Generalized Bayesian Inference under Prior-Data Conflict

Gero Walter

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München
vorgelegt am 14.08. 2013

München 2013

Generalized Bayesian Inference under Prior-Data Conflict

Gero Walter

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von
Gero Walter
geboren am 26.06. 1979 in Tübingen

München, den 14.08. 2013

Erstgutachter: Prof. Dr. Thomas Augustin
Zweitgutachter: Prof. Dr. Frank P.A. Coolen
Tag der Disputation: 25.10. 2013

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass diese Dissertation von mir selbstständig und ohne unerlaubte Beihilfe angefertigt ist.

München, den 14. August 2013

Gero Walter

Danksagung

Ich möchte allen einen Dank aussprechen, die mich während Zeit der Entstehung dieser Dissertation begleitet haben.

Thomas Augustin möchte ich herzlich danken für die enge Zusammenarbeit, für Enthusiasmus und Ermutigungen, und ganz besonders dafür, dass er trotz der arbeitsintensiven administrativen Zumutungen, die über ihn hereingebrochen sind, immer für mich Zeit gefunden hat.

Frank Coolen möchte ich danken für die spannende und inspirierende Zusammenarbeit, sowie für seine Einladungen ans Department of Mathematical Sciences in Durham, die mir wissenschaftliche Fortschritte und interessante Einblicke in Kunst, Natur und Kultur ermöglicht haben.

Meinen weiteren Prüfern Christian Heumann und Helmut Küchenhoff sei gedankt für ihre Zeit und ihr Interesse; meinen Koautoren Matthias Troffaes und Manuel Eugster Dank für ihre Hilfe und die erfolgreiche Zusammenarbeit; den Reviewern der in diese Arbeit eingeflossenen Artikel herzlichen Dank für ihre hilfreichen und anregenden Anmerkungen.

Herzlichen Dank an meine Kollegen in der Arbeitsgruppe für die interessanten Diskussionen, Anregungen, Feedback und andere gemeinsame Aktivitäten. Insbesondere meinen „Büromitbewohnern“ möchte ich auch dafür danken, dass sie mein gelegentliches Grummeln ertragen haben.

Den Doktoranden-Kollegen am Institut danke ich für die schöne gemeinsame Zeit, für spontane Hilfe und die Beruhigung, mit all den Doktorarbeits-Problemen nicht alleine zu sein.

Brigitte Maxa, Elke Höfner, Christa Jürgensonn und den Sekretariatsmitarbeiterinnen möchte ich danken, dass sie — wie alle anderen Kollegen — dafür gesorgt haben, dass ich mich am Institut für Statistik zuhause gefühlt habe.

Meiner Familie und ganz besonders Mirjam vielen vielen Dank für all die Unterstützung.

Acknowledgements

I would like to express my gratitude for everyone who stood by me while I was writing this thesis.

Many thanks to Thomas Augustin for his close collaboration, inspiring enthusiasm and constant encouragement. Despite the heavy load of administrative challenges you had to juggle, thank you for always finding time for me.

I would like to thank Frank Coolen for his fascinating and inspiring collaboration and his invitations to the Department of Mathematical Sciences in Durham, that helped me advance scientifically and gave me fascinating new insights into art, nature, and culture.

Thank you also to my other examiners, Christian Heumann and Helmut Küchenhoff, for their time and interest. Many thanks to my coauthors Matthias Troffaes and Manuel Eugster for their help and fruitful teamwork. I would also like to thank the reviewers of the papers I included in this thesis for their very stimulating and helpful comments.

Thank you very much to my work group colleagues for the support, the animated discussions, the feedback and for all the other things we did together. I would especially like to thank my office mates for putting up with my occasional muttering.

To my fellow doctoral students at the department, a warm thank you for the time spent together and for the reassurance provided when the inevitable obstacles and doubts came up.

Thank you Brigitte Maxa, Elke Höfner, Christa Jürgensonn and the secretaries, for making me feel at home at the department, much like all the other members of staff.

To my family and Mirjam, thank you very, very much for all the support.

Contents

List of Figures

List of Tables

Abstract

This thesis is concerned with the generalisation of Bayesian inference towards the use of imprecise or interval probability, with a focus on model behaviour in case of prior-data conflict.

Bayesian inference is one of the main approaches to statistical inference. It requires to express (subjective) knowledge on the parameter(s) of interest not incorporated in the data by a so-called prior distribution. All inferences are then based on the so-called posterior distribution, the subsumption of prior knowledge and the information in the data calculated via Bayes' Rule.

The adequate choice of priors has always been an intensive matter of debate in the Bayesian literature. While a considerable part of the literature is concerned with so-called non-informative priors aiming to eliminate (or, at least, to standardise) the influence of priors on posterior inferences, inclusion of specific prior information into the model may be necessary if data are scarce, or do not contain much information about the parameter(s) of interest; also, shrinkage estimators, common in frequentist approaches, can be considered as Bayesian estimators based on informative priors.

When substantial information is used to elicit the prior distribution through, e.g, an expert's assessment, and the sample size is not large enough to eliminate the influence of the prior, *prior-data conflict* can occur, i.e., information from outlier-free data suggests parameter values which are surprising from the viewpoint of prior information, and it may not be clear whether the prior specifications or the integrity of the data collecting method (the measurement procedure could, e.g., be systematically biased) should be questioned. In any case, such a conflict should be reflected in the posterior, leading to very cautious inferences, and most statisticians would thus expect to observe, e.g., wider credibility intervals for parameters in case of prior-data conflict. However, at least when modelling is based on conjugate priors, prior-data conflict is in most cases completely averaged out, giving a false certainty in posterior inferences.

Here, imprecise or interval probability methods offer sound strategies to counter this issue, by mapping parameter uncertainty over *sets* of priors resp. posteriors instead of over single distributions. This approach is supported by recent research in economics, risk analysis and artificial intelligence, corroborating the multi-dimensional nature of uncertainty and concluding that standard probability theory as founded on Kolmogorov's or de Finetti's framework may be too restrictive, being appropriate only for describing one dimension, namely ideal stochastic phenomena.

The thesis studies how to efficiently describe sets of priors in the setting of samples from an exponential family. Models are developed that offer enough flexibility to express a wide range of (partial) prior information, give reasonably cautious inferences in case of prior-

data conflict while resulting in more precise inferences when prior and data agree well, and still remain easily tractable in order to be useful for statistical practice. Applications in various areas, e.g. common-cause failure modeling and Bayesian linear regression, are explored, and the developed approach is compared to other imprecise probability models.

Zusammenfassung

Das Thema dieser Dissertation ist die Generalisierung der Bayes-Inferenz durch die Verwendung von unscharfen oder intervallwertigen Wahrscheinlichkeiten. Ein besonderer Fokus liegt dabei auf dem Modellverhalten in dem Fall, dass Vorwissen und beobachtete Daten in Konflikt stehen.

Die Bayes-Inferenz ist einer der Hauptansätze zur Herleitung von statistischen Inferenzmethoden. In diesem Ansatz muss (eventuell subjektives) Vorwissen über die Modellparameter in einer sogenannten Priori-Verteilung (kurz: Priori) erfasst werden. Alle Inferenzaussagen basieren dann auf der sogenannten Posteriori-Verteilung (kurz: Posteriori), welche mittels des Satzes von Bayes berechnet wird und das Vorwissen und die Informationen in den Daten zusammenfasst.

Wie eine Priori-Verteilung in der Praxis zu wählen sei, ist dabei stark umstritten. Ein großer Teil der Literatur befasst sich mit der Bestimmung von sogenannten nichtinformativen Prioris. Diese zielen darauf ab, den Einfluss der Priori auf die Posteriori zu eliminieren oder zumindest zu standardisieren. Falls jedoch nur wenige Daten zur Verfügung stehen, oder diese nur wenige Informationen in Bezug auf die Modellparameter bereitstellen, kann es hingegen nötig sein, spezifische Priori-Informationen in ein Modell einzubeziehen. Außerdem können sogenannte Shrinkage-Schätzer, die in frequentistischen Ansätzen häufig zum Einsatz kommen, als Bayes-Schätzer mit informativen Prioris angesehen werden.

Wenn spezifisches Vorwissen zur Bestimmung einer Priori genutzt wird (beispielsweise durch eine Befragung eines Experten), aber die Stichprobengröße nicht ausreicht, um eine solche informative Priori zu überstimmen, kann sich ein Konflikt zwischen Priori und Daten ergeben. Dieser kann sich darin äußern, dass die beobachtete (und von eventuellen Ausreißern bereinigte) Stichprobe Parameterwerte impliziert, die aus Sicht der Priori äußerst überraschend und unerwartet sind. In solch einem Fall kann es unklar sein, ob eher das Vorwissen oder eher die Validität der Datenerhebung in Zweifel gezogen werden sollen. (Es könnten beispielsweise Messfehler, Kodierfehler oder eine Stichprobenverzerrung durch *selection bias* vorliegen.) Zweifellos sollte sich ein solcher Konflikt in der Posteriori widerspiegeln und eher vorsichtige Inferenzaussagen nach sich ziehen; die meisten Statistiker würden daher davon ausgehen, dass sich in solchen Fällen breitere Posteriori-Kredibilitätsintervalle für die Modellparameter ergeben. Bei Modellen, die auf der Wahl einer bestimmten parametrischen Form der Priori basieren, welche die Berechnung der Posteriori wesentlich vereinfachen (sogenannte konjugierte Priori-Verteilungen), wird ein solcher Konflikt jedoch einfach ausgemittelt. Dann werden Inferenzaussagen, die auf einer solchen Posteriori basieren, den Anwender in falscher Sicherheit wiegen.

In dieser problematischen Situation können Intervallwahrscheinlichkeits-Methoden einen fundierten Ausweg bieten, indem Unsicherheit über die Modellparameter mittels *Mengen*

von Prioris beziehungsweise Posterioris ausgedrückt wird. Neuere Erkenntnisse aus Risikoforschung, Ökonometrie und der Forschung zu künstlicher Intelligenz, die die Existenz von verschiedenen Arten von Unsicherheit nahelegen, unterstützen einen solchen Modellansatz, der auf der Feststellung aufbaut, dass die auf den Ansätzen von Kolmogorov oder de Finetti basierende übliche Wahrscheinlichkeitsrechung zu restriktiv ist, um diesen mehrdimensionalen Charakter von Unsicherheit adäquat einzubeziehen. Tatsächlich kann in diesen Ansätzen nur eine der Dimensionen von Unsicherheit modelliert werden, nämlich die der idealen Stochastizität.

In der vorgelegten Dissertation wird untersucht, wie sich Mengen von Prioris für Stichproben aus Exponentialfamilien effizient beschreiben lassen. Wir entwickeln Modelle, die eine ausreichende Flexibilität gewährleisten, sodass eine Vielfalt von Ausprägungen von partiellem Vorwissen beschrieben werden kann. Diese Modelle führen zu vorsichtigen Inferenzaussagen, wenn ein Konflikt zwischen Priori und Daten besteht, und ermöglichen dennoch präzisere Aussagen für den Fall, dass Priori und Daten im Wesentlichen übereinstimmen, ohne dabei die Einsatzmöglichkeiten in der statistischen Praxis durch eine zu hohe Komplexität in der Anwendung zu erschweren. Wir ermitteln die allgemeinen Inferenzeigenschaften dieser Modelle, die sich durch einen klaren und nachvollziehbaren Zusammenhang zwischen Modellunsicherheit und der Präzision von Inferenzaussagen auszeichnen, und untersuchen Anwendungen in verschiedenen Bereichen, unter anderem in sogenannten common-cause-failure-Modellen und in der linearen Bayes-Regression. Zudem werden die in dieser Dissertation entwickelten Modelle mit anderen Intervallwahrscheinlichkeits-Modellen verglichen und deren jeweiligen Stärken und Schwächen diskutiert, insbesondere in Bezug auf die Präzision von Inferenzaussagen bei einem Konflikt von Vorwissen und beobachteten Daten.

1. Introduction

In this introductory chapter, we will first deal with some preliminaries in Section 1.1, where we will give an overview on the contents of this thesis, declare the sources these contents are based on, and present some notational conventions. In Section 1.2, we will discuss some basic fundamentals that frame the work we want to accomplish. There, we will describe the basics of statistical inference using parametric models, and give a brief introduction to the Bayesian approach to statistical inference. Section 1.3 then presents a motivating example, illustrating the advantages in uncertainty modelling that can be gained from using imprecise probability models, thus serving as a preview on the general concepts we will then introduce in the later chapters.

1.1. Preliminaries

1.1.1. Overview

In this thesis, a generalisation of Bayesian inference towards the use of imprecise or interval probability is investigated. A general framework for models based on sets of conjugate priors is established, and some new models within this framework are proposed. These models are then compared to some other models based on sets of priors discussed in the literature, focussing on model behaviour in case of prior-data conflict.

With the fundamentals of Bayesian inference based on parametric distributions and conjugate priors covered in Section 1.2, a motivating example in Section 1.3, considering the reliability analysis problem of common-cause failure modelling, serves to show the potential of generalised Bayesian inference using sets of conjugate priors.

Chapter 2 then gives a general introduction to the methodology of imprecise probability models, presenting the general approach to generalised Bayesian inference with lower previsions or sets of priors. Furthermore, motives for the use of imprecise probability methods are discussed, among which prior-data conflict (encountered already in the motivating example) and weakly informative priors are the central topics guiding our assessments of the specific imprecise probability models covered then in Chapter 3.

There, we first present a general framework for generalised Bayesian inference using sets of conjugate priors, giving a superstructure for some notable imprecise probability models which have been central to the development and application of imprecise probability methods in statistical inference (Section 3.1). Some important favourable inference properties for these models are demonstrated, and a number of models that can be subsumed under

this framework are discussed with respect to the handling of prior-data conflict and the possibility to model weak prior information.

Section 3.2 briefly discusses some alternative models based on sets of priors, and compares these to models of the framework from Section 3.1. The remainder of Chapter 3 reproduces two works that suggest novel models that provide a sophisticated handling of prior-data conflict (Sections 3.3 and 3.5), and gives a short overview on a software implementation (Section 3.4).

Chapter 4 concludes the thesis, giving a summary and discussion of the central achievments, and sketching some opportunities for applications and avenues for further research.

The Appendix (Chapter A) provides some supplemental material. Section A.1 studies prior-data conflict sensitivity in Bayesian linear regression, while Section A.2 describes an informal rationale and some first technical results for a novel approach that, in addition to prior-data conflict sensitivity, leads to favourable behaviour in case of strong agreement between prior information and data.

1.1.2. Sources

This thesis is partly based on previously published works where the author of this thesis was the first or second author. These works, also listed in the Bibliography, are given below.

Augustin, T., G. Walter, and F. Coolen (2013). "Statistical Inference". In: *Introduction to Imprecise Probabilities*. Ed. by T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes. In preparation. Wiley (cit. on pp. 3, 43, 47, 63).

Troffaes, M., G. Walter, and D. Kelly (2013). *A Robust Bayesian Approach to Modelling Epistemic Uncertainty in Common-Cause Failure Models*. Preprint available at `http://arxiv.org/abs/1301.0533`. Accepted for publication at: Reliability Engineering & System Safety (cit. on pp. 3, 19, 23).

Walter, G. (2012). *A Technical Note on the Dirichlet-Multinomial Model — The Dirichlet Distribution as the Canonically Constructed Conjugate Prior*. Tech. rep. 131. Department of Statistics, LMU Munich. URL: `http://epub.ub.uni-muenchen.de/14068/` (cit. on p. 3).

Walter, G. and T. Augustin (2009a). *Bayesian linear regression — different conjugate models and their (in)sensitivity to prior-data conflict*. Tech. rep. 69. Substantially extended version of Walter and Augustin (2010). Department of Statistics, LMU Munich. URL: `http://epub.ub.uni-muenchen.de/11050/1/tr069.pdf` (cit. on pp. 3, 133).

Walter, G. and T. Augustin (2009b). "Imprecision and Prior-data Conflict in Generalized Bayesian Inference". In: *Journal of Statistical Theory and Practice* 3. Reprinted in Coolen-Schrijner, Coolen, Troffaes, Augustin, et al. (2009), pp. 255–271. ISSN: 1559-8616 (cit. on pp. 3, 9, 28, 55, 59, 64–66, 79, 104, 106, 119, 136, 153).

Walter, G., T Augustin, and F. Coolen (2011). "On Prior-Data Conflict in Predictive Bernoulli Inferences". In: *ISIPTA'11: Proceedings of the Seventh International Symposium on Imprecise Probabilities: Theories and Applications*. Ed. by F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger. SIPTA, pp. 391–400. URL: `http://www.sipta.org/isipta11/proceedings/046.html` (cit. on pp. 3, 55, 59, 66, 103).

In detail, these works have been used in this thesis as described below.

In **Chapter 1**, Section 1.2 is based on Augustin, Walter, and Coolen (2013), using §§1.1–1.5 and §4.1. Section 1.2.3.5 is instead based on Walter (2012). Section 1.3 is based on Troffaes, Walter, and Kelly (2013), except Section 1.3.1, which was newly written for this thesis.

In **Chapter 2**, Section 2.1 was newly written for this thesis, except Section 2.1.3, which is based on Augustin, Walter, and Coolen (2013, §§4.2, 4.4). Also, Section 2.1.4 uses parts of Augustin, Walter, and Coolen (2013, §5.2), and Augustin, Walter, and Coolen (2013, §6.1). Section 2.2 was newly devised for this thesis, under use of Augustin, Walter, and Coolen (2013, §2.3) for the last paragraph in Section 2.2.3.1, and Sections 2.2.3.2 and 2.2.3.3. Section 2.2.4 uses parts of Augustin, Walter, and Coolen (2013, §§2.2, 2.4), and a paragraph from Augustin, Walter, and Coolen (2013, §7).

In **Chapter 3**, Section 3.1 is based on Augustin, Walter, and Coolen (2013, §4.3). Section 3.2 was newly written for this thesis, using some minor parts of Augustin, Walter, and Coolen (2013, §§4.2, 4.4). Section 3.3 is a slightly abriged reproduction of Walter and Augustin (2009b), with a minor change of notation towards the one introduced in Section 1.2.3.1. Section 3.4 gives a short overview on a software implementation of the model presented in Section 3.3, and was newly written for this thesis. Section 3.5 consists of Walter, Augustin, and Coolen (2011), again with a slight change of notation for consistency with the rest of the material presented in this thesis. Here, it was also possible to add a few explanatory paragraphs that could not appear in the original publication due to page restrictions.

Chapter 4 was newly written for this thesis.

In the **Appendix** (Chapter A), Section A.1 consists of Walter and Augustin (2009a), which is a substantially extended version of Walter and Augustin (2010), reproduced here with a slight change in notation and some added comments. Section A.2 was instead written newly for this thesis.

1.1.3. Notation

Scalars are denoted by italic letters (x, θ), whereas vectors are denoted by bold italic letters ($\boldsymbol{x}$, $\boldsymbol{\theta}$). Matrices are written in bold regular (i.e., non-italic) uppercase letters, like $\mathbf{X}$, $\mathbf{Z}$, and transposed matrices are marked by a raised uppercase sans serif 'T' ($\mathbf{X}^{\mathsf{T}}$). The trace of a matrix is denoted by $\mathrm{tr}(\mathbf{X})$; unit or identity matrices are denoted by $\mathbf{I}$, sometimes with

an added subscript indicating their size (in the case that the size may be not obvious or in order to emphasise it), such that $\mathbf{I}_p$ denotes a unit matrix of size $p \times p$.

For statistical models, samples, i.e. realisations of random variables, are denoted by lowercase letters (x, $\boldsymbol{x}$), random quantities by uppercase letters (X, $\boldsymbol{X}$); however, as this thesis is mostly concerned with Bayesian methods, the strict distrinction between random variables and 'fixed' quantities as made in frequentist statistics is not maintained throughout most of the thesis.

Sets or spaces are given as calligraphic uppercase letters ($\mathcal{X}$, $\mathcal{Y}$, $\mathcal{M}$). Some special sets are denoted as follows: the real numbers by $\mathbb{R}$, positive real numbers by $\mathbb{R}_{>0}$, nonnegative real numbers by $\mathbb{R}_{\geq 0}$, and the space of q-dimensional tuples of real numbers by $\mathbb{R}^q$.

In the Bayesian setting, prior and posterior probability distribution functions, i.e. densities on parameters, are usually denoted by lowercase letter p; sample model densities (probability distribution functions on observable quantities) are denoted by lowercase letter f: $f(x)$, $p(\theta)$.

This distinction for densities on parameters or samples is not maintained for the associated probability measures and cumulative distribution functions: cumulative distribution functions are denoted by uppercase letter F, e.g., $\mathrm{F}(x) := \int_{-\infty}^{x} f(u)\,\mathrm{d}u$, or $\mathrm{F}(\theta) := \int_{-\infty}^{\theta} p(\psi)\,\mathrm{d}\psi$; probability measures, e.g. for subsets A of a sample space Ω, or a subset Θ_1 of the parameter space Θ, are denoted by uppercase P, i.e. $\mathrm{P}(A) = \sum_{\omega \in A} f(\omega)$ if Ω is countable, or $\mathrm{P}(\Theta_1) = \int_{\Theta_1} p(\theta)\,\mathrm{d}\theta$ for continuous Θ.

Expectation and variance of a random quantity X are denoted by $\mathrm{E}[X]$ and $\mathrm{Var}(X)$, respectively. In a Bayesian setting, quantities identifying the distribution (with respect to which expectation and variance are calculated) are added in the argument, separated by a vertical line, as in $f(x \mid \theta)$, or $\mathrm{E}[\theta \mid n^{(0)}, y^{(0)}]$.

Other notational conventions are declared upon introduction of the concepts they are representing.

1.2. Some Fundamentals

In this section, we will briefly introduce our notion of statistical inference, and discuss models that are used to describe random samples. Then, we will give a short introduction into the basic principles of Bayesian inference based on conjugate priors.

1.2.1. Statistical Inference

Statistical inference is about learning from data. It is basically concerned with inductive reasoning, i.e., establishing a general rule from observations. As is long known as the problem of induction (Hume 2000), it is impossible to justify inductive reasoning by pure reason, and therefore one cannot infer general statements (laws) with absolute truth from single observations. The statistical remedy for this inevitable and fundamental dilemma of any type of inductive reasoning is (postulated, maybe virtual) *randomness* of the sampling process that generates the data. If, and only if, the sample is, or can be understood as, drawn randomly, probability theory allows to quantify the error of statistical propositions concluded from the sample.

Specifically, to model the randomness, a *statistical model* is formulated. It is a tuple $(\mathcal{X}, \mathcal{Q})$, consisting of the *sample space* $\mathcal{X}$, i.e. the domain of the random quantity X under consideration, and a set $\mathcal{Q}$ of probability distributions,[1] collecting all probability distributions that are judged to be potential candidates for the distribution of X. In this setting $\mathcal{Q}$ is called *sampling model* and every element $\mathrm{P} \in \mathcal{Q}$ *(potential) sampling distribution*. The inferential task is to learn the true element $\mathrm{P}^* \in \mathcal{Q}$ from multiple observations of the random process producing X.

1.2.2. Parametric Models

In this thesis, generally, so-called *parameteric models* are considered, where $\mathcal{Q}$ is parametrised by a parameter ϑ of finite dimension, assuming values in the so-called *parameter space* Θ, $\Theta \subseteq \mathbb{R}^q$, $q < \infty$, i.e. $\mathcal{Q} = (\mathrm{P}_\vartheta)_{\vartheta \in \Theta}$. Here, the different sampling distributions P_ϑ are implicitly understood as belonging to a specific class of distributions, the basic type of which is assumed to be known completely (e.g., normal distributions, see Example 1.1 below), and only some characteristics ϑ (e.g., the mean) of the distributions are unknown.

Throughout, we will assume (as is the case for all common applications) that the underlying candidate distributions P_ϑ of the random quantity X are either discrete or absolutely continuous with respect to the Lebesgue measure (see, e.g., Karr 1993, pp. 32f, 38 for some technical details) for every $\vartheta \in \Theta$. Then it is convenient to express every P_ϑ in the discrete case by its *mass function* f_ϑ, with $f_\vartheta(x) := \mathrm{P}_\vartheta(X = x), \forall x \in \mathcal{X}$, and in

[1] Most models of statistical inference rely on σ-additive probability distributions. Therefore, technically, in addition an appropriate (σ-)field $\sigma(\mathcal{X})$, describing the domain of the underlying probability measure, has to be specified. In most applications there are straightforward canonical choices for $\sigma(\mathcal{X})$, and thus σ-fields are not explicitly discussed here.

the continuous case by its *probability density function* (pdf) f_ϑ, where f_ϑ is such that $\mathrm{P}_\vartheta(X \in [a,b]) = \int_a^b f_\vartheta(x)\,\mathrm{d}x$.

An *i.i.d. sample of size* n (where *i.i.d.* abbreviates independent, identically distributed) *based on the parametric statistical model* $(\mathcal{X}, (p_\vartheta)_{\vartheta\in\Theta})$ is a vector

$$\boldsymbol{X} = (X_1, \ldots, X_n)^\mathsf{T}$$

of independent random quantities X_i with the same distribution P_ϑ. Then $\boldsymbol{X}$ is defined on $\mathcal{X}^n$ with probability distribution $\mathrm{P}_\vartheta^{\otimes n}$ as the n-dimensional product measure describing the independent observations. For Bayesian approaches as discussed here, independence is often replaced by exchangeability (see, e.g., Bernardo and Smith 2000, §4.2). $\mathrm{P}_\vartheta^{\otimes n}$ thus has the probability mass or density function

$$f_\vartheta(x_1, \ldots, x_n) := \prod_{i=1}^n f_\vartheta(x_i)\,.$$

The term *sample* is then also used for the concretely observed value(s) $\boldsymbol{x} = (x_1, \ldots, x_n)^\mathsf{T}$.

Now we will present two examples for basic parametric models that will be repeatedly discussed further on.

Example 1.1 (Normal distribution)**.** *A common model for observations that in principle can assume any value on the real line is the* normal distribution *with parameters* μ *and* σ^2*, also called the* Gaussian distribution*. Typical examples for data of this kind are scores in intelligence testing, or technical measurements in general.*[2]

For each observation x_i*,* $i = 1, \ldots, n$*, the normal probability density is*

$$f_{(\mu,\sigma^2)}(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ -\frac{1}{2\sigma^2}(x_i - \mu)^2 \Big\}\,,$$

with the two parameters $\mu \in \mathbb{R}$ *and* $\sigma^2 \in \mathbb{R}_{>0}$ *being in fact the mean and the variance of (the distribution of)* x_i*, respectively. As a shortcut, we write* $x_i \sim \mathrm{N}(\mu, \sigma^2)$*.*

With the independence assumption, the density of $\boldsymbol{x} = (x_1, \ldots, x_n)$ *amounts to*

$$f_{(\mu,\sigma^2)}(\boldsymbol{x}) = \prod_{i=1}^n f_{(\mu,\sigma^2)}(x_i) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \Big\}\,. \tag{1.1}$$

Later on, we restrict considerations to the case where the variance is known to be equal to σ_0^2*, denoted by* $x_i \sim \mathrm{N}(\mu, \sigma_0^2)$*. Inference may thus concern the parameter* μ *directly, or future observations* $x_{n+1}, x_{n+2}, \ldots$ *in a chain of i.i.d. observations.*

[2] The normal distribution is distinguished by the central limit theorem (see, e.g., Karr 1993, §7.3, or Breiman 1968, §9), stating that, under regularity conditions, the distribution of an appropriately scaled sum of n standardized random variables converges to a normal distribution for $n \to \infty$.

Example 1.2 (Multinomial distribution)**.** *The* multinomial distribution *is a common model for samples where only a limited number of distinct values can be observed. These distinct values are often named* categories *(hence the term* categorical data*), and are usually numbered from* 1 *to* k*, without imposing any natural ordering on these values. We have therefore a discrete distribution, giving the probability for observing certain category counts* $(n_1, \ldots, n_k) = \boldsymbol{n}$ *in a sample of* n *observations in total. Thus,* $\sum_{j=1}^{k} n_j = n$*.*

We start the definition of the multinomial distribution by decomposing the collection of n *observations into its constituents, single observations of either of the categories* $1, \ldots, k$*. Such a single observation, often named* multivariate Bernoulli observation*, can be encoded as a vector* $\boldsymbol{x}_i$ *of length* k*, where the* j*-th element,* x_{ij}*, equals* 1 *if category* j *has been observed, and all other elements being* 0*. Given the vectorial parameter* $\boldsymbol{\theta}$ *of length* k*, where the component* θ_j *models the probability of observing category* j *in a single draw (therefore* $\sum_{j=1}^{k} \theta_j = 1$*), the probability for observing* $\boldsymbol{x}_i$ *can be written as*

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \prod_{j=1}^{k} \theta_j^{x_{ij}} \, .$$

Assuming independence, the probability for observing a certain sequence $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ *of* n *observations can thus be written as*

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \propto \prod_{i=1}^{n} \prod_{j=1}^{k} \theta_j^{x_{ij}} = \prod_{j=1}^{k} \theta_j^{\sum_{i=1}^{n} x_{ij}} = \prod_{j=1}^{k} \theta_j^{n_j} \, ,$$

where $n_j = \sum_{i=1}^{n} x_{ij}$ *tells us how often category* j *was observed in the sample.*

For the probability to observe a certain category count $(n_1, \ldots, n_k) = \boldsymbol{n}$*, we have to account for the different possible orderings in* $\mathbf{x}$ *leading to the same count vector* $\boldsymbol{n}$*. Therefore,*

$$f_{\boldsymbol{\theta}}(\boldsymbol{n}) = \binom{n}{n_1, \ldots, n_k} \prod_{j=1}^{k} \theta_j^{n_j} = \frac{n!}{n_1! \cdot \ldots \cdot n_k!} \prod_{j=1}^{k} \theta_j^{n_j} \, . \tag{1.2}$$

As a shortcut, we write $\boldsymbol{n} \sim \mathrm{M}(\boldsymbol{\theta})$*.*

1.2.3. Statistical Inference with the Bayesian Paradigm

As the inference models discussed in this thesis are all based on the Bayesian approach to statistical inference, we will now give a short introduction to the basic principles of Bayesian inference.

The Bayesian approach requires (possibly subjective) knowledge on the parameter ϑ to be expressed by a probability distribution on[3] Θ, with the probability mass or density

[3] Again we implicitly assume that Θ is complemented by an appropriate σ-field $\sigma(\Theta)$.

function $p(\vartheta)$ called *prior distribution*. Indeed, the basic assumption in the Bayesian approach is that *any* prior information about Θ can be sufficiently expressed by a (precise) prior $p(\theta)$.[4] Interpreting the elements $f_\vartheta(\boldsymbol{x})$ of the sampling model as conditional distributions of the sample given the parameter, denoted by $f(\boldsymbol{x} \mid \vartheta)$ and called *likelihood*, turns the problem of statistical inference into a problem of probabilistic deduction, where the *posterior distribution*, i.e. the distribution of the parameter given the sample data, can be calculated by Bayes' Rule.[5] Thus, in the light of the sample $\boldsymbol{x} = (x_1, \ldots, x_n)$, the prior distribution is updated by Bayes' Rule to obtain the posterior distribution with density or mass function

$$p(\vartheta \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \vartheta) \cdot p(\vartheta)\,. \tag{1.3}$$

The posterior distribution is understood as comprising all the information from the sample and the prior knowledge. It therefore underlies all further inferences on the parameter ϑ, like point estimators, interval estimators, or the *posterior predictive distribution*, which is the distribution of further observations based on $p(\vartheta \mid \boldsymbol{x})$ (see Eq. (1.8) below).

1.2.3.1. Regular Conjugate Families of Distributions

Traditional Bayesian inference is frequently based on so-called *conjugate priors* related to a specific likelihood. Such priors have the convenient property that the posterior resulting from (1.3) belongs to the same class of parametric distributions as the prior, and thus only the parameters have to be updated, which makes calculation of the posterior and thus the whole Bayesian inference easily tractable.[6]

Fortunately, there are general results guiding the construction of conjugate priors in several models used most frequently in practice, namely in the case where the sample distribution belongs to a so-called *(regular) canonical exponential family* (e.g., Bernardo and Smith 2000, pp. 202 and 272f). This indeed covers many sample distributions relevant in a statistician's everyday life, like Normal and Multinomial models, Poisson models, or Exponential and Gamma models. After presentation of the general framework, we will discuss its instantiation for the Normal and the Multinomial sampling models as introduced in Examples 1.1 and 1.2 above.

A sample distribution (from now on understood directly as the distribution $\mathrm{P}_\vartheta^{\otimes n}$ of an i.i.d. sample $\boldsymbol{x}$ of size n) is said to belong to the *(regular) canonical exponential family* if

[4] This assumption is refuted most prominently by Walley (1991), whose theory of Bayesian inference without a need for precise priors will be discussed in Section 2.1.

[5] Gillies (1987, 2000) argues that Bayes' Theorem was in fact developed in order to confront the problem of induction as posed by Hume (2000).

[6] This motivation for the use of conjugate priors can be founded on formal arguments. As will be explained below, the posterior expectation of the parameter of interest is actually a linear function of a sufficient statistic of the data and the prior expectation. It turns out that, under some regularity conditions, requiring such linearity of posterior expectation implies the use of conjugate priors (Bernardo and Smith 2000, p. 276).

its density or mass function satisfies the decomposition

$$f(\boldsymbol{x} \mid \vartheta) \propto \exp\left\{\langle \psi, \tau(\boldsymbol{x})\rangle - n\mathbf{b}(\psi)\right\}, \tag{1.4}$$

where $\psi \in \Psi \subset \mathbb{R}^q$ is a transformation of the (possibly vectorial) parameter $\vartheta \in \Theta$, and $\mathbf{b}(\psi)$ a scalar function of ψ (or, in turn, of ϑ). $\tau(\boldsymbol{x})$ is a function of the sample $\boldsymbol{x}$ that fulfills $\tau(\boldsymbol{x}) = \sum_{i=1}^{n} \tau^*(x_i)$, with $\tau^*(x_i) \in \mathcal{T} \subset \mathbb{R}^q$, while $\langle \cdot, \cdot \rangle$ denotes the scalar product.[7]

From these ingredients, a conjugate prior on ψ can be constructed as[8]

$$p(\psi \mid n^{(0)}, y^{(0)})\,\mathrm{d}\psi \propto \exp\left\{n^{(0)}\left[\langle y^{(0)}, \psi\rangle - \mathbf{b}(\psi)\right]\right\}\mathrm{d}\psi\,, \tag{1.5}$$

where $n^{(0)}$ and $y^{(0)}$ are now the parameters by which a certain prior can be specified. We will refer to priors of the form (1.5) as *canonically constructed priors*. The domain of $y^{(0)}$ is $\mathcal{Y}$, the interior of the convex hull of $\mathcal{T}$; the scalar $n^{(0)}$ must take strictly positive values for the prior to be *proper* (i.e., integrable to 1).

An interpretation for these parameters will be given shortly. First, let us calculate the posterior density for ψ. The prior parameters $y^{(0)}$ and $n^{(0)}$ are updated to their posterior values $y^{(n)}$ and $n^{(n)}$ in the following way:

$$y^{(n)} = \frac{n^{(0)}}{n^{(0)} + n} \cdot y^{(0)} + \frac{n}{n^{(0)} + n} \cdot \frac{\tau(\boldsymbol{x})}{n}\,, \qquad\qquad n^{(n)} = n^{(0)} + n\,, \tag{1.6}$$

such that the posterior can be written as

$$p(\psi \mid \boldsymbol{x}, n^{(0)}, y^{(0)}) =: p(\psi \mid n^{(n)}, y^{(n)}) \propto \exp\left\{n^{(n)}\left[\langle y^{(n)}, \psi\rangle - \mathbf{b}(\psi)\right]\right\}\mathrm{d}\psi\,. \tag{1.7}$$

In this setting, $y^{(0)}$ and $y^{(n)}$ can be seen as the parameter describing the main characteristics of the prior and the posterior, and thus we will call them *main prior* and *main posterior parameter*, respectively. $y^{(0)}$ can also be understood as a prior guess for the random quantity $\tilde{\tau}(\boldsymbol{x}) := \tau(\boldsymbol{x})/n$ summarizing the sample, as $\mathrm{E}[\tilde{\tau}(\boldsymbol{x}) \mid \psi] = \nabla\mathbf{b}(\psi)$, where in turn $\mathrm{E}[\nabla\mathbf{b}(\psi) \mid n^{(0)}, y^{(0)}] = y^{(0)}$ (e.g., Bernardo and Smith 2000, Prop. 5.7, p. 275).

Characteristically, $y^{(n)}$ is a weighted average of this prior guess $y^{(0)}$ and the sample 'mean' $\tilde{\tau}(\boldsymbol{x})$, with weights $n^{(0)}$ and n, respectively.[9] Therefore, $n^{(0)}$ can be seen as "prior strength" or "pseudocounts", reflecting the weight one gives to the prior as compared to the sample size n. To make this more explicit, $n^{(0)}$ can be interpreted as the size of an imaginary sample that corresponds to the trust on the prior information in the same way as the sample size of a real sample corresponds to the trust in conclusions based on such a real sample (Walter and Augustin 2009b, p. 258; see Section 3.3.2).

[7] It would be possible, and indeed is often done in the literature, to consider a single observation x in Eq. (1.4) only, as the conjugacy property does not depend on the sample size. However, we find our version with n-dimensional i.i.d. sample $\boldsymbol{x}$ more appropriate for a statistical treatment.

[8] In our notation, $^{(0)}$ denotes prior parameters; $^{(n)}$ posterior parameters.

[9] This weighted average property of Bayesian updating with conjugate priors is an important issue we comment on in Sections 3.1.4 and 3.3.3.2. See also Section A.1.2 for an illustration of this issue for the Normal-Normal and Multinomial-Dirichlet models.

The posterior $p(\psi \mid n^{(n)}, y^{(n)})$ can be transformed back to a distribution on ϑ in order to deal with a commonly known parameter or distribution family for it (as we will do, e.g., in Sections 1.2.3.3 and 1.2.3.5 below). Besides the posterior itself, also the posterior predictive distribution

$$f(\boldsymbol{x}^* \mid \boldsymbol{x}, n^{(0)}, y^{(0)}) = \int f(\boldsymbol{x}^* \mid \psi) p(\psi \mid n^{(n)}, y^{(n)}) \, \mathrm{d}\psi \,, \tag{1.8}$$

the distribution of future samples $\boldsymbol{x}^*$ after having seen a sample $\boldsymbol{x}$, forms the basis for the different inference tasks. Next, we will briefly describe a taxonomy of inference tasks.

1.2.3.2. Inference Tasks

We may structure the different inference tasks by the type of statement one wants to infer from the data. As such, this taxonomy is not exclusive to Bayesian inference methods, and neither to the parametric models considered in Section 1.2.2, but it will be formulated in terms of parameters in a Bayesian setting here.

We distinguish two groups of inferences, namely

1. static conclusions, and
2. predictive conclusions.

Static conclusions refer directly to properties of the sampling model, typically to its parameter(s). The following procedures, which are based directly on the posterior (1.7) in the Bayesian paradigm, are the most common:

1a) *Point estimators*, where a certain parameter value is selected to describe the sample.

1b) *Interval estimators*, where the information is condensed in a certain subset of the parameter space Θ, typically in an interval when $\Theta \subseteq \mathbb{R}$.

1c) *Hypotheses tests*, where the information in the sample is only used to decide between two mutually exclusive statements about parameter(s) called *hypotheses*, usually denoted by H_0 and H_1.

Predictive conclusions instead summarize the information by statements on properties of typical further units, either by describing the whole distribution (as with the posterior predictive (1.8)), or by certain summary measures. Similar to static conclusions, one can thus consider, e.g.,

2a) *Point prediction*, where a certain sample value is selected as the most likely to occur. This is especially useful in the case of discrete sampling distributions, where this procedure amounts to classification of further sample units.

2b) *Interval prediction*, where instead a certain subset of the sample space $\mathcal{X}$ is determined, into which furter sample units are likely to fall. An example are prediction bands in regression analysis.

Both static and predictive conclusions can in fact be formally understood as special cases of *decision making*, where, more generally, the conclusion is to select certain utility maximising or loss minimising acts from a set of possible acts. We will flesh this out to some extent in the examples below.[10]

The concretion of the framework for Bayesian inference with canonical conjugate priors as presented in Section 1.2.3.1 is now demonstrated for the sampling models discussed in Examples 1.1 and 1.2. As a first simple example, we will consider inference in the Binomial model, being the special case of the Multinomial model with only two categories. Then, we will briefly turn to the Normal model, before we present the more complex considerations for the Multinomial model with $k > 2$ categories. The latter model is then used in Section 1.3 for common-cause failure modeling, which will serve as a real-world example illustrating the powers and shortcomings of standard Bayesian inference, ultimately motivating the shift to imprecise Bayesian inference.

1.2.3.3. The Beta-Binomial Model

As the special case of the multinomial model (1.2) with only two categories, we will consider the Binomial model

$$f(\boldsymbol{x} \mid \theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s}\,, \tag{1.9}$$

where $\boldsymbol{x}$, the vector of n observations, is composed of scalar x_i's being either 0 or 1, denoting 'failure' or 'success' in an experiment with these two outcomes. $s = \sum_{i=1}^{n} x_i$ is the number of successes, and the (unknown) parameter $\theta \in (0,1)$ is the probability for 'success' in a single trial. (1.9) can be written in the canonical exponential family form (1.4):

$$f(\boldsymbol{x} \mid \theta) \propto \exp\Big\{ \log\Big(\frac{\theta}{1-\theta}\Big) s - n\big(-\log(1-\theta)\big)\Big\}\,.$$

We have thus $\psi = \log(\theta/(1-\theta))$, $\mathbf{b}(\psi) = -\log(1-\theta)$, and $\tau(\boldsymbol{x}) = s$. The function $\log(\theta/(1-\theta))$ is known as the *logit*, denoted by $\mathrm{logit}(\theta)$.

From these ingredients, a conjugate prior on ψ can be constructed along (1.5), leading here to

$$p\Big(\log\Big(\frac{\theta}{1-\theta}\Big) \mid n^{(0)}, y^{(0)}\Big)\,\mathrm{d}\psi \propto \exp\Big\{n^{(0)}\Big[y^{(0)} \log\Big(\frac{\theta}{1-\theta}\Big) + \log(1-\theta)\Big]\Big\}\,\mathrm{d}\psi\,.$$

This prior, transformed to the parameter of interest θ,

$$p(\theta \mid n^{(0)}, y^{(0)})\,\mathrm{d}\theta \propto \theta^{n^{(0)}y^{(0)}-1}(1-\theta)^{n^{(0)}(1-y^{(0)})-1}\,\mathrm{d}\theta\,,$$

is a Beta distribution with parameters $n^{(0)}y^{(0)}$ and $n^{(0)}(1-y^{(0)})$, in short,

$$\theta \sim \mathrm{Beta}(n^{(0)}y^{(0)}, n^{(0)}(1-y^{(0)}))\,.$$

[10] For more details, see, e.g., Robert (2007, §2), where loss functions typical for statistical settings are described in §2.5, pp. 77ff.

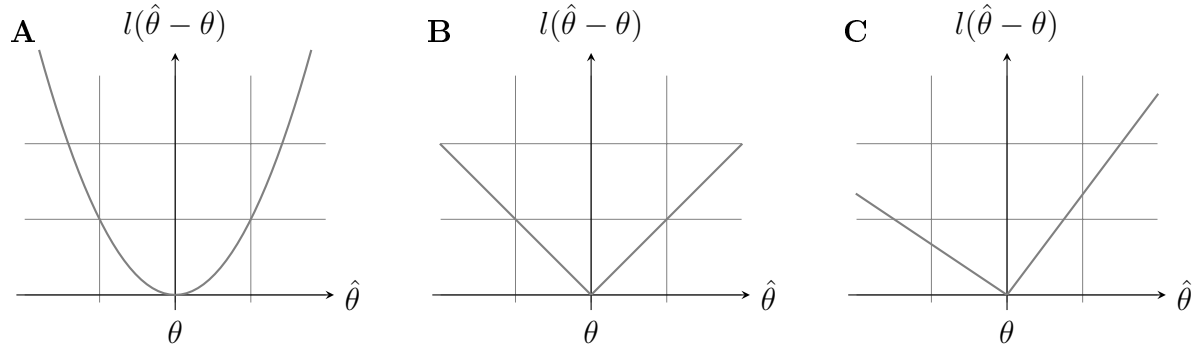

Figure 1.1.: The quadratic (left), absolute (center), and check (right) loss functions.

The combination of a Binomial sampling model with this conjugate Beta prior is called *Beta-Binomial model*. Here, $y^{(0)} = \mathrm{E}[\theta \mid n^{(0)}, y^{(0)}]$ can be interpreted as prior guess of θ, while $n^{(0)}$ governs the concentration of probability mass around $y^{(0)}$, with large values of $n^{(0)}$ giving high concentration of probability mass. Due to conjugacy, the posterior on θ is a Beta$(n^{(n)}y^{(n)}, n^{(n)}(1 - y^{(n)}))$, where the posterior parameters $n^{(n)}$ and $y^{(n)}$ are given by (1.6).

A **point estimator** for θ can be extracted from the posterior distribution $p(\theta \mid \boldsymbol{x})$ by considering Θ as the set of possible acts, and choosing a *loss function*. The loss function l gives a functional form for the severity of deviations of an estimator to its goal; here, it values the distance of a point estimator $\hat{\theta}$ to θ.

The *quadratic loss function* $l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$ values small deviations relatively low, whereas large deviations are given a high weight (see Figure 1.1 A). As can be shown (see, e.g., Casella and Berger 2002, pp. 352f), the quadratic loss function leads to the posterior expectation as the Bayesian point estimator. Here, $\mathrm{E}[\theta \mid \boldsymbol{x}, n^{(0)}, y^{(0)}] = \mathrm{E}[\theta \mid n^{(n)}, y^{(n)}] = y^{(n)}$, and so the posterior expectation of θ is a weighted average of the prior expectation $\mathrm{E}[\theta \mid n^{(0)}, y^{(0)}] = y^{(0)}$ and the sample proportion s/n, with weights $n^{(0)}$ and n, respectively.

Taking the *absolute loss function* $l(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$ leads to the median of the posterior distribution as the point estimator (see Figure 1.1 B). Here, $\mathrm{med}(\theta \mid n^{(n)}, y^{(n)})$ has no closed form solution, and must be determined numerically. More generally, taking the *check function* as the loss function,

$$l(\hat{\theta}, \theta) = \begin{cases} 2q(\hat{\theta} - \theta) & \text{if } \hat{\theta} - \theta \geq 0 \\ 2(q-1)(\hat{\theta} - \theta) & \text{if } \hat{\theta} - \theta < 0 \end{cases},$$

a tilted version of the absolute loss function (see Figure 1.1 C), leads to the quantile $q \in (0, 1)$ of the posterior as point estimate.[11]

[11] The check function is usually given without the factor 2, as it is not relevant for the optimisation. We have included it to make the relation to the absolute loss function more clear; for $q = 0.5$, the check function becomes here indeed the absolute loss function.

The *indicator loss function*

$$l(\hat{\theta}, \theta) = \begin{cases} 0 & |\hat{\theta} - \theta| \leq \epsilon \\ 1 & \text{else} \end{cases},$$

for $\epsilon \to 0$, leads to the maximum of the posterior, often abbreviated as MAP (maximum a posteriori) estimator (see, e.g., Bernardo and Smith 2000, §5.1.5, p. 257, or Robert 2007, §4.1.2, p. 166). For a Beta$(n^{(n)}y^{(n)}, n^{(n)}(1-y^{(n)}))$, the mode is

$$\operatorname{mode} p(\theta \mid n^{(n)}, y^{(n)}) = \frac{n^{(n)}y^{(n)} - 1}{n^{(n)} - 2} = \frac{n^{(0)}y^{(0)} - 1 + s}{n^{(0)} - 2 + n},$$

and thus is a weighted average of the prior mode $\frac{n^{(0)}y^{(0)}-1}{n^{(0)}-2}$ $(n^{(0)} > 2)$ and the sample proportion s/n, with weights $n^{(0)} - 2$ and n, respectively.

Note that asymptotic optimality properties of maximum likelihood estimators (consistency, efficiency) are usually preserved for these Baysian point estimators (e.g., Robert 2007, Note 1.8.4, pp. 48f).

In the Bayesian approach, **interval estimation** is rather simple, as the posterior distribution delivers a direct measure of probability for arbitrary subsets of the parameter space Θ. Mostly, so-called *highest posterior density* (HPD) intervals are considered, where for a given probability level γ the shortest interval covering this probability mass is calculated. For unimodal densities, this is equivalent to finding a threshold α such that the probability mass for the set of all θ with $p(\theta \mid n^{(n)}, y^{(n)}) \geq \alpha$ equals γ, hence the name.[12] For the Beta posterior, a HPD interval for θ must be determined by numeric optimisation. For approximately symmetric (around $\frac{1}{2}$) Beta posteriors, a good approximation is the symmetric credibility interval, delimited by the $\frac{1-\gamma}{2}$- and the $\frac{1+\gamma}{2}$-quantile of the posterior.

The **testing of hypotheses** concerning the parameter of interest θ can be done by comparing posterior probabilities of two (disjunct) subsets of the parameter space. Like in frequentist Neyman-Pearson testing, these are often denoted by Θ_0 and Θ_1, but unlike there, in Bayesian testing the hypotheses $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ play a symmetric role. Therefore, it is also possible to express evidence *in favour of* H_0, whereas frequentist tests are constructed such that they can express conclusive evidence only when H_0 is rejected.[13] However, *point hypotheses*, where one of Θ_0 or Θ_1 consists of a single element of the parameter space only (usually, $\Theta_0 = \{\theta_0\}$),[14] require a special treatment if the prior on Θ is absolutely continuous, as is the case for the priors (1.5) considered here, because then, $\mathrm{P}(\theta \in \Theta_0) = 0$ for any θ_0. Such an inference task can be considered in terms of a problem of *model selection*, where $\theta = \theta_0$ vs. $\theta \neq \theta_0$ decides between two different statistical models,

[12] See, e.g., Bernardo and Smith (2000, §5.1.5, pp. 259f), or Robert (2007, Def. 5.5.3, p. 260).

[13] In Neyman-Pearson testing, only the probability for the *error of the first kind*, denoted by α, of rejecting H_0 although it is true, is set to a low predefined level, whereas the probability for the *error of the second kind*, denoted by β, of accepting H_0 although it is false, may be very large. $1 - \beta$, the probability of correctly rejecting H_0, is also known as the *power* of a test.

[14] Such a testing problem is often called *two-sided* in Neyman-Pearson testing.

to each of which a prior probability is assigned (e.g., Robert 2007, §5.2.4). An overview on Bayesian testing, including a detailed comparison with classical testing procedures, is given in Robert (2007, §5.2–5.4).

Especially in model selection problems, evidence against, or in favor of, a hypothesis is often not expressed in posterior probability for hypotheses, but by means of the so-called *Bayes factor*, arising when considering odds instead of probability:[15]

$$\frac{p(H_1 \mid \boldsymbol{x})}{p(H_0 \mid \boldsymbol{x})} = \frac{f(\boldsymbol{x} \mid H_1)}{f(\boldsymbol{x} \mid H_0)} \cdot \frac{p(H_1)}{p(H_0)}$$

Here, the factor

$$B_{10} := \frac{f(\boldsymbol{x} \mid H_1)}{f(\boldsymbol{x} \mid H_0)},$$

translating prior to posterior odds, is the Bayes factor for comparing H_1 to H_0.[16]

Improper priors are problematic in Bayesian testing (Robert 2007, §5.2.5) and should be avoided for parameters the test decides upon (Kass and Raftery 1995, p. 782).[17] We see this as a strong argument against the use of improper priors. An example where improper priors seem inadequate also for parameter estimation will be given in Section 1.3.4; comments on improper priors from the viewpoint of generalised Bayesian inference are given in Section 3.1.2, item V, and in Section 3.1.3.

The **posterior predictive** distribution, giving the probability for s^* successes in n^* future trials after having seen s successes in n trials, is

$$f(s^* \mid n^{(n)}, y^{(n)}) = \binom{n^*}{s^*} \frac{\mathrm{B}\left(s^* + n^{(n)}y^{(n)}, n^* - s^* + n^{(n)}(1 - y^{(n)})\right)}{\mathrm{B}\left(n^{(n)}y^{(n)}, n^{(n)}(1 - y^{(n)})\right)},$$

known as the *Beta-Binomial distribution*.[18]

1.2.3.4. The Normal-Normal Model

The normal density (1.1), here with the variance σ^2 known to be equal to σ_0^2, also adheres to the exponential family form:

$$f(\boldsymbol{x} \mid \mu, \sigma_0^2) \propto \exp\left\{\frac{\mu}{\sigma_0^2}\sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma_0^2}\right\}.$$

[15] See, e.g., Robert (2007, §5.2.2, Def. 5.2.5, p. 227), or Kass and Raftery (1995, p. 776).

[16] Usually, more than two hypotheses are considered in model selection, by comparing competing models $H_1, H_2, \ldots$ to a null model H_0 with help of Bayes factors $B_{10}, B_{20}, \ldots$.

[17] Walley (1991, §5.5.4 (j)) gives an instructing example for the problems that arise in testing with so-called non-informative priors.

[18] Section 3.5 discusses imprecise Bayesian inference in the Beta-Binomial model, studying the probability of the next observation to be a success in dependence on the number of successes s in n past observations.

So we have here $\psi = \frac{\mu}{\sigma_0^2}$, $\mathbf{b}(\psi) = \frac{\mu^2}{2\sigma_0^2}$, and $\tau^*(x_i) = x_i$. From these ingredients, a conjugate prior can be constructed with (1.5), leading to

$$p\Big(\frac{\mu}{\sigma_0^2} \mid n^{(0)}, y^{(0)}\Big)\,\mathrm{d}\frac{\mu}{\sigma_0^2} \propto \exp\Big\{n^{(0)}\Big(\langle y^{(0)}, \frac{\mu}{\sigma_0^2}\rangle - \frac{\mu^2}{2\sigma_0^2}\Big)\Big\}\,\mathrm{d}\frac{\mu}{\sigma_0^2}\,.$$

This prior, transformed to the parameter of interest μ and with the square completed,

$$p(\mu \mid n^{(0)}, y^{(0)})\,\mathrm{d}\mu \propto \frac{1}{\sigma_0^2}\exp\Big\{-\frac{n^{(0)}}{2\sigma_0^2}(\mu - y^{(0)})^2\Big\}\,\mathrm{d}\mu\,,$$

is a normal distribution with mean $y^{(0)}$ and variance $\frac{\sigma_0^2}{n^{(0)}}$, i.e. $\mu \sim \mathrm{N}(y^{(0)}, \frac{\sigma_0^2}{n^{(0)}})$.[19]

With (1.6), the parameters for the posterior distribution are

$$y^{(n)} = \mathrm{E}[\mu \mid n^{(n)}, y^{(n)}] = \frac{n^{(0)}}{n^{(0)}+n}\cdot y^{(0)} + \frac{n}{n^{(0)}+n}\cdot \bar{x} \tag{1.10}$$

$$\frac{\sigma_0^2}{n^{(n)}} = \mathrm{Var}(\mu \mid n^{(n)}, y^{(n)}) = \frac{\sigma_0^2}{n^{(0)}+n}\,. \tag{1.11}$$

The posterior expectation of μ thus is a weighted average of the prior expectation $y^{(0)}$ and the sample mean $\bar{x}$, with weights $n^{(0)}$ and n, respectively. The effect of the update step on variance is that it decreases by the factor $n^{(0)}/(n^{(0)}+n)$.

Here, all of the three standard choices of loss functions mentioned for the Beta-Binomial model lead to the same **point estimator** $\hat{\mu} = y^{(n)}$, as in normal distributions mean, median, and mode coincide.

As **interval estimation**, the HPD interval can be calculated, due to symmetry of the normal posterior, as $[z^{(n)}_{\frac{1-\gamma}{2}}, z^{(n)}_{\frac{1+\gamma}{2}}]$, where, e.g., $z^{(n)}_{\frac{1-\gamma}{2}}$ is the $\frac{1-\gamma}{2}$-quantile of the normal distribution with mean $y^{(n)}$ and variance $\frac{\sigma_0^2}{n^{(n)}}$.

The **testing of hypotheses** about μ works again by comparing posterior probabilities of two disjunct subsets of the parameter space. Note that the frequentist analogue to such a test is the (one-sided) one-sample Z-test (or Gaussian test).

The **posterior predictive** distribution for n^* future observations denoted by $\boldsymbol{x}^*$ is again a normal distribution, $\boldsymbol{x}^* \mid n^{(n)}, y^{(n)} \sim \mathrm{N}\left(y^{(n)}, \frac{\sigma_0^2}{n^{(n)}}(n^{(n)}+n^*)\right)$, centered at the posterior mean (1.10), and with variance increasing with the posterior variance (1.11) and the number of observations to be predicted.

[19] The conjugate prior if both μ and σ^2 are unknown is the so-called normal-inverse gamma distribution, a combination of the normal distribution above and the inverse gamma distribution (see, e.g., Bernardo and Smith 2000, pp. 119, 431). This prior is a special case of the priors for Bayesian regression discussed in Section A.1, where $\mathbf{X} = (1, \ldots, 1)^\mathsf{T}$ and $\boldsymbol{\beta} = \mu$. When instead of the variance σ^2 the precision $\kappa = 1/\sigma^2$ is considered, this prior can be written as a normal-gamma prior (see, e.g., Bernardo and Smith 2000, pp. 136, 434).

1.2.3.5. The Dirichlet-Multinomial Model

The construction of the canonical conjugate prior by Eq. (1.5) for the Multinomial model $\mathrm{M}(\boldsymbol{\theta})$ with $k > 2$ is more complex than for the case $k = 2$ as covered in Section 1.2.3.3. It is a well-known result that this construction leads to the commonly used Dirichlet prior; however, in the literature the construction is usually not derived in detail.[20] We will cover the construction of $p(\boldsymbol{\psi} \mid n^{(0)}, y^{(0)})$, and also the transformation to $p(\boldsymbol{\theta} \mid n^{(0)}, y^{(0)})$, in more detail here.

We will use the formulation of $\mathrm{M}(\boldsymbol{\theta})$ as the multivariate Bernoulli distribution like in Example 1.2. The distribution of a single multivariate Bernoulli observation is equivalent to a Multinomial distribution with sample size 1, and i.i.d. repetitions of a multivariate Bernoulli distribution lead to the Multinomial distribution. Since i.i.d. repetitions do not interfere with conjugacy (as mentioned in Footnote 7, page 9.), we may construct the canonical conjugate prior by considering a single multivariate Bernoulli observation.

To do without side conditions like $\sum_{j=1}^{k} \theta_j = 1$ as needed in Example 1.2, we will define here as a single multivariate Bernoulli observation distinguishing $k + 1$ categories $j = 0, 1, \ldots, k$ (instead of k categories $j = 1, \ldots, k$ as in Example 1.2) a vector $\boldsymbol{x}$ with k components indexed $1, \ldots, k$ such that

$$\boldsymbol{x} \in \{0,1\}^k \cap \left\{\boldsymbol{x} : \sum_{j=1}^{k} x_j \in \{0,1\}\right\},$$

and $x_0 := 1 - \sum_{j=1}^{k} x_j$.

The parameter vector is treated in the same way, i.e., we consider now $\boldsymbol{\theta}$ with k components $\theta_j, j = 1, \ldots, k$ such that

$$\boldsymbol{\theta} \in (0,1)^k \cap \{\boldsymbol{\theta} : 0 < \sum_{j=1}^{k} \theta_j < 1\},$$

and thus $\theta_0 := 1 - \sum_{j=1}^{k} \theta_j$.

Formulating the density (1.2) accordingly, and rewriting it towards (1.4), we get

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{\theta}) &= \left(\prod_{j=1}^{k} \theta_j^{x_j}\right)\left(1 - \sum_{j=1}^{k} \theta_j\right)^{1-\sum_{j=1}^{k} x_j} = \theta_0 \prod_{j=1}^{k} \left(\frac{\theta_j}{\theta_0}\right)^{x_j} \\ &= \exp\left\{\sum_{j=1}^{k} x_j \ln\left(\frac{\theta_j}{\theta_0}\right) - \big(-\ln(\theta_0)\big)\right\}. \end{aligned}$$

With $\boldsymbol{\psi}$ and $\mathbf{b}(\boldsymbol{\psi})$ derived from the sample model as

$$\psi_j = \ln\left(\frac{\theta_j}{\theta_0}\right),\ j = 1, \ldots, k \qquad \text{and} \qquad \mathbf{b}(\boldsymbol{\psi}) = -\ln(\theta_0),$$

[20] E.g., Quaeghebeur and Cooman (2005, Table 1) tabulate, without proof, priors constructed for a number of sample models. Note that the first version of the paper contains a sign error in the $\mathbf{b}(\boldsymbol{\psi})$ column for both the Binomial (Bernoulli) and the Multinomial (multivariate Bernoulli) sampling model.

the conjugate prior is at first constructed as a density over $\boldsymbol{\psi}$, dropping the upper index $^{(0)}$ in $n^{(0)}$ and the vectorial $\boldsymbol{y}^{(0)}$ for ease of notation:

$$p(\boldsymbol{\psi} \mid n, \boldsymbol{y})\ \mathrm{d}\boldsymbol{\psi} \propto \exp\left\{ n \left[\sum_{j=1}^{k} y_j \ln\left(\frac{\theta_j}{\theta_0}\right) - \big(-\ln(\theta_0)\big) \right] \right\} \mathrm{d}\boldsymbol{\psi}\,.$$

Written as a density over $\boldsymbol{\theta}$, we have

$$p(\boldsymbol{\theta} \mid n, \boldsymbol{y})\ \mathrm{d}\boldsymbol{\theta} \propto \exp\left\{ n \left[\sum_{j=1}^{k} y_j \ln\left(\frac{\theta_j}{\theta_0}\right) - \big(-\ln(\theta_0)\big) \right] \right\} \cdot \left| \det\left(\frac{\mathrm{d}\boldsymbol{\psi}}{\mathrm{d}\boldsymbol{\theta}}\right) \right| \ \mathrm{d}\boldsymbol{\theta}\,,$$

with the elements of the Jacobian matrix $\dfrac{\mathrm{d}\boldsymbol{\psi}}{\mathrm{d}\boldsymbol{\theta}}$ being

$$\begin{aligned}
\frac{\mathrm{d}\psi_i}{\mathrm{d}\theta_i} &= \frac{1}{d\theta_i} \ln\left(\frac{\theta_i}{1-\sum_{j=1}^{k}\theta_j}\right) = \frac{1-\sum_{j=1}^{k}\theta_j}{\theta_i} \cdot \frac{1-\sum_{j=1}^{k}\theta_j + \theta_i}{(1-\sum_{j=1}^{k}\theta_j)^2} = \frac{\theta_0+\theta_i}{\theta_0\theta_i} \\
\frac{\mathrm{d}\psi_h}{\mathrm{d}\theta_i} &= \frac{1}{d\theta_i} \ln\left(\frac{\theta_h}{1-\sum_{j=1}^{k}\theta_j}\right) = \frac{1-\sum_{j=1}^{k}\theta_j}{\theta_h} \cdot \frac{\theta_h}{(1-\sum_{j=1}^{k}\theta_j)^2} = \frac{1}{\theta_0}\,, \quad h \neq i
\end{aligned}$$

Thus,

$$\begin{aligned}
\det\left(\frac{\mathrm{d}\boldsymbol{\psi}}{\mathrm{d}\boldsymbol{\theta}}\right) &= \det\begin{pmatrix} \frac{\theta_0+\theta_1}{\theta_0\theta_1} & \frac{1}{\theta_0} & \cdots & \frac{1}{\theta_0} \\ \frac{1}{\theta_0} & \frac{\theta_0+\theta_2}{\theta_0\theta_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{\theta_0} \\ \frac{1}{\theta_0} & \cdots & \frac{1}{\theta_0} & \frac{\theta_0+\theta_k}{\theta_0\theta_k} \end{pmatrix} \\
&= \left(\frac{1}{\theta_0}\right)^k \det\begin{pmatrix} \frac{\theta_0}{\theta_1}+1 & 1 & \cdots & 1 \\ 1 & \frac{\theta_0}{\theta_2}+1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & \frac{\theta_0}{\theta_k}+1 \end{pmatrix} \\
&\overset{*}{=} \left(\frac{1}{\theta_0}\right)^k \prod_{j=1}^{k} \frac{\theta_0}{\theta_j} \cdot \left(1 + \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \frac{\theta_1}{\theta_0} & & 0 \\ & \ddots & \\ 0 & & \frac{\theta_k}{\theta_0} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\right) \\
&= \prod_{j=1}^{k} \frac{1}{\theta_j} \cdot \left(1 + \sum_{i=1}^{k} \frac{\theta_i}{\theta_0}\right) = \left(\prod_{j=1}^{k} \frac{1}{\theta_j}\right) \frac{\theta_0 + \sum_{i=1}^{k}\theta_i}{\theta_0} \\
&= \left(\prod_{j=1}^{k} \frac{1}{\theta_j}\right) \frac{1}{\theta_0} = \prod_{j=0}^{k} \frac{1}{\theta_j}\,,
\end{aligned}$$

where equality * holds by the theorem

$$\det(\mathbf{A} + \boldsymbol{a}\,\boldsymbol{a}^{\mathsf{T}}) = \det(\mathbf{A})(1 + \boldsymbol{a}^{\mathsf{T}}\mathbf{A}^{-1}\boldsymbol{a})$$

for all column vectors $\boldsymbol{a}$ and appropriately sized, invertible matrices $\mathbf{A}$ (Rao et al. 2008, Theorem A 16 (x), Appendix A3, p. 494).

With $\det\left(\frac{\mathrm{d}\psi}{\mathrm{d}\boldsymbol{\theta}}\right) = \prod_{j=0}^{k} \frac{1}{\theta_j}$, we get

$$\begin{aligned}
p(\boldsymbol{\theta} \mid n, \boldsymbol{y}) &\propto \exp\left\{ n \left[\sum_{j=1}^{k} y_j \ln\left(\frac{\theta_j}{\theta_0}\right) - \big(-\ln(\theta_0)\big) \right] \right\} \cdot \left| \prod_{j=0}^{k} \frac{1}{\theta_j} \right| \\
&= \exp\left\{ n \left[\sum_{j=1}^{k} y_j \Big(\ln(\theta_j) - \ln(\theta_0) \Big) + \ln(\theta_0) \right] - \sum_{j=0}^{k} \ln(\theta_j) \right\} \\
&= \exp\left\{ n \left[\sum_{j=1}^{k} y_j \ln(\theta_j) + \ln(\theta_0) \underbrace{\left(1 - \sum_{j=1}^{k} y_j \right)}_{=:y_0} \right] - \sum_{j=0}^{k} \ln(\theta_j) \right\} \\
&= \exp\left\{ n \left[\sum_{j=0}^{k} y_j \ln(\theta_j) \right] - \sum_{j=0}^{k} \ln(\theta_j) \right\} \\
&= \exp\left\{ \sum_{j=0}^{k} (n\,y_j - 1) \ln(\theta_j) \right\} = \exp\left\{ \sum_{j=0}^{k} \ln\left(\theta_j^{n\,y_j - 1}\right) \right\} \\
&= \prod_{j=0}^{k} \theta_j^{n\,y_j - 1}\,,
\end{aligned}$$

which is the core of a Dirichlet density over $\boldsymbol{\theta}$. Therefore, the Dirichlet distribution is the canonically constructed conjugate prior to the multivariate Bernoulli. Due to the considerations from the beginning of this section, we see that the Dirichlet distribution is the canonically constructed conjugate prior also to the Multinomial sample model with arbitrary sample sizes.

The Dirichlet distribution can be seen as a direct generalisation of the Beta distribution, and we will speak of the *Dirichlet-Multinomial model* as the analogue to the Beta-Binomial model from Section 1.2.3.3.

Returning to the notation from Example 1.2, where k categories $j = 1, \ldots, k$ are considered, we have thus

$$p(\boldsymbol{\theta} \mid n^{(0)}, \boldsymbol{y}^{(0)}) \propto \prod_{j=1}^{k} \theta_j^{n^{(0)} y_j^{(0)} - 1} \tag{1.12}$$

as the prior. The vectorial $\boldsymbol{y}^{(0)}$ is an element of the interior of the $k-1$-dimensional unit simplex Δ, thus $\forall j\ y_j^{(0)} \in (0,1)$, $\sum_{j=1}^{k} y_j^{(0)} = 1$, in short $\boldsymbol{y}^{(0)} \in \mathrm{int}(\Delta)$. Here, the main

posterior parameter is calculated as

$$y_j^{(n)} = \frac{n^{(0)}}{n^{(0)}+n} y_j^{(0)} + \frac{n}{n^{(0)}+n} \cdot \frac{n_j}{n}\,, \quad j = 1, \ldots, k,$$

and is thus again a weighted average of the main prior parameter (which can be interpreted as prior guess for $\boldsymbol{\theta}$, as $\mathrm{E}[\boldsymbol{\theta} \mid n^{(0)}, \boldsymbol{y}^{(0)}] = \boldsymbol{y}^{(0)}$) and the fractions of observations in each category, with weights $n^{(0)}$ and n, respectively. $n^{(0)}$ again governs the concentration of probability mass around $y^{(0)}$, with larger values of $n^{(0)}$ leading to higher concentrations.

With these tangible intuitions for $n^{(0)}$ and $\boldsymbol{y}^{(0)}$, we will denote the Dirichlet prior directly in terms of the canonical parameters, i.e. by $\boldsymbol{\theta} \sim \mathrm{Dir}(n^{(0)}, \boldsymbol{y}^{(0)})$. The canonical parameters relate to the commonly used parameter, often denoted by $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$, not to be confused with $\boldsymbol{\alpha}$ as discussed in Section 1.3.3, via

$$y_j^{(0)} = \frac{\alpha_j}{\sum_{l=1}^k \alpha_l} \qquad\qquad n^{(0)} = \sum_{l=1}^k \alpha_l\,.$$

We will now present a problem in reliability analysis for which the most common model relies on the Dirichlet-Multinomial model. The issues encountered in this application that is based on Troffaes, Walter, and Kelly (2013) will motivate the use of sets of distributions as prior models, illustrating the advantages in uncertainty modelling that can be gained from using imprecise probability models.[21] The models discussed there will then be integrated into a general framework of inference using sets of canonical conjugate priors in Section 3.1.

[21] Basic aspects of the theory of imprecise probability, and further motivations for using it, will be given in Chapter 2.

1.3. Dirichlet-Multinomial Model for Common-Cause Failure

In reliability analysis of systems with redundant components, *common-cause failure* refers to simultaneous failure of several redundant components due to a common or shared root cause, like extreme environmental conditions (e.g., fire, flood, or earthquake) (Høyland and Rausand 1994, p. 325). It has been recognized as the dominant factor to the unreliability of redundant systems, and its modeling has become an important part of reliability analysis following the Reactor Safety Study (U.S. Nuclear Regulatory Commission 1975), which was prepared in the wake of the Three Mile Island accident, where a partial core meltdown in a nuclear power plant took place (e.g., Walker 2005).

1.3.1. An Example for the Need to Model Common-Cause Failure

For a nuclear power plant, a common-cause failure analysis can relate to, e.g., the diesel generators that, if in an emergency the off-site power supply is cut off, provide the electricity to power the emergency core cooling systems. Emergency core cooling systems are needed to transfer away residual heat emitted from the core after shutdown, in case the normal heat removal process[22] is not available. When residual heat is not removed, it can overheat the core, which could lead to a (partial) core meltdown, and subsequently to a possibly catastrophic release of radioactive material.[23]

Due to their critical role for the functioning of emergency cooling systems, and thus for the safety systems of nuclear power plants in general (Ishack et al. 1986, pp. 1, 4), there are usually several diesel generators installed in a nuclear power plant, each of which can supply enough energy to power the cooling systems on its own. The number of diesel generators installed is typically in the range of two to four per reactor (see, e.g., Ishack et al. (1986), pp. 121, 145). If the nuclear power plant has several reactors, diesel generators may be shared between reactors (Chopra et al. 2004, p. 7). The Fukushima Daiichi nuclear disaster is a recent example of an accident involving common cause failure of diesel generators. In this case, all 12 available diesel generators at reactors 1 to 6 ceased to function due to a tsunami wave flooding the rooms where they were installed (Weightman et al. 2011, p. 31). The tsunami wave had been caused by the Tōhoku earthquake (e.g., Ritsema, Lay, and Kanamori 2012), which had prompted the reactors to shutdown automatically, and in

[22] Heat from the core is transferred to steam generators, producing steam that drives turbines linked to generators producing electricity. Usually, the depressurized steam exiting the turbines is then condensated to water, which is fed back into the steam generators.

[23] The residual heat that is present in the core of a nuclear power plant after shutdown of the nuclear chain reaction is called *decay heat*. It results from secondary decay processes, i.e. from the decay of fission products produced during normal operation of the power plant. Although comprising only a small fraction of the energy output during normal operation (where the energy stems from the primary fission process (United States Department of Energy 1993, Module 4, p. 33)), depending on the design of the reactor, the decay heat may be enough to damage the core of a reactor significantly (U.S. Nuclear Regulatory Commission 1975, pp. VIII-9 and VIII-25f).

doing so, switching to the diesel generators for power supply.[24]

The arguably most widely used model for common-cause failure is the so-called *Basic Parameter Model*. The *alpha-factor* parametrisation of this model uses a multinomial distribution as its aleatory model for observed failures (Mosleh et al. 1988). As seen in Section 1.2.3.5, the conjugate prior to the multinomial model is the Dirichlet distribution. In the standard Bayesian approach, the analyst specifies the parameters $(n^{(0)}, \boldsymbol{y}^{(0)})$ of a precise Dirichlet distribution to model epistemic uncertainty in the alpha-factors, which are the parameters of the multinomial sample model.

We will first describe the Basic Parameter Model in its standard form, and subsequently present its reparametrisation in terms of alpha-factors.

1.3.2. The Basic Parameter Model

Consider a system that consists of k components. Throughout, we make the following standard assumptions: (i) repair is immediate, and (ii) failures follow a Poisson process.

For simplicity, we assume that all k components are exchangeable, in the sense that they have identical failure rates. More precisely, we assume that all events involving *exactly* j components failing have the same failure rate, which we denote by q_j. This model is called the *basic parameter model*, and we write $\boldsymbol{q}$ for $(q_1, \ldots, q_k)$.

For example, if we have three components, A, B, and C, then the rate at which we see only A failing is equal to the rate at which we see only B failing, and is also equal to the rate at which we see only C failing; this failure rate is q_1. Moreover, the rate at which we observe only A and B jointly failing is equal to the rate at which we observe only B and C jointly failing, and also equal to the rate at which we observe only A and C jointly failing; this failure rate is q_2. The rate at which we see all three components jointly failing is q_3.

In case of k identical components without common-cause failure modes, thus each failing independently at rate λ, we would have[25]

$$q_1 = \lambda \qquad \text{and} \qquad q_j = 0 \text{ for } j \geq 2.$$

The fact that we allow arbitrary values for the q_j reflects the lack of independence, and whence, our modelling of common-cause failures. At this point, it is worth noting that we do not actually write down a statistical model for all possible common-cause failure modes—we could do so if this information was available, and in fact, this could render the basic parameter model obsolete, and allow for more detailed inferences. In essence, the basic parameter model allows us to statistically model lack of independence between component failures, without further detail as to where dependencies arise from: all failure modes are lumped together, so to speak.

It is useful to note that it is possible, and sometimes necessary, to relax the exchangeability assumption to accommodate specific asymmetric cases. For example, when components

[24] All six off-site power lines were cut off, also due to the earthquake (Weightman et al. 2011, p. 31).

[25] This is due to our Poisson assumption, and the assumption of immediate repair: independent Poisson processes never generate events simultaneously when we observe failure times precisely.

are in different state of health, single failures would clearly not have identical failure rates. Because the formulas become a lot more complicated, we stick to the exchangeable case here.[26]

Clearly, to answer typical reliability questions, such as for instance "what is the probability that two or more components fail in the next month?", we need $\boldsymbol{q}$. In practice, the following three issues commonly arise. First, $\boldsymbol{q}$ is rarely measured directly, as failure data is often collected only per component. Secondly, when direct data about joint failures is available, typically, this data is sparse, because events involving more than two components failing simultaneously are usually quite rare. Thirdly, there are usually two distinct sources of failure data, one usually very large data set related to failure per component, and one usually much smaller data set related to joint failures. For these reasons, it is sensible to reparametrise the model in terms of parameters that can be more easily estimated, as follows.

1.3.3. The Alpha-Factor Model

The alpha-factor parametrisation of the basic parameter model (Mosleh et al. 1988) starts out with considering the total failure rate of a component q_t, which could involve failure of any number of components, that is, this is the rate obtained by looking at just one component, ignoring everything else. Clearly,

$$q_t = \sum_{j=1}^{k} \binom{k-1}{j-1} q_j. \tag{1.13}$$

For example, again consider a three component system, A, B, and C. The rate at which A fails is then the rate at which only A fails (q_1), plus the rate at which A and B, or A and C fail ($2q_2$), plus the rate at which all three components fail (q_3).

Next, the alpha-factor model introduces α_j—the so-called alpha-factor—which denotes the probability of *exactly* j of the k components failing given that failure occurs; in terms of relative frequency, α_j is the fraction of failures that involve *exactly* j failed components. We write $\boldsymbol{\alpha}$ for $(\alpha_1, \ldots, \alpha_k)$. Clearly,

$$\alpha_j = \frac{\binom{k}{j} q_j}{\sum_{\ell=1}^{k} \binom{k}{\ell} q_\ell}. \tag{1.14}$$

For example, again consider A, B, and C. Then the rate at which exactly one component fails is $3q_1$ (as we have three single components, each of which failing with rate q_1), the rate at which exactly two components fail is $3q_2$ (as we have three combinations of two components, each combination failing with rate q_2), and the rate at which all components fail is q_3. Translating these rates into fractions, we arrive precisely at Eq. (1.14).

[26] A discussion of the asymmetric case, i.e., without the assumption of exchangeability of components, can be found in Troffaes and Blake (2013).

It can be shown that (e.g., Mosleh et al. 1988, p. C-10f)

$$q_j = \frac{1}{\binom{k-1}{j-1}} \frac{j\alpha_j}{\sum_{\ell=1}^{k} \ell\alpha_\ell}\, q_t\,. \tag{1.15}$$

Eqs. (1.13), (1.14), and (1.15) establish a one-to-one link between the basic parameter model (with parameters $\boldsymbol{q}$) and the alpha-factor model (with parameters q_t, $\boldsymbol{\alpha}$). The benefit of the alpha-factor model over the basic parameter model lies in its distinction between the total failure rate of a component q_t, for which there is generally a lot of information, and common-cause failures modelled by $\boldsymbol{\alpha}$, for which there is generally very little information.

Next, we will show how we can use the Dirichlet-Multinomial model from Section 1.2.3.5 to estimate the alpha factors $\boldsymbol{\alpha}$. Estimating the marginal failure rate q_t is possible in the framework from Section 1.2.3.1 as well, when failures are assumed to follow a Poisson process. Then, the Gamma prior usually specified for the intensity parameter is of the form (1.5). We will not cover such a *Gamma-Poisson model* here further, but details for this model can be found in Troffaes, Walter, and Kelly (2013). There, also the combination of the Dirichlet-Multinomial model for $\boldsymbol{\alpha}$ and the Gamma-Poisson model for q_t to make inferences on $\boldsymbol{q}$ is discussed.

1.3.4. Dirichlet Prior for Alpha-Factors

Suppose that we have observed a sequence of n failure events, where we have counted the number of components involved with each failure event, say n_j of the n observed failure events involved *exactly* j failed components. As in Example 1.2, we write $\boldsymbol{n}$ for $(n_1, \ldots, n_k)$. With the alpha-factors taking the role of the parameter $\boldsymbol{\theta}$ in (1.2), the multinomial sample model for $\boldsymbol{n}$ has the form

$$f(\boldsymbol{n} \mid \boldsymbol{\alpha}) \propto \prod_{j=1}^{k} \alpha_j^{n_j}\,.$$

As mentioned already, in this application of multinomial sample model, the n_j are usually very low for $j \geq 2$, with zero being quite common for larger j. In such cases, standard techniques such as maximum likelihood for estimating the alpha-factors fail to produce sensible inferences. For any inference to be reasonably possible, it has been recognized (Mosleh et al. 1988) that we have to rely on epistemic information, that is, information which is not just described by the data.

A standard way to include epistemic information in the model is through specification of the conjugate Dirichlet prior (1.12) for the alpha-factors (Mosleh et al. 1988), i.e., using the Dirichlet-Multinomial model for inferences about $\boldsymbol{\alpha}$:

$$p(\boldsymbol{\alpha} \mid n^{(0)}, \boldsymbol{y}^{(0)}) \propto \prod_{j=1}^{k} \alpha_j^{n^{(0)} y_j^{(0)} - 1}$$

To recap, $n^{(0)} > 0$ and $\boldsymbol{y}^{(0)} \in \Delta$, where Δ is the $(k-1)$-dimensional unit simplex:

$$\Delta = \left\{ (y_1^{(0)}, \ldots, y_k^{(0)}) \colon y_1^{(0)} \geq 0, \ldots, y_k^{(0)} \geq 0, \sum_{j=1}^{k} y_j^{(0)} = 1 \right\}$$

Calculating the posterior density for $\boldsymbol{\alpha}$ then results in

$$p(\boldsymbol{\alpha} \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}) = p(\boldsymbol{\alpha} \mid n^{(n)}, \boldsymbol{y}^{(n)}) \propto \prod_{j=1}^{k} \alpha_j^{n^{(n)} y_j^{(n)} - 1} = \prod_{j=1}^{k} \alpha_j^{n^{(0)} y_j^{(0)} + n_j - 1} .$$

Of typical interest is for instance the posterior expectation of the probability α_j of observing j of the k components failing due to a common cause given that failure occurs,

$$\begin{aligned} \mathrm{E}[\alpha_j \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] &= \int_{\Delta} \alpha_j p(\boldsymbol{\alpha} \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}) \, \mathrm{d}\boldsymbol{\alpha} \\ &= \frac{n^{(0)} y_j^{(0)} + n_j}{n^{(0)} + n} = \frac{n^{(0)}}{n^{(0)} + n} y_j^{(0)} + \frac{n}{n^{(0)} + n} \frac{n_j}{n} , \end{aligned} \tag{1.16}$$

where $n = \sum_{j=1}^{k} n_j$ is the total number of observations.

Eq. (1.16) provides the interpretation for the hyperparameters $n^{(0)}$ and $\boldsymbol{y}^{(0)}$ as mentioned in Section 1.2.3.5, discussed here in terms of the alpha-factor model:

- If $n = 0$, then $\mathrm{E}[\alpha_j \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = y_j^{(0)}$, so $y_j^{(0)}$ is the prior expected chance of observing j of the k components failing due to a common cause, given that failure occurs.
- $\mathrm{E}[\alpha_j \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}]$ is a weighted average of $y_j^{(0)}$ and n_j/n (the proportion of j-component failures in the n observations), with weights $n^{(0)}$ and n, respectively. The parameter $n^{(0)}$ thus determines how much data is required for the posterior to start moving away from the prior. If $n \ll n^{(0)}$ then the prior will weigh more; if $n = n^{(0)}$, then prior and data will weigh equally; and if $n \gg n^{(0)}$, then the data will weigh more. In particular, $\mathrm{E}[\alpha_j \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = y_j^{(0)}$ if $n = 0$ (as already mentioned), and $\mathrm{E}[\alpha_j \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] \to \frac{n_j}{n}$ as $n \to \infty$.

1.3.5. Usual Handling of Epistemic Information for Alpha-Factors

Crucial to reliable inference in the alpha-factor model is proper modelling of epistemic information about failures, which is in the above approach expressed through the choice of $(n^{(0)}, \boldsymbol{y}^{(0)})$. By means of an example taken from Kelly and Atwood (2011), we will demonstrate that posterior inferences rely crucially on this choice of hyperparameters, particularly when faced with zero counts.

Consider a system with four redundant components ($k = 4$). The probability of j out of k failures, given that failure has happend, was denoted by α_j. We assume that the analyst's prior expectation $\mu_{\mathrm{spec},j}$ for each α_j is:

$$\mu_{\mathrm{spec},1} = 0.950 \qquad \mu_{\mathrm{spec},2} = 0.030 \qquad \mu_{\mathrm{spec},3} = 0.015 \qquad \mu_{\mathrm{spec},4} = 0.005 \tag{1.17}$$

We have 36 observations, in which 35 showed one component failing, and 1 showed two components failing:

$$n_1 = 35 \qquad n_2 = 1 \qquad n_3 = 0 \qquad n_4 = 0$$

Atwood (1996) studied priors for the binomial model which maximise entropy (and whence, are 'non-informative'[27]) whilst constraining the mean to a specific value. Although these priors are not conjugate, Atwood (1996) showed that they can be well approximated by Beta distributions, which are conjugate. Kelly and Atwood (2011) applied this approach to the multinomal model with conjugate Dirichlet priors, by choosing a constrained non-informative prior for the marginals of the Dirichlet—which are Beta. This leads to an over-specified system of equalities, which can be solved via least-squares optimisation.

For the problem we are interested in, $\mu_{\text{spec},1}$ is close to 1. In this case, the solution of the least-squares problem turns out to be close to:

$$\begin{aligned} y_j^{(0)} &= \mu_{\text{spec},j} \text{ for all } j \in \{1, \ldots, k\} \\ n^{(0)} &= \frac{1}{2(1 - \mu_{\text{spec},1})} \end{aligned} \tag{1.18}$$

For our example, this means that $n^{(0)} = 10$ (Kelly and Atwood 2011, p. 400, §3). Using Eq. (1.16), under this prior (Kelly and Atwood 2011, p. 401, §3.1):

$$\mathrm{E}[\alpha_1 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{9.5 + 35}{10 + 36} = 0.967 \qquad \mathrm{E}[\alpha_2 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{0.3 + 1}{10 + 36} = 0.028$$
$$\mathrm{E}[\alpha_3 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{0.15 + 0}{10 + 36} = 0.003 \qquad \mathrm{E}[\alpha_4 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{0.05 + 0}{10 + 36} = 0.001$$

Kelly and Atwood (2011, p. 402, §4) compare these results against a large number of other choices of priors, and note that the posterior resulting from Eq. (1.18) seems too strongly influenced by the prior, particularly in the presence of zero counts. For instance, the uniform prior is a Dirichlet distribution with hyperparameters $y_j^{(0)} = 0.25$ and $n^{(0)} = 4$, which gives:

$$\mathrm{E}[\alpha_1 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{1 + 35}{4 + 36} = 0.9 \qquad \mathrm{E}[\alpha_2 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{1 + 1}{4 + 36} = 0.05$$
$$\mathrm{E}[\alpha_3 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{1 + 0}{4 + 36} = 0.025 \qquad \mathrm{E}[\alpha_4 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{1 + 0}{4 + 36} = 0.025$$

Jeffrey's prior is again a Dirichlet distribution with hyperparameters $y_j^{(0)} = 0.125$ and $n^{(0)} = 4$, which gives:

$$\mathrm{E}[\alpha_1 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{0.5 + 35}{4 + 36} = 0.8875 \qquad \mathrm{E}[\alpha_2 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{0.5 + 1}{4 + 36} = 0.0375$$

[27] We will comment on so-called *non-informative* priors in Section 3.1.2, item V, and in Section 3.1.3.

$$\mathrm{E}[\alpha_3 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{0.5+0}{4+36} = 0.0125 \qquad \mathrm{E}[\alpha_4 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{0.5+0}{4+36} = 0.0125$$

The degree of variation in the posterior under different priors is evidently somewhat alarming. In the next section, we thus aim to robustify the model by using sets of priors instead of a single Dirichlet prior to model epistemic information. We will argue that bounds, rather than precise values, should be assigned for $(n^{(0)}, \boldsymbol{y}^{(0)})$, and that in this application, assigning an interval for the learning parameter $n^{(0)}$ is especially important. This effects to using the imprecise probability models that are at the core of this thesis. A systematic presentation of these models, together with more arguments for employing them in important statistical inference tasks, is given in Chapter 3.

1.3.6. Cautious Epistemic Information for Alpha-Factors

As a preview on the general concept of Bayesian inference with sets of conjugate priors, we will now demonstrate how inferences on alpha-factors can benefit from considering sets of priors based on sets of hyperparameters.

For Bayesian inference in the Dirichlet-Multinomial model when no prior information is available, Walley (1996a) proposes as a so-called *near-ignorance prior*[28] a set of Dirichlet priors, with hyperparameters constrained to the set:

$$\mathcal{H} = \{(n^{(0)}, \boldsymbol{y}^{(0)})\colon \boldsymbol{y}^{(0)} \in \Delta\}$$

for some fixed value of $n^{(0)}$, which determines the learning speed of the model (Walley 1991, p. 218, §5.3.2; Walley 1996a, p. 9, §2.3).[29]

When prior information is available, more generally, we may assume that we can specify a subset $\mathcal{H}$ of $(0, +\infty) \times \Delta$. Following Walley's suggestions (Walley 1991, p. 224, §5.4.3; Walley 1996a, p. 32, §6), we take

$$\mathcal{H} = \left\{(n^{(0)}, \boldsymbol{y}^{(0)})\colon n^{(0)} \in [\underline{n}^{(0)}, \overline{n}^{(0)}],\, \boldsymbol{y}^{(0)} \in \Delta,\, y_j^{(0)} \in [\underline{y}_j^{(0)}, \overline{y}_j^{(0)}]\right\} \tag{1.19}$$

where the analyst has to specify the bounds $[\underline{y}_j^{(0)}, \overline{y}_j^{(0)}]$ for each $j \in \{1, \ldots, k\}$, and $[\underline{n}^{(0)}, \overline{n}^{(0)}]$.[30]

As $y^{(n)}$ is linear in $y^{(0)}$ (see Eq. (1.6)), the posterior lower and upper expectations of α_j are:

$$\underline{\mathrm{E}}[\alpha_j \mid \mathcal{H}, \boldsymbol{n}] = \min\left\{\frac{\underline{n}^{(0)}\underline{y}_j^{(0)} + n_j}{\underline{n}^{(0)} + n}, \frac{\overline{n}^{(0)}\underline{y}_j^{(0)} + n_j}{\overline{n}^{(0)} + n}\right\} = \begin{cases} \frac{\underline{n}^{(0)}\underline{y}_j^{(0)} + n_j}{\underline{n}^{(0)} + n} & \text{if } \underline{y}_j^{(0)} \geq n_j/n \\ \frac{\overline{n}^{(0)}\underline{y}_j^{(0)} + n_j}{\overline{n}^{(0)} + n} & \text{if } \underline{y}_j^{(0)} \leq n_j/n \end{cases} \tag{1.20}$$

[28] See Sections 3.1.2 and 3.1.3 for a more detailed discussion of the IDM and near-ignorance priors.

[29] Our notation relates to Walley's as $n^{(0)} \leftrightarrow s$, $y_j^{(0)} \leftrightarrow t_j$.

[30] We will discuss models based on this type of prior parameter set in Sections 3.3 and 3.5.2.2. Different choices of parameter sets are discussed in general in Section 3.1.1.

$$\overline{\mathrm{E}}[\alpha_j \mid \mathcal{H}, \boldsymbol{n}] = \max\left\{\frac{\underline{n}^{(0)}\underline{y}_j^{(0)} + n_j}{\underline{n}^{(0)} + n}, \frac{\overline{n}^{(0)}\overline{y}_j^{(0)} + n_j}{\overline{n}^{(0)} + n}\right\} = \begin{cases} \frac{\overline{n}^{(0)}\overline{y}_j^{(0)}+n_j}{\overline{n}^{(0)}+n} & \text{if } \overline{y}_j^{(0)} \geq n_j/n \\ \frac{\underline{n}^{(0)}\overline{y}_j^{(0)}+n_j}{\underline{n}^{(0)}+n} & \text{if } \overline{y}_j^{(0)} \leq n_j/n \end{cases} \tag{1.21}$$

For the model to be of any use, we must be able to elicit the bounds for the prior parameters. The interval $[\underline{y}_j^{(0)}, \overline{y}_j^{(0)}]$ simply represents bounds on the prior expectation of the chance α_j.

1.3.6.1. Fixed Learning Parameter

Typically, the learning parameter $n^{(0)}$ is taken to be 2 (not without controversy; see insightful discussions in Walley (1996a)). One might therefore be tempted to using the same prior expectations $y_j^{(0)}$ for the α_j as above (Eq. (1.17)), with $n^{(0)} = 2$, resulting in the following posterior expectations:

$$\mathrm{E}[\alpha_1 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{1.9 + 35}{2 + 36} = 0.971 \qquad \mathrm{E}[\alpha_2 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{0.06 + 1}{2 + 36} = 0.028$$
$$\mathrm{E}[\alpha_3 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{0.03 + 0}{2 + 36} = 0.0007 \qquad \mathrm{E}[\alpha_4 \mid n^{(0)}, \boldsymbol{y}^{(0)}, \boldsymbol{n}] = \frac{0.01 + 0}{2 + 36} = 0.0002$$

Whence, for this example, it is obvious that $n^{(0)} = 2$ is an excessively poor choice: the posterior expectations in case of zero counts are pulled way too much towards zero. One might suspect that this is partly due to the strong prior information, that is, the knowledge of $y_j^{(0)}$. However, even if we interpret the prior expectations (1.17) as bounds, say:

$$[\underline{y}_1^{(0)}, \overline{y}_1^{(0)}] = [0.950, 1] \tag{1.22a}$$
$$[\underline{y}_2^{(0)}, \overline{y}_2^{(0)}] = [0, 0.030] \tag{1.22b}$$
$$[\underline{y}_3^{(0)}, \overline{y}_3^{(0)}] = [0, 0.015] \tag{1.22c}$$
$$[\underline{y}_4^{(0)}, \overline{y}_4^{(0)}] = [0, 0.005] \tag{1.22d}$$

we still find:

$$\big[\underline{\mathrm{E}}[\alpha_1 \mid \mathcal{H}, \boldsymbol{n}], \overline{\mathrm{E}}[\alpha_1 \mid \mathcal{H}, \boldsymbol{n}]\big] = [0.971, 0.974]$$
$$\big[\underline{\mathrm{E}}[\alpha_2 \mid \mathcal{H}, \boldsymbol{n}], \overline{\mathrm{E}}[\alpha_2 \mid \mathcal{H}, \boldsymbol{n}]\big] = [0.026, 0.028]$$
$$\big[\underline{\mathrm{E}}[\alpha_3 \mid \mathcal{H}, \boldsymbol{n}], \overline{\mathrm{E}}[\alpha_3 \mid \mathcal{H}, \boldsymbol{n}]\big] = [0, 0.0007]$$
$$\big[\underline{\mathrm{E}}[\alpha_4 \mid \mathcal{H}, \boldsymbol{n}], \overline{\mathrm{E}}[\alpha_4 \mid \mathcal{H}, \boldsymbol{n}]\big] = [0, 0.0002]$$

Clearly, only the posterior inferences about α_1 (and perhaps also α_2) seem reasonable. We conclude that *the imprecise Dirichlet model with $n^{(0)} = 2$ learns too fast from the data in case of zero counts.*

On the one hand, when counts are sufficiently far from zero, the posterior probability with $n^{(0)} = 2$, and perhaps even $n^{(0)} = 1$ or $n^{(0)} = 0$,[31] seem appropriate. For zero counts,

[31] As mentioned on page 9, taking $n^{(0)} \leq 0$ would result in an improper prior.

however, a larger value of $n^{(0)}$ seems mandatory. Therefore, it seems logical to pick an interval for $n^{(0)}$.

A related argument for choosing an interval for $n^{(0)}$, in case of an informative set of priors, is provided by Walley (1991, p. 225, §5.4.4): a larger value of $\overline{n}^{(0)}$ ensures that the posterior does not move away too fast from the prior, which is particularly important for zero counts, and the difference between $\underline{n}^{(0)}$ and $\overline{n}^{(0)}$ effectively results in greater posterior imprecision if $n_j/n \notin [\underline{y}_j^{(0)}, \overline{y}_j^{(0)}]$.[32]

To see this, note that, if $\underline{y}_j^{(0)} \leq n_j/n \leq \overline{y}_j^{(0)}$, it follows from Eqs. (1.20) and (1.21) that both lower and upper posterior expectation are calculated using $\overline{n}^{(0)}$. When $n_j/n \leq \underline{y}_j^{(0)}$ (or $\overline{y}_j^{(0)} \leq n_j/n$), the lower (upper) posterior expectation is calculated using $\underline{n}^{(0)}$ instead, which is nearer to n_j/n due to the lower weight $\underline{n}^{(0)}$ for the prior bound $\underline{y}_j^{(0)}$ ($\overline{y}_j^{(0)}$). The increased imprecision reflects the conflict between the prior assignment $[\underline{y}_j^{(0)}, \overline{y}_j^{(0)}]$ and the observed fraction n_j/n, and this is referred to as *prior-data conflict*.[33]

1.3.6.2. Interval for Learning Parameter

We follow Good (1965, p. 19) (as suggested by Walley 1991, Note 5.4.1, p. 524), and reason about posterior expectations of hypothetical data to elicit $\underline{n}^{(0)}$ and $\overline{n}^{(0)}$; also see Walley (1991, p. 219, §5.3.3) for further discussion on elicitation on $n^{(0)}$—our approach is similar, but simpler for the case under study. We assume that $\overline{y}_1^{(0)} = 1$ and $\underline{y}_j^{(0)} = 0$ for all $j \geq 2$.

The upper probability of multiple ($j \geq 2$) failed components in trial $m+1$, given single-component failures ($j = 1$) in all of the first m trials, is

$$\overline{\mathrm{E}}[\alpha_j \mid \mathcal{H}, n_1 = m, n = m] = \frac{\overline{n}^{(0)}\overline{y}_j^{(0)}}{\overline{n}^{(0)} + m}\,, \quad j \geq 2\,.$$

(Note: there is no prior-data conflict in this case.) Whence, for the above probability to reduce to $\overline{y}_j^{(0)}/2$ (i.e., to reduce the prior upper probability by half), we need that $m = \overline{n}^{(0)}$. In other words, $\overline{n}^{(0)}$ *is the number of one-component failures required to reduce the upper probabilities of multi-components failure by half.*

Conversely, the lower probability of single-component failure ($j = 1$) in trial $m+1$, given only multiple ($j \geq 2$) failed components in the first m trials, is

$$\underline{\mathrm{E}}[\alpha_1 \mid \mathcal{H}, n_1 = 0, n = m] = \frac{\underline{n}^{(0)}\underline{y}_1^{(0)}}{\underline{n}^{(0)} + m}\,.$$

[32] We will show in Sections 3.1.4 and 3.3.4 that this argument, generalising the issue of zero counts to the issue of prior-data conflict (see Footnote 33 below), makes sense also for the general case of canonically constructed priors in Eq. (1.5).

[33] The issue of prior-data conflict, and models that allow sensitivity of inferences in its presence, are discussed in more detail in Sections 2.2.3.3 and 3.1.4. Walter and Augustin (2009b), reproduced in Section 3.3, is the publication centered around this idea.

(Note: there is strong prior-data conflict in this case.) In other words, $\underline{n}^{(0)}$ *is the number of multi-component failures required to reduce the lower probability of one-component failure by half.* Note that, in this case, a few alternative interpretations present themselves. First, for $j \geq 2$,

$$\overline{\mathrm{E}}[\alpha_j \mid \mathcal{H}, n_j = m, n = m] = \frac{\underline{n}^{(0)}\overline{y}_j^{(0)} + m}{\underline{n}^{(0)} + m}\,.$$

In other words, $\underline{n}^{(0)}$ *is also the number of* j*-component failures required to increase the upper probability of* j *components failing to* $(1+\overline{y}_j^{(0)})/2$ (generally, this will be close to $1/2$, provided that $\overline{y}_j^{(0)}$ is close to zero). Secondly, as $\underline{y}_j^{(0)} = 0$ for $j \geq 2$, we get, for $j \geq 2$,

$$\underline{\mathrm{E}}[\alpha_j \mid \mathcal{H}, n_j = m, n = m] = \frac{m}{m + \overline{n}^{(0)}}\,.$$

so $\overline{n}^{(0)}$ *is also the number of multi-component failures required to increase the lower probability of multi-component failures from zero to a half.*

Any of these counts seem well suited for elicitation, and are easy to interpret. As a guideline, we suggest the following easily remembered rules:

- $\overline{n}^{(0)}$ is the number of one-component failures required to reduce the upper probabilities of multi-component failures by half, and
- $\underline{n}^{(0)}$ is the number of multi-component failures required to reduce the lower probability of one-component failures by half.

Taking the above interpretation, the difference between $\overline{n}^{(0)}$ and $\underline{n}^{(0)}$ reflects the fact that the rate at which we reduce upper probabilities is less than the rate at which we reduce lower probabilities, and thus reflects a level of caution in our model.

Coming back to our example, reasonable values are $\underline{n}^{(0)} = 1$ (if we immediately observe multi-component failures, we might be quite keen to reduce our lower probability for one-component failure) and $\overline{n}^{(0)} = 10$ (we are happy to halve our upper probabilities of multi-component failures after observing 10 one-component failures). With these values, when taking for $y_j^{(0)}$ the values given in Eq. (1.17), we find the following posterior lower and upper expectations of α_j:

$$\begin{aligned}
\big[\underline{\mathrm{E}}[\alpha_1 \mid \mathcal{H}, \boldsymbol{n}], \overline{\mathrm{E}}[\alpha_1 \mid \mathcal{H}, \boldsymbol{n}]\big] &= [0.967, 0.972] \\
\big[\underline{\mathrm{E}}[\alpha_2 \mid \mathcal{H}, \boldsymbol{n}], \overline{\mathrm{E}}[\alpha_2 \mid \mathcal{H}, \boldsymbol{n}]\big] &= [0.0278, 0.0283] \\
\big[\underline{\mathrm{E}}[\alpha_3 \mid \mathcal{H}, \boldsymbol{n}], \overline{\mathrm{E}}[\alpha_3 \mid \mathcal{H}, \boldsymbol{n}]\big] &= [0.00041, 0.00326] \\
\big[\underline{\mathrm{E}}[\alpha_4 \mid \mathcal{H}, \boldsymbol{n}], \underline{\mathrm{E}}[\alpha_4 \mid \mathcal{H}, \boldsymbol{n}]\big] &= [0.00014, 0.00109]
\end{aligned}$$

These bounds indeed reflect caution in inferences where zero counts have occurred ($j = 3$ and $j = 4$), with upper expectations considerably larger as compared to the model with fixed s, while still giving a reasonable expectation interval for the probability of one-component failure.

If we desire to specify our initial bounds for $y_j^{(0)}$ more conservatively, as in Eqs. (1.22), we find similar results:

$$\begin{aligned}
\left[\underline{\mathrm{E}}[\alpha_1 \mid \mathcal{H}, \boldsymbol{n}], \overline{\mathrm{E}}[\alpha_1 \mid \mathcal{H}, \boldsymbol{n}]\right] &= [0.967, 0.978] \\
\left[\underline{\mathrm{E}}[\alpha_2 \mid \mathcal{H}, \boldsymbol{n}], \overline{\mathrm{E}}[\alpha_2 \mid \mathcal{H}, \boldsymbol{n}]\right] &= [0.0270, 0.0283] \\
\left[\underline{\mathrm{E}}[\alpha_3 \mid \mathcal{H}, \boldsymbol{n}], \overline{\mathrm{E}}[\alpha_3 \mid \mathcal{H}, \boldsymbol{n}]\right] &= [0, 0.00326] \\
\left[\underline{\mathrm{E}}[\alpha_4 \mid \mathcal{H}, \boldsymbol{n}], \overline{\mathrm{E}}[\alpha_4 \mid \mathcal{H}, \boldsymbol{n}]\right] &= [0, 0.00109]
\end{aligned}$$

1.3.6.3. Conclusion

We have seen that choosing bounds, rather than precise values, for the hyperparameters of the Dirichlet prior for the alpha-factors leads to more cautious inferences. This is especially important here, as inferences are strongly sensitive to the choice of prior, due to the zero counts that are common in this application of the Dirichlet-Multinomial model. We also concluded that assigning an interval for the learning parameter was necessary to arrive at reasonable posterior expectation intervals. Still, the method is relatively easy to use, as we identified simple ways to elicit bounds for the hyperparameters by reasoning on hypothetical data; essentially, the analyst needs to specify how quickly he is willing to learn from various sorts of hypothetical data.

In the above application, Bayesian inference in the Dirichlet-Multinomial model was generalised to imprecise or interval probability methods. Next, in Chapter 2, we will give a general introduction to the methodology of imprecise or interval probability, discuss further motivations for the use of interval probability methods for statistical inference, and demonstrate in Chapter 3 that inferences in canonical conjugates as described in Section 1.2.3.1 can be generalised in the same way as was done above.

2. Imprecise Probability as Foundation of Generalised Bayesian Inference

After having seen a detailed example for Bayesian inference using sets of conjugate priors in Section 1.3, in this chapter we will now give a general introduction to the methodology of generalised Bayesian inference, before we give a systematic discussion of models for generalised Bayesian inference with sets of conjugate priors in the central Section 3.1 of this thesis.

In Section 2.1 below, we will try to outline the general theory of imprecise or interval probability, describing its main formulations and interpretations, and discuss the generalised Bayesian inference procedure. Section 2.2 then gives at first some general motives for the use of imprecise probability methodology, concluding with the motives especially relevant in the context of the Bayesian approach to statistical inference, among which prior-data conflict and weakly informative priors will receive special attention in the model discussion in Chapter 3.

2.1. Imprecise or Interval Probability

This Section will give a condensed introduction to the main theoretic concepts in interval or imprecise probability as needed for the topics discussed in this thesis. Here, we take a decidedly epistemic view on interval or imprecise probability, as we will argue in Section 2.2 that imprecise probability distributions are often a better tool for expressing prior beliefs than precise probability distributions.

2.1.1. General Concept and Basic Interpretation

The central idea of imprecise or interval probability (Walley 1991; Weichselberger 2001; Coolen, Troffaes, and Augustin 2011) is to replace the usual, precise probability measure $\mathrm{P}(A)$ for events A[1] with a *lower* and *upper probability*, denoted by $\underline{\mathrm{P}}(A)$ and $\overline{\mathrm{P}}(A)$, respectively, satisfying

$$0 \leq \underline{\mathrm{P}}(A) \leq \overline{\mathrm{P}}(A) \leq 1\,. \tag{2.1}$$

In this setting, a usual probability measure forms the extreme case $\underline{\mathrm{P}}(A) = \overline{\mathrm{P}}(A) = \mathrm{P}(A)$, when there is enough information to determine the distribution on the sample space Ω in precise stochastic terms. On the other extreme, when $\underline{\mathrm{P}}(A) = 0$ and $\overline{\mathrm{P}}(A) = 1$, we have no information at all on the probability for A to occur, and intermediate cases $0 \leq \underline{\mathrm{P}}(A) < \overline{\mathrm{P}}(A) \leq 1$ represent different degrees of knowledge on this probability.

Therefore, interval or imprecise probability adds another modeling dimension: While usual, precise probability measures can be used to model phenomena when there is perfect stochastical information, like, e.g., in a lottery where the number of winning tickets (and the total number of tickets) is precisely known, imprecise probability measures can account for cases where there is uncertainty about the probabilities themselves, just like in a lottery where the number of winning tickets is not exactly known. Non-stochastic uncertainty about model features like probabilities is often called *ambiguity*,[2] forming a crucial part of the human decision process, and there are studies suggesting that humans process ambiguity in a way differing from pure probabilistic reasoning (Hsu et al. 2005).

In contrast to a probability measure $\mathrm{P}(A)$, the set functions $\underline{\mathrm{P}}(A)$ and $\overline{\mathrm{P}}(A)$ do not adhere to the additivity axiom of Kolmogorov's (1933) formalisation of probability as a normed measure, and thus are also known as *non-additive probabilities*. There is also a link to *fuzzy measures*, which are also non-additive measures (see, e.g., Denneberg 1997).

In general, $\underline{\mathrm{P}}(A)$ may be understood as accounting for evidence certainly in favour of A, and $\overline{\mathrm{P}}(A)$ accounting for all evidence speaking not strictly against A. The difference of $\overline{\mathrm{P}}(A)$ and $\underline{\mathrm{P}}(A)$ thus allows for inconclusive evidence that may not speak unanimously in favor of or against A, respectively. As evidence strictly against A can be seen as evidence certainly in favour of $A^{\complement}$, the complement of A, it is mostly assumed that $\underline{\mathrm{P}}(A^{\complement}) = 1 - \overline{\mathrm{P}}(A)$,

[1] Events of interest A are taken to be subsets of the sample space Ω, forming a σ-algebra, a non-empty collection of sets including countable unions and intersections of subsets of Ω.

[2] See Section 2.2.2.

and thus it suffices to determine either of $\underline{P}$ or $\overline{P}$, the other one being defined through this relation.

There are currently two main approaches to a general theory of statistical inference with interval or imprecise probability:

(i) The theory by Weichselberger (2000; 2001) regards probability intervals $[\underline{P}(A), \overline{P}(A)]$ for all or some events A as the basic entity (Coolen, Troffaes, and Augustin 2011, p. 646), from which an interval-valued distribution on the sample space Ω is constructed. His approach is axiomatic, in the sense that it replaces Kolmogorov's (1933) additivity axiom by two axioms,[3] from which the theory is derived, while imposing no specific interpretation on these constructs. This theory of interval probability was developed as the foundation of a concept of *logical probability* (Weichselberger 2007), where probability is not assigned to events, but to logical conclusions (from a premise to a consequence), with the aim to arrive at a theory of statistical inference which allows for fiducial-like probability statements, e.g., probability statements on parameters similar to those derived from a posterior distribution in a Bayesian setting, but without the need to specify a prior distribution (see, e.g., Hannig, Yver, and Lee 2011).

(ii) In contrast, the theory by Walley (1991; 2000) aims to generalise the Bayesian approach to statistical inference, adopting a strictly subjective, behavioural interpretation for imprecise probability as lower and upper betting rates (see below), and extending the Bayesian inference paradigm (as discussed in Section 1.2.3) to imprecise probability distributions. In generalising de Finetti (1937; 1970), the basic entities are lower and upper *previsions*, i.e. expectation functionals, for *gambles*, i.e. random quantities, instead of lower and upper probabilities for events. This is due to the fact that unlike in the theory of precise probability—where the definition of a distribution via expectations (often denoted *linear previsions* in the imprecise probability literature) is eqivalent to a definition via a precise probability distribution—the definition of an imprecise distribution via lower and upper previsions is more general than a definition via lower and upper probability for events (Walley 2000, p. 132).

As this thesis is concerned with a generalisation of Bayesian inference based on sets of conjugate priors, the approach by Walley, and its reliance on a subjective, epistemic interpretation of (imprecise) probability as (bounds for) betting rates, is now described in more detail.

2.1.2. Main Formulations

The main mathematical formulations for imprecise distributions in the theory by Walley (1991; 2000) are

[3] The first states that (2.1) holds, the second that the set $\mathcal{M}$ consisting of all usual, precise distributions $P(\cdot)$ with $\underline{P}(A) \leq P(A) \leq \overline{P}(A)$, for all $A \in \Omega$, is non-empty. The second axiom guarantees that there exists at least one precise probability distribution which is compatible with an interval-valued probability distribution, and thus rules out the case of contradictory assignments of $[\underline{P}(A), \overline{P}(A)]$.

(i) lower previsions,

(ii) sets of (precise) probability distributions, and

(iii) sets of desirable gambles.

It is possible to switch between these formulations, although (ii) can be slightly more general than (i), and to a greater extent, (iii) is more general than (ii) (Walley 2000).

We will first introduce lower previsions and the most important rationality requirements guiding their assessment and use, then have a brief look at sets of desirable gambles as the most comprehensive formulation. Sets of probability distributions, as the formulation used to describe inference models in this thesis, are then explained in relation to the other formulations, and the generalised Bayesian inference procedure as used in Section 1.3.6 is justified formally.

2.1.2.1. Lower Previsions

As mentioned in Section 2.1.1, the basic entities in the theory by Walley (1991) are lower and upper *previsions*. These are functions on *gambles*, or *random variables*, defined as bounded mappings from Ω to $\mathbb{R}$. A gamble X can be understood as uncertain reward or payout, where the reward $X(\omega)$ depends on $\omega \in \Omega$, the unknown 'state of the world' from the possibility space Ω. The reward is measured in units of utility assumed to form a linear scale (Walley 1991, §2.2).

A *lower prevision* or *lower expectation* is then a mapping $\underline{\mathrm{E}} : \mathcal{K} \to \mathbb{R}$, where $\mathcal{K}$ is a set of gambles.[4] Central to Walley's theory is the interpretation of $\underline{\mathrm{E}}[X]$ as the subject's supremum buying price for X, that is, the subject is disposed to pay at most the fixed amount $\underline{\mathrm{E}}[X]$ in exchange for the uncertain reward X.[5] $\underline{\mathrm{E}}[X]$ thus expresses the subject's state of knowledge about the value of X, factoring in the propensity of all possible $\omega \in \Omega$ with their specific payouts $X(\omega)$. The *upper prevision* $\overline{\mathrm{E}}[X]$ is the infimum selling price for X, i.e., the fixed amount the subject is willing to receive in exchange for X (Walley 1996b, p. 9). Walley's theory is based on this behavioral, epistemic interpretation of previsions, and all rationality criteria and inference procedures are deduced from this root (Walley 1996b, p. 5).

The theory allows thus a zone of indeterminacy, by $\underline{\mathrm{E}}[X] < \overline{\mathrm{E}}[X]$, for prices of X at which the subject is neither willing to buy nor to sell the gamble X, as illustrated in Figure 2.1. This is in contrast to the usual epistemic operationalisation of subjective

[4] In Walley's central monography (Walley 1991), his papers and the imprecise probability literature in general, lower previsions are usually denoted by $\underline{\mathrm{P}}$. Furthermore, events $A \subset \Omega$ are notationally identified with the indicator function $I_A(\omega)$, being a gamble with payout 1 if $\omega \in A$ and else 0, such that $\underline{\mathrm{P}}(A)$ denotes the lower probability of A. In order to follow conventions in statistical literature, however, lower previsions are denoted by $\underline{\mathrm{E}}$ here, and $\underline{\mathrm{P}}$ refers exclusively to lower probabilities.

[5] More precisely, the price the subject is disposed to pay for X is strictly less than $\underline{\mathrm{E}}[X]$. For sake of readability, this mathematically important distinction is not rigorously maintained in this brief treatment, as it is hardly relevant for the interpretation of the results in the later parts of this thesis.

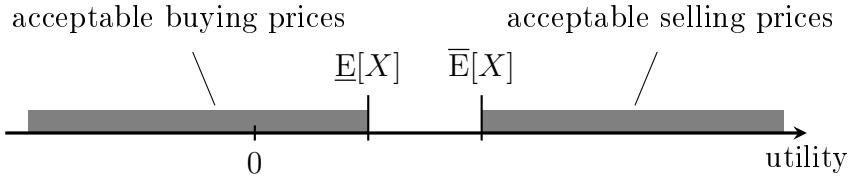

Figure 2.1.: Illustration of $\underline{\mathrm{E}}$ and $\overline{\mathrm{E}}$ as supremum buying and infimum selling prices for a gamble X on a linear utility scale.

Bayesian probability, which implies that $\underline{\mathrm{E}}[X]$ and $\overline{\mathrm{E}}[X]$ must coincide at a unique fair price $\mathrm{E}[X]$. This requirement is refuted by Walley, and called by him *the Bayesian dogma of precision* (Walley 1991, §5).

A lower and upper prevision are called *conjugate*[6] if the relation $\overline{\mathrm{E}}[X] = -\underline{\mathrm{E}}[-X]$ holds for all $X \in \mathcal{K}$.[7] Then, it suffices to consider only either of $\underline{\mathrm{E}}$ or $\overline{\mathrm{E}}$, and in the literature $\underline{\mathrm{E}}$ is chosen, the theory consequently being referred to as the theory of *coherent lower previsions* (see, e.g., Augustin et al. 2013, §3.2).

2.1.2.2. Coherence and Avoiding Sure Loss

Before we come to the concept of *coherence*, a weaker rationality requirement put forward by Walley is that a lower prevision $\underline{\mathrm{E}}$ should fulfill the property of *avoiding sure loss* (Walley 1991, §2.4). This means that a subject, whose state of information about the occurrence of 'states of the world' from the possibility space Ω is encoded in his or her choice of $\underline{\mathrm{E}}$, acting by buying and selling gambles accordingly, should choose $\underline{\mathrm{E}}$ such that there is no combination of gambles that would result in a net loss, whatever $\omega \in \Omega$. As an analogy, to violate *avoiding sure loss* can be regarded like a logical inconsistency in a set of propositions, i.e., there exist at least two propositions in the set of propositions (the analogy to $\underline{\mathrm{E}}$) which are contradictory (Walley 1991, §2.4, footnote 1).

Continuing this analogy, $\underline{\mathrm{E}}$ violating the stronger property of *coherence* is like the "failure to deduce all [...] logical implications" (Walley 1991, §2.4, footnote 1) from the set of propositions. As an example, if a subject assesses $\underline{\mathrm{E}}$ such that $\underline{\mathrm{E}}[X] = \underline{\mathrm{E}}[Y] = \frac{1}{4}$, and also that $\underline{\mathrm{E}}[X+Y] = \frac{1}{4}$, $\underline{\mathrm{E}}$ avoids sure loss, but is not coherent, because one can imply from $\underline{\mathrm{E}}[X] = \underline{\mathrm{E}}[Y] = \frac{1}{4}$ that $\underline{\mathrm{E}}[X+Y]$ must be at least $\frac{1}{2}$ (Walley 1991, p. 67). Coherence is a "normative requirement of consistency" (Walley 2000, p. 130) that is a consequence of few basic rationality requirements. In fact, if $\mathcal{K}$ is a linear space of gambles, i.e., $\mathcal{K}$ is closed with respect to addition and multiplication with constants, then coherence is equivalent to the three following conditions (e.g., Walley 1996b, p. 11):

(i) $\underline{\mathrm{E}}[X] \geq \inf_{\omega\in\Omega} X(\omega)$, i.e., one is always prepared to buy a gamble for less than its

[6] Conjugacy in this sense should not be confused with conjugacy for prior distributions as discussed in Section 1.2.3.1.

[7] This implies the relation $\underline{\mathrm{P}}(A^{\complement}) = 1 - \overline{\mathrm{P}}(A)$ of lower and upper probabilities for events $A \subset \Omega$ as mentioned in Section 2.1.1.

minimum possible reward. (Being prepared to pay more than $\inf_{\omega\in\Omega} X(\omega)$ amounts to making a commitment to the possibility that other states ω' with $X(\omega') > \inf_{\omega\in\Omega} X(\omega)$ may occur.)

(ii) $\underline{\mathrm{E}}[\lambda X] = \lambda\underline{\mathrm{E}}[X]$ for all $X \in \mathcal{K}$, $\lambda > 0$. $\underline{\mathrm{E}}$ should be such that if it postulates that one should be prepared to buy a gamble X at a price of at most $\underline{\mathrm{E}}[X]$, then one should also buy gambles that are fractions or multiples of X at prices of at most the original price divided or multiplied accordingly. This illustrates that prices are to be understood on a linear utility scale (as mentioned above) and should not be considered as plain monetary prices, for which this condition may not be reasonable.

(iii) $\underline{\mathrm{E}}[X + Y] \geq \underline{\mathrm{E}}[X] + \underline{\mathrm{E}}[Y]$. This property of superlinearity or superadditivity implies that the supremum acceptable buying price for X and Y combined should be at least the sum of supremum acceptable buying prices for each of X and Y individually. It could be, e.g., that X and Y balance out each other in a way that buying both of them at once makes the transaction less risky, such that a higher price limit for $X + Y$ is acceptable. Usual precise expectations, for which additivity $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$ must hold, cannot directly accommodate such reasoning.

Consequently, in analogy to deducing all logical implications from a set of propositions, there is a technique, called *natural extension*, that adjusts a given assessment of lower and upper previsions (on some set of gambles $\mathcal{K}$, which must avoid sure loss) to make it coherent, in the least committal way as possible[8] (see, e.g., Walley 1996b, pp. 15ff). This is accomplished by a linear program, and may involve raising $\underline{\mathrm{E}}[X]$ for some $X \in \mathcal{K}$ in order to make them coherent to the values of $\underline{\mathrm{E}}[Y]$, supremum buying prices for some other gambles $Y \in \mathcal{K}$ as previously assessed, or newly defining $\underline{\mathrm{E}}[Z]$ for gambles Z not considered during assessment—explaining the name 'extension'.

2.1.2.3. Sets of Desirable Gambles

Sets of desirable gambles (Walley 2000, §6; Quaeghebeur 2013) are an alternative formulation for probability assessments as expressed by a lower prevision $\underline{\mathrm{E}}$. They are a mathematically convenient formulation and as such a very useful tool in the development of the theory of imprecise probability. However, a detailed understanding of this concept is not necessary for the purpose of this thesis, and the account below is intended to give some insight into this central concept in the theory of imprecise probability.

The *gambles* here are again a description of an uncertain reward as above, and all bounded functions from Ω to $\mathbb{R}$ are forming the space $\mathcal{L}$ of all gambles. The assessment in this formulation lies in the designation of a subset $\mathcal{D}$ of all gambles $\mathcal{L}$ as *desirable gambles*, i.e., as supplying an uncertain reward (in utilities) for which, from the subject's state of information about the propensity of occurence of the different $\omega \in \Omega$, the potential benefits outweigh the potential losses. Thus, all gambles X that will never incur a loss,

[8] This means that there may be other adjustments of $\underline{\mathrm{E}}$ that are compatible with the initial assessments, but are less conservative, i.e., give higher supremum buying prices.

$\{X : X \geq 0\}$, where $X \geq 0$ denotes that $X(\omega) \geq 0\ \forall \omega \in \Omega$, should be contained in a coherent set of desirable gambles $\mathcal{D}$. Indeed, the notion of *coherence* for lower previsions translates to sets of desirable gambles as follows: $\mathcal{D} \subset \mathcal{L}$ is coherent if and only if (Walley 2000, p. 137)

(D1) $0 \notin \mathcal{D}$ (0 is the gamble X with $X(\omega) = 0\ \forall \omega \in \Omega$),

(D2) if $X \in \mathcal{L}$ and $X > 0$ then $X \in \mathcal{D}$
($X > 0$ denotes that $X \geq 0$ and $\exists \omega \in \Omega : X(\omega) > 0$),

(D3) if $X \in \mathcal{D}$ and $c \in \mathbb{R}_{>0}$ then $cX \in \mathcal{D}$,

(D4) if $X \in \mathcal{D}$ and $Y \in \mathcal{D}$ then $X + Y \in \mathcal{D}$.

In $\mathcal{L}$, a coherent set of gambles $\mathcal{D}$ is thus a convex cone containing all positive gambles $\{X : X > 0\}$ but not the zero gamble.[9]

A coherent set of desirable gambles can be derived from a coherent lower prevision $\underline{\mathrm{E}}$ by (Walley 2000, p. 139)

$$\mathcal{D} = \left\{ X \in \mathcal{L} : X > \sum_{i=1}^{n} c_i (X_i - \underline{\mathrm{E}}[X_i] + \epsilon) \text{ for some } n \in \mathbb{N}, c_i > 0, \epsilon > 0, X_i \in \mathcal{L} \right\}.$$

If we adopt $\underline{\mathrm{E}}$ as the lower prevision expressing our beliefs about Ω, gambles $X_i - \underline{\mathrm{E}}[X_i] + \epsilon$ with $\epsilon > 0$ should be desirabe for us; considering the price $\underline{\mathrm{E}}[X_i]$ as the supremum acceptable buying price for X_i, the gamble $X_i - \underline{\mathrm{E}}[X_i]$ should—in our view—be at least as good as the zero gamble.

Conversely, a coherent lower prevision can be deduced from a coherent set of desirable gambles by $\underline{\mathrm{E}}[X] := \sup\{c > 0 : X - c \in \mathcal{D}\}$. If we make $c > 0$ as large as possible such that the gamble $X - c$ is still acceptable to us, then $\underline{\mathrm{E}}[X]$ simply is our supremum acceptable buying price for X.

As already mentioned, sets of desirable gambles are a more general formulation as compared to lower previsions, illustrated by the fact that there can be several sets $\mathcal{D}$ leading to the same $\underline{\mathrm{E}}$ (Walley 2000, p. 139).[10]

2.1.2.4. Sets of Probability Distributions

We now turn to the formulation of imprecise probability assessments applied in this thesis, sets of (precise) probability distributions, which are also called *credal sets* (e.g., Walley 2000, p. 136).

[9] As the weaker property, a set of gambles $\mathcal{D}$ avoids sure loss if $\mathrm{posi}(\mathcal{D}) \cap \{X : X(\omega) < 0\ \forall \omega \in \Omega\} = \emptyset$. Here, $\mathrm{posi}(\mathcal{D}) := \{\sum_{i=1}^{n} c_i X_i : c_i \in \mathbb{R}_{>0}, X_i \in \mathcal{D}, n \in \mathbb{N}\}$ is the positive hull of $\mathcal{D}$, i.e. the set of gambles resulting from application of (D3) and (D4) on $\mathcal{D}$.

[10] Another formulation equivalent to sets of desirable gambles is the description as *partial preference orderings*, where a partial ordering of the gambles in $\mathcal{L}$ is given to express beliefs on Ω (Walley 2000, p. 138).

There is a one-to-one correspondence between coherent lower previsions and non-empty, closed and convex sets of (precise) probability distributions (Walley 1991, §3.6.1). When dropping the conditions of convexity and closure, sets of probability distributions can be slightly more general than coherent lower previsions (Walley 2000, §5). However, the nature of this slight increase in generality is not relevant here, although we will consider the question of convexity more closely later.

Given a coherent lower prevision $\underline{\mathrm{E}}$, the corresponding set of distributions $\mathcal{M}$ is closed and convex, and consists of all probability distributions whose expectations E_p dominate $\underline{\mathrm{E}}$, i.e.,[11]

$$\mathcal{M} = \{p(\omega) : \mathrm{E}_p[X] \geq \underline{\mathrm{E}}[X] \ \forall X \in \mathcal{L}\}\,.$$

In fact, for all $p \in \mathcal{M}$ and $X \in \mathcal{L}$, the relation $\underline{\mathrm{E}}[X] \leq \mathrm{E}_p[X] \leq \overline{\mathrm{E}}[X]$ holds; the set $\mathcal{M}$ consists of all probability distributions whose expectations are compatible with the bounds defined by the lower prevision $\underline{\mathrm{E}}$ and its conjugate upper prevision $\overline{\mathrm{E}}$.

Conversely, given a non-empty set of probability distributions $\mathcal{M}$, where $\mathcal{M}$ needs not necessarily be closed or convex, the corresponding coherent lower prevision, for any gamble $X \in \mathcal{L}$, is defined by

$$\underline{\mathrm{E}}[X] = \inf_{p \in \mathcal{M}} \mathrm{E}_p[X]\,,$$

and in this case, $\underline{\mathrm{E}}$ is called the *lower envelope* of E_p, $p \in \mathcal{M}$ (Walley 1991, p. 132).

There are very important relations between the notions of avoiding sure loss and coherence on the one hand, and the formulation of imprecise probability assessments via sets of probability distributions on the other hand. The two equivalences below are known as the *lower envelope theorem* (Walley 1991, §3.3.3).

(a) $\underline{\mathrm{E}}$ avoids sure loss if and only if the corresponding set $\mathcal{M}$ is non-empty. Thus, an assessment $\underline{\mathrm{E}}$ for which there is no compatible probability distribution must incur a sure loss, and cannot be considered as reasonable. On the contrary, any imprecise probability distribution defined by assigning a set $\mathcal{M}$ of probability distributions avoids sure loss.

(b) Furthermore, $\underline{\mathrm{E}}$ is coherent if and only if it can be described as the lower envelope based on its corresponding $\mathcal{M}$. Therefore, all coherent lower previsions are characterised as lower envelopes based on some set of precise distributions $\mathcal{M}$, and imprecise probability assignments established via such a set $\mathcal{M}$ are, by design, coherent.

Although the one-to-one correspondence mentioned above holds only for closed and convex sets $\mathcal{M}$, there is nothing in the theory preventing us to consider open or non-convex sets $\mathcal{M}$ as our probability model, because lower and upper previsions derived from $\mathcal{M}$ are nevertheless coherent.

[11] The probability distributions contained in a credal set $\mathcal{M}$ will be referred to by their probability density or mass functions $p(\cdot)$ in place of their probability measures $\mathrm{P}(\cdot)$, as we will consider mostly the densities in our later discussions.

2.1.2.5. Conditioning and the Generalised Bayes' Rule

To use imprecise probability distributions for statistical inference in the same way as usual Bayesian inference employs precise probability distributions, we need a notion of *conditioning* or *updating* for imprecise probability distributions. In analogy to the procedure described in Section 1.2.3, the objective is to express prior knowledge on a parameter of interest ϑ by an imprecise prior distribution, and all inferences shall be based on the (imprecise) posterior distribution derived from it. As in Bayesian inference with precise distributions, the now imprecise prior should be conditioned on the observed data $\boldsymbol{x}$.

A coherent lower prevision can be conditioned on an event B by using the so-called *Generalised Bayes' Rule* (GBR, Walley 1991, §6.4), by which the conditional coherent lower prevision $\underline{\mathrm{E}}[X \mid B]$ based on a lower prevision $\underline{\mathrm{E}}[X]$ can be derived. $\underline{\mathrm{E}}[X \mid B]$ is then also coherent to $\underline{\mathrm{E}}[X]$, i.e., it is a model satisfying the rationality criteria as discussed for $\underline{\mathrm{E}}[X]$ now also for the beliefs expressed in $\underline{\mathrm{E}}[X]$ contingent on B.[12]

In the formulation via a credal set, the Generalised Bayes' Rule is equivalent to conditioning each distribution in the credal set on B via Bayes' Rule (Walley 1991, §6.4.2), and the set of conditioned distributions is thus an equivalent model for $\underline{\mathrm{E}}[X \mid B]$. This important result is known as the *lower envelope theorem for conditional previsions.*

2.1.3. Generalised Bayesian Inference Procedure

Walley has thus established a general framework for coherent statistical inference under imprecise probabilities. It allows to transfer the basic aspects of traditional Bayesian inference to the generalised setting, as the fundamental paradigms of Bayesian inference as discussed in Section 1.2.3 are maintained. Prior knowledge on the parameter, expressed by a now imprecise prior distribution $\underline{\mathrm{E}}(\cdot)$ with credal set $\mathcal{M}$, is updated in the light of the observed sample $\boldsymbol{x}$ to the posterior $\underline{\mathrm{E}}(\cdot \mid \boldsymbol{x})$, with the credal set $\mathcal{M}_{|\boldsymbol{x}}$, and this statistical inference is again understood as a deductive process, obtained directly by conditioning on the observed sample, now according to the Generalized Bayes' Rule that ensures coherence of this inferential process. For practical implementation of the Generalized Bayes' Rule, the lower envelope theorem for conditional previsions mentioned above is of particular relevance. The prior credal set $\mathcal{M}$ is updated element by element to obtain the posterior credal set

$$\mathcal{M}_{|\boldsymbol{x}} = \{p(\cdot \mid \boldsymbol{x}) : p(\cdot) \in \mathcal{M}\} \ , \tag{2.2}$$

consisting of all posterior distributions (represented by their density or mass functions $p(\cdot \mid \boldsymbol{x})$) obtained by traditional Bayesian updating of elements of the prior credal set.

[12] The adequacy of the Generalised Bayes' Rule for statistical inference procedures has been questioned in the literature, and there is doubt that it may reasonably represent learning. We will comment on this topic in Section 2.1.3.2.

2.1.3.1. Relation to Bayesian Sensitivity Analysis

Walley's lower envelope theorem also establishes a close connection to robust Bayesian approaches and Bayesian sensitivity analysis (see, e.g., Berger et al. 1994; Ríos Insua and Ruggeri 2000; Ruggeri, Ríos Insua, and Martín 2005) based on sets of distributions. In fact, Walley's theory of lower previsions can be seen as providing a formal framework for these approaches (see, e.g., Berger et al. 1994, §1.1). However, there is a basic difference in the interpretation of the underlying sets of probability distributions. While in the imprecise probability context a credal set is understood as an entity of its own, the robust Bayesian approach emphasizes the single elements in the set, and very often discusses the effects of deviations from a certain central element.[13]

As a consequence, for the robust Bayesian point of view it is quite natural and common to impose some further regularity conditions on the elements in the set of distributions, like additional smoothness constraints or unimodality of the underlying densities.[14] Since lower and upper posterior probabilities are determined by the extreme points of the underlying credal sets, this distinction may indeed matter substantially in practice.

In essence, robust Bayesian inference and Bayesian sensitivity analysis understand robustness and insensitivity mostly as desirable properties, while the imprecise probability framework may use such behavior actively in modelling, in particular in the context of prior-data conflict (see Sections 2.2.3.3 and 3.1.4).

2.1.3.2. Critique

The seemingly self-evident character of the Generalised Bayes' Rule has been questioned in the literature. Walley (1991, p. 335) notes that, although the theory of coherence suggests that "[...the GBR] is a reasonable updating strategy, [...] there is nothing in the theory that requires You to [...] construct conditional previsions [...] through the GBR" and to "[...] adopt [...them] as Your new unconditional prevision". He is also very clear about the fact (see Walley 1991, p. 334) that "there is a role for other updating strategies, not because the updated beliefs constructed through the GBR are unjustified, but because they are often indeterminate". Indeed, Walley (1991, §6.11.1) lists twelve items summarizing "[...] the reasons for which the GBR may fail to be applicable or useful as an updating strategy."

One central argument is that the notion of coherence, of which the Generalised Bayes' Rule is a consequence, may not be adequate to represent learning, especially when observed data is rather surprising, and should completely overturn prior beliefs about the data generating process. As posterior beliefs derived from the Generalised Bayes' Rule must be coherent with both prior beliefs and data, they may be rather imprecise in case of very vague prior beliefs or unexpected data. Essentially, the Generalised Bayes' Rule does not allow to abandon prior beliefs in the light of surprising data, and both too vague or too

[13] These so-called *neighbourhood models* are briefly discussed in Section 3.2.1.1.

[14] However, in Section 3.1, in the case of $\mathcal{M}^{(0)}$ consisting of parametric distributions only, we will follow a similar route, as the canonical conjugates typically are unimodal and smooth.

specific, but (in the light of the data) inappropriate prior beliefs may influence posterior beliefs to an intolerable extent.[15]

Indeed, it has been debated that—intuitively—posterior inferences derived via the Generalised Bayes' Rule are often too imprecise. As such possibly counterintuitive bounds result from elements in the prior credal set under which the observed sample is rather unlikely, in order to maintain the understanding of updating as conditioning on the sample, several authors therefore suggested to reduce the prior credal sets in dependence on the sample. A radical way to achieve this is to consider in the conditioning process only those priors under which the observed sample has highest probability (see, e.g., Held, Kriegler, and Augustin 2008; see also Walley and Fine 1982), resulting in a procedure that is related to Dempster's rule of conditioning (see, e.g., Destercke and Dubois 2013a, §3.2). Other ways to address this issue are update rules where imprecision is updated in accordance with an information measure (Coolen 1993a; Coolen 1994; see also Coolen 1993b), or hierarchical models, like the model suggested by Cattaneo (2008) in the (profile) likelihood context (see the short description in Section 2.2.4.2), a variant of which also allows to incorporate prior information by using a so-called prior likelihood.

Another approach to tackle these issues is to refine the concept of coherence itself, and to replace or complement it with a notion that accounts for possible change of a subject's beliefs in the light of new evidence. This concept of *temporal coherence* (Goldstein 1985) attempts to frame, informally stated, only current beliefs about future beliefs, and not the future beliefs themselves (that may take into account new evidence not according to coherence). For an inference approach based on this notion, see Goldstein and Wooff (2007); Troffaes and Goldstein (2013) give some first results on consequences of temporal coherence for inference with coherent lower previsions.

A very important foundational critique of the Generalised Bayes' Rule relates to issues from a decision theoretic point of view. It has been shown that the decision theoretic justification of Bayes' Rule as producing prior risk minimising decision functions does not extend to the case of sets of priors. Updating by Generalised Bayes' Rule does no longer necessarily lead to optimal decision functions and thus, as one could argue, also not to optimal inference procedures; see, in particular, Augustin (2003) for a detailed discussion, Noubiap and Seidel (2001) and Augustin (2004) for algorithms to calculate optimal decision functions.

These important criticisms notwithstanding, the models discussed in this thesis will rely on the inference procedure described above, based on the Generalised Bayes' Rule to obtain $\mathcal{M}_{|\boldsymbol{x}}$. As will be discussed in Section 3.1, the suggested models nevertheless provide very attractive inference features. Furthermore, we will sketch in Section 4.3 (see also some first technical considerations in Section A.2) how models along this approach can be modified to cater for 'bonus precision' if prior and data coincide especially well.

[15] However, as we will see in Section 3.1.2, items I. and II., in our model, data can outweigh prior beliefs, leading to reasonable inferences.

2.1.4. A Brief Glance on Related Concepts

The theory of imprecise probability outlined in Sections 2.1.2 and 2.1.3 above is of course not the only attempt to complement probability theory as tool for handling uncertainty;[16] there are many other concepts with a similar aim, like possibility distributions (which often take the form of a fuzzy interval), or fuzzy probability measures, which are also known as capacities (see Section 2.2 below).

In fact, many of these concepts are closely linked to lower previsions or sets of probability distributions, and a large part can indeed be seen as special cases of generic lower previsions, or as sets of probability distributions with certain restrictions (Destercke and Dubois 2013b, Fig. 5.5). A considerable part of the imprecise probability literature explores these links, of which Destercke and Dubois (2013a; 2013b) give a concise overview.

Although generally less expressive than coherent lower previsions, these special cases can nevertheless be useful tools, as they may be more easy to handle or elicit for the problem at hand, or simply be more easy to communicate to the practitioners involved (Destercke and Dubois 2013b, §1).

2.1.4.1. Belief Functions

An important special case of coherent lower previsions are *belief functions* (see, e.g., Destercke and Dubois 2013a, §2). Here, the corresponding upper probability (related through conjugacy as mentioned in Section 2.1.1) is often considered explicitely, denoted as *plausibility function*. Belief and plausibility functions can be based on a *probability mass assignment* $m(\cdot)$, valuing the occurrence propensity for *subsets* E of the sample space $\mathcal{X}$ by weights $m(E) \geq 0$ which sum up to 1, i.e. $\sum_{E \subseteq \mathcal{X}} m(E) = 1$.[17] The *belief function* is then defined, for any $A \subseteq \mathcal{X}$, by

$$\mathrm{Bel}(A) = \sum_{E \subseteq A, E \neq \emptyset} m(E)\,,$$

collecting the probability mass assignments for all sets that necessarily imply A; the *plausibility function* is defined by

$$\mathrm{Pl}(A) = \sum_{E \cap A \neq \emptyset} m(E)\,,$$

collecting the probability mass assignments for all sets that are compatible with A (have common elements with A, i.e., do not contradict A).

This approach can be useful when, in the case of discrete sample spaces $\mathcal{X}$, it is not possible to obtain precise observations $x \in \mathcal{X}$, but only subsets $E \subset \mathcal{X}$ as observations. Consider, e.g., a poll where participants are allowed to name sets of political parties or

[16] A number of motives for going beyond usual probability measures will be discussed in Section 2.2.

[17] In a probability distribution, instead only singletons, or one-element subsets, may receive weights $m(\cdot) > 0$. Probability mass assignments can also be seen as providing a probability distribution on the power set of the sample space $\mathcal{X}$.

candidates they intend to vote, instead of being forced to name only one, even if they are currently undecided between, say, three parties or candidates.

Belief functions are indeed a special case of coherent lower previsions. If $m(\emptyset) = 0$, then belief functions coincide with ∞-monotone capacities, which are a special case of lower previsions (e.g., Destercke and Dubois 2013a, §2.1). Only if $m(\emptyset) > 0$, which does not seem like a reasonable choice in most circumstances anyway, the set of corresponding probability distributions $\mathcal{M}$ is empty.

2.1.4.2. Examples for Frequentist Approaches

To avoid the impression that imprecise probability models are necessarily related to the Bayesian approach to inference, we will now also mention briefly some non-Bayesian approaches.

Augustin, Walter, and Coolen (2013, §5) introduce some frequentist approaches to inference using imprecise probability. A field that is so far comparatively widely developed is the theory of statistical hypothesis testing, based on the so-called *Huber-Strassen theorem* (Huber and Strassen 1973, Theorem 4.1). It allows to test between two hypotheses H_0 and H_1, each representing a coherent lower prevision[18] (instead of a probability distribution as in classical Neyman-Pearson testing),

$$H_0: \text{ `}\underline{P}_0 \text{ is true'} \qquad \text{versus} \qquad H_1: \text{ `}\underline{P}_1 \text{ is true'},$$

by proving the existence of a so-called *globally least favourable pair* of probability distributions $p_0 \in \mathcal{M}_0$ and $p_1 \in \mathcal{M}_1$, where $\mathcal{M}_0$ and $\mathcal{M}_1$ are the credal sets corresponding to $\underline{P}_0$ and $\underline{P}_1$, respectively. This least favourable pair is representative for the testing problem regarding $\underline{P}_0$ and $\underline{P}_1$, in the sense that a test that is optimal for distinguishing p_0 and p_1 is also optimal in testing $\underline{P}_1$ against $\underline{P}_0$, independently of significance level α and sample size n.

An important, swiftly evolving, approach for predictive inferences using frequentist imprecise probability is the framework of *nonparametric predictive inference* (NPI, Augustin and Coolen 2004; Coolen 2006).[19] It is based on Hill's (1968) assumption $A_{(n)}$, giving a direct conditional probability for a future observable random quantity based on observed values of related random quantities. Suppose that $X_1, \ldots, X_n, X_{n+1}$ are continuous and exchangeable random quantities. Let the ordered observed values of $X_1, \ldots, X_n$ be denoted by $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$, and let $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$ for ease of notation. For a future observation X_{n+1}, based on n observations, the assumption $A_{(n)}$ (Hill 1968) is

$$\mathrm{P}\left(X_{n+1} \in (x_{(j-1)}, x_{(j)})\right) = \frac{1}{n+1} \quad \text{for } j = 1, 2, \ldots, n+1\,.$$

[18] More precisely, Huber and Strassen (1973) proved the existence of least favourable pairs for capacities, while Augustin (1998) proved existence of such pairs for the more general class *F-Wahrscheinlichkeit*, which is the equivalence of coherent lower previsions in the framework by Weichselberger (2001).

[19] See also `www.npi-statistics.com`.

$A_{(n)}$ does not assume anything else, and is a post-data assumption related to exchangeability. The frequentist nature of $A_{(n)}$ can easily be understood by considering a simple case like $n = 2$, for which $A_{(2)}$ states that the third observation has equal probability to be the minimum, median or maximum of all three observations, no matter the values x_1 and x_2. For repeated applications in situations with exchangeable random quantities, this post-data assumption is clearly seen to hold with a frequency interpretation of probability. For one-off applications, such an inference can be considered reasonable if one has no information at all about the location of X_3 relative to x_1 and x_2, or if one explicitly does not wish to use any such information. $A_{(n)}$ is also closely related to simple random sampling, and for the case with $n = 2$ it just implies that the minimum, mean and maximum of the three random quantities each have the same probability to be X_3.[20]

Inferences based on $A_{(n)}$ are predictive and nonparametric, and avoid the use of so-called *counterfactuals*, which play an important role in classical inferences like hypothesis testing, and which are often criticised by opponents of frequentist statistics (see, e.g., Dawid 2000). $A_{(n)}$ is not sufficient to derive precise probabilities for many events of interest, but it provides optimal bounds for probabilities for all events of interest involving X_{n+1}. These bounds are coherent lower (and upper) probabilities (Augustin and Coolen 2004).

NPI is a framework of statistical theory and methods that use these $A_{(n)}$-based lower and upper probabilities, and also considers several variations of $A_{(n)}$ which are suitable for different inferences. For example, NPI has been presented for Bernoulli data, multinomial data, and lifetime data with right-censored observations; NPI also enables inference for multiple future observations, with their interdependence explicitly taken into account. NPI provides a solution to some central goals formulated for objective (Bayesian) inference, which cannot be obtained when using precise probabilities (Coolen 2006). NPI is also exactly calibrated (Lawless and Fredette 2005), which is a strong consistency property, and it never leads to results that are in conflict with inferences based on empirical probabilities.

Inferential problems for which NPI solutions have recently been presented or are being developed include aspects of medical diagnosis with the use of ROC curves, robust classification, inference on competing risks, quality control and acceptance sampling. To pick out an important application, a generalisation of the Kaplan-Meier estimator for survival functions (Kaplan and Meier 1958) was developed by Coolen and Yan (2004), which effectively expresses the uncertainty inherent in the estimation by lower and upper bounds. We think this is a striking example for the potential of imprecise probability methods, allowing to draw conclusions from data without the need to add overreaching assumptions.

The nature of $A_{(n)}$ as a minimal model assumption also opens the possibility to study the effect of common further modelling assumptions, by comparing NPI-based inferences with the results of classical procedures. As we will repeatedly argue in Sections 2.2.1, 2.2.3.1 and 2.2.4.1 below, often the (imprecise) answers offered by imprecise probability methods are already enough to solve the substantial question at hand, and adding further assumptions to enable precise answers (as is the usual approach) is unnecessary, with the serious risk of giving spuriously precise answers when these further assumptions are

[20] For a detailed study of $A_{(n)}$ see Hill (1988).

unrealistic or unjustified.

Having studied some theoretical foundations of imprecise probability and inference approaches based on them, we will now turn to the reasons motivating the use of imprecise probability in statistical inference. Again, special attention is paid to motives from a Bayesian perspective (Section 2.2.3).

2.2. Motives for the Use of Imprecise Probability Methods

Although strictly negated by advocates of Bayesian methods (e.g., by Lindley 1987), the need to go beyond usual probability measures has been recognized for a long time,[21] and in more recent times most prominently by scientists involved in the development of *expert systems* (see, e.g., Walley 1996b).

Expert systems have the task to formally represent the knowledege of one or several experts in a field, properly modeling the reasoning of these experts in order to aid non-experts in their decisions. A prototypical example is MYCIN (Shortliffe 1976), which was developed to assist physicians in diagnosing bacterial infections. The reasoning to be modeled by expert systems typically involes "uncertainty, partial ignorance and incomplete or conflicting information", and thus provides "an especially good testing ground for theories of uncertainty because they aim to formalise and automate as much as possible of the reasoning process" (Walley 1996b, p. 2).

We will first discuss a central, or fundamental, motive in Section 2.2.1, along with a number of motives that can be seen as derived from this central motive. Before we discuss motives specific to the Bayesian approach in Section 2.2.3, we briefly discern, in consequence of the central motive, the notions of *risk* and *ambiguity* in Section 2.2.2.

2.2.1. The Fundamental Motive

The fundamental motive at play in expert systems, and in the study of uncertainty in artificial intelligence in general (see, e.g., Lawry et al. 2006), is the view that usual probability measures are not expressive enough to model reasoning under uncertainty, and that, in fact, probability is fit to cover only one aspect of the uncertainty involved. The exclusive role of probability as a methodology for handling uncertainty has eloquently been rejected, e.g., by Klir and Wierman (1999, p. 1):

> For three hundred years [...] uncertainty was conceived solely in terms of probability theory. This seemingly unique connection between uncertainty and probability is now challenged [... by several other] theories, which are demonstrably capable of characterizing situations under uncertainty. [...]
>
> [...] it has become clear that there are several distinct types of uncertainty. That is, it was realized that uncertainty is a multidimensional concept. [... That] multidimensional nature of uncertainty was obscured when uncertainty was conceived solely in terms of probability theory, in which it is manifested by only one of its dimensions.

Weichselberger (2001, §1.4) identifies this multidimensionality as a *meta motive*, which underlies a number of more manifest motives for the use of imprecise probability tools (Weichselberger 2001, p. 92):

[21] Hampel (2009b) and Weichselberger (2001, §1) give a historical overview on the development of ideas related to non-additive measures and interval probability.

- The existence of pairs of events that may be not comparable in terms of probability measures, as neither one of them can be seen as more probable than the other, nor should they be seen as excactly equiprobable.

- Incomplete information with respect to a probability measure leads naturally to lower and upper bounds for probabilities of events. These bounds may be, however, already informative enough to decide upon the substantive question at hand.[22]

- Uncertainty in subjective probability assessments, which may manifest in differing buying and selling prices (as discussed in Section 2.1.2). Indeed, Walley's (1991) central motive is that subjective beliefs to be modelled by a prior are usually imprecise, such that lower previsions, rather than probability distributions, are the adequate model (see Section 2.2.3.1).

- The provision of basic tools that generalise probability measures, namely capacities (Choquet 1954), also known as fuzzy measures (e.g., Murofushi and Sugeno 1989), and the Choquet integral, which allows integration according to capacities. These tools open up a wealth of modelling opportunities.[23]

- The need to model approximate adherence to a central probability distribution, which is often done through *neighbourhood models* (see Section 3.2.1.1). This is the dominant model type in robust Bayesian approaches (see, e.g., Berger et al. 1994).

- The available data may not be sufficient to excactly determine the parameters of a usual proability model. Instead, only ranges for these parameters can be identified. A fastly growing literature is devoted to approaches along these lines, especially in econometrics, where it is called *partial identification* (e.g., Manski 2003), and in biometrics, where it is known as *systematic sensitivity analysis* (e.g., Vansteelandt et al. 2006).

- Sequences of experiments that do not warrant the assumption of convergence for relative frequencies, because, e.g., the notion of replication may not be satisfied closely enough as necessary for the validity of results from classical probability theory. In similar veins, imprecise probability may allow to model imperfect randomisation schemes, where, e.g., "symmetry of sample units cannot be fully established" (Augustin, Walter, and Coolen 2013, §2.5).

[22] An important example for such an approach is the theory of *nonparametric predictive inference* (NPI, see, e.g., Coolen 2011), where minimal model assumptions lead to a nonparametric model supplying lower and upper probability bounds for events involving future observations. As mentioned in Section 2.1.4.2, a striking application of this theory is a generalisation of the Kaplan-Meier estimator for survival functions (Kaplan and Meier 1958), which effectively expresses the uncertainty inherent in the estimation by lower and upper bounds (Coolen and Yan 2004).

[23] In fact, certain subclasses of capacities are a special case of lower previsions (see, e.g., Destercke and Dubois 2013b, Fig. 5.5).

In summary, the flexible, multidimensional perspective on uncertainty makes imprecise probabilities capable of mirroring the quality of knowledge. Only well supported knowledge is expressed by comparatively precise models, including the traditional concept of probability as the special case of perfect stochastic information, while highly imprecise (or even vacuous) models are used in the situation of scarce (or no) prior knowledge.

2.2.2. Risk and Ambiguity

The multidimensionality of uncertainty is most often accounted for by distinguishing two specific dimensions, commonly denoted by *risk* and *ambiguity* (e.g., by Ellsberg 1961).

- Risk is the uncertainty involved in ideal lotteries, i.e., when the stochastic mechanism driving the phenomenon under uncertainty (the data generating process) is known completely. Precise probability distributions are the adequate tool to model such uncertainty, arising, e.g., in a lottery where the number of winning tickets is exactly known.

- Ambiguity arises when there is insufficient information about the stochastic mechanism, or information about it is lacking completely, and covers thus non-stochastic uncertainty. Uncertain phenomena where there is no information at all about ocurrence or no-ocurrence of events would constitute a situation of pure ambiguity, like a lottery for which there is no information whatsoever about the number of winning tickets.

Ellsberg's (1961) seminal experiment demonstrated that, in contrast to the then dominant theoretical frameworks for rational decisions under uncertainty (most prominently, Savage 1954), not all decisions can be framed in terms of *risk*, and that aspects of *ambiguity* do form a crucial part of the decision process and should not be glossed over.[24]

Indeed, in our view, real-life situations often pose mixtures of risk and ambiguity, and should be modeled by lower previsions or sets of probability distributions. In the latter formulation, the stochastic aspect is dealt with by the single distributions included in the set, whereas ambiguity is expressed by the magnitude of the set itself, which manifests, e.g., in the length of intervals for probabilities of events derived from the set. Sets of probability distributions are thus, in our view, an adequate model to characterise, e.g., a lottery where there is some limited information about the number of winning tickets, like the information that there should be about 5 to 10 winning tickets per 100 tickets.[25]

[24] More recently, the study by Hsu et al. (2005) even suggests that decision problems involving ambiguity are processed by different cerebral mechanisms as those involving pure risk.

[25] We will comment on the seemingly attractive alternative of putting a second-order distribution on the set of winning probabilities at Section 2.2.4.

2.2.3. Motives from a Bayesian Perspective

From a decidedly Bayesian perspective as taken in this thesis, we first want to point out foundational motives for the use of imprecise probability methods, before we come to two specific motives in Sections 2.2.3.2 and 2.2.3.3.

2.2.3.1. Foundational Motives

We agree with Berger et al. (1994, §1.1) in the view that there are very strong foundational arguments for the subjective Bayesian approach to statistical inference, which, however, hold only "if it is assumed that one can make arbitrarily fine discriminations in judgement about unknowns and utilities" (Berger et al. 1994, p. 303). It is the implications of this extremely challenging requirement which advocates of Bayesian inference with precise priors like Lindley (1987) do not seem to appreciate enough. In fact, it is hardly imaginable how such "arbitrarily fine discriminations" could be made in practice even for the most basic inference tasks. The foundational arguments for Bayesian inference are thus worthless, unless the requirement of arbitrarily fine precision can be relaxed. Indeed, the framework by Walley (1991) does just that: preserving the foundational arguments, especially the notion of coherence, while allowing for incomplete and imprecise prior judgements.[26]

In fact, Walley consequently concludes that precise priors are unnecessary for coherent inference, and that even the assumption of an underlying 'ideal' precise prior (unobtainable because of, e.g., time constraints in elicitation) is unjustified, and calls this the *dogma of ideal precision* (Walley 1991, §5.9).[27]

In a decision theoretic context (which can be seen as a superstructure for statistical inference tasks, see Section 1.2.3.2), Walley argues that the strife to obtain precise prior distributions is unnecessary also because often, imprecise probability will suffice to determine an optimal decision, and in the contrary case, they adequately reflect the lack of information inherent in the decision problem. Then, precise decisions are in fact based on some arbitrary modelling choices,[28] but this arbitrariness is obscured by the method insisting on precision, such that the precision of the method is merely a spurious one. Consequently, it is better to acknowledge indecision instead of hiding it, as—if it must—a decision can be made nevertheless, with the same arbitrariness as is hidden in a spuriously precise method (Walley 1991, §5.7).

Concretely, the power to differentiate between different degrees of partial knowledge distinguishes imprecise probabilities as an attractive modelling tool in statistics. In particular

[26] However, keep in mind the caveats for an inference procedure based on the Generalised Bayes' Rule as adressed in Section 2.1.3.2.

[27] When disagreeing with such a view, remarking that precise Bayesian methods can be adequate in many inference settings, Walley replies that nevertheless, "[...] we need a theory of imprecise probabilites to tell us when precision is a poor idealisation" (Walley 1991, §5.8.1, item 3, p. 250).

[28] E.g., "the decisions which result from definite rules such as maximising entropy or assigning a 'non-informative' second-order distribution will depend on the arbitrary choice of a possibility space Ω" (Walley 1991, §5.6, footnote 16, p. 531). We will comment on second-order distributions and hierarchical Bayesian models in Section 2.2.4.

it allows to overcome two severe practical limitations inherent to Bayesian inference based on precise probabilities, briefly discussed below.

2.2.3.2. Weakly Informative Priors

A first important issue is the proper modelling of no (or extremely vague) prior knowledge. In traditional statistics, so-called *non-informative priors* have been proposed, which, by different criteria, eventually all select a single traditional prior distribution, turning ignorance into a problem with rather precise probabilistic information. Furthermore, these 'non-informative' priors are usually improper, leading to severe problems in hypotheses testing (as mentioned in Section 1.2.3.3). We will comment on this issue in more detail in Section 3.1.2, item V, and in Section 3.1.3.

2.2.3.3. Prior-Data Conflict

Moreover, increasing imprecision to make the conclusions more cautious is the natural way to handle *prior-data conflict*, i.e. when outlier-free data are observed that nevertheless are rather unexpected under the prior assumptions. For instance, Evans and Moshonov (2006, p. 893) warn that if "[...] the observed data is surprising in the light of the sampling model and the prior, then we must be at least suspicious about the validity of inferences drawn." While there is no way to express this caution ('being suspicious') in precise probabilistic models, the imprecise probability models at the core of this thesis (see Chapter 3) were developed specifically to take care of this issue in an appropriate way. A more detailed discussion on the issue of prior-data conflict will thus be given in Section 3.1.2, item IV, and in Section 3.1.4; these treatments are in turn based on the studies in Sections 3.3 and 3.5. Some illustrative examples of prior-data conflict can be found in Section A.1.2, while the effect on Bayesian linear regression estimates is discussed in Sections A.1.3 and A.1.4.

As a topic related to prior-data conflict, decreasing imprecision in sequential learning may express naturally that the accumulating information is non-conflicting and stable.[29]

2.2.4. Critique and Discussion of Some Alternatives

We will briefly discuss here some critiques on imprecise probability methods, and touch on hierarchical models as an alternative. Other models based on sets of prior distributions will instead be discussed in Section 3.2.

2.2.4.1. Objections to Imprecise Probability Models

It is often argued that imprecise probability methods are more difficult to apply than precise models, and lead to complicated and cumbersome inference procedures.

While the criticism of higher complexity is certainly true for the mathematical foundations (lower previsions are indeed more complex than linear previsions or precise probability

[29] See the comment in footnote 10, page 61.

distributions), we believe that increased effort (if actually substantial) in modeling and application for imprecise probability methods is rewarded with a more realistic description of the uncertainty involved. Imprecise probability models lead to more reliable conclusions, by explicitely communicating the uncertainties inherent in the model or the data, in contrast to the often seemingly precise results from precise probability methods, which are frequently obtained due to a multitude of, often heroic, model assumptions. This 'overprecision' is often offset in statistical practice by 'taking models not too seriously', and to understand them as crude approximations to reality only. Box and Draper's statement "Essentially, all models are wrong, but some of them are useful" (Box and Draper 1987, p. 424) has become a frequently cited dictum, often understood as a guiding paradigm to statistical modeling. Taking aside the problematic assumption of 'continuity' of statistical procedures inherent in this view,[30] employing precise models and then discounting their precise results seems somehow circuitous.

We think models are preferable that, by allowing for imprecision, can be taken seriously in their implications. This is also the position of Manski, whose "Law of decreasing credibility",[31]

> "The credibility of inferences decreases with the strength of the assumptions maintained" (Manski 2003, p. 1),

can be understood as a compelling motto for statistical inference with imprecise probabilities, bringing credibility, or reliability, of conclusions explicitly into the argument, and encouraging us to impose justified assumptions only.

Furthermore, the perceived difficulty of applying imprecise probability methods in practice is, at least in the subjective Bayesian inference paradigm, often negligible. Sets of prior distributions are relatively easy to handle (see, e.g., the models in Section 3.1), and the crucial point in practical analyses instead is often the choice of the prior. For this, e.g., probabilities or quantiles must be elicited by questioning an expert in the field under study. To us it seems hardly questionable that it will be easier for such an expert to provide ranges instead of precise numbers.[32] What happens in practice already is that experts are often much more comfortable to provide ranges instead of precise numbers,[33] and again, it seems unnecessary and circuitous to derive precise probability statements from such imprecise assessments. We thus consider the common objection to imprecise Bayesian approaches that eliciting two numbers instead of one must be more difficult a misconception.[34]

[30] 'Continuity' is understood here in the sense that small perturbations of the underlying model do not destroy the substantive conclusions drawn from the data. Results about the robustness of statistical models have undermined this assumption of 'model continuity' (see Huber (1981) and Hampel et al. (1986) for monographs on robust statistics).

[31] "Credibility" is used here in a non-technical sense.

[32] See a simple illustration of this in Walley (1991, §5.8, footnote 7, p. 535).

[33] See, e.g., the study by Rinderknecht, Borsuk, and Reichert (2011), where the imprecision naturally present in experts' assessments is not ignored, but adequately modelled.

[34] See, e.g., Walley (1991, §5.8.2) for a more detailed discussion of this objection.

2.2.4.2. Hierarchical Models

Another obvious approach when considering a set of probability distributions $\mathcal{M}$ as the model for prior information is to rate the elements in $\mathcal{M}$, and devise a *second-order distribution* or a *hyperprior* on $\mathcal{M}$. Indeed, this is a common approach, known as *hierarchical Bayesian modelling*, where usually, $\mathcal{M}$ is a family of parametric distributions indexed by a *hyperparameter* ϕ, and the hyperprior over ϕ is a non-informative prior.

As Walley (1991, §5.10.4, p. 260) notes, "hierarchical models can be very useful when the hyperparameter ϕ has a clear meaning". However, this is very often not the case, and ϕ serves "merely as an index for the unknown, uninterpreted [ideal, precise prior] p_T." Refuting the need for such an ideal precise prior p_T (as noted in Section 2.2.3.1), Walley (1991, §5.10.2) demonstrates that such an approach leads to incoherence when the priors in $\mathcal{M}$ are meant to model behavioural dispositions, as is the case for subjective Bayesian approaches.

Leaving aside the serious problems that can arise when ignorance about ϕ is modelled by a non-informative prior,[35] a second-order distribution approach actually defines, via averaging over the elements in $\mathcal{M}$, a precise prior on the first level, leading to precise posterior inferences, thus obscuring the uncertainty in the model that was the reason for devising a second-order distribution in the first place.

In our view, a much better alternative for hierarchical modelling was suggested by Cattaneo (2008). Here, information on the plausibility of elements in $\mathcal{M}$ is modelled by a possibility distribution, such that uncertainty in posterior inferences is reflected again by a possibility distribution, which takes the form of a normalised likelihood function and can be interpreted, like usual likelihoods, as a measure of relative plausibility.[36]

In a non-Bayesian context, the model consists of a set of sample distributions $\mathcal{M}$ (it is possible to take a whole family of sample distributions as $\mathcal{M}$), and of the likelihood function $\mathrm{lik} : \mathcal{M} \to \mathbb{R}_{\geq 0}$ as the second level, describing the plausibility of its elements based on the sample. Inference on certain deduced quantities $\vartheta = g(p), p \in \mathcal{M}$ then can be based on the so-called *profile likelihood* $\mathrm{proflik}(\vartheta) = \sup_{\{p \in \mathcal{M} | g(p) = \vartheta\}} \mathrm{lik}(p)$. A variant of this model allows to incorporate prior information by using a so-called prior likelihood. The approach can serve as a basis for a direct likelihood-based decision theory (Cattaneo 2007; Cattaneo 2012). It has been successfully applied in graphical models (e.g., Antonucci, Cattaneo, and Corani 2011, see also Corani et al. 2013, §4.3), and in the context of regression with

[35] See, e.g., Walley (1991, §5.5.4 (h)), or Pericchi and Nazaret (1988, p. 367).

[36] The possibility distributions for quantities of interest can also be seen as *fuzzy numbers*, such that, e.g., the coverage probability for a certain parameter subset Θ_1 is not a real number, but a fuzzy probability, expressing the uncertainty for that probability by relative plausibility values (Cattaneo 2008). The hierarchical model has thus close relations to approaches using fuzzy numbers and distributions, with the advantage of a clear interpretation of the possibility functions as likelihoods. Furthermore, this approach has a sound foundation for combination rules and deduction of inferences in probability theory. In contrast, in the fuzzy literature, the interpretation of fuzzy numbers, or membership functions in general, usually explained as expressing 'degrees of membership', remains often unclear, the rules for combining different inference sources are controversially disputed, and the fuzzy literature disagrees on how to update fuzzy probability models in the light of data (Cattaneo 2008, §3).

imprecise data (Cattaneo and Wiencierz 2012).

After having described the theoretical foundations of, and having motivated the switch to, imprecise probability models in general, in Chapter 3 we will now present models for Bayesian inference with imprecise proabilities that are practicable and account for the modelling opportunities discussed above.

3. Generalised Bayesian Inference with Sets of Conjugate Priors in Exponential Families

The imprecise probability models based on conjugate priors we will discuss in this chapter are an important model class, and have been central to the development and application of imprecise probability methods in statistical inference. They have lead to the so-called *Imprecise Dirichlet model* (IDM, see Section 3.1.3 below) by Walley (1996a), and more generally to powerful imprecise probability models for inference based on i.i.d. exponential family sampling models by Quaeghebeur and Cooman (2005) and Quaeghebeur (2009). These models were extended by Walter and Augustin (2009b, see Section 3.3) and Walter, Augustin, and Coolen (2011, see Section 3.5) to allow in addition an elegant handling of prior-data conflict.

The chapter is structured as follows. Section 3.1 attempts to give a systematic overview on these models, and illustrates some characteristic modelling opportunities of generalised Bayesian inference, while Section 3.2 briefly discusses some alternative models based on sets of priors. Section 3.3 then presents one model in detail, namely the so-called generalised iLUCK-model developed in Walter and Augustin (2009b), and illustrates its application to the Normal-Normal and Dirichlet-Multinomial models. Section 3.4 gives a short overview on a software implementation of generalised iLUCK-models, the add-on package `luck` (Walter and Krautenbacher 2013) for the statistical programming environment **R** (R Core Team 2013). Finally, Section 3.5 presents two attempts to further refine inference behaviour in the presence of prior-data conflict. While the first approach considers more sophisticated shapes for prior parameter sets, the second is a fundamentally different approach that, combining inferences from two arbitrary distinct models, nevertheless shows fascinating similarities to the first approach.

3.1. Model Overview and Discussion

This section gives a systematic overview on imprecise probability models for inference in canonical exponential families based on sets of canonical conjugate priors. It integrates the approaches discussed in Sections 3.3 and 3.5, and further approaches discussed in the literature, into a general framework, and reviews their inference properties.

In Section 3.1.1, a general framework that characterises these models is elaborated, and models discussed in the literature that can be subsumed under this framework are described. Section 3.1.2 presents a number of inference properties that this framework provides, along with two further criteria characterising the unique modelling opportunities that generalised Bayesian inference can offer: sensitivity to prior-data conflict, and the case of weakly informative priors. The models from Section 3.1.1 are then discussed in the light of the two criteria in Sections 3.1.3 and 3.1.4, providing a summary of model strengths and weaknesses.

3.1.1. The General Framework

Consider inference based on samples from a regular canonical exponential family (1.4) using the conjugate prior (1.5) as discussed in Section 1.2.3.1. One specifies a prior parameter set $\mathbb{\Pi}^{(0)}$ of $(n^{(0)}, y^{(0)})$ values and takes as imprecise prior—described via the credal set $\mathcal{M}^{(0)}$—the set of traditional priors with $(n^{(0)}, y^{(0)}) \in \mathbb{\Pi}^{(0)}$. The credal set $\mathcal{M}^{(n)}$ of posterior distributions,[1] obtained by updating each element of $\mathcal{M}^{(0)}$ via Bayes' Rule, then can be described as the set of parametric distributions with parameters varying in the set of updated parameters $\mathbb{\Pi}^{(n)} = \{(n^{(n)}, y^{(n)}) | (n^{(0)}, y^{(0)}) \in \mathbb{\Pi}^{(0)}\}$.

Alternatively, $\mathcal{M}^{(0)}$ can be defined as the set of all convex mixtures of parametric priors with $(n^{(0)}, y^{(0)}) \in \mathbb{\Pi}^{(0)}$. In this case, the set of priors corresponding to $\mathbb{\Pi}^{(0)}$ considered above gives the set of extreme points for the actual convex set $\mathcal{M}^{(0)}$. Updating this convex prior credal set with the Generalized Bayes' Rule results in a set $\mathcal{M}^{(n)}$ of posterior distributions that is again convex, and $\mathcal{M}^{(n)}$ conveniently can be obtained by taking the convex hull of the set of posteriors defined by the set of updated parameters $\mathbb{\Pi}^{(n)}$.

To see this, consider a mixture distribution $p_m(\vartheta \mid n_1^{(0)}, y_1^{(0)}, n_2^{(0)}, y_2^{(0)}, \kappa)$ based, without loss of generality, on only two parametric components with parameters $(n_1^{(0)}, y_1^{(0)})$ and $(n_2^{(0)}, y_2^{(0)})$, respectively:

$$p_m(\vartheta \mid n_1^{(0)}, y_1^{(0)}, n_2^{(0)}, y_2^{(0)}, \kappa) := \kappa\, p(\vartheta \mid n_1^{(0)}, y_1^{(0)}) + (1-\kappa)\, p(\vartheta \mid n_2^{(0)}, y_2^{(0)})\,.$$

Denoting the marginals by

$$f_1(\boldsymbol{x}) = \int_\Theta f(\boldsymbol{x} \mid \vartheta) p(\vartheta \mid n_1^{(0)}, y_1^{(0)})\, \mathrm{d}\vartheta\,,$$

[1] To emphasize the dependence of the sample size n, the posterior credal set is denoted by $\mathcal{M}^{(n)}$ instead of $\mathcal{M}_{|\boldsymbol{x}}$ as in (2.2).

$$
\begin{aligned}
f_2(\boldsymbol{x}) &= \int_\Theta f(\boldsymbol{x} \mid \vartheta) p(\vartheta \mid n_2^{(0)}, y_2^{(0)}) \,\mathrm{d}\vartheta \,, \\
f_m(\boldsymbol{x}) &= \int_\Theta f(\boldsymbol{x} \mid \vartheta) p_m(\vartheta \mid n_1^{(0)}, y_1^{(0)}, n_2^{(0)}, y_2^{(0)}, \kappa) \,\mathrm{d}\vartheta \\
&= \kappa \int_\Theta f(\boldsymbol{x} \mid \vartheta) p(\vartheta \mid n_1^{(0)}, y_1^{(0)}) \,\mathrm{d}\vartheta + (1-\kappa) \int_\Theta f(\boldsymbol{x} \mid \vartheta) p(\vartheta \mid n_2^{(0)}, y_2^{(0)}) \,\mathrm{d}\vartheta \\
&= \kappa \int_\Theta f_1(\boldsymbol{x}) p(\vartheta \mid n_1^{(n)}, y_1^{(n)}) \,\mathrm{d}\vartheta + (1-\kappa) \int_\Theta f_2(\boldsymbol{x}) p(\vartheta \mid n_2^{(n)}, y_2^{(n)}) \,\mathrm{d}\vartheta \\
&= \kappa f_1(\boldsymbol{x}) + (1-\kappa) f_2(\boldsymbol{x}) \,,
\end{aligned}
$$

the posterior corresponding to $p_m(\vartheta \mid n_1^{(0)}, y_1^{(0)}, n_2^{(0)}, y_2^{(0)}, \kappa)$, derived via Bayes' Rule, can then be written as

$$
\begin{aligned}
& p_m(\vartheta \mid n_1^{(0)}, y_1^{(0)}, n_2^{(0)}, y_2^{(0)}, \kappa, \boldsymbol{x}) \\
&\quad = \frac{f(\boldsymbol{x} \mid \vartheta)}{f_m(\boldsymbol{x})} \Big(\kappa \, p(\vartheta \mid n_1^{(0)}, y_1^{(0)}) + (1-\kappa) \, p(\vartheta \mid n_2^{(0)}, y_2^{(0)}) \Big) \\
&\quad = \kappa \, \frac{f(\boldsymbol{x} \mid \vartheta)}{f_m(\boldsymbol{x})} p(\vartheta \mid n_1^{(0)}, y_1^{(0)}) + (1-\kappa) \, \frac{f(\boldsymbol{x} \mid \vartheta)}{f_m(\boldsymbol{x})} p(\vartheta \mid n_2^{(0)}, y_2^{(0)}) \\
&\quad = \kappa \, \frac{f_1(\boldsymbol{x})}{f_m(\boldsymbol{x})} p(\vartheta \mid n_1^{(n)}, y_1^{(n)}) + (1-\kappa) \, \frac{f_2(\boldsymbol{x})}{f_m(\boldsymbol{x})} p(\vartheta \mid n_2^{(n)}, y_2^{(n)}) \\
&\quad =: p_m(\vartheta \mid n_1^{(n)}, y_1^{(n)}, n_2^{(n)}, y_2^{(n)}, \kappa^*) \,,
\end{aligned} \tag{3.1}
$$

where

$$
\kappa^* = \kappa \, \frac{f_1(\boldsymbol{x})}{f_m(\boldsymbol{x})} = \frac{\kappa \, f_1(\boldsymbol{x})}{\kappa \, f_1(\boldsymbol{x}) + (1-\kappa) \, f_2(\boldsymbol{x})} \,.
$$

The posterior of the κ-mixture distribution, based on parametric priors with parameters $(n_1^{(0)}, y_1^{(0)})$ and $(n_2^{(0)}, y_2^{(0)})$, is thus the κ^*-mixture of the two parametric posteriors with updated parameters $(n_1^{(n)}, y_1^{(n)})$ and $(n_2^{(n)}, y_2^{(n)})$, respectively.

Now, the convex hull of $p(\vartheta \mid n_1^{(0)}, y_1^{(0)})$ and $p(\vartheta \mid n_2^{(0)}, y_2^{(0)})$ is given by the set of all mixture distributions based on $p(\vartheta \mid n_1^{(0)}, y_1^{(0)})$ and $p(\vartheta \mid n_2^{(0)}, y_2^{(0)})$, where κ varies in $[0,1]$. As for any $\kappa \in [0,1]$, the corresponding κ^* is again in $[0,1]$, and it holds that $\{\kappa^* \mid \kappa \in [0,1]\} = [0,1]$, the convex hull of $p(\vartheta \mid n_1^{(n)}, y_1^{(n)})$ and $p(\vartheta \mid n_2^{(n)}, y_2^{(n)})$ is indeed equivalent to the set of updated mixtures $p_m(\vartheta \mid n_1^{(0)}, y_1^{(0)}, n_2^{(0)}, y_2^{(0)}, \kappa, \boldsymbol{x})$ for $\kappa \in [0,1]$.

Through complete induction, this result holds also for any finite number of mixture components, such that the set of updated mixture distributions $\mathcal{M}^{(n)}$ can thus be constructed as the convex hull of the parametric posteriors with $(n^{(n)}, y^{(n)}) \in \mathbb{\Pi}^{(n)}$.

When the credal set $\mathcal{M}^{(0)}$ is taken to contain all finite mixtures of parametric priors, $\mathcal{M}^{(0)}$ is very flexible and contains, through the mixture distributions, a wealth of distributional shapes.[2] Nevertheless, maximisation and minimisation over $\mathcal{M}^{(n)}$ is quite feasible for

[2]Indeed, if the parametric distributions are normal distributions and $\mathbb{\Pi}^{(0)}$ is large enough, it can be

quantities that are *linear* in the parametric posteriors contained in $\mathcal{M}^{(n)}$, as then we can be sure that suprema and infima are attained at the extreme points of $\mathcal{M}^{(n)}$, which are the parametric distributions generated by $\mathbb{\Pi}^{(n)}$, enabling us to search over $\mathbb{\Pi}^{(n)}$ only. Extremes for prior and posterior quantities of interest that are not linear in the parametric distributions may be very difficult to obtain, or the model could even be useless for them (because the model may give too imprecise, or even vacuous, bounds). Expectations are linear in the parametric distributions, while variances are not.[3]

For both cases, the relationship between the parameter sets $\mathbb{\Pi}^{(0)}$ and $\mathbb{\Pi}^{(n)}$ and the credal sets $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(n)}$ will allow us to discuss different models $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(n)}$ by considering the corresponding parameter sets $\mathbb{\Pi}^{(0)}$ and $\mathbb{\Pi}^{(n)}$.[4]

As an example, in the precise Beta-Bernoulli model as discussed in Section 1.2.3.3, the posterior predictive probability for the event that a future single draw is a success is equal to $y^{(n)}$, and so we get, for an imprecise model with $\mathcal{M}^{(0)} = \{p(\theta \mid n^{(0)}, y^{(0)}) | (n^{(0)}, y^{(0)}) \in \mathbb{\Pi}^{(0)}\}$, the lower and upper probability

$$\underline{y}^{(n)} := \inf_{\mathbb{\Pi}^{(n)}} y^{(n)} = \inf_{\mathbb{\Pi}^{(0)}} \frac{n^{(0)} y^{(0)} + s}{n^{(0)} + n},$$
$$\overline{y}^{(n)} := \sup_{\mathbb{\Pi}^{(n)}} y^{(n)} = \sup_{\mathbb{\Pi}^{(0)}} \frac{n^{(0)} y^{(0)} + s}{n^{(0)} + n}.$$

Special imprecise probability models are now obtained by specific choices of $\mathbb{\Pi}^{(0)}$. We distinguish the following types of models:

(a) $n^{(0)}$ is fixed, while $y^{(0)}$ varies in a set $\mathcal{Y}^{(0)}$.
The IDM (Walley 1996a), as well as its generalisation to all sample distributions of the canonical exponential form (1.4) by Quaeghebeur and Cooman (2005) are of this type. The approach by Boratyńska (1997) also belongs to this category, as she specifies bounds for $n^{(0)} y^{(0)}$ while holding $n^{(0)}$ constant (see Benavoli and Zaffalon 2012, p. 1973).

(b) $n^{(0)}$ varies in a set $\mathcal{N}^{(0)}$, while $y^{(0)}$ is fixed.
This type of model is rarely discussed in the literature, but is mentioned by Walley (1991) in §7.8.3 and in §1.1.4, footnote no. 10. Both instances assume the Normal-Normal model as described in Section 1.2.3.4, where the set of priors is spanned by normal distributions with a fixed mean $y^{(0)}$ and a range of variances $\sigma_0^2/n^{(0)}$.

assumed that $\mathcal{M}^{(0)}$ contains a very wide range of priors, as mixtures of normal distributions are dense in the space of well-behaved probability distributions (see, e.g., Priebe and Marchette 2000, p. 44, or Ferguson 1983).

[3] In the context of decision making (as mentioned in Sections 1.2.3.2 and 1.2.3.3), the Bayes criterion selects as optimal acts those acts which minimise posterior risk, where posterior risk is the expected loss under the posterior distribution. With credal sets, the optimal acts are usually determined by minimising the upper posterior risk (see, e.g., Huntley, Hable, and Troffaes 2013, §3.2). The posterior risk, being an expectation, is thus a quantity linear in the posterior distribution, such that its upper bound for posteriors in $\mathcal{M}^{(n)}$ can be easily determined even if $\mathcal{M}^{(n)}$ is taken to contain convex combinations.

[4] Note that, although, by the general framework, the credal sets $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(n)}$ may be defined as convex hulls, the parameter sets $\mathbb{\Pi}^{(0)}$ and $\mathbb{\Pi}^{(n)}$ generating them need not necessarily be so, and typically are not convex, indeed. See, e.g., Figure 3.9 on page 107.

(c) Both $n^{(0)}$ and $y^{(0)}$ vary in a set $\{(n^{(0)}, y^{(0)}) \mid n^{(0)} \in \mathcal{N}^{(0)}, y^{(0)} \in \mathcal{Y}^{(0)}\}$.
This type of model is first discussed in Walley (1991, §5.4.3) for the Beta-Bernoulli model, and was later generalised by Walter and Augustin (2009b) to sample distributions of the canonical exponential form (1.4).[5] We have used a model of this type in Section 1.3.6.2, and will discuss and illustrate this approach in more detail in Section 3.3. It should be noted here that while the prior parameter set is a Cartesian product of $\mathcal{N}^{(0)} = [\underline{n}^{(0)}, \overline{n}^{(0)}]$ and $\mathcal{Y}^{(0)}$, the posterior parameter set is not. This is due to Eq. (1.6), which results in different ranges for $y^{(n)}$ depending on the value of $n^{(0)}$ used in the update step.[6]

(d) Both $n^{(0)}$ and $y^{(0)}$ vary in other sets $\mathbb{\Pi}^{(0)} \subset (\mathbb{R}_{>0} \times \mathcal{Y})$.
In this type, also in the prior parameter set the range of $y^{(0)}$ may depend on $n^{(0)}$, as in Walter, Augustin, and Coolen (2011, §2.3, see Section 3.5.2.3), or, vice versa, the range of $n^{(0)}$ may depend on the value of $y^{(0)}$, as in Benavoli and Zaffalon (2012).

3.1.2. Properties and Criteria

Before discussing the approaches mentioned above in some detail, we will describe some properties that all of them have in common. These properties are due to the update mechanism (1.6) for $n^{(0)}$ and $y^{(0)}$ and the resulting size and position of $\mathbb{\Pi}^{(n)}$, being a direct consequence of the (Generalised) Bayes' Rule in the setting of canonical exponential families. Remember that $n^{(0)}$ is incremented with n, while $y^{(n)}$ is a weighted average of $y^{(0)}$ and the sample statistic $\tilde{\tau}(\boldsymbol{x})$, with weights $n^{(0)}$ and n, respectively. Thus, while the (absolute) stretch of $\mathbb{\Pi}^{(n)}$ in the $n^{(0)}$ resp. $n^{(n)}$ dimension will not change during updating, the stretch in the $y^{(0)}$ resp. $y^{(n)}$ dimension will do so. When speaking of the size of $\mathbb{\Pi}^{(n)}$, we will thus refer to the stretch in the main parameter dimension, also denoted by $\Delta_y(\mathbb{\Pi}^{(n)})$.

I. The larger n relative to (values in the range of) $n^{(0)}$, ceteris paribus (c.p.) the smaller $\mathbb{\Pi}^{(n)}$, i.e. the more precise the inferences. Vice versa, the larger the $n^{(0)}$ value(s) as compared to n, c.p. the larger $\mathbb{\Pi}^{(n)}$, and the more imprecise the inferences. Thus, a high weight on the imprecise prior $\mathcal{M}^{(0)}$ will lead to a more imprecise posterior $\mathcal{M}^{(n)}$.[7]

II. In particular, for $n \to \infty$, the stretch of $y^{(n)}$ in $\mathbb{\Pi}^{(n)}$ will converge towards zero, i.e. $\Delta_y(\mathbb{\Pi}^{(n)}) \to 0$, with the limit located at $\tilde{\tau}(\boldsymbol{x})$. For inferences based mainly

[5] More precisely, Walter and Augustin (2009b, see Section 3.3) proposed a direct extension of the model framework discussed in Section 3.1 by the definition of so-called *LUCK-models*, which were already used in Walter, Augustin, and Peters (2007) to generalise Bayesian linear regression. These *LUCK-models* utilize the fact that the central properties of the model framework (see Section 3.1.2 below) rely on the specific form of the update step (1.6) only. Thus, these properties can be generalised to settings that are not based on i.i.d. observations from canonical exponential family distributions, but nevertheless follow the same update step.

[6] This change of shape of the parameter $\mathbb{\Pi}^{(n)}$ set is illustrated in Figure 3.9, page 107. We discuss the phenomenon of shape change in Section 3.5.2.2, and again in Sections 4.3 and A.2.1.

[7] For model types with fixed $n^{(0)}$, if $n^{(0)} = n$, then $\Delta_y(\mathbb{\Pi}^{(n)}) = \Delta_y(\mathbb{\Pi}^{(0)})/2$, i.e. the width of the posterior expectation interval is half the width of the prior interval. This fact may also guide the choice of $n^{(0)}$.

on $y^{(n)}$, this leads to a convergence towards the 'correct' conclusions. This applies, e.g., to point estimators like the posterior mean or median, which converge to the 'true' parameter, to interval estimates (HPD intervals) contracting around this point estimate (length of interval $\to 0$), and to the probability that a test decides for the 'right' hypothesis, which converges to 1.[8] This property, that holds also for traditional (precise) Bayesian inferences, is similar to the consistency property often employed in frequentist statistics.

III. The larger $\Delta_y(\mathbb{\Pi}^{(0)})$, the larger c.p. $\Delta_y(\mathbb{\Pi}^{(n)})$; a more imprecise prior $\mathcal{M}^{(0)}$ will naturally lead to a more imprecise posterior $\mathcal{M}^{(n)}$, which carries over to the inferences.

Apart from the above properties that are guaranteed in all of the four model types (a) – (d), one might want the models to adhere to (either of) the following additional criteria:

IV. **Prior-data conflict sensitivity.** In order to mirror the state of posterior information, posterior inferences should, all other things equal, be more imprecise in the case of prior-data conflict. Reciprocally, if prior information and data coincide especially well, an additional gain in posterior precision may be warranted. Such models could deliver (relatively) precise answers when the data confirm prior assumptions, while rendering much more cautionary answers in the case of prior-data conflict, thus leading to cautious inferences if, and only if, caution is needed.

Most statisticians using precise priors would probably expect a more diffuse posterior in case of prior-data conflict. However, in the canonical conjugate setting of Eq. (1.5), which is often used when data are scarce and only strong prior beliefs allow for a reasonably precise inference answer, this is usually not the case. E.g., for the Normal-Normal model, the posterior variance (1.11) is not sensitive to the location of $\bar{\boldsymbol{x}}$, and decreases by the factor $n^{(0)}/(n^{(0)}+n)$ for any $\bar{\boldsymbol{x}}$, thus giving a false certainty in posterior inferences in case of prior-data conflict. In the Beta-Binomial model, the posterior variance $y^{(n)}(1-y^{(n)})/(n^{(0)}+n)$ depends on the location of s/n, but in a similar way as the prior variance $y^{(0)}(1-y^{(0)})/n^{(0)}$ depends on $y^{(0)}$, mirroring only the fact that Beta distributions centered at the margins of the unit interval are constrained in their spread. Thus, there is no systematic reaction to prior-data conflict also in this case.

In the imprecise Bayesian framework as discussed here, prior-data conflict sensitivity translates into having a larger $\mathbb{\Pi}^{(n)}$ (leading to a larger $\mathcal{M}^{(n)}$) in case of prior-data conflict, and, mutatis mutandis, into a smaller $\mathbb{\Pi}^{(n)}$ if prior and data coincide especially well.

V. **Possibility of weakly or non-informative priors.** When only very weak or (almost) no prior information is available on the parameter(s) one wishes to learn about,

[8] In the binomial and normal example, the posteriors in $\mathcal{M}^{(n)}$ will concentrate all their probability mass at $y^{(n)} \to \tilde{\tau}(\boldsymbol{x})$, and as $\tilde{\tau}(\boldsymbol{x}) \to \theta$ in probability, all these inference properties follow.

it should be possible to model this situation adequately. The traditional Bayesian approach to this problem, so-called *non-informative priors*, are, due to their nature as single, i.e. precise, probability distributions, not expressive enough; a precise probability distribution necessarily produces a precise value for $P(\vartheta \in A)$ for any $A \subseteq \Theta$, which seems incompatible with the notion of prior ignorance about ϑ. Furthermore, in the literature there are several, often mutually incompatible, approaches to construct non-informative priors, such as Laplace's prior, Jeffreys' prior, or reference priors (see, e.g., Bernardo and Smith 2000, §5.6.2). Most of these precise priors seem to convey mainly a state of indifference instead of ignorance (Rüger 1999, p. 271).[9]

The approaches mentioned in (a) – (d) are now discussed in detail, sorted according to two basic scenarios regarding the intended use: When no prior information on ϑ is available (or nearly so), so-called *near-ignorance priors* are used to model the state of prior ignorance. When, in contrast, there is substantial prior information available, the challenge is to model this information adequately in the prior, while ensuring easy handling and prior-data conflict sensitivity.[10]

3.1.3. The IDM and other Prior Near-Ignorance Models

The *Imprecise Dirichlet model* (IDM) was developed by Walley (1996a) as a model for inferences from multinomial data when no prior information is available.[11] As indicated by its name, it uses as imprecise prior a (near-) noninformative set of Dirichlet priors, which

[9] E.g., in the designation of the uniform prior as a non-informative prior by the principle of insufficient reason (i.e., taking Laplace's prior), it is argued that there is no reason to favor one parameter value over another, and thus, all of them should get the same probability resp. density value. For analysts restricting themselves to precise priors, this argument leads necessarily to the uniform prior. When considering imprecise priors, however, the principle of insufficient reason does not uniquely determine a certain prior. It only states that the probability resp. density *interval* should be equal for all parameter values, but that interval may be any interval $\subseteq [0, 1]$. We may thus realise that the principle of insufficient reason actually implies *indifference* between parameter values only, and that other considerations are needed to distinguish a certain imprecise prior as (nearly) non-informative; usually, it is postulated that (a certain class of) inferences based on the prior should be (nearly) vacuous (see, e.g., Benavoli and Zaffalon 2012), and specific information (e.g., symmetry of parameter values) would be needed to reduce the size of the prior credal set. See, in particular, Weichselberger (2001, §4.3) formulating two symmetry principles extending the principle of insufficient reason, and the closely related discussion of weak and strong invariance in Miranda and Cooman (2013, §3). For a critique on non-informative priors from an imprecise probability viewpoint see, e.g., Walley (1991, §5.5); their partition dependence is also discussed in the context of elicitation (see Smithson 2013, §3).

[10] As another possible modelling aim, in situations when data is revealed to the analyst sequentially in distinct batches, it might also be useful if the model is able to resonate unusual patterns or extreme differences between the batches. This actually effects to doubting the i.i.d. assumptions on which these models are founded. This could be useful in the area of *statistical surveillance* (see, e.g., Frisén 2011), where, e.g., the number of cases of a certain infectious disease is continuously monitored, with the aim to detect epidemic outbreaks in their early stages.

[11] The imprecise Beta-Binomial model from Walley (1991, §5.3.2) can be seen as a precursor to the IDM, covering the special case of two categories.

is obtained by choosing $\mathcal{Y}^{(0)}$ as the whole interior of the unit simplex Δ, with $n^{(0)}$ fixed,[12]

$$\mathcal{M}^{(0)} = \left\{ p(\theta \mid n^{(0)}, y^{(0)}) \,\middle|\, 0 < y_j^{(0)} < 1 \; \forall j, \sum_{j=1}^{k} y_j^{(0)} = 1 \right\} .$$

The prior credal set is thus determined by the choice of $n^{(0)}$. Walley (1996a) argues for choosing $n^{(0)} = 1$ or $n^{(0)} = 2$, where, in the latter case, inferences from the IDM encompass both frequentist and Bayesian results based on standard choices for noninformative priors. For any choice of $n^{(0)}$, this imprecise prior expresses a state of ignorance about θ, as, for all $j = 1, \ldots, k$,

$$\underline{\mathrm{E}}[\theta_j] = \underline{y}_j^{(0)} = \inf_{y_j^{(0)} \in \mathcal{Y}^{(0)}} y_j^{(0)} = 0 \,,$$
$$\overline{\mathrm{E}}[\theta_j] = \overline{y}_j^{(0)} = \sup_{y_j^{(0)} \in \mathcal{Y}^{(0)}} y_j^{(0)} = 1 \,,$$

and probabilities for events regarding θ_j are vacuous, i.e. $[\underline{\mathrm{P}}, \overline{\mathrm{P}}](\theta_j \in A) = (0, 1)$, for any $A \subset [0, 1]$.[13]

The posterior credal set $\mathcal{M}^{(n)}$ is then the set of all Dirichlet distributions with parameters $n^{(n)}$ and $y^{(n)}$ obtained by (1.6),

$$\mathcal{M}^{(n)} = \left\{ p(\theta \mid n^{(n)}, y^{(n)}) \;\middle|\; 0 < y_j^{(0)} < 1 \; \forall j, \sum_{j=1}^{k} y_j^{(0)} = 1 \right\} .$$

For any event A_J that the next observation belongs to a subset J of the categories, $J \subseteq \{1, \ldots, k\}$, the posterior lower and upper probabilities are given by

$$\underline{\mathrm{P}}(A_J) = \frac{n(A_J)}{n^{(0)} + n} \qquad\qquad \overline{\mathrm{P}}(A_J) = \frac{n^{(0)} + n(A_J)}{n^{(0)} + n} \,,$$

where $n(A_J) = \sum_J n_j$ is the number of observations from the category subset J.

The IDM is motivated by a number of inference principles put forward by Walley (1996a, §1), most notably the *representation invariance principle* (RIP, see Walley 1996a, §2.9): Inferences based on the IDM are invariant under different numbers of categories considered in the sample space.[14] The usefulness of the RIP has been controversially discussed (see,

[12] As mentioned before, our notation relates to Walley's (1996a) as $t_j \leftrightarrow y_j^{(0)}$, $s \leftrightarrow n^{(0)}$, $t_j^* \leftrightarrow y_j^{(n)}$.

[13] However, the IDM may give non-vacuous prior probabilities for some more elaborate events. An example (Walley 1996a, p. 14) is the event (A_J, A_K) that the next observation belongs to a category subset $J \subseteq \{1, \ldots, k\}$, and the observation following that belongs to a category subset K, where $J \cap K = \emptyset$ and $|n(A_J) - n(A_K)| < n^{(0)}$.

[14] In the example discussed in Walley (1996a), where colored marbles are drawn from a bag, it does not matter, e.g., for prior and posterior probabilities for "red" as the next draw, whether one considers the categorization {red, other} or {red, blue, other}.

e.g., the discussion to Walley 1996a), and alternative imprecise probability models that do not rely on it have been developed.[15]

Due to its tractability, the IDM has been employed in a number of applications. Walley (1996a) offers an application to data from medical studies; Bernard (2005) details an extension of the IDM to contingency table data that was briefly covered in Walley (1996a). In 2009, a special issue of the *International Journal of Approximate Reasoning* (Bernard 2009) was devoted to the IDM. Since its introduction, the IDM has found applications in, e.g., reliability analysis (e.g., Coolen 1997; Utkin and Kozine 2010; Utkin, Zatenko, and Coolen 2010; Li et al. 2011), or operations research (e.g., Utkin 2006); however, the IDM has had an especially strong impact in the area of artificial intelligence, namely in the construction of classification methods (including, e.g., pattern recognition) and in inference based on graphical models, see, e.g., Corani et al. (2013) and Antonucci, Campos, and Zaffalon (2013) and their references. These IDM-based methods, in turn, are used in a vast variety of tasks from all kinds of subjects, such as medicine (e.g., Zaffalon, Wesnes, and Petrini 2003), agriculture (e.g., Zaffalon 2005), or geology (e.g., Antonucci, Salvetti, and Zaffalon 2007). The IDM can be regarded as the most influential imprecise probability model so far.

In the IDM, satisfying near-ignorance for the prior and still having non-vacuous posterior probabilities is possible because the domain of the prior main parameter $y^{(0)}$ is bounded ($\mathcal{Y} = \text{int}(\Delta)$). For most conjugate priors to exponential family distributions, $\mathcal{Y}$ is not bounded, and thus, trying to reach prior ignorance in the same way as in the IDM, by taking $\mathcal{Y}^{(0)} = \mathcal{Y}$ for a fixed $n^{(0)}$, would lead to vacuous posterior probabilities.[16] Instead, as was shown by Benavoli and Zaffalon (2012), for conjugate priors to one-parameter exponential family distributions, one needs to vary $n^{(0)}$ in conjunction with $y^{(0)}$ to get both prior near-ignorance and non-vacuous posterior probabilities. In essence, the term $n^{(0)}y^{(0)}$ appearing in (1.6) must be bounded while letting $\mathcal{Y}^{(0)} = \mathcal{Y}$, which effects to a prior parameter set $\mathbb{\Pi}^{(0)}$ where the range of $n^{(0)}$ depends on $y^{(0)}$.

To summarize, imprecise probability methods allow for a much more adequate modeling of prior ignorance than non-informative priors, the traditional Bayesian approach to this problem, can deliver. Instead of the somehow awkward choice of a certain non-informativeness approach, to define an imprecise non-informative prior, the analyst just needs to specify one parameter (or two, as partly in Benavoli and Zaffalon (2012)) determining the learning speed of the model, namely $n^{(0)}$ for the IDM.

[15] See, e.g., Coolen and Augustin (2005; 2009) for an alternative based on the NPI approach (see Section 2.1.4.2). Important differences between this model and the IDM are also briefly discussed at the end of Augustin, Walter, and Coolen (2013, §6.1). Other alternatives to the IDM are discussed by Bickis (2009, see Section 3.2.2), and by Mangili and Benavoli (2013).

[16] From (1.6), it follows that for $y^{(0)} \to \infty$ we get $y^{(n)} \to \infty$ if $n^{(0)}$ is fixed.

3.1.4. Substantial Prior Information and Sensitivity to Prior-Data Conflict

Models intended specifically for use in situations with substantial prior information are presented in Walley (1991, footnote no. 10 in §1.1.4, and §7.8.3), Quaeghebeur and Cooman (2005), and Walter and Augustin (2009b);[17] the IDM can be modified by not taking $\mathcal{Y}^{(0)} = \mathrm{int}(\Delta)$, but a smaller set $\mathcal{Y}^{(0)}$ fitting the prior information, as was done in Section 1.3.6.1.

Generalising the IDM approach to conjugate priors for sample distributions of the canonical form (1.4), Quaeghebeur and Cooman (2005) proposed an imprecise prior $\mathcal{M}^{(0)}$ based on $\mathbb{\Pi}^{(0)} = \mathcal{Y}^{(0)} \times n^{(0)}$. E.g., in the Normal-Normal model as described in Section 1.2.3.4, one can take as imprecise prior all convex mixtures of normals with mean in $\mathcal{Y}^{(0)} = [\underline{y}^{(0)}, \overline{y}^{(0)}]$ and a fixed variance $\sigma_0^2/n^{(0)}$. $\mathcal{Y}^{(n)}$, the posterior set of expectations (or modes, or medians) of μ, is then bounded by

$$\underline{y}^{(n)} = \inf_{\mathbb{\Pi}^{(0)}} y^{(n)} = \inf_{\mathcal{Y}^{(0)}} \frac{n^{(0)} y^{(0)} + n\bar{x}}{n^{(0)} + n} = \frac{n^{(0)} \underline{y}^{(0)} + n\bar{x}}{n^{(0)} + n} \tag{3.2}$$

$$\overline{y}^{(n)} = \sup_{\mathbb{\Pi}^{(0)}} y^{(n)} = \sup_{\mathcal{Y}^{(0)}} \frac{n^{(0)} y^{(0)} + n\bar{x}}{n^{(0)} + n} = \frac{n^{(0)} \overline{y}^{(0)} + n\bar{x}}{n^{(0)} + n}\,. \tag{3.3}$$

The lower (upper) posterior expectation of μ is thus a weighted average of the lower (upper) prior expectation and the sample mean, with weights $n^{(0)}$ and n, respectively. As mentioned above, Quaeghebeur and de Cooman's (2005) model for Bernoulli or multinomial data leads to the IDM; because $\mathcal{Y}$, the domain of $y^{(0)}$, is not bounded in the general case, the model is normally used to express substantial prior information.[18]

More generally in case of a one-parameter exponential family, $\mathbb{\Pi}^{(0)}$ is fully described by the three real parameters $\underline{y}^{(0)}$, $\overline{y}^{(0)}$, and $n^{(0)}$, which are straightforward to elicit; furthermore, also $\mathbb{\Pi}^{(n)}$ is fully described by $\underline{y}^{(n)}$, $\overline{y}^{(n)}$, and $n^{(n)}$, and many inferences will be expressible in terms of these three parameters only. Models of this kind allow for a simple yet powerful imprecise inference calculus, where the amount of ambiguity in the prior information can be represented by the magnitude of the set $\mathcal{Y}^{(0)}$, with $n^{(0)}$ determining the learning speed.

The downside of this easily manageable model is that it is insensitive to prior-data conflict, as the imprecision for the main posterior parameter,

$$\Delta_y(\mathbb{\Pi}^{(n)}) = \overline{y}^{(n)} - \underline{y}^{(n)} = \frac{n^{(0)}(\overline{y}^{(0)} - \underline{y}^{(0)})}{n^{(0)} + n}\,, \tag{3.4}$$

does not depend on the sample $\boldsymbol{x}$. Imprecision is thus the same for any sample $\boldsymbol{x}$ of size n, whenever prior information about μ as encoded in $\mathcal{Y}^{(0)}$ is in accordance with data information $\tilde{\tau}(\boldsymbol{x})$ or not. The relation of $\Delta_y(\mathbb{\Pi}^{(n)})$ (which determines the precision of posterior

[17] See Section 3.3 for a more detailed coverage and examples.

[18] However, it could be used for near-ignorance prior situations in case of other sampling models where $\mathcal{Y}^{(0)}$ can encompass the whole domain without causing posterior vacuousness. This applies, e.g., to circular distributions like the von Mises distribution, where the mean direction angle μ has the domain $(-\pi, \pi]$, see Quaeghebeur (2009, §B.1.4) and Mardia and El-Atoum (1976).

inferences) to the inferential situation at hand is loosened, as possible conflict between prior and data is not reflected by increased imprecision. In that sense, the IDM with prior information and the model by Quaeghebeur and Cooman (2005) do not utilize the full expressive power of imprecise probability models, behaving similar to precise conjugate models by basically ignoring prior-data conflict.

To counter this unwanted behaviour, Walter and Augustin (2009b) suggested that imprecise priors to canonical sample distributions (1.4) should be based on parameter sets of the form

$$\amalg^{(0)} = [\underline{n}^{(0)}, \overline{n}^{(0)}] \times \mathcal{Y}^{(0)}\,, \tag{3.5}$$

as Walley (1991, §5.4.3) had already implemented for the Beta-Binomial model. Then, (3.2) and (3.3) become

$$\underline{y}^{(n)} = \begin{cases} \dfrac{\overline{n}^{(0)}\underline{y}^{(0)} + n\bar{x}}{\overline{n}^{(0)} + n} & \bar{x} \geq \underline{y}^{(0)} \\ \dfrac{\underline{n}^{(0)}\underline{y}^{(0)} + n\bar{x}}{\underline{n}^{(0)} + n} & \bar{x} < \underline{y}^{(0)} \end{cases}, \qquad \overline{y}^{(n)} = \begin{cases} \dfrac{\overline{n}^{(0)}\overline{y}^{(0)} + n\bar{x}}{\overline{n}^{(0)} + n} & \bar{x} \leq \overline{y}^{(0)} \\ \dfrac{\underline{n}^{(0)}\overline{y}^{(0)} + n\bar{x}}{\underline{n}^{(0)} + n} & \bar{x} > \overline{y}^{(0)} \end{cases}.$$

If $\underline{y}^{(0)} < \bar{x} < \overline{y}^{(0)}$, both $\underline{y}^{(n)}$ and $\overline{y}^{(n)}$ are calculated using $\overline{n}^{(0)}$; when $\bar{x}$ falls into $[\underline{y}^{(0)}, \overline{y}^{(0)}]$, the range of prior expectations for the mean, prior information gets maximal weight $\overline{n}^{(0)}$ in the update step (1.6), leading to the same results as for a model with fixed $n^{(0)} = \overline{n}^{(0)}$. If, however, $\bar{x} < \underline{y}^{(0)}$, then $\underline{y}^{(n)}$ is calculated using $\underline{n}^{(0)}$, giving less weight to the prior information that turned out to be in conflict with the data. Thus, as $\underline{y}^{(n)}$ is a weighted average of $\underline{y}^{(0)}$ and $\bar{x}$, with weights $n^{(0)}$ and n, respectively, $\underline{y}^{(n)}$ will be lower (nearer towards $\bar{x}$) as compared to an update using $\overline{n}^{(0)}$, resulting in increased imprecision $\Delta_y(\amalg^{(n)})$ compared to the situation with $\underline{y}^{(0)} < \bar{x} < \overline{y}^{(0)}$. In the same way, there is additional imprecision $\Delta_y(\amalg^{(n)})$ if $\bar{x} > \overline{y}^{(0)}$.[19]

Indeed, (3.4) can then be written as

$$\Delta_y(\amalg^{(n)}) = \frac{\overline{n}^{(0)}(\overline{y}^{(0)} - \underline{y}^{(0)})}{\overline{n}^{(0)} + n} + \inf_{y^{(0)} \in \mathcal{Y}^{(0)}} |\tilde{\tau}(\boldsymbol{x}) - y^{(0)}| \frac{n(\overline{n}^{(0)} - \underline{n}^{(0)})}{(\underline{n}^{(0)} + n)(\overline{n}^{(0)} + n)}\,,$$

such that we have the same $\Delta_y(\amalg^{(n)})$ as for a model with $\amalg^{(0)} = \mathcal{Y}^{(0)} \times \overline{n}^{(0)}$ when $\tilde{\tau}(\boldsymbol{x}) \in \mathcal{Y}^{(0)}$, whereas $\Delta_y(\amalg^{(n)})$ increases if $\tilde{\tau}(\boldsymbol{x}) \notin \mathcal{Y}^{(0)}$, the increase depending on the distance of $\tilde{\tau}(\boldsymbol{x})$ to $\mathcal{Y}^{(0)}$, as well as on $\underline{n}^{(0)}$, $\overline{n}^{(0)}$, and n. This model is described in more detail in Section 3.3, along with illustrative examples (see Figure 3.5 for the Dirichlet-Multinomial model, and Figure 3.4 for the Normal-Normal model).

Models with $\amalg^{(0)}$ as in (3.5), i.e., belonging to model type (c), are sensitive to prior-data conflict, where prior-data conflict is operationalised as $\tilde{\tau}(\boldsymbol{x}) \notin \mathcal{Y}^{(0)}$. There is no such direct

[19] In the above, $\bar{x}$ can be replaced by $\tilde{\tau}(\boldsymbol{x})$ to hold for canonical priors (1.5) in general. See Section 3.3.4 for an definition in general terms.

mechanism for a gain in precision when prior and data information coincide especially well.[20] However, $\mathcal{Y}^{(0)}$ could be chosen relatively small such that it mirrors this situation, considering as the neutral situation $\tilde{\tau}(\boldsymbol{x})$ being not too far away from $\mathcal{Y}^{(0)}$, and taking as prior-data conflict situations when $\tilde{\tau}(\boldsymbol{x})$ is in a greater distance to $\mathcal{Y}^{(0)}$.[21]

As mentioned in Section 3.1.1, page 59, the parameter set $\mathbb{\Pi}^{(n)}$ resulting from updating $\mathbb{\Pi}^{(0)}$ as in (3.5) by (1.6) is not a Cartesian product of $\mathcal{Y}^{(n)}$ and $\mathcal{N}^{(n)}$, i.e., a *rectangle* set in case of one-dimensional $y^{(0)}$, as was the case for $\mathbb{\Pi}^{(0)}$. It might therefore be necessary to minimize and maximize over $\mathbb{\Pi}^{(n)}$ itself if inferences depend on $y^{(n)}$ and $n^{(n)}$ simultaneously. If, e.g., $n^{(n)}y^{(n)}$ must be minimised to determine the posterior lower bound of a characteristic of interest, $\min_{\mathbb{\Pi}^{(n)}} n^{(n)}y^{(n)}$ may not be found at $\underline{y}^{(0)}$, i.e., $\min_{\mathbb{\Pi}^{(n)}} n^{(n)}y^{(n)} \neq \min_{\mathcal{N}^{(0)}} n^{(n)}\underline{y}^{(n)}$.

The model of type (b), where $\mathbb{\Pi}^{(0)} = \mathcal{N}^{(0)} \times y^{(0)}$, briefly mentioned only in Walley (1991, footnote no. 10 in §1.1.4, and §7.8.3), also leads to a more complex description of $\mathbb{\Pi}^{(n)}$ as compared to the models with $\mathbb{\Pi}^{(0)} = n^{(0)} \times \mathcal{Y}^{(0)}$ (type (a)).

In principle, $\mathbb{\Pi}^{(0)}$ could have any form fitting the prior information at hand (type (d)). On close inspection, a rectangular shape for $\mathbb{\Pi}^{(0)}$ may not be appropriate in many situations. One could, e.g., argue that the $y^{(0)}$ interval should be narrower at $\underline{n}^{(0)}$ than at $\overline{n}^{(0)}$, because we might be able to give quite a precise $y^{(0)}$ interval for a low prior strength $\underline{n}^{(0)}$, whereas for a high prior strength $\overline{n}^{(0)}$, we should be more cautious with our elicitation of $y^{(0)}$ and thus give a wider interval; interestingly, one could also argue the other way round based on similarly convincing arguments.[22] To fully specify $\mathbb{\Pi}^{(0)}$ along these lines, lower and upper bounds for $y^{(0)}$ must be given for all intermediate values of $n^{(0)}$ between $\underline{n}^{(0)}$ and $\overline{n}^{(0)}$, e.g., by some functional form $\underline{y}^{(0)}(n^{(0)})$ and $\overline{y}^{(0)}(n^{(0)})$. The choice of such general forms is not straightforward, as it may heavily influence the posterior inferences, and it may be very difficult to elicit as a whole. One such choice is discussed in Walter, Augustin, and Coolen (2011, §2.3, see Section 3.5.2.3) for the Binomial case, developed with the intention to create a smoother reaction to prior-data conflict than in the model with rectangle $\mathbb{\Pi}^{(0)}$.

In summary, there is a trade-off between easy description and handling of $\mathbb{\Pi}^{(0)}$ on one side, and modeling accuracy and fulfillment of desired properties on the other:

- The model in Quaeghebeur and Cooman (2005), which takes $\mathbb{\Pi}^{(0)} = n^{(0)} \times \mathcal{Y}^{(0)}$, is very easy to handle, as the posterior parameter set $\mathbb{\Pi}^{(n)} = n^{(n)} \times \mathcal{Y}^{(n)}$ has the same form as $\mathbb{\Pi}^{(0)}$, and it often suffices to consider the two elements $(n^{(0)}, \underline{y}^{(0)})$ and $(n^{(0)}, \overline{y}^{(0)})$ to find posterior bounds for inferences. It is, however, insensitive to prior-data conflict.

- The model by Walter and Augustin (2009b) is sensitive to prior-data conflict, but this advantage is payed for by a more complex description of $\mathbb{\Pi}^{(n)}$.

- More general set shapes $\mathbb{\Pi}^{(0)} \subset \mathbb{R}_{>0} \times \mathcal{Y}$ are possible, but may be difficult to elicit

[20] However, first ideas and some preliminary results for a set shape allowing this are given in Sections 4.3 and A.2, respectively.

[21] See also Section 3.5.4 for this idea.

[22] See, e.g., the rationale discussed at the beginning of Section 3.5.2.3.

and complex to handle.[23]

To conclude, when substantial prior information is available, that, however, is not permitting the identification of a single prior distribution, imprecise probability models allow for adequate modeling of partial information and prior-data conflict sensitivity, and will ultimately result in more reliable inferences.

Before we will study two choices for $\mathbb{M}^{(0)}$ in more detail in Sections 3.3 and 3.5 (Section 3.4 describes a software implementation of the model from Section 3.3), we will now discuss some other models based on sets of priors in Section 3.2.

[23] For an example see, as mentioned above, the work in Section 3.5, specifically in Section 3.5.2.3. First ideas and some preliminary results for another approach to a set shape allowing also more precise inferences when prior and data coincide especially well are given in Sections 4.3 and A.2, respectively.

3.2. Alternative Models Using Sets of Priors

In Section 3.1, we have discussed a specific class of models based on parametrically constructed sets of conjugate priors, described important properties of inferences based on this class in general terms, and illustrated the potential of generalised Bayesian inference methods for the situations of prior near-ignorance and substantial prior information. Now, we will discuss two alternative frameworks for inference based on sets of priors in Section 3.2.1, and consider a number of inference models that could serve as an alternative to the models discussed in Section 3.1. The two frameworks based on sets of priors in alternative to the model framework discussed in this thesis are *neighbourhood models* (Section 3.2.1.1) and the *density-ratio class* (Section 3.2.1.2).[24] Then, we will briefly describe a few concrete inference models, some of which are based on these two model frameworks, distinguishing between models that are based on conjugate priors (Section 3.2.2) and models that are not (Section 3.2.3).

3.2.1. Some Alternative Model Frameworks

3.2.1.1. Neighbourhood Models

An important class of models that make use of sets of priors are *neighborhood models*. These are typically considered in the robust Bayesian approach (see, e.g., Berger et al. 1994; Ríos Insua and Ruggeri 2000), where a certain prior distribution P_0 is singled out as a potential model for prior information, but, due to lack of confidence in this choice, a neigbourhood around P_0 is considered, consisting of distributions 'near' P_0. The rationale for this approach is to ensure robustness of the Bayesian analysis based on a single prior P_0 by checking that small deviations from P_0 do not lead to large deviations in posterior inferences. As mentioned in Section 2.1.3.1, imprecise probability offers a formal, not casuistic framework for such *Bayesian sensitivity analysis*; however, interpretation of the sets of priors, and the modelling intention is different, especially with respect to the inference situations we perceive as important modelling opportunities for generalised Bayesian inference.[25] We will thus touch only briefly on neighbourhood models, picking out two typical examples, although many different kinds of neighbourhood models are discussed in the literature (see, e.g., the surveys by Berger, Ríos Insua, and Ruggeri (2000) and Ruggeri, Ríos Insua, and Martín (2005)).

A typical example is the ε-contamination class (see, e.g., Berger et al. 1994, §4.3.2), which can be informally described as follows: In a (virtual) sample distribution, not all data are distributed according to P_0; instead, $100 \cdot \varepsilon\%$ of the data is distributed according to any distribution from a set $\mathcal{Q}$, and depending on the choice for $\mathcal{Q}$, a variety of ε-contamination

[24] By some authors, the density ratio class is considered a neighbourhood model where instead of one central distribution P_0 two distributions are considered (e.g., Pericchi and Walley 1991, §4.3). We think, however, that the density ratio class is better characterised as a separate model framework.

[25] These are (i) the possibility of modeling prior near-ignorance (see Sections 2.2.3.2 and 3.1.3), and (ii) prior-data conflict sensitivity in case of informative priors (see Sections 2.2.3.3 and 3.1.4).

classes can be defined.[26]

Another example for a neighbourhood model is the *odds-ratio model*. It tries to model approximate adherence to a central probability law with distribution P_0 by giving the following constraints for pairs of events A and B:

$$\frac{\mathrm{P}(A)}{\mathrm{P}(B)} \leq (1-\epsilon)\frac{\mathrm{P}_0(A)}{\mathrm{P}_0(B)}\,, \quad A, B \subseteq \Omega$$

The set of distributions P compatible with these restrictions forms then an odds-ratio model with parameter ϵ, and can be represented by a lower prevision $\underline{\mathrm{E}}_\epsilon$. When such a model is taken as an imprecise prior in Bayesian inference, the set of posteriors can again be expressed as an odds-ratio model (Destercke and Dubois 2013b, §7.2). Other neighbourhood models, like, e.g., variants of the ε-contamination class mentioned above, may instead not be closed under Bayesian updating.

3.2.1.2. The Density Ratio Class

The *density ratio class*, also known as *interval of measures* (DeRobertis and Hartigan 1981; Berger 1990), provides also an interesting model framework for Bayesian inference using sets of priors. Here, instead of generating the set of prior distributions by varying their parameters in a set (as in Section 3.1), the set of priors $\mathcal{M}$ is defined by bounding the probability density functions $p(\vartheta) \in \mathcal{M}$ via a lower bounding function $l(\vartheta)$ and an upper bounding function $u(\vartheta)$.[27]

A set of (prior) distributions on ϑ is defined by

$$\mathcal{M}_{l,u} = \{p(\vartheta) : \exists c \in \mathbb{R}_{>0} : l(\vartheta) \leq cp(\vartheta) \leq u(\vartheta)\}\,, \tag{3.6}$$

where $l(\vartheta)$ and $u(\vartheta)$ are bounded non-negative functions (i.e., non-normalised densities) for which $l(\vartheta) \leq u(\vartheta)\ \forall\ \vartheta$. $l(\vartheta)$ and $u(\vartheta)$ are often called *lower and upper density functions*, and only need to be known up to a multiplicative constant. If $l(\vartheta) > 0\ \forall\ \vartheta$, then (3.6) can also be written as

$$\mathcal{M}_{l,u} = \left\{p(\cdot) : \frac{p(\vartheta)}{p(\vartheta')} \leq \frac{u(\vartheta)}{l(\vartheta')}\ \forall\ \vartheta, \vartheta'\right\}\,,$$

hence the name 'density ration class'.

The density ratio class defines a certain type of credal sets; thus, as discussed in Section 2.1.2, it can also be expressed via an associated coherent lower prevision $\underline{\mathrm{E}}_{l,u}$. The

[26] For $\mathcal{Q}$ taken as 'all distributions', the ε-contamination class is also called 'linear-vacuous mixture' in the imprecise probability literature (e.g., Destercke and Dubois 2013b, §7.3), constituting an important special case of coherent lower previsions.

[27] A very accessible presentation of density-ratio classes with parametric bounding shapes $l(\vartheta)$ and $u(\vartheta)$, along with a method for elicitation from an expert providing quantiles (or quantile ranges) for a number of probability values, is given in Rinderknecht, Borsuk, and Reichert (2011). We discuss this model in Section 3.2.3.

density ratio class has a number of advantageous properties, especially as compared to many neighbourhood models (see, e.g., Rinderknecht, Borsuk, and Reichert 2011, §2.3); most importantly, it has the convenient property of invariance under Bayesian updating. The set of posteriors derived from $\mathcal{M}_{l,u}$ through the Generalised Bayes' Rule (i.e., by updating element by element, see Sections 2.1.2.5 and 2.1.3) can again be expressed as a density ratio class, with $l(\vartheta)$ and $u(\vartheta)$ updated according to Bayes' Rule (DeRobertis and Hartigan 1981).

Although this invariance property is advantageous, a consequence of it is that also the ratio $u(\vartheta)/l(\vartheta)$ is constant under updating, such that posterior imprecision, as measured by the magnitude of $\mathcal{M}_{l,u}$ is the same as prior imprecision, for any sample size n (see, e.g., Rinderknecht 2011, §4.2.2). This is in strong contrast to the behaviour of the models discussed in Section 3.1, where $\mathcal{M}^{(0)}$ converges to a one-element set for $n \to \infty$.

It is important to note here that even if the bounding functions $l(\vartheta)$ and $u(\vartheta)$ are defined parametrically (or, as in the approaches by Coolen (1993a; 1994) described below, even as conjugates), $\mathcal{M}_{l,u}$ does not contain only these parametric densities (or conjugate densities). Instead, $\mathcal{M}_{l,u}$ contains a variety of shapes, where (if $l(\vartheta)$ and $u(\vartheta)$ are not proportional) the tail behaviour can vary between that of $l(\vartheta)$ and $u(\vartheta)$.[28]

The density ratio class is thus similar to sets of priors discussed in the model framework from Section 3.1.1 where $\mathcal{M}^{(0)}$ is taken as all convex mixtures of parametric priors with parameters in $\mathbb{\Pi}^{(0)}$. Then, also $\mathcal{M}^{(0)}$ contains a variety of shapes that, however, do not allow for substantially different tail behaviour, but it also has the same property of invariance under Bayesian updating, because $\mathcal{M}^{(n)}$ can be constructed as the set of all convex mixtures of distributions with parameters in $\mathbb{\Pi}^{(n)}$.[29]

3.2.2. Some Approaches Based on Conjugate Priors

We discuss here some approaches based on conjugate priors. First, a density ratio class model using conjugate distributions for the bounding functions $l(\vartheta)$ and $u(\vartheta)$ by Coolen (1993a; 1994) is described in Section 3.2.2.1. Then, a brief summary of an approach by Bickis (2009) is given, who constructs a conjugate to the multinomial sample distribution where a specific correlation structure for the categories can be specified (Section 3.2.2.2). Afterwards, some results of the study by Pericchi and Walley (1991) are reported in Section 3.2.2.3, who compare models for inference on the mean of a normal distribution with known variance that are based on a number of conjugate and non-conjugate sets of priors.

3.2.2.1. The Model by Coolen (1993a; 1994)

Coolen (1993a) proposed an interesting model for sample distributions from the one-parameter exponential family using the density ratio class. In this model, the prior bounding functions $l(\vartheta)$ and $u(\vartheta)$ from (3.6) are linked through the relation $u(\vartheta) = c_0 \cdot l(\vartheta)$, where $c_0 \geq 1$ is independent of ϑ. Defining $l(\vartheta \mid \psi^{(0)})$ as (proportional to) the conjugate prior,

[28] See, e.g., Rinderknecht, Borsuk, and Reichert (2011, §3.2), or Pericchi and Walley (1991, §4.3).
[29] See Equation (3.1) in Section 3.1.1.

with hyperparmeter $\psi^{(0)}$, calculation of the posterior lower bounding function $l(\vartheta \mid \boldsymbol{x}, \psi^{(0)})$ is straight-forward, by[30]

$$l(\vartheta \mid \boldsymbol{x}, \psi^{(0)}) = l(\vartheta \mid \psi^{(0)}) f(\boldsymbol{x} \mid \vartheta) = l(\vartheta \mid \psi^{(n)}) \,,$$

and the posterior upper bounding function is defined as

$$u(\vartheta \mid \boldsymbol{x}, \psi^{(0)}) =: \frac{c_n}{c_0} u(\vartheta \mid \psi^{(0)}) f(\boldsymbol{x} \mid \vartheta) = c_n l(\vartheta \mid \psi^{(n)}) \,,$$

where c_n is introduced to allow the magnitude of $\mathcal{M}_{l,u}$ to decrease in dependence of the samples size n.[31] c_n must thus be chosen as a function decreasing in n, and Coolen (1993a) suggests a functional form with $c_n \to 1$ for $n \to \infty$, containing a parameter ξ that has a meaning similar to $n^{(0)}$, in the sense that if $n = \xi$, an information measure (suggested by Walley 1991, §5.3.7) is doubled.[32]

To use the model, one has to elicit the parameter(s) of the conjugate prior $\psi^{(0)}$, along with c_0 and ξ. The model is quite easy to handle, as the density ratio class provides relatively simple formulas for, e.g., lower and upper cumulative density functions, and lower and upper predictive densities. These formulas are easy to calculate if the involved integrals are easy to obtain, as is the case for a conjugate choice of $l(\vartheta)$. However, this model is insensitive to prior-data conflict, owing to the requirement of c_0 and c_n to be independent of ϑ and $\boldsymbol{x}$ (except for the sample size n); furthermore, as mentioned by Coolen (1993a, p. 341), there could be many other functional forms for c_n that were equally reasonable as the one suggested in the paper.

Coolen (1994) presents a further study of this model with a focus on predictive inferences, considering the special case of Bernoulli observations. There, $l(\theta)$ is proportional to a Beta(α, β), and $u(\theta)$ is defined as

$$u(\theta) = l(\theta) + c_0^* \cdot a(\theta) \,,$$

where $a(\theta)$ is proportional to a Beta(μ, λ), and c_0^* is again a factor that determines the prior imprecision.[33] For the case $\mu = \alpha$, $\beta = \lambda$, and $c_0^* = c_n^* = c$ for all n, formulas are derived that allow to study imprecision in posterior predicitive probabilities analytically. This model gives interesting insights into the dependence of imprecision on s, but the fact that posterior imprecision is the very same as prior imprecision if $s/n = \alpha/(\alpha + \beta)$ (i.e., data and prior assignments are perfectly in line) suggests that models with constant c_n^* should be avoided, and that instead c_n^* should decrease with n (Coolen 1994, p. 160). Furthermore, our conjecture is that the influence of s on posterior imprecision is mainly due to the restriction that the two bounding functions have to be proportional to Beta

[30] Remember that $l(\cdot)$ needs to be known up to a multiplicative constant only.

[31] Note that $c_n \neq c_0$ means that we actually do not update $\mathcal{M}$ according to the Generalised Bayes' Rule.

[32] Imprecision for models with fixed $n^{(0)}$ (Section 3.1.1, item a) is halfed when $n^{(0)} = n$, see Section 3.1.2, item I.

[33] The relation between c_0 and c_0^* is thus $c_0^* = c_0 - 1$.

densities. For the case of $s/n = \alpha/(\alpha+\beta)$, imprecision decreases for any $s \neq n/2$ (Coolen 1994, Table 1). This is very similar to the phenomenon that, in the Multinomial-Dirichlet model, the posterior variance of θ_j decreases for any n_j if $y_j^{(0)} = 1/2$ (see Section A.1.2.2).

Afterwards, Coolen (1994, §4) illustrates the general case with $\mu \neq \alpha$, $\beta \neq \lambda$, and c_n^* as in Coolen (1993a) with some numeric examples, showing a reasonable behaviour of the model. Unfortunately, there are no theoretical results for model behaviour of the general case (as we were able to give in Section 3.1.2). Elicitation of the parameters $(\alpha, \beta, \mu, \lambda, c_0^*, \xi)$ for an informative prior would possibly involve some elaborate pre-posterior procedures (Coolen 1994, p. 163), as the influence of different choices of $l(\theta)$ and $a(\theta)$ on $\mathcal{M}_{l,u}$ is not straightforward to ascertain.[34]

3.2.2.2. The Model by Bickis (2009)

Bickis (2009) suggests a multivariate logit-normal model as an alternative to the IDM for an application where neighbouring category probabilities θ_j are correlated. Instead of the conjugate Dirichlet prior for $\boldsymbol{\theta}$ (see Section 1.2.3.5), a multivariate normal prior is assumed for the (element-wise) logits of $\boldsymbol{\theta}$, i.e., $\log(\boldsymbol{\theta}/(1-\boldsymbol{\theta})) \sim \mathrm{N}_k(\mu\mathbf{1}, \sigma^2\mathbf{M})$, where $\mathbf{1} = (1, \ldots, 1)$, and $\mathbf{M}$ is a $(k \times k)$ matrix giving the correlation structure. The resulting posterior in terms of $\boldsymbol{\theta}$ is itself not tractable, but can be approximated by an exponential family that can be seen as the convex hull of the logit-normal and dirichlet families (Bickis 2009, p. 189), and that contains the Dirichlet distribution for the limit $\sigma^2 \to \infty$. Although a conjugate prior in the sense that there is a simple update step for the hyperparameters, posterior inferences for this prior are derived by simulation, as the posterior is not analytically tractable.

In the paper, a near-noninformative set of priors is constructed by means of a set of hyperparameters, used for an interesting application to estimate a discrete hazard function for which it is useful to mirror an autocorrelation structure in $\mathbf{M}$. By giving a numeric example, Bickis (2009) shows that posterior inferences based on this set of priors can be calculated using relatively simple algorithms. A number of interesting research questions present themselves for this model, regarding the potential for applications where no correlation structure must be assumed, possible differences to inferences based on the IDM, and the behaviour in case of prior-data conflict when an informative set of priors is chosen.

3.2.2.3. Some of the Models Studied by Pericchi and Walley (1991)

By studying credibility intervals for an unknown mean μ for samples from a normal distribution with known variance σ^2, Pericchi and Walley (1991) give a neat overview on a number of approaches based on sets of priors, a part of which are based on conjugate priors. As we do in Section 3.1, this overview makes the distinction of modelling near-noninformativeness versus models for substantial prior information, the latter of which are

[34] An alternative for eliciting α, β, μ, λ, and c_0^* could be to use the elicitation method by Rinderknecht, Borsuk, and Reichert (2011), who require the analyst to specify quantile intervals $[q_l^p, q_u^p]$ for some given probability levels $p \in (0,1)$, and fit $\mathcal{M}_{l,u}$ such that the corresponding cumulative density functions do not exceed the constraints given by the quantile intervals.

investigated with respect to prior-data conflict sensitivity.[35]

For modelling near-noninformativeness, Pericchi and Walley (1991, §3) present several 'translation-invariant' models, i.e., models whose posterior inferences do not depend on the location of $\bar{x}$. This includes a model where $\mathcal{M}$ consists of conjugate normal distributions $\mu \sim \mathrm{N}(\mu_0, \nu^2)$ for which $|\mu - \mu_0| \leq c\nu^2/2\sigma$, and where $\nu \to \infty$. It is considered inferior to a model where $\mathcal{M}$ consists of all double exponential distributions with a fixed variance but the mean varying in $\mathbb{R}$, showing the potential for models going beyond conjugate distributions.

Pericchi and Walley (1991, §4) denote all models that are considered for substantial prior information as 'neighbourhood models', and two variants of the ε-contamination class (see Section 3.2.1.1) are studied that, however, are not satisfactorily according to the desiderata the authors had established in §2. The model they advocate in §5 for this situation is instead a density ratio class that can be seen as a special case of the approach by Rinderknecht (2011, see below), where $l(\vartheta)$ is proportional to the conjugate normal distribution, and $u(\vartheta)$ is the improper uniform density $u(\vartheta) \propto 1$ (Pericchi and Walley 1991, §4.3). It shows a favourable behaviour similar to the model we argued for in Section 3.1.4 (which is discussed in more detail in Section 3.3 below).

3.2.3. Some Other Approaches Using Sets of Priors

In this section, we will discuss two approaches based on sets of priors that are, in contrast to the approaches discussed in Section 3.2.2, not based on conjugate distributions. In Section 3.2.3.1, we will give a short overview on Rinderknecht (2011, §4), discussing a density ratio class model based on not necessarily conjugate bounding functions. In Section 3.2.3.2, we will comment on an approach based on the discretisation of the parameter space (Whitcomb 2005).[36]

3.2.3.1. The Model by Rinderknecht (2011)

Rinderknecht (2011, §4) presents a model for inferences based on sets of priors of the form of a density-ratio class, where the bounding functions $l(\vartheta)$ and $u(\vartheta)$ are parametric, but not necessarily conjugate (we already mentioned his work with respect to elicitation in footnote 27, page 69). It is demonstrated that also marginals derived from a density ratio class with a multivariate parameter take again the form of density ratio classes, and can be calculated straight-forwardly. Also, it is shown that deduced quantities (like probabilistic predictions) derived from the set of posteriors can be framed as a density ratio class. However, while for quantities that are a deterministic function of the model parameters the resulting density ratio class is exact, the bounding functions for probabilistic predictions

[35] In fact, Pericchi and Walley (1991) was our inspiration to make this distinction in the first place.

[36] A further, very recent, contribution on Bayesian inference with sets of priors is Mangili and Benavoli (2013), modelling prior near-ignorance on the unit simplex by several sets of non-conjugate priors which provide an alternative to the IDM.

are too wide, thus providing a conservative approximation for the exact set of posterior predictive distributions (Rinderknecht 2011, §4.2.4).

The theoretical results are implemented and demonstrated via an example, a model for prediction of a certain type of river biomass that depends on six model parameters. Prior sets for these parameters were elicited from an expert, and combined with data from surveys of different streams. As the priors seem to be non-conjugate (there is no reference to the respective likelihoods in the publication), joint and marginal posteriors, along with biomass predicitions, were calculated using Markov Chain Monte Carlo (MCMC) techniques (see, e.g., Gilks 1998). Interestingly, Rinderknecht (2011, §4.3) shows that MCMC results for a single precise distribution can be used to approximate the set of posterior distributions, such that computational burden for the density ratio class is only marginally more extensive than in case of a precise prior.

Results are reasonably precise if prior information for only one of the six model parameters is modelled by a density ratio class, and for the other by precise probability distributions. If prior information for each of the six parameters is modelled by a density ratio class and combined independently, results are higly imprecise and of no practical use; the posterior marginals are even much more imprecise than their prior counterparts.[37] This undesirable phenomenon is called *dilation* (see Seidenfeld and Wasserman 1993). Here, it seems to result primarily from a kind of 'curse of dimension', but may also be due to the fact that, as noted above, the magnitude of $\mathcal{M}$ does not decrease with n for density ratio classes updated via the Generalised Bayes' Rule.

Although quite attractive for problems with a one-dimensional parameter (or where there is sufficient information to model further parameters by precise priors), the model is currently inadequate for higher-dimensional problems. This could be adressed through the development of multivariate elicitation procedures, eliciting also the dependence stucture for model parameters, or by replacing the independent combination of marginal prior sets by another strategy (as mentioned in Rinderknecht 2011, §5.2). A solution could be to factor a multivariate prior $p(\vartheta_1, \ldots, \vartheta_p)$ recursively by

$$\begin{aligned} p(\vartheta_1, \ldots, \vartheta_p) &= p(\vartheta_1 \mid \vartheta_2, \ldots, \vartheta_p) p(\vartheta_2, \ldots, \vartheta_p) \\ &= p(\vartheta_1 \mid \vartheta_2, \ldots, \vartheta_p) p(\vartheta_2 \mid \vartheta_3, \ldots, \vartheta_p) \cdots p(\vartheta_{p-1} \mid \vartheta_p) p(\vartheta_p)\,, \end{aligned}$$

where usually the dependencies can be reduced by a large degree through assumptions of conditional independence. This is the approach in probabilistic graphical models, where (in)dependencies between model parameters are visualised by a graph. Important guidance could be drawn from the vivid research conducted in the area of imprecise graphical models, also known as *imprecise Bayesian* or *credal networks*, especially with respect to independence concepts and efficient calculations (for a recent overview, see Antonucci, Campos, and Zaffalon 2013).

More importantly, however, we find the model unsatisfactory because it offers no clear mechanism to model posterior imprecision in dependence of sample size. The model is thus

[37] In the example, the marginal posterior lower bounding functions $l(\vartheta \mid \boldsymbol{x})$ are indistinguishable from zero if plotted in the same coordinate system as their respective the upper bounding functions $u(\vartheta \mid \boldsymbol{x})$, such that the posterior set of distributions contains also nearly uniform distributions.

inable to model prior near-ignorance, and does not exhibit a natural decrease in imprecision as information accumulates.[38]

Also, except for a very specific special case (Pericchi and Walley 1991, §4.3, as discussed above), for this model there are so far no studies or general results regarding the behaviour in case of prior-data conflict. This could be a promising topic of research, given the model described in Pericchi and Walley (1991, §4.3) shows favourable behaviour, while the model by Coolen (1993a) does not. Our conjecture is that the difference, or the ratio, of $u(\vartheta)$ to $l(\vartheta)$ must vary with ϑ to provide prior-data conflict sensitivity, where Pericchi and Walley (1991, §4.3) represents an extreme case, combining a $l(\vartheta)$ proportional to the light-tailed normal distribution with $u(\vartheta) \propto 1$ that gives the most heavy tails possible.

3.2.3.2. The Model by Whitcomb (2005)

Whitcomb (2005) studies imprecise Bayesian inference in discrete parameter spaces. The inference procedure is illustrated with three examples where substantial prior information is combined with relatively few observations, and its results are discussed with a focus on the relation of prior to posterior imprecision.

Contrary to the other studies mentioned so far, Whitcomb considers a discretised parameter space, i.e., there is only a finite number of values $\vartheta_1, \ldots, \vartheta_m$ the parameter ϑ can assume, and uses a linear programming formulation of the Generalised Bayes' Rule to derive posterior inferences from discrete prior distributions (Whitcomb 2005, §3). These discrete priors are derived from expert elicitations, given either as lower and upper bounds for $p(\vartheta_j), j = 1, \ldots, m$, or as lower and upper bounds for probability ratios $p(\vartheta_j)/p(\vartheta_{j'}), j \in \{1, \ldots, m\} \backslash j'$, where one of the parameter values $\theta_1, \ldots, \theta_m$ serves as pivot. From the latter, Whitcomb then derives lower and upper bounds for $p(\vartheta_j), j = 1, \ldots, m$, such that in both cases the set of prior probability functions is given by

$$\mathcal{M} = \{p(\vartheta) \mid \underline{p}_j \leq p(\vartheta_j) \leq \overline{p}_j \ \forall j = 1, \ldots, m\},$$

where $\underline{p}_j$ and $\overline{p}_j$ are the lower and upper (indirectly) elicited bounds for $p(\vartheta_j)$, respectively. As a summary measure for imprecision, Whitcomb chooses $\Delta = \sum_{j=1}^{m} \overline{p}_j - \underline{p}_j$.

We focus here on the example of a reliability analysis problem regarding the mean time to failure θ of a technical component (Whitcomb 2005, §4.1), studying the influence of several hypothetical data sets on posterior imprecision. The hypothetical data is assumed to come from a life testing experiment with observations following an exponential distribution, such that each data set can be represented by its exponential likelihood. Likelihoods based on a mean in agreement and in conflict with the elicited prior are considered,[39] with three

[38] However, a possibly attractive approach could be to combine ideas from Coolen (1993a) and Rinderknecht (2011) in a density ratio class model with non-Bayesian updating where c_n is defined such that posterior inferences also reflect prior-data conflict.

[39] However, as Krautenbacher (2011, §4.3) shows, the mean meant to be in agreement with the prior specifications is actually outside the interval $\left[\underline{\mathrm{E}}[\theta], \overline{\mathrm{E}}[\theta]\right]$ derived from the prior probability specifications. We discuss this work below.

sample sizes $n = 2, 6, 10$ in each group. For the resulting six posteriors, imprecision in the probability intervals for the discrete parameter values is compared to likewise imprecision in the prior.

In accordance to intuition, imprecision Δ decreases with the sample size for those likelihoods based on a mean in agreement with the elicited prior, the posterior probability intervals being more precise than the respective prior intervals. For the likelihoods based on a mean in contrast with the prior, the picture is different. Here, for $n = 2$ and $n = 6$, posterior imprecision is instead larger than prior imprecision, mirroring the conflict of information from prior and data. Δ is largest for $n = 6$, indicating that in this case prior and data seem to have a similar weight.[40] Only for the largest sample size $n = 10$, posterior imprecision is less than prior imprecision, indicating that the prior is now overwhelmed by the data. Strangely, posterior imprecision for this likelihood is even less than for the likelihood based on the same sample size and a mean in agreement with the prior. This seems somehow unintuitive, and could be due to specifics of the elicited prior probability intervals, or could be an artefact of the relatively low number of distinct parameter values $m = 7$.

Krautenbacher (2011, §4) compared the results from this example with a model of the framework from Section 3.1 (model type c, see also Section 3.3.4 below). To derive a prior parameter set $\mathbb{M}^{(0)}$ from the expert assessments given by Whitcomb (2005, Table I), he suggested and implemented an algorithm that determined $\mathbb{M}^{(0)}$ by searching over a parameter grid, starting from a precise distribution in accordance with the prior probability intervals. His results are generally similar to those of Whitcomb (2005, §4.1), but with some interesting differences in the details.

As the conjugate prior model is formulated in terms of $\lambda = 1/\theta$, Krautenbacher first calculates expectation intervals and (approximate) unions of highest density intervals for λ based on the discrete prior and posteriors of Whitcomb. These are then compared with expectation intervals and unions of highest density intervals for λ derived from the determined $\mathbb{M}^{(0)}$.[41]

In Whitcomb's model, imprecision as regarded through highest density and expectation intervals for λ is naturally somewhat different from the imprecision Δ based on θ. All posterior expectation intervals are shorter than the prior expectation interval; for the prior-data agreement case, expectation interval results are similar to those based on Δ, with $\left[\overline{\mathrm{E}}[\lambda] - \underline{\mathrm{E}}[\lambda]\right]$ decreasing according to sample size. For the prior-data conflict situations, however, similarities vanish: while Δ increased for $n = 2$ and $n = 6$, and dropped sharply for $n = 10$, $\left[\overline{\mathrm{E}}[\lambda] - \underline{\mathrm{E}}[\lambda]\right]$ is decreasing with n just like in the prior-data agreement case, but with now even shorter intervals. However, with regards to the skewness of Whitcombs's posteriors in terms of λ (depicted in Krautenbacher 2011, Abb. 24, p. 62), the expectation intervals are probably misleading here. Unions of highest density intervals for λ, which are considered as an alternative, can be determined only very coarsely, as the number

[40] In contrast to the models discribed in Section 3.1 where the parameter $n^{(0)}$ clearly communicates the weight of the prior as compared to the sample, the prior is here defined non-parametric and does not entail any parameters by which the weight of an elicited prior can be gauged.

[41] See also Example 3.3, page 86, for calculation of unions of highest density intervals in these models.

		Whitcomb (2005, §4.1)			Krautenbacher (2011, §4)		
	n	HD interval		Δ_{HD}	HD interval		Δ_{HD}
prior	0	0.00	2.00	2.00	0.05	1.01	0.96
pda	2	0.10	1.00	0.90	0.12	1.07	0.95
	6	0.20	1.00	0.80	0.22	1.09	0.87
	10	0.25	1.00	0.75	0.30	1.09	0.79
pdc	2	0.20	2.00	1.80	0.14	1.39	1.25
	6	0.33	2.00	1.67	0.34	2.01	1.67
	10	0.50	2.00	1.50	0.54	2.52	1.98

Table 3.1.: Highest density intervals for λ based on the discrete model (Whitcomb 2005, §4.1, left) and on the conjugate model (Krautenbacher 2011, §4, right), as given in Krautenbacher (2011, Tab. 2, Tab. 3). Δ_{HD} gives the length of the HD interval; 'pda' indicates posterior intervals for data in agreement with the prior, 'pdc' indicates posterior intervals for the situation of prior-data conflict.

of distinct parameter values in Whitcomb's model is very low. Nevertheless, results for these are more intuitive. Starting from length 2.00 of the prior HD interval, posterior HD intervals get shorter with growing sample size in both groups, and the intervals in case of prior-data conflict are always larger than their counterparts in the prior-data agreement case (see Table 3.1).

In the conjugate-based model, expectation intervals instead behave as expected. Krautenbacher derived the prior parameter set as $\mathbb{\Pi}^{(0)} = [2.39, 2.85] \times [2.91, 4.08]$; however, the set of priors $\mathcal{M}^{(0)}$ based on $\mathbb{\Pi}^{(0)}$ fit the constraints posed by Whitcomb's prior probability intervals not very well (see Krautenbacher 2011, Abb. 20); the conjugate Gamma distributions do not seem to capture all aspects of the prior information in this case. As is clear from the general properties described in Section 3.1.4, imprecision as measured by $\left[\overline{\mathrm{E}}[\lambda] - \underline{\mathrm{E}}[\lambda]\right]$ decreases with sample size, with slightly higher imprecision in the prior-data conflict case. Owing to the caveat above, we will not compare these with the expectation intervals from Whitcomb's model, and instead look more closely on the HD intervals from both models, as given in Table 3.1.

Although the prior HD interval of the discrete model has double the length of the conjugate model, posterior HD intervals are surprisingly similar in length and position for the prior-data agreement case. The conjugate-based posterior HD interval lengths in case of prior-data conflict are, in contrast to the discrete model, growing with n; from comparison with the expectation intervals that behave as expected, we think that this unintuitive result can be appropriated to peculiarities of the Gamma distribution.[42]

In summary, both models show more or less adequate results, with the discrete model

[42] While for the prior variance holds $\mathrm{Var}(\lambda) \in [0.028, 0.070]$, the variance for the posterior with $n = 10$ in case of prior-data conflict ranges in $[0.120, 0.263]$. The upper variance thus almost quadruples, leading to a very wide posterior HD interval.

showing some unintuitive behaviour with respect to the expectation intervals, and the conjugate model with respect to HD intervals. Although the conjugate model behaves, as ensured by the general results, 'right' in terms of expectation intervals, the picture can be different for other inferences, depending on the functional form of the conjugate distributions. The example here makes it clear that, in considering expectations only (as is often done in the literature on imprecise probability methods, being based on the notion of previsions), important effects on inference can be obscured. On the other hand, a prerequisite for a conjugate model to give meaningful posterior inferenes is that the parametric priors are indeed a good model for the prior information at hand. For this example, this seems not to be the case.

A model based on a discretised parameter space offers more flexibility, at the cost of higher computational complexity. For the approach by Whitcomb, it also difficult to discern the effects of a chosen prior in interplay with a parametric likelihood.[43] This could be tackled by considering more refined elicitation strategies involving 'pre-posterior' elements, in which the analyst is asked to indicate what she is willing to learn from hypothetical data (see Section 3.5.2.3, where such a strategy is used to develop a shape for $\amalg^{(0)}$).

The question whether a discrete, nonparameteric model or instead a parametric, often continuous, model should be used is widespread in statistical inference in general. In traditional statistics, absolutely continuous distributions are usually employed when inference using discrete distributions becomes too complex, typically approximating a nonparametric model with a parametric one.[44] Similarly, the algorithms to compute posterior credal sets and inferences for discrete models, often framed via the alternative model formulation as conditional lower previsions (see, e.g., Troffaes and Hable 2013), may easily become unfeasible for large m; the alternative is then to consider sets of continuous prior distributions like in the model framework from Section 3.1.

After having reviewed some models in alternative to the model framework from Section 3.1, we will now study some examples for models from this framework in more detail. Section 3.3 discusses models of type (c) (p. 59), and a software implementation is briefly described in Section 3.4. Section 3.5 then presents a model of type (d) (p. 59), along with a fundamentally different approach that combines (posterior) inferences from two different models.

[43] As mentioned above, there is no natural summary measure giving the weight of the information encoded in the prior in comparison in the data, as is given by $n^{(0)}$ in the conjugate models.

[44] As an example, consider the test for independence in contingency tables. Fisher's exact test, a nonparametric test using a permutation argument (thus resulting in a discrete distribution), can become difficult to calculate for large samples. An alternative is then the chi-squared test that is based on the continuous, one-parametric $\chi^2(df)$ distribution, which, for large samples, is a good approximation of the distribution of the χ^2 test statistic then used to determine the test decision.

3.3. Imprecision and Prior-Data Conflict in Generalised Bayesian Inference

This section reproduces the work "Imprecision and Prior-Data Conflict in Generalised Bayesian Inference", published as a peer-reviewed article in the *Journal of Statistical Theory and Practice*, and reprinted as a book chapter in *Imprecision in Statistical Theory and Practice* (Walter and Augustin 2009b). As such, it is reproduced here almost verbatim, except for some minor shortenings, especially in the Introduction (Section 3.3.1), and the addition of some comments and footnotes linking this work to other parts of this thesis. Furthermore, the notation was changed slightly towards the one introduced in Section 1.2.3.1 (most importantly, writing $n^{(n)}$ and $y^{(n)}$ for the canonical posterior parameters instead of $n^{(1)}$ and $y^{(1)}$, respectively), and citations were updated and changed to the style employed throughout this thesis.

Here, we first study the class of imprecise probability models of type (a) from the framework established in Section 3.1.1 (see p. 59). As mentioned there, this type of models includes the Imprecise Dirichlet Model (IDM, Walley 1996a) under prior information, and more generally the framework of Quaeghebeur and Cooman (2005) for imprecise inference in canonical exponential families. We demonstrate that such models, in their originally proposed form, prove to be insensitive to the extent of prior-data conflict. We then propose an extension, namely by employing prior parameter sets of type (c), that reestablishes the natural relationship between knowledge and imprecision: the higher the discrepancy between the observed sample and what was expected from prior knowledge, the higher the imprecision in the posterior, producing cautious inferences if, and only if, caution is needed. Our approach is illustrated by some examples and simulation results.

3.3.1. Introduction

As discussed in Sections 2.1 and 2.2, imprecise probability provides a powerful methodology to handle the multidimensional nature of uncertainty neglected by the traditional concept of probability. Specifically, imprecise probability models allow to explicitely reflect the amount of knowledge they stand for, and thus promise to offer a vivid tool for handling situations of prior-data conflict in (generalised) Bayesian inference.

The most common — closely related — mathematical tools in the theory of imprecise probability are non-additive set-functions, interval-valued probabilities and sets of classical probabilities (often called credal sets).[45] The width of the interval or the "magnitude" of the set are then seen as a measure for the imprecision in the probabilistic assignment, allowing to take into account ambiguity (non-stochastic uncertainty) in statistical inference and decision making, where in the tradition of Ellsberg's (1961) seminal experiments ambiguity has proven to be a constitutive element.[46]

[45] See Section 2.1.2 for a short exposition of these mathematical tools and their relations.
[46] See the distinction of *risk* and *ambiguity* in Section 2.2.2.

Many imprecise probability models are constructed by a generalisation of Bayesian approaches.[47] The Bayesian paradigm relies on the assumption that the complete knowledge on the parameter ϑ of a statistical model can be expressed by a probability distribution on the parameter space Θ. Learning from a sample observation $\boldsymbol{x}$ is then performed by updating the *prior distribution*, describing the knowledge before having seen the sample $\boldsymbol{x}$, to obtain the *posterior distribution* via Bayes' rule. The posterior distribution, often shortened to 'posterior', then is understood to subsume the complete knowledge on ϑ after having seen the sample, and therefore it underlies exclusively all inferences drawn from the data. If the prior $p(\vartheta)$ is a precise probability distribution, then the posterior $p(\vartheta \,|\, x)$ can be directly calculated via Bayes' rule[48]

$$p(\vartheta \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \vartheta) \cdot p(\vartheta)\,,$$

where $f(\boldsymbol{x} \mid \vartheta)$ denotes the sample model, also called *likelihood* in this context.

Imprecise probabilities allow to overcome the "dogma of precision" (Walley 1991, §5) underlying common statistical models, in particular with respect to the often rather arbitrary choice of the prior distribution.[49] The ambiguity in the prior specification, or positively formulated, the quality of prior knowledge, can be modelled by considering sets $\mathcal{M}$ of prior distributions. The most straightforward way to proceed is then to update this set $\mathcal{M}$ element by element via Bayes' rule to obtain the set $\mathcal{M}_{|\boldsymbol{x}}$ of posterior distributions. This updating procedure is understood as self-evident in the robust Bayesian framework (e.g., Ríos Insua and Ruggeri 2000), and can be moreover justified by deriving it from general coherence arguments in the theory developed by Walley (1991), where it is referred to as the *Generalised Bayes' Rule*.[50] Although there are some strong arguments for a plurality of learning rules,[51] we will strictly rely on this approach in this contribution.

Probably the most popular model along this line is the Imprecise Dirichlet Model (IDM) for handling categorical data, introduced by Walley (1996a)[52] (see in particular Bernard (2005) and Bernard (2009) for an overview on further developments). This model has been extended by Quaeghebeur and Cooman (2005) to generalised Bayesian inference from canonical exponential families, not only covering the omnipresent Normal and Multinomial models as described in the Examples below, starting with Examples 3.3 and 3.4, but also almost all other sample distributions relevant in a Statistician's everyday life, such as Poisson models, often used in ecology and insurance mathematics, or exponential models and gamma models common in reliability and survival analysis. In Walter, Augustin, and

[47] See the discussion of the Bayesian approach to statistical inference in Section 1.2.3.

[48] See Equation (1.3).

[49] See the description of lower previsions, Walley's central formulation of imprecise probability, in Section 2.1.2.1, and the motives from a Bayesian perspective in Section 2.2.3.

[50] See Sections 2.1.2.5 and 2.1.3.

[51] See, e.g., Coolen and Augustin (2009), Weichselberger (2007), Cattaneo (2007), Held, Kriegler, and Augustin (2008), Held (2007), Coolen and Augustin (2009), Augustin and Coolen (2004), and Augustin (2003). We discuss some of these approaches in Section 3.2, and the critical aspects of the Generalised Bayes' Rule discussed in others in Section 2.1.3.2 and again in Section 4.3.

[52] See Section 3.1.3.

Peters (2007), Quaeghebeur and de Cooman's model was further generalised to a class of models (called LUCK-models there and throughout this paper) which includes linear regression models.

While up to now the main focus in generalized Bayesian inference has been on robustness issues given perturbations of an ideal prior (e.g., Ríos Insua and Ruggeri 2000) or on models for prior near ignorance (as in the IDM), the present paper is devoted to the important problem of handling prior-data conflict. *Prior-data conflict*[53] is a generic term to name situations in which informative prior beliefs and trusted data[54] are conflicting. A formalisation of this concept for the models considered in this paper is given in Definition 3.3. If prior and likelihood are precise, then in updating via Bayes' rule, prior-data conflict is averaged out, and there is no way to see whether the posterior arose from a situation with or without substantial prior-data conflict.[55]

Generalised Bayesian models, in contrast, promise to solve this problem in an elegant way: With the magnitude of the set $\mathcal{M}_{|\boldsymbol{x}}$ mapping the posterior ambiguity, high prior-data conflict should, ceteris paribus, lead to a large $\mathcal{M}_{|\boldsymbol{x}}$ resulting in high imprecision in the posterior probabilities, while in the case of no prior-data conflict $\mathcal{M}_{|\boldsymbol{x}}$, and thus the imprecision, should be much smaller.

Although Walley (1991, §5.2.2) explicitly emphasizes this possibility to express prior-data conflict as one of the main motivations for the paradigmatic change from precise to imprecise probability, surprisingly little attention has been paid to this issue. The very rare exceptions include two short sections in Walley (1991, p. 6 and §5.4), and the papers by Pericchi and Walley (1991), Coolen (1994) and Whitcomb (2005). Moreover, even the powerful models mentioned above, including the IDM under prior information and Quaeghebeur and de Cooman's extension, are, in their originally proposed form, not able to take into account prior-data conflict, and therefore do not fully utilize the expressive power of imprecise probabilities.

By extending these models, and some of their generalizations in this paper, we overcome their serious deficiency with respect to prior-data conflict. We arrive at powerful inferences where the degree of prior-data conflict is transferred into a corresponding amount of imprecision in the posterior quantities, resulting in reliable inferences by being cautious whenever caution is needed, also providing a neat basis for decision making.[56] Our method can be shown to extend the basic ingredients used in Walley (1991) for the normal case and the IDM to the general class considered here.

The remainder of this section is structured as follows: In Section 3.3.2, we provide the formal background by distinguishing a wide class of classical, precise probability models where Bayesian inference has a particular form, being directly suitable to a generalisation to imprecise probabilities by varying the linearly updated main parameter (Section 3.3.3). We

[53] See also the discussion of prior-data conflict in Sections 2.2.3.3 and 3.1.4.

[54] Our development neglects here the additional problems of outliers. So we simply assume that the data are not spoilt by outliers, or that outliers have been removed in advance.

[55] See also the comment in Section 1.2.3.1 on this central averaging property evident from Equation (1.6).

[56] See Troffaes (2007) for a review of decision criteria based on imprecise probabilities.

illustrate these models and then show in Section 3.3.4 how these models can be extended to deal with prior-data conflict in a more sensible way, by additionally varying another parameter. Section 3.3.5 illustrates our procedure in two important special cases: inferences on the mean of a normal distribution and in the IDM.

3.3.2. Traditional Bayesian Inference and LUCK-models

In many cases, traditional Bayesian inference is based on so-called conjugate priors (related to a specific likelihood). These distributions have the convenient property that the posterior resulting from Bayes' Rule (1.3) belongs to the same class of parametric distributions as the prior. The posterior thus remains easily tractable, and updating can be described in terms of parameters only. For describing the imprecise probability model used later on more easily, we want to distinguish certain standard situations (called *models with 'Linearly Updated Conjugate prior Knowledge' (LUCK)* here) of Bayesian updating with classical (traditional, precise) probabilities, where prior and posterior fit nicely together in the sense that (i) they belong to the same class of parametric distributions, and, in addition, (ii) the updating of one parameter ($y^{(0)}$ below) of the prior is linear given a second parameter ($n^{(0)}$). More precisely, we return to the following definition, originally introduced in Walter, Augustin, and Peters (2007):[57]

Definition 3.1 (LUCK-models)**.** *Consider traditional Bayesian inference on a parameter ϑ based on a sample $\boldsymbol{x}$ as described in Section 1.2.3 (Equation (1.3)), and let the prior $p(\vartheta)$ be characterized by the (vectorial) parameter $\vartheta^{(0)}$. The pair $\big(p(\vartheta), p(\vartheta \mid \boldsymbol{x})\big)$ is said to constitute a* LUCK-model of size $n \in \mathbb{N}$ with respect to the likelihood $f(\boldsymbol{x} \mid \vartheta)$ in the natural parameter ψ with prior parameters $n^{(0)} \in \mathbb{R}_{>0}$ and $y^{(0)}$ and sample statistic $\tau(\boldsymbol{x})$ *iff $p(\vartheta)$ and $p(\vartheta \mid \boldsymbol{x})$ can be rewritten in the following way:*

$$p(\vartheta) \propto \exp\left\{n^{(0)}\left[\langle \psi, y^{(0)}\rangle - \mathbf{b}(\psi)\right]\right\} \tag{3.7}$$

and

$$p(\vartheta \mid \boldsymbol{x}) \propto \exp\left\{n^{(n)}\left[\langle \psi, y^{(n)}\rangle - \mathbf{b}(\psi)\right]\right\}, \tag{3.8}$$

with

$$y^{(n)} = \frac{n^{(0)}}{n^{(0)}+n} \cdot y^{(0)} + \frac{n}{n^{(0)}+n} \cdot \frac{\tau(\boldsymbol{x})}{n}, \qquad n^{(n)} = n^{(0)} + n, \tag{3.9}$$

where ϑ is transformed to ψ and $\mathbf{b}(\psi)$, and $\vartheta^{(0)}$ to $n^{(0)}$ and $y^{(0)}$.

[57] Compare to the model presented in Section 1.2.3.1. There, the likelihood $f(\boldsymbol{x} \mid \vartheta)$ was considered to have the functional form (1.4). Here, $f(\boldsymbol{x} \mid \vartheta)$ may have instead any form, given that the update step from prior to posterior distribution adheres to Equations (3.7) – (3.9). The model discussed in Section 1.2.3.1 is a special case, as Equations (3.7) and (3.9) are the same as Equations (1.5) and (1.6), respectively. This is again elaborated in Example 3.1 below.

$y^{(0)}$ and $y^{(n)}$ can be seen as the parameter describing the main characteristics of the prior and the posterior, respectively, and so later on, $y^{(0)}$ and $y^{(n)}$ will be called *main prior* and *main posterior parameter*. In the models considered here, $y^{(0)}$ can also be understood as a prior guess for the random quantity $\tilde{\tau}(\boldsymbol{x}) := \tau(\boldsymbol{x})/n$ summarizing the sample. According to the left part of (3.9) these two different sources of information are linearly combined to obtain the main posterior parameter:

$$y^{(n)} = \frac{n^{(0)}}{n^{(0)} + n} \cdot y^{(0)} + \frac{n}{n^{(0)} + n} \cdot \tilde{\tau}(\boldsymbol{x}) \,. \tag{3.10}$$

This relation also equips $n^{(0)}$ with a vivid interpretation as "prior strength" or as "pseudo-counts", reflecting the weight one gives to the prior with respect to the sample. So, $n^{(0)}$ can be interpreted as the size of an imaginary sample that corresponds to the trust on the prior information in the same way as the sample size of a real sample corresponds to the trust in conclusions based on such a real sample. As a preparation for the generalizations considered later, let us turn to some characteristic examples, also illustrating the interpretations of $y^{(0)}$ and $n^{(0)}$.

Example 3.1 (Bayesian Inference in Exponential Families)**.** *In the case of independently and identically distributed (i.i.d.) observations* $\boldsymbol{x} = (x_1, \ldots, x_n)$ *from* regular canonical exponential families *(Bernardo and Smith 2000, p. 202 and p. 272f), a general result (see, e.g., ibid., Proposition 5.4) is available on how to construct conjugate priors. A prior obtained by this method then constitutes a LUCK-model of size n (the sample size) with the sample statistic of the whole sample being the sum of statistics for each sample element (which can be concluded from the canonical form of the likelihood), so* $\tau(\boldsymbol{x}) = \sum_{i=1}^{n} \tau^*(x_i)$*, and* $\tilde{\tau}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \tau^*(x_i)$*.*[58] *This is effectively the framework of Bayesian inference with regular conjugate priors as presented in Section 1.2.3.1. To perceive the generality of this result, recall that many of the sample models most often used in practice form an exponential family, as shown earlier in this thesis for the Binomial distribution (Section 1.2.3.3), the Normal or Gaussian distribution (Section 1.2.3.4), and the Multinomial distribution (Section 1.2.3.5).*

Example 3.2 (Bayesian Inference in Linear Regression)**.** *As shown by Walter (2006), Walter, Augustin, and Peters (2007), and Walter (2007), the importance of LUCK-models is not limited to the i.i.d. case but also provides a formal superstructure containing in particular the practically important case of linear regression models, modelling the (linear) influence of certain variables (called covariates, confounders, regressors, stimulus or independent variables) on a certain outcome (also called response or dependent variable).*[59]

[58] Quaeghebeur and Cooman (2005) consider this special case of LUCK-models in their seminal work motivating the generalisations presented here.

[59] See Section A.1 for alternative models that adhere to the regular conjugate framework of Section 1.2.3.1.

3.3.3. Imprecise Priors for Inference in LUCK-models

Our definition of LUCK-models was inspired by the work of Quaeghebeur and Cooman (2005), who develop a general approach for inference with imprecise priors, which was proven in Walter, Augustin, and Peters (2007) to be generalisable to arbitrary LUCK-models. Quaeghebeur and de Cooman's seminal idea was that the seemingly strange parameterisation in terms of $y^{(0)}$ and $n^{(0)}$ in (3.7) and (3.8) is perfectly suitable to be generalised to credal sets of priors. The crucial point is that $y^{(0)}$, the main prior parameter[60] is updated *linearly*, when the prior strength $n^{(0)}$ is taken as fixed. This makes an easily tractable imprecise calculus possible: When sets of priors are defined via sets of main parameters $y^{(0)}$, and these sets of $y^{(0)}$ are defined by lower and upper bounds, the lower and upper bounds of the sets of posterior main parameters $y^{(n)}$ can be obtained directly from (3.9).

3.3.3.1. iLUCK-models

In constructing a conjugate prior to a given likelihood, Quaeghebeur and de Cooman strictly rely on the method described in Section 1.2.3.1 (Example 3.1), but their technique to construct an *imprecise* conjugate prior, i.e. a set of priors, by considering a set of $y^{(0)}$, does not depend on this derivation, but rather on the linearity of the updating of values of $y^{(0)}$ given $n^{(0)}$ when the set of posterior distributions is calculated. As the LUCK-models capture exactly this property, it is possible to construct *imprecise* conjugate priors for arbitrary LUCK-models according to Quaeghebeur and de Cooman's technique.[61] We give a general exposition of this model class, illustrate it by continuing the previous examples, and elaborate their serious deficiency with respect to the handling of prior-data conflict, which then will be overcome in Section 3.3.4.

Due to the linearity of the updating for fixed $n^{(0)}$, minimisation and maximisation problems on the set of posteriors can be reduced to minimisation and maximisation problems on the set of priors when the parameter $y^{(n)}$ (or a linear function of it) is the quantity of interest. This very same update procedure is used in the Imprecise Dirichlet Model (IDM), which is based on the Dirichlet-Multinomial model as presented in Section 1.2.3.5. Just as Walley required s to be fixed in the IDM, Quaeghebeur and de Cooman consider sets of $y^{(0)}$ but only a single value for the other prior parameter, denoted by $n^{(0)}$ here. As we will show in detail in Section 3.3.3.2, the resulting models necessarily ignore prior-data conflict. However, as a preparation for our generalisation presented in Section 3.3.4, we want to present the model with varying $y^{(0)}$ but fixed $n^{(0)}$, which will be called iLUCK-model (for *imprecise* LUCK-model), in more detail:[62]

[60] In the Normal-Normal model: the prior mean for μ; in the Dirichlet-Multinomial model: the vector of prior expected values for the category probabilities $\boldsymbol{\theta}$.

[61] See Walter (2006); Walter, Augustin, and Peters (2007); Walter (2007) for an application of this idea to obtain linear regression models with imprecise prior distributions.

[62] In the systematic of model types developed in Section 3.1.1, iLUCK-models correspond to models of type (a).

Definition 3.2 (iLUCK-models)**.** *Consider the situation of Definition 3.1, and a set of LUCK-models* $\big(p(\vartheta), p(\vartheta \mid \boldsymbol{x})\big)$ *(with respect to the likelihood* $f(\boldsymbol{x} \mid \vartheta)$ *in the natural parameter* ψ *with prior parameters* $n^{(0)} \in \mathbb{R}_{>0}$ *and* $y^{(0)}$ *and sample statistic* $\tau(\boldsymbol{x})$*), produced by* $y^{(0)}$ *varying in some set* $\mathcal{Y}^{(0)} \subset \mathcal{Y}$*, where the parameter space* $\mathcal{Y}$ *is taken as the convex hull (without the boundary) of the range of* $\tau(\boldsymbol{x})$*. Let furthermore the credal sets* $\mathcal{M}$ *and* $\mathcal{M}_{|\boldsymbol{x}}$ *consist of all convex mixtures obtained by this variation of* $p(\vartheta)$ *and* $p(\vartheta \mid \boldsymbol{x})$*. Then* $\big(\mathcal{M}, \mathcal{M}_{|\boldsymbol{x}}\big)$ *is called the corresponding* imprecise LUCK-model (iLUCK-model) *based on* $\mathcal{Y}^{(0)}$ *and* $n^{(0)}$*.*

Remark 3.1. *Note that if* $\mathcal{M}$ *is used as an imprecise prior, by construction,* $\mathcal{M}_{|\boldsymbol{x}}$ *is the corresponding imprecise posterior. Although the imprecise prior contains not only the parametric distributions, but also arbitrary convex mixtures of them, it is nevertheless easy to obtain the imprecise posterior: Since it is sufficient to update the extreme points and the updating process is linear in* $\mathcal{Y}^{(0)}$*, the imprecise posterior* $\mathcal{M}_{|\boldsymbol{x}}$ *is simply obtained as the set of all convex mixtures of posteriors* $p(\vartheta \mid \boldsymbol{x})$ *arising from* (3.8) *by varying* $y^{(n)}$ *in* $\mathcal{Y}^{(n)}$*, where*

$$\mathcal{Y}^{(n)} = \left\{ \frac{n^{(0)} y^{(0)} + \tau(\boldsymbol{x})}{n^{(0)} + n} \;\middle|\; y^{(0)} \in \mathcal{Y}^{(0)} \right\} = \frac{n^{(0)}}{n^{(0)} + n} \cdot \mathcal{Y}^{(0)} + \frac{n}{n^{(0)} + n} \cdot \tilde{\tau}(\boldsymbol{x})\,. \tag{3.11}$$

In generalisation of (3.10), $\mathcal{Y}^{(n)}$ can actually be seen as a shifted and rescaled version of $\mathcal{Y}^{(0)}$, which allows us to keep the vivid interpretation of $n^{(0)}$ as "prior strength" or as "pseudocounts", as it plays again the same role for the prior as n for the sample.[63]

Remark 3.2. *In iLUCK-models, the "magnitude" of* $\mathcal{Y}^{(0)}$ *and* $\mathcal{Y}^{(n)}$ *naturally reflects the imprecision in the prior and the posterior, respectively. Consequently, we will define*[64]*, with*

$$\underline{y}^{(i)} := \inf\big\{y^{(i)} \;\big|\; y^{(i)} \in \mathcal{Y}^{(i)}\big\} \;\text{ and }\; \overline{y}^{(i)} := \sup\big\{y^{(i)} \;\big|\; y^{(i)} \in \mathcal{Y}^{(i)}\big\},\; i = 0, n\,, \tag{3.12}$$

the main parameter prior imprecision and main parameter posterior imprecision $\mathrm{MPI}^{(0)}$ *and* $\mathrm{MPI}^{(n)}$ *by*

$$\mathrm{MPI}^{(i)} := \overline{y}^{(i)} - \underline{y}^{(i)}, \quad i = 0,\, n. \tag{3.13}$$

A natural tool to summarize basic properties of the updating process is to look at

$$\mathrm{PG} := \mathrm{MPI}^{(0)} - \mathrm{MPI}^{(n)}\,, \tag{3.14}$$

which is called main parameter precision gain *here.*

Taking the Normal-Normal and the Dirichlet-Multinomial model as concretisations of Example 3.1, inference with iLUCK-models is now illustrated.

[63] $\mathcal{Y}^{(0)}$ must be bounded, as for any $\overline{y}^{(0)} = \infty$, it holds that $\overline{y}^{(n)} = \infty$ as well. For the IDM, introducing explicit bounds is not necessary, as the parameter space $\mathcal{Y}$ itself is already bounded, being the unit simplex.

[64] If the main parameter is multidimensional (denoted by $\boldsymbol{y}^{(i)}, i = 0, n$), then, throughout the paper, the infimum, the supremum, and the measure $\mathrm{MPI}^{(i)}$ and related quantities, are to be understood as defined component by component. Natural choices for real-valued measures derived from vector-valued $\mathrm{MPI}^{(i)}$ would be to consider appropriate norms.

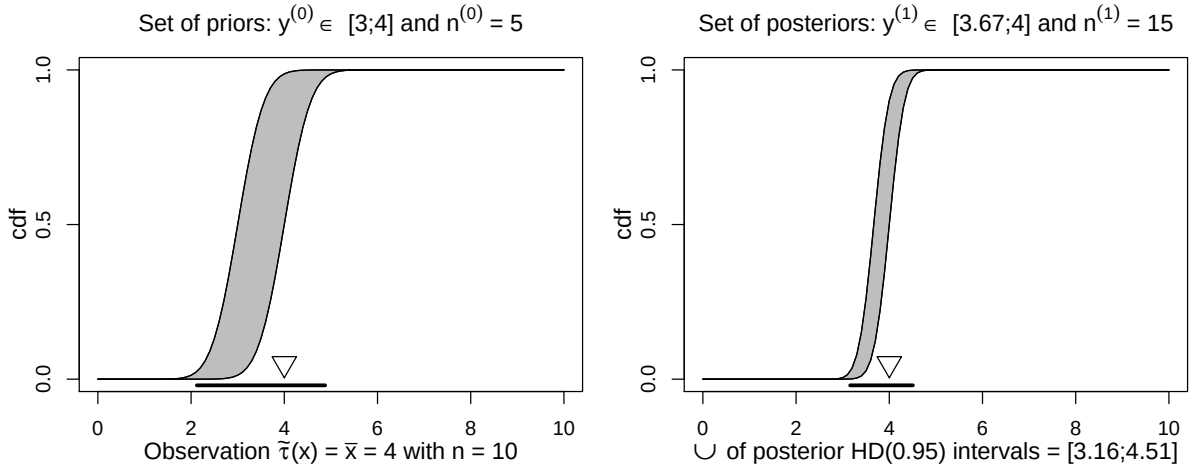

Figure 3.1.: Prior (left) and posterior (right) credal sets for a sample from $\mathrm{N}(\mu, 1)$ drawn as sets of normal cdfs. (Example 3.3 in the situation of no prior data conflict.)

Example 3.3 (Normal-Normal Model)**.** *In the Normal-Normal model as presented in Section 1.2.3.4, $y^{(0)}$ corresponds to the expected value for μ; the choice of $\mathcal{Y}^{(0)}$ in application should thus be easy. To simplify notation, we will assume here and later on that $\sigma_0^2 = 1$. Let us assume $\mathcal{Y}^{(0)} = [\underline{y}^{(0)}; \overline{y}^{(0)}] = [3; 4]$, and for fixing $n^{(0)}$, suppose further that we are not very certain about this prior range for μ, but still think it is quite a reasonable assumption, and so we base it on 5 pseudo observations by choosing $n^{(0)} = 5$, giving a value of the variance for the prior distribution on μ of $\frac{1}{5}$. Updating this prior with the i.i.d. sample $\boldsymbol{x} \in \mathbb{R}^n$ yields*

$$\underline{y}^{(n)} = \frac{n^{(0)}\underline{y}^{(0)} + \sum_{i=1}^{n} x_i}{n^{(0)} + n}, \qquad \overline{y}^{(n)} = \frac{n^{(0)}\overline{y}^{(0)} + \sum_{i=1}^{n} x_i}{n^{(0)} + n}, \qquad n^{(n)} = n^{(0)} + n\,.$$

To make this concrete, consider a sample of size $n = 10$ with $\tilde{\tau}(\boldsymbol{x}) = \bar{x} = 4$. Then $\mathcal{Y}^{(n)} = [\frac{55}{15}; \frac{60}{15}] \approx [3.67; 4]$, $\mathrm{MPI}^{(n)} = \frac{1}{3}$ and $n^{(n)} = 15$. The posterior credal set consists therefore of all convex combinations of normal distributions with means in $[3.67; 4]$ and variance $\frac{1}{15}$. Prior and posterior beliefs can be illustrated by the union of credal intervals calculated as highest density (HD) intervals[65] for all distributions in the corresponding credal set. As the normal distributions with mean $y^{(0)} \in \mathcal{Y}^{(0)}$ are the extreme points of the prior credal set, and the normal distributions are stochastically ordered with respect to the mean, the prior union is the interval from the lowest lower border of HD intervals (calculated from $\mathrm{N}(\underline{y}^{(0)}, \frac{1}{n^{(0)}})$) to the highest upper border (calculated from $\mathrm{N}(\overline{y}^{(0)}, \frac{1}{n^{(0)}})$). For a probability weight $\gamma = 0.95$, we get $[2.123; 4.877]$. The posterior union of HD intervals is $[3.161; 4.506]$ and, covering a much smaller range as a priori, shows the decreasing of

[65] See the concept of highest posterior density (HPD) intervals as mentioned in Section 1.2.3.3, which is used here also to illustrate the prior state of knowledge.

uncertainty obtained by the update step, also reflected in a main parameter precision gain of $\mathrm{PG} = \frac{2}{3}$. *This update step is illustrated in Figure 3.1, where the prior and posterior credal set are displayed by the normal cumulative distribution functions, the black lines indicating the functions defined by the vertices of* $\mathcal{Y}^{(0)}$ *and* $\mathcal{Y}^{(n)}$, *respectively. The observation* $\tilde{\tau}(\boldsymbol{x})$ *is marked by the point of the triangle in both graphs, and the prior and posterior union of HD intervals are marked by a thick line in the graph for the prior and posterior set, respectively.*

Example 3.4 (Dirichlet-Multinomial Model)**.** *An iLUCK-model based on the Dirichlet-Multinomial Model as discussed in Section 1.2.3.5 is, for* $\mathcal{Y}^{(0)} = \mathcal{Y}$, *equivalent to the imprecise Dirichlet model (IDM, see Section 3.1.3), and was considered in Section 1.3.6.1 for the common-cause failure application.*

In the usual applications of the IDM, the aim is to start with prior ignorance; this is modelled by choosing $\mathcal{Y}^{(0)}$ *as the unit simplex. For* $n^{(0)}$ *values of 1 or 2 are suggested. Here, we must rely on the interpretation of* $n^{(0)}$ *as prior strength, as there is no interpretation in terms of other parameters as in Example 1a. Considering prior knowledge for a three-category multinomial model suggesting that extreme values for* θ_1 *and* θ_2 *are implausible, one could choose* $\mathcal{Y}^{(0)} = \{y_1^{(0)} \in [0.2; 0.8] \times y_2^{(0)} \in [0.2; 0.8] \times y_3^{(0)} \in [0; 0.6]\}$, *where the upper bound for* $y_3^{(0)}$ *is a result of the unit simplex constraint* $\sum_{j=1}^{k} y_j^{(0)} = 1$. *In addition, we choose again* $n^{(0)} = 5$ *as in Example 3.3. Considering a sample of size* 5, *where* 3 *observations are of category 1, and* 2 *of category 2, we get* $n^{(n)} = 10$ *and the ranges* $y_1^{(n)} \in [0.4; 0.7]$, $y_2^{(n)} \in [0.3; 0.6]$, *and* $y_3^{(n)} \in [0; 0.3]$ *for the posterior class probabilities. In analogy to Example 3.3, this update step is illustrated with the left and center graph of Figure 3.2, where prior and posterior credal sets are represented by cutouts from a plane in the three-dimensional parameter space. Each point in the plane cutout for the prior set on the left graph represents a certain combination of* $y_1^{(0)}$, $y_2^{(0)}$, *and* $y_3^{(0)}$ *by the magnitude of coordinates. The same applies for the posterior set depicted in the center graph. Some additional lines were drawn to make locating the cutouts in space more easy.*

Remark 3.3. *Inference in* iLUCK-models *has the following important properties, where the first three items in essence generalize results that have been discussed in the literature for the IDM, where, as already said above,* $n^{(0)}$ *usually is denoted by* s.[66]

i) The larger $n^{(0)}$ *relative to* n, *the more weight is placed on the prior knowledge expressed by* $\mathcal{Y}^{(0)}$, *resulting in wider posterior expectation intervals and larger* $\mathrm{MPI}^{(n)}$.

ii) For growing sample sizes n, *the set* $\mathcal{Y}^{(n)}$ *will converge towards a one-element set, as the weight of the 'imprecise'* $\mathcal{Y}^{(0)}$ *decreases with respect to the 'precise' sample* $\tilde{\tau}(\boldsymbol{x})$, *ultimately resulting in an (almost) precise posterior with* $\mathrm{MPI}^{(n)} = 0$ *just as in classical methods.*

[66] See the discussion of inference properties in Section 3.1.2.

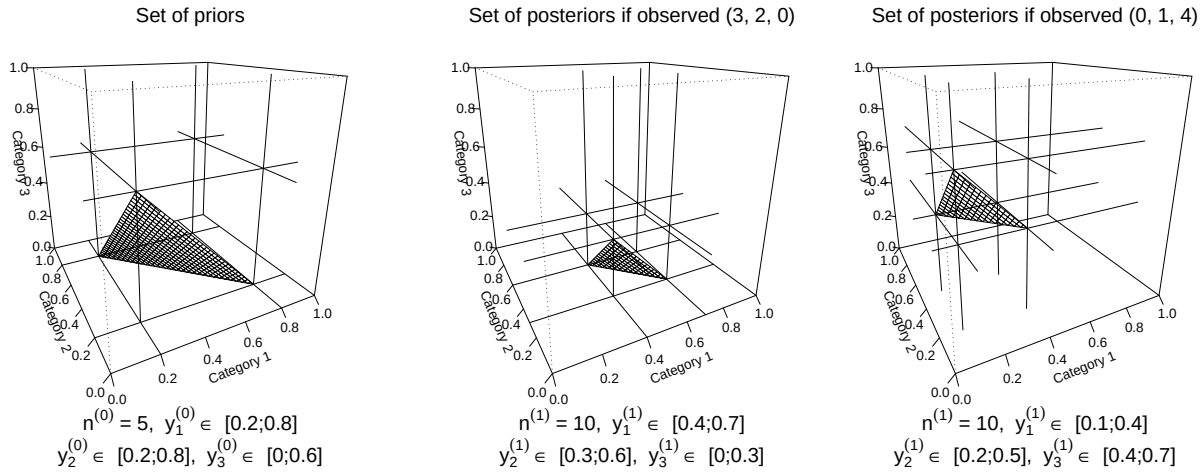

Figure 3.2.: Prior (left) and posterior (center, right) credal sets for samples from $\mathrm{M}(\boldsymbol{\theta})$ in accordance with (center) and, as studied in Section 3.3.3.2, contrary to (right) prior beliefs. Note that both posterior sets have the same shape and size and differ only in location, in contrast to the ones depicted in Figure 3.5.

iii) *In particular, for $n^{(0)} = n$, the width of the posterior expectation interval is half the width of the prior interval, i.e., $\mathrm{MPI}^{(n)} = \frac{1}{2}\mathrm{MPI}^{(0)}$. This property is easily derived from (3.15) below, and provides another vivid interpretation of $n^{(0)}$.*

iv) *Ceteris paribus, a smaller choice of $\mathcal{Y}^{(0)}$ will result in a smaller $\mathcal{Y}^{(n)}$, leading to more precise inference statements as opposed to the choice of a larger $\mathcal{Y}^{(0)}$.*[67]

v) *For the main posterior parameter imprecision, we obtain:*

$$\mathrm{MPI}^{(n)} = \frac{n^{(0)}\left(\overline{y}^{(0)} - \underline{y}^{(0)}\right)}{n^{(0)} + n}. \tag{3.15}$$

While items i) – iv) demonstrate the intuitively appealing behavior of iLUCK-models, we are seriously concerned with the fact that $\mathrm{MPI}^{(n)}$ is independent of $\tilde{\tau}(\boldsymbol{x})$, and, as studied in more detail in the next subsection, therefore insensitive to prior-data conflict.

3.3.3.2. iLUCK-models and Prior-Data Conflict

In the linear setting of iLUCK-models, the generic concept of prior-data conflict can be formalized by considering the distance of the observed quantity $\tilde{\tau}(\boldsymbol{x})$ to its nearest prior guess $y^{(0)} \in \mathcal{Y}^{(0)}$:

[67] This item has not been widely considered in the context of the IDM, which typically is used to describe inference from a state of prior ignorance.

Definition 3.3 ((Degree of) Prior-Data Conflict)**.** *For iLUCK-models, the degree of prior-data conflict can be defined as*

$$\Delta\left(\tilde{\tau}(\boldsymbol{x}); \underline{y}^{(0)}, \overline{y}^{(0)}\right) := \inf\left\{\left|\tilde{\tau}(\boldsymbol{x}) - y^{(0)}\right| : \underline{y}^{(0)} \leq y^{(0)} \leq \overline{y}^{(0)}\right\} . \tag{3.16}$$

Consequently, if $\Delta\left(\tilde{\tau}(\boldsymbol{x}); \underline{y}^{(0)}, \overline{y}^{(0)}\right) > 0$*, we have an instance of prior-data conflict.*[68]

This Definition can be illustrated by, e.g., Example 3.3, where $\tilde{\tau}(\boldsymbol{x}) = \bar{x}$ is the sample mean. A sample mean outside $\mathcal{Y}^{(0)}$, the a priori assumed interval of means for the normal distribution on μ, is an instance of prior-data conflict. If a sample mean of 8 was observed in the numerical example discussed above, where $[\underline{y}^{(0)}; \overline{y}^{(0)}] = [3; 4]$ had been assumed, we would obtain $\Delta(\,) = 4$, formalizing our intuition that prior-data conflict is at hand.

As we argued in Sections 2.2.3.3 and 3.3.1, imprecise probability models that allow to take prior information into account should lead to more imprecision if prior-data conflict is present than in situations where it is not. It is easy to see that iLUCK-models do not fulfill this property because the main parameter posterior imprecision in (3.15) does not depend on the sample statistic $\tilde{\tau}(\boldsymbol{x})$. Thus, for any sample of size n, an iLUCK-model leads to the same main parameter posterior precision gain whether the sample supports the prior assumptions modelled in $\mathcal{Y}^{(0)}$ or it confronts them. As a consequence of the Bayesian paradigm that all inference is only allowed to depend on the posterior, this holds also for all derived quantities like HD intervals. To make this concrete, let us continue Examples 3.3 and 3.4.

Example 3.5 (Normal-Normal Model, continued)**.** *Assume, in the situation considered in Section 3.3.3.1, that the sample had led to* $\tilde{\tau}(\boldsymbol{x}) = 8$*, suggesting the mean* μ *to be nearer to* 8 *than to the range* $[3; 4]$ *assumed for it before having seen the sample. Then* $\mathcal{Y}^{(n)} = [\frac{95}{15}; \frac{100}{15}] \approx [6.33; 6.67]$ *and again* $n^{(n)} = 15$*. The posterior range of expected values for* μ *has now moved towards the value suggested by the sample, but we have the same posterior main parameter imprecision* $\frac{1}{3}$ *as in the case with* $\tilde{\tau}(\boldsymbol{x}) = 4$*, which was not conflicting with the prior assumptions. Intuitively, this fact gets maybe more perplexing when considering the union of posterior HD intervals: For* $\tilde{\tau}(\boldsymbol{x}) = 4$*, it is* $[3.161; 4.506]$*, covering a length of* 1.345*; for* $\tilde{\tau}(\boldsymbol{x}) = 8$*, it is* $[5.827; 7.172]$*, covering the very same length, and therefore – although being completely surprised by the outcome of the sample – we would not be more cautious. These disturbing results are illustrated in Figure 3.3.*

Example 3.6 (Dirichlet-Multinomial Model, continued)**.** *Here, the same phenomenon occurs: Having observed categories 1 to 3 now* 0*,* 1*, and* 4 *times, respectively,* $\mathcal{Y}^{(n)}$ *covers the same area, despite of the clear conflict of these observations with the prior knowledge expressed in* $\mathcal{Y}^{(0)}$*, and has only moved in location, as can be seen in the right graph of Figure 3.2.*

[68] Instead of $\Delta(\,) > 0$, one could also consider some threshold $\Delta(\,) > \varepsilon > 0$ as a criterion for prior-data conflict, making Definition 3.3 also reasonable for LUCK-models. However, with respect to Remark 3.5 below, we prefer $\varepsilon = 0$.

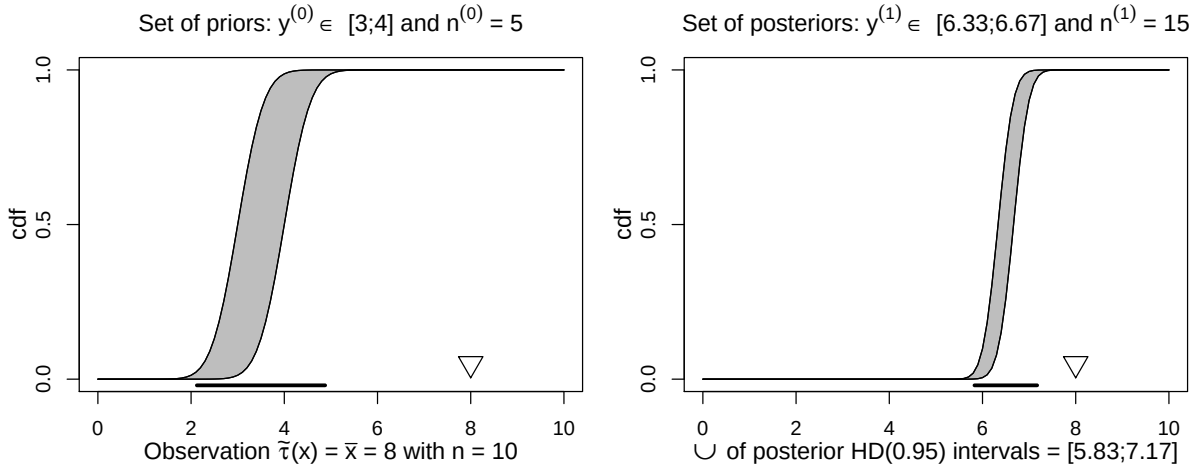

Figure 3.3.: In iLUCK-models, an observation contrary to prior beliefs leads to the same amount of imprecision as an observation in accordance with prior beliefs (see Figure 3.1). Generalized iLUCK-models overcome this deficiency, as visualized in Figure 3.4.

3.3.4. Improved Imprecise Priors for Inference in LUCK-models

There do exist some imprecise probability models that react on prior-data conflict in reasonable ways. The approaches by Pericchi and Walley (1991) and Coolen (1994) are restricted to specific one-parameter situations with a different type of models for the prior. Whitcomb (2005), on the other hand, by partially relying on models from the exponential family, considers closely related classes of underlying distributions without using their specific structure. The important hint for properly handling prior-data conflict in LUCK-models is again obtained from Walley (1991). In §5.4, he discusses prior-data conflict in the imprecise Beta-Binomial model,[69] which is a special case of the Imprecise Dirichlet Model recalled in Example 3.4. His successful idea is to vary the hyperparameter s in addition. Additionally, he also briefly mentions the normal case underlying Example 3.3 (Walley 1991, §1.1.5 (k)), where he suggests to use an imprecise prior with an imprecise variance. In the light of the general framework of LUCK-models, both exemplary solutions can be subsumed under the idea to use an imprecise prior strength, thus weighting prior and sample information in (3.11) in a more flexible way, i.e. to vary additionally the prior strength parameter $n^{(0)}$ in some set $\mathcal{N}^{(0)}$. Indeed this way to proceed will turn out to be successful. We start by describing the powerful generalisation we want to propose in more detail and then discuss some of its basic properties.[70]

[69] See Section 1.2.3.3.

[70] In the systematic of model types developed in Section 3.1.1, generalised iLUCK-models correspond to models of type (c).

Definition 3.4 (generalised iLUCK-models)**.** *Consider the situation of Definitions 3.1 and 3.2, and a set of LUCK-models* $\big(p(\vartheta), p(\vartheta \mid \boldsymbol{x})\big)$ *that is produced by* $y^{(0)}$ *varying in some set* $\mathcal{Y}^{(0)} \subset \mathcal{Y}$ *and, in addition,* $n^{(0)}$ *varying in a set* $\mathcal{N}^{(0)} \subset \mathbb{R}_{>0}$*. Let furthermore again the credal sets* $\mathcal{M}$ *and* $\mathcal{M}_{|\boldsymbol{x}}$ *consist of all convex mixtures obtained from this variation of* $p(\vartheta)$ *and* $p(\vartheta \mid \boldsymbol{x})$*. Then* $\big(\mathcal{M}, \mathcal{M}_{|\boldsymbol{x}}\big)$ *is called the corresponding* generalised iLUCK-model *based on* $\mathcal{Y}^{(0)}$ *and* $\mathcal{N}^{(0)}$*.*

Remark 3.4. *Again, by construction, when* $\mathcal{M}$ *is used as an imprecise prior,* $\mathcal{M}_{|\boldsymbol{x}}$ *is the corresponding posterior. It is obtained as the set of all convex mixtures of distributions defined by the parameter set*

$$\left\{ \big(n^{(n)}, y^{(n)}\big) \;\middle|\; n^{(n)} = n^{(0)} + n,\ y^{(n)} = \frac{n^{(0)} y^{(0)} + \tau(\boldsymbol{x})}{n^{(0)} + n},\ n^{(0)} \in \mathcal{N}^{(0)},\ y^{(0)} \in \mathcal{Y}^{(0)} \right\}.$$

Note that the set of posterior parameters is not a cartesian product, i.e., rectangular. However, extreme values for the main posterior parameter $y^{(n)}$ *are still easy to derive:*[71]

$$\underline{y}^{(n)} = \begin{cases} \frac{\overline{n}^{(0)} \underline{y}^{(0)} + \tau(\boldsymbol{x})}{\overline{n}^{(0)} + n} & \text{if} \quad \tilde{\tau}(\boldsymbol{x}) \geq \underline{y}^{(0)} \\ \frac{\underline{n}^{(0)} \underline{y}^{(0)} + \tau(\boldsymbol{x})}{\underline{n}^{(0)} + n} & \text{if} \quad \tilde{\tau}(\boldsymbol{x}) < \underline{y}^{(0)} \quad \Longleftrightarrow \quad \textit{prior-data conflict} \end{cases} \tag{3.17}$$

$$\overline{y}^{(n)} = \begin{cases} \frac{\overline{n}^{(0)} \overline{y}^{(0)} + \tau(\boldsymbol{x})}{\overline{n}^{(0)} + n} & \text{if} \quad \tilde{\tau}(\boldsymbol{x}) \leq \overline{y}^{(0)} \\ \frac{\underline{n}^{(0)} \overline{y}^{(0)} + \tau(\boldsymbol{x})}{\underline{n}^{(0)} + n} & \text{if} \quad \tilde{\tau}(\boldsymbol{x}) > \overline{y}^{(0)} \quad \Longleftrightarrow \quad \textit{prior-data conflict.} \end{cases} \tag{3.18}$$

For minimising and maximising $y^{(n)}$, the value $\underline{n}^{(0)}$ is used only in the situation of prior-data conflict, that is, if $\tilde{\tau}(\boldsymbol{x}) \notin \mathcal{Y}^{(0)}$. When no prior-data conflict occurs, the extreme values are attained for $\overline{n}^{(0)}$. Therefore, the fixed value of $n^{(0)}$ in iLUCK-models in the spirit of Quaeghebeur and Cooman (2005) can be seen as the upper border of an implicit set $\mathcal{N}^{(0)}$, and considering only a fixed $n^{(0)}$ means that prior-data conflict is neglected.

On the other hand, if observations are such that $\tilde{\tau}(\boldsymbol{x}) \in \big[\underline{y}^{(0)}; \overline{y}^{(0)}\big]$, then no prior-data conflict is present, and inference in generalised iLUCK-models leads to very similar results as inference in iLUCK-models.[72]

[71] As $\frac{d}{dy^{(0)}} y^{(n)} = \frac{n^{(0)}}{n^{(0)}+n} \geq 0 \ \forall n, n^{(0)}, \tau(\boldsymbol{x})$, it holds that $y^{(n)}$ is growing in $y^{(0)}$ regardless of the value of $n^{(0)}$, and thus, for obtaining $\overline{y}^{(n)}$, we must insert $\overline{y}^{(0)}$, and for obtaining $\underline{y}^{(n)}$, we must insert $\underline{y}^{(0)}$ in (3.9), just as in iLUCK-models, where $n^{(0)}$ is fixed. Then, with $\frac{d}{dn^{(0)}} y^{(n)} = \frac{y^{(0)} n - \tau(\boldsymbol{x})}{(n^{(0)}+n)^2}$, we see that $y^{(n)}$ is growing in $n^{(0)}$ if $y^{(0)} > \tilde{\tau}(\boldsymbol{x})$ and decreasing in $n^{(0)}$ if $y^{(0)} < \tilde{\tau}(\boldsymbol{x})$, leading to Equations (3.17) and (3.18).

[72] From (3.17) and (3.18) it gets clear why varying $n^{(0)}$ in addition does change the update step in the desired way. When $\tilde{\tau}(\boldsymbol{x}) < \underline{y}^{(0)}$, then $\overline{n}^{(0)}$ is still used to calculate $\overline{y}^{(n)}$, but $\underline{n}^{(0)}$ to calculate $\underline{y}^{(n)}$. The use of $\underline{n}^{(0)}$ instead of $\overline{n}^{(0)}$ results in a lower value for $\underline{y}^{(0)}$, as then more weight is given to $\tilde{\tau}(\boldsymbol{x})$ with respect to $\underline{y}^{(0)}$. (Equation (3.11) makes this most visible.) The same type of reasoning applies for the case $\tilde{\tau}(\boldsymbol{x}) > \overline{y}^{(0)}$, where $\overline{n}^{(0)}$ is still used to calculate $\underline{y}^{(n)}$, but $\underline{n}^{(0)}$ to calculate $\overline{y}^{(n)}$.

Remark 3.5. *By construction, the attractive properties of inferences by iLUCK-models formulated in i) to iv) of Remark 3.3 are still satisfied in our extended model. Now also in addition prior-data conflict is handled in a convincing way: Considering the relationship between the main parameter posterior imprecision defined as in* (3.13) *and the degree of prior-data conflict as defined in* (3.16),[73] *the very same results as derived by Walley (1991, p. 224) for the two-category Dirichlet-Multinomial model hold in general for any generalised iLUCK-model:*

$$\mathrm{MPI}^{(n)} = \frac{\overline{n}^{(0)}(\overline{y}^{(0)} - \underline{y}^{(0)})}{\overline{n}^{(0)} + n} + \Delta\left(\tilde{\tau}(\boldsymbol{x}); \underline{y}^{(0)}, \overline{y}^{(0)}\right) \frac{n(\overline{n}^{(0)} - \underline{n}^{(0)})}{(\overline{n}^{(0)} + n)(\underline{n}^{(0)} + n)}.$$

It is only through $\Delta(\,)$ *that* $\mathrm{MPI}^{(n)}$ *is depending on the actual shape of observation* $\boldsymbol{x}$ *besides its size* n*, and increasing it if prior-data conflict occurs. When no prior-data conflict is present,* $\Delta\left(\tilde{\tau}(\boldsymbol{x}); \underline{y}^{(0)}, \overline{y}^{(0)}\right) = 0$*, and we have the same amount of posterior imprecision in substantial parameters as in* iLUCK-models*, given by* (3.15).

Consequently, Walley's (1991, §5.4, footnote 3) observation that the factor to $\Delta(\,)$ *gets maximal if* $n = \left(\underline{n}^{(0)}\overline{n}^{(0)}\right)^{\frac{1}{2}}$ *remains valid. Then, the main parameter posterior imprecision is maximal for a given degree of conflict, implicitly telling that then the weight of the prior and sample must be the same, as more weight on any of them compared to the other (preferring one source of information to the other) would lead to a less wide* $\mathcal{Y}^{(n)}$*. This fact gives additional orientation for choosing* $\mathcal{N}^{(0)}$*, by considering the global strength of prior knowledge, being equivalent to* $\underline{n}^{(0)} := \left(\underline{n}^{(0)}\overline{n}^{(0)}\right)^{\frac{1}{2}}$.

3.3.5. Illustration of the Generalised iLUCK-model

The theoretical considerations in Section 3.3.4 are now illustrated by means of Examples 3.3 – 3.6, and some simulated data for larger sample sizes.

Example 3.7 (Normal-Normal model, continued)**.** *For the case of the Normal-Normal Model (Examples 3.3 and 3.5), the behavior of an appropriate generalised iLUCK-model is shown for the situations previously modeled with an iLUCK-model as depicted in Figures 3.1 and 3.3. In Figure 3.4, the top row shows the updating in absence of prior-data conflict, whereas the lower row displays the update step in presence of prior-data conflict. Again, the vertices of the credal set, the set of normal distribution functions, is represented by the shaded area, and the lines indicate the distributions that are obtained by updating the four extreme distributions from* $\mathcal{N}^{(0)} \times \mathcal{Y}^{(0)}$.

Contrary to the iLUCK-model, also a range of variances is considered in the prior, allowing for reasonable posterior inference, as can be seen in the right hand graphs: when the prior model is consistent with the observation $\tilde{\tau}(\boldsymbol{x}) = \bar{x}$*, a similar union of posterior HD intervals is obtained as for the model displayed in Figure 3.1; when instead prior*

[73] Remark 3.2 and Definition 3.3 can directly be applied also to generalised iLUCK-models.

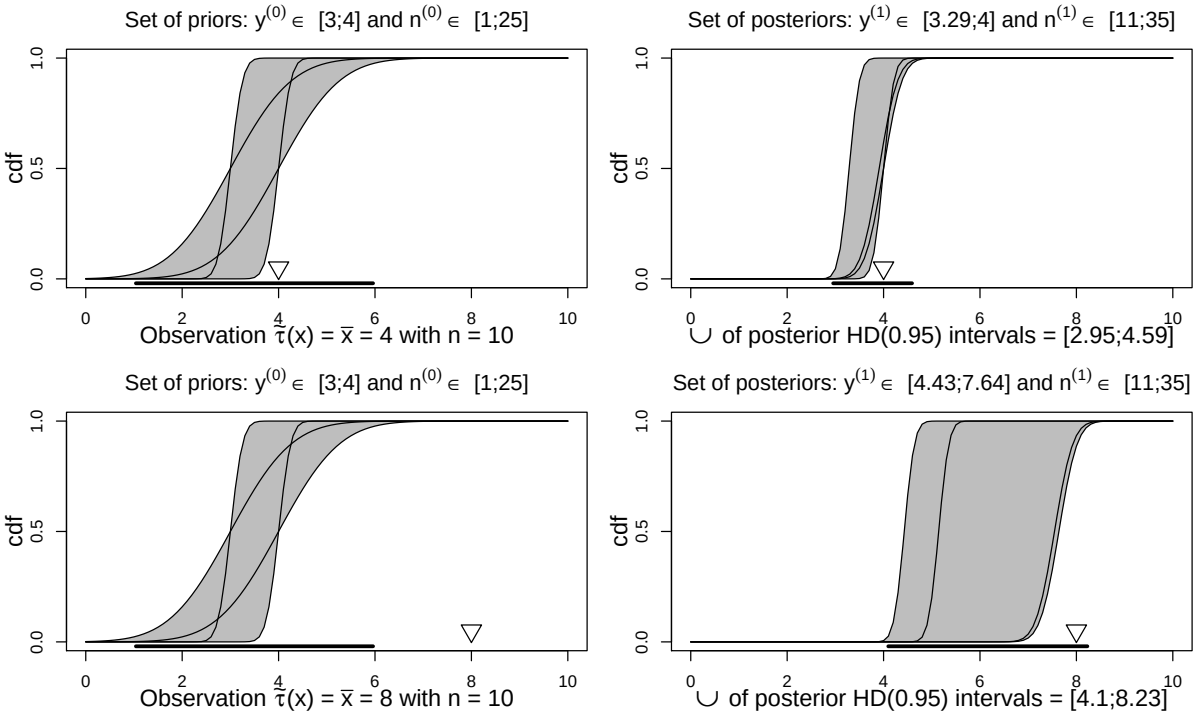

Figure 3.4.: Prior (left) and posterior (right) credal sets for samples from $\mathrm{N}(\mu, 1)$ in accordance with (upper) and contrary to (lower) prior beliefs. With generalised iLUCK-models, the posterior set in the latter case is significantly larger than in the former, leading to more cautious inference as desired.

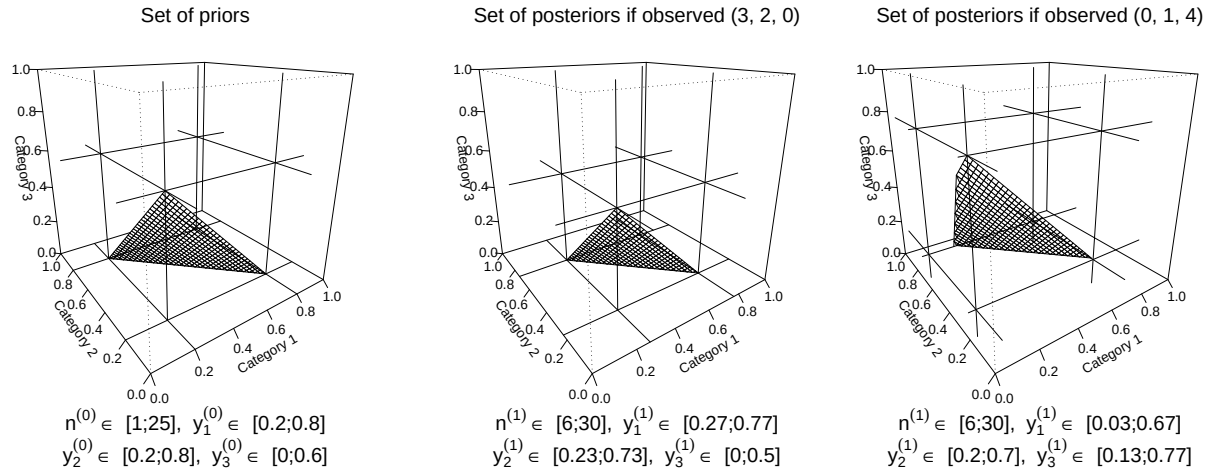

Figure 3.5.: Prior (left) and posterior (center, right) credal sets for samples from $\mathrm{M}(\boldsymbol{\theta})$ in accordance with (center) and contrary to (right) prior beliefs. As in Figure 3.4, with generalised iLUCK-models the sets differ in size according to the degree of prior-data conflict induced by the observations.

assumptions and data are conflicting, the posterior credal set reflects our uncertainty on which to trust, being substantially larger than the prior credal set. Therefore, the union of posterior HD intervals as indicated by the thick line in the lower right graph is not only wider than in the absence of prior-data conflict as seen in the top right graph, but not much shorter than its prior counterpart, which is $[1.040; 5.960]$. *In order to give comparable results,* $\mathcal{N}^{(0)}$ *was chosen to give the same global prior strength as the iLUCK-model for Example 3.3 by fixing* $\underline{n}^{(0)} = 5$ *and a minimal prior strength* $\underline{n}^{(0)} = 1$, *resulting in* $\overline{n}^{(0)} = 25$.

Example 3.8 (Dirichlet-Multinomial model, continued)**.** *For the Dirichlet-Multinomial model (Examples 3.4 and 3.6), the behavior of the generalised iLUCK-model is visualized in Figure 3.5. Here, the same set of main parameters as in Figure 3.2 is updated, leading to plane cutouts (symbolising posterior parameter sets) that differ not only in location as before, but also in size (and shape) for the two cases. For the center graph, prior assumptions on the category probabilities are in accordance with the observations 3, 2 and 0 for category one, two, and three, respectively, and thus the posterior plane cutout is a subset of the prior plane. In contrast, when observations 0, 1, and 4 are made, being in conflict with the prior assumptions for category one and three, the resulting posterior parameter intervals for those two categories do actually get wider, as can be seen in the right graph, and thus, inference drawn in this case is more cautious.*

Example 3.9 (Larger sample sizes)**.** *Sufficiently precise inference can be drawn nevertheless when the sample size gets much larger with respect to the prior strength, giving*

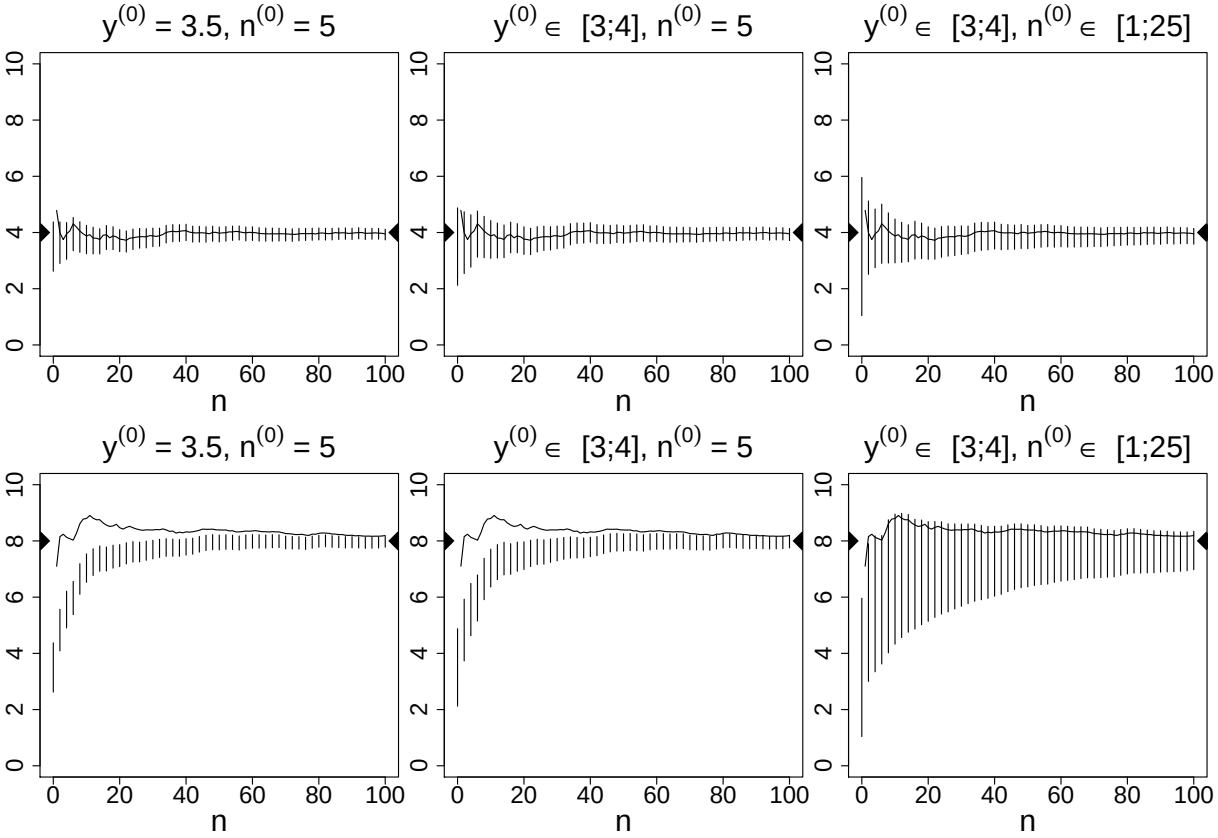

Figure 3.6.: (Unions of) 95%-posterior HD intervals (vertical lines) for observations in accordance (upper row, drawn from $\mathrm{N}(4,1)$) and in conflict (lower row, drawn from $\mathrm{N}(8,1)$) with prior assumptions based on a single prior (left), an iLUCK-model (center) and a generalised iLUCK-model (right) in dependence on the sample size n. The development of the sample mean is indicated by the wiggly line. Through the averaging, the HD intervals in the lower left and center graph are not enlarged but only shifted and thus do not cover the sample mean.

data more weight than the prior guess. When no prior-data conflict occurs, the imprecision obtained with generalised iLUCK-models is not substantially larger than obtained with iLUCK-models or even with a classical, precise prior. These three model classes are compared by means of the Normal-Normal model (Examples 3.3, 3.7 and 3.7) in Figure 3.6, where the first column gives posterior HD (further denoted by HPD) intervals for a precise prior, the second column unions of HPD intervals for the iLUCK-model, and the third column unions of HPD intervals for the generalised iLUCK-model, in all three cases drawn as vertical black lines for sample sizes $n = 2, 4, 6, \ldots, 100$.

In the upper row, these prior models are updated successively with observations drawn from a $\mathrm{N}(4, 1)$ *being in accordance with the prior assumptions, whereas for the lower row, observations in conflict with the prior assumptions were simulated by observations drawn from a* $\mathrm{N}(8, 1)$. *In the upper row, the (unions of) intervals tend to the sample mean indicated by the wiggly line more or less uniformly for all three models. Naturally, the most precise prior gives the most precise posterior inferences, but the HPD interval lengths do not differ excessively between the model classes, especially when the sample size n approaches* 100.

The lower row demonstrates the deficiency that the classical and the iLUCK-model unfortunately share. For them, observations confronting the prior assumptions lead only to an adjustment in location of the intervals through the averaging, but not in their length, giving a false certainty in posterior inference. Here, the generalised iLUCK-model is more truthful to the character of the situation, giving very wide unions of HPD intervals for low sample sizes covering both the a priori assumed range and the sample mean. With growing sample size, more confidence is given to the data, and thus, the unions of HPD intervals tend the more to the sample mean the larger the sample gets.

3.3.6. Concluding Remarks

In this work, we considered generalised Bayesian inference in a wide class of imprecise probability models. We first demonstrated that, in their originally proposed form, these models do not react to prior-data conflict and so do not utilise the full expressive power of imprecise probabilities. We extended these models such that the natural relationship between knowledge and imprecision is reestablished: the higher the discrepancy between prior assumptions and sample observations, the more cautious the posterior inference. We compare the previous modelling and our extension in two running examples covering two widely used situations, inference from a scaled normal and from a multinomial distribution, corroborating that our extension shows promising behavior.

Further research will start with a more detailed study of the behavior of the proposed model, also taking into consideration that it may have some problems with dealing with either minor or very extreme degrees of conflict.[74] Then careful attention should also be paid to other approaches for generalised Bayesian inference in exponential families (Coolen 1993a; Boratyńska 1997) under prior data conflict, as well as to a thorough comparison of

[74] See the study in Section 3.5.

the inference developed here with the discretisation-based models considered by Whitcomb (2005), and with the approaches of Pericchi and Walley (1991) and Coolen (1994), who use different prior classes.[75]

Far beyond these further developments, it should well be remembered that this study consciously confined the whole argumentation to a certain Bayesian setting: we studied how far one can go *if* one relies strictly on the Generalised Bayes' Rule, transferring sets of priors to sets of posteriors element by element via Bayes' Rule (1.3). Alternative learning rules, including the approaches enumerated in footnote 51, p. 80,[76] could prove very important as, to use one of the referees' felicitous words, "Bayesian methods cannot allow for surprise, the same is true for robustified or generalised Bayesian methods, although these may hide this shortcoming better. One could argue, therefore, that they cannot be used to represent learning."

[75] A discussion of the approaches by Coolen (1993a; 1994), Whitcomb (2005), and Pericchi and Walley (1991) is given in Section 3.2; The approach by Boratyńska (1997) was mentioned in Section 3.1.1, where we concluded that her approach belongs to models of type (a) that are insensitive to prior-data conflict.

[76] See also the critique regarding the Generalised Bayes' Rule in Section 2.1.3.2.

3.4. The `luck` Package

This secion gives a short overview on a software implementation of the models discussed in Section 3.3, the add-on package `luck` (Walter and Krautenbacher 2013) for the statistical programming environment **R**. Author and maintainer of this package is Gero Walter, with Norbert Krautenbacher as a second author, who contributed code for an application to exponentially distributed data.

Introduction. In recent years, the statistical programming environment **R** (R Core Team 2013) is used more and more frequently for every-day statistical analyses in academia and inside corporations. It has also become a de facto standard among statisticians for software implementations of novel methods, which can be easily distributed by means of so-called *packages* via the online repository 'cran' (Comprehensive **R** Archive Network, `http://www.cran.R-project.org`). The present implementation is programmed in a class system of **R** that reflects the hierarchy between the unified description in terms of LUCK-models[77] and the concrete application of this framework by, e.g., considering inference on the mean of scaled normal distribution using sets of conjugate priors (as described in Example 3.3). This hierarchical structure makes extensions of the package in order to enable inference in a wide class of sample distributions very easy.

Object-oriented Programming. Such hierarchical structures can be implemented using the object-oriented paradigm for programming. The central tools in this paradigm are *classes* and *methods*. *Classes* are used to represent general concepts that should be modelled in software. Such a concept is structured through a class by defining the traits that examples for the general concept should have. Classes thus provide the blueprint for *objects*, by defining a number of *attributes* or *slots* these objects have. Given concrete values for the slots of the class, a concrete object can be created according to the blueprint provided by the class definition. These concrete realisations of a class are called *instances*. *Methods* then provide functions to manipulate such instances, the class description guaranteeing a standardised input for the functions. However, the most prominent feature of the object-oriented paradigm is that hierarchies between concepts can be modelled by *inheritance*. More specific concepts can be modelled as special cases of a general concept, and the blueprint for such specialised objects are called *subclasses*, which *inherit* the traits of the more general class. Moreover, also methods can be specialised to reflect the class hierarchies, by giving more specific output for subclasses. The relations between classes are usually depicted in so-called *UML graphs*. Figure 3.7 gives an example for such a graph.

The Basic Classes. The package `luck` uses the S4 class system of **R**, providing the basic framework for the definition, display and updating of sets of priors based on LUCK-models

[77] As described in footnote 5, page 59, LUCK-models generalise the framework of canoncial conjugate priors from Section 1.2.3.1, by requiring only the form of the update step (1.6), and not the specific functional form of (1.4) for $f(\boldsymbol{x} \mid \vartheta)$ (see Definition 3.1).

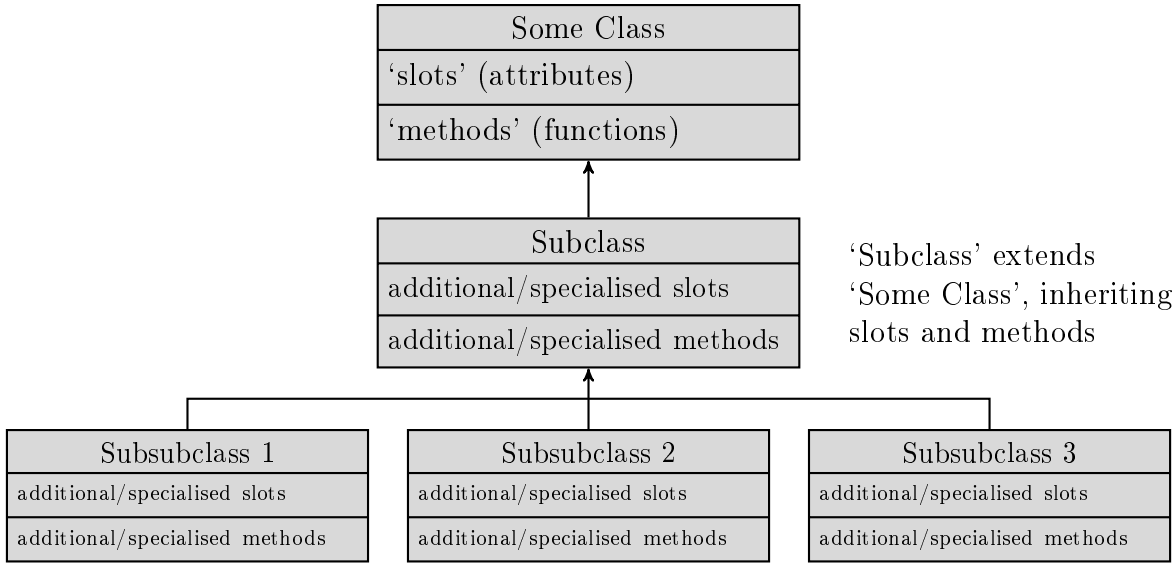

Figure 3.7.: Illustration of class hierarchies in object-oriented software. Each class is drawn as a rectangle with three parts, the top part giving the name of the class. The two lower parts name the slots and the methods for the class, respectively. A class that inherts from another class is linked to this other class with an arrow. Such graphs are called *UML diagrams*.

by defining the central class `LuckModel`. Inferences in a concrete sample distribution are then carried out using lean subclasses that make the 'translation' between the (abstract) LUCK-model framework and a concrete sample distribution.

For defining prior parameter sets $\mathbb{\Pi}^{(0)}$, the class `LuckModel` provides the slots `n0` for $n^{(0)}$, and `y0` for $y^{(0)}$, respectively, the contents of which are defined internally as intervals, but where the lower and upper bound may coincide, such that both $\mathcal{N}^{(0)}$ and $\mathcal{Y}^{(0)}$ can be either an interval or a singleton.[78] From the model types discussed in Section 3.1.1, type (a), (b), and (c) can thus be implemented by choosing `n0` and `y0` accordingly as singletons or intervals. To calculate the set of posterior distributions, the class `LuckModel` also provides a `data` slot. This must contain an object of class `LuckModelData`, providing the data in the needed form ($\tau(\boldsymbol{x})$ and n).

Posterior Parameter Sets. As the posterior parameter sets $\mathbb{\Pi}^{(n)}$ resulting from updating all the priors in a prior parameter set $\mathbb{\Pi}^{(0)}$ are, in the most general case of $\mathbb{\Pi}^{(0)} = [\underline{n}^{(0)}, \overline{n}^{(0)}] \times [\underline{y}^{(0)}, \overline{y}^{(0)}]$, not cartesian products of $[\underline{n}^{(n)}, \overline{n}^{(n)}]$ and $[\underline{y}^{(n)}, \overline{y}^{(n)}]$ anymore, posterior sets are not explicitely represented as `LuckModel` objects. Whenever posterior quantities are of interest (specified in functions or methods for `LuckModel` objects by the option

[78] In case of $\boldsymbol{y}^{(0)} \in \mathbb{R}^k$, $\mathcal{Y}^{(0)}$ may be a multidimensional interval, i.e., a cartesian product of intervals $[\underline{y}_j^{(0)}, \overline{y}_j^{(0)}]$, $j = 1, \ldots, k$, (see Example 3.4), with $\underline{y}^{(0)}$ and $\overline{y}^{(0)}$ denoting the vectors of these elementwise lower and upper bounds, respectively.

`"posterior = TRUE"`), the range of these quantities are calculated by minimising and maximising over $\mathbb{\Pi}^{(n)}$, which in turn can be done by a box-constrained optimisation over $\mathbb{\Pi}^{(0)}$ that is a cartesian product. This is why the data object (`LuckModelData`) is directly included in the `LuckModel` object. For the box-constrained optimisation, a helper function called `wrapOptim` is used, such that all cases (none, one or both of `n0` and `y0` are interval-valued) can be treated in the same way.[79]

Subclasses for Concrete Distributions. As mentioned above, the class `LuckModel` in fact only implements the general superstructure as given in the (abstract) definition of LUCK-models (see Definition 3.1). For inference in a concrete distribution family like the Normal distribution, this general framework must be concretised by defining a subclass, inheriting from `LuckModel`, for this specific distribution family, along with a subclass of `LuckModelData` to specify how $\tau(\boldsymbol{x})$ is calculated for this distribution family. Currently, this has been done for data from a scaled normal distribution, i.e., $x_i \sim \mathrm{N}(\mu, 1)$,[80] with the classes `ScaledNormalLuckModel` and `ScaledNormalData`, and for data from an exponential distribution, i.e., $x_i \sim \mathrm{Exp}(\lambda)$,[81] with the classes `ExponentialLuckModel` and `ExponentialData`. Figure 3.8 shows the UML diagram, illustrating the hierarchical structure of the package.

Implemented Methods. For illustrations of `LuckModel` objects and inferences based on them, some methods have been written. First, there are methods to display and print plain `LuckModel` objects (existing only on the superstructure level):

- So-called *constructor functions* for the `LuckModel` and `LuckModelData` class make the creation of LUCK-models more easy. They can be supplied with a number of different arguments, which are checked on consistency upon creation of the object.
- Due to the S4 implementation, so-called *accessor* and *replacement functions* are defined, regulating the access and the replacement of object slots.
- The `show` method prints the contents of a `LuckModel` object in more readable form. If the object contains a `LuckModelData object`, this is printed along as well, resorting on a `show` method for `LuckModelData` objects.[82]
- The `plot` method for `LuckModel` objects represents the parameter sets graphically.

Secondly, there are methods for working with and displaying the resulting sets of prior or posterior distributions for concrete data distributions, along with two exemplary inference methods.

[79] The function `optim` provided by **R** for general-purpose multivariate optimisation is not recommended for univariate optimisation.

[80] See Example 3.3, p. 86f.

[81] See the study by Krautenbacher (2011), who contributed the code for this distribution family. The results of this study are briefly discussed in Section 3.2.3.2.

[82] `show` methods are the S4 equivalent to `print` methods for S3 objects.

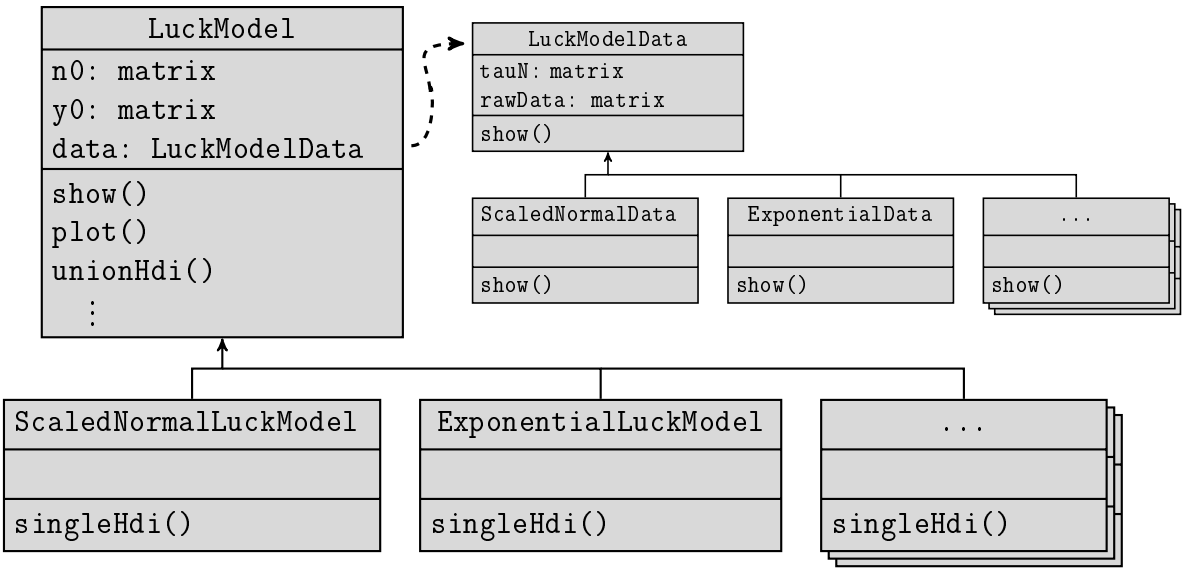

Figure 3.8.: UML diagram for the `luck` package, illustrating the class hierarchy. In UML diagrams, the slots and methods that subclasses inherit are not specify explicitely; only new slots and methods are displayed. The class of each slot is given next to its name. The `data` slot in `LuckModel` is of class `LuckModelData`, indicated by the dashed arrow.

- The constructor functions are modified, for example to check if the input fits to the data or prior distribution, or to allow the simulation of data according to the data distribution when creating the `LuckModelData` object.

- The accessor and replacement functions need not be specified again for the specialized classes, as those can be taken from the general classes, i.e., these functions are 'inherited' from the respective `LuckModel` or `LuckModelData` class. An exception is the function for replacing the raw data in the `LuckModelData object`, as there, the input must be checked to fit the sample space domain. As an example, in case of the `ExponentialData` class, data must be strictly positive.

- The `show` methods for `LuckModel` and `LuckModelData` are specialised in order to explain to the user the meaning of the canonical parameters $n^{(0)}$ and $y^{(0)}$.

- `unionHdi` calculates the union of highest density intervals for a specialised `LuckModel` object.[83] This method relies on a function `singleHdi`, that gives a highest density interval for a single parameter combination (n, y) for the respective conjugate prior or posterior distribution. Therefore, for each specialised LuckModel class, `singleHdi` must be provided.

[83] Highest density (HD) intervals were discussed in Section 1.2.3.3, while unions of highest density intervals were considered, e.g., in Example 3.3, p. 86f.

- `cdfplot` plots the set of cumulative density functions for specialised `LuckModel` objects. Again, this method relies on a function `singleCdf`, returning values of the cumulative density function for a single parameter combination (n, y) for the respective conjugate prior distribution.

3.5. On Prior-Data Conflict in Predictive Bernoulli Inferences

This section reproduces the work "On Prior-Data Conflict in Predictive Bernoulli Inferences", published as a peer-reviewed contribution to *ISIPTA'11: Proceedings of the Seventh International Symposium on Imprecise Probabilities: Theories and Applications* (Walter, Augustin, and Coolen 2011). As such, it is reproduced here almost verbatim, except for some minor shortenings, especially in the Introduction (Section 3.5.1), and the addition of some comments and footnotes linking this work to other parts of this thesis. Furthermore, the notation was changed slightly to assure consistency with the rest of the material presented in this thesis. Specifically, the success probability in Bernoulli trials is denoted by θ instead of p.

Here, we discuss the ability of different approaches to deal with prior-data conflict by focussing on the prototypical example of Bernoulli trials. We study a generalised Bayesian setting, including Walley's Imprecise Beta-Binomial model (a model of type (a) in the framework of Section 3.1.1) and his extension to handle prior data conflict (which is of type (c), and called pdc-IBBM here). We then investigate a model of type (d) by proposing an alternative shape for prior parameter sets, chosen in a way that shows improved behaviour in the case of prior-data conflict, and describe its influence on the posterior predictive distribution. Thereafter we present a new approach (outside the framework of Section 3.1.1), consisting of an imprecise weighting of two originally separate inferences, one of which is based on an informative imprecise prior, whereas the other one is based on an uninformative imprecise prior. This approach deals with prior-data conflict in a fascinating way.

3.5.1. Introduction

Imprecise probability has shown to be a powerful methodology to cope with the multidimensional nature of uncertainty (see the discussion in Sections 2.1 and 2.2). Imprecision allows the quality of information, on which probability statements are based, to be modeled. Well supported knowledge is expressed by comparatively precise models, while highly imprecise (or even vacuous) models reflect scarce (or no) knowledge on probabilities. This flexible, multidimensional perspective on uncertainty modeling has intensively been utilized in generalised Bayesian inference to overcome the criticism of the arbitrariness of the choice of single prior distributions in traditional Bayesian inference.[84] In addition, only imprecise probability models react reliably to the presence of prior-data conflict, i.e. situations where "the prior [places] its mass primarily on distributions in the sampling model for which the observed data is surprising" (Evans and Moshonov 2006, p. 894). Lower and

[84] This criticism, subsumed by Walley (1991, §5) as the "dogma of precision", is discussed more detailed in Section 2.2.3.

upper probabilities[85] allow a specific reaction to prior-data conflict, and offer reasonable inferences if the analyst wishes to stick to his prior assumptions: starting with the same level of ambiguity in the prior specification, wide posterior intervals can reflect conflict between prior and data, while no prior-data conflict will lead to narrow intervals.[86] Ideally, the model could provide an extra 'bonus' of precision if prior assumptions are very strongly supported by the data. Such a model would have the advantage of (relatively) precise answers when the data confirm prior assumptions, while still rendering more cautionary answers in the case of prior-data conflict, thus leading to cautious inferences if, and only if, caution is needed.

Although Walley (1991, p. 6) explicitly emphasizes this possibility to express prior-data conflict as one of the main motivations for imprecise probability, it has received surprisingly little attention. Rare exceptions include two short sections in Walley (1991, p. 6 and §5.4) and the papers by Pericchi and Walley (1991), Coolen (1994) and Whitcomb (2005). The popular IDM (Walley 1996a; Bernard 2009) and its generalisation to exponential families (Quaeghebeur and Cooman 2005) do not reflect prior-data conflict. Walter and Augustin (2009b, see Section 3.3) used the basic ideas of Walley (1991, §5.4) to extend the approach of Quaeghebeur and Cooman (2005) to models that show sensitivity to prior-data conflict.

In this work, a deeper investigation of the issue of prior-data conflict is undertaken, focusing on the prototypic special case of predictive inference in Bernoulli trials:[87] We are interested in the posterior predictive probability for the event that a future Bernoulli random quantity will have the value 1, also called a 'success'. This event is not explicitly included in the notation, i.e. we simply denote its lower and upper probabilities by $\underline{P}$ and $\overline{P}$, respectively. This future Bernoulli random quantity is assumed to be exchangeable with the Bernoulli random quantities whose observations are summarised in the data, consisting of the number n of observations and the number s of these that are successes. In our analysis of this model, we will often treat s as a a real-valued observation in $[0, n]$, keeping in mind that in reality it can only take integer values, but the continuous representation is convenient for our discussions, in particular in our predictive probability plots (PPP), where for given n, $\underline{P}$ and $\overline{P}$ are discussed as functions of s.

Section 3.5.2.1 describes a general framework for generalised Bayesian inference in this setting. The method presented in Walley (1991, §5.4.3), called 'pdc-IBBM' in this paper, is considered in detail in Section 3.5.2.2 and we show that its reaction to prior-data conflict can be improved by suitable modifications of the underlying imprecise priors. A basic proposal along these lines is discussed in Section 3.5.2.3, with further alternatives sketched in Section 3.5.2.4. Section 3.5.3 addresses the problem of prior-data conflict from a completely different angle. There, we combine two originally separate inferences, one based on an informative imprecise prior and one on an uninformative imprecise prior, by

[85] See Section 2.1.2 for a short exposition of lower and upper previsions, and related mathematical tools for handling uncertainty in statistical inference.

[86] See the discussions of prior-data conflict sensitivity in Sections 2.2.3.3 and 3.1.4, and our contribution centerd around this issue in Section 3.3.

[87] See also the discussion of the Beta-Binomial model in Section 1.2.3.3, and the imprecise Dirichlet-Multinomial model discussed in several examples in Section 3.3.

an imprecise weighting scheme. We conclude the contribution with a brief comparison of the different approaches in Section 3.5.4.

3.5.2. Imprecise Beta-Binomial Models

3.5.2.1. The Framework

The traditional Bayesian approach for our basic problem is the Beta-Binominal model, which expresses prior beliefs about the probability θ of observing a 'success' by a Beta distribution. With[88] $f(\theta \mid n^{(0)}, y^{(0)}) \propto \theta^{n^{(0)}y^{(0)}-1}(1-\theta)^{n^{(0)}(1-y^{(0)})-1}$, $y^{(0)} = \mathrm{E}[\theta \mid n^{(0)}, y^{(0)}]$ can be interpreted as prior guess of θ, while $n^{(0)}$ governs the concentration of probability mass around $y^{(0)}$, also known as 'pseudo counts' or 'prior strength'.[89] These denominations are due to the role of $n^{(0)}$ in the update step: With s successes in n draws observed, the posterior parameters are[90]

$$n^{(n)} = n^{(0)} + n, \qquad y^{(n)} = \frac{n^{(0)}y^{(0)} + s}{n^{(0)} + n}. \tag{3.19}$$

Thus $y^{(n)}$ is a weighted average of the prior parameter $y^{(0)}$ and the sample proportion s/n, and potential prior data conflict is simply averaged out.

Overcoming the dogma of precision, formulating generalised Bayes updating in this setting is straightforward. By Walley's Generalised Bayes' Rule (Walley 1991, §6)[91] the imprecise prior $\mathcal{M}^{(0)}$, described by convex sets of precise prior distributions, is updated to the imprecise posterior $\mathcal{M}^{(n)}$ obtained by updating $\mathcal{M}^{(0)}$ element-wise. In particular, the convenient conjugate analysis used above can be extended:[92] One specifies a prior parameter set $\Pi^{(0)}$ of $(n^{(0)}, y^{(0)})$ values and takes as imprecise prior the set $\mathcal{M}^{(0)}$ consisting of all convex mixtures of Beta priors with $(n^{(0)}, y^{(0)}) \in \Pi^{(0)}$. In this sense, the set of Beta priors corresponding to $\Pi^{(0)}$ gives the set of extreme points for the actual convex set of priors $\mathcal{M}^{(0)}$. Updating $\mathcal{M}^{(0)}$ with the Generalised Bayes' Rule results in the convex set $\mathcal{M}^{(n)}$ of posterior distributions, that conveniently can be obtained by taking the convex hull of the set of Beta posteriors, which in turn are defined by the set of updated parameters $\Pi^{(n)} = \{(n^{(n)}, y^{(n)}) \mid (n^{(0)}, y^{(0)}) \in \Pi^{(0)}\}$. This relationship between the sets $\Pi^{(0)}$ and $\Pi^{(n)}$ and the sets $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(n)}$ will allow us to discuss different models $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(n)}$ by depicting the corresponding parameter sets $\Pi^{(0)}$ and $\Pi^{(n)}$. When interpreting our results, care will be needed with respect to convexity. Although $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(n)}$ are convex, the

[88] Our notation relates to Walley's (1991) as $n^{(0)} \leftrightarrow s_0$, $y^{(0)} \leftrightarrow t_0$.

[89] As in previous parts of the thesis, $^{(0)}$ denotes prior parameters; $^{(n)}$ posterior parameters.

[90] Compare to Section 1.2.3.1, Equation (1.6): the model is prototypic for conjugate Bayesian analysis in canonical exponential families, for which updating of the parameters $n^{(0)}$ and $y^{(0)}$ can be written as (3.19).

[91] For more details on the Generalised Bayes' Rule and the generalised Bayesian inference procedure, see Sections 2.1.2.5 and 2.1.3, respectively.

[92] See the discussion of the general framework for generalised Bayesian inference based on sets of conjugate priors in Section 3.1.1.

parameter sets $\mathbb{\Pi}^{(0)}$ and $\mathbb{\Pi}^{(n)}$ generating them need not necessarily be so. Indeed, convexity of the parameter set is not necessarily preserved in the update step: Convexity of $\mathbb{\Pi}^{(0)}$ does not imply convexity of $\mathbb{\Pi}^{(n)}$.

Throughout, we are interested in the posterior predictive probability $[\underline{P}, \overline{P}]$ for the event that a future draw is a success. In the Beta-Bernoulli model, this probability is equal to $y^{(n)}$, and we get

$$\underline{P} = \underline{y}^{(n)} := \min_{\mathbb{\Pi}^{(n)}} y^{(n)} = \min_{\mathbb{\Pi}^{(0)}} \frac{n^{(0)} y^{(0)} + s}{n^{(0)} + n}, \tag{3.20}$$

$$\overline{P} = \overline{y}^{(n)} := \max_{\mathbb{\Pi}^{(n)}} y^{(n)} = \max_{\mathbb{\Pi}^{(0)}} \frac{n^{(0)} y^{(0)} + s}{n^{(0)} + n}. \tag{3.21}$$

3.5.2.2. Walley's pdc-IBBM

Special imprecise probability models are now obtained by specific choices of $\mathbb{\Pi}^{(0)}$. If one fixes $n^{(0)}$ and varies $y^{(0)}$ in an interval $[\underline{y}^{(0)}, \overline{y}^{(0)}]$, Walley's (1991, §5.3) model with learning parameter $n^{(0)}$ is obtained, which typically is used in its near-ignorance form $[\underline{y}^{(0)}, \overline{y}^{(0)}] \to (0, 1)$, denoted as the imprecise Beta (Binomal/Bernoulli) model (IBBM)[93], which is a special case of the popular Imprecise Dirichlet (Multinomial) Model (Walley 1996a; Walley and Bernard 1999). Unfortunately, in this basic form with fixed $n^{(0)}$, the model is insensitive to prior-data conflict (Walter and Augustin 2009b, p. 263, see Section 3.3.3.2). Walley (1991, §5.4) therefore generalised this model by additionally varying $n^{(0)}$. In his extended model, called *pdc-IBBM* in this work, the set of priors is defined via the set of prior parameters $\mathbb{\Pi}^{(0)} = [\underline{n}^{(0)}, \overline{n}^{(0)}] \times [\underline{y}^{(0)}, \overline{y}^{(0)}]$, being a two-dimensional interval, or a rectangle set. Studying inference in this model, it is important to note that the set of posterior parameters $\mathbb{\Pi}^{(n)}$ is not rectangular anymore. The resulting shapes are illustrated in Figure 3.9: For the prior set $\mathbb{\Pi}^{(0)} = [1, 5] \times [0.4, 0.7]$—thus assuming a priori the fraction of successes to be between 40% and 70% and rating these assumptions with at least 1 and at most 5 pseudo observations—the resulting posterior parameter sets $\mathbb{\Pi}^{(n)}$ are shown for data consisting of 3 successes in 6 draws (left) and with all 6 draws successes (right). We call the left shape *spotlight*, and the right shape *banana*. In both graphs, the elements of $\mathbb{\Pi}^{(n)}$ yielding $\underline{y}^{(n)}$ and $\overline{y}^{(n)}$, and thus $\underline{P}$ and $\overline{P}$, are marked with a circle.

The transition point between the *spotlight* and the *banana* shape in Figure 3.9 is the case when $\frac{s}{n} = \overline{y}^{(0)}$. Then $\overline{y}^{(n)}$, being a weighted average of $\overline{y}^{(0)}$ and $\frac{s}{n}$, is attained for all $n^{(0)} \in [\underline{n}^{(0)}, \overline{n}^{(0)}]$, and the top border of $\mathbb{\Pi}^{(n)}$ in the graphical representation of Figure 3.9 is constant. Likewise, $\underline{y}^{(n)}$ is constant if $\frac{s}{n} = \underline{y}^{(0)}$. Therefore, (3.20) and (3.21) can be subsumed as

$$\underline{P} = \begin{cases} \frac{\overline{n}^{(0)} \underline{y}^{(0)} + s}{\overline{n}^{(0)} + n} & \text{if } s \geq n \cdot \underline{y}^{(0)} =: S_1 \\ \frac{\underline{n}^{(0)} \underline{y}^{(0)} + s}{\underline{n}^{(0)} + n} & \text{if } s \leq n \cdot \underline{y}^{(0)} =: S_1 \end{cases},$$

[93] We use 'IBBM' also for the model with prior information.

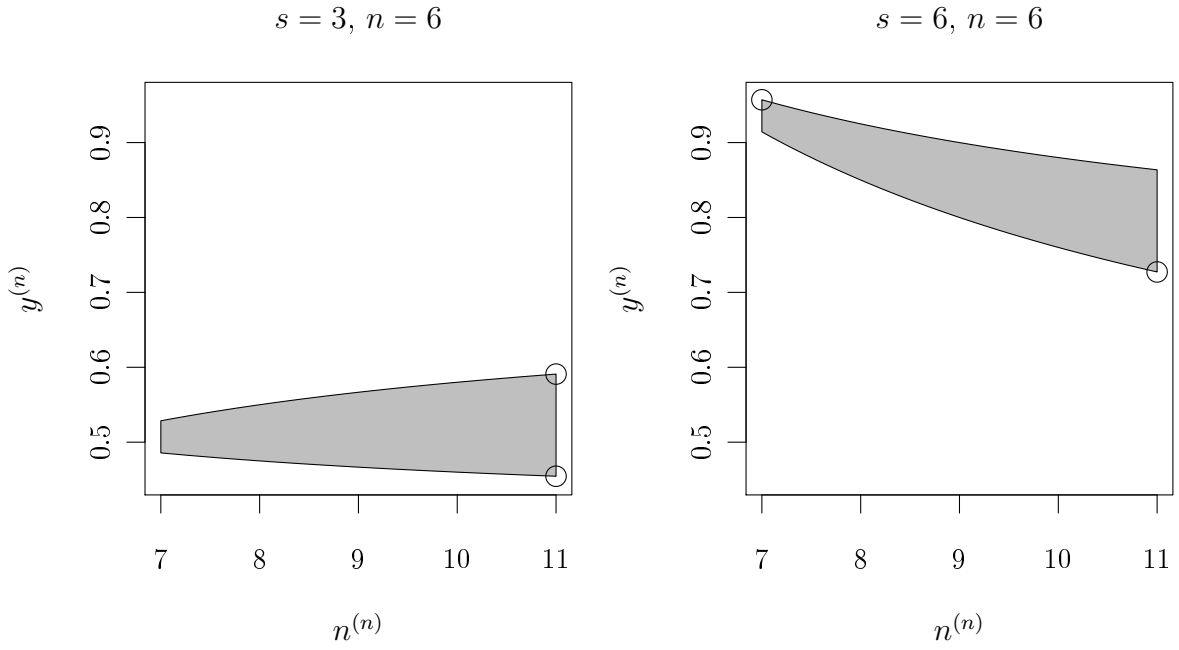

Figure 3.9.: Posterior parameter sets $\mathbb{\Pi}^{(n)}$ for rectangular $\mathbb{\Pi}^{(0)}$. Left: *spotlight* shape; right: *banana* shape.

$$\overline{\mathrm{P}} = \begin{cases} \frac{\overline{n}^{(0)}\overline{y}^{(0)}+s}{\overline{n}^{(0)}+n} & \text{if } s \leq n \cdot \overline{y}^{(0)} =: S_2 \\ \frac{\underline{n}^{(0)}\underline{y}^{(0)}+s}{\underline{n}^{(0)}+n} & \text{if } s \geq n \cdot \overline{y}^{(0)} =: S_2 \end{cases} .$$

The interval $[S_1, S_2]$ gives the range of expected successes $[n \cdot \underline{y}^{(0)}, n \cdot \overline{y}^{(0)}]$ and will be called 'Total Prior-Data Agreement' interval, or TPDA. For s in the TPDA, we are 'spot on': $\underline{y}^{(n)}$ and $\overline{y}^{(n)}$ are attained for $\overline{n}^{(0)}$ and $\mathbb{\Pi}^{(n)}$ has the *spotlight* shape. But if the observed number of successes is outside TPDA, $\mathbb{\Pi}^{(n)}$ goes *bananas* and either $\underline{\mathrm{P}}$ or $\overline{\mathrm{P}}$ is calculated with $\underline{n}^{(0)}$.

To summarise, the predictive probability plot (PPP), displaying $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ for $s \in [0, n]$, is given in Figure 3.10. For the pdc-IBBM, the specific values are

$$\begin{aligned} S_1 &= n\underline{y}^{(0)} & A &= \frac{\underline{n}^{(0)}\underline{y}^{(0)}}{\underline{n}^{(0)}+n} & C &= \frac{\overline{n}^{(0)}\underline{y}^{(0)}+n}{\overline{n}^{(0)}+n} \\ S_2 &= n\overline{y}^{(0)} & B &= \frac{\overline{n}^{(0)}\overline{y}^{(0)}}{\overline{n}^{(0)}+n} & D &= \frac{\underline{n}^{(0)}\overline{y}^{(0)}+n}{\underline{n}^{(0)}+n} \\ E_1 &= \underline{y}^{(0)} & E_2 &= \frac{\overline{n}^{(0)}\overline{y}^{(0)}+n\underline{y}^{(0)}}{\overline{n}^{(0)}+n} & \text{sl. } 1 &= \frac{1}{\overline{n}^{(0)}+n} \\ F_2 &= \overline{y}^{(0)} & F_1 &= \frac{\overline{n}^{(0)}\underline{y}^{(0)}+n\overline{y}^{(0)}}{\overline{n}^{(0)}+n} & \text{sl. } 2 &= \frac{1}{\underline{n}^{(0)}+n} . \end{aligned}$$

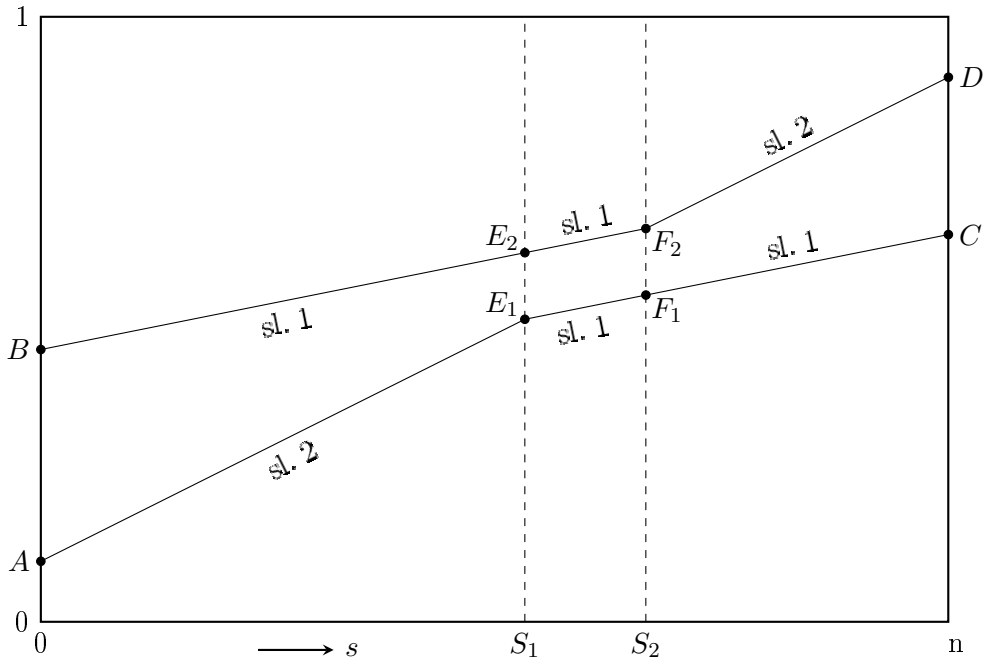

Figure 3.10.: $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ for models in Sections 3.5.2.2 and 3.5.2.3.

As noted by Walley (1991, p. 224), the posterior predictive imprecision $\Delta = \overline{\mathrm{P}} - \underline{\mathrm{P}}$ can be calculated as

$$\Delta = \frac{\overline{n}^{(0)}(\overline{y}^{(0)} - \underline{y}^{(0)})}{\overline{n}^{(0)} + n} + \frac{\overline{n}^{(0)} - \underline{n}^{(0)}}{(\underline{n}^{(0)} + n)(\overline{n}^{(0)} + n)} \Delta(s, \mathbb{\Pi}^{(0)}) \,,$$

where $\Delta(s, \mathbb{\Pi}^{(0)}) = \inf\{|s - ny^{(0)}| : y^{(0)} \in [\underline{y}^{(0)}, \overline{y}^{(0)}]\}$ is the distance of s to the TPDA. If $\Delta(s, \mathbb{\Pi}^{(0)}) \neq 0$, we have an effect of additional imprecision as desired, increasing linearly in s, because $\mathbb{\Pi}^{(n)}$ is going *bananas*. However, when considering the fraction of observed successes instead of s, the onset of this additional imprecision immediately if $\frac{s}{n} \notin [\underline{y}^{(0)}, \overline{y}^{(0)}]$ seems very abrupt. Moreover, and even more severe, it happens irrespective of the number of trials n. When updating successively, this means that all single Bernoulli observations, being either 0 or 1, have to be treated as if being in conflict (except if $\overline{y}^{(0)} = 1$ and $s = n$ or if $\underline{y}^{(0)} = 0$ and $s = 0$). Furthermore, regarding $s/n = 7/10$ as an instance of prior-data conflict when $\overline{y}^{(0)} = 0.6$ had been assumed seems somewhat picky. To explore possibilities to amend this behaviour, alternative approaches are explored next.

3.5.2.3. *Anteater* Shape Prior Sets

Choosing a two-dimensional interval $\mathbb{\Pi}^{(0)}$ seems logical, but the resulting inference is not fully satisfactory in case of prior data conflict. Recall that $\mathbb{\Pi}^{(0)}$ is used to produce $\mathcal{M}^{(0)}$, which then is processed by the Generalised Bayes rule. Any shape can be chosen for $\mathbb{\Pi}^{(0)}$,

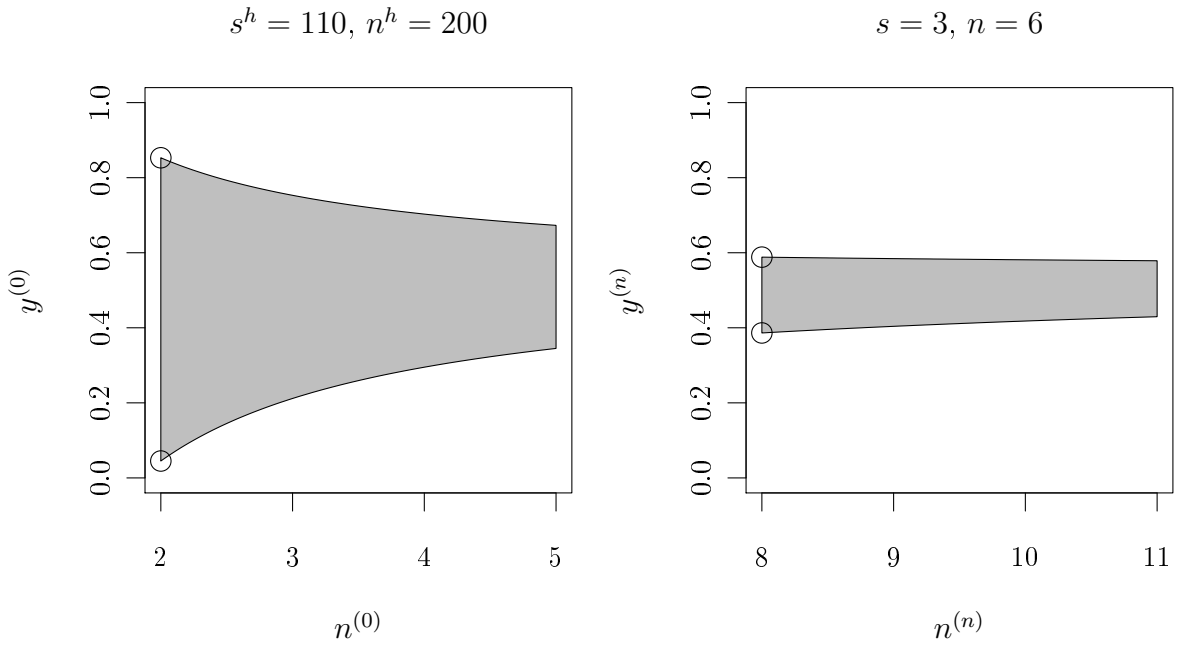

Figure 3.11.: $\mathbb{\Pi}^{(0)}$ and $\mathbb{\Pi}^{(n)}$ for the *anteater* shape.

including the composure of single pairs $(n^{(0)}, y^{(0)})$. Here, we investigate an alternative shape, with $y^{(0)}$ a function of $n^{(0)}$, aiming at a more advanced behaviour in the case of prior-data conflict. To elicit $\mathbb{\Pi}^{(0)}$, one could consider a thought experiment:[94] Given the hypothetical observation of s^h successes in n^h trials, which values should $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ take? In other words, what would one like to learn from data s^h/n^h in accordance with prior beliefs? As a simple approach, we can define $\mathbb{\Pi}^{(0)}$ such that $\underline{\mathrm{P}} = \underline{c}$ and $\overline{\mathrm{P}} = \overline{c}$ are constants in $n^{(n)} = n^{(0)} + n^h$. Then, the lower and upper bounds for $y^{(0)}$ must be

$$\begin{aligned} \underline{y}^{(0)}(n^{(0)}) &= \frac{(n^h + n^{(0)})\underline{c} - s^h}{n^{(0)}}, \\ \overline{y}^{(0)}(n^{(0)}) &= \frac{(n^h + n^{(0)})\overline{c} - s^h}{n^{(0)}}, \end{aligned} \tag{3.22}$$

for $n^{(0)}$ in an interval $[\underline{n}^{(0)}, \overline{n}^{(0)}]$ derived by the range $[\underline{n}^{(n)}, \overline{n}^{(n)}]$ one wishes to attain for $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ given the n^h hypothetical observations.[95] The resulting shape of $\mathbb{\Pi}^{(0)}$ is as in Figure 3.11 (left) and called *anteater* shape. Rewriting (3.22), $\mathbb{\Pi}^{(0)}$ is now defined by

$$\mathbb{\Pi}^{(0)} = \left\{ (n^{(0)}, y^{(0)}) \,\middle|\, n^{(0)} \in [\underline{n}^{(0)}, \overline{n}^{(0)}],\ y^{(0)}(n^{(0)}) \in \left[\underline{c} - \frac{n^h}{n^{(0)}} \left(\frac{s^h}{n^h} - \underline{c} \right),\ \overline{c} + \frac{n^h}{n^{(0)}} \left(\overline{c} - \frac{s^h}{n^h} \right) \right] \right\}.$$

With the reasonable choice of $\underline{c}$ and $\overline{c}$ such that $\underline{c} \leq s^h/n^h \leq \overline{c}$, $\mathbb{\Pi}^{(0)}$ can be interpreted as follows: The range of $y^{(0)}$ protrudes over $[\underline{c}, \overline{c}]$ on either side far enough to ensure $\underline{\mathrm{P}} = \underline{c}$

[94] This strategy is also known as 'pre-posterior' analysis in the Bayesian literature.

[95] For the rest of this section, we tacitly assume that n^h, s^h, $n^{(0)}$ and $\underline{c}/\overline{c}$ are chosen such that $\underline{y}^{(0)} \geq 0$ resp. $\overline{y}^{(0)} \leq 1$ to generate Beta distributions as priors.

and $\overline{\mathrm{P}} = \overline{c}$ if updated with $s = s^h$ for $n = n^h$, the amount of protrusion decreasing in $n^{(0)}$ as the movement of $y^{(0)}(n^{(0)})$ towards s^h/n^h is slower for larger values of $n^{(0)}$. As there is a considerable difference in behaviour if $n > n^h$ or $n < n^h$, these two cases are discussed separately.

If $n > n^h$, the PPP graph in Figure 3.10 holds again, now with the values

$$A = \frac{\underline{c}(\underline{n}^{(0)} + n^h) - s^h}{\underline{n}^{(0)} + n} \qquad S_1 = s^h + \underline{c}(n - n^h)$$

$$B = \frac{\overline{c}(\overline{n}^{(0)} + n^h) - s^h}{\overline{n}^{(0)} + n} \qquad S_2 = s^h + \overline{c}(n - n^h)$$

$$C = \frac{\underline{c}(\overline{n}^{(0)} + n^h) - s^h + n}{\overline{n}^{(0)} + n} \qquad \text{sl. } 1 = 1/(\overline{n}^{(0)} + n)$$

$$D = \frac{\overline{c}(\underline{n}^{(0)} + n^h) - s^h + n}{\underline{n}^{(0)} + n} \qquad \text{sl. } 2 = 1/(\underline{n}^{(0)} + n)$$

$$E_1 = \underline{c} \qquad E_2 = \underline{c} + \frac{\overline{n}^{(0)} + n^h}{\overline{n}^{(0)} + n}(\overline{c} - \underline{c}) = \overline{c} - \frac{n - n^h}{\overline{n}^{(0)} + n}(\overline{c} - \underline{c})$$

$$F_2 = \overline{c} \qquad F_1 = \overline{c} - \frac{\overline{n}^{(0)} + n^h}{\overline{n}^{(0)} + n}(\overline{c} - \underline{c}) = \underline{c} + \frac{n - n^h}{\overline{n}^{(0)} + n}(\overline{c} - \underline{c})\,.$$

As for the pdc-IBBM, the TPDA boundaries S_1 and S_2 mark the transition points where either $\underline{y}^{(n)}$ or $\overline{y}^{(n)}$ are constant in $n^{(0)}$. We now have

$$\frac{S_1}{n} = \underline{c} + \frac{n^h}{n}\Big(\frac{s^h}{n^h} - \underline{c}\Big), \qquad \frac{S_2}{n} = \overline{c} - \frac{n^h}{n}\Big(\overline{c} - \frac{s^h}{n^h}\Big),$$

so this TPDA is a subset of $[\underline{c}, \overline{c}]$. The *anteater* shape is, for $n > n^h$, even more strict than the pdc-IBBM, as, e.g., $\underline{y}^{(0)}(\overline{n}^{(0)}) = \underline{c} - \frac{n^h}{\overline{n}^{(0)}}\big(\frac{s^h}{n^h} - \underline{c}\big) < \frac{S_1}{n}$.

The situation for $n < n^h$ is illustrated in Figure 3.12, where A, B, C, D, E_1, F_2 and slopes 1 and 2 are the same as for $n > n^h$, but

$$E_2 = \underline{c} + \frac{\underline{n}^{(0)} + n^h}{\underline{n}^{(0)} + n}(\overline{c} - \underline{c}) = \overline{c} + \frac{n^h - n}{\underline{n}^{(0)} + n}(\overline{c} - \underline{c})\,,$$

$$F_1 = \overline{c} - \frac{\underline{n}^{(0)} + n^h}{\underline{n}^{(0)} + n}(\overline{c} - \underline{c}) = \underline{c} - \frac{n^h - n}{\underline{n}^{(0)} + n}(\overline{c} - \underline{c})\,.$$

Note that now $S_2 < S_1$, so the TPDA is $[S_2, S_1]$. In this interval, $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ are now calculated with $\underline{n}^{(0)}$; for $s \notin [S_2, S_1]$ the same situation as for $n > n^h$ applies, with the bound nearer to s/n calculated with $\underline{n}^{(0)}$ and the other with $\overline{n}^{(0)}$.

The upper transition point S_1 can now be between $\overline{y}^{(0)}(\underline{n}^{(0)})$ and $\overline{y}^{(0)}(\overline{n}^{(0)})$, and having S_1 decreasing in n now makes sense: the smaller n, the larger S_1, i.e. the more tolerant is the *anteater* set. The switch over S_1 (with s/n increasing) is illustrated in the three

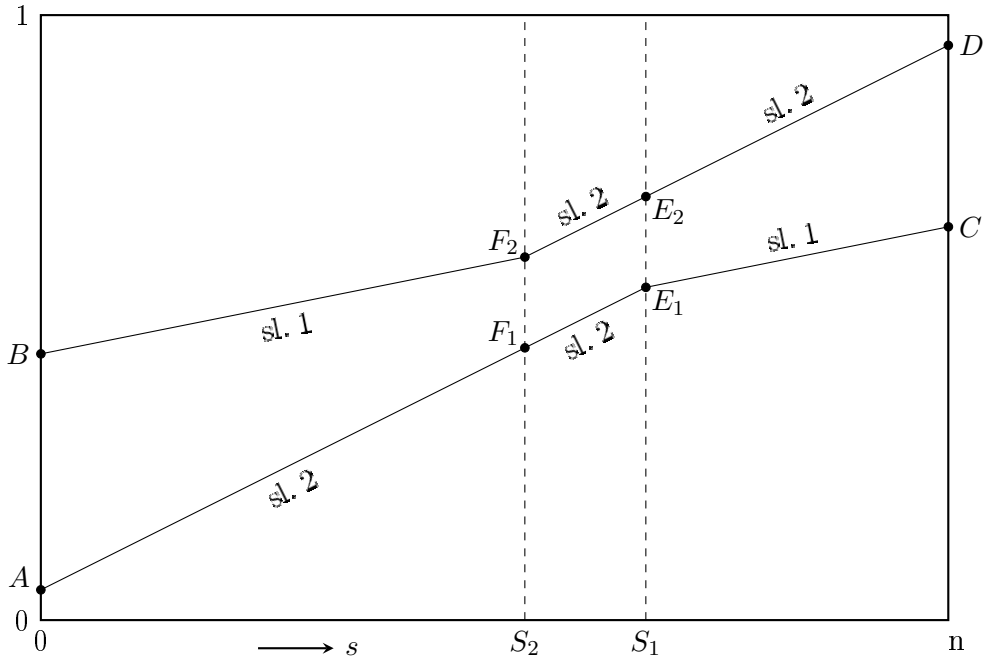

Figure 3.12.: $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ for the *anteater* shape if $n < n^h$.

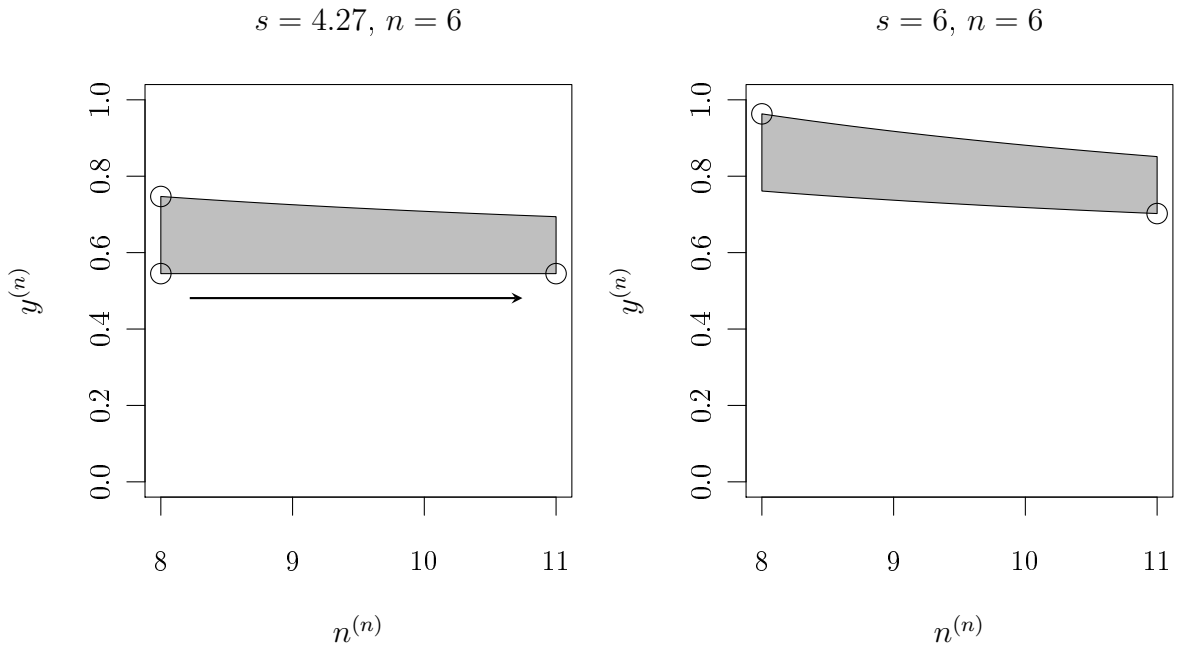

Figure 3.13.: Posterior parameter sets $\mathbb{\Pi}^{(n)}$ for *anteater* prior sets $\mathbb{\Pi}^{(0)}$. Left: the transition point where the lower contour of the posterior parameter set is attained for all $n^{(n)}$, right: the *banana* shape.

	$\mathbb{\Pi}^{(n)}$ shape		
$n > n^h$	$s < S_1$ banana	$s \in [S_1, S_2]$ spotlight	$s > S_2$ banana
$n = n^h$	$s < s^h$ banana	$s = s^h$ rectangular	$s > s^h$ banana
$n < n^h$	$s < S_2$ banana	$s \in [S_2, S_1]$ anteater	$s > S_1$ banana

Table 3.2.: Shapes of $\mathbb{\Pi}^{(n)}$ if $\mathbb{\Pi}^{(0)}$ has the *anteater* shape.

graphs in Figures 3.11 (right) and 3.13 (left, right): First, $\mathbb{\Pi}^{(0)}$ from Figure 3.11 (left) is updated with $s/n = 3/6 < S_1/n$, leading again to an *anteater* shape, and so we get $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ from the elements of $\mathbb{\Pi}^{(n)}$ at $\underline{n}^{(n)}$, as marked with circles. Second, the transition point is reached for $s = S_1 = 4.27$, and now $\underline{\mathrm{P}}$ is attained for any $n^{(n)} \in [\underline{n}^{(n)}, \overline{n}^{(n)}]$, as emphasized by the arrow. Third, as soon as s exceeds S_1 (in the graph: $s/n = 6/6$), it holds that $\underline{y}^{(n)}(\underline{n}^{(n)}) > \underline{y}^{(n)}(\overline{n}^{(n)})$, and $\underline{\mathrm{P}}$ is now attained at $\overline{n}^{(n)}$. As for the pdc-IBBM, for s outside the TPDA $\mathbb{\Pi}^{(n)}$ goes *bananas*, leading to additional imprecision. The imprecision $\Delta = \overline{\mathrm{P}} - \underline{\mathrm{P}}$ if $n < n^h$ is

$$\Delta = \frac{\underline{n}^{(0)} + n^h}{\underline{n}^{(0)} + n}(\overline{c} - \underline{c}) + \frac{\overline{n}^{(0)} - \underline{n}^{(0)}}{(\underline{n}^{(0)} + n)(\overline{n}^{(0)} + n)}\Delta(s, n, \boldsymbol{c}),$$

where $\Delta(s, n, \boldsymbol{c}) = n\left|c^* - \frac{s}{n}\right| - n^h\left|c^* - \frac{s^h}{n^h}\right|$, and $c^* = \arg\max_{c \in [\underline{c}, \overline{c}]} \left|\frac{s}{n} - c\right|$ is the boundary of $\boldsymbol{c} := [\underline{c}, \overline{c}]$ with the largest distance to s/n. For $s \in [S_2, S_1]$, $\Delta(s, n, \boldsymbol{c}) = 0$, giving a similar structure as for the pdc-IBBM, except that $\Delta(s, n, \boldsymbol{c})$ does not directly give the distance of s/n to $\mathbb{\Pi}^{(0)}$ but is based on $[\underline{c}, \overline{c}]$. The imprecision increases again linearly with s, but now also with n. The distance of s/n to the opposite bound of $[\underline{c}, \overline{c}]$ (weighted with n) is discounted by the distance of s^h/n^h to the same bound (weighted with n^h). In essence, $\Delta(s, n, \boldsymbol{c})$ is thus a reweighted distance of s/n to s^h/n^h. The more dissimilar these fractions are, the larger the posterior predictive imprecision is.

For $n = n^h$, $S_1 = S_2 = s^h$, so the TPDA is reduced to a single point. In this case, the *anteater* shape can be considered as an equilibrium point, with any $s \neq s^h$ leading to increased posterior imprecision. In this case, the weights in $\Delta(s, n, \boldsymbol{c})$ coincide, and so the posterior imprecision depends directly on $|s - s^h|$.

For $n > n^h$ the transition behaviour is as for the pdc-IBBM: As long as $s \in [S_1, S_2]$, $\mathbb{\Pi}^{(n)}$ has the *spotlight* shape, where both $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ are calculated with $\overline{n}^{(n)}$; Δ for $s \in [S_1, S_2]$ is thus calculated with $\overline{n}^{(n)}$ as well. If, e.g., $s > S_2$, $\overline{\mathrm{P}}$ is attained with $\underline{n}^{(n)}$, and $\Delta(s, n, \boldsymbol{c})$ gives directly the distance of s/n to s^h/n^h, the part of which is inside $[\underline{c}, \overline{c}]$ is weighted with n, and the remainder with n^h. Table 3.2 provides an overview of the possible shapes of $\mathbb{\Pi}^{(n)}$.

3.5.2.4. Intermediate Résumé

Despite the (partly) different behaviour inside the TPDA, both pdc-IBBM and the *anteater* shape display only two different slopes in their PPPs (Figures 3.10 and 3.12), with either $\underline{n}^{(n)}$ or $\overline{n}^{(n)}$ used to calculate $\underline{P}$ and $\overline{P}$. It is possible to have shapes such that for some s other values from $[\underline{n}^{(n)}, \overline{n}^{(n)}]$ are used. As a toy example, consider $\mathbb{\Pi}^{(0)} = \{(1, 0.4), (3, 0.6), (5, 0.4)\}$, so consisting only of three parameter combinations $(n^{(0)}, y^{(0)})$. $\overline{P}$ is then derived as $\overline{y}^{(n)} = \max\{\frac{0.4+s}{1+n}, \frac{1.8+s}{3+n}, \frac{2+s}{5+n}\}$, leading to

$$\overline{y}^{(n)} = \begin{cases} \frac{0.4+s}{1+n} & \text{if } s > 0.7n + 0.3 \\ \frac{1.8+s}{3+n} & \text{if } 0.1n - 1.5 < s < 0.7n + 0.3 \\ \frac{2+s}{5+n} & \text{if } s < 0.1n - 1.5 \end{cases} .$$

So, in a PPP we would observe the three different slopes $1/(1+n)$, $1/(3+n)$ and $1/(5+n)$ depending on the value of s. Our conjecture is therefore that with carefully tailored sets $\mathbb{\Pi}^{(0)}$, an arbitrary number of slopes is possible, and so even smooth curvatures. Using a thought experiment as for the *anteater* shape, $\mathbb{\Pi}^{(0)}$ shapes can be derived to fit any required behaviour. Another approach for constructing a $\mathbb{\Pi}^{(0)}$ that is more tolerant with respect to prior-data conflict could be as follows: As the onset of additional imprecision in the pdc-IBBM is caused by the fact that $\overline{y}^{(n)}(\underline{n}^{(n)}) > \overline{y}^{(n)}(\overline{n}^{(n)})$ as soon as $s/n > \overline{y}^{(0)}$, we could define the $y^{(0)}$ interval at $\underline{n}^{(0)}$ to be narrower than the $y^{(0)}$ interval at $\overline{n}^{(0)}$, so that the *banana* shape results only when s/n exceeds $\overline{y}^{(0)}(\overline{n}^{(0)})$ far enough. Having a narrower $y^{(0)}$ interval at $\underline{n}^{(0)}$ than at $\overline{n}^{(0)}$ could also make sense from an elicitation point of view: We might be able to give quite a precise $y^{(0)}$ interval for a low prior strength $\underline{n}^{(0)}$, whereas for a high prior strength $\overline{n}^{(0)}$ we must be more cautious with our elicitation of $y^{(0)}$, i.e. giving a wider interval. The rectangular shape for $\mathbb{\Pi}^{(0)}$ as discussed in Section 3.5.2.2 seems thus somewhat peculiar. One could also argue that if one has substantial prior information, but acknowledges that this information may be wrong, one should not reduce the weight of the prior $n^{(0)}$ on the posterior while keeping the same informative interval of values of $y^{(0)}$.

Generally, the actual shape of a set $\mathbb{\Pi}^{(0)}$ influences the inferences, but for a specific inference, only a few aspects of the set are relevant. So, while a detailed shape of a prior set may be very difficult to elicit, it may not even be that relevant for a specific inference. A further general issue seems unavoidable in the generalised Bayesian setting as developed here, namely the dual role of $n^{(0)}$. On the one hand, $n^{(0)}$ governs the weighting of prior information $y^{(0)}$ with respect to the data s/n, as mentioned in Section 3.5.2.1: The larger $n^{(0)}$, the more $\underline{P}$ and $\overline{P}$ are dominated by $\underline{y}^{(0)}$ and $\overline{y}^{(0)}$. On the other hand, $n^{(0)}$ governs also the degree of posterior imprecision: the larger $n^{(0)}$, the larger c.p. Δ. A larger $n^{(0)}$ thus leads to more imprecise posterior inferences, although a high weight on the supplied prior information should boost the trust in posterior inferences if s in the TPDA, i.e. the prior information turned out to be appropriate. In the next section, we thus develop a different approach separating these two roles: Now, two separate models for predictive inference, each resulting in different precision as governed by $n^{(0)}$, are combined with an imprecise weight α taking the role of regulating prior-data agreement.

3.5.3. Weighted Inference

We propose a variation of the Beta-Binomial model that is attractive for prior-data conflict and has small yet fascinating differences with the models in Sections 3.5.2.2 and 3.5.2.3. We present a basic version of the model in Section 3.5.3.1, followed by an extended version in Section 3.5.3.2. Opportunities to generalise the model are mentioned in Section 3.5.3.3.

3.5.3.1. The Basic Model

The idea for the proposed model is to combine the inferences based on two models, each part of an imprecise Bayesian inferential framework using sets of prior distributions, although the inferences can also result from alternative inferential methods. The combination is not achieved by combining the two sets of prior distributions into a single set, but by combining the posterior predictive *inferences* by imprecise weighted averaging. When the weights assigned to the two models can vary over the whole range $[0,1]$ we actually return to imprecise Bayesian inference with a prior set, as considered in this subsection. In Section 3.5.3.2 we restrict the values of the model weights. The basic model turns out to be relevant from many perspectives, in particular to highlight similarities and differences with the methods presented in Sections 3.5.2.2 and 3.5.2.3, and it is a suitable starting point for more general models. These aspects will be discussed in Subsection 3.5.3.3.

We consider the combination of the imprecise posterior predictive probabilities $[\underline{\mathrm{P}}^i, \overline{\mathrm{P}}^i]$ and $[\underline{\mathrm{P}}^u, \overline{\mathrm{P}}^u]$ for the event that the next observation is a success with

$$\underline{\mathrm{P}}^i = \frac{s^i + s}{n^i + n + 1} \qquad \text{and} \qquad \overline{\mathrm{P}}^i = \frac{s^i + s + 1}{n^i + n + 1}\,, \tag{3.23}$$

$$\underline{\mathrm{P}}^u = \frac{s}{n+1} \qquad \text{and} \qquad \overline{\mathrm{P}}^u = \frac{s+1}{n+1}\,. \tag{3.24}$$

The superscript i indicates 'informative', in the sense that these lower and upper probabilities relate to an 'informative' prior distribution reflecting prior beliefs of similar value as s^i successes in n^i observations. The superscript u indicates 'uninformative', which can be interpreted as absence of prior beliefs. These lower and upper probabilities can for example result from Walley's IBBM, with $\underline{\mathrm{P}}^i$ and $\overline{\mathrm{P}}^i$ based on the prior set with $n^{(0)} = n^i + 1$ and $y^{(0)} \in \left[\frac{s^i}{n^i+1}, \frac{s^i+1}{n^i+1}\right]$, and $\underline{\mathrm{P}}^u$ and $\overline{\mathrm{P}}^u$ on the prior set with $n^{(0)} = 1$ and $y^{(0)} \in [0,1]$. There are other methods for imprecise statistical inference that lead to these same lower and upper probabilities, including Nonparametric Predictive Inference for Bernoulli quantities (Coolen 1998)[96], where the s^i and n^i would only be included if they were actual observations, for example resulting from a second data set that one may wish to include in the 'informative' model but not in the 'uninformative' model.

The proposed method combines these lower and upper predictive probabilities by imprecise weighted averaging. Let $\alpha \in [0,1]$, we define

$$\underline{\mathrm{P}}_\alpha = \alpha\underline{\mathrm{P}}^i + (1-\alpha)\underline{\mathrm{P}}^u, \qquad \overline{\mathrm{P}}_\alpha = \alpha\overline{\mathrm{P}}^i + (1-\alpha)\overline{\mathrm{P}}^u, \tag{3.25}$$

[96] See the short description in Section 2.1.4, and for more resources also `www.npi-statistics.com`.

and as lower and upper predictive probabilities for the event that the next Bernoulli random quantity is a success,[97]

$$\underline{\mathrm{P}} = \min_{\alpha\in[0,1]} \underline{\mathrm{P}}_\alpha \qquad \text{and} \qquad \overline{\mathrm{P}} = \max_{\alpha\in[0,1]} \overline{\mathrm{P}}_\alpha \,.$$

Allowing α to take on any value in $[0,1]$ reduces this method to the IBBM with a single prior set, as discussed in Section 3.5.2, with the prior set simply generated by the union of the two prior sets for the 'informative' and the 'uninformative' models as described above. For all s these minimum and maximum values are obtained at either $\alpha = 0$ or $\alpha = 1$. With switch points $S_1 = (n+1)\frac{s^i}{n^i} - 1$ and $S_2 = (n+1)\frac{s^i}{n^i}$, they are equal to

$$\underline{\mathrm{P}} = \begin{cases} \underline{\mathrm{P}}^u = \frac{s}{n+1} & \text{if } s \leq S_2 \\ \underline{\mathrm{P}}^i = \frac{s^i+s}{n^i+n+1} & \text{if } s \geq S_2 \end{cases}, \qquad \overline{\mathrm{P}} = \begin{cases} \overline{\mathrm{P}}^i = \frac{s^i+s+1}{n^i+n+1} & \text{if } s \leq S_1 \\ \overline{\mathrm{P}}^u = \frac{s+1}{n+1} & \text{if } s \geq S_1 \end{cases}.$$

The PPP graph for this model is displayed in Figure 3.14. Note that the lower probability $\underline{\mathrm{P}}$ is made up of two line segments, one from $s = 0$ to $s = S_2$ and a second line segment, with smaller slope, for the larger values until $s = n$. The upper probability $\overline{\mathrm{P}}$ is also made up of two line segments, one from $s = 0$ to $s = S_1$ and a second line segment, with larger slope, for the larger values until $s = n$. These switch points are the same for the special cases of the weighted inference method as discussed in detail in this section.

The upper probability for $s = S_1$ and the lower probability for $s = S_2$ are both equal to $\frac{s^i}{n^i}$. The TPDA contains only a single possible value of s (except if S_1 and S_2 are integer), namely the one that is nearest to $\frac{s^i}{n^i}$. The specific values for this basic case are

$$\begin{aligned} A &= 0 & B &= \frac{s^i+1}{n^i+n+1} & C &= \frac{s^i+n}{n^i+n+1} \\ D &= 1 & E &= \frac{s^i}{n^i} - \frac{1}{n+1} & F &= \frac{s^i}{n^i} + \frac{1}{n+1} \\ & & \text{sl. } 1 &= \frac{1}{n^i+n+1} & \text{sl. } 2 &= \frac{1}{n+1}\,. \end{aligned}$$

If s is in the TPDA, it reflects optimal agreement of the 'prior data' (n^i, s^i) and the (really observed) data (n, s), so it may be a surprise that both the lower and upper probabilities in this case correspond to $\alpha = 0$, so they are fully determined by the 'uninformative' part of the model. This is an important aspect, it will be discussed in more detail and compared to the methods of Section 3.5.2 in Subsection 3.5.3.3. For s in the TPDA, both $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ increase with slope $\frac{1}{n+1}$, and imprecision $\Delta = \frac{1}{n+1}$.

Figure 3.14, with the specific values for this basic case given above, illustrates what happens for values of s outside this TPDA. Moving away from the TPDA in either direction, the imprecision increases, as was also the case in the models in Section 3.5.2. For s decreasing towards 0, this is effectively due to the smaller slope of the upper probability,

[97] While in (3.20) and (3.21), prior and sample information are imprecisely weighted, here informative and uninformative models are combined.

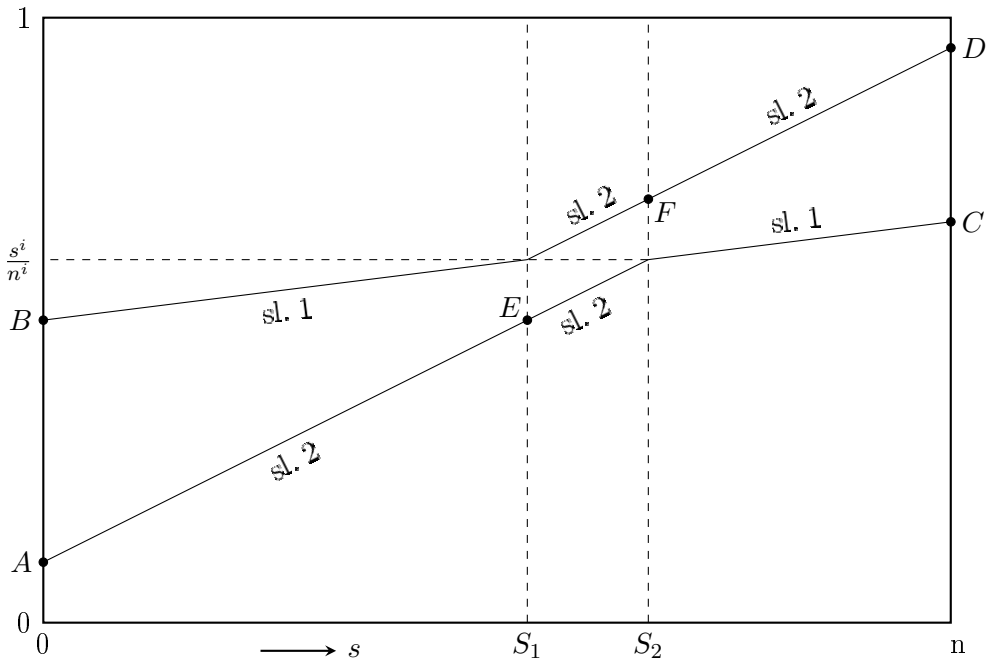

Figure 3.14.: $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ for the weighted inference model.

while for s increasing towards 1 it is due to the smaller slope of the lower probability. For $s \in [0, S_1]$, the imprecision is $\Delta = \frac{s^i+1}{n^i+n+1} - \frac{sn^i}{(n^i+n+1)(n+1)}$. For $s \in [S_2, n]$, the imprecision is $\Delta = \frac{1}{n+1} - \frac{s^i}{n^i+n+1} + \frac{sn^i}{(n^i+n+1)(n+1)}$. For the two extreme possible cases of prior data conflict, with either $s^i = n^i$ and $s = 0$ or $s^i = 0$ and $s = n$, the imprecision is $\Delta = \frac{n^i+1}{n^i+n+1}$. For this combined model with $\alpha \in [0, 1]$, we have $\underline{\mathrm{P}} \leq \frac{s}{n} \leq \overline{\mathrm{P}}$ for all s, which is attractive from the perspective of objective inference.

There are important further aspects of this model to be discussed, in particular how to meaningfully choose s^i and n^i, but this is postponed until Section 3.5.3.3, where we will start discussing these aspects by again focussing first on this basic model followed by discussion of an extended version of this model, which is introduced next.

3.5.3.2. The Extended Model

We extend the basic model from Subsection 3.5.3.1, perhaps remarkably by reducing the interval for the weighting variable α. We assume that $\alpha \in [\alpha_l, \alpha_r]$ with $0 \leq \alpha_l \leq \alpha_r \leq 1$. We consider this an extended version of the basic model as there are two more parameters that provide increased modelling flexibility. It is important to remark that, with such a restricted interval for the values of α, this weighted model is no longer identical to an IBBM with a single set of prior distributions. One motivation for this extended model is that the basic model seemed very cautious by not using the informative prior part if s is in the TPDA. For $\alpha_l > 0$, the informative part of the model influences the inferences for all

values of s, including the one in the TPDA. As a consequence of taking $\alpha_l > 0$, however, the line segment $(s, \frac{s}{n})$ with $s \in [0, n]$ will not always be in between the lower and upper probabilities anymore, specifically not at, and close to, $s = 0$ and $s = n$, as follows from the results presented below.

The lower and upper probabilities resulting from the two models that are combined by taking an imprecise weighted average are again as given by formulae (3.23)–(3.24), with the weighted averages $\underline{P}_\alpha$ and $\overline{P}_\alpha$, for any $\alpha \in [\alpha_l, \alpha_r]$, again given by (3.25). This leads to the lower and upper probabilities for the combined inference

$$\underline{P} = \min_{\alpha \in [\alpha_l, \alpha_r]} \underline{P}_\alpha \qquad \text{and} \qquad \overline{P} = \max_{\alpha \in [\alpha_l, \alpha_r]} \overline{P}_\alpha \,.$$

The lower and upper probabilities have, as function of s, the generic forms presented in Figure 3.14, with $[S_1, S_2] = \left[(n+1)\frac{s^i}{n^i} - 1, (n+1)\frac{s^i}{n^i}\right]$ as in Section 3.5.3.1. The specific values for Figure 3.14 are

$$\begin{aligned} A &= \tfrac{\alpha_l s^i}{n^i+n+1} & B &= \tfrac{1}{n+1} + \tfrac{\alpha_r[s^i(n+1)-n^i]}{(n^i+n+1)(n+1)} \\ D &= 1 - \tfrac{\alpha_l(n^i-s^i)}{n^i+n+1} & C &= \tfrac{n}{n+1} - \tfrac{\alpha_r[(n^i-s^i)(n+1)-n^i]}{(n^i+n+1)(n+1)} \\ \text{sl. } 1 &= \tfrac{n^i+n+1-\alpha_r n_i}{(n^i+n+1)(n+1)} & E &= \tfrac{s^i}{n^i} - \tfrac{1}{n+1}\left[1 - \tfrac{\alpha_l n^i}{n^i+n+1}\right] \\ \text{sl. } 2 &= \tfrac{n^i+n+1-\alpha_l n_i}{(n^i+n+1)(n+1)} & F &= \tfrac{s^i}{n^i} + \tfrac{1}{n+1}\left[1 - \tfrac{\alpha_l n^i}{n^i+n+1}\right] . \end{aligned}$$

The increase in imprecision when s moves away from the TPDA can again be considered as caused by the informative part of the model, which is logical as the uninformative part of the model cannot exhibit prior-data conflict. Maximum prior-data conflict occurs again if $s^i = 0$ and $s = n$, in which case $\underline{P} = \frac{n}{n+1} - \frac{\alpha_r n^i n}{(n^i+n+1)(n+1)}$ and $\overline{P} = 1 - \frac{\alpha_l n^i}{n^i+n+1}$, or if $s^i = n^i$ and $s = 0$, when $\underline{P} = \frac{\alpha_l n^i}{n^i+n+1}$ and $\overline{P} = \frac{1}{n+1} + \frac{\alpha_r n^i n}{(n^i+n+1)(n+1)}$.

The possibility to choose values for α_l and α_r provides substantially more modelling flexibility compared to the basic model presented in Section 3.5.3.1. One may, for example, wish to enable inferences solely based on the informative part of the model, hence choose $\alpha_r = 1$, but ensure that this part has influence on the inferences in all situations, with equal influence to the uninformative part in case of TPDA. This latter aspect can be realized by choosing $\alpha_l = 0.5$. When compared to the situation in Section 3.5.3.1, this choice moves, in Figure 3.14, A and D away from 0 and 1, respectively, but does not affect B and C. It also brings E and F a bit closer to the corresponding upper and lower probabilities, respectively, hence reducing imprecision in the TPDA.

3.5.3.3. Weighted Inference Model Properties

The basic model presented in Section 3.5.3.1 fits in the Bayesian framework, but its use of prior information is different to the usual way in Bayesian statistics. The lower and upper probabilities are mainly driven by the uninformative part, which, e.g., implies that $\underline{P} \leq \frac{s}{n} \leq \overline{P}$ for all values of s. While in (imprecise, generalised) Bayesian statistics any

part of the model that uses an informative prior can be regarded as adding information to the data, the informative part of the basic model leads to more careful inferences when there is prior-data conflict. Figure 3.14 shows that, for the basic case of Section 3.5.3.1, the points A and D are based only on the uninformative part of the model, but the points B and C are based on the informative part of the model.

Prior-data conflict can be of different strength, one would expect to only talk about 'conflict' if consideration is required, hence the information in the prior and in the data should be sufficiently strong. The proposed method in Section 3.5.3.1 takes as starting point inference that is fully based on the data, it uses the informative prior part of the model to widen the interval of lower and upper probabilities in the direction of the value $\frac{s^i}{n^i}$. For example, if one observed $s = 0$, the upper probability of a success at the next observation is equal to $\frac{s^i+1}{n^i+n+1}$, which reflects inclusion of the information in the prior set for the informative part of the model that is most supportive for this event, equivalent to $s^i + 1$ successes in $n^i + 1$ observations. As such, the effect of the prior information is to weaken the inferences by increasing imprecision in case of prior-data conflict.

One possible way in which to view this weighted inference model is as resulting from a multiple expert or information source problem, where one wishes to combine the inferences resulting individually from each source. The basic model of Section 3.5.3.1 leads to the most conservative inference such that no individual model or expert disagrees, while the restriction on weights provides a guaranteed minimum level for the individual contributions to the combined inference.

It should be emphasized that the weighted inference model has wide applicability. The key idea is to combine, by imprecise weighting, the actual inferences resulting from multiple models, and as such there is much scope for the use and further development of this approach. The individual models could even be models such as those described in Sections 3.5.2.2 and 3.5.2.3, although that would lead to more complications. If the individual models are coherent lower and upper probabilities, i.e. provide separately coherent inferences, then the combined inference via weighted averaging and taking the lower and upper envelopes is also separately coherent.[98]

In applications, it is often important to determine a sample size (or more general design issues) before data are collected. If one uses a model that can react to prior-data conflict, this is likely to lead to a larger data requirement. One very cautious approach is to choose n such that the maximum possible resulting imprecision does not exceed a chosen threshold. In the models presented here, this maximum imprecision will always occur for either $s = 0$ or $s = n$, whichever is further away from the TPDA. In such cases, a preliminary study has shown an attractive feature if one can actually sample sequentially. If some data are obtained with success proportion close to s^i/n^i, the total data requirement (including these first observations) to ensure that the resulting maximum imprecision cannot exceed the same threshold level is substantially less than had been the case before any data were available. This would be in line with intuition, and further research towards this and related aspects seems promising, including of course the further data need in case first

[98] This follows, e.g., from Walley (1991, §2.6.3f).

sampled data is in conflict with (n^i, s^i), and the behaviour of the models of Section 3.5.2 in such cases.

The weighted inference method combines the inferences based on two models, and can be generalised to allow more than two models and different inferential methods. It is also possible to allow more imprecision in each of the models that are combined, leading to more parameters in the overall model that can be used to control the behaviour of the inferences. Similar post-inference combination via weighted averaging, but with precise weights, has been presented in the frequentist statistics literature (Hjort and Claeskens 2003; Longford 2003), where the weights are actually determined based on the data and a chosen optimality criterion for the combined inference. In Bayesian statistics, estimation or prediction inferences based on different models can be similarly combined using Bayes factors (Kass and Raftery 1995), which are based on both the data (via the likelihood function) and prior weightings for the different models.[99] In our approach, we do not use the data or prior beliefs about the models to derive precise weights for the models, instead we cautiously base our combined lower and upper predictive probabilities on those of the individual models with a range of possible weights. This range is set by the analyst and does not explicitly take the data or prior beliefs into account, but it provides flexibility with regard to the relative importance given to the individual models.

3.5.4. Insights and Challenges

We have discussed two different classes of inferential methods to handle prior-data conflict in the Bernoulli case. These can be generalised to the multinomial case corresponding to the IDM. It also seems possible to extend the approaches to continuous sampling models like the normal or the gamma distribution, by utilizing the fact that the basic form of the updating of $n^{(0)}$ and $y^{(0)}$ in (3.19) underlying (3.20) and (3.21) is valid for arbitrary canonical exponential families (see Quaeghebeur and Cooman 2005; Walter and Augustin 2009b, and Section 3.1). Further insight into the weighting method may also be provided by comparing it to generalised Bayesian analysis based on sets of conjugate priors consisting of nontrivial mixtures of two Beta distributions. There, however, the posterior mixture parameter depends on the other parameters. For a deeper understanding of prior-data conflict, it may also be helpful to extend our methods to coarse data, in an analogous way to Utkin and Augustin (2007) and Troffaes and Coolen (2009), and to look at other model classes of prior distributions, most notably at contamination neighbourhoods. Of particular interest here may be to combine both types of prior models, considering contamination neighbourhoods of our exponential family based-models with sets of parameters, as developed in the Neyman-Pearson setting by Augustin (2002, § 5).

The models presented here address prior-data conflict in different ways, either by fully utilizing the prior information in a way that is close to the traditional Bayesian method, where this information is added to data information, or by not including them initially

[99] See also the brief characterisation of Bayes factors in the context of hypotheses testing and model selection in Section 1.2.3.3.

as in Section 3.5.3. All these models show the desired increase of imprecision in case of prior-data conflict. It may be of interest to derive methods that explicitly respond to (perhaps surprisingly) strong prior-data agreement.[100] One possibility to achieve this with the methods presented here is to consider the TPDA as this situation of strong agreement in which one wants imprecision reduced further than compared to an 'expected' situation, and to choose the prior set (Section 3.5.2) or the two inferential models (Section 3.5.3) in such a way to create this effect. This raises interesting questions for elicitation, but both approaches provide opportunities for this and we consider it as an important topic for further study.

Far beyond further extensions one has, from the foundational point of view, to be aware that there are many ways in which people might react to prior-data conflict, and we may perhaps at best hope to catch some of these in a specific model and inferential method. This is especially important when the conflict is very strong, and indeed has to be considered as full contradiction of modeling assumptions and data, which may lead to a revision of the whole system of background knowledge in the light of surprising observations, as Hampel argues.[101] In this context, applying the weighting approach to the NPI-based model for categorical data (Coolen and Augustin 2009) may provide some interesting opportunities, as it explicitly allows to consider not yet observed and even undefined categories (Coolen and Augustin 2005).

There is another intriguing way in which one may react to prior-data conflict, namely by considering the combined information to be of less value than either the real data themselves or than both information sources. Strong prior beliefs about a high success rate could be strongly contradicted by data, as such leading to severe doubt about what is actually going on. The increase of imprecision in case of prior-data conflict in the methods presented in this paper might be interpreted as reflecting this, but there may be other opportunities to model such an effect. It may be possible to link these methods to some popular approaches in frequentist statistics, where some robustness can be achieved or where variability of inferences can be studied by round robin deletion of some of the real observations. This idea may open up interesting research challenges for imprecise probability models, where the extent of data reduction could perhaps be related to the level of prior-data conflict. Of course, such approaches would only be of use in situations with substantial amounts of real data, but as mentioned before, these are typically the situations where prior-data conflict is most likely to be of sufficient relevance to take its modelling seriously. As (imprecise, generalised) Bayesian methods all work essentially by *adding* information to the real data, it is unlikely that such new methods can be developed within the Bayesian framework, although there may be opportunities if one restricts the inferences to situations where one has at least a pre-determined number of observations to ensure that posterior distributions are proper. For example, one could consider allowing the prior strength parameter $n^{(0)}$ in the IBBM to take on negative values, opening up a rich field for research and discussions.

[100] See our suggestion for a parameter set shape that allows such a behaviour in Section 4.3 and Section A.2.
[101] See in particular the discussion of the structure and role of background knowledge in Hampel (2009a).

4. Concluding Remarks

We will conclude this thesis with a short summary of, and discussion of the central achievments in, this thesis in Sections 4.1 and 4.2, respectively. In Section 4.3, we will sketch some opportunities for applications and avenues for further research.

4.1. Summary

After a detailed motivating example and application in Section 1.3, we presented some theoretical foundations of imprecise or interval probability in Section 2, with a focus on the approaches leading the way towards a generalisation of Bayesian inference. There, we also gave some general motives for using imprecise probability methods, afterwards focussing again on the Bayesian setting, where a foundational motive[1] and two specific motives were described: the case of weakly, or (near-) non-informative priors, and the issue of *prior-data conflict* (Section 2.2.3.3), where strong prior beliefs are in conflict with trusted data, but data is too sparse to overrule prior beliefs.

After a short view on some approaches understanding themselves as alternative to imprecise probability in Section 2.2.4, we proposed a general framework for imprecise Bayesian inference based on sets of conjugate priors in Section 3.1,[2] serving as a bracket and reference point for most of the models investigated in this thesis. Some general results regarding inference properties of models in this framework were presented (we will comment on these properties in the discussion below), and the two issues of weakly informative priors and prior-data conflict, problematic in usual Bayesian inference, could subsequently be considered as modelling opportunities of generalised Bayesian inference:

(i) Imprecise Bayesian methods allow to model weak prior information more adequately than the so-called non-informative priors usually employed (see item V. on page 60, and Section 3.1.3).

(ii) As we find that a common class of Bayesian inference procedures (those based on the canonical conjugates described in Section 1.2.3.1) provides insuffients inferences

[1] The Bayesian approach to inference requires an unattainable precision in prior assessments for the foundational arguments for its use (e.g., coherence) to be valid (see Section 2.2.3.1).

[2] Based on the parametrisation of canonically constructed conjugate priors described in Section 1.2.3.1, where a prior is identified through a strength parameter $n^{(0)}$ and a main parameter $y^{(0)}$, the framework considers sets of priors $\mathcal{M}^{(0)}$ induced by sets of parameters $\mathbb{\Pi}^{(0)}$, which consist of pairs $(n^{(0)}, y^{(0)})$. The set of posteriors $\mathcal{M}^{(n)}$, obtained by updating each distribution in the set of priors $\mathcal{M}^{(0)}$, can then be conveniently described through the set of posterior parameters $\mathbb{\Pi}^{(n)}$, consisting of pairs of updated parameters $(n^{(n)}, y^{(n)})$.

in case of prior-data conflict (see Section A.1.2 for some basic examples, and Section A.1.3 and A.1.4 for the case of Bayesian linear regression), we developed imprecise probability methods that overcome this deficiency, by mapping ambiguity in posterior inferences now also in dependence of prior-data conflict (see Sections 3.1.4, and specifically 3.3). In these models, a conflict of prior beliefs and data entails larger posterior parameter sets $\mathbb{\Pi}^{(n)}$, leading to cautious inferences if, and only if, caution is needed.

Section 3.2 considered some alternative models based on sets of priors, and discussed their inference properties in comparison to the results in Section 3.1. Section 3.3 then motivated, developed, and illustrated a central model type from Section 3.1 in more detail; afterwards, Section 3.4 briefly described a software implementation of this model framework.

Finally, Section 3.5 presented attempts to further refine inference behaviour in the presence of prior-data conflict. In a first approach, the model discussed in Section 3.3 was generalised by further tayloring the prior parameter set $\mathbb{\Pi}^{(0)}$. The second, fundamentally different, approach attained prior-data conflict sensitivity by combining *inferences* based on two different priors, an informative prior (expressing prior beliefs) and an uninformative prior (a near-ignorance prior), through an imprecise weighting scheme.

As supplemental material, the Appendix (Chapter A) contains

(i) a study of prior-data conflict sensitivity in Bayesian linear regression, presenting a simplified prior model that gives interesting insights into the updating step for the regression parameters and offers opportunities for inferences based on sets of priors (Section A.1);

(ii) some first technical results characterising a new prior parameter set shape, described informally in Section 4.3 below (Section A.2).

4.2. Discussion

As the model overview in Section 3.1 gives already a detailed discussion of the imprecise probability models considered in this thesis,[3] we will try to emphasize some central points here only.

The inference properties described in Section 3.1.2 for the generalised Bayesian model framework established in Section 3.1.1 are quite remarkable in their generality, and illustrate the ability of imprecise probability models to realistically model partial information

[3] General motives for the use of imprecise probability models were discussed in Section 2.2, with a focus on motives from a Bayesian perspective in Section 2.2.3. These motives can be subsumed as follows: Generalised Bayesian models allow, in contrast to classical Bayesian models, a realistic modeling of partial prior information, and can account for model uncertainty in a preferable way. Critique on, and possible alternatives to, Generalised Bayesian inference models were discussed in Section 2.2.4, while alternatives to the models covered in Section 3.1 are discussed in Section 3.2.

for a wide field of inference tasks. In these models, ambiguity in prior specifications influences ambiguity in posterior inferences in a natural and comprehensible way.[4]

Regarding the further model criterion 'prior-data conflict sensitivity', a central concept in this thesis, we concluded that the shape of the prior parameter set $\mathbb{\Pi}^{(0)}$ crucially influences model behaviour, and noted at the end of Section 3.1.4 that there is a clear trade-off between easy handling of $\mathbb{\Pi}^{(0)}$ or $\mathbb{\Pi}^{(n)}$, and the desired model property. Models with fixed prior strength parameter $n^{(0)}$, like the IDM under prior information and the model by Quaeghebeur and Cooman (2005), are very easy to handle, with $n^{(0)}$, the lower ($\underline{y}^{(0)}$) and upper ($\overline{y}^{(0)}$) bound of the prior main parameter as the only parameters to elicit, and $\mathbb{\Pi}^{(n)}$ being characterised again by the three values $n^{(n)}$, $\underline{y}^{(n)}$ and $\overline{y}^{(n)}$.[5] However, these models are insensitve to prior-data conflict, and the model with $\mathbb{\Pi}^{(0)} = [\underline{n}^{(0)}, \overline{n}^{(0)}] \times [\underline{y}^{(0)}, \overline{y}^{(0)}]$ presented in Section 3.3.4 mitigates this deficiency, but at the cost of a more complex shape of $\mathbb{\Pi}^{(n)}$, not being a cartesian product of $[\underline{n}^{(n)}, \overline{n}^{(n)}]$ and $[\underline{y}^{(n)}, \overline{y}^{(n)}]$ anymore. The more refined behaviour discussed in Section 3.5 is achieved by a more complex choice of $\mathbb{\Pi}^{(0)}$, where the range of $y^{(0)}$ values depends on $n^{(0)}$ via the contour functions $\underline{y}^{(0)}(n^{(0)})$ and $\overline{y}^{(0)}(n^{(0)})$ (see Section 3.5.2.3).

Revisiting and continuing the discussion of parameter set shapes in Section 3.5.2.4, the general framework in principle allows arbitrary shapes for the prior parameter set $\mathbb{\Pi}^{(0)}$. However, freely elicting this set shape (e.g., by allowing for arbitrary contour functions $\underline{y}^{(0)}(n^{(0)})$ and $\overline{y}^{(0)}(n^{(0)})$) from limited prior information might be very difficult. If only specific inferences are of interest, it is possible that only a few aspects of the shape are relevant, such that elicitation of the entire shape might actually not be necessary. Furthermore, the derivation of posterior inferences from such complex prior shapes can be difficult as well.

In general, models based on sets of canonical conjugate priors form, in our view, a 'sweet spot' in the realm of (imprecise) statistical models, by allowing sophisticated model behaviour combined with easy elicitation and computation. In particular, these models are distinguished by the following characteristics:

- They are relatively easy to elicit: there is a simple and straightforward interpretation of the model parameters $n^{(0)}$ and $y^{(0)}$ as prior strength and main parameter, respectively (see Sections 1.2.3.1 and 3.3.2). In general terms, the weighted average structure of the update step (1.6) and the model behaviour resulting from it is relatively clear.

- They are easy to apply in a variety of inference problems: it is possible to construct conjugate priors as needed for the problem at hand, often resulting in closed form solutions for many quantities of interest.

[4] In contrast, as mentioned in Section 3.2.3.2, in imprecise probability models for discrete parameter spaces, it is for example difficult to judge the weight of a certain prior model as compared to the data.

[5] For sake of a clarity, we consider only one-dimensional main parameters $y^{(0)}$ in the discussion here and below, although the model framework allows for multidimensional $y^{(0)}$, as illustrated by the examples involving the Dirichlet-Multinomial model.

- Nevertheless, they give reasonable and sophisticated inferences: as discussed above, uncertainty is adequately mirrored by the size of the set of posteriors $\mathcal{M}^{(n)}$, fulfilling the inference properties I.–III. with the potential to implement prior-data conflict sensitivity or near-noninformativeness, as required by the inference task at hand.[6]

- They allow for flexible modelling, as it is possible to choose or tweak the prior set shape according to inference needs, where, however, the trade-off mentioned above should be taken into account. Specifically, we think the model presented in Section 3.3.4 (with $\mathbb{\Pi}^{(0)} = [\underline{n}^{(0)}, \overline{n}^{(0)}] \times [\underline{y}^{(0)}, \overline{y}^{(0)}]$)[7] constitutes a sensible compromise between model complexity and inference properties.

However, some doubts with respect to the rectangular shape of $\mathbb{\Pi}^{(0)}$ are nevertheless permitted. From a strictly behavioral point of view (e.g., if $\mathcal{M}^{(0)}$ should express an expert's prior beliefs), taking $\mathbb{\Pi}^{(0)} = [\underline{n}^{(0)}, \overline{n}^{(0)}] \times [\underline{y}^{(0)}, \overline{y}^{(0)}]$, is, as we mentioned in Sections 3.1.4 and 3.5.2.4, a somewhat peculiar choice, as it entails that we assign the same interval for the main parameter $y^{(0)}$ over a range of prior strengths.

Instead, it might be more reasonable to assume that the expert is able to give a more precise interval estimate for lower prior strengths, while being more cautious in his prior assignments for higher prior strengths by choosing a wider interval for $y^{(0)}$ at $\overline{n}^{(0)}$.[8]

However, it seems equally reasonable to choose a shorter interval for $y^{(0)}$ at $\overline{n}^{(0)}$ than at $\underline{n}^{(0)}$, by the considerations described at the beginning of Section 3.5.2.3, involving a thought experiment, or 'pre-posterior thinking': If we designate bounds for $y^{(n)}$ constant in $n^{(n)}$ and a range of $n^{(n)}$ values to express what we want to learn from certain hypothetical data, the corresponding $\mathbb{\Pi}^{(0)}$ derived by doing a 'backwards update' has, due to the update mechanism (1.6), a wider range for low values of $n^{(0)}$, and a narrower range for high $n^{(0)}$ values.[9] For an actual sample size smaller than the hypothetical one, this 'anteater' shape

[6] We think this is a major advantage to the alternative models discussed in Section 3.2, especially to models based on the density ratio class. As noted in Section 3.2.1.2, the magnitude of $\mathcal{M}_{l,u}$ does not decrease with n in density ratio class models that are updated according to the Generalised Bayes' Rule.

[7] This kind of prior set shape was also employed in the motivational example from Section 1.3, where we argued for an interval-valued $n^{(0)}$ at the end of Section 1.3.6.1, and demonstrated the merits of such a parameter shape for this specific situation of prior-data conflict in Section 1.3.6.2.

[8] As described in Section 3.5.2.4, such a choice can also lead to a more 'tolerant' behaviour in case of prior-data conflict. If, e.g., $\overline{y}^{(0)}(\underline{n}^{(0)}) < \overline{y}^{(0)}(\overline{n}^{(0)})$ was chosen and $\tilde{\tau}(\boldsymbol{x}) > \overline{y}^{(0)}(\overline{n}^{(0)})$, the higher weight $\overline{n}^{(0)}$ for $\overline{y}^{(0)}(\overline{n}^{(0)})$ in the update step would move $\overline{y}^{(n)}(\overline{n}^{(n)})$ more slowly towards $\tilde{\tau}(\boldsymbol{x})$ as compared to $\overline{y}^{(n)}(\underline{n}^{(n)})$, but for moderate degrees of prior-data conflict, the faster movement of $\overline{y}^{(n)}(\underline{n}^{(n)})$ would be offset by the 'head start' of $\overline{y}^{(0)}(\overline{n}^{(0)})$. As the upper bound for $y^{(n)}$ is attained at $\overline{y}^{(n)}(\overline{n}^{(n)})$ if $\tilde{\tau}(\boldsymbol{x}) < \overline{y}^{(0)}(\overline{n}^{(0)})$ (i.e., data is in agreement with the prior assignements), the effect of extra imprecision would only appear as soon as $\overline{y}^{(n)}(\underline{n}^{(n)})$ 'overtakes' $\overline{y}^{(n)}(\overline{n}^{(n)})$ in the move towards $\tilde{\tau}(\boldsymbol{x})$, i.e., when $\overline{y}^{(n)}(\underline{n}^{(n)}) > \overline{y}^{(n)}(\overline{n}^{(n)})$.

[9] For this model, to elicit the pre-posterior bounds, certain hypothetical data have to be chosen. Section 3.5.2.3 touches only very briefly on these hypothetical data (that are a part of the model parameters) and how to choose them, and there is ample opportunity for research here. One could also further modify the model by allowing to choose the 'pre-posterior $\mathbb{\Pi}^{(n)}$' more freely, with bounds for $y^{(n)}$ not constant in $n^{(n)}$.

is more 'tolerant' as the rectangular shape.[10] (A suggestion for another more sophisticated shape of $\mathbb{\Pi}^{(0)}$, with the aim to allow for extra precision if prior and data coincide especially well, is discussed in the outlook below.)

However, the favourable properties given in general terms in Section 3.1 should not obscure the restrictions that are imposed by the canonical conjugate priors the framework is based on.

The study by Krautenbacher (2011, see the discussion in Section 3.2.3.2) reminds us that fitting a prior parameter set $\mathbb{\Pi}^{(0)}$ to available prior information can be non-trivial, and that also the generally well-understood conjugate distributions may exhibit unintuitive inference behaviours when focussing on prior and posterior expectations only, as the model framework tempts us to do. It is thus advisable to not loose sight of the distributions in $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(n)}$ in their entirety, e.g., by considering unions of highest density intervals for different credibility levels.

The restrictions imposed by canonical conjugate priors can be mitigated, or even completely overcome, when choosing $\mathcal{M}^{(0)}$ not to contain parametric distributions only, but to comprise all finite mixtures of parametric distributions, i.e., considering $\mathcal{M}^{(0)}$ as the convex hull of the parametric distributions.[11] However, as mentioned in Section 3.1.1, then only inferences that are linear in the parametric posteriors are easy to obtain, and the model may deliver very imprecise, or even vacuous, results for nonlinear functions of $p(\vartheta \mid n^{(n)}, y^{(n)})$ like, e.g., the posterior variance.

It seems that models based on the density ratio class, the model framework most similar to this variant of the model framework of Section 3.1.1 (similar in that they both allow non-parametric priors in the set of prior distributions, although the model is generated by use of parametric distributions), do not have this issue of tractability of nonlinear inferences (see the discussion of the model by Rinderknecht (2011, §4) in Section 3.2.3.1); while this seems a major strength of the model, it is in our view offset by a fundamental weakness, the lack of a clear mechanism by which imprecision in coherent posterior inferences can be modelled in dependence of sample size.[12]

Another, more fundamental, handicap of the model framework from Section 3.1 is the double role of $n^{(0)}$ as mentioned at the end of Section 3.5.2.4. There, the issue is framed in terms of the imprecise Beta-Binomial model, but it is actually valid for the general case of updating in imprecise probability models based on canonical conjugate priors: On the

[10] The 'anteater' shapes are somewhat similar to shapes that would result if we required, for information on the main prior parameter $y^{(0)}$ symmetrical around 0, $|n^{(0)} \cdot y^{(0)}|$ to be constant. Such a shape was proposed by Benavoli and Zaffalon (2012) for canonical priors to one-parameter exponential family likelihoods, with the aim to generate near-noninformative prior sets $\mathcal{M}^{(0)}$.

[11] As noted in footnote 2, p. 57, if the parametric distributions are normal distributions and $\mathbb{\Pi}^{(0)}$ is large enough, it can be assumed that $\mathcal{M}^{(0)}$ contains a very wide range of priors, as mixtures of normal distributions are dense in the space of well-behaved probability distributions.

[12] As mentioned in Sections 3.2.1.2 and 3.2.3.1, imprecision as measured by the magnitude of $\mathcal{M}_{l,u}$ is the same for any sample size n if $\mathcal{M}_{l,u}$ is updated according to Bayes' Rule. Only if the requirement of coherence, the foundation of the Generalised Bayes' Rule, is dropped, density ratio class models are possible that allow for imprecision to depend on sample size n. See also footnote 38 on page 75, where we suggested a model based on a combination of ideas from Coolen (1993a) and Rinderknecht (2011).

one hand, $n^{(0)}$ governs the weighting of prior information $y^{(0)}$ with respect to the data $\tilde{\tau}(\boldsymbol{x})$; the larger $n^{(0)}$, the more the values of $\underline{y}^{(n)}$ and $\overline{y}^{(n)}$ are dominated by $\underline{y}^{(0)}$ and $\overline{y}^{(0)}$, respectively. On the other hand, $n^{(0)}$ governs also the degree of posterior imprecision: the larger $n^{(0)}$, the larger c.p. $\text{MPI}^{(n)} = \overline{y}^{(n)} - \underline{y}^{(n)}$. A larger $n^{(0)}$ thus leads to more imprecise posterior inferences, although a high weight on the supplied prior information should boost the trust in posterior inferences if $\tilde{\tau}(\boldsymbol{x}) \in [\underline{y}^{(0)}, \overline{y}^{(0)}]$, i.e., if prior beliefs turned out to be appropriate.

In Section 3.5.3, an approach separating these two roles of $n^{(0)}$ was developed, by considering two separate models, an uninformative and an informative model,[13] with individual levels of precision induced by different choices of $n^{(0)}$. Inference intervals based on these two models are then combined using an imprecise weight, forming the actual inference interval.

This model of *weighted inference* is a very general approach, being applicable for a wide variety of inference tasks and accomodating all sorts of models that provide interval-valued inferences. Combining favourable properties for the inference situation discussed in Section 3.5 with feasible elicitation and handling (see Section 3.5.3.3), it is an approach going beyond the model framework of Section 3.1.1, and thus the general properties described in Section 3.1.2 do not necessarily hold.

In summary, we think the model framework from Section 3.1.1 exploits the full expressive power of imprecise probability in a very elegant way, by allowing a realistic description and treatment of model uncertainty, and thus, as expressed in Section 2.2.4.1, enable us to avoid heroic model assumptions, and the spuriously precise inferences they entail.

We also think that there is still further potential to models based on sets of conjugate priors. As described in the Outlook below, further advances are within reach that, in addition to the properties I.–III. and prior-data conflict sensitivity (see Section 3.1.2), allow for more precise posterior inferences when prior and data coincide especially well.[14]

4.3. Outlook

Here, we will summarise central ideas discussed in several concluding sections (3.3.6, 3.5.4, A.1.5), carrying some of them further, presenting opportunities for applications and potential avenues for further research. In particular, an idea for modelling of *strong prior-data agreement* will be explained in more detail.

We will subsume our ideas for further research and developments in three steps: First, we will consider some potential applications and areas of study for the currently existing models in the framework from Section 3.1.1. Then, we will sketch some ideas to extend the presently available models, including a discussion of a novel parameter set shape that allows to cater for *strong prior-data agreement* (see the technical details in Section A.2).

[13] These two models could also be denoted by 'cautious' and 'bold', respectively. While the cautious model tries to use only a minimal amount of information, the bold model goes for more daring assumptions.

[14] We already mentioned this this modelling goal in Section 3.5.4, where we spoke of *strong prior-data agreement*.

These further developments still follow the lines of generalised Bayesian inference using sets of priors as described in Sections 2.1.2.5 and 2.1.3 that ensure coherence of inferences by updating the set of priors via the Generalised Bayes' Rule. In the last part of this outlook, we will look instead beyond this framework of coherent Bayesian inference, by discussing some general thoughts about updating and learning in the context of statistical inference.

As seen in Section 3.2, there are a number of alternative models for statistical inference using sets of priors. Besides the study by Krautenbacher (2011), no studies comparing inferences from the models described in Setions 3.3 and 3.5 with those based on the alternative models named in Section 3.2 (most importantly, models based on the density ratio class) in detail have yet been conducted to our knowledge. Such studies could be rewarding by giving more detailed insight into strengths and weaknesses of the respective models. Further insight could also be gained from a comparison to the hierarchical model developed by Cattaneo (2008, see Section 2.2.4.2).

There are of course many opportunities for application of the models described in this thesis, and we will mention just two interesting general cases here.

The canonical conjugate prior for Bayesian linear regression constructed in Section A.1.4 could be used for an imprecise regression analysis based on sets of priors. Although being probably the most important concept in modern statistical inference, in the imprecise probability literature, so far only very few contributions considering regression analysis have been published.[15] Work in this direction could contribute to a major step towards a wider application of imprecise probability models in statistical practice.[16] With priors based on $\mathbb{I}\mathrm{I}^{(0)}$ of type (c), the model could offer favourable inference properties in situations that are prone to prior-data conflict (see the discussion in Section A.1.5).

Another potentially fruitful area of application is models for *statistical surveillance* (for a brief overview see, e.g., Frisén 2011). Here, the aim is to monitor a data-generating process over time to detect changes in the process as early as possible, without generating too many 'false alarms'. A typical example is disease monitoring, where the number of cases of a certain infectious disease reported by clinicians is continuously monitored on the national level, with the aim to detect epidemic outbreaks in their early stages. Such outbreak detection is usually analysed using likelihood ratios, comparing new observations with a model derived from previous observations. As mentioned in footnote 10, page 61, we could also consider this problem in terms of prior-data conflict: if a new batch of m observations does not fit our current model (a set $\mathcal{M}^{(n)}$ subsuming prior information and previous data of size n), the posterior model, based on $\mathcal{M}^{(n+m)}$, resonates this by increased imprecision, triggering the alarm. We would expect rectangular parameter set shapes to perform quite well in such a task, but also other set shapes could prove useful here.

Now, we will present some ideas for further development within the framework described in Section 3.1.1, i.e., for models based on sets of canonical conjugate priors that are updated

[15] Among the few exceptions are Walter, Augustin, and Peters (2007), Utkin (2010), Utkin and Coolen (2011), Cattaneo and Wiencierz (2012), and Utkin and Wiencierz (2013).

[16] In contrast, classification models are a thriving subject area in imprecise probability. For a recent overview see, e.g., Corani et al. (2013).

via the Generalised Bayes' Rule; afterwards, we will also consider some potential approaches outside this framework.

As a first important development, the models discussed in this thesis rely on a precise sampling model, but the generalised Bayesian inference framework also allows for imprecise sampling models (see, in particular, Walley 1991, §8.5), also discussed under the notion of *likelihood robustness* (e.g., Shyamalkumar 2000) in the robust Bayesian framework, not idealising the sampling model in the same way as we ceased to idealise the prior model.

Another important area for further study concerns the case of multidimensional $y^{(0)}$. Elicitation of such high-dimensional parameter sets $\mathbb{\Pi}^{(0)}$ poses interesting challenges. The simple, 'hyper-rectangle' set suggested in Section 1.3.6, Equation (1.19), although an effective choice for the application considered there, might not be adequate in all circumstances. However, elicitation of more general sets $\mathcal{Y}^{(0)} \times [\underline{n}^{(0)}, \overline{n}^{(0)}]$, or even arbitrary subsets of $\mathcal{Y} \times \mathbb{R}_{>0}$, can be very difficult, and further inquiries into this problem could be rewarding, also in connection to the application of the prior constructed in Section A.1.4, as problems in regression typically involve many covariates.

Also, for $\mathbb{\Pi}^{(0)}$ as in Equation (1.19), prior-data conflict is mirrored by increased imprecision for each dimension separately (see also Example 3.8, Figure 3.5). It is an open question if other prior set shapes entail different consequences with respect to multidimensional prior-data conflict sensitivity. A possible approach to simplify elicitation in high-dimensional cases (that could also be useful for lower-dimensional $y^{(0)}$) is to not consider $\mathcal{N}^{(0)} = [\underline{n}^{(0)}, \overline{n}^{(0)}]$, but $\mathcal{N}^{(0)} = \{\underline{n}^{(0)}, \overline{n}^{(0)}\}$, i.e., taking only a pair of $n^{(0)}$ values. This would not only simplify elicitation, but also make computation of posterior inferences probably more feasible. However, similar to the discussion of the rectangular shape in Section 4.2 above, the behavioural implications of such a choice of parameter set would have to be studied in detail.

Before we will discuss more general arguments about learning and updating, we will now describe a novel idea for generating parameter prior sets $\mathbb{\Pi}^{(0)}$ that, in addition to prior-data conflict sensitivity, allow to mirror *strong prior-data agreement* (i.e., the case when prior and data coincide especially well) by increased posterior precision.

As mentioned in Sections 3.1.4 and 3.5.2.4, the shape of $\mathbb{\Pi}^{(0)}$ determines the shape of $\mathbb{\Pi}^{(n)}$ and, via $\mathcal{M}^{(n)}$, has a crucial influence on posterior inferences. In general lines, this influence is clear (see the studies on the behaviour of rectangular shapes in Sections 3.3 and 3.5), but it is nevertheless quite difficult to elicit more general set *shapes* and to ascertain their consequences.

Although the updating of the canonical parameters according to (1.6), i.e., the weighted average update step for $y^{(0)}$, and the increment step for $n^{(0)}$, seems very intuitive (and is central to the behaviour of the model, see the list of inference properties in Section 3.1.2), the shape change of $\mathbb{\Pi}^{(0)}$ to $\mathbb{\Pi}^{(n)}$ through this update step, and its effects on posterior inferences, is difficult to grasp. This is due to the fact that, while the update of a single coordinate $(n^{(0)}, y^{(0)})$ is a simple shift, this shift is different for each of the coordinates in a set; in fact, the shift of the $y^{(0)}$ component changes with both $n^{(0)}$ and $y^{(0)}$.[17] These

[17] As given by (1.6), for the $n^{(0)}$ coordinate holds $n^{(0)} \mapsto n^{(0)} + n$, while for the $y^{(0)}$ coordinate holds

different shifts for the coordinates within a set lead to the change of shape from $\mathbb{\Pi}^{(0)}$ to $\mathbb{\Pi}^{(n)}$.

As this shape change is so problematic for understanding of the update step, a different parametrisation of the canonical priors in which each coordinate has the same shift in updating would be advantageous, such that updating of parameter sets could be expressed as a shift of the entire set within the parameter space.

In fact, such a parametrisation has been developed by Miķelis Bickis (2011, personal communication), and he is currently preparing a manuscript elaborating the details of his findings. In this parametrisation, a canonical prior is represented by a coordinate $(\eta_0^{(0)}, \eta_1^{(0)})$, where $\eta_1^{(0)}$ replaces the main prior parameter $y^{(0)}$, while $\eta_0^{(0)}$ is just a shifted version of $n^{(0)}$.[18] In the parametrisation from Section 1.2.3.1, $y^{(0)}$ had the convenient property of being equal to $\mathrm{E}\left[\mathrm{E}[\tilde{\tau}(\boldsymbol{x}) \mid \psi] \mid n^{(0)}, y^{(0)}\right]$, giving the (prior) expectation, or a prior guess, for the mean sample statistic $\tilde{\tau}(\boldsymbol{x})$. Naturally, $\eta_1^{(0)}$ cannot have this property, but in the transformed space for the Beta-Binomial model, the points on rays emanating from the coordinate $(-2, 0)$ will give a constant expectation of $\mathrm{E}[\tilde{\tau}(\boldsymbol{x}) \mid \psi]$, such that interpretation of parameter sets in terms of $(\eta_0^{(0)}, \eta_1^{(0)})$ is still relatively easy. With the set shape unchanged by the update, tailoring shapes for desired inference behaviour is much easier in this representation. Indeed, Frank Coolen and the author of this thesis have devised a set shape that allows for both prior-data conflict sensitivity and more precise inferences in case of strong prior-data agreement, a behaviour deemed very desirable in Section 3.5.4 and in the discussion of the Generalised Bayes' Rule in Section 2.1.3.2. We are thus able to offer a solution to this issue that allows to remain within the generalised Bayesian framework by using the Generalised Bayes' Rule for updating.

As our preliminary studies show some very appealing results in case of the Beta-Binomial model, for which a possible parametrisation of our shape is discussed in Section A.2, we are confident that these encouraging results also hold for the Normal-Normal model and in the general case of canonical conjugates. A joint publication of Miķelis Bickis, Frank Coolen and the author of this thesis is planned that will elaborate these findings in more detail.

Concluding the outlook, we will now turn to more general thoughts about updating and learning in the context of statistical inference using sets of priors.

As we wrote in Section 3.3.6, it should well be remembered that the models considered in this thesis consciously confined the whole argumentation to a certain Bayesian setting: we studied how far one can go *if* one relies strictly on the Generalised Bayes' Rule, transferring sets of priors to sets of posteriors element by element via Bayes' Rule (1.3). Considering the criticisms regarding the Generalised Bayes' Rule discussed in Section 2.1.3.2, alternative learning rules could thus provide superior model behaviour.[19]

$y^{(0)} \mapsto \frac{n^{(0)}y^{(0)}+\tau(\boldsymbol{x})}{n^{(0)}+n} = y^{(0)} + \frac{\tau(\boldsymbol{x})-ny^{(0)}}{n^{(0)}+n}$.

[18] This parametrisation is described in more detail in Section A.2, along with some first technical results that confirm the desired properties for the novel shape suggested there.

[19] In Section 3.2, we discussed some of the approaches mentioned in footnote 51, page 80, that consider alternative learning rules.

However, we argue that some aspects of inferences based on the Generalised Bayes' Rule that are critizised in the literature can be confronted, or at least mitigated, by a careful choice of the sets of priors.

We consider this, to some extent, to be the case for the critique that the Generalised Bayes' Rule is too inflexible when prior information is confronted with surprising data, as it insists on coherence of posterior inferences with prior assumptions that, in the light of the data, may turn out to be inadequate. The models discussed in Sections 3.1.4 and 3.3.4 mitigate this issue by allowing for prior-data conflict sensitivity. The resulting posterior sets are still, in essence, a compromise between the prior set and the data, but mirror the conflict between them by increased imprecision. This increased imprecision can then serve as a 'warning light', highlighting the doubts with regards to the posterior model in such a conflict, which may ultimatively motivate the analyst to reconsider her prior assignments.

Another critique based on the perceived inflexibility of the Generalised Bayes' Rule relates to posterior inferences being often 'too imprecise'. For the case of weak prior information, this has been confronted by approaches by Walley (1996a), Bickis (2009), Benavoli and Zaffalon (2012), and, most recently, by Mangili and Benavoli (2013). Although starting from a set of near-noninformative priors, these approaches lead to reasonably precise posterior inferences. Our approach for a novel parameter set $\Pi^{(0)}$, described above and in Section A.2, may be able to fend off this criticism also for informative priors by offering especially precise inferences when prior information and data are in accordance.

Other critical aspects of the Generalised Bayes' Rule are more difficult to tackle. Especially the caveat by Augustin (2003), showing that the decision theoretic justification of Bayes' Rule as producing prior risk minimising decision functions does not extend to the case of sets of priors, should motivate us to look beyond the Generalised Bayes' Rule. The issue here is that the Generalised Bayes' Rule may not lead to optimal inference procedures, as the optimality of the corresponding decision functions is not guaranteed.[20] From this angle, the desire for alternative learning rules gains a more solid footing in our view.

Indeed, we already considered some models going beyond the generalised Bayesian framework. Apart from our suggestion in footnote 38, page 75, that could lead to an interesting density ratio class model combining sophisticated elicitation with reasonable handling of imprecision, we think that a very attractive approach to inference where prior information can be included in the reasoning is the model by Cattaneo (2008, see Section 2.2.4.2). Along the lines of Antonucci, Cattaneo, and Corani (2011), it would even be possible to apply the hierarchical modelling to the framework of Section 3.1.1.

An approach beyond the Generalised Bayes' Rule that we studied in more detail is the weighting model from Section 3.5.3. As mentioned in Section 4.2 above, it was motivated by the dual role of $n^{(0)}$ in the framework from Section 3.1.1, controlling both posterior precision and deviation from prior assignments that can lead to unintuitive results in case of strong prior data agreement (see Section 3.5.2.4). Another idea for models outside the generalised Bayesian framework related to the role of $n^{(0)}$ was discussed in Section 3.5.4, where we

[20] For a brief overview on decision theory in the context of imprecise probability methods, see Huntley, Hable, and Troffaes (2013).

argued that a possible way to react to prior-data conflict could be to consider the combined information of prior and data to be of less value than either the data themselves or than both information sources separately. When strong prior beliefs collide with contradicting data, this could lead to severe doubt about what is actually going on. To model this behaviour, one could consider allowing the prior strength parameter $n^{(0)}$ to take on negative values, opening up a rich field for research and discussions.

We also hope that further studies like, e.g., those in temporal coherence as mentioned in Section 2.1.3.2, refining the concept of coherence towards allowing very substantial revisions in case of surprising data, will pave the way for models resonating the reasoning of Hampel (2009a; 2011), who argues that in order to represent learning, our models must allow to revise the whole system of background knowledge in the light of surprising observations.

A. Appendix

A.1. Bayesian Linear Regression: Different Conjugate Models and Their (In)Sensitivity to Prior-Data Conflict

This section reproduces the work "Bayesian Linear Regression — Different Conjugate Models and Their (In)Sensitivity to Prior-Data Conflict", published as technical report no. 69 at the Department of Statistics of Ludwig-Maximilians-University Munich (LMU) (Walter and Augustin 2009a). This technical report is a substantially extended version of a contribution to "Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir" (Walter and Augustin 2010). As such, it is reproduced here almost verbatim, except for some minor rewording, and the addition of some comments and footnotes linking this work to other parts of this thesis. Furthermore, the notation was changed slightly towards the one introduced in Section 1.2.3.1 (most importantly, denoting posterior parameters with upper index $^{(n)}$ instead of $^{(1)}$, e.g., writing $y^{(n)}$ instead of $y^{(1)}$), and citations were updated and changed towards the style employed throughout this thesis.

Here, two approaches for Bayesian linear regression modeling based on conjugate priors are considered in detail, namely the standard approach, described, e.g., in Fahrmeir et al. (2013), and an alternative adoption of the general construction procedure for exponential family sampling models. We recognize that — in contrast to some standard i.i.d. models like the scaled normal model and the Beta-Binomial or Dirichlet-Multinomial model, where prior-data conflict is completely ignored — the models may show some reaction to prior-data conflict, however in a rather unspecific way. Finally we briefly sketch the extension to a corresponding imprecise probability model, where, by considering sets of prior distributions instead of a single prior, prior-data conflict can be handled in a very appealing and intuitive way.

A.1.1. Introduction

Regression analysis is a central tool in applied statistics that aims to answer the omnipresent question how certain variables (called covariates, confounders, regressors, stimulus or independent variables, here denoted by $\boldsymbol{x}$) influence a certain outcome (called response or dependent variable, here denoted by z). Due to the complexity of real-life data situations, basic linear regression models, where the expectation of the outcome z_i simply equals the linear predictor $\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}$, have been generalized in numerous ways, ranging from

generalised linear models (Fahrmeir and Tutz 2001, see also Fahrmeir and Kaufmann 1985 for classical work on asymptotics) for non-normal distributions of $z_i \mid \boldsymbol{x}_i$, or linear mixed models allowing the inclusion of clustered observations, over semi- and nonparametric models (Kauermann, Krivobokova, and Fahrmeir 2009; Fahrmeir and Raach 2007; Scheipl and Kneib 2009), up to generalised additive (mixed) models and structured additive regression (Fahrmeir and Kneib 2009; Fahrmeir and Kneib 2006; Kneib and Fahrmeir 2007).

Estimation in such highly complex models may be based on different estimation techniques such as (quasi-) likelihood, general estimation equations (GEE) or Bayesian methods. Especially the latter offer in some cases the only way to attain a reasonable estimate of the model parameters, due to the possibility to include some sort of prior knowledge about these parameters, for instance by "borrowing strength" (e.g., Higgins and Whitehead 1996).

The tractability of large scale models with their ever increasing complexity of the underlying models and data sets should not obscure that still many methodological issues are a matter of debate. Since the early days of modern Bayesian inference one central issue has, of course, been the potentially strong dependence of the inferences on the prior. In particular in situations where data is scarce or unreliable, the actual estimate obtained by Bayesian techniques may rely heavily on the shape of prior knowledge, expressed as prior probability distributions on the model parameters. Recently, new arguments came into this debate by new methods for detecting and investigating *prior-data conflict* (Evans and Moshonov 2006; Bousquet 2008), i.e. situations where "[...] the observed data is surprising in the light of the sampling model and the prior, [so that] we must be at least suspicious about the validity of inferences drawn." (Evans and Moshonov 2006, p. 893)[1]

The present contribution investigates the sensitivity of inferences on potential prior-data conflict: What happens in detail to the posterior distribution, and the estimates derived from it, if prior knowledge and what the data indicates are severely conflicting? If the sample size n is not sufficiently large to discard the possibly erroneous prior knowledge and thus to rely on data only, prior-data conflict should affect the inference and should — intuitively and informally — result in an increased degree of uncertainty in posterior inference. Probably most statisticians would thus expect a higher variance of the posterior distribution in situations of prior-data conflict.

However, this is by no means automatically the case, in particular when adopting conjugate prior models, which are often used when data are scarce, where only strong prior beliefs allow for a reasonably precise answer in inference. Two simple and prominent examples of complete insensitivity to prior-data conflict are recalled in Section A.1.2: i.i.d. inferences on the mean of a scaled normal distribution and on the probability distribution of a categorical variable by the Dirichlet-Multinomial model.[2]

Sections A.1.3 and A.1.4 extend the question of (in)sensitivity to prior-data to regression models. We confine attention to linear regression analysis with conjugate priors, because

[1] See also the discussion on prior-data conflict in Sections 2.2.3.3, 3.1.4, and in the paper reproduced in Section 3.3.

[2] See the description of these two models in Sections 1.2.3.4 and 1.2.3.5, respectively.

— contrary to the more advanced regression model classes — the linear model still allows a fully analytical access, making it possible to understand potential restrictions imposed by the model in detail. We discuss and compare two different conjugate models:

(i) the standard conjugate prior (SCP, Section A.1.3) as described in Fahrmeir et al. (2013) or, in more detail, in O'Hagan (1994); and

(ii) a conjugate prior, called "canonically constructed conjugate prior" (CCCP, Section A.1.4) in the following, which is derived by a general method used to construct conjugate priors to sample distributions that belong to a certain class of exponential families, described, e.g., in Bernardo and Smith (2000).[3]

Whereas the former is the more general prior model, allowing for a very flexible modeling of prior information (which might be welcome or not), the latter allows only a strongly restricted covariance structure for $\boldsymbol{\beta}$, however offering a clearer insight in some aspects of the update process.

In a nutshell, the result is that both conjugate models do react to prior-data conflict by an enlarged factor to the variance-covariance matrix of the distribution on the regression coefficients $\boldsymbol{\beta}$; however, this reaction is unspecific, as it affects the variance and covariances of all components of $\boldsymbol{\beta}$ in a uniform way — even if the conflict occurs only in one single component.

Probably such an unspecific reaction of the variance is the most a (classical) Bayesian statistician can hope for, and traditional probability theory based on precise probabilities can offer. Indeed, Kyburg (1987) notes that

> [...] there appears to be no way, within the theory, of distinguishing between the cases in which there are good statistical grounds for accepting a prior distribution, and cases in which the prior distribution reflects merely ungrounded personal opinion.

and the same applies, in essence, to the posterior distribution.

A more sophisticated modeling would need a more elaborated concept of imprecision than is actually provided by looking at the variance (or other characteristics) of a (precise) probability distribution. Indeed, recently the theory of imprecise probabilities (Walley 1991; Weichselberger 2001) is gaining strong momentum.[4] It emerged as a general methodology to cope with the multidimensional character of uncertainty, also reacting to recent insights and developments in decision theory[5] and artificial intelligence,[6] where the exclusive role of probability as a methodology for handling uncertainty has eloquently been rejected (Klir and Wierman 1999):

[3] This is the regular conjugate framework of Section 1.2.3.1.

[4] See Section 2.1 for a short exposition of the theoretical foundations and motivations for the use of imprecise probability in statistical inference.

[5] See Hsu et al. (2005) for a neuro science corroboration of the constitutive difference of stochastic and non-stochastic aspects of uncertainty in human decision making, in the tradition of Ellsberg's (1961) seminal experiments.

[6] See, e.g., Walley (1996b) for the use of imprecise probability methods in expert systems.

> For three hundred years [...] uncertainty was conceived solely in terms of probability theory. This seemingly unique connection between uncertainty and probability is now challenged [... by several other] theories, which are demonstrably capable of characterizing situations under uncertainty. [...]
>
> [...] it has become clear that there are several distinct types of uncertainty. That is, it was realized that uncertainty is a multidimensional concept. [... That] multidimensional nature of uncertainty was obscured when uncertainty was conceived solely in terms of probability theory, in which it is manifested by only one of its dimensions.

Current applications include, among many other, risk analysis, reliability modeling and decision theory, see Augustin et al. (2009), Coolen et al. (2011) and Coolen-Schrijner et al. (2009) for recent collections on the subject.[7] As a welcome byproduct, imprecise probability models also provide a formal superstructure on models considered in robust Bayesian analysis (Ríos Insua and Ruggeri 2000), and frequentist robust statistic in the tradition of Huber and Strassen (1973), see also Augustin and Hable (2010) for a review.

By considering *sets* of distributions, and corresponding interval-valued probabilities for events, imprecise probability models allow to express the quality of the underlying knowledge in an elegant way. The higher the ambiguity, the larger c.p. the sets. The traditional concept of probability is contained as a special case, appropriate if and only if there is perfect stochastic information. This methodology allows also for a natural handling of prior-data conflict. If prior and data are in conflict, the set of posterior distributions are enlarged, and inferences become more cautious.[8]

In Section A.1.5, we briefly report that the CCCP model has a structure that allows a direct extension to an imprecise probability model along the lines of Quaeghebeur and de Cooman's (2005) imprecise probability models for i.i.d. exponential family models. Extending the models further by applying arguments from Walter and Augustin (2009b, see Section 3.3) yields a powerful generalisation of the linear regression model that is also capable of a component-specific reaction to prior-data conflict.

A.1.2. Prior-data Conflict in the i.i.d. Case

As a simple demonstration that conjugate models might not react to prior-data conflict reasonably, inference on the mean of data from a scaled normal distribution and inference on the category probabilities in multinomial sampling will be described in the following two subsections.

A.1.2.1. Samples from a scaled Normal distribution

The conjugate distribution to an i.i.d.-sample $\boldsymbol{x}$ of size n from a scaled normal distribution with mean μ, denoted by $\mathrm{N}(\mu, 1)$ is a normal distribution with mean $\mu^{(0)}$ and variance ${\sigma^{(0)}}^2$[9].

[7] See also, e.g., the list of applications of the IDM given in Section 3.1.3.

[8] For more details on the topic of imprecision and prior-data conflict, see Section 3.3.

[9] Here, and in the following, parameters of a prior distribution will be denoted by an upper index $^{(0)}$, whereas parameters of the respective posterior distribution by an upper index $^{(n)}$.

The posterior is then again a normal distribution with the following updated parameters:[10]

$$\mu^{(n)} = \frac{\frac{1}{n}}{\frac{1}{n} + {\sigma^{(0)}}^2}\mu^{(0)} + \frac{{\sigma^{(0)}}^2}{\frac{1}{n} + {\sigma^{(0)}}^2}\bar{x} = \frac{\frac{1}{{\sigma^{(0)}}^2}}{\frac{1}{{\sigma^{(0)}}^2} + n}\mu^{(0)} + \frac{n}{\frac{1}{{\sigma^{(0)}}^2} + n}\bar{x} \tag{A.1}$$

$${\sigma^{(n)}}^2 = \frac{{\sigma^{(0)}}^2 \cdot \frac{1}{n}}{{\sigma^{(0)}}^2 + \frac{1}{n}} = \frac{1}{\frac{1}{{\sigma^{(0)}}^2} + n}\,. \tag{A.2}$$

The posterior expectation (and mode) is thus a simple weighted average of the prior mean $\mu^{(0)}$ and the estimation from data $\bar{x}$, with weights $1/{\sigma^{(0)}}^2$ and n, respectively.[11] The variance of the posterior distribution is getting smaller automatically.

Now, in a situation where data is scarce, but with prior information one is very confident about, one would choose a low value for ${\sigma^{(0)}}^2$, thus resulting in a high weight for the prior mean $\mu^{(0)}$ in the calculation of $\mu^{(n)}$. The posterior distribution will be centered around a mean between $\mu^{(0)}$ and $\bar{x}$, and it will be even more pointed as the prior, because ${\sigma^{(n)}}^2$ is considerably smaller than ${\sigma^{(0)}}^2$, the factor to ${\sigma^{(0)}}^2$ in (A.2) being quite smaller than one.

The posterior basically would thus say that one can be quite sure that the mean μ is around $\mu^{(n)}$, regardless if $\mu^{(0)}$ and $\bar{x}$ were near to each other or not, where the latter would be a strong hint for prior-data conflict. The posterior variance does not depend on this; the posterior distribution is thus insensitive to prior-data conflict.

Even if one is not so confident about one's prior knowledge and thus assigning a relatively large variance to the prior, the posterior mean is less strongly influenced by the prior mean, but the posterior variance still is getting smaller, no matter if the data support the prior information or not.

The same insensitivity appears also in the widely used Dirichlet-Multinomial model as presented in the following subsection:

A.1.2.2. Samples from a Multinomial distribution

Given a sample of size n from a multinomial distribution, with probabilities θ_j for categories or classes $j = 1, \ldots, k$, subsumed in the vectorial parameter $\boldsymbol{\theta}$ (with $\sum_{j=1}^{k} \theta_j = 1$), the conjugate prior on $\boldsymbol{\theta}$ is a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$. Written in terms of the canonical parameters $n^{(0)}$ and $y^{(0)}$ as in Section 1.2.3.5, $\alpha_j = n^{(0)} \cdot y_j^{(0)}$, such that $\sum_{j=1}^{k} y_j^{(0)} = 1$, $(y_1^{(0)}, \ldots, y_k^{(0)})^\mathsf{T} =: \boldsymbol{y}^{(0)}$. Recall that the components of $\boldsymbol{y}^{(0)}$ have a direct interpretation as prior class probabilities, whereas $n^{(0)}$ is a parameter indicating the confidence in the values of $\boldsymbol{y}^{(0)}$, similar to the inverse variance as in Section A.1.2.1 (the quantity $n^{(0)}$ will appear also in Section A.1.4).[12]

[10] This is the Normal-Normal model from Section 1.2.3.4, where $\sigma_0^2 = 1$, $y^{(0)} = \mu^{(0)}$, and $n^{(0)} = 1/{\sigma^{(0)}}^2$.

[11] The reason for using these seemingly strange weights will become clear later.

[12] If $\boldsymbol{\theta} \sim \mathrm{Dir}(n^{(0)}, \boldsymbol{y}^{(0)})$, then $\mathrm{Var}(\theta_j) = \frac{y_j^{(0)}(1-y_j^{(0)})}{n^{(0)}+1}$. If $n^{(0)}$ is high, then the variances of $\boldsymbol{\theta}$ will become low, thus indicating high confidence in the chosen values of $\boldsymbol{y}^{(0)}$.

As seen in Section 1.2.3.5, the posterior distribution, obtained after updating via Bayes' Rule with a sample vector $\boldsymbol{n} = (n_1, \ldots, n_k)$, $\sum_{j=1}^{k} n_j = n$ collecting the observed counts in each category, is a Dirichlet distribution with parameters

$$y_j^{(n)} = \frac{n^{(0)}}{n^{(0)} + n} y_j^{(0)} + \frac{n}{n^{(0)} + n} \cdot \frac{n_j}{n}\,, \qquad\qquad n^{(n)} = n^{(0)} + n\,.$$

The posterior class probabilities $\boldsymbol{y}^{(n)}$ are calculated as a weighted mean of the prior class probabilities $\boldsymbol{y}^{(0)}$ and $\frac{n_j}{n}$, the proportion in the sample, with weights $n^{(0)}$ and n, respectively; the confidence parameter $n^{(0)}$ is incremented by the sample size n.

Also here, there is no systematic reaction to prior-data conflict. The posterior variance for each class probability θ_j is

$$\mathrm{Var}(\theta_j \mid \boldsymbol{n}) = \frac{y_j^{(n)}(1 - y_j^{(n)})}{n^{(n)} + 1} = \frac{y_j^{(n)}(1 - y_j^{(n)})}{n^{(0)} + n + 1}\,.$$

The posterior variance depends heavily on $y_j^{(n)}(1 - y_j^{(n)})$, having values between 0 and $\frac{1}{4}$, which do not change specifically to prior data conflict. The denominator increases from $n^{(0)} + 1$ to $n^{(0)} + n + 1$.

Imagine a situation with strong prior information suggesting a value of $y_j^{(0)} = 0.25$, so one could choose $n^{(0)} = 5$, resulting in a prior class variance of $\frac{1}{32}$. Consider a sample of size $n = 10$ with all observations belonging to class j (thus $n_j = 10$), being in clear contrast to the prior information. The posterior class probability is then $y_j^{(n)} = 0.75$, resulting the enumerator value of the class variance to remain constant. Therefore, due to the increasing denominator, the variance decreases to $\frac{3}{256}$, in spite of the clear conflict between prior and sample information. Of course, one can also construct situations where the variance increases, but this happens only in case of an update of $y_j^{(0)}$ towards $\frac{1}{2}$. If $y_j^{(0)} = \frac{1}{2}$, the variance will decrease for any degree of prior-data conflict.

A.1.3. The Standard Approach for Bayesian Linear Regression (SCP)

The regression model is noted as follows:

$$z_i = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i\,, \quad \boldsymbol{x}_i \in \mathbb{R}^p\,,\ \boldsymbol{\beta} \in \mathbb{R}^p\,,\ \varepsilon_i \sim \mathrm{N}(0, \sigma^2)\,,$$

where z_i is the response, $\boldsymbol{x}_i$ the vector of the p covariates for observation i, and $\boldsymbol{\beta}$ is the p-dimensional vector of adjacent regression coefficients.

The vector of regressors $\boldsymbol{x}_i$ for each observation i is generally considered to be non-stochastic, thus it holds that $z_i \sim \mathrm{N}(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}, \sigma^2)$, or, for n i.i.d. samples, $\boldsymbol{z} \sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, i.e., $\boldsymbol{z} \in \mathbb{R}^n$, the column vector of the responses z_i, has a multivariate normal distribution with vector of expectations $\mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the *design matrix* (of which row i is the vector of covariates $\boldsymbol{x}_i^{\mathsf{T}}$ for observation i), and matrix of variances and covariances $\sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity or unit matrix of size n.

Without loss of generality, one can either assume $x_{i1} = 1\ \forall i$ such that the first component of $\boldsymbol{\beta}$ is the intercept parameter,[13] or consider only centered responses $\boldsymbol{z}$ and standardized covariates to make the estimation of an intercept unnecessary.

In Bayesian linear regression analysis, the distribution of the response $\boldsymbol{z}$ is interpreted as a distribution of $\boldsymbol{z}$ given the parameters $\boldsymbol{\beta}$ and σ^2, and prior distributions on $\boldsymbol{\beta}$ and σ^2 must be considered. For this, it is convenient to split the joint prior on β and σ^2 as $p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta} \mid \sigma^2)p(\sigma^2)$ and to consider conjugate distributions for both parts, respectively.

In the literature, the proposed conjugate prior for $\boldsymbol{\beta} \mid \sigma^2$ is a normal distribution with expectation vector $\boldsymbol{m}^{(0)} \in \mathbb{R}^p$ and variance-covariance matrix $\sigma^2\mathbf{M}^{(0)}$, where $\mathbf{M}^{(0)}$ is a symmetric positive definite matrix of size $p \times p$. The prior on σ^2 is an inverse gamma distribution (i.e., $1/\sigma^2$ is gamma distributed) with parameters $a^{(0)}$ and $b^{(0)}$, in the sense that

$$p(\sigma^2) \propto \frac{1}{(\sigma^2)^{a^{(0)}+1}} \exp\left\{ -\frac{b^{(0)}}{\sigma^2} \right\}.$$

The joint prior on $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)^\mathsf{T}$ is then denoted as a normal-inverse gamma (NIG) distribution. The derivation of this prior and the proof of its conjugacy can be found, e.g., in Fahrmeir et al. (2013), or in O'Hagan (1994), the latter using a different parameterisation of the inverse gamma part, where $a^{(0)} = \frac{d}{2}$ and $b^{(0)} = \frac{a}{2}$.

For the prior model, it holds thus that (if $a^{(0)} > 1$ resp. $a^{(0)} > 2$)

$$\begin{aligned} \mathrm{E}[\boldsymbol{\beta} \mid \sigma^2] &= \boldsymbol{m}^{(0)}, \qquad & \mathrm{Var}(\boldsymbol{\beta} \mid \sigma^2) &= \sigma^2\mathbf{M}^{(0)}, \\ \mathrm{E}[\sigma^2] &= \frac{b^{(0)}}{a^{(0)} - 1}, \qquad & \mathrm{Var}(\sigma^2) &= \frac{(b^{(0)})^2}{(a^{(0)} - 1)^2(a^{(0)} - 2)}. \end{aligned} \tag{A.3}$$

As σ^2 is considered as nuisance parameter, the unconditional distribution on $\boldsymbol{\beta}$ is of central interest, because it subsumes the shape of prior knowledge on $\boldsymbol{\beta}$ as expressed by the choice of parameters $\boldsymbol{m}^{(0)}$, $\mathbf{M}^{(0)}$, $a^{(0)}$ and $b^{(0)}$. It can be shown that $p(\boldsymbol{\beta})$ is a multivariate noncentral t distribution with $2a^{(0)}$ degrees of freedom, location parameter $\boldsymbol{m}^{(0)}$ and dispersion parameter $\frac{b^{(0)}}{a^{(0)}}\mathbf{M}^{(0)}$, such that

$$\mathrm{E}[\boldsymbol{\beta}] = \boldsymbol{m}^{(0)}, \qquad \mathrm{Var}(\boldsymbol{\beta}) = \frac{b^{(0)}}{a^{(0)} - 1}\mathbf{M}^{(0)} = \mathrm{E}[\sigma^2]\mathbf{M}^{(0)}. \tag{A.4}$$

The joint posterior distribution $p(\boldsymbol{\theta} \mid z)$, due to conjugacy, is then again a normal-inverse gamma distribution with the updated parameters

$$\begin{aligned} \boldsymbol{m}^{(n)} &= \left(\mathbf{M}^{(0)^{-1}} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\left(\mathbf{M}^{(0)^{-1}}\boldsymbol{m}^{(0)} + \mathbf{X}^\mathsf{T}\boldsymbol{z}\right), \\ \mathbf{M}^{(n)} &= \left(\mathbf{M}^{(0)^{-1}} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}, \\ a^{(n)} &= a^{(0)} + \frac{n}{2}, \end{aligned}$$

[13] usually denoted by β_0; however, we stay with the numbering $1, \ldots, p$ for the components of $\boldsymbol{\beta}$.

$$b^{(n)} = b^{(0)} + \frac{1}{2}\left(\boldsymbol{z}^\mathsf{T}\boldsymbol{z} + \boldsymbol{m}^{(0)^\mathsf{T}}\mathbf{M}^{(0)^{-1}}\boldsymbol{m}^{(0)} - \boldsymbol{m}^{(n)^\mathsf{T}}\mathbf{M}^{(n)^{-1}}\boldsymbol{m}^{(n)}\right) .$$

The properties of the posterior distributions can thus be analyzed by inserting the updated parameters into (A.3) and (A.4).

A.1.3.1. Update of $\boldsymbol{\beta} \mid \sigma^2$

The normal distribution part of the joint prior is updated as follows:

$$\begin{aligned}\mathrm{E}[\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{z}] &= \boldsymbol{m}^{(n)} \\ &= \left(\mathbf{M}^{(0)^{-1}} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\left(\mathbf{M}^{(0)^{-1}}\boldsymbol{m}^{(0)} + \mathbf{X}^\mathsf{T}\boldsymbol{z}\right) \\ &= (\mathbf{I} - \mathbf{A})\,\boldsymbol{m}^{(0)} + \mathbf{A}\,\hat{\boldsymbol{\beta}}_{\mathrm{LS}} ,\end{aligned}$$

where $\mathbf{A} = \left(\mathbf{M}^{(0)^{-1}} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}$. The posterior estimate of $\boldsymbol{\beta} \mid \sigma^2$ thus can be seen as a matrix-weighted mean of the prior guess and the least-squares estimate $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$. The larger the diagonal elements of $\mathbf{M}^{(0)}$ (i.e., the weaker the prior information), the smaller the elements of $\mathbf{M}^{(0)^{-1}}$ and thus the 'nearer' is $\mathbf{A}$ to the identity matrix, so that the posterior estimate is nearer to the least-squares estimate.

The posterior variance of $\boldsymbol{\beta} \mid \sigma^2$ is

$$\mathrm{Var}(\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{z}) = \sigma^2\mathbf{M}^{(n)} = \sigma^2\left(\mathbf{M}^{(0)^{-1}} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1} .$$

As the elements of $\mathbf{M}^{(n)^{-1}}$ get larger with as compared to $\mathbf{M}^{(0)^{-1}}$, the elements of $\mathbf{M}^{(n)}$ will, roughly speaking, become smaller than those of $\mathbf{M}^{(0)}$, so that the variance of $\boldsymbol{\beta} \mid \sigma^2$ decreases.

Therefore, the updating of $\boldsymbol{\beta} \mid \sigma^2$ is obviously insensitive to prior-data conflict, because the posterior distribution will not become flatter in case of a large distance between $\mathrm{E}[\boldsymbol{\beta}]$ and $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$. Actually, as O'Hagan (1994) derives, for any $\phi = \boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}$, $\boldsymbol{a} \in \mathbb{R}^p$, i.e., any linear combination of elements of $\boldsymbol{\beta}$, it holds that $\mathrm{Var}(\phi \mid \sigma^2, \boldsymbol{z}) \leq \mathrm{Var}(\phi \mid \sigma^2)$, becoming a strict inequality if $\mathbf{X}$ has full rank. In particular, the variance of each β_i decreases automatically with the update step.

A.1.3.2. Update of σ^2

It can be shown (O'Hagan 1994) that

$$\mathrm{E}[\sigma^2 \mid \boldsymbol{z}] = \frac{2a^{(0)} - 2}{2a^{(0)} + n - 2}\,\mathrm{E}[\sigma^2] + \frac{n - p}{2a^{(0)} + n - 2}\hat{\sigma}^2_{\mathrm{LS}} + \frac{p}{2a^{(0)} + n - 2}\hat{\sigma}^2_{\mathrm{PDC}} , \tag{A.5}$$

where $\hat{\sigma}^2_{\mathrm{LS}} = \frac{1}{n-p}(\boldsymbol{z} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}})^\mathsf{T}(\boldsymbol{z} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}})$ is the least-squares based estimate for σ^2, and

$$\hat{\sigma}^2_{\mathrm{PDC}} = \frac{1}{p}(\boldsymbol{m}^{(0)} - \hat{\boldsymbol{\beta}}_{\mathrm{LS}})^\mathsf{T}\left(\mathbf{M}^{(0)} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)^{-1}(\boldsymbol{m}^{(0)} - \hat{\boldsymbol{\beta}}_{\mathrm{LS}}) .$$

For the latter it holds that $\mathrm{E}[\hat{\sigma}^2_{\mathrm{PDC}} \mid \sigma^2] = \sigma^2$; the posterior expectation of σ^2 can thus be seen as a weighted mean of three estimates:

(i) the prior expectation for σ^2,

(ii) the least-squares estimate, and

(iii) an estimate based on a weighted squared difference of the prior mean $\boldsymbol{m}^{(0)}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, the least-squares estimate for $\boldsymbol{\beta}$.

The weights depend on $a^{(0)}$ (one prior parameter for the inverse gamma part), the sample size n, and the dimension of $\boldsymbol{\beta}$, respectively. The role of the first weight gets more plausible when remembering the formula for the prior variance of σ^2 in (A.3), where $a^{(0)}$ appears in the denominator. A larger value of $a^{(0)}$ means thus smaller prior variance, in turn giving a higher weight for $\mathrm{E}[\sigma^2]$ in the calculation of $\mathrm{E}[\sigma^2 \mid \boldsymbol{z}]$. The weight to $\hat{\sigma}^2_{\mathrm{LS}}$ corresponds to the classical degrees of freedom, $n-p$. With the the sample size approaching infinity, this weight will dominate the others, such that $\mathrm{E}[\sigma^2 \mid \boldsymbol{z}]$ approaches $\hat{\sigma}^2_{\mathrm{LS}}$.

Similar results hold for the posterior mode instead of the posterior expectation.

Here, the estimate $\hat{\sigma}^2_{\mathrm{PDC}}$ allows some reaction to prior-data conflict: it measures the distance between $\boldsymbol{m}^{(0)}$ (prior) and $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ (data) estimates for $\boldsymbol{\beta}$, with a large distance resulting basically in a large value of $\hat{\sigma}^2_{\mathrm{PDC}}$ and thus an enlarged posterior estimate for σ^2.

The weighting matrix for the distances is playing an important role as well. The influence of $\mathbf{M}^{(0)}$ is as follows: for components of $\boldsymbol{\beta}$ one is quite certain about the assignment of $\boldsymbol{m}^{(0)}$, the respective diagonal elements of $\mathbf{M}^{(0)}$ will be low, so that these diagonal elements of the weighting matrix will be high. Therefore, large distances in these dimensions will increase $\hat{\sigma}^2_{\mathrm{PDC}}$ strongly. An erroneously high confidence in the prior assumptions on $\boldsymbol{\beta}$ is thus penalised by an increasing posterior estimate for σ^2. The influence of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ interprets as follows: covariates with a low spread in x-values, giving an unstable base for the estimate $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, will result in low diagonal elements of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$. Via the double inverting, those diagonal elements of the weighting matrix will remain low and thus give the difference a low weight. Therefore, $\hat{\sigma}^2_{\mathrm{PDC}}$ will not excessively increase due to a large difference in dimensions where the location of $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ is to be taken cum grano salis.

As to be seen in the following subsection, the behavior of $\mathrm{E}[\sigma \mid \boldsymbol{z}]$ is of high importance for posterior inferences on $\boldsymbol{\beta}$.

A.1.3.3. Update of $\boldsymbol{\beta}$

The posterior distribution of $\boldsymbol{\beta}$ is again a multivariate t, with expectation

$$\mathrm{E}[\boldsymbol{\beta} \mid \boldsymbol{z}] = \mathrm{E}\left[\,\mathrm{E}[\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{z}] \mid \boldsymbol{z}\right] = \boldsymbol{m}^{(n)}$$

as described in Section A.1.3.1, and variance

$$\begin{aligned}
\mathrm{Var}[\beta \mid \boldsymbol{z}] &= \frac{b^{(n)}}{a^{(n)}-1}\mathbf{M}^{(n)} \\
&= \mathrm{E}[\sigma^2 \mid \boldsymbol{z}]\,\mathbf{M}^{(n)} \\
&= \left(\frac{2a^{(0)}-2}{2a^{(0)}+n-2}\,\mathrm{E}[\sigma^2] + \frac{n-p}{2a^{(0)}+n-2}\hat{\sigma}^2_{\mathrm{LS}} + \frac{p}{2a^{(0)}+n-2}\hat{\sigma}^2_{\mathrm{PDC}}\right)
\end{aligned} \tag{A.6}$$

$$\cdot \left(\mathbf{M}^{(0)^{-1}} + \mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}$$
$$= \left(\frac{2a^{(0)}-2}{2a^{(0)}+n-2}\,\mathrm{E}[\sigma^2] + \frac{n-p}{2a^{(0)}+n-2}\hat{\sigma}^2_{\mathrm{LS}} + \frac{p}{2a^{(0)}+n-2}\hat{\sigma}^2_{\mathrm{PDC}}\right)$$
$$\cdot \left(\mathbf{M}^{(0)} - \mathbf{M}^{(0)}\mathbf{X}^{\mathsf{T}}(\mathbf{I}+\mathbf{X}\mathbf{M}^{(0)}\mathbf{X}^{\mathsf{T}})^{-1}\mathbf{X}\mathbf{M}^{(0)}\right),$$

not being directly expressible as a function of $\mathrm{E}[\sigma^2]\mathbf{M}^{(0)}$, the prior variance of β.

Due to the effect of $\mathrm{E}[\sigma^2 \mid \boldsymbol{z}]$, the posterior variance-covariance matrix of $\boldsymbol{\beta}$ can increase in case of prior data conflict, if the rise of $\mathrm{E}[\boldsymbol{\beta} \mid \boldsymbol{z}]$ (due to an even stronger rise of $\hat{\sigma}^2_{\mathrm{PDC}}$) can overcompensate the decrease in the elements of $\mathbf{M}^{(n)}$. However, we see that the effect of prior-data conflict on the posterior variance of $\boldsymbol{\beta}$ is *globally* and not component-specific; it influences the variances for *all* components of $\boldsymbol{\beta}$ with the same amount, even if the conflict was confined only to some or even just one single component. Taking it to the extremes, if the prior assignment $\boldsymbol{m}^{(0)}$ was (more or less) correct in all but one component, with that one being far out, the posterior variances will increase for all components, also for the ones with prior assignments that have turned out to be basically correct.

A.1.4. An Alternative Approach for Conjugate Priors in Bayesian Linear Regression (CCCP)

In this section, a prior model for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ will be constructed along the general construction method for sample distributions that form a linear, canonical exponential family (see the canical conjugates framework in Section 1.2.3.1, and, e.g., Bernardo and Smith 2000). As shown for the examples in Sections 1.2.3.3, 1.2.3.4 and 1.2.3.5, the method is typically used for the i.i.d. case, but the likelihood arising from $\boldsymbol{z} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ will be shown to follow the specific exponential family form as well.

The canonically constructed conjugate prior (CCCP) model will also result in a normal-inverse gamma distribution, but with a fixed variance-covariance structure. The CCCP model is thus a special case of the SCP model, which – as will be detailed in this subsection – offers some interesting further insights into the structure of the update step.

The likelihood arising from the distribution of $\boldsymbol{z}$,

$$\begin{aligned}
&f(\boldsymbol{z} \mid \boldsymbol{\beta}, \sigma^2) \\
&\quad = \prod_{i=1}^{n} f(z_i \mid \boldsymbol{\beta}, \sigma^2) \\
&\quad = \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(z_i - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})^2\right\} \\
&\quad = \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}} \exp\Big\{-\frac{1}{2\sigma^2}(\boldsymbol{z} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{z} - \mathbf{X}\boldsymbol{\beta})\Big\} \\
&\quad = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\Big\{-\frac{n}{2}\log(\sigma^2)\Big\}
\end{aligned}$$

$$\exp\Big\{-\frac{1}{2\sigma^2}\boldsymbol{z}^\mathsf{T}\boldsymbol{z}+\frac{1}{2\sigma^2}\boldsymbol{z}^\mathsf{T}\mathbf{X}\boldsymbol{\beta}+\frac{1}{2\sigma^2}(\mathbf{X}\boldsymbol{\beta})^\mathsf{T}\boldsymbol{z}-\frac{1}{2\sigma^2}(\mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{X}\boldsymbol{\beta})\Big\}$$

$$= \underbrace{\frac{1}{(2\pi)^{\frac{n}{2}}}}_{\mathbf{a}(\boldsymbol{z})=\prod_{i=1}^n \mathbf{a}(z_i)} \exp\Big\{\underbrace{\begin{pmatrix}\frac{\boldsymbol{\beta}}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{pmatrix}^\mathsf{T}}_{\boldsymbol{\psi}_1}\underbrace{\begin{pmatrix}\mathbf{X}^\mathsf{T}\boldsymbol{z}\\ \frac{1}{2}\boldsymbol{z}^\mathsf{T}\boldsymbol{z}\end{pmatrix}}_{\boldsymbol{\tau}_1(\boldsymbol{z})} \underbrace{-\frac{1}{\sigma^2}}_{\psi_2}\underbrace{\frac{1}{2}\boldsymbol{z}^\mathsf{T}\boldsymbol{z}}_{\tau_2(\boldsymbol{z})} \underbrace{-\Big(\frac{1}{2\sigma^2}\boldsymbol{\beta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta}+\frac{n}{2}\log(\sigma^2)\Big)}_{n\mathbf{b}(\boldsymbol{\psi})}\Big\},$$

indeed corresponds to the canonical exponential family form[14]

$$f(\boldsymbol{z}\mid\boldsymbol{\psi})=\mathbf{a}(z)\cdot\exp\{\langle\psi,\tau(z)\rangle-n\cdot\mathbf{b}(\boldsymbol{\psi})\},$$

where $\boldsymbol{\psi}=\boldsymbol{\psi}(\boldsymbol{\beta},\sigma^2)$ is a certain function of $\boldsymbol{\beta}$ and σ^2, the parameters of interest. $\tau(\boldsymbol{z})$ is a $p+1$-dimensional sufficient statistic of $\boldsymbol{z}$ used in the update step. Here, we have

$$\boldsymbol{\psi}=\begin{pmatrix}\frac{\boldsymbol{\beta}}{\sigma^2}\\ -\frac{1}{\sigma^2}\end{pmatrix},\qquad \tau(\boldsymbol{z})=\begin{pmatrix}\mathbf{X}^\mathsf{T}\boldsymbol{z}\\ \frac{1}{2}\boldsymbol{z}^\mathsf{T}\boldsymbol{z}\end{pmatrix},\qquad \mathbf{b}(\boldsymbol{\psi})=\frac{1}{2n\sigma^2}\boldsymbol{\beta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta}+\frac{1}{2}\log(\sigma^2). \tag{A.7}$$

According to the general construction method, a conjugate prior for $\boldsymbol{\psi}$ can be obtained from these ingredients by the following equation:[15]

$$p(\boldsymbol{\psi})=\mathbf{c}(n^{(0)},\boldsymbol{y}^{(0)})\cdot\exp\left\{n^{(0)}\cdot[\langle\boldsymbol{\psi},\boldsymbol{y}^{(0)}\rangle-\mathbf{b}(\boldsymbol{\psi})]\right\}, \tag{A.8}$$

where $n^{(0)}$ and $\boldsymbol{y}^{(0)}$ are the parameters that define the concrete prior distribution of its distribution family; whereas $\boldsymbol{\psi}$ and $\mathbf{b}(\boldsymbol{\psi})$ were identified in (A.7). $\mathbf{c}(\cdot)$ corresponds to a normalisation factor for the prior.[16]

Here, the conjugate prior writes as

$$p(\boldsymbol{\psi})d\boldsymbol{\psi}=\mathbf{c}(n^{(0)},\boldsymbol{y}^{(0)})\exp\left\{n^{(0)}\Big[\boldsymbol{y}^{(0)\mathsf{T}}\begin{pmatrix}\frac{\boldsymbol{\beta}}{\sigma^2}\\ -\frac{1}{\sigma^2}\end{pmatrix}-\frac{1}{2n\sigma^2}\boldsymbol{\beta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta}-\frac{1}{2}\log(\sigma^2)\Big]\right\}d\boldsymbol{\psi}.$$

As this is a prior on $\boldsymbol{\psi}$, but we want to arrive at a prior on $\boldsymbol{\theta}=(\boldsymbol{\beta}^\mathsf{T},\,\sigma^2)^\mathsf{T}$, we must transform the density $p(\boldsymbol{\psi})$:

$$p(\boldsymbol{\theta})d\boldsymbol{\theta}=p(\boldsymbol{\psi})d\boldsymbol{\psi}\cdot\left|\det\left(\frac{d\boldsymbol{\psi}}{d\boldsymbol{\theta}}\right)\right|$$

For the transformation, we need the determinant of the Jacobian matrix $\frac{d\boldsymbol{\psi}}{d\boldsymbol{\theta}}$. As it holds that

$$\frac{d\psi_i}{d\theta_j}=\frac{1}{d\beta_j}\,\frac{\beta_i}{\sigma^2}=\begin{cases}0 & i\neq j\\ \frac{1}{\sigma^2} & i=j\end{cases}\qquad\qquad \forall i,\,j\in\{1,\ldots,p\},$$

[14] In Equation 1.4, the integration constant $\mathbf{a}(\cdot)$ was omitted.

[15] See Equation 1.5, where the integration constant $\mathbf{c}(\cdot)$ was omitted as well.

[16] When applying the general construction method to the two examples from Section A.1.2, the priors as presented there will result, where $y^{(0)}=\mu^{(0)}$ and $n^{(0)}=1/{\sigma^{(0)}}^2$ for the prior to the scaled normal model. For details of the derivation, see Sections 1.2.3.4 and 1.2.3.5, respectively.

$$\begin{aligned}
\frac{d\psi_{p+1}}{d\theta_j} &= \frac{1}{d\beta_j}\left(-\frac{1}{\sigma^2}\right) = 0 && \forall j \in \{1,\dots,p\}\,,\\
\frac{d\psi_i}{d\theta_{p+1}} &= \frac{1}{d\sigma^2}\frac{\beta_i}{\sigma^2} = -\frac{\beta_i}{(\sigma^2)^2} && \forall i \in \{1,\dots,p\}\,,\\
\frac{d\psi_{p+1}}{d\theta_{p+1}} &= \frac{1}{d\sigma^2}\left(-\frac{1}{\sigma^2}\right) = \frac{1}{(\sigma^2)^2}\,,
\end{aligned}$$

we get

$$\left|\det\left(\frac{d\boldsymbol{\psi}}{d\boldsymbol{\theta}}\right)\right| = \left|\det\begin{pmatrix} \frac{1}{\sigma^2}\mathbf{I} & -\frac{\boldsymbol{\beta}}{(\sigma^2)^2} \\ \mathbf{0} & \frac{1}{(\sigma^2)^2}\end{pmatrix}\right| = \frac{1}{(\sigma^2)^{p+2}}\,,$$

where $\mathbf{I}$ is the $p \times p$ identity matrix, and $\mathbf{0}$ a p-dimensional row vector of zeroes. Therefore, the prior on $\boldsymbol{\theta} = (\boldsymbol{\beta},\, \sigma^2)^\mathsf{T}$ is

$$\begin{aligned}
&p(\boldsymbol{\theta})d\boldsymbol{\theta}\\
&= p(\boldsymbol{\psi})d\boldsymbol{\psi}\cdot\left|\det\left(\frac{d\boldsymbol{\psi}}{d\boldsymbol{\theta}}\right)\right|\\
&= \mathbf{c}(n^{(0)}, \boldsymbol{y}^{(0)})\\
&\quad \exp\left\{n^{(0)}\boldsymbol{y}_1^{(0)\mathsf{T}}\frac{\boldsymbol{\beta}}{\sigma^2} - n^{(0)}y_2^{(0)}\frac{1}{\sigma^2} - \frac{n^{(0)}}{2n\sigma^2}\boldsymbol{\beta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta} - \frac{n^{(0)}}{2}\log(\sigma^2) - (p+2)\log(\sigma^2)\right\},
\end{aligned} \tag{A.9}$$

where $\boldsymbol{y}^{(0)} = \left(\boldsymbol{y}_1^{(0)\mathsf{T}}, y_2^{(0)}\right)^\mathsf{T}$, and $\boldsymbol{y}_1^{(0)} \in \mathbb{R}^p$, $y_2^{(0)} \in \mathbb{R}_{>0}$.

$\boldsymbol{\theta}$ can now be shown to follow a normal-inverse gamma distribution by comparing coefficients. In doing that, some attention must be paid to the terms proportional to $-1/\sigma^2$ (appearing as $-\log(\sigma^2)$ in the exponent), because these can appear in both the normal distribution $p(\boldsymbol{\beta} \mid \sigma^2)$ and in the inverse gamma $p(\sigma^2)$ distribution. Furthermore, it is necessary to complete the square for the normal part, resulting in an additional term for the inverse gamma part.

The density of a normal distribution on $\boldsymbol{\beta} \mid \sigma^2$ with a mean vector[17] $\overline{\boldsymbol{m}}^{(0)} = \overline{\boldsymbol{m}}^{(0)}(n^{(0)}, \boldsymbol{y}^{(0)})$ and a variance-covariance matrix $\sigma^2\overline{\mathbf{M}}^{(0)} = \sigma^2\overline{\mathbf{M}}^{(0)}(n^{(0)}, \boldsymbol{y}^{(0)})$, both to be seen as functions of the canonical parameters $n^{(0)}$ and $\boldsymbol{y}^{(0)}$, has the following form:

$$\begin{aligned}
p(\boldsymbol{\beta} \mid \sigma^2) &= \frac{1}{(2\pi)^{\frac{p}{2}}(\sigma^2)^{\frac{p}{2}}}\exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta} - \overline{\boldsymbol{m}}^{(0)}\right)^\mathsf{T}\overline{\mathbf{M}}^{(0)^{-1}}\left(\boldsymbol{\beta} - \overline{\boldsymbol{m}}^{(0)}\right)\right\}\\
&= \frac{1}{(2\pi)^{\frac{p}{2}}}\exp\left\{\overline{\boldsymbol{m}}^{(0)\mathsf{T}}\overline{\mathbf{M}}^{(0)^{-1}}\frac{\boldsymbol{\beta}}{\sigma^2} - \frac{1}{2\sigma^2}\boldsymbol{\beta}^\mathsf{T}\overline{\mathbf{M}}^{(0)^{-1}}\boldsymbol{\beta}\right.\\
&\qquad\qquad \left. - \frac{1}{2\sigma^2}\overline{\boldsymbol{m}}^{(0)\mathsf{T}}\overline{\mathbf{M}}^{(0)^{-1}}\overline{\boldsymbol{m}}^{(0)} - \frac{p}{2}\log(\sigma^2)\right\}.
\end{aligned}$$

[17] We will denote the parameters of the canonically constructed prior (CCCP) by an overlined version of the parameters of the standard conjugate prior (SCP) in order to emphasise the different meanings.

Comparing coefficients with the terms from (A.9) depending on $\boldsymbol{\beta}$, we get

$$\overline{\mathbf{M}}^{(0)^{-1}} = \frac{n^{(0)}}{n}\,\mathbf{X}^\mathsf{T}\mathbf{X}\,, \qquad\qquad \overline{\boldsymbol{m}}^{(0)} = n\,(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\,\boldsymbol{y}_1^{(0)}\,,$$

where the latter derives from

$$\begin{aligned} & n^{(0)}\boldsymbol{y}_1^{(0)^\mathsf{T}}\frac{\boldsymbol{\beta}}{\sigma^2} \stackrel{!}{=} \overline{\boldsymbol{m}}^{(0)^\mathsf{T}}\frac{n^{(0)}}{n\sigma^2}\,\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta} \\ \iff \qquad & n^{(0)}\boldsymbol{y}_1^{(0)^\mathsf{T}} \stackrel{!}{=} \overline{\boldsymbol{m}}^{(0)^\mathsf{T}}\frac{n^{(0)}}{n}\,\mathbf{X}^\mathsf{T}\mathbf{X} \\ \iff \quad & \boldsymbol{y}_1^{(0)^\mathsf{T}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\,n \stackrel{!}{=} \overline{\boldsymbol{m}}^{(0)^\mathsf{T}}\,. \end{aligned}$$

We must thus complete the square in the exponent with

$$\begin{aligned} & -\frac{1}{2\sigma^2}\overline{\boldsymbol{m}}^{(0)^\mathsf{T}}\overline{\mathbf{M}}^{(0)^{-1}}\overline{\boldsymbol{m}}^{(0)} + \frac{1}{2\sigma^2}\overline{\boldsymbol{m}}^{(0)^\mathsf{T}}\overline{\mathbf{M}}^{(0)^{-1}}\overline{\boldsymbol{m}}^{(0)} \\ & \qquad = -\frac{1}{2\sigma^2}\left(n\cdot n^{(0)}\boldsymbol{y}_1^{(0)^\mathsf{T}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{y}_1^{(0)}\right) - \frac{1}{\sigma^2}\left(-\frac{n^{(0)}n}{2}\boldsymbol{y}_1^{(0)^\mathsf{T}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{y}_1^{(0)}\right)\,, \end{aligned}$$

such that the joint density of $\boldsymbol{\beta}$ and σ^2 reads as

$$\begin{aligned} p(\boldsymbol{\beta},\sigma^2) = \mathbf{c}(n^{(0)},\boldsymbol{y}^{(0)}) & \\ \exp\Big\{ & \underbrace{n^{(0)}\boldsymbol{y}_1^{(0)^\mathsf{T}}\frac{\boldsymbol{\beta}}{\sigma^2} - \frac{n^{(0)}}{2n\sigma^2}\boldsymbol{\beta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta} - \frac{1}{2\sigma^2}\left(n\cdot n^{(0)}\boldsymbol{y}_1^{(0)^\mathsf{T}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{y}_1^{(0)}\right) - \frac{p}{2}\log(\sigma^2)}_{\text{to } p(\boldsymbol{\beta}\mid\sigma^2) \text{ (normal distribution)}} \\ & \underbrace{-\frac{1}{\sigma^2}\left(-\frac{n^{(0)}n}{2}\boldsymbol{y}_1^{(0)^\mathsf{T}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{y}_1^{(0)}\right) - n^{(0)}y_2^{(0)}\frac{1}{\sigma^2} - \left(\frac{n^{(0)}+p}{2}+2\right)\log(\sigma^2)}_{\text{to } p(\sigma^2) \text{ (inverse gamma distribution)}}\Big\}\,. \end{aligned} \tag{A.10}$$

Therefore, one part of the conjugate prior (A.10) reveals as a multivariate normal distribution with mean vector $\overline{\boldsymbol{m}}^{(0)} = n\,(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\,\boldsymbol{y}_1^{(0)}$ and variance-covariance matrix $\sigma^2\overline{\mathbf{M}}^{(0)} = \frac{n\sigma^2}{n^{(0)}}\,(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$, i.e.

$$\boldsymbol{\beta} \mid \sigma^2 \sim \mathrm{N}_p\left(n\,(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\,\boldsymbol{y}_1^{(0)}\,,\ \frac{n\sigma^2}{n^{(0)}}\,(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)\,. \tag{A.11}$$

The other terms in (A.10) can be directly identified with the core of an inverse gamma distribution with parameters

$$\begin{aligned} & \overline{a}^{(0)} = \frac{n^{(0)}+p}{2}+1 \qquad \text{and} \\ & \overline{b}^{(0)} = n^{(0)}y_2^{(0)} - \frac{n^{(0)}}{2}\boldsymbol{y}_1^{(0)^\mathsf{T}}n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{y}_1^{(0)} = n^{(0)}y_2^{(0)} - \frac{1}{2}\overline{\boldsymbol{m}}^{(0)^\mathsf{T}}\overline{\mathbf{M}}^{(0)^{-1}}\overline{\boldsymbol{m}}^{(0)}\,, \end{aligned}$$

$$\text{i.e.,} \quad \sigma^2 \sim \text{IG}\left(\frac{n^{(0)}+p+2}{2}\,,\; n^{(0)}y_2^{(0)} - \frac{n^{(0)}}{2}\boldsymbol{y}_1^{(0)\mathsf{T}} n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{y}_1^{(0)}\right). \tag{A.12}$$

We have thus derived the CCCP distribution on $(\boldsymbol{\beta}, \sigma^2)$, which can be expressed either in terms of the canonical prior parameters $n^{(0)}$ and $\boldsymbol{y}^{(0)}$, or in terms of the prior parameters from Section A.1.3, $\overline{\boldsymbol{m}}^{(0)}$, $\overline{\mathbf{M}}^{(0)}$, $\overline{a}^{(0)}$ and $\overline{b}^{(0)}$.

As already noted, $\overline{\mathbf{M}}^{(0)} = \frac{n}{n^{(0)}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ can be seen as a restricted version of $\mathbf{M}^{(0)}$. $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ is known as the variance-covariance structure from the least squares estimate $\text{Var}(\boldsymbol{\beta}) = \hat{\sigma}^2_{\text{LS}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$, and is here the fixed prior variance-covariance structure for $\boldsymbol{\beta} \mid \sigma^2$. Confidence in the prior assignment is expressed by the choice of $n^{(0)}$: With $n^{(0)}$ chosen large relative to n, strong confidence in the prior assignment of $\overline{\boldsymbol{m}}^{(0)}$ can be expressed, whereas a low value of $n^{(0)}$ will result in a less pointed prior distribution on $\boldsymbol{\beta} \mid \sigma^2$.

$\overline{\boldsymbol{m}}^{(0)}$ can be chosen freely from $\mathbb{R}^p$, just like $\boldsymbol{m}^{(0)}$ for the SCP, because $\boldsymbol{y}_1^{(0)} \in \mathbb{R}^p$; values for $\overline{a}^{(0)}$ are instead restricted by $\overline{a}^{(0)} > p/2 + 1$, as $n^{(0)}$ must be postive.

The condition $y_2^{(0)} > 0$ does not actually restrict the choice of $\overline{b}^{(0)}$, as the second term $-\frac{1}{2}\overline{\boldsymbol{m}}^{(0)\mathsf{T}}\overline{\mathbf{M}}^{(0)^{-1}}\overline{\boldsymbol{m}}^{(0)}$, containing a quadratic form with a postive definite matrix, is always negative, such that the first term $n^{(0)}y_2^{(0)}$ must be positive anyway in order to allow a positive value for $\overline{b}^{(0)}$ (which is needed to make the prior proper).

As seen in Section 1.2.3.1, the update step for a canonically constructed prior, expressed in terms of $n^{(0)}$ and $\boldsymbol{y}^{(0)}$, possesses a convenient form: In the prior (A.8), the parameters $n^{(0)}$ and $\boldsymbol{y}^{(0)}$ must simply be replaced by their updated versions $n^{(n)}$ and $\boldsymbol{y}^{(n)}$, which calculate as

$$\begin{aligned} y_j^{(n)} &= \frac{n^{(0)}y_j^{(0)} + \tau(\boldsymbol{z})_j}{n^{(0)}+n}\,, \quad \forall j \in \{1,\ldots,p+1\}\,, \\ n^{(n)} &= n^{(0)} + n\,. \end{aligned}$$

In the following, we will describe what this means for the update steps of $\boldsymbol{\beta} \mid \sigma^2$, σ^2, and β, and compare these results with those for the SCP.

A.1.4.1. Update of $\boldsymbol{\beta} \mid \sigma^2$

As $\boldsymbol{y}^{(0)}$ and $\boldsymbol{y}^{(n)}$ are not directly interpretable, it is certainly easier to express prior beliefs on $\boldsymbol{\beta}$ via the mean vector $\overline{\boldsymbol{m}}^{(0)}$ of the prior distribution of $\beta \mid \sigma^2$ just as in the SCP model. As the transformation $\overline{\boldsymbol{m}}^{(0)} \mapsto \boldsymbol{y}^{(0)}$ is linear, this poses no problem:

$$\begin{aligned} \text{E}[\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{z}] = \overline{\boldsymbol{m}}^{(n)} &= n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\,\boldsymbol{y}_1^{(n)} \\ &= n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\left(\frac{n^{(0)}}{n^{(0)}+n}\,\boldsymbol{y}_1^{(0)} + \frac{n}{n^{(0)}+n}\cdot\frac{1}{n}(\mathbf{X}^\mathsf{T}\boldsymbol{z})\right) \\ &= n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\frac{n^{(0)}}{n^{(0)}+n}\cdot\frac{1}{n}(\mathbf{X}^\mathsf{T}\mathbf{X})\overline{\boldsymbol{m}}^{(0)} + n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\frac{n}{n^{(0)}+n}\cdot\frac{1}{n}(\mathbf{X}^\mathsf{T}\boldsymbol{z}) \\ &= \frac{n^{(0)}}{n^{(0)}+n}\,\text{E}[\boldsymbol{\beta} \mid \sigma^2] + \frac{n}{n^{(0)}+n}\hat{\boldsymbol{\beta}}_{\text{LS}}\,. \end{aligned} \tag{A.13}$$

The posterior expectation for $\boldsymbol{\beta} \mid \sigma^2$ is here a scalar-weighted mean of the prior expectation and the least squares estimate, with weights $n^{(0)}$ and n, respectively. The role of $n^{(0)}$ in the prior variance of $\boldsymbol{\beta} \mid \sigma^2$ is directly mirrored here. As described for the generalised setting in Section 1.2.3.1 (see also Section 3.3.2), $n^{(0)}$ can be seen as a parameter describing the "prior strength" or expressing "pseudocounts". In line with this interpretation, high values of $n^{(0)}$ as compared to n result here in a strong influence of $\overline{\boldsymbol{m}}^{(0)}$ for the calculation of $\overline{\boldsymbol{m}}^{(n)}$, whereas for small values of $n^{(0)}$, $\mathrm{E}[\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{z}]$ will be dominated by the value of $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$.

The variance of $\boldsymbol{\beta} \mid \sigma^2$ is updated as follows:

$$\mathrm{Var}(\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{z}) = \frac{n\sigma^2}{n^{(n)}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \frac{n\sigma^2}{n^{(0)} + n}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\,.$$

Here, $n^{(0)}$ is updated to $n^{(n)}$, and thus the posterior variances are automatically smaller than the prior variances, just as in the SCP model.

A.1.4.2. Update of σ^2

For the assignment of the parameters $\overline{a}^{(0)}$ and $\overline{b}^{(0)}$ to define the inverse gamma part of the joint prior, only $y_2^{(0)}$ is left to choose, as $n^{(0)}$ and $\boldsymbol{y}_1^{(0)}$ are already assigned via the choice of $\overline{\boldsymbol{m}}^{(0)}$ and $\overline{\mathbf{M}}^{(0)}$. To choose $y_2^{(0)}$, it is convenient to consider the prior expectation of σ^2 (alternatively, the prior mode of σ^2 could be considered as well):

$$\begin{aligned}
\mathrm{E}[\sigma^2] = \frac{\overline{b}^{(0)}}{\overline{a}^{(0)} - 1} &= \frac{n^{(0)} y_2^{(0)} - \frac{1}{2}\overline{\boldsymbol{m}}^{(0)\mathsf{T}}\overline{\mathbf{M}}^{(0)^{-1}}\overline{\boldsymbol{m}}^{(0)}}{\frac{n^{(0)}+p}{2} + 1 - 1} \\
&= \frac{2}{n^{(0)} + p}\left(n^{(0)} y_2^{(0)} - \frac{n^{(0)}}{2}\boldsymbol{y}_1^{(0)\mathsf{T}} n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{y}_1^{(0)}\right) \\
&= \frac{2n^{(0)}}{n^{(0)} + p} y_2^{(0)} - \frac{1}{n^{(0)} + p}\overline{\boldsymbol{m}}^{(0)\mathsf{T}}\overline{\mathbf{M}}^{(0)^{-1}}\overline{\boldsymbol{m}}^{(0)}\,.
\end{aligned}$$

A value of $y_2^{(0)}$ dependent on the value of $\mathrm{E}[\sigma^2]$ can thus be chosen by the linear mapping

$$y_2^{(0)} = \frac{n^{(0)} + p}{2n^{(0)}}\,\mathrm{E}[\sigma^2] + \frac{1}{2n^{(0)}}\overline{\boldsymbol{m}}^{(0)\mathsf{T}}\overline{\mathbf{M}}^{(0)^{-1}}\overline{\boldsymbol{m}}^{(0)}\,.$$

For the posterior expected value of σ^2, there is a similar decomposition as for the SCP model, and furthermore two other possible decompositions offering interesting interpretations of the update step of σ^2. The three decompositions are presented in the following.

Decomposition Including an Estimate of σ^2 Through the Null Model. In a first decomposition, the posterior variance of σ^2 can be written as:

$$\mathrm{E}[\sigma^2 \mid \boldsymbol{z}] = \frac{\overline{b}^{(n)}}{\overline{a}^{(n)} - 1} = \frac{2n^{(n)}}{n^{(n)} + p} y_2^{(n)} - \frac{1}{n^{(n)} + p}\overline{\boldsymbol{m}}^{(n)\mathsf{T}}\overline{\mathbf{M}}^{(n)^{-1}}\overline{\boldsymbol{m}}^{(n)}$$

$$
\begin{aligned}
&= \frac{2n^{(0)}}{n^{(0)}+n+p} y_2^{(0)} + \frac{1}{n^{(0)}+n+p} \boldsymbol{z}^\mathsf{T}\boldsymbol{z} - \frac{1}{n^{(0)}+n+p} \overline{\boldsymbol{m}}^{(n)\mathsf{T}} \overline{\mathbf{M}}^{(n)^{-1}} \overline{\boldsymbol{m}}^{(n)} \\
&= \frac{n^{(0)}+p}{n^{(0)}+n+p} \mathrm{E}[\sigma^2] + \frac{n-1}{n^{(0)}+n+p} \frac{1}{n-1} \boldsymbol{z}^\mathsf{T}\boldsymbol{z} \\
&\qquad + \frac{1}{n^{(0)}+n+p} \left(\overline{\boldsymbol{m}}^{(0)\mathsf{T}} \overline{\mathbf{M}}^{(0)^{-1}} \overline{\boldsymbol{m}}^{(0)} - \overline{\boldsymbol{m}}^{(n)\mathsf{T}} \overline{\mathbf{M}}^{(n)^{-1}} \overline{\boldsymbol{m}}^{(n)} \right),
\end{aligned}
\tag{A.14}
$$

and so can be seen as a weighted average of the prior expected value, $\frac{1}{n-1}\boldsymbol{z}^\mathsf{T}\boldsymbol{z}$, and a term depending on prior and posterior estimates for $\boldsymbol{\beta}$, with weights $n^{(0)}+p$, $n-1$ and 1, respectively. When adopting the centered $\boldsymbol{z}$, standardized $\mathbf{X}$ approach, $\frac{1}{n-1}\boldsymbol{z}^\mathsf{T}\boldsymbol{z}$ is the estimate for σ^2 under the null model, that is, if $\boldsymbol{\beta} = \mathbf{0}$.

Contrary to what a cursory inspection might suggest, the third term's influence, having the constant weight of 1, will not vanish for $n \to \infty$, as the third term does not approach a constant.[18]

The third term reflects the change in information about $\boldsymbol{\beta}$:

(i) If we are very uncertain about the prior beliefs on $\boldsymbol{\beta}$ as expressed in $\overline{\boldsymbol{m}}^{(0)}$ and thus assign a small value for $n^{(0)}$ as compared to n, we will get relatively large variances and covariances in $\overline{\mathbf{M}}^{(0)}$ by the factor $n/n^{(0)} > 1$ to $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$, resulting in a small term $\overline{\boldsymbol{m}}^{(0)\mathsf{T}} \overline{\mathbf{M}}^{(0)^{-1}} \overline{\boldsymbol{m}}^{(0)}$. After updating, the elements in $\overline{\mathbf{M}}^{(n)}$ become smaller automatically due to the updated factor $n/(n^{(0)}+n)$ to $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$.

If the values of $\overline{\boldsymbol{m}}^{(n)}$ do not differ much from the values in $\overline{\boldsymbol{m}}^{(0)}$, then the term $\overline{\boldsymbol{m}}^{(n)\mathsf{T}} \overline{\mathbf{M}}^{(n)^{-1}} \overline{\boldsymbol{m}}^{(n)}$ would be larger than its prior counterpart $\overline{\boldsymbol{m}}^{(0)\mathsf{T}} \overline{\mathbf{M}}^{(0)^{-1}} \overline{\boldsymbol{m}}^{(0)}$, ultimately reducing the posterior expectation for σ^2 through the third term being negative.

If $\overline{\boldsymbol{m}}^{(n)}$ does significantly differ from $\overline{\boldsymbol{m}}^{(0)}$, then the term $\overline{\boldsymbol{m}}^{(n)\mathsf{T}} \overline{\mathbf{M}}^{(n)^{-1}} \overline{\boldsymbol{m}}^{(n)}$ can actually be smaller than $\overline{\boldsymbol{m}}^{(0)\mathsf{T}} \overline{\mathbf{M}}^{(0)^{-1}} \overline{\boldsymbol{m}}^{(0)}$, and thus give a larger value of $\mathrm{E}[\sigma^2 \mid \boldsymbol{z}]$ as compared with the situation $\overline{\boldsymbol{m}}^{(n)} \approx \overline{\boldsymbol{m}}^{(0)}$.

(ii) On the contrary, large values for $n^{(0)}$ as compared to n, indicating high trust in prior beliefs on $\boldsymbol{\beta}$, lead to small variances and covariances in $\overline{\mathbf{M}}^{(0)}$ by the factor $n/n^{(0)}$ to $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$, resulting in a larger term $\overline{\boldsymbol{m}}^{(0)\mathsf{T}} \overline{\mathbf{M}}^{(0)^{-1}} \overline{\boldsymbol{m}}^{(0)}$ as compared to the case with low $n^{(0)}$. After updating, variances and covariances in $\overline{\mathbf{M}}^{(n)}$ will become even smaller, amplifying the term $\overline{\boldsymbol{m}}^{(n)\mathsf{T}} \overline{\mathbf{M}}^{(n)^{-1}} \overline{\boldsymbol{m}}^{(n)}$ even more if $\overline{\boldsymbol{m}}^{(n)} \approx \overline{\boldsymbol{m}}^{(0)}$, ultimately reducing the posterior expectation for σ^2 more than in the situation with low $n^{(0)}$.

[18] Although $\overline{\boldsymbol{m}}^{(n)}$ approaches $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, and $\overline{\boldsymbol{m}}^{(0)}$ is a constant, $\overline{\mathbf{M}}^{(0)^{-1}}$ and $\overline{\mathbf{M}}^{(n)^{-1}}$ are increasing for growing n, with $\overline{\mathbf{M}}^{(n)^{-1}}$ increasing faster than $\overline{\mathbf{M}}^{(0)^{-1}}$. The third term will thus eventually turn negative, reducing the null model variance that has weight $n-1$.

If, however, the values of $\overline{\boldsymbol{m}}^{(n)}$ do differ significantly from the values in $\overline{\boldsymbol{m}}^{(0)}$, the term $\overline{\boldsymbol{m}}^{(n)\mathsf{T}}\overline{\mathbf{M}}^{(n)^{-1}}\overline{\boldsymbol{m}}^{(n)}$ can be smaller than $\overline{\boldsymbol{m}}^{(0)\mathsf{T}}\overline{\mathbf{M}}^{(0)^{-1}}\overline{\boldsymbol{m}}^{(0)}$ also here, and even more so as compared to the situation with low $n^{(0)}$, giving eventually an even larger posterior expectation for σ^2.

Decomposition Similar to the SCP Model. A decomposition similar to the one in Section A.1.3.2 can be derived by considering the third term from (A.14) in more detail:

$$
\begin{aligned}
&\overline{\boldsymbol{m}}^{(0)\mathsf{T}}\overline{\mathbf{M}}^{(0)^{-1}}\overline{\boldsymbol{m}}^{(0)} - \overline{\boldsymbol{m}}^{(n)\mathsf{T}}\overline{\mathbf{M}}^{(n)^{-1}}\overline{\boldsymbol{m}}^{(n)} \\
&\quad= n^{(0)}\cdot n\cdot \boldsymbol{y}_1^{(0)\mathsf{T}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{y}_1^{(0)} - n^{(n)}\cdot n\cdot \boldsymbol{y}_1^{(n)\mathsf{T}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{y}_1^{(n)} \\
&\quad= n^{(0)}\cdot n\cdot \boldsymbol{y}_1^{(0)\mathsf{T}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{y}_1^{(0)} - (n^{(0)}+n)\cdot n\frac{n^{(0)}\boldsymbol{y}_1^{(0)\mathsf{T}} + \boldsymbol{z}^\mathsf{T}\mathbf{X}}{n^{(0)}+n}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\frac{n^{(0)}\boldsymbol{y}_1^{(0)} + \mathbf{X}^\mathsf{T}\boldsymbol{z}}{n^{(0)}+n} \\
&\quad= \left(n^{(0)}\cdot n - \frac{n\cdot n^{(0)^2}}{n^{(0)}+n}\right)\boldsymbol{y}_1^{(0)\mathsf{T}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{y}_1^{(0)} - \frac{2n^{(0)}\cdot n}{n^{(0)}+n}\boldsymbol{y}_1^{(0)\mathsf{T}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{z} \\
&\qquad\qquad - \frac{n}{n^{(0)}+n}\boldsymbol{z}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{z} \\
&\quad= \frac{n}{n^{(0)}+n}\left[\overline{\boldsymbol{m}}^{(0)\mathsf{T}}\overline{\mathbf{M}}^{(0)^{-1}}\overline{\boldsymbol{m}}^{(0)} - 2\overline{\boldsymbol{m}}^{(0)\mathsf{T}}\overline{\mathbf{M}}^{(0)^{-1}}\hat{\boldsymbol{\beta}}_{\mathrm{LS}} - \frac{n}{n^{(0)}}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}^\mathsf{T}\overline{\mathbf{M}}^{(0)^{-1}}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}\right] \\
&\quad= \frac{n}{n^{(0)}+n}\left[\left(\overline{\boldsymbol{m}}^{(0)} - \hat{\boldsymbol{\beta}}_{\mathrm{LS}}\right)^\mathsf{T}\overline{\mathbf{M}}^{(0)^{-1}}\left(\overline{\boldsymbol{m}}^{(0)} - \hat{\boldsymbol{\beta}}_{\mathrm{LS}}\right) - \left(\frac{n}{n^{(0)}}+1\right)\hat{\boldsymbol{\beta}}_{\mathrm{LS}}^\mathsf{T}\overline{\mathbf{M}}^{(0)^{-1}}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}\right] \\
&\quad= \frac{n}{n^{(0)}+n}\left(\overline{\boldsymbol{m}}^{(0)} - \hat{\boldsymbol{\beta}}_{\mathrm{LS}}\right)^\mathsf{T}\overline{\mathbf{M}}^{(0)^{-1}}\left(\overline{\boldsymbol{m}}^{(0)} - \hat{\boldsymbol{\beta}}_{\mathrm{LS}}\right) - \boldsymbol{z}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{z}\,.
\end{aligned}
$$

Thus, we get

$$
\begin{aligned}
\mathrm{E}[\sigma^2 \mid \boldsymbol{z}] &= \frac{n^{(0)}+p}{n^{(0)}+n+p}\,\mathrm{E}[\sigma^2] + \frac{1}{n^{(0)}+n+p}\left(\boldsymbol{z}^\mathsf{T}\boldsymbol{z} - \boldsymbol{z}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{z}\right) \\
&\quad + \frac{1}{n^{(0)}+n+p}\cdot\frac{n}{n^{(0)}+n}(\overline{\boldsymbol{m}}^{(0)} - \hat{\boldsymbol{\beta}}_{\mathrm{LS}})^\mathsf{T}\overline{\mathbf{M}}^{(0)^{-1}}(\overline{\boldsymbol{m}}^{(0)} - \hat{\boldsymbol{\beta}}_{\mathrm{LS}}) \\
&= \frac{n^{(0)}+p}{n^{(0)}+n+p}\,\mathrm{E}[\sigma^2] + \frac{n-p}{n^{(0)}+n+p}\cdot\underbrace{\frac{1}{n-p}(z - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}})^\mathsf{T}(z - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}})}_{\hat{\sigma}^2_{\mathrm{LS}}} \\
&\quad + \frac{p}{n^{(0)}+n+p}\cdot\underbrace{\frac{n}{n^{(0)}+n}\frac{1}{p}(\overline{\boldsymbol{m}}^{(0)} - \hat{\boldsymbol{\beta}}_{\mathrm{LS}})^\mathsf{T}\overline{\mathbf{M}}^{(0)^{-1}}(\overline{\boldsymbol{m}}^{(0)} - \hat{\boldsymbol{\beta}}_{\mathrm{LS}})}_{=:\overline{\sigma}^2_{\mathrm{PDC}}}\,. \qquad \text{(A.15)}
\end{aligned}
$$

The posterior expectation for σ^2 can therefore be seen also here as a weighted average of the prior expected value, the estimation $\hat{\sigma}^2_{\mathrm{LS}}$ resulting from least squares methods, and $\overline{\sigma}^2_{\mathrm{PDC}}$, an estimate for σ^2 similar to $\hat{\sigma}^2_{\mathrm{PDC}}$,[19] with weights $n^{(0)}+p$, $n-p$ and p, respectively.

[19] $\mathrm{E}[\overline{\sigma}^2_{\mathrm{PDC}} \mid \sigma^2] = \sigma^2$ computes very similar to the calculations given in O'Hagan (1994, p. 249) and is given below.

As in the update step for $\beta \mid \sigma^2$, $n^{(0)}$ is guarding the influence of the prior expectation on the posterior expectation. Just as in the decomposition for the SCP model, the weight for $\hat{\sigma}^2_{\text{LS}}$ will dominate the others when the sample size approaches infinity. Also for the CCCP model, $\overline{\sigma}^2_{\text{PDC}}$ is getting large if prior beliefs on β are skewed with respect to "what the data says", eventually inflating the posterior expectation of σ^2. The weighting of the differences is similar as well: High prior confidence in the chosen value of $\overline{\boldsymbol{m}}^{(0)}$ as expressed by a high value of $n^{(0)}$ will give a large $\overline{\mathbf{M}}^{(0)^{-1}}$, and thus penalising erroneous assignments stronger as compared to a lower value of $n^{(0)}$. Again, $\mathbf{X}^\mathsf{T}\mathbf{X}$, the matrix structure in $\overline{\mathbf{M}}^{(0)^{-1}}$, weighs the differences for components with covariates having a low spread weaker due to the instability of the respective component of $\hat{\boldsymbol{\beta}}_{\text{LS}}$ under such conditions.

Now we give the proof that $\mathrm{E}[\overline{\sigma}^2_{\text{PDC}} \mid \sigma^2] = \sigma^2$. As a preparation, it holds that

$$\begin{aligned}
\mathrm{E}[\hat{\boldsymbol{\beta}}_{\text{LS}} \mid \sigma^2] &= \mathrm{E}\left[\,\mathrm{E}[\hat{\boldsymbol{\beta}}_{\text{LS}} \mid \boldsymbol{\beta}, \sigma^2]\middle|\sigma^2\right] = \mathrm{E}[\beta \mid \sigma^2] = \overline{\boldsymbol{m}}^{(0)}\,, \quad \text{and} \\
\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\text{LS}} \mid \sigma^2) &= \mathrm{E}\left[\,\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\text{LS}} \mid \boldsymbol{\beta}, \sigma^2)\middle|\sigma^2\right] + \mathrm{Var}\left(\,\mathrm{E}[\hat{\boldsymbol{\beta}}_{\text{LS}} \mid \boldsymbol{\beta}, \sigma^2]\middle|\sigma^2\right) \\
&= \mathrm{E}[\sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \mid \sigma^2] + \mathrm{Var}(\boldsymbol{\beta} \mid \sigma^2) \\
&= \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} + \frac{n\sigma^2}{n^{(0)}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \frac{n^{(0)}+n}{n^{(0)}}\sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\,.
\end{aligned}$$

With this in mind, we can now derive

$$\begin{aligned}
&\mathrm{E}\left[\left(\overline{\boldsymbol{m}}^{(0)} - \hat{\boldsymbol{\beta}}_{\text{LS}}\right)^\mathsf{T}\overline{\mathbf{M}}^{(0)^{-1}}\left(\overline{\boldsymbol{m}}^{(0)} - \hat{\boldsymbol{\beta}}_{\text{LS}}\right)\middle|\sigma^2\right] \\
&\quad= \mathrm{E}\left[\,\mathrm{tr}\left(\overline{\mathbf{M}}^{(0)^{-1}}\left(\overline{\boldsymbol{m}}^{(0)} - \hat{\boldsymbol{\beta}}_{\text{LS}}\right)\left(\overline{\boldsymbol{m}}^{(0)} - \hat{\boldsymbol{\beta}}_{\text{LS}}\right)^\mathsf{T}\right)\middle|\sigma^2\right] \\
&\quad= \mathrm{tr}\left(\overline{\mathbf{M}}^{(0)^{-1}}\,\mathrm{E}\left[\left(\overline{\boldsymbol{m}}^{(0)} - \hat{\boldsymbol{\beta}}_{\text{LS}}\right)\left(\overline{\boldsymbol{m}}^{(0)} - \hat{\boldsymbol{\beta}}_{\text{LS}}\right)^\mathsf{T}\middle|\sigma^2\right]\right) \\
&\quad= \mathrm{tr}\left(\frac{n^{(0)}}{n}(\mathbf{X}^\mathsf{T}\mathbf{X}) \cdot \frac{n^{(0)}+n}{n^{(0)}}\sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right) \\
&\quad= \mathrm{tr}\left(\frac{n^{(0)}+n}{n}\sigma^2\mathbf{I}_p\right) = \frac{n^{(0)}+n}{n} \cdot p \cdot \sigma^2\,,
\end{aligned}$$

such that the factor $\frac{n}{n^{(0)}+n}\frac{1}{p}$ in $\overline{\sigma}^2_{\text{PDC}}$ cancels out, and indeed $\mathrm{E}[\overline{\sigma}^2_{\text{PDC}} \mid \sigma^2] = \sigma^2$.

Decomposition with Estimates of σ^2 Through Prior and Posterior Residuals. A third interpretation of $\mathrm{E}[\sigma^2 \mid \boldsymbol{z}]$ can be derived by another reformulation of the third term in (A.14):

$$\begin{aligned}
&\overline{\boldsymbol{m}}^{(0)^\mathsf{T}}\overline{\mathbf{M}}^{(0)^{-1}}\overline{\boldsymbol{m}}^{(0)} - \overline{\boldsymbol{m}}^{(n)^\mathsf{T}}\overline{\mathbf{M}}^{(n)^{-1}}\overline{\boldsymbol{m}}^{(n)} \\
&\quad= \frac{n^{(0)}}{n}\overline{\boldsymbol{m}}^{(0)^\mathsf{T}}\mathbf{X}^\mathsf{T}\mathbf{X}\overline{\boldsymbol{m}}^{(0)} - \frac{n^{(n)}}{n}\overline{\boldsymbol{m}}^{(n)^\mathsf{T}}\mathbf{X}^\mathsf{T}\mathbf{X}\overline{\boldsymbol{m}}^{(n)} \\
&\quad= \frac{n^{(0)}}{n}(\boldsymbol{z} - \mathbf{X}\overline{\boldsymbol{m}}^{(0)})^\mathsf{T}(\boldsymbol{z} - \mathbf{X}\overline{\boldsymbol{m}}^{(0)}) - \frac{n^{(n)}}{n}(\boldsymbol{z} - \mathbf{X}\overline{\boldsymbol{m}}^{(n)})^\mathsf{T}(\boldsymbol{z} - \mathbf{X}\overline{\boldsymbol{m}}^{(n)})
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{n^{(n)}}{n}\boldsymbol{z}^\mathsf{T}\boldsymbol{z} - \frac{n^{(0)}}{n}\boldsymbol{z}^\mathsf{T}\boldsymbol{z} + \frac{n^{(0)}}{n}2\boldsymbol{z}^\mathsf{T}\mathbf{X}\overline{\boldsymbol{m}}^{(0)} - \frac{n^{(n)}}{n}2\boldsymbol{z}^\mathsf{T}\mathbf{X}\overline{\boldsymbol{m}}^{(n)} \\
&= \frac{n^{(0)}}{n}(z - \mathbf{X}\overline{\boldsymbol{m}}^{(0)})^\mathsf{T}(z - \mathbf{X}\overline{\boldsymbol{m}}^{(0)}) - \frac{n^{(n)}}{n}(z - \mathbf{X}\overline{\boldsymbol{m}}^{(n)})^\mathsf{T}(z - \mathbf{X}\overline{\boldsymbol{m}}^{(n)}) + \boldsymbol{z}^\mathsf{T}\boldsymbol{z} - 2\boldsymbol{z}^\mathsf{T}\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{LS}}\,.
\end{aligned}
$$

With this, we get

$$
\begin{aligned}
\mathrm{E}[\sigma^2 \mid \boldsymbol{z}] = \frac{n^{(0)}+p}{n^{(0)}+n+p}\,\mathrm{E}[\sigma^2] + \frac{n^{(0)}+p}{n^{(0)}+n+p}\underbrace{\frac{n^{(0)}}{n}\cdot\frac{1}{n^{(0)}+p}(\boldsymbol{z} - \mathbf{X}\overline{\boldsymbol{m}}^{(0)})^\mathsf{T}(\boldsymbol{z} - \mathbf{X}\overline{\boldsymbol{m}}^{(0)})}_{=:{\sigma^{(0)}}^2,\ \text{as}\ \mathrm{E}[{\sigma^{(0)}}^2\mid\sigma^2]=\sigma^2} \\
+ \frac{2(n-p)}{n^{(0)}+n+p}\hat{\sigma}^2_{\text{LS}} - \frac{n^{(n)}+p}{n^{(0)}+n+p}\underbrace{\frac{n^{(n)}}{n}\cdot\frac{1}{n^{(n)}+p}(\boldsymbol{z} - \mathbf{X}\overline{\boldsymbol{m}}^{(n)})^\mathsf{T}(\boldsymbol{z} - \mathbf{X}\overline{\boldsymbol{m}}^{(n)})}_{=:{\sigma^{(n)}}^2,\ \text{as}\ \mathrm{E}[{\sigma^{(n)}}^2\mid\sigma^2,\boldsymbol{z}]=\mathrm{E}[{\sigma^{(n)}}^2\mid\sigma^2]=\sigma^2}\,.
\end{aligned}
\tag{A.16}
$$

Here, the calculation of $\mathrm{E}[\sigma^2 \mid \boldsymbol{z}]$ is based again on $\mathrm{E}[\sigma^2]$ and $\hat{\sigma}^2_{\text{LS}}$, but now complemented with two special estimates: ${\sigma^{(0)}}^2$, an estimate based on the "prior residuals" $z - \mathbf{X}\overline{\boldsymbol{m}}^{(0)}$, and a respective posterior version ${\sigma^{(n)}}^2$, based on $z - \mathbf{X}\overline{\boldsymbol{m}}^{(n)}$.

However, $\mathrm{E}[\sigma^2 \mid \boldsymbol{z}]$ is only "almost" a weighted average of these ingredients, as the weights sum up to $n^{(0)} - p + n$ instead of $n^{(0)} + p + n$. Especially strange is the negative weight for ${\sigma^{(n)}}^2$, actually making the factor to ${\sigma^{(n)}}^2$ in (A.16) result to -1.

A possible interpretation would be to group $\mathrm{E}[\sigma^2]$ and ${\sigma^{(0)}}^2$ as prior-based estimations with joint weight $2(n^{(0)} + p)$, and $\hat{\sigma}^2_{\text{LS}}$ as data-based estimation with weight $2(n - p)$. Together, these estimations have a weight of $2(n^{(0)} + n)$, being almost (neglecting the missing $2p$) a "double estimate" that is corrected back to a "single" estimate with the posterior-based estimate ${\sigma^{(n)}}^2$.

A.1.4.3. Update of $\boldsymbol{\beta}$

As for the SCP model, the posterior on $\boldsymbol{\beta}$, being the most relevant distribution for inferences, is a multivariate t with expectation $\overline{\boldsymbol{m}}^{(n)}$ as described in Section A.1.4.1. For $\mathrm{Var}(\beta \mid \boldsymbol{z})$, one gets different formulations, depending on the formula for $\mathrm{E}[\sigma^2 \mid \boldsymbol{z}]$:

$$
\mathrm{Var}(\boldsymbol{\beta} \mid \boldsymbol{z}) \;=\; \frac{\overline{b}^{(n)}}{\overline{a}^{(n)} - 1}\overline{\mathbf{M}}^{(n)} = \mathrm{E}[\sigma^2 \mid \boldsymbol{z}]\,\frac{n}{n^{(n)}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \tag{A.17}
$$

$$
\begin{aligned}
\overset{(A.14)}{=}\; &\frac{n^{(0)}+p}{n^{(0)}+n+p}\frac{n^{(0)}}{n^{(n)}}\underbrace{\mathrm{E}[\sigma^2]\frac{n}{n^{(0)}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}}_{\mathrm{Var}(\boldsymbol{\beta})} + \frac{n-1}{n^{(0)}+n+p}\frac{n}{n^{(n)}}\frac{1}{n-1}\boldsymbol{z}^\mathsf{T}\boldsymbol{z}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \\
&+ \frac{1}{n^{(0)}+n+p}\frac{n}{n^{(n)}}\left(\overline{\boldsymbol{m}}^{(0)\mathsf{T}}\overline{\mathbf{M}}^{(0)^{-1}}\overline{\boldsymbol{m}}^{(0)} - \overline{\boldsymbol{m}}^{(n)\mathsf{T}}\overline{\mathbf{M}}^{(n)^{-1}}\overline{\boldsymbol{m}}^{(n)}\right)(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}
\end{aligned}
$$

$$\overset{(A.15)}{=} \frac{n^{(0)}+p}{n^{(0)}+n+p}\frac{n^{(0)}}{n^{(n)}}\underbrace{\mathrm{E}[\sigma^2]\frac{n}{n^{(0)}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}}_{\mathrm{Var}(\boldsymbol{\beta})} + \frac{n-p}{n^{(0)}+n+p}\frac{n}{n^{(n)}}\underbrace{\hat{\sigma}^2_{\mathrm{LS}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}}_{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{LS}})} + \frac{p}{n^{(0)}+n+p}\frac{n}{n^{(n)}}\overline{\sigma}^2_{\mathrm{PDC}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$$

$$\overset{(A.16)}{=} \frac{n^{(0)}+p}{n^{(0)}+n+p}\frac{n^{(0)}}{n^{(n)}}\underbrace{\mathrm{E}[\sigma^2]\frac{n}{n^{(0)}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}}_{\mathrm{Var}(\boldsymbol{\beta})} + \frac{n^{(0)}+p}{n^{(0)}+n+p}\frac{n^{(0)}}{n^{(n)}}\underbrace{\sigma^{(0)2}\frac{n}{n^{(0)}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}}_{=:\mathrm{Var}^{(0)}(\boldsymbol{\beta})} + \frac{2(n-p)}{n^{(0)}+n+p}\frac{n}{n^{(n)}}\underbrace{\hat{\sigma}^2_{\mathrm{LS}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}}_{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{LS}})} - \frac{n^{(n)}+p}{n^{(0)}+n+p}\underbrace{\sigma^{(n)2}\frac{n}{n^{(n)}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}}_{=:\mathrm{Var}^{(n)}(\boldsymbol{\beta})}.$$

In these equations, it is possible to isolate $\mathrm{Var}(\boldsymbol{\beta})$, $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{LS}})$ and, in the formulation with (A.16), the newly defined $\mathrm{Var}^{(0)}(\boldsymbol{\beta})$ and $\mathrm{Var}^{(n)}(\boldsymbol{\beta})$. However, all three versions do not constitute a weighted average, even when the corresponding formula for $\mathrm{E}[\sigma^2 \mid \boldsymbol{z}]$ has this property.

Just as in the SCP model, $\mathrm{Var}(\beta \mid \boldsymbol{z})$ can increase if the automatic decrease of the elements in $\overline{\mathbf{M}}^{(n)}$ is overcompensated by a strong increase of $\mathrm{E}[\sigma^2 \mid \boldsymbol{z}]$. Again, this reaction to prior-data conflict is unspecific because it depends on $\mathrm{E}[\sigma^2 \mid \boldsymbol{z}]$ alone, and affects all elements of variance-covariance matrix in the same way.

A.1.5. Discussion and Outlook

For both the SCP and CCCP model, $\mathrm{E}[\boldsymbol{\beta} \mid \boldsymbol{z}]$ can be seen as a weighted average of $\mathrm{E}[\boldsymbol{\beta}]$ and $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, such that the posterior distribution on $\boldsymbol{\beta}$ will be centered around a mean somewhere between $\mathrm{E}[\boldsymbol{\beta}]$ and $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, with the location depending on the respective weights. The weights for the CCCP model appear especially intuitive: $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ is weighted with the sample size n, whereas $\mathrm{E}[\boldsymbol{\beta}]$ has the weight $n^{(0)}$, reflecting the "prior strength" or "pseudocounts".

Due to this, prior-data conflict may at most affect the variances only. Indeed, for both prior models, $\mathrm{E}[\sigma^2 \mid \boldsymbol{z}]$ can increase in the presence of prior-data conflict, as shown by the decompositions in Sections A.1.3.2 and A.1.4.2.

Through the formulations (A.6) and (A.17) for $\mathrm{Var}(\boldsymbol{\beta} \mid \boldsymbol{z})$, respectively, it can be seen that the posterior distribution on $\boldsymbol{\beta}$ can in fact become less pointed than the prior when prior-data conflict is present. Nevertheless, the effect might not be as strong as desired: In the formulations (A.5) and (A.15), respectively, the effect is based only on one term of the decomposition, and furthermore may be foiled through the automatic decrease of $\mathbf{M}^{(n)}$ and $\overline{\mathbf{M}}^{(n)}$. Probably the most problematic finding is that this (possibly weak) reaction affects the whole variance-covariance matrix uniformally, and thus, in both models, the reaction to prior-data conflict is by no means component-specific.

Therefore, the prior models lack the capability to mirror the appropriateness of the prior assignments for each covariate individually. As the SCP model is already the most general

approach in the class of conjugate priors, this non-specificity feature seems inevitable in Bayesian linear regression based on precise conjugate priors.

In fact, as argued in Section A.1.1, a more sophisticated and specific reaction to prior-data conflict is only possible by extending considerations beyond the traditional concept of probability. Imprecise probabilities, as a general methodology to cope with the multidimensional nature of uncertainty, appears promising here. For generalised Bayesian approaches, the possibility to mirror the quality of prior knowledge is one of the main reasons for the paradigmatic skip from classical probability to interval or imprecise probability.[20] In this framework, ambiguity in the prior specification can be modeled by considering sets $\mathcal{M}$ of prior distributions. In the most common approach based on Walley's (1991) Generalised Bayes' Rule (see Section 2.1.2.5), posterior inference is then based on a set of posterior distributions $\mathcal{M}_{|\boldsymbol{z}}$, resulting from updating the distributions in the prior set element by element.[21]

Of particular computational convenience are again models based on conjugate priors, as developed for the Dirichlet-Multinomial model by Walley (1996a), see also Bernard (2009),[22] and for i.i.d. exponential family sampling models by Quaeghebeur and Cooman (2005), which were extended by Walter and Augustin (2009b)[23] to allow an elegant handling of prior-data conflict: With the magnitude of the set $\mathcal{M}_{|\boldsymbol{z}}$ mapping the posterior ambiguity, high prior-data conflict leads, ceteris paribus, to a large $\mathcal{M}_{|\boldsymbol{z}}$, resulting in high imprecision in the posterior probabilities, and cautious inferences based on it, while in the case of no prior-data conflict $\mathcal{M}_{|\boldsymbol{z}}$, and thus the imprecision, is much smaller.[24]

The essential technical ingredient to derive this class of models is the general construction principle, described in Section 1.2.3.1, underlying the CCCP model from Section A.1.4, and thus that model can be extended directly to a powerful corresponding imprecise probability model.[25] A detailed development is beyond the scope of this contribution.

[20] See the discussion of this motive in Section 2.2.3.3.

[21] See Section 2.1.3.

[22] The IDM is discussed in more detail in Section 3.1.3.

[23] See Section 3.3 for a reproduction of this work.

[24] See the overview on imprecise probability models based on this class of conjugate priors in Section 3.1.

[25] For σ^2 fixed, the model from Section A.1.3 can also be comprised under the more general structure described in Section 3.3.2, that also can be extended to imprecise probabilities, see Walter, Augustin, and Peters (2007), and Walter (2006) for details.

A.2. A Parameter Set Shape for Strong Prior-Data Agreement Modelling

In this section, we will explain an approach for parameter set shapes that allow for extra precision in case of strong prior-data agreement as discussed in Section 4.3. First, we will briefly characterise the novel parametrisation of canoncial conjugate priors this approach relies on. To keep things simple, we restrict ourselves here for the case of the Beta-Binomial model (see Section 1.2.3.3), but the approach is generalisable to arbitrary canonical conjugate priors.[26] Then we will suggest a shape in this parametrisation that accomplishes both prior-data conflict sensitivity and 'bonus precision' in case of strong prior-data agreement. We present a parametric description for such a shape and show that it will indeed lead to the desired properties.

A.2.1. A Novel Parametrisation of Canonical Conjugate Priors

In the parametrisation in terms of $n^{(0)}$ and $y^{(0)}$, described in Section 1.2.3.1, a conjugate prior is updated to its respective posterior by a shift in the parameter space, given by (1.6):

$$n^{(0)} \mapsto n^{(0)} + n\,, \qquad y^{(0)} \mapsto \frac{n^{(0)}}{n^{(0)}+n} \cdot y^{(0)} + \frac{n}{n^{(0)}+n} \cdot \frac{\tau(\boldsymbol{x})}{n} = y^{(0)} + \frac{\tau(\boldsymbol{x}) - n y^{(0)}}{n^{(0)}+n}\,.$$

We see thus that, while the shift for the n coordinate is the same for all elements $(n^{(0)}, y^{(0)})$ in a prior parameter set $\mathbb{\Pi}^{(0)}$, the shift in the y coordinate depends on $n^{(0)}$, and the location of $y^{(0)}$ itself. Due to this, the shape of $\mathbb{\Pi}^{(0)}$ will change during the update step.

This shape change to some extent obscures the posterior inference properties of a certain shape of the prior parameter set $\mathbb{\Pi}^{(0)}$. Therefore, a different parametrisation of the canonical priors in which each coordinate has the same shift in updating would be advantageous. Then, updating of parameter sets could be expressed as a shift of the entire set within the parameter space.

A parametrisation developed by Miķelis Bickis (2011, personal communication) achieves just that. He is currently preparing a manuscript elaborating the details of his findings, and we will present here a preview on the results for the Beta-Binomial case.

In this parametrisation, a canonical prior is represented by a coordinate $(\eta_0^{(0)}, \eta_1^{(0)})$,[27] where $\eta_1^{(0)}$ replaces the main prior parameter $y^{(0)}$, while $\eta_0^{(0)}$ is just a shifted version of $n^{(0)}$. The relation of $(\eta_0^{(0)}, \eta_1^{(0)})$ to $(n^{(0)}, y^{(0)})$ is as follows:

$$n^{(0)} = \eta_0^{(0)} + 2\,, \qquad\qquad y^{(0)} = \frac{\eta_1^{(0)}}{\eta_0^{(0)}+2} + \frac{1}{2}\,. \tag{A.18}$$

[26] For a more detailed derivation of this parametrisation, we have to refer to a future publication of Miķelis Bickis.

[27] Again, we denote prior parameters with superscript $^{(0)}$, and posterior parameters with superscript $^{(n)}$.

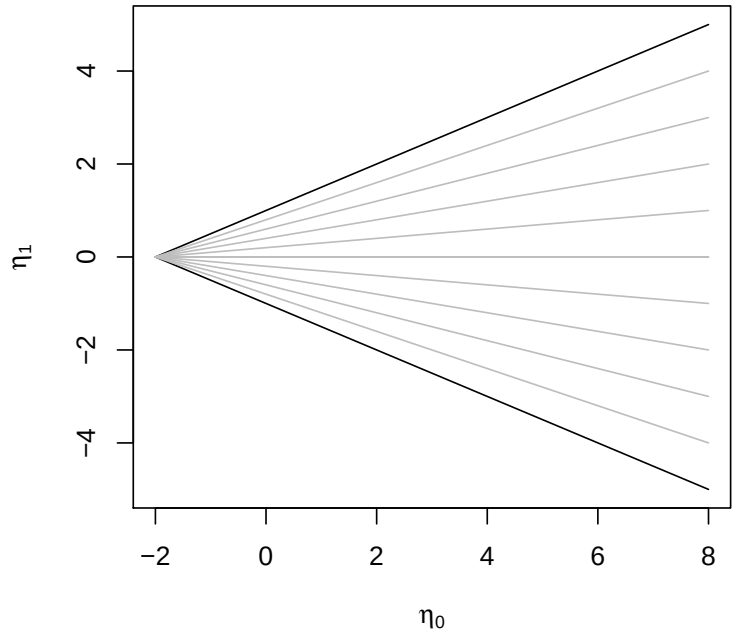

Figure A.1.: Bounds for the domain of η_0 and η_1 for the Beta-Binomial model (black), with rays of constant expectation for $y_c = \{0.1, 0.2, \ldots, 0.9\}$ (grey).

The domain of η_0 and η_1 in case of the Beta-Binomial model is

$$\mathrm{H} = \left\{ (\eta_0, \eta_1) \middle| \eta_0 > -2,\ |\eta_1| < \frac{1}{2}(\eta_0 + 2) \right\}, \tag{A.19}$$

and the update step in terms of η_0 and η_1 is given by

$$\begin{aligned} \eta_0^{(n)} &= \eta_0^{(0)} + n\,, \\ \eta_1^{(n)} &= \eta_1^{(0)} + \frac{1}{2}(s - (n - s)) = \eta_1^{(0)} + s - \frac{n}{2}\,, \end{aligned} \tag{A.20}$$

where s is the number of successes in the n Bernoulli trials. A 'success' in a Bernoulli trial thus leads to a step of 1 in the η_0 direction and of $+\frac{1}{2}$ in the η_1 direction, while a 'failure' leads to step of 1 in the η_0 direction and of $-\frac{1}{2}$ in the η_1 direction.[28]

As we wrote in Section 4.3, η_1 cannot have the convenient property of being equal to the expectation of the mean sample statistic $\tilde{\tau}(\boldsymbol{x})$ (here, s/n), as was the case for y.[29] However, from (A.18) we can derive that coordinates $(\eta_0, \eta_1) \in \mathrm{H}$ satifying

$$\eta_1 = f(\eta_0) = (\eta_0 + 2)(y_c - \frac{1}{2}) \tag{A.21}$$

[28] As in Section 3.5, we will treat s as a a real-valued observation in $[0, n]$ because the continuous representation is convenient for our discussions, keeping in mind that in reality it can only take integer values.

[29] For $y^{(0)}$, it holds that $y^{(0)} = \mathrm{E}\left[\mathrm{E}[\tilde{\tau}(\boldsymbol{x}) \mid \psi] \mid n^{(0)}, y^{(0)}\right]$, as mentioned in Section 1.2.3.1.

will have a constant expectation y_c. The domain H, and these *rays of constant expectation* emanating from the coordinate $(-2, 0)$, are depicted in Figure A.1.

A.2.2. Informal Rationale for Boat-Shaped Parameter Sets

When Mikelis Bickis presented this parametrisation of conjugate priors at a talk (Bickis 2011), both Frank Coolen and the author of this thesis had independently the same basic idea for a set shape that allows for both prior-data conflict sensitivity and more precise inferences in case of strong prior-data agreement. The basic idea for this shape is described informally below, while a suggestion for a parametrisation of such a shape is described and discussed in Section A.2.3.

In the parametrisation in terms of $(n^{(0)}, y^{(0)})$ and $(n^{(n)}, y^{(n)})$, posterior inferences become more precise, because the stretch in the main parameter dimension y, denoted by $\Delta_y(\amalg^{(n)})$, tends to 0 for $n \to \infty$ (see the discussion in Section 3.1.2). In the domain H as depicted in Figure A.1, instead the rays of constant expectation fan out for growing n, while a parameter set will retain its size in updating. Increased precision in a posterior parameter set $\mathrm{H}^{(n)}$, which is just its prior counterpart $\mathrm{H}^{(0)}$ shifted to the right, is given by the fact the more $\mathrm{H}^{(n)}$ is located to the right, the fewer rays of constant expectation $\mathrm{H}^{(n)}$ will intercept. Imprecision in terms of $\mathrm{E}\left[\mathrm{E}[\tilde{\tau}(\boldsymbol{x}) \mid \psi] \mid n^{(n)}, y^{(n)}\right] = y^{(n)}$ can thus be imagined as the size of the 'shadow' that a set $\mathrm{H}^{(n)}$ casts when considering a light source in $(-2, 0)$ (the point from which the rays of constant expectation emanate). In short, the smaller this shadow, the more precise the inferences.

In the context of the model from Section A.2.1, we will denote by $\underline{y}^{(n)}$ and $\overline{y}^{(n)}$ the bounds of this shadow, i.e.,

$$\underline{y}^{(n)} := \min_{(\eta_0^{(n)}, \eta_1^{(n)}) \in \mathrm{H}^{(n)}} y^{(n)} = \min_{(\eta_0^{(n)}, \eta_1^{(n)}) \in \mathrm{H}^{(n)}} \frac{\eta_1^{(n)}}{\eta_0^{(n)} + 2} + \frac{1}{2},$$

$$\overline{y}^{(n)} := \max_{(\eta_0^{(n)}, \eta_1^{(n)}) \in \mathrm{H}^{(n)}} y^{(n)} = \max_{(\eta_0^{(n)}, \eta_1^{(n)}) \in \mathrm{H}^{(n)}} \frac{\eta_1^{(n)}}{\eta_0^{(n)} + 2} + \frac{1}{2},$$

and we call the coordinates $\arg\min_{(\eta_0, \eta_1) \in \mathrm{H}^{(n)}} y^{(n)}$ and $\arg\max_{(\eta_0, \eta_1) \in \mathrm{H}^{(n)}} y^{(n)}$ the *lower* and *upper touchpoints* of $\mathrm{H}^{(n)}$ responsible for the shadow $[\underline{y}^{(n)}, \overline{y}^{(n)}]$. Mutatis mutandis, the same definitions can be made for the prior set $\mathrm{H}^{(0)}$.

Due to the fanning out of rays, most shapes for $\mathrm{H}^{(0)}$ will lead to decreasing imprecision for increasing n. Indeed, models of type (a) from Section 3.1.1, where $\amalg^{(0)} = n^{(0)} \times [\underline{y}^{(0)}, \overline{y}^{(0)}]$, are represented here again by a line segment $\mathrm{H}^{(0)} = \eta_0^{(0)} \times [\underline{\eta}_1^{(0)}, \overline{\eta}_1^{(0)}]$, such that the posterior touchpoints are, for any s and n, $(\eta_0^{(n)}, \underline{\eta}_1^{(n)})$ and $(\eta_0^{(0)}, \overline{\eta}_1^{(n)})$, where $\underline{\eta}_1^{(n)}$ and $\overline{\eta}_1^{(n)}$ are the updated versions of $\underline{\eta}_1^{(0)}$ and $\overline{\eta}_1^{(0)}$, respectively. Due to (A.20), it holds that $\overline{\eta}_1^{(n)} - \underline{\eta}_1^{(n)} = \overline{\eta}_1^{(0)} - \underline{\eta}_1^{(0)}$; therefore, imprecision decreases here because a line segment of fixed size will cast a smaller shadow when further to the right, as illustrated in Figure A.2.

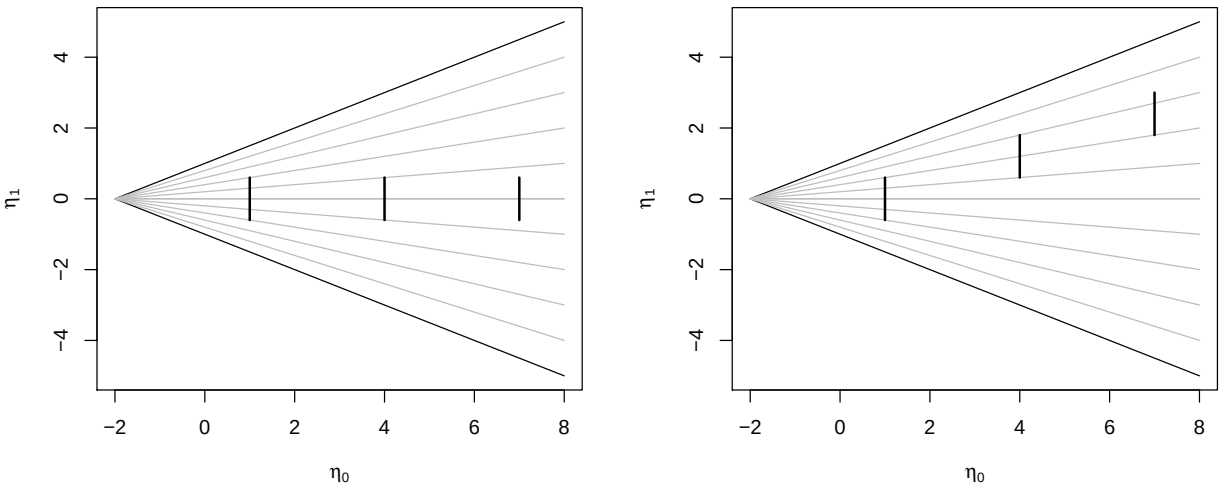

Figure A.2.: Parameter set $\mathrm{H}^{(0)} = \eta_0^{(0)} \times [\underline{\eta}_1^{(0)}, \overline{\eta}_1^{(0)}]$ and respective posterior sets $\mathrm{H}^{(n)}$ for $s/n = 0.5$ (left) and $s/n = 0.9$ (right). Note that all sets have the same size, imprecision decreasing only through their position on the η_0 axis.

For prior-data conflict sensitivity, we need shapes that cover a range of η_0 values, for the same reasons as in the framework of Section 3.1.1, where only sets with a range of $n^{(0)}$ values offered this property. Sets that are elongated along the rays of constant expectation will behave here similar to the rectangular shapes of Section 3.1.1. When shifted along its respective ray of constant expectation, imprecision will be reduced as the shadow of the set will become smaller just as described above for line segments. When such a shape is instead shifted away from its ray of constant expectation, imprecision will be increased, as a prolonged shape that is now turned away from its ray will cast a larger shadow.[30]

A set $\mathrm{H}^{(0)}$ allowing for less imprecison in case of strong prior-data agreement must also be able to cast a smaller shadow if the update shift goes into the direction of its ray, but we will enhance this effect by considering now also the properties of the canonical posteriors the coordinates of $\mathrm{H}^{(n)}$ represent.

We have seen that for the conjugate distributions themselves, $n^{(0)}$ is generally a parameter determining the spread of the distribution (e.g., in the Normal-Normal model (see Section 1.2.3.4), $n^{(0)}$ was the inverse variance), such that we will have more precise inferences if the shadow bounds $\underline{y}^{(n)}$ and $\overline{y}^{(n)}$ are attained at higher values of η_0, leading to lower variances in the 'critical' distributions at the boundary of the posterior expectation interval $[\underline{y}^{(n)}, \overline{y}^{(n)}]$. For this to happen, we need a shape for which the touchpoints responsible for $\underline{y}^{(n)}$ and $\overline{y}^{(n)}$ are attained at higher values of η_0 in case of strong prior-data agreement. Shapes that accomplish this must have a curvature along their length in the direction of the constant rays of expectation. The shape we suggest thus looks like a bullett, or like a boat with a transom stern (see, e.g., Figure A.3).

[30] This will become clear from the depiction of boatshape sets in Figure A.4.

A.2.3. The Boatshape

In this section, we will suggest a parametrisation for such a shape. The definition, along with some first graphical examples, is given in Section A.2.3.1, and we discuss some first technical results for this shape in Sections A.2.3.2 – A.2.3.4.

A.2.3.1. Basic Definition

We will now present a parametrisation for such a boat-shaped parameter set $\mathrm{H}^{(0)}$. To keep things simple, we will consider here and in the follwing only prior sets that are symmetric around the η_0 axis, i.e., centered around $y_c = 0.5$, expressing the prior information that we deem a fraction of successes of $\frac{s}{n} = \frac{1}{2}$ as the most probable.[31]

For the contours of $\mathrm{H}^{(0)}$, we suggest an exponential function as the functional form, where the 'prow' of the set is located at $(\underline{\eta}_0, 0)$. The lower and the upper contour $\underline{c}^{(0)}(\eta_0)$ and $\overline{c}^{(0)}(\eta_0)$ are defined as

$$\begin{aligned} \underline{c}^{(0)}(\eta_0) &= -a\left(1 - e^{-b(\eta_0 - \underline{\eta}_0)}\right), \\ \overline{c}^{(0)}(\eta_0) &= \quad a\left(1 - e^{-b(\eta_0 - \underline{\eta}_0)}\right), \end{aligned}$$

where a and b are parameters controlling the shape. We will also need the respective derivations with respect to η_0, given by

$$\begin{aligned} \frac{d}{d\eta_0}\underline{c}^{(0)}(\eta_0) &= -abe^{-b(\eta_0 - \underline{\eta}_0)}, \\ \frac{d}{d\eta_0}\overline{c}^{(0)}(\eta_0) &= \quad abe^{-b(\eta_0 - \underline{\eta}_0)}. \end{aligned}$$

For this basic situation, given the parameters $\underline{\eta}_0$, $\overline{\eta}_0$, a, and b, $\mathrm{H}^{(0)}$ is thus defined as

$$\mathrm{H}^{(0)} = \{(\eta_0, \eta_1) \colon \underline{\eta}_0 \leq \eta_0 \leq \overline{\eta}_0, \underline{c}^{(0)}(\eta_0) \leq \eta_1 \leq \overline{c}^{(0)}(\eta_0)\}. \tag{A.22}$$

A prior boatshape set with $\underline{\eta}_0 = 1$, $\overline{\eta}_0 = 6$, $a = 2$, and $b = 0.8$ is depicted in Figure A.3, where the left graph shows this set as defined in terms of (η_0, η_1), and the right graph shows the set from the left transformed into the space $\mathcal{N} \times \mathcal{Y}$.

We have as yet no appealing formal description for the role of the parameters a and b. Informally, a determines the half-width of the set; the width, i.e., the size in the η_1 dimension, would be a if $\overline{\eta}_0 \to \infty$. b instead determines the 'bulkyness' of the shape. Together with $\underline{\eta}_0$, a and b determine the prior interval for the expected success probability $[\underline{y}^{(0)}, \overline{y}^{(0)}]$. For fixed $\underline{\eta}_0$ and a, increasing b leads to a wider prior expectation interval. For $[\underline{y}^{(0)}, \overline{y}^{(0)}]$, the choice of $\overline{\eta}_0$ is irrelevant.[32]

[31] The general case of sets $\mathrm{H}^{(0)}$ with central ray $y_c \neq 0.5$ is discussed informally in Section A.2.4.

[32] $\overline{\eta}_0$ plays only a role in determining when the 'unhappy learning' phase starts (see end of Section A.2.3.4).

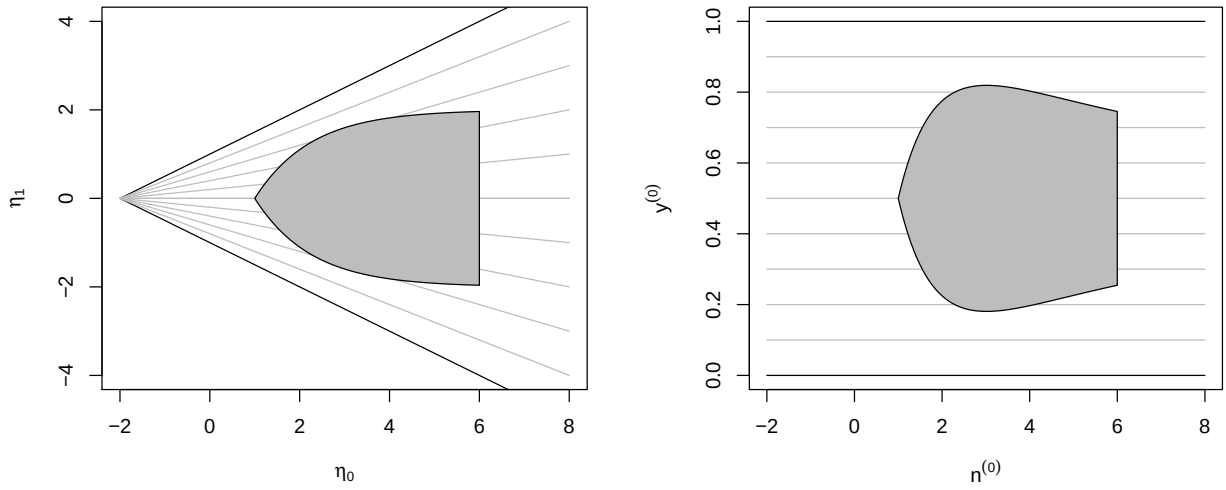

Figure A.3.: Boatshape prior set in the parametrisation via (η_0, η_1) (left) and via $(n^{(0)}, y^{(0)})$ (right), with parameters $\underline{\eta}_0 = 1$, $\overline{\eta}_0 = 6$, $a = 2$, and $b = 0.8$.

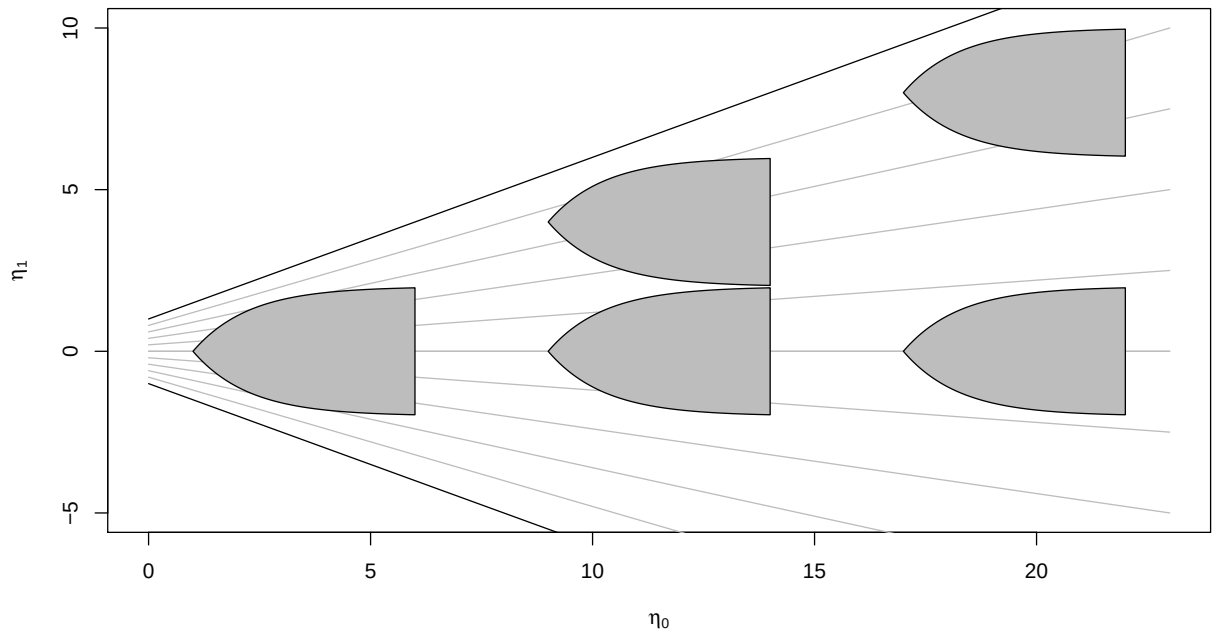

Figure A.4.: Boatshape prior and posterior sets for data in accordance and in conflict with the prior. The prior set is the same as in Figure A.3. While the posterior sets for $\frac{s}{n} = 0.5$ move along the ray for $y_c = 0.5$, the posterior sets for $\frac{s}{n} = 1$ are shifted away from the ray for $y_c = 0.5$, resulting in increased posterior imprecision. Note that lower and upper touchpoints are in the middle of the contour for the prior and the posterior resulting for data $\frac{s}{n} = \frac{4}{8}$, while at least one touchpoint is at the end for all other sets. (see also Figure A.5).

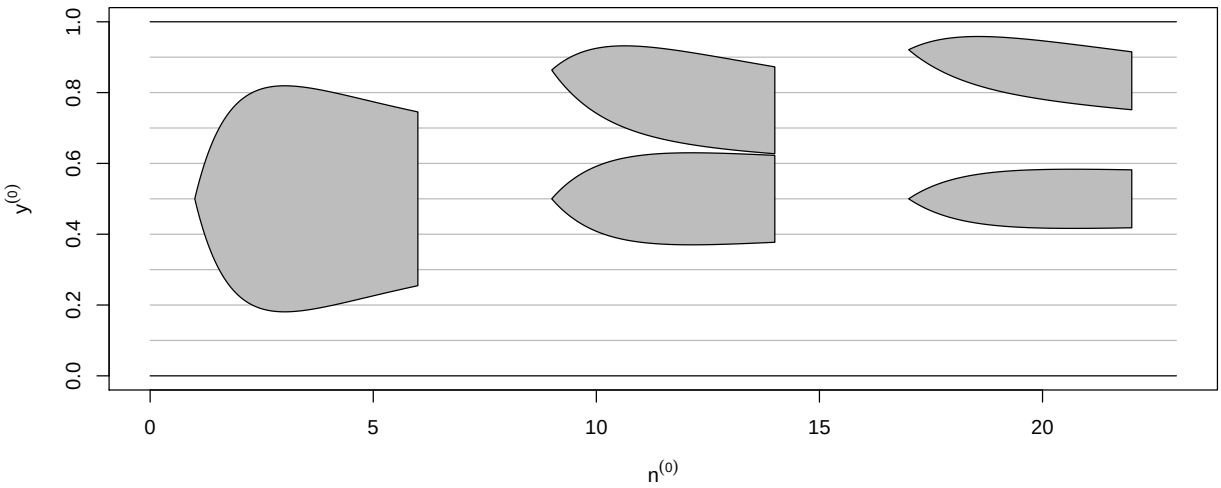

Figure A.5.: Boatshape prior and posterior sets from Figure A.4 in the parametrisation via $(n^{(0)}, y^{(0)})$. Note that in the strong prior-data agreement case, posterior sets based on a rectangular prior set with the same prior main parameter imprecision would be larger than the ones depicted here, illustrating the extra gain in precision.

A.2.3.2. Finding the Touchpoints for the Basic Set

In contrast to models discussed in Section 3.1, where $\underline{y}^{(0)}$ and $\overline{y}^{(0)}$ were at either ends of a set $\mathbb{\Pi}^{(0)}$, here, for a set (A.22), the touchpoints $\underline{y}^{(0)}$ and $\overline{y}^{(0)}$ are not necessarily at $\underline{\eta}_0^{(0)}$ or $\overline{\eta}_0^{(0)}$.[33] Instead, the rays of constant expectation (A.21) touching the parameter set must be determined to find $\underline{y}^{(0)}$ and $\overline{y}^{(0)}$. To do this, the tangent equations for the lower and the upper contour function depending on η_0 are determined. As all rays of constant expectation pass through the point $(-2, 0)$, the tangent that passes through this point is determined by inserting this point into the tangent equation, and the resulting equation is solved for η_0. The resulting points $(\eta_0^l, \underline{c}^{(0)}(\eta_0^l))$ and $(\eta_0^u, \overline{c}^{(0)}(\eta_0^u))$ then give the lower and upper touchpoints of the parameter set, respectively, and can be transformed to $\underline{y}^{(0)}$ and $\overline{y}^{(0)}$ by using (A.18).

As the basic set is symmetrical to the η_0 axis, we have $\eta_0^u = \eta_0^l$, and it suffices to find, e.g., η_0^u, by considering the upper contour tangent.

We denote the tangent in contour point $(\eta_0, \overline{c}^{(0)}(\eta_0))$ by

$$\bar{t}_{\eta_0}(x) = dx + i\,,$$

where $d = \frac{d}{d\eta_0}\overline{c}^{(0)}(\eta_0)$ and i such that $\bar{t}_{\eta_0}(x)$ goes through the point $(\eta_0, \overline{c}^{(0)}(\eta_0))$:

$$\bar{t}_{\eta_0}(x) = dx + i \quad \Longleftrightarrow$$

[33] See, e.g., the prior in Figure A.3.

$$
\begin{aligned}
\overline{c}^{(0)}(\eta_0) &= \frac{d}{d\eta_0}\overline{c}^{(0)}(\eta_0)\eta_0 + i \\
i &= \overline{c}^{(0)}(\eta_0) - \frac{d}{d\eta_0}\overline{c}^{(0)}(\eta_0)\eta_0 \\
&= a - ae^{-b(\eta_0-\underline{\eta}_0)} - \eta_0 abe^{-b(\eta_0-\underline{\eta}_0)} \\
&= a - a(1+b\eta_0)e^{-b(\eta_0-\underline{\eta}_0)} \\
\Longrightarrow \quad \bar{t}_{\eta_0}(x) &= abe^{-b(\eta_0-\underline{\eta}_0)}x + a - a(1+b\eta_0)e^{-b(\eta_0-\underline{\eta}_0)} \\
&= a - a\big(1+b(\eta_0-x)\big)e^{-b(\eta_0-\underline{\eta}_0)}
\end{aligned}
$$

Now, let us find the touchpoint $(\eta_0^u, \overline{c}^{(0)}(\eta_0^u))$ whose tangent goes through $(-2, 0)$, as this gives us $\overline{y}^{(0)}$. We insert $(-2, 0)$ into the tangent equation and solve for η_0.

$$
\begin{aligned}
a - a\big(1+b(\eta_0^u+2)\big)e^{-b(\eta_0^u-\underline{\eta}_0)} &\stackrel{!}{=} 0 \\
1 + b(\eta_0^u+2) &\stackrel{!}{=} e^{b(\eta_0^u-\underline{\eta}_0)} \qquad \text{(A.23)}
\end{aligned}
$$

This equation has only one solution for $\eta_0^u > \underline{\eta}_0$, that is, however, not available in closed form.

As a general rule, the nearer η_0^u is to $\underline{\eta}_0$, the larger $\frac{d}{d\eta_0}\overline{c}^{(0)}(\eta_0^u)$, that is, $\overline{y}^{(0)}$ is more away from $\frac{1}{2}$. Here, this means that the larger η_0^u, the more imprecise is the prior parameter set.

A.2.3.3. Strong Prior-Data Agreement Property

We will now prove the essential property that sets (A.22) will lead to especially precise inferences when data are strongly suporting prior information.

For a prior parameter set $\mathrm{H}^{(0)} = \eta_0^{(0)} \times [\underline{\eta}_1^{(0)}, \overline{\eta}_1^{(0)}]$ symmetric around 0, the prior upper expected value $\overline{y}^{(0)}$ results from the transformation (A.18) of the point $(\eta_0^{(0)}, \overline{\eta}_1^{(0)})$. The posterior upper expected value $\overline{y}^{(n)}$, given data that coincide especially well with the prior, i.e., data with $s = \frac{n}{2}$, will then be found at the point $(\eta_0^{(0)} + n, \overline{\eta}_1^{(0)})$, because in this case, the set does not move in the vertical (η_1) direction. As $y^{(0)}$ is decreasing in η_0 and η_1 is constant, $\overline{y}^{(n)}$ will be lower than $\overline{y}^{(0)}$, i.e., imprecision is reduced.

Imprecision is, however, even more strongly reduced for the boatshape parameter set (A.22). Say, we define the prior parameter set such that the prior upper touchpoint is at the η_0 coordinate $\eta_0^u = \eta_0^{(0)}$. For this shape, the η_0 coordinate for the posterior upper touchpoint $\eta_0^{u(n)}$ will be larger than the updated $\eta_0^{(0)}$, i.e., $\eta_0^{u(n)} > \eta_0^{(0)} + n$ (as will be shown below), and thus $\overline{y}^{(n)}$ is lower. Although the η_1 coordinate will be slightly larger at the point $(\eta_0^{u(n)}, \overline{c}(\eta_0^{u(n)}))$ as compared to the point $(\eta_0^{(0)} + n, \overline{\eta}_1^{(0)})$, the corresponding $\overline{y}^{(n)}$ is still lower, as it holds that

$$
\frac{d}{d\eta_0}\overline{c}(\eta_0^{u(n)}) < \frac{d}{d\eta_0}\overline{c}(\eta_0^{(0)} + n)
$$

because $\frac{d}{d\eta_0}\overline{c}(\eta_0)$ is decreasing in η_0, and a smaller slope for the tangent through $(-2,0)$ is equivalent to a lower $\overline{y}^{(n)}$. This is the desired reduction in imprecision for the case of strong prior-data agreement, also depicted exemplarily in Figures A.4 and A.5.

The property $\eta_0^{u(n)} > \eta_0^{(0)} + n$ of the boatshape set will be shown below. Due to symmetry of prior and posterior parameter shape around the η_0 axis, $\eta_0^{u(n)} = \eta_0^{l(n)}$, i.e., the touchpoint at the upper contour (giving $\overline{y}^{(n)}$) is equal to the touchpoint at the lower contour (giving $\underline{y}^{(n)}$), and thus, the argument formulated in terms of $\eta_0^{u(n)}$ holds also for $\eta_0^{l(n)}$.

The upper exponential contour for the posterior boatshape, updated with $s = \frac{n}{2}$, has its 'prow' now at $(\underline{\eta}_0 + n, 0)$, and is defined by the function

$$\begin{aligned}\overline{c}(\eta_0) &= a\left(1 - e^{-b(\eta_0 - n - \underline{\eta}_0)}\right)\\ \frac{d}{d\eta_0}\overline{c}(\eta_0) &= abe^{-b(\eta_0 - n - \underline{\eta}_0)}\,.\end{aligned}$$

The tangent in contour point $(\eta_0, \overline{c}(\eta_0))$ is

$$\overline{t}_{\eta_0}(x) = a - a\big(1 + b(\eta_0 - x)\big)e^{-b(\eta_0 - n - \underline{\eta}_0)}\,.$$

Again, we insert $(-2,0)$ into this tangent equation and solve for η_0.

$$\begin{aligned}a - a\big(1 + b(\eta_0^{u(n)} + 2)\big)e^{-b(\eta_0^{u(n)} - n - \underline{\eta}_0)} &\stackrel{!}{=} 0\\ 1 + b(\eta_0^{u(n)} + 2) &\stackrel{!}{=} e^{b(\eta_0^{u(n)} - n - \underline{\eta}_0)}\,.\end{aligned} \tag{A.24}$$

We compare now (A.24) to (A.23) and conclude that indeed $\eta_0^{u(n)} > \eta_0^{(0)} + n$.

In Figure A.6, the two exponential graphs have the same curvature, the right one is the same as the left, only being shifted to the right by n. The value of $\eta_0^{(0)} + n$, defined as the abscissa of the intersection of the left exponential and the linear function, being shifted to the right by n, would thus be on the right exponential curve. Because $\eta_0^{u(n)}$ results from the intersection of the right exponential curve and the linear function, it is necessarily larger than $\eta_0^{u(0)} + n$, as the linear function is increasing.

A.2.3.4. General Update with $s > \frac{n}{2}$

Let us now consider the update of the basic boatshape (A.22) in the general case $s \neq \frac{n}{2}$. Due to symmetry of the prior set, we can, without loss of generality, consider again only the case $s > \frac{n}{2}$.

For the prior set, being symmetric around the η_0 axis, both touchpoints are located at the same η_0 coordinate, the resulting $\underline{y}^{(0)}$ and $\overline{y}^{(0)}$ having the same distance to 0.5. Regarding the posterior set, according to (A.20), η_0 coordinates are incremented by n, while η_1 coordinates are incremented by $s + \frac{n}{2}$. That is, if $s \neq \frac{n}{2}$, the updated set is no longer symmetric around the η_0 axis, such that we must consider the lower and upper contours separately.

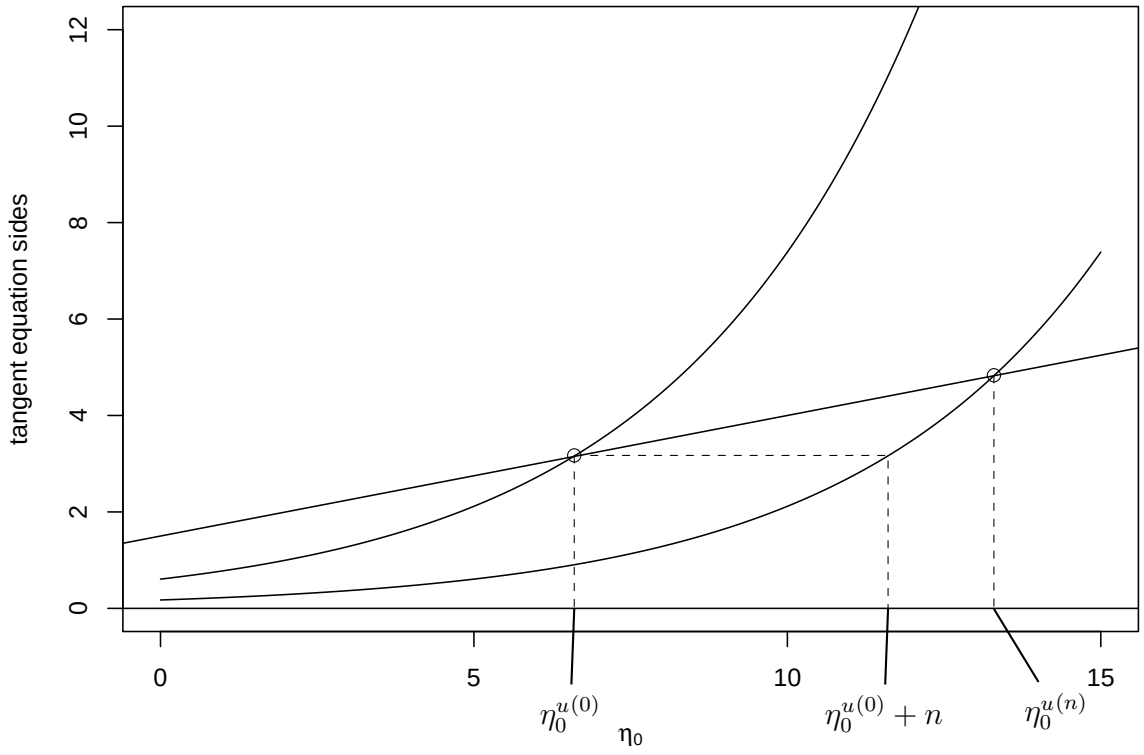

Figure A.6.: Illustration for the argument that $\eta_0^{u(n)} > \eta_0^{u(0)} + n$.

The upper and lower contours and their respective derivatives for the updated boatshape set are now

$$\begin{aligned}
\overline{c}(\eta_0) &= s - \frac{n}{2} + a - ae^{-b(\eta_0 - n - \underline{\eta}_0)}\,, \\
\frac{d}{d\eta_0}\overline{c}(\eta_0) &= abe^{-b(\eta_0 - n - \underline{\eta}_0)}\,, \\
\underline{c}(\eta_0) &= s - \frac{n}{2} - a + ae^{-b(\eta_0 - n - \underline{\eta}_0)}\,, \\
\frac{d}{d\eta_0}\underline{c}(\eta_0) &= -abe^{-b(\eta_0 - n - \underline{\eta}_0)}\,.
\end{aligned}$$

The upper and lower tangents in contour point $(\eta_0, c(\eta_0))$ are now given by

$$\begin{aligned}
\overline{t}_{\eta_0}(x) &= s - \frac{n}{2} + a - a\big(1 + b(\eta_0 - x)\big)e^{-b(\eta_0 - n - \underline{\eta}_0)}\,, \\
\underline{t}_{\eta_0}(x) &= s - \frac{n}{2} - a + a\big(1 + b(\eta_0 - x)\big)e^{-b(\eta_0 - n - \underline{\eta}_0)}\,.
\end{aligned}$$

Inserting again $(-2, 0)$, we get the equations defining the η_0 coordinates $\eta_0^{u(n)}$ and $\eta_0^{l(n)}$ that give us $\overline{y}^{(n)}$ and $\underline{y}^{(n)}$, respectively:

$$\frac{a}{s - \frac{n}{2} + a}\big(1 + b(\eta_0^u + 2)\big) \stackrel{!}{=} e^{b(\eta_0^u - n - \underline{\eta}_0)}\,, \tag{A.25}$$

$$\frac{a}{\frac{n}{2} - s + a}\big(1 + b(\eta_0^u + 2)\big) \stackrel{!}{=} e^{b(\eta_0^u - n - \underline{\eta}_0)}\,. \tag{A.26}$$

We see thus that the picture from Figure A.6 holds here as well, except that the linear function (left hand side of equations (A.25) and (A.26)) is changed in slope and intercept by a factor. (Equivalently, we can consider it to be rotated around the root $-2 - \frac{1}{b}$.) For $s = \frac{n}{2}$, this factor is 1 for both the lower and the upper touchpoint, resulting in the situation of strong prior-data agreement as considered in Section A.2.3.3, where $\eta_0^{u(n)} = \eta_0^{l(n)}$ moved to the right.

Due to symmetry, we will consider the case $s > \frac{n}{2}$ only to describe $\eta_0^{u(n)}$ and $\eta_0^{l(n)}$.

Description of $\eta_0^{u(n)}$. The factor to the linear function $\frac{a}{s - \frac{n}{2} + a}$ in (A.25) is smaller than 1 and decreasing in s. Thus, the larger s, the smaller the factor, the most extreme case being $s = n$, where the factor is $\frac{a}{\frac{n}{2} + a}$. As the linear function's slope will be less steep (the intercept is lowered as well), the intersection with the exponential function moves to the left, i.e. $\eta_0^u(s) < \eta_0^u(\frac{n}{2})$ for $\frac{n}{2} < s \leq n$. This means that $\overline{y}^{(n)}(s) > \overline{y}^{(n)}(\frac{n}{2})$ in general. However, decrease of $\eta_0^u(s)$ is limited by $\underline{\eta}_0 + n$. When the intersection point reaches the left end of the shape at $\underline{\eta}_0 + n$, the gradual increase of $\overline{y}^{(n)}$ through the changing tangent slope for $\underline{\eta}_0 + n \leq \eta_0^u(s) \leq \eta_0^u(\frac{n}{2})$ is replaced by a different change mechanism, where increase of $\overline{y}^{(n)}$ is solely due to increase in the η_1 direction. Due to (A.18), $\overline{y}^{(n)}$ is then linear in s.

Description of $\eta_0^{l\,(n)}$. In (A.26), the factor to the linear function is $\frac{a}{\frac{n}{2}-s+a}$. Here, we have to distinguish the two cases $\frac{n}{2} \leq s < \frac{n}{2}+a$ and $s \geq \frac{n}{2}+a$. In the first case, the factor is larger than 1 and increasing in s. Therefore, the intersection of the linear function with the exponential function will move towards the right, i.e., we will have a larger $\eta_0^{l\,(n)}$, and $\underline{y}^{(n)}$ increases. In the second case, the factor is undefined (for $s = \frac{n}{2}+a$) or negative (for $s >= \frac{n}{2}+a$). Either way, there will be no intersection of the linear function with the exponential function for any $\eta_0 > \underline{\eta}_0 + n$ (For $s \to \frac{n}{2}+a$, the slope $\to \infty$). In fact, for $s \geq \frac{n}{2}+a$, the whole shape is above the η_0 axis, and the touchpoint must be thus at $\overline{\eta}_0 + n$. Actually, $\overline{\eta}_0 + n$ will be the touchpoint already at some $\frac{n}{2} \leq s < \frac{n}{2}+a$, when the intersection point arrives at $\overline{\eta}_0 + n$. At this point, gradual increase of $\underline{y}^{(n)}$ resulting from the movement of $\eta_0^{l\,(n)}$ along the set towards the right is replaced by a linear increase in s. Again, this linear increase is due to the η_1 coordinate being incremented according to (A.20), and from (A.18) we see that $\underline{y}^{(n)}$ is linear in η_1.

Synthesis. For $s > \frac{n}{2}$, both $\overline{y}^{(n)}$ and $\underline{y}^{(n)}$ will at first increase gradually with s, as $\eta_0^{u(n)}$ moves to the left, and $\eta_0^{l\,(n)}$ moves to the right. We will call such updating of the prior parameter set, where neither posterior touchpoints are at the left or the right end of the set, as 'happy learning'.

At some s^u, $\eta_0^{u(n)}$ will arrive at $\underline{\eta}_0 + n$, and at some s^l, $\eta_0^{l\,(n)}$ will arrive at $\overline{\eta}_0 + n$. Whether $s^l < s^u$ or the other way round depends on the choice of parameters $\underline{\eta}_0, \overline{\eta}_0, a$ and b. Either way, once s is larger than either of s^l or s^u, we switch to "unhappy learning", where data s is very much out of line with our prior expectations as expressed by the prior parameter set $\mathrm{H}^{(0)}$. Ultimately, when $s > s^u$ and $s > s^l$, both $\overline{y}^{(n)}$ and $\underline{y}^{(n)}$ will increase linearly in s, but with different slopes. $\overline{y}^{(n)}$ will increase with slope $\frac{1}{\underline{\eta}_0+n+2}$, whereas $\underline{y}^{(n)}$ will increase with a lower slope $\frac{1}{\overline{\eta}_0+n+2}$.

A.2.4. Discussion and Outlook

Taking advantage of a novel parametrisation derived by Miķelis Bickis that was shortly sketched in Section A.2.1, we proposed a prior parameter set shape with the aim to model strong prior-data agreement. Our preliminary studies show some very appealing results for the Beta-Binomial model. Our conjectures about boat-shaped parameter sets in the parametrisation via (η_0, η_1), described in Section A.2.2, could be confirmed in our preliminary studies subsumed in Section A.2.3.

In these studies, we confined ourselves to sets symmetric around the η_0 axis, thus expressing prior information suggesting values of $\frac{s}{n}$ close to $\frac{1}{2}$. As we mentioned in Section A.2.2, prior sets symmetric around rays of constant expectation (A.21) may express prior information with a stress on $\frac{s}{n} = y_c$; these can be obtained by rotating the set (A.22) such that its symmetry axis is on the ray of constant expectation with y_c. Then, basically everything should work the same as described above, except that for y_c near to 0 or 1, one would have to take care to respect the bounds of the parameter space. Informally, we can also

think of this as rotating the whole parameter space (the wedge) under the boat-set until its symmetry axis is aligned to y_c.

So far, we have only vague intuitions for the role of the parameters a and b that, together with $\underline{\eta}_0$ and $\overline{\eta}_0$, define the shape. An idea to find elicitation rules is to investigate also here 'pre-posterior' strategies, by letting the analyst reason on hypothetical counts and what she would like to learn from them.

Related to this, the concrete behaviour during 'happy learning' is difficult to pinpoint exactly, as there are no closed form solutions for $\eta_0^{u(n)}$ and $\eta_0^{l(n)}$. Also, the exact threshold for s where we transfer from 'happy learning' to 'unhappy learning' (where strong prior-data conflict indicated by a linear increase of $\overline{y}^{(n)}$ and $\underline{y}^{(n)}$) is not available in closed form. We plan to study these aspects of the model by numeric examples, drawing $\underline{y}^{(n)}$ and $\overline{y}^{(n)}$ against s for some exemplary choices of $a, b, \underline{\eta}_0$ and $\overline{\eta}_0$, similarly to the *predictive probability plots* given in Section 3.5.[34]

Regarding more advanced considerations on elicitation, if the severity of deviations in the upper ($\frac{s}{n} > y_c$) and in the lower ($\frac{s}{n} < y_c$) direction differ for the analyst (e.g., she wants to be less imprecise for $\frac{s}{n} < y_c$, although she still thinks $\frac{s}{n} \approx y_c$), a and b could be different for the lower and the upper contour. Furthermore, in this parametrisation, it is possible to elicit sets $\mathrm{H}^{(0)}$ that are near-noninformative with respect to $y^{(0)}$, but nevertheless can express preferences towards a certain success fraction by being symmetric around y_c. Such an approach could be similar to the priors suggested by Atwood (1996) and Kelly and Atwood (2011) mentioned in Section 1.3.5, opening up interesting research questions regarding the relation of near-noniformativeness to situations with substantial prior information.

[34] Examples for these plots are Figures 3.10, 3.12, and 3.14.

Bibliography

Antonucci, A., M. Cattaneo, and G. Corani (2011). "Likelihood-Based Naive Credal Classifier". In: *ISIPTA '11: Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*. Ed. by F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger. Innsbruck: SIPTA, pp. 21–30 (cit. on pp. 52, 130).

Antonucci, A., A. Salvetti, and M. Zaffalon (2007). "Credal networks for hazard assessment of debris flows". In: *Advanced Methods for Decision Making and Risk Management in Sustainability Science*. Ed. by J. Kropp and J. Scheffran. New York: Nova Science Publishers, pp. 125–132 (cit. on p. 63).

Antonucci, A., C. de Campos, and M. Zaffalon (2013). "Probabilistic Graphical Models". In: *Introduction to Imprecise Probabilities*. Ed. by T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes. In preparation. Wiley (cit. on pp. 63, 74).

Atwood, C. (1996). "Constrained noninformative priors in risk assessment". In: *Reliability Engineering and System Safety* 53, pp. 37–46 (cit. on pp. 25, 166).

Augustin, T. (1998). *Optimale Tests bei Intervallwahrscheinlichkeit*. Göttingen: Vandenhoeck and Ruprecht (cit. on p. 43).

Augustin, T. (2002). "Neyman-Pearson testing under interval probability by globally least favorable pairs: Reviewing Huber-Strassen theory and extending it to general interval probability". In: *Journal of Statistical Planning and Inference* 105, pp. 149–173 (cit. on p. 119).

Augustin, T. (2003). "On the suboptimality of the generalized Bayes rule and robust Bayesian procedures from the decision theoretic point of view: A cautionary note on updating imprecise priors". In: *ISIPTA '03: Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications*. Ed. by J.-M. Bernard, T. Seidenfeld, and M. Zaffalon. Waterloo: Carleton Scientific, pp. 31–45 (cit. on pp. 41, 80, 130).

Augustin, T. (2004). "Optimal decisions under complex uncertainty—basic notions and a general algorithm for data-based decision making with partial prior knowledge described by interval probability". In: *ZAMM. Zeitschrift für Angewandte Mathematik und Mechanik. Journal of Applied Mathematics and Mechanics* 84, pp. 678–687. URL: `http://dx.doi.org/10.1002/zamm.200410151` (cit. on p. 41).

Augustin, T. and F. Coolen (2004). "Nonparametric predictive inference and interval probability". In: *Journal of Statistical Planning and Inference* 124, pp. 251–272 (cit. on pp. 43, 44, 80).

Augustin, T. and R. Hable (2010). "On the impact of robust statistics on imprecise probability models: a review". In: *Safety, Reliability and Risk of Structures, Infrastructures and Engineering Systems. Proceedings of the 10th International Conference on Structural*

Safety and Reliability, ICOSSAR, 13–17 September 2009, Osaka, Japan. Ed. by M. S. Hitoshi Furuta Dan Frangopol (cit. on p. 136).

Augustin, T., F. Coolen, S. Moral, and M. Troffaes, eds. (2009). *ISIPTA '09: Proceedings of the Sixth International Symposium on Imprecise Probabilities: Theories and Applications*. SIPTA. Durham, UK (cit. on p. 136).

Augustin, T., F. Coolen, G. de Cooman, and M. Troffaes, eds. (2013). *Introduction to Imprecise Probabilities*. In preparation. Wiley (cit. on p. 35).

Augustin, T., G. Walter, and F. Coolen (2013). "Statistical Inference". In: *Introduction to Imprecise Probabilities*. Ed. by T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes. In preparation. Wiley (cit. on pp. 3, 43, 47, 63).

Benavoli, A. and M. Zaffalon (2012). "A model of prior ignorance for inferences in the one-parameter exponential family". In: *Journal of Statistical Planning and Inference* 142, pp. 1960–1979. DOI: 10.1016/j.jspi.2012.01.023. URL: http://www.sciencedirect.com/science/article/pii/S0378375812000535 (cit. on pp. 58, 59, 61, 63, 125, 130).

Berger, J. (1990). "Robust Bayesian analysis: sensitivity to the prior". In: *Journal of Statistical Planning and Inference* 25, pp. 303–328 (cit. on p. 69).

Berger, J., D. Ríos Insua, and F. Ruggeri (2000). "Bayesian Robustness". In: *Robust Bayesian Analysis*. Ed. by D. Ríos Insua and F. Ruggeri. Berlin: Springer, pp. 1–31 (cit. on p. 68).

Berger, J., E. Moreno, L. Pericchi, M. Bayarri, et al. (1994). "An overview of robust Bayesian analysis". In: *TEST* 3, pp. 5–124 (cit. on pp. 40, 47, 49, 68).

Bernard, J.-M. (2005). "An introduction to the imprecise Dirichlet model for multinomial data". In: *International Journal of Approximate Reasoning* 39, pp. 123–150 (cit. on pp. 63, 80).

Bernard, J.-M. (2009). "Special issue on the imprecise Dirichlet model". In: *International Journal of Approximate Reasoning* 50, pp. 201–268 (cit. on pp. 63, 80, 104, 153).

Bernardo, J. and A. Smith (2000). *Bayesian Theory*. Chichester: Wiley (cit. on pp. 6, 8, 9, 13, 15, 61, 83, 135, 142).

Bickis, M. (2009). "The imprecise logit-normal model and its application to estimating hazard functions". In: *Journal of Statistical Theory and Practice* 3. Reprinted in Coolen-Schrijner, Coolen, Troffaes, Augustin, et al. (2009), pp. 183–195 (cit. on pp. 63, 70, 72, 130).

Bickis, M. (2011). "The geometry of imprecise updating". Talk at GEOMIP-11: Workshop on Geometry of Imprecise Probability and Related Statistical Methods. Abstract available at http://maths.dur.ac.uk/stats/people/fc/geomip11.html (cit. on pp. 129, 154, 156).

Boratyńska, A. (1997). "Stability of Bayesian inference in exponential families". In: *Statistics & Probability Letters* 36, pp. 173–178 (cit. on pp. 58, 96, 97).

Bousquet, N. (2008). "Diagnostic of prior-data agreement in applied Bayesian analysis". In: 35, pp. 1011–1029 (cit. on p. 134).

Box, G. and N. Draper (1987). *Empirical Model-building and Response Surface*. New York: Wiley. ISBN: 0-471-81033-9 (cit. on p. 51).

Breiman, L. (1968). *Probability*. Reading, Mass: Addison-Wesley (cit. on p. 6).

Casella, G. and R. Berger (2002). *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning. ISBN: 9780534243128 (cit. on p. 12).

Cattaneo, M. (2007). "Statistical Decisions Based Directly on the Likelihood Function". PhD thesis. ETH Zürich. URL: http://dx.doi.org/10.3929/ethz-a-005463829 (cit. on pp. 52, 80).

Cattaneo, M. (2008). "Fuzzy Probabilities Based on the Likelihood Function". In: *Soft Methods for Handling Variability and Imprecision*. Ed. by D. Dubois, M. Lubiano, H. Prade, M. Gil, et al. Vol. 48. Springer, pp. 43–50 (cit. on pp. 41, 52, 127, 130).

Cattaneo, M. (2012). *Likelihood decision functions*. Tech. rep. 128. Department of Statistics, LMU Munich. URL: http://epub.ub.uni-muenchen.de/13713/ (cit. on p. 52).

Cattaneo, M. and A. Wiencierz (2012). "Likelihood-based Imprecise Regression". In: *International Journal of Approximate Reasoning* 53.8, pp. 1137 –1154. URL: http://www.sciencedirect.com/science/article/pii/S0888613X12000862 (cit. on pp. 53, 127).

Chopra, O., M. Duong, A. Faya, and T. Saito (2004). *Design of Emergency Power Systems for Nuclear Power Plants: Safety Guide*. Tech. rep. NS-G-1.8. Vienna, Austria: International Atomic Energy Agency (cit. on p. 20).

Choquet, G. (1954). "Theory of capacities". In: *Annales de l'Institut Fourier* 5, pp. 131–295 (cit. on p. 47).

Coolen, F. (1993a). "Imprecise conjugate prior densities for the one-parameter exponential family of distributions". In: *Statistics & Probability Letters* 16, pp. 337–342 (cit. on pp. 41, 70–72, 75, 96, 97, 125).

Coolen, F. (1993b). "Statistical Modeling of Expert Opinions using Imprecise Probabilities". Available from http://alexandria.tue.nl/extra3/proefschrift/PRF9B/9305256.pdf. PhD thesis. Eindhoven Technical University (cit. on p. 41).

Coolen, F. (1994). "On Bernoulli experiments with imprecise prior probabilities". In: *The Statistician* 43, pp. 155 –167 (cit. on pp. 41, 70–72, 81, 90, 97, 104).

Coolen, F. (1997). "An imprecise Dirichlet model for Bayesian analysis of failure data including right-censored observations". In: *Reliability Engineering and System Safety* 56, pp. 61–68 (cit. on p. 63).

Coolen, F. (1998). "Low structure imprecise predictive inference for Bayes' problem". In: *Statistics & Probability Letters* 36, pp. 349–357 (cit. on p. 114).

Coolen, F. (2006). "On nonparametric predictive inference and objective Bayesianism". In: *Journal of Logic, Language and Information* 15, pp. 21–47 (cit. on pp. 43, 44).

Coolen, F. (2011). *Nonparametric Predictive Inference*. In: *International Encyclopedia of Statistical Science*. Ed. by M. Lovric. Berlin: Springer, pp. 968–970 (cit. on p. 47).

Coolen, F. and T. Augustin (2005). "Learning from multinomial data: A nonparametric predictive alternative to the Imprecise Dirichlet Model". In: *ISIPTA '05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*. Ed. by F. Cozman, B. Nau, and T. Seidenfeld. Manno: SIPTA, pp. 125–134 (cit. on pp. 63, 120).

Coolen, F. and T. Augustin (2009). "A nonparametric predictive alternative to the Imprecise Dirichlet Model: The case of a known number of categories". In: *International Journal of Approximate Reasoning* 50, pp. 217–230 (cit. on pp. 63, 80, 120).

Coolen, F., M. Troffaes, and T. Augustin (2011). *Imprecise Probability.* In: *International Encyclopedia of Statistical Science.* Ed. by M. Lovric. Berlin: Springer, pp. 645–648 (cit. on pp. 32, 33).

Coolen, F. and K. Yan (2004). "Nonparametric predictive inference with right-censored data". In: *Journal of Statistical Planning and Inference* 126, pp. 25–54 (cit. on pp. 44, 47).

Coolen, F., G. de Cooman, T. Fetz, and M. Oberguggenberger, eds. (2011). *ISIPTA '11: Proceedings of the Seventh International Symposium on Imprecise Probabilities: Theories and Applications.* SIPTA. Innsbruck, Austria (cit. on p. 136).

Coolen-Schrijner, P., F. Coolen, M. Troffaes, T. Augustin, et al. (2009). *Imprecision in Statistical Theory and Practice.* Greensboro, NC, USA: Grace Scientific Publishing LLC (cit. on pp. 2, 136, 168, 177).

Corani, G., J. Abellán, A. Masegosa, S. Moral, et al. (2013). "Classification". In: *Introduction to Imprecise Probabilities.* Ed. by T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes. In preparation. Wiley (cit. on pp. 52, 63, 127).

Dawid, A. (2000). "Causal inference without counterfactuals (with discussion)". In: *Journal of the American Statistical Association* 95, pp. 407–424 (cit. on p. 44).

de Finetti, B. (1937). "La prévision: Ses lois logiques, ses sources subjectives". In: *Annales de l'Institut Henri Poincaré* 7. English translation in Kyburg and Smokler 1964, pp. 1–68 (cit. on p. 33).

de Finetti, B. (1970). *Teoria delle Probabilità.* Turin: Einaudi (cit. on p. 33).

Denneberg, D. (1997). *Non-additive measure and integral.* Dordrecht, NL: Kluver Academic Publishers. ISBN: 0-7923-2840-X (cit. on p. 32).

DeRobertis, L. and J. Hartigan (1981). "Bayesian Inference using intervals of measures". In: *The Annals of Statistics* 9, pp. 235–244 (cit. on pp. 69, 70).

Destercke, S. and D. Dubois (2013a). "Other Uncertainty Theories based on Capacities". In: *Introduction to Imprecise Probabilities.* Ed. by T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes. In preparation. Wiley (cit. on pp. 41–43).

Destercke, S. and D. Dubois (2013b). "Special Cases". In: *Introduction to Imprecise Probabilities.* Ed. by T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes. In preparation. Wiley (cit. on pp. 42, 47, 69).

Ellsberg, D. (1961). "Risk, ambiguity, and the Savage axioms". In: *Quarterly Journal of Economics* 75, pp. 643–669 (cit. on pp. 48, 79, 135).

Evans, M. and H. Moshonov (2006). "Checking for Prior-Data Conflict". In: *Bayesian Analysis* 1, pp. 893–914 (cit. on pp. 50, 103, 134).

Fahrmeir, L. and H. Kaufmann (1985). "Consistency and asymptotic normality of the maximum-likelihood estimator in generalized linear-models". In: *Annals of Statistics* 13, pp. 342–368 (cit. on p. 134).

Fahrmeir, L. and T. Kneib (2006). "Structured additive regression for categorial space-time data: A mixed model approach". In: *Biometrics* 62 (1), pp. 109–118 (cit. on p. 134).

Fahrmeir, L. and T. Kneib (2009). "Propriety of posteriors in structured additive regression models: Theory and empirical evidence". In: *Journal of Statistical Planning and Inference* 139 (3), pp. 843–859 (cit. on p. 134).

Fahrmeir, L. and A. Raach (2007). "A Bayesian semiparametric latent variable model für mixed responses". In: *Psychometrika* 72, pp. 327–346 (cit. on p. 134).

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer (cit. on p. 134).

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression. Models, Methods and Applications*. New York: Springer (cit. on pp. 133, 135, 139).

Ferguson, T. (1983). "Bayesian density estimation by mixtures of normal distributions". In: *Recent Advances in Statistics* 24, pp. 287–302 (cit. on p. 58).

Frisén, M. (2011). *Surveillance*. In: *International Encyclopedia of Statistical Science*. Ed. by M. Lovric. Berlin: Springer, pp. 1577–1579 (cit. on pp. 61, 127).

Gilks, W., ed. (1998). *Markov Chain Monte Carlo in Practice*. Boca Raton, FL: Chapman & Hall/CRC (cit. on p. 74).

Gillies, D. (1987). "Was Bayes a Bayesian?" In: *Historia Mathematica* 14, pp. 325–346. DOI: 10.1016/0315-0860(87)90065-6 (cit. on p. 8).

Gillies, D. (2000). *Philosophical Theories of Probability*. New York: Routledge (cit. on p. 8).

Goldstein, M. (1985). "Temporal coherence". In: *Bayesian Statistics* 2, pp. 231–248 (cit. on p. 41).

Goldstein, M. and D. Wooff (2007). *Bayes Linear Statistics: Theory and Methods*. Chichester: Wiley (cit. on p. 41).

Good, I. (1965). *The estimation of probabilities*. Cambridge (MA): MIT Press (cit. on p. 28).

Hampel, F. (2009a). "How can we get new knowledge?" In: *ISIPTA '09: Proceedings of the Sixth International Symposium on Imprecise Probabilities: Theories and Applications*. Ed. by T. Augustin, F. Coolen, S. Moral, and M. Troffaes, pp. 219–227 (cit. on pp. 120, 131).

Hampel, F. (2009b). "Nonadditive Probabilities in Statistics". In: *Journal of Statistical Theory and Practice* 3, pp. 11–23 (cit. on p. 46).

Hampel, F. (2011). "Potential Surprises". In: *ISIPTA '11: Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*. Ed. by F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger. Innsbruck: SIPTA, pp. 199–207. URL: http://www.sipta.org/isipta11/proceedings/papers/s050.pdf (cit. on p. 131).

Hampel, F., E. Ronchetti, P. Rousseeuw, and W. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley (cit. on p. 51).

Hannig, J., H. Yver, and T. Lee (2011). *Fiducial Inference*. In: *International Encyclopedia of Statistical Science*. Ed. by M. Lovric. Berlin: Springer, pp. 515–519 (cit. on p. 33).

Held, H. (2007). "Quantile-Filtered Bayesian Learning for the Correlation Class". In: *ISIPTA '07: Proceedings of the Fifth International Symposium on Imprecise Probabilities*

and Their Applications. Ed. by G. de Cooman, J. Vejnarová, and M. Zaffalon, pp. 223–232 (cit. on p. 80).

Held, H., E. Kriegler, and T. Augustin (2008). "Bayesian Learning for a Class of Priors with Prescribed Marginals". In: *International Journal of Approximate Reasoning* 49, pp. 212–233 (cit. on pp. 41, 80).

Higgins, J. and A. Whitehead (1996). "Borrowing strength from external trials in a meta-analysis". In: *Statistics in Medicine* 15 (24), pp. 2733–2749 (cit. on p. 134).

Hill, B. (1968). "Posterior distribution of percentiles: Bayes' theorem for sampling from a population". In: *Journal of the American Statistical Association* 63, pp. 677 –691 (cit. on p. 43).

Hill, B. (1988). "De Finetti's theorem, induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion)". In: *Bayesian Statistics 3*. Ed. by D. Lindley, J. Bernardo, M. DeGroot, and A. Smith. Oxford University Press, pp. 211–241 (cit. on p. 44).

Hjort, N. and G. Claeskens (2003). "Frequentist model average estimators". In: *Journal of the American Statistical Association* 98, pp. 879–899 (cit. on p. 119).

Høyland, A. and M. Rausand (1994). *System reliability theory: models and statistical methods*. A Wiley interscience publication. New York, NY: Wiley. ISBN: 0-471-59397-4 (cit. on p. 20).

Hsu, M., M. Bhatt, R. Adolphs, D. Tranel, et al. (2005). "Neural Systems Responding to Degrees of Uncertainty in Human Decision-Making". In: *Science*. New Series 310.5754, pp. 1680–1683. ISSN: 00368075. URL: `http://www.jstor.org/stable/3842970` (cit. on pp. 32, 48, 135).

Huber, P. (1981). *Robust Statistics*. New York: Wiley (cit. on p. 51).

Huber, P. and V. Strassen (1973). "Minimax tests and the Neyman-Pearson lemma for capacities". In: *The Annals of Statistics* 1, pp. 251–263 (cit. on pp. 43, 136).

Hume, D. (2000). *A treatise of human nature*. Ed. by D. Norton and M. Norton. The Oxford Philosophical Texts Edition of David Hume. Oxford, UK: Clarendon Press (cit. on pp. 5, 8).

Huntley, N., R. Hable, and M. Troffaes (2013). "Decision Making". In: *Introduction to Imprecise Probabilities*. Ed. by T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes. In preparation. Wiley (cit. on pp. 58, 130).

Ishack, G., B. Fourest, K. Kotthoff, J. Macleod, et al., eds. (1986). *CSNI Specialist Meeting on Operating Experience Relating to On-Site Electric Power Sources*. CSNI Report No. 115. Paris, France: OECD Nuclear Energy Agency (cit. on p. 20).

Kaplan, E. and P. Meier (1958). "Nonparametric Estimation from Incomplete Observations". In: *Journal of the American Statistical Association* 53, pp. 457–481 (cit. on pp. 44, 47).

Karr, A. (1993). *Probability*. New York: Springer (cit. on pp. 5, 6).

Kass, R. and A. Raftery (1995). "Bayes factors". In: *Journal of the American Statistical Association* 90.430, pp. 773–795 (cit. on pp. 14, 119).

Kauermann, R., T. Krivobokova, and L. Fahrmeir (2009). "Some asymptotic results on generalized penalized spline smooting". In: *Journal of the Royal Statistical Society. Series B. Methodological* 71, pp. 487–503 (cit. on p. 134).

Kelly, D. and C. Atwood (2011). "Finding a minimally informative Dirichlet prior distribution using least squares". In: *Reliability Engineering and System Safety* 96.3, pp. 398–402. ISSN: 0951-8320. DOI: 10.1016/j.ress.2010.11.008 (cit. on pp. 24, 25, 166).

Klir, G. and M. Wierman (1999). *Uncertainty-based Information. Elements of Generalized Information Theory*. Heidelberg: Physika (cit. on pp. 46, 135).

Kneib, T. and L. Fahrmeir (2007). "A mixed model approach for geoadditive hazard regression for interval-censored survival times". In: 34 (1), pp. 207–228 (cit. on p. 134).

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. English translation: *Foundations of Probability*, Chelsea, Providence, RI, 1950. Berlin: Springer (cit. on pp. 32, 33).

Krautenbacher, N. (2011). "Ein Beitrag zur generalisierten Bayes-Inferenz: Erweiterung und Anwendung der Implementierung der generalized iLUCK-models (A contribution to generalized Bayesian Inference: Extension and Application of the implementation of generalized iLUCK-models)". Diplomarbeit (Diploma Thesis). Department of Statistics, LMU Munich (cit. on pp. 75–77, 100, 125, 127).

Kyburg, H. (1987). "Logic of Statistical Reasoning". In: *Encyclopedia of Statistical Sciences*. Ed. by S. Kotz, N. Johnson, and C. Read. Vol. 5. New York: Wiley-Interscience, pp. 117–122 (cit. on p. 135).

Kyburg, H. and H. Smokler, eds. (1964). *Studies in Subjective Probability*. Second edition (with new material) 1980. New York: Wiley (cit. on p. 170).

Lawless, J. and M. Fredette (2005). "Frequentist prediction intervals and predictive distributions". In: *Biometrika* 92, pp. 529–542 (cit. on p. 44).

Lawry, J., E. Miranda, A. Bugarin, S. Li, et al., eds. (2006). *Soft Methods in Integrated Uncertainty Modelling*. Berlin/Heidelberg: Springer (cit. on p. 46).

Li, C.-Y., X. Chen, X.-S. Yi, and J.-Y. Tao (2011). "Interval-valued reliability analysis of multi-state systems". In: *IEEE Transactions on Reliability* 60, pp. 323–330. ISSN: 0018-9529. DOI: 10.1109/TR.2010.2103670 (cit. on p. 63).

Lindley, D. (1987). "The Probability Approach to the Treatment of Uncertainty in Artificial Intelligence and Expert Systems". In: *Statistical Science* 2.1, pp. 17–24 (cit. on pp. 46, 49).

Longford, N. (2003). "An alternative to model selection in ordinary regression". In: *Statistics and Computing* 13, pp. 67–80 (cit. on p. 119).

Mangili, F. and A. Benavoli (2013). "New prior near-ignorance models on the simplex". In: *ISIPTA '13: Proceedings of the Eighth International Symposium on Imprecise Probabilities: Theories and Applications*. Ed. by F. Cozman, T. Denoeux, S. Destercke, and T. Seidenfeld, pp. 213–222. URL: http://www.sipta.org/isipta13/proceedings/papers/s021.pdf (cit. on pp. 63, 73, 130).

Manski, C. (2003). *Partial Identification of Probability Distributions*. New York: Springer (cit. on pp. 47, 51).

Mardia, K. and S. El-Atoum (1976). "Bayesian Inference for the Von Mises Fisher Distribution". In: *Biometrika* 63, pp. 203–206 (cit. on p. 64).

Miranda, E. and G. de Cooman (2013). "Structural Judgements". In: *Introduction to Imprecise Probabilities*. Ed. by T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes. In preparation. Wiley (cit. on p. 61).

Mosleh, A., K. Fleming, G. Parry, H. Paula, et al. (1988). *Procedures for treating common cause failures in safety and reliability studies: Procedural framework and examples*. Tech. rep. NUREG/CR-4780. Newport Beach, CA (USA): PLG Inc. (cit. on pp. 21–23).

Murofushi, T. and M. Sugeno (1989). "An interpretation of fuzzy measures and the Choquet integral as an integral with respect to a fuzzy measure". In: *Fuzzy sets and Systems* 29, pp. 201–227 (cit. on p. 47).

Noubiap, R. and W. Seidel (2001). "An algorithm for calculating Γ-minimax decision rules under generalized moment conditions". In: *Annals of Statistics* 29, pp. 1094–1116 (cit. on p. 41).

O'Hagan, A. (1994). *Bayesian Inference*. Vol. 2B. Kendall's Advanced Theory of Statistics. London: Arnold (cit. on pp. 135, 139, 140, 149).

Pericchi, L. and W. Nazaret (1988). "On being imprecise at the higher levels of a hierarchical linear model". In: *Bayesian Statistics 3*. Ed. by J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, pp. 361–375 (cit. on p. 52).

Pericchi, L. and P. Walley (1991). "Robust Bayesian credible intervals and prior ignorance". In: *International Statistical Review* 59, pp. 1–23 (cit. on pp. 68, 70, 72, 73, 75, 81, 90, 97, 104).

Priebe, C. and D. Marchette (2000). "Alternating Kernel and Mixture Density Estimation". In: *Computational Statistics & Data Analysis* 35, pp. 43–65 (cit. on p. 58).

Quaeghebeur, E. (2009). "Learning from Samples Using Coherent Lower Previsions". PhD thesis. Ghent University (available from http://hdl.handle.net/1854/LU-495650) (cit. on pp. 55, 64).

Quaeghebeur, E. (2013). "Desirability". In: *Introduction to Imprecise Probabilities*. Ed. by T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes. In preparation. Wiley (cit. on p. 36).

Quaeghebeur, E. and G. de Cooman (2005). "Imprecise probability models for inference in exponential families". In: *ISIPTA '05. Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*. Ed. by F. Cozman, R. Nau, and T. Seidenfeld. Manno: SIPTA, pp. 287–296 (cit. on pp. 16, 55, 58, 64–66, 79, 80, 83, 84, 91, 104, 119, 123, 136, 153).

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/` (cit. on pp. 55, 98).

Rao, C., H. Toutenburg, Shalabh, and C. Heumann (2008). *Linear Models and Generalizations*. With contributions by Michael Schomaker. Berlin: Springer. ISBN: 978-3-540-74226-5 (cit. on p. 18).

Rinderknecht, S. (2011). "Contribution to the Use of Imprecise Scientific Knowledge in Decision Support". PhD thesis. ETH Zürich. URL: http://dx.doi.org/10.3929/ethz-a-007313975 (cit. on pp. 70, 73–75, 125, 175).

Rinderknecht, S., M. Borsuk, and P. Reichert (2011). "Eliciting density ratio classes". In: *International Journal of Approximate Reasoning* 52. Reprinted as Rinderknecht (2011, §2), pp. 792 –804. DOI: 10.1016/j.ijar.2011.02.002 (cit. on pp. 51, 69, 70, 72).

Ríos Insua, D. and F. Ruggeri, eds. (2000). *Robust Bayesian Analysis*. Springer (cit. on pp. 40, 68, 80, 81, 136).

Ritsema, J., T. Lay, and H. Kanamori (2012). "The 2011 Tohoku earthquake". In: *Elements* 8.3, pp. 183–188. DOI: 10.2113/gselements.8.3.183 (cit. on p. 20).

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. 2nd. Springer texts in statistics. New York, NY: Springer (cit. on pp. 11, 13, 14).

Rüger, B. (1999). *Test- und Schätztheorie. Band I: Grundlagen*. München, Wien: Oldenbourg (cit. on p. 61).

Ruggeri, F., D. Ríos Insua, and J. Martín (2005). "Robust Bayesian Analysis". In: *Handbook of Statistics. Bayesian Thinking: Modeling and Computation*. Ed. by D. Dey and C. Rao. Vol. 25. Elsevier, pp. 623 –667. DOI: 10.1016/S0169-7161(05)25021-6 (cit. on pp. 40, 68).

Savage, L. (1954). *The Foundations of Statistics*. New York: John Wiley & Sons Inc. (cit. on p. 48).

Scheipl, F. and T. Kneib (2009). "Locally adaptive Bayesian P-splines with a Normal-Exponential-Gamma prior". In: *Computational Statistics & Data Analysis* 53 (10), pp. 3533–3552 (cit. on p. 134).

Seidenfeld, T. and L. Wasserman (1993). "Dilation for sets of probabilities". In: *The Annals of Statistics* 21, pp. 1139–1154 (cit. on p. 74).

Shortliffe, E. (1976). *Computer-based medical consultations: MYCIN*. New York: Elsevier. ISBN: 0-4440-0179-4 (cit. on p. 46).

Shyamalkumar, N. (2000). "Likelihood robustness". In: *Robust Bayesian Analysis*. Ed. by D. Ríos Insua and F. Ruggeri. New York: Springer, pp. 127–143 (cit. on p. 128).

Smithson, M. (2013). "Elicitation". In: *Introduction to Imprecise Probabilities*. Ed. by T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes. In preparation. Wiley (cit. on p. 61).

Troffaes, M. (2007). "Decision making under uncertainty using imprecise probabilities". In: *International Journal of Approximate Reasoning* 45, pp. 17–29 (cit. on p. 81).

Troffaes, M. and S. Blake (2013). "A Robust Data Driven Approach to Quantifying Common-Cause Failure in Power Networks". In: *ISIPTA '13: Proceedings of the Eighth International Symposium on Imprecise Probabilities: Theories and Applications*. Ed. by F. Cozman, T. Denoeux, S. Destercke, and T. Seidenfeld, pp. 311–317. URL: http://www.sipta.org/isipta13/proceedings/papers/s031.pdf (cit. on p. 22).

Troffaes, M. and F. Coolen (2009). "Applying the Imprecise Dirichlet Model in cases with partial observations and dependencies in failure data". In: *International Journal of Approximate Reasoning* 50, pp. 257–268 (cit. on p. 119).

Troffaes, M. and M. Goldstein (2013). "A Note on the Temporal Sure Preference Principle and the Updating of Lower Previsions". In: *ISIPTA '13: Proceedings of the Eighth International Symposium on Imprecise Probabilities: Theories and Applications*. Ed. by F. Cozman, T. Denoeux, S. Destercke, and T. Seidenfeld, pp. 319–328. URL: http://www.sipta.org/isipta13/proceedings/papers/s032.pdf (cit. on p. 41).

Troffaes, M. and R. Hable (2013). "Computation". In: *Introduction to Imprecise Probabilities*. Ed. by T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes. In preparation. Wiley (cit. on p. 78).

Troffaes, M., G. Walter, and D. Kelly (2013). *A Robust Bayesian Approach to Modelling Epistemic Uncertainty in Common-Cause Failure Models*. Preprint available at http://arxiv.org/abs/1301.0533. Accepted for publication at: Reliability Engineering & System Safety (cit. on pp. 3, 19, 23).

United States Department of Energy (1993). *DOE Fundamentals Handbook: Nuclear physics and reactor theory*. DOE fundamentals handbook. U.S. Department of Energy. URL: http://books.google.de/books?id=sITYQgAACAAJ (cit. on p. 20).

U.S. Nuclear Regulatory Commission (1975). *Reactor safety study: an assessment of accident risk in U.S. commercial nuclear power plants*. NUREG-75/014 (WASH-1400) (cit. on p. 20).

Utkin, L. (2006). "A method for processing the unreliable expert judgments about parameters of probability distributions". In: *European Journal of Operational Research* 175, pp. 385 –398. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2005.04.041 (cit. on p. 63).

Utkin, L. (2010). "Regression analysis using the imprecise Bayesian normal model". In: *International Journal of Data Analysis Techniques and Strategies* 2, pp. 356–372 (cit. on p. 127).

Utkin, L. and T. Augustin (2007). "Decision making under imperfect measurement using the imprecise Dirichlet model." In: *International Journal of Approximate Reasoning* 44.3, pp. 322–338 (cit. on p. 119).

Utkin, L. and F. Coolen (2011). "Interval-valued regression and classification models in the framework of machine learning". In: *ISIPTA '11: Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*. Ed. by F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger. Innsbruck: SIPTA, pp. 371–380. URL: http://www.sipta.org/isipta11/proceedings/papers/s008.pdf (cit. on p. 127).

Utkin, L. and I. Kozine (2010). "On new cautious structural reliability models in the framework of imprecise probabilities". In: *Structural Safety* 32, pp. 411–416. ISSN: 0167-4730. DOI: 10.1016/j.strusafe.2010.08.004 (cit. on p. 63).

Utkin, L. and A. Wiencierz (2013). "An imprecise boosting-like approach to regression". In: *ISIPTA '13: Proceedings of the Eighth International Symposium on Imprecise Probabilities: Theories and Applications*. Ed. by F. Cozman, T. Denoeux, S. Destercke, and T. Seidenfeld, pp. 345–354. URL: http://www.sipta.org/isipta13/proceedings/papers/s035.pdf (cit. on p. 127).

Utkin, L., S. Zatenko, and F. Coolen (2010). "New Interval Bayesian Models for Software Reliability Based on Non-homogeneous Poisson Processes". In: *Automation and Remote*

Control 71, pp. 935–944. ISSN: 0005-1179. DOI: 10.1134/S0005117910050218 (cit. on p. 63).

Vansteelandt, S., E. Goetghebeur, M. Kenward, and G. Molenberghs (2006). "Ignorance and uncertainty regions as inferential tools in a sensitivity analysis". In: *Statistica Sinica* 16, p. 953 (cit. on p. 47).

Walker, S. (2005). *Three Mile Island: a nuclear crisis in historical perspective*. Berkeley: Univ. of California Press. ISBN: 0-520-24683-7 (cit. on p. 20).

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall. ISBN: 0-412-28660-2 (cit. on pp. 8, 14, 26, 28, 32–35, 38–40, 47, 49, 51, 52, 58, 59, 61, 64–66, 71, 80, 81, 90, 92, 103–106, 108, 118, 128, 135, 153).

Walley, P. (1996a). "Inferences from multinomial data: Learning about a bag of marbles". In: *Journal of the Royal Statistical Society, Series B* 58.1, pp. 3–34 (cit. on pp. 26, 27, 55, 58, 61–63, 79, 80, 104, 106, 130, 153).

Walley, P. (1996b). "Measures of uncertainty in expert systems". In: *Artificial Intelligence* 83, pp. 1–58 (cit. on pp. 34–36, 46, 135).

Walley, P. (2000). "Towards a unified theory of imprecise probability". In: *International Journal of Approximate Reasoning* 24, pp. 125–148 (cit. on pp. 33–38).

Walley, P. and J.-M. Bernard (1999). *Imprecise Probabilistic Prediction for Categorical Data*. Technical Report CAF-9901. Université Paris 8 (cit. on p. 106).

Walley, P. and T. Fine (1982). "Towards a frequentist theory of upper and lower probability". In: *Annals of Statistics* 10, pp. 741–761 (cit. on p. 41).

Walter, G. (2006). "Robuste Bayes-Regression mit Mengen von Prioris — Ein Beitrag zur Statistik unter komplexer Unsicherheit". Diplomarbeit (Diploma Thesis). Department of Statistics, LMU Munich. URL: http://www.stat.uni-muenchen.de/~thomas/team/diplomathesis_GeroWalter.pdf (cit. on pp. 83, 84, 153).

Walter, G. (2007). *The Normal Regression Model as a LUCK-model*. Discussion Paper. http://www.stat.uni-muenchen.de/~thomas/team/isipta07_proof.pdf (cit. on pp. 83, 84).

Walter, G. (2012). *A Technical Note on the Dirichlet-Multinomial Model — The Dirichlet Distribution as the Canonically Constructed Conjugate Prior*. Tech. rep. 131. Department of Statistics, LMU Munich. URL: http://epub.ub.uni-muenchen.de/14068/ (cit. on p. 3).

Walter, G. and T. Augustin (2009a). *Bayesian linear regression — different conjugate models and their (in)sensitivity to prior-data conflict*. Tech. rep. 69. Substantially extended version of Walter and Augustin (2010). Department of Statistics, LMU Munich. URL: http://epub.ub.uni-muenchen.de/11050/1/tr069.pdf (cit. on pp. 3, 133).

Walter, G. and T. Augustin (2009b). "Imprecision and Prior-data Conflict in Generalized Bayesian Inference". In: *Journal of Statistical Theory and Practice* 3. Reprinted in Coolen-Schrijner, Coolen, Troffaes, Augustin, et al. (2009), pp. 255–271. ISSN: 1559-8616 (cit. on pp. 3, 9, 28, 55, 59, 64–66, 79, 104, 106, 119, 136, 153).

Walter, G. and T. Augustin (2010). "Bayesian linear regression — different conjugate models and their (in)sensitivity to prior-data conflict". In: *Statistical Modelling and Regression Structures*. Ed. by T. Kneib and G. Tutz. Festschrift in Honour of Lud-

wig Fahrmeir. Springer, pp. 59–78. ISBN: 978-3-7908-2412-4. URL: http://www.springerlink.com/content/ut13855471268053/ (cit. on pp. 2, 3, 133, 177).

Walter, G., T Augustin, and F. Coolen (2011). "On Prior-Data Conflict in Predictive Bernoulli Inferences". In: *ISIPTA'11: Proceedings of the Seventh International Symposium on Imprecise Probabilities: Theories and Applications*. Ed. by F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger. SIPTA, pp. 391–400. URL: http://www.sipta.org/isipta11/proceedings/046.html (cit. on pp. 3, 55, 59, 66, 103).

Walter, G., T. Augustin, and A. Peters (2007). "Linear Regression Analysis under Sets of Conjugate Priors". In: *ISIPTA'07, Proceedings of the Fifth International Symposium on Imprecise Probabilities and their Applications*. Ed. by G. de Cooman, J. Vejnarová, and M. Zaffalon. SIPTA, pp. 445–455. URL: http://www.sipta.org/isipta07/proceedings/041.html (cit. on pp. 59, 80, 82–84, 127, 153).

Walter, G. and N. Krautenbacher (2013). luck: ***R*** *package for Generalized iLUCK-models*. URL: http://luck.r-forge.r-project.org/ (cit. on pp. 55, 98).

Weichselberger, K. (2000). "The theory of interval probability as a unifying model for uncertainty". In: *International Journal of Approximate Reasoning* 24, pp. 149–170 (cit. on p. 33).

Weichselberger, K. (2001). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept*. Physica, Heidelberg (cit. on pp. 32, 33, 43, 46, 61, 135).

Weichselberger, K. (2007). "The logical concept of probability: Foundation and interpretation". In: *ISIPTA '07: Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications*. Ed. by G. de Cooman, J. Vejnarová, and M. Zaffalon. Prague: Action M Agency for SIPTA (cit. on pp. 33, 80).

Weightman, M., P. Jamet, J. Lyons, and S. Samaddar (2011). *IAEA International Fact Finding Expert Mission of the Fukushima Dai-Ichi NPP Accident Following the Great East Japan Earthquake and Tsunami*. Tech. rep. http://www-pub.iaea.org/MTCD/Meetings/PDFplus/2011/cn200/documentation/cn200_Final-Fukushima-Mission_Report.pdf. Vienna, Austria: International Atomic Energy Agency (cit. on pp. 20, 21).

Whitcomb, K. (2005). "Quasi-Bayesian analysis using imprecise probability assessments and the generalized Bayes' rule". In: *Theory and Decision* 58, pp. 209–238 (cit. on pp. 73, 75–77, 81, 90, 97, 104).

Zaffalon, M. (2005). "Credible classification for environmental problems". In: *Environmental Modelling & Software* 20, pp. 1003–1012 (cit. on p. 63).

Zaffalon, M., K. Wesnes, and O. Petrini (2003). "Reliable diagnoses of dementia by the naive credal classifier inferred from incomplete cognitive data". In: *Artificial Intelligence in Medicine* 29, pp. 61–79 (cit. on p. 63).