

---

# Gaussian Processes with Probabilistic Programming

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

Abstract

## 1 Introduction

MCMC lends itself to Bayesian interpretations of Gaussian Processes since they can provide a vehicle to express otherwise intractable integrals necessary for a fully Bayesian representation.

## 2 Gaussian Processes

In the following, we will introduce GP related theory and notations. We will exclusively work on two variable regression problems. Let the data be real-valued scalars  $\{x_i, y_i\}_{i=1}^n$  (complete data will be denoted by column vectors  $\mathbf{x}, \mathbf{y}$ ). GPs present a non-parametric way to express prior knowledge on the space of possible functions  $f$  that we assume to have generated the data.  $f$  is assumed latent and the GP prior is given by a multivariate Gaussian with mean and covariance  $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ , where  $m(\mathbf{x})$  is a function of the mean of all functions that map to  $y_i$  at  $x_i$  and  $k(\mathbf{x}, \mathbf{x}')$  is a kernel or covariance function that summarizes the covariance of all functions that map to  $y_i$  at  $x_i$ . We can absorb the mean function into the covariance function so without loss of generality we can set the mean to zero. The marginal likelihood can be expressed as:

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{x}) p(\mathbf{f}|\mathbf{x}) d\mathbf{f} \quad (1)$$

where the prior is Gaussian  $\mathbf{f}|\mathbf{x} \sim \mathcal{N}(0, k(\mathbf{x}, \mathbf{x}'))$ . For a zero mean Gaussian Process this results in a Gaussian posterior  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  with mean:

$$\boldsymbol{\mu} = \mathbf{K}(\mathbf{x}, \mathbf{x}^*) \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*)^{-1} \mathbf{y} \quad (2)$$

and covariance

$$\boldsymbol{\Sigma} = \mathbf{K}(\mathbf{x}, \mathbf{x}) + \mathbf{K}(\mathbf{x}, \mathbf{x}^*) \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*)^{-1} \mathbf{K}(\mathbf{x}^*, \mathbf{x}). \quad (3)$$

where  $\mathbf{K}$  is a covariance function. The covariance function covers general high-level properties of the observed data such as linearity, periodicity and smoothness. The most widely used type of covariance function is the squared exponential covariance function:

$$k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right) \quad (4)$$

where  $\sigma$  and  $\ell$  are hyper-parameters.  $\sigma$  is a scaling factor and  $\ell$  is the typical length-scale. Smaller variations can be achieved by exchanging these hyper-parameters. Below, we see how we can express simple GP smoothing with a few

lines of Venture code while allowing users to custom design covariance functions.

Listing 1: GP Smoothing

```
[ASSUME l 1]
[ASSUME sf 2]

 $k(x, x') := \sigma^2 \exp(-\frac{(x-x')^2}{2\ell^2})$ 

[ASSUME f VentureFunction(k, σ, ℓ) ]
[ASSUME SE make-se (apply-function f l sf) ]
[ASSUME (make-gp 0 SE ) ]
```

In the case where hyper-parameters are unknown they can be found deterministically by optimizing the marginal likelihood using a gradient based optimizer. Non-deterministic, Bayesian representations of this case are also known (Neal, 1997). Extending the program described in listing 1 for a Bayesian treatment of hyper-parameters is simple using the build in stochastic procedure that simulates drawing samples from a gamma distribution:

Listing 2: Bayesian GP Smoothing

```
[ASSUME l (gamma 1 3)]
[ASSUME sf (gamma 1 2)]

 $k(x, x') := \sigma^2 \exp(-\frac{(x-x')^2}{2\ell^2})$ 

[ASSUME f VentureFunction(k, σ, ℓ) ]
[ASSUME SE make-se (apply-function f l sf) ]
[ASSUME (make-gp 0 SE ) ]
```

Larger variations are achieved by changing the type of the covariance function structure. A different type could be a linear covariance function:

$$k(x, x') = \sigma^2(x - \ell)(x' - \ell). \quad (5)$$

Note that covariance function structures are compositional. We can add covariance functions if we want to model globally valid structures

$$k_3(x, x') = k_1(x, x') + k_2(x, x') \quad (6)$$

and we can multiply covariance functions if the data is best explained by local structure

$$k_4(x, x') = k_1(x, x') \times k_2(x, x'); \quad (7)$$

both,  $k_3$  and  $k_4$  are valid covariance function structures. This leads to an infinite space of possible structures that could potentially explain the observed data best (e.g. Fig. 4.1). In the following, we will refer to covariance functions that are not composite as base covariance functions. Note that this form of composition can be easily expressed in Venture, for example if one wishes to add a linear and a periodic kernel:

### Listing 3: LIN $\times$ PER

```

[ASSUME l (gamma 1 3)]
[ASSUME sf (gamma 1 2)]
[ASSUME a (gamma 2 2)]

 $k_{LIN}(x, x') = \sigma_1^2 \exp(-\frac{(x-x')^2}{2\ell^2})$ 

 $k_{PER}(x, x') := \sigma_2^2 \exp(-\frac{2 \sin^2(\pi(x-x')/p)}{\ell^2})$ 

[ASSUME fLIN VentureFunction( $k_{LIN}, \sigma_1$ ) ]
[ASSUME fPER VentureFunction( $k_{PER}, \sigma_2, \ell, p$ ) ]
[ASSUME LIN (make-LIN (apply-function fLIN a)) ]
[ASSUME PER (make-PER (apply-function fPER l sf)) ]
[ASSUME (make-gp 0 (function-times LIN PER)) ]

```

Knowledge about the composite nature of covariance functions is not new, however, until recently, the choice and the composition of covariance functions were done ad-hoc. The Automated Statistician Project came up with an approximate search over the possible space of kernel structures (Duvinaud et al., 2013; Lloyd et al., 2014).

## 2.1 A Bayesian interpretation

In the following, we will explore a Bayesian representation of GP. The probability of the hyper-parameters of a GP with assumptions as above and given covariance function structure  $\mathbf{K}$  can be described as:

$$P(\boldsymbol{\theta} \mid \mathbf{D}, \mathbf{K}) = \frac{P(\mathbf{D} \mid \boldsymbol{\theta}, \mathbf{K}) P(\boldsymbol{\theta} \mid \mathbf{K})}{P(\mathbf{D} \mid \mathbf{K})}. \quad (8)$$

We are interested in the case where covariance structure is not given. Our probabilistic programming based MCMC framework approximates the following intractable integrals of the expectation for the prediction:

$$\mathbb{E}[y^* \mid x^*, D, \mathbf{K}_\Omega^s] = \iint f(x^*, \boldsymbol{\theta}, \mathbf{K}) P(\boldsymbol{\theta} \mid \mathbf{D}, \mathbf{K}) P(\mathbf{K} \mid \Omega, s, n) d\boldsymbol{\theta} d\mathbf{K}. \quad (9)$$

This is done by sampling from the posterior probability distribution of the hyper-parameters and the possible kernel:

$$y^* \approx \frac{1}{T} \sum_{t=1}^T f(x^* \mid \boldsymbol{\theta}^{(t)}, \mathbf{K}^{(t)}). \quad (10)$$

## 3 Stochastic Processes

In order to provide the sampling of the kernel, we introduce a stochastic process to the SP that simulates the grammar for algebraic expressions of kernel algebra. Here, we start with a set of possible kernels and draw a random subset. For this subset of size  $n$ , we sample a set of possible operators that operate on the base kernels.

The marginal probability of a kernel structure which allows us to sample is characterized by the probability of a uniformly chosen subset of the set of  $n$  possible covariance functions times the probability of sampling a global or a local structure which is given by a binomial distribution:

$$P(\mathbf{K} \mid \Omega, s, n) = P(\Omega \mid s, n) \times P(s \mid n) \times P(n), \quad (11)$$

with

$$P(\Omega \mid s, n) = \binom{n}{r} p_{+\times}^k (1 - p_{+\times})^{n-k} \quad (12)$$

and

$$P(s \mid n) = \frac{n!}{s!} \quad (13)$$

where  $P(n)$  is a prior on the number of base kernels used. It is possible to also assign a prior for the probability to sample global or local priors, however, we have assigned complete uncertainty to this with the binomial  $p = 0.5$ .

## 4 Experiments

### 4.1 Structure Learning

We defined a set of covariance structures so that we could reproduce results for covariance function structure learning as in the Automated Statistician. Our results are very similar to what has been reported by previous work Duvenaud et al. (2013).

### 4.2 Log-Likelihood

### 4.3 Residuals

## References

- Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J., and Ghahramani, Z. (2013). Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1166–1174.
- Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J., and Ghahramani, Z. (2014). Automatic construction and natural-language description of nonparametric regression models. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.
- Neal, R. M. (1997). Monte carlo implementation of gaussian process models for bayesian regression and classification. *arXiv preprint physics/9701026*.

216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269

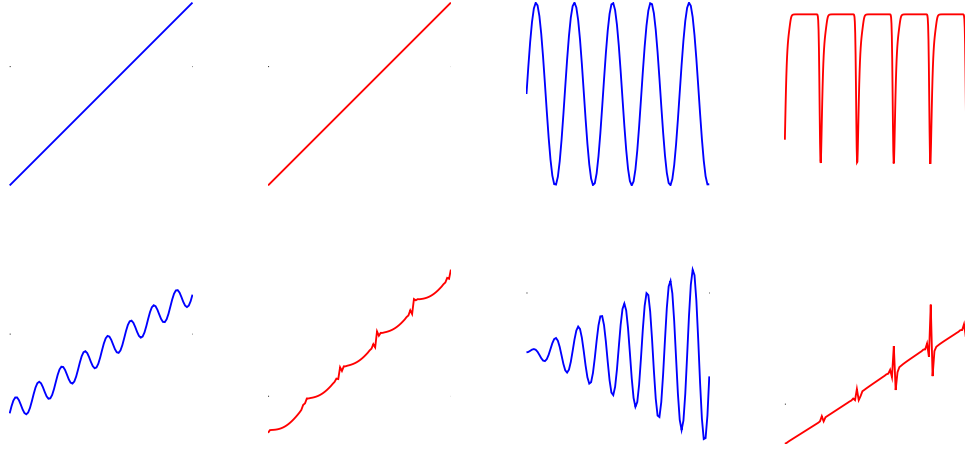


Figure 1: Composition of covariance functions (blue, left) and samples from the distribution of curves they can produce (red, right).

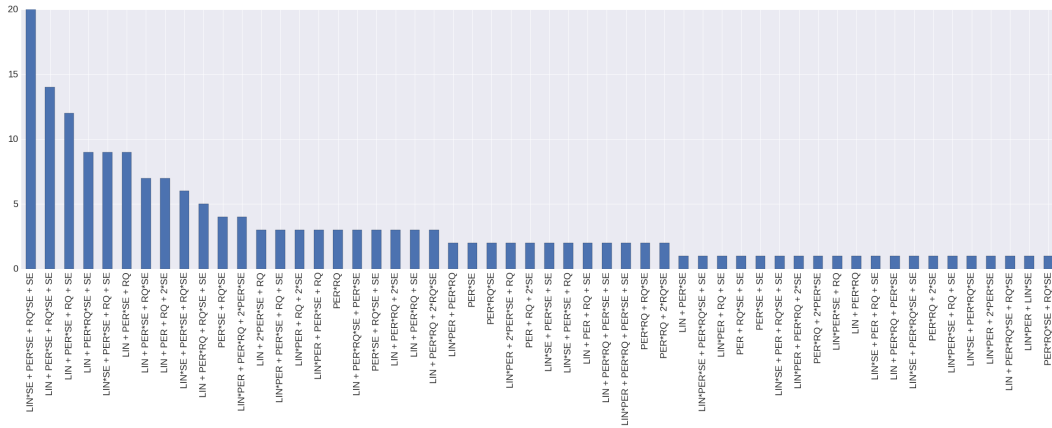


Figure 2: Preliminary results from the cross-validation on the CO2 data. Note that Duvenaud et al. (2013) report  $\text{LIN} \times \text{SE} + \text{PER} \times \text{SE} + \text{RQ} \times \text{SE}$ . We have run a leave one out cross-validation on this data set. Above we see the preliminary results on 181 validations (of a total of  $545 \times 2$  runs).