Such lognormal-like distributions are also present in various other parts of the nervous system [...] and could represent a ubiquitous feature of neuronal networks." (Berg, 2017, p. 3)

Furthermore, the location of the distribution, i.e. the mean and the entire range, is also compatible with other postulates of medical research (cf. Roxin et al., 2011, p. 16220). The configured model of the PPC hence leads to neuronal activities which are close to those reported in real measurements.

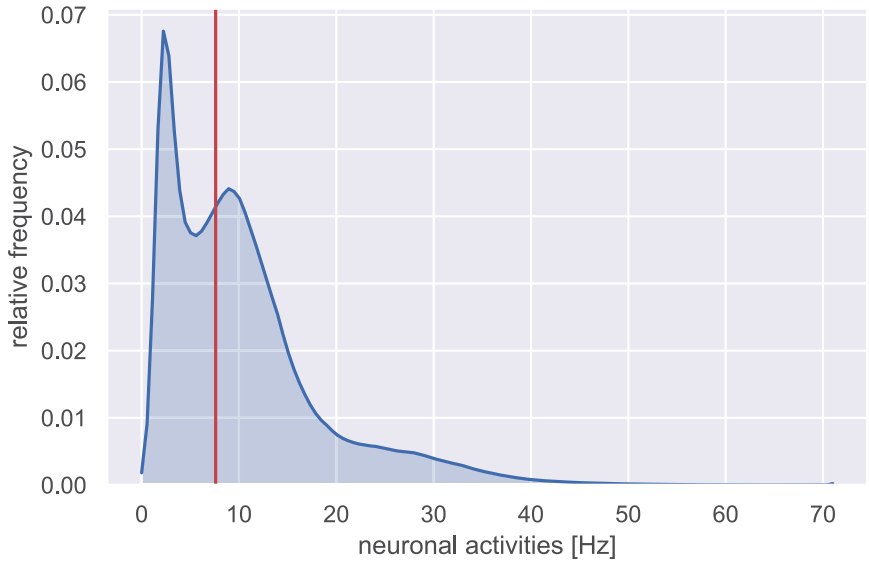


Figure 6.25: Theoretical neuron frequencies during decision-making using the PPC model and the MLD as decoder function

In summary, the following statements can be made about the adequacy and plausibility of the PPC model (fitted on data of human decision-making):

1. The PPC model is capable of replicating real uncertainty in decision-making with high precision.

2. It naturally leads to plausible and medically interpretable correlations and thus to implications about the human brain as they are reported in the field of neuroscience.

3. This model reproduces the same frequency distribution as measured in reality and reported in the standard literature.

All these findings are strong indications that the PPC model represents a possible neuronal explanation for the phenomenon of human uncertainty. None of these findings was initially expected or even a logical consequence. Of course, it can be argued that the goodness of fit is only a logical consequence since a maximum likelihood method was used which is in general a good fitting tool. This is, however, not tenable in many respects. Maximum likelihood estimation is a tool to adapt any parametric model

to real data (e.g. fitting a Pareto or uniform distribution over a normally distributed feature), but whether this model fits well on reality in the end can not be enforced by using this tool. Besides, the maximum likelihood method has not been employed for parameter fitting but to map a noisy population response onto a point estimator regarding a preselected cognition vector. However, it is remarkable that this model of population activity together with an ML-based decoder resulted into distributions which equal the real measured feedback in terms of shape and uncertainty (i.e. equality in all their moments). It could have also been that the real feedback distributions cannot be reproduced at all. Also, it is not inherent in the model to autonomously operate energy optimisation or to organise high frequencies (i.e. more attention) as a trigger for less uncertainty in decision-making. Even the frequency distribution that is generated for the case of decision-making fits remarkably well into the range of what can be considered normal for the brain. At this point, it would also have been possible that this model works well but produces frequencies in the MHz or GHz band.

Finally, a cognitive model has been developed for this thesis which is based on theories from neuroscience (i.e. the Bayesian brain hypothesis). This model reproduces non-vanishing human uncertainty unambiguously (i.e. quantitatively convincing) and precisely (i.e. qualitatively convincing) whilst generating medically consistent correlations and frequencies (i.e. plausibly convincing). Of course, these findings do not prove the correctness of the PPC model in such a way that it can be understood as the brain's real implementation of cognition. Such a claim would be impossible due to the axiomatic construction of science. This shall be illustrated by an example: Physics can be considered as one of the oldest fields of science which gave birth to plenty of theories and understandings about nature. Yet, none of physics' theories actually represents the truth. There is no point mass or elementary particle, because no one has ever seen an electron or a vibrating string. Physics is thus a mere collection of fictitious assumptions to simplify the complexity of nature along with built-on mathematical descriptions of observable phenomenons that are consistent with other theories (cf. Kircher et al., 2009, pp. $4, 31, 41, 754$). The same standard must also be applied to the PPC model developed in this thesis: It has been relied upon hypothetical assumptions and a mathematical formulation has been used to create a theory that reproduces a phenomenon, makes predictions which can indeed be observed in reality, and that does not contradict previous measurements. This model hence describes a phenomenon with full compliance of scientific standards but does not represent the absolute truth.

At this point, however, there is sufficient evidence that a model of human uncertainty has been found which indeed adopts an inner perspective. This was important to show, because it demonstrates the appropriateness for this object of comparison as opposed to purely phenomenological models. Both approaches, i.e. the inner and the outer perspective, hold the potential to significantly impact the design of future systems.

6.8 Predicting with Neurological and Behavioural Models

To this point, a neuroscientific model has been operationalised and adapted for unreliable decision-making. This model has been employed to translate user behaviour into possible neuronal features. Its rationality was substantiated through various indications, in particular by providing data-driven neurological interpretations and comparisons with medical publications. In this section, this user model is applied to different recommendation techniques in order to predict future uncertain user behaviour. The prediction quality is then compared with the quality of the same techniques when these are trained directly on a purely behavioural user model instead. An equivalence of feedback representation can be assumed when the quality of both systems does not differ significantly. In this case, the neuronal features can be seen to have the same representative power to the target variable as does the real user behaviour. Using this interpretation allows investigating whether the associated cognition vectors still represent real user behaviour or not (classification validity). In mathematical terms: Let

$$\{\xi\} \sim_M \{(\mu, \sigma)\} :\Longleftrightarrow H_0 : \mathcal{D}_M(\xi) = \mathcal{D}_M(\mu, \sigma) \text{ not rejected} \tag{6.36}$$

be an equivalence relation where $\{\xi\}$ is the set of assigned cognition vectors, $\{(\mu, \sigma)\}$ the set of real user behaviour, $\mathcal{D}$ a distributed prediction quality distance (e.g. RMSE) and M a specific machine learning method. This definition makes sense, as it is indeed irrelevant for the quality of method M which of the equivalent feature spaces is finally used. For the upcoming investigations,

$$F_N(k) := (\xi_{u,i})_{\substack{u=0,\ldots,66 \\ i=0,\ldots,(k-1),(k+1),\ldots,4}} \tag{6.37}$$

represents the neuronal feature matrix with omission of item k. Each instance (row) is a single feature vector for a specific user and each column is a neuronal feature.

For example, $F_N(4)$ looks like

$$\begin{pmatrix} n_{0,0} & g_{0,0} & w_{0,0} & o_{0,0} & s_{0,0} & \cdots & n_{0,3} & g_{0,3} & w_{0,3} & o_{0,3} & s_{0,3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{66,0} & g_{66,0} & w_{66,0} & o_{66,0} & s_{66,0} & \cdots & n_{66,3} & g_{66,3} & w_{66,3} & o_{66,3} & s_{66,3} \end{pmatrix}$$

where the u-th row $F_N(4)_{u,\bullet}$ is the feature vector of user u. Analogously,

$$F_S(k) := ((\mu, \sigma)_{u,i})_{\substack{u=0,\ldots,66 \\ i=0,\ldots,(k-1),(k+1),\ldots,4}} \tag{6.38}$$

denotes the statistical feature matrix with omission of item k where the u-th row $F_S(k)_{u,\bullet}$ is the feature vector of user u. For instance, $F_S(4)$ looks like

$$\begin{pmatrix} \mu_{0,0} & \sigma_{0,0} & \cdots & \mu_{0,3} & \sigma_{0,3} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mu_{66,0} & \sigma_{66,0} & \cdots & \mu_{66,3} & \sigma_{66,3} \end{pmatrix}.$$

Cosine Similarity Recommendation

Grouping users or items according to the direction of corresponding feature vectors is very common in recommender systems engineering. This will be the first method to be tested for feature equivalence between the neurological and the behavioural (statistical) model. The so-called cosine similarity is defined as the cosine of the angle $\vartheta = \sphericalangle(v, w)$ between two vectors $v, w \in \mathbb{R}^n$ from the same vector space (feature space). By using the canonical definition of the inner product $\langle v, w \rangle$ of vector spaces, it follows that

$$S_C(v, w) := \cos(\vartheta) = \frac{\langle v, w \rangle}{\|v\|_2 \cdot \|w\|_2}. \tag{6.39}$$

The image set of this mapping is $\mathrm{im}(S_C) = [-1, 1]$ where $S_C(v, w) = 0$ indicates no similarity between v and w.

To build a collaborative filtering recommender system, the cosine similarity is used to associate each user with a group of similar users. The mean rating is then computed for the target item from all users within this particular similarity group. This mean acts as a predictor for the original user. In more detail: Let k denote the item for which a prediction is required. The item-related features are excluded from the respective feature vectors. For each user $u = 0, \ldots, 66$, there is a group

$$G(u, k) := \{u^* \quad | \quad S_C(F_N(k)_{u,\bullet}\ ,\ F_N(k)_{u^*,\bullet}) > 0.9\ ;\ u \neq u^*\} \tag{6.40}$$

of users $u^* = 0, \ldots, 66$ with $u \neq u^*$ whose absolute cosine similarity score is greater than 0.9. The predictor $\pi_{u,k}$ for a possible rating from user u to item k is defined as the mean

$$\pi_{u,k} := \frac{1}{|G(u,k)|} \sum_{u^* \in G(u,k)} \mu_{u^*,k} \tag{6.41}$$

of expectations $\mu_{u^*,k}$ from all group members' rating distributions for item k. The item-related prediction quality of this procedure is evaluated by

$$\mathrm{RMSE}_N(k) = \sqrt{\frac{1}{67} \sum_{u=0}^{66} \left(\mathfrak{F}_{u,k} - \pi_{u,k}\right)^2} \tag{6.42}$$

where $\mathfrak{F}_{u,k} \sim \mathcal{N}(\mu_{u,k}, \sigma_{u,k})$ represents the real feedback distribution from user u to item k. This procedure is analogously applied to the statistical model.

The RMSE distributions of both the behavioural (statistical) and the neuroscientific model are shown in Fig. 6.26. It can be seen that there are almost no differences between those distributions in terms of location and scattering. Only item $k = 1$ and item $k = 4$ have an insignificant location shift in favour of the behavioural model. The same tests introduced and used in Sec. 3.4 are also applied at this point to execute a statistical analysis with a significance level of $\alpha = 0.05$. In particular:

- A KS-test checks whether both samples are based on the same random variable or not. The equality of underlying random variables can not be rejected significantly.
- Welch's t-test is used to check for the equality of means. Again, this equality can not be rejected significantly.
- Levene's test checks for homoscedasticity, i.e. the equality of variances. This equality can not be rejected significantly.

Certainly, if the null hypothesis is not rejected, it must not automatically be accepted as being true. However, these results show that potential differences are not large enough for being detected while not exceeding a specific chance of error with respect to the available sample size. In this light and along with a visual comparison of those distributions in Fig. 6.26, both user models can be considered to produce nearly the same prediction quality (with only insignificant differences). Both user models are hence equivalent with respect to this specific collaborative filtering model for prediction.

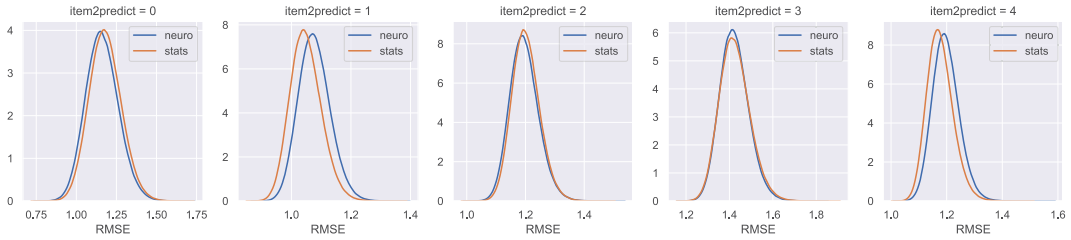


Figure 6.26: Item-related RMSE distributions for the CF approach using the cosine similarity

Machine Learning Regressors

Another frequently used approach for prediction is given by adaptive regression models. Such models assume a mathematical function with variable parameters (degrees of freedom) whose values have to be determined through optimisation against given data. Usually, the entire data set is split into a training set and a testing set, respectively. For the upcoming analysis, the common ratio of 70/30 is chosen, i.e. the training set consists of 70% of the entire data record and the prediction quality is then tested on the remaining 30%. Additionally, selecting data points for the training set is done randomly. The entire prediction process is then repeated several times to cover quality fluctuations due to (un)fortunate training set selection.

Let k denote the item to be predicted and let M be a particular regressor model. The task is to find a functional dependency between a user's feature vector $F_N(k)_{u,\bullet}$ and the target variable, i.e. the mean rating $\mu_{u,k}$ for item k or its standard deviation $\sigma_{u,k}$, respectively. To this end, a model M is fitted via

$$M\colon F_N(k)_{u,\bullet} \mapsto \mu_{u,k} \quad \text{or} \quad M\colon F_N(k)_{u,\bullet} \mapsto \sigma_{u,k} \tag{6.43}$$

where all feature vectors come from the training set. After all of M's degrees of freedom have been determined, the predictors for the remaining test set are given as $\pi_{u,k} := M(F_N(k)_{u,\bullet})$. The prediction quality of M can then be assessed through the RMSE. This analysis involves the following regression models with their standard implementations provided by scikit-learn (cf. Pedregosa et al., 2011):

- **Support Vector Machine (SVM)**
- **Decision Tree (Dtree)**
- **ElasticNet Regression (ElasNet)**

	KS-test		Welch's t-test		Levene's test	
	mean	std	mean	std	mean	std
MLP	n. rejected	n. rejected	n. rejected	n. rejected	n. rejected	n. rejected
SVM	n. rejected	**rejected**	n. rejected	n. rejected	n. rejected	**rejected**
DTree	n. rejected	n. rejected	n. rejected	n. rejected	n. rejected	n. rejected
RForest	n. rejected	n. rejected	n. rejected	n. rejected	n. rejected	n. rejected
ElasNet	n. rejected	n. rejected	n. rejected	n. rejected	n. rejected	n. rejected

Table 6.6: Hypothesis testing for the RMSE distributions resulting from the neuroscientific and the behavioural user model

with Scikit's default settings as well as

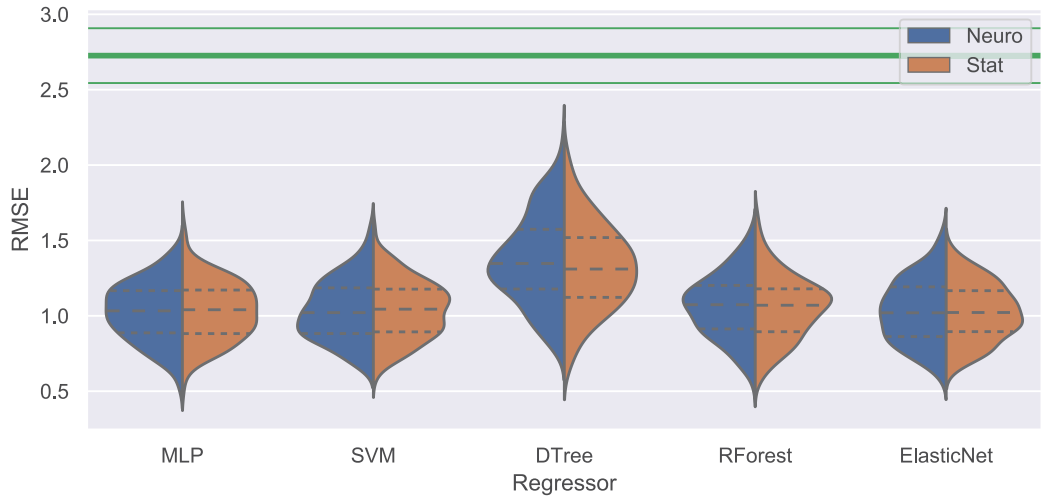
- **Random Forest (RForest)** with ten underlying decision trees
- **Multi Layer Perceptron (MLP)** with logistic activation function and ten hidden layers with sizes of 100, 80, 60, 50, 40, 30, 25, 20, 15, 10 neurons

and Scikit's default settings otherwise.

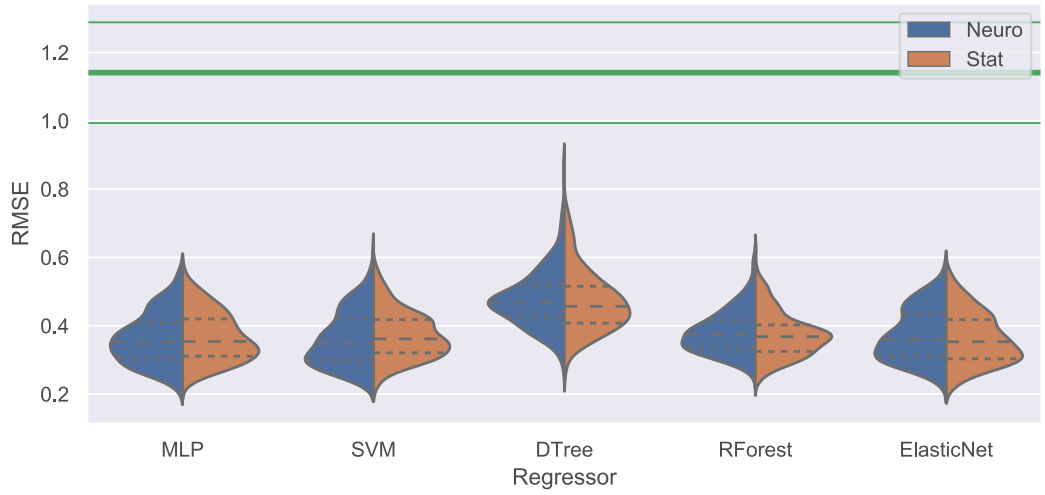
Figure 6.27 depicts the RMSE distributions of all regression models in a violin plot together with a user model partition to facilitate a visual comparison. It is notable that all distributions based on the neuroscientific and the behavioural user model exhibit almost complete overlaps while differing from the control group ($\pi_{u,k} = 1\ const.$). Already with this visual comparison it is apparent that both user models are equivalent in terms of prediction quality. The corresponding distributions for both user models were tested to determine whether they originate from an identical underlying random variable (KS-test), have equal expectations (Welch's t test), and exhibit equal variances (Levene's test). The results are shown in Tab. 6.6. Only for the SVM with the standard deviation as the target variable, both resulting RMSE distributions possess a significantly different shape and variance. For all the other regressors, both user models do not induce significant differences in prediction quality. Therefore, the neuronal features are equivalent to the behavioural features according to Eq. 6.36 in (almost) all cases.

Interpretation

The results obtained were neither certain from the outset, nor are they even inherent in the model. The neuroscientific feature space has got a higher dimensionality and thus



(a) target: feedback expectation



(b) target: feedback variance

Figure 6.27: RMSE distributions for different regressor models. The green lines represent the mean (thick line) and the σ-interval (thin lines) of a reference recommender ($\pi_{u,k} = 1\,const.$).

provides much more degrees of freedom. Moreover, the input data has undergone a serious transformation. Under these conditions, one would have expected medium or even large prediction quality differences rather than equality. This equality is furthermore prevalent in (almost) all regression models and in CF-clustering, respectively. The high frequency of model equivalence across multiple prediction techniques is a strong indication for a latent characteristic of the neuroscientific user model which has to be interpreted. A possible interpretation is that the associated cognition vectors still represent the real user behaviour instead of just being mathematical artefacts. This being said, one may conclude that the information injection which is needed to expand a two-dimensional into a five-dimensional space is realistic and not biasing the true decision-making. In other words, the additional information coming from the assumed cognitive process within the PPC model is in absolute concordance with real human behaviour. Following this line of argument, just another strong hint has been found for the plausibility and applicability of the PPC model for a user rating scenario.

6.9 Chapter Summary

The reason for investigating the neurological origins of human uncertainty comprises two dimensions.

The main reason was that contemporary solution strategies (of turning human uncertainty into possible benefits for predictive data mining) are merely statistical and technical procedures (cf. Ch. 5). Instead of approaching this problem only phenomenologically, one can also take an inner perspective which is based on human cognition as represented in theories of psychology and neuroscience. Since such a strategy had not yet been pursued specifically for human uncertainty in the field of predictive data mining (at least as far as is known), there was hence a research gap that needed to be addressed for the overarching concept of human-like systems. A comparison between the results of the inner and the outer perspective was promising new insights about the future design of such systems. For both models, a holistic prediction of feedback distributions with uncertainty representation can be achieved. This is a great advantage compared to the current research standards, where only individual draws are taken into account and accepted as being credible. Initial results indicate that empirical models work just as well as theoretical models with regard to specific human characteristics. On the contrary, the integration of new human attributes can be directly implemented when

theoretical models are used. For example, the population frequencies could be modelled explicitly as a function of stress (cf. Vanitha and Krishnan, 2016) or fatigue (cf. Saroj et al., 2003) since such dependencies have already been investigated in medical research. In a purely empirical model, these correlations have to be learned again for each use case and each human characteristic. This would generate knowledge that is only valid in specific situations and under specific circumstances. As opposed to this, theoretical models offer a universal theory that may be applied to all use cases by deduction. This might be an advantage which keeps future systems simple and clear.

The secondary reason arose from the fact that doubts remained after the measurement of human uncertainty in Ch. 3 as to whether it is indeed inherent in human beings or merely induced by the measurement itself. Following the latter opinion may imply that decisions are actually reliable and that uncertainty only emerges when environmental factors change, e.g. when a rating has to be repeated several times with different preceding items. The plausibility of this hypothesis is invalidated, since the pdf-rating (no altered item histories) generates probability densities that do not differ significantly from those of the re-rating (altered item histories). Even though the measurement method can be excluded as the origin of feedback uncertainty, this is no evidence or indication that this uncertainty is an inherent characteristic of human beings.

This question was implicitly addressed by pursuing the main goal described above. The theory of probabilistic population codes provides a popular and adequate cognition model from which unreliable user feedback arises naturally. Foremost, all epistemological quality criteria that have been laid down in Sec. 6.1 are fulfilled. The model is able to explain human uncertainty and has demonstrated a remarkable accuracy for reproducing all uncertain feedback distributions. Due to the implementation as a user model, it is simple to integrate into existing systems (and performs equally well compared to a behavioural model). The parameters that have been learned from real user behaviour thereby led to interpretations which are supported by the relevant medical literature. For example, the frequency distribution of common neuron populations is virtually replicated and possible parameter correlations also point to documented facts: (1) increased attention leads to less uncertainty, (2) a population holistically employs each neuron for information encoding including noise correlations, and (3) the system performs self-initiated energy optimisation. This theory is also well connected with other (fundamental) models such as tuning curve models, Poisson-like noise, agency, place cells, and cue integration. Furthermore, neither the literature search in Sec. 6.2 nor the own research

was able to identify internal inconsistencies at any time. The theory of probabilistic population codes hence provides a solid basis for considering the phenomenon of uncertainty as inherent in human beings rather than ascribing it to an unfortunate method of measurement. It can hence be assumed that all data sets containing explicit user feedback are probably affected by human uncertainty. Consequently, all impacts of human uncertainty on the assessment of prediction techniques do most probably apply in reality, albeit it often remains undiscovered for uncertainty has seldom been recorded so far.

A supposed weakness of this model may be its transformation from a two-dimensional into a five-dimensional vector space despite the absence of further information. This criticism is only conjectural, as the information is obtained from the cognitive process as represented by the tuning curve shapes, the population set-up, the distribution of neuronal noise, and the utilised decoder function. As this thesis advocates the development of more human-like systems, it will also part with the paradigm that only pure quantitative data (e.g. user ratings) is considered to be valuable. Following this line of argument rather involves not denying the possibility that the human nature or the functioning of the brain can constitute a source of valuable information. The results in this chapter indeed foster the interpretation that the postulated cognition process is almost completely in harmony with human behaviour, i.e. the neuronal description of a user is equivalent to considering the exhibited behaviour. This is a notable indication that the information injection truly represents higher cognition and decision-making.

7 | Discussion

The purpose of this chapter is to recapitulate the results of this thesis, to draw conclusions for the overall picture of user data in predictive data mining, and to raise novel ideas to produce human-like systems. Some parts of this chapter are mainly based on my work Jasberg and Sizov (2019). In particular, Sec. 7.1 and 7.2 have been published almost verbatim there. Moreover, Sec. 7.1 has also been published in Jasberg and Sizov (2018a) almost verbatim. All these sections underwent contentual changes such as the underlying storyline or the addition of the PPC concept.

The essence of this thesis is to evaluate current developments in the field of predictive data mining, especially the predominant role of accuracy-driven research and the subordinated role of user models and their representation of human characteristics. Therefore, four research objectives have been formulated in Ch. 1, i.e.

A) revealing the existence of human uncertainty in a realistic use case,

B) demonstrating the impact of human uncertainty and effects of neglecting,

C) introducing methods for adequately dealing with human uncertainty,

D) substantiating a human-inherent origin and propose more human systems.

7.1 Results and Insights

Research goals A, B, and C collectively demonstrate the importance of data quality for the field of predictive data mining, especially for the use case of recommendation and personalisation. In particular, there are two important factors of data reliability that have to be considered:

1. People make decisions in dependence of their current contextual situations, that is, users tend to give different feedback under different circumstances. A comprehensive research has developed on the context-dependency of user feedback focusing on the impact of specific surroundings (cf. Hu et al., 2014; Zhao et al., 2016, 2017). These dependencies can often be associated with constant biases (e.g. correlations) that can be ruled out of affected data records.

2. Human decision-making is subject to some natural variability, i.e. even when a constant situational context can be assumed, users give different feedback when the feedback task is repeated only a few moments later. In this thesis, this phenomenon is referred to as human uncertainty. Even if the repetition – which can technically be interpreted as a context change as well – is omitted in order to have the probability density directly entered instead, exactly the same uncertainty can be observed.

The second factor is in the main scope of this dissertation. The core of this present work is to prove the existence of human uncertainty in explicit user feedback and to report on its impact on comparative assessments of prediction engines and personalisation approaches. The key messages to be shared are as follows:

User feedback are distributions: Based on the latest research in the field of neuroscience, cognition may be based on inner distributions which are constantly updated by a complicated generative process within the human cortex. In consequence, results of human decision-making yield a certain degree of volatility and must be seen as a distribution itself. This volatility – which is denoted human uncertainty in our context – can be explained by the irregular release of neuromodulators like dopamine and acetylcholine. This volatility of user feedback has been independently discovered in a simple user study.

Metrics of distributions become distributions themselves: Based on latest research in metrology (the science of accurate measurement), the uncertainty of quantities propagates with respect to a specific mathematical model when composed quantities are computed. In a probabilistic sense, the composed quantity is distributed by a probability density which emerges as a convolution of all arguments' densities. Typical approaches to determine a resulting distribution are Monte-Carlo simulations as well as the Gaussian Error Propagation.

Every ranking is subject to an error probability: Transferred to the comparative assessment of data mining approaches, the results of well-established accuracy metrics turn out to be distributions rather than single scores. It is not unusual that two such distributions have an intersection, i.e. there is a probability of ranking inversions when only single draws are considered. This error is strongly dependent on human uncertainty but also on the prediction quality. The better two systems perform, the more they have to differ from each other in order to enable a statistically sound ranking.

Improvements are limited: There is a special case of ranking error. When improving a system, it is possible to achieve a specific level of accuracy, so that further improvements must be so much better to be recognised that they lie outside the defined range and are thus impossible to achieve. This limit is called magic barrier and marks an offset on an accuracy metric that can not be undercut. Therefore, this barrier represents a natural limit for the improvement of prediction systems.

Existing solutions are not satisfying: Trying to reduce uncertainty that is already existent in data sets has not been successful since reported solutions are logically inconsistent and origin from simply ignoring unpleasant data. Rather, one has to start accepting the uncertain nature of user data and acknowledge this uncertainty in related (comparative) assessments. This acceptance should be reflected within research contributions and further efforts should be made to predict uncertainty beforehand. This can be achieved by using adequate user models.

As already stated at the beginning, there have been certain doubts as to whether human uncertainty is indeed human-inherent and random by nature or just produced by the measurement procedures applied in this thesis. For this reason, a follow-up

research was started which should address this question and simultaneously provide hints for possible design policies concerning human-like systems. With regard to research objective D, the following insights can be reported:

A human origin is supported by neuroscience: The model of probabilistic population codes is a good candidate for explaining human uncertainty by natural noise within the nervous system. When trained on the RETRAIN record, this model has demonstrated an astonishing accuracy for reproducing all uncertain feedback distributions. Also, recommendation analyses foster the fact that the postulated cognition process is completely in harmony with human behaviour, i.e. the cognitive description of a user is equivalent to considering his demonstrated behaviour. All deduced assumptions from this model are supported by the medical literature that has been reviewed. The frequency distribution of gauged neuron populations is virtually replicated and possible parameter correlations also point to well documented facts: (1) increased concentration leads to less uncertainty, (2) a population holistically employs each neuron for information encoding, and (3) the system performs self-initiated energy optimisation.

This model is easy to implement into modern systems: The utilised theory was translated into a user model, i.e. a mathematical representation of a user's past behaviour. In doing so, neuronal parameters are used to span a high-dimensional feature space. Therefore, one can simply perform computations directly on this new space and proven methods of machine learning remain unchanged. Various techniques of machine learning have been tested using this new feature space and computations always ran smoothly. Only the necessary pre-computation, i.e. the user classification is time-consuming and computationally intensive.

Behavioural models perform equally well: The PPC model works well for computer science and, on an epistemic level, also for theoretical neuroscience or computational neuroscience, respectively. From the computer science perspective, the behavioural model performs equally well, i.e. both models lead to the same accuracy when feedback distributions are holistically predicted. This confirms all postulates about the cognition process since the induced information injection is coherent to reality. However, the PPC model provides no additional benefit in terms of prediction so far when compared to a simpler model.

Now that all the facts and insights from all previous chapters have been collected, these can be merged and interpreted in the light of this thesis' scopes and topics.

7.2 Interpretation

As already explained in Ch. 1, recommender systems are based on the storage of explicit and implicit feedback in a user model as well as on machine learning algorithms that use this user model as a learning set. The study of Enríquez et al. (2019) shows that current research focuses almost exclusively on optimising the accuracy of machine learning algorithms. Conversely, it can be concluded that no great importance is assigned to the user model and explicit user feedback. Ricci et al. (2010) even describe that the user model is often a mere collection of user feedback, which may seem very simplistic for realistic problems. The results of this thesis show that the assessment of data quality can indeed be of considerable relevance. The lack of reliability of (explicit) user feedback means that the true measurand remains unknown and only tendencies can be recognised with a particular degree of uncertainty. Any technical processing of this feedback has been shown to propagate this uncertainty. Accordingly, an important interpretation for the research of recommender systems is: The credibility of knowledge that is gained about people based on their interactions with digital systems is limited.

Furthermore, the results show that the continuous improvements in accuracy through machine learning optimisation should be questioned sensibly since it can lead to more or less high probabilities of errors during (comparative) assessments. As explained above, the current status quo includes a predominant role of accuracy-driven research and a subordinated role of user models and their representation of human characteristics. This imbalance implies that those probabilities of error cannot be evaluated at all. Accordingly, another interpretation of this thesis' results is: With the current orientation of research, it may not be possible to distinguish real improvements from false improvements. Based on the absence of plausible solutions, it can be assumed that there is no real solution to the probabilistic nature and its inconveniences for (comparative) assessments. This is supported by the fact that physicists still explicitly measure uncertainty and always publish results together with a propagated uncertainty. So the solution might be to explicitly account for uncertainty in order to avoid possible misinterpretations. This methodology is exactly what might enrich research in the field of recommender systems for the future. Overall, everything points to the following interpretation: All

results presented above serve as an indication that the contemporary research about recommender systems may require another important dimension, i.e. the credibility evaluation of reported research findings.

So far, the results of this thesis have been interpreted for the example of recommender systems, i.e. systems that predict each user's preference for particular items and present the most preferred items as recommendations. Against this background, what interpretations can be derived for the entire research area of predictive data mining? Even in this general case, according to Weiss and Indurkhya, it is always about the prediction of a target variable based on causal or correlative relationships with other variables from a large data pool (cf. Weiss and Indurkhya, 1998, p. 7). For the general case of predictive data mining, Weiss and Indurkhya mention fraud detection, (online) marketing, healthcare outcomes, and investment analysis as the most common use cases (cf. Weiss and Indurkhya, 1998, p. 7). Surely, key variables might be present in all these cases, which derive from human behaviour or decisions. This means that regardless of whether the target variable itself represents human characteristics (e.g. user preferences), data based on human behaviour, including decision-making, needs to be considered every now and then. Consequently, these key variables are likely to be subject to some degree of human uncertainty as well. In summary, the findings described in this thesis indeed demonstrate that uncertainty is very likely to be existent in any database containing implicit or explicit human feedback which impacts predictive data mining in general. One possible example might be the prediction of terminations inside a company based on employee satisfaction, performance ratings from superiors, as well as other variables (e.g. number of previous promotions, number of vacation days, etc.). In particular, employee satisfaction and performance ratings represent human feedback that is most likely subject to uncertainty. In this case, prediction quality would also be a distribution and possible optimisations must include a thorough probabilistic analysis. These explanations allow the interpretation that most of the predictive data mining is likely to be affected by the phenomena presented here since human feedback is conceivable in many use cases.

Finally, what future perspectives do these results and interpretations unveil? In addition to broadcasting statistical methods, a suitable solution might be to predict uncertainty, e.g. in order to allow for a differentiation between difficult and easy variables. Said and Bellogín (2018) has already initialised this research by using uncertainty to classify users into easy and difficult ones. By sensibly composing learning sets with

different fractions of simple and difficult users, the authors were able to reduce the overall uncertainty of accuracy. The prediction of uncertainty further allows asking additional strategic questions, e.g. how much uncertainty is one willing to accept, or whether a planned (but not yet realised) innovation will ultimately cause enough optimisation to be statistically sound. In this light, it makes sense to develop more complicated user models (or data models in general). In the next section, a new perspective of future system design will be proposed with the goal to further sensitise predictive data mining for human beings, i.e. to better explain individual user behaviour and to stronger account for human characteristics.

7.3 Systems with Empathy for the Human Nature

The demand for human-like systems has coined the name "systems with empathy for the human nature" which has been frequently used in conference talks and private communications. According to Oxfords Advanced Learner's Dictionary, the term *empathy* can be defined as "the ability to understand another person's feelings, experience, etc." (Oxford University Press, nd) and can hence be understood synonym to *sensitivity* or *intuition* for the human nature in general. The demand for systems with empathy for the human nature epitomise an important claim: Systems of predictive data mining should seek to further understand human beings by means of their individual psyche, the neurological foundations of behaviour, emotional states, and life circumstances.

At this point, the question arises why such systems might be needed in the future. Sizov elucidates that the "appropriate interpretation of collective [human] feedback requires the development of suitable models that [...] 'explain' observations" (Sizov, 2017b, p. 869). However, Enríquez et al. (2019) demonstrates that the majority of contemporary research solely reports on improving machine learning techniques, allowing for the interpretation that the explanation of individual human beings is often reduced to solely finding optimal weights of a machine learning model. This conclusion has also be drawn by Sizov who states that "model components and parameters [themselves] are often interpreted as an 'explanation' of observations" (Sizov, 2017b, p. 869). The rationale for reflecting on this methodology can be summarised as follows:

> "Although these systems are useful for both users and service providers, the main downside is the limited interpretability and explainability of the

> data. Such limitations in both interpretability and explainability translate in using data without understanding the root-cause of behaviors." (Ferwerda et al., nd)

This quotation addresses two points that are important for the argumentation in this section. The first point is the usefulness of the current methodology and the fact that algorithms of predictive data mining apparently work well. The successful use of such algorithms on Amazon, Netflix, Spotify etc. is obvious and is not to be disproved in this thesis. However, this thesis revealed that the true extent of prediction quality remains unknown and that the credibility of detecting further improvements must hence be questioned. It has been demonstrated that comparative assessments are often not as straightforward as initially assumed and may be subject to error probabilities similar to a coin toss. Hence, it may not be possible to distinguish real improvements from false ones. At this point, certainty can probably only be generated if features such as human uncertainty are explicitly taken into account, i.e. if more attention is paid to the human being and his characteristics. The second point addresses interpretability and explainability, which according to Ferwerda et al. (nd) does not exist in the described methodology and thus, the true reason for behaviour may not be identified. This assertion is supported by the research of Sizov who was able to prove by hypothesis testing on a data set with collected uncertainty information that a "considerable fraction of users exhibits some (unfitting) behaviour that contradicts the [tuned] model" (Sizov, 2017b, p. 870). This is an important signal that a deeper understanding of the human being itself should be sought.

The main question is how such systems need to be designed in order to better explain human behaviour. A proposal is made by Ferwerda et al. who summarise that

> "recent work has thus started to adopt a more theory-driven approach by including psychological theories and models to improve personalized systems. These systems take advantage of psychological theories/models to explain and predict behaviors of users, and allow for a deeper understanding of users' behavior, preferences, and needs, which in turn also lead to more generalizable results." (Ferwerda et al., nd)

This idea was applied in this thesis to cover the phenomenon of human uncertainty. Respective prototypes (i.e. proofs-of-concept) have been developed using two disparate approaches, namely:

- a psychological or behavioural model based on additional observation of the particular human characteristic that is in scope. In this thesis, the mean and variance was used to represent a user's feedback.

- a theoretical neuroscientific model based on the theory of probabilistic population codes that transforms a user's feedback distribution into possible underlying neuronal states of this user.

Both systems allow for a holistic prediction of feedback distributions with uncertainty representation. This is a major advantage to the current research standards in which only single draws are considered and assumed to be absolutely credible. The results demonstrate that both approaches fulfil their tasks of representing and predicting human uncertainty with equal quality. In terms of runtime and resource efficiency, the behavioural model has proven to be more useful for fast computations. However, the theoretical model has been implemented in this thesis for the first time and there is still an enormous potential for technical improvement. In fact, not all technical possibilities have been used to increase efficiency since the focus has clearly been on the epistemic nature of this model. Nevertheless, the very first results show that empirical models for particular human characteristics perform just as well as theoretical ones. In contrast, the integration of new human variables can be implemented directly when using theoretical models. For example, the population frequencies could be modelled explicitly as a function of stress (cf. Vanitha and Krishnan, 2016) or fatigue (cf. Saroj et al., 2003) since such dependencies have already been investigated in medical research. In a purely empirical behaviourist model, these correlations need to be re-learned for each use case (e.g. ratings, skipped songs, click behaviour, etc.) and every human characteristic. One thus receives a collection of knowledge that is valid only in special situations and under specific circumstances. In contrast, theoretical models offer a universal theory that can be applied to all use cases by deduction. The selection of a particular situation (stress level, fatigue, emotional state, etc.) would determine initial parameter settings of this model which are then tuned by considering the behaviour of all users considered. This might be an advantage that keeps future systems simple and clear.

The final question regarding systems with empathy for the human nature is how they might look like in the near future and what obstacles need to be overcome. As a matter of fact, the phenomenon of human uncertainty is just one single feature taken from a plethora of human characteristics. Yet another good example for a human characteristic

impacting predictive data mining is the fact that "different users tend to have different internal scales" (Koren and Sill, 2011, p. 117) leading to the phenomenon that "one user can take '3 stars' as similar to '4 stars', while another user strongly relates '3 stars' to low quality, being similar to '1-2 stars'" (Koren and Sill, 2011, p. 117). This ultimately challenges the validity of numerical user feedback and restrains the comparability. Given the difference between prediction and subsequent user feedback, aggregating these discrepancies from different users may not be appropriate because the users involved may perceive them differently. This example demonstrates that there are still human characteristics waiting to be found that might be of importance in future predictive data mining. Yet, these features might be neither obvious nor easy to find, and knowledge about their implications and manifestations has to be built eventually. Despite these inherent challenges, this can still be an open field of undiscovered treasures. Future systems with empathy for the human nature could possibly take into account more of these psychological and neuronal features as well as social and ambient factors. The possibilities for capturing psycho-social, ambient, and even neuronal features already exist nowadays: Let's think about the ever-increasing amounts of data through mobile devices and the Internet of Things (useful for geotagging), or let's think of fitness tracker (useful for stress detection) or even neurofeedback meditation via mobile apps where EEG signals are measured and evaluated to tell a user about the individual degree of relaxation. Using this data in terms of empathy for the human nature may require new ideas for its implementation including models that explain human behaviour in the light of this information. This involves an interdisciplinary collaboration between neuroscience, psychology, applied computer science, and applied mathematics.

Regarding the state of research on systems with empathy for the human nature, it can be noted that the first steps towards this direction have already been taken. At this point, the preliminary work of many authors (elaborating on data quality in recommender systems) would have to be mentioned once again (see Ch. 2). These (phenomenological) contributions – and future contributions as well – only need to be consolidated with a view to developing new systems with empathy for the human nature. This thesis presents a possible way to achieve this consolidation and to design respective systems by transferring neural theories of cognition and mind into the user model itself. Based on the findings of Enríquez et al. (2019), it can be said that the research community advocating the introduction of such mindsets and methodologies is still too small and disproportionate to the mainstream (conducting accuracy-driven

optimisation of machine learning). It would be desirable that more researchers consider adopting an additional focus on explaining the individual user in the light of human nature. After all, although the first steps in this direction have been taken, there is certainly still a lot of research to be done.

7.4 Recommendations for Further Research

The recommendations for further research basically comprise two dimensions: On the one hand, they are intended to show ways to overcome the limitations of the research conducted in this dissertation. On the other hand, they aim to present possibilities of how systems with empathy for the human nature can be further promoted. Therefore, this section is divided into two parts. The first part focuses on this dissertation and describes new research questions that might arise from it (local recommendations). The second part considers the dissertation itself only as a part of a larger project and describes the possible exploration of this big picture (global recommendations).

Local Recommendations

Like any research, the work presented in this dissertation has certain limitations. These will be addressed thematically and ideas for further research will be derived therefrom.

First and foremost, the RETRAIN study was conducted within the DACH countries and is thus limited to a German-speaking cultural area. It is hence unclear whether there are cultural or location-dependent differences in human uncertainty and whether the measured uncertainty is indeed indicative for international data sets. A total of 67 users participated in the RETRAIN study and human uncertainty was measured for five items so that a total of 335 user-item pairs are available. This number of observations is rather small and the results of this study are therefore only indicative. This mostly affects the share of uncertain user ratings within the overall data set as well as the distribution of human uncertainty. Nonetheless, the analysis of comparative assessments is not affected, as this was theoretically derived for an unspecified number of observations. This data set is also too small for a phenomenological analysis of uncertainty effects on real recommender systems or basic machine learning techniques. Typical data sets for such research contain several million observations. For example, the latest MovieLens data set currently contains “25 million ratings and one million tag

applications applied to 62,000 movies by 162,000 users" (GroupLens, nd). Even the oldest and smallest record of MovieLens from 1998 still has "100,000 ratings from 1000 users on 1700 movies" (GroupLens, nd). The Yahoo! R2 data set even contains "over 717 million ratings of 136 thousand songs given by 1.8 million users of Yahoo! Music services" (Yahoo! Research, nd). It is hence essential for further research that the existing data set is extended. Data sizes in the order of MovieLens and Yahoo! is beyond the reach of an online experiment, but a data size of 100k ratings can be considered as realistic. The RETRAIN study gathered ratings with uncertainty information for 335 user-item pairs within one week and cost about 400 EUR in total. An extrapolation shows that more than 1300 ratings could be obtained in a month and the expected cost of 1600 EUR is still acceptable for a university chair. If this study were carried out in 77 countries (about 40% of all countries) and the costs were borne by one chair in each country, a common data set with 100k ratings could be created and used for international research. The pdf-rating procedure could help to reduce the effort and might enable the gathering of uncertainty information for more user-item pairs in finite time. Such a data set would be very valuable for studying the impact of human uncertainty on recommendation and machine learning in general. Results obtained from such data can be considered as representative rather than indicative. Furthermore, the exploration of local and cultural differences can – aside from the epistemic value – be used to fine-tune future systems.

Concerning the measurement of human uncertainty, it must be said that the existing methods do not suffice for multiple reasons: For the re-rating procedure, it has already been proven that users quickly get tired and no longer provide meaningful feedback (cf. Sizov, 2017a, p. 897), so that four to five re-ratings mark the maximum at which this procedure measures validly. With only five observations per user-item pair, the confidence interval for the mean and the standard deviation of each feedback distribution is too large. For this reason, deduced statements are only indicative, e.g. for the comparative assessment of recommender systems. The pdf-rating procedure is perceived as being too complicated and the currently used transformation to determine the statistics for a feedback distribution also induces too large confidence intervals. Furthermore, the reliability of this method has not yet been tested, which is imperative to catch up on. This measurement method hence needs to be explored in more detail and also needs to be provided with an improved (i.e. a simpler) user interface. Another aspect that belongs to the measurement methods is the employed feedback scale. The RETRAIN study used a 5-star scale as it is used by Amazon. Other scales such as those with two,

three, or ten response options have not yet been investigated. Consequently, a possible dependence between the employed feedback scale and human uncertainty is currently unclear and needs further investigation. Therefore, a follow-up study is required which should not only include repeated pdf-ratings but also a combination of re-ratings and pdf-ratings for different scales and with different user interfaces. Besides, only explicit user feedback was addressed in this thesis. However, implicit user feedback is probably subject to missing reliability as well. For example, Mao et al. (2019) examined the reliability of user click behaviour on a page with search results and inferred the induced uncertainty of item-relevance estimation. Based on the repeated simulation of user clicks, the authors concluded that

> "user clicks carry implicit relevance feedback that is valuable for improving the ranking performance of Web search engines. However, the click signal is noisy and affected by different kinds of behavioral biases [...], making it systematically different from true relevance." (Mao et al., 2019, p. 125)

Indeed, by using Bayesian click models, it is possible to determine the uncertainty of relevance estimation that originates from unreliable user clicks (cf. Mao et al., 2019, p. 126). At this point, Mao et al. have provided the proof-of-concept that implicit user feedback must also be represented by distributions. The authors also provided approaches to derive such distributions from observed user behaviour. Further research may take this result as a starting point and aim to combine different kinds of implicit feedback with given explicit feedback in the sense of Bayesian cue integration. In other words, implicit and explicit feedbacks may be aggregated to a general preference probability distribution. The advantage of such modelling is that Bayesian cue integration prefers more precise information, i.e. the narrowest distribution. Consequently, it may be possible to reduce the variance of a resulting preference distribution through different information cues which then positively affects, for example, the comparative assessment of future systems.

Another point relates to human uncertainty itself. By comparing the re-rating procedure and pdf-rating procedure, a methodological induction of this phenomenon could be excluded on an indicative level. Likewise, the existence of an adequate neuronal model for uncertain decision-making provides further indication that this phenomenon is indeed human-inherent. New experiments could help to overcome the indicative character of these findings and provide more clarity concerning the existence and the

extent of human uncertainty. To this end, a collaboration with a neurological or psychological research institute would be particularly beneficial. This is because the act of memorising new information is located in the human hippocampus (cf. Birbaumer and Schmidt, 2018, p. 653) and, therefore, patients with lesions in this brain area cannot transfer new information into long-term memory (cf. Speckmann et al., 2019, p. 277). Having access to such participants for further study may allow to finally exclude the assumption that human uncertainty originates from a changed item history during repeated evaluations. In other words, each repeated item presentation would be like rating this item for the very first time. In practice, having access to such patients is not even necessary at all. Karnath and Thier (2012) describe that in transcranial magnetic stimulation (TMS) a highly focused magnetic field can be generated near the skull, which induces a small current at the brain surface (cf. Karnath and Thier, 2012, pp. 26–27). This technique allows to create so-called transient (i.e. temporary) virtual lesions and is already in use for neuroscientific research (cf. Karnath and Thier, 2012, p. 27). From an ethical point of view, this technique is completely acceptable as it is non-invasive and virtually harmless: "TMS can be employed in almost any healthy volunteer who meets a few basic health-related criteria" (Glimcher and Fehr, 2014, p. 95). Moreover, TMS has already been shown to interfere with memory functions: Siebner and Ziemann describe that repeated TMS treatment (rTMS) can prevent the formation of memory in the mouse model for about 30 minutes (cf. Siebner and Ziemann, 2007, p. 390). By completely excluding memory effects from a repeated rating task, it would additionally be possible to discover the true extent of uncertainty (which might be larger than in the RETRAIN study) and it would certainly allow for more rating trials since fatigue and boredom would not arise. This may allow testing the assumption of normality for the feedback distributions using a larger number of observations.

From the field of predictive data mining, the research of this thesis focused mainly on recommender systems along with their comparative assessment. Given a larger data record, one could also explore possible effects on basic machine learning techniques. This is an important step since these techniques can be found in almost every application of predictive data mining nowadays. Such research probably furnishes more general conclusions for a broader spectrum of use cases. When focusing on recommendations, one basically considers very similar variables, i.e. user ratings are inferred from other user ratings. However, how does human uncertainty affect other use cases in which variables are causally related but not quite similar? Such an example may be the

prediction of termination probabilities within a company based on multiple variables including human estimations. Let's assume that employee satisfaction measured by an online questionnaire is taken into account. This information is probably uncertain, but its impact on the uncertainty of a predicted termination probability remains unclear. One variable's value certainly may indicate the extent of the other variable through correlation, but this correlation may not be one-to-one. Mathematically speaking, there is a downstream random variable which maps the uncertain employee satisfaction somehow to the interval $[0, 1]$ of termination probability. Accordingly, the uncertainty propagation of a still unknown mathematical model must be taken into account. Such transfer effects do not exist in the case of recommender systems when variables of the same kind are considered. It becomes even more complicated when other variables are added (which is a more realistic case), e.g. the evaluation by a supervisor along with non-uncertain variables such as 'holidays used per year' and 'overtime per year'. Although the effects of all these variables on the probability of termination can be presumed, the aggregation is not obvious (which is why machine learning is used for such tasks). However, the uncertainty propagation is not obvious either so that the existing theory needs a further extension. Moreover, rating a film trailer represents a relatively unimportant decision that may not require much reflection. It is still unknown whether and to what extent human uncertainty occurs in more important decisions, e.g. in choosing a partner on Parship where matching algorithms are used based on an online questionnaire (cf. PE Digital GmbH, nd). Such an experiment would constitute the first step towards a holistic view of human uncertainty for a broader range of predictive data mining.

Regarding the model of probabilistic population codes (PPC), further research is required on runtime and resource efficiency. More efficient algorithms may allow using a finer discretisation of neuronal parameter spaces and support resolving smaller differences of best-fitting cognition vectors. This may improve clustering analyses providing new user insights and better prediction performance. Another starting point for further research focuses on the additivity of agents in terms of Bayesian cue integration. This can be used to examine the aggregation of sub-ratings for a particular item. For example, one could collect ratings for usefulness, workmanship, ergonomics, durability, and delivery time. If the participants additionally specify how important they perceive each of these sub-aspects, this can be used to determine how a unifying feedback density may constitute as a linear combination of all sub-densities. The so-gained knowledge

may allow converting a given user feedback into possible sub-feedbacks in a way that is plausible in the light of neuroscience, i.e. that presumably reflects real cognition. This may help to identify alternative user clusters, e.g. users who appreciate ergonomics or those who prefer durability. Such results may further support revealing new user insights. Further research may also focus on the integration of stress, fatigue, and other tuning curve families into the PPC model, immediately followed by reassessing a possible superiority to the behavioural model. Up to this point, only bell-shaped tuning curves have been considered although sigmoid-shaped tuning curves are also discussed in neuroscience literature (cf. Dayan and Abbott, 2001, p. 15 of Ch. 1). During the research for this thesis, sigmoid-shaped tuning curves have also been implemented for a while. The results demonstrated that the population activity forms monotonically increasing or decreasing curves, even with asymptotes depending on the particular configuration qua cognition vector. Such curves are, for example, used in utility theory (cf. Glimcher and Fehr, 2014, pp. 5–6). Moreover, specific configurations provoke the population response to take the form of a complete sigmoid-function, which is often employed in prospect theory (cf. Glimcher and Fehr, 2014, pp. 113 and 119). Both of these theories are basic to decision-making from the perspective of economics (cf. Glimcher and Fehr, 2014, pp. 113 and 119). Since PPCs come into question as a model of decision-making, the additional ability for physiologically representing functions for utility and value would be a remarkable coincidence. The applicability of the PPC model to these use cases deserves further investigation. In doing so, the principle used in this dissertation can be transferred: The curves for utility and value can be determined for a particular decision-maker through experiments. These can then be mapped onto possible cognition vectors using the same simulation as described in Ch. 6 while the neurological implications can be compared with findings reported in the scientific literature. As this model was recently proven to work well for explaining perception and motor control, the findings in this thesis revealed an even broader applicability. Everything that is known about this theory to this point indicates that it could perhaps be a universal theory for the human brain, i.e. a single mechanism that acts as a solution strategy for all brain tasks. It is therefore imperative to investigate this assumption through the further research described above (but not exclusively).

Global Recommendations

The global recommendations are intended to describe a possible exploration of the broader field beyond the research done in this dissertation. The contextual framework of this dissertation is given by what is called systems with empathy for the human nature, i.e. systems that address the peculiarities of human beings. Human uncertainty has only been one example of this. Apart from this topic, there is certainly undiscovered knowledge that can yet make a huge contribution to future systems with empathy for the human nature. For example, Koren and Sill mention the phenomenon of different individual interpretations of scales (cf. Koren and Sill, 2011, p. 117).

The final question is: Which other human peculiarities and phenomena still exist and how to discover them along with their impact on predictive data mining? This was accomplished in this thesis by the transfer of knowledge from different scientific disciplines, in particular by transferring the concept of measurement uncertainty from metrology and physics to the subject of measuring user feedback. The same principle can surely be applied to find additional aspects of human nature. For this reason, a global recommendation for further research can only be to establish an interdisciplinary collaboration in the form of a research alliance. Based on previous experience, especially with regard to the study of literature on the topic of this dissertation, a cooperation of the following disciplines appears to be very fruitful:

- computer science
- neuroscience
- computational neuroscience
- psychology
- psychometrics
- economics

For example, a similar association exists between economics, psychology, and neuroscience since the late 1990s and gave birth to the new field of neuroeconomics (cf. Glimcher and Fehr, 2014, pp. xii–xiii). It is reported that this "converging group [...] quickly generated a set of meetings and conferences that fostered a growing sense of interdisciplinary collaboration" (Glimcher and Fehr, 2014, p. xiii) which has induced a significant benefit for all these disciplines to this day (cf. Glimcher and Fehr, 2014, p. xxvii). If a consolidation of all these different concepts of decision-making turned out to be successful in creating synergy effects, this can surely be extended to decision-making while interacting with information systems. The beginning can be very similar

to that of neuroeconomics. By announcing special issues and organising conferences, a community is gradually being established. In this community, parties with common interests and goals can coalesce and form research alliances. The question is how such a research alliance, particularly related to predictive data mining, can be equally successful. Such an undertaking can be realised, e.g. in the form of a funded project on the topic "systems with empathy for the human nature". The results of this dissertation together with the results of Sizov (2017a,b) can provide a good basis to indicate the necessity and relevance of such a funded project. If there is a chair for each of the above-mentioned disciplines, each of which is granted two doctoral positions for a period of six years, intensive interdisciplinary research is ensured.

In doing so, which factors of success would play a major role? First of all, a common basis for research must be created. This requires the elaboration of possible research intersections and the rigorous construction of transitions. Especially for the case under consideration in this thesis, this is certainly possible as the intersections are sufficiently pronounced: One possible transition between predictive data mining and (computational) neuroscience has already been emphasised in Ch. 6. In addition, the possibility of modelling utility and value was highlighted as a way to integrate the basic theories of economics. However, economics itself offers numerous experimental proofs that people make decisions violating rationality under specific circumstances (Glimcher and Fehr, 2014, p. 109). Such experimental detection of behavioural peculiarities is, among other topics, the core subject of psychology and psychometrics. Altogether, each of the mentioned fields offers a sufficient number of theories to be integrated into predictive data mining, but there are also many different data sets that can be consolidated in this framework. Another factor for success, which also supports the first one, is to develop a common understanding of science and methodology. This step is not trivial and requires tolerance, respect, trust as well as being open-minded. These conditions are not always met. For example, the first alliances of economics, psychology, and neuroscience faced the problem that this kind of fusion

> "was controversial within their parent disciplines. Many neurobiologists outside the emerging neuroeconomic community argued that the complex normative models of economics would be of little value for understanding the behavior of real humans and animals. Many economists [...] argued that algorithmic-level studies of decision making were unlikely to improve the

> predictive power of the revealed preference approach." (Glimcher and Fehr, 2014, p. xxii)

For this reason, it is beneficial for such alliances to emerge from networking activities at conferences with an interdisciplinary focus. In this way, an open mind is more likely to be found and the common goals are probably shared throughout the participants. This definition of common goals represents yet another factor of success. On this basis, research plans can be developed involving scheduled reports and periodic research colloquia to which representatives of other scientific disciplines (e.g. physics) may also be invited. This regular communication is particularly important with regard to the early identification of further research opportunities across individual disciplines. For example, research results in the field of psychology can open up completely new research perspectives in the field of predictive data mining. In such research arrangements, consolidation of resources may also be organised to go beyond the sharing of data records, e.g. by granting access to a high-performance cluster or TMS equipment for other departments (depending on legal regulations). Finally, how can those collaborations furnish systems with empathy for the human nature? In Glimcher and Fehr (2014), this question has implicitly been answered for the subject of computational psychology:

> "What are the steps of computational modeling? The first step is to take a conceptual theoretical framework, and reformulate its assumptions into a more rigorous mathematical or computer language formalism. But often the conceptual theory is insufficient or too weak to completely specify a model, or it is missing important details. In this case, the second step is to make additional detailed assumptions [...] which complete the model in order to generate precise quantitative predictions. [...] Computational models almost always contain parameters whose values are initially unknown, and the third step in computational modeling is to estimate these parameter values from some of the observed data. The fourth step is to compare the predictions of competing models with respect to their ability to explain the empirical results to determine which model provides a better representation of the cognitive/neural system that we are trying to represent. The last step is often to start over by reformulating the theoretical framework and constructing new models in light of the feedback obtained from new experimental results." (Glimcher and Fehr, 2014, p. 50)

This description perfectly reflects the procedure of transforming the theory of probabilistic population codes into a user model for recommender systems as previously presented. Therefore, these steps constitute a proven concept for application when further peculiarities of human behaviour need to be transformed as well.

This new dimension of user knowledge is substantially different from solely being able to infer likes. It may be possible to create user models with stored personality structures, thinking patterns, cognitive dispositions, and emotional tendencies. The possible applications are manifold and especially those use cases will benefit where knowledge about the human being itself is essential (e.g. dating services, business recruiting, insurance pricing, employment service, etc.). Further research in this area can therefore only be strongly recommended.

Bibliography

Ahmad, S., Cui, Y., and Hawkins, J. (2017). The HTM spatial pooler – a neocortical algorithm for online sparce distributed coding. *Frontiers in Computational Neuroscience*, 11:1 – 15. **[cited on page 11.]**

Alais, D. and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3):257–262. **[cited on page 114.]**

Amatriain, X., Pujol, J. M., and Oliver, N. (2009a). I like it... I like it not: Evaluating user ratings noise in recommender systems. In *User Modeling, Adaptation, and Personalization*, pages 247–258, Berlin, Heidelberg. Springer Berlin Heidelberg. **[cited on pages 29, 30, 31, 32, 33, 34, 36, 41, and 93.]**

Amatriain, X., Pujol, J. M., Tintarev, N., and Oliver, N. (2009b). Rate it again: Increasing recommendation accuracy by user re-rating. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 173–180, New York, NY, USA. ACM. **[cited on pages 16, 29, 30, 43, 51, and 93.]**

Averbeck, B. B., Latham, P. E., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366. **[cited on page 167.]**

Battaglia, P. W., Jacobs, R. A., and Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A*, 20(7):1391–1397. **[cited on page 114.]**

Bauer, J., Dávila-Chacón, J., and Wermter, S. (2015). Modeling development of natural multi-sensory integration using neural self-organisation and probabilistic population codes. *Connection Science*, 27(4):358–376. **[cited on page 115.]**

Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience*, 34(10):3632–3645. **[cited on page 117.]**

Bear, M. F., Connors, B. W., and Paradiso, M. A. (2018). *Neurowissenschaften : Ein grundlegendes Lehrbuch für Biologie, Medizin und Psychologie*. Springer Spektrum, Berlin Heidelberg, 3rd edition. **[cited on page 109.]**

Beck, J., Ma, W., Latham, P., and Pouget, A. (2007). Probabilistic population codes and the exponential family of distributions. In *Computational Neuroscience: Theoretical Insights into Brain Function*, volume 165 of *Progress in Brain Research*, pages 509–519. Elsevier. **[cited on pages 115 and 128.]**

Berg, R. W. (2017). Neuronal population activity in spinal motor circuits: Greater than the sum of its parts. *Frontiers in neural circuits*, 11:103–103. **[cited on page 167.]**

BIPM (2004). What is metrology? `https://web.archive.org/web/20110927012931/http://www.bipm.org/en/convention/wmd/2004/`. Last accessed on Apr 07, 2020. **[cited on page 34.]**

Birbaumer, N. and Schmidt, R. F. (2018). *Biologische Psychologie*. Springer, Heidelberg. **[cited on page 192.]**

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, NY, USA. **[cited on pages 26 and 97.]**

Bobadilla, J., Ortega, F., Hernando, A., and Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, 46:109–132. **[cited on page 26.]**

Chapin, J. K. (2004). Using multi-neuron population recordings for neural prosthetics. *Nature Neuroscience*, 7(5):452–455. **[cited on page 117.]**

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum, Hillsdale, NJ, USA, 2nd edition. **[cited on page 136.]**

Dayan, P. and Abbott, L. (2001). *Theoretical Neuroscience - Computational and Mathematical Modeling of Neural Systems*. Computational Neuroscience. MIT Press, Cambridge. **[cited on pages 105, 112, 117, 119, 120, 122, and 194.]**

Döring, N. and Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Springer, Berlin Heidelberg, 5th edition. **[cited on pages 133 and 135.]**

Doya, K., Ishii, S., Pouget, A., and Rao, R. P. (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. Computational Neuroscience. MIT Press, Cambridge. **[cited on pages 38, 105, 116, 118, 119, 120, 121, and 132.]**

Dresler, M. (2011). *Kognitive Leistungen - Intelligenz und mentale Fähigkeiten im Spiegel der Neurowissenschaft*. Spektrum Akademischer Verlag, Heidelberg. **[cited on page 165.]**

Enríquez, J., Morales-Trujillo, L., Calle-Alonso, F., Domínguez-Mayo, F., and Lucas-Rodríguez, J. (2019). Recommendation and classification systems: A systematic mapping study. *Scientific Programming*. **[cited on pages 3, 4, 16, 18, 24, 25, 77, 183, 185, and 188.]**

Erdmann, A., Erdmann, U., Martens, A., Müller, O., and Paul, A. (2015). *Neurobiologie*. Grüne Reihe. Schroedel, Braunschweig. **[cited on page 165.]**

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433. **[cited on page 114.]**

Faisal, A., Selen, L., and Wolpert, D. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, 9:292–303. **[cited on pages 38, 108, 109, 113, and 121.]**

Ferwerda, B., Tkalcic, M., and Chen, L. (n.d.). Psychological models for personalized human-computer interaction (HCI). `https://www.frontiersin.org/research-topics/11465`. Last accessed on Apr 07, 2019. **[cited on page 186.]**

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138. **[cited on pages 105, 106, 107, 109, 110, and 111.]**

Friston, K., Schwartenbeck, P., Fitzgerald, T., Moutoussis, M., Behrens, T., and Dolan, R. (2013). The anatomy of choice: active inference and agency. *Frontiers in Human Neuroscience*, 7:1–18. **[cited on pages 107 and 112.]**

Glimcher, P. W. and Fehr, E. (2014). *Neuroeconomics*. Academic Press, San Diego, 2nd edition. **[cited on pages 192, 194, 195, 196, and 197.]**

GroupLens (n.d.). Movielens. `https://grouplens.org/datasets/movielens/`. Last accessed on Apr 07, 2020. **[cited on page 190.]**

Heinrich, B., Hopf, M., Lohninger, D., Schiller, A., and Szubartowicz, M. (2019). Data quality in recommender systems: the impact of completeness of item content data on prediction accuracy of recommender systems. *Electronic Markets*, pages 1–21. **[cited on page 26.]**

Henze, N. (2013). *Stochastik für Einsteiger - Eine Einführung in die faszinierende Welt des Zufalls*. Springer Spektrum, Wiesbaden, 10th edition. **[cited on page 57.]**

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53. **[cited on pages 3, 4, 16, 26, 27, 28, 30, 34, 37, and 77.]**

Herrero, J., Gieselmann, Marc anda Sanayei, M., and Thiele, A. (2013). Attention-induced variance and noise correlation reduction in macaque v1 is mediated by nmda receptors. *Neuron*, 78(4):pp. 729–739. **[cited on page 167.]**

Hill, W., Stead, L., Rosenstein, M., and Furnas, G. (1995). Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, pages 194–201. ACM Press/Addison-Wesley Publishing Co. **[cited on pages 28, 29, 30, 31, 32, 35, 36, 44, 51, and 86.]**

Hillis, J. M., Watt, S. J., Landy, M. S., and Banks, M. S. (2004). Slant from texture and disparity cues: Optimal cue combination. *Journal of Vision*, 4(12):967–992. **[cited on page 114.]**

Hu, L., Sun, A., and Liu, Y. (2014). Your neighbors affect your ratings: On geographical neighborhood influence to rating prediction. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 345–354, New York, NY, USA. ACM. **[cited on page 180.]**

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243. **[cited on pages 119 and 120.]**

Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Research*, 39(21):3621–3629. **[cited on page 114.]**

Janczyk, M. and Pfister, R. (2013). *Inferenzstatistik verstehen*. Springer Spektrum, Berlin, Heidelberg. **[cited on page 52.]**

Jannach, D., Zanker, M., Felfernig, A., and Friedrich, G. (2010). *Recommender Systems: An Introduction*. Cambridge University Press, Cambridge. **[cited on pages 3, 4, 22, and 23.]**

Jasberg, K. and Sizov, S. (2017a). Assessment of prediction techniques: The impact of human uncertainty. In *Proceedings of the International Conference on Web Information Systems Engineering*, WISE'17, pages 106–120, Cham. Springer International Publishing. **[cited on pages 63 and 91.]**

Jasberg, K. and Sizov, S. (2017b). The magic barrier revisited: Accessing natural limitations of recommender assessment. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, pages 56–64, New York, NY, USA. ACM. **[cited on pages 1 and 63.]**

Jasberg, K. and Sizov, S. (2017c). Probabilistic perspectives on collecting human uncertainty in predictive data mining. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, UMAP '17, pages 104–112, New York, NY, USA. ACM. **[cited on pages 1 and 41.]**

Jasberg, K. and Sizov, S. (2018a). Human uncertainty and ranking error: Fallacies in metric-based evaluation of recommender systems. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, SAC '18, pages 1358–1365, New York, NY, USA. ACM. **[cited on pages 1, 63, 91, and 179.]**

Jasberg, K. and Sizov, S. (2018b). Neuroscientific user models: The source of uncertain user feedback and potentials for improving web personalisation. In *Proceedings of the International Conference on Web Information Systems Engineering*, WISE'18, pages 422–437, Cham. Springer International Publishing. **[cited on pages 101 and 137.]**

Jasberg, K. and Sizov, S. (2018c). Unsicherheiten menschlicher Entscheidungsfindung in Empfehlungssystemen - oder: Was wir von den klassischen Naturwissenschaften

übernehmen können. *Information - Wissenschaft & Praxis*, 69(1):21–30. **[cited on pages 41 and 63.]**

Jasberg, K. and Sizov, S. (2019). Human uncertainty in explicit user feedback and its impact on the comparative evaluations of accurate prediction and personalisation. *Behaviour & Information Technology*, pages 1–34. **[cited on pages 1, 41, 63, 91, and 179.]**

JCGM (2008a). Guide to the expression of uncertainty in measurement. Technical report, BIPM. **[cited on pages 9, 16, 35, 36, 37, 42, and 105.]**

JCGM (2008b). International vocabulary of metrology – basic and general concepts and associated terms. Technical report, BIPM. **[cited on page 35.]**

JCGM (2008c). Supplement 1 to the GUM - propagation of distributions using a monte carlo method. Technical report, BIPM. **[cited on pages 16, 17, 36, 37, 42, and 64.]**

Karnath, H.-O. and Thier, P. (2012). *Kognitive Neurowissenschaften*. Springer, Berlin Heidelberg, 3rd edition. **[cited on page 192.]**

Kircher, E., Girwidz, R., and Häußler, P. (2009). *Physikdidaktik - Theorie und Praxis*. Springer, Berlin Heidelberg, 2nd edition. **[cited on pages 101, 102, 103, 104, and 169.]**

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., and Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504. **[cited on page 28.]**

Knill, D. C. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719. **[cited on pages 38, 117, and 128.]**

Knill, D. C. and Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research*, 43(24):2539–2558. **[cited on page 114.]**

Kohavi, R. and Thomke, S. H. (2017). The surprising power of online experiments: Getting the most out of A/B and other controlled tests. `https://hbr.org/2017/09/`

the-surprising-power-of-online-experiments. Last accessed on Apr 07, 2020. **[cited on page 16.]**

Koren, Y. (2008). Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 426–434, New York, NY, USA. ACM. **[cited on page 85.]**

Koren, Y. and Sill, J. (2011). Ordrec: An ordinal model for predicting personalized item rating distributions. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 117–124, New York, NY, USA. ACM. **[cited on pages 3, 4, 25, 26, 30, 103, 188, and 195.]**

Krapp, M. and Nebel, J. (2011). *Methoden der Statistik*. Vieweg + Teubner, Wiesbaden. **[cited on page 52.]**

Ku, H. (1966). Notes on the use of propagation of error formulas. *Journal of Research of the National Bureau of Standards*, 70(4):263–273. **[cited on pages 16, 17, 36, and 72.]**

Kubat, M. (2015). *An Introduction to Machine Learning*. Springer, Berlin Heidelberg. **[cited on page 26.]**

Lee, L. (2000). Measures of distributional similarity. *arXiv:cs/0001012*. **[cited on pages 73, 133, and 134.]**

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151. **[cited on pages 73 and 134.]**

Ma, W. J., Beck, J. M., E., L. P., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438. **[cited on pages 39, 42, 110, 114, 115, 116, and 119.]**

Ma, W. J. and Pouget, A. (2009). Population codes: Theoretic aspects. *Encyclopedia of neuroscience*, 7:749–755. **[cited on pages 113, 114, 118, 120, 125, 126, 131, and 132.]**

Mallot, H. A. (2013). *Computational Neuroscience - A First Course*, volume 2 of *Series in Bio-/Neuroinformatics*. Springer, Heidelberg. **[cited on pages 17 and 120.]**

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA. **[cited on page 26.]**

Mao, J., Chu, Z., Liu, Y., Zhang, M., and Ma, S. (2019). Investigating the reliability of click models. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR'19, pages 125–128, New York, NY, USA. ACM. **[cited on page 191.]**

McNee, S. M., Riedl, J., and Konstan, J. A. (2006). Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, pages 1097–1101, New York, NY, USA. ACM. **[cited on pages 5, 6, and 27.]**

Moreno-Bote, R. (2014). Poisson-like spiking in circuits with probabilistic synapses. *PLOS Computational Biology*, 10(7):1–13. **[cited on pages 115 and 121.]**

Netflix Inc. (n.d.). The netflix prize rules. `http://www.netflixprize.com/rules.html`. Last accessed Apr 07, 2020. **[cited on pages 27 and 82.]**

Niven, J. E. and Laughlin, S. B. (2008). Energy limitation as a selective pressure on the evolution of sensory systems. *Journal of Experimental Biology*, 211(11):1792–1804. **[cited on pages 152 and 166.]**

O'Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1):171–175. **[cited on pages 112 and 123.]**

Oxford University Press (n.d.). Oxford Learner's Dictionaries. `https://www.oxfordlearnersdictionaries.com/definition/english/empathy`. Last accessed on Apr 07, 2020. **[cited on page 185.]**

PE Digital GmbH (n.d.). Das Parship-Prinzip®. `https://www.parship.de/tour/parship-prinzip/`. Last accessed on Apr 07, 2020. **[cited on page 193.]**

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. **[cited on page 173.]**

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, 16(9):1170–1178. **[cited on pages 111, 114, and 115.]**

Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, 1(2):125–132. **[cited on pages 112, 113, 117, 118, 120, 121, 124, and 132.]**

Ramachandran, K. and Tsokos, C. (2009). *Mathematical Statistics With Applications*. Elsevier Academic Press. **[cited on page 53.]**

Rawat, B. and Dwivedi, S. K. (2019). Selecting appropriate metrics for evaluation of recommender systems. *International Journal of Information Technology and Computer Science*, 1:14–23. **[cited on page 26.]**

Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B. (2010). *Recommender Systems Handbook*. Springer-Verlag, Berlin Heidelberg. **[cited on pages 2, 3, 4, 22, 23, 26, and 183.]**

Roxin, A., Brunel, N., Hansel, D., Mongillo, G., and van Vreeswijk, C. (2011). On the distribution of firing rates in networks of cortical neurons. *Journal of Neuroscience*, 31(45):16217–16226. **[cited on page 167.]**

Roxy, P. and Devore, J. L. (2011). *Statistics: The Exploration and Analysis of Data*. Cengage Learning, Boston, 7th edition. **[cited on page 57.]**

Said, A. and Bellogín, A. (2018). Coherence and inconsistencies in rating behavior: estimating the magic barrier of recommender systems. *User Modeling and User-Adapted Interaction*, 28(2):97–125. **[cited on pages 31, 98, 103, and 184.]**

Said, A., Jain, B. J., Narr, S., and Plumbaum, T. (2012). Users and noise: The magic barrier of recommender systems. In *User Modeling, Adaptation, and Personalization*, pages 237–248, Berlin, Heidelberg. Springer. **[cited on pages 30, 34, 37, 77, and 85.]**

Saroj, K. L., Craig, A., Boord, P., Kirkup, L., and Nguyen, H. (2003). Development of an algorithm for an EEG-based driver fatigue countermeasure. *Journal of Safety Research*, 34(3):321–328. **[cited on pages 177 and 187.]**

Savage, N. (2019). Marriage of mind and machine: Bringing together artificial intelligence and neuroscience promises to yield benefits for both fields. *Nature*, 571:15 – 17. **[cited on page 11.]**

Schurz, G. (2015). *Wahrscheinlichkeit*. Grundthemen Philosophie. De Gruyter, Berlin Boston. **[cited on page 36.]**

Siebner, H. R. and Ziemann, U. (2007). *Das TMS-Buch: Handbuch der transkraniellen Magnetstimulation*. Springer, Heidelberg. **[cited on page 192.]**

Sizov, S. (2017a). Comparative assessment of rating prediction techniques under response uncertainty. In *Proceedings of the International Conference on Web Intelligence*, pages 891–898, New York, NY, USA. ACM. **[cited on pages 7, 190, and 196.]**

Sizov, S. (2017b). Mining ordinal data under human response uncertainty. In *Proceedings of the International Conference on Web Intelligence*, WI '17, pages 869–876, New York, NY, USA. ACM. **[cited on pages 4, 5, 6, 15, 18, 19, 96, 97, 98, 103, 164, 185, 186, and 196.]**

Speckmann, E.-J., Hescheler, J., and Köhling, R. (2019). *Physiologie: Das Lehrbuch*. Elsevier, München, 7th edition. **[cited on page 192.]**

Statistisches Bundesamt (2019). Bevölkerung im Alter von 15 Jahren und mehr nach allgemeinen und beruflichen Bildungsabschlüssen nach Jahren. `https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Bildungsstand/Tabellen/bildungsabschluss.html`. Last accessed on Mar 11, 2020. **[cited on pages 48 and 51.]**

Takemura, K. (2014). *Behavioral Decision Theory - Psychological and Mathematical Descriptions of Human Choice Behavior*. Springer, Tokyo Heidelberg. **[cited on page 105.]**

Taylor, J. R. (1997). *Introduction to error analysis, the study of uncertainties in physical measurements*. University Science Books, Mill Valley, CA, USA, 2nd edition. **[cited on page 36.]**

Trommershauser, J., Kording, K., and Landry, M. S. (2011). *Sensory Cue Integration*. Computational Neuroscience Series. Oxford University Press, Beschreibung Oxford. **[cited on page 114.]**

van Beers, R. J., Sittig, A. C., and van der Gon, J. J. (1999). Integration of proprioceptive and visual position-information: An experimentally supported model. *Journal of Neurophysiology*, 81(3):1355–1364. **[cited on page 114.]**

Vanitha, V. and Krishnan, P. (2016). Real time stress detection system based on EEG signals. *Biomedical Research – Computational Life Sciences and Smarter Technological Advancement*, Special Issue: S271-S275:271–275. **[cited on pages 177 and 187.]**

Walck, C. (1996). Hand-book on statistical distributions for experimentalists. University of Stockholm. **[cited on page 74.]**

Walker, E. Y., Cotton, R. J., Ma, W. J., and Tolias, A. S. (2020). A neural basis of probabilistic computation in visual cortex. *Nature Neuroscience*, 23(1):122–129. **[cited on page 128.]**

Weiss, S. M. and Indurkhya, N. (1998). *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann Publishers, San Francisco. **[cited on pages 1, 21, 22, 102, and 184.]**

Yahoo! Research (n.d.). Ratings and classification data. `https://webscope.sandbox.yahoo.com/catalog.php?datatype=r`. Last accessed on Apr 07, 2020. **[cited on page 190.]**

Yang, Z., Zhang, Z.-K., and Zhou, T. (2012). Anchoring bias in online voting. *arXiv:1209.0057*. **[cited on pages 3, 12, 32, and 35.]**

Yu, A. J. and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–692. **[cited on pages 105 and 107.]**

Zemel, R. S., Dayan, P., and Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Computation*, 10(2):403–430. **[cited on page 112.]**

Zhang, H.-R., Min, F., Wu, Y.-X., Fu, Z.-L., and Gao, L. (2018). Magic barrier estimation models for recommended systems under normal distribution. *Applied Intelligence*, 48(12):4678–4693. **[cited on page 28.]**

Zhao, G., Qian, X., and Kang, C. (2017). Service rating prediction by exploring social mobile users' geographic locations. *IEEE Transactions on Big Data*, 3(1):67–78. **[cited on page 180.]**

Zhao, G., Qian, X., Lei, X., and Mei, T. (2016). Service quality evaluation by exploring social users' contextual information. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3382–3391. **[cited on page 180.]**

Ziegler, C.-N., McNee, S. M., Konstan, J. A., and Lausen, G. (2005). Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 22–32, New York, NY, USA. ACM. **[cited on pages 27 and 28.]**