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

Luca Rendsburg¹

¹University of Tübingen

Agustinus Kristiadi¹

²Max Planck Institute for Intelligent Systems, Tübingen

Philipp Hennig^{1,2}

Ulrike von Luxburg^{1,2}

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).

¹While 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.

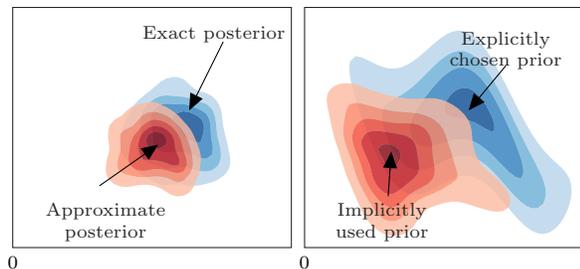
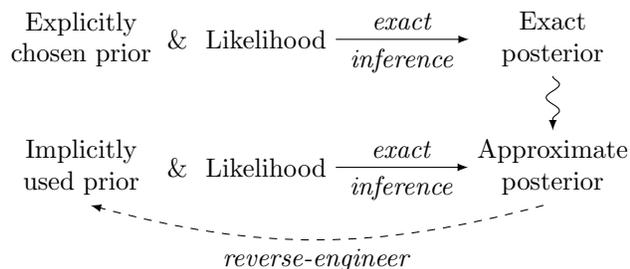


Figure 1: **Left:** a posterior approximation is biased towards solutions of small norm. **Right:** the approximation corresponds to the exact posterior under another implicitly defined prior, which is itself biased towards solutions of small norm.

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:



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' theorem² $\pi_y(\theta) \propto_\theta q(\theta|y)/f(y|\theta)$. Unfortunately, π_y generally depends on the observation y . This means that the approximations to different observations can

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

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 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 implementation 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 alternately 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

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

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

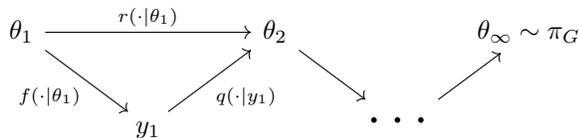


Figure 2: Schematic diagram of samples from the Gibbs chain (Definition 1) with auxiliary variables y_t . The distribution of θ_t converges to the Gibbs prior π_G .

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

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

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

$$\pi_G(\theta) = \int_{\mathcal{Y}} g(y)q(\theta|y) dy \quad (3)$$

$$= \int_{\mathcal{Y}} \tilde{g}(y)f(y|\theta)\pi_y(\theta) dy, \quad (4)$$

where $g(y) = \int_{\Theta} \pi_G(\tilde{\theta})f(y|\tilde{\theta})d\tilde{\theta}$ and $\tilde{g}(y) = g(y)/\int_{\Theta} \pi_y(\tilde{\theta})f(y|\tilde{\theta})d\tilde{\theta}$ 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). \square

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 chain³

Data: Likelihood f , approximate inference method q , number of steps T

Result: Correlated samples $(\theta_1, \dots, \theta_T)$ from π_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).

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

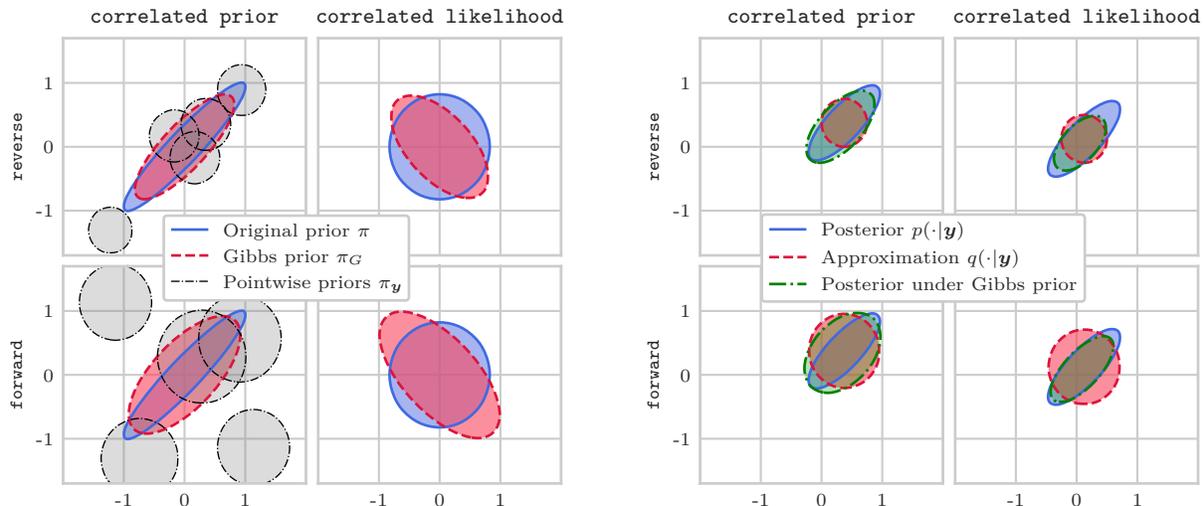
3.4 How to Use the Gibbs Prior

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 $g : \Theta \rightarrow \mathbb{R}$.

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.



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

(b) **Posterior distributions.** Posterior, its approximation, and posterior under the Gibbs prior at fixed \mathbf{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 $(1 \ -1)^\top$. 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

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

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 $\mathbf{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^{n \times d}$. 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}_{MF} , which consists of all distributions on \mathbb{R}^d with independent components. For the objective we consider the commonly used reverse KL divergence

$$q(\cdot | \mathbf{y}) := \arg \min_{q \in \mathcal{Q}_{\text{MF}}} \text{KL}(q \| p(\cdot | \mathbf{y})) \quad (6)$$

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

$$q(\cdot | \mathbf{y}) := \arg \min_{q \in \mathcal{Q}_{\text{MF}}} \text{KL}(p(\cdot | \mathbf{y}) \| q) \quad (7)$$

(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 $p(\cdot | \mathbf{y})$, the approximations $q(\cdot | \mathbf{y})$, the pointwise priors $\pi_{\mathbf{y}}$, 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}_{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 $(\pi_{\mathbf{y}})_{\mathbf{y} \in \mathcal{Y}}$ concentrates at a single distribution. Figure 3a shows $\pi_{\mathbf{y}}$ for various \mathbf{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 mod-

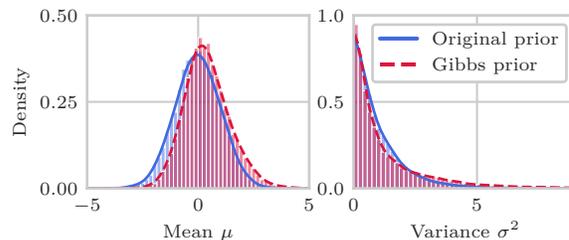


Figure 4: Marginal distributions of prior and Gibbs prior for the sum of log-normals model. A comparison shows that the approximation overestimates μ and puts more mass on extreme values for σ^2 .

els. For more details and convergence monitoring of the Gibbs chains see Appendix E.

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 $f: \Theta \rightarrow \mathbb{R}$. 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

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

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

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

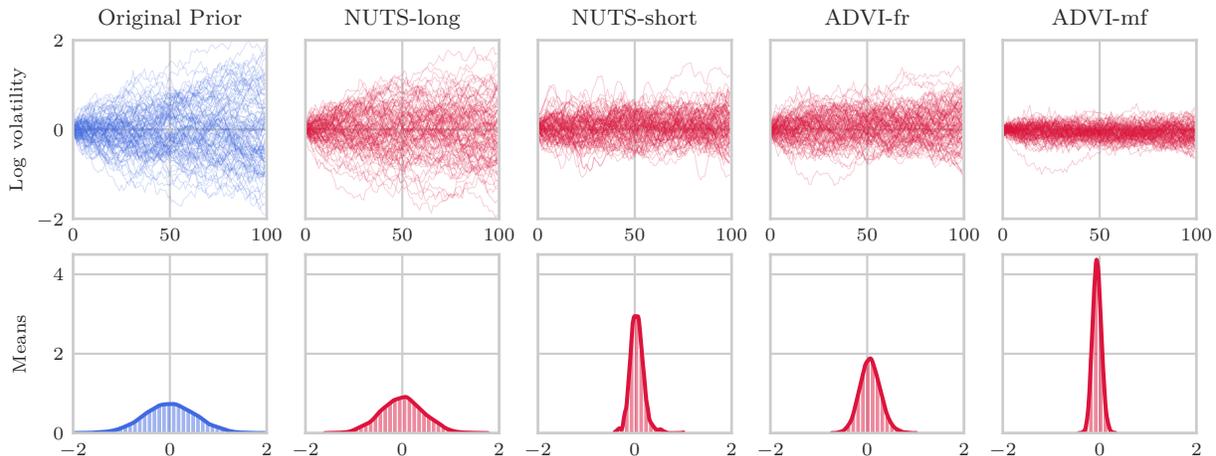


Figure 5: **Top row:** Samples of $\theta \in \mathbb{R}^{100}$ from original prior (blue) and Gibbs priors (red) under various approximations. **Bottom row:** Histograms of the summary statistic $\theta \mapsto 1/100 \sum_{i=1}^{100} \theta_i$, which is the mean value of a time series. Methods that 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 \(2014\)](#):

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

where $\theta_0 = 0, \sigma = .09, \nu = 12$, and $T = 100$. The latent parameters $\theta = (\theta_1, \dots, \theta_T)$ follow a Gaussian random walk and describe the log volatility of the returns $y = (y_1, \dots, y_T)$, 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

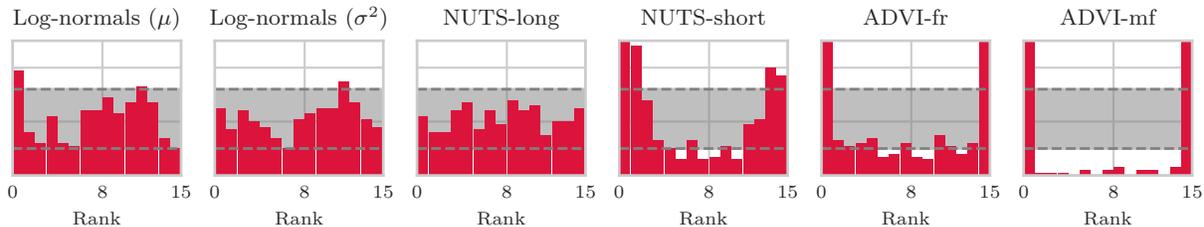


Figure 6: Histograms of rank statistics for the baseline Talts et al. (2018). First two histograms are for Section 5.1 with coordinates as summary statistics, other histograms are for Section 5.2 with the mean. Gray band shows a 99% confidence interval under the exact posterior. Deviations from uniformity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

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

Luca Rendsburg¹Agustinus Kristiadi¹Philipp Hennig^{1,2}Ulrike von Luxburg^{1,2}¹University of Tübingen²Max Planck Institute for Intelligent Systems, Tübingen

A INCOMPATIBLE CONDITIONALS AND PSEUDO-GIBBS SAMPLING

When treating approximations as exact posteriors we inevitably face the problem of compatibility, which we shortly introduce in this paragraph. A bivariate model can be specified explicitly through its joint distribution $p(\theta, y)$, for example as in Bayesian models with a marginal $p(\theta)$ (prior) and a conditional distribution $p(y|\theta)$ (likelihood). Alternatively, the model can be specified implicitly through its conditional distributions $p(\theta|y)$ and $p(y|\theta)$. Joint modeling simplifies theoretical analysis because closed-form expressions are available, whereas conditional modeling is less accessible, but more flexible and interpretable. However, an arbitrary pair of conditional distributions can be *incompatible*, meaning that there exists no joint distribution which produces these conditionals, and if it exists it does not have to be unique (Arnold and Press, 1989). Arnold et al. (2001) argues that “in general, reasonable-seeming conditional models will not be compatible with any single joint distribution”. For example, consider the following Bayesian model with real-valued latent variables θ and observations y : for an improper prior $\pi(\theta) = 1$ and a Gaussian likelihood $f(y|\theta) = \mathcal{N}(y|\theta, 1)$ Bayes’ theorem yields the posterior $p(\theta|y) = \mathcal{N}(\theta|y, 1)$. Even though both conditional distributions—the likelihood and the posterior—are proper distributions, there exists no proper joint distribution because the corresponding marginal π is improper. Hence, the conditionals f and p are incompatible. But incompatibility is no all-or-nothing property: even if there exists no joint distribution, one might still look for the joint distribution that is “most” compatible with the given conditionals, which leads to notions such as near-compatibility and ε -compatibility (Arnold et al., 2002). There exist algorithms for assessing the compatibility of conditional distributions (Kuo and Wang, 2011; Kuo et al., 2017) based on fractions of conditional densities, but most of this theory is restricted to discrete settings.

Gibbs sampling is one of the most natural ways of accessing the joint distribution of a conditionally specified model. It works by using the conditional distributions to define a time-reversible Markov chain whose stationary distribution is the joint distribution (Hastings, 1970; Geman and Geman, 1984). Gibbs sampling is well-understood and theoretically sound if the conditionals are compatible, but what happens if they are incompatible? Despite the fact that no joint distribution exists, the Markov chain defined by the conditionals can still converge to a unique stationary distribution, which represents a compromise between the incompatible conditionals (Muré, 2019). In this case, Gibbs sampling is called pseudo-Gibbs sampling. Pseudo-Gibbs samplers are widely used, for example in dependency networks (Heckerman et al., 2001) and missing data imputation (Van Buuren et al., 2006; Hughes et al., 2014). Characterizing the stationary distribution of a pseudo-Gibbs sampler is ongoing research (Chen and Ip, 2014; Kuo and Wang, 2019; Muré, 2019).

B THE GIBBS PRIOR IS A SUMMARY STATISTIC

B.1 Different Approximations can have the same Gibbs Prior

This section shows that there are fewer Gibbs priors than conditional distributions because different conditionals can define the same Gibbs chain. For the sake of simplicity, we consider the finite setting with latent space $\Theta = [n] = \{1, \dots, n\}$ and observation space $\mathcal{Y} = [m]$ with $n, m \in \mathbb{N}$. Here the likelihood is given by the stochastic matrix $F \in \mathbb{R}^{n \times m}$ and the approximation by another stochastic matrix $Q \in \mathbb{R}^{m \times n}$, that is, F and Q have non-negative entries and their rows sum to 1. The Gibbs chain from Definition 1 is defined via the transition matrix $P = FQ \in \mathbb{R}^{n \times n}$, which is again a stochastic matrix, and the Gibbs prior is a probability vector $\pi_G \in \mathbb{R}^n$.

The next proposition shows that different approximations $Q \neq \tilde{Q}$ can define the same Gibbs chain. In particular, they define the same Gibbs prior.

Proposition 3. *For $n \geq 2$ let $F \in \mathbb{R}^{n \times m}$, $Q \in \mathbb{R}^{m \times n}$ be stochastic matrices with entries in $(0, 1)$ and $\ker F \neq \{0\}$. Then there exists a stochastic matrix $\tilde{Q} \in \mathbb{R}^{m \times n}$ with $Q \neq \tilde{Q}$ that satisfies*

$$FQ = F\tilde{Q}. \quad (8)$$

In particular, both Markov chains have the same stationary distribution.

Proof. The main idea is to define a suitable perturbation W such that $\tilde{Q} = Q + W$ is a stochastic matrix that satisfies Eq. (8).

Let $0 \neq x_0 \in \ker F$ and $0 \neq w \in \mathbf{1}^\perp = \{x \in \mathbb{R}^n \mid x^\top \mathbf{1} = 0\}$, where $\mathbf{1} \in \mathbb{R}^n$ denotes the vector whose entries are all 1. The vectors x_0 and w can be chosen non-zero by the assumptions $\ker F \neq \{0\}$ and $n \geq 2$. With these vectors, we define the perturbation matrix $W := x_0 w^\top \in \mathbb{R}^{m \times n}$ and $\tilde{Q} := Q + W \in \mathbb{R}^{m \times n}$. First observe that $Q \neq \tilde{Q}$, because $x_0, w \neq 0$ implies $W \neq 0$. Using $x_0 \in \ker F$, we verify Eq. (8) by computing

$$F\tilde{Q} = F(Q + W) = FQ + \underbrace{Fx_0}_{=0} w^\top = FQ.$$

It remains to show that \tilde{Q} is a stochastic matrix. We may assume that w was chosen such that the first condition $0 \leq \tilde{Q} = Q + W = Q + x_0 w^\top$ holds; otherwise, w can be scaled by an arbitrarily small constant such that this inequality is satisfied, which is always possible because $Q > 0$ by assumption. The other condition is that the rows of \tilde{Q} sum to 1, which we verify with $Q\mathbf{1} = \mathbf{1}$ (because Q is a stochastic matrix) and $w \in \mathbf{1}^\perp$ by computing

$$\tilde{Q}\mathbf{1} = \underbrace{Q\mathbf{1}}_{=\mathbf{1}} + x_0 \underbrace{w^\top \mathbf{1}}_{=0} = \mathbf{1}.$$

For the second statement we only need to verify that the stationary distribution of the Markov chain defined with the transition matrix $P = FQ$ indeed uniquely exists. This is the case because the assumptions $F, Q > 0$ imply $P > 0$, hence the corresponding Markov chain is positive recurrent with finite state space. This implies the existence of a unique stationary distribution. \square

Example 4. An example for Proposition 3 with $n = 2$ and $m = 3$ is given by the matrices

$$F = \begin{pmatrix} .1 & .4 & .5 \\ .3 & .2 & .5 \end{pmatrix}, \quad Q = \begin{pmatrix} .2 & .8 \\ .4 & .6 \\ .5 & .5 \end{pmatrix}, \quad \tilde{Q} = \begin{pmatrix} .1 & .9 \\ .3 & .7 \\ .6 & .4 \end{pmatrix},$$

which satisfy $Q \neq \tilde{Q}$ and

$$FQ = \begin{pmatrix} .43 & .57 \\ .39 & .61 \end{pmatrix} = F\tilde{Q}.$$

B.2 A Weaker Notion of Compatibility between Conditional Distributions is Sufficient

In this section, we argue that the notion of compatibility between conditional distributions is actually stricter than necessary for assessing whether the Gibbs prior provides a useful explanation. We do so by introducing a weaker notion of compatibility under which the Gibbs prior retains a strong interpretation. First, we recap the setting as presented in the main paper. For a Bayesian model with likelihood F we are given approximations Q to the true posterior, and our goal is to assess their inductive bias in terms of an a priori preference for solutions. We propose to consider another fully Bayesian model \mathcal{M}_G , specified with the same likelihood F and the Gibbs prior π_G . The Gibbs prior π_G can then be used to reason about the inductive bias of Q if the conditional distributions F and Q are compatible, because then the posteriors under \mathcal{M}_G coincide with Q .

However, even when they are different, the Bayesian model \mathcal{M}_G can accurately describe inference based on Q . This is achieved by considering the whole pipeline of inference instead of inference based on a fixed observation: starting with an unknown true latent parameter θ , we observe some data through the likelihood $y \sim f(\cdot|\theta)$, based on which we use the approximations to estimate the latent parameter $\theta' \sim q(\cdot|y)$. This process is summarized in the probabilities of estimating θ' if the true parameter is θ , which are precisely the transition probabilities of the Gibbs chain. Defining the same Gibbs chain as the approximations is therefore sufficient for the Bayesian model \mathcal{M}_G to qualify as an interpretable reformulation. This leads to the following weaker notion of compatibility between conditional distributions:

Definition 5 (Weak compatibility of conditional distributions). For two families of conditional distributions $F = (f(\cdot|\theta))_{\theta \in \Theta}$ on \mathcal{Y} and $Q = (q(\cdot|y))_{y \in \mathcal{Y}}$ on Θ , let π_G denote the corresponding Gibbs prior from Definition 1. Let $P_G = (p_G(\cdot|y))_{y \in \mathcal{Y}}$ denote the posteriors under to the Bayesian model specified by π_G and F . Then F and Q are called *weakly compatible*, if the Gibbs chain of F and Q coincides with the Gibbs chain of F and P_G , that is,

$$\mathbb{E}_{Y \sim f(\cdot|\theta)} [q(\theta'|Y)] = \mathbb{E}_{Y \sim f(\cdot|\theta)} [p_G(\theta'|Y)] \quad \forall \theta \in \Theta.$$

As the naming suggests, weak compatibility of two conditional distributions is strictly weaker than compatibility: compatibility trivially implies weak compatibility, whereas Proposition 3 shows that the converse is not true. Since the Gibbs prior can be used to reason about conditionals that are only weakly compatible, this means that it is useful in more situations than what compatibility suggests. In particular, there exist different conditional distributions Q which justifiably get assigned the same Gibbs prior, because they yield the same Gibbs chain.

C MEASURING THE DEGREE OF COMPATIBILITY

Most existing literature focuses on the question whether families of conditional distributions are exactly compatible (Kuo and Wang, 2011; Kuo et al., 2017). However, in the context of this paper, the more relevant question is how incompatible they are, requiring a practical way to measure the degree of compatibility. This leads to notions such as near-compatibility and ε -compatibility (Arnold et al., 2002) and is based on computing some distance between joint distributions, which involve the conditional distributions (Ghosh and Balakrishnan, 2015).

In this section, we present a practical way of measuring the degree of compatibility between likelihood and approximations. Recall that the Gibbs chain from Definition 1 is based on alternate sampling from likelihood and approximation as depicted in Figure 2. A sequence of this chain has the form $(\theta_1, y_1, \theta_2, y_1, \dots)$ with $\theta_t \in \Theta$ and $y_t \in \mathcal{Y}$. We then defined the Gibbs prior π_G as the limiting distribution of the θ_t , but we can analogously consider the limiting distribution of the \mathcal{Y} -components y_t . To this end, we generalize Definition 1 by looking viewing the Gibbs chain as a Markov chain on $\Theta \times \mathcal{Y}$:

Definition 6 (Gibbs chain (extension to Definition 1)). 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 $\Theta \times \mathcal{Y}$ whose transition function is given by

$$r((\theta', y') | (\theta, y)) = f(y'|\theta)q(\theta'|y').$$

This chain is called the *Gibbs chain*. The projection onto the Θ -components is a Markov chain on Θ , any stationary distribution of which is called a *Gibbs prior* and denoted by π_G . The projection onto the \mathcal{Y} -components is a Markov chain on \mathcal{Y} , any stationary distribution of which is and denoted by p_G .

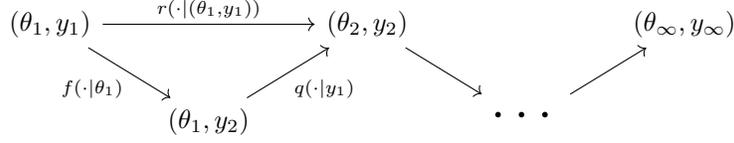


Figure 7: Schematic diagram of samples from the Gibbs chain from Definition 6, where one step first updates y with f and then θ with q . The distribution of θ_t converges to the Gibbs prior π_G and the distribution of y_t converges to p_G .

Kuo and Wang (2019) also studied this Gibbs chain for discrete distributions. In the main paper we specified a joint distribution on $\Theta \times \mathcal{Y}$ with the Gibbs prior π_G as the Θ -marginal and the likelihood F as the corresponding conditional. The main observation for measuring the degree of compatibility between F and Q is that we can also specify a joint distribution from the other direction, that is, with p_G as the \mathcal{Y} -marginal and Q as the corresponding conditional. We abbreviate those two joint distributions with $\pi_G F$ and $p_G Q$. They coincide if and only if F and Q are compatible. It is therefore natural to measure the degree of compatibility via some divergence between them.

A practical algorithm We can obtain (correlated) samples from the two joint distributions $\pi_G F$ and $p_G Q$ with the same Gibbs chain used for obtaining samples from the Gibbs prior π_G . Simulating the Gibbs chain is described by Figure 7. It yields a sequence $(\theta_1, y_1, \theta_2, y_2, \dots)$, where the marginal distributions of θ_t and y_t converge to π_G and p_G , respectively. Since the components are updated alternately with the conditional distributions, we can pair the entries to obtain samples from the joint distributions. However, the order of the pairing is important:

- $(\theta_t, y_t) \sim \pi_G F$, because $\theta_t \sim \pi_G$ (for large t) and $y_t \sim f(\cdot|\theta_t)$
- $(\theta_{t+1}, y_t) \sim p_G Q$, because $y_t \sim p_G$ (for large t) and $\theta_{t+1} \sim q(\cdot|y_t)$

This leads to basically the same algorithm as Algorithm 1 in the main paper, except that the auxiliary variables y_t are stored as well and paired accordingly with no computational overhead:

Algorithm 2: Simulating the Gibbs chain (samples from the joint distributions $\pi_G F$ and $p_G Q$)

Data: Likelihood f , approximate inference method q , number of steps T

Result: Correlated samples $(\theta_t, y_t)_{t=0}^{T-1}$ from π_G and $(\theta_{t+1}, y_t)_{t=0}^{T-1}$ from p_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 $q(\cdot|y_t)$

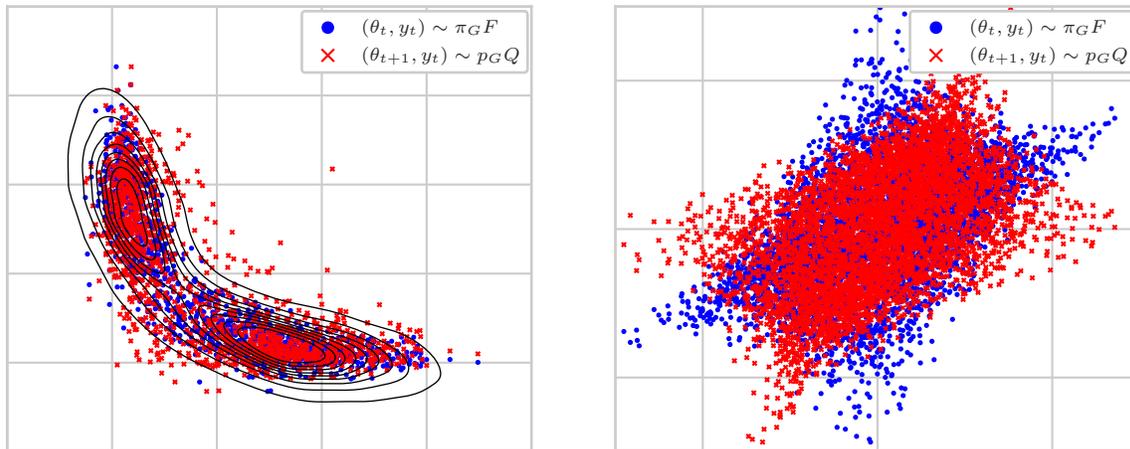
end

Example 7 (Gaussian conditional distributions). We demonstrate Algorithm 2 for two pairs of Gaussian conditional distributions with $\Theta = \mathcal{Y} = \mathbb{R}$, one compatible and one incompatible. The first example is taken from Arnold et al. (2001) and given by the conditional distributions

$$f(y|\theta) = \mathcal{N}\left(y \mid \frac{4}{1+\theta^2}, \frac{1}{1+\theta^2}\right) \quad \text{and} \quad q(\theta|y) = \mathcal{N}\left(\theta \mid \frac{4}{1+y^2}, \frac{1}{1+y^2}\right).$$

These conditionals are compatible with the bivariate joint density

$$p(\theta, y) = \exp\left((1, \theta, \theta^2) \begin{pmatrix} c & 4 & -1/2 \\ 4 & 0 & 0 \\ -1/2 & 0 & -1/2 \end{pmatrix} \begin{pmatrix} 1 \\ y \\ y^2 \end{pmatrix}\right),$$



(a) Correlated samples from $\pi_G F$ and $p_G Q$ for two compatible conditionals with the underlying joint distribution as contours. The conditionals being compatible is equivalent to all three joint distributions coinciding.

(b) Correlated samples from $\pi_G F$ and $p_G Q$ for two incompatible conditionals. The conditionals being incompatible is equivalent to those two joint distributions being different.

Figure 8: Samples from Algorithm 2 for the Gaussian conditional distributions of Example 7.

where $c \in \mathbb{R}$ plays the role of the normalizing constant. The corresponding samples from Algorithm 2 are shown in Figure 8a together with a contour plot of the true joint density $p(\theta, y)$. All three joint densities $\pi_G F$, $p_G Q$, and p overlap, confirming the compatibility of these conditional distributions.

The other example is given by the conditional densities

$$f(y|\theta) = \mathcal{N}\left(y \middle| \frac{\theta}{2}, \frac{1}{1+\theta^2}\right) \quad \text{and} \quad q(\theta|y) = \mathcal{N}\left(\theta \middle| \frac{y}{2}, \frac{1}{1+y^2}\right).$$

These conditional distributions are incompatible (Arnold et al. (2001) gives a full characterization of compatible Gaussian conditional distributions). Therefore the joint distributions $\pi_G F$ and $p_G Q$ cannot coincide exactly. This is confirmed by Figure 8b, which shows samples from the two joint distributions. Based on these samples we could now measure some kind of divergence between the two distributions to assess the degree of compatibility.

D RESULTS AND PROOFS FOR THE GAUSSIAN TOY EXAMPLE

D.1 Distributions in the Gaussian Toy Example

This section computes the distributions of interest for the Bayesian model defined in Eq. (5) from Section 4. This includes the posterior, the approximations, and the pointwise prior in Proposition 8, as well as the Gibbs prior in Theorem 9.

Proposition 8 (Posterior, approximations, and pointwise priors). *Consider the Bayesian model defined in Eq. (5) and let $\theta \in \Theta$ and $\mathbf{y} = (y_1, \dots, y_n) \in \mathcal{Y}$.*

(i) *The posterior distribution is given by*

$$p(\theta|\mathbf{y}) = \mathcal{N}(\theta|\mu_n, \Sigma_n), \quad (9)$$

where $\Sigma_n = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}$, $\mu_n = \Sigma_n (\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{\mathbf{y}})$, and $\bar{\mathbf{y}} = 1/n \sum_{j=1}^n y_j$.

(ii) *The mean field variational approximation is given by*

$$q(\theta|\mathbf{y}) = \mathcal{N}(\theta|\mu_n, \Lambda_n), \quad (10)$$

where

$$\Lambda_n = \begin{cases} \text{diag}(\Sigma_n^{-1})^{-1}, & \text{for } q \text{ defined via Eq. (6) (reverse)} \\ \text{diag}(\Sigma_n), & \text{for } q \text{ defined via Eq. (7) (forward)} \end{cases}. \quad (11)$$

Hereby, the diag operator keeps the diagonal entries of a matrix and sets all off-diagonal entries to 0.

(iii) Whether the pointwise prior $\pi_{\mathbf{y}}$ is a proper distribution depends on the matrix $\Lambda_n^{-1} - n\Sigma^{-1}$. If it is positive definite, then

$$\pi_{\mathbf{y}}(\theta) \propto \mathcal{N}\left(\theta \mid \mu_{\mathbf{y}}, \tilde{\Sigma}\right),$$

where $\tilde{\Sigma} = (\Lambda_n^{-1} - n\Sigma^{-1})^{-1}$ and $\mu_{\mathbf{y}} = \tilde{\Sigma}(\Lambda_n^{-1}\mu_n - n\Sigma^{-1}\bar{\mathbf{y}})$. Otherwise, $\pi_{\mathbf{y}}$ is improper. In particular, $\pi_{\mathbf{y}}$ is always proper in the setting **correlated prior**.

Theorem 9 (Gibbs prior). The Gibbs marginal π_G to the Bayesian model defined in Eq. (5) is given by

$$\pi_G(\theta) = \mathcal{N}(\theta \mid \mu_0, \Sigma_G). \quad (12)$$

Hereby, μ_0 is the mean of the prior distribution π and Σ_G satisfies the Lyapunov equation

$$A\Sigma_G A^\top - \Sigma_G + B = 0 \quad (13)$$

where $A = n\Sigma_n\Sigma^{-1}$ and $B = \Lambda_n + n\Sigma_n\Sigma^{-1}\Sigma_n$ with Σ_n and Λ_n defined as in Proposition 8.

Proof of Proposition 8. Recall the density function of a multivariate normal distribution $\mathcal{N}(\theta \mid \mu, \Sigma)$, which is given by

$$\mathcal{N}(\theta \mid \mu, \Sigma) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\right) \propto_\theta \exp\left(-\frac{1}{2}[\theta^\top \Sigma^{-1}\theta - 2\theta^\top \Sigma^{-1}\mu]\right).$$

Proof of (i). First observe that up to proportionality the likelihood $f(\mathbf{y} \mid \theta)$ as a function of θ depends only on the average observation $\bar{\mathbf{y}} = 1/n \sum_{j=1}^n y_j$

$$\begin{aligned} f(\mathbf{y} \mid \theta) &= \prod_{j=1}^n \mathcal{N}(y_j \mid \theta, \Sigma) \propto_\theta \prod_{j=1}^n \exp\left(-\frac{1}{2}[\theta^\top \Sigma^{-1}\theta - 2\theta^\top \Sigma^{-1}y_j]\right) \\ &= \exp\left(-\frac{1}{2}[\theta^\top n\Sigma^{-1}\theta - 2\theta^\top n\Sigma^{-1}\bar{\mathbf{y}}]\right) \\ &\propto_\theta \mathcal{N}(\bar{\mathbf{y}} \mid \theta, 1/n\Sigma). \end{aligned} \quad (14)$$

Together with $\pi(\theta) = \mathcal{N}(\theta \mid \mu_0, \Sigma_0)$, Bayes' theorem yields

$$\begin{aligned} p(\theta \mid \mathbf{y}) &\propto_\theta \pi(\theta)f(\mathbf{y} \mid \theta) \propto_\theta \exp\left(-\frac{1}{2}[\theta^\top \Sigma_0^{-1}\theta - 2\theta^\top \Sigma_0^{-1}\mu_0]\right) \exp\left(-\frac{1}{2}[\theta^\top n\Sigma^{-1}\theta - 2\theta^\top n\Sigma^{-1}\bar{\mathbf{y}}]\right) \\ &= \exp\left(-\frac{1}{2}[\theta^\top (\Sigma_0^{-1} + n\Sigma^{-1})\theta - 2\theta^\top (\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{\mathbf{y}})]\right) \\ &= \exp\left(-\frac{1}{2}[\theta^\top \Sigma_n^{-1}\theta - 2\theta^\top \Sigma_n^{-1}\mu_n]\right) \\ &\propto \mathcal{N}(\theta \mid \mu_n, \Sigma_n). \end{aligned}$$

Note that $\Sigma_0^{-1} + n\Sigma^{-1}$ is positive definite as the sum of two positive definite matrices, and therefore $\Sigma_n = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}$ is positive definite as well.

Proof of (ii). By definition of the mean-field variational family, every variational density factorizes as $q(\theta \mid \mathbf{y}) = \prod_{j=1}^m q_j(\theta_j)$. For the setting **forward** we refer to (Bishop, 2006, Section 10.1.2), where it is shown that the optimal q_j simply coincide with the marginal densities of the posterior $q_j(\theta_j) = p(\theta_j \mid \mathbf{y}) = \mathcal{N}(\theta_j \mid (\mu_n)_j, (\Sigma_n)_{j,j})$.

For the other setting **reverse** let \mathbb{E}_{-j} denote the expectation over all latent variables θ_i except θ_j with respect to the factorized distribution $\prod_{i \neq j} q_i(\theta_i)$. To simplify the following computation, we abbreviate $\Sigma_n^{-1} =: \Gamma$ and $\mu_n =: \mu$ (now μ_k refers to the k -th component of μ_n). We use that the optimal solution satisfies the recursive update rule

$$\begin{aligned}
 q_j(\theta_j) &\propto_{\theta_j} \exp(\mathbb{E}_{-j} \log p(\theta, \mathbf{y})) && \text{(Bishop (2006))} \\
 &\propto_{\theta_j} \exp(\mathbb{E}_{-j} \log p(\theta | \mathbf{y})) \\
 &\propto_{\theta_j} \exp\left(-\frac{1}{2} \mathbb{E}_{-j} \left[(\theta - \mu)^\top \Gamma (\theta - \mu) \right]\right) && (p(\theta | \mathbf{y}) = \mathcal{N}(\theta | \mu, \Gamma^{-1}) \text{ by Eq. (9)}) \\
 &\propto_{\theta_j} \exp\left(-\frac{1}{2} \left(\Gamma_{j,j} (\theta_j - \mu_j)^2 + 2(\theta_j - \mu_j) \sum_{k \neq j} \Gamma_{j,k} (m_k - \mu_k) \right)\right) && (m_k := \mathbb{E}_{q_k} \theta_k) \\
 &\propto_{\theta_j} \exp\left(-\frac{1}{2\Gamma_{j,j}^{-1}} \left(\theta_j - \mu_j + \frac{1}{\Gamma_{j,j}} \sum_{k \neq j} \Gamma_{j,k} (m_k - \mu_k) \right)^2\right) \\
 &\propto_{\theta_j} \mathcal{N}\left(\theta_j \middle| \mu_j - \frac{1}{\Gamma_{j,j}} \sum_{k \neq j} \Gamma_{j,k} (m_k - \mu_k), \Gamma_{j,j}^{-1}\right) \\
 &= \mathcal{N}(\theta_j | m_j, \Gamma_{j,j}^{-1}). && \text{(Definition of } m_j)
 \end{aligned}$$

This shows that the solutions q_j are normally distributed and have the claimed variance $\Gamma_{j,j}^{-1} = (\Sigma_n^{-1})_{j,j}^{-1}$. However, their means m_j are only recursively determined and to conclude the proof, we need to show that $m_j = \mu_j$. The last equation in the previous computation gives the recursive relation

$$\begin{aligned}
 m_j &= \mu_j - \frac{1}{\Gamma_{j,j}} \sum_{k \neq j} \Gamma_{j,k} (m_k - \mu_k) && \forall j = 1, \dots, m \\
 \Leftrightarrow & \frac{1}{\Gamma_{j,j}} \sum_{k=1}^m \Gamma_{j,k} m_k = \frac{1}{\Gamma_{j,j}} \sum_{k=1}^m \Gamma_{j,k} \mu_k && \forall j = 1, \dots, m \\
 \Leftrightarrow & \langle \Gamma_j, \mathbf{m} \rangle = \langle \Gamma_j, \mu \rangle && \forall j = 1, \dots, m \\
 \Leftrightarrow & \Gamma \mathbf{m} = \Gamma \mu,
 \end{aligned}$$

where Γ_j denotes the j -th row of Γ and \mathbf{m} is the vector containing all m_k . Since Γ is positive definite, the last equality implies $\mathbf{m} = \mu$ and concludes the proof.

Proof of (iii). We can compute the pointwise prior $\pi_{\mathbf{y}}(\theta)$ with its definition in Eq. (1) with Eq. (10) for $q(\theta | \mathbf{y})$ and Eq. (14) for $f(\mathbf{y} | \theta)$ as

$$\begin{aligned}
 \pi_{\mathbf{y}}(\theta) &\propto_{\theta} \frac{q(\theta | \mathbf{y})}{f(\mathbf{y} | \theta)} \propto_{\theta} \frac{\mathcal{N}(\theta | \mu_n, \Lambda_n)}{\mathcal{N}(\bar{\mathbf{y}} | \theta, 1/n\Sigma)} \propto_{\theta} \frac{\exp\left(-\frac{1}{2} [\theta^\top \Lambda_n^{-1} \theta - 2\theta^\top \Lambda_n^{-1} \mu_n]\right)}{\exp\left(-\frac{1}{2} [\theta^\top n\Sigma^{-1} \theta - 2\theta^\top n\Sigma^{-1} \bar{\mathbf{y}}]\right)} \\
 &= \exp\left(-\frac{1}{2} [\theta^\top (\Lambda_n^{-1} - n\Sigma^{-1}) \theta - 2\theta^\top (\Lambda_n^{-1} \mu_n - n\Sigma^{-1} \bar{\mathbf{y}})]\right). && (15)
 \end{aligned}$$

If $\Lambda_n^{-1} - n\Sigma^{-1}$ is positive definite, we can continue the computation

$$\begin{aligned}
 \pi_{\mathbf{y}}(\theta) &\propto_{\theta} \exp\left(-\frac{1}{2} [\theta^\top (\Lambda_n^{-1} - n\Sigma^{-1}) \theta - 2\theta^\top (\Lambda_n^{-1} \mu_n - n\Sigma^{-1} \bar{\mathbf{y}})]\right) \propto_{\theta} \exp\left(-\frac{1}{2} [\theta^\top \tilde{\Sigma}^{-1} \theta - 2\theta^\top \tilde{\Sigma}^{-1} \mu_{\mathbf{y}}]\right) \\
 &\propto_{\theta} \mathcal{N}\left(\theta \middle| \mu_{\mathbf{y}}, \tilde{\Sigma}\right).
 \end{aligned}$$

If $\Lambda_n^{-1} - n\Sigma^{-1} =: S$ is not positive definite, then we can show that $\pi_{\mathbf{y}}$ is improper. In this case, S has an eigenvalue $\lambda \leq 0$ and corresponding eigenvector $v \in \mathbb{R}^d$ with $\|v\| = 1$. Consider the hypercylinder A of points whose distance to the axis $\mathbb{R}v$ is at most 1, formally defined as

$$A := \{\theta \in \mathbb{R}^d \mid \theta = tv + w, \text{ where } t \in \mathbb{R}, w \in v^\perp, \|w\| = 1\},$$

where $v^\perp = \{w \in \mathbb{R}^d : \langle v, w \rangle = 0\}$. Abbreviate $\gamma := \Lambda_n^{-1} \mu_n - n \Sigma^{-1} \bar{\mathbf{y}}$ and collect all constants in $C > 0$ (can change at different steps). Then we can lower bound the right hand side of Eq. (15) on A via

$$\begin{aligned}
 \frac{q(\theta|\mathbf{y})}{f(\mathbf{y}|\theta)} &= C \exp\left(-\frac{1}{2}\theta^\top S\theta + \langle \theta, \gamma \rangle\right) \\
 &= C \exp\left(-\frac{1}{2}(tv + w)^\top S(tv + w) + \langle tv + w, \gamma \rangle\right) && (\theta = tv + w \in A) \\
 &= C \exp\left(\underbrace{-\frac{1}{2}\lambda t^2}_{\geq 0} + \langle v, \gamma \rangle t + \langle w, \gamma \rangle - \frac{1}{2}w^\top S w\right) && (Sv = \lambda v, \|v\| = 1, \langle v, w \rangle = 0) \\
 &\geq C \exp\left(\langle v, \gamma \rangle t + \langle w, \gamma \rangle - \frac{1}{2}w^\top S w\right) \\
 &\geq C \exp(\langle v, \gamma \rangle t). && (\langle w, \gamma \rangle - \frac{1}{2}w^\top S w \text{ is bounded for } \|w\| \leq 1)
 \end{aligned}$$

Using this lower bound, we can lower bound the integral over A through

$$\int_A \frac{q(\theta|\mathbf{y})}{f(\mathbf{y}|\theta)} d\theta \geq C \int_{\mathbb{R}} \exp(\langle v, \gamma \rangle t) dt = \infty.$$

Hence $\pi_{\mathbf{y}}$ is improper.

The last statement is that $\pi_{\mathbf{y}}$ is always proper in the setting **correlated prior** ($\Sigma = I$), which means we need to show that S has strictly positive eigenvalues.

We treat the cases **reverse** and **forward** separately. For **reverse**, it is

$$S = \Lambda_n^{-1} - n\Sigma^{-1} = \text{diag}(\Sigma_0^{-1} + nI) - nI = \text{diag}(\Sigma_0^{-1}). \quad (\text{diag is a linear operator})$$

The diagonal entries of the symmetric positive definite matrix Σ_0^{-1} are lower bounded by its smallest eigenvalue $\lambda_{\min}(\Sigma_0^{-1}) > 0$, which follows from the Courant–Fischer–Weyl min-max principle. Since these diagonal entries are the eigenvalues of S , this implies that S is positive definite. For the other case **forward**, we have

$$S = \Lambda_n^{-1} - n\Sigma^{-1} = \text{diag}\left(\left(\Sigma_0^{-1} + n\right)^{-1}\right) - n.$$

We again need to bound the diagonal elements of $(\Sigma_0^{-1} + n)^{-1}$ with its eigenvalues. A similar argument as above yields

$$\lambda_{\max}\left(\text{diag}\left(\left(\Sigma_0^{-1} + n\right)^{-1}\right)\right) \leq \lambda_{\max}\left(\left(\Sigma_0^{-1} + n\right)^{-1}\right) = \frac{1}{\lambda_{\min}(\Sigma_0^{-1}) + n}. \quad (16)$$

With this, the eigenvalues of S are bounded by

$$\begin{aligned}
 \lambda_{\min}(S) &= \lambda_{\min}\left(\text{diag}\left(\left(\Sigma_0^{-1} + n\right)^{-1}\right) - n\right) = \frac{1}{\lambda_{\max}\left(\text{diag}\left(\left(\Sigma_0^{-1} + n\right)^{-1}\right)\right)} - n \\
 &\geq \frac{1}{\frac{1}{\lambda_{\min}(\Sigma_0^{-1}) + n}} - n && (\text{Eq. (16)}) \\
 &= \lambda_{\min}(\Sigma_0^{-1}) > 0.
 \end{aligned}$$

□

To prove Theorem 9 we require some general properties of Gaussian densities in Lemma 10 and Lemma 11 because the proofs consist mainly of rearranging Gaussian densities. Next we compute the transition function of the Markov chain from Definition 1 in Proposition 12. We then prove Theorem 9 by guessing that the stationary distribution is Gaussian and verifying the stationary equation.

Lemma 10 (Some properties of Gaussians). Let $\mathcal{N}(x|\mu, \Sigma)$ denote the density of a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ on \mathbb{R}^d at $x \in \mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and positive definite covariance $\Sigma \in \mathbb{R}^{d \times d}$. Then the following equalities hold

- (i) $\mathcal{N}(x+y|\mu, \Sigma) = \mathcal{N}(x|\mu-y, \Sigma)$ and $\mathcal{N}(x|\mu, \Sigma) = \mathcal{N}(\mu|x, \Sigma)$ for $x, y \in \mathbb{R}^d$.
- (ii) Let $A \in \mathbb{R}^{d \times d}$ be non-singular. Then $\mathcal{N}(Ax|\mu, \Sigma) = C_{A, \Sigma} \mathcal{N}(x|A^{-1}\mu, A^{-1}\Sigma A^{-T})$, where $C_{A, \Sigma} \in \mathbb{R}$ is a constant that depends only on A and Σ .
- (iii) Let $\mu_1, \mu_2 \in \mathbb{R}^d$ and $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d \times d}$ positive definite. Then the convolution of two Gaussian densities corresponds to the sum of two independent Gaussians, i.e., for $z \in \mathbb{R}^d$ it holds

$$\int_{\mathbb{R}^d} \mathcal{N}(z-x|\mu_1, \Sigma_1) \mathcal{N}(x|\mu_2, \Sigma_2) dx = [\mathcal{N}(\cdot|\mu_1, \Sigma_1) * \mathcal{N}(\cdot|\mu_2, \Sigma_2)](z) = \mathcal{N}(z|\mu_1 + \mu_2, \Sigma_1 + \Sigma_2).$$

Proof. Points (i) and (iii) are trivial. For (ii), we compute

$$\begin{aligned} \mathcal{N}(Ax|\mu, \Sigma) &= C_{\Sigma} \exp\left(-\frac{1}{2}(Ax - \mu)^{\top} \Sigma^{-1} (Ax - \mu)\right) = C_{\Sigma} \exp\left(-\frac{1}{2}(x - A^{-1}\mu)^{\top} A^{\top} \Sigma^{-1} A (x - A^{-1}\mu)\right) \\ &= C_{A, \Sigma} \mathcal{N}(x|A^{-1}\mu, A^{-1}\Sigma A^{-T}). \end{aligned}$$

Note that $A^{-1}\Sigma A^{-T}$ is positive definite as well: symmetry is obvious and it holds

$$x^{\top} A^{-1} \Sigma A^{-T} x = (A^{-T} x)^{\top} \Sigma (A^{-T} x) > 0$$

for $x \neq 0$ because A is non-singular and Σ is positive definite. \square

The next lemma computes an integral that appears both in computing the transition function of the Gibbs chain and in computing its stationary distribution.

Lemma 11. Let $\theta', a, \mu_2 \in \mathbb{R}^d$, $A \in \mathbb{R}^{d \times d}$ non-singular, and $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d \times d}$ positive definite. Then it is

$$\int_{\mathbb{R}^d} \mathcal{N}(\theta'|a + Ax, \Sigma_1) \mathcal{N}(x|\mu_2, \Sigma_2) dx = \mathcal{N}(\theta'|a + A\mu_2, \Sigma_1 + A\Sigma_2 A^{\top}).$$

Proof. We start by rephrasing the first density

$$\begin{aligned} \mathcal{N}(\theta'|a + Ax, \Sigma_1) &= \mathcal{N}(Ax|\theta' - a, \Sigma_1) && \text{(Lemma 10, (i))} \\ &= C_{A, \Sigma_1} \mathcal{N}(x|A^{-1}(\theta' - a), A^{-1}\Sigma_1 A^{-T}) && \text{(Lemma 10, (ii))} \\ &= C_{A, \Sigma_1} \mathcal{N}(A^{-1}\theta' - x|A^{-1}a, A^{-1}\Sigma_1 A^{-T}). && \text{(Lemma 10, (i))} \end{aligned}$$

This yields

$$\begin{aligned} \int_{\mathbb{R}^d} \mathcal{N}(\theta'|a + Ax, \Sigma_1) \mathcal{N}(x|\mu_2, \Sigma_2) dx &\propto_{\theta'} \int_{\mathbb{R}^d} \mathcal{N}(A^{-1}\theta' - x|A^{-1}a, A^{-1}\Sigma_1 A^{-T}) \mathcal{N}(x|\mu_2, \Sigma_2) dx \\ &= \mathcal{N}(A^{-1}\theta'|A^{-1}a + \mu_2, A^{-1}\Sigma_1 A^{-T} + \Sigma_2) && \text{(Lemma 10, (iii))} \\ &\propto_{\theta'} \mathcal{N}(\theta'|a + A\mu_2, \Sigma_1 + A\Sigma_2 A^{\top}). && \text{(Lemma 10, (ii))} \end{aligned}$$

\square

We can now compute the transition function of the Gibbs chain from Definition 1.

Proposition 12. The transition function of the Gibbs chain is given by Gaussian distributions

$$r(\theta'|\theta) = \mathcal{N}(\theta'|a + A\theta, B), \tag{17}$$

where $\theta, \theta' \in \mathbb{R}^d$, $a = \Sigma_n \Sigma_0^{-1} \mu_0$, $A = n \Sigma_n \Sigma^{-1}$, and $B = \Lambda_n + n \Sigma_n \Sigma^{-1} \Sigma_n$.

Proof. Let $\theta, \theta' \in \mathbb{R}^d$. By definition, the transition function of the Gibbs chain is given by

$$r(\theta'|\theta) = \int_{\mathbb{R}^{n \times d}} q(\theta'|\mathbf{y})f(\mathbf{y}|\theta) d\mathbf{y} = \int_{\mathbb{R}^d} q(\theta'|\bar{\mathbf{y}})\bar{f}(\bar{\mathbf{y}}|\theta) d\bar{\mathbf{y}},$$

where we have transformed the integral to the mean $\bar{\mathbf{y}}$, because q only depends on \mathbf{y} through $\bar{\mathbf{y}}$. The corresponding push-forward measure is given by $\bar{f}(\bar{\mathbf{y}}|\theta) = \mathcal{N}(\bar{\mathbf{y}}|\theta, 1/n\Sigma)$. Using Proposition 8 and expressing μ_n with a and A , we have that $q(\theta'|\bar{\mathbf{y}}) = \mathcal{N}(\theta'|\mu_n, \Lambda_n) = \mathcal{N}(\theta'|a + A\bar{\mathbf{y}}, \Lambda_n)$. Putting everything together, we get

$$\begin{aligned} r(\theta'|\theta) &= \int_{\mathbb{R}^d} \mathcal{N}(\theta'|a + A\bar{\mathbf{y}}, \Lambda_n) \mathcal{N}\left(\bar{\mathbf{y}}\left|\theta, \frac{1}{n}\Sigma\right.\right) d\bar{\mathbf{y}} \\ &= \mathcal{N}\left(\theta'\left|a + A\theta, \Lambda_n + A\frac{1}{n}\Sigma A^\top\right.\right). \end{aligned} \quad (\text{Lemma 11})$$

The equality $A\frac{1}{n}\Sigma A^\top = n\Sigma_n\Sigma^{-1}\Sigma_n$ concludes the proof. \square

We are now ready to prove Theorem 9.

Proof of Theorem 9. The proof is based on guessing that the stationary distribution is Gaussian. We first show that Gaussian distributions are closed under taking a step with the transition function and then derive the parameters of the stationary distribution based on the stationary equation.

Let $p(\theta) = \mathcal{N}(\theta|m, M)$ with $m \in \mathbb{R}^d$ and $M \in \mathbb{R}^{d \times d}$ positive definite. Using Lemma 11 and Proposition 12, the distribution after one step Rp is given by

$$Rp(\theta') = \int_{\mathbb{R}^d} r(\theta'|\theta)p(\theta) d\theta = \int_{\mathbb{R}^d} \mathcal{N}(\theta'|a + A\theta, B) \mathcal{N}(\theta|m, M) d\theta = \mathcal{N}(\theta'|a + Am, B + AMA^\top). \quad (18)$$

If a p satisfies the stationary equation $p = Rp$, then it is the stationary distribution $p = \pi_G$. Using that Rp is again Gaussian, Eq. (18) shows that this is satisfied if and only if

$$m = a + Am \quad \text{and} \quad M = B + AMA^\top.$$

The solution for the mean equation is obtained by rearranging and plugging in the definitions of a , A and Σ_n :

$$\begin{aligned} m &= (I - A)^{-1}a = (I - n\Sigma_n\Sigma^{-1})^{-1}\Sigma_n\Sigma_0^{-1}\mu_0 = (\Sigma_n^{-1} - n\Sigma^{-1})\Sigma_0^{-1}\mu_0 \\ &= (\Sigma_0^{-1} + n\Sigma^{-1} - n\Sigma^{-1})^{-1}\Sigma_0^{-1}\mu_0 \\ &= \mu_0. \end{aligned}$$

The stationary equation for the covariance matrix M is equivalent to the Lyapunov equation

$$AMA^\top - M + B = 0,$$

which has a unique solution. This concludes the proof. \square

D.2 Numerical Evaluation of the Biases in the Gaussian Toy Example

This section presents numerical values that back up the statements about the biases from Section 4.2.

The first bias is the compactness of the mean-field approximations. Table 1 shows the compactness of all relevant distributions as measured by the entropy, which is given by $d/2(1 + \ln(2\pi)) + 1/2\ln(\det \Sigma)$ for a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$. As stated in the main paper, under the setting **forward** approximations q are less compact than the exact posterior p . This is reflected by the priors as the Gibbs prior π_G is less compact than the exact prior π . Under the setting **reverse**, this trend is reversed: approximations are more compact than the exact posterior, and the Gibbs prior is more compact than the exact prior.

The second bias is the loss of correlation under the mean-field approximations. Table 2 shows the correlation between different components θ_1 and θ_2 of the 2-dimensional latent variable $\theta = (\theta_1, \theta_2)$ only for the prior distributions, because the components of the approximations are by definition uncorrelated. Recall that the

exact posterior distribution was the same in both settings, but under `correlated prior` the posterior correlation was due to prior correlation, whereas under `correlated likelihood` it was due to likelihood correlation. In the setting `correlated prior`, the Gibbs prior is less correlated than the prior. In the setting `correlated likelihood`, the prior is uncorrelated, but the Gibbs prior is negatively correlated to “cancel out” the positive correlation of the likelihood covariance.

Table 1: Compactness of various distributions across settings as measured by the entropy. First value is under the setting `correlated prior` and second value is under the setting `correlated likelihood`. Note that the covariance of exact and approximate posterior does not depend on the observation.

Entropy	forward	reverse
Prior π	2.24 / 2.84	2.24 / 2.84
Gibbs prior π_G	2.82 / 3.15	2.21 / 2.52
Exact posterior p	1.50 / 1.50	1.50 / 1.50
Approximate posterior q	1.97 / 1.97	1.02 / 1.02

Table 2: Correlation of prior distributions across settings as measured by the covariance $\text{Cov}(\theta_1, \theta_2)$ between the components of $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$. First value is under the prior distribution π and second value is under the corresponding Gibbs prior π_G .

Covariance $\text{Cov}(\theta_1, \theta_2)$	forward	reverse
<code>correlated prior</code>	1.45 / 0.91	1.45 / 0.74
<code>correlated likelihood</code>	0 / -1.13	0 / -0.52

E EXPERIMENTAL DETAILS

In this section, we give more details on the Bayesian models and approximations to their posteriors which are considered in Section 5. We used the python library `numpyro` (Phan et al., 2019) for the posterior approximation methods Laplace, NUTS, and ADVI.

Baseline Talts et al. (2018) We allocate this baseline the same resources as the corresponding Gibbs chain in terms of draws from the posterior. That is, if our Gibbs chain runs for M steps, the baseline repeats N draws $(\tilde{\theta}, \tilde{y})$ with $\theta_1, \dots, \theta_L \sim q(\cdot | \tilde{y})$ such that $N \cdot L \approx M$. Specifically, we choose $N = 323$ and $L = 31$. For the histograms in Figure 6, we re-binned once to reduce noise.

Convergence monitoring We monitor the convergence of our Gibbs chains with two standard measures, the Gelman-Rubin diagnostic \hat{R} (Gelman and Rubin, 1992) and the lag- k autocorrelation. Both are shown in Figure 9 for all experiments from Section 5. The Gelman-Rubin diagnostic \hat{R} uses multiple chains to compute the ratio of between-chain variance to within-chain variance. A ratio $\hat{R} \approx 1$ indicates convergence. The top row of Figure 9 shows that this value is reached quickly in all cases except for NUTS-long, which takes longer to converge. Potential explanations are that convergence is generally slower in the high-dimensional setting ($d = 100$ for NUTS-long compared to $d = 2$ for Log-normals) and that the other less accurate methods introduce additional bias that promotes faster convergence. The lag- k autocorrelation is defined as the correlation of a sequence with its shifted version by k steps. A high autocorrelation of a Markov chain indicates slow mixing and thus slower convergence. The bottom row of Figure 9 shows the autocorrelation for the Gibbs chains, which eventually oscillate around 0 due to finite sample noise. The autocorrelation gets close to 0 quickly for the low-dimensional setting log-normals and for the less accurate methods NUTS-short and ADVI-mf in the high-dimensional setting. Only the more accurate methods in the high-dimensional setting NUTS-long and ADVI-fr take longer to reach 0. This hints towards slower convergence and is in line with Talts et al. (2018), who predict slow convergence of the Gibbs chain when the parameters are strongly correlated to the observations. In particular, we expect this to be the case for a large number of observations.

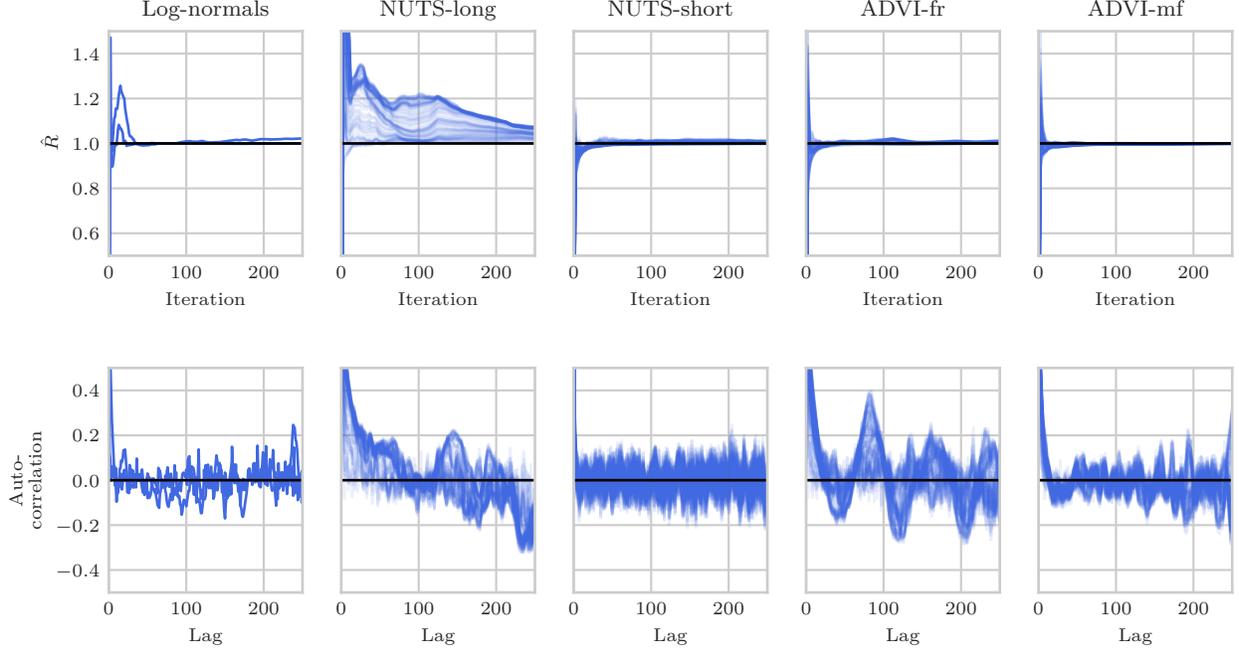


Figure 9: Gelman-Rubin diagnostic \hat{R} (**top row**) and lag- k autocorrelation (**bottom row**) for the Gibbs chains from Section 5 with one curve per dimension. Section 5.1 (**first column**) has $d = 2$ dimensions and Section 5.2 (**other columns**) have $d = 100$ dimensions. Values $\hat{R} \approx 1$ or lag- k autocorrelation ≈ 0 indicate convergence of the Gibbs chain.

E.1 Sum of log-normals

The Bayesian model has latent parameters $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}$, on which we place a prior $\pi(\theta)$ with independent marginal distributions $\mu \sim \mathcal{N}(0, 1)$ and $\sigma^2 \sim \text{Gamma}(1, 1)$. The likelihood $f(y|\theta)$ for an observation $y > 0$ is given by an L -fold convolution of a log-normal distribution, that is, $y|\theta \sim \text{LogNormal}^{*L}(\mu, \sigma^2)$.

To obtain the approximation $q(\theta|y)$ to the true posterior $p(\theta|y)$ of this model, we employ the following two-step procedure:

1. Define an approximate likelihood $\tilde{f}(y|\theta)$ as the Fenton-Wilkinson approximation to the true likelihood f , which is another log-normal distribution with matching first two moments. Specifically, $\tilde{f}(\cdot|\theta)$ describes the distribution $\text{LogNormal}(\alpha, \beta^2)$, where

$$\begin{aligned} \alpha &= \mu + \log L + 0.5(\sigma^2 - \beta^2), \\ \beta^2 &= \log \left[\frac{\exp \sigma^2 - 1}{L} + 1 \right]. \end{aligned}$$

2. Define $q(\theta|y)$ as the Laplace-approximation to the posterior of this new model $\tilde{p}(\theta|y) \propto \pi(\theta)\tilde{f}(y|\theta)$. This means that $q(\cdot|y)$ describes a bivariate normal distribution $\mathcal{N}(\theta_y^*, \Sigma_y)$ with

$$\begin{aligned} \theta_y^* &= \arg \max_{\theta} \pi(\theta)\tilde{f}(y|\theta) \\ \Sigma_y &= -H_{\log \tilde{p}}^{-1}, \end{aligned}$$

where $H_{\log \tilde{p}}$ describes the Hessian matrix of $\theta \mapsto \log(\pi(\theta)\tilde{f}(y|\theta))$.

E.2 Stochastic Volatility

This model is a simplified model of the one described in Hoffman and Gelman (2014), who place additional prior distributions on the parameters σ , ν , and θ_0 . We made the simplifying choice $\theta_0 = 0$. The other hyperparameters $\sigma = .09$ and $\nu = 12$ were chosen by taking the posterior means under S&P500 dataset. The posterior was approximated with NUTS where the priors were $\sigma \sim \text{Exp}(50)$ and $\nu \sim \text{Exp}(0.1)$, following Hoffman and Gelman (2014).

Measuring compactness and divergence Table 3 supplements our statements about the bias of the approximation methods in Section 5.2. Regarding compactness, we can confirm that the methods NUTS-short, ADVI-fr, and ADVI-mf are overly compact compared to the original prior. Regarding divergence, we see that the method NUTS-long is closest to the original prior and the restrictive method ADVI-mf is farthest. The more powerful versions yield Gibbs priors that are closer to the original prior, that is, NUTS-long is closer than NUTS-short and ADVI-fr is closer than ADVI-mf.

Table 3: Compactness and distance to original prior for the approximation methods of Section 5.2. Compactness is measured by the Frobenius norm of the empirical covariance matrix and distance to the original prior is measured by the maximum mean discrepancy under the Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2/2)$.

	Original prior	NUTS-long	NUTS-short	ADVI-fr	ADVI-mf
Compactness	34.18	23.81	3.37	6.09	1.21
Distance to original prior	0	0.014	0.035	0.021	0.179