Gaussian Processes for Particle Physicists

Mikael Kuusela

Department of Statistics and Data Science,
Carnegie Mellon University

PHYSTAT Tutorial

July 20, 2022

Outline

1. Definition and basic properties

2. Mean functions, covariance functions and parameter estimation

3. Applications in particle physics

Outline

1. Definition and basic properties

2. Mean functions, covariance functions and parameter estimation

3. Applications in particle physics

Introduction

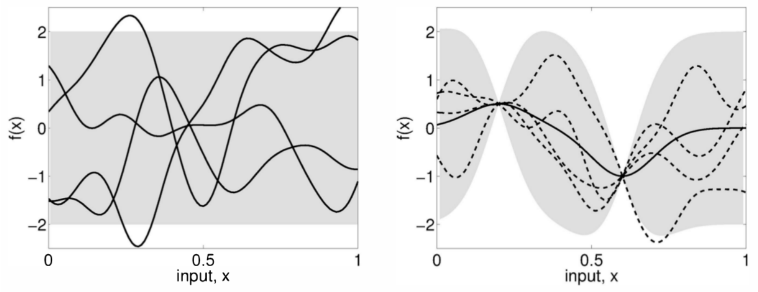


Figure from Rasmussen and Williams (2006)

Gaussian processes are a versatile class of statistical models for *random functions*.

They enable learning from data in situations involving random or unknown functions.

Gaussian processes are popular because 1) they provide a plausible model for various real-world phenomena, 2) they provide *useful* inferences, and 3) they are relatively easy to work with.

Multivariate Gaussian distribution

A random vector $\boldsymbol{y} \in \mathbb{R}^n$ has an *n*-variate Gaussian distribution, denoted by $\boldsymbol{y} \sim N(\boldsymbol{m}, \boldsymbol{\Sigma})$, if its pdf is given by

$$p(\boldsymbol{y}|\boldsymbol{m}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{m})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{m})\right)$$

This is parameterized by the *mean vector* $\boldsymbol{m} \in \mathbb{R}^n$ and the symmetric and positive definite *covariance matrix* $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ so that

$$\begin{aligned} \mathbb{E}[y_i] &= m_i \quad && \text{for all } i = 1, \ldots, n \\ \mathrm{Cov}[y_i, y_j] &= \Sigma_{ij} \quad && \text{for all } i, j = 1, \ldots, n \end{aligned}$$

Multivariate Gaussian distribution

Multivariate Gaussian random vectors have a number of nice properties.

For example, consider the decomposition

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}, \quad \boldsymbol{m} = \begin{bmatrix} \boldsymbol{m}_1 \\ \boldsymbol{m}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

Then the marginal distribution of $\boldsymbol{y}_1$ is

$$\boldsymbol{y}_1 \sim N(\boldsymbol{m}_1, \boldsymbol{\Sigma}_{11})$$

and the conditional distribution of $\boldsymbol{y}_1$ given $\boldsymbol{y}_2$ is

$$(\boldsymbol{y}_1 | \boldsymbol{y}_2) \sim N(\boldsymbol{m}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{m}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$$

By rearranging the elements of $\boldsymbol{y}$, we can have any subset of elements in the component $\boldsymbol{y}_1$ and the remaining elements in the component $\boldsymbol{y}_2$. In other words:

- Any subset of elements of $\boldsymbol{y}$ have a multivariate Gaussian distribution
- Any subset of elements of $\boldsymbol{y}$ conditioned on the rest have a multivariate Gaussian distribution

Gaussian processes: Definition

Now, imagine that n is very large. We then have a large collection of random variables

$$\{y_1, y_2, \ldots, y_{n-1}, y_n\} = \{y_i\}_{i=1}^{n},$$

indexed by the discrete index i, whose joint behavior is described by the multivariate Gaussian distribution.

A Gaussian process is an infinite-dimensional generalization of this to a collection of random variables indexed on a continuum.

Definition

A *Gaussian process* is a random function $f(\boldsymbol{x})$ whose values $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)$ at any finite set of inputs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ follow a multivariate Gaussian distribution.

Gaussian processes: Definition

Definition

A *Gaussian process* is a random function $f(\boldsymbol{x})$ whose values $f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n)$ at any finite set of inputs $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ follow a multivariate Gaussian distribution.

A Gaussian process is parameterized by the *mean function* $m(\boldsymbol{x})$ and the *covariance function* $k(\boldsymbol{x}_1, \boldsymbol{x}_2)$ so that

$$\begin{aligned} m(\boldsymbol{x}) &= \mathbb{E}[f(\boldsymbol{x})], \quad \text{for all } \boldsymbol{x} \\ k(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \mathrm{Cov}[f(\boldsymbol{x}_1), f(\boldsymbol{x}_2)], \quad \text{for all } \boldsymbol{x}_1, \boldsymbol{x}_2. \end{aligned}$$

We then denote $f \sim GP(m(\boldsymbol{x}), k(\boldsymbol{x}_1, \boldsymbol{x}_2))$.

The covariance function $k(\boldsymbol{x}_1, \boldsymbol{x}_2)$ has to be such that the covariance matrix of $[f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n)]^\mathsf{T}$ for any inputs $\boldsymbol{x}_i$, $i = 1, \dots, n$, is positive definite.

Functions with this property are called *positive definite*. There are various well-known families of positive definite functions, but it's good to keep in mind that not all bivariate functions are valid covariance functions.

Gaussian processes: Inference

Let $f \sim GP(m(\boldsymbol{x}), k(\boldsymbol{x}_1, \boldsymbol{x}_2))$ and assume that we get to observe

$$y_1 = f(\boldsymbol{x}_1), y_2 = f(\boldsymbol{x}_2), \ldots, y_n = f(\boldsymbol{x}_n).$$

What can we then say about $y_* = f(\boldsymbol{x}_*)$ at some unobserved location $\boldsymbol{x}_*$?

Gaussian processes: Inference

Let $f \sim GP(m(\boldsymbol{x}), k(\boldsymbol{x}_1, \boldsymbol{x}_2))$ and assume that we get to observe

$$y_1 = f(\boldsymbol{x}_1), y_2 = f(\boldsymbol{x}_2), \ldots, y_n = f(\boldsymbol{x}_n).$$

What can we then say about $y_* = f(\boldsymbol{x}_*)$ at some unobserved location $\boldsymbol{x}_*$?

Since y_* is a random quantity, statisticians call this *prediction* of y_* (as opposed to *estimation* of a fixed parameter).

Denote $\boldsymbol{y}_n = [y_1, \ldots, y_n]^{\mathsf{T}}$. Then, by definition:

$$\begin{bmatrix} y_* \\ \boldsymbol{y}_n \end{bmatrix} = \begin{bmatrix} y_* \\ y_1 \\ \vdots \\ y_n \end{bmatrix} \sim N(\boldsymbol{m}, \boldsymbol{\Sigma}), \quad \text{where} \quad \boldsymbol{m} = \begin{bmatrix} m(\boldsymbol{x}_*) \\ m(\boldsymbol{x}_1) \\ \vdots \\ m(\boldsymbol{x}_n) \end{bmatrix} = \begin{bmatrix} m(\boldsymbol{x}_*) \\ \boldsymbol{m}_n \end{bmatrix}$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} k(\boldsymbol{x}_*, \boldsymbol{x}_*) & k(\boldsymbol{x}_*, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_*, \boldsymbol{x}_n) \\ k(\boldsymbol{x}_1, \boldsymbol{x}_*) & k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_*) & k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix} = \begin{bmatrix} k(\boldsymbol{x}_*, \boldsymbol{x}_*) & \boldsymbol{k}_*^{\mathsf{T}} \\ \boldsymbol{k}_* & \boldsymbol{K}_n \end{bmatrix}$$

Gaussian processes: Inference

Then, by the properties of the multivariate Gaussian distribution, the conditional distribution of y_* given $\boldsymbol{y}_n$ is

$$(y_* | \boldsymbol{y}_n) \sim N(m(\boldsymbol{x}_*) + \boldsymbol{k}_*^\mathsf{T} \boldsymbol{K}_n^{-1}(\boldsymbol{y}_n - \boldsymbol{m}_n), k(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^\mathsf{T} \boldsymbol{K}_n^{-1} \boldsymbol{k}_*)$$

Since we are trying to predict y_* given $\boldsymbol{y}_n$, this is also known as the *predictive distribution* of y_*. We can directly extract from this the predictive mean

$$\hat{y}_* = \mathbb{E}[y_* | \boldsymbol{y}_n] = m(\boldsymbol{x}_*) + \boldsymbol{k}_*^\mathsf{T} \boldsymbol{K}_n^{-1}(\boldsymbol{y}_n - \boldsymbol{m}_n)$$

and the predictive variance

$$\hat{\sigma}_*^2 = \mathrm{Var}[y_* | \boldsymbol{y}_n] = k(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^\mathsf{T} \boldsymbol{K}_n^{-1} \boldsymbol{k}_*.$$

We can then predict y_* using $\hat{y}_*$. A standard result from statistical learning theory says that this is the mean squared error optimal predictor of y_*.

The $(1 - \alpha)$ predictive uncertainty is given by $[\hat{y}_* - z_{1-\alpha/2}\hat{\sigma}_*, \hat{y}_* + z_{1-\alpha/2}\hat{\sigma}_*]$, which has correct coverage assuming that the model is correct.

Gaussian processes: Inference

As a result, we conclude that y_* should be predicted using

$$\hat{y}_* = m(\boldsymbol{x}_*) + \boldsymbol{k}_*^\mathsf{T} \boldsymbol{K}_n^{-1}(\boldsymbol{y}_n - \boldsymbol{m}_n)$$

and the uncertainty of the prediction at level $(1-\alpha)$ is given by

$$[\hat{y}_* - z_{1-\alpha/2}\hat{\sigma}_*, \hat{y}_* + z_{1-\alpha/2}\hat{\sigma}_*].$$

This has various names depending on the context, including *kriging* (spatial statistics / geostatistics), *objective mapping* (oceanography) or *optimal interpolation* (atmospheric science).

Gaussian processes: Inference

Notice also that we can repeat the same calculation for other $\boldsymbol{x}_*$'s to obtain pointwise predictions of $f(\boldsymbol{x})$ on a fine grid, for example.

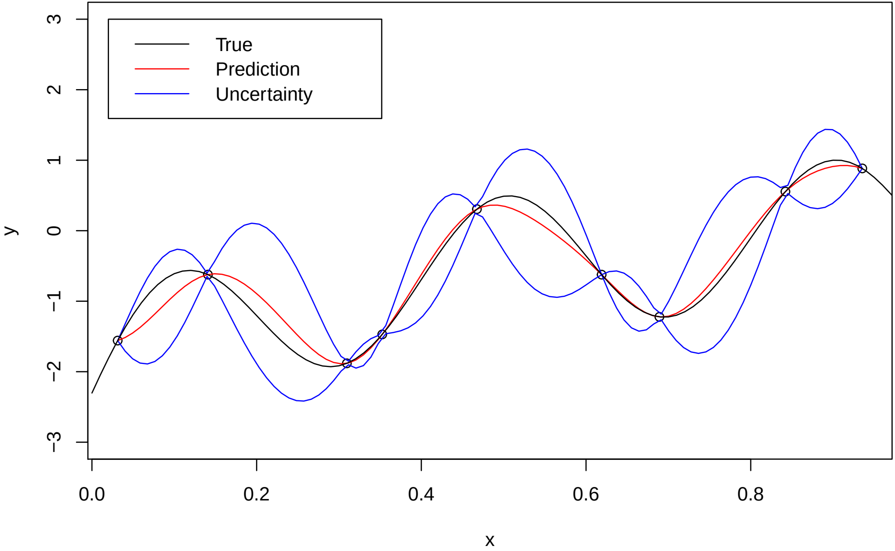
We can also repeat the calculation for the vector

$$[y_{1*}, \ldots, y_{p*}, y_1, \ldots, y_n]^\mathsf{T} = [f(\boldsymbol{x}_{1*}), \ldots, f(\boldsymbol{x}_{p*}), f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]^\mathsf{T}$$

to obtain the predictive distribution of $y_{1*}, \ldots, y_{p*}$ given $y_1, \ldots, y_n$, which also provides us the predictive covariance between different locations $\boldsymbol{x}_{i*}$.

Key observation: Because finite evaluations of a Gaussian process follow a multivariate Gaussian distribution, we immediately know how to make a finite number of predictions given a finite number of observations.

Illustration



Gaussian process regression

In practice, we do not necessarily want to force the function f to go through the observations $y_1, \ldots, y_n$.

Therefore, the following *Gaussian process regression* model is commonly employed:

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i,$$

where $f \sim GP(m(\boldsymbol{x}), k(\boldsymbol{x}_1, \boldsymbol{x}_2))$, $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ and f is independent of the ε_i's.

The extra term ε_i is called the *nugget effect* and corresponds to measurement error, unexplained variation or microscale variation, depending on the context.

One might then be interested in predicting either $f_* = f(\boldsymbol{x}_*)$ or $y_* = f(\boldsymbol{x}_*) + \varepsilon_*$

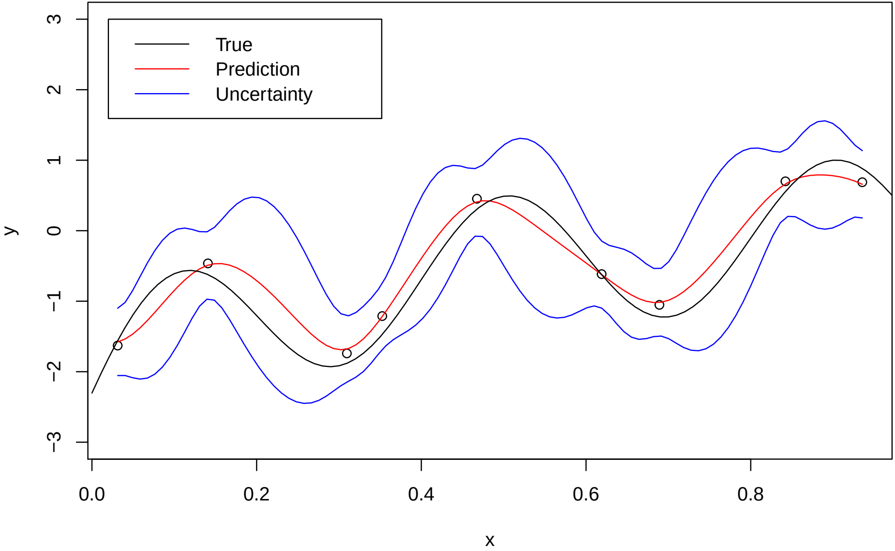
The predictive distribution in the first case is

$$(f_*|\boldsymbol{y}_n) \sim N(m(\boldsymbol{x}_*)+\boldsymbol{k}_*^\mathsf{T}(\boldsymbol{K}_n+\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y}_n-\boldsymbol{m}_n), k(\boldsymbol{x}_*, \boldsymbol{x}_*)-\boldsymbol{k}_*^\mathsf{T}(\boldsymbol{K}_n+\sigma^2\boldsymbol{I})^{-1}\boldsymbol{k}_*)$$

The latter case is otherwise the same but the predictive variance is

$$\mathrm{Var}[y_*|\boldsymbol{y}_n] = \mathrm{Var}[f_*|\boldsymbol{y}_n] + \sigma^2 = k(\boldsymbol{x}_*, \boldsymbol{x}_*)+\sigma^2-\boldsymbol{k}_*^\mathsf{T}(\boldsymbol{K}_n+\sigma^2\boldsymbol{I})^{-1}\boldsymbol{k}_*$$

Illustration



Outline

1. Definition and basic properties
2. Mean functions, covariance functions and parameter estimation
3. Applications in particle physics

Gaussian process modeling

A Gaussian process $f \sim GP(m(\boldsymbol{x}), k(\boldsymbol{x}_1, \boldsymbol{x}_2))$ is parameterized by the *mean function* $m(\boldsymbol{x})$ and the *covariance function* $k(\boldsymbol{x}_1, \boldsymbol{x}_2)$

In order to model data using a GP, one therefore needs to decide how to choose these functions.

A significant portion of GP literature revolves around this question.

There is some ambiguity with regards to what portion of the data should be explained using $m(\boldsymbol{x})$ and what portion using $k(\boldsymbol{x}_1, \boldsymbol{x}_2)$, especially if there is only a single realization of f

- "One person's mean structure is another person's covariance structure"

Some authors claim that one can simply set $m(\boldsymbol{x}) = 0$ without loss of generality, but it's not quite that simple

In practice, we tend to use certain parametric classes of functions for both:

$$m(\boldsymbol{x}) = m(\boldsymbol{x}; \boldsymbol{\beta}), \quad k(\boldsymbol{x}_1, \boldsymbol{x}_2) = k(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta})$$

Choice of the mean function

The mean function $m(\boldsymbol{x})$ should be flexible enough to model the average shape of the random function $f(\boldsymbol{x})$, but also rigid enough to not fit the stochastic high-frequency fluctuations that might be present in the data

It might sound like it is difficult to strike a balance here, but luckily the final predictions are usually quite robust against modest misspecification of the mean

Common choices for $m(\boldsymbol{x}; \boldsymbol{\beta})$:

- Linear in $\boldsymbol{x}$ and $\boldsymbol{\beta}$: $m(\boldsymbol{x}; \boldsymbol{\beta}) = \boldsymbol{x}^{\mathsf{T}} \boldsymbol{\beta}$
- Splines (especially in 1D): $m(x; \boldsymbol{\beta}) = \sum_{i=1}^{p} \beta_i B_i(x)$, where $B_i(\cdot)$ are B-spline basis functions
- Nonlinear (in both $\boldsymbol{x}$ and $\boldsymbol{\beta}$) regression functions (e.g., neural nets)

Choice of the covariance function

Recall that $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathrm{Cov}[f(\boldsymbol{x}_1), f(\boldsymbol{x}_2)]$.

Which bivariate function $k(\cdot, \cdot)$ to use? (Remember that $k(\cdot, \cdot)$ needs to be positive definite.)

A common assumption is to say that $k(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is *stationary* (i.e., translation invariant): $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = k(\boldsymbol{x}_1 - \boldsymbol{x}_2)$

Furthermore, it is common to assume *isotropy*

$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = k(\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|)$$

or *geometric anisotropy*

$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = k(\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_{\boldsymbol{A}}),$$

where $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_{\boldsymbol{A}} = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\mathsf{T}} \boldsymbol{A} (\boldsymbol{x}_1 - \boldsymbol{x}_2)}$ for a positive definite matrix $\boldsymbol{A}$

Choice of the covariance function

Let's focus on the case with geometric anisotropy. Denote $s = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_{\boldsymbol{A}}$.

At this point, we need to choose the matrix $\boldsymbol{A}$ and the function $k(s)$.

Here $\boldsymbol{A}$ controls the length scales and orientation of the dependence in $f(\boldsymbol{x})$ over $\boldsymbol{x}$.

The function $k(s)$ controls the remaining properties of the random field $f(\boldsymbol{x})$, such as smoothness, periodicity, etc.

Choice of the covariance function

Popular models for $k(s)$ include:

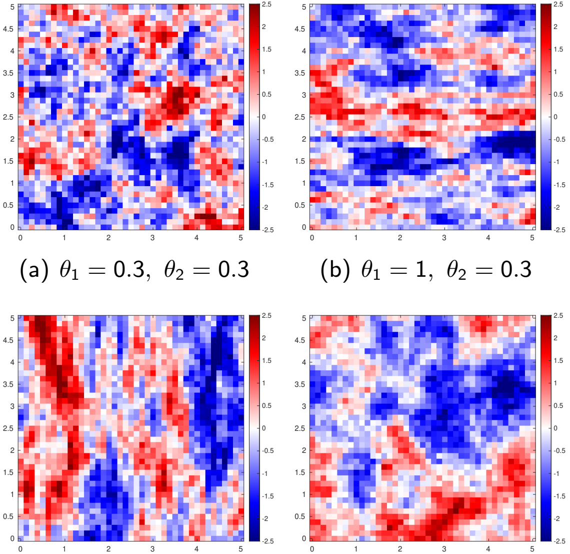
- Exponential: $k(s) = \phi \exp(-s), \ \phi > 0$
  - $f(\boldsymbol{x})$ continuous but not differentiable
- Squared exponential: $k(s) = \phi \exp(-s^2), \ \phi > 0$
  - $f(\boldsymbol{x})$ infinitely differentiable
- Matérn: $k(s) = \phi \frac{2^{1-\nu}}{\Gamma(\nu)} s^\nu K_\nu(s), \ \phi > 0$, where $\nu > 0$ is a smoothness parameter and K_ν is a modified Bessel function
  - $f(\boldsymbol{x})$ k times differentiable if and only if $\nu > k$
  - Gives exponential for $\nu = \frac{1}{2}$ and squared exponential for $\nu \to \infty$
  - Has simplified form when ν is half integer, i.e., $\nu = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots$

For example, if we pick $\boldsymbol{A} = \operatorname{diag}(1/\theta_1^2, \ldots, 1/\theta_d^2)$ and let $k(s)$ be exponential, then we have the following covariance model

$$
\begin{aligned}
&k(\boldsymbol{x}_1, \boldsymbol{x}_2; \phi, \theta_1, \ldots, \theta_d) \\
&\quad = \phi \exp\left(-\sqrt{\left(\frac{x_{11} - x_{21}}{\theta_1}\right)^2 + \left(\frac{x_{12} - x_{22}}{\theta_2}\right)^2 + \cdots + \left(\frac{x_{1d} - x_{2d}}{\theta_d}\right)^2} \right)
\end{aligned}
$$

parameterized by $\phi, \theta_1, \ldots, \theta_d > 0$

Illustration: Effect of covariance length scales



(a) $\theta_1 = 0.3,\ \theta_2 = 0.3$

(b) $\theta_1 = 1,\ \theta_2 = 0.3$

(c) $\theta_1 = 0.3,\ \theta_2 = 1$

(d) $\theta_1 = 1,\ \theta_2 = 1$

Parameter estimation

Let $\boldsymbol{\theta}$ denote the vector of covariance parameters that affect the data-data covariance $\boldsymbol{K}_n$ so that $\boldsymbol{K}_n(\boldsymbol{\theta})$

Then the unknown parameters of the model are $(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)$ and we wish to learn these parameters using the observed data $\boldsymbol{y}_n$

Various techniques for estimating these parameters exist, but the most common approach is to use maximum likelihood.

Since $\boldsymbol{y}_n$ follows a multivariate Gaussian, the log-likelihood of $(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)$ is

$$\begin{aligned}\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) &= \log p(\boldsymbol{y}_n | \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) \\ &= -\frac{1}{2}\Big[n \log(2\pi) + \log\det\left(\boldsymbol{K}_n(\boldsymbol{\theta}) + \sigma^2 \boldsymbol{I}\right) \\ &\qquad + (\boldsymbol{y}_n - \boldsymbol{m}_n(\boldsymbol{\beta}))^{\mathsf{T}} (\boldsymbol{K}_n(\boldsymbol{\theta}) + \sigma^2 \boldsymbol{I})^{-1} (\boldsymbol{y}_n - \boldsymbol{m}_n(\boldsymbol{\beta}))\Big]\end{aligned}$$

The estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2)$ are those values that maximize $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)$

For linear mean functions, $\boldsymbol{\beta}$ can be solved in closed-form (for given $(\boldsymbol{\theta}, \sigma^2)$), but to solve $(\boldsymbol{\theta}, \sigma^2)$ one needs to typically use numerical optimization

Outline

1 Definition and basic properties

2 Mean functions, covariance functions and parameter estimation

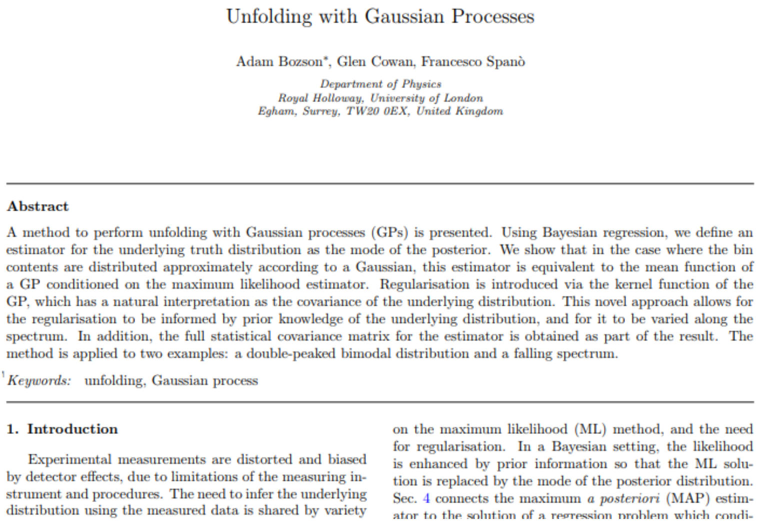
3 Applications in particle physics

Gaussian processes in particle physics

Some uses of Gaussian processes in HEP:

- Bayesian prior for an unknown function f
- Modeling of background shapes
- Bayesian optimization
- Emulators/surrogates for computationally intensive simulations
- ...

Gaussian processes and unfolding



Unfolding with Gaussian Processes

Adam Bozson*, Glen Cowan, Francesco Spanò

Department of Physics
Royal Holloway, University of London
Egham, Surrey, TW20 0EX, United Kingdom

Abstract

A method to perform unfolding with Gaussian processes (GPs) is presented. Using Bayesian regression, we define an estimator for the underlying truth distribution as the mode of the posterior. We show that in the case where the bin contents are distributed approximately according to a Gaussian, this estimator is equivalent to the mean function of a GP conditioned on the maximum likelihood estimator. Regularisation is introduced via the kernel function of the GP, which has a natural interpretation as the covariance of the underlying distribution. This novel approach allows for the regularisation to be informed by prior knowledge of the underlying distribution, and for it to be varied along the spectrum. In addition, the full statistical covariance matrix for the estimator is obtained as part of the result. The method is applied to two examples: a double-peaked bimodal distribution and a falling spectrum.

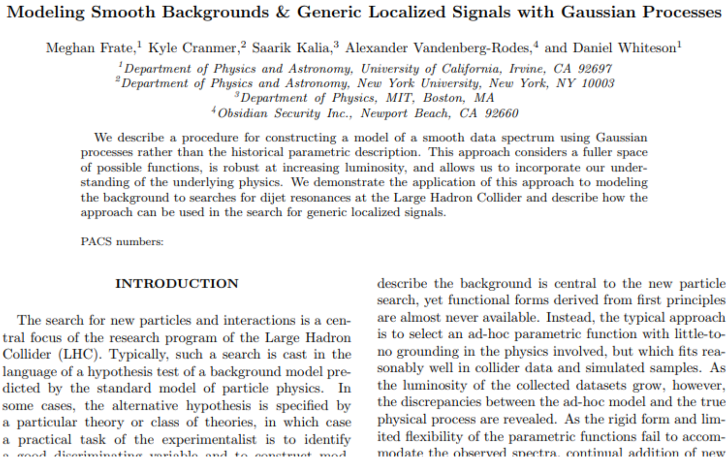
Keywords: unfolding, Gaussian process

1. Introduction

Experimental measurements are distorted and biased by detector effects, due to limitations of the measuring instrument and procedures. The need to infer the underlying distribution using the measured data is shared by variety on the maximum likelihood (ML) method, and the need for regularisation. In a Bayesian setting, the likelihood is enhanced by prior information so that the ML solution is replaced by the mode of the posterior distribution. Sec. 4 connects the maximum *a posteriori* (MAP) estimator to the solution of a regression problem which condi-

[arXiv:1811.01242]

Gaussian processes for background modeling



Modeling Smooth Backgrounds & Generic Localized Signals with Gaussian Processes

Meghan Frate,[1] Kyle Cranmer,[2] Saarik Kalia,[3] Alexander Vandenberg-Rodes,[4] and Daniel Whiteson[1]

[1]Department of Physics and Astronomy, University of California, Irvine, CA 92697
[2]Department of Physics and Astronomy, New York University, New York, NY 10003
[3]Department of Physics, MIT, Boston, MA
[4]Obsidian Security Inc., Newport Beach, CA 92660

We describe a procedure for constructing a model of a smooth data spectrum using Gaussian processes rather than the historical parametric description. This approach considers a fuller space of possible functions, is robust at increasing luminosity, and allows us to incorporate our understanding of the underlying physics. We demonstrate the application of this approach to modeling the background to searches for dijet resonances at the Large Hadron Collider and describe how the approach can be used in the search for generic localized signals.

PACS numbers:

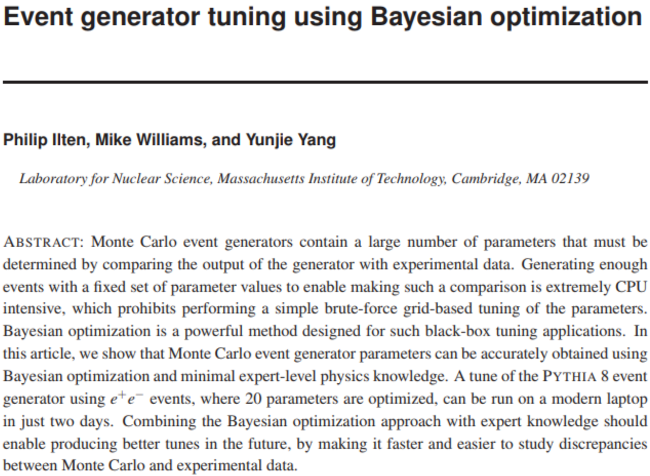
INTRODUCTION

The search for new particles and interactions is a central focus of the research program of the Large Hadron Collider (LHC). Typically, such a search is cast in the language of a hypothesis test of a background model predicted by the standard model of particle physics. In some cases, the alternative hypothesis is specified by a particular theory or class of theories, in which case a practical task of the experimentalist is to identify a good discriminating variable and to construct mod-

describe the background is central to the new particle search, yet functional forms derived from first principles are almost never available. Instead, the typical approach is to select an ad-hoc parametric function with little-to-no grounding in the physics involved, but which fits reasonably well in collider data and simulated samples. As the luminosity of the collected datasets grow, however, the discrepancies between the ad-hoc model and the true physical process are revealed. As the rigid form and limited flexibility of the parametric functions fail to accommodate the observed spectra, continual addition of new

[arXiv:1709.05681]

Bayesian optimization in HEP



Event generator tuning using Bayesian optimization

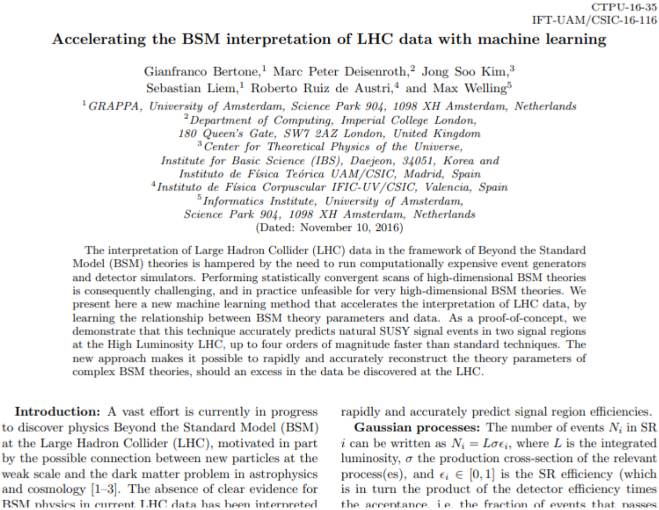
Philip Ilten, Mike Williams, and Yunjie Yang

Laboratory for Nuclear Science, Massachusetts Institute of Technology, Cambridge, MA 02139

ABSTRACT: Monte Carlo event generators contain a large number of parameters that must be determined by comparing the output of the generator with experimental data. Generating enough events with a fixed set of parameter values to enable making such a comparison is extremely CPU intensive, which prohibits performing a simple brute-force grid-based tuning of the parameters. Bayesian optimization is a powerful method designed for such black-box tuning applications. In this article, we show that Monte Carlo event generator parameters can be accurately obtained using Bayesian optimization and minimal expert-level physics knowledge. A tune of the PYTHIA 8 event generator using e^+e^- events, where 20 parameters are optimized, can be run on a modern laptop in just two days. Combining the Bayesian optimization approach with expert knowledge should enable producing better tunes in the future, by making it faster and easier to study discrepancies between Monte Carlo and experimental data.

[arXiv:1610.08328]

Surrogate models using Gaussian processes



CTPU-16-35
IFT-UAM/CSIC-16-116

Accelerating the BSM interpretation of LHC data with machine learning

Gianfranco Bertone,[1] Marc Peter Deisenroth,[2] Jong Soo Kim,[3]
Sebastian Liem,[1] Roberto Ruiz de Austri,[4] and Max Welling[5]

[1] *GRAPPA, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, Netherlands*
[2] *Department of Computing, Imperial College London,*
180 Queen's Gate, SW7 2AZ London, United Kingdom
[3] *Center for Theoretical Physics of the Universe,*
Institute for Basic Science (IBS), Daejeon, 34051, Korea and
Instituto de Física Teórica UAM/CSIC, Madrid, Spain
[4] *Instituto de Física Corpuscular IFIC-UV/CSIC, Valencia, Spain*
[5] *Informatics Institute, University of Amsterdam,*
Science Park 904, 1098 XH Amsterdam, Netherlands

(Dated: November 10, 2016)

The interpretation of Large Hadron Collider (LHC) data in the framework of Beyond the Standard Model (BSM) theories is hampered by the need to run computationally expensive event generators and detector simulators. Performing statistically convergent scans of high-dimensional BSM theories is consequently challenging, and in practice unfeasible for very high-dimensional BSM theories. We present here a new machine learning method that accelerates the interpretation of LHC data, by learning the relationship between BSM theory parameters and data. As a proof-of-concept, we demonstrate that this technique accurately predicts natural SUSY signal events in two signal regions at the High Luminosity LHC, up to four orders of magnitude faster than standard techniques. The new approach makes it possible to rapidly and accurately reconstruct the theory parameters of complex BSM theories, should an excess in the data be discovered at the LHC.

Introduction: A vast effort is currently in progress to discover physics Beyond the Standard Model (BSM) at the Large Hadron Collider (LHC), motivated in part by the possible connection between new particles at the weak scale and the dark matter problem in astrophysics and cosmology [1–3]. The absence of clear evidence for BSM physics in current LHC data has been interpreted

rapidly and accurately predict signal region efficiencies.

Gaussian processes: The number of events N_i in SR i can be written as $N_i = L\sigma\epsilon_i$, where L is the integrated luminosity, σ the production cross-section of the relevant process(es), and $\epsilon_i \in [0, 1]$ is the SR efficiency (which is in turn the product of the detector efficiency times the acceptance, i.e. the fraction of events that passes

[arXiv:1611.02704]

Additional reading

The following textbooks are good starting points for learning more:

- C.E. Rasmussen and C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006
- M.L. Stein, Interpolation of spatial data: Some theory for kriging, Springer, 1999
- N.A.C. Cressie, Statistics for spatial data, Revised edition, John Wiley & Sons, 1993
- C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006
- J. Mockus, Bayesian Approach to Global Optimization: Theory and Applications, Kluwer, 1989

Backup

Gaussian processes in Earth science

In Earth sciences, the following Gaussian process model is often used for interpolating atmospheric or oceanographic observations:

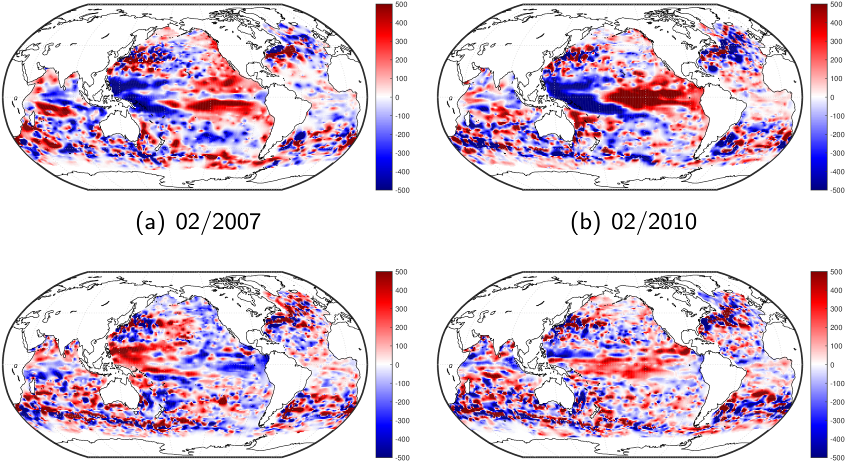
$$
\begin{aligned}
y_{i,j} &= f_i(x_{\mathrm{lat},i,j}, x_{\mathrm{lon},i,j}, t_{i,j}) + \varepsilon_{i,j}, \\
f_i &\overset{\mathrm{iid}}{\sim} \mathrm{GP}(m, k), \quad \varepsilon_{i,j} \overset{\mathrm{iid}}{\sim} N(0, \sigma^2),
\end{aligned}
$$

where

- $y_{i,j}$ is some observed quantity (for example, temperature, humidity, CO_2 concentration,...)
- $i = 1, \ldots, n$ refers to years and $j = 1, \ldots, m_i$ to observations in the ith year
- $x_{\mathrm{lat},i,j}$, $x_{\mathrm{lon},i,j}$ and $t_{i,j}$ are the latitude, longitude and time of $y_{i,j}$

Key point: This is a fully frequentist model. It is quite sensible to model the year-to-year variations in these fields as a Gaussian process.

Upper ocean heat content anomalies



(a) 02/2007

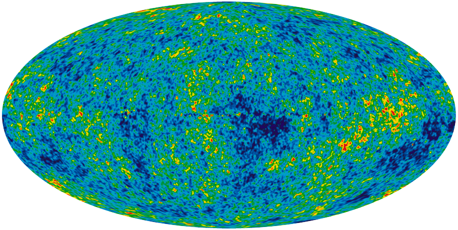
(b) 02/2010

(c) 02/2013

(d) 02/2015

Monthly ocean heat content anomalies interpolated from *in situ* oceanographic float data using locally stationary Gaussian processes

Gaussian processes in cosmology



Cosmic microwave background temperature fluctuations from WMAP

Standard cosmological models imply that CMB is a Gaussian random field (i.e., a Gaussian process with 2 input dimensions)

Observational evidence of non-Gaussianity would have important implications for theories of the early Universe

Key point: Here we have a function that by physical arguments is known to be a Gaussian process. Hypothetically one can imagine observing multiple realizations of this random function (in practice there is of course just a single realization).