Flexible Prior Elicitation via the Prior Predictive Distribution

Marcelo Hartmann[1,*] **Georgi Agiashvili**[1] **Paul Bürkner**[2] **Arto Klami**[1]
[1] Department of Computer Science, University of Helsinki
[2] Department of Computer Science, Aalto University
∗ marcelo.hartmann@helsinki.fi

Abstract

The prior distribution for the unknown model parameters plays a crucial role in the process of statistical inference based on Bayesian methods. However, specifying suitable priors is often difficult even when detailed prior knowledge is available in principle. The challenge is to express quantitative information in the form of a probability distribution. Prior elicitation addresses this question by extracting subjective information from an expert and transforming it into a valid prior. Most existing methods, however, require information to be provided on the unobservable parameters, whose effect on the data generating process is often complicated and hard to understand. We propose an alternative approach that only requires knowledge about the observable outcomes – knowledge which is often much easier for experts to provide. Building upon a principled statistical framework, our approach utilizes the prior predictive distribution implied by the model to automatically transform experts judgements about plausible outcome values to suitable priors on the parameters. We also provide computational strategies to perform inference and guidelines to facilitate practical use.

1 INTRODUCTION

The Bayesian approach for statistical inference is widely used both in statistical modeling and in general-purpose machine learning. It builds on the simple and intuitive rule that allows updating one's *prior beliefs* about the state of the world through newly made observations (i.e., data) to obtain *posterior beliefs* in a fully probabilistic manner. Nowadays, the Bayesian approach can routinely be used in a vast number of applications due to combination of powerful inference algorithms and probabilistic programming languages (Meent et al., 2018), such as Stan (Carpenter et al., 2017).

Despite available computational tools, the task of designing and building the model can still be difficult. Often, the user building the model can safely be assumed to have good knowledge of the phenomenon they are modeling. However, they additionally need to have sufficient statistical knowledge in order to formulate the domain assumptions in terms of probabilistic models which are sensible enough to obtain valid inference. This is by no means an easy task for the majority of users. Hence, the model building process is often highly iterative, requiring frequent modifications of modeling assumptions, for example, based on predictive checks and model comparisons; see Daee et al. (2017), Schad et al. (2019) and Sarma and Kay (2020) for attempts of formalising the modeling workflow.

We focus on one particular stage of the modeling process, namely the problem of specifying priors for the model parameters. The prior distribution lies at the heart of the Bayesian paradigm and must be designed coherently to make Bayesian inference operational (e.g., see Kadane and Wolfson, 1998). The practical difficulty, though, even for more experienced users, is the encoding of one's actual prior beliefs in form of parametric distributions. The parameters may not even have direct interpretation, and the effect of the prior on the data generating mechanism can be quite involved and show large disparity with respect to what the user's prior beliefs over the data distribution could be (Kadane et al., 1980).

The existing literature addresses this issue via *expert knowledge elicitation*. This is understood as the process of extracting the expert's information (knowledge or opinion) related to quantities or events that are uncertain, and expressing them in the form of a probability distribution, the prior. See, for example, the works by Lindley

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

(1983) and Gelfand et al. (1995) for early ideas and introduction. See Garthwaite et al. (2005) and O'Hagan (2019) for detailed reviews of expert elicitation procedures and guidelines.

The majority of the knowledge elicitation literature is on eliciting information with respect to the *parameters* of the model, that is, asking the expert to make statements about plausible values of the parameters. The early works do this within specific parametric prior families, whereas more recently, O'Hagan and Oakley (2004) and Gosling (2005) have proposed nonparametric approaches based on Gaussian processes (O'Hagan, 1978), allowing more more flexibility. Even though the prior itself can be of flexible form, the elicitation process is typically carried out on a parameter-by-parameter basis so that each parameter receives its own independent univariate prior. As a result, the implied joint prior on the whole set of parameters is often unreasonable. Although Moala and O'Hagan (2010) generalized the approach of Gosling (2005) to multivariate priors, the resulting process is difficult for experts, since they are required to express high-dimensional joint probabilities. Hence, its practical use is basically limited to just two dimensions.

Independently of whether we assign individual or joint priors on the model parameters, any prior can only be understood in the context of the model it is part of (e.g., Gelman et al., 2017; Simpson et al., 2017). This point may be obvious but its practical implications are far reaching. Subject matter experts, who may understandably lack in-depth knowledge of statistical modeling, are left with the task of assigning sensible priors on parameters whose scale and real-world implications are hard to grasp even for statistical experts.

For this reason, Kadane et al. (1980) and Akbarov (2009) argue that prior elicitation should be conducted using observable quantities, by asking statements related to the *prior predictive distribution*, that is, the distribution of the data as predicted by the model conditioned on the parameters' prior, instead of directly referring to the prior on the unobservable parameters. After eliciting the prior predictive distribution, the information can then be transformed into priors on the parameters by a suitable methodology. The logic of using the prior predictive distribution is that the expert should always have an understanding about plausible values of the observable variables based on their own domain knowledge – even if they may not fully understand the statistical model and the role of parameters used to represent the underlying data generating mechanism. After all, what is an expert if they do not understand their own data?

From a predictive viewpoint, Kadane et al. (1980), Kadane and Wolfson (1998), Geisser (1993), and Akbarov (2009) present practical methods for recovering the prior distribution via expert's information on the prior predictive distribution. Those methods are based on specifying particular moments of the prior predictive distribution for a Gaussian linear regression model, or on providing prior predictive probabilities for fixed subregions of the sample space where the prior distribution is assumed to be univariate. In the latter case, the strategy is to perform least-squares minimization between theoretical probabilities and those probabilities quantified by the expert. However, in the sense of O'Hagan and Oakley (2004), these approaches neglect the fact that the expert's information itself can be uncertain and provide no measure for whether the chosen predictive model is able to reproduce the expert's probabilistic judgements well enough. That is to say, existing methods do not take into account imprecisions in probabilistic judgements when constructing the prior predictive distribution, nor do they provide a principled framework which would guide the experts to select a predictive model and/or prior distribution matching their knowledge (Jeffreys and Zellner, 1980).

Our contribution addresses the question of prior elicitation via prior predictive distributions using a principled statistical framework which 1) makes prior elicitation independent on the specific structure of the probabilistic model from the users' viewpoint, 2) handles complex models with many parameters and potentially multivariate priors, 3) fully accounts for uncertainty in experts/users probabilistic judgements on the data, and 4) provides a formal quality measure indicating if the chosen predictive model is able to reproduce experts' probabilistic judgements. Our work provides both the theoretical basis as well as flexible tools that allow the modeller to express their knowledge in terms of the probability of the data while taking into account the uncertainty in their judgements.

In Section 2, we highlight basic foundation of the Bayesian statistical paradigm, we introduce the prior predictive distribution used throughout the paper to represent expert's opinions about data which would be observed from an experiment. Sections 3 and 4 introduce the methodology to tackle imprecise probabilistic judgements via a principled statistical framework, and general computational procedures to recover the hyperparameters of a prior distribution. The development is interleaved with practical examples illustrating the core concepts and demonstrating its practical use – via concrete instantiations for multivariate prior elicitation for generalized linear models and a small-scale user study comparing the proposed methodology for classical prior elicitation directly on model parameters. We close the paper in Section 5, where conclusions and potential future di-

rections are presented.

2 NOTATION AND PRELIMINARIES

2.1 BAYESIAN APPROACH TO STATISTICAL INFERENCE

The process of performing Bayesian statistical inference usually starts by building a joint probability distribution of observable variables/measurements $\boldsymbol{Y}$ and unobservable parameters $\boldsymbol{\theta}$. The corresponding marginal distribution with respect to $\boldsymbol{\theta}$ is referred to as the prior distribution and the marginal distribution with respect to $\boldsymbol{Y}$ is referred to as the prior predictive distribution. According to the Bayesian paradigm, the prior distribution should be designed independently of the measurement outcomes, that is to say, it must reflect our prior knowledge about the parameters $\boldsymbol{\theta}$ before seeing the actual independent measurements $\mathbf{y}_1, \mathbf{y}_2, \ldots$ (i.e., realizations of $\boldsymbol{Y}$) obtained in the experiments (Berger, 1993; O'Hagan, 2004). After having obtained the measurements, the posterior distribution of $\boldsymbol{\theta}$ arises from the joint distribution by conditioning on $\mathbf{y}_1$, $\mathbf{y}_2$, ... (O'Hagan, 2004).

2.2 PRIOR PREDICTIVE DISTRIBUTION

Let $\boldsymbol{Y} = [Y_1 \ldots Y_S]$ be a S-dimensional vector of observable variables and denote the sample space Ω as a subset of $\mathbb{R}^S$. Hereafter we denote by $\boldsymbol{Y} \,|\, \boldsymbol{\theta} \sim \pi_{\boldsymbol{Y}\,|\,\boldsymbol{\theta}}$ our data probability distribution conditioned on the parameters. We also write $\boldsymbol{\theta} \sim \pi_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^D$ and $\pi_{\boldsymbol{\theta}}$ belongs to a given family of parametric distributions, say $\mathcal{F}_{\boldsymbol{\lambda}}$ indexed by a hyperparameter vector $\boldsymbol{\lambda}$. Then, by marginalizing out the parameters $\boldsymbol{\theta}$, the prior predictive distribution is given by

$$\pi_{\boldsymbol{Y}}(\mathbf{y} \,|\, \boldsymbol{\lambda}) = \int_{\Theta} \pi_{\boldsymbol{Y}\,|\,\boldsymbol{\theta}}(\mathbf{y} \,|\, \boldsymbol{\theta}) \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} \,|\, \boldsymbol{\lambda}) \,\mathrm{d}\boldsymbol{\theta}\,. \quad (1)$$

The prior predictive distribution is not to be confused with the marginal likelihood of observed data, which is obtained by marginalization over $\boldsymbol{\theta}$ of the observed data's sampling distribution times the prior (e.g., Jeffreys and Zellner, 1980).

Given any subset $A \subseteq \Omega$, the prior predictive probability of A, denoted as $\mathbb{P}(\boldsymbol{Y} \in A|\, \boldsymbol{\lambda})$, can be obtained by exchanging the order of integration via the Fubini-Tonelli theorem (Folland, 2013) as

$$\begin{aligned} \mathbb{P}_{A|\,\boldsymbol{\lambda}} &:= \int_A \pi_{\boldsymbol{Y}}(\mathbf{y} \,|\, \boldsymbol{\lambda}) \,\mathrm{d}\mathbf{y} \\ &= \mathbb{E}_{\boldsymbol{\theta}} \left(\mathbb{P}_{\boldsymbol{Y}\,|\,\boldsymbol{\theta}}(\boldsymbol{Y} \in A|\, \boldsymbol{\theta}) \right). \end{aligned} \quad (2)$$

See supplementary materials for details. The hyperparameter vector $\boldsymbol{\lambda}$, which defines a particular prior from the set of all priors $\mathcal{F}_{\boldsymbol{\lambda}}$, will be treated as constant. Hence, no prior needs to be assigned to it. Instead, the values of $\boldsymbol{\lambda}$ will be obtained during the prior predictive elicitation method presented below.

3 PRIOR PREDICTIVE ELICITATION

Our approach follows Gosling (2005) by approaching the elicitation process as a problem of statistical inference where the information to be provided by the expert is in the form of probabilistic judgements about the data. However, the solution itself is novel. From an high-level perspective, our elicitation methodology for any Bayesian model can be summarized as follows:

1. Define the parametric generative model for observable data $\boldsymbol{Y}$ compose of a probabilistic model conditioned on the parameters $\boldsymbol{\theta}$ and a (potentially multivariate) prior distribution for the parameters. The prior distribution depends on hyperparameters $\boldsymbol{\lambda}$ essentially defining the prior which we seek to obtain (see Section 2).
2. Partition the data space into exhaustive and mutually exclusive data categories. For each of these categories, ask the expert what they believe is the probability of the data falling in that category.
3. Model the elicited probabilities from Step 2 as a function of the hyperparameters $\boldsymbol{\lambda}$ from Step 1 while taking into account that the expert information is itself of probabilistic nature and has inherent uncertainty.
4. Perform iterative optimization of the model from Step 3 to obtain an estimate for $\boldsymbol{\lambda}$ describing the expert opinion best within the chosen parametric family of prior distributions.
5. Evaluate how well the predictions obtained from the optimal prior distribution of Step 4 can describe the elicited expert opinion.

In the remainder of this section, we first introduce the basic formalism for modelling the users' beliefs in Section 3.1, provide a key consistency result in Section 3.2, then demonstrate how it can be applied to predictive problems in Section 3.3, and finally discuss the interfaces for the actual knowledge elicitation procedure in Section 3.4. Each part is concluded by an example illustrating the concept.

3.1 MODELING EXPERT OPINIONS

Our assumption is that the output elicitation procedure provides information as probabilistic assignments re-

garding the data vector $\boldsymbol{Y}$ falling within a fixed set of mutually exclusively and exhaustive events $\mathbf{A}$, instead of trying to specify a full density/mass function for the prior predictive distribution, task which would be almost impossible (see Goldstein and Wooff, 2007).

Such collection of assignments is easier to provide, can be considered as the data available for inferring the prior, and is not to be confused by actual measurement data following the generative model. Our focus here is in the mathematical machinery required for converting this information into prior distributions, not taking any stance on how the information is collected from the expert. However, we will briefly discuss the elicitation process itself in Section 3.4.

Let $\mathbf{A} = \{A_1, \ldots, A_n\}$ be a partition of the sample space Ω. Throughout the elicitation procedure, the expert supplies their opinions regarding the quantities $\mathbb{P}_{A_i|\,\boldsymbol{\lambda}}$ for all $i = 1, \ldots, n$. The expert's judgements themselves are not fully deterministic and retain some uncertainty. Also, the expert may be more comfortable to make statements for certain partitions of Ω than for others.

To account for the uncertainty in the probability quantifications of $\mathbb{P}_{A_i|\,\boldsymbol{\lambda}}$, we assume that the obtained judgements $\mathbf{p} = [p_1 \cdots p_n]$ follow a Dirichlet distribution (Ferguson, 1973) with base measure given by the prior predictive probabilities $\mathbb{P}_{A_i|\,\boldsymbol{\lambda}}$ and precision parameter α. Hence, for any chosen partition $\mathbf{A}$ of size n, we denote the distribution of $\mathbf{p}$ as

$$\mathbf{p}\,|\alpha, \boldsymbol{\lambda} \sim \mathcal{D}(\alpha, [\mathbb{P}_{A_1|\,\boldsymbol{\lambda}} \cdots \mathbb{P}_{A_n|\,\boldsymbol{\lambda}}]), \tag{3}$$

where $\mathcal{D}(\cdot)$ stands for Dirichlet distribution and whose multivariate density function reads

$$\mathcal{D}(\mathbf{p}\,|\alpha, \boldsymbol{\lambda}) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{n} \Gamma(\alpha\, \mathbb{P}_{A_i|\,\boldsymbol{\lambda}})} \prod_{i=1}^{n} p_i^{\alpha \mathbb{P}_{A_i|\,\boldsymbol{\lambda}} - 1}. \tag{4}$$

Naturally, we require $\sum_{i=1}^{n} \mathbb{P}_{A_i|\,\boldsymbol{\lambda}} = 1$. The Dirichlet density (4) accounts for the uncertainty inherent to the numerical quantification of the probability vector $\mathbf{p}$ due to, for example, biases introduced through the mechanisms of elicitation processes (the way in which questions are made), practical imperfection (imprecision) of experts' judgements in probabilistic terms or poor judgements on the effect of parameters in the output of the model. For details and in-depth discussion, see O'Hagan and Oakley (2004), O'Hagan (2019) and Sarma and Kay (2020) .

The hyperparameter α measures how well the prior predictive probability model is able to represent (or reproduce) the probability data provided in the elicitation process. The larger the values of α, the less variance around the expected value $\mathbb{P}_{A_i|\,\boldsymbol{\lambda}}$. For practical use of this principle, we can find the maximum likelihood estimate (MLE) $\hat{\alpha}$ of α, which can be directly understood in terms of the deviance between the prior predictive probability and the expert's opinion. More specifically, we have

$$\hat{\alpha} \approx \frac{n/2 - 1/2}{\mathrm{KL}(\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}} \;||\; \mathbf{p})} \tag{5}$$

where $\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}} = [\mathbb{P}_{A_i|\,\boldsymbol{\lambda}} \cdots \mathbb{P}_{A_n|\,\boldsymbol{\lambda}}]^\top$ and $\mathrm{KL}(\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}} \;||\; \mathbf{p})$ is the Kullback-Leibler divergence between the two distributions. The practical interpretation is that for small KL values, we would not be able discriminate the prior predictive probability from the probability data provided by the expert. See supplementary materials for the proof of Equation (5).

Example: Consider a generative model given by $Y|\theta \sim \mathcal{N}(\theta, \sigma^2)$ and $\theta \sim \frac{1}{2}\mathcal{N}(\mu_1, \sigma_1^2) + \frac{1}{2}\mathcal{N}(\mu_2, \sigma_2^2)$. This yields the prior predictive distribution $Y \sim \frac{1}{2}\mathcal{N}(\mu_1, \sigma^2 + \sigma_1^2) + \frac{1}{2}\mathcal{N}(\mu_2, \sigma^2 + \sigma_2^2)$ with hyperparameters $\boldsymbol{\lambda} = [\mu_1,\ \mu_2,\ \sigma^2,\ \sigma_1^2,\ \sigma_2^2]^\top$. For a set $A = (a, b] \subset \mathbb{R}$, the prior predictive probability is $\mathbb{P}_{A|\,\boldsymbol{\lambda}} = \sum_{k=1}^{2} \frac{1}{2}\Phi\big((a - \mu_k)/\sqrt{\sigma^2 + \sigma_k^2}\big) - \frac{1}{2}\Phi\big((b - \mu_k))/\sqrt{\sigma^2 + \sigma_k^2}\big)$. Figure 1 illustrates the effect of the α parameter for a given partition $\mathbf{A}$ with $n = 10$. For each $\alpha \in \{1, 15, 50, 100, 300, 1000\}$. we generated $\mathbf{p}$ by sampling from (3), using fixed hyperparameter values of $\mu_1 = -\mu_2 = 2$ and $\sigma^2 = \sigma_1^2 = \sigma_2^2 = 1$.

3.2 CONSISTENCY WITH RESPECT TO PARTITIONING

Even though we work in a Bayesian context looking to recover a prior distribution, the core procedure of our method applies classical statistical inference. Given a numerical vector of probabilities from the elicitation process, the goal is to show that we are able to find the value of certain parameters (in this case the hyperparameters $\boldsymbol{\lambda}$ and concentration α parameter) of the Dirichlet probabilistic model (3) which would have most likely generated this particular data (of user's subjective knowledge). In other words, we are aiming to obtain the maximum likelihood estimator (MLE).

To study the MLE, we consider the limit where the partitioning is made increasingly more fine grained by increasing n towards infinity. However, we still only obtain information from the user once (i.e., for a single partitioning). That is, the user is providing more and more information about the probabilities, but does not repeat the procedure multiple times. As we will show below, the MLE is consistent under these circumstances, providing the true $\boldsymbol{\lambda}$ when $n \to \infty$, under reasonable assumptions.

Recall that equations (3) and (4) represent the probabilistic model of $\mathbf{p}$ conditioned on the parameters $\boldsymbol{\eta} = (\boldsymbol{\lambda}, \alpha)$.

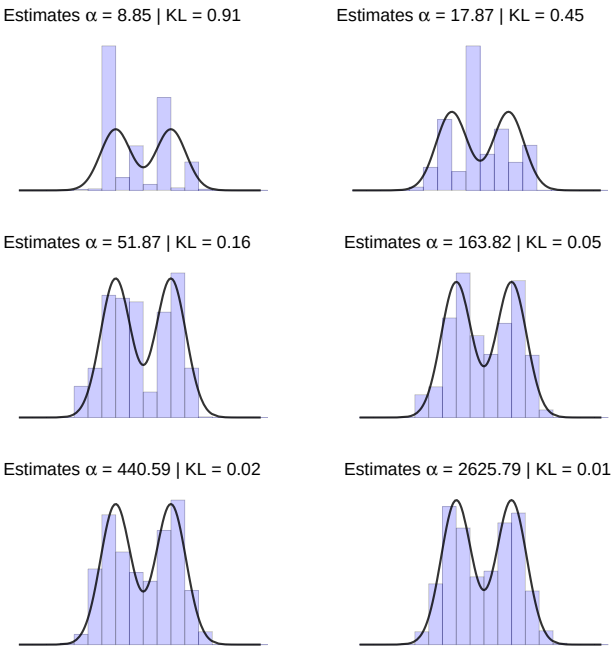

Figure 1: Illustration of the role of the concentration parameter α. Large values correspond to scenarios where the prior predictive distribution (solid line) is able to represent expert's opinions (bars) accurately. That is, α provides an accuracy diagnostic for our method with higher values indicating higher accuracy.

.

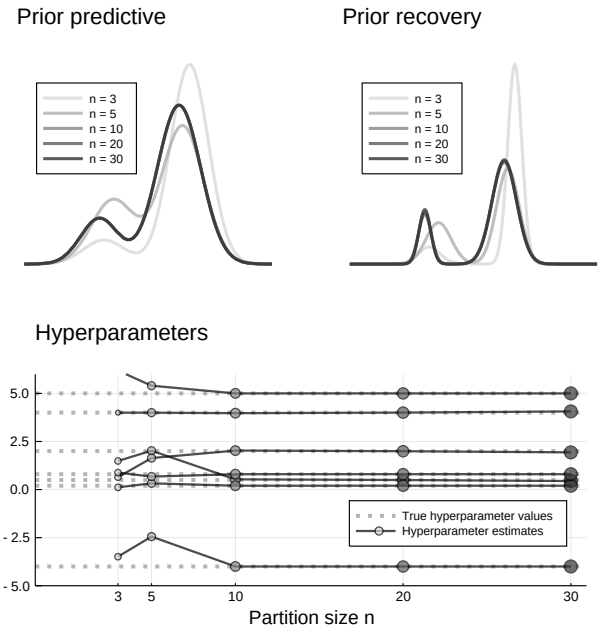

Figure 2: Consistency of the MLE for $\boldsymbol{\lambda}$. **Bottom:** All six hyperparameter values converge to the true values as the number of partitions n increases (each line corresponds to one hyperparameter), here converging already roughly for $n = 10$. **Top:** Both the estimated prior distribution (left) and the corresponding prior predictive distribution (right) converge towards the respective true distributions, depicted as black lines. .

Suppose that there exists a true prior distribution of the expert has hyperparameter values $\boldsymbol{\lambda}_0$ and denote $\boldsymbol{\eta}_0 = (\boldsymbol{\lambda}_0, \alpha_0)$, where α_0 measures the noise in her/his statements . Take the size of the partition n to be large and denote the log-likelihood as $T_{\boldsymbol{\eta}}(\mathbf{p}) = \log \mathcal{D}(\mathbf{p}\,|\alpha, \boldsymbol{\lambda})$ with expectation $Q_{\boldsymbol{\eta}_0}(\boldsymbol{\eta}) = \mathbb{E}_{\mathcal{D}}(T_{\boldsymbol{\eta}}(\mathbf{p}))$.

We show that the expected log-likelihood is maximized at $\boldsymbol{\eta}_0$. By Jensen's inequality, we know that

$$\mathbb{E}_{\mathcal{D}}\left[-\log \frac{\mathcal{D}(\mathbf{p}\,|\alpha, \boldsymbol{\lambda})}{\mathcal{D}(\mathbf{p}\,|\alpha_0, \boldsymbol{\lambda}_0)}\right] > -\log \mathbb{E}\left[\frac{\mathcal{D}(\mathbf{p}\,|\alpha, \boldsymbol{\lambda})}{\mathcal{D}(\mathbf{p}\,|\alpha_0, \boldsymbol{\lambda}_0)}\right] = 0, \qquad (6)$$

yielding

$$Q_{\boldsymbol{\eta}_0}(\boldsymbol{\eta}_0) = \mathbb{E}_{\mathcal{D}}(T_{\boldsymbol{\eta}_0}(\mathbf{p})) > \mathbb{E}_{\mathcal{D}}(T_{\boldsymbol{\eta}}(\mathbf{p})) = Q_{\boldsymbol{\eta}_0}(\boldsymbol{\eta}),$$

which holds for all $\boldsymbol{\eta}$. The expectation $\mathbb{E}_{\mathcal{D}}(\cdot)$ is taken with respect to the distribution (4). The technical condition to ensure uniqueness of the MLE is that the probabilistic model (4) must be identifiable[1]. That is, equality of likelihoods must imply equality of parameters: $\mathcal{D}(\mathbf{p}\,|\alpha_1, \boldsymbol{\lambda}_1) = \mathcal{D}(\mathbf{p}\,|\alpha_2, \boldsymbol{\lambda}_2) \Rightarrow \boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$ for all $\mathbf{p}$. Otherwise we may encounter multiple maxima and thus the prior distribution in the set $\mathcal{F}_{\boldsymbol{\lambda}}$ is not unique.

Example: Extending the earlier example, consider a more general generative model where the prior distribution is now $\theta \sim w_1\mathcal{N}(\mu_1, \sigma_1^2) + w_2\mathcal{N}(\mu_2, \sigma_2^2)$ yielding the prior predictive distribution $Y \sim w_1\mathcal{N}(\mu_1, \sigma^2+\sigma_1^2) + w_2\mathcal{N}(\mu_2, \sigma^2 + \sigma_2^2)$, where w_1 and w_2 are weights summing up to 1 and the hyperparameters are given by $\boldsymbol{\lambda} = [\mu_1, \mu_2, \sigma^2, \sigma_1^2, \sigma_2^2, w_1, w_2]$.

Suppose α is fixed and the true prior distribution has hyperparameters $\boldsymbol{\lambda}_0$. We run an experiment where probability vectors are generated from (3) with increasing partition sizes. Figure 2 shows that, as the partition size increases, the estimates $\hat{\boldsymbol{\lambda}}$ converge to $\boldsymbol{\lambda}_0$, which means the prior distribution is recovered from single-sample elicitation of probability data.

[1] In practise, this may not be an issue when fitting the model. However, we believe it is important to understand the theoretical properties of the inference process so that we can avoid problems in the optimisation procedures.

3.3 COVARIATE-DEPENDENT MODELS AND MULTIVARIATE PRIORS

Next, we demonstrate how the proposed approach can be used for concrete modelling problems, by detailing the procedure for the widely-used family of generalized linear models (GLM; Nelder and Wedderburn, 1972). As GLMs typically have several parameters – one parameter per predicting covariate plus an intercept and potentially a dispersion parameter – direct specification of the parameters' joint prior is often difficult. However, our prior predictive approach can handle this situation elegantly.

In the case of a GLM, our elicitation method requires the selection of sets of covariate values for which the expert is comfortable to express probability judgements about plausible realizations of $\boldsymbol{Y}$. More formally, for each set of covariates $\mathbf{x}_j = [x_{j,1} \cdots x_{j,C}]$, $j = 1, \ldots, J$, the expert provides probability judgements $\mathbf{p}_j = [p_{j,1} \cdots p_{j,n_j}]$ with $\sum_{i_j=1}^{n_j} p_{j,i_j} = 1$, where n_j is the partition size for covariate set j implying the partition $\mathbf{A}_j = \{A_{j,1}, \ldots, A_{j,n_j}\}$. Under the assumption of the judgement $\mathbf{p}_j$ being pairwise conditionally independent, we can express the likelihood function of α and $\boldsymbol{\lambda}$ as

$$\mathcal{D}(\mathbf{p}_1, \ldots, \mathbf{p}_J \,|\alpha, \boldsymbol{\lambda}) = \frac{\Gamma(\alpha)^J}{\prod_{j=1}^{J} \prod_{i_j=1}^{n_j} \Gamma(\alpha\, \mathbb{P}_{A_{j,i_j}|\,\boldsymbol{\lambda},\mathbf{x}_j})} \times \prod_{j=1}^{J} \prod_{i_j=1}^{n_j} p_{j,i_j}^{\alpha \mathbb{P}_{A_{j,i_j}|\,\boldsymbol{\lambda},\mathbf{x}_j} - 1} \tag{7}$$

where $\mathbb{P}_{A_{j,i_j}|\,\boldsymbol{\lambda},\mathbf{x}_j}$ is the prior predictive probability for the set A_{j,i_j} related to covariate set $\mathbf{x}_j$.

Importantly, there is no need for the partitions themselves or their size to be the same throughout the sets of covariate values: For each j, the expert can create any partition they are most comfortable with making judgements about. This feature provides much more freedom to the expert in expressing their knowledge of the data compared to alternative methods. For example, to obtain a prior distribution for logistic regression model, the method of Bedrick et al. (1997) requires the user to provide a fixed number of probabilities just enough to make the Jacobians appearing in their method invertible.

Example: Here we consider a generative model for binary data in the presence of a vector of covariates. The observable variable conditioned on the parameters is distributed according to a Bernoulli model and we take a multivariate Gaussian distribution as the prior distribution for the vector of parameters in the predictor function. This can be formalized as

$$Y|\,\boldsymbol{\theta} \sim \mathcal{B}\big(\Phi(\mathbf{x}^\top \boldsymbol{\theta})\big) \qquad \boldsymbol{\theta} \sim \mathcal{N}_D(\boldsymbol{\mu}, \Sigma) \tag{8}$$

yielding the prior predictive distribution

$$Y \sim \mathcal{B}\Big(p(\mathbf{x}, \boldsymbol{\lambda})\Big) \tag{9}$$

with $p(\mathbf{x}, \boldsymbol{\lambda}) = \Phi(\mathbf{x}^\top \boldsymbol{\mu} \,/\sqrt{1 + \mathbf{x}^\top \,\Sigma\, \mathbf{x}}\,)$.

The notation $\mathcal{N}_D(\cdot, \cdot)$ stands for a D-dimensional Gaussian distribution and $\mathcal{B}(\cdot)$ for the Bernoulli distribution. The hyperparameter vector $\boldsymbol{\lambda} = [\boldsymbol{\mu}, \Sigma]$, consists of the prior means $\boldsymbol{\mu} = [\mu_1, \cdots, \mu_D]$ and prior covariance matrix Σ. We fix the partitioning throughout the covariate set as $A_{j,1} = \{0\}$, $A_{j,2} = \{1\}$ since $Y \in \Omega = \{0, 1\}$. Equation (2) simplifies to $\mathbb{P}_{A_1|\,\boldsymbol{\lambda}} = 1 - p(\mathbf{x}, \boldsymbol{\lambda})$ and $\mathbb{P}_{A_2|\,\boldsymbol{\lambda}} = p(\mathbf{x}, \boldsymbol{\lambda})$.

The notation $\mathcal{N}_D(\cdot, \cdot)$ stands for a D-dimensional Gaussian distribution and $\mathcal{B}(\cdot)$ for the Bernoulli distribution. The hyperparameter vector $\boldsymbol{\lambda} = [\boldsymbol{\mu}, \Sigma]$, consists of the prior means $\boldsymbol{\mu} = [\mu_1, \cdots, \mu_D]$ and prior covariance matrix Σ. We fix the partitioning throughout the covariate set as $A_{j,1} = \{0\}$, $A_{j,2} = \{1\}$ since $Y \in \Omega = \{0, 1\}$. Equation (2) simplifies to $\mathbb{P}_{A_1|\,\boldsymbol{\lambda}} = 1 - p(\mathbf{x}, \boldsymbol{\lambda})$ and $\mathbb{P}_{A_2|\,\boldsymbol{\lambda}} = p(\mathbf{x}, \boldsymbol{\lambda})$.

The parametrisation of the covariance matrix follows the separation strategy suggested by Barnard et al. (2000) on an unconstrained space as presented by Kurowicka and Cooke (2003). That is, the covariance matrix is rewritten as $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_D^2)\;\; R \;\; \mathrm{diag}(\sigma_1^2, \ldots, \sigma_D^2)$ where $(\sigma_1^2, \ldots, \sigma_D^2)$ are the variances and R is the correlation matrix.

In the simulation experiment, we vary the dimension $D \in \{2, 3, 4, 5, 6\}$ and the number of sets of covariates $J \in \{3, 5, 15, 30, 80\}$. For each D we randomly pick a true value for $\boldsymbol{\lambda}$, and for each covariate set, we draw random probabilities of success/failure from the Dirichlet probability model. Hence, the likelihood is given by (7). We repeat the procedure for each D and J where the hyperparameters $\boldsymbol{\lambda}$ are fixed with respect to J.

To show the convergence with respect to the estimates of Σ obtained from the expert judgements, we compare the logarithm of the Frobenius norm between the estimated covariance matrix and the true covariance matrix (Fig. 3). For sufficiently large J, roughly from $J = 15$ onwards, we are able to accurately elicit multivariate priors up to 5-6 dimensional priors – this is a significant improvement over earlier methods that have been limited to univariate or at most bivariate priors (Moala and O'Hagan, 2010). For increasing D from 2, 3, 4, 5 to 6, the respective number of hyperparameters in the vector $\boldsymbol{\lambda}$ becomes 5, 9, 14, 20 to 27, explaining the increased elicitation difficulty for large D.

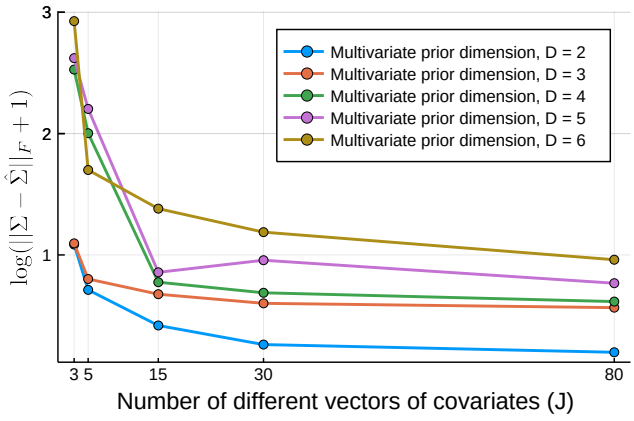

Figure 3: Convegence of the covariance matrix estimates for multivariate prior elicitation for binary linear regression as a function of the number of covariates J for which the user provides probability estimates, measured using the logarithm of the Frobenius norm of the difference between the true covariance matrix and the estimate. The coloured lines refer to the dimensionality D of the prior distribution, showing that we can effectively elicit multivariate priors of reasonable dimensionality, with naturally increasing difficulty for larger D.

3.4 PRIOR ELICITATION IN PRACTICE

Using the machinery above requires obtaining the probability judgements $\mathbf{p}$ from the user. The method itself is general, and can be used as part of any practical Bayesian modelling workflow when linked to any particular elicitation interface. We have implemented an extension of the SHELF interface (Oakley and O'Hagan, 2019) as a reference, by replacing the direct parameter elicitation components with variants that query the user for the prior predictive probabilities. This readily provides practical elicitation methods for the user to specify probabilities by utilizing probability quantiles or roulette chips. This means that probability ratios for events are provided and then individual probabilities are recovered under the natural constraint $\sum_{i=1}^{n} p_i = 1$. Hence, the user can choose the way of providing information they feel most comfortable with. Besides graphical interfaces, the elicitation can be carried out by the modeller interviewing a domain expert. Experienced modellers may also choose to simply express some particular priors via providing $\mathbf{p}$ while designing the model.

Example: To evaluate the applicability of our method in practice, we conducted a small user study of $N = 5$ doctoral students of computer science with reasonable statistical knowledge. The task was to elicit priors of a human growth model (see Preece and Baines, 1978, model 1, Section 2) with a six-dimensional hyperparameter vector $\boldsymbol{\lambda}$. We queried the users for $n_j = 6$ probabilities and $J = 4$ covariates, each corresponding to stature distribution of males at the age of $t \in \{0, 2.5, 10, 17.5\}$ years. We chose this model because everyone can be expected to have a rough understanding of the observed data and hence can act as an expert. As a baseline, we used a standard elicitation procedure which queries the prior distributions for each parameter directly (again with $n = 6$). Some of these parameters are intuitive (e.g., stature as adult) while some control the quantitative behaviour of the model in a non-trivial way. The model was implemented in `brms` (Bürkner, 2017) to demonstrate compatibility with existing modelling tools. Gradient-free optimization (see next section) was used for converting the elicited probabilities into priors. Table 1 shows exemplary for one user how the prior predictive distribution corresponding to $\boldsymbol{\lambda}$ elicited with the proposed method matches well with results of Preece and Baines (1978). When applying direct parameter elicitation, the match was clearly worse because the user was unable to provide reasonable estimates for parameters without an intuitive meaning, despite being provided an explanation of the model and its parameters. In a standardized interview, all users reported that they were more comfortable providing probability judgements for the observables than for the parameters, and that they were more confident that the resulting prior matches their actual subjective prior. See supplementary materials for details of the model and user study, as well as results for all users.

Table 1: Result of a real prior elicitation experiment for one user, characterized by statistics of the prior distribution. The proposed approach (Predictive) better matches the parameters found by fitting the model to actual data Preece and Baines (Reference; 1978), compared to direct parameter elicitation (Parametric). This is visible in the lower α estimate as well. The reference column excludes b due to their use of a non-probabilistic model.

		Predictive		**Parametric**	
Parameter	Reference	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$	$\mathbb{E}[\cdot]$	$\mathbb{V}(\cdot)$
h_1	174.6	174.5	0.8	176.2	105.3
h_{t_*}	162.9	162.8	4.2	129.1	33.6
s_0	0.1	0.1	< 0.1	1.2	1.13
s_1	1.2	3.3	0.21	1.2	1.13
t_*	14.6	13.4	0.01	12.5	0.57
b	–	15.79	12.9	1.97	4.57
α	–	6.9	–	1.2	–

4 ON THE LEARNING ALGORITHMS

Having characterized the problem itself and its asymptotic properties, we now turn our attention to the computational problem of estimating the hyperparameter vector $\boldsymbol{\lambda}$ and the uncertainty parameter α in practice. We start by mentioning basic notions for the type of models and properties over which our method is able to accommodate and systematise general purpose model independent computer algorithms.

The methodology presented in Section 3 supports both discrete and continuous components in the observables variables $\boldsymbol{Y}$, or combinations of both. It also works for any data dimension S and any parameter dimension D. Interesting cases are when $S = 1$ and $D > 1$, meaning that, as we have showed previously, we can recover a multivariate prior distribution from probability judgements of 1-dimensional observable variable. This is novel in the recent literature.

For arbitrary S, where we would possibly work with a multivariate distribution over a vector of observable variables, probabilities for a generic rectangular set $A = \bigtimes_{s=1}^{S}(a_s, b_s]$ can be formulated via the cumulative distribution function of the prior predictive distribution (1) as follows. Let $I = (a, b]$ be an interval, g some function with $g : \mathbb{R}^S \to \mathbb{R}$, and Δ_I^s the difference operator with $\Delta_I^s = g(y_1, \ldots, y_{s-1}, b) - g(y_1, \ldots, y_{s-1}, a)$. Then, equation (2) takes the general form

$$\begin{aligned}\mathbb{P}_{A|\,\boldsymbol{\lambda}} &= \int_{a_1}^{b_1} \cdots \int_{a_S}^{b_S} \pi_{\boldsymbol{Y}\,|\,\boldsymbol{\lambda}}(y_1, \ldots, y_S)\mathrm{d}y_1 \ldots \mathrm{d}y_S \\ &= \Delta_{I_1}^1 \Delta_{I_2}^2 \cdots \Delta_{I_S}^S F_{\boldsymbol{Y}\,|\,\boldsymbol{\lambda}}(y_1, \ldots, y_S), \qquad (10)\end{aligned}$$

where $F_{\boldsymbol{Y}\,|\,\boldsymbol{\lambda}}(\cdot)$ is the cumulative distribution function of the prior predictive distribution (1). Cases in which $S > 1$ appear, for example, in lifetime analysis or Markovian models. In lifetime analysis, components of electronic equipments are dependent and there is a need to consider bivariate models in the first level of the generative model (Lawless, 2011). Markovian models are widely used to model natural phenomena such as population growth, climate, traffic, and language models in which multiple measurement variables naturally occur (Kijima, 1997).

Natural gradients for closed-form cases: If equation (10) is available in closed-form, usual gradient-based optimisation algorithms are applicable. We recommend using natural gradients (Amari, 1998), which have been widely applied for statistical machine learning problems (e.g., see Girolami and Calderhead, 2011). In this case, the Fisher information matrix for $\boldsymbol{\lambda}$ can be computed in closed-form using results from the original parametrisation of the Dirichlet distribution (Ferguson, 1973) as

$$H_{\boldsymbol{\lambda}} = (\tfrac{d}{d\boldsymbol{\lambda}}\,\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}})^{\top}\, H_{\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}}}\, (\tfrac{d}{d\boldsymbol{\lambda}}\,\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}}), \qquad (11)$$

where $H_{\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}}} = \alpha^2(\mathrm{diag}(\psi'(\alpha\,\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}})) - \psi'(\alpha)\mathbb{1}\mathbb{1}^{\top})$ is the Fisher information matrix of the standard Dirichlet distribution, $\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}} = [\mathbb{P}_{A_1|\,\boldsymbol{\lambda}} \cdots \mathbb{P}_{A_n|\,\boldsymbol{\lambda}}]^{\top}$, and $\frac{d}{d\boldsymbol{\lambda}}\,\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}} = \left[\frac{d}{d\lambda_1}\,\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}} \;\cdots\; \frac{d}{d\lambda_M}\,\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}}\right]^{\top}$. The function $\psi'(\cdot)$ is the the derivative of the digamma function and $\frac{d}{d\lambda_M}\mathbb{P}$ is the derivative of the vector $\mathbb{P}$ with respect to an element in the vector of hyperparameters $\boldsymbol{\lambda}$. Due to the closed-form expression, we can use natural gradients with almost no additional computational cost. The only extra step is the calculation of $\frac{d}{d\lambda_M}\mathbb{P}$ which can be obtained easily with automatic differentiation regardless of the chosen generative model.model.

Stochastic natural gradients optimization: If (10) cannot be expressed in closed-form but the equation (4) or (7) are differentiable with respect to $\boldsymbol{\lambda}$, one can use gradient-based optimization with *reparametrisation gradients* and automatic differentiation. The elements of $\mathbb{P}$ are expected values with respect to the prior distribution (2), and the goal is then to find a pivotal function for the prior (see Casella and Berger, 2001, page 427, Section 9.2.2) and obtain Monte-Carlo estimates of it (which is not difficult once we can use the representation (10)) and gradients $\frac{d}{d\lambda_M}\mathbb{P}$ with very low computational cost according to Figurnov et al. (2018).

When the generative model has a higher level hierarchical structure, such as $\boldsymbol{Y}\,|\,\boldsymbol{\theta}_1 \sim \pi(\mathbf{y}\,|\,\boldsymbol{\theta}_1)$, $\boldsymbol{\theta}_1\,|\,\boldsymbol{\theta}_2 \sim \pi(\boldsymbol{\theta}_1\,|\,\boldsymbol{\theta}_2)$, …, $\boldsymbol{\theta}_L\,|\,\boldsymbol{\lambda} \sim \pi(\boldsymbol{\theta}_L\,|\,\boldsymbol{\lambda})$, we can show that the elements of $\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}}$ and $\frac{d}{d\lambda_M}\,\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}}$ can also be computed efficiently together with a stochastic estimation of the hyperparameters' Fisher information matrix. That is

$$\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}} = \mathbb{E}_{X_L}\left(\mathbb{E}_{X_{L-1}} \cdots \left(\mathbb{E}_{X_1}\left(\mathbb{P}_{A|f_1(\boldsymbol{\lambda})}\right)\right)\right) \qquad (12)$$

where X_ℓ are pivotal quantities with respect to distributions $\pi(\boldsymbol{\theta}_\ell\,|\,\boldsymbol{\theta}_{\ell+1})$ for $\ell = 1, \ldots, L$ and $f_1(\boldsymbol{\lambda})$ is a function which depends only on the hyperparameters $\boldsymbol{\lambda}$. Gradients are estimated similarly as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda_m}\,\mathbb{P}_{\boldsymbol{A}|\,\boldsymbol{\lambda}} = \mathbb{E}_{X_L}\Bigg(&\mathbb{E}_{X_{L-1}} \cdots \\ &\cdots \left(\mathbb{E}_{X_1}\left(\frac{\mathrm{d}f_1}{\mathrm{d}\lambda_m}\frac{\mathrm{d}}{\mathrm{d}f_1}\mathbb{P}_{A|f_1(\boldsymbol{\lambda})}\right)\right)\Bigg) \qquad (13)\end{aligned}$$

The equations above can be plugged into (11) to obtain an estimation for the hyperparameters' Fisher information matrix. The proof and detailed explanations are provided in the supplementary materials.

Gradient-free optimization: Finally, for completely arbitrary models, we can step outside of gradient-based optimization and use general-purpose global optimization tools for determining $\boldsymbol{\lambda}$. Methods such as Bayesian optimization and Nelder-Mead only require the ability to evaluate the objective (10), and many practical optimization libraries (e.g. `optimR`) provide extensive range of practical alternatives. For models with relatively small number of hyperparameters, we have found such tools to work well in practice. However, whenever either of the gradient-based methods described above is applicable, we recommend using them due to substantially improve efficiency.

Optimization of α: Finally, besides $\boldsymbol{\lambda}$, we usually want to estimate α as well which quantifies the uncertainty as explained in Section 3.1. One can either directly optimise (4) for $(\boldsymbol{\lambda}, \alpha)$ together, or switch optimisation of (4) for $\boldsymbol{\lambda}$ with fixed α with optimization of (4) for α with fixed $\boldsymbol{\lambda}$. This may be easier since we have an approximate closed-form expression for α provided in the supplementary materials.

5 DISCUSSION AND CONCLUSIONS

Prior elicitation is an important stage in the Bayesian modeling workflow (Schad et al., 2019), especially for hierarchical models whose parameters have a complex relationship with the observed data. Standard prior elicitation strategies, such as O'Hagan and Oakley (2004); Moala and O'Hagan (2010), do not really help in such scenarios, since the expert still needs to express information in terms of probability distribution of the model's parameters. The idea of eliciting knowledge in terms of the observable data is not new – in fact, it dates back to Kadane et al. (1980). However, to our knowledge we proposed the first practical formulation that accounts for uncertainty in the expert's judgements of the prior predictive distribution, with easy, general, and complete implementation that allows eliciting both univariate and multivariate prior distributions more efficiently.

We demonstrated the general formalism in several practical contexts, ranging from simple conceptual illustrations and technical verifications to real elicitation examples. In particular, we showed that multivariate priors (of reasonable dimensionality) can be elicited in context of generalized linear models based on relatively small collection of probability judgements for different covariate sets. The approach can be coupled with existing modelling tools and used for eliciting prior information from real users, as demonstrated for the human growth model of Preece and Baines (1978) implemented in `brms` (Bürkner, 2017). Even though we only carried out a simplified and small-case experiment, the results already indicate that even users familiar with statistical modelling were more comfortable expressing knowledge of the observed data rather than model parameters, and that the resulting priors better matched their beliefs.

The obvious continuation of this work would consider tighter integration of the method into a principled Bayesian workflow, coupled with more extensive user studies. We also look forward to extend our method to cases of multiple experts opinions about the same observable variables. As a first attempt, we could consider the same predictive model and distinct α's for multiple experts. However, more work is needed in that regard.

Acknowledgements

This work was supported by the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI; Grants 320181, 320182, 320183) and the Technology Industries of Finland Centennial Foundation (grant 70007503; Artificial Intelligence for Research and Development).

References

Akbarov, A. (2009) *Probability elicitation: Predictive approach*. Ph.D. thesis, University of Salford.

Amari, S. (1998) Natural Gradient Works Efficiently in Learning. *Neural Computation (communicated by Steven Nowlan and Erkki Oja)*, **10**, 251–276.

Barnard, J., McCulloch, R. and Meng, X.-L. (2000) Modelling covariance matrices in terms of standart deviations and correlations, with applications to shrinkage. *Statistical Sinica*, **4**, 1281–1311.

Bedrick, E. J., Christensen, R. and Johnson, W. (1997) Bayesian binomial regression: Predicting survival at a trauma center. *The American Statistician*, **51**, 211–218.

Berger, J. O. (1993) *Statistical decision theory and Bayesian analysis*. Springer series in Statistics. Springer-Verlag, 2nd ed edn.

Bürkner, P.-C. (2017) BMRS: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, **80**, 1–28.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017) Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, **76**, 1–32.

Casella, G. and Berger, R. L. (2001) *Statistical Inference*. Duxbury Press, 2 edn.

Daee, P., Peltola, T., Soare, M. and Kaski, S. (2017) Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. *Machine Learning*, **106**, 1599–1620.

Ferguson, T. S. (1973) A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, **1**, 209–230.

Figurnov, M., Mohamed, S. and Mnih, A. (2018) Implicit reparameterization gradients. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, 439–450.

Folland, G. (2013) *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley.

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.

Geisser, S. (1993) *Predictive inference: An introduction*. Springer.

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.

Gelman, A., Simpson, D. and Betancourt, M. (2017) The prior can often only be understood in the context of the likelihood. *Entropy*, **19**, 555.

Girolami, M. and Calderhead, B. (2011) Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods. *Journal of the Statistical Royal Society B*, **73**, 123–214.

Goldstein, M. and Wooff, D. (2007) *Bayes Linear Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. Wiley.

Gosling, J. (2005) *Elicitation: A nonparametric view*. Ph.D. thesis, University of Sheffield.

Jeffreys, H. and Zellner, A. (1980) *Bayesian analysis in econometrics and statistics: essays in honor of Harold Jeffreys*. Studies in Bayesian econometrics. North-Holland Pub. Co.

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

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.

Kijima, M. (1997) *Markov Processes for Stochastic Modeling*. Stochastic Modeling Series. Taylor & Francis.

Kurowicka, D. and Cooke, R. (2003) A parameterization of positive definite matrices in terms of partial correlation vines. *Linear Algebra and its Applications*, **372**, 225–251.

Lawless, J. (2011) *Statistical Models and Methods for Lifetime Data*. Wiley Series in Probability and Statistics. Wiley.

Lindley, D. (1983) Reconciliation of probability distributions. *Operations Research*, **31**, 866–880.

Meent, J.-W. V. D., Paige, B., Yang, H. and Wood, F. (2018) An introduction to probabilistic programming. *ArXiv*.

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.

Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, **135**, 370–384.

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).

O'Hagan, A. (1978) Curve fitting and optimal design for prediction. *Journal of Royal Statistical Society B*, **40**, 1–42.

— (2004) *Kendall's Advanced Theory of Statistics: Bayesian Inference*. Oxford University Press.

— (2019) Expert knowledge elicitation: Subjective but scientific. *The American Statistician*, **73**, 69–81.

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.

Preece, M. A. and Baines, M. J. (1978) A new family of mathematical models describing the human growth. *Annals of Human Biology*, **5**, 1–24.

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.

Schad, D. J., Betancourt, M. and Vasishth, S. (2019) Toward a principled Bayesian workflow in cognitive science. *ArXiV:1904.12765*.

Simpson, D., Rue, H., Martins, T., Riebler, A. and Sørbye, S. (2017) Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, **32**, 1–28.