On modeling data from visual psychophysics: A Bayesian graphical model approach

vorgelegt von
Diplom-Informatikerin (Bioinformatik)
Hannah Martina Helen Dold
aus Freiburg im Breisgau

Von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
Dr. rer. nat.

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. Manfred Opper
Berichter: Prof. Dr. Klaus Obermayer
Berichter: Prof. Felix Wichmann, DPhil

Tag der wissenschaflichen Aussprache: 14.01.2013

Berlin 2013

D 83

To my parents Susanne and Wolfgang

and my grandparents Hanne and Helmut

Summary

Computational models are an essential tool in the area of visual psychophysics. Experiments are derived from them, they provide a framework for thinking, and they formalize our understanding of psychological processes. Models in visual psychophysics can be either mechanistic, inspired by nature and imitating psychological processes, or they can be statistical, exclusively describing a function that maps input to output. If a model is defined as a Bayesian graphical model, Bayesian inference allows to estimate the parameter posterior distribution. In addition to the classical point estimate, the posterior distribution provides further diagnostics of the model that influence the conclusions drawn from the model and its parameters. In this thesis I show how we can model visual processing at two very different levels using a Bayesian approach.

A multi-resolution filter model with a static nonlinearity is the standard model for early spatial vision. Its parameters are assumed to be biologically plausible. Conclusions about a model that are based on specific estimated parameter values are only valid if the parameters are constrained by the data at hand. I show that Bayesian statistics led to the discovery of weaknesses in the vision model formulation and missing experimental data. Visual inspection of the posterior suggested starting points for improvements, and allowed me to tailor the early vision model to its founding psychophysical data.

The psychometric function relates stimulus intensity to behavioral performance and the function's parameters are estimated with data from a single experimental condition. I present data that suggests that in some situations a parameter is shared across conditions which would require all conditions to be estimated simultaneously. Instead of building a more complex model, I present how Bayes rule can formulate the inference procedure for psychometric functions to fit many experimental conditions in an effectively simultaneous fashion. This joint procedure allows us to estimate a common parameter across conditions on the basis of established routines.

A common issue in modeling, for example, early spatial vision or a number of psychometric functions is that several experimental conditions have to be combined. This requires access to substantial amounts of data. When I started working on this thesis there was no adequate tool for handling psychophysical data. This, combined with vastly differing storage formats and contents between laboratories and experiments, makes a reanalysis or meta-analysis of earlier experiments difficult. I developed a new software package for Python that stores and manages experimental data including annotations in a

database. It is open-source, operating system independent, and does neither require knowledge of the structured query language, nor of databases.

This thesis demonstrates how close the interaction of modeling, theory, and experiment is, and how important a good experimental basis is for accurate modeling. Well formulated models, elaborate statistical procedures, and extensive data sets—combined they are able to work as engine for further research.

Zusammenfassung

Computermodelle sind ein wichtiges Instrument im Bereich der visuellen Psychophysik. Aus ihnen werden Experimente abgeleitet, sie bieten eine Gedankenstruktur und sie drücken unser Verständnis psychologischer Vorgänge in Formeln aus. Modelle in der visuellen Psychophysik sind entweder mechanistisch, von der Natur inspiriert und psychologische Vorgänge imitierend, oder sie sind statistisch, ausschließlich eine Zuordnung von Eingabe zu Ausgabe beschreibend. Für ein Modell, das als Bayes'sches graphisches Modell definiert ist, kann die A-posteriori-Wahrscheinlichkeit der Parameter mittels Bayes'scher Inferenz geschätzt werden. Zusätzlich zu klassischen Punktschätzern bietet die A-posteriori-Verteilung weitere Modelldiagnostiken, welche von Modell und Parametern gezogene Schlussfolgerungen beeinflussen. Ich zeige in dieser Dissertation wie visuelle Prozesse auf zwei sehr unterschiedlichen Ebenen mit einem Bayes'schen Ansatz modelliert werden können.

Ein Multiskalenmodell mit einer statischen Nichtlinearität ist das Standardmodell für die frühe Mustererkennung. Man nimmt an, dass seine Parameter biologisch plausibel sind. Schlussfolgerungen über das Modell, die auf bestimmten geschätzten Parameterwerten beruhen, sind nur dann gültig, wenn die Parameter von den vorliegenden Daten beschränkt sind. Ich zeige, wie Bayes'sche Statistik zur Entdeckung von Schwachstellen im Sehmodell und von fehlenden experimentellen Daten führte. Die visuelle Überprüfung der A-posteriori-Verteilung deutete Ansätze zur Verbesserung an und erlaubte mir das Sehmodell auf seine zu Grunde liegenden Psychophysikdaten anzupassen.

Die psychometrische Funktion beschreibt die Beziehung zwischen Reizstärken und Verhaltensleistungen wobei die Funktionsparameter mit Daten einer einzigen Experimentalbedingung geschätzt werden. Die von mir aufgezeigten experimentellen Daten legen in manchen Situationen einen gemeinsamen Parameter über Bedingungen hinweg nahe, der eine gleichzeitige Schätzung aller Bedingungen erforderlich macht. An Stelle der Konstruktion eines komplizierteren Modells präsentiere ich, wie das Bayestheorem den Inferenzprozess auf psychometrischen Funktionen so formulieren kann, dass er gleichwertig zu einer gleichzeitigen Anpassung mehrerer Experimentalbedingungen wird. Das gemeinsame Verfahren erlaubt einen über Bedingungen geteilten Parameter auf Grundlage etablierter Routinen.

Eine Gemeinsamkeit bei der Modellierung etwa von früher Mustererkennung oder mehren psychometrischen Funktionen ist, dass einige Experimentalbedingungen zusammengefasst werden. Dies bedarf Zugang zu einer erheblichen Menge an Daten. Als ich diese Arbeit begonnen habe, existierte

kein passendes Hilfsprogramm um Psychophysikdaten zu handhaben. Dies, verbunden mit stark unterschiedlichen Speicherformaten und Inhalten zwischen Arbeitsgruppen und Experimenten, erschwert erneute Auswertungen oder Metastudien früherer Experimente. Ich habe eine neues Programmpaket für Python entwickelt, welches experimentelle Daten und Anmerkungen in einer Datenbank speichert und verwaltet. Es ist frei zugänglich, vom Betriebssystem unabhängig und setzt weder Wissen zur strukturierten Abfragesprache (SQL) noch zu Datenbanken voraus.

Diese Dissertation stellt dar, wie eng das Zusammenspiel von Modellierung, Theorie und Experiment und wie wichtig eine gute experimentelle Grundlage für akkurate Modellierung ist. Gut formulierte Modelle, sorgfältig ausgearbeitete statistische Analysen und umfangreiche Datensätze – zusammen sind sie in der Lage als Motor für weitere Forschung zu dienen.

List of Publications

Manuscripts in preparation

Towards an image-driven early spatial vision model: A Bayesian graphical modeling approach.
Hannah M. H. Dold and Felix A. Wichmann
This manuscript is based on the content of the first chapter *Towards an image-driven early spatial vision model*

Separate Bayesian inference reveals model properties shared between multiple experimental conditions.
Hannah M. H. Dold, Felix A. Wichmann and Ingo Fründ
The second chapter *Joint Bayesian estimation of several psychometric functions* originates from a manuscript to be submitted soon. The study was devised, designed, conducted, analyzed and the original manuscript written to equal parts with Ingo Fründ.

Published conference abstracts

Hannah M. H. Dold, Ingo Fründ and Felix A. Wichmann (2011). *Separate Bayesian inference reveals model properties shared between multiple experimental conditions*.
Front. Comput. Neurosci. Conference Abstract: BC11 : Computational Neuroscience & Neurotechnology Bernstein Conference & Neurex Annual Meeting 2011;
doi: 10.3389/conf.fncom.2011.53.00107

Hannah M. H. Dold, Sven Dähne and Felix A. Wichmann (2010). *Effects of arbitrary structural choices on the parameters of early spatial vision models.*
Journal of Vision 10(7): 1370; doi:10.1167/10.7.1370

Software packages

Hannah M. H. Dold and Rike-Benjamin Schuppner (2012). *XDAPY—Experimental data with Pyhton*
available at http://xdapy.github.com/
The project was originally devised, designed and implemented by Hannah Dold. Rike-Benjamin Schuppner joined the project in December 2010, restructured the source code and added new functionality.

Acknowledgements

I WOULD LIKE TO USE THIS UNIQUE OPPORTUNITY to express my sincere appreciation and gratitude to all those people who have supported me during the writing of this thesis.

FIRST OF ALL I thank my supervisor, Prof. Felix Wichmann, for introducing me to research, for his very worthwhile scientific advices and for his positive ethos.

I HAD THE LUXURY to focus on my research due to the support of the Technische Universität Berlin and the Bernstein Center for Computational Neuroscience Berlin. I also thank the latter for providing outstanding training opportunities and for the devotion of its coordinator Vanessa Casagrande.

THIS MANUSCRIPT was substantially improved by the feedback and comments of Frank Jäkel, Fiona Sammler and Katharina Zeiner. Not only I, but also future readers, will thank you for your time and efforts.

THE MEMBERS OF THE *Modelling of Cognitive Processes* group in the years from 2008 to 2012 excelled in creating a wonderful work environment with inspiring discussions. Thank you all for the lovely company on the many accumulated miles from lab to Mensa and back.

MY SPECIAL THANKS go to Sven Dähne and Kai Görgen with whom I implemented a modeling framework for vision, to Tiziano Zito for general programming advices and to Rike-Benjamin Schuppner for the programming skills he demonstrated during the realization of Xdapy.

TO SHARE AN OFFICE and collaborate with Ingo Fründ was always a great pleasure, and I would not want to do without his friendship.

OF ALL THE FRIENDS who accompanied me during the past years, who have been there for me, and who I would like to thank, I restrict myself to naming only three more because they offered me a home away from home: Alexander Ecker, Anne Kling, and Stefanie Klott.

AND LAST, but in no way least, I acknowledge my beloved partner Sebastian Grau who allowed me to head off on my own to follow my dream and who continued to believe in me during periods of sleepless, dreamless nights. For both, I am more grateful than words can express.

Contents

Introduction

Organisms have to interact with their environment. Sunflowers turn towards the sun for photosynthesis. Bats hunt their prey at absolute darkness relying on echolocation. Many bacteria direct their limited movements towards nutrients. But in order to interact with the environment in the best possible way, these organisms first have to perceive their environment. For humans, sight is the most dominant sense. We use it to navigate, to perceive dangers, to communicate nonverbally, to prepare grasping, and to discriminate eatable from uneatable objects. The fascination with visual processes dates back to the antique times, when scientists, astronomers, and philosophers started investigating the visual system. They were mainly occupied with the optics of the eye, the lens, the glass body and how light passes through them, how the light is focused, reflected and broken. In the last century interests have moved away from the periphery towards the circuits that propagate the visual information from the retina to the cortex and that process the visual information. Thereby vision research diverged into several branches with a focus on motion, color, or spatial structures. My thesis falls into the field of spatial vision, the perception of luminance and spatial structures.

Human pattern and contrast perception, summarized as spatial vision, has been studied extensively since the middle of the 20th century using a wide variety of experimental techniques. Our current understanding is a mosaic of insights from psychophysics, functional MRI, single and multi-cell recordings, and modeling. Even though it is impossible to disentangle the many influences, this thesis will focus on psychophysics and modeling. Other methodological fields are referenced to where they profoundly shaped the thinking of psychophysicists. This allows me to investigate how far psychophysics alone is enough to support

our view on spatial vision. The level of description that is aimed at is more algorithmic than biological: A model where the stimulus is initially convolved with a variety of independent spatial filters which differ in their peak spatial frequency sensitivity and in orientation tuning. The filter outputs are transformed through a nonlinear response function. Subsequently, the outputs are combined and fed into a decision mechanism. Noise could be present at several processing stages.

The motivation for a model based on spatial filters is based on Fourier theory. According to Fourier theory a signal can be represented as a series of orthogonal wave functions. Applied to the visual domain this means that an image can be decomposed into a unique subset of spatial frequencies. A Fourier transformation as basic operation on the visual input is appealing, since it allows for the representation of all stimuli in a common framework.

The first to explicitly investigate human vision from a system analysis point of view was the television engineer Otto Schade in the 1950s. He measured the detection thresholds of systematically manipulated sinusoidal gratings. He introduced an approach that was later adapted by many scientists who investigated spatial vision. Schade could therefore also be called a psychophysicist.[1]

[1] Westheimer, G. (2001). The Fourier theory of vision. *Perception*, *30*, 531–541

A very different methodology led to crucial insights on the neural substrate for spatial structure and pattern processing. Hubel and Wiesel found orientation selective cells in the cat primary visual cortex, V1, when sweeping a bar pattern over the neuron's receptive area.[2] These neurons that react to oriented structures in images promised to be neuronal detectors for oriented structures such as edges or lines. Reverse correlation techniques made it possible to map the receptive fields of these V1 neurons in more spatial detail.[3] The receptive fields are composed of alternating, elongated regions with excitatory and inhibitory properties. Due to their striking resemblance to Gabor functions, Gabors became the functional description of the neurons.[4] A Gabor is the product of a sine wave and a Gaussian window or aperture.

[2] Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*, 106–154

[3] Ringach, D., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, *28*, 147–166

[4] Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, *2*(7), 1160–1169

Campbell and Robson investigated how sine-waves—such as the stimuli used by Schade— and square-waves—similar to the bar pattern used by Hubel and Wiesel—relate to each other. They conducted an experiment with stimuli being sinusoids and

combinations of sinusoids that suggests that the perception of patterns is predicted by the perception of the individual frequency components.[5] More specifically they showed that a sine-wave and square-wave of the same fundamental frequency can be discriminated from each other only if the contrast levels of the higher harmonics are above their individual thresholds. This finding along with subsequent similar approaches nourished the expectation to find a cortical code that resembles Fourier transformation in its nature. By definition each image can be decomposed into a unique subset of frequency components and the detectability of each component can be measured. If the detectability of the full image may be computed by generalization from the component responses, then the internal processes at early stages of the visual system could be an analog to the Fourier decomposition.

[5] Campbell, F. W., & Robson, J. G. (1968). Applicaton of Fourier analysis to the visibility of gratings. *Journal of Physiology*, *197*, 551–566

Subsequent studies investigating vision as a Fourier analyzer often found parallels to Fourier analysis and results that could be explained by a Fourier-like processing followed by a mechanism that operates on the decomposition. In her book "Visual Pattern Analysers"[6], Norma Graham summarized the research in the field conducted prior to 1989. Her working hypothesis is that the input image is decomposed in a Fourier-like process into separate processing streams, so-called channels, that are specific for spatial frequency, orientation, or phase. The book names many basic properties—dimensions—of the channels. Phase, position, orientation, spatial frequency, spatial extent are defined in the spatial domain and have equivalents in the temporal domain. Then there is mean luminance, contrast, and even binocularity.[7] Graham collects the relevant studies of the time and processes them strategically from a modeler's perspective. The assumptions underlying a model are spelled out explicitly whenever a model is formulated. One important assumption is that observers base their response only on the most active channel or the channel in which they know that the signal appears. Graham presents experiments that fall into five categories based on the experimental paradigm employed. With respect to this work three categories are relevant and two are less so. The later two embrace experiments that exploit uncertainty of the observer about the target signal[8] and experiment that require that the signal is not only detected but identified from others[9]. The studies that are closer related to this work use adaptation paradigms, summation paradigms, or they explore interactions between filter properties. An observer who is exposed to a specific

[6] Graham, N. V. S. (1989). *Visual Pattern Analyzers*. Issue 16 of Oxford Psychology Series, Oxford University Press

[7] Without doubt temporal aspects and binocularity are important. But they add much more complexity over the purely spatial luminance properties, so that their effects are not further considered in the present work.

[8] Davis, E. T., & Graham, N. (1981). Spatial frequency uncertainty effects in the detection of sinusoidal gratings. *Vision research*, *21*, 705–712; and Davis, E. T., Kramer, P., & Graham, N. (1983). Uncertainty about spatial frequency, spatial position, or contrast of visual patterns. *Pereption & Psychophysics*, *33*(1), 20–28

[9] Thomas, J. P. (1985). Detection and identification: how are they related? *Journal of the Optical Society of America A*, *2*(9), 1457–1467

stimulus for an extended duration becomes less sensitive to the stimulus through adaptation. If probed with another stimulus, the extend to which the observer is sensitive to the test stimulus allows conclusions about the similarity of the stimuli with respect to the channel properties. A sensitivity of a stimulus that is very different from the adapted stimulus should not be effected by the adaptation procedure. The sensitivity should be affected, if the probe stimulus is processed by the same mechanisms as the adapted stimulus.[10] The general idea of summation experiments is similar to adaptation only that here additional channels are activated by presenting an extra stimulus.[11] If the properties of the channels would be independent of each other, each dimension could be analyzed individually and the analysis would be straightforward. Of course, nature did not make the case as simple and therefore all experiments that target the properties of the channel become more difficult to interpret. Over all, Graham points out many properties that a model for early vision should ultimately posses. However, each single topic results in a particular model apt to provide an explanation for that topic only. She does not yet condense the results to a single general model that is valid across experimental paradigms.

Modeling has steadily gained in importance since the rise of personal computing devices in the 1980s. Computational models evolved conjointly with the experiments seeking to predict their results. Many publications combined experiments and models in the form of energy detection models[12], transducer models,[13] and models based on contrast-gain control[14].

[10] Blakemore, C., & Campbell, F. W. (1969). On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *Journal of Physiology*, *203*, 237–260; and Georgeson, M. A., & Harris, M. G. (1984). Spatial selectivity of contrast adaptation: models and data. *Vision Research*, *24*(7), 729–741

[11] Graham, N. V., & Nachmias, J. (1971). Detection of grating patterns containing two spatial frequencies: a comparison of single-channel and multiple-channels models. *Vision Research*, *11*(3), 251–259

[12] Burton, G. J. (1981). Contrast discrimination by the human visual system. *Biological Cybernetics*, *40*(1), 27–38

[13] Legge, G. E., & Foley, J. M. (1980). Contrast masking in human vision. *Journal of the Optical Society of America A*, *70*(12), 1458–1471

[14] Foley, J. M. (1994). Human luminance pattern-vision mechanisms: masking experiments require a new model. *Journal of the Optical Society of America A*, *11*(6), 1710–1719; and Watson, A. B., & Solomon, J. A. (1997). Model of Visual Contrast Gain Control and Pattern Masking. *Journal of the Optical Society of America A*, *14*(9), 2379–2391

In order to find a global description for spatial vision—neglecting temporal aspect and binocularity—several investigators from different institutes and research organizations decided to build a joint data set.[15] The modelfest project is unique in its agreement to combine the forces of several labs in the field of vision research to achieve a common goal: create a database to test vision models. The 43 stimuli range from Gabors with different spatial frequencies and aspect ratios to a checkerboard and a city skyline. An adaptive procedure was used to determine the detection threshold for each pattern. The final data set comprises detection thresholds for 14 observers with 4 repetitions for each of the 43 stimuli. The original project was experimental, but a few years later, in 2005, Watson and Ahumada reanalyzed the modelfest data.[16] Their model used many processing steps including channels, aperture, contrast sensitivity function, an oblique filter, and Minkowski pooling. The authors start with a subset of the processing steps and increase the complexity by adding components. They also test several functional descriptions for the contrast sensitivity function. They find that the application of a contrast sensitivity function in the Fourier domain renders a further usage of channels unnecessary. They start their analysis with a simple model that contains only few processing steps and then add further mechanisms to improve the model's prediction. The model's complexity is not further extended as soon as the fit does not improve any more. A detailed statistical inspection of the resulting model is missing, however.

[15] Carney, T., Klein, S. A., Tyler, C. W., Silverstein, A. D., Beutter, B., Levi, D., Watson, A. B., Reeves, A. J., Norcia, A. M., Chen, C.-C., Makous, W., & Eckstein, M. P. (1999). Development of an image/threshold database for designing and testing human vision models. *Proceedings Vol. 3644 Human Vision and Electronic Imaging IV, Bernice E. Rogowitz; Thrasyvoulos N. Pappas, Editors,*, (pp. 542–551)

[16] Watson, A. B., & Ahumada, A. J. (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision*, *5*(9), 717–740

In 1999 Felix Wichmann submitted his doctoral thesis on *Some Aspects of Modelling Human Spatial Vision: Contrast Discrimination*.[17] He conducted experiments that investigated the human ability to discriminate sinusoidal gratings for a wide range of contrasts and for three presentation times. He suggests that an experimental basis that spans different pedestal contrasts would be necessary to constrain the nonlinearity (called transducer function in his original work) that follows spatial filtering in the classical definition of early vision models. This inspired the first main chapter of this thesis which investigates the interaction between experimental data and modeling and shows that his intuition was correct.

[17] Wichmann, F. A. (1999). *Some aspects of modelling human spatial vision: Contrast discrimination*. Unpublished doctoral dissertation, The University of Oxford, Oxford, UK

For a long time, the majority of models predicted data for a single type of experiment, limiting prediction power. Itti, Koch, and Braun take the modeling one step further by asking the question, if a single model can account for different aspects of spatial vision.[18]

[18] Itti, L., Koch, C., & Braun, J. (2000). Revisiting spatial vision: toward a unifying model. *Journal of the Optical Society of America A*, *17*(11), 1899–1917

They use results from five psychophysical tasks: (1) contrast discrimination, (2) spatial frequency and (3) orientation discrimination as a function of contrast, (4) threshold elevation as function of mask orientation, or (5) mask spatial frequency. All results are obtained in a two-alternative forced choice procedure. For the tuning functions in experiment (4) and (5) the mask was presented only in the distractor interval and not in the interval containing the signal. The model is a version of the linear nonlinear model class with linear Gabor filter, a transducer function, contrast gain control, and a decision stage based on Fisher's linear discriminant. Since they did not vary the target spatial frequency, a mechanism for contrast sensitivity is not necessary. Most recently, Goris et al.[19] put forward a promising new approach incorporating insights from population coding and neurophysiology. In a follow up study they showed that this model can predict several spatial vision phenomena from summation, adaptation, and uncertainty.[20] They claim that normalization and maximum-likelihood decoding are crucial for the model. Given the many parameters and variables in the model it is not clear however, where the power of the model comes from and which mechanisms are truly necessary to predict the experiments.

[19] Goris, R. L. T., Wichmann, F. A., & Henning, G. B. (2009). A neurophysiologically plausible population-code model for human contrast discriminations. *Journal Of Vision*, *9*(7), 1–22

[20] Goris, R. L. T., Putzeys, T., Wagemans, J., & Wichmann, F. (2011). A neural population model for pattern detection. *Journal of Vision*, *11*(11), 1165

The direction of modeling in the vision sciences goes to combining several experimental conditions or tasks. In the modeling studies cited above each stage of the model is originally inspired by experimental findings from different areas. All groups take great care to describe the model components and to fit the model to the data. However, the investigating groups use a performance measure that is solely based on the error minimization between model output and data. The so-called residual error only tells about the model ability to replicate the experimental measurements, however. In machine learning this residual error would be called the training error since it is evaluated on the same data as the parameters were estimated. Overfitting is then likely to happen and generalization ability to unseen data or the number of model parameters are not taken into account. Only Itti et al.[14] at least mention the variances of their model's parameter estimates. But neither of the groups indicate which kind of experiment is crucial to constrain which parameter or if all parameters are well constrained. However, information about the model's adequateness given the data is the type of information that should have been gathered to strengthen the model formalism—the model structure.

SUCH A MODEL OF EARLY VISION IS A MECHANISTIC SUMMARY OF THE DATA, since the complexity of all the data sets are combined in a single formalism. But more than that, behind the model is the hope of a general representation of visual information at an early stage that is more or less hard-wired and task independent. The model has the pretense to be inspired by nature and to reflect the processing that eventually leads to perception. In my thesis I will first apply Bayesian estimation methods in the area of vision models instead of the previous frequentist approaches. The methodological advances in Bayesian statistics allow me to investigate the model's internal processing and its accordance with experimental findings. In addition I will use full psychometric functions and not only thresholds at a fixed behavioral performance level. In Bayesian inference error assessment is based on likelihoods instead of residual error. This fits well with the extension to psychometric functions because likelihoods take the variances at different performance levels into account.

IN THE FIRST MAIN CHAPTER, *Towards an image-driven early spatial vision model*, the model as a whole will be discussed from a modeling perspective rather than a mechanistic or theoretical perspective. First, a short introduction to *Bayesian model analysis* motivates the need of the parameter posterior distribution. I chose the Metropolis-Hastings sampling algorithm to approximate the true posterior distribution and will therefore explain the algorithm and its challenges. The sampled posterior distribution can be analyzed by rather plain and simple visualizations. I will show two distributions to exemplify the method: an inconspicuous distribution that can directly be interpreted and a second distribution that exhibits problematic properties. The second section concerns *The implementation of an early spatial vision model*. Each of the processing stages will be discussed separately: The purpose of the model stage is explained, further relevant literature is reviewed, and details of the implementation are discussed. Here, I will also explain how the model can be specified as a graphical model which facilitates the subsequent application of the Bayesian estimation procedure. The following section, *The identifiability of vision model parameters*, contains the main findings. But as importantly, the section demonstrates that a Bayesian model analysis is more than a parameter estimation procedure. It is a tool to understand the model and its connection to the experiment. It can provide hints on structural changes that are necessary to the model or on relevant experimental data that

should be gathered. The chapter's results are finally related to the current literature and further extensions and prospects disclosed in the *Discussion*.

A model that summarizes data must of course not operate in a psychological mechanistic setting. It could reflect any mapping from experimental manipulation to the observed result. In the area of psychophysics, the prevailing summary of a single data set is the psychometric function. It describes the change in performance with varying stimulus intensity. The model describing such data has two critical parameters: the stimulus intensity that elicits a predefined performance, the threshold, and the rate of performance change with increases in stimulus intensity, the slope. Most psychophysical studies contain several conditions. For example a study might target the differences in threshold between conditions assuming constant slope, e.g. in order to determine a contrast sensitivity function dependent on stimulus spatial frequency. However, the standard psychometric function can only account for one condition at the time. Combining several experimental conditions in a joint analysis that takes the equality between parameters into account can result in a complex model with a single slope variable and at least a threshold variable per condition. There are standard routines to fit a single psychometric function, while no such routines are available to fit the model with joint slope.

In the second main chapter, *Joint Bayesian estimation of several psychometric functions*, a routine is presented that uses a sequence of single psychometric functions to fit several conditions quasi-simultaneously. A posteriori independence of the parameters is the theoretical assumption that must be made for the procedure to truly equal the full model. The procedure is explained in the section titled *Separate sampling for joint inference* for an arbitrary model. Then, the method is again illustrated on an *Example from perceptual psychology: the psychometric function* and also the *Evaluation of the method* is tailored to the specific case of psychometric functions. *Another example and statistical tests* indicate the limits of the procedure and provide guidance for when the joint procedure is legitimate to use and when it should be avoided. A short *Discussion* summarizes the chapter.

The goal to embrace a significant amount of data in a single analysis is present in the first chapters. The data is not large in

terms of paper it can be printed on or in terms of bytes. The data is however rather complex in the way it is structured. It was collected in different experiments, under varying conditions or for several observers. This is not a special case in science anymore where the tendency towards meta studies and unifying studies increases. This type of studies requires discipline in the treatment of data, since the scientist who performs the meta study might not be the scientist conducting the study in the first place, or there might be several months or even years between the experiments and the analysis.

THE THIRD MAIN CHAPTER is concerned with *The storage of experimental data*. It documents a software project that accompanied the modeling and data analysis. The software package is called *Xdapy* which is an abbreviation for **ex**perimental **da**ta with **py**thon. As the name suggests, I designed it to use python as the only programming language which users need to know. The chapter initially explains how a *Mapping the world onto objects* is achieved. Experiments are structured into logical units—objects—that contain data as well as annotations. Then by *Mapping objects onto a database* all information can be stored in an uniform way. In this section I will derive and explain the necessary database structures. The software interface to the database finally allows to *Managing experimental data with python: Xdapy*. The sections explains the setup of the system and its work routines with an example, user documentation, and example code. A brief *Discussion* completes the chapter.

IN THE *Conclusion* I will discuss the similarities and differences between the model kinds employed in the first and second main chapter and comment on the importance of extensive data sets, solid models, and an elaborate statistical estimation method.

Towards an image-driven early spatial vision model

In the general introduction I reviewed literature relating to spatial vision. The predominant model for spatial vision contains a linear filter stage followed by a point-wise nonlinearity and a decision mechanism. In this chapter I will estimate the parameters of a light-weight spatial vision model with Bayesian statistics instead of using the frequentist statistics as has been done previously. Thereby, I will confirm that the Bayesian approach is more powerful than pure parameter estimation and present implications for the modeling of vision experiments. The chapter begins with a short introduction to Bayesian inference and its accompanied model analysis. Subsequently, I comment in detail on the model specifications. Then, the Bayesian procedure is applied to the model and finally the results are discussed.

Bayesian model analysis

Bayesian estimation is the field of Bayesian statistics that uses the probability of parameters given the observed data to obtain a parameter estimate. This posterior distribution, $P(\theta|D)$, is determined using Bayes' rule

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{\mathcal{P}(D)} \propto P(D|\theta)P(\theta).$$

Since the probability of the data is constant once it is observed, only the probability of the data given the parameters—the likelihood—and the prior probability of the parameter need to be computed. The likelihood can be calculated from the model. The prior probability can be based on real prior knowledge or based on theoretical considerations. Bayesian estimation is related to maximum likelihood (ML) estimation[21] and maximum a posteriori (MAP) estima-

[21] $\hat{\theta} = \arg\max_\theta P(D|\theta)$

tion[22]. Both alternatives however are simple point estimates and are not based on the full posterior distribution.

[22] $\hat{\theta} = \arg\max_\theta \frac{P(D|\theta)P(\theta)}{P(D)}$

The Bayes estimate is obtained by minimizing the posterior expected loss given a specific loss function.[23] The quadratic loss function is a common function which results in the minimization of the mean squared error. In this case the Bayes estimate is the mean of the posterior distribution. For a 0/1 loss function the Bayes estimate is the mode of the posterior distribution and equals the MAP estimate. With a linear loss function the estimate results in the distribution's median. In the following section mean estimates are reported[24] in line with general practice in Bayesian statistics. If the calculation of the minimal posterior expected loss is difficult to obtain analytically, a sampling algorithm can be used to approximate the posterior distribution from which the Bayes estimate is inferred. So-called Markov Chain Monte Carlo (MCMC) algorithms are a class of such sampling algorithms. They generate a MCMC chain, a sequence of random samples in which a further sample depends only on the current sample and not on previous samples. These samples characterize the posterior distribution. Here, the Metropolis-Hastings algorithm[25] is used to generate the MCMC chain. In each step of the algorithm a new sample is proposed and used or discarded according to an acceptance rule.

[23] Hoff, P. D. (2011). *A first Course in Bayesian Statistical Methods*. Heidelberg: Springer

[24] $\hat{\theta} = E[P(\theta|D)]$

[25] Kroese, D. P., Taimre, T., & Botev, Z. I. (2011). *Handbook of Monte Carlo Methods*. Hoboken: John Wiley & Sons, Inc

The acceptance rule computes the fraction $a = \frac{P(x')}{P(x_i)}$ between the probability of the proposed sample, x', and the current sample, x_i. The proposed sample is always accepted as x_{i+1} if $a > 1$ and otherwise it is accepted with probability a. The chain converges towards the posterior distribution by always accepting a new sample for which the posterior probability increased with respect to the current sample. Then, it assures that the whole posterior is sampled by also accepting samples that do not strictly move towards the center of the distribution. In that case, a slightly less probable sample is more likely to be accepted than a sample with a larger decrease in posterior probability. Finally, the acceptance rule also assures that the chain does not remain in a local optimum by rarely accepting a new sample with even a significant decrease in its posterior probability.

The proposal distribution generates new samples. It has to be specified by the modeler and its parameters can be regarded as hyperparameters. In this thesis a multivariate Gaussian proposal

distribution is chosen. In each sampling step the distribution's mean vector is set to the last sample that was accepted. Thereby, the proposal distribution moves in parameter space along with the sampling. The covariance matrix of the distribution is critical since it specifies the step sizes along the dimensions. Here, the hyperparameters are tuned iteratively. In each iteration a MCMC chain is drawn and with that chain the hyperparameters are adjusted, such that the marginal proposal distribution along a dimension spans a range that exceeded the range of the marginal posterior distribution[26]. Thus, the optimal proposal distribution reflects properties of the posterior distributions. Unfortunately this dependance is a chicken and egg problem, because the posterior distribution is to be estimated and the estimation depends on the proposal distribution. The iterative approach tries to bypass this problem. If the parameters of the model are independent of each other, the variances are the only hyperparameters of the proposal distribution. However, if the model parameters are correlated, an isotropic proposal distribution is not sufficient because the acceptance rate of samples will be very low. Thus, in addition to the variances, the covariances between samples are computed in each iteration of the hyperparameter tuning and are used to parametrize the proposal distribution for the following sampling iteration. The hyperparameters are finally accepted if they do not change anymore with further iterations or if the acceptance rate within the iteration drops clearly. A good proposal distribution leads to a chain that is well mixed meaning that the sample of the chain are identically and independently distributed. In practice this is mostly not achieved directly and chains must be postprocessed before being used for parameter estimation.

[26] Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110–120

Chains are postprocessed to remove dependencies between subsequent samples and initial parts in which the chain is not balanced around the optimum. First, a thinning process removes dependencies by using not all the samples but only every n^{th} sample—a *thinned* chain. Second, instationary parts which are common at the beginning of the sampling when the chain moves from the tails of the parameter posterior distribution towards the center—the burn-in process–are ignored. As a sanity check several chains should be drawn and the user has to ascertain that the different chains converge to the same distribution. The number of samples that need to be omitted through thinning and removing

the burn-in depends on the chain. In this work, I will always show three postprocessed chains.

The advantage of a Bayesian estimator over other estimation procedures, e.g. gradient descent, is that it provides samples from the posterior distribution $P(\theta \mid D)$ which is the probability density of the parameter vector θ given the data D. That means that through the data the probable parameter space is restricted to particular regions. However, the probability density provides more valuable information than only a point estimate as obtained with classic ML. First of all, the posterior confidence interval for a particular parameter can be obtained. Second, the shape of the marginal distribution can be visualized and third, correlations between parameters can be computed.[27] Why are these criteria of interest?

[27] Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis*. Boca Raton: Chapman & Hall, 2 ed; and Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions*. New York: Springer, 3 ed

All three criteria increase the modelers confidence in the parameter estimates of the model. First, credible intervals denote the steadiness of the estimate and say that the true parameter falls with a given certainty in the denoted interval. Second, the shape of the posterior distribution is of interest since it allows to draw conclusion about the error surface which might contain more than one optimum with similar likelihoods if the posterior distribution is not unimodal. If the marginal distribution of one parameter has more than one mode, those modes can be accompanied by trends along dimensions of other parameters. Third, pairwise correlations between parameters allow to unravel trends between the respective parameters. I am going go visualize correlations by scatter plots of a first parameter against another. The interactions between parameters are important because they can point to possible problems in the model formulation and might lead to a simpler formulation of the model. The law of parsimony, also known as Occam's razor, would favor such a simpler explanation as long as the explanatory power is sufficient. A further possibility to identify correlations is by inspecting the principal components of the distribution. More precisely, if an eigenvalue is larger than expected for uncorrelated data, the pattern of the corresponding eigenvector displays relevant correlational structure in addition to the pairwise correlation plots.

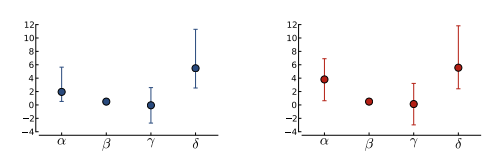
I will contrast the criteria for two artificial posterior distributions to exemplify the ideas. The reference distribution consists of a four-dimensional density function with marginals that correspond

to a) a gamma distribution, b) a beta distribution, c) a t-distribution and d) another gamma distribution. This distribution exemplifies an uncritical posterior distribution and the visualizations of its diagnostics are subsequently shown in blue colors. In contrast, a second four-dimensional probability density is generated with potentially unwanted characteristics such as a bimodal marginal density and correlations between dimensions (Table 1). The diagnostics of this distribution are subsequently presented in red colors.

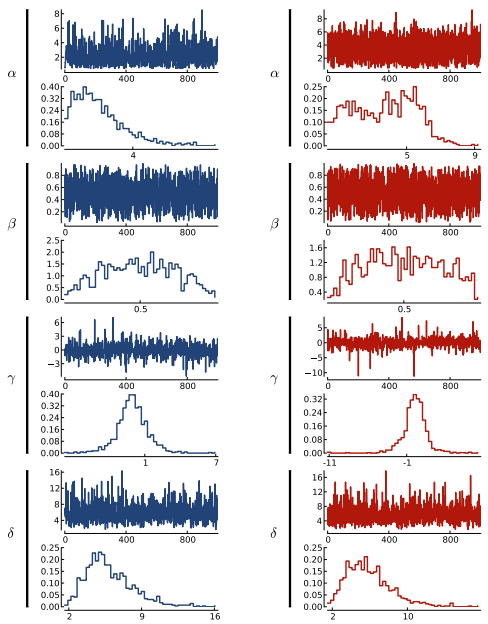
	α	β	γ	δ
α	1.0	0.4	0.2	0.0
β	0.4	1.0	-0.8	-0.8
γ	0.2	-0.8	1.0	0.0
δ	0.0	-0.8	0.0	1.0

Table 1: Correlations of the second artificial posterior distribution

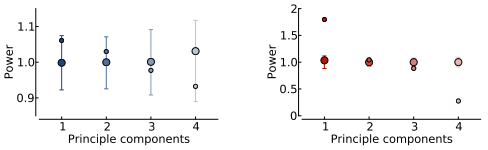
Figure 1 contrasts the two artificial posterior distributions with respect to the suggested criteria. The confidence intervals are shown in Figure 1a. The means and confidence intervals of the distributions do not differ strongly. Therefore, both posterior distributions are inconspicuous just by inspecting the confidence intervals and without further analysis the same conclusions could be drawn from both distributions which would be wrong as shows the next figure part. Figure 1b shows the MCMC sampling traces and the marginal posteriors. In this artificial example, the samples are independent random variates of the four-dimensional density function and not real MCMC samples. Each trace presents the samples along one dimension. As expected, histograms of the marginalized samples follow the specified marginal distribution. The two density functions are clearly different along the first dimension which is labeled α. Here, the second distribution, plotted in red, shows two peaks. In a real modeling case, such a bimodal distribution is probably unwanted since it shows that the corresponding parameter is bistable and that a simple interpretation of the parameter might not be possible. Both diagnostics discussed so far depend on the analysis of each dimension of the parameter space in isolation. Interactions between parameters can be looked at through the pairwise correlation between parameters. Figure 1e shows each parameter plotted against the other parameters. For the independent distribution, all pairwise correlations are around zero and therefore the samples are aligned along the main axes. Linear regressions show no trends and the coefficients of determination R^2 are zero, which indicates that the variability in one parameter is not linearly accounted for by the other parameter. The correlation plots of the second distribution show linear trends and R^2 values that are higher than expected for uncorrelated samples especially for the pair β -γ and β - δ. To quantify whether the value of R^2 is higher than expected by sampling variance alone, a bootstrap procedure is applied. 1000 independent draws from both marginal



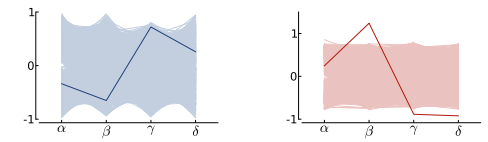
(a) Confidence intervals



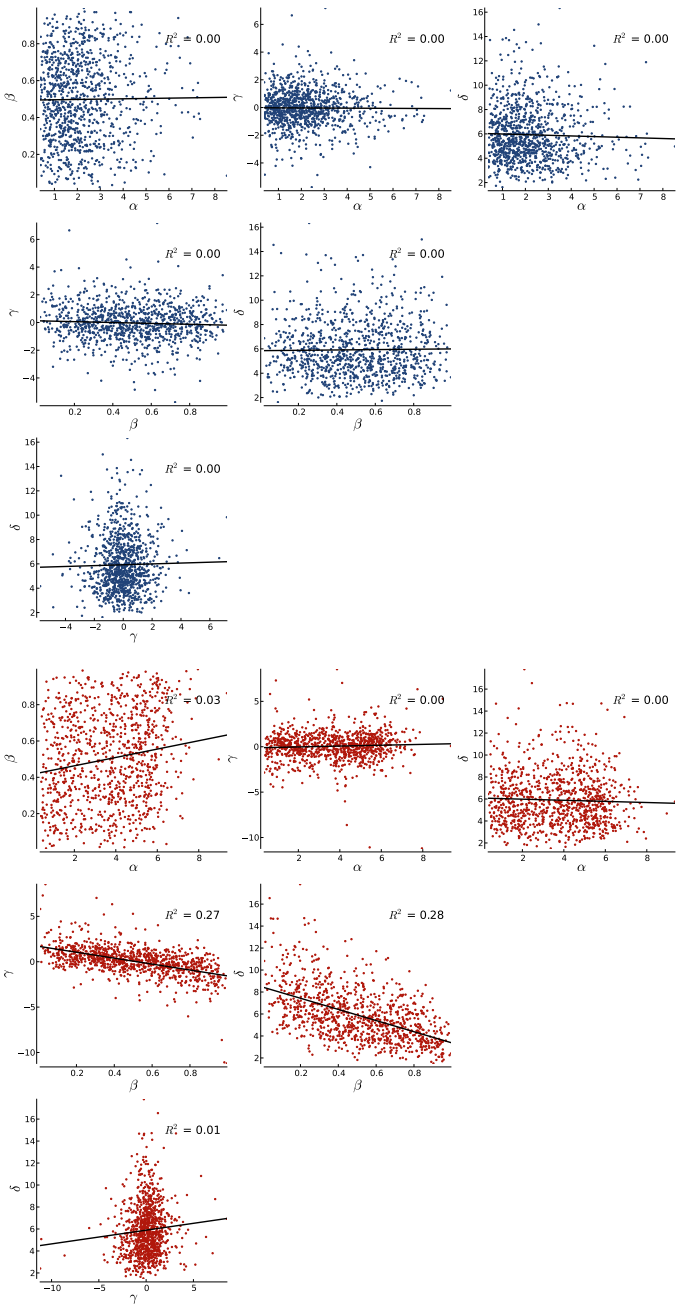
(b) Sampling traces and marginal posteriors



(c) Principal component power



(d) First principal component



(e) Pairwise correlations

Figure 1: Diagnostic of the two example posterior densities

histograms are taken and the R^2 computed. Table 2 and Table 3 list the observed R^2 and the bootstraped range of R^2 for both example posterior distributions.

	β	γ	δ
α	0.00 $\in$(0.00,0.01)	0.00 $\in$ (0.00,0.01)	0.00 $\in$ (0.00,0.01)
β		0.00 $\in$ (0.00,0.01)	0.00 $\in$ (0.00,0.01)
γ			0.00 $\in$ (0.00,0.01)

Table 2: R^2 values of the first example distribution and expected ranges for a distribution without correlations

	β	γ	δ
α	0.03 $\notin$ (0.00,0.01)	0.00 $\in$ (0.00,0.01)	0.00 $\in$ (0.00,0.01)
β		0.27 $\notin$ (0.00,0.01)	0.28 $\notin$ (0.00,0.01)
γ			0.01 $\in$ (0.00,0.01)

Table 3: R^2 values of the second example distribution and expected ranges for a distribution without correlations

The pairwise correlation plots as in Figure 1e are a rather intuitive way to visualize dependencies. However, especially for high-dimensional posterior distributions, the number of pairwise analyses is large. For n parameters there are $n * (n-1)/2$ pairwise combinations. Here, I suggest an alternative analysis which originated in a discussion with Matthias Bethge[28]. It provides a faster impression on the correlation structure of the posterior distribution. This analysis is based on principal component analysis, the eigenvalue decomposition of the covariance matrix. If the eigenvalue of the first principal component exceeds the expected value for uncorrelated data (again computed through bootstrapping), this principal component contains the main structure in the posterior. Figure 1c shows the expected value for uncorrelated bootstrapped data with the same marginals, the 95% confidence intervals for this value, and the observed value. If the observed value stems from an uncorrelated distribution it falls within the confidence region. However, the first principal component exceeds the expected range in the case of correlated dimensions. When plotting the first principal component in Figure 1d this component remains within the range spanned by other bootstrapped principal components in one case and surmounts the range in the other case. The illustration in red for the posterior distribution with correlated dimensions shows that effect is strongest for β, γ, and δ. In addition, it shows that the effects of β and γ, and β and δ are of opposite directions. The correlation between them could therefore be negative. Likewise the effects of γ and δ are of the same sign and a correlation between them could therefore be positive. Since this analysis is based on the full covariance matrix, it can not show on which pairwise correlations the effects are actually based. But given the information

[28] Centre for Integrative Neuroscience, Otfried-Müller-Str. 25, 72076 Tübingen, Deutschland

from the principal component analysis it is clear which pairwise correlations to inspect. This could prove especially valuable for a large number of parameters.

The visual inspections of the posterior distribution are designed to investigate the properties of the posterior distribution and not the quality of the sampling procedure. Since the samples give rise to the posterior distribution, samples and posterior can not be disentangled and only posteriors that are well sampled are informative. That means that the quality of the MCMC chain is important when analyzing the posterior distribution. Several statistics have been developed to judge the quality of sampling procedures. The independence of the chain can be assessed by its auto correlation. For convergence two statistics are commonly used: Geweke's convergence diagnostic[29] and Gelman and Rubin's convergence diagnostic[30]. The former tests the equality of the means from the first 10% of the chain and the last 50% of the chain. The later assumes a normal distributed posterior distribution. It requires several chains which started with a sample from an overdispersed distribution and compares within chain variance to between chain variance. Those indexes are useful for automated analysis and to provide a general rating of the chain's quality. However, most modelers do not have a good intuition for those scalars that summarize a complex statistic. Similar to the discussion about hypothesis testing with p-values[31] and whether a p = 0.049 should lead to a different conclusion than p = 0.051 it is not trivial to decide which value of the above statistic is good enough to accept a chain. If the raw data is provided in form of multiple sampling traces, the reader can apply both convergence statistics graphically. She can check if the several chains sample in the same region and if the sampled region is consistent at the beginning and end of a chain. Then based on the chain appearance she can decide herself if the quality meets her requirements. The following section will therefore always provide three thinned and steady state sampling traces in all the figures where applicable.

[29] Geweke, J. (1992). Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In J. Bernardo, J. Berger, A. Dawid, & A. Smith (Eds.) *Bayesian Statistics 4*. Oxford: Oxford University Press

[30] Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472

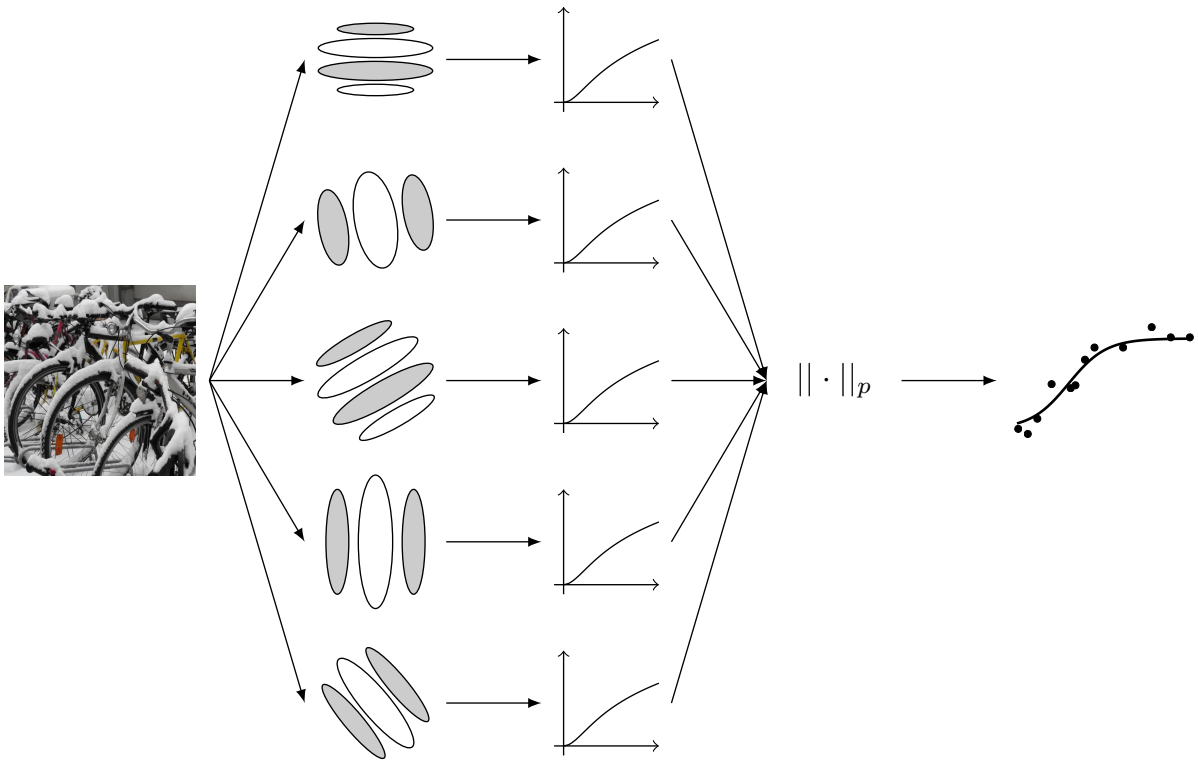
[31] Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist*, *49*(12), 997–1003; and Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804

The diagnostics pictured in Figure 1 do not primarily assess the chain quality, but help the modeler learn about the model and its parametrization. It allows true insights about model and data as illustrated next.

The implementation of an early spatial vision model

IN THIS SECTION I will address questions, considerations, and concerns regarding the implementation of different model stages. I will explain the solutions chosen for the present work including critical implementation details. The following paragraphs are ordered according to their appearance in the model (Figure 2), starting with the image as input and the necessary preprocessing to standardize experimental conditions and images. The second paragraph describes the filters used to split the input into the different channels of a multiresolution model. Then I will describe the type of nonlinearity that is applied on the filter responses. Subsequently, the decision rule is explained that recombines the responses from all channels to a single scalar decision variable. Finally, I will present how this decision variable is linked to the behavioral responses measured experimentally.

Figure 2: The stages of the model



Grayscale images as model input

A GENERAL APPLICABLE MODEL of the early visual system should be image driven. Thus, in addition to fitting existing data, the model can serve as hypothesis generator for further experimentation on more than abstract mathematically defined stimuli, but stimuli including complex stimuli such as natural images or any other stimulus.[32] An image-driven model has the further advantage that it can combine experiments with differences in their experimental viewing condition which could stem from different stimulus resolutions or different effective stimulus sizes.

[32] The endeavor of identifying the model, however, will be based on abstract stimuli for reasons of experimental manipulation and control, and practically also because of the availability of data.

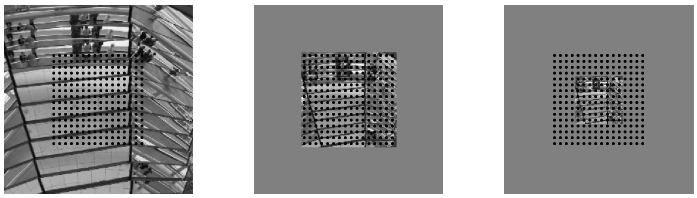
TO RENDER THE MODEL independent of these variables, two possible approaches could be followed. The first approach limits the additional information that comes with every pixel by introducing correlations and thereby saturating performance. The second limits the overall amount of information by sampling only in a small portion of the visual field and is implemented as a kind of artificial retina. Correlations add more variability to the model and saturate the model only in the limits with many samples. Before saturation is reached, the sampling still matters. Therefore, an artificial retina without correlations is adopted right away to keep the model as clear and simple as possible.

THE HUMAN FOVEA spans approximately 5.2 degrees of visual angle and the peak cones density is around 1.6×10^5 per mm^2.[33] The artificial retina used at the first stage of the model is inspired by nature and extends with 2.5 degrees slightly over the rod-free foveal region of 1.7 degrees. Thereby, it covers the central high density region but does not extend into the periphery with low cone density. Given the large correlations between adjacent receptor activity and the low resolution of the stimuli used in experiments, a regular sampling grid of 256×256 samples shall be sufficient especially given its advantage in computational speed. The model output will depend on the settings of the artificial retina. As long as the sampling is high enough to cover the structures of the stimulus, the general findings of the Bayesian model analysis are not affected. However, the concrete values of the model's parameter estimate depend on the setting and not too much relevance should be attributed to values of this specific setting.

[33] Wandell, B. A. (1995). *Foundations of Vision*. Sinauer Associates, Inc., flyleaf

(a) Image presented with different resolutions

(b) Samples from (a)



(c) Image presented at different viewing distances

(d) Samples from (c)

Figure 3: Dependance of pooling on image size and image resolution

FIGURE 3 ILLUSTRATES how the same image, presented in different resolutions or at different retinal extends, results in different inputs to the model. Figure 3a shows the sampling of a stimulus that expands over four degrees of visual angle. This corresponds to a stimulus presented at a distance of one meter and a size of seven centimeters in horizontal and vertical dimension. The resolution of the three images in Figure 3a is divided in halves from left to right. The grid of dots extends over the image region of 2.5 degree visual

angle in which the just proposed artificial retina would sample the example image. The resolution is reduced to a 16×16 grid for visibility. The gray values at the the position of the grid points would be taken to form the image that is actually used as model input. The samples drawn from the differently resolved images are presented below each image in the second row (Figure 3b). If the same image is seen with increasing viewing distances, the retinal area that is spanned through the image reduces. In Figure 3c the images extend over four, two, and one degree visual angle and again the samples are presented in the following row (Figure 3d). The differences between the sampled images are less dramatic for different resolutions than for the different retinal extends which are exaggerated here. To summarize, the extend and resolution of the presented stimulus must be taken into account for a generalized model input. The size and resolution of the image matter especially whenever non-pointwise operations are performed.

[34] Graham, N. V. S. (1989). *Visual Pattern Analyzers*. Issue 16 of Oxford Psychology Series, Oxford University Press

Not only is visual acuity highest in the fovea and declines towards the periphery. The sensitivity to spatial frequencies also drops with eccentricity, but at different rates depending on the frequency.[34] To account for this phenomenon in a multilayer model it would not only be necessary to grade the responses of all layers with eccentricity, but to weight the peripheral areas according to the layer. The model complexity would increase enormously. Since the experiments that I will use and describe in more detail later presented stimuli at around 3 degree visual angle, a model for foveal vision is sufficient .

[35] Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7), 1160–1169

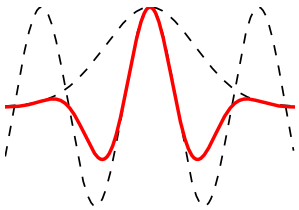


Figure 4: The gabor is the product of a sine wave and a Gaussian

Choosing the filters

The first stage of the vision model is the filter bank. Gabors are commonly used as filters. They gained popularity because of their similarity to receptive fields of V1 cells.[35] The Gabor is no more than a Gaussian windowed sinusoid. It is specified by the parameters of the sinusoid—orientation, spatial frequency, phase—and the Gaussian function that determines the spatial extent of the Gabor. Figure 4 shows a sinusoid and a Gaussian as dashed lines. Multiplied together they result in the Gabor shown as solid line. Further, Gabors were specified to have a bandwidth at half amplitude of 1.4 octaves which is the average measured in cat and monkey[36] and in humans[37].

[36] De Valois, R. L., & De Valois, K. K. (1980). Spatial vision. *Annual Review of Psychology*, *31*, 309–341

[37] Blakemore, C., & Campbell, F. W. (1969). On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *Journal of Physiology*, *203*, 237–260

THE FILTER BANK contains several Gabors that differ with respect to their spatial frequency, orientation, and phase. The different filters we use are spaced in octaves of spatial frequency. More specifically, I choose filters with 1, 2, 4, 8, and 16 cycles per degree. Orientation and phase match the stimulus properties.

EACH FILTER IS NORMALIZED to produce the maximal response to a sinusoidal grating matched to the filter's property, e.g. with the same spatial frequency, orientation, and phase. This normalization assures that the filter producing the maximal response is the one with matched properties. Figure 5 sketches the response strength of some normalized filters to sinusoids of different frequencies. The main diagonal contains the responses of filters to the matched sinusoids which is the maximal response that can be obtained. The figure shows both that the maximal response of a filter is to the matched sinusoid and that for a sinusoid the maximal response is produced by the corresponding filter.

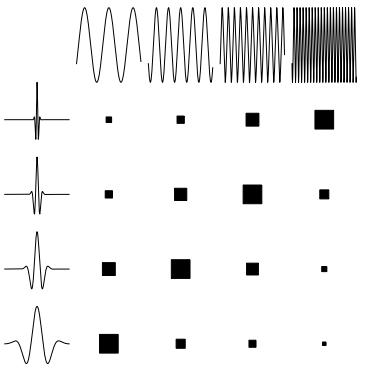


Figure 5: Maximal responses of Gabor filters to matched sinusoids

AN ALTERNATIVE to a filter bank is a filter pyramid. The difference between the filter bank and the pyramid is that for pyramids the occurrence of spatial frequencies and orientation is selected based on theoretical considerations to reduce reconstruction error. Pyramids and filterbanks are overcomplete representations, but a pyramid is less so due to downsampling and thus computationally more efficient than a filter bank. One popular example of pyramids is the Simoncelli's[38] steerable pyramid. However, despite its disadvantages, I selected Gabors as filters to connect to the majority of literature in the area of spatial vision models.

The contrast sensitivity function

A BASIC FINDING OF PSYCHOPHYSICS is that spatial frequencies are not all equally detectable. For long presentation times (>1 sec) frequencies between 3 and 4 cycles per degree are perceived best and detection performance drops for higher and lower spatial frequencies. The finding has been replicated many times[39] and related results are also observable with EEG[40] and fMRI[41] where responses are modulated with spatial frequency and contrast.

[38] Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: a flexible architecture for multi-scale derivative computation. *2nd IEEE International Conference on Image Processing*, *III*, 444–447

[39] Campbell, F. W., & Robson, J. G. (1968). Applicaton of Fourier analysis to the visibility of gratings. *Journal of Physiology*, *197*, 551–566; and Kelly, D. H. (1979). Motion and vision. ii. stabilized spatio-temporal threshold surface. *Journal of the Optical Society of America A*, *69*, 1340–1349

[40] Allen, D., Norcia, A. M., & Tyler, C. W. (1986). Comparative study of electrophysiological and psychophysical measurement of the contrast sensitivity function in humans. *American Journal of Optometry and Physiological Optics*, *63*(6), 442–449

[41] Boynton, G., Demb, J. B., Glover, G., & Heeger, D. J. (1999). Neuronal basis of contrast discrimination. *Vision Research*, *39*, 257–269

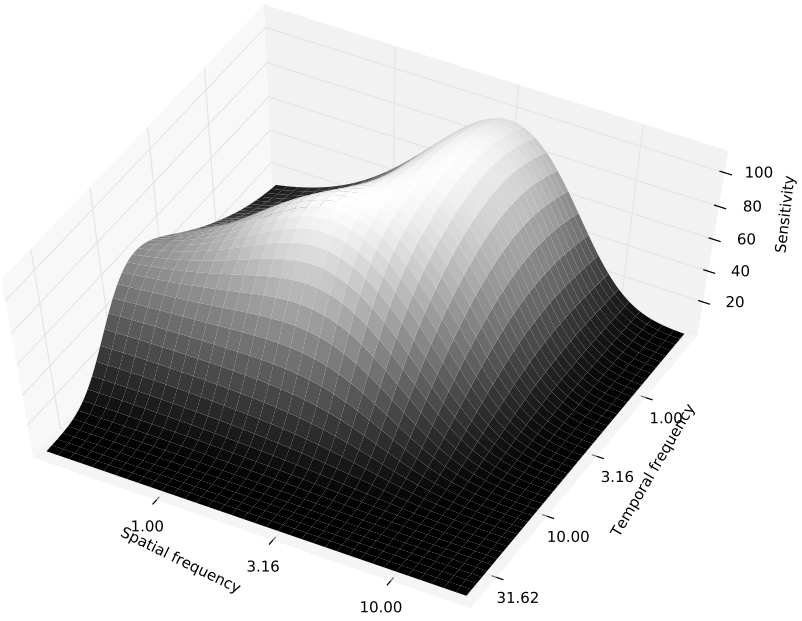


Figure 6: Contrast sensitivity function according to Kelly

The simplest mechanisms that could account for such a finding is either that the mechanisms sensitive to the different spatial frequencies are themselves differently sensitive or that the mechanisms are identical but that the number of mechanisms differs across spatial frequencies. The formalism could be implemented in my modeling either case be a scaling of the response of a filter mechanism by a constant gain, c.

Kelly formalized the contrast sensitivity function for a stationary grating as

$$c = \frac{1}{2} k \mathrm{v} \alpha^2 \exp^{\frac{-2\alpha}{\alpha_{\max}}}$$

with spatial frequency f,
velocity v,
$\alpha = 2\pi \mathrm{f}$
$\alpha_{\max} = 45.9/(\mathrm{v}+2)$
$k = 6.1 + 7.3|\log(\frac{\mathrm{v}}{3})|^3$
This formula is derived from contrast detection experiments with different observers. It only depends on the spatial frequency, f, and the velocity of the gratings, v. However, the scaling of the function is arbitrary.

I WILL GET SLIGHTLY AHEAD OF A PURE MODEL DESCRIPTION and anticipate the specific contrast sensitivity functions of two observers with initials FAW and GBH. The functions are based on the contrast detection experiments summarized on page 33 and page 47 and are taken from Wichmann (1999)[42]. For both observers the best fitting gain in a least squares sense is estimated. The data and resulting function are shown in Figure 7. Since the stimuli were stationary the velocity was fixed to 0.15 as recommended by Kelly. The original function as specified by Kelly drops too fast for high spatial frequencies. An additional variable μ is introduced that allows a shallower decline: $c = \frac{1}{2}kv\alpha^2 \exp^{\frac{-2\mu\alpha}{\alpha_{\max}}}$. The variable μ is fitted simultaneously for both observers with an optimal value of 0.875. The values of the optimal Kelly function evaluated at the spatial frequencies of the filters are set to account for contrast threshold. This means that the contrast sensitivity of the model is predefined and not estimated together with the other parameters.

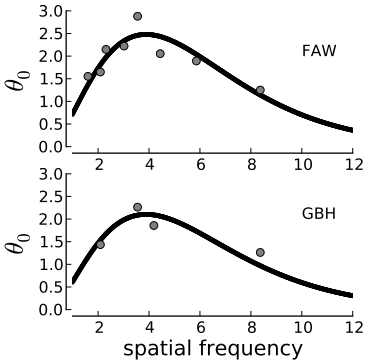


Figure 7: 75% correct response detection threshold and best fitting contrast sensitivity function for observer FAW and GBH

[42] Wichmann, F. A. (1999). *Some aspects of modelling human spatial vision: Contrast discrimination*. Unpublished doctoral dissertation, The University of Oxford, Oxford, UK

The static nonlinearity

IN 1966 the researchers Naka and Rushton recorded action potentials from the fish retina while manipulating input intensity[43] and found that the response function can be described by

$$R(x) = R_{\max}\frac{x^2}{x^2+\gamma},$$

where x is the input intensity and the semisaturation constant, γ, denotes the reflection point of this sigmoidal function. $R_{\max}$ is the maximal response that was observed in the experiments. In the current context the function is used to describe the response of the visual system depending on the input's contrast.

[43] Naka, K. I., & Rushton, W. A. H. (1966). S-potentials from luminosity units in the retina of fish (cyprinidae). *Journal of Physiology*, *185*(3), 587–599

A SIMILAR FUNCTION was used by Heeger in his single neuron model presented in 1992.[44] Shortly afterward, Foley adopted the notation to psychophysical models in the more or less original form.[45] Wichmann used the generalized form $R(x) = \frac{x^{\kappa+\eta}}{x^{\kappa}+\gamma}$ and used the data from discrimination experiments to determine the parameters of the function.[46] Here, almost the same function will be used, because of its generality $R(x) = \frac{x^{\kappa+\eta}}{x^{\kappa}+\gamma^{\kappa}}$. The exponents κ and η and the constant γ determine the shape of the transducer. The general formulation allows the function to equal either a sigmoid for $\gamma > 0$ and $\eta \geq 0$, an exponential for $\gamma = 0$, or linear relationship

[44] Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*, 181–197

[45] Foley, J. M. (1994). Human luminance pattern-vision mechanisms: masking experiments require a new model. *Journal of the Optical Society of America A*, *11*(6), 1710–1719

[46] Wichmann, F. A. (1999). *Some aspects of modelling human spatial vision: Contrast discrimination*. Unpublished doctoral dissertation, The University of Oxford, Oxford, UK

for $\gamma = 0$, $\kappa = 0$, and $\eta = 1$. This variable definition allows to find the transducer that is supported by the data and does not impose a fixed form.

Pooling rule

For a basic detection or discrimination tasks, the spatial vision community assumes that no higher cortical processing is necessary as pointed out by Graham[47]. Behavior could be directly based on the activity of the channels. A general read-out mechanism is necessary to map the activity of the channels onto a single response. A candidate mechanism for stimulus independent pooling is Minkowski pooling, which is another name for the L^p-norm

$$||x||_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

This norm is very powerful and has been used earlier in its general form.[48] Other previous suggestions on how pooling is done are special cases for $p = 2$ and $p = \infty$.

$$||x||_2 := \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}$$

The L^2-norm squares responses before summing.[49] Goris et al. argue that this is the optimal rule considered from a signal-to-noise perspective.[50] By scaling each response with its signal-to-noise ratio given the specific stimulus, the informative responses are enhanced and the uninformative responses are suppressed. Given signal independent noise sources this is proportional to a weighting with the mechanisms mean activity which is just the response of the model.

$$\begin{aligned} ||x||_\infty &:= \lim_{p \to \infty} \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \\ &= \max(|x_1|, \ldots, |x_n|) \end{aligned}$$

A L^∞-norm is equivalent to a maximum operation. It is thus an implementation of the maximum rule which determines the response only based on the activity of the maximally responsive mechanism.[51]

A Minkowski norm as pooling rule faces two main arguments against its use. From a psychological perspective, it seems

[47] Graham, N. V. S. (1989). *Visual Pattern Analyzers*. Issue 16 of Oxford Psychology Series, Oxford University Press

[48] Quick, R. F. (1974). A vector-magnitue model of contrast detection. *Kybernetik*, *16*, 65–67; Graham, N. (1977). Visual detection of aperiodic spatial stimuli by probability summation along narrowband channels. *Vision Research*, *17*, 637–652; and Bowne, S. F. (1990). Contrast discrimination cannot explain spatial frequency, orientation or temporal frequency discrimination. *Vision Research*, *30*(3), 449–461

[49] Watson, A. B., & Ahumada, A. J. (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision*, *5*(9), 717–740

[50] Goris, R. L. T., Wichmann, F. A., & Henning, G. B. (2009). A neurophysiologically plausible population-code model for human contrast discriminations. *Journal Of Vision*, *9*(7), 1–22

[51] Graham, N. V. S. (1989). *Visual Pattern Analyzers*. Issue 16 of Oxford Psychology Series, Oxford University Press; and Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025

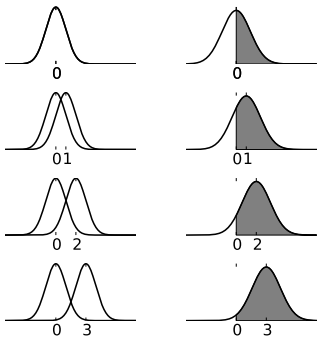
too simplistic to capture the complex behavior supposedly involved in decision-making. Mathematically, the norm is very sensitive to large values possibly being outliers. Yet, despite its drawbacks, the L^p metric is a pleasant choice, first of all, since it allows to interpret the transduced responses as coordinates in a metric space.[52] The origin of that space is the stimulus that elicits no response in neither of the filters: a homogeneous blank stimulus. The sensitivity of all other stimuli is expressed through the distance of the stimulus to the origin. In addition, it is simple to compute and, last but not least, it is the de facto standard in the field. Thus, I will use the general Minkowski norm as decision rule and estimate $p = \beta$ from the data.

[52] Tversky, A., & Krantz, D. H. (1970). The dimensional representation and the metric structure of similarity data. *Journal of Mathematical Psychology*, *7*, 572–596

Differencing and link function

Psychophysical experiments that target the early visual system measure behavior in terms of correct responses. This behavior is an all-or-nothing response for a single trial. However, for nontrivial conditions the response is variable over repetitions and binomially distributed. Each trial in a two-alternative forced choice task (2afc) contains two stimuli, spatially or temporally separated. The observer's task is then to denote the interval that contains the signal, e.g. the higher contrast. The model also needs to discern signal pattern from noise pattern and thereby additionally reflect the observers response distribution.

Signal detection theory provides a principled way to implement this function. Signal detection theory assumes that each stimulus intensity elicits an internal response—a decision variable—that can not directly be observed by the experimenter. But, the strength of this response—the magnitude of the decision variable—determines the behavioral performance. The response strength is not deterministic for a given stimulus due to noise in the system. In a single trial only one response value is measured, but the distribution can be accessed across trials for stimuli that are close to threshold. Thus, not only one internal response is associated to a stimulus intensity, but a response distribution. For a 2afc task two internal response distributions on the decision axis are available, one that stems from the signal trials and one that stems from the distractor trials, also called noise trials. The theory predicts the ability to discriminate noise and signals in terms of correct



(a) signal and noise distributions (b) difference distribution

Figure 8: Link function motivated by Signal Detection Theory

responses based on the difference and variance of the signal and noise distributions. The mathematical derivation is as follows.

The distributions of signal, $X_s \sim \mathcal{N}(\mu_{signal}, \sigma_{signal})$, and noise, $X_{noise} \sim \mathcal{N}(\mu_{noise}, \sigma_{noise})$, are commonly assumed to be Gaussian with unit variance and a mean of 0 for the noise distribution and a mean that depends on stimulus intensity for the signal distribution

$$X_{noise} \sim \mathcal{N}(0,1) \quad X_{signal} \sim \mathcal{N}(\mu_{signal}, 1).$$

This situation is shown in Figure 8a for different stimulus intensities. To form a decision the difference between the two distributions is computed and a criterion applied on the difference distribution. The best possible detection performance is observed for a criterion set at 0. Then, the observer would respond that a signal is present whenever the difference of the distributions is larger than 0. The correct response probability for that criterion therefore equals the area under the difference curve for positive values. Figure 8b sketches the distribution of differences between signal and noise and highlights the area of correct responses. The correct response probability for a fixed signal distribution is

$$\begin{aligned}
P(x=1|X=1) &= P(X_{signal} > X_{noise}) && (1)\\
&= P(X_{signal} - X_{noise} > 0) && (2)\\
&= 1 - \Phi\left(\frac{0 - (\mu_{signal} - \mu_{noise})}{\sqrt{\sigma_{signal} + \sigma_{noise}}}\right) && (3)\\
&= 1 - \Phi\left(\frac{\mu_{noise} - \mu_{signal}}{\sqrt{\sigma_{signal} + \sigma_{noise}}}\right) && (4)\\
&= \Phi\left(\frac{\mu_{signal}}{\sqrt{2}}\right). && (5)
\end{aligned}$$

where Φ is the cumulative normal distribution. If the signal intensity is very large, the internal response is large and almost the full support of the difference distribution resides on the positive side. Thus, the correct response probability is close to 1. If no signal is present in either interval, the internal distributions of signal and noise are equivalent and the difference distribution is centered around the origin. The correct response probability is 0.5. The link function that relates decision variables to correct responses is a sigmoid. It is called probit with the Gaussian noise model from signal detection theory.

THE ASSUMPTIONS OF SIGNAL DETECTION THEORY about the noise model are rather strong and the derivative of the difference distributions only works as nicely if the noise is Gaussian. Therefore, instead of adopting a fixed assumption about the shape of the noise on the decision variable, I implemented a more flexible approach. The boundary conditions for any link function are as follows. The specific input to the function, its argument, is defined by the magnitude of the difference between the signal model response and noise model response. The function values asymptote at the experimentally observed correct response values of 0.5 and 1. Thus, any function that fulfills those boundaries could be used as link function. Initially, I decided to use a logistic psychometric function with a fixed threshold at 1 and a fixed slope of $1/4$.[53] This reflection point agrees with that of other studies that only measured thresholds. The slope is rather steep and results in a function that almost dropped to guessing rate for indistinguishable signal and noise values. In this implementation the slope is not dependent on the signal intensity. I introduced this new flexible approach due to the following advantages:

[53] Intermediate results will lead to a modification of this initial choice.

1. The link function can be simply substituted if required by the data.
2. Explicitly defined signal and noise distributions are not necessary.
3. The variability in responses might be due to variable channel responses or a noisy read-out mechanism.
4. The link function can still be interpreted in the sense of signal detection theory.
5. Lapse rates can be incorporated in the estimation.

THE LAPSE RATE is the reduction of the upper asymptotes if performance does not saturate with 100% correct responses for strong stimulus intensities. The variance of binomial distributions is dependent on the performance level, p, through the relationship $np(1-p)$ where n is the number of trials. As a consequence the variance for performance levels close to 1 is negligible and thus bears strong constraints. Those constraints are not taken into account if the upper asymptote is not shifted towards the saturating performance.

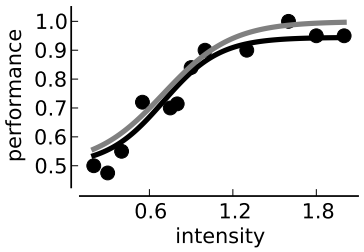


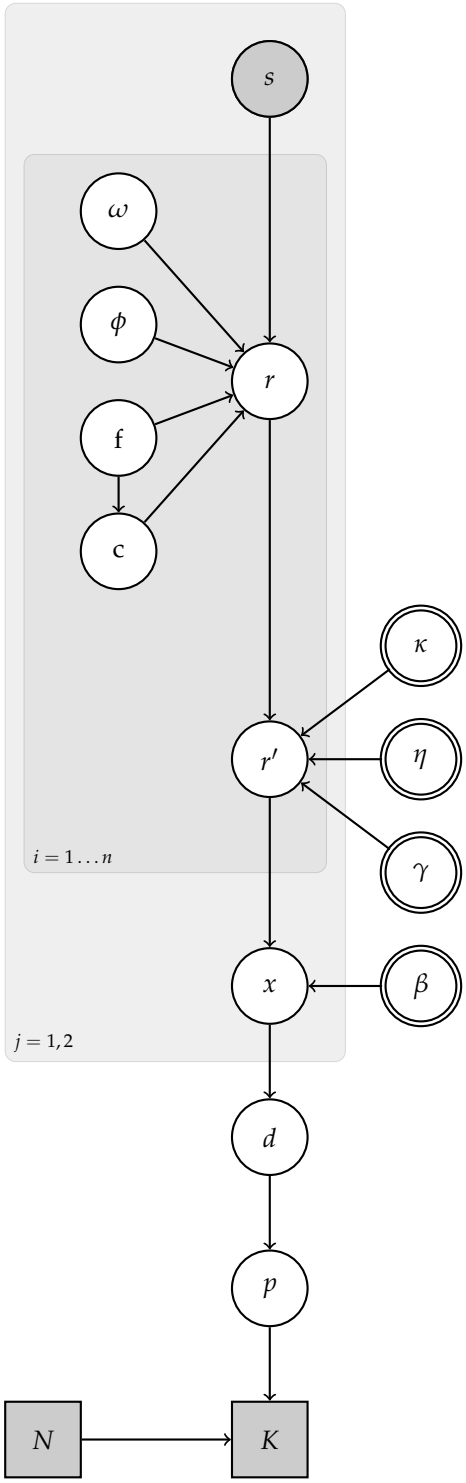
Figure 9: Psychometric function estimated with (black) and without (gray) lapse rate asymptote.

If we choose to incorporate lapse rates a further consequence for modeling arises. The lapse rate is usually not constant for all conditions because it depends on many factors such as stimulus intensity and observer fatigue. Therefore each data set should be equipped with its own lapse rate parameter and the number of parameters increases dramatically. Fortunately, the lapse rate is an auxiliary variable with respect to the model selection and its number is independent of the model. To simplify the parameter estimation, the lapse rate is determined prior to the estimation of the vision model parameters. The best fitting psychometric function is computed for each condition and the lapse rate only inserted during estimation.

Graphical model

To use Bayesian inference as parameter estimator, the model needs to be specified appropriately. I will describe the model as a graphical model and contrast it to the common illustrative version as depicted in Figure 2. The illustrative sketch of the vision model provides a comprehensive and intuitive representation for the model. But, it does not contain much information about the underlying functions and the actual implementation of the model. A graphical model on the other hand supplies that information in form of variables, their relationships, and functions. It consists of nodes which are linked by edges to denote the processing chain in the model. Normally each node is named by a unique name or variable that is the result of the previous processing steps. The formulae linking one node to the other are provided with the graph. The input and the output are observed variables that are known to the modeler. Nodes with this property are filled in a graphical model. Furthermore, stochastic and deterministic variables are differentiated by a single frame or double frames. Deterministic variables are fully defined by the state of their ancestors, stochastic variables are random instances of a given distribution, the so-called prior distribution. The combination of graph, functions, and priors are essential for the full specification of the graphical model.

In order to construct a graphical model as shown in Figure 10, it is most intuitive to start with what we know as an experimenter: the observed variables. The observer has repeatedly seen images and responded with button presses. Therefore, observed



$g = c \exp(\frac{-2.77x^2}{1.4^2}) \cos(2\pi \mathrm{f} x + \phi)$

$\omega \in [90]$

$\phi \in [0]$

$r_{ij} = s_j * g(\omega_i, \phi_i, \mathrm{f}_i, c_i)$

$\mathrm{f} \in [1, 2, 4, 8, 16]$

$c_i = \frac{1}{2} k v (2\pi \mathrm{f})^2 \exp^{\frac{-4\mu\pi\mathrm{f}}{\alpha \max}}$

$\kappa \sim \text{Uniform}\,(0, 50)$

$r'_{ij} = \frac{r_{ij}^{\kappa+\eta}}{r_{ij}^{\kappa} + \gamma^{\kappa}}$

$\eta \sim \text{Uniform}\,(0, 5)$

$\gamma \sim \text{Uniform}\,(0, 1)$

$x_j = \|r'_j\|_\beta$

$\beta \sim \text{Uniform}\,(0, 20)$

$d = |x_1 - x_2|$

$p = 1/(1 + \exp(\frac{d-1}{\sigma}))$

$\sigma = 0.25$

$K \sim \text{Binomial}\,(P, N)$

Figure 10: The graphical model

variables are the image, s, the number of repetitions, N, and the number of correct responses, K. Everything else happens in the head of the observer or in the model and cannot be observed. In the previous section each step of the model was explained and mathematical descriptions given. Now these steps need to be embedded in the graphical modeling framework. The two stimuli presented in a single trial, s_1 and s_2, are processed by the same processing cascade denoted by the outer panel with index j. They are first mapped into a filter space by convolution with n Gabor filters shown in the inner panel. A set of filters g_i is specified by their spatial frequencies f, the orientations ω and phase ϕ. The filters are scaled depending on their spatial frequency according to Kelly's function. In the next processing step the filter outputs r_{ij} are transduced by a Naka-Rushton sigmoid. The parametrization of the Naka-Rushton nonlinearity uses three parameters κ, η, and γ. Each parameter is a random variable. Since no prior knowledge is assumed, the prior distributions for all three parameters are uniform. The numerous transduced filter responses $r\prime_{ij}$ are combined for each stimulus individually to form a single decision variable x_j by computing the L^p-norm with p= β. The two variables on the decision axis are finally combined to a single response by the magnitude of their difference. The resulting variable d is related to $d\prime$ in signal detection theory if the noise and signal distributions were Gaussian. The decision variable is linked to the correct response probability, p, in a single trial by the logistic function. The number of observed correct responses stems from a binomial distribution with the parameters p and N.

With a Bayesian inference procedure the posterior distributions of the stochastic variables are estimated given empirical data from psychophysical experiments. Specific estimates, e.g. mean, median or mode, confidence intervals, and other statistics about the model and its parameters can be obtained from the parameter posterior distributions.

The identifiability of vision model parameters

In the two preceding sections I explained the concepts of Bayesian model analysis and the structure of an early vision model. In the following section I will demonstrate that, in addition to the model implementation and to the analysis chosen, a third critical factor determines the success when analyzing a model: the data.

Past studies that estimated the parameters of a vision model were based on a single[54] or several data sets[55]. They use these data sets and estimate the parameters. Then they present the reader that the models are capable of explaining the data. Itti et al provide variance estimate for their parameters. However, then they do not pursue the enormous variances that are obtained for some parameters. In the other studies, the model is not analyzed further based on the estimation results. The following pages document a reiterative process between a Bayesian analysis of the model, the model's predictions assessment, and resulting adaptations of the analysis by either adapting the model or the data foundation. Thereby, it is not only possible to show that the model is able to explain the data, but also to point out if the model parameters can be identified and to provide hints for the reasons why the identification is a success or failure.

In a first attempt, I will instantiate a model, use detection data as a foundation, estimate the parameters, and inspect the model's predictions. In so far, it is merely a repetition of and link to previous attempts.

All experimental data shown in this section was provided by Wichmann and were used previously in his doctoral thesis[56]. The detection data set is collected under a temporal two-alternative forced choice detection paradigm in which the observer's task was to detect a stimulus from a homogenous background of mean luminance. The stimuli extended over 3.1 degree visual angle and were presented for 79 ms. The stimulus resolution was 256×256 pixels and a raised cosine window function was used. The experiment consists of 8 conditions which differ in spatial frequencies and ranged from 1.6 cycles per degree to 8.4 cycles per degree. For each experimental condition a full psychometric function was measured. The signal contrasts were measured in a blocked procedure, that

[54] Watson, A. B., & Solomon, J. A. (1997). Model of Visual Contrast Gain Control and Pattern Masking. *Journal of the Optical Society of America A*, *14*(9), 2379–2391; and Watson, A. B., & Ahumada, A. J. (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision*, *5*(9), 717–740

[55] Itti, L., Koch, C., & Braun, J. (2000). Revisiting spatial vision: toward a unifying model. *Journal of the Optical Society of America A*, *17*(11), 1899–1917; and Goris, R. L. T., Wichmann, F. A., & Henning, G. B. (2009). A neurophysiologically plausible population-code model for human contrast discriminations. *Journal Of Vision*, *9*(7), 1–22

[56] Wichmann, F. A. (1999). *Some aspects of modelling human spatial vision: Contrast discrimination*. Unpublished doctoral dissertation, The University of Oxford, Oxford, UK

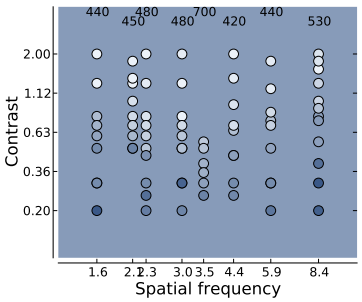
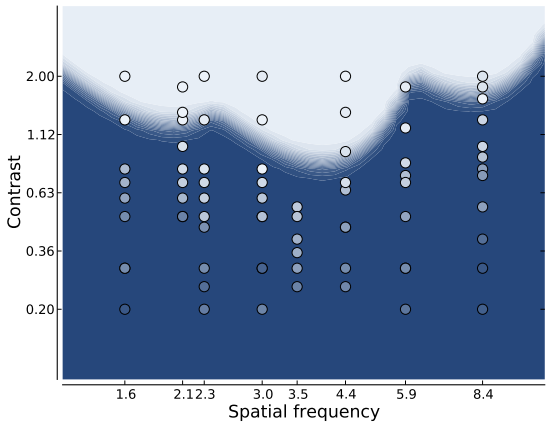


Figure 11: Experimental results from a contrast detection experiment

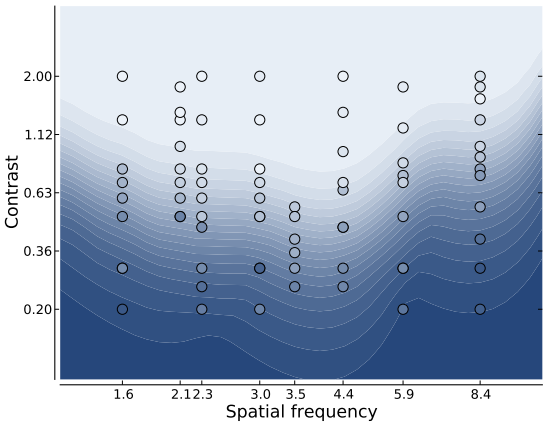
means that all trials with stimuli of the same signal contrast were presented in succession. The median trial number for one signal contrast was 50 trials. Blocks that were collected while the observer was still learning were excluded. A total of 3940 trials are analyzed in this data set. Figure 11 presents the results. The signal contrast is plotted against the spatial frequencies. The proportion of correct responses determines the color of the data point. A performance of 0.4 corresponds to a dark blue and perfect performance of 1 corresponds to light blue. The middle blue background of the figure corresponds to a performance level of 0.75. Thus, the data points that are lighter than the background are above the 75% correct response threshold and points darker than background are below threshold. The U-shaped detection pattern is characteristic for this type of experiment and the inverse of the contrast sensitivity function described on page 23. The numbers written in the upper part of Figure 11 are the cumulative numbers of trials for each condition.

The model is composed of different stages that include sampling of the retinal image, extending over 2.5 degree with 205 samples per visual angle, and a filter bank with 5 Gabor filters with spatial frequencies of $1, 2, 4, 8$, and 16 cycles per degree. Each filter is matched to stimulus orientation and phase. The contrast sensitivity function is estimated on the same data set by interpolating the thresholds of the measured spatial frequencies. A static nonlinearity $\frac{x^{\kappa+\eta}}{x^{\kappa}+\gamma^{\kappa}}$ transduces the filter responses which are pooled through the L^{β}-norm. The final internal response which resulted from differencing between Minkowski results is mapped onto observed probabilities through a link function.

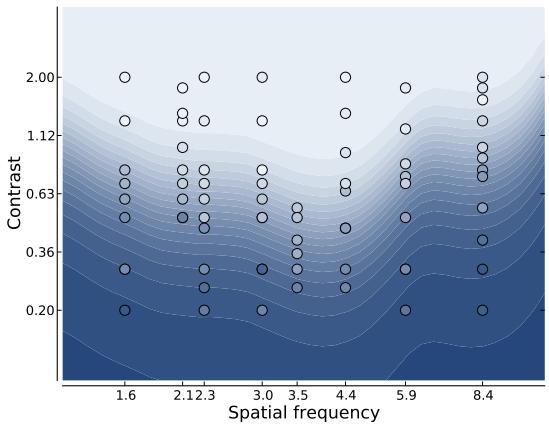
A Metropolis-Hastings algorithm is used to sample from the model's parameter distribution. In total 1.100.000 samples were drawn. The first 100.000 were excluded and the remaining 1.000.000 were thinned by a factor of 100. The mean of the remaining 10.000 samples was taken as estimate and is shown in the first row of Table 4 labeled as Model 1a. Additionally, the table lists the deviance that was evaluated at n signal contrasts (blocks). The predictions of this model instance are shown in Figure 12a. The data is presented in a fashion equivalent to Figure 11 with the only difference that the contour in the background shows the predictions. The contour gradient ranges from chance, 0.5, to perfectly correct responses, 1. That means the lapses are not visualized in the contour representation. During the sampling procedure however, they are taken into



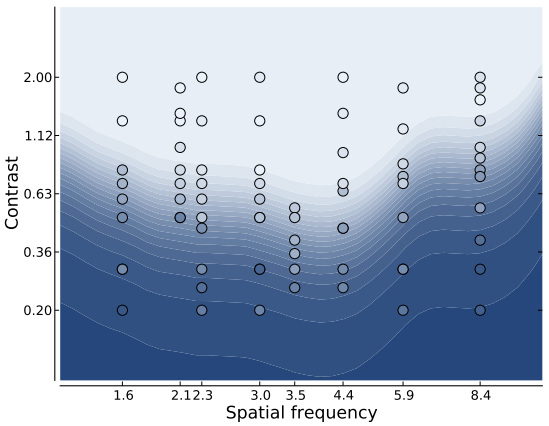
(a) Model 1a



(b) Model 1b



(c) Model 1c

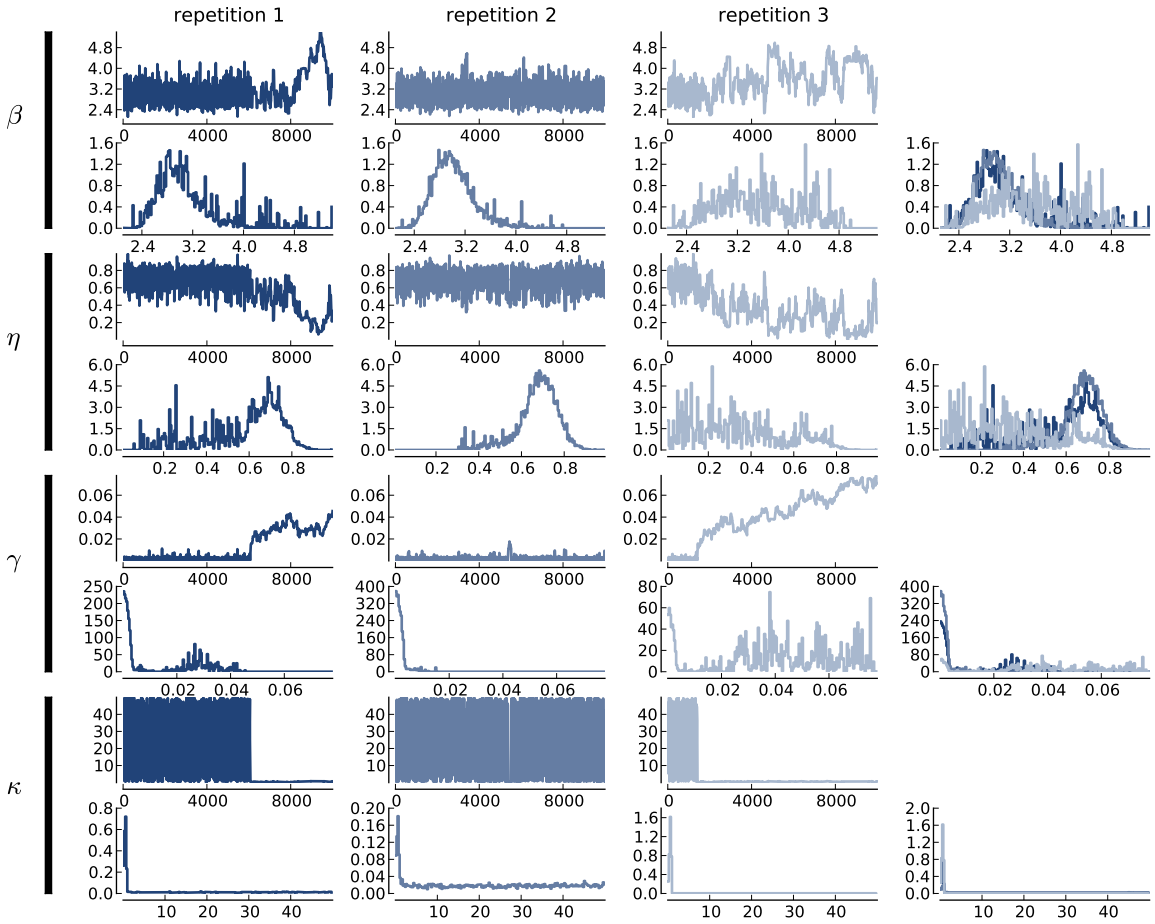


(d) Model 2

Figure 12: Model predictions for contrast detection experiment

account for each data set individually. The gradient is very strong and its shape shows both the contrast sensitivity function and the spatial frequency tuning of the filters. The gradient shows dips at which the model is more sensitive compared to neighboring spatial frequencies. Those dips are located at 2, 4, and 8 cycles per degree, the frequencies contained in the filter bank. There would be two more dips one at 1 cycle per degree and one at 16 cycles per degree that fall outside the shown frequency range. The general U-shaped structure of the gradient reflects contrast sensitivity. Only the fast transition of the gradient does not reflect the shallower performance rise of the data.

Figure 13: Marginal posterior distributions and MCMC chains obtained with contrast detection data (FAW)



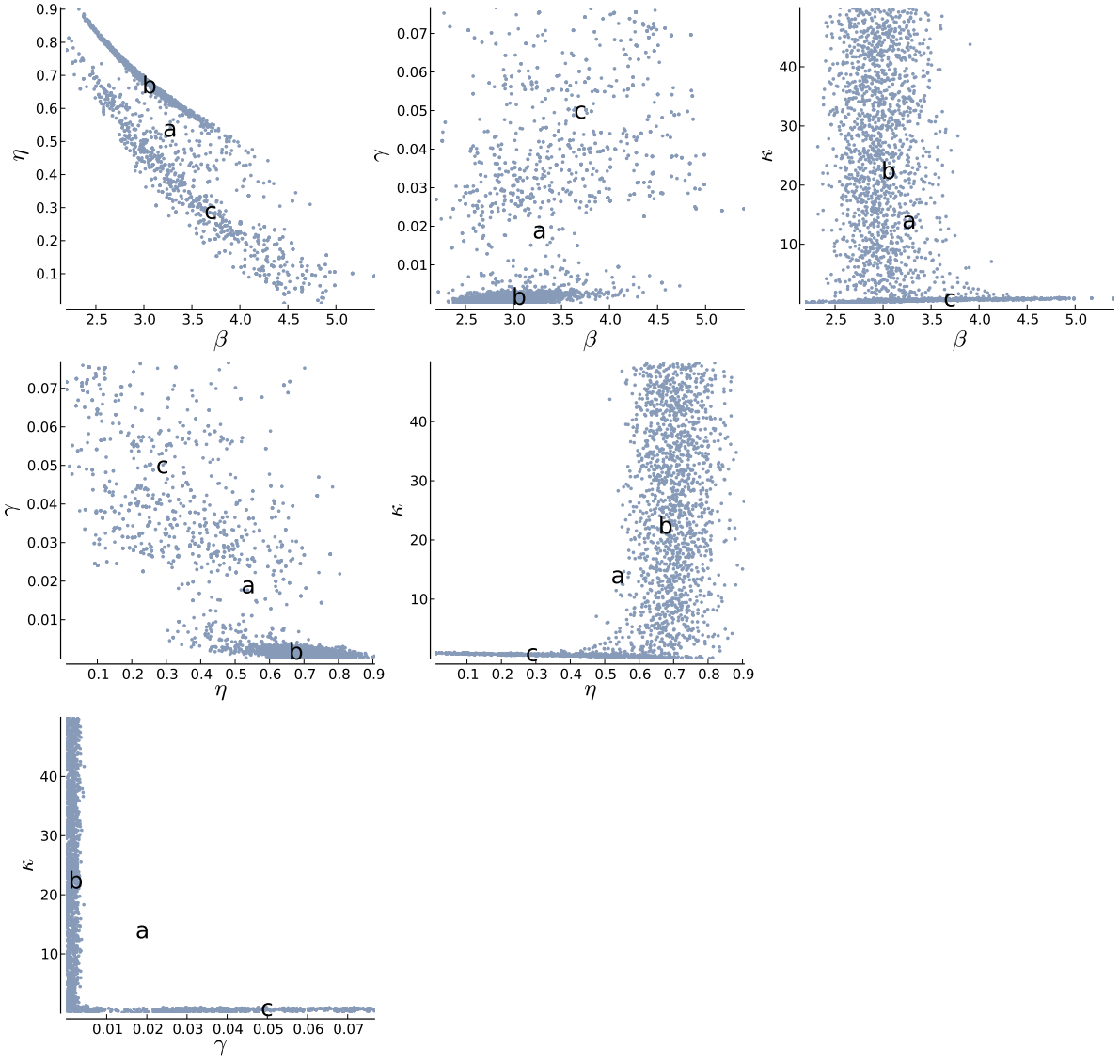


Figure 14: Pairwise correlations obtained with contrast detection data (FAW)

Given the samples from the MCMC procedure, the analysis need not end with the prediction of the results. The diagnostics of the MCMC samples can be used to search for the origin of the gradient missestimation. Figure 13 shows three traces and their histograms from repetitions of the sampling procedure. The repetitions are arranged in columns and a row for each of the four parameters. The last column overlays the histograms of the repetitions. It is clear at first sight that something did not work as expected. The reason is most dominant for parameter κ shown in the last row. Their are two states in which the sampler remains. The first ranges from 0 to 50 which is the region that is allowed by the uniform prior. The second sampling state is confined between 0 and 1. The other parameters also exhibit two states, but less extreme than κ and they do not form independent traces. The two sampling states result in a bimodal structure of the parameter posterior distribution which is also apparent in Figure 14. The projection of the samples on two parameter dimensions is shown for each parameter pair. The letter **a** indicates the mean estimate of all samples. In each subplot two clusters or regions can be determined. A mean taken across clusters is prone to lead to estimates that reside neither in one nor the other cluster. Such an estimate has a low likelihood and is the source for the poor model predictions in Figure 12a and the high deviance in Table 4.

Table 4: Mean estimates and deviances obtained on the detection data set. Parameters of the link function are not estimated.

	nonlinearity	pooling	link	deviance (n=68)
Model 1a	$\kappa = 13.64$	$\beta = 3.26$	*logistic*	1311
	$\eta = 0.53$		$\alpha = 1$	
	$\gamma = 0.02$		$\sigma = 0.25$	
Model 1b	$\kappa = 22.5$	$\beta = 3.04$	*logistic*	99
	$\eta = 0.67$		$\alpha = 1$	
	$\gamma = 0.002$		$\sigma = 0.25$	
Model 1c	$\kappa = 0.58$	$\beta = 3.69$	*logistic*	128
	$\eta = 0.28$		$\alpha = 1$	
	$\gamma = 0.05$		$\sigma = 0.25$	
Model 2		$\beta = 2.1$	*logistic*	112
			$\alpha = 1$	
			$\sigma = 0.25$	

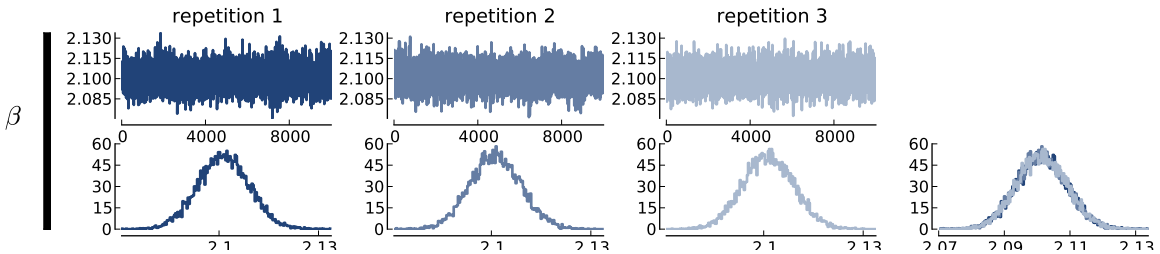
If the means of individual clusters are chosen as estimates, the predictions are not as poor after all. The predictions for the cluster means denoted as **b** and **c** in Figure 14 are presented in the contours of Figures 12b and 12c. Thus, the sampling procedure did not truly fail. The pitfall is hidden in the parametrization and the formulation of the model in conjunction with the data set at hand.

The problem in the model formulation can be deducted from the parameter posterior distributions. The function of the nonlinearity is $\frac{x^{\kappa+\eta}}{x^{\kappa}+\gamma^{\kappa}}$. The samples in cluster b combine rather large values for κ with already small values for γ. The factor that is added in the denominator becomes negligible and the fraction can be reduced by κ. All that is left from the nonlinearity is x^{η} and even η can be left out of the computations since it melts into the parameter β that is applied in the subsequent stage.

$$\frac{x^{\kappa+\eta}}{x^{\kappa}+\underbrace{\gamma^{\kappa}}_{>>\kappa,\gamma<1=0}} = \frac{x^{\kappa+\eta}}{x^{\kappa}} = x^{\eta}$$

Accordingly, a model without static nonlinearity (model 2 in Table 4) should be equally good in predicting the contrast detection data. Figure 12d shows the prediction of such a model and Figure 15 the sampling traces of the only remaining parameter β. The mixing of the traces is very good and the deviance is in the range of the overparametrerized models 1b and 1c.

Figure 15: Same as Figure 13, but for the model without static nonlinearity.



To EXPLAIN, why detection data is not enough to constrain the model and which kind of experiment is needed instead, it is necessary to understand the information that a detection experiment offers. As a matter of fact, a detection experiment is closely related to estimating a derivative. Assume that two stimuli are presented in a single trial, x_0 and x_S. These input quantities are processed by the function f and result in output quantities $f(x_0) = y_0$ and $f(x_S) = y_0 + h$ which are related to the observer's performance. A sketch of this scenario is drawn in Figure 16. It is possible to compute the derivative with respect to y for the system's function f.[57] The limit $\lim_{h\to 0}$ can not be measured experimentally. Instead one can think of the different contrast values that are measured in a detection experiment for a single stimulus to serve as estimates for

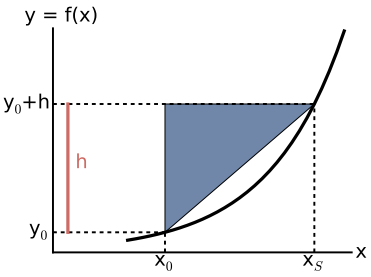


Figure 16: The difference quotient

[57]
$$f^{-1\prime}(y_0) = \lim_{h\to 0} \frac{f^{-1}(y_0+h)-f^{-1}(y_0)}{h} = \lim_{h\to 0} \frac{x_S-x_0}{h}$$

[58]

$$f^{-1*}(y_0) = \lim_{h \to \theta} \frac{f^{-1}(y_0+h) - f^{-1}(y_0)}{h}$$
$$= \lim_{h \to \theta} \frac{x_S - x_0}{h}$$

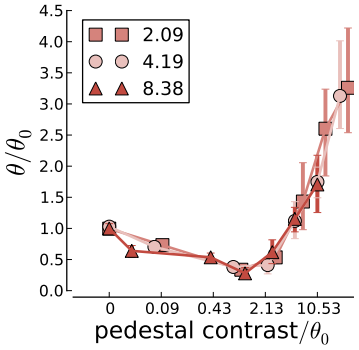


Figure 17: 75% correct response threshold for detection of sinusoids of different spatial frequency on pedestals presented for 79 ms

[59] Bird, C. M., Henning, G. B., & Wichmann, F. A. (2002). Contrast discrimination with sinusoidal gratings of different spatial frequency. *Journal of the Optical Society of America A*, *19*(7), 1267–1273

[60] Nachmias, J., & Kocher, E. C. (1970). Visual detection and discrimination of luminance increments. *Journal of the Optical Society of America*, *60*, 382–389; and Nachmias, J., & Sansbury, R. V. (1973). Grating contrast: Discrimination may be better than detection. *Vision research*, *14*, 1039–1042

the limit $\lim_{h \to \theta}$, where θ is a fixed predefined performance level.[58] As a consequence, even if several contrasts are measured, they only serve to determine a single limit. With respect to the progression of f with increasing contrasts, the measurements in a detection experiment only provide information about the behavior around zero contrast. This is not even enough to know whether the contrast response function is linear or nonlinear. To determine the shape of the contrast response function, it would be necessary to evaluate the limit $\lim_{h \to \theta}$ for a wider range of reference contrasts x_0. This is exactly what is done in contrast discrimination experiments.

Bird, Henning and Wichmann reported data from contrast discrimination experiments which they repeated for several spatial frequencies.[59] The data was also collected in a two-alternative forced choice task. Both intervals contained a stimulus, one having higher contrast than the other. The remaining stimulus properties such as spatial frequency and orientation were identical. The stimulus of lower contrast is called pedestal or mask and its contrast is the pedestal contrast. The observer's task is to respond to the stimulus of higher contrast. Experimental conditions differ with respect to the pedestal contrast that ranged from 0% (equivalent to the detection case) to 25.6% contrast. Figure 17 is a reanalysis of Bird et al's data. It shows the 75% correct response discrimination thresholds normalized with the 75% correct response detection thresholds. Errorbars denote the 90% confidence intervals and each color codes for a different spatial frequency. The figure shows that the amount of additional contrast needed to detect a signal on a low contrast mask is even less than for detection and it increases rapidly with higher pedestal contrasts. This effect is called the dipper effect and it reflects the contrast response function.[60] Bird et al. found that the dipper's shape is consistent across spatial frequencies. Differences between spatial frequencies are a consequence of the contrast sensitivity function. If corrected for that, namely by normalizing with detection threshold, θ_0, the shape of the dipper function remains constant. Therefore, the nonlinearity is invariant with respect to spatial frequency. The consequence of this study for the model is that responses are scaled by the contrast sensitivity of the respective filter prior to the application of the nonlinearity. Then, it should also be sufficient to estimate the nonlinearity with contrast discrimination data of a single spatial frequency.

ACCORDINGLY, the analysis is repeated with a second data set that consists of contrast discrimination data of 8.37 cycles per degree. If above explanation and the experimental finding of Bird et al. is correct, this data should be able to constrain the parameters. The spatial frequency is 8.4 cycles per degree in all conditions and a total of 3250 trials are analyzed in this data set. Other experimental variables are the same as for the first data set. The results are shown in Figure 18. Here, the signal contrast is plotted against the pedestal contrast. Again the achieved performance is coded by color and ranges from 0.4 in dark red to 1 in light red. The middle red background corresponds again to a 0.75 performance level.

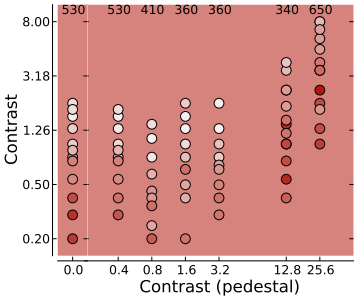


Figure 18: Experimental results from a contrast discrimination experiment

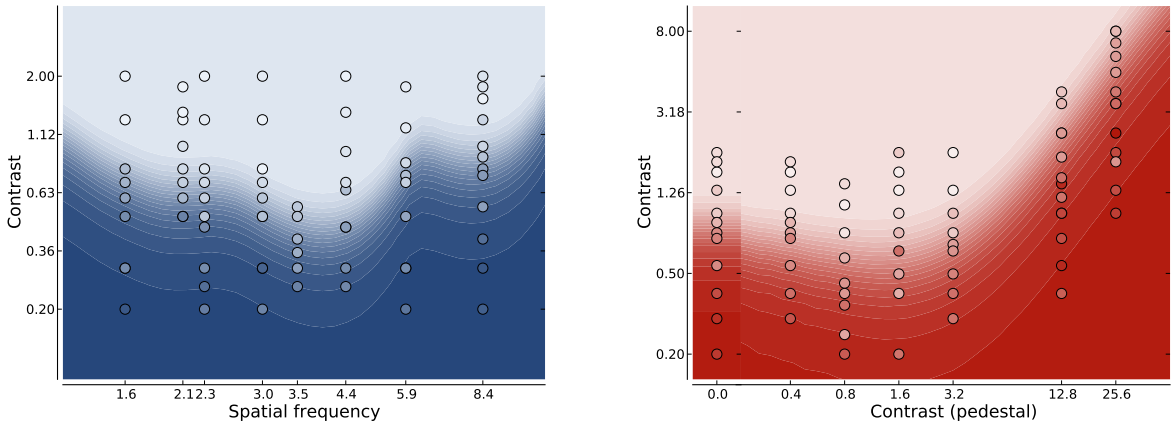


Figure 19: Model prediction of contrast detection and discrimination experiment

FIGURE 19 shows the observed data and the model predictions, the response probability p, in the background. Again the darkest color in the background corresponds to a performance of 0.5 and the lightest to 1. Both the pattern of the inverse contrast sensitivity function and the dipper effect are predicted. However, the predictions especially for the discrimination data are too low for low signal contrasts. This failure can be directly attributed to the link function. The model was instantiated with a fixed logistic link function having its inflection point at 1 and a scale of 0.25. The inflection point is arbitrarily set on the decision axis and the specific value 1 was chosen to facilitate comparison with studies published prior to the present work. The value of the scale parameter follows directly from the inflection point. Earlier I postulated that the link function must approach 0.5 for decreasing stimulus intensities. The

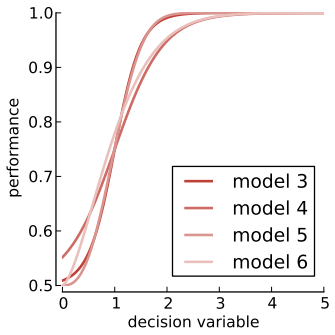


Figure 20: Different link functions with estimates from Table 5

Table 5: Mean estimates and deviances obtained on the contrast discrimination data set. Parameters of the link function are only estimated for model 4 and model 6

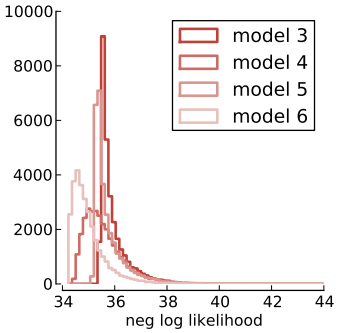


Figure 21: Negative log likelihood distribution of samples

scale value was selected accordingly. The predictions in Figure 19 challenge the choice of link function since it drops faster to chance than the data. To obtain a better fit, the link function must be shallower in the lower part. With a logistic link function this means to give up the necessity of a true guessing rate. Alternatively, a link function that matches the limit behavior per definition could have been chosen such as the cumulative distribution function of a Weibull distribution. The family of Weibull distributions are bound to positive values and they can represent many different shapes also shapes with a slower rise in the lower tail. To compare the performance of different link functions, the estimation procedure was repeated for four models differing only in the link function using the contrast discrimination data set.

	nonlinearity	pooling	link	deviance(n=71)
Model 3	$\kappa = 1.4$	$\beta = 3.33$	*logistic*	116
	$\eta = 0.16$		$\alpha = 1$	
	$\gamma = 0.04$		$\sigma = 0.25$	
Model 4	$\kappa = 1.59$	$\beta = 3.22$	*logistic*	86
	$\eta = 0.2$		$\alpha = 1$	
	$\gamma = 0.03$		$\sigma = 0.46$	
Model 5	$\kappa = 1.39$	$\beta = 3.33$	weibull	115
	$\eta = 0.15$		$\alpha = 1.14$	
	$\gamma = 0.04$		$\sigma = 2.66$	
Model 6	$\kappa = 1.62$	$\beta = 3.29$	*weibull*	84
	$\eta = 0.2$		$\alpha = 1.14$	
	$\gamma = 0.03$		$\sigma = 1.54$	

The first model fitted to contrast discrimination data, model 3, has a fixed logistic link function and is the model that was used to generate Figure 19. The shape parameter of model 4 was not fixed and estimated together with the other model parameters. Model 5 and 6 use a Weibull link function. The Weibull link function also contains two parameters, one for position and one for the shape. The Weibull function with the best fit to the logistic function from model 3 was determined and used to instantiate model 5. Model 5 can therefore be thought of as a control, since it should produce similar results as model 3. For model 6 the shape parameter was estimated together with the other model parameters. Table 5 contains an overview of the four models with different link functions and the parameter estimates rounded to two digital places. The last column contains the deviances of the model calculated on the discrimination data set. Figure 20 visualizes the link functions of the four models and Figure 21 contains histograms of likelihoods

obtained during the sampling process. As expected, model 3 and model 5 are almost equivalent in likelihoods, deviance and parameter estimates. The two models with adaptable shape parameter are more likely than the models with fixed shape. Figure 22 show the estimates of the models' parameters including confidence intervals. The confidence intervals of model 3 and model 5 are nearly overlapping. In general, the nonlinearity and pooling parameters do not differ tremendously. The best results in terms of likelihood and deviance are obtained for the Weibull link with estimated shape parameter. I will therefore only discus model 6's prediction and sampling diagnostic in detail.

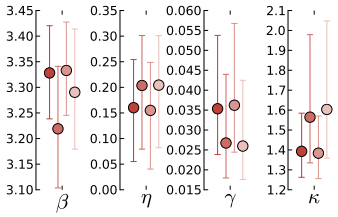


Figure 22: Mean estimates and 95% confidence intervals of parameter estimates from model 3 (dark) to model 6 (very light)

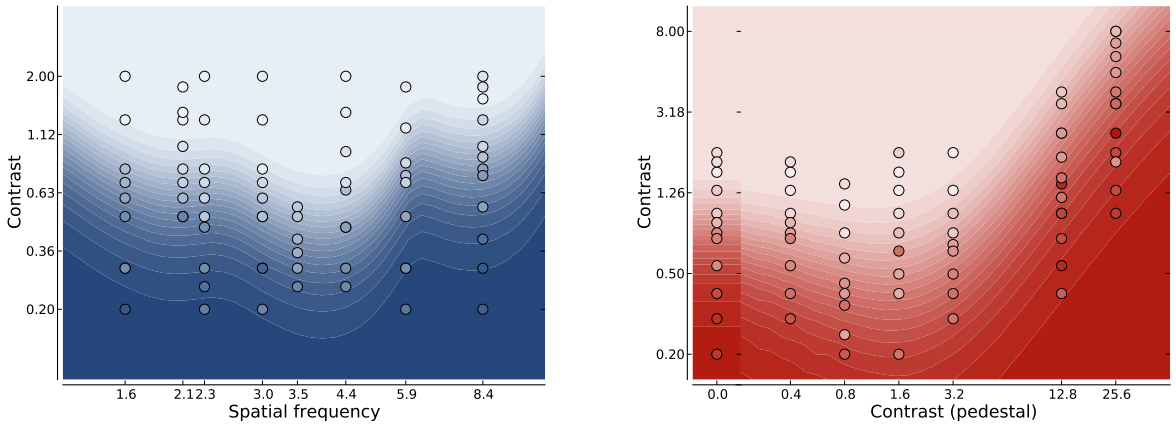


Figure 23: Model 6's prediction of contrast detection and discrimination experiment

An overlay of model predictions and experimental data is shown in Figure 23. As expected because of model 6's similarity to model 3, the general pattern of contrast sensitivity and dipper effect are well reflected. The improved likelihood can indeed be attributed to the better fit of the low signal contrasts. Figure 24 shows three MCMC repetitions with the same arrangement as Figure 13 and an extra row for the shape parameter called σ. The posterior distributions are well-defined and all histograms are constrained. The last column contains overlays of the histograms from the three runs and they are almost identical.

Given the promising marginal parameter distributions, the correlations between the parameters remain to be analyzed. I will first discuss the principle components of the posterior in

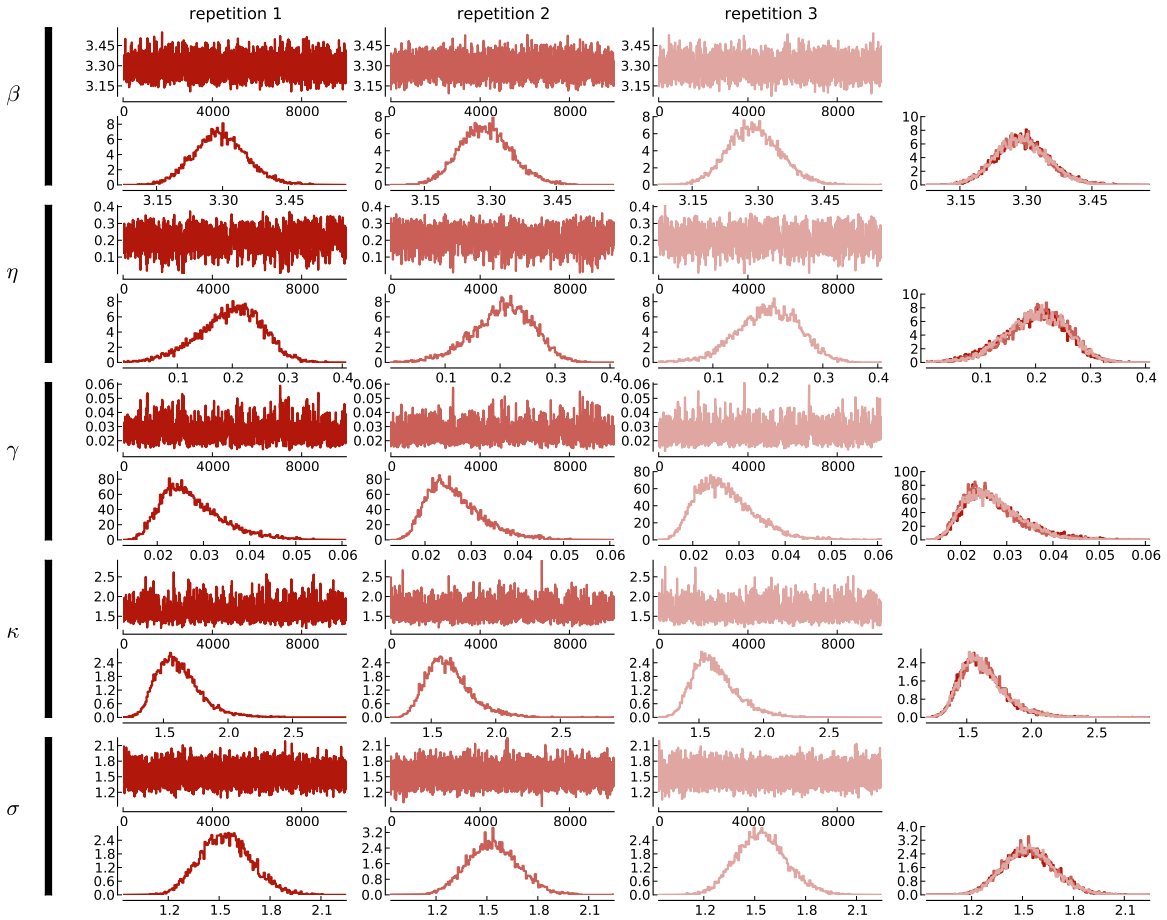


Figure 24: Marginal posterior distributions and MCMC chains obtained with contrast discrimination data (FAW) and Weibull variable link

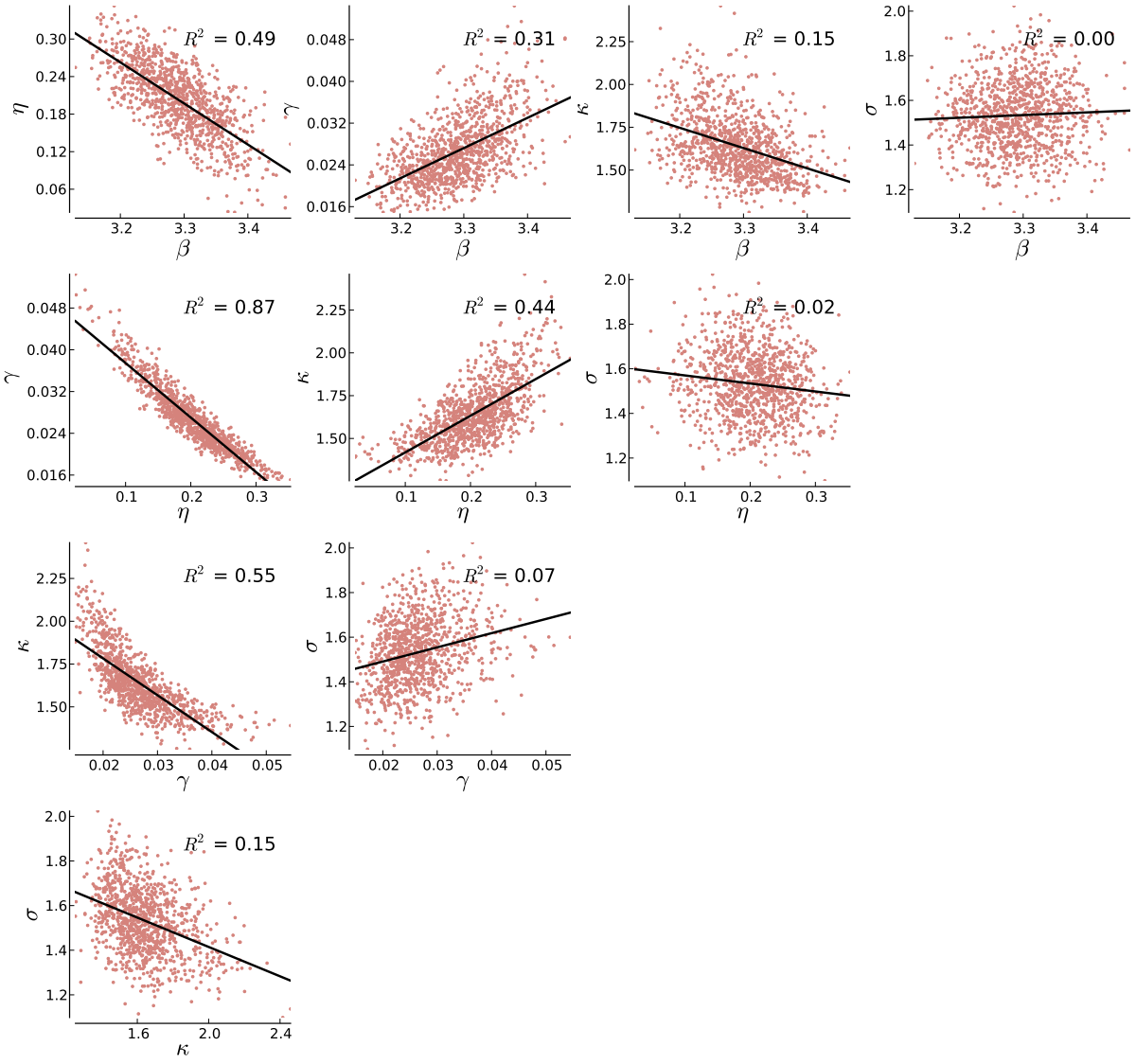
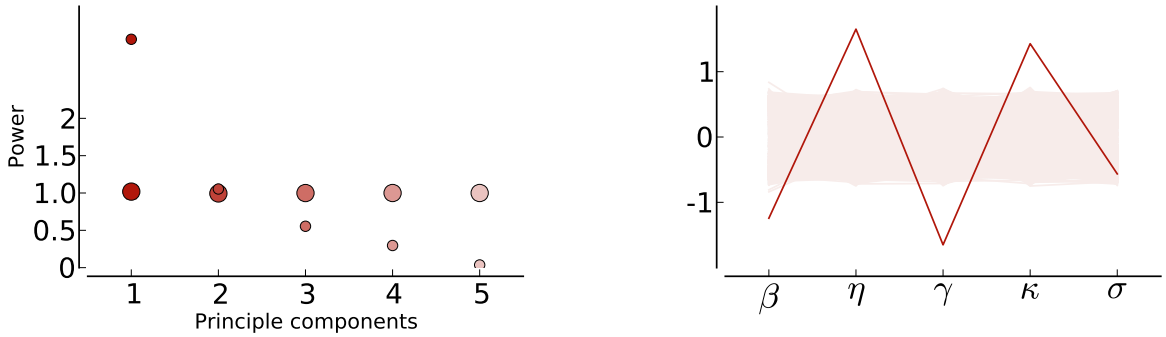


Figure 25: Pairwise correlations obtained with contrast discrimination data (FAW) and Weibull variable link



(a) Power of principle components

(b) First principle component

Figure 26: Principal component of parameter distribution obtained with Weibull variable link

Figure 26, derive their information about the parameters and proceed with the pairwise scatter plots in Figure 25. The first principle components is very strong as depicted in Figure 26a. It resides far outside of the confidence intervals of the power of the principal components for an uncorrelated distributions. The confidence intervals are so small that they disappear behind the mean at power 1. This first principal component shown in Figure 26b should therefore capture the correlation structure well. The first principal component says that parameters that fall on the same side of the bootstrapped range for uncorrelated distributions are positively correlated. This are the pairs of β-γ and η-κ. The parameters that fall on opposite sides are negatively correlated. This are the pairs of β-η , β-κ, γ-η and γ-κ. The parameter σ remains inside of the bootstrapped range and can be regarded as independent of the other parameters. The scatter plots in Figure 25 show two-dimensional projections from the parameter space. The trends already present in the first principal component are confirmed and supplemented with more details. Linear regressions are computed to quantify the dependencies and the R^2 are provided in the figures. The shape variable σ shows only minor correlations with other parameters. Since the previous stages should only determine that a stimulus at threshold is mapped to a predefined value on the decision axis, this result was expected and its occurrence affirmative. The performance changes around threshold are only accounted for by the link functions shape. The shape should therefore also not correlate strongly with other parameters. This is also the reason that the other parameters did not vary tremendously between model 4 to

model 6. The parameter β, the pooling exponent, is correlated with the parameters of the nonlinearity. Its strongest correlation is with the exponent in the numerator η. A change in η has the strongest effect on the magnitude of the transduced filter responses and β needs to counterbalance that effect to maintain the result in the area of the threshold decision variable. The parameters of the nonlinearity η, γ, and κ are all strongly correlated. The largest correlation is observed between the parameters η and γ with $R^2 = 0.87$. That means that 87% of the variance in one parameter can be explained by the other parameter. Such strong correlations within parameters of one processing stage hint to an inappropriate parametrization. Either the three parameters can be disentangled or fewer parameters might be sufficient.

	nonlinearity	pooling	link	deviance(n=64)
Model 6	$\kappa = 2.55$ $\eta = 0.31$ $\gamma = 0.02$	$\beta = 3.00$	weibull $\alpha = 1.14$ $\sigma = 1.51$	94

Table 6: Mean estimates and deviances obtained on the contrast discrimination data set for observer GBH. Parameters of the link function are estimated.

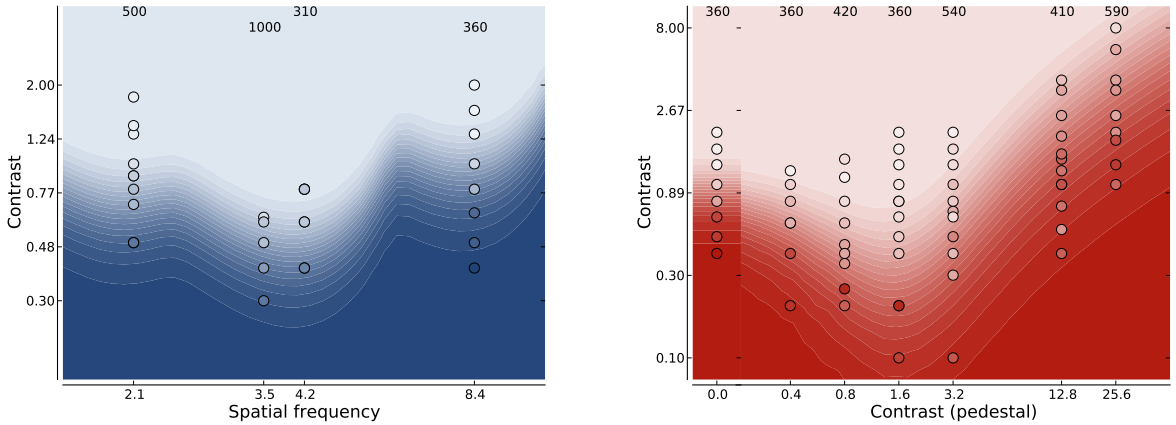


Figure 27: Model 6's prediction of contrast detection and discrimination experiment (GBH)

The estimation procedure was repeated for the most sensible model, model 6, and data from a second observer, GBH, to corroborate the findings obtained so far. The mean estimates of the procedure (Table 6) were used to predict the experimental data. The predictions are shown in Figure 27 with the data as circles and the predictions as background contours. The data sets of observer GBH comprise 3040 trials in the discrimination and 2170 trials in

the detection case.The predictions capture both contrast detection and contrast discrimination data well. Contrast detection experiments were only conducted for four different spatial frequencies, but they are sufficient to estimate the contrast sensitivity function according to Kelly's function which is shown in Figure 7. The contrast discrimination experiments were used to estimate the model parameters in the MCMC procedure which was equivalent to the routine described for observer FAW.

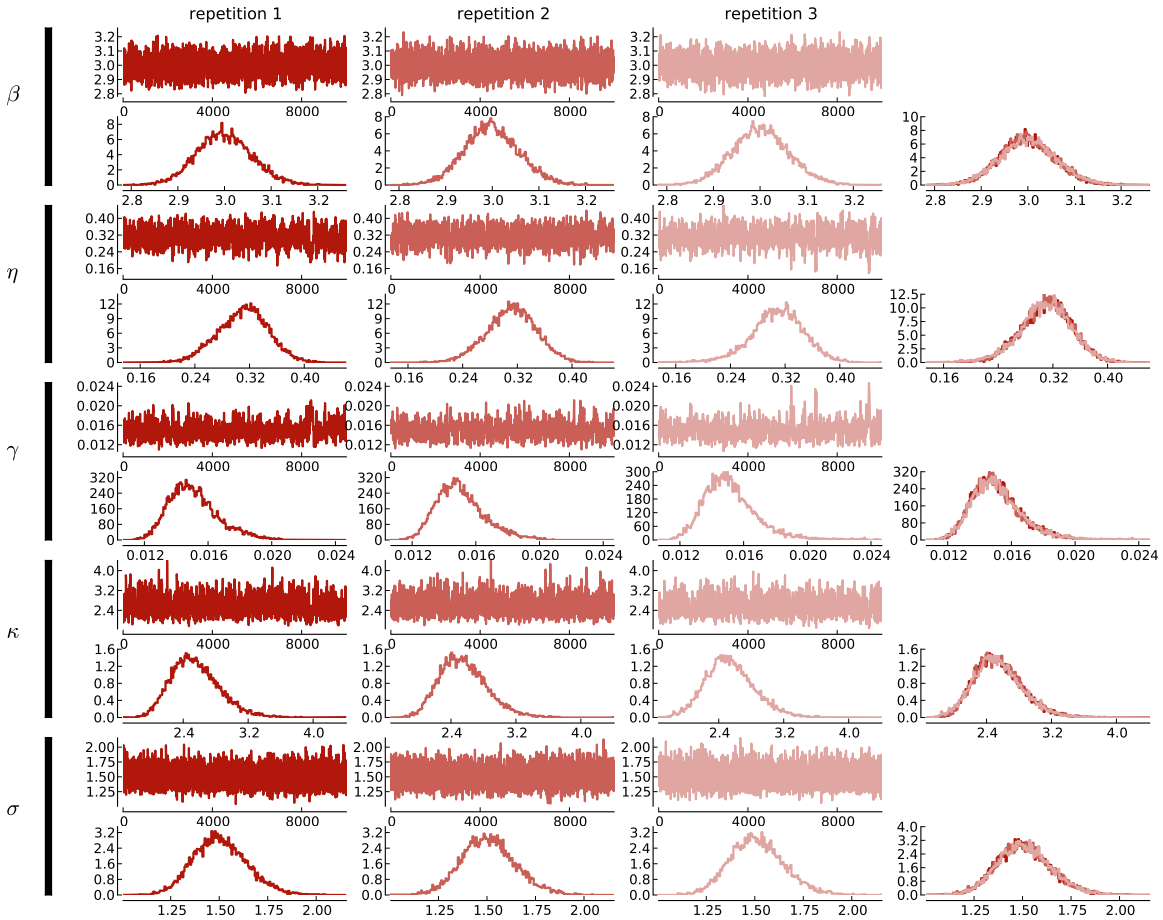


Figure 28: Marginal posterior distributions and MCMC chains obtained with contrast discrimination data (GBH) and model 6

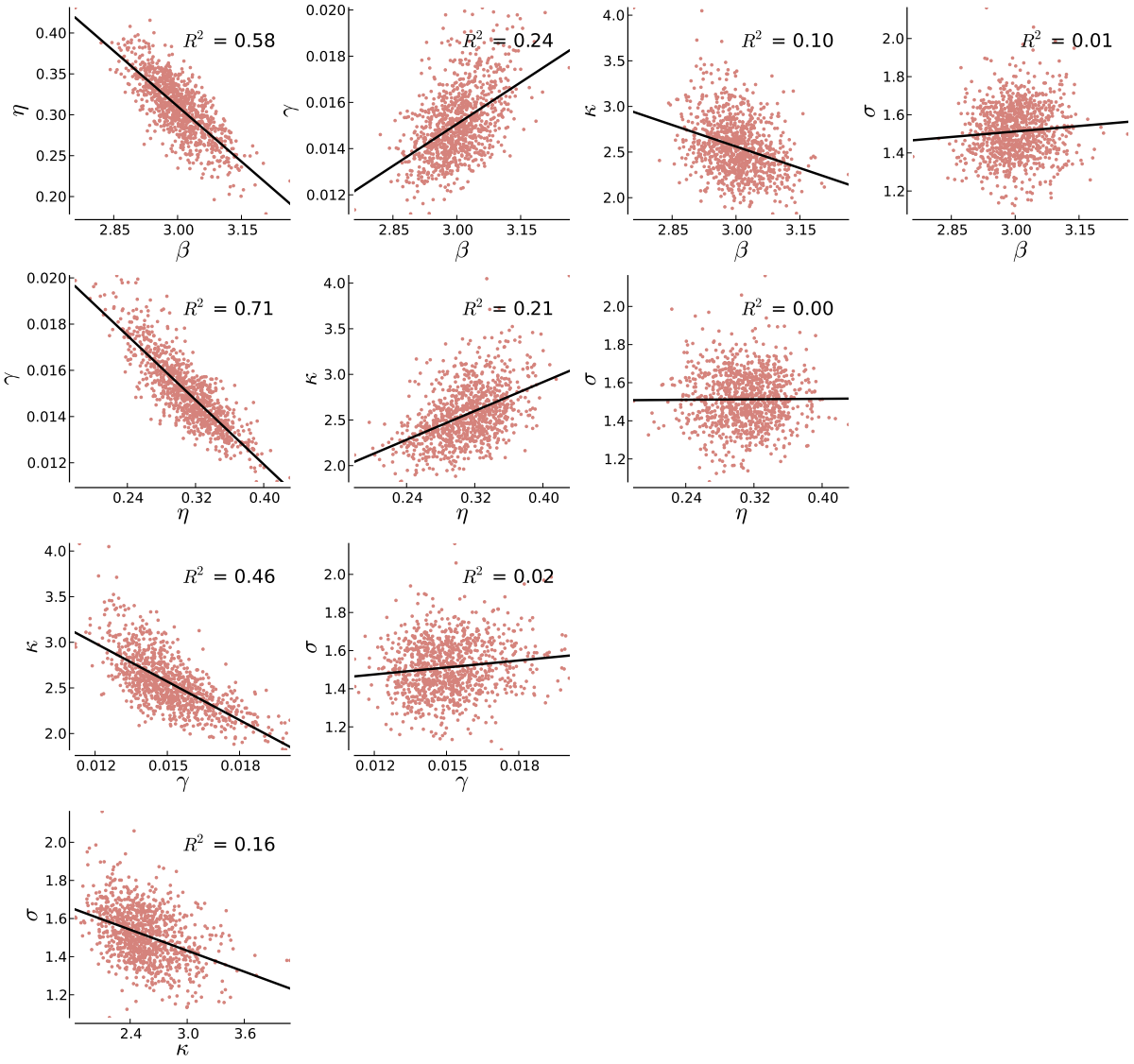
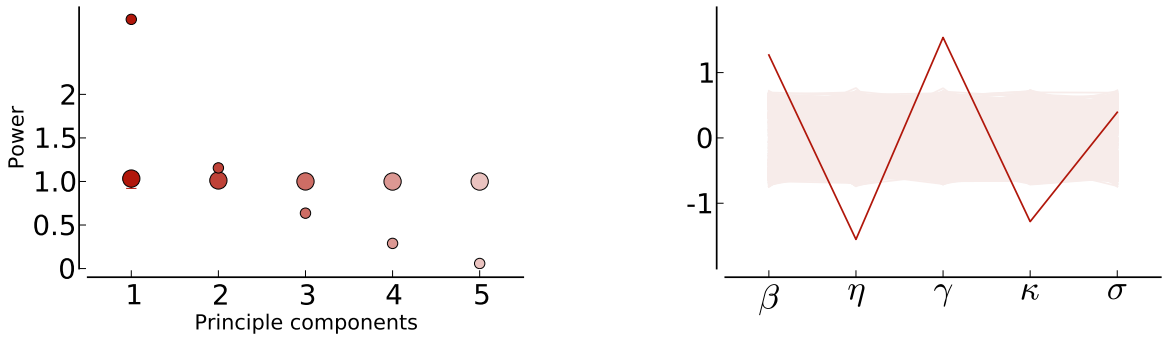


Figure 29: Pairwise correlations obtained with contrast discrimination data (GBH) and model 6

The traces and marginal histograms of three sampling repetitions are presented in Figure 28. The last column shows the histograms of all runs superimposed. Convergence is apparently good and the marginal posterior distributions well sampled.



(a) Power of principle components

(b) First principle component

Figure 30: Principal component of parameter distribution (GBH)

The correlations between parameters remain to be analyzed. Pairwise scatter plots including regression lines and R^2 values are show in Figure 29. The shapes of the distributions and the relative R^2 values are comparable to Figure 25. The description from page 46 for the principal components of observer FAW would fit equally well to Figure 30 of observer GBH. This is how similar the correlation structures are. Here as well, the most prominent observations are the near independence of the link function's shape parameter σ from the other parameters, the strong correlations between the parameters of the nonlinearity and the dependance of the exponent η and the pooling parameter β.

Finally, the well-behaved marginal distributions suggest a comparison of the estimates across observers through confidence intervals. Normally, confidence intervals would provide an insufficient representation of the distributions, because they neglect correlations—a quality they inherit from the marginal. However, since we have already seen that the correlation structure is comparable across observers, confidence intervals are in the current case still a valid comparison. Figure 31 shows parameter estimates and errorbars indicating the 95% confidence intervals. Observer GBH is denoted in blue (right) and observer FAW in red (left). Except for

parameter σ, the analysis does not support a common parameter space across observers. The same generic model fits the experimental data of both observers, but individual differences do not allow a common parameter set. A discussion about concrete parameter values is postponed to the subsequent section within the general discussion.

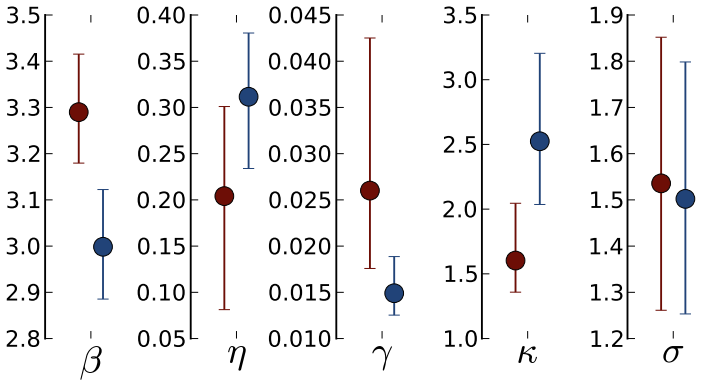


Figure 31: Parameter comparison between estimates for FAW (red) and GBH (blue)

Discussion

This chapter contains two main ingredients: psychophysical data sets and a basic early vision model. Through a Bayesian method the parameters of an early vision model were estimated and their distribution analyzed. In this section I will summarize the chapter's main results and discuss their implications. I will mention limitations of the approach and line out further directions that could be taken to expand on this work.

The last section is characterized through a recycling of processing steps from the description of a data set, the effort to estimate model parameters, success or failure with the parameter estimation, and the restart with further assumptions or data. The early vision model applied stems from the class of linear, nonlinear cascade models with a linear filter stage, a static nonlinearity and a decision function. It is consistent with the literature of different experimental fields. The model parameters were estimated with a Bayesian sampling algorithm, a Markov chain Monte Carlo method called Metropolis-Hastings algorithm. The visual analysis of the estima-

tion procedure included the sampling traces, the marginal posterior distribution and pairwise parameter correlations. These criteria and the goodness-of-fit measured in deviance between model predictions and data were the main criteria used to evaluate the model. The model was fit to data from two different experiments. The first data set was collected with a pure contrast detection paradigm and the second with a contrast discrimination paradigm.

The MCMC chain obtained with the contrast detection data set is bistable and results in a posterior distribution with two clusters. The cluster means hint to an overparametrization of the model. The contrast detection data can be equally well predicted with a model that lacks a static nonlinearity. It seems that detection data is inherently inapt to describe any function that is defined on the contrast dimension. In addition to the estimation results, I derived a theoretical reasoning based on derivatives. The argument postulates that contrast discrimination data collected for several pedestal contrasts has the capacity to constrain the static nonlinearity. A repetition of the estimation procedure confirmed this prognosis. The MCMC chains obtained with contrast discrimination data are sampling a unimodal and constrained posterior distribution. The mean estimates capture the main effects in the data well. A slight misprediction at low signal contrasts could be abolished trough the change from a logistic link function to the cumulative density function of a Weibull distribution. The only remaining unpleasant property of the posterior distribution are rather large correlations between the parameters of the nonlinearity.

I have shown that a strong emphasis on specific parameter values, or model implementation details, requires a very cautious interpretation as long as the posterior distributions are not well constrained. In the case of more complex models the interpretation of a single parameter value as evidence for a structural property remains difficult even with a good model fit and constrained posterior distributions due to the assumptions used for the model. Nonetheless, I will briefly comment on the discussion whether the visual system uses signal energy for contrast detection. For a model with a maximum operator as pooling stage, L^{∞}, and a preceding nonlinearity an exponent of 2 is interpreted as extracting stimulus energy and hence the model is termed energy model. As soon as the maximum pooling stage is replaced through Minkowski pooling the interpretation of the previous nonlinearity becomes difficult

trough the additional Minkowski exponent. In case of the model without nonlinearity—introduced to clarify the deficits of contrast detection experiments—the Minkowski pooling parameter β was close to 2 and may be taken as evidence for energy extraction. In the final model with exponents in the nonlinearity and in the pooling, however, the interpretation is not as clear. One interpretation would treat the nonlinearity with its exponents not as energy detector but as a noise process, then the Minkowski parameter could still be interpreted as energy detector—but with values above 3 (see Figure 31), they are too large for energy detection. The alternative interpretation keeps the nonlinearity to detect stimulus energy and the Minkowski exponent would not reflect stimulus energy anymore but maybe a decision energy so to speak. Then, the values I obtained for κ, the exponent in the nonlinearity, are close to 2 which is consistent with stimulus energy. The Minkowski exponent β has a value of about 3 and is thus larger than decision energy, and larger as the optimal value reasoned from a signal-to-noise perspective[61]. Yet it would be too small to be a maximum operator which is theoretically given with $\beta = \infty$, but practically already with $\beta = 5$. Furthermore, the values referenced here only hold across observers. As can be seen in Figure 31, the values do not match an energy interpretation anymore once individual observers are inspected. Given the different parameter sets across observers and given the dependance of these results on the setting at the initial stages of the model, this discussion may not be meaningful after all. At the very least, however, it shows that the model instances of this chapter resemble the model instances of previous studies and do not operate in a fully different region of the parameter space.

[61] Goris, R. L. T., Wichmann, F. A., & Henning, G. B. (2009). A neurophysiologically plausible population-code model for human contrast discriminations. *Journal Of Vision*, *9*(7), 1–22

The Bayesian estimation approach turned out to be a helpful tool to pin down the grasp of experimental data. It showed that even a contrast detection experiment with eight spatial frequency conditions, evaluated at a total of 68 different signal contrasts, and containing 3940 trials of a single observer is not enough to constrain the nonlinearity. Even if it is the most extensive data set from a single observer that has been used for that purpose. To collect detection data for many more observers and a wider variety of stimulus pattern as was done by the modelfest group[62] would not help either, because it is the wrong kind of experimental paradigm to do so. The solution came only with another 3250 trials spread over 71 signal contrasts and 7 pedestal contrasts from a contrast discrimination experiment. This data set constrained the

[62] They measured detection thresholds of 43 stimuli in 4 repetitions with 16 observers

parameters. However, despite the seemingly large data set and only few parameters, these parameters are highly correlated. The problem is that mechanistic models are not developed to possess good statistical properties which complicates every statistical approach. In general, if a model is too general for the available data, one can either reduce the model complexity to match the data or collect the necessary data.

With respect to the model that suffices to explain the detection data—the model without nonlinearity—, the results are similar to the results of Watson et al.[63] They concluded that the application of an amplitude modifying contrast sensitivity function in the Fourier domain renders an additional channel setup unnecessary. Here, the filter and the scaling of responses within the filter according to the detection threshold of the filter's optimal spatial frequency is the equivalent of the Fourier space implementation from Watson. The channels do not effect the further processing steps. The number of processing steps in my model implementation is smaller than that of Watson and they seem to be enough for our contrast detection data set.

[63] Watson, A. B., & Ahumada, A. J. (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision*, *5*(9), 717–740

Here, the model was constructed with the goal to obtain an image-driven model that remains low in complexity. This is critical, because it may be that instead of analyzing a complex system, one is stuck with analyzing a complex model of a complex system. To understand the internal processes of the model, less is more. As a consequence, the model used here is purposely more minimalistic than other model implementations. In order to increase the model complexity by adding further processing steps, the data that supports those steps would also be needed. Otherwise the model returns to being not constrained and nothing is gained in terms of algorithmic understanding.

The current model is more complex than those of other studies with respect to one detail. Here, it is not only a single threshold for a fixed performance level that is estimated but the full psychometric functions. To do so, a function that links the model output on a decision level to the behavioral response was added. An asymmetric sigmoid captured the data best. Interestingly, the credible intervals of the shape of the cumulative Weibull density function are almost the same for the two observers. Furthermore, the shape parameter possesses the favorable quality to be only hardly correlated with

other parameters. As a consequence, this extra level is justified by the data.

THIS STUDY could be continued in several future directions. For example as a next step the specific assumptions and choices of the model could be further investigated. Alternatively, the model could be applied to further experimental paradigms and if necessary be extended in its functionality. I will therefore first suggest some potential substitutions in the current model and then elaborate on possible model extensions. Last I will reference two methodological approaches that are promising in combination with the procedure advocated here.

SINCE THE NOTION OF THE LINEAR, NONLINEAR CASCADE MODELS IS VAGUE, the freedom in selecting a specific version of the model class is large. Even given a specific version the actual implementation can be debated and questioned. Of course the assumptions and decisions for the current model implementation are not irrevocable. Some mechanisms might even be exchangeable. For example a Garbor has a positive dc component in frequency space. If it were not for the Gabor's spread in the community, a filter with a cosine-shaped frequency content[64] or a steerable pyramid[65] would be theoretically superior. On the basis of psychophysical data it is probably impossible to prefer one filter system over the other. Furthermore, a switch to the other filter systems is unlikely to alter the conclusions from the previous sections.

[64] Kekre, H., Sahasrabudhe, S., & Goyal, N. (1983). Raised cosine function for image restoration. *Signal Processing*, *5*(1), 61 – 73

[65] Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: a flexible architecture for multi-scale derivative computation. *2nd IEEE International Conference on Image Processing, III*, 444–447

HERE, the architecture of the front-end, namely the filter stage, is defined only once with a given set of filters and their characteristics such as optimal frequency, optimal orientation, bandwidth, and functional shape. The filter properties including the total number of filters could be included in the estimation process.[66]

[66] The optimization of these parameters complicate the estimation procedure. One would maybe first apply a maximum likelihood optimization routine followed by a full Bayesian analysis.

THE PREVIOUS ANALYSIS HAS SHOWN that the parameters of the nonlinearity are still highly correlated. Its functional formulation was chosen to be general in order to allow the data to constrain its shape. Now that the shape is constrained, one could reparametrize the function. This would help to facilitate the estimation procedure if the new parametrization shows less correlations. The function might even be reduced to contain only two parameters.

The type of adaptations that concern alternatives for model components are rather trivial and might be approached without further experimental data. More challenging are model extensions since those need to be founded on further experiments. I will shortly discuss the extensions that I consider most promising: adaptable filters, contrast gain control, noise, and an optimal linear decoder as readout mechanism.

Adaptable filters provide additional functional flexibility and could be achieved trough constant volume operators proposed by Cornsweet and Yellot[67]. These operators change the size of the spatial summation area inverse to illumination. Further evidence for such a mechanism comes from experiments regarding the spatial frequency tuning of macaque V1 neurons[68]. They show that spatial frequency tuning is sharpened with decreasing stimulus contrast and an correlated expansion of spatial summation. Furthermore, Sceniak et al show that the reduction of spatial frequency bandwidth is not symmetric around the preferred spatial frequency. An effect that is hard to reconcile with classical Gabor filters that can only be manipulated symmetrically.

[67] Cornsweet, T. N., & Yellott Jr, J. I. (1985). Intensity-dependent spatial summation. *Journal of the Optical Society of America A*, *2*(10), 1769–1786

[68] Sceniak, M. P., Hawken, M. J., & Shapley, R. (2002). Contrast-dependent changes in spatial frequency tuning of macaque V1 neurons: effects of a changing receptive field size. *Journal of Neurophysiology*, *88*, 1363–1373

Contrast gain control is an normalization process. The nonlinearity that is applied after the filter is then not only operating point wise and independent of its neighbors, but the responses of filters with similar properties interact. Lateral inhibition is a physiological example for interactions[69] and divisive inhibition a possible implementation[70]. For divisive inhibition a filter response is divided by the responses from neighboring filters with similar properties. How strong the filters influence each other is given by a weighting function. The parameters of the weighting function are then additional variables to be estimated. Depending on the available data they include spatial frequency tuning width, orientation tuning width, and gain.

[69] Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J., & Davis, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. *Nature*, *378*, 492–496; and Blakemore, C., & Tobin, E. A. (1972). Lateral inhibition between orientation detectors in the cat's visual cortex. *Experimental Brain Research*, *15*, 439–440

[70] Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*, 181–197; and Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature neuroscience*, *4*(8), 819–825

The role of noise in the model could be investigated. There are external noise sources, early sensor noise, and late decision noise. The model formulation from the previous chapters only include noise that becomes affective at the decision stage and is not dependent on signal strength. There is evidence for level dependent noise from neurophysiology.[71] The variance in firing rates increases with increasing mean like a Poisson process. Similar effects have been reported

[71] Dean, A. F. (1981). The variability of discharge of simple cells in the cat striate cortex. *Experimental Brain Research*, *44*, 437–440

in psychophysics.[72] However, most recently it hast been argued that the decision of fixed or variable noise can not be determined on the modeling. Since the transducer function compensates for the noise model[73]. Sensor noise could be independent, but a correlated noise source is likely.[74] The advantage of correlated noise is that the information gain saturates with the number of filters and one could derive a maximal filter number. The difficulty with noise processes is that they are difficult to separate. If the performance limiting, dominant noise source is early in the processing chain, then the subsequent weaker sources or nonlinearities can not be investigated.[75]

[72] Tolhurst, D. J., Movshon, J. A., & Dean, A. F. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research*, *23*(8), 775–785; and Georgeson, M. A., & Meese, T. S. (2006). Fixed or variable noise in contrast discrimination? The jury's still out... *Vision Research*, *46*(25), 4294–4303

[73] García-Pérez, M. A., & Alcalá-Quintana, R. (200). Fixed vs. variable noise in 2AFC contrast discrimination: lessons from psychometric functions. *Spatial Vision*, *22*(4), 273–300

[74] Henning, G. B., Bird, C. M., & Wichmann, F. A. (2002). Contrast discrimination with pulse trains in pink noise. *Journal of the Optical Society of America A*, *19*(7), 1259–1266

[75] so-called Birdshall's theorem

An alternative readout mechanism is the last suggestion as a model extension. The Minkowski pooling rule is flexible and covers a few specific mechanisms used in previous studies. However, it is still a naive rule, that does not use the reliability of the filter or which filters are most informative for the task at hand. Alternatively, a rule could use the signal-to-noise ratio to incorporate the filter reliability[76]. Or an optimal linear decoder could be implemented for a fixed front-end[77]. The disadvantage of the just proposed decision rules is that they need to be retrained for each task and can not be applied out of the box such as Minkowski pooling.

[76] Goris, R. L. T., Wichmann, F. A., & Henning, G. B. (2009). A neurophysiologically plausible population-code model for human contrast discriminations. *Journal Of Vision*, *9*(7), 1–22

[77] Eliasmith, C., & Anderson, C. H. (2003). *Neural Engineering: Computation, representation and dynamics in neurobiological systems.*. MIT Press

Each of these suggestions requires data from experiments that manipulate the right dimensions. More data of the same paradigms would not help. The present case showed that the static nonlinearity could not be constrained through contrast detection experiments despite the many conditions, blocks, and trials. Only the contrast discrimination experiments provided the required restrictiveness. Which experiments were restrictive enough for above suggestions is difficult to tell. However, to approach them with the data presented so far would probably be pointless. Bayesian model analysis could be used to complement the model with an extension. To do so, one would choose an extension and recollect the data that provided evidence for this extension or design a psychophysical task that targets its mechanism. Next, one would predict the data without the extension being implemented and eventually reestimate the parameters. If the model fails to predict the data, an additional mechanism with the respective functionality should be implemented and the estimation procedure repeated. Along that way the predictions and parameter posterior distributions provide control and guidance, and might lead to further insights.

However, it needs to be said, that for a vision model operating in image space with many pixels, the computation times were already large with only 4 or 5 parameters that needed to be estimated. It took between 5 and 6 hours to compute the full MCMC chain with 1.010.000 samples on a *Six-Core AMD Opteron(tm) Processor* 8431. This large sample size was necessary because the samples were highly correlated and the acceptance rate rather low with only 5% to 10%. Such an highly inefficient procedure could be avoided if the samples could have been drawn in a less correlated fashion. The correlated structure of the parameter space is the main problem here. To improve sampling efficiency this correlation structure would need to be reflected in the proposal distribution as already briefly mentioned above. Unfortunately, the hyperparameter tuning is time consuming, not standardized and depends on the experience and efforts of the modeler. This is one major disadvantage of MCMC sampling in my opinion. The simplest shortcut to reduce sampling times is to reduce the operations computed in the model. Some costly operations can not be avoided but for example the resolution and extent of the image could be reduced. If the number of final values as input to the filter stage is reduced from 256×256 to 64×64 the computation times are smaller than expected because array operations in python apparently do not scale linearly with array size[78]. The problem with input reduction is to find the limit for which the input is still capturing the necessary structures. Another chicken and egg problem if the necessary structures are not yet known. Therefore, to be conservative the results presented here are obtained with the same resolution as the stimulus presented to the observer during experiments. The case of time limiting computations becomes even more severe the larger the number of parameters to be estimated.

[78] python2.6 on debian

If the model grows in complexity and if a promising region in parameter space to start the sampling is missing, then another method could be used first to determine an initial model state. For example Pinto et al. proposed an approach which they compare to high-throughput screening in other fields.[79] They created instances of models from a model class by chance and computed model predictions. The model that delivered the best results was isolated. This new procedure is promising to contrast model alternatives and reject possible, but unlikely, model variants. A detailed model identification analysis can be performed on the apparently optimal model structure.

[79] Pinto, N., Doukhan, D., DiCarlo, J. J., & Cox, D. D. (2009). A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Computational Biology*, *5*(11), 1–12

THE ULTIMATE CHALLENGE for any model is to make predictions that can be verified experimentally. That is to generalize from the training samples to hitherto unseen data. This approach is similar to cross validation in machine learning. To apply classical cross validation a part of the already available data is reserved for testing purposes. But here the implications can be even stronger since the model serves as hypothesis generator and predicts data that would not have been tested before hand. This test is very strong if the model predictions of the model to be tested differ from the current view or predictions of similar models.[80]

MODELING THE EARLY VISUAL SYSTEM has taken different routes from biological plausibility, to behavioral plausibility to theoretical considerations. For example Goris et al.[81] designed a model that is inspired by many findings from neurophysiology. It is setup as realistic as possible including noise, correlations, a large number of neuronal mechanisms, gain control, and so forth. However, such a detailed model becomes difficult to analyze and it stays unclear how the internal processes interact. The other extreme is to study the purely mathematical and statistical processes that could build the foundation of vision.[82] For decades the belief in Fourier-like vision was so strong, that it dominated many less axiomatic approaches to vision like the Gestalt school of vision. For the abstraction from individual findings and experiments to general processes, many kind of experiments and methodologies should be investigated at once in meta studies. Here a small step in that direction was taken by using a generally accepted model, data sets from different experiments, and combining them with the powerful tools of Bayesian model analysis. Because in the long run, a theory for vision is needed— say the natural laws of seeing.

[80] Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*(3), 499–518

[81] Goris, R. L. T., Wichmann, F. A., & Henning, G. B. (2009). A neurophysiologically plausible population-code model for human contrast discriminations. *Journal Of Vision*, *9*(7), 1–22

[82] Sinz, F., & Bethge, M. (2008). The conjoint effect of divisive normalization and orientation selectivity on redundancy reduction in natural images. In *Advances in neural information processing systems 21, Twenty-Second Annual Conference on Neural Information Processing Systems*, (pp. 1521–1528)

Joint Bayesian estimation of several psychometric functions

In the previous section I presented a full-grown vision model. The model architecture developed over decades through many iterations of experimental findings and model adjustments by various researchers. These findings allowed to specify the functions that span different experimental conditions. Now assume that a new series of experiments was conducted and the experimental manipulation across conditions targets sensitivity changes. The function that relates the sensitivity of the conditions is not yet known. In this case we can not just design a stage of the vision model with the new relationship. First, we have to explore the experimental data. In order to explore sensitivity changes, it is state-of-the-art to fit a psychometric function—definition following below—to the data of each condition. Then, differences between the psychometric functions are compared. Sensitivity, expressed as threshold relations, depends on the performance level at which the threshold is estimated. This complicates the analysis. If the slope were equal across conditions, however, reporting of a single performance level suffices. We would therefore like to estimate all psychometric functions with a common slope if the data allows to do so.

In the simplest case, all conditions could be handled within the framework of generalized linear models[83] which provides numerically efficient ways of estimation and has well established procedures to check for goodness-of-fit. Estimation of generalized linear models is also easy for multiple dependent variables and thus for multiple conditions. For generalized linear models to be applicable, the dependent variable, a probability, must take the lowest plausible value of zero and the highest plausible value will be

[83] McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. Boca Raton: Chapman & Hall/CRC, 2 ed; and Dobson, A. J., & Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. Boca Raton: Chapman & Hall/CRC, 3 ed

one. In the context of behavioral studies however, observers might have a certain probability to guess correctly, even if the stimulus was too weak to be detected by the eye[84] or they miss a percentage of short stimuli because of eye blinks. Then, the lowest plausible value for the dependent variable can actually be larger than zero and the highest plausible value smaller than one. If the shifted asymptotes are not estimated, this might result in estimation biases for the actual parameters of the psychometric function. Including these parameters renders the likelihood function of the model non-concave and in many cases multimodal—a fact that seriously complicates model estimation. Consequently, software that can be used to perform inference in such models typically employs methods for global optimization such as grid searches or Monte Carlo procedures.[85] These fitting routines are designed for, and work well, with the estimation of a single condition. Extending these models to higher dimensions is not trivial and the parameter space that needs to be searched by these global optimization routines grows exponentially with the number of added conditions. Numerical stability and/or efficiency might be sacrificed. Furthermore, goodness-of-fit becomes more difficult to judge as well as other routines, e.g. determining influential observations in the data set[86]. From a practical point of view it can be said that fitting a psychometric function for individual cases is a standard routine for a psychophysicist. A deviation from standards needs to be justified since the standard was tailored to the problem by the needs of the field. Our goal was therefore the ability to handle several data sets simultaneously by extending the common routines, and not by changing to a more powerful methodology.

[84] Treutwein, B., & Strasburger, H. (1999). Fitting the psychometric function. *Perception & Psychophysics*, *61*(1), 87–106

[85] Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313

[86] Fründ, I., Haenel, N. V., & Wichmann, F. A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision*, *11*(6(16)), 1–19

Instead of fitting all conditions in a common model,[87] we suggest an alternative approach that still models each condition individually. This requires that information for all other conditions is incorporated into the inference of each of these isolated simple dose-rate models. Bayesian statistics allows for a very natural way to include external information into the inference process. In Bayesian statistics, the external information is typically incorporated in the form of a "prior" probability distribution because it describes all the information available to an experimenter before he or she has seen the data that are actually analyzed. Here we propose a method to derive prior distributions that integrate information from other experimental conditions and pose an implicit constraint to force a desired parameter to be (close to) equal across

[87] De Lean, A., Munson, I. J., & Rodbard, D. (1978). Simultaneous analysis of families of sigmoidal curves: application to bioassay, radioligand assay, and physiological dose-response curves. *American Journal of Physiology*, *235*(2), E97–E102

conditions. The subsequent sections first introduce the method, apply it on an example and subsequently evaluate the method.

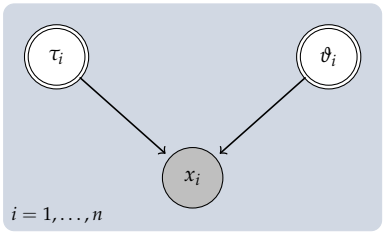
Separate sampling for joint inference

IN ORDER TO EXPLAIN THE IDEAS behind joint inference, we imagine that n data sets were collected experimentally, one data set per experimental condition. Each data set, $x_i, i = 1, \ldots, n$, can be described by the same model M with parameter vector $\theta = (\vartheta, \tau)$, but the specific parameter values of M might differ. Note that ϑ and τ can be scalar as well as multidimensional. The standard analysis treats each data set individually; each condition is analyzed separately. We will refer to this collection of fitted models as the *isolated models*. The graphical model of M is depicted in Figure 32a. The observed variables, $x_i, i = 1, \ldots, n$, are shown by the filled node. They represent the data sets collected for n different conditions. The random variables $\tau_i, \vartheta_i, i = 1, \ldots, n$ are random variables that are drawn as double circles. The plate in the background groups variables that belong to condition i only. Each node represents a factor in the joint distribution of model variables. This factor is the conditional distribution of the variable conditioned on the variables from which it receives incoming arrows. For parameter posterior distributions of model M follows that

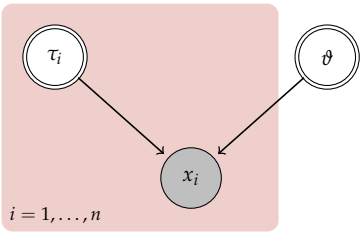
$$p(\vartheta_i, \tau_i | x_i) = p(\vartheta_i | x_i) p(\tau_i | x_i), \quad i = 1, \ldots, n. \tag{6}$$

This equation is used to estimate each condition in isolation.

THE GOAL OF JOINT INFERENCE however is to fit all conditions simultaneously, because the experimenter suspected that one parameter, say ϑ, is shared across conditions. The graphical model for the joint analysis is shown in Figure 32b and we will refer to this model as the joint model. Such a situation can arise, for example, if the system described by the model has some parameters that are dependent on the state of the system, here $\tau_i, i = 1, \ldots, n$, and some that are state independent, ϑ. We will show next how the computation of the isolated models in a first step can serve to fit the full model. The main assumption for the joint inference procedure is that the parameters in θ are a posteriori independent. The assumption's appropriateness is further investigated in the next section. Given the assumed a posteriori independence of the model



(a) The isolated models



(b) The joint model

Figure 32: Graphical models illustrating the different approaches.

parameters, we can write the joint parameter in the joint model via the marginal. The method is illustrated with $n = 2$ data sets.

$$
\begin{aligned}
p(\vartheta|x_1, x_2) &\propto \int p(\tau_1, \tau_2, \vartheta|x_1, x_2) d\tau_1 d\tau_2 && (7)\\
&\propto \int p(x_1, x_2|\tau_1, \tau_2, \vartheta) p(\tau_1, \tau_2, \vartheta) d\tau_1 d\tau_2 && (8)\\
&= \int p(x_1|\tau_1, \vartheta) p(x_2|\tau_2, \vartheta) p(\tau_1) p(\tau_2) p(\vartheta) d\tau_1 d\tau_2 && (9)\\
&= \int p(x_1|\tau_1, \vartheta) p(\tau_1) p(x_2|\tau_2, \vartheta) p(\tau_2) p(\vartheta) d\tau_1 d\tau_2 && (10)\\
&\propto \int \overbrace{p(x_1|\tau_1, \vartheta)}^{\text{likelihood}} \overbrace{p(\tau_1) p(\tau_2, \vartheta|x_2)}^{\text{priors}} d\tau_1 d\tau_2. && (11)
\end{aligned}
$$

Here, we used Bayes Theorem, the a priori independence of model parameters, and finally reorganized the terms to arrive at an expression that we will use next to sample from the posterior. This expression suggests a reinterpretation in the form of likelihood and a prior terms: The "likelihood" only contains the first data set x_1. The second data set x_2 appears in one of the prior terms. The term that works as a prior for ϑ resembles the posterior of the isolated model applied to the second data set (equation (6)). If the joint model is a correct description of the data, then the shared parameter ϑ equals the parameters of the isolated models and $\vartheta = \vartheta_1 = \vartheta_2$. The previous equation can therefore be rewritten by replacing ϑ by ϑ_1 and ϑ_2 to arrive at,

$$
p(\vartheta|x_1, x_2) \propto \int p(x_1|\tau_1, \vartheta_1) p(\tau_1) \overbrace{p(\tau_2, \vartheta_2|x_2)}^{\text{posterior from isolated model}} d\tau_1 d\tau_2. \qquad (12)
$$

As a result the posterior of ϑ in the full model that is based on all data sets simultaneously can be estimated in a two step procedure. The first step is to determine the posterior of the isolated model on the second data set $p(\tau_2, \vartheta_2|x_2)$ and determine its marginal $p(\vartheta_2|x_2)$ from (6). In the second step we estimate the parameters of the isolated model from the first data set $p(\tau_1, \vartheta_1|x_1)$ using the marginal $p(\vartheta_2|x_2)$ as a prior for ϑ_1. Thus, it is possible to reduce the formulation of the full model to a sequence of isolated models.

The terms in equation (10) may be ordered differently. This way, we can apply the isolated models in a different sequence.

$$
p(\vartheta|x_1, x_2) \propto \int \overbrace{\underbrace{p(\tau_1, \vartheta_1|x_1)}_{\text{priors}} p(\tau_2) \underbrace{p(x_2|\tau_2, \vartheta_2)}_{\text{likelihood}}}^{\text{posterior from isolated model}} d\tau_1 d\tau_2 \qquad (13)
$$

Theoretically, the order should not matter and the marginal posterior distributions computed with different orders are equivalent. This can be used as a sanity check. If the full model is an adequate description of the data and ϑ is shared across conditions, the marginal posteriors of the parameters obtained in different orders should overlap.

To SUMMARIZE THE APPROACH for n data sets: In a first step the parameter posterior distributions for each of the n conditions are determined in isolation. The marginal posterior distributions of the shared parameter from $n-1$ conditions are multiplied and this product is the prior for a second round of inference on the n^{th} condition. In contrast to the first step, the second step introduces information from all other conditions into the inference procedure. This way, the second step of inference implicitly performs inference on all conditions simultaneously. In the next section, we will illustrate the strategy for a concrete example from perceptual psychology.

Example from perceptual psychology: the psychometric function

The psychometric function relates the performance of an observer to the intensity of a stimulus. Here, intensity can be the sound pressure of an auditory tone or the contrast of a visual stimulus. Performance is typically expressed in terms of the probability that the observer correctly detects a predefined target stimulus.

WE ANALYZE PSYCHOMETRIC FUNCTION DATA from a single observer in an experiment by Felix Wichmann[88]: The observer performed a two-alternative forced choice task in which he had to monitor two time intervals. Each interval lasted 79 ms. In one of these two intervals, a weak sinusoidal target grating with a spatial frequency of 8.37 degree visual angle was presented. The observers task was to identify which one of these two intervals contained the target grating. At each contrast level of the target grating, either 40 or 50 responses were collected. Performance was measured as the fraction of trials in which the observer correctly identified the interval that contained the target grating. We analyze data that were collected in two different experimental conditions: First, a "masking" grating of low contrast (Michelson contrast[89] of 1.6%) was presented in both intervals. The mask was in phase with the target grating such that the task was essentially to identify the interval in which the grating had higher contrast. We will

[88] Wichmann, F. A. (1999). *Some aspects of modelling human spatial vision: Contrast discrimination*. Unpublished doctoral dissertation, The University of Oxford, Oxford, UK

[89] If L_{max} and L_{min} are the maximal and minimal luminance of the stimulus, then the Michelson contrast is $\frac{L_{\text{max}}-L_{\text{min}}}{L_{\text{max}}+L_{\text{min}}}$.

refer to this condition as the "low contrast mask" condition. In the second condition, the mask had a high contrast (Michelson contrast of 51.2%). We will refer to this condition as the "high contrast mask" condition. Wichmann reports no strong changes in slope for different masking contrasts.

In order to model these data, the responses were assumed to be binomially distributed with aprobability of success given by [90]

$$\Psi(x) = \gamma + \frac{1-\gamma-\lambda}{1+\exp(-z_0(x-\alpha)/\beta)}, \quad z_0 = \log(9), \quad \gamma := \frac{1}{2}.$$

This model has three free parameters α, β, λ. The parameter λ describes the upper asymptote of the model and is treated as a nuisance parameter. Although λ is usually not of scientific interest, omitting λ from the model introduces potential estimation bias in the other parameters.[91] The remaining two parameters α and β are psychologically interesting: α is the stimulus intensity at which the psychometric function is halfway between the lower asymptote (γ) and the upper asymptote $1-\lambda$. Thus, α is often reported as the *threshold* and $1/\alpha$ can be considered a measure of sensitivity. The other parameter of interest is β, which is related to the slope of the psychometric function: If β is large, the psychometric function is very shallow, if β is small, the psychometric function is very steep. By incorporating the constant z_0 into the equation, β gives the range of stimulus intensities on which the psychometric function rises from a performance level 10% above the lower asymptote to 10% below the upper asymptote. The percentages refer to the compressed range of performances spanned by the psychometric function, that is the range of performances between the lower asymptote and the upper asymptote. The model is first fitted to each condition individually. The procedure requires proper prior distributions for all model parameters. We chose a Gaussian distribution for the threshold parameter α and a Beta distribution for the asymptote λ. The inverse Gamma distribution is a commonly used conjugate prior for scale parameters. Therefore, we chose an inverse Gamma distribution as prior for the slope parameter. The resulting marginal posterior distributions of the parameters are again well described by a Gaussian, a Beta, and a Gamma distribution. In the next step of the joint procedure the marginal posterior distributions of several conditions are multiplied. In the appendix on *Products of exponential family distributions* we derive these products that serve as priors in the second and final inference step.

[90] Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313; Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision*, *5*, 478–492; and Fründ, I., Haenel, N. V., & Wichmann, F. A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision*, *11*(6(16)), 1–19

[91] Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313

FIGURE 33 ILLUSTRATES the results of the analysis. Figure 33A and 33B display the data (dots) as well as the Bayesian posterior mean estimate of the corresponding psychometric function. The functions provide visually good fits although the high contrast mask (Figure 33B) data scatter slightly more around the fitted function than in the low contrast mask (Figure 33A) condition. The deviance residuals plotted below the curve capture this observation well. Figure 33C shows marginal histograms of samples from the posterior determined using MCMC. We tried to summarize the samples by fitting them with a parametric model. The solid lines in the second row are maximum likelihood fits of Gamma distributions to the samples from the marginal posterior distribution. We observe that the histogram for the "low contrast mask" condition (Figure 33C light color) is very similar to the histogram for the "high contrast mask" condition (Figure 33C darker color). Furthermore, the histograms are very well approximated by a fitted Gamma distributions. This suggests that the fitted Gamma distributions can be taken as parametric summaries of the posterior samples. With this prerequisite, a simultaneous fit of the psychometric functions in the two masking conditions might succeed. Indeed after applying the procedure presented in the previous section, the fits remain very good. The joint mean a posteriori fits in Figure 33D and E fit the data nearly as well as for separate inference. Note that even the residual plots in the bottom part of Figure 33D and 33E are very similar to those for isolated fits (Figure 33A and 33B). Also, the a posteriori histograms in Figure 33F are highly overlapping. It should however be noted that neither the histograms nor the fitted parametric summaries are exactly the same. Once the experimenter has decided to accept the simultaneous inference to provide valid results, the posterior samples stem from the same distribution, the joint posterior distribution. In that case, differences between histograms only reflect the sampling errors during posterior sampling. We will now evaluate this approach.

Evaluation of the method

THE PREVIOUS SECTION illustrated the method through an example from perceptual psychology. In this section we will use the same model as in the example to evaluate the method with respect to two questions that concern the general applicability. First, by definition our method requires that the posterior parameter distributions

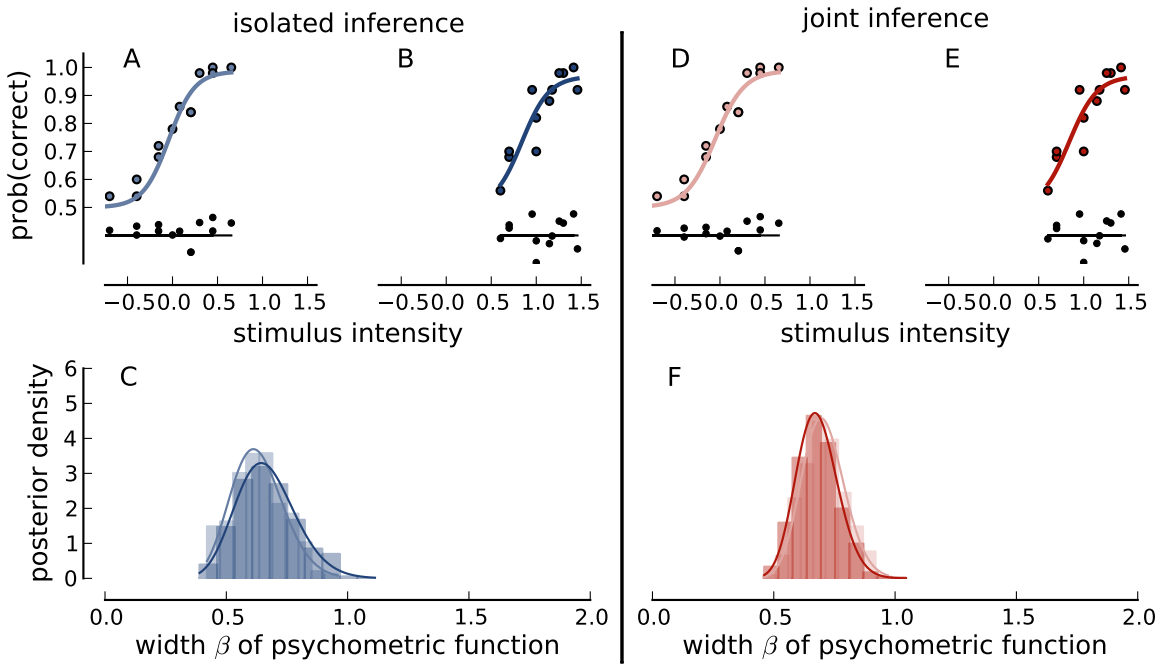


Figure 33: Data from different experimental conditions, psychometric functions, deviance residuals and marginal posterior distributions that suggest parameter equality.

after the first sampling round can be represented by its marginal distributions without loss of information. This is only true if the parameters are a posterior independent. Here, we investigate how crucial this independence assumption is. Second, we study the success of the method. Success means to achieve an overlap between the marginal posteriors without impairing deviance. We address both questions by simulating data from functions with known parameters and applying the method on pairs of these synthetic data sets.

The synthetic data sets used in this section were all generated from the same underlying psychometric function ($\alpha = 2$, $\beta = 1$). That means, by design the joint inference procedure is legitimate to use. The data sets differ with respect to their sampling scheme—the intensities at which the psychometric function is evaluated—and the number of responses per stimulus intensity (trials). Both were chosen randomly for each data set. The number of trials per intensity block ranged between 20 and 200. Six intensity levels were chosen randomly to sample the psychometric function. It was

assured that they covered certain intervals in the asymptotes and rising part of the psychometric function. Thereby, the data sets differ in the amounts of correlation between the parameters. We observed that properly sampled psychometric functions[92] exhibit only minor correlations between parameters. Thus, the assumption will typically be justified in practice. We quantified how well the procedure works dependent on the correlation between α and β. This seems sufficient, since λ is only a nuisance parameter. For the quantification we chose two statistics, one that captures goodness-of-fit and one that captures the overlap between the posterior distributions of the joint parameter. The procedure works well if the goodness-of-fit is nearly the same between isolated and simultaneous fits, and if the overlap between the posterior distributions increases.

[92] Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313

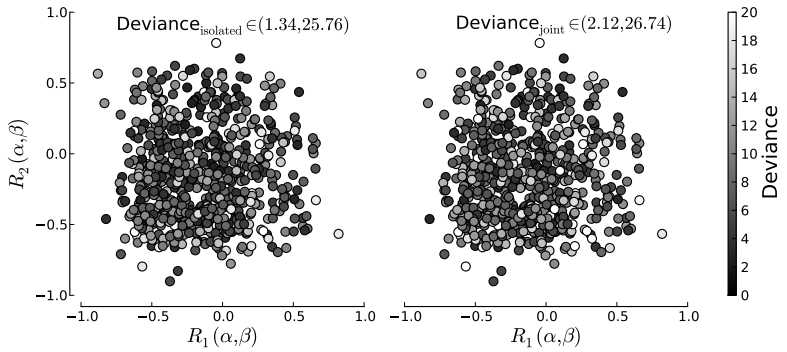


Figure 34: Deviance as a function of parameter correlations for the isolated models(left) and joint model(right).

Goodness-of-fit of a single condition was quantified by deviance:

$$D = -2 \sum_{i=1}^{6} k_i \log\left(\frac{p_i}{y_i}\right) + (n_i - k_i) \log\left(\frac{1-p_i}{1-y_i}\right)$$

with the number of intensity levels, i, the number of trials, n, the number of correct responses, k, the model prediction, p, and the observed performance, y. To compare the fit of the isolated models with the fit of the full model, the deviance of the n isolated models are just summed: $\mathbf{D} = \sum_{i=1}^{n} D_i$. The full model is fitted n times. Each fit with a different data set in the likelihood term and subsequently all deviances are also summed up. Figure 34 shows goodness-of-fit as a function of the correlation between the two parameters. The first panel presents deviance sum $\mathbf{D}$ of both data

sets fitted in isolation. The data is plotted against the correlation value R_1 of the first data set and the correlation R_2 of the second data set. The better the fit, the smaller deviance and the darker the color. The second panel presents **D** for the same data sets, but after the simultaneous fitting procedure. The color pattern of Figure 34 shows that, first, there is no trend with correlation and, second, deviance remains nearly unchanged for different correlations between the parameters α and β. For some combinations, deviance becomes slightly worse.

The overlap between two distributions is quantified with a statistic based on the first and third quartile of the distributions. We prefer this statistic over other options such as KL-divergence, because it is a simple and robust measure with respect to the mass of the sample distribution. The exact shape of the distributions and the tails are not that important. Figure 36 illustrates the statistic. Let the quartiles be Q_1 and Q_3 for one distribution and Q_1' and Q_3' for the second distribution. The overlap is computed by:

$$q := \frac{\min(Q_3, Q_3') - \max(Q_1, Q_1')}{\max(Q_3, Q_3') - \min(Q_1, Q_1')}.$$

This means, that if the distributions are very similar and the quartiles fall on the same values, then the overlap is 1 (Figure 36A). If the interquartile ranges overlap partly, the result is positive. The overlap is 0, if one interquartile range starts where the other ends (Figure 36B) and grows negative with the limit of -1 if the distributions diverge (see Figure 36C for an extreme example).

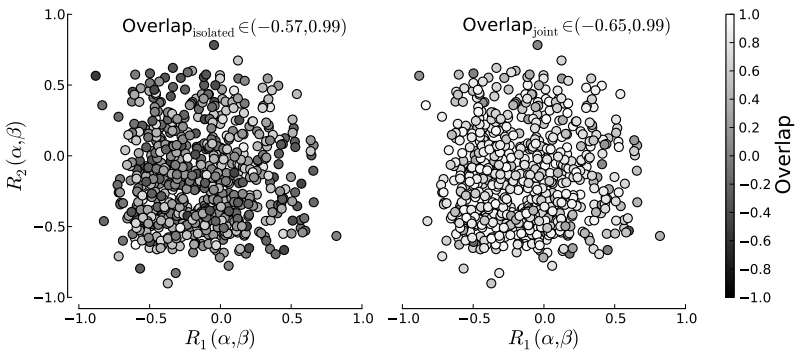


Figure 35: Overlap as a function of parameter correlations for the isolated models(left) and joint model(right).

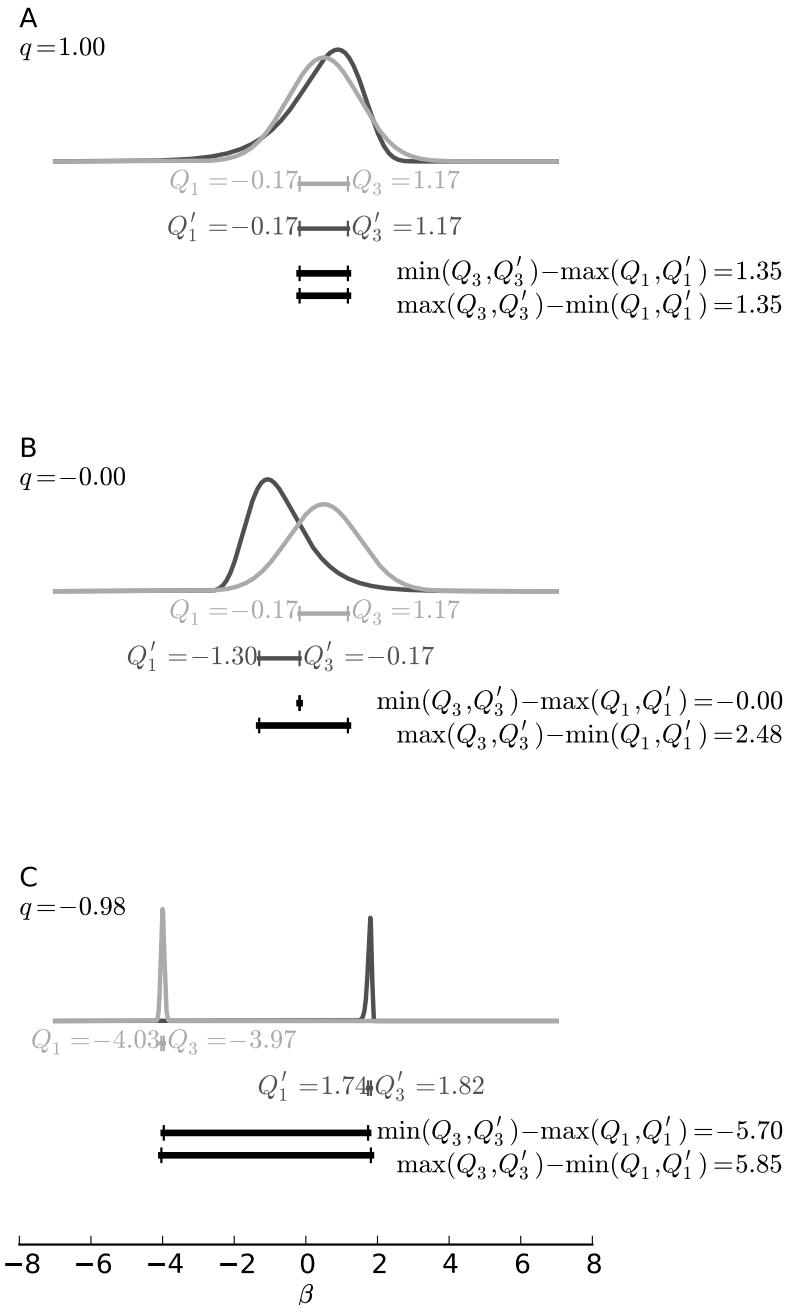


Figure 36: Overlap statistic explained

Figure 35 presents the overlap q between the posterior distribution of the width parameter β as a function of the correlation structure of the data sets. Again the results are shown for both: isolated and joint fits. The lighter the color, the greater the overlap. As for deviance the figure shows no trend of the overlap dependent on the correlation between α and β. We would like to point out, that the initial overlap between the marginal posterior distributions is rather low, even if—as in this example—the generating functions of the data sets were the same. This is due to the rather large variance of binomially distributed data, especially for few trials. With the simultaneous fitting the overlap increases strongly and results in mainly positive indices.

The simulations plotted in Figure 34 and Figure 35 showed that neither deviance nor overlap are sensitive to the assumption of parameter independence. This allows us to summarize the results across correlation and present them as histograms in Figure 37. The first panel shows histograms of deviance as obtained with isolated and joint fitting in middle blue and dark red, respectively. The second panel shows histograms of overlap also for isolated and joint fitting using the same color code. The deviance histograms are very similar while the overlap histogram shifts clearly towards larger values for the joint fit. In combination, the results presented so far suggest that the method is robust in the case of data from the same underlying function.

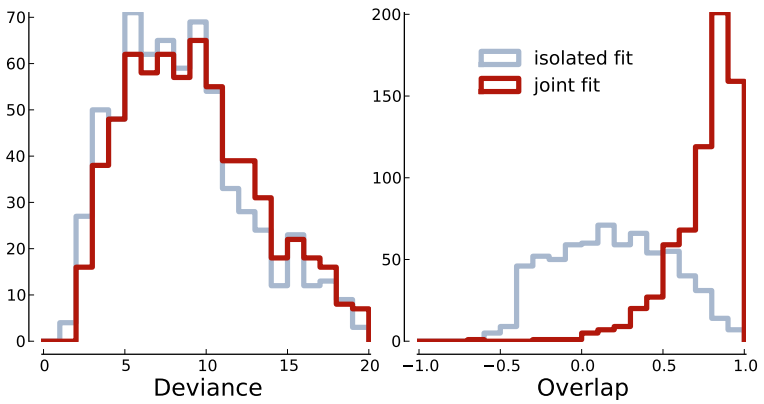


Figure 37: Histograms of the data shown in Figure 34 and Figure 35

Another example and statistical tests

FOR THE EXAMPLE presented on page 68 it was reasonable that the parameter, on which the method was applied, did not differ between data sets collected in different experimental conditions. For the simulated data it was even guaranteed by design. We have shown that in this case the joint fitting procedure resulted in model fits that were as good as the benchmarks obtained in isolated fits with the additional gain of approximate equality of one of the parameter posterior distributions. Consistent with these results, Wichmann[93] reports no strong changes in the width of the psychometric function for different masking contrasts. However, if observers had to discriminate a target grating from a homogeneous background—the "no mask" condition—he reports a decrease in width. This section presents data and statistical analysis of a scenario in which parameter equality is not guaranteed.

[93] Wichmann, F. A. (1999). *Some aspects of modelling human spatial vision: Contrast discrimination*. Unpublished doctoral dissertation, The University of Oxford, Oxford, UK

AGAIN WE ANALYZE DATA that were collected in two different experimental conditions: We reuse the data previously termed the "low contrast mask" condition and add the "no mask" condition. Indeed, the psychometric function in the no mask condition (Figure 38A) is slightly steeper than in the two masking conditions (for example "low contrast mask condition" in Figure 38B). Also the histograms (Figure 38C) are quite different for the no mask condition as compared to the low contrast mask conditions. In general, β tends to be lower in the "no mask" condition. This reflects the previous result by Wichmann on the same data that psychometric function slopes were markedly different if a mask was present or not. It is clear that these two marginal distributions are considerably different. There is only little overlap between these two distributions. Nonetheless, we can use the method to force the two posteriors to be (approximately) equal. The psychometric functions corresponding to the resulting mean a posteriori estimates are shown in Figure 38D and E and the respective a posteriori distributions in Figure 38F. The procedure results in posteriors that are closer to each other. However, the fit quality is much worse than for the two separate fits. This can be seen by comparison of the psychometric functions: In the no mask conditions (Figure 38D), the fitted function is consistently above the recorded data points at low signal contrast, while in the low contrast mask condition the fitted function is consistently below the recorded data points (Figure 38E). In such a case joint fit-

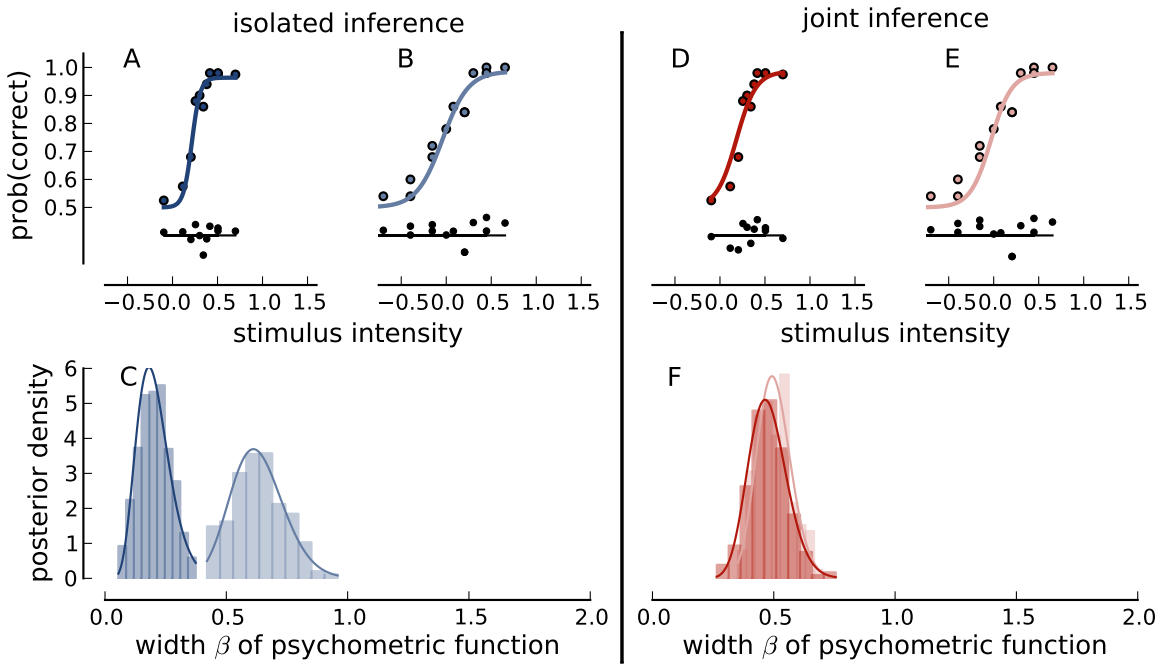


Figure 38: Data from different experimental conditions, psychometric functions, deviance residuals, and marginal posterior distributions with markedly different parameter distributions.

ting should not be applied and we will next describe a quantitative measure which evaluates the appropriateness of the joint model.

We use model selection to decide whether the joint model or the isolated models provide a better description of the data. Again artificial data sets were generated. In the previous simulations we observed that large correlations only occurred if the psychometric function were not well constrained by the data points. For example, if no data was collected in the raising part or in one of the asymptotes. Realistic sampling schemes, as one would demand for meaningful experimental data, do not yield to large parameter correlations. We took advantage of that observation and selected only data sets with correlations of less than ± 0.5. In contrast to the previous simulations, here the simulated data sets differed with respect to the widths of their generating psychometric functions. We studied the appropriateness of the two model alternatives by treating the model itself as another parameter and determining the marginal posterior distribution of this parameter. To do so, we selected a pair of simulated data sets and derived samples for

the isolated model as well as the joint model. In the next step, we considered the posterior distribution the common space of models and parameters. In the appendix on page 111, we show how the marginal model distribution in this space can be obtained through Gibbs sampling and that it is even possible to approximate the stationary distribution analytically. The models posterior probabilities obtained with data sets having no difference between the width of the generating functions, then we expect the model posterior probability of the joint model to be at least equal to the model posterior probability of the isolated models. The joint model could even be favored because it is simpler. Simplicity in this context is expressed in the area covered by the prior distributions of all parameters together. If the functions that generated the data sets had truly different slopes, we would like our method to prefer the isolated model. Of course, it is impossible to discriminate "equal slopes" from "very similar but not equal slopes" on finite data sets. Thus, if the width (or equivalently slope) difference between the generating psychometric functions of two data sets is sufficiently small, we would like our method to consistently prefer the joint model.

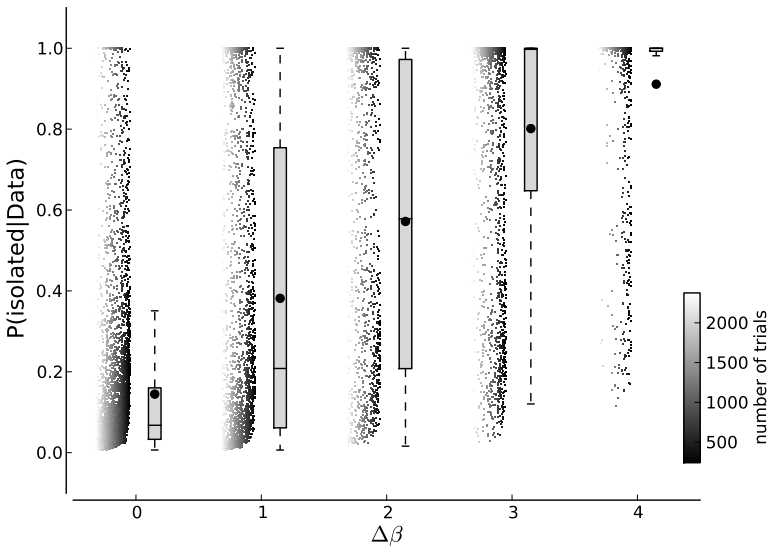


Figure 39: Model selection between isolated and joint model

FIGURE 39 shows boxplots and the mean of the isolated models' posterior probability depending on the true width difference between the generating parameter, $\Delta\beta$. The scattered values are the raw

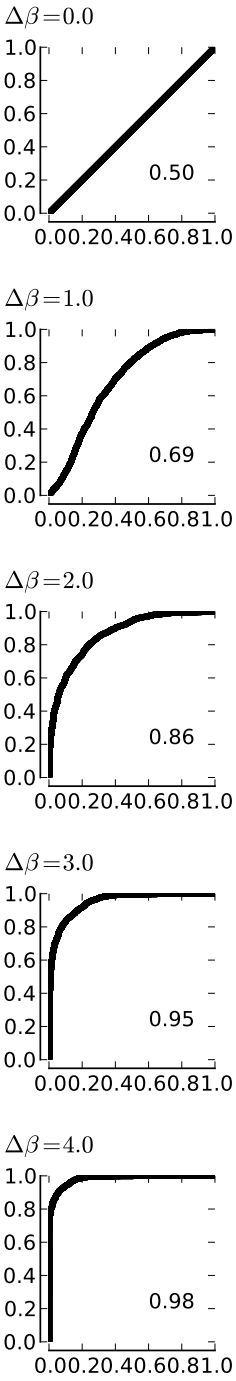


Figure 40: Receiver operating characteristics plotting Hits versus False Alarms

results colored by the number of trials in the data sets. Applied on our simulations we find that the isolated models' posterior probabilities accumulate below values of 0.2 if $\Delta\beta = 0$. With increasing $\Delta\beta$ the main support of the box plot and individual results in the scatter shift towards 1. The probability of the isolated models increases with $\Delta\beta$ as is expected. We had a closer look on the simulations where the data sets were generated with very different slopes but where the model posterior favored a joint analysis. Many of those data sets contained samples that did not describe the psychometric function well. Either these data sets were lacking samples in the raising part or samples in one of the asymptotes. In a real scientific experiment psychometric functions with this property would not be tolerated. The consequence of such data is that the prior from the other condition has no conflicting data and a joint fit is feasible. The strong scatter of the model posteriors therefore stems from the limited number of data samples—here six—combined with an unfortunate positioning of intensity values. The shift of the distributions from 0 to 1 with increasing width is slower than it would be with realistically sampled psychometric functions.

Of course, it is important to show that the model comparison works as expected. However, as a scientist one is interested in the separability between the simulations that allow the joint procedure and the simulations that do not. Therefore, we also computed the "area under the curve" (AUC)— a measure for linear separability between two distributions analyzed in a receiver operating characteristic (ROC). Figure 40 shows the characteristic computed for each pair of $\Delta\beta = 0$ and $\Delta\beta > 0$. A single receiver operating characteristic plots hits, $\mathcal{P}(P_{joint}|\Delta\beta = 0) < crit$, dependent on false alarms, $\mathcal{P}(P_{joint}|\Delta\beta = x) < crit$. The value of the AUC is shown inside the ROC. A value of 0.5 indicates that the distributions are completely overlapping and separability is impossible. A value of 1 indicates perfectly separable distribution. The AUC increases quickly with $\Delta\beta$. The observed AUC values as well underestimate the power of the procedure for better defined psychometric functions.

Coming back to the examples, Bayesian model selection as described above gives the following results. In the first example with the low and high contrast mask the slopes appeared equal. Here, the posterior probability should favor the joint model and indeed, the posterior probability for the joint model is 0.977. In the second example presented at the beginning of this section,

one condition did not contain a mask at all. We claimed that the slopes were different in this case. Here, the model selection strongly supports the isolated models with a model posterior probability of 0.997. Therefore, and consistent with the visual inspection of psychometric functions, the joint procedure should be only applied on the masked conditions, but not in the second example with and without a mask condition.

Discussion

We have presented a Bayesian approach to perform joint inference. By joint inference we mean to perform inference on the basis of several data sets simultaneously. The main difference to other procedures is that the data sets are fitted individually, but all available data sets are taken into account either directly through the likelihood function or through the prior. Thereby the computational and logical effort of the fitting remains manageable because the true complexity is hidden.

We have demonstrated the joint inference procedure with a specific application from perceptual psychology. Several data sets were requested to be explained by the same model class with the supposition that one parameter, the slope, is equivalent in all data sets. In general, the procedure allows a direct model comparison between the joint and isolated models to test the assumption of parameter equality. Here, we used a non-frequentist model selection criterion based on Bayes factors which are readily interpretable. The joint inference procedure can also be applied in other contexts than psychology, because the psychometric function belongs to the generic class of dose-rate-models.

The so-called dose-rate-model is an important model in many natural sciences. Dose-rate-models describe the probability that some event occurs as a function of some independent "dose" variable. In the psychological example the dose corresponds to the contrast of the pattern and the rate to the probability that an observer detects the pattern. Further examples can be found in medicine—the probability of therapeutic success as a function of dose of some medication, or in toxicology—the fraction of test animals that die after application of some toxic substance[94]. The problem of reduced asymptotes is also present in these areas. In

[94] Eaton, D. L., & Klaassen, C. D. (2001). Principles of toxicology. In C. D. Klaassen (Ed.) *Casarett & Doul's Toxicology: The Basic Science of Poisons*. McGraw-Hill

medicine, there might be spontaneous remissions, and in toxicology, some of the test animals might be resistant to the toxic substance. In these cases, the asymptotic levels (spontaneous remissions, guesses, ...) need to be estimated, too. Joint inference as I have described in this chapter could be a suitable choice for all these applications.

In the case of more complex, large models joint inference could also be applicable. Here, experiments must be set up to manipulate only a single or a few selected model parameters, assuming that other parameters of the model remain constant across the different experimental conditions. Thus, there are some parameters in the model that are determined by the general experimental setup; there are other parameters in the model that can only be estimated through a single condition. In this case the complex model can be divided into simpler modules which are tested separately and fitted jointly. The additional benefits we expect from joint inference being applied is that the computational overhead is low and that standard procedures for Bayesian inference can be adopted.

The storage of experimental data

Two real data sets were used as examples in the last section. In addition, 12000 artificial data sets were generated to quantify and illustrate the statistical procedures through simulations. In general, experimental research generates large amounts of data and even more data may be produced during the analysis. As a consequence a worrying amount of researchers' time is devoted to the storage, organization, and retrieval of data. In an attempt to increase transparency and reproducibility, a number of journals now require that experimental data is submitted in addition to the manuscripts. Therefore, a transparent and straightforward data management system steadily gains in importance and is no longer a question of personal preference, but a necessary condition for state of the art experiments.

Usually, the data itself are not self-explanatory and are thus associated with documentations, explanations, and parameter settings, subsequently called annotations or meta-data. The individual researcher decides which information to store at the time and which format to use for data and annotation. This decision depends on personal preferences and may even depend on the hypothesis, or worse, the lack of computer science expertise of at least some scientists[95]. Thus, content and format can differ tremendously between experiments and even more between researchers. Several unfavorable consequences accompany the inconsistency in data management. First, researchers are wasting valuable time and resources in reorganizing data instead of pursuing the research question. Second, only the person who collected the data has complete knowledge of the format, storage location, and the content's meaning. If this person leaves the lab, the data become uninterpretable and potentially worthless. But even if the annotations were sufficient for the original purpose of the experiment, they might

[95] Wilson, G. V. (2006). Where's the real bottleneck in scientific computing? *American Scientist*, *94*, 5–6

be incomplete if the data were to be reanalyzed from a different point of view. Perhaps most importantly, a joint or meta-analysis of several experiments can require unnecessary effort, if the storage formats of experiments differ. Given modern data management tools, the current practices of experimental data storage are outdated and need to be readdressed.

Figure 41: Visual psychophysics experiment: a sinusoidal grating is presented to the observer holding a response box in her hands.

The goal of the work presented in this chapter is to provide experimenters with a unified, intuitive tool to store data and annotations while still holding them logically separate. The system should enforce a complete and consistent description of the data while at the same time it should be adaptable for different experimental requirements. Furthermore, the tool should provide fast and easy data access for one or many users without storing redundant information and without loosing information that may become important in future.

Such a system must fulfill at least three basic requirements. First, it has to be entirely open source, thus available at no charge and alterable by the user. Second, the tool must run on the standard operating systems available today (Mac OS X, Windows, and Linux). And finally, implementation details have to be hidden by a straightforward interface, so that users do not need special programming knowledge such as structured query language (SQL). In the next sections I will explain how data and information associated to the experiment can be structured first from an intuitive user point of view and then from a database storage perspective. Then, I will present the interface that I designed to operate between these two ends and that adheres to the just presented requirements.

Mapping the world onto objects

How can all the data about an experiment including annotations be structured? To approach this question we will inspect a typical visual psychophysics experiment in Figure 41. An observer sits in front of a monitor and holds a response box in her hands. A stimulus is presented on the screen and depending on the task, the observer presses a button. A computer controls stimulus presentation as well as response registration. This scene is part of a larger context. The observer participates in a full experiment which probably requires that the person observes several sessions likely at

different days. During a session the observer sees many stimuli and a single repetition of stimulus presentation with response is called a trial. For subsequent analysis, all information present in the scene should be stored as annotations in addition to the data. Temporal, material, or logical units—such as the observer, the equipment, the experimental design, a specific session, or a single trial—implicitly divide annotations and data into clusters. The objects emerge out of these clusters. Moreover, most experiments have an inherent, hierarchical structure. A trial, for example, belongs to a session within an experiment as sketched in Figure 42.

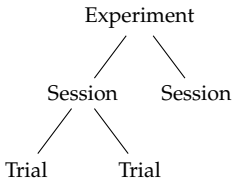


Figure 42: Hierarchical structure

NOT ALL OBJECTS are related in a strict top-to-bottom hierarchy. The same observer can of course participate in several experiments. While not all trials collected in an experiment must be from the same observer, it is not immediately clear what to do with the observer's cluster of information. The observer cannot be inserted in the hierarchy as an object by itself, because the hierarchy would be ambiguous. If we added the observer information directly to another object instead of creating an observer object, then the observer information were entered repeatedly to different object instances. Ideally however, information is only entered once. The solution adopted here is to allow a second kind of relation between objects that do not require a hierarchy. Thus, the observer object can be created the same way as other objects, only that it is then attached as a whole to other objects. It would be attached to the highest object in the hierarchy that contains only children where that observer is correct. In the example with experiment, session and trial the observer would be attached to the session since only one observer participated in the session and all trials of the sessions belong to the same observer (Figure 43). The stimulus is another object that can occur repeatedly within an experiment. It is therefore reasonable to create an object and to attach it outside the actual hierarchy. Depending on the paradigm the stimulus would probably be attached to the trial and not to the session. This second kind of relations are subsequently called contexts.

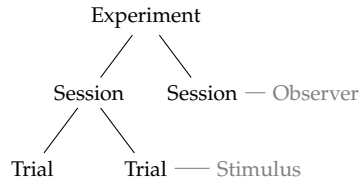


Figure 43: Context structure

THE FIRST STEP in organizing experimental data therefore is to extract elementary units and to identify relations between these units. Then, all relevant properties of those units need to be listed including all sorts of information thus far rarely recorded but perhaps important for meta-analyses. Next, these units with their

properties need to be represented as objects manageable by the computer.

Mapping objects onto a database

Intuitively, a different object class with specific properties could be created for each unit type. However, assuming a higher level of abstraction, each object has only three basic properties: an object type, a corresponding list of annotations and a corresponding list of data.[96] An object of type session for example would contain an annotation list with at least the date inside and possibly eye movement traces in the data list. Other objects, such as a setup object, would contain only annotations, thus having an empty data list. The advantage of such an abstract view is that all objects can be treated similarly by a shared set of functions. Furthermore, objects are not limited to a fixed set of annotations, but are flexibly extendable. Even if an annotation is added at a later point in time, no fundamental programming changes would be required. This point is essential since the paths experimental sciences take are impossible to predict.

[96] Ecker, A. S., Berens, P., & Tolias, A. S. (2007). A data management system for electrophysiological data analysis. *Project Report, Karl Steinbuch Scholarship*

The next critical step is to decide how to store the objects on a computer. In theory, objects could be saved, amongst others, in text files, XML files, or in a database. Text files constitute the most basic option. They are readable by humans and supported by probably every programming language on every operating system. But, text files can not be queried. In contrast, XML files can be queried and are at least partly readable. The disadvantage of XML files is that they can only represent hierarchical data. A database stores its content not in a human readable manner. However, this disadvantage is diminished by its powerful query capabilities and by its possibility to store flat, hierarchical, and even circular data. I therefore decided to adopt a database for psychophysical data.

The objects structure is mapped on a database structure by an Entity-Relationship Model[97] and an Entity-Relationship Diagram (ER-Diagram) visualizes the conceptual schema of the database in Figure 44. Again, no concrete objects are mapped on a database schema, but the common abstract design of experimental objects. Each experimental object is distributed to three primary *entity sets*. More precisely, one entity set exists for each basic object

[97] Chen, P. P.-S. (1976). The entity-relationship model–toward a unified view of data. *ACM Transactions on Database Systems*, *1*(1), 9–36; and Silberschatz, A., Korth, H., & Sudarshan, S. (2006). *Database System Concepts*. New York: McGraw-Hill, 5 ed

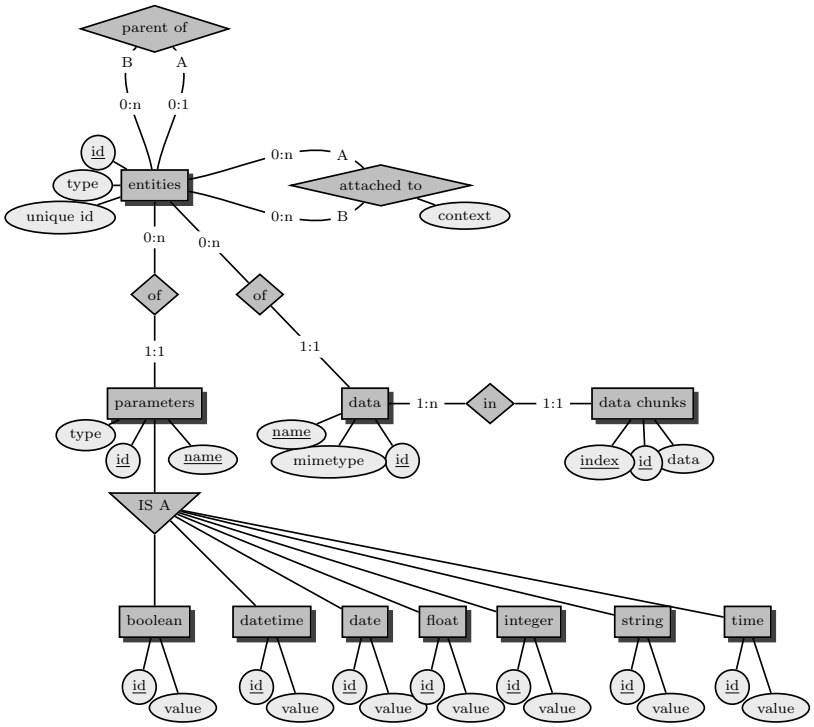


Figure 44: Entity-Relationship-Diagram visualizing the conceptual schema of the database. Three primary entity sets 'entity', 'parameters', and 'data' (rectangles) represent the basic properties of objects. Attributes are shown as ellipses and are underlined if they form a primary key or an unique constraint. A relationship between entity sets is shown as a diamond. Cardinality constraints are depicted by the notation on connecting lines. Parameters are of different computational types and stored accordingly in an inheritance tree grouped beneath the main 'parameters' entity set.

property—the object type, the annotation list and the data list—called 'entities', 'parameters' and 'data', respectively. Inside the database an entity set is a table with a collection of attributes as columns. In the ER-diagram rectangles depict entity sets and ellipses describe attributes. A new observer object, a new session object or a new trial object requires a new entry in the major entity set 'entity'. The attributes of the entity set 'entities' are an unique numerical identifier, an unique general identifier and the object type. The difference between two entries of the same type, for example two observers, are only the identifiers, the object type would be `Observer` in either case. The first numerical identifier is called the *primary key*. They are used to separate different table entries and used to make connections between tables. Primary keys are usually underlined attributes in ER-diagrams. The second general identifier has a different purpose. It is not defined to ease database processing but to provide a unique handle on each object that is saved. It is either a randomly generated unique id or defined by the user. Depending on the annotations or data associated with

the object, new entries are necessary in the 'parameters' and 'data' entity sets as well.

Each annotation is represented through a name that is characteristic and speaking for the annotation and its value. Since the values are of different database types such as integers, floats, strings and others, the parameters entity set is implemented as an inheritance tree. The main entity 'parameters' contains the name of the parameter, an identifier and a field that denotes the database type of the value. One table is created for the main database types. In these tables the values are then stored and can be referenced from the main table. Thus, the main table itself does not contain any value. The full pair name-value is restored only through the identifier and the value type which tells the subordinate table actually containing the value. This separation into an inheritance tree—visualized by a triangle—is important for advanced query capabilities and not related to the conceptual structure of the database.

All recorded data are stored in a binary format. The entity set 'data' and a single data entry is related to the correct object by a *foreign key*. The mimetype of the object is stored as well in the data table. Thereby, the user knows how to handle the specific data item. For example python numpy arrays can be stored through a numpy routine. The numpy mimetype would be 'npy' and the user would apply the numpy.load routine to reload the data. A numpy array with a shape of 100×100 results in a file size of about 79 kilobyte and 7.7 megabyte for a 1000×1000 array.[98] Since the maximal size of the data column is fixed, a maximal size of 5 megabyte, for example, would be too small to fit the larger array. Since the size is defined at a time when the actual data size may not yet be known, we added a second hidden layer in the binary data handling with the entity set 'data chunks'. The size of a data chunk is also fixed, but if some data is larger than the data chunk allows, then it is divided into parts. A single part is saved per chunk including an index indicating its position in the data. This extra functionality allows to add data that is larger than the maximal size allowed by the database column. The maximal size therefore does not need to be set to exceedingly large values just to prevent complications. In contrary, it should be set in the range of the expected data size or even smaller. Python's and the database's memory pose a further constraint on the file size. If the files to be loaded are very large, the application can fail due

[98] numpy version 1.4.1

to limited memory. The memory consuming processes are split into parts through the data chunks and failure are circumvented. The database structure explained so far contains all the tables that store the information about a single object. I will describe next how relations between tables are implemented.

RELATIONSHIPS between entity sets are shown by diamonds and connecting lines. In Figure 44 the numbers on the connecting lines depict the minimal and maximal number of relations in which an entity may participate. An entry in 'entities' may have zero or many data entries shown by the $0 : n$ notation on the line leaving the rectangle 'entity' in direction of 'data'. A 'data' entry is unique and must belong to exactly one 'entities' entry as specified by the $1 : 1$ notation. This kind of relationship is called a one-to-many relationship and it can be implemented by an extra column in the data table, that references to the corresponding row in the entities table through a foreign key containing the entities id. Similar one-to-many relationships are implemented for parameters-of-entities and data-chunks-in-data relations. The database structures that are necessary to save single object are completed with the tables and one-to-many relationships. The relationships that are used to implement the relations between objects at a database level are more complex. Both the hierarchical and non-hierarchical relations only reference to or within the 'entities' table. The hierarchical structure in experimental data is also a one-to-many relationship in which an entity B is the parent of another entity A. An entity has $0 : n$ children and $0 : 1$ parent. The non-hierarchical relations must be implemented in a many-to-many relationship which cannot be implemented through direct foreign keys and requires a *relationship set*. A *relationship set* is a table itself and is also shown as a diamond in Figure 44. It possesses the attribute called context. The example from the previous section, with an observer object being attached to a session object, would be implemented here with a meaningful context attribute. A speaking attribute would be "observes" since a person observes in a session. The relationship is a many-to-many relationship since the observer can of course be attached to other sessions as well. A session might have different objects attached in other contexts. The hardware setup might be implemented in an object and attached to the session with a context "equips", since the hardware equips a session.

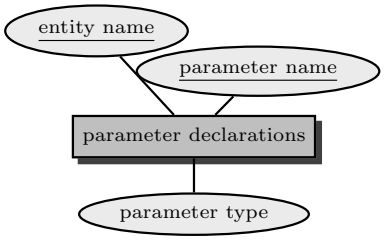


Figure 45: Table 'parameter declarations' stores the object and parameter declarations independent of the object's content

The database structure to store experimental objects is complete. It is general, flexible, and does not impose any constraints on the stored objects. Consistency checks are only performed at a database level which means that the database types are checked. Whether the objects that are stored are consistent with another object of the same object type already stored at an earlier day is not enforced. However, consistency at the object level is also important. For example, would we like to preclude clerical mistakes. The database saves a parameter called "contrast" or "contast" and would not complain about a difference. During a search for the keyword "contrast", the wrongly spelled parameter could not be found. To prevent this kind of error a further table was introduced that is independent on the main database structures. Figure 45 shows the corresponding ER-Diagram. This table holds combinations of an entity type, an parameter name and the parameters type, e.g. (`Observer`, "contrast", float). The interface that connects database and application verifies the consistency with the help of this additional table. The interface called Xdapy is introduced in the following section.

Managing experimental data with Python: Xdapy

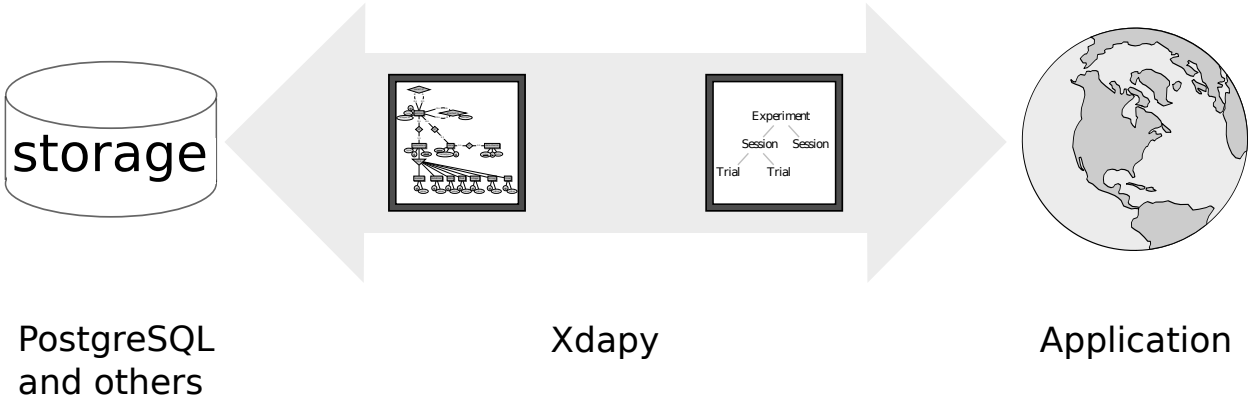


Figure 46: Figure of Xdapy

Xdapy is the intermediate layer between application and database. On the one hand Xdapy knows about the database structure and where the database is located. It manages a copy of the

structures/tables of the database according to the schema in Figure 44. On the other hand it knows about the user defined objects. Unnoticed by the user, it maps between objects and the structures of the database. The user has to define her objects mirroring the experimenter's world following Xdapy's definitions in her application. Examples can be found on page 89. Two classes provide Xdapy's main functionality: the mapper and the connection.

THE MAPPER handles the content processing. It controls which objects can be used with the database. That means the mapper can only handle objects that are either already defined in the structure table or that the user has registered in the mapper in the current session. The mapper saves and loads objects when running experiments and recording data. During the data analysis, the mapper is used to find specific information about the experiments. An application communicates exclusively with the mapper.

THE CONNECTION manages database related processing and setup. With the installation of Xdapy, the connection is used once to create the tables in the database. Then, in a regular Xdapy work routine, its purpose is to connect, as the name suggests, with the database and to provide working capacities, so-called sessions, to the mapper. In the remaining chapter, I will first explain how Xdapy is installed, configured and tested. Then I will use the example from Figure 43 to define objects and create, save, and load object instances.

Install, configure, and test Xdapy

WE DECIDED to use Python as programming language[99] because of its growing number of users in the neurosciences. Xdapy depends on a few third party products, mainly on the software package SQLAlchemy[100] and the relational database management system PostgreSQL[101] or SQLite[102]. The choice of Postgres for the primary implementation was influenced by the popularity of this open source database, its availability for many operating systems and its wide distribution. The full functionality of Xdapy has only been tested with Postgresql. However, the light-weighted alternative—SQLite—should work reasonably well. The test procedures are for example based on SQLite. In general, it is possible to use Xdapy with different relational database management systems (RDMS). However, a few adaptations may be necessary.

[99] http://www.python.org/

[100] http://www.sqlalchemy.org/

[101] http://www.postgresql.org/

[102] http://www.sqlite.org/

The following five main software packages must be installed:

- PostgreSQL / SQLite
- Python 2.6 / 2.7
- psycopg2 2.4.1 or newer
- SQLAlchemy 0.7 or newer
- Xdapy

The database management systems, SQLAlchemy, and psychopg2—a Python-PostgreSQL database adapter—can be downloaded from their respective web pages and are installed following the accompanying installation advices. If the database is supposed to run on a remote server, only the PostgreSQL client software has to be installed on the user's machine. Python is installed on most operating systems. Otherwise it can also be reinstalled following the instructions on python.org. Xdapy's source code can be downloaded from the git repository https://github.com/xdapy/xdapy.git and is ready to use.

A few configurations are necessary after the installation. First, the directory of Xdapy must be added to the PYTHONPATH. Next, a PostgreSQL user must be created and at least one database initialized to use Xdapy. However, a second database for testing is recommended and for the demo below even a third database may be initialized. Again many tutorials and documentations for these processes are online and no further details are provided here especially since the instructions might change with future releases. To facilitate database access from Xdapy the default connections are provided in an initialization file *~/.xdapy/engine.ini*. Figure 47 displays an example file that contains three possible connections. The default option must be specified. It is the link to the database that Xdapy uses to store the application's data. Furthermore, a test option can be provided to run Xdapy's tests after the installation. Here, we suggest to use an in-memory SQLite database. A third SQLite database is used in the following demonstration.

```
[default]
url = postgresql://hannah@
      localhost/xdapy
[test]
url = sqlite://
[demo]
url = sqlite:///demo.db
```

Figure 47: Content of *~/.xdapy/engine.ini* for default database access

A small test battery should be run to test if the installation and configuration was successful. After changing to the directory that contains Xdapy, either nosetests automatically run all tests or starting with python2.7 unittest works as well.

```
cd path/to/xdapy
nosetests
#or
python -m unittest discover
```

The following configuration already requires a functioning setup and is the first action on the new database: the creation of tables.

```
from xdapy import Connection

connection = Connection.profile("demo")
connection.create_tables()
```

We opened a connection to the `demo` profile defined in engine.ini and create the tables required by Xdapy. Tables are only created once. A second issue of the command can overwrite existing tables including their content.

Define and register object classes

THE OBJECT CLASSES that will contain the data and metadata are defined next. All objects are derived from class 'Entity' and the name of the class is the type of the object. Their parameters are defined in a Python dictionary, declared_params, with the parameter name as key and the parameter type as value.

```
from xdapy import Entity

class Experiment(Entity):
    declared_params = {
        'project': 'string',
        'experimenter': 'string'
    }

class Observer(Entity):
    declared_params = {
        'birthyear': 'integer',
        'initials': 'string',
        'handedness': 'string',
        'glasses': 'boolean'
    }

class Session(Entity):
    declared_params = {
        'count': 'integer',
        'date': 'date',
        'experimentalcondition': 'string',
    }

class Trial(Entity):
    declared_params = {
        'count': 'integer',
```

```
        'stimulus': 'integer',
        'answercorrect': 'boolean'
    }
```

Xdapy does not enforce that each parameter from the declared parameters dictionary is defined in every object instance. If such an enforcement is desired by the user, he has to implement it in the application. However, no other parameter except for the declared ones can be stored with an object. Each object has a unique general identifier by which it can be distinguished from other objects. The general identifier is a unique field that is assigned to every object instance when saved in the database. If not defined explicitly it is a random string based identifier. For objects that are repeatedly used, such as observers in the current use case, it is useful to define another intuitive identifier to simplify the object access. Such special identifiers are defined by a method called `gen_unique_id`. For the observer the unique identifier could be the initials and birthyear, e.g. GTF1801.

```
class Observer(Entity):
    declared_params = {
        'birthyear': 'integer',
        'initials': 'string',
        'handedness': 'string',
        'glasses': 'boolean'
    }

    def gen_unique_id():
        return "%s%s"%(self.params["initials"],
                       self.params["birthyear"])
```

In order to save an object instance with a specific general identifier, the properties of the object that are used to build the identifier must be defined prior to saving.

Every use of Xdapy requires a registration of the objects in the mapper. If the object definitions are stored in a file, say objects.py, then they can simply be imported and registered in the mapper. The registration process is a security step to ensure compatibility between different applications of Xdapy.

```
from xdapy import Connection, Mapper
from objects import Trial, Experiment, Session, Observer

connection = Connection.profile("demo")
m = Mapper(connection)
m.register(Observer)
m.register(Experiment)
m.register(Trial)
m.register(Session)
```

Once the object classes are registered Xdapy is ready to work with data.

Create and save object instances

After the initialization of the mapper and after registering the object classes, new objects can be created and saved using the mapper. The parameters can either be supplied during initialization or subsequently by adding them to the dictionary `params`.

```
e1 = Experiment(project='Main', experimenter="John_Doe")
m.save(e1)

e2 = Experiment(experimenter="John_Doe")
e2.params['project'] = "Control"
m.save(e2)
```

Once an object is created it should also be saved. By saving an object its parameters, data and relations are saved as well. Here some more objects for a realistic use case:

```
o1 = Observer(handedness="right", birthyear=1986,
              initials='LD', glasses=False)
o2 = Observer(handedness='right', birthyear=1982,
              initials='SG', glasses=True)
m.save(o1)
m.save(o2)

import datetime
s1 = Session(count=1, date=datetime.date.today(),
              experimentalcondition='outline')
s2 = Session(count=2, date=datetime.date.today(),
              experimentalcondition='filled')
m.save(s1)
m.save(s2)

t1 = Trial(count=1, stimulus=20, answercorrect=True)
t2 = Trial(count=2, stimulus=36, answercorrect=True)
t3 = Trial(count=3, stimulus=8, answercorrect=False)
t4 = Trial(count=4, stimulus=87, answercorrect=False)
t5 = Trial(count=5, stimulus=26, answercorrect=True)
t6 = Trial(count=6, stimulus=74, answercorrect=True)
t7 = Trial(count=7, stimulus=20, answercorrect=False)
```

Creating objects, adding parameters, and saving them is straightforward. However, the goal to make data understandable is not completely matched. What are the numbers in the field of the stimulus parameter in the trial object? Is it the size of the stimulus, its contrast, or an identifier? If it were stimulus size or contrast, the parameter name is not well chosen. If it were a stimulus identifier, the information about the stimulus itself is fully missing. A

stimulus probably contains many parameters and these must be annotated for the sake of reproducibility. At best, the parameters used to generate the stimulus and a representation of the stimulus would be saved in the database. Depending on the experiment, the stimulus might be used repeatedly in different trials. It would therefore make sense to combine the stimulus parameters and only link them to the trials in which the stimulus was used. In practice that means to create a stimulus object and to create context relations between trials and stimuli. A stimulus commonly used in visual psychophysics and already defined on page 22 is the Gabor. A Gabor is defined by the parameters of a sine wave and the parameters of a Gaussian aperture. For the current illustration the true number of parameters is slightly reduced to horizontal frequency, vertical frequency, variance, contrast, and the number of sampling points in pixels.

```
class Stimulus(Entity):
    declared_params = {
        'pixel': 'integer',
        'contrast': 'float',
        'horizontal_frequency': 'float',
        'vertical_frequency': 'float',
        'variance': 'float'
    }
```

In addition, the stimulus could be complemented with a real image in binary form. Similar to an objects parameter dictionary, a data dictionary is provided by the superclass Entity. If the image data is contained in file 'image.jpg', it can be inserted as follows:

```
stimulus = Stimulus()
m.save(stimulus)

f = open("image.jpg")
stimulus.data["image"].put(f)
stimulus.data["image"].mimetype = "jpg"
m.save(stimulus)
f.close()
```

Afterward the data can be retrieved also via a file.

```
out = open("out.jpg", "w")
stimulus.data["image"].get(out)
out.close()
```

If the data is not already in a binary format, e.g. a numpy array, it can first be saved through the numpy serialization routine in a temporary file and can be inserted afterward.

```
from tempfile import TemporaryFile
import numpy as np

#insert binary data
tmpout = TemporaryFile()
narray = np.random.rand(5,5)
np.save(tmpout, narray)
tmpout.seek(0) #change from write to read rights
stimulus.data['array'].put(tmpout)
stimulus.data['array'].mimetype = 'npy'
m.save(stimulus)
tmpout.close()

#retrieve binary data
tmpout = TemporaryFile()
stimulus.data['array'].get(tmpout)
tmpout.seek(0)
k = np.load(tmpout)
```

Alternatively the serialization routines from pickle might be used.

```
import cPickle

#insert binary data
stimulus.data['array'].put(cPickle.dumps(narray))
stimulus.data['array'].mimetype = 'pcl'
m.save(stimulus)

#retrieve binary data
k = cPickle.loads(stimulus.data['image'].get_string())
```

The difference between data and parameters is that parameters can be used in a query, data can not. That means that every information that is necessary to retrieve objects from the database must be saved as a parameter. The data variables can be used to save files directly such as source code or raw data. We just learned how objects are saved in the database. So far however, there are no links between objects. Would Session, s1, belong to the main experiment or to the control?

Relate objects instances

There are two possibilities to link objects. The first is by building a hierarchy through parent and child relationships. With such a relationship we define for example that session s1 belongs to the main experiment e1 or we add the trials to a session.

```
#Hierarchy through parent or child definition
s1.parent(e1)
s2.parent(e1)
m.save(e1)

```

```
#save single child
s1.children.append(t1)
s1.children.append(t2)
#save several children at once
s2.children += [t3, t4, t5, t6, t7]
m.save(s1)
m.save(s2)
```

The second is by creating a context. An observer, for example, is not directly in the experimental data hierarchy, since an observer can participate in several sessions even of different experiments. Through a context, the observer is linked to a session and as a result, all information of that session or in the hierarchy below the session belongs to that observer.

```
s1.context = {'observed':[o1]}

t1.attach('presented', stimulus)
```

Given the hierarchy in which session s1 is embedded, the first line indicate that session s1 and all the trial from t1 to t7 were observed by observer o1. A context can be either set directly via the context attribute or via the `attach` method.

Load object instances

Assuming that all sessions of Experiment e1 and e2 were recorded and stored, the question is how to extract the data from the database for analysis. There are several possibilities to query the database. The simplest options are the mapper methods `find_all` and `find_first`.

```
o = m.find_first(Observer, filter=
                 {"birthyear": range(1970, 1985)})
print o.params
>>>>{u'birthyear': 1982, u'handedness': u'right',
     u'glasses': True, u'initials': u'SG'}

o = m.find_all(Observer, filter={"initials": ["%D%"]})
print o[0].params
>>>>{u'birthyear': 1986, u'handedness': u'right',
     u'glasses': False, u'initials': u'LD'}
```

There is a further way to query. In contrast to the other possibilities it does not return objects or a list of objects, but a SQLAlchemy query structure that has not loaded the objects from the database yet. The query structure can be iterated and then loads and returns the objects on demand.

```
o = m.find(Observer, filter={'glasses':False})
```

The following procedure lists the stimulus number and response of the trials from the `outline` condition in the `main` experiment in which observer `MM` responded correctly.

```
sessions = m.find_complex("Session",
    {("_context", "observed_by"): {"_any": [("Observer",
                                        {"initials": "LD"}),]},
    "_parent":("Experiment":{"project":"Main"})})
    "experimentalcondition":'outline'})
trials = m.find_complex("Trial",{"_parent": {"_any": sessions}})
for trial in trials:
    print trial.params['stimulus'], trial.params['answercorrect']

>>>> 20 True
>>>> 36 True
```

Admittedly, the syntax of complex searches is not easy to read. Repeatedly used searches could therefore be encapsulated in regular python methods.

Discussion

To SUMMARIZE, Xdapy provides the functionality to save any hierarchical data with cross-references in a SQL database. The structure and content of the data is solely determined by the user. The user application also has to handle obligatory data fields and consistency. This separation is necessary to have a general framework on the one hand and consistent data on the other hand. The advantage of Xdapy over other storage approaches is its flexibility. The user can define and save the object including the parameters and the relation between the objects without any knowledge of SQL. The retrieval of data also only requires that the user exploits the structures he developed. Since the formats at the database level and the application level are independent of each other, new objects can be added without changes at the database level.

AROUND THE SAME TIME AS THIS PROJECT STARTED the German Neuroinformatics Node (G-Node) also started a data management and sharing project [103]. This shows the need for such systems. The spirit of G-Node's system and the project presented here are similar. However several design decisions were taken differently. First, G-Node requires the user to store the data remotely on the G-node site. This has the advantage of providing all the infrastructure for the user. For the data storage system I developed, the user can retain full control over its data and store it on its own machine

[103] https://portal.g-node.org/data/

or its own server. The user can thereby use a database without knowledge of SQL. However, the user herself or a system administrator needs to maintain the database managements system and its access. Second, G-Node focuses on electrophysiological data and tries to provide a standard for that specific type of data. Here, any kind of data can be stored. It is still in the user's responsibility to decide which information is stored and to develop the structures that contain that information. Of course, this freedom is tied to the possibility of misuse. Misuse in the sense that annotations might not be stored at the level of detail that should be required. I believe that with time, increased usage, and gained experience standards will develop within research groups that are tailored first to their research questions and that will later on develop across groups and fields.

The needs of scientists differ and different tools might be necessary to satisfy all of them. Having a choice between different tools that serve the same purpose—manage, store, and annotate data—is therefore important. So far, Xdapy is the beginning of a data storage tool in psychophysics and it will hopefully prove to meet the needs of many psychophysicists.

Conclusion

Computational modeling constitutes an essential scientific tool across disciplines. Modeling is typically based on empirical data, but the purpose of the model and how it is used can differ dramatically. The extreme cases call "model serves data" and "data serves model". In most applications these two cases refer to statistical versus mechanistic models. The "model serves data" case encompasses applications in which the model is used to describe data, to interpolate between data or to predict data. Here, the modeling part is mainly model fitting and it is often sufficient to use statistical models since the main focus is placed on the data. On the other hand, in the "data serves model" case the focus is on the model itself. The data is used to identify the model or model components or to choose between competing models. Whenever conclusions are drawn from the model and its parameters it is of utmost importance that these parameters are meaningful. This is only the case if they are constrained by the data at hand. The two applications presented in this thesis both belong to the "data serves model" category. The first case presented in the chapter *Towards an image-driven early spatial vision model* is clear because of the mechanistic nature of the vision model. The data is used to pin down the underlying mechanisms. To allow for a proper interpretation of the necessary model structures, it is critical to obtain a well fitting model that, in addition, is well defined. The classification is less obvious in the case of the *Joint Bayesian estimation of several psychometric functions*. The psychometric function is a special case that would traditionally be regarded as a statistical model. However, since the conclusions are drawn based on the parameters of the model, as is done when comparing thresholds across conditions, the model also falls into the "data serves model" class. In this thesis I used Bayesian graphical models to formulate both model cases that are based on data from visual psychophysics. The gain that accompanies the Bayesian

approach differs between the two cases and I next will discuss the advantages and promises for either case.

The idea of early vision as a merely Fourier-style process is outdated. Too many experimental findings were incompatible with simple linear operations.[104] Yet, the current standard model for human spatial pattern sensitivity remains a multi-resolution model and this linear nonlinear cascade model is implicitly or explicitly present in many studies and standard text books in visual neuroscience. Somehow, a large gap opened between supporting and contradictory results which is also reflected in many advocates and vehement opponents of the early vision literature.[105,106] The model evolved mainly from experiment to mechanism. An experiment suggested a possible mechanism, that mechanism was useful also as predictor in other experiments and thereby gained in evidence. Several mechanisms were then put together to form the model. As a consequence the model can be thought of as a post-hoc explanation. The similarity between psychophysical channels and simple cells in the primary visual cortex nourished the confidence in the model.[107] A reason for some researchers to stick to the model till this day is that each processing stage is supported by experimental findings and the belief in an early general representation of visual information. To increase the acceptance of the model further evidence in favor of the model is needed and Bayesian model analysis seems to be a good tool to provide it. The analysis of the model's parameter posterior distribution has a twofold appeal. Once it is possible to show that the model is not too general, but actually tailored to the recorded data, the model can be more easily accepted. This can be shown by presenting the posterior distributions in addition to the model prediction. Second, as demonstrated here as well, the posterior distributions provide insights about the adequateness of the model. If the distribution has irregularities, these can be taken as starting point for further investigations. Irregularities might lead to a different parametrization of the model, to a model simplification or to the gathering of more experimental data. They can even provide the relevant parameter range to be address through experimental manipulation and facilitate the design of experiments. The Bayesian approach thereby creates new experiments and modeling ideas.

How creative Bayesian inference can be is also reflected in the second chapter, but on a different level compared to the first.

[104] Henning, G. B. (1975). Some experiments bearing on the hypothesis that the visual system analyses spatial patterns in independent bands of spatial frequency. *Vision research*, *15*, 887–897; Derrington, A. M., & Henning, G. B. (1989). Some observations on the masking effects of two-dimensional stimuli. *Vision Research*, *29*(2), 241–246; and Meese, T. S., & Holmes, D. J. (2007). Spatial and temporal dependencies of cross-orientation suppression in human vision. *Proceedings of the Royal Society B*, *274*, 127–136

[105] Westheimer, G. (2001). The Fourier theory of vision. *Perception*, *30*, 531–541

[106] personal impression

[107] Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978b). Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *Journal of Physiology*, *283*, 53–77; Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978a). Receptive field organization of complex cells in the cat's striate cortex. *Journal of Physiology*, *283*, 79–99; and Carandini, M. (2006). What simple and complex cells compute. *Journal of Physiology*, *577*(2), 463–466

The classical formulation of Bayes' rule allows to inject prior knowledge into the inference process. With the joint estimation procedure the formulation is turned into an inference process that adds information obtained at the same time—it performs simultaneous inference. To recall the application: Two psychometric functions were measured. The results indicated that the slopes of the psychometric functions could be equal. Therefore an analysis was necessary that accounts for this equality and preferably also validates if the slopes only appear to be equal or if they can be legitimately said to be equal. Thus, the procedure is designed to analyze several experimental conditions in a joint model. However, in contrast to the vision model, the joint analysis does not rely on a full mechanistic model that specifies how all conditions are linked to each other, for example through a contrast sensitivity function. The joint model still takes each condition separately and most parameters adhere in a one-to-one relation to a single condition only. Additionally, the joint analysis defines that some parameter, e.g. the slope, is equal across the condition. This supplementary definition is then not expressed in a full model which grows with every condition, but in the definition of the prior. As a result the analysis remains low in complexity. If a robust Bayesian standard procedure is available for individual conditions, as it is for psychometric function fitting, then this procedure can be kept for the joint analysis. This reduces tool development efforts and increases user confidence. Arguably, the biggest advantage is that many experimental conditions can be treated jointly in meta-analyses without additional efforts.

Our detailed knowledge in the area of vision has multiplied and local theories, or hypotheses, exist to explain these details. A major challenge is to converge those local theories into a common comprehensive theory. Meta-analyses are a straightforward way towards this goal. Despite their importance however, several reasons make it surprisingly difficult to perform a meta-analysis. The necessary data is most likely not collected from the person who conducts the meta-analyses. Years could pass between the gathering of the data and their inclusion into a meta-analysis. The reanalysis may require information about the procedures that was not originally stored, since the documentation was tailored to the needs of the primary analysis. An example from auditory psychophysics demonstrates these problems: Tone-in-noise detection experiments were already conducted many years ago and at the time evaluated in blocks.[108] Today, one would like to analyze these experiment

[108] Fletcher, H. (1938). Loudness, masking and their relation to the hearing process and the problem of noise measurement. *The Journal of the Acoustical Society of America*, *9*, 275–293

with a trial-to-trial precision.[109] In order to do so, the classical experiments had to be repeated painstakingly instead of just being reanalyzed.[110] Xdapy, the data management tool presented in the previous chapter, has the potential to ease experimental data access and storage in the future. It strongly encourages the user to annotate the data consistently and simplifies extensive annotating. At the moment, the tool is still in its infancy. However, if received well by researchers in psychophysics and perhaps the behavioral neurosciences at large, it will save the valuable time that researchers otherwise would have to spend on data handling, thus easing data sharing and promoting meta-analyses.

In sum, this thesis substantiates the trinity of model, data, and estimation method. First, I used Bayesian methods to tailor an early vision model to its founding psychophysical data. Second, I presented how Bayes rule can formulate simple modular models to fit many experimental conditions instead of a complex model. Finally, I presented a software package to save, store, retrieve, and potentially share experimental data.

[109] Schönfelder, V. H., & Wichmann, F. A. (2012b). Sparse regularized regression identifies behaviorally-relevant stimulus features from psychophysical data. *The Journal of the Acoustical Society of America*, *131*(5), 3953–3969

[110] Schönfelder, V. H., & Wichmann, F. A. (2012a). Identification of Stimulus Cues in Narrow-Band Tone-in-Noise Detection using Sparse Observer Models (in preparation)

Bibliography

Allen, D., Norcia, A. M., & Tyler, C. W. (1986). Comparative study of electrophysiological and psychophysical measurement of the contrast sensitivity function in humans. *American Journal of Optometry and Physiological Optics*, *63*(6), 442–449.

Bird, C. M., Henning, G. B., & Wichmann, F. A. (2002). Contrast discrimination with sinusoidal gratings of different spatial frequency. *Journal of the Optical Society of America A*, *19*(7), 1267–1273.

Blakemore, C., & Campbell, F. W. (1969). On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *Journal of Physiology*, *203*, 237–260.

Blakemore, C., & Tobin, E. A. (1972). Lateral inhibition between orientation detectors in the cat's visual cortex. *Experimental Brain Research*, *15*, 439–440.

Bowne, S. F. (1990). Contrast discrimination cannot explain spatial frequency, orientation or temporal frequency discrimination. *Vision Research*, *30*(3), 449–461.

Boynton, G., Demb, J. B., Glover, G., & Heeger, D. J. (1999). Neuronal basis of contrast discrimination. *Vision Research*, *39*, 257–269.

Burton, G. J. (1981). Contrast discrimination by the human visual system. *Biological Cybernetics*, *40*(1), 27–38.

Campbell, F. W., & Robson, J. G. (1968). Applicaton of Fourier analysis to the visibility of gratings. *Journal of Physiology*, *197*, 551–566.

Carandini, M. (2006). What simple and complex cells compute. *Journal of Physiology*, *577*(2), 463–466.

Carney, T., Klein, S. A., Tyler, C. W., Silverstein, A. D., Beutter, B., Levi, D., Watson, A. B., Reeves, A. J., Norcia, A. M., Chen, C.-C., Makous, W., & Eckstein, M. P. (1999). Development of an image/threshold database for designing and testing human vision models. *Proceedings Vol. 3644 Human Vision and Electronic Imaging IV, Bernice E. Rogowitz; Thrasyvoulos N. Pappas, Editors,*, (pp. 542–551).

Chen, P. P.-S. (1976). The entity-relationship model–toward a unified view of data. *ACM Transactions on Database Systems*, *1*(1), 9–36.

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist*, *49*(12), 997–1003.

Cornsweet, T. N., & Yellott Jr, J. I. (1985). Intensity-dependent spatial summation. *Journal of the Optical Society of America A*, *2*(10), 1769–1786.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, *2*(7), 1160–1169.

Davis, E. T., & Graham, N. (1981). Spatial frequency uncertainty effects in the detection of sinusoidal gratings. *Vision research*, *21*, 705–712.

Davis, E. T., Kramer, P., & Graham, N. (1983). Uncertainty about spatial frequency, spatial position, or contrast of visual patterns. *Pereption & Psychophysics*, *33*(1), 20–28.

Dean, A. F. (1981). The variability of discharge of simple cells in the cat striate cortex. *Experimental Brain Research*, *44*, 437–440.

De Lean, A., Munson, I. J., & Rodbard, D. (1978). Simultaneous analysis of families of sigmoidal curves: application to bioassay, radioligand assay, and physiological dose-response curves. *American Journal of Physiology*, *235*(2), E97–E102.

Derrington, A. M., & Henning, G. B. (1989). Some observations on the masking effects of two-dimensional stimuli. *Vision Research*, *29*(2), 241–246.

De Valois, R. L., & De Valois, K. K. (1980). Spatial vision. *Annual Review of Psychology*, *31*, 309–341.

Dobson, A. J., & Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. Boca Raton: Chapman & Hall/CRC, 3 ed.

Eaton, D. L., & Klaassen, C. D. (2001). Principles of toxicology. In C. D. Klaassen (Ed.) *Casarett & Doul's Toxicology: The Basic Science of Poisons*. McGraw-Hill.

Ecker, A. S., Berens, P., & Tolias, A. S. (2007). A data management system for electrophysiological data analysis. *Project Report, Karl Steinbuch Scholarship*.

Eliasmith, C., & Anderson, C. H. (2003). *Neural Engineering: Computation, representation and dynamics in neurobiological systems.*. MIT Press.

Fletcher, H. (1938). Loudness, masking and their relation to the hearing process and the problem of noise measurement. *The Journal of the Acoustical Society of America*, *9*, 275–293.

Foley, J. M. (1994). Human luminance pattern-vision mechanisms: masking experiments require a new model. *Journal of the Optical Society of America A*, *11*(6), 1710–1719.

Fründ, I., Haenel, N. V., & Wichmann, F. A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision*, *11*(6(16)), 1–19.

García-Pérez, M. A., & Alcalá-Quintana, R. (200). Fixed vs. variable noise in 2AFC contrast discrimination: lessons from psychometric functions. *Spatial Vision*, *22*(4), 273–300.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis*. Boca Raton: Chapman & Hall, 2 ed.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Georgeson, M. A., & Harris, M. G. (1984). Spatial selectivity of contrast adaptation: models and data. *Vision Research*, *24*(7), 729–741.

Georgeson, M. A., & Meese, T. S. (2006). Fixed or variable noise in contrast discrimination? The jury's still out... *Vision Research*, *46*(25), 4294–4303.

Geweke, J. (1992). Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In J. Bernardo,

J. Berger, A. Dawid, & A. Smith (Eds.) *Bayesian Statistics 4*. Oxford: Oxford University Press.

Goris, R. L. T., Putzeys, T., Wagemans, J., & Wichmann, F. (2011). A neural population model for pattern detection. *Journal of Vision*, *11*(11), 1165.

Goris, R. L. T., Wichmann, F. A., & Henning, G. B. (2009). A neurophysiologically plausible population-code model for human contrast discriminations. *Journal Of Vision*, *9*(7), 1–22.

Graham, N. (1977). Visual detection of aperiodic spatial stimuli by probability summation along narrowband channels. *Vision Research*, *17*, 637–652.

Graham, N. V., & Nachmias, J. (1971). Detection of grating patterns containing two spatial frequencies: a comparison of single-channel and multiple-channels models. *Vision Research*, *11*(3), 251–259.

Graham, N. V. S. (1989). *Visual Pattern Analyzers*. Issue 16 of Oxford Psychology Series, Oxford University Press.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*, 181–197.

Henning, G. B. (1975). Some experiments bearing on the hypothesis that the visual system analyses spatial patterns in independent bands of spatial frequency. *Vision research*, *15*, 887–897.

Henning, G. B., Bird, C. M., & Wichmann, F. A. (2002). Contrast discrimination with pulse trains in pink noise. *Journal of the Optical Society of America A*, *19*(7), 1259–1266.

Hoff, P. D. (2011). *A first Course in Bayesian Statistical Methods*. Heidelberg: Springer.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*, 106–154.

Itti, L., Koch, C., & Braun, J. (2000). Revisiting spatial vision: toward a unifying model. *Journal of the Optical Society of America A*, *17*(11), 1899–1917.

Kekre, H., Sahasrabudhe, S., & Goyal, N. (1983). Raised cosine function for image restoration. *Signal Processing*, *5*(1), 61 – 73.

Kelly, D. H. (1979). Motion and vision. ii. stabilized spatio-temporal threshold surface. *Journal of the Optical Society of America A*, *69*, 1340–1349.

Kroese, D. P., Taimre, T., & Botev, Z. I. (2011). *Handbook of Monte Carlo Methods*. Hoboken: John Wiley & Sons, Inc.

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision*, *5*, 478–492.

Legge, G. E., & Foley, J. M. (1980). Contrast masking in human vision. *Journal of the Optical Society of America A*, *70*(12), 1458–1471.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. Boca Raton: Chapman & Hall/CRC, 2 ed.

Meese, T. S., & Holmes, D. J. (2007). Spatial and temporal dependencies of cross-orientation suppression in human vision. *Proceedings of the Royal Society B*, *274*, 127–136.

Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978a). Receptive field organization of complex cells in the cat's striate cortex. *Journal of Physiology*, *283*, 79–99.

Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978b). Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *Journal of Physiology*, *283*, 53–77.

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*(3), 499–518.

Nachmias, J., & Kocher, E. C. (1970). Visual detection and discrimination of luminance increments. *Journal of the Optical Society of America*, *60*, 382–389.

Nachmias, J., & Sansbury, R. V. (1973). Grating contrast: Discrimination may be better than detection. *Vision research*, *14*, 1039–1042.

Naka, K. I., & Rushton, W. A. H. (1966). S-potentials from luminosity units in the retina of fish (cyprinidae). *Journal of Physiology*, *185*(3), 587–599.

Pinto, N., Doukhan, D., DiCarlo, J. J., & Cox, D. D. (2009). A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Computational Biology*, *5*(11), 1–12.

Quick, R. F. (1974). A vector-magnitue model of contrast detection. *Kybernetik*, *16*, 65–67.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Ringach, D., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, *28*, 147–166.

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110–120.

Sceniak, M. P., Hawken, M. J., & Shapley, R. (2002). Contrast-dependent changes in spatial frequency tuning of macaque V1 neurons: effects of a changing receptive field size. *Journal of Neurophysiology*, *88*, 1363–1373.

Schönfelder, V. H., & Wichmann, F. A. (2012a). Identification of Stimulus Cues in Narrow-Band Tone-in-Noise Detection using Sparse Observer Models (in preparation).

Schönfelder, V. H., & Wichmann, F. A. (2012b). Sparse regularized regression identifies behaviorally-relevant stimulus features from psychophysical data. *The Journal of the Acoustical Society of America*, *131*(5), 3953–3969.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature neuroscience*, *4*(8), 819–825.

Silberschatz, A., Korth, H., & Sudarshan, S. (2006). *Database System Concepts*. New York: McGraw-Hill, 5 ed.

Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J., & Davis, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. *Nature*, *378*, 492–496.

Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: a flexible architecture for multi-scale derivative computation. *2nd IEEE International Conference on Image Processing*, *III*, 444–447.

Sinz, F., & Bethge, M. (2008). The conjoint effect of divisive normalization and orientation selectivity on redundancy reduction in natural images. In *Advances in neural information processing systems 21, Twenty-Second Annual Conference on Neural Information Processing Systems*, (pp. 1521–1528).

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions*. New York: Springer, 3 ed.

Thomas, J. P. (1985). Detection and identification: how are they related? *Journal of the Optical Society of America A*, *2*(9), 1457–1467.

Tolhurst, D. J., Movshon, J. A., & Dean, A. F. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research*, *23*(8), 775–785.

Treutwein, B., & Strasburger, H. (1999). Fitting the psychometric function. *Perception & Psychophysics*, *61*(1), 87–106.

Tversky, A., & Krantz, D. H. (1970). The dimensional representation and the metric structure of similarity data. *Journal of Mathematical Psychology*, *7*, 572–596.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wandell, B. A. (1995). *Foundations of Vision*. Sinauer Associates, Inc., flyleaf.

Watson, A. B., & Ahumada, A. J. (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision*, *5*(9), 717–740.

Watson, A. B., & Solomon, J. A. (1997). Model of Visual Contrast Gain Control and Pattern Masking. *Journal of the Optical Society of America A*, *14*(9), 2379–2391.

Westheimer, G. (2001). The Fourier theory of vision. *Perception*, *30*, 531–541.

Wichmann, F. A. (1999). *Some aspects of modelling human spatial vision: Contrast discrimination*. Unpublished doctoral dissertation, The University of Oxford, Oxford, UK.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313.

Wilson, G. V. (2006). Where's the real bottleneck in scientific computing? *American Scientist*, *94*, 5–6.

Appendix

Products of exponential family distributions

We consider two probability distributions P_1 and P_2 from the same exponential family with densities f_1 and f_2. We define a third density $f_3 := f_1 \cdot f_2$. Under what conditions is the distribution P_3 with density f_3 from the same exponential family as P_1 and P_2?

First, we write out $f_i,\ i = 1, 2$

$$f_i(x) = h(\vartheta_i)g(x)\exp(\eta(\vartheta_i)T(x)). \tag{14}$$

With this, we observe that

$$f_3(x) = h(\vartheta_1)h(\vartheta_2)g(x)^2\exp((\eta(\vartheta_1) + \eta(\vartheta_2))T(x)).$$

Thus, if the following three conditions are fulfilled, P_3 will be from the same exponential family as P_1 and P_2:

$$h(\vartheta_1)h(\vartheta_2) = h(u(\vartheta_1, \vartheta_2)), \tag{15}$$

$$g(x)^2 = g(x), \tag{16}$$

$$\eta(\vartheta_1) + \eta(\vartheta_2) = \eta(v(\vartheta_1, \vartheta_2)), \tag{17}$$

with appropriate functions u and v. Equation (15) describes a normalization constant – if (17) and (16) are true, then (15) will be true, too. Thus, it is sufficient to check conditions (16) and (17). Equation (16) is equivalent to $g(x) = 1$.

Writing a normal distribution in the form of (14) results in $g(x) = 1$ and

$$\eta(\mu, \sigma^2) = (-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}).$$

By writing out equation (17), we see that

$$\eta(\mu_1, \sigma_1^2) + \eta(\mu_2, \sigma_2^2) = (-\frac{1}{2}(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}), \frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}).$$

From this, we immediately see that

$$\frac{1}{\sigma_3^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}.$$

With that, some algebra results in

$$\mu_3 = \left(\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}\right) \cdot \sigma_3^2 = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}.$$

For the gamma distribution, the situation is actually simpler. Again, we have $g(x) = 1$ and

$$\eta(\alpha, \beta) = \left(-\frac{1}{\beta}, \alpha - 1\right).$$

From this, we see that

$$\begin{aligned} \frac{1}{\beta_3} &= \frac{1}{\beta_1} + \frac{1}{\beta_2}, \\ \alpha_3 &= \alpha_1 + \alpha_2 + 1. \end{aligned}$$

For the beta distribution, we finally obtain

$$\eta(\alpha, \beta) = (\alpha - 1, \beta - 1), \quad g(x) = 1,$$

and thus

$$\alpha_3 = \alpha_1 + \alpha_2 + 1.$$

A similar formula holds for β_3.

Determining model posteriors

We used model posterior probabilities to determine whether to prefer a joint fit or the two isolated fits. To do so, we considered a "meta model" with an additional parameter $\mathcal{M}$ that indicated whether the data were generated from the joint model $\mathcal{M} = 1$ or the isolated models $\mathcal{M} = 2$. After having generated samples from the isolated models and from the joint model, we can determine the marginal distribution of $\mathcal{M}$ using Gibbs sampling. Denote the N samples of all parameters from the models estimated with the joint procedure by $\{\theta_i^{(1)}\}_{i=1}^N$, and the samples of parameters from the isolated models by $\{\theta_i^{(2)}\}_{i=1}^N$. To perform Gibbs sampling in the meta model with parameters $\{\mathcal{M}, \theta\}$, we need to consider the full conditional distributions

$$f(\theta|\mathcal{M}, D), \tag{18}$$

$$f(\mathcal{M}|\theta, D) = (f(\mathcal{M} = 1|\theta^{(1)}, D), f(\mathcal{M} = 2|\theta^{(2)}, D)) = (p_1, p_2), \tag{19}$$

where D denotes the total set of data that have been collected. Once we have determined posterior samples for the models under consideration, we can generate samples from equation (18) by resampling from $\{\theta_i^{(k)}\}_{i=1}^N$. In addition, the parameter p_1 and p_2 in equation (19) are given by

$$p_k \propto P(D|\theta^{(k)}, \mathcal{M} = k)P(\theta^{(k)}|\mathcal{M} = k).$$

With the additional requirement that $p_1 + p_2 = 1$, these two can easily be determined for any parameter sample $\theta_*^{(k)}$. By alternating between equations (18) and (19), we can generate a Markov chain in the parameter space of the meta model. The stationary distribution of the Markov chain will approximate the posterior distribution of the meta model. We can further derive an analytical expression for the marginal distribution over $\mathcal{M}$. To this, we consider the transition probabilities

$$P(\mathcal{M}_k \to \mathcal{M}_\ell) = \int P(\mathcal{M}_\ell|\theta)P(\theta|\mathcal{M}_k)\,\mathrm{d}\theta \approx \sum_i \frac{g_\ell(\theta_i^{(k)})}{\sum_j g_j(\theta_i^{(k)})},$$

where g_j is the unnormalized posterior of model j. Denote the transition matrix with entries $T_{k\ell} = P(\mathcal{M}_k \to \mathcal{M}_\ell)$ by $\mathbf{T}$, then the stationary distribution π should satisfy

$$\pi\mathbf{T} = \pi,$$

with

$$\pi_1 + \pi_2 = 1.$$

Thus for a 2×2 transition matrix, we have to solve the following linear equation system

$$\pi_1 T_{11} + \pi_2 T_{12} = \pi_1 \tag{20}$$

$$\pi_1 T_{21} + \pi_2 T_{22} = \pi_2 \tag{21}$$

$$\pi_1 + \pi_2 = 1. \tag{22}$$

Using the first and the last of these equations, we can see that

$$\pi_2 = \frac{1 - T_{11}}{1 + T_{12} - T_{11}}, \quad \pi_1 = 1 - \pi_2 = \frac{T_{22}}{1 + T_{12} - T_{11}}.$$