Discovering Inductive Bias with Gibbs Priors

Eduard Vladimirov

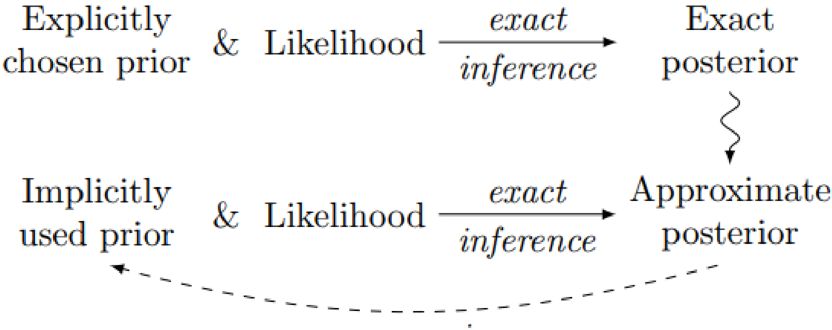
MIPT, 2023

March 12, 2024

Motivation

Main idea

The problem: diagnosing approximate Bayesian inference methods in terms of their inductive bias
The solution: Gibbs prior as a natural solution to the problem and as a diagnostic tool. It is based on pseudo-Gibbs sampling



Background

Идея MCMC

Пусть имеется однородная марковская цепь с функцией плотности вероятности перехода между состояниями $q(\mathbf{Z}_{i+1}|\mathbf{Z}_i)$.

- Возьмем некоторое $p_0(\mathbf{Z})$ и сгенерируем $\mathbf{Z}_0 \sim p_0(\mathbf{Z})$;
- Генерируем $\mathbf{Z}_{i+1} \sim q(\mathbf{Z}_{i+1}|\mathbf{Z}_i)$, $i = 0, 1, \ldots$;
- Выбрасываем первые m_0 наблюдений (и прореживаем, если нужна НОР (i.i.d) выборка).

Figure: Markov Chain Monte-Carlo

Схема Гиббса (Gibbs)

$p(\mathbf{Z}) \propto \tilde{p}(\mathbf{Z})$, $\mathbf{Z} \in \mathbb{R}^n$.
Считаем, что одномерные условные распределения $p(z_j|\mathbf{Z}_{\setminus j})$ легко нормируемы.

- Имеем $\mathbf{Z}_i$, хотим получить $\mathbf{Z}_{i+1}$;
- $z_{i+1}^1 \sim p(z^1|z_i^2, \ldots, z_i^n)$;
  $z_{i+1}^2 \sim p(z^2|z_{i+1}^1, z_i^3, \ldots, z_i^n)$;
  $\ldots$
  $z_{i+1}^n \sim p(z^n|z_{i+1}^1, z_{i+1}^2, \ldots, z_{i+1}^{n-1})$.

Figure: Gibbs schema

Existing approaches

Divergence-based Diagnostics	**True Posterior-based Diagnostics**
Stein discrepancies between the posterior and its approximation.	distortion map for posterior cumulative distribution functions to the identity.
symmetric KL divergence between the approximation and another baseline approximation.	compare average posterior means and covariances to prior means and covariances.
the symmetric KL divergence between the true joint distribution $p(y)p(\theta\|y)$ and its approximation $p(y)q(\theta\|y)$.	distribution of posterior quantiles, tested for uniformity; corrected by Talts et al. (2018).
	test for uniformity of p-values related to the coverage property; this method is extended by Rodrigues et al. (2018).

Table: Categorization of diagnostics in literature

Designation and pointwise-prior

Let $(f(\cdot|\theta))$ be the likelihood and $(q(\cdot|y))$ the approximations to the posteriors $(p(\cdot|y))$. It is reasonable to define the implicit prior to the approximations by fixing an observation y and simply reverting Bayes' theorem

$$\pi_y(\theta) \propto q(\theta|y)/f(y|\theta).$$

Unfortunately, π_y generally depends on the observation y. This means that the approximations to different observations can correspond to different implicit priors, in which case no single distribution $\tilde{\pi}$ satisfies $q(\theta|y) \propto \tilde{\pi}(\theta)f(y|\theta)$.

Gibbs prior

Definition

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

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

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

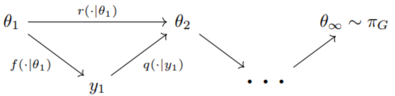


Figure 2: Schematic diagram of samples from the Gibbs chain (Definition 1) with auxiliary variables y_t. The distribution of θ_t converges to the Gibbs prior π_G.

Figure: Sampling from Gibbs chain

Algorithm 1: Simulating the Gibbs chain[3]

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

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

Figure: Simulating the Gibbs chain

Existence and uniqueness of Gibbs priors

Theorem

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.

(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.

Gaussian Toy Model

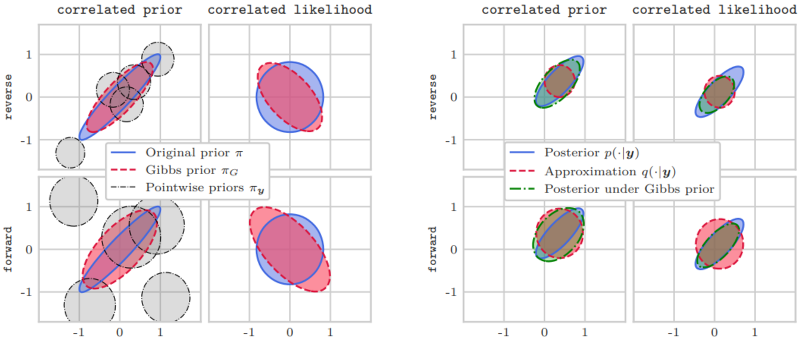
Problem: estimate the mean $\theta \in \mathbb{R}^d$ of a d-dimensional Gaussian distribution with known covariance matrix based on n independent samples $y_1, \dots, y_n \in \mathbb{R}^d$. Placing a Gaussian prior on θ yields the Bayesian model

$$\theta \sim \mathcal{N}(\mu_0, \Sigma_0),$$

$$y_i|\theta \overset{\text{indep}}{\sim} \mathcal{N}(\theta, \Sigma), \quad i = 1, \dots, n,$$

where $\mu_0 \in \mathbb{R}^d$ and $\Sigma_0, \Sigma \in \mathbb{R}^{d \times d}$ are positive definite.

Gaussian Toy Model



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

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

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

Figure: Prior and posterior distributions

Baseline

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 $f : \Theta \to \mathbb{R}$. Under random samples $\tilde{\theta} \sim \pi, \tilde{y} \sim f(\cdot|\tilde{\theta})$, and $\theta_1, \dots, \theta_L \sim q(\cdot|\tilde{y})$, the rank of $f(\tilde{\theta})$ in $\{f(\theta_1), \dots, f(\theta_L)\}$ is computed. This is repeated over multiple draws of $(\tilde{\theta}, \tilde{y})$, which gives a histogram of the ranks. Since the histogram is uniform under the exact posterior, any deviations from uniformity indicate an approximation mismatch.

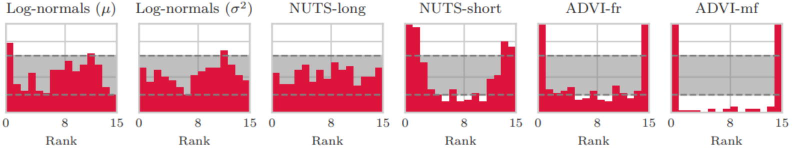


Figure 6: Histograms of rank statistics for the baseline Talts et al. (2018). First two histograms are for Section 5.1 with coordinates as summary statistics, other histograms are for Section 5.2 with the mean. Gray band shows a 99% confidence interval under the exact posterior. Deviations from uniformity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

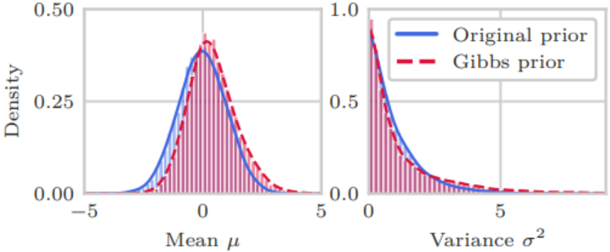
Figure: Baseline

Sum of Log-Normals

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

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

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



Stochastic Volatility

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

$$\theta_i|\theta_{i-1} \sim \mathcal{N}(\theta_i, \sigma^2), \quad i = 1, \dots, T,$$

$$y_i \overset{\text{indep}}{\sim} \text{StudentT}(\nu, 0, \exp(\theta_i)), \quad i = 1, \dots, T,$$

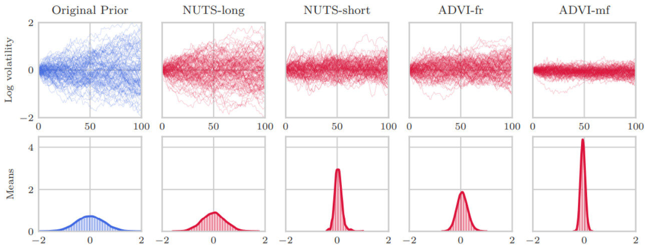


Figure 5: **Top row:** Samples of $\theta \in \mathbb{R}^{100}$ from original prior (blue) and Gibbs priors (red) under various approximations. **Bottom row:** Histograms of the summary statistic $\theta \mapsto 1/100 \sum_{i=1}^{100} \theta_i$, which is the mean value of a time series. Methods that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

Literature

1. **Main article** Discovering Inductive Bias with Gibbs Priors: A Diagnostic Tool for Approximate Bayesian Inference.