Bayesian Statistics

Fabio Sigrist

ETH Zurich, Autumn Semester 2019

Today's topics

- Sequential Monte Carlo
- Approximate Bayesian computation
- Gaussian processes

Sequential Monte Carlo

Basic idea sequential Monte Carlo

- **Idea**: Instead of sampling from one target π, ones samples from a sequence of related targets

$$\pi_0, \pi_1, \ldots, \pi_n = \pi$$

- This is done by applying importance sampling in a sequential manner

- For instance, we can take
  1. π_k as the posterior of θ given the first k observations
  2. Or $\pi_k(x) \propto \pi(x)^{\phi_k}, \quad 0 \leq \phi_0 < \phi_1 < \ldots \phi_n = 1$

     In this case, π_0 is close to a uniform distribution, and we can use rejection sampling to simulate from π_0

Basic idea sequential Monte Carlo

- Assume that at step k, $X^{k,t}$, $t = 1, \ldots, N$, is a sample from π_k
- The sample $(X^{k,t})$ is sequentially modified to obtain a sample $(X^{n,t})$ from the original target π
- As in importance sampling, one can generate weighted samples or equally weighted ones by applying a resampling step. We first focus on the later case

Sequential Monte Carlo

- Assume that we can efficiently generate an **initial sample ($X^{0,t}$) from π_0**, e.g., using rejection or importance sampling

- **Propagation step**: At step k,assume that $(X^{k-1,t})$ is a sample from π_{k-1} and propagate this sample using a transition kernel p_k:

$$Y^{k,t} \sim p_k(X^{k-1,t}, y)dy, \quad \text{independently for } t = 1, 2, \ldots N$$

We then have

$$Y^{k,t} \sim \int \pi_{k-1}(x)p_k(x, y)dx \cdot dy$$

- **Importance sampling step**: $Y^{k,t}$ does not have the correct density π_k but we can correct for this by applying importance sampling with weights

$$w^{k,t} \propto \frac{\pi_k(Y^{k,t})}{\int \pi_{k-1}(x)p_k(x, Y^{k,t})dx}.$$

This results in a sample $(X^{k,t})$ from π_k

Sequential Monte Carlo algorithm with resampling

- **Problem**: $\int \pi_{k-1}(x) p_k(x, y) dx$ is not available analytically
- **Solution**: replace the importance sampling step by

$$X^{k,t} = Y^{k,I^t}, \quad \mathbb{P}(I^t = s) \propto \frac{\pi_k(Y^{k,s}) q_{k-1}(Y^{k,s}, X^{k-1,s})}{\pi_{k-1}(X^{k-1,t}) p_k(X^{k-1,t}, Y^{k,t})}$$

  - q_{k-1} is an arbitrary **auxiliary backward transition kernel**

- $(X^{k,t})$ is a sample from π_k

  See blackboard for derivation

Sequential Monte Carlo algorithm: weighted version

- If one is willing to use **weighted samples** at all stages, a resampling step is not needed
- Simply update the weights $w^{k,t}$

$$w^{k,t} \propto w^{k-1,t} \frac{\pi_k(X^{k,t}) q_{k-1}(X^{k,t}, X^{k-1,t})}{\pi_{k-1}(X^{k-1,t}) p_k(X^{k-1,t}, X^{k,t})}$$

- **Drawback**: this sequential multiplication leads very quickly to unbalanced weights
- Resampling helps to concentrate the computing effort in those region of the space where the densities π_k have their main mass

Choice of transition kernel p_k

Potential **choices** for the **propagation kernel p_k** include:

- Independent moves: $p_k(x, y) = p_k(y)$
- Random walk moves: $p_k(x, y)$ proposes samples that are symmetric around the mean x. E.g., Gaussian random walk
- MCMC moves: set $p_k(x, y)$ as the density of an MCMC kernel with π_k as invariant distribution

Choice of transition kernel q_k

- One can show that optimal **choice** (weights have minimal variance) for the auxiliary **backward transition kernel q_k** is

$$q_k(y, x) = \frac{\nu_{k-1}(x) p_k(x, y)}{\nu_k(y)},$$

where

$$\nu_k(y) = \int \pi_{k-1}(x) p_k(x, y) dx$$

which, however, is not available

- One (of several) potential choices for the kernel q_k is

$$q_k(y, x) = \frac{\pi_k(x) p_k(x, y)}{\pi_k(y)}$$

Approximate Bayesian computation

Approximate Bayesian computation

- For some models, **evaluating the likelihood $f(x \mid \theta)$ is complicated or even impossible** $\Rightarrow$ MCMC cannot be used to sample from the posterior

- Often, simulating

$$X \sim f(x \mid \theta)dx$$

is much easier and we can therefore generate pairs

$$(\theta^t, X^t) \sim \pi(\theta)f(x \mid \theta)d\theta dx$$

Approximate Bayesian computation: discrete X

- If ***X* is discrete**, we can use **rejection sampling** to simulate from the density proportional to

$$\pi(\theta) f(x \mid \theta) 1_{[x = x_{obs}]}$$

whose marginal is the posterior. I.e., we simply accept only pairs (θ^t, X^t) such that $X^t = x_{obs}$

Approximate Bayesian computation: continuous X

- For **continuous X**, one can **use the same idea** and replace the point mass at x_{obs} by a **distribution which is concentrated near x_{obs}**

$$\pi(\theta) f(x \mid \theta) \exp(-d(x, x_{obs})/\varepsilon)$$

where d is a metric on the space of observations

- Instead of working with a fixed ε, one can also choose a sequence $\varepsilon_n \to 0$ with a rather large ε_0 and use a sequential Monte Carlo algorithm to produce samples of the corresponding targets

Introduction to Gaussian processes

Gaussian process

- A **Gaussian process** is a collection of random variables $\{Z(s); s \in D \subset \mathbb{R}^d\}$ for which any finite-dimensional distribution is Gaussian. It is specified by

- A **mean function**

$$\begin{aligned} m\colon D &\longrightarrow \mathbb{R} \\ s &\longmapsto m(s) \end{aligned}$$

- A **covariance function**

$$\begin{aligned} m\colon D \times D &\longrightarrow \mathbb{R} \\ (s, s') &\longmapsto C(s, s') \end{aligned}$$

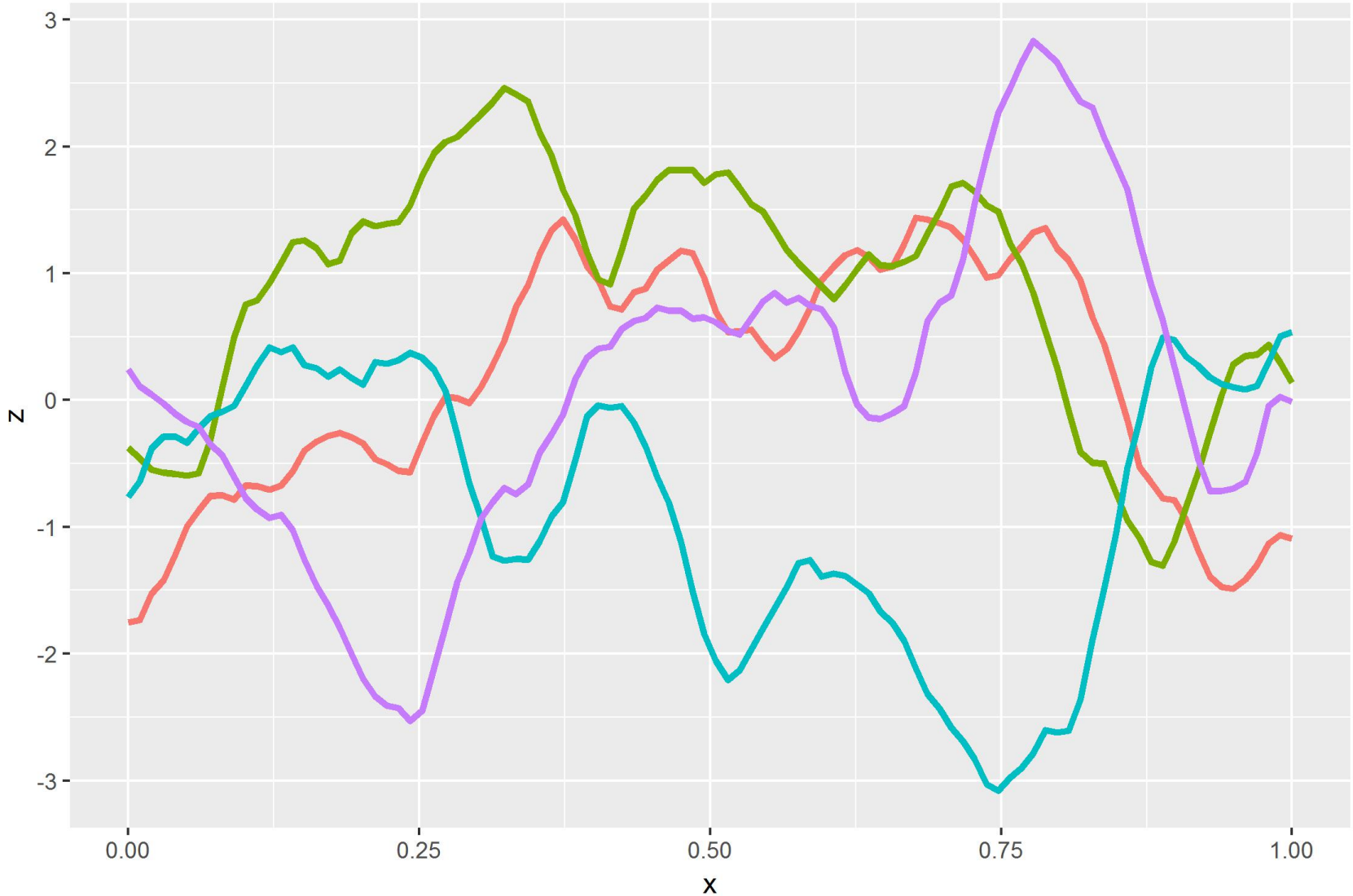
where C must be symmetric and positive definite:

$$C(s, s') = C(s', s) \quad \forall s, s' \in D$$

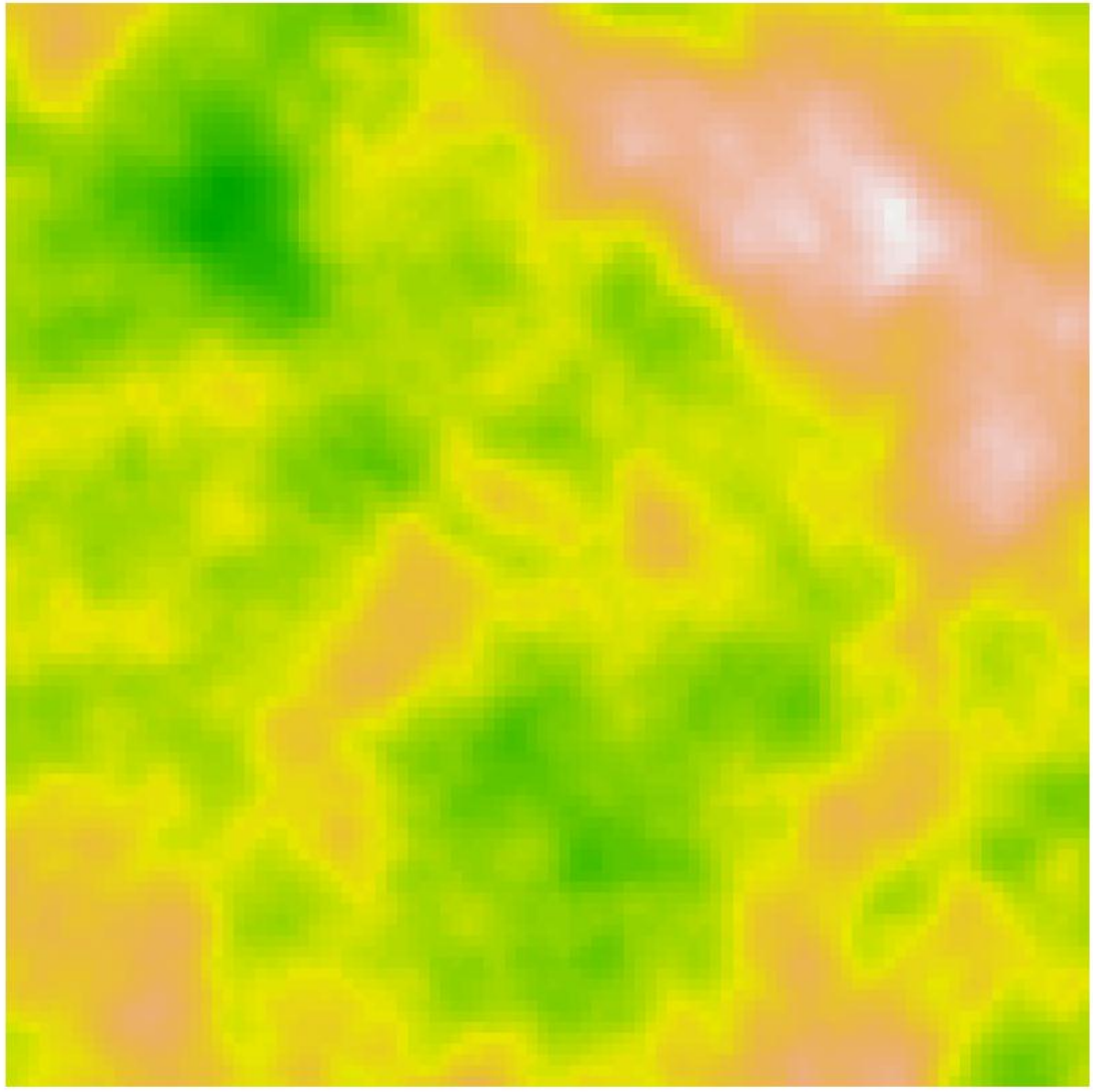
and

$$\sum_{i,j=1}^{n} C(s_i, s_j)\beta_i\beta_j \geq 0 \quad \forall s_i \in D, n \in \mathbb{N}, \beta_i \in \mathbb{R}$$

Example: Samples from a 1D Gaussian process



Example: A sample from a 2D Gaussian process



Example of a Gaussian process model

- Assume that $\boldsymbol{Z(s)}, \boldsymbol{s \in [0,1]}$, **follows a Gaussian process** with

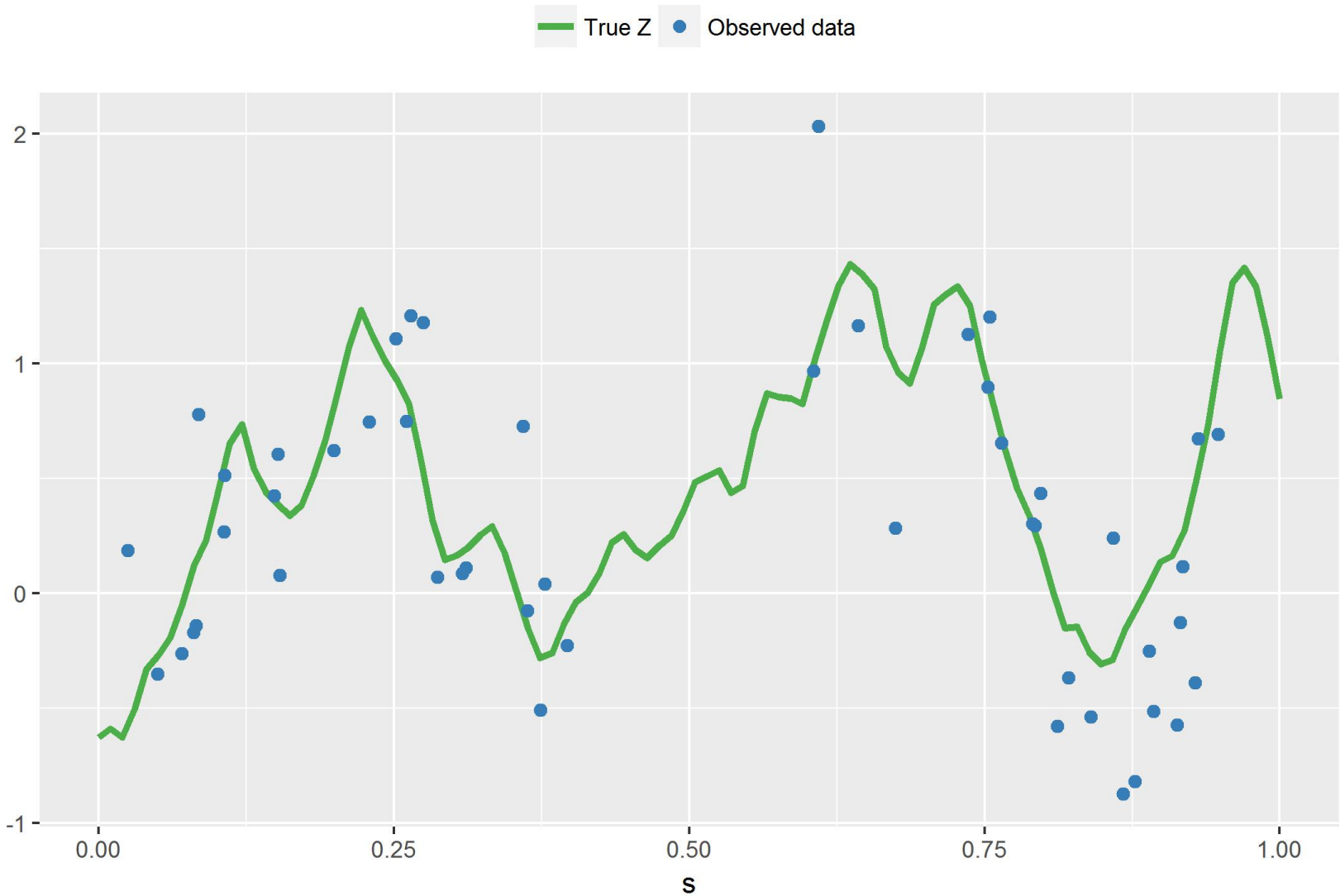
$$m(s) = 0 \text{ and } C(s, s') = \sigma^2 \left(1 + \sqrt{3}\frac{|s - s'|}{\rho}\right) \exp\left(-\frac{|s - s'|}{\rho}\right) *$$

- We **observe data** $y = (y(s_1), \ldots, y(s_n))$ at n location s_i

$$Y(s_i) = Z(s_i) + \varepsilon(s_i), \quad i = 1, \ldots, n, \quad \varepsilon(s_i) \text{ i.i.d. } \sim N(0, \sigma_\varepsilon^2)$$

*This is a so-called Matérn covariance function with smoothness parameter $\nu = 1.5$. Other common covariance function choices include the exponential covariance $C(s, s') = \sigma^2 \exp(-\frac{|s-s'|}{\rho})$ and the Gaussian covariance $C(s, s') = \sigma^2 \exp(-\frac{|s-s'|^2}{2\rho^2})$

Sample from a Gaussian process $Z(s)$ and observed data y



Bayesian inference for Gaussian processes

- In order to complete the specification of a Bayesian model, we assign a prior $\pi(\theta)$ to the parameters $\theta = (\sigma^2, \rho, \sigma_\varepsilon^2)$

- We then consider the following tasks:
  1. Determine the **posterior $\pi(\theta \mid y)$**

  2. Determine the **joint posterior $\pi(\theta, Z \mid y)$**, where $Z = (Z(s_1), \ldots, Z(s_n))$

  3. Determine the **posterior predictive distribution $\pi(Z^{pred} \mid y)$**, where $Z^{pred} = (Z(s_1^{pred}), \ldots, Z(s_{n'}^{pred}))$

     The latter is referred to as **Gaussian process regression** or **kriging**

See R examples for how to do this in RStan

For more details: Gelfand, A. E., Diggle, P., Guttorp, P., & Fuentes, M. (2010). Handbook of spatial statistics. CRC press

Illustration posterior predictive distribution (PPD) for Z

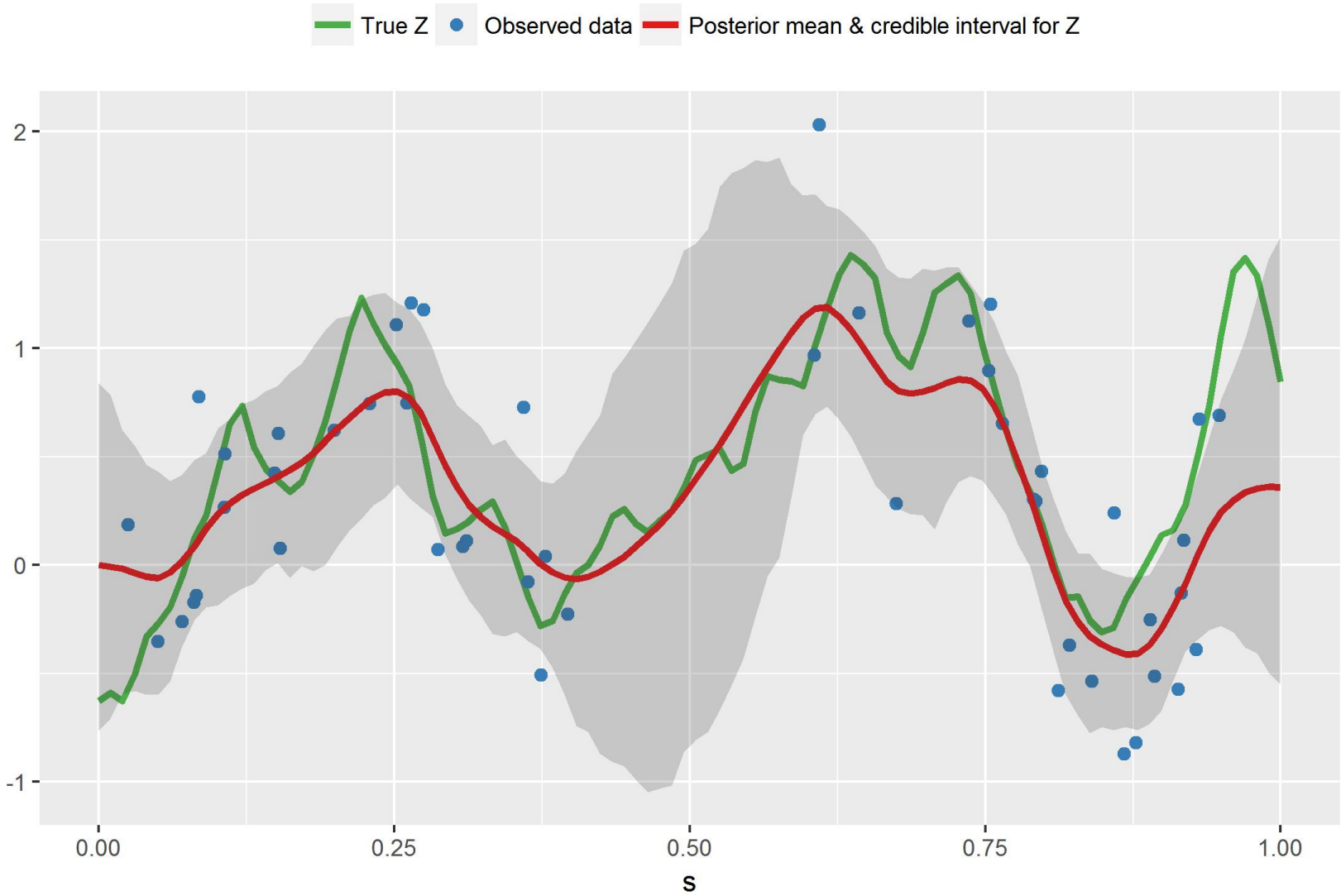


Illustration posterior predictive distribution (PPD) for Z