Discovering Inductive Bias with Gibbs Priors: A Diagnostic Tool for Approximate Bayesian Inference

Luca Rendsburg[1] **Agustinus Kristiadi**[1] **Philipp Hennig**[1,2] **Ulrike von Luxburg**[1,2]

[1]University of Tübingen [2]Max Planck Institute for Intelligent Systems, Tübingen

Abstract

Full Bayesian posteriors are rarely analytically tractable, which is why real-world Bayesian inference heavily relies on approximate techniques. Approximations generally differ from the true posterior and require diagnostic tools to assess whether the inference can still be trusted. We investigate a new approach to diagnosing approximate inference: the approximation mismatch is attributed to a change in the inductive bias by treating the approximations as exact and reverse-engineering the corresponding prior. We show that the problem is more complicated than it appears to be at first glance, because the solution generally depends on the observation. By reframing the problem in terms of incompatible conditional distributions we arrive at a natural solution: the *Gibbs prior*. The resulting diagnostic is based on pseudo-Gibbs sampling, which is widely applicable and easy to implement. We illustrate how the Gibbs prior can be used to discover the inductive bias in a controlled Gaussian setting and for a variety of Bayesian models and approximations.

1 INTRODUCTION

Bayesian inference is based on the posterior distribution $p(\theta|y)$ over latent variables θ given an observation y. Bayes' theorem gives an explicit formula for computing the posterior, but is often infeasible in practice because the latent space is too large to work with, the appearing integrals are intractable, or the likelihood function cannot be evaluated. In these cases, practitioners revert to approximating the posterior instead. This approach comprises a cornucopia of methods, which can be divided into two groups. The first group consists of deterministic approximation methods that compute a feasible approximating distribution [1] $q(\theta|y)$ to the exact posterior $p(\theta|y)$ and includes methods such as variational inference (Hinton and van Camp, 1993; Jordan et al., 1999; Blei et al., 2017; Hoffman et al., 2013; Ranganath et al., 2014; Kucukelbir et al., 2017), Laplace approximations (Spiegelhalter and Lauritzen, 1990; MacKay, 1992; Rue et al., 2009, 2017; Daxberger et al., 2021), and expectation propagation (Minka, 2001). The second group consists of stochastic sampling methods that generate samples from (an approximation to) the posterior and includes methods such as Markov chain Monte Carlo (Casella and George, 1992; Hoffman and Gelman, 2014; Bardenet et al., 2017) and approximate Bayesian computation (Diggle and Gratton, 1984; Sisson et al., 2018; Beaumont, 2019). For a general introduction to approximate methods in Bayesian inference see Bishop (2006). While approximate methods make Bayesian inference feasible, they come at the cost of a distortion in the posterior. The resulting approximate inference can deviate significantly from exact Bayesian inference. This calls for diagnostic tools to assess whether the result can still be trusted. Most existing diagnostics suffer from one or more of the following weaknesses: they are specific to a particular setting, they require evaluating the density of the approximation, which is unavailable for sampling-based methods, or they are restricted to the marginal distributions of a multivariate posterior. An overview of diagnostic tools is given in Section 2.

Existing diagnostics describe the difference to exact Bayesian inference by assessing the mismatch between approximation and true posterior. In contrast, we investigate a new perspective for diagnostic tools: we describe the approximate inference directly by attribut-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

[1] While standard notation for the approximation is $q(\theta)$, it will be useful in the context of this paper to think of it as a conditional distribution.

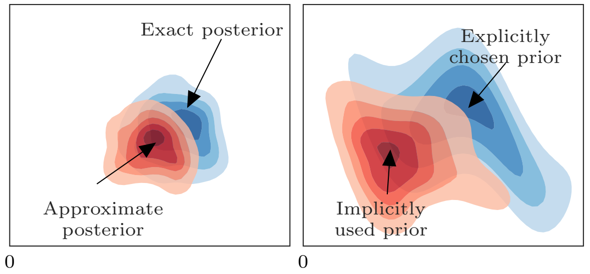


Figure 1: **Left:** a posterior approximation is biased towards solutions of small norm. **Right:** the approximation corresponds to the exact posterior under another implicitly defined prior, which is itself biased towards solutions of small norm.

ing this mismatch to a change in the inductive bias. In a fully Bayesian setting, the inductive bias is specified explicitly by the model, which consists of the prior (a priori preference for solutions) and the likelihood (data generating process). Approximating the posterior can introduce additional bias that is not reflected in the model specification. We fix the likelihood and only allow the prior to change. The main idea of this work is to treat the approximation as an exact posterior to the same likelihood and reverse-engineer the corresponding implicitly used prior:

$$\begin{tikzcd}
\text{Explicitly chosen prior} \;\&\; \text{Likelihood} \arrow[r, "\textit{exact}", "\textit{inference}"'] & \text{Exact posterior} \arrow[d, squiggly] \\
\text{Implicitly used prior} \;\&\; \text{Likelihood} \arrow[r, "\textit{exact}", "\textit{inference}"'] & \text{Approximate posterior} \arrow[l, dashed, bend left, "\textit{reverse-engineer}"]
\end{tikzcd}$$

This implicit prior describes the inductive bias of the approximation in terms of an a priori preference for solutions. Figure 1 shows an example of inference based on posterior approximations that are biased towards solutions of small norm. This corresponds to effectively using a different prior with more mass on solutions of small norm than the explicitly chosen prior.

Let $(f(\cdot|\theta))_\theta$ be the likelihood and $(q(\cdot|y))_y$ the approximations to the posteriors $(p(\cdot|y))_y$. It is reasonable to define the implicit prior to the approximations by fixing an observation y and simply reverting Bayes' theorem[2] $\pi_y(\theta) \propto_\theta q(\theta|y)/f(y|\theta)$. Unfortunately, π_y generally depends on the observation y. This means that the approximations to different observations can correspond to different implicit priors, in which case no single distribution $\tilde{\pi}$ satisfies $q(\theta|y) \propto_\theta \tilde{\pi}(\theta) f(y|\theta)$. We only have the following weaker interpretation:

> Inference based on the approximate posteriors $(q(\cdot|y))_y$ is exact Bayesian inference with the same likelihood $(f(\cdot|\theta))_\theta$, but the prior is chosen from the family $(\pi_y)_y$ depending on the observation y.

Of course, the prior should not depend on the observation if we want to interpret it as the a priori preference for solutions. To understand the inductive bias of the approximations, we need an observation-independent distribution to compromise between this family of priors. We look at this problem through the lens of incompatible conditional distributions (Arnold and Press, 1989). This yields a natural solution based on pseudo-Gibbs sampling, which we call the *Gibbs prior.* An introduction to incompatible conditionals and pseudo-Gibbs sampling is given in Appendix A.

Observation-(in)dependent diagnostics A diagnostic can either treat an approximation under a *fixed observation* $q(\cdot|y)$ or assess the average behavior of the approximation method *across observations* $(q(\cdot|y))_y$. These different tasks can show opposing behavior because an approximation can be good on specific instances but bad in general, or vice versa. Diagnosing a single approximation helps to understand and improve the inference under the fixed observation, but does not inform about how the approximation method performs in other cases. In our setting, this task is performed by the distributions π_y. However, we are interested in the systematic bias of the whole approximation method, which is why we search for an observation-independent compromise between the π_y. This kind of diagnostic does not guarantee the same behavior on any fixed observation, but helps to understand the method itself.

Contributions

- We investigate the novel approach of diagnosing approximate Bayesian inference methods in terms of their inductive bias. We show that this requires a compromise and reframe it as a problem of incompatible conditional distributions.
- We propose the Gibbs prior as a natural solution to the above problem (Section 3) and as a diagnostic tool. It is based on pseudo-Gibbs sampling, which is widely applicable and easy to implement.
- We demonstrate how the Gibbs prior can be used to discover the inductive bias of approximate Bayesian inference methods in a Gaussian toy example (Section 4) and two intractable Bayesian models (Section 5).

[2] Note that π_y can be improper, that is, not integrable.

2 RELATED WORK

We divide the literature for diagnostics into two broad categories, depending on how they assess an approximation mismatch. Diagnostics in the first category compute a divergence between (quantities related to) the posterior and its approximation. Gorham and Mackey (2015, 2017) compute Stein discrepancies between the posterior and its approximation. Cusumano-Towner and Mansinghka (2017) compute the symmetric KL divergence between the approximation and another baseline approximation. Domke (2021) computes the symmetric KL divergence between the true joint distribution $p(y)p(\theta|y)$ and its approximation $p(y)q(\theta|y)$. Huggins et al. (2020) use the Wasserstein distance to bound the error of posterior point estimates. Diagnostics in the second category consider derived quantities that are known exactly under the true posterior and test whether they deviate under the approximations. Xing et al. (2020) compare a distortion map for posterior cumulative distribution functions to the identity. Yu et al. (2021) compare average posterior means and covariances to prior means and covariances. Cook et al. (2006) initiate another line of work based on the distribution of posterior quantiles, which is tested for uniformity; a corrected implemenation is presented by Talts et al. (2018). Yao et al. (2018) relax the uniformity test of Cook et al. (2006) and only test for symmetry. They also present another diagnostic based on Pareto-smoothed importance sampling. Prangle et al. (2014) test for uniformity of p-values related to the coverage property; this method is extended by Rodrigues et al. (2018). Our diagnostic also falls into this category where the Gibbs prior is compared to the original prior. The above diagnostics can also be divided by whether they analyze approximation methods for fixed or general observations. Our goal of diagnosing average approximation behavior is shared by Domke (2021); Yu et al. (2021); Cook et al. (2006); Talts et al. (2018); Yao et al. (2018).

Our diagnostic is based on sampling alternatingly from likelihood and approximation. The same technique was originally used by Geweke (2004) under the name *successive-conditional simulator* with the same goal of diagnosing approximations. Although both diagnostics are based on the same technique, they apply it differently: Geweke (2004) uses the simulator without reference to compatibility for generating tuples $(\tilde{\theta}_i, \tilde{y}_i)_i$, which are tested against samples from the Bayesian model $(\theta_i, y_i)_i$ to assess whether the approximations are exact; we focus on the marginal values $(\tilde{\theta}_i)_i$ that describe the implicitly used prior to assess the inductive bias. Our diagnostic is also similar in spirit to Joshi and Ruggeri (2020) who link distortions in the likelihood to distortions in the prior.

3 METHOD

3.1 Preliminaries

Let $\pi(\theta)$ be a proper prior distribution on a space of latent variables $\theta \in \Theta$ and $f(y|\theta)$ a positive likelihood on a space of observations $y \in \mathcal{Y}$. The corresponding posterior distribution is denoted by $p(\theta|y)$. For every fixed y let $q(\theta|y)$ denote the approximation to the posterior given by the approximate method in question. For sampling-based methods this distribution cannot be evaluated because it is specified only implicitly through samples, which suffices for our diagnostic. We denote the families of distributions as $F := (f(\cdot|\theta))_{\theta\in\Theta}$, $P := (p(\cdot|y))_{y\in\mathcal{Y}}$, and $Q := (q(\cdot|y))_{y\in\mathcal{Y}}$. The families F and Q are called *compatible* if there exists a joint distribution on $\Theta \times \mathcal{Y}$ which has F and Q as conditionals. They are called *incompatible* if they are not compatible (Arnold and Press, 1989).

Our goal is to understand the inductive bias of inference based on the approximations $q(\theta|y)$ in terms of an a priori preference for solutions. The bias is fully encoded in the original prior $\pi(\theta)$ if the approximation is perfect. However, a mismatch $q(\theta|y) \neq p(\theta|y)$ can introduce additional bias, which is not captured by the original prior. The main idea of this paper is to treat the approximation as an exact posterior and look for the corresponding prior distribution $\tilde{\pi}(\theta)$. This new prior describes the combination of explicitly encoded bias $\pi(\theta)$ and implicitly incurred bias because of approximation mismatch. We can then compare those priors to gain insights into how the approximation changes the inductive bias.

3.2 Assessing the Inductive Bias of Posterior Approximations with Gibbs Priors

This section describes the problem of finding a prior to the approximations from the perspective of incompatible conditionals. We first motivate the problem by considering fixed observations and then propose a solution based on pseudo-Gibbs sampling.

For a fixed observation $y \in \mathcal{Y}$, the implicit *pointwise prior* π_y corresponding to $q(\cdot|y)$ is defined via

$$\pi_y(\theta) \propto_\theta \frac{q(\theta|y)}{f(y|\theta)} \,. \tag{1}$$

This describes the inductive bias of the approximation $q(\cdot|y)$ for a fixed observation, but it is not necessarily the same across different observations. The pointwise prior π_y will depend on y if and only if the conditional families F and Q are incompatible, which is a simple consequence of the definition. Informally, the scatter of the family $(\pi_y)_{y\in\mathcal{Y}}$ is an indicator for the degree of compatibility: in the compatible case, all π_y

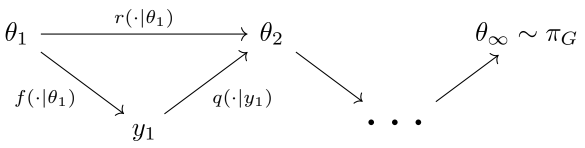


Figure 2: Schematic diagram of samples from the Gibbs chain (Definition 1) with auxiliary variables y_t. The distribution of θ_t converges to the Gibbs prior π_G.

are concentrated at some distribution $\pi_y \equiv \tilde{\pi}$, which is the implicit prior to the approximations. As the compatibility decreases, $(\pi_y)_{y\in\mathcal{Y}}$ gets more scattered (see Figure 3a). One possible measure of incompatibility is discussed in Appendix C. As a sanity check, observe that a perfect approximation $Q = P$ recovers the original prior $\pi = \pi_y$ for every y.

Ideally, the inductive bias of approximate inference could be explained by a single prior independent from the observation, like a prior in fully Bayesian inference. But as the above considerations show, this is not possible if the family $(\pi_y)_{y\in\mathcal{Y}}$ contains different members who offer conflicting explanations. Therefore, we search for a compromise that reasonably represents the different π_y. We do so by looking at the situation from the perspective of conditional distributions: a joint distribution on $\Theta \times \mathcal{Y}$ (Bayesian model) is specified indirectly through the conditionals F (likelihood) and Q (posterior approximations). We want to obtain the corresponding Θ-marginal (prior). A standard way to access the joint distribution via its conditionals is Gibbs sampling (Geman and Geman, 1984; Casella and George, 1992). Gibbs sampling starts with any initial point (θ_0, y_0) in the joint space and alternatingly updates θ given y and then y given θ. Under some assumptions, this vector converges to a sample from the joint distribution. Although Gibbs sampling assumes that the involved conditionals are compatible, it can be used the same way if they are incompatible. In this case it is referred to as *pseudo-Gibbs sampling*, a term coined by Heckerman et al. (2001). Pseudo-Gibbs sampling leads us to the following candidate prior:

Definition 1 (Gibbs prior). For two families of distributions $(f(\cdot|\theta))_{\theta\in\Theta}$ on $\mathcal{Y}$ and $(q(\cdot|y))_{y\in\mathcal{Y}}$ on Θ consider the discrete-time Markov chain on Θ whose transition function is given by

$$r(\theta'|\theta) = \mathbb{E}_{Y\sim f(\cdot|\theta)}\left[q(\theta'|Y)\right] . \qquad (2)$$

This chain is called the *Gibbs chain*. Any stationary distribution of this Markov chain is called a *Gibbs prior* and denoted by π_G.

The Gibbs chain is illustrated in Figure 2. A single step of the chain according to Eq. (2) can be simulated with an auxiliary variable y: first sample from the likelihood $y \sim f(\cdot|\theta)$ and then from the approximation $\theta' \sim q(\cdot|y)$. Under the caveat of incompatibility, we have the following intuition for the Gibbs prior:

> The Gibbs prior describes the a priori preference for solutions of the approximate inference method.

A simple reformulation of the stationarity condition for π_G offers two alternative representations

$$\pi_G(\theta) = \int_{\mathcal{Y}} g(y) q(\theta|y)\,\mathrm{d}y \qquad (3)$$

$$= \int_{\mathcal{Y}} \tilde{g}(y) f(y|\theta) \pi_y(\theta)\,\mathrm{d}y\,, \qquad (4)$$

where $g(y) = \int_\Theta \pi_G(\tilde{\theta}) f(y|\tilde{\theta})\,\mathrm{d}\tilde{\theta}$ and $\tilde{g}(y) = g(y)/\int_\Theta \pi_y(\tilde{\theta}) f(y|\tilde{\theta})\,\mathrm{d}\tilde{\theta}$ are weighting functions and Eq. (4) requires all π_y to be proper. Eq. (3) shows that the Gibbs prior is a mixture of the pointwise approximations. This suggests that consistent trends between approximations and posteriors are reflected in the Gibbs prior, for example underestimation of the norm as in Figure 1. Eq. (4) relates back to our original motivation of a compromise between $(\pi_y)_{y\in\mathcal{Y}}$ and shows that the Gibbs prior is a mixture of these distributions, reweighted by the likelihood.

Proposition 2 (Existence and uniqueness of Gibbs priors). *Consider two families of distributions $F = (f(\cdot|\theta))_{\theta\in\Theta}$ on $\mathcal{Y}$ and $Q = (q(\cdot|y))_{y\in\mathcal{Y}}$ on Θ. Let M be the corresponding Gibbs chain from Definition 1.*

(i) If F and Q are compatible with joint distribution $p(\theta, y)$, then the marginal $p(\theta)$ is a Gibbs prior. If M is additionally irreducible, then it is the only Gibbs prior.

(ii) If Θ and $\mathcal{Y}$ are finite, then there exists a Gibbs prior. If additionally F or Q are positive, then the Gibbs prior is unique.

Proof (sketch). The first statement of part (i) is a standard Gibbs sampling result; it can be proven by verifying the detailed balance equation for $p(\theta)$, which implies that M is a reversible Markov chain and $p(\theta)$ a stationary distribution. The statement about uniqueness is trivial, because Gibbs priors are defined as stationary distributions of M. A list of sufficient criteria in different settings is given in Arnold and Press (1989). Part (ii) concerns the existence of a (unique) stationary distribution. This condition is a standard result for finite Markov chains, for more general cases see Norris and Norris (1998). □

Proposition 2 admits additional interpretations in our Bayesian setting, where F is the likelihood and Q some approximation to the posterior. Part (i) states that if

Q is the exact posterior under some other prior $\tilde{\pi}$, then this prior is recovered by the Gibbs prior $\pi_G = \tilde{\pi}$. Part (ii) shows that Gibbs priors exist under much weaker assumptions than compatibility of F and Q. There are only few other results about the Gibbs chain and its Gibbs priors in the general incompatible case. Muré (2019) shows that Gibbs priors are an optimal compromise between incompatible conditionals among a restricted set of distributions. For discrete distributions, Kuo and Wang (2019) show that the transitions of the Gibbs chain can be interpreted as iterative projections with respect to the KL divergence.

3.3 Sampling from the Gibbs Prior

Algorithm 1: Simulating the Gibbs chain[3]

```
Data: Likelihood $f$, approximate inference
      method $q$, number of steps $T$
Result: Correlated samples $(\theta_1, \dots, \theta_T)$ from $\pi_G$
$\theta_0 \leftarrow$ Arbitrary initialization, e. g. sample from $\pi(\cdot)$
for $t \leftarrow 0$ to $T-1$ do
    $y_t \leftarrow$ Randomly sample from $f(\cdot|\theta_t)$
    $q(\cdot|y_t) \leftarrow$ Approximation to $p(\cdot|y_t)$
    $\theta_{t+1} \leftarrow$ Randomly sample from $q(\cdot|y_t)$
end
```

Algorithm 1 describes how to obtain a sequence of correlated samples from the Gibbs prior. Since it is defined as the stationary distribution of the Gibbs chain, this is achieved by simply simulating the chain as in Figure 2. This approach is very generally applicable because it only requires sampling from the approximate posteriors, but not evaluating their density. The complexity depends largely on the complexity of computing the approximations to the posterior, which has to be redone every step for a different observation. The number of steps needed to assure convergence depends on the mixing speed of the Markov chain. Under the exact posterior, the Gibbs chain mixes fast if there are few observations. Informally, the posterior $p(\theta|y) \propto \pi(\theta) f(y|\theta)$ relies heavily on the the prior π (the stationary distribution) which ensures that the chain converges to its stationary distribution quickly. When there are many observations, the posterior concentrates and the high correlation between parameters and observations leads to slow mixing. In that sense, Algorithm 1 is more practical under few observations; this case is arguably more interesting because posterior inference gets easier as the number of observations increases. To ensure that the resulting samples actually correspond to the Gibbs prior, we recommend to monitor convergence of the Gibbs chain (Roy, 2020).

3.4 How to Use the Gibbs Prior

There are two principled ways of using the Gibbs prior to diagnose an approximate inference method. The first way is to assess the quality of the approximation by quantifying the distance to the original prior π with some divergence measure $D(\pi_G, \pi)$, or testing the hypothesis $H_0 : \pi_G = \pi$. A large discrepancy between π_G and π indicates a bad approximation, because a perfect approximation would yield $\pi_G = \pi$. The second way is to understand the inductive bias that the approximation imposes by examining the shift in mass from π to π_G. A direct comparison might not be enlightening if the latent space Θ is large; instead, one could visualize their differences (Lloyd and Ghahramani, 2015) or compare the distribution of summary statistics $g \colon \Theta \to \mathbb{R}$.

Note that there are caveats to this interpretation of the Gibbs prior due to incompatibility of likelihood and approximations. Thinking of the Gibbs prior as the effectively used prior for approximate inference becomes less valid for stronger incompatibility, because the family of pointwise priors $(\pi_y)_{y\in\mathcal{Y}}$ requires a stronger compromise. This is also demonstrated in the next section.

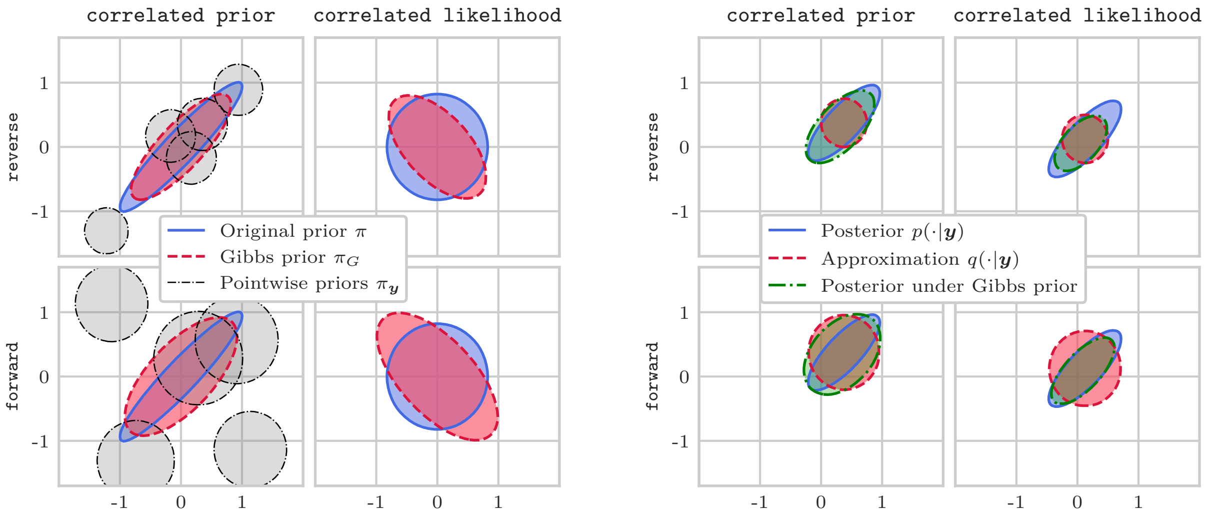
Summary We conclude this section by summarizing the three broad cases that can occur when comparing the Gibbs prior π_G with the original prior π:

1. $\pi_G \approx \pi$: the Gibbs prior is close to the original prior, which suggests that the approximations do not introduce additional bias. In particular, this is the case when the approximations are close to the true posterior. The reverse implication is not necessarily true (Appendix B.1).

2. $\pi_G \neq \pi$: the Gibbs prior differs from the original prior, which implies that the approximations differ from the true posterior. This means that the approximations introduce additional bias, which can be assessed by interpreting the Gibbs prior as the effectively used prior. The validity of this interpretation depends on the compatibility between likelihood and approximations.

3. The Gibbs chain in Algorithm 1 does not converge. This can have multiple reasons: the approximations are good but the prior π is improper, the approximations are bad, or the chain was not run long enough. We recommend to use the diagnostic conservatively and dismiss it in these cases to avoid falsely rejecting a good approximation. To exclude the last case of running the Gibbs chain not long enough, the convergence of the chain should be monitored.

[3] Code available at https://github.com/tml-tuebingen/gibbs-prior-diagnostic



(a) **Prior distributions.** Original prior, Gibbs prior, and pointwise priors for different $\boldsymbol{y}$ (same in both plots).

(b) **Posterior distributions.** Posterior, its approximation, and posterior under the Gibbs prior at fixed $\boldsymbol{y}$.

Figure 3: Distributions of interest for the variational inference settings described in Section 4.1 with $d = 2$ and $n = 1$. The setting `correlated prior` uses $\Sigma_0 = I$ and a Σ which is strongly correlated along $(1 \quad 1)^\top$. For `correlated likelihood` Σ_0 and Σ are interchanged. Colored areas show superlevel density sets with mass 0.3.

4 ILLUSTRATIVE TOY EXAMPLE

We now give a simple example to demonstrate the concepts from the previous section.

4.1 Gaussian Toy Model

Consider the problem of estimating the mean $\theta \in \mathbb{R}^d$ of a d-dimensional Gaussian distribution with known covariance matrix based on n independent samples $y_1, \dots, y_n \in \mathbb{R}^d$. Placing a Gaussian prior on θ yields the Bayesian model

$$\begin{aligned} \theta \quad &\sim \quad \mathcal{N}(\mu_0, \Sigma_0)\,, \\ y_i|\theta \quad &\overset{\text{indep.}}{\sim} \quad \mathcal{N}(\theta, \Sigma)\,, \quad i = 1, \dots, n\,, \end{aligned} \tag{5}$$

where $\mu_0 \in \mathbb{R}^d$ and $\Sigma_0, \Sigma \in \mathbb{R}^{d \times d}$ are positive definite. The observations are collected in a matrix $\boldsymbol{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^{n \times d}$. We consider four different settings for variational inference in this model, which are determined by the following two choices:

Correlated posterior We choose the prior and likelihood covariance matrices such that the posterior distribution has correlated components. This can be achieved by either a correlated prior and isotropic likelihood (referred to as `correlated prior`) or an isotropic prior and a correlated likelihood (referred to as `correlated likelihood`).

Variational approximation We consider the mean field variational approximation (Bishop, 2006). This method approximates the posterior with the variational family $\mathcal{Q}_{\mathrm{MF}}$, which consists of all distributions on $\mathbb{R}^d$ with independent components. For the objective we consider the commonly used reverse KL divergence

$$q(\cdot|\boldsymbol{y}) := \operatorname*{arg\,min}_{q \in \mathcal{Q}_{\mathrm{MF}}} \mathrm{KL}(q \,\|\, p(\cdot|\boldsymbol{y})) \tag{6}$$

(referred to as `reverse`) or the forward KL divergence

$$q(\cdot|\boldsymbol{y}) := \operatorname*{arg\,min}_{q \in \mathcal{Q}_{\mathrm{MF}}} \mathrm{KL}(p(\cdot|\boldsymbol{y}) \,\|\, q) \tag{7}$$

(referred to as `forward`).

These settings are simple enough so that all distributions of interest are Gaussians and can be computed in closed form. This includes the posteriors $p(\cdot|\boldsymbol{y})$, the approximations $q(\cdot|\boldsymbol{y})$, the pointwise priors $\pi_{\boldsymbol{y}}$, and the Gibbs prior π_G. For details see Appendix D, which also provides numerical justifications for the following arguments about biases.

4.2 Bias Discovery Using the Gibbs Prior

Both approximations `reverse` and `forward` have two known biases, compactness and loss of correlation (Turner and Sahani, 2011). These biases can now also be discovered with the Gibbs prior. Figure 3a shows the priors and Gibbs priors and Figure 3b shows the corresponding posteriors and approximations.

Bias: compactness One known bias of mean field variational inference is the compactness of the approximations as measured by the entropy (Turner and Sahani, 2011): comparing the approximations to the true posterior in Figure 3b shows that they are too compact for `reverse` and not compact enough for `forward`. The same behavior can be observed on the prior level: the Gibbs prior is more compact than the prior for `reverse` and less compact for `forward`.

Bias: loss of correlation The variational approximations cannot capture any correlation between the coordinates by definition of the variational family $\mathcal{Q}_{\text{MF}}$. This bias is easily understood on the posterior level, but it is less obvious what this means in terms of an a priori preference for solutions. In fact, this corresponding preference depends on the source of the posterior correlation and cannot be explained by the posterior alone. For `correlated prior`, the posterior correlation is caused by the prior correlation. Uncorrelated approximations therefore correspond to an uncorrelated prior. The Gibbs priors confirm this intuition by being less correlated than the prior. For `correlated likelihood`, the posterior correlation is caused by the likelihood correlation. Here, the Gibbs priors show that the approximations correspond to a prior whose correlation is orthogonal to the likelihood correlation. Intuitively, the orthogonal correlations of prior and likelihood "cancel out" to produce uncorrelated posteriors.

4.3 Is the Gibbs Prior a Prior?

The approximations are exact posteriors under the Gibbs prior if and only if the approximations are compatible to the likelihood. Equivalently, this is the case when the family of pointwise priors $(\pi_{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}$ concentrates at a single distribution. Figure 3a shows $\pi_{\boldsymbol{y}}$ for various $\boldsymbol{y}$. For `correlated prior` they differ strongly and for `correlated likelihood` they are improper and therefore not shown. In both settings, this implies that the conditionals are incompatible as is typically the case. This is confirmed by Figure 3b, which shows that the posteriors under the Gibbs prior do not exactly coincide with the approximations. Despite these incompatibilities, this example shows that the Gibbs prior can discover inductive biases of the approximate methods. The Gibbs prior should therefore be thought of as a summary statistic for the inductive bias (see Appendix B for more details).

5 EXPERIMENTS

We experiment with the Gibbs prior as a diagnostic tool for various approximations in two Bayesian models. For more details and convergence monitoring of the Gibbs chains see Appendix E.

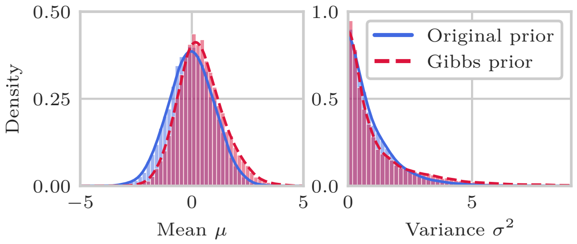


Figure 4: Marginal distributions of prior and Gibbs prior for the sum of log-normals model. A comparison shows that the approximation overestimates μ and puts more mass on extreme values for σ^2.

Baseline We compare our findings to the diagnostic Talts et al. (2018). This diagnostic is based on the stationarity equation of the prior π under the Gibbs chain, but only considers 1-step transitions with some test statistics $f\colon \Theta \to \mathbb{R}$. Under random samples $\tilde{\theta} \sim \pi$, $\tilde{y} \sim f(\cdot|\tilde{\theta})$, and $\theta_1, \dots \theta_L \sim q(\cdot|\tilde{y})$, the rank of $f(\tilde{\theta})$ in $\{f(\theta_1), \dots, f(\theta_L)\}$ is computed. This is repeated over multiple draws of $(\tilde{\theta}, \tilde{y})$, which gives a histogram of the ranks. Since the histogram is uniform under the exact posterior, any deviations from uniformity indicate an approximation mismatch. We allocate this method the same computational resources in terms of posterior draws as our Gibbs chain.

5.1 Sum of log-normals

Setup Our first model describes the sum of $L = 10$ independent samples from a log-normal distribution and is given by

$$\mu \sim \mathcal{N}(0,1), \quad \sigma^2 \sim \text{Gamma}(1,1)\,,$$

$$x_l|\theta = (\mu,\sigma^2) \overset{\text{indep.}}{\sim} \text{LogNormal}(\mu,\sigma^2), \quad y = \sum_{l=1}^{L} x_l\,.$$

Since the corresponding likelihood is infeasible we approximate the posterior in a two-step procedure: first, we replace the likelihood by its Fenton-Wilkinson approximation (Fenton, 1960), which is another log-normal distribution with matching first two moments, and then we use a Laplace approximation to the posterior of this new model.

Bias discovery To discover the bias of this approximation we simulate the Gibbs prior based on 10,000 iterations of Algorithm 1 and show it alongside the original prior in Figure 4. The first observation is that the Gibbs prior does not coincide with the original

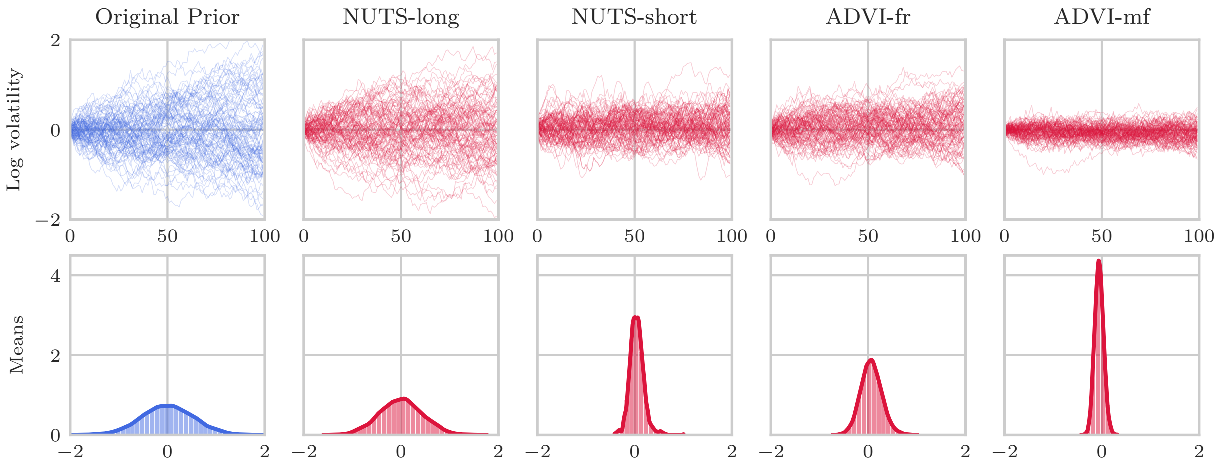


Figure 5: **Top row:** Samples of $\theta \in \mathbb{R}^{100}$ from original prior (blue) and Gibbs priors (red) under various approximations. **Bottom row:** Histograms of the summary statistic $\theta \mapsto 1/100 \sum_{i=1}^{100} \theta_i$, which is the mean value of a time series. Methods that are closer to the prior introduce less bias.

prior, which implies that the approximation is not exact. Furthermore, the deviation between the two distributions is systematic. For the mean μ, the Gibbs prior has a similar shape as the original prior, but is shifted to the right. This implies that the approximations systematically overestimate μ. For the variance σ^2, the Gibbs prior puts more mass on extreme values, which means that there is no systematic under- or overestimation. Compare these findings to Rodrigues et al. (2018) who consider a fixed approximation to an observation y drawn from $\theta = (0, 1)$. They confirm that μ is overestimated, but also find that σ^2 is underestimated. This does not contradict our findings, because they analyze the approximation to a *fixed* observation, while we analyze the approximations *across* observations. The other baseline Talts et al. (2018) is shown in the first two histograms of Figure 6 for the coordinates of $\theta = (\mu, \sigma^2)$ as summary statistics, that is, $f_i(\theta) = \theta_i$. The histogram for μ exceeds the confidence region at the smallest rank, which also suggests overestimation. For σ^2, the deviation from uniformity is not strong enough to deduce a systematic approximation mismatch.

5.2 Stochastic Volatility

Setup Stochastic volatility models are used in mathematical finance for time series to describe the latent variation of trading price (called the returns). We consider a model similar to Hoffman and Gelman (2014):

$$\theta_i|\theta_{i-1} \sim \mathcal{N}(\theta_i, \sigma^2), \quad i = 1, \ldots, T\,,$$
$$y_i \overset{\text{indep.}}{\sim} \text{StudentT}(\nu, 0, \exp \theta_i), \quad i = 1, \ldots, T\,,$$

where $\theta_0 = 0, \sigma = .09, \nu = 12$, and $T = 100$. The latent parameters $\theta = (\theta_1, \ldots, \theta_T)$ follow a Gaussian random walk and describe the log volatility of the returns $y = (y_1, \ldots, y_T)$, which are independent given θ. As posterior inference methods, we investigate the Hamiltonian Monte Carlo method NUTS (Hoffman and Gelman, 2014) with different number of steps (10 for NUTS-short and 40 for NUTS-long) and the variational inference method ADVI (Kucukelbir et al., 2017), which comes in a less powerful mean-field (ADVI-mf) and more powerful full-rank (ADVI-fr) variant.

Bias discovery For each approximation method, we can again use the corresponding Gibbs prior in two ways: we test *whether* it deviates from the original prior to assess exactness of the approximation, and if it does, we inspect *how* it deviates to assess the systematic bias. Figure 5 shows samples from original prior and Gibbs priors under the approximations alongside the distribution of means for each time series as a summary statistic. Each Gibbs chain was simulated for 10,000 steps, which took 13 hours for ADVI-fr and roughly 5 hours for the other methods on a GPU. We observe that the Gibbs prior for the long MCMC chain is almost identical to the prior, which confirms that this method is accurate; the Gibbs prior for the corresponding short chain is further away from the prior and closer to the initialization of the chain because it has not fully converged. The method ADVI-mf shows a strong deviation from the prior by concentrating on less extreme values of the latent variables. This indicates that the approximation is overly compact compared to the true posterior. The same

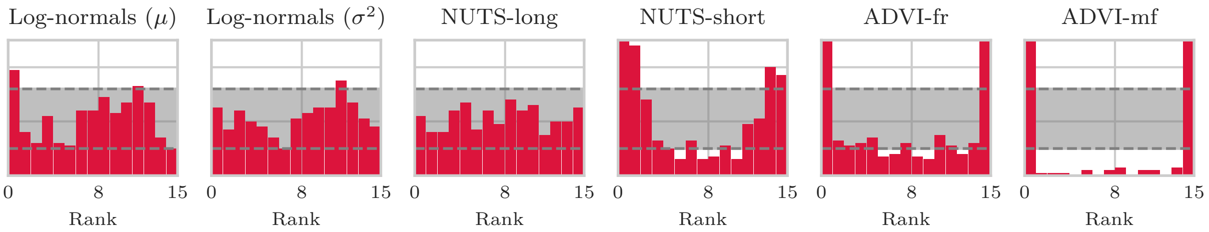


Figure 6: Histograms of rank statistics for the baseline Talts et al. (2018). First two histograms are for Section 5.1 with coordinates as summary statistics, other histograms are for Section 5.2 with the mean. Gray band shows a 99% confidence interval under the exact posterior. Deviations from uniformity indicate approximation mismatch.

phenomenon was already observed for mean field variational inference in Section 4. It can also be observed for ADVI-fr, but is less pronounced because the method is strictly more powerful. The baseline Talts et al. (2018) is shown in the last four histograms of Figure 6 for the same summary statistic as in Figure 5, the mean value of θ. For NUTS-long, the histogram stays within the confidence region, which confirms that this method is accurate. The other three methods show a ∪-shape, which is most pronounced for ADVI-mf. This indicates that the methods are overly compact and is in line with our findings. While this baseline can in principle also discover systematic approximation mismatches in terms of over-/underestimation and compactness, the Gibbs prior provides a more complete and nuanced picture.

6 CONCLUSION AND FUTURE WORK

Conclusion We describe a novel diagnostic approach for assessing the inductive bias of approximate Bayesian inference methods. A reformulation of this problem leads to a natural solution, which we call the Gibbs prior. We demonstrate how it can be used to discover the inductive bias in various examples.

Future work The Gibbs prior compromises between many pointwise priors. The precise nature of this compromise is intricate, offering several avenues for future analysis. While we introduced the Gibbs prior in the context of approximate Bayesian methods, it can be defined for any generative method returning a distribution over latent variables given an observation. Another direction is using the pointwise priors as observation-dependent diagnostics. They do not suffer from incompatibility, but can be more challenging to sample from if the approximation density is unknown.

Broader impact Recently, there has been a surge of interest in interpretable and explainable machine learning algorithms. One principled way of explaining an algorithm is to inspect its inductive bias, which describes the preferred solutions independent of the data. While the inductive bias is specified only implicitly for most algorithms, it is made explicit in Bayesian inference through prior and likelihood. Unfortunately, this transparency is concealed for approximate Bayesian inference, because approximations introduce additional hidden bias. We present a method to uncover this inductive bias again, which opens up a new paradigm for the practical evaluation of approximate inference.

Acknowledgements

This work has been supported by the German Research Foundation through the Cluster of Excellence “Machine Learning – New Perspectives for Science” (EXC 2064/1 number 390727645), the BMBF Tübingen AI Center (FKZ: 01IS18039A), and the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

References

B. C. Arnold and S. J. Press. Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405):152–156, 1989.

B. C. Arnold, E. Castillo, and J. M. Sarabia. Conditionally specified distributions: An introduction (with comments and a rejoinder by the authors). *Statistical Science*, 16(3):249 – 274, 2001.

B. C. Arnold, E. Castillo, and J. M. Sarabia. Exact and near compatibility of discrete conditional distributions. *Computational statistics & data analysis*, 40(2):231–252, 2002.

R. Bardenet, A. Doucet, and C. Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43, 2017.

M. A. Beaumont. Approximate Bayesian computation. *Annual Review of Statistics and Its Application*, 6 (1):379–403, 2019.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518): 859–877, 2017.

G. Casella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

S.-H. Chen and E. Ip. Behavior of the Gibbs sampler when conditional distributions are potentially incompatible. *Journal of Statistical Computation and Simulation*, 85:1–10, 2014.

S. R. Cook, A. Gelman, and D. B. Rubin. Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692, 2006.

M. Cusumano-Towner and V. K. Mansinghka. Aide: An algorithm for measuring the accuracy of probabilistic inference algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

E. Daxberger, A. Kristiadi, A. Immer, R. Eschenhagen, M. Bauer, and P. Hennig. Laplace redux-effortless Bayesian deep learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

P. J. Diggle and R. J. Gratton. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):193–227, 1984.

J. Domke. An easy to interpret diagnostic for approximate inference: Symmetric divergence over simulations. *arXiv preprint arXiv:2103.01030*, 2021.

L. Fenton. The sum of log-normal probability distributions in scatter transmission systems. *IRE Transactions on Communications Systems*, 8(1):57–67, 1960.

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

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, PAMI-6(6):721–741, 1984.

J. Geweke. Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association*, 99(467):799–804, 2004.

I. Ghosh and N. Balakrishnan. Study of incompatibility or near compatibility of bivariate discrete conditional probability distributions through divergence measures. *Journal of Statistical Computation and Simulation*, 85(1):117–130, 2015.

J. Gorham and L. Mackey. Measuring sample quality with Stein's method. In *Neural Information Processing Systems (NeurIPS)*, 2015.

J. Gorham and L. Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning (ICML)*, 2017.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1: 49–75, 2001.

G. E. Hinton and D. van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, 1993.

M. D. Hoffman and A. Gelman. The no-u-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(4):1303–1347, 2013.

J. Huggins, M. Kasprzak, T. Campbell, and T. Broderick. Validated variational inference via practical posterior error bounds. In *Artificial Intelligence and Statistics (AISTATS)*, 2020.

R. A. Hughes, I. R. White, S. R. Seaman, J. R. Carpenter, K. Tilling, and J. A. Sterne. Joint modelling rationale for chained equations. *BMC medical research methodology*, 14(1):1–10, 2014.

M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

C. Joshi and F. Ruggeri. Duality between approximate Bayesian methods and prior robustness. *arXiv preprint arXiv:2004.00796*, 2020.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017.

K.-L. Kuo and Y. J. Wang. A simple algorithm for checking compatibility among discrete conditional distributions. *Computational Statistics & Data Analysis*, 55(8):2457–2462, 2011.

K.-L. Kuo and Y. J. Wang. Pseudo-Gibbs sampler for discrete conditional distributions. *Annals of the Institute of Statistical Mathematics*, 71(1):93–105, 2019.

K.-L. Kuo, C.-C. Song, and T. J. Jiang. Exactly and almost compatible joint distributions for high-dimensional discrete conditional distributions. *Journal of Multivariate Analysis*, 157:115–123, 2017.

J. R. Lloyd and Z. Ghahramani. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

D. J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4 (3):448–472, 1992.

T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence (UAI)*, 2001.

J. Muré. Optimal compromise between incompatible conditional probability distributions, with application to Objective Bayesian Kriging. *ESAIM: P&S*, 23:271–309, 2019.

J. R. Norris and J. R. Norris. *Markov chains*. Cambridge University Press, 1998.

D. Phan, N. Pradhan, and M. Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

D. Prangle, M. G. B. Blum, G. Popovic, and S. A. Sisson. Diagnostic tools for approximate Bayesian computation using the coverage property. *Australian & New Zealand Journal of Statistics*, 56(4):309–329, 2014.

R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Artificial Intelligence and Statistics (AISTATS)*, 2014.

G. Rodrigues, D. Prangle, and S. Sisson. Recalibration: A post-processing method for approximate Bayesian computation. *Computational Statistics & Data Analysis*, 126:53–66, 2018.

V. Roy. Convergence diagnostics for Markov chain Monte Carlo. *Annual Review of Statistics and Its Application*, 7(1):387–412, 2020.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

H. Rue, A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren. Bayesian computing with inla: A review. *Annual Review of Statistics and Its Application*, 4(1):395–421, 2017.

S. A. Sisson, Y. Fan, and M. Beaumont. *Handbook of approximate Bayesian computation*. CRC Press, 2018.

D. J. Spiegelhalter and S. L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20(5):579–605, 1990.

S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

R. E. Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models. In *Bayesian Time series models*, chapter 5, pages 109–130. Cambridge University Press, 2011.

S. Van Buuren, J. P. Brand, C. G. Groothuis-Oudshoorn, and D. B. Rubin. Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 76(12):1049–1064, 2006.

H. Xing, G. Nicholls, and J. (Kate) Lee. Distortion estimates for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence (UAI)*, 2020.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Yes, but did it work?: Evaluating variational inference. In *International Conference on Machine Learning (ICML)*, 2018.

X. Yu, D. J. Nott, M.-N. Tran, and N. Klein. Assessment and adjustment of approximate inference algorithms using the law of total variance. *Journal of Computational and Graphical Statistics*, 0(0): 1–14, 2021.