Bayesian Analysis (2024) **19**, Number 4, pp. 1129–1161

Prior Knowledge Elicitation: The Past, Present, and Future*

Petrus Mikkola[†], Osvaldo A. Martin[†,**], Suyog Chandramouli[†,‡], Marcelo Hartmann[‡], Oriol Abril Pla[‡], Owen Thomas[§], Henri Pesonen[§], Jukka Corander[§,‡,††], Aki Vehtari[†], Samuel Kaski[†,‖,‡‡], Paul-Christian Bürkner[¶,‡‡] and Arto Klami[‡,‡‡,§§]

Abstract. Specification of the prior distribution for a Bayesian model is a central part of the Bayesian workflow for data analysis, but it is often difficult even for statistical experts. In principle, prior elicitation transforms domain knowledge of various kinds into well-defined prior distributions, and offers a solution to the prior specification problem. In practice, however, we are still fairly far from having usable prior elicitation tools that could significantly influence the way we build probabilistic models in academia and industry. We lack elicitation methods that integrate well into the Bayesian workflow and perform elicitation efficiently in terms of costs of time and effort. We even lack a comprehensive theoretical framework for understanding different facets of the prior elicitation problem.

Why are we not widely using prior elicitation? We analyse the state of the art by identifying a range of key aspects of prior knowledge elicitation, from properties of the modelling task and the nature of the priors to the form of interaction with the expert. The existing prior elicitation literature is reviewed and categorized in these terms. This allows recognizing under-studied directions in prior elicitation research, finally leading to a proposal of several new avenues to improve prior elicitation methodology.

Keywords: prior elicitation, prior distribution, informative prior, Bayesian

arXiv: 2112.01380

*This work was supported by the Academy of Finland (Flagship program: Finnish Center for Artificial Intelligence FCAI), by the Technology Industries of Finland Centennial Foundation, by the Jane and Aatos Erkko Foundation, European Research Council grant 742158 (SCARABEE, Scalable inference algorithms for Bayesian evolutionary epidemiology), and by the UKRI Turing AI World-Leading Researcher Fellowship, EP/W002973/1. Partially funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 – 390740016.

[†]Helsinki Institute of Information Technology, Department of Computer Science (PM,OAM,AK,SK) and Department of Information and Communication Engineering (SC), Aalto University, Finland, petrus.mikkola@aalto.fi; osvaldo.martin@aalto.fi; suyog.chandramouli@aalto.fi; samuel.kaski@aalto.fi

[‡]Helsinki Institute of Information Technology, Department of Computer Science (MH,OA,AK,SC) and Department of Mathematics and Statistics (JC), University of Helsinki, Finland, marcelo.hartmann@helsinki.fi; oriol.abrilpla@helsinki.fi; arto.klami@helsinki.fi

[§]Institute of Basic Medical Sciences, University of Oslo (JC,HP) and Akershus Universitetssykehus (OT), Norway, o.m.t.thomas@medisin.uio.no; h.e.pesonen@medisin.uio.no; jukka.corander@medisin.uio.no

[¶]Cluster of Excellence SimTech, University of Stuttgart, Germany, paul-christian.buerkner@simtech.uni-stuttgart.de

[‖]Department of Computer Science, University of Manchester, UK

[**]Instituto de Matemática Aplicada San Luis, CONICET-UNSL, Argentina

[††]Parasites and Microbes, Wellcome Sanger Institute, UK

[‡‡]Equal contribution.

[§§]Corresponding author, arto.klami@helsinki.fi

 https://doi.org/10.1214/23-BA1381

workflow, domain knowledge.

1 Introduction

Bayesian statistics uses probabilistic models, formalized as a set of interconnected random variables following some assumed probability distributions, for describing observations. Designing a suitable model for a given data analysis task requires both significant statistical expertise and domain knowledge, and is typically carried out as an iterative process that involves repeated testing and refinement. This process can be formulated as the *Bayesian workflow* to aid the modeller work in a more reproducible and documentable manner; see Gelman et al. (2020) for a recent detailed formalization partitioning the process into numerous sub-workflows focusing on different facets of the process, such as model specification, inference and model validation.

We focus on one central part of that Bayesian workflow: the choice of prior distributions for the parameters of the model. In particular, we discuss approaches to *eliciting* knowledge from a domain expert to be converted into prior distributions suitable for use in a probabilistic model, rather than assuming the analyst can specify the priors directly. The fundamental goal of this *expert knowledge or prior elicitation* process (defined in Section 2.1) is to help practitioners design models that better capture the essential properties of the system or process under study. Good elicitation tools could also help in the additional goal of fostering wide-spread adoption of probabilistic modelling by reducing the required statistical expertise. An ideal prior elicitation approach would simultaneously make model specification faster, easier, and better at representing the knowledge of the expert. It is hoped that availability of good prior elicitation tools would qualitatively transform the process of prior specification within the Bayesian modelling workflow, analogously to what *probabilistic programming* languages and their efficient model-agnostic algorithms have done for model specification and inference (e.g. Stan Development Team, 2021; Salvatier et al., 2016; Ge et al., 2018).

Prior elicitation has a long history dating back to the 1960s (Winkler, 1967), and excellent textbook accounts (O'Hagan et al., 2006), surveys and reviews (Garthwaite et al., 2005; O'Hagan, 2019) are available. Despite the established problem formulation and broad scientific literature on methods for eliciting priors in different special cases – often for some particular model family – we are still lacking practical tools that would routinely be used as part of the modelling workflow. While a few actively developed tools for interactive prior elicitation exist and are used in selected domains, exemplified by `SHELF` (Oakley and O'Hagan, 2019) and `makemyprior` (Hem et al., 2021), their active user-base remains a tiny fraction of people regularly applying probabilistic models. Instead, practitioners often use rather ad hoc procedures to specify and modify the priors (e.g. Sarma and Kay, 2020), building on personal expertise and experience, ideally learned by following literature on prior recommendations – for instance by Stan Development Team (2021), on logistic regression (Gelman et al., 2008; Ghosh et al., 2018), on hierarchical models (Gelman, 2006; Simpson et al., 2017; Chung et al., 2015), on Gaussian random fields (Fuglstad et al., 2019), or on autoregressive processes (Sørbye and Rue, 2017).

We discuss reasons for the still limited impact of prior elicitation research on prior specification in practice, and propose a range of research directions that need to be pursued to change the situation. Our main claim is that we are still fairly far from having practical prior elicitation tools that could significantly influence the way probabilistic models are built in academia and industry. To improve over the current state, coordinated research involving expertise from multiple disciplines is needed. This paper is both our call for experts to join these efforts, and a concrete guide for future research. Consequently, the paper is written both for people already developing prior elicitation techniques and for people working on specific complementary problems, who we are encouraging to contribute to the common goal. For people looking for practical methods for prior elicitation in their own modelling problems, we unfortunately cannot yet provide very concrete solutions, but we are looking for your feedback on the requirements and desired goals.

As will be clarified later, several interconnected elements hinder the uptake of prior elicitation methods. Some of these are purely *technical* properties of the elicitation algorithms, relating to limited scope in terms of models that prevents their use in general probabilistic programming, or ability to only address univariate priors, sequentially, rather than jointly eliciting all priors of a model. Some are more *practical*, such as many of the approaches still being too difficult for non-statistical experts to use, and lack of good open source software that integrates well with the current probabilistic programming tools used for other parts of the modelling workflow. Finally, some aspects are more *societal*: The concrete value of prior elicitation has not yet been adequately demonstrated in highly visible case studies, and hence end-users do not know to request better approaches, and decision-makers have not invested resources for them.

Critically, these issues are highly interconnected. For building large-scale demonstrations of the practical value of prior elicitation in visible applications, we would already need to have high-quality software that integrates with existing modelling workflows, as well as elicitation methods capable of efficiently eliciting priors for models of sufficient complexity. Given that the field is currently falling short of achieving any of these aspects, we argue that significant coordinated effort is needed before we can make concrete recommendations on best practices for elicitation in any given instance. We can largely work in parallel towards mitigating these issues, but it is important to do this in a coordinated manner, typically so that researchers with complementary scientific expertise work together to address the most closely connected elements. For instance, an ideal team for designing the software tools would combine at least computer engineers, statisticians, interface designers and cognitive scientists, to guarantee that the most important aspects for all dimensions are accounted for.

To proceed towards practical recommendations, we start by identifying seven key dimensions that characterize the prior elicitation challenge and possible solutions for it, to provide a coherent framework for discussing the matter. We inspect prior elicitation from the perspectives of (1) properties of the prior distribution itself, (2) the model family and the prior elicitation method's dependence on it, (3) the underlying elicitation space, (4) how the method interprets the information provided by the expert, (5) computation, (6) the form and quantity of interaction with the expert(s), and

(7) the assumed capability of the expert, both in terms of their domain knowledge and statistical understanding. We discuss all of these fundamental dimensions in detail (Section 2.3), identifying several practical guidelines on how specific characteristics for each of them influence the desired properties for the elicitation method. We also provide a review of existing elicitation methods to highlight gaps in the available literature, but for more comprehensive reviews at earlier stages of the literature, we recommend consulting O'Hagan et al. (2006) and Garthwaite et al. (2005).

Building on this framework, we proceed to make recommendations for future research, by characterizing in more detail the current blockers listed above, and outlining our current suggestions on what kind of research is needed to resolve the issues. These recommendations are necessarily on a relatively high abstraction level, but we hope they still provide a tangible starting point for people coming from outside the current prior elicitation research community. In particular, we discuss easy-to-use software that integrates with open probabilistic programming platforms as a necessary requirement for practical impact, already outlining a possible architecture and key components for such a system. We emphasize the need for considerably extended user evaluation for verifying that the methods have practical value.

2 Prior elicitation

2.1 What is prior elicitation?

Specifying prior probability distributions over variables of interest (such as model's parameters) is an essential part of Bayesian inference. These distributions represent available information regarding values of the variables prior to considering the current data at hand. *Prior elicitation* is one way to specify priors and refers to the process of eliciting the subjective knowledge of domain experts in a structured manner and expressing this knowledge as prior probability distributions (Garthwaite et al., 2005; O'Hagan et al., 2006). This involves not only actually gathering the information from an expert, but also any computational methods that may be needed to transform the collected information into well-defined prior probability distributions.

While prior elicitation is the focus of our article, it is only one of many ways to specify informative priors. Alternatively, analysts may directly specify priors based on a variety of other information sources including relevant literature or databases when the parameters have fairly concrete real-world referents (Gelman and Shalizi, 2013). For instance, in medicine, data-based priors have been widely adopted (Bartoš et al., 2021), while there are situations where prior elicitation is preferred, such as with parameter settings that are unverifiable from the data to hand (Dallow et al., 2018). When historical data are available, priors can be specified by 'borrowing' from that data, known as *historical borrowing* (Viele et al., 2014), using hierarchical modelling (Pocock, 1976; Spiegelhalter et al., 2004; Neuenschwander et al., 2010, 2016; Hobbs et al., 2011; Schmidli et al., 2014) or through power priors (Ibrahim and Chen, 2000; Ibrahim et al., 2015; Psioda and Ibrahim, 2019).

Besides encoding domain knowledge, there are other grounds for specifying priors.

For instance, priors can be chosen such that they affect the information in the likelihood as weakly as possible (*noninformative priors*), yield smoother and more stable inferences (*regularizing priors*), or yield 'asymptotically acceptable' posterior inference (*reference priors*) (Gelman et al., 2017; Kass and Wasserman, 1996). In particular, one may require a prior to ensure posterior consistency (Rousseau, 2016; Moreno et al., 2015). While we acknowledge the validity of these approaches as well, we do not discuss them in more detail in this article due to our specific goal of investigating the state of prior elicitation, not prior specification in general. However, we repeat the general observation that practically flat priors, such as normal$(0, 10^6)$, sometimes used by practitioners should be avoided, due to problems in posterior inference (Carlin, 2000; van Dongen, 2006; Gelman, 2006; Gelman et al., 2013, 2017; Gelman and Yao, 2020; Smid and Winter, 2020).

Most parameters of theory-driven, physics-based models have a clear meaning outside the model itself. For example, the weight of a star has meaning outside the statistical model that is used for weight estimation from astronomical data. However, for less precise theories and corresponding models, say, in the social sciences, parameters often only have meaning within the context of the model they are part of (Gelman et al., 2017). Accordingly, for the latter kind of models, prior elicitation procedures need to take into account that the distribution being elicited is part of a model and cannot simply be viewed in isolation. The Bayes rule connects the prior $p(\theta)$ to the posterior $p(\theta|y)$ within the context of the likelihood $p(y|\theta)$,

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}, \tag{2.1}$$

where the observables and the parameters are denoted by y and θ, respectively. The goal of prior elicitation is to elicit $p(\theta)$ from an expert. In line with Gelman et al. (2017), we note that the likelihood $p(y|\theta)$ partially determines the scale and the range of reasonable values for θ. In that respect, prior elicitation differs from the elicitation for evidence-based decision-making (e.g. Kennedy et al., 2008; Brownstein et al., 2019) or expert systems (e.g. Studer et al., 1998; Wilson and Corlett, 2005), where the objective is to elicit a probability distribution (not paired with any likelihood) that represents uncertainty on the parameters of a decision model (Grigore et al., 2016) or the node probability tables of a Bayesian network (Nunes et al., 2018). We note, however, that whether prior elicitation should depend on the (sampling) model is still under community debate and there is no universally accepted answer yet.

A common elicitation process involves two persons, called expert and analyst. We follow the convention that the expert is referred to as a female and the analyst as a male (Oakley and O'Hagan, 2007), and use the term *analyst* instead of facilitator to emphasize that the analyst can play many roles simultaneously (O'Hagan et al., 2006, Section 2.2.1), for instance, as a statistician and a facilitator. The *facilitator* is an expert in the process of elicitation. He can take an active role such as manage dialogue between the expert(s) or a more passive role such as assisting in the elicitation between the expert and an elicitation algorithm. Not all elicitation methods require a human facilitator, but instead he/it is built into the elicitation software (see an interesting

alternative definition by Kahle et al., 2016). The *expert* refers to the domain expert, who is also called a substantive expert. She has relevant knowledge about the uncertain quantities of interest, such as the model parameters or observables. For more about the definition and recruitment of the experts, see Bolger (2018).

2.2 Why isn't the use of prior elicitation widespread (yet)?

Priors can have significant effect on the outcome of the whole modelling process and support is clearly needed for their specification (Robert, 2007; O'Hagan, 2019), yet prior elicitation techniques are not routinely used within practical Bayesian workflows. The most natural explanation for this is that the current solutions are simply not sufficient for the needs of the people building statistical models and doing practical data analysis. We are not aware of structured literature looking into these aspects systematically, and hence we provide here our evaluation of the main reasons why prior elicitation has not yet entered daily use in the statistical modelling community. The goal here is to provide a high-level overview of the main issues we have identified based on both the scientific literature and our experiences while interacting with the modelling community, in particular the user bases of `Stan` (Stan Development Team, 2021), `brms` (Bürkner, 2017), `PyMC` (Salvatier et al., 2016) and `Bambi` (Capretto et al., 2020). Not all claims of this subsection are supported by direct scientific evidence.

As briefly mentioned in the Introduction, we believe the reasons for limited use of prior elicitation are multifaceted and highly interconnected. We believe the three primary reasons, all of approximately equal importance, are:

- **Technical:** We do not know how to design accurate, computationally efficient, and general methods for eliciting priors for arbitrary models.
- **Practical:** We lack good tools for elicitation that would integrate seamlessly to the modelling workflow, and the cost of evaluating elicitation methods is high.
- **Societal:** We lack convincing examples of prior elicitation success stories, needed for attracting more researchers and resources.

By the *technical* dimension we refer to the quality and applicability of the prior elicitation methods and interfaces, for instance in terms of what kinds of models and priors are supported, and how accurate and efficient the algorithms are. An ideal solution would work in general cases, provide an easy interface for the expert to provide information, accurately reproduce the true knowledge of an expert, and be computationally efficient and reliable to be incorporated into the modelling workflow. In Section 2.4 we will summarize the current literature and discuss the limitations of the current technical solutions, effectively concluding that we do not yet have prior elicitation techniques that would reach a sufficient level of technical quality in general cases.

By the *practical* dimension we refer to concrete tools ready to be used by practitioners. On a rough level, a prior elicitation method consists of some interface for interacting

with the expert and the computational algorithm for forming the prior. Often the interfaces proposed for the task have been fairly general, but the majority of the research on the computational algorithms has been dedicated to methods that are only applicable for specific models or forms of priors. Their practical value remains limited. Even though some examples of model-agnostic elicitation methods exist and some initiatives have been developed for extended periods of time, we are still nowhere near a point where prior elicitation tools would routinely be used as a part of the modelling process. Besides the technical reasons mentioned above, one major reason is that the tools have not been integrated as parts of the broadly used modelling ecosystems, but rather as isolated tools with their own interface conventions, modelling languages, and internal data formats. To put it briefly, a person building a model e.g. in `Stan` cannot launch an elicitation interface to elicit priors for their specific model, and in the extreme case there might not even exist any tools applicable to their model. In Section 3.5, we will outline directions for overcoming this practical challenge.

Another practical issue concerns evaluation of prior elicitation methods. Even though the basis of evaluating the elicitation methodologies is well established (see Section 3.4), the practical value of prior elicitation is extremely difficult and costly to evaluate. Already isolated studies demonstrating e.g. improved task completion time, compared to manual prior specification, for some prototypical model require careful empirical experimentation with human users. While this is a common practice in human computer interaction research, for statisticians it requires quite notable additional effort and expertise. More importantly, for the real cases of interest the evaluation setup is unusually complex because the modelling process itself is a highly complex iterative process that requires statistical expertise and takes a long time, possibly weeks or months. Any empirical evaluation of the value of prior elicitation requires enrolling high-level experts who are tasked to carry out complex operations with systems that are unfamiliar to them, and possible significant individual differences in the way models are built necessitate large user bases for conclusive evidence. This can only be done once the practical software is sufficiently mature, and even then is both difficult and expensive. The problem is naturally not unique to prior elicitation, but instead resembles e.g. the cost of evaluating the effect of new medical practices that require medical professionals testing new procedures that may also result in worse treatments, or evaluation of new educational policies and practices. However, justifying the cost is often easier for these tasks that are considered critically important for the society.

Following the above discussion on cost of evaluation, we believe that there is a significant *societal* argument explaining the limited use of prior elicitation. As detailed in this article, the task is challenging and consequently requires significant resources spanning several scientific fields, combining fundamental statistical and algorithmic research with cognitive science and human-computer interaction for forming the solid basis with high-quality software integration and costly evaluation. This requires significant resources, yet the current research is driven solely by academia and the field has remained somewhat small.

To some extent this can be attributed to the long history of avoiding strong subjective priors in quest for objective scientific knowledge or fair and transparent decision-making.

Audiences struggling to accept subjective priors in the first place are best convinced by maximally clear examples that leave no room for additional layers of complexity, such as prior elicitation procedures. Follow-up research encouraged by these examples is likely to follow similar practices even when they could benefit from improved processes for prior specification. We hence argue that lack of broad interest more specifically on prior elicitation is largely because the value of prior elicitation has not been concretely demonstrated in breakthrough applications of societal importance. Without such demonstrations, the level of interest for these tools will remain low outside the statistics research community. However, already isolated examples of significant scientific or economical breakthroughs building on explicit use of prior elicitation could lead to increase in both research funding (e.g. in the form of public-private partnerships for applying the technology) and in particular in interest for open source software development. To some extent these efforts are shared with the general task of convincing researchers and decision-makers that use of subjective priors is scientifically valid and valuable, but additional effort is needed in demonstrating the value of prior elicitation, in the form of examples where it results in improved models or offers a more cost-efficient, reliable and reproducible process.

This argumentation, unfortunately, is very circular in nature. To boost interest in developing better prior elicitation methods, we would need a high-profile demonstration of their value, but establishing that demonstration would require access to high-quality solutions that integrate well with the modelling tools. However, it is important to realize that the demonstrator can likely be done well before having a robust general-purpose solution. Instead, it is sufficient to have proper software and interface integration of prior elicitation with one modelling ecosystem that is already used for addressing societally important modelling questions, combined with elicitation algorithms that work for the specific types of models needed and can later be extended for even more general models without changing the interfaces. For instance, Bayesian models developed within the `Stan` ecosystem played a significant role in modelling the effect of various interventions had on the spread of COVID-19 (Flaxman et al., 2020), and demonstrating the value of prior elicitation in such a context would likely have been sufficient for raising the awareness of this research direction.

2.3 Prior elicitation hypercube

The interdisciplinary nature of the prior elicitation problem, and therefore scattered coverage of the topic, makes it difficult to obtain an overall perspective to the current state of research. To provide a frame of reference, we identify seven key dimensions that characterize the prior elicitation problem. Together the dimensions form a *prior elicitation hypercube*, depicted in Figure 1, that both helps discuss the current literature in a more structured manner and enables identifying understudied directions. The first two dimensions (D1–D2) cover the Bayesian model itself (prior and likelihood). Dimensions D3–D5 specify key proprieties of an elicitation algorithm, such as in which space the elicitation is conducted (D3), how the expert's input is modelled (D4), and how to deal with the computational issues (D5). The last dimensions D6–D7 cover what is assumed about the expert(s) and the interaction with them. The current prior

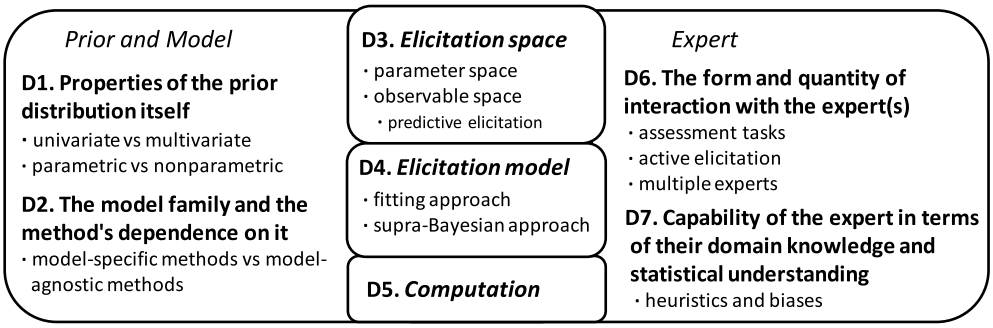

Figure 1: **Prior Elicitation Hypercube**. The seven dimensions (D1-D7) of the hypercube.

elicitation literature is reviewed and categorized in terms of these dimensions in the Supplement (Mikkola et al., 2023). For convenience, the section numbers of the supplementary material are preceded by S, so that e.g. Section S1 refers to the first section of the Supplement.

D1: Properties of the prior distribution itself. Two properties of the prior distribution have attained considerable attention in the literature: dimensionality and parametric vs nonparametric nature. Dimensionality is about the number of parameters: is the prior univariate or multivariate? Eliciting multivariate (joint) distributions is a more complex task than eliciting univariate (marginal) distributions (O'Hagan et al., 2006). It is not enough to elicit the prior in a parameter-by-parameter manner because it is the joint behaviour that affects inferences, and hence it is the joint distribution that must be considered (Gelman et al., 2017). Maybe because of the challenge of eliciting multivariate priors, univariate elicitation has been studied more, even though most models have more than one parameter and hence multiparameter prior elicitation is really needed (Gelman et al., 2020, Section 7.3).

The second property is about whether the prior comes from some parametric family or is nonparametric. The main strand of the prior elicitation literature is about the elicitation of parametric prior distributions, in which case the follow-up question is to which parametric family the prior belongs. The family is determined by the choice of the analyst rather than as a result of the elicitation, although it is often chosen so that it does not conflict with the elicitation data. The choice of the family is closely connected to the likelihood/model (see Section S4), since the natural conjugate family is often considered. On the other end, there is an important line of research on nonparametric prior elicitation that has been built upon Gaussian processes (Oakley and O'Hagan, 2007).

D2: The model family and the method's dependence on it. The underlying probabilistic model and the data analysis task in which it is applied, significantly impact the choice of the prior elicitation method. A bulk of the prior elicitation research addresses the elicitation of parameters of some specific model or model class. We call these

types of methods *model-specific*, and they are reviewed in Section S4. In contrast, in our literature review we found a relatively small number of *model-agnostic* prior elicitation methods, but to name some, we refer the reader to Gelfand et al. (1995); Oakley and O'Hagan (2007); Hartmann et al. (2020). To promote adoption of prior elicitation in applications, it is highly desirable that a prior elicitation method is not model-specific, or at least is applicable to a wide range of models, and we strongly encourage research in this direction as acknowledged earlier by Kadane and Wolfson (1998). Furthermore, the underlying data analysis task may indicate which parameters are of interest, and thus need to be elicited (in the context of a chosen Bayesian model), and which may be less relevant (Clemen and Reilly 2001, p.292; Stefan et al. 2020).

D3: Elicitation space. Prior elicitation is about eliciting expert knowledge to form a prior distribution for model parameters. Hence, it is not surprising that the majority of prior elicitation research is focused on querying values of parameters, or quantities directly related to parameters, from the experts. In this case, we say that the underlying elicitation space is the *parameter space*. This implies that the expert has to have at least some intuition about the meaning of the parameters (interpretability) and about their natural scales. However, this cannot be assumed in all cases. The elicitation of parameters of Bayesian neural networks serves as an extreme example. Models of this type can have thousands of parameters without any interpretation attached to them. In many cases it may be more beneficial to query the expert about something else, such as model observables. In this case we say that the underlying elicitation space is the *observable space*. The *model observables* are variables (e.g. model outcomes) that can be observed and directly measured, in contrast to *latent variables* (e.g. model parameters) that only exist within the context of the model and are not directly observed. Kadane and Wolfson (1998) made a similar dichotomy where they called elicitation in the parameter space *structural elicitation*, and elicitation where the expert makes judgments about "the dependent variable given various values of the predictor variables", *predictive elicitation*. Predictive elicitation is a type of elicitation in the observable space. In general, elicitation in observable space does not require the model to have both dependent and independent variables (e.g. Coolen, 1992; Hughes and Madden, 2002; Gaoini et al., 2009), the existence of a "regression likelihood" (Kadane and Wolfson, 1998, p.5), nor the prior predictive distribution (S2.1). For instance, an 'elicitation likelihood' can be used for connecting the expert's knowledge on observables to the parameters of interest (see Section 3.1).

D4: Elicitation model. There are fundamental differences between elicitation methods in terms of how the information provided by the expert is interpreted. Since early prior elicitation research (Winkler, 1967; Bunn, 1978), the dominant approach has been "fitting a simple and convenient distribution to match the elicited summaries" (O'Hagan et al., 2006). This '*fitting approach*' does not assume any specific mechanism on how the expert data are generated, and for instance, inconsistencies in the data are reconciled by least-square minimization. *Overfitting* in elicitation means eliciting more summaries than needed to fit a parametric distribution (O'Hagan et al., 2006; Hosack et al., 2017), in which case inconsistencies may appear. Overfitting itself is desirable because it allows for imprecision in the elicited summaries, and the fitted compromise

prior may be expected in practice to yield a more faithful representation of the expert's knowledge (O'Hagan et al., 2006).

There is an alternative to the fitting approach, and how inconsistencies are dealt with. The elicitation of an expert's knowledge can be treated as any other Bayesian inference problem where the analyst's posterior belief about the expert's knowledge is updated in the light of received expert data (Lindley et al., 1979; Gelfand et al., 1995; O'Hagan and Oakley, 2004; Gosling, 2005; Oakley and O'Hagan, 2007; Daneshkhah et al., 2006; Gosling et al., 2007; Oakley et al., 2010; Moala and O'Hagan, 2010; Micallef et al., 2017; Hartmann et al., 2020). The analyst has his own prior over the expert belief, and there is an *elicitation likelihood* that allows the analyst's posterior, which would be the elicited expert's prior, to be inferred from the elicitation data. This standpoint is similar to supra-Bayesian pooling found in the literature of aggregating knowledge of multiple experts (Section S6). We follow the latter terminology, even if there is only a single expert to be elicited, and say that such an elicitation method follows the *supra-Bayesian approach*. In this approach, inconsistencies in the elicited data are accounted for by a noise mechanism built into the elicitation likelihood.

D5: Computation. Computation is needed in many parts of an elicitation algorithm, such as in constructing prior from the elicited data and in active elicitation (Section S5), and the computational aspects need to be accounted for in practical tools. One-shot (D6) elicitation methods that follow the fitting approach (D4) and solely operate in the parameter space are often computationally efficient and can easily be incorporated into a practical workflow. In contrast, iterative (D6) predictive (D3) elicitation methods that operate in both spaces and require repeated computation of the prior predictive distribution require considerably more attention in terms of computational efficiency, both because of increased computational cost and the need for fast response time for convenient user experience.

D6: The form and quantity of interaction with the expert(s). The sixth dimension is about the interaction between the expert(s) and the analyst. On the one hand, the form of *assessment tasks* that an expert performs (and similar aspects relating to the interaction modality with a single expert) is important. On the other hand, if there is more than one expert, the format in which the experts interact is also important. For instance, the behavioural aggregation method used in the SHELF protocol (Oakley and O'Hagan, 2019) encourages the experts to discuss their opinions, and to settle upon group consensus judgments, to which a single prior distribution is fitted (O'Hagan, 2019). Eliciting the knowledge of a group of experts, and how to combine the elicited information into a single *aggregate distribution*, is a well established topic.

Concerning a single expert, there are choices to be made about the interaction modality of the elicitation. The expert can be either queried in a one-shot manner (*one-shot elicitation*), or iteratively where the expert's input affects what is queried next (*iterative elicitation*). For instance, a prior elicitation algorithm that exploits active elicitation (Section S5) is iterative. We distinguish iterative elicitation from *interactive elicitation* that entails interaction with the elicitation system (Kadane et al., 1980), such as the system updating a visualization of a prior distribution based on a slider position controlled by the expert (Jones and Johnson, 2014). It is not obvious at all in which form the

information should be elicited from the expert. Several things need to be taken simultaneously into account, such as what assessment tasks are informative, computationally feasible, and, most importantly, encourage "a thoughtful, auditable and relevant answer that is not affected or biased in some way by the giver's psychology" (Hanea et al., 2021). Thus, the design of the assessment tasks is key, as mathematically equivalent assessment tasks are not necessarily psychologically equivalent (O'Hagan et al., 2006). For instance, the impact of the *visualization* of assessment tasks on elicitation has been studied by Hullman et al. (2018), Kim et al. (2019, 2020), and Sarma and Kay (2020). Research on different assessment tasks is reviewed in Section S1. Since the assessment task should also consider psychological and cognitive aspects of a person being elicited (O'Hagan, 2019), this topic is also related to the next dimension.

D7: Capability of the expert in terms of their domain knowledge and statistical understanding.

Perhaps the most challenging and researched issue in prior elicitation is that most people are unable to express their prior knowledge in terms of probabilities, although they will do so if asked, but their answers may be based on very superficial thinking. (Hanea et al., 2021; Kahneman, 2011). If the expert has no solid statistical training, she may not be able to provide reliable probabilistic assessments. In that case, we can resort to assessment tasks that do not require probabilistic input, such as querying likely hypothetical samples (Casement and Kahle, 2018). If the expert has only vague domain knowledge, the elicitation algorithm should validate the provided information, for instance, by using 'seed variables' as in Cooke's method (Cooke, 1991). Even if the expert has both excellent statistical and domain knowledge, she may be inclined to commit to popular cognitive biases and to use cognitive shortcuts (heuristics) in her reasoning, as well documented by Tversky and Kahneman (1974). This line of research is known as heuristics and biases in prior elicitation, and it is intrinsically connected to psychology (Hogarth, 1975). We provide only entry-points to this broad researched field in Section S7.

2.4 Overview on past literature

We reviewed the current main lines of research in prior elicitation through the lens of the prior elicitation hypercube (Section 2.3). The literature review can be found in the Supplement, with sections referenced to using S1, S2 and so on, and a summary of the main findings is presented here.

We observed that there are regions in the prior elicitation hypercube that are well understood. Elicitation of a univariate parametric prior is an extensively studied topic. Certain descriptive elements of the prior distribution, known as *summaries*, such as quantiles, are typically queried from the expert. These univariate elicitation methods commonly differ in the type of elicited summary, the order the summaries are elicited in, and the framing of the corresponding assessment tasks (visual, gamble, etc.). The leading principle in thinking of the aforementioned aspects and designing elicitation methods in general, has been to minimize cognitive biases and so-called heuristics (Section S7) which expert probabilistic judgments may be subject to (O'Hagan, 2019). There are widely

accepted protocols on how to deal with these biases, and how to conduct elicitation with single (Section S1) and multiple experts (Section S6). However, not all methods take them properly into account.

The research on elicitation in the space of observables is abundant (Section S2), but with a serious limitation. Namely, almost all the research is model-specific. Some prior and model families have been studied extensively (Section S4), with significant attention e.g. on elicitation of priors for generalized linear models. From the perspective of priors, there have been several works on the specific cases of Beta and Dirichlet distributions (Section S1.3). When these priors are considered together with their conjugate likelihood, which allows for a complete sampling model, the assessment tasks are often in the space of observables. If this is not the case, then the assessment tasks are in the space of parameters.

There are also distinct research lines on scoring rules (Section S1.4), nonparametric elicitation (Section S3), and active elicitation (Section S5). *Active elicitation* research refers to several articles on how active learning (Cohn et al., 1994) can be applied in prior elicitation to help make most out of the limited elicitation budget due to costly human effort. *Nonparametric prior elicitation* research is mostly built upon a supra-Bayesian elicitation framework, where the expert's subjective density is assumed to follow a Gaussian process (Oakley and O'Hagan, 2007). *Scoring rules* are a class of devices for eliciting and evaluating probabilities (Murphy and Winkler, 1970). They encourage the expert to make careful assessments.

Despite the fact that multivariate prior elicitation has been studied from many perspectives, many of these methods are difficult to apply beyond text book examples. In particular, the methods do not scale well to high-dimensional parameter spaces. Copula-based elicitation requires assessment of parameter dependencies, which is cognitively challenging (Garthwaite et al., 2005, Sec. 2.3) and scales poorly (e.g. Gaussian copula requires specification of a covariance matrix, Clemen and Reilly, 1999, with $dim(\boldsymbol{\theta})(dim(\boldsymbol{\theta}) + 1)/2$ elements). Nonparametric Gaussian process elicitation that in principle could work with higher dimensions has been empirically demonstrated only for two parameters (Moala and O'Hagan, 2010). Predictive elicitation with generalized linear models (Kadane et al., 1980; Bedrick et al., 1996) does not help either. Although the original method by Kadane et al. (1980) can handle linear regression on at least four covariates, scaling the method to hundreds of covariates is out of question due to the increasing number of elicitation queries. Furthermore, the independence assumption of covariates is troublesome in some predictive methods (Garthwaite and Dickey, 1988; Bedrick et al., 1996). The fundamental challenge for these methods, and for multivariate methods in general, is how to find assessment tasks that are both feasible for the expert and informative enough to identify the complex joint prior distribution of parameters. Moreover, an inference algorithm is needed that can form a prior from the elicited data.

3 Where should we be going?

We have discussed some limitations of the current prior elicitation research (Sections 2.2 and 2.4). In this section, we discuss possible solutions. We propose five promising avenues

(Sections 3.1-3.5) to help in solving the technical, practical, and societal challenges described in Section 2.2; we believe research on these avenues will increase the adoption of prior elicitation techniques.

Technical solutions: We believe that an elicitation method should support elicitation both in the parameter and observable space, should be model-agnostic, and should be sample-efficient since human effort is costly. In Section 3.1, we propose an approach for prior elicitation that takes these objectives into account. We also believe that elicitation is easier when the prior is globally joint. These globally joint priors are discussed in Section 3.3, but essentially, they let elicitation be reduced to just a few interpretable hyperparameters.

Practical solutions: To help make model building easier, faster and better in reflecting expert knowledge, we need to integrate prior elicitation into the Bayesian workflow (Section 3.2). And this requires software able to inter-operate with already existing tools for Bayesian modelling, including probabilistic programming languages (Section 3.5). The software needs to support model-agnostic elicitation, otherwise there will be problems with integration into the Bayesian workflow, because a change in the model specification could preclude prior elicitation.

Societal solutions: We emphasize the need for considerably extended user evaluation, required for verifying that the methods have practical value (Section 3.4), and the need of case studies showing the advantages that a careful prior elicitation process can bring to the modelling process.

3.1 Bayesian treatment of the expert in prior elicitation

In this section, we propose a unified approach to prior elicitation that brings together several elicitation methods. The approach allows the expert to provide her response in both the parameter and observable space (D3), and supports sample-efficient elicitation (D6) by treating the expert in a Bayesian fashion.

In the supra-Bayesian approach, elicitation of an expert's knowledge is treated as any other Bayesian inference problem where the analyst's posterior belief about the expert's knowledge is updated in the light of received expert data (see the discussion in D4). We propose viewing the prior elicitation event itself as an interplay of the expert and the analyst with the following characteristics:

Analyst Poses queries to the expert and gathers the expert's input into a dataset $\mathcal{D}$. The analyst's goal is to infer the expert's distribution of the parameters θ, conditional on the expert's input data, $p(\theta|\mathcal{D})$.

Expert Based on her domain expertise, the expert answers to the analyst's queries. The expert's input is modelled through the *user model* $p(z|q)$ that is the conditional probability of the expert's input z given the analyst's query q. That is, $\mathcal{D}$ consists of N samples $(z_i, q_i)_{i=1}^N$, and all the q_i are treated as fixed.

Expert data can be provided in multiple elicitation spaces, all of which can be combined to derive a single prior within the user model. For example, we can elicit expert data in

both the observable space (data $\mathcal{D}_{\mathcal{Y}}$) and in the parameter space (data $\mathcal{D}_{\Theta}$). The analyst's goal is then to infer the distribution of the parameters conditional on the expert's input data, that is $p(\theta|\mathcal{D}_{\mathcal{Y}}, \mathcal{D}_{\Theta})$. We assume that the analyst updates his knowledge according to Bayes' rule. Hence, he treats elicitation as a posterior inference problem,

$$p(\theta|\mathcal{D}_{\mathcal{Y}}, \mathcal{D}_{\Theta}) = \frac{p(\mathcal{D}_{\mathcal{Y}}|\theta)p(\mathcal{D}_{\Theta}|\theta)p(\theta)}{p(\mathcal{D}_{\mathcal{Y}}, \mathcal{D}_{\Theta})}, \tag{3.1}$$

given the elicitation likelihoods $p(\mathcal{D}_{\mathcal{Y}}|\theta)$ and $p(\mathcal{D}_{\Theta}|\theta)$, and the analyst's prior belief on the expert's knowledge $p(\theta)$. In Equation 3.1, we have assumed $\mathcal{D}_{\mathcal{Y}}$ and $\mathcal{D}_{\Theta}$ to be conditionally independent given θ. The likelihoods $p(\mathcal{D}_{\mathcal{Y}}|\theta)$ and $p(\mathcal{D}_{\Theta}|\theta)$ account for the uncertainty inherent to the elicitation process due to the mechanism how the expert quantifies her knowledge on θ. Hence, the conditional independence assumption essentially states that: given that there exists a fixed parameter vector θ that the expert thinks to be 'true', the mechanism how the expert reveals her knowledge on θ is independent between the two elicitation spaces.

The analyst's prior $p(\theta)$ can be taken to be one of the 'objective' priors mentioned in Section 2.1. Besides the prior, the framework only requires specifying $p(z|q, \theta)$ which describes, at individual query q level, how the expert would respond if she thinks that θ is true. This $p(z|q, \theta)$ is also the likelihood for a single data-point (z, q), since q is treated as fixed without a probability distribution assigned to it. The user model can be obtained by marginalization, $p(z|q) = \int p(z|q, \theta)p(\theta)d\theta$.

The proposed approach can be readily extended to support both sample-efficient elicitation (via active elicitation) and AI-assisted elicitation.

Active elicitation. How to make the most out of the limited budget of N expert's inputs? In other words, what is an optimal strategy to select a sequence of queries $(q_i)_{i=1}^{N}$? This is where the user model comes to play. When the analyst poses a query q, he anticipates that the expert's input z is distributed according to $p(z|q)$. The analyst applies the user model to choose the most informative queries. For instance, if the analyst wants to maximize the expected information gain of $p(\theta|\mathcal{D})$ with respect to a new query q, then the user model is needed for anticipating the corresponding yet unseen response z, which involves taking expectation over $p(z|q)$.

AI-assisted elicitation. One important thing to note is that the analyst (or here facilitator) need not manually select the next queries, but the whole elicitation process can be supervised by an 'artificial facilitator' – an *AI-assistant*. For instance, the AI-assistant can be as simple as consisting only of a user model combined with an active learning criterion for selecting next queries. However, in principle, it is possible to extend the functionalities and capabilities of the AI-assistant to take into account, for instance, the expert's biases and incapabilities of providing informative input for some queries.

Through the following examples, we illustrate how the proposed approach brings together prior elicitation methods found in the literature:

- *Quantiles with mixture beta assumption* (Gelfand et al., 1995). $\mathcal{D}$ is a set of quantiles of the prior distribution of parameters. The elicitation space is the parameter

space, $\mathcal{D} = \mathcal{D}_\Theta$. The likelihood $p(\mathcal{D}_\Theta|\theta)$ equals Eq. (4) in Gelfand et al. (1995), and it is derived from a few assumptions, one being that the expert's input is a transformation of a mixture of beta-distributed random variables. The authors proposed using Markov chain Monte Carlo for sampling from the posterior $p(\theta|\mathcal{D}_\Theta)$.

- *Judgements about plausible outcomes* (Hartmann et al., 2020). $\mathcal{D}$ is a set of prior predictive probabilities where the expert provides $\mathrm{P}(A_i|\lambda)$ for all $i = 1, \ldots, n$, given a partition $\mathbf{A} = \{A_1, \ldots, A_n\}$ of the observable space and hyperparameter vector λ of a parametric prior $p(\theta|\lambda)$. The elicitation space is the observable space, $\mathcal{D} = \mathcal{D}_\mathcal{Y}$. Hartmann et al. (2020) assumed a Dirichlet likelihood and used maximum likelihood estimation to estimate λ.

- *Judgements about parameter values and relevance, using active elicitation* (Daee et al., 2017). The assumed model-specific setup considers a linear regression with a sparsity-inducing spike-and-slab prior (George and McCulloch, 1993) on the regression coefficients. $\mathcal{D}$ is a set of judgements on regression coefficient values and relevance. The elicitation space is the parameter space, $\mathcal{D} = \mathcal{D}_\Theta$. The elicitation likelihood $p(\mathcal{D}_\Theta|\theta)$ and the analyst's prior $p(\theta)$ can be written as a product of Normal and Bernoulli distributions (Daee et al., 2017, Appendix A).

  The active elicitation approach in the paper mixes the regression data and the elicitation data. The proposed active elicitation criterion maximizes the information gain between the posterior predictive distribution and the posterior predictive distribution with a new expert's data point (z, q). The posterior predictive distribution is conditional to both the observational and elicitation data.

3.2 Bayesian modelling workflow

Having to choose a prior distribution can be portrayed both as a burden and a blessing. We choose to affirm that it is a necessity. If you are not choosing your priors yourself, then someone else is inevitably doing it for you, and the automatic assignment of flat priors is not a good idea (Carlin, 2000; van Dongen, 2006; Gelman, 2006; Gelman et al., 2013, 2017; Smid and Winter, 2020; Martin et al., 2021). Under some scenarios, we can rely on default priors and default models. For instance, we may simply need to use a given model for routine inference over new datasets. However, having the flexibility to alter model assumptions could be advantageous, and priors are just one form of assumptions. Thus, adopting a Bayesian workflow for prior elicitation should help to reduce the burden and increase the blessing.

We need a Bayesian workflow, rather than mere Bayesian inference, for several reasons (Gelman et al., 2020): Bayesian modelling can be challenging and generally requires exploration and iteration over alternative models, including different priors, in order to achieve inference that we can trust. Even more, for complex problems we typically do not know ahead of time what model(s), that is, the combination of prior and likelihood, we want to fit and even if so, we would still want to understand the fitted model(s) and its relation to the data. Such understanding can often best be achieved by comparing inferences from a series of related models and evaluating when and how conclusions are similar or not.

One common practical approach to modelling starts with a template model (see discussion by Gelman et al., 2020) with default priors. A need for a more carefully designed prior may be revealed only after careful analysis of the first models, and it may be motivated by unrealistic results, computational problems, or the need for incorporating domain knowledge into a model. In other words, the choice of prior, as with other modelling decisions, is often informed by iterative model exploration. Prior elicitation is thus a central part of a Bayesian workflow, and is not restricted to happen only at the beginning of the workflow.

A useful workflow does not just follow all pre-described steps, but also omits them when they are unnecessary, in order to help allocate finite resources where they are most needed. For example, for simple parametric models and informative data, the likelihood can dominate the prior and the gain from prior elicitation could be negligible. Thus, in many cases it may be sensible to start with some common default priors or priors weakly informed by some summary statistics of the data (e.g. by centering and normalizing the covariate and target values in regression), and then assess the need for more careful prior elicitation using prior diagnostic tools (Kallioinen et al., 2021).

In that sense, knowing when to perform prior elicitation is central to a prior elicitation workflow. A good general heuristic is "in situations where prior information is appreciable, and the data are limited" as O'Hagan et al. (2006) have put it. Then, whether we should perform prior elicitation can be reformulated into: Is it worthwhile to spend resources to incorporate domain knowledge? Or more nuanced: How much information do we need to gather, and how accurate should that information be? In many instances, getting the order of magnitude right and/or obtaining a prior that works to remove nonsensical outcomes may be sufficient. Furthermore, the level of accuracy does not need to be the same for all the parameters in a model, as refining a few or even just one prior can translate into considerably better inference.

Informative priors are useful for inducing strong regularization, namely shrinkage priors such as horseshoe, regularized horseshoe, R2D2, spike-and-slab, and global-local-shrinkage priors. These are applied, for example, in genetic association studies (Guan and Stephens, 2011) where there are a lot of covariates of which only very few are actually relevant and comparably small data sets, making inference without regularization very hard if not possible ($p \gg n$ problems, see Peng et al. (2013)). Outside such shrinkage priors and Bayesian trial design (since it is almost always a small-data scenario, see Yuan et al. (2016)), there can be more nuanced scenarios where informative priors via prior elicitation are crucial. For example, there can be gaps in time-series data in which case the expert may provide structural information in a form of a prior distribution that helps to fill gaps in the posterior distribution, or the expert knowledge may help to extrapolate from one group in the data to another (e.g. see Siivola et al., 2021).

In line with the current literature, we have so far discussed prior elicitation with regard to the choice of distributions and their parameters. This definition can be naturally extended to prior elicitation over models, which could provide a new sub-field for prior elicitation or a sister field of model elicitation. As evaluating over the entire range of conceivable models is unfeasible, answering questions such as: "Is a linear model adequate?", "Do we need to extrapolate and perform predictions outside the observed

domain?", and similar ones would help us to narrow down options and save resources. Restricting the search to a few options early on will help, even if we later choose to expand the set of models.

Finally, a prior elicitation workflow should include one step to assess that the incorporated information is actually useful and an evaluation of the sensitivity of the results to the prior choice, including possible prior-data conflicts (Depaoli et al., 2020; Gelman et al., 2020; Lopes and Tobias, 2011; Al-Labadi and Evans, 2017; Evans and Moshonov, 2006; Reimherr et al., 2021; Berger, 1990; Berger et al., 1994; Canavos, 1975; Hill and Spall, 1994; Skene et al., 1986; Jacobi et al., 2018; Roos et al., 2015; Pérez et al., 2006; Giordano et al., 2018; Bornn et al., 2010; Ho, 2020; Kallioinen et al., 2021).

3.3 Developing better priors

One direction to improve prior elicitation is to develop priors for which elicitation is easier *per se*. In this context, 'easier' can mean one of at least three perspectives: (a) easier to understand for experts (D7), (b) computationally easier (D5), and/or (c) leaving fewer degrees of freedom, that is, fewer hyperparameters to elicit. Perspective (a) is especially relevant for direct elicitation in the parameter space, while perspective (b) is mostly relevant for indirect elicitation in the observable space due to computational requirements of the translation procedure to the parameter space ((D3); see Section S2). Both of these perspectives tend to go hand in hand with the perspective (c) because fewer required choices often make the priors easier to understand for experts due to reduced cognitive load, and reduce computational requirements due to a lower-dimensional target space of the translation. Accordingly, if we focus on (c), we can have the justified hope that other advantages will naturally follow in the process.

Reducing the number of hyperparameters comes with the initial (model-building) choice of what matters to be elicited and what is acceptable to just fix to a constant or forced to be of the same value (equality constraint). This line of reasoning leads to the notion of *joint* hyperparameters where the individual priors all depend on a much smaller (or highly structured) set of hyperparameters, jointly shared across parameters. Any kind of *hierarchical prior* follows this logic by design (Bürkner, 2017). For example, consider a simple hierarchical linear model across observations i with intercepts a_j varying across a total number of J groups:

$$
\begin{aligned}
y_i &\sim \text{normal}(\mu_i, \sigma) \\
\mu_i &= a_{j[i]} \\
a_j &\sim \text{normal}(a, \tau) \\
a &\sim \text{normal}(\mu_a, \sigma_a) \\
\tau &\sim \text{Gamma}(\alpha_\tau, \beta_\tau)
\end{aligned}
$$

Focusing on the priors for a_j, we have essentially reduced the problem of finding a total of J priors, each with one or more hyperparameters, to just choosing four hyperparameters, namely the location μ_a and scale σ_a of the normal prior on the joint mean a as well

as the shape α_τ and rate β_τ of the Gamma prior on the joint standard deviation τ.[1] However, such hierarchical priors are only *locally joint* in the sense that they do not encompass all or even most parameters but only a subset. This becomes apparent if we extend the above model by additional additive terms, for example,

$$\mu_i = a_{j[i]} + b_i + c_i + d_i,$$

with each term having their own mutually independent set of parameters and corresponding hyperparameters.

It would be desirable to develop priors that are *globally joint* in that they span most or even all parameters leaving just a few hyperparameters to choose. With the purpose of preventing overfitting and facilitating variable selection in high-dimensional linear regression models on comparably sparse data, several *hierarchical shrinkage priors* have been developed that fulfil these properties (Bhattacharya et al., 2015; Piironen et al., 2017; Zhang et al., 2020). However, they do not yet generalize much beyond linear regression settings and their usefulness in the context of prior elicitation has not been studied so far. If we can extend these priors to more complicated models and find parameterizations with intuitive hyperparameters, such globally joint priors could prove extremely valuable in making prior elicitation more practical and widely applicable.

3.4 Evaluating prior elicitation

When any new prior elicitation method is proposed, a natural question that arises is whether it works as desired. Similarly, when a variety of prior elicitation methods are available for a given context, the practitioner wonders which one is better. Such questions concern the evaluation of prior elicitation methods. There are multiple desiderata for prior elicitation. Johnson et al. (2010a), for instance, categorize these into (i) validity – whether the elicitation captures the true belief of the expert, (ii) reliability – whether repeated elicitations reproduce the same priors, (iii) responsiveness – whether the elicitation is sensitive to changes in beliefs, and (iv) feasibility, which refers to the costs or resources required for elicitation. Many of these desiderata may seem as being at odds with each other, but they are all relevant for the eventual goal of supporting the building of good models with available resources.

In an ideal scenario, any researcher or user of prior elicitation methods would easily be able to compare the pros and cons of existing off-the-shelf methods for her problem, or even test new ones in small-scale user studies. So far, there have been very few projects where multiple prior elicitation methods have been empirically compared (Winkler, 1967; Johnson et al., 2010b; Grigore et al., 2016), and these have been in very application specific contexts. There is a need for more general and standard validation paradigms for prior elicitation, and the prior elicitation field has no equivalents to practices such as using benchmark datasets for comparing machine-learning algorithms,

[1] In hierarchical models, it is common to also call a and τ 'hyperparameters' although they are not set by the analyst or expert but rather estimated from the data along with other model parameters. To avoid confusion, we continue to restrict the use of 'hyperparameters' to parameters chosen in the elicitation process, which are thus fixed during model fitting.

e.g. Deng et al. (2009); LeCun et al. (2010). We think this is a particularly challenging topic to work on because we also lack good metrics for evaluation. The simple metrics that have been widely used in this context may not be valid measures of the quantities we care about. For example, (i) an expert's subjective feedback about elicited priors may be subject to the kind of biases that also distort their priors, (ii) task completion time is considered to be a proxy for cognitive effort, but the elicitation may be finished inaccurately and in a hurried manner due to the cognitive strain it produces, and so on. Prior elicitation metrics can be potentially improved by incorporating research from areas such as Psychology and Human-Computer Interaction. Improved metrics, increased comparative work and the development of standardized validation paradigms or platforms would be essential as the prior-elicitation field makes more progress. In addition, many proposed evaluation metrics are model-specific, but we also need more general methods that can be used across the board in a model-agnostic manner.

Among the different criteria for prior elicitation, assessing faithfulness, accuracy, or validity may be the hardest. From this perspective, the aim of prior elicitation is to accurately capture subjective knowledge of experts/users. However, there are many sources of distortions in priors elicited by an expert including their cognitive biases while making judgments in uncertain settings, and measurement noise introduced by the prior elicitation method, for example, by eliciting probability distributions over discretized intervals (Miller III and Rice, 1983; Parmar et al., 1994; Tan et al., 2003), especially when there are a smaller number of intervals or bins. A promising empirical approach to evaluating faithfulness of prior elicitation would involve validating elicited priors against an expected ground truth. For instance, one could train participants on data produced by a specified model with specified priors, and see how well the true parameter priors are recovered by the elicitation methods. Such methods could be the basis for developing test-beds for prior elicitation evaluation.

Model-specificity and training efforts in paradigms to evaluate faithfulness can also be bypassed by comparing elicited results against a 'gold standard' model-agnostic method, which is known to have higher accuracy. While the nature of such baseline methods would be a topic of future research, there may be some viable candidates. A very promising perspective in psychology treats human judgements as a result of sampling from their subjective probabilities. This viewpoint has been successfully applied in the Markov chain Monte Carlo with people (MCMCP) approach (Sanborn and Griffiths, 2008; Sanborn et al., 2010) and its variants (Hsu et al., 2012; León-Villagrá et al., 2020; Harrison et al., 2020) to elicit beliefs about how stimuli from a multidimensional stimulus space (e.g. n-dimensional stick figures) maps onto a target category (e.g. 'shape of a cat'). In MCMCP participants take the place of an MCMC acceptance function, and repeatedly accept or reject proposals regarding the category membership of the sampled stimuli. The adaptive nature of MCMC ensures that proposals are over time increasingly sampled from parts of the stimulus space representing the participants' subjective representation of the category. The participants' prior beliefs are then constructed as the stationary distribution of the Markov chain that their judgments eventually converge to. The performance of MCMCP and its variants, on natural categories as well as trained artificial categories make us believe that similar sampling-based methods may

have promise in the prior elicitation field both, for obtaining faithful priors, and for acting as model-agnostic baseline methods in paradigms that assess faithfulness.

When evaluating the accuracy of prior elicitation, we may also want to consider the effect of the elicited prior on the predictions or decisions made based on the model. In some scenarios, it is possible that even coarse elicitation processes can obtain practically useful information and further refinement of the elicitation may not bring additional benefits. Also, even if there is a significant bias in the elicited prior, that bias may have negligible effect on the end result. It can thus be useful to evaluate sensitivity and robustness of inference with respect to the elicited prior and its potential aspects that are difficult to elicit. For example, it is difficult for humans to estimate tiny probabilities, which is reflected in the difficulties of determining the tail shape of the elicited prior. A bias in the elicited prior and too thin tails can lead to strong prior sensitivity or prior-data conflict (Al-Labadi and Evans, 2017; Evans and Moshonov, 2006; Kallioinen et al., 2021; Bürkner, 2021). On the other hand, thick tailed priors may lead to ignoring the otherwise correctly elicited prior information.

3.5 Software for prior elicitation

The absence of general software for prior elicitation that integrates well with existing probabilistic programming languages and tools is hindering the adoption of Bayesian methods outside our core community, and is thus eventually detrimental to their wider development. As with other tools designed to help with the Bayesian workflow, a general design guideline is to avoid automated solutions that could result in the user not paying proper attention to their choices. Ideal software for prior elicitation should take into account the strengths and weaknesses of both humans and computers. Numerical tasks that are computationally demanding, error-prone or even tedious should be automatized as much as possible, while allowing the user to retain control of important decisions and, ideally, the user should be helped to take informed decisions and avoid mistakes. For example, a prior elicitation tool should help users to incorporate domain knowledge while preventing them to become overconfident about their own opinions, and it should easily integrate with other tools to perform prior sensitive checks, for example.

In addition to these general guidelines, there are several desirable features that a software for prior elicitation could have, such as being open source and having a simple and intuitive interface suitable for non-specialists. At least one part of such interface should be visual to enable better input from humans and to perform validation of the proposed priors, and some level of interactive visualization capability would further help to obtain information from experts. Furthermore, switching between different types of visualization (kernel density estimates plots, quantile dotplots, histograms, etc.) would also be valuable as would be the possibility to add user-defined transformations before visualization. For example, Sarma and Kay (2020) describe how different visualizations could lead to different strategies for prior elicitation, and that most participants in their study primarily used a combination of strategies for determining their choice of priors. In addition, research shows that even people with statistical training can have problems correctly interpreting probability densities (Section S1.1), and so alternative representations like quantile dotplots may be preferred (Kay et al., 2016).

Prior elicitation software could be written agnostic of the underlying programming language, or at least interoperable with as many languages as possible, in order to avoid duplication of efforts. Building on top of already present open source libraries related to Bayesian workflow and uncertainty visualization like `ggdist` (Kay, 2021), `Bayesplot` (Gabry et al., 2019; Gabry and Mahr, 2021) and `ArviZ` (Kumar et al., 2019) would help to achieve this goal. Moreover, working on top of such libraries could help to maintain modularity, which is especially desirable at the present state of development of the software for prior elicitation. Modularity would also help to reduce computational costs, if experimentation with visualizations and transformations can be made independent of the model. By specifying each task as distinctly as possible and dividing work, the community can generate and maintain software more easily, while at the same time encouraging research in prior elicitation on one or several dimensions of the research hypercube.

Given that we still need more research to assert which elicitation space (D3) is more appropriate for a given research problem, building software in a modular fashion should allow users to switch between the parameter and the observable space as needed. Similarly, the type of assessment task (D6) should be something that can be chosen by the user (e.g. as in `SHELF` or `MATCH`). It is also important to develop software that supports model-agnostic prior elicitation (D2), otherwise there will be problems with integration into the Bayesian workflow (Section 3.2) because a change in the model specification could preclude prior elicitation.

4 Conclusion

This paper covered the state of the prior elicitation today, focusing on discussing reasons for the somewhat limited impact the research has had on practice. We identified bottlenecks at different levels and argued that significant coordinated effort covering several scientific disciplines will be needed to transform the current state and make prior elicitation a routine part of the practical modelling workflow. In summary, we make the following concrete calls to arms:

1. **We need to focus on elicitation matters that answer to the needs of practical modelling workflow.** Compared to past research, the efforts should be re-directed more towards (a) elicitation methods that are agnostic of the model and prior, (b) elicitation strategies (e.g. active elicitation) that are efficient from the perspective of the modeller and compatible with iterative model-building, and (c) formulations that make elicitation of multivariate priors easier, for instance by designing hierarchical priors that are simpler to elicit.

2. **We need better open software** that integrates seamlessly into the current modelling workflow, and that is sufficiently modular so that new elicitation algorithms can be quickly taken into use and evaluated in concrete modelling cases. The elements not specific to elicitation algorithms (e.g. visualization of the priors, the language used for specifying the models and desired prior families) should be implemented using existing libraries whenever possible, and the tools should be open source.

3. **We need cost-efficient and well-targeted evaluation techniques** for supporting development of new methods and validating their relative quality and value in practical tasks. In ideal case, we would like to see a testbed for prior elicitation techniques that enable easy evaluation of alternative methods in varying situations with feasible experimentation cost, as well as practical ways of collecting information about efficiency of elicitation methods in real use cases.

4. **We need spearhead examples** that clearly demonstrate the value of prior elicitation in applications of societal interest to increase enthusiasm beyond the current niche. These examples need to be ones where use of subjective prior knowledge is useful without a doubt and additionally prior elicitation either improves the value of the model over carefully crafted priors or results in clear cost reductions or improved robustness via a more efficient process (e.g. for cases where the priors need to be specified repeatedly or for several parallel cases).

For the first two we already outline concrete directions in Section 3. We hypothesize that addressing all four foci will transform the status of prior elicitation, by providing the required infrastructure, public interest and funding for speeding up future development.

Supplementary Material

Supplementary Material and Literature Review for "Prior knowledge elicitation: The past, present, and future" (DOI: 10.1214/23-BA1381SUPP; .pdf). In this supplementary material, we present the current main lines of research in prior elicitation through the lens of the prior elicitation hypercube (Section 2.3).

References

Al-Labadi, L. and Evans, M. (2017). "Optimal robustness results for relative belief inferences and the relationship to prior-data conflict." *Bayesian Analysis*, 12(3): 705–728. MR3655873. doi: https://doi.org/10.1214/16-BA1024. 1146, 1149

Bartoš, F., Gronau, Q. F., Timmers, B., Otte, W. M., Ly, A., and Wagenmakers, E.-J. (2021). "Bayesian model-averaged meta-analysis in medicine." *Statistics in Medicine*. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9170 MR4352765. doi: https://doi.org/10.1002/sim.9170. 1132

Bedrick, E. J., Christensen, R., and Johnson, W. (1996). "A New Perspective on Priors for Generalized Linear Models." *Journal of the American Statistical Association*, 91(436): 1450–1460. MR1439085. doi: https://doi.org/10.2307/2291571. 1141

Berger, J. O. (1990). "Robust Bayesian analysis: sensitivity to the prior." *Journal of statistical planning and inference*, 25(3): 303–328. MR1064429. doi: https://doi.org/10.1016/0378-3758(90)90079-A. 1146

Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B., et al. (1994). "An overview of robust Bayesian analysis." *Test*, 3(1): 5–124. 1146

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). "Dirichlet–Laplace priors for optimal shrinkage." *Journal of the American Statistical Association*, 110(512): 1479–1490. MR3449048. doi: https://doi.org/10.1080/01621459.2014.960967. 1147

Bolger, F. (2018). "The selection of experts for (probabilistic) expert knowledge elicitation." In *Elicitation*, 393–443. Springer. MR3700927. 1134

Bornn, L., Doucet, A., and Gottardo, R. (2010). "An efficient computational approach for prior sensitivity analysis and cross-validation." *Canadian Journal of Statistics*, 38(1): 47–64. MR2676929. doi: https://doi.org/10.1002/cjs.10045. 1146

Brownstein, N. C., Louis, T. A., O'Hagan, A., and Pendergast, J. (2019). "The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making." *The American Statistician*, 73(sup1): 56–68. PMID: 31057338. MR3925709. doi: https://doi.org/10.1080/00031305.2018.1529623. 1133

Bunn, D. W. (1978). "Estimation of a Dirichlet prior distribution." *Omega*, 6(4): 371–373. 1138

Bürkner, P.-C. (2021). "Specifying Priors in a Bayesian Workflow." URL https://paul-buerkner.github.io/data/prior_specification_bayesian_workflow.pdf 1149

Bürkner, P.-C. (2017). "brms: An R Package for Bayesian Multilevel Models using Stan." *Journal of Statistical Software*, 80(1): 1–28. 1134, 1146

Canavos, G. C. (1975). "Bayesian estimation: A sensitivity analysis." *Naval Research Logistics Quarterly*, 22(3): 543–552. MR0408154. doi: https://doi.org/10.1002/nav.3800220310. 1146

Capretto, T., Piho, C., Kumar, R., Westfall, J., Yarkoni, T., and Martin, O. A. (2020). "Bambi: A simple interface for fitting Bayesian linear models in Python." URL https://arxiv.org/abs/2012.10754 1134

Carlin, B. P. (2000). *Bayes and empirical Bayes methods for data analysis*. Chapman & Hall/CRC. 1133, 1144

Casement, C. J. and Kahle, D. J. (2018). "Graphical prior elicitation in univariate models." *Communications in Statistics - Simulation and Computation*, 47(10): 2906–2924. MR3874066. doi: https://doi.org/10.1080/03610918.2017.1361981. 1140

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., and Dorie, V. (2015). "Weakly Informative Prior for Point Estimation of Covariance Matrices in Hierarchical Models." *Journal of Educational and Behavioral Statistics*, 40(2): 136–157. 1130

Clemen, R. T. and Reilly, T. (1999). "Correlations and Copulas for Decision and Risk Analysis." *Management Science*, 45(2): 208–224. URL http://www.jstor.org/stable/2634871 1141

Clemen, R. T. and Reilly, T. (2001). *Making Hard Decisions with DecisionTools*. Duxbury/Thomson Learning. 1138

Cohn, D., Atlas, L., and Ladner, R. (1994). "Improving generalization with active learning." *Machine learning*, 15(2): 201–221. 1141

Cooke, R. M. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science (Environmental Ethics and Science Policy)*. Oxford University Press. MR1136548. 1140

Coolen, F. (1992). *Elicitation of expert knowledge and assessment of imprecise prior densities for lifetime distributions*. Memorandum COSOR. Technische Universiteit Eindhoven. 1138

Daee, P., Peltola, T., Soare, M., and Kaski, S. (2017). "Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction." *Machine Learning*, 106(9): 1599–1620. MR3694045. doi: https://doi.org/10.1007/s10994-017-5651-7. 1144

Dallow, N., Best, N., and Montague, T. H. (2018). "Better decision making in drug development through adoption of formal prior elicitation." *Pharmaceutical Statistics*, 17(4): 301–316. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.1854 1132

Daneshkhah, A., Oakley, J., and O'Hagan, A. (2006). "Nonparametric prior elicitation with imprecisely assessed probabilities." Technical report, Citeseer. MR2380570. doi: https://doi.org/10.1093/biomet/asm031. 1139

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: A large-scale hierarchical image database." In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee. 1148

Depaoli, S., Winter, S. D., and Visser, M. (2020). "The Importance of Prior Sensitivity Analysis in Bayesian Statistics: Demonstrations Using an Interactive Shiny App." *Frontiers in Psychology*, 11. 1146

Evans, M. and Moshonov, H. (2006). "Checking for prior-data conflict." *Bayesian analysis*, 1(4): 893–914. MR2282210. doi: https://doi.org/10.1016/j.spl.2011.02.025. 1146, 1149

Flaxman, S., Mishra, S., and et al., A. G. (2020). "Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe." *Nature*, 584: 257–261. 1136

Fuglstad, G.-A., Simpson, D., Lindgren, F., and vard Rue, H. (2019). "Constructing Priors that Penalize the Complexity of Gaussian Random Fields." *Journal of the American Statistical Association*, 114(525): 445–452. MR3941267. doi: https://doi.org/10.1080/01621459.2017.1415907. 1130

Gabry, J. and Mahr, T. (2021). "bayesplot: Plotting for Bayesian Models." R package version 1.8.0. URL https://mc-stan.org/bayesplot/ 1150

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). "Visualization in Bayesian workflow." *J. R. Stat. Soc. A*, 182: 389–402. MR3902665. doi: https://doi.org/10.1111/rssa.12378. 1150

Gaoini, E., Dey, D., and Ruggeri, F. (2009). *Bayesian modeling of flash floods using generalized extreme value distribution with prior elicitation*. University of Connecticut, Department of Statistics. MR2756085. 1138

Garthwaite, P. H. and Dickey, J. M. (1988). "Quantifying Expert Opinion in Linear Regression Problems." *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(3): 462–474. URL http://www.jstor.org/stable/2345708 MR0970980. 1141

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). "Statistical Methods for Eliciting Probability Distributions." *Journal of the American Statistical Association*, 100: 680–701. MR2170464. doi: https://doi.org/10.1198/016214505000000105. 1130, 1132, 1141

Ge, H., Xu, K., and Ghahramani, Z. (2018). "Turing: a language for flexible probabilistic inference." In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, 1682–1690. URL http://proceedings.mlr.press/v84/ge18b.html 1130

Gelfand, A. E., Mallick, B. K., and Dey, D. K. (1995). "Modeling Expert Opinion Arising as a Partial Probabilistic Specification." *Journal of the American Statistical Association*, 90: 598–604. MR1340512. 1138, 1139, 1143, 1144

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian analysis*, 1(3): 515–534. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 1130, 1133, 1144

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC. MR3235677. 1133, 1144

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). "A weakly informative default prior distribution for logistic and other regression models." *The Annals of Applied Statistics*, 2(4): 1360 – 1383. MR2655663. doi: https://doi.org/10.1214/08-AOAS191. 1130

Gelman, A. and Shalizi, C. R. (2013). "Philosophy and the practice of Bayesian statistics." *British Journal of Mathematical and Statistical Psychology*, 66(1): 8–38. MR3044854. doi: https://doi.org/10.1111/j.2044-8317.2011.02037.x. 1132

Gelman, A., Simpson, D., and Betancourt, M. (2017). "The Prior Can Often Only Be Understood in the Context of the Likelihood." *Entropy*, 19(10): 555. 1133, 1137, 1144

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). "Bayesian Workflow." *arXiv*. ArXiv: 2011.01808. URL http://arxiv.org/abs/2011.01808 1130, 1137, 1144, 1145, 1146

Gelman, A. and Yao, Y. (2020). "Holes in Bayesian statistics." *Journal of Physics G: Nuclear and Particle Physics*, 48(1): 014002. 1133

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 1144

Ghosh, J., Li, Y., and Mitra, R. (2018). "On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression." *Bayesian Analysis*, 13(2): 359–383. MR3780427. doi: https://doi.org/10.1214/17-BA1051. 1130

Giordano, R., Broderick, T., and Jordan, M. I. (2018). "Covariances, Robustness, and Variational Bayes." *Journal of Machine Learning Research*, 19(51): 1–49. URL http://jmlr.org/papers/v19/17-670.html MR3874159. 1146

Gosling, J. (2005). "Elicitation: A nonparametric view." Ph.D. thesis, University of Sheffield. MR2224093. doi: https://doi.org/10.1111/j.1740-9713.2005.00100.x. 1139

Gosling, J. P., Oakley, J. E., and O'Hagan, A. (2007). "Nonparametric elicitation for heavy-tailed prior distributions." *Bayesian Anal.*, 2(4): 693–718. MR2361971. doi: https://doi.org/10.1214/07-BA228. 1139

Grigore, B., Peters, J., Hyde, C., and Stein, K. (2016). "A comparison of two methods for expert elicitation in health technology assessments." *BMC medical research methodology*, 16(1): 1–11. 1133, 1147

Guan, Y. and Stephens, M. (2011). "Bayesian variable selection regression for genome-wide association studies and other large-scale problems." *The Annals of Applied Statistics*, 5(3): 1780 – 1815. MR2884922. doi: https://doi.org/10.1214/11-AOAS455. 1145

Hanea, A. M., Nane, G. F., Bedford, T., and French, S. (2021). *Expert Judgement in Risk and Decision Analysis*, volume 293. Springer Nature. MR4238539. doi: https://doi.org/10.1007/978-3-030-46474-5_1. 1140

Harrison, P., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., Larrouy-Maestri, P., and Jacoby, N. (2020). "Gibbs sampling with people." *Advances in Neural Information Processing Systems*, 33. 1148

Hartmann, M., Agiashvili, G., Bürkner, P., and Klami, A. (2020). "Flexible prior elicitation via the prior predictive distribution." In *Conference on Uncertainty in Artificial Intelligence*, 1129–1138. PMLR. 1138, 1139, 1144

Hem, I. G., Fuglstad, G.-A., and Riebler, A. (2021). "Makemyprior: Intuitive Construction of Joint Priors for Variance Parameters in R." *arXiv:2105.09712 [stat]*. MR4171145. doi: https://doi.org/10.1214/19-BA1185. 1130

Hill, S. D. and Spall, J. C. (1994). "Sensitivity of a Bayesian analysis to the prior distribution." *IEEE transactions on systems, man, and cybernetics*, 24(2): 216–221. MR1267381. doi: https://doi.org/10.1109/21.281421. 1146

Ho, P. (2020). "Global Robust Bayesian Analysis in Large Models." *FRB Richmond Working Paper*. 1146

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., and Sargent, D. J. (2011). "Hierarchical commensurate and power prior models for adaptive incorporation of historical infor-

mation in clinical trials." *Biometrics*, 67(3): 1047–1056. MR2829239. doi: https://doi.org/10.1111/j.1541-0420.2011.01564.x. 1132

Hogarth, R. M. (1975). "Cognitive Processes and the Assessment of Subjective Probability Distributions." *Journal of the American Statistical Association*, 70(350): 271–289. URL http://www.jstor.org/stable/2285808 1140

Hosack, G. R., Hayes, K. R., and Barry, S. C. (2017). "Prior elicitation for Bayesian generalised linear models with application to risk control option assessment." *Reliability Engineering & System Safety*, 167: 351–361. Special Section: Applications of Probabilistic Graphical Models in Dependability, Diagnosis and Prognosis. 1138

Hsu, A., Martin, J., Sanborn, A., and Griffiths, T. (2012). "Identifying representations of categories of discrete items using Markov chain Monte Carlo with People." *Proceedings of the Annual Meeting of the Cognitive Science Society*, 34(34). URL https://escholarship.org/uc/item/3943355b 1148

Hughes, G. and Madden, L. (2002). "Some methods for eliciting expert knowledge of plant disease epidemics and their application in cluster sampling for disease incidence." *Crop Protection*, 21(3): 203–215. 1138

Hullman, J., Kay, M., Kim, Y.-S., and Shrestha, S. (2018). "Imagining Replications: Graphical Prediction & Discrete Visualizations Improve Recall & Estimation of Effect Uncertainty." *IEEE Transactions on Visualization and Computer Graphics*, 24(1): 446–456. 1140

Ibrahim, J. G. and Chen, M.-H. (2000). "Power Prior Distributions for Regression Models." *Statistical Science*, 15(1): 46–60. URL http://www.jstor.org/stable/2676676 MR1842236. doi: https://doi.org/10.1214/ss/1009212673. 1132

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). "The power prior: theory and applications." *Statistics in Medicine*, 34(28): 3724–3749. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6728 MR3422144. doi: https://doi.org/10.1002/sim.6728. 1132

Jacobi, L., Joshi, M. S., and Zhu, D. (2018). "Automated sensitivity analysis for Bayesian inference via Markov chain Monte Carlo: Applications to Gibbs sampling." *Available at SSRN 2984054*. 1146

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T., and Feldman, B. M. (2010a). "Methods to elicit beliefs for Bayesian priors: a systematic review." *Journal of clinical epidemiology*, 63(4): 355–369. 1147

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T., Grosbein, H. A., and Feldman, B. M. (2010b). "A valid and reliable belief elicitation method for Bayesian priors." *Journal of Clinical Epidemiology*, 63(4): 370–383. 1147

Jones, G. and Johnson, W. O. (2014). "Prior elicitation: Interactive spreadsheet graphics with sliders can be fun, and informative." *The American Statistician*, 68(1): 42–51. MR3303833. doi: https://doi.org/10.1080/00031305.2013.868828. 1139

Kadane, J. and Wolfson, L. J. (1998). "Experiences in elicitation." *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1): 3–19. 1138

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S., and Peters, S. C. (1980). "Interactive Elicitation of Opinion for a Normal Linear Model." *Journal of the American Statistical Association*, 75: 845–854. MR0600966. 1139, 1141

Kahle, D., Stamey, J., Natanegara, F., Price, K., and Han, B. (2016). "Facilitated prior elicitation with the wolfram CDF." *Biometrics & Biostatistics International Journal*, 3. 1133

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan. 1140

Kallioinen, N., Paananen, T., Bürkner, P.-C., and Vehtari, A. (2021). "Detecting and diagnosing prior and likelihood sensitivity with power-scaling." *arXiv preprint arXiv:2107.14054*. 1145, 1146, 1149

Kass, R. E. and Wasserman, L. (1996). "The Selection of Prior Distributions by Formal Rules." *Journal of the American Statistical Association*, 91(435): 1343–1370. URL http://www.jstor.org/stable/2291752 1133

Kay, M. (2021). *ggdist: Visualizations of Distributions and Uncertainty*. R package version 2.4.1. URL https://mjskay.github.io/ggdist/ 1150

Kay, M., Kola, T., Hullman, J. R., and Munson, S. A. (2016). "When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems." In *Proceedings of the 2016 chi conference on human factors in computing systems*, 5092–5103. 1149

Kennedy, M., Anderson, C., O'Hagan, A., Lomas, M., Woodward, I., Gosling, J. P., and Heinemeyer, A. (2008). "Quantifying Uncertainty in the Biospheric Carbon Flux for England and Wales." *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 171(1): 109–135. URL http://www.jstor.org/stable/30130733 MR2412649. doi: https://doi.org/10.1111/j.1467-985X.2007.00489.x. 1133

Kim, Y.-S., Kayongo, P., Grunde-McLaughlin, M., and Hullman, J. (2020). "Bayesian-assisted inference from visualized data." *IEEE Transactions on Visualization and Computer Graphics*, 27(2): 989–999. 1140

Kim, Y.-S., Walls, L. A., Krafft, P., and Hullman, J. (2019). "A bayesian cognition approach to improve data visualization." In *Proceedings of the 2019 chi conference on human factors in computing systems*, 1–14. 1140

Kumar, R., Carroll, C., Hartikainen, A., and Martin, O. (2019). "ArviZ a unified library for exploratory analysis of Bayesian models in Python." *Journal of Open Source Software*, 4(33): 1143. URL https://doi.org/10.21105/joss.01143 1150

LeCun, Y., Cortes, C., and Burges, C. (2010). "MNIST handwritten digit database." 1148

León-Villagrá, P., Otsubo, K., Lucas, C., and Buchsbaum, D. (2020). "Uncovering Category Representations with Linked MCMC with People." In *CogSci*. 1148

Lindley, D. V., Tversky, A., and Brown, R. V. (1979). "On the Reconciliation of Probability Assessments." *Journal of the Royal Statistical Society. Series A (General)*, 142(2): 146–180. MR0547236. doi: https://doi.org/10.2307/2345078. 1139

Lopes, H. F. and Tobias, J. L. (2011). "Confronting prior convictions: On issues of prior sensitivity and likelihood robustness in Bayesian analysis." *Annu. Rev. Econ.*, 3(1): 107–131. 1146

Martin, O. A., Kumar, R., and Lao, J. (2021). *Bayesian Modeling and Computation in Python*. Boca Raton: Chapman and Hall/CRC, 1st edition. 1144

Micallef, L., Sundin, I., Marttinen, P., Ammad-ud din, M., Peltola, T., Soare, M., Jacucci, G., and Kaski, S. (2017). "Interactive Elicitation of Knowledge on Feature Relevance Improves Predictions in Small Data Sets." In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, 547–552. New York, NY, USA: Association for Computing Machinery. URL https://doi.org/10.1145/3025171.3025181 1139

Mikkola, P., Martin, O.A., Chandramouli, S., Hartmann, M., Abril Pla, O., Thomas, O., Pesonen, H., Corander, J., Vehtari, A., Kaski, S., Bürkner, P.-C. and Klami, A. (2023). "Supplementary Material for "Prior Knowledge Elicitation: The Past, Present, and Future"." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1381SUPP. 1137

Miller III, A. C. and Rice, T. R. (1983). "Discrete approximations of probability distributions." *Management science*, 29(3): 352–362. 1148

Moala, F. and O'Hagan, A. (2010). "Elicitation of multivariate prior distributions: A nonparametric Bayesian approach." *Journal of Statistical Planning and Inference*, 140: 1635–1655. MR2606706. doi: https://doi.org/10.1016/j.jspi.2010.01.004. 1139, 1141

Moreno, E., Girón, J., and Casella, G. (2015). "Posterior Model Consistency in Variable Selection as the Model Dimension Grows." *Statistical Science*, 30(2): 228–241. Publisher: Institute of Mathematical Statistics. MR3353105. doi: https://doi.org/10.1214/14-STS508. 1133

Murphy, A. H. and Winkler, R. L. (1970). "Scoring rules in probability assessment and evaluation." *Acta psychologica*, 34: 273–286. 1141

Neuenschwander, B., Capkun-Niggli, G., Branson, M., and Spiegelhalter, D. J. (2010). "Summarizing historical information on controls in clinical trials." *Clinical Trials*, 7(1): 5–18. 1132

Neuenschwander, B., Roychoudhury, S., and Schmidli, H. (2016). "On the use of co-data in clinical trials." *Statistics in Biopharmaceutical Research*, 8(3): 345–354. 1132

Nunes, J., Barbosa, M., Silva, L., Gorgônio, K., Almeida, H., Perkusich, A., Nunes, J., Barbosa, M., Silva, L., Gorgônio, K., Almeida, H., and Perkusich, A. (2018). *Issues in the Probability Elicitation Process of Expert-Based Bayesian Networks*. IntechOpen. Publication Title: Enhanced Expert Systems. 1133

Oakley, J. E., Daneshkhah, A., and O'Hagan, A. (2010). "Nonparametric prior elicitation using the Roulette method." Technical report, School of Mathematics and Statistics, University of Sheffeld, UK. 1139

Oakley, J. E. and O'Hagan, A. (2007). "Uncertainty in Prior Elicitations: A Nonparametric Approach." *Biometrika*, 94. MR2380570. doi: https://doi.org/10.1093/biomet/asm031. 1133, 1137, 1138, 1139, 1141

Oakley, J. E. and O'Hagan, A.(2019). "SHELF: The Sheffield Elicitation Framework (Version 4.0). School of Mathematics and Statistics, University of Sheffield, UK (http://tonyohagan.co.uk/shelf)." 1130, 1139

O'Hagan, A. (2019). "Expert Knowledge Elicitation: Subjective but Scientific." *The American Statistician*, 73(sup1): 69–81. MR3925710. doi: https://doi.org/10.1080/00031305.2018.1518265. 1130, 1134, 1139, 1140

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, Ltd. 1130, 1132, 1133, 1137, 1138, 1139, 1140, 1145

O'Hagan, A. and Oakley, J. E. (2004). "Probability is perfect, but we can't elicit it perfectly." *Reliability Engineering & System Safety*, 85: 239–248. 1139

Parmar, M. K., Spiegelhalter, D. J., Freedman, L. S., and Committee, C. S. (1994). "The CHART trials: Bayesian design and monitoring in practice." *Statistics in medicine*, 13(13-14): 1297–1312. 1148

Peng, B., Zhu, D., Ander, B. P., Zhang, X., Xue, F., Sharp, F. R., and Yang, X. (2013). "An Integrative Framework for Bayesian Variable Selection with Informative Priors for Identifying Genes and Pathways." *PLOS ONE*, 8(7): 1–16. URL https://doi.org/10.1371/journal.pone.0067672 1145

Pérez, C., Martín, J., and Rufo, M. J. (2006). "MCMC-based local parametric sensitivity estimations." *Computational Statistics & Data Analysis*, 51(2): 823–835. MR2297491. doi: https://doi.org/10.1016/j.csda.2005.09.005. 1146

Piironen, J., Vehtari, A., et al. (2017). "Sparsity information and regularization in the horseshoe and other shrinkage priors." *Electronic Journal of Statistics*, 11(2): 5018–5051. MR3738204. doi: https://doi.org/10.1214/17-EJS1337SI. 1147

Pocock, S. J. (1976). "The combination of randomized and historical controls in clinical trials." *Journal of Chronic Diseases*, 29(3): 175–188. 1132

Psioda, M. A. and Ibrahim, J. G. (2019). "Bayesian clinical trial design using historical data that inform the treatment effect." *Biostatistics*, 20(3): 400–415. MR3973117. doi: https://doi.org/10.1093/biostatistics/kxy009. 1132

Reimherr, M., Meng, X.-L., and Nicolae, D. L. (2021). "Prior sample size extensions for assessing prior impact and prior-likelihood discordance." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. MR4294538. doi: https://doi.org/10.1111/rssb.12414. 1146

Robert, C. P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, volume 2. Springer. MR2723361. 1134

Roos, M., Martins, T. G., Held, L., and Rue, H. (2015). "Sensitivity analysis for Bayesian hierarchical models." *Bayesian Analysis*, 10(2): 321–349. MR3420885. doi: https://doi.org/10.1214/14-BA909. 1146

Rousseau, J. (2016). "On the frequentist properties of Bayesian nonparametric methods." *Annual Review of Statistics and Its Application*, 3: 211–231. 1133

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). "Probabilistic programming in Python using PyMC3." *PeerJ Computer Science*, 2: e55. 1130, 1134

Sanborn, A. and Griffiths, T. L. (2008). "Markov chain Monte Carlo with people." In *Advances in Neural Information Processing Systems*, 1265–1272. 1148

Sanborn, A., Griffiths, T. L., and Shiffrin, R. M. (2010). "Uncovering mental representations with Markov chain Monte Carlo." *Cognitive psychology*, 60(2): 63–106. 1148

Sarma, A. and Kay, M. (2020). "Prior Setting in Practice: Strategies and Rationales Used in Choosing Prior Distributions for Bayesian Analysis." In *Conference on Human Factors in Computing Systems*, 1–12. 1130, 1140, 1149

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., and Neuenschwander, B. (2014). "Robust meta-analytic-predictive priors in clinical trials with historical control information." *Biometrics*, 70(4): 1023–1032. MR3295763. doi: https://doi.org/10.1111/biom.12242. 1132

Siivola, E., Weber, S., and Vehtari, A. (2021). "Qualifying drug dosing regimens in pediatrics using Gaussian processes." *Statistics in Medicine*, 40(10): 2355–2372. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8907 MR4242800. doi: https://doi.org/10.1002/sim.8907. 1145

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors." *Statistical Science*, 32(1): 1–28. MR3634300. doi: https://doi.org/10.1214/16-STS576. 1130

Skene, A., Shaw, J., and Lee, T. (1986). "Bayesian modelling and sensitivity analysis." *Journal of the Royal Statistical Society: Series D (The Statistician)*, 35(2): 281–288. 1146

Smid, S. C. and Winter, S. D. (2020). "Dangers of the defaults: A tutorial on the impact of default priors when using Bayesian SEM with small samples." *Frontiers in Psychology*, 11: 3536. 1133, 1144

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*, volume 13. John Wiley & Sons. 1132

Stan Development Team (2021). "Stan Modeling Language Users Guide and Reference Manual, Version 2.28." URL http://mc-stan.org/ 1130, 1134

Stefan, A., Evans, N., and Wagenmakers, E.-J. (2020). "Practical challenges and methodological flexibility in prior elicitation." *Psychol Methods*. 1138

Studer, R., Benjamins, V., and Fensel, D. (1998). "Knowledge engineering: Principles and methods." *Data & Knowledge Engineering*, 25(1): 161–197. URL https://www.sciencedirect.com/science/article/pii/S0169023X97000566 1133

Sørbye, S. H. and Rue, H. v. (2017). "Penalised Complexity Priors for Stationary Autoregressive Processes." *Journal of Time Series Analysis*, 38(6): 923–935. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jtsa.12242 MR3714116. doi: https://doi.org/10.1111/jtsa.12242. 1130

Tan, S.-B., Chung, Y.-F. A., Tai, B.-C., Cheung, Y.-B., and Machin, D. (2003). "Elicitation of prior distributions for a phase III randomized controlled trial of adjuvant therapy with surgery for hepatocellular carcinoma." *Controlled clinical trials*, 24(2): 110–121. 1148

Tversky, A. and Kahneman, D. (1974). "Judgement under Uncertainty: Heuristics and Biases." *Science*, 185: 1124–1131. 1140

van Dongen, S. (2006). "Prior specification in Bayesian statistics: Three cautionary tales." *Journal of Theoretical Biology*, 242(1): 90–100. URL https://www.sciencedirect.com/science/article/pii/S0022519306000609 MR2266729. doi: https://doi.org/10.1016/j.jtbi.2006.02.002. 1133, 1144

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S., et al. (2014). "Use of historical control data for assessing treatment effects in clinical trials." *Pharmaceutical statistics*, 13(1): 41–54. 1132

Wilson, J. R. and Corlett, N. (2005). *Knowledge Elicitation: Methods, Tools and Techniques*. CRC press. 1133

Winkler, R. L. (1967). "The assessment of prior distributions in Bayesian analysis." *Journal of the American Statistical Association*, 62(319): 776–800. MR0220368. 1130, 1138, 1147

Yuan, Y., Nguyen, H. Q., and Thall, P. F. (2016). *Bayesian designs for phase I-II clinical trials*. CRC Press Boca Raton, FL. 1145

Zhang, Y. D., Naughton, B. P., Bondell, H. D., and Reich, B. J. (2020). "Bayesian Regression Using a Prior on the Model Fit: The R2-D2 Shrinkage Prior." *Journal of the American Statistical Association*, 0(0): 1–13. MR4436318. doi: https://doi.org/10.1080/01621459.2020.1825449. 1147