JHEP

Published for SISSA by Springer

Received: *February 24, 2022*
Revised: *December 14, 2022*
Accepted: *February 6, 2023*
Published: *February 23, 2023*

Model selection and signal extraction using Gaussian Process regression

Abhijith Gandrakota,[a] **Amit Lath,**[b] **Alexandre V. Morozov**[b] **and Sindhu Murthy**[c]

[a] *Fermi National Accelerator Laboratory, Batavia, IL, U.S.A.*
[b] *Department of Physics & Astronomy, Rutgers, The State University of New Jersey, 136 Frelinghuysen Rd., Piscataway, NJ, U.S.A.*
[c] *Department of Physics, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, U.S.A.*

E-mail: abhijith@fnal.gov, lath@physics.rutgers.edu, morozov@physics.rutgers.edu, sindhum@andrew.cmu.edu

Abstract: We present a novel computational approach for extracting localized signals from smooth background distributions. We focus on datasets that can be naturally presented as binned integer counts, demonstrating our procedure on the CERN open dataset with the Higgs boson signature, from the ATLAS collaboration at the Large Hadron Collider. Our approach is based on Gaussian Process (GP) regression — a powerful and flexible machine learning technique which has allowed us to model the background without specifying its functional form explicitly and separately measure the background and signal contributions in a robust and reproducible manner. Unlike functional fits, our GP-regression-based approach does not need to be constantly updated as more data becomes available. We discuss how to select the GP kernel type, considering trade-offs between kernel complexity and its ability to capture the features of the background distribution. We show that our GP framework can be used to detect the Higgs boson resonance in the data with more statistical significance than a polynomial fit specifically tailored to the dataset. Finally, we use Markov Chain Monte Carlo (MCMC) sampling to confirm the statistical significance of the extracted Higgs signature.

Keywords: Hadron-Hadron Scattering , Beyond Standard Model, Higgs Physics, Unfolding

ArXiv ePrint: 2202.05856

Open Access, © The Authors.
Article funded by SCOAP[3].

https://doi.org/10.1007/JHEP02(2023)230

JHEP02(2023)230

Contents

1 Introduction **1**

2 Model selection and signal extraction procedure **3**
- 2.1 Gaussian process regression 3
- 2.2 Model selection 4
- 2.3 The Poisson likelihood 6
- 2.4 Gaussian process kernels 6
- 2.5 Functional fit 7

3 Datasets **7**

4 Model selection for background-only fits **8**

5 Signal extraction **9**
- 5.1 Synthetic datasets for testing statistical significance of signal extraction 11
- 5.2 Test for systematic biases in signal extraction 11
- 5.3 Posterior distributions of signal parameters and significance analysis 14

6 Summary **16**

JHEP02(2023)230

1 Introduction

Analyzing data from physical experiments or observations often involves fitting computational models with the goal of extracting a signal in the presence of both background effects and random noise. For example, such a setting appears naturally in the analysis of X-ray crystalline sample diffraction patterns containing contributions from both from distinctive Bragg peaks and diffuse background scattering [1, 2], inference of transiting exoplanet parameters in astronomy [3, 4], and the discovery of the Higgs boson and search for the new physics at the Large Hadron Collider (LHC) at CERN [5]. The data from LHC and similar experiments usually comes in the form of binned integer counts [6]. Traditionally, modeling binned integer-valued data is performed under the assumption of the Poisson distribution, by employing a parametric fit [6]. The fitted models are subsequently used to estimate the background contributions and extract the signal of interest. However, the choice of parametric functions is often ad-hoc and the degree of model complexity requires a delicate balancing act between overfitting and underfitting the data [7–9].

For data analyses of the type performed on the LHC bin counts, the optimal complexity of the model is usually evaluated by performing likelihood-ratio tests [10, 11]. Most analyses that employ this technique test it on a fraction of the full dataset, usually 10% or less. This

process is called "blinding" and is meant to reduce biases in the analysis. However, the model selection process must be repeated periodically as more data becomes available, and often the functional form employed in the model has to be updated as well [12]. Using non-parametric methods such as Gaussian Process regression is an effective way of alleviating these concerns [7, 13–16].

Gaussian Process (GP) regression is a well-established machine learning technique [7, 14] commonly used in diverse fields such as astrophysics, gravitational wave detection, and high energy physics [17–20]. In particular, in high energy physics GP regression was used to model the smooth continuum background from quantum chromodynamics (QCD) in searches for dijet resonances in LHC data [20]. The authors argued that using GP for background estimation was more robust with respect to increasing luminosity compared to parametric fitting methods. GP regression's advantages over more conventional methods, which employ a linear expansion over a fixed set of basis functions such as polynomials or Gaussians, are due to its non-parametric flexibility. Instead of explicit basis functions, GP regression is defined in terms of kernel functions which specify the degree of correlation between two points in the dataset. The GP approach allows us to perform inference using a much broader class of functions, including those which would otherwise require an infinite basis set [7]. GP regression is robust with respect to the size of the dataset [20].

Nevertheless, the flexibility of GP regression can be a double-edged sword. In GP regression, kernel functions typically depend on several hyperparameters that are varied to fit the data, typically through non-Bayesian techniques such as maximizing the marginal likelihood of the observed data [7, 14]. The type of the kernel function and the values of its hyperparameters determine the success in capturing certain features in the data, such as periodic oscillations and long-term trends [14]. Thus, the universality and the power of the GP approach are only fully realized when both the kernel type and kernel hyperparameter values are chosen judiciously. In other words, a method is required that can constrain the flexibility of the GP regression in a controlled manner.

Previous work in this area has focused on "kernel learning" to address the issues of flexibility and robustness, with several techniques proposed that aim at constructing composite kernels for Support Vector Machines [21], Relevance Vector Machines [22], and GP regression [23, 24] using libraries of base kernels. Semiparametric regression attempts to combine interpretability of parametric models with flexibility of non-parametric models by combining the two approaches in a single framework [25]. However, none of the above approaches focus specifically on integer count data or on the processes that are naturally viewed as localized signals superimposed on a smooth background distribution. A previous application of GP regression to LHC data [20] employed both standard and custom-built kernels motivated by physical considerations. In contrast, in this work we have developed both a model selection procedure suitable for GP regression and a principled approach for estimating the statistical significance of the extracted signal.

New signals in physical observations of particle resonances in LHC data often appear as localized features ("bumps") superimposed on a smooth background. Accurate modeling of the background spectrum is therefore essential to both extracting the signal and assessing its statistical significance. In this paper, we present a rigorous approach to model selection

JHEP02(2023)230

in GP regression applied to binned integer data, which we expect to be a superposition of a localized signal and a smooth background of an unknown functional form. We exploit the flexibility of the GP regression by determining the kernel hyperparameters through the fit to background-only data, with the signal window masked out. These parameters are subsequently used to extrapolate the background contribution across the signal window, enabling us to separately measure the background and signal contributions.

We describe procedures for kernel type selection based on both Bayesian and Akaike information criteria. We also propose a method for estimating the statistical significance of the signal by performing a hypothesis test with the background model as the null hypothesis and the background+signal model as the alternative hypothesis. We illustrate our procedure by successfully detecting the Higgs boson resonance in the di-photon Higgs sample from the ATLAS experiment at the LHC [26]. We show that, compared to parametric fits, GP regression leads to the Higgs boson signature with higher statistical significance. Our computational pipeline can be applied for background estimation and signal detection in any dataset where a weak localized signal is superimposed on a smooth background distribution with an unknown functional form.

2 Model selection and signal extraction procedure

2.1 Gaussian process regression

In regression, the observed data $y = (y_1 \dots y_N)$ is modeled by $z = (z(x_1) \dots z(x_N))$:

$$y_k = z(x_k) + \epsilon_k, \tag{2.1}$$

where $k = 1 \dots N$, $X = (x_1 \dots x_N)$ is a vector of input variables, and ϵ_k is a random noise variable independently sampled from a Gaussian distribution $\mathcal{N}(\epsilon|0, \sigma_i^2)$ for each data point, where σ_i^2 is the noise variance in bin i. In other words, $p(y|z) = \mathcal{N}(y|z, \Sigma)$, where Σ is a diagonal $N \times N$ matrix with $\Sigma_{ii} = \sigma_i^2$. Since in our case the observed data y is given by integer counts in N bins and X is given by the centers of the bins with the integer counts, GP model predictions have to be rounded. Alternatively, GP real-valued predictions can be used as Poisson or multinomial rates to generate integer event counts.

In the GP framework, the model $z(x)$ is not represented as an explicit linear expansion over a set of pre-determined basis functions $\phi(x)$ such as polynomials. Instead, the dependence of z on X is given probabilistically by a multi-dimensional Gaussian distribution: $p(z|X) = \mathcal{N}(z|m, K)$, where $m = (m(x_1) \dots m(x_N))$ is a vector of values of the mean function $m(x)$ for all datapoints and K is the Gram matrix. The elements of the Gram matrix are values of the kernel function $k(x, x')$ evaluated for all pairs of input variables: $K_{ij} = k(x_i, x_j)$. The marginal likelihood $p(y|X)$, integrated over all possible models, is then given by [7, 27]:

$$p(y|X) = \int dz\ p(y|z)p(z|X) = \mathcal{N}(y|m, C), \tag{2.2}$$

where the covariance matrix $C = K + \Sigma$.

JHEP02(2023)230

Thus, GP regression is defined by the mean function $m(x)$ and the kernel function $k(x, x')$ which determines the degree of systematic correlation between any two datapoints [7, 14]. In general, the kernel function depends on a set of n hyperparameters $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$, whose number and meaning depend on the kernel type. The hyperparameters of a given kernel are usually optimized by maximizing the marginal likelihood in eq. (2.2), a non-Bayesian procedure [7, 14]. Ordinarily, kernel hyperparameters would also include σ_i^2, which represent the amount of experimental noise in each bin. However, since this would introduce too many hyperparameters, we estimate σ_i^2 directly from the data using Garwood intervals, which allow us to extract two-sided confidence intervals from the number of events in each bin, under the assumption that the events are Poisson-distributed [28, 29]. Note that the Garwood intervals converge to $[y_i - \sqrt{y_i}, y_i + \sqrt{y_i}]$ when $y_i \gg 1$, yielding $\sigma_i \simeq \sqrt{y_i}$ as expected from the Poisson statistics; in case of slightly asymmetric intervals, we have chosen the larger value as an estimate of σ_i. Thus, σ_i^2 are estimated independently and are not treated as hyperparameters in our approach.

The strength of the GP approach stems from the fact that the joint marginal probability of observing a set of datapoints is Gaussian (eq. (2.2)). Moreover, the predictive probability $p(\tilde{y}_i|y)$, the conditional probability distribution of observing a real-valued "count" $\tilde{y}_i$ in bin i given a dataset with N previous observations, is also Gaussian, with the mean $f(x_i)$ and the variance $V(x_i)$ given by:

$$\begin{aligned} f(x_i) &= m(x_i) + \tilde{k}^T C^{-1} \tilde{y}_i, \\ V(x_i) &= \alpha - \tilde{k}^T C^{-1} \tilde{k}, \end{aligned} \tag{2.3}$$

where $\tilde{k} = (k(x_1, x_i), \ldots, k(x_N, x_i))$ and $\alpha = k(x_i, x_i) + \sigma_i^2$.

In this work, we consider two types of GP regression: with $m(x_i) = 0$, $\forall i$ for modeling the background-only distribution, and with the Gaussian mean function (the signal model function) for modeling signal+background datasets, where the signal component is represented by:

$$m(x_i) = \frac{A}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right). \tag{2.4}$$

Here, A defines the signal strength, while μ and σ represent signal mean and width, respectively. The rounded value of A can be interpreted as the total number of signal events. Note that when the Gaussian mean function is introduced, the set of model hyperparameters θ needs to be augmented with $\{A, \mu, \sigma\}$. Throughout this work, we use the words "prediction" and "predicted" to signify the output of computational models fitted to either experimentally observed or synthetically generated datasets, rather than *ab initio* theoretical predictions made before any data is seen. We hope that this usage, which is widespread in the machine learning literature, does not create any confusion since we do not discuss fundamental theoretical predictions (e.g., those based on the Standard Model) here.

2.2 Model selection

A key issue in GP regression is the choice of a kernel and the derivation of the optimal set of hyperparameters $\hat{\theta}$ for it. Typically, the optimal set of hyperparameters is obtained by

JHEP02(2023)230

maximizing the marginal log-likelihood $\log p(y|X,\theta,K_i)$ [7, 14], where $p(y|X,\theta,K_i)$ is given by eq. (2.2) and its dependence on the set of hyperparameters θ and the kernel type K_i are made explicit for clarity. Since this step is non-Bayesian, a question of kernel selection arises, which needs to take into account kernel complexity, quantified by both the amount of smoothing provided by a given kernel and the number of kernel hyperparameters. A standard way for carrying out the model comparison is based on the Bayesian Information Criterion (BIC) for the marginal log-likelihood, which provides an approximation to the model evidence used in Bayesian model selection [7, 30]:

$$-\log p(y|X,K_i) \simeq -\log p(y|X,\hat{\theta},K_i) + \frac{n}{2}\log N \equiv \mathrm{BIC}^{\mathrm{naive}}, \tag{2.5}$$

where $p(y|X,K_i)$ is the model evidence (the product of the marginal likelihood and the hyperparameter priors integrated over θ), n is the number of model parameters, and N is the number of datapoints. Note that the second term on the right-hand side penalizes model complexity, such that lower BIC scores are more preferable. The derivation of BIC relies on a number of approximations whose validity depends on the details of the system under consideration. Specifically, the derivation employs the Laplace approximation to estimate the integral over the hyperparameters and assumes that N is so large (or the Gaussian prior distribution over the hyperparameters is so broad) that the effects of the hyperparameter priors are negligible, resulting in:

$$-\log p(y|X,K_i) \simeq -\log p(y|X,\hat{\theta},K_i) + \frac{1}{2}\log|H| \equiv \mathrm{BIC}, \tag{2.6}$$

where $H = -\vec{\nabla}_\theta\vec{\nabla}_\theta \log p(y|X,\theta,K_i)|_{\hat{\theta}}$ is the Hessian in the model hyperparameter space evaluated at the hyperparameter values $\hat{\theta}$ that maximize the marginal log-likelihood. If N is large and the Hessian has full rank, the second term on the right-hand side can be roughly approximated as $\frac{1}{2}\log|H| \simeq \frac{n}{2}\log N$, yielding eq. (2.5).

An alternative, non-Bayesian approach to model selection is based on the Akaike Information Criterion (AIC), which accounts for the fact that the log-likelihood computed on a training dataset provides an estimate of the prediction error that is too optimistic, because the same data is being used to fit the model and assess its error [31, 32]. To account for this optimism, a correction term is added which is based on the sum of covariances between the observed datapoint and the newly generated datapoint for each input variable x_i. It can be shown that the sum of covariances is proportional to the number of degrees of freedom in the $N \to \infty$ limit, resulting in the following expression for AIC:

$$\mathrm{AIC} \equiv -2\log p(y|X,\hat{\theta},K_i) + 2d, \tag{2.7}$$

where d is the number of degrees of freedom in the model. Thus AIC provides an estimate of the log-likelihood that would have resulted if another dataset were independently generated at the same values of input variables (an in-sample estimate). In the case of GP regression, d is replaced by d_{eff} in eq. (2.7), where d_{eff} is the effective number of degrees of freedom for the GP regression with a given kernel type, which captures the amount of smoothing induced by the GP fit [14, 33]:

$$d_{\mathit{eff}}(\hat{\theta}) = \mathrm{tr}[K(\hat{\theta})(K(\hat{\theta}) + \Sigma)^{-1}], \tag{2.8}$$

JHEP02(2023)230

where the dependence of the Gram matrix on the optimal kernel hyperparameters $\hat{\theta}$ is made explicit for clarity. Note that similar to BIC, lower AIC values are preferable.

Choosing the appropriate kernel type is crucial to the success of GP regression, since different kernels emphasize different correlation structures in the data. In practice, kernels are often handcrafted manually using simple comparison metrics such as $\mathrm{BIC}^{\mathrm{naive}}$. In some cases, composite kernels are constructed automatically using kernel engineering techniques (see e.g. ref. [23]). Here, we propose a kernel selection technique which is based on the consensus between AIC and BIC measures of model complexity. This framework allows us to choose a specific kernel type on the basis of the consensus between Bayesian and non-Bayesian criteria.

2.3 The Poisson likelihood

Since our data consists of integer counts in N bins, we have also employed a Poisson likelihood model to generate integer predictions in each bin. Specifically, we assume that the mean of the predictive probability $f(x_i)$ (eq. (2.3)) provides the rate for the Poisson process in each bin [20]:

$$\log L_{\mathcal{P}} = \sum_{i=1}^{N} \left[y_i - f(x_i) - y_i \log\left(\frac{y_i}{f(x_i)}\right)\right]. \tag{2.9}$$

Note that $f(x_i)$ implicitly depends on the kernel type and the optimized hyperparameter values $\hat{\theta}$. Eq. (2.9) can be used both to generate integer counts and compute the log-likelihood of the observed counts. The Poisson log-likelihood can also be used instead of the GP marginal log-likelihood to compute BIC (eq. (2.6)), $\mathrm{BIC}^{\mathrm{naive}}$ (eq. (2.5)), and AIC (eq. (2.7)).

2.4 Gaussian process kernels

We used the GP package from scikit-learn (https://scikit-learn.org/stable/) to implement our GP regression procedure. In this paper, we have explored three kernels to model the background distribution: the Radial Basis Function kernel (RBF), the Matérn kernel with $\nu = 5/2$ (Matern), and the second-order polynomial kernel (Poly2). The three kernel functions are defined as follows:

$$k_{\mathrm{RBF}}(x, x') = \sigma_0 \exp\left[-\frac{(x - x')^2}{2l^2}\right], \tag{2.10}$$

where σ_0 is the amplitude and l is the length scale of the covariance function;

$$k_{\mathrm{Matern}}(x, x') = \sigma_0 \left[1 + \frac{\sqrt{5}}{l} d(x, x') + \frac{5}{3l} d(x, x')^2\right] \exp\left[-\frac{\sqrt{5}}{l} d(x, x')\right], \tag{2.11}$$

where σ_0 is the covariance amplitude as in k_{RBF}, l is a positive parameter characterizing the covariance, and $d(x, x')$ is the Euclidean distance between datapoints x and x';

$$k_{\mathrm{Poly2}}(x, x') = (\sigma_0^2 + x \cdot x')^2, \tag{2.12}$$

where σ_0 sets the magnitude of the zeroth-order term in the polynomial expansion. Thus the RBF, Matern and Poly2 kernels depend on 2, 2 and 1 hyperparameter, respectively.

JHEP02(2023)230

2.5 Functional fit

For comparison, we also employ a fourth-order parametric polynomial fit with explicit basis functions, typically used to model the background distribution [26]:

$$f(x_i) = \sum_{p=0}^{4} w_p x_i^p + m(x_i), \tag{2.13}$$

where w_p are the fitting coefficients and $m(x_i)$ is either set to 0 for background-only fits with the signal window masked out, or given by eq. (2.4) for signal+background fits on the entire dataset. The fits were carried out using ROOT data analysis software [34], by maximizing the Poisson log-likelihood in eq. (2.9). For the background-only fit, the values of the fitting coefficients are $w_0 = 1.84 \times 10^5 \pm 1.60 \times 10^2$, $w_1 = -4.49 \times 10^3 \pm 1.80 \times 10^0$, $w_2 = 4.22 \times 10^1 \pm 1.00 \times 10^{-2}$, $w_3 = -1.79 \times 10^{-1} \pm 9.00 \times 10^{-4}$, $w_4 = 2.84 \times 10^{-4} \pm 4.0 \times 10^{-7}$. For the background+signal fit, the fitting coefficients are $w_0 = 1.64 \times 10^5 \pm 6.16 \times 10^4$, $w_1 = -3.87 \times 10^3 \pm 1.86 \times 10^3$, $w_2 = 3.52 \times 10^1 \pm 2.10 \times 10^1$, $w_3 = -1.43 \times 10^{-1} \pm 1.05 \times 10^{-1}$, $w_4 = 2.19 \times 10^{-4} \pm 1.94 \times 10^{-4}$, while the signal contribution is described using $\{A, \mu, \sigma\} = \{443 \pm 199, 124.5 \pm 0.8, 2.3 \pm 0.9\}$. All the parameter uncertainties have been estimated via Hessian analysis routines implemented in MINUIT [35] (https://iminuit.readthedocs.io/en/stable/citation.html) and available through ROOT. The value of A and its uncertainty have been rounded to correspond to the integer number of events. The datasets on which the fits have been performed are described in more detail below.

3 Datasets

We use the di-photon sample from the open dataset made available by the ATLAS collaboration at LHC [26]. We use the selection criteria as documented in ref. [26] to create a di-photon invariant mass distribution, $m_{\gamma\gamma}$, that shows the Higgs boson decay. The Higgs boson decay is a localized bump on top of the smooth background distribution, traditionally modeled by a polynomial [20]. The di-photon distribution consists of integer event counts y_i in $N = 30$ bins. Since in this work we focus on the datasets in which we expect to find a localized signal whose location is approximately known, we first mask out the region containing the signal. We expect the signal to be localized with a characteristic width that is relatively small compared to the background energy scales [20]. Indeed, in new resonance searches one typically scans for the signal at multiple windows within the full range of the dataset, with a prior expectation for the signal width.

To search for a signal in a specific window, we use the entire range of data with the signal window masked out to determine the optimal parameters for the background-only GP regression fit. In new resonance searches, this process could be repeated for multiple masked-out signal windows. Here, we model the signal using a simple Gaussian whose mean μ and width σ are approximately known to be 125 GeV and 2.5 GeV, respectively [5, 26]. Thus, in the background-only fits we mask out a signal window $\pm 2\sigma$ around the signal mean μ; all data outside of this window are assumed to belong to the background distribution

JHEP02(2023)230

and are therefore fit using GP regression with $m(x_i) = 0$. Specifically, we optimize the parameters θ of a given kernel by maximizing the marginal log-likelihood (eq. (2.2)), which yields the optimal set of hyperparameters $\hat{\theta}$. Given $\hat{\theta}$, the predictive distribution is then provided by eq. (2.3) with $m(x_i) = 0$. The resulting hyperparameter values are $\{\sigma_0, l\} = \{5.39 \times 10^7, 46.0\}$ for the RBF kernel, $\{\sigma_0, l\} = \{1.12 \times 10^8, 124.0\}$ for the Matern kernel, and $\sigma_0 = 29.3$ for the Poly2 kernel.

4 Model selection for background-only fits

To determine which kernel type best represents our data, we have carried out model selection using BIC and AIC model comparison measures, summarized in table 1. Our main goal is to develop a practical, transparent approach to model selection in signal+background problems. A secondary goal is to investigate how $\mathrm{BIC}^{\mathrm{naive}}$ compares with BIC. The former involves additional approximations, while the latter is harder to implement numerically since it requires numerical evaluation of the Hessian.

We find that the marginal log-likelihood is much worse for Poly2 than for either RBF or Matern. This disadvantage is too substantial to be offset by the fact that Poly2 uses one less hyperparameter. Furthermore, both $\mathrm{BIC}^{\mathrm{naive}}_{\mathrm{GL}}$ (eq. (2.5)) and $\mathrm{BIC}_{\mathrm{GL}}$ (eq. (2.6)) rank the three kernels in the same way, giving a slight edge to RBF over Matern. Note that this rank is the same as with the marginal log-likelihood without BIC corrections. However, Matern is slightly favored over RBF when considering Poisson log-likelihoods with rates provided by the mean of the GP predictive distribution (eq. (2.9)). This preference for Matern holds when the Poisson log-likelihoods are augmented with model complexity corrections to produce $\mathrm{BIC}^{\mathrm{naive}}_{\mathrm{PL}}$ scores. Next, we have considered the effect of adding the AIC penalty to the Poisson log-likelihoods ($\mathrm{AIC}_{\mathrm{PL}}$ in table 1). The AIC penalty effectively accounts for the amount of smoothing caused by each kernel type [33, 36]. We observe that RBF is favored over Matern according to the $\mathrm{AIC}_{\mathrm{PL}}$ scores.

To summarize these findings in a succinct way, we propose a high-level voting scheme in which the kernel type is chosen on the basis of 1,2,3 rankings produced by $\mathrm{BIC}^{\mathrm{naive}}_{\mathrm{GL}}$, $\mathrm{BIC}^{\mathrm{naive}}_{\mathrm{PL}}$, and $\mathrm{AIC}_{\mathrm{PL}}$ scores. We exclude $\mathrm{BIC}_{\mathrm{GL}}$ from the vote because it is perfectly correlated with $\mathrm{BIC}^{\mathrm{naive}}_{\mathrm{GL}}$ and therefore conveys the same message. The sums of the rankings are 4 for RBF, 5 for Matern, and 9 for Poly2. Thus, RBF and Matern have nearly the same scores, while Poly2 is ranked considerably lower. Our ranking scheme is simple and intuitive but, as with all ranking schemes, it does not take into account the magnitude of the differences in the model selection scores. Alternative schemes are possible depending on user preferences and the system under consideration.

Finally, we compare Func4 with the three GP regression models. Using the same ranking scheme on the table 1 columns for which Func4 scores are available ($\mathrm{AIC}_{\mathrm{PL}}$ and $\mathrm{BIC}^{\mathrm{naive}}_{\mathrm{PL}}$), we obtain the following ranking sums: 3 for RBF, 4 for Matern, 5 for Func4, and 8 for Poly2. Thus, Func4 is less preferable than RBF or Matern. With $\mathrm{BIC}^{\mathrm{naive}}_{\mathrm{PL}}$, Func4 is strongly disfavored due to its larger number of fitting parameters. With $\mathrm{AIC}_{\mathrm{PL}}$, Func4 ranks second, just behind RBF. This is despite the fact that Poisson log-likelihoods slightly favor Func4 over RBF or Matern GP fits (table 1). Considering all the evidence together,

JHEP02(2023)230

Model	$\log\lvert H\rvert$	n	d	-log(PL)	-log(GL)	$\mathrm{BIC}^{\mathrm{naive}}_{\mathrm{GL}}$	$\mathrm{BIC}_{\mathrm{GL}}$	$\mathrm{AIC}_{\mathrm{PL}}$	$\mathrm{BIC}^{\mathrm{naive}}_{\mathrm{PL}}$
Poly2	-0.531	1	2.99	38.02	87.52	89.22	87.25	82.02	39.72
RBF	0.417	2	4.68	8.95	72.15	75.55	72.36	27.26	12.35
Matern	2.906	2	5.67	8.69	72.30	75.70	73.75	28.72	12.09
Func4	—	5	5	8.65	—	—	—	27.30	17.15

Table 1. *Model comparison for background-only fits.* $|H|$ is the determinant of the Hessian of the marginal log-likelihood (eq. (2.2)) evaluated at $\hat{\theta}$; n is the number of model parameters; d is the number of model degrees of freedom ($d = d_{eff}$ for GP regression, $d = n$ for Func4); log(PL) is the Poisson likelihood (eq. (2.9)) computed using either $f(x_i)$ evaluated at the optimal values of hyperparameters $\hat{\theta}$ from GP regression (Poly2,RBF,Matern) or found by maximum likelihood, with $f(x_i)$ given by eq. (2.13) (Func4); log(GL) is the marginal log-likelihood (eq. (2.2)). BIC (eq. (2.6)) and $\mathrm{BIC}^{\mathrm{naive}}$ (eq. (2.5)) were computed using log(GL) for BIC and both log(GL)/log(PL) for $\mathrm{BIC}^{\mathrm{naive}}$. AIC values (eq. (2.7)) were computed using log(PL).

we conclude that GP regression with the RBF kernel is the best way to model our data, although the preference of RBF over Matern is fairly slight.

To better understand AIC and BIC-based model selection, we have carried out visual comparisons of the four different models, by plotting the mean predictive distribution of the GP regression with RBF, Matern and Poly2 kernels ($f(x_i)$ in eq. (2.3) with $m(x_i) = 0$), and the maximum-likelihood (ML) Func4 fit (eq. (2.13)) (figure 1). All four models were fit to the event counts outside of the signal window. It is clear from the upper panel of figure 1 that RBF, Matern and Func4 produce very similar fits, whereas Poly2 fits the data poorly. This is also evident from the residuals (figure 1, lower panel), whose magnitudes are consistently larger for Poly2 but very similar for the other three models. As expected, when RBF, Matern and Func4 background-only fits are subtracted from the data, which contains the Higgs boson signal, the corresponding residuals show a distinct bump within the signal window; outside of that window, deviations of the residuals from zero are almost always within the error bars. Thus, visual inspection could be used to rule out Poly2 but not to differentiate between RBF, Mattern, and Func4, which therefore require BIC/AIC analysis presented above.

5 Signal extraction

In order to extract the signal superimposed on top of the background distribution, we have carried out GP regression with the RBF kernel and $m(x_i)$ given by eq. (2.4) using the entire dataset (figure 2). Importantly, the kernel parameters were kept at the values $\hat{\theta}$ obtained via the previous fit to the background distribution, with the signal window masked out. Thus, the kernel hyperparameters are responsible for modeling the background, while the Gaussian parameters in eq. (2.4) are responsible for describing the signal. The resulting signal parameters are $\{A_{\mathrm{RBF}}, \mu_{\mathrm{RBF}}, \sigma_{\mathrm{RBF}}\} = \{473 \pm 123, 124.7 \pm 0.6, 2.4 \pm 0.4\}$, with the uncertainties estimated using Hessian-based error analysis routines implemented in MINUIT [35] (https://iminuit.readthedocs.io/en/stable/citation.html). For comparison, we

JHEP02(2023)230

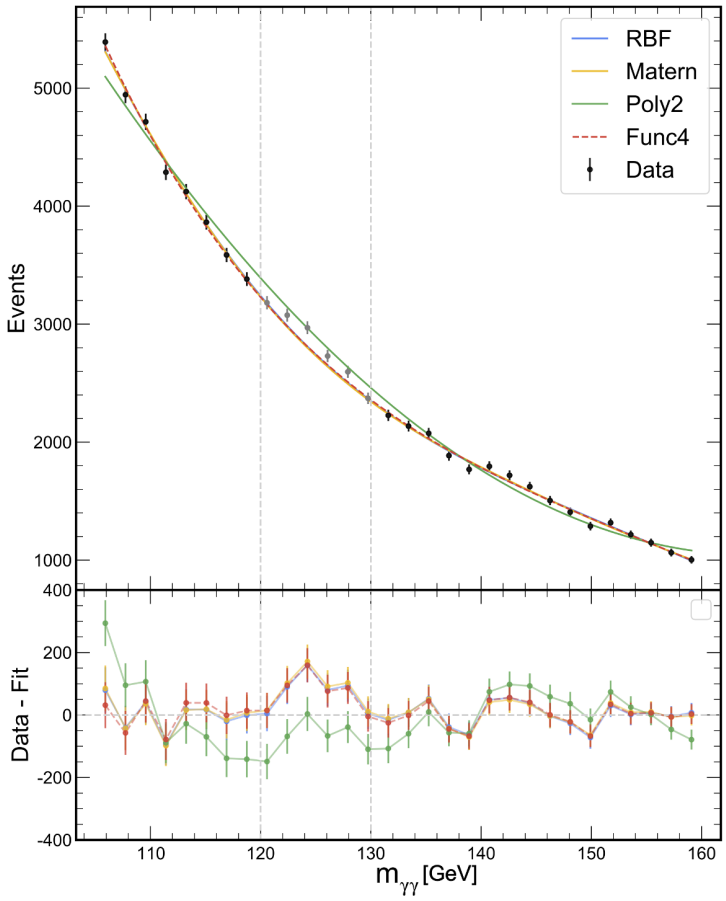


Figure 1. *Fits to the background distribution.* Upper panel: shown are the mean predictive distributions produced by GP regression with RBF, Matern and Poly2 kernels, as well as the ML fit with the fourth-order polynomial (Func4). All models were fit on the background-only data outside of the 125 ± 5 GeV window where the signal is located. Event counts y_i are shown with error bars constructed using Garwood intervals [28, 29] (black markers: background only, grey markers: background+signal). Lower panel: residuals $y_i - f(x_i)$ for each of the four models in the upper panel, with y_i error bars from the upper panel. Residual values are connected by lines of the same color and type as in the upper panel, to guide the eye. In both panels, vertical dashed lines indicate the boundaries of the signal window.

JHEP02(2023)230

have also refit the Func4 model (eq. (2.13)) on the entire dataset, by adding a Gaussian function (eq. (2.4)) to capture the signal contribution. As mentioned above, the Gaussian fitting parameters are $\{A_{\text{Func4}}, \mu_{\text{Func4}}, \sigma_{\text{Func4}}\} = \{443 \pm 199, 124.5 \pm 0.8, 2.3 \pm 0.9\}$ in this case. We observe that both RBF and Func4 are capable of capturing the approximately Gaussian signal which remains after subtracting the background distribution. Indeed, the correspondence between the mean predictive distribution for RBF, the ML Func4 fit, and the Higgs simulation (described in ref. [26]) is very high (figure 2). However, we note that the GP approach with the RBF kernel results in $A_{\text{RBF}} = 473 \pm 123$ (3.85σ) events above the background, compared to $A_{\text{Func4}} = 443 \pm 199$ (2.23σ) events from the Func4 fit. Thus, the probability of zero signal counts is 6.01×10^{-5} for RBF and 1.30×10^{-2} for Func4, indicating a more significant prediction of the signal presence with the GP RBF fit.

5.1 Synthetic datasets for testing statistical significance of signal extraction

To investigate further the statistical significance of the observed signal and the potential systematic biases, we have created 5000 toy datasets based on the GP fit with the RBF kernel and $m(x_i) = 0$ to the background-only data outside the signal window. This fit has generated an effective integer number of counts due to the background only, $N_{\text{eff}} = [\sum_{i=1}^{N} f(x_i)]$, where the square brackets indicate the rounding operation. Next, we sampled from the GP predictive probability $\mathcal{N}(\tilde{y}_i | f(x_i), V(x_i))$, producing real-valued background "counts" $\tilde{y}_i$ in each bin i. Finally, we used $\tilde{y}_i / \sum_{i=1}^{N} \tilde{y}_i$ as probabilities in a multinomial sampling process, generating a synthetic histogram of integer event counts. Each synthetic histogram was constrained to have N_{eff} counts. Note that our toy datasets include both the uncertainty inherent in GP regression and the uncertainty related to generating integer event counts from the underlying model. In order to create a full background+signal test set, we have added signal counts from the Higgs simulations [26] to each of the 5000 background datasets. Thus the signal component is fixed, while the background component varies from dataset to dataset according to the background model uncertainties.

5.2 Test for systematic biases in signal extraction

To test the robustness of the fit, we check for potential biases in our background estimation procedure. Namely, for each of the 5000 background+signal toy datasets described above, we carry out a GP fit with the Gaussian mean function (eq. (2.4)) on the entire dataset, while keeping the kernel hyperparameter values fixed at $\hat{\theta}$, values found by the previously described fit to the background distribution, with the signal window masked out. This procedure generates a set of 5000 predicted signal strength values, $\{A_{\text{fit}}\}$, which can be compared with the corresponding exact value $A_{\text{true}} = 450$, the sum of the event counts added to the background-only counts in order to create the combined background+signal toy datasets. Specifically, we compute a distribution of differences between fitted and exact signal strengths [37] (figure 3):

$$\Delta A = A_{\text{fit}} - A_{\text{true}}. \tag{5.1}$$

We observe that the empirical distribution is well described by a Gaussian with $\mu = -30.43$ and $\sigma = 157.97$. Although this value of μ is just a fraction of σ, it is non-zero

JHEP02(2023)230

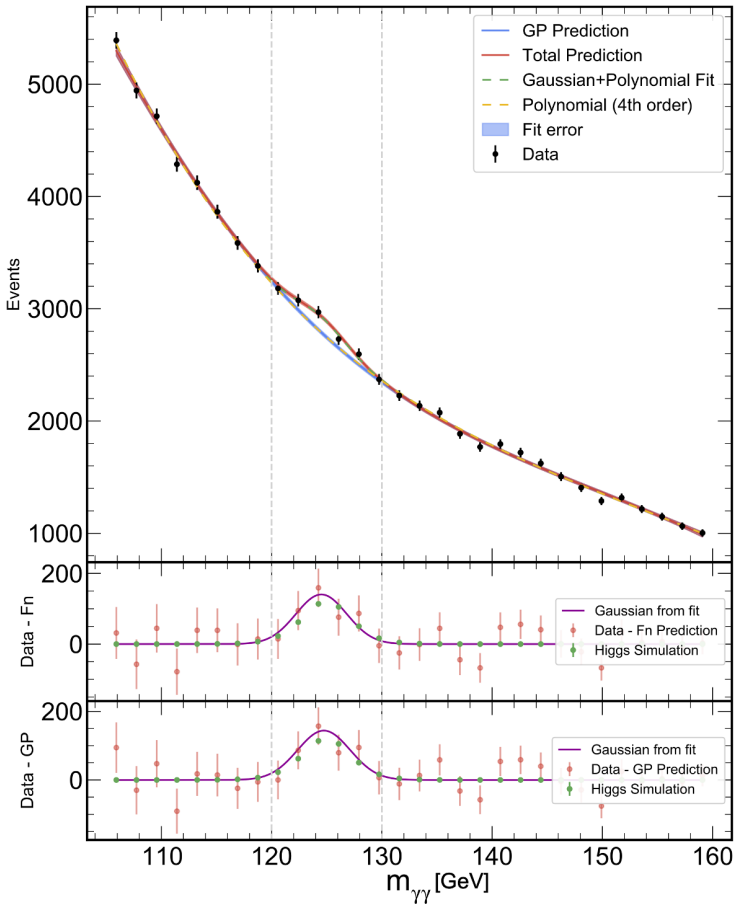


Figure 2. *Extracting the Higgs boson signal from the data.* Upper panel: RBF[Bkg] and Func4[Bkg] are the background fits with the masked-out signal window (same as figure 1). RBF[Bkg+signal] and Func4[Bkg+signal] are the fits on the entire dataset, with kernel hyperparameters (RBF) and polynomial parameters (Func4) kept at their background-only values. Event counts y_i are shown with error bars constructed using Garwood intervals [28, 29] (black markers). Middle panel: Gaussian signal predicted using Func4 background, residuals $y_i - f(x_i)$ with respect to the Func4 background-only model, with y_i error bars, and the results of the Higgs simulation described in ref. [26]. Lower panel: same as the middle panel but with RBF instead of Func4. In all panels, vertical dashed lines indicate the boundaries of the signal window.

JHEP02(2023)230

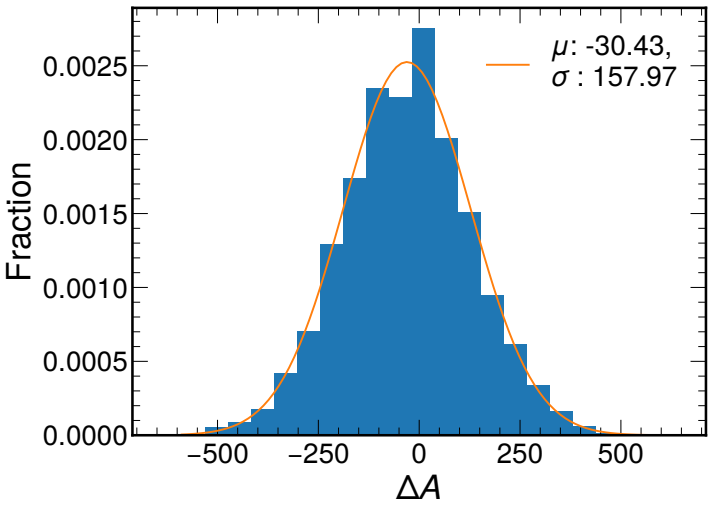


Figure 3. *Systematic biases in signal extraction.* Shown is a normalized histogram of the differences between fitted and exact signal strengths (eq. (5.1)) obtained by generating 5000 toy datasets and carrying out GP regression with the RBF kernel as described in the text (blue bars). Orange curve: a Gaussian fit to the histogram, which yields $\mu = -30.43$, $\sigma = 157.97$.

Signal Width (GeV)	Scale Factor X	$\langle \Delta A \rangle$	$\langle \Delta A \rangle / A_{\text{true}}$	$\sigma(\Delta A)$
2.40	0.5	-28.26	-0.13	108.21
2.40	1.0	-30.43	-0.07	157.97
2.40	2.0	-56.65	-0.06	230.45
2.40	10.0	-129.43	-0.03	503.79

Table 2. *Dependence of the systematic bias on the size of the data sample.* The scale factor is denoted by X, $\Delta A = A_{\text{fit}} - A_{\text{true}}$ is the difference between fitted and exact signal strengths, and $\langle \Delta A \rangle$ and $\sigma(\Delta A)$ are the mean and the standard deviation of ΔA over 5000 independently generated synthetic datasets.

with a high level of statistical significance: $\mu = -30.43 \pm \sigma/\sqrt{5000} = -30.43 \pm 2.23$. Thus, our two-step background+signal reconstruction procedure leads to a relatively slight but systematic underestimation of the signal counts. Since the average magnitude of the systematic bias is small, we conclude that the signal contribution can be reliably extracted from the underlying smooth background distribution.

To explore how our predictions depend on the total size of the data sample, we have redone the systematic bias analysis in figure 3 with 5000 toy datasets in which both the total number of background counts and the total injected signal strength A_{true} were rescaled by a factor X (table 2). We observe that although both the mean value of the bias $\langle \Delta A \rangle$ and its standard deviation $\sigma(\Delta A)$ grow with the size of the data sample, the ratio of $\langle \Delta A \rangle$ to the true signal strength A_{true} decreases. Thus, the relative bias of GP RBF regression becomes smaller as more and more data are collected.

JHEP02(2023)230

Signal Width (GeV)	Signal Strength A_{true}	$\langle \Delta A \rangle$	$\langle \Delta A \rangle / A_{\text{true}}$	$\sigma(\Delta A)$
2.00	450	-25.28	-0.06	126.51
2.40	450	-30.43	-0.07	157.97
3.00	450	-52.30	-0.12	181.92
3.50	450	-59.77	-0.13	225.83
2.40	225	-17.07	-0.08	145.93
2.40	450	-30.43	-0.07	157.97
2.40	675	-31.67	-0.05	159.74
2.40	900	-39.14	-0.04	160.03

Table 3. *Dependence of the systematic bias on the signal width and strength.* The difference between fitted and exact signal strengths is denoted by $\Delta A = A_{\text{fit}} - A_{\text{true}}$; $\langle \Delta A \rangle$ and $\sigma(\Delta A)$ are the mean and the standard deviation of ΔA over 5000 independently generated synthetic datasets.

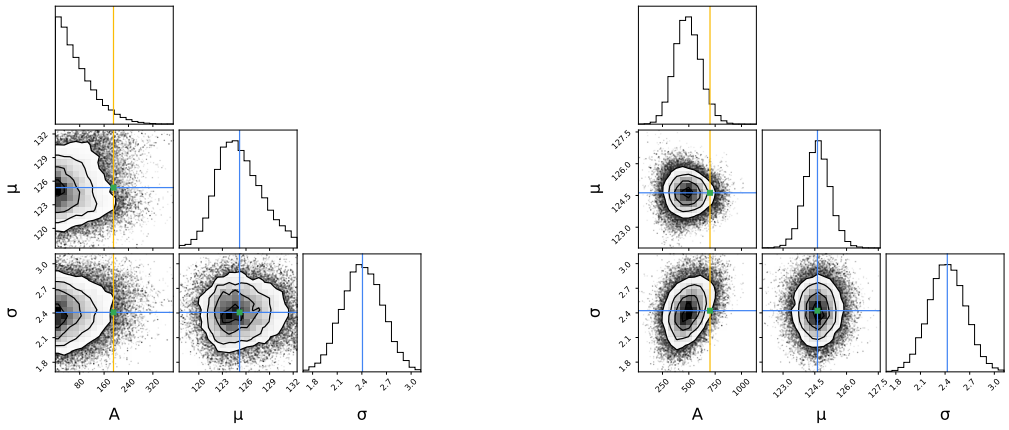
Next, we have studied how the systematic bias depends on the signal strength and width (table 3). As in figure 3, we have generated 5000 toy datasets with both background and signal counts. While the total number of background counts was unaltered, we have varied the width and the strength of the injected signals. We observe that the relative bias, quantified as $\langle \Delta A \rangle / A_{\text{true}}$ in table 3, increases with the signal width. This is not surprising since wider signals are spread over more bins and therefore it is harder to separate them from the background. When the signal strength is increased, the relative bias decreases — it is easier to extract stronger signals from the background.

5.3 Posterior distributions of signal parameters and significance analysis

We have investigated the posterior distributions of signal-characterizing parameters by carrying out Markov Chain Monte Carlo (MCMC) sampling [38] of the Poisson log-likelihood (eq. (2.9)). Routinely employed in Bayesian analysis, MCMC sampling of posterior probabilities is similar to studying model parameter sensitivity and estimating confidence intervals in frequentist statistics [39, 40]. Poisson rates $f(x_i)$ depend on the hyperparameter values $\hat{\theta}$ obtained via the previously described background-only fit and on the mean function $m(x_i)$, whose parameters $\{A, \mu, \sigma\}$ were sampled from the following priors: the prior for A is uniform in the $[0, +\infty)$ range, while the prior for μ is Gaussian, with the 124.7 GeV mean and 0.02×124.7 GeV standard deviation. The σ prior is also Gaussian, with the 2.4 GeV mean and 0.1×2.4 GeV standard deviation. The mean values are consistent with the Higgs simulation [26] and with the fits presented in figure 2. The 0.02 and 0.1 scaling factors in the priors are motivated by the ATLAS studies [5]. MCMC was implemented using the Emcee package [41] (https://emcee.readthedocs.io), with 10^4 samples in each of 12 independent MC trajectories.

Figure 4 shows MCMC posterior distributions of the three parameters characterizing the signal: overall signal strength A, the mean position of the signal peak μ, and the width of the signal peak σ. In figure 4a, MCMC sampling is based on a synthetic dataset without

JHEP02(2023)230



(a) (b)

Figure 4. *Posterior distributions of signal parameters obtained by MCMC sampling.* Shown are posterior probabilities in the corner plot format (https://corner.readthedocs.io/en/latest/) [42]. Diagonal entries: marginalized posterior probabilities $P(A)$, $P(\mu)$, $P(\sigma)$, off-diagonal entries: joint posterior probabilities $P(\mu, A)$, $P(\sigma, A)$, $P(\mu, \sigma)$, with the contours indicating 2D sigma levels. Blue lines mark the median values in the posterior distributions, yellow lines mark the 95% quantile values. Panel (a): MCMC analysis of a randomly chosen synthetic dataset (out of 5000 toy datasets with background counts only). Panel (b): MCMC analysis of the experimentally observed background+signal counts. In both panels, 10^4 samples are shown.

any signal added. The dataset was randomly chosen among the 5000 background-only test sets described above. As expected, $P(A)$, the marginalized posterior probability for signal strength, is highest when A is close to zero and falls off rapidly as A increases, while $P(\mu)$ and $P(\sigma)$ appear Gaussian. Moreover, correlations between pairs of parameters are weak: $r_{\mu,A} = 0.065$, $r_{\sigma,A} = 0.065$, $r_{\sigma,\mu} = 0.030$, where $r_{x,y}$ is a linear correlation coefficient between x and y. In contrast, when the real data containing both the background counts and Higgs boson events is analyzed, the maximum posterior probability value of A is located around 500 counts, consistent with the earlier Hessian-based error analysis of GP regression with the RBF kernel (figure 4b). Indeed, a Gaussian fit of $P(A)$ in figure 4b yields 485 ± 121 Higgs boson events, very close to the 473 ± 123 Higgs boson events obtained earlier using the GP regression framework. Thus there is a clear signature of signal counts in the real data. Interestingly, sampling of the joint probability $P(\sigma, A)$ reveals a positive correlation between signal strength A and signal width σ model parameters ($r_{\sigma,A} = 0.321$), such that stronger signals tend to have larger widths. The other two correlations remain weak: $r_{\mu,A} = -0.048$, $r_{\sigma,\mu} = 0.019$.

To provide a more quantitative estimate of the statistical significance of the signal strength observed in real data, we have plotted a histogram of the 95% confidence levels for A for all 5000 background-only toy datasets (figure 5) [43, 44]. The value observed with the actual data is $3.15\tilde{\sigma}$ above the median, where $\tilde{\sigma}$ is the distance between the median and the 84% quantile, and is larger than 99.84% of the values empirically observed in the histogram.

JHEP02(2023)230

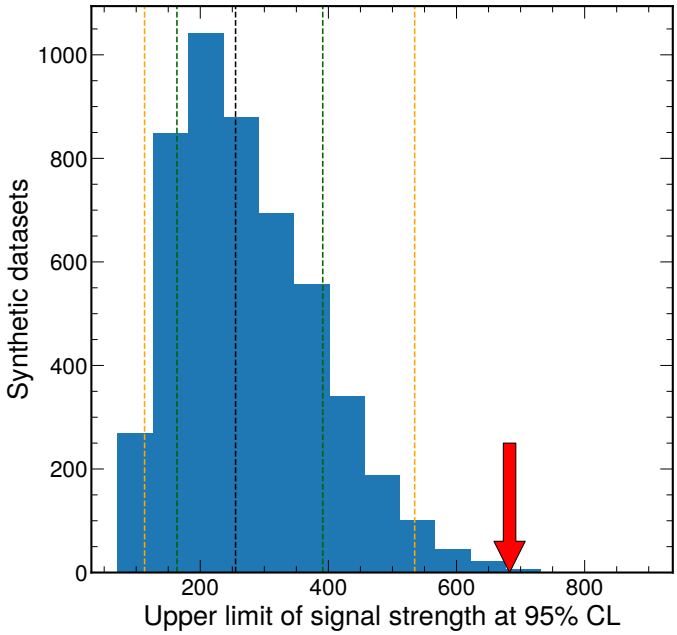


Figure 5. *Distribution of the 95% quantiles of the signal strength A in background-only datasets.* Shown is the histogram of the 95% quantiles (confidence levels (CL) for the upper limit of the signal strength) for A, inferred from 5000 synthetic datasets with background-only counts using MCMC sampling. The black dashed line indicates the 50% quantile, or the median, of the 95% CL distribution (i.e., the median number of Higgs boson events observed above background, reported at 95% CL for the upper limit of the signal strength). From left to right, the yellow lines show 2.5% and 97.5% quantiles and the green lines show 16% and 84% quantiles, respectively. The red arrow indicates the 95% CL value obtained from the background+signal dataset (cf. figure 4b).

6 Summary

In this work, we have developed a procedure for using Gaussian Process (GP) regression to extract localized signals with an approximately known position and width from smooth background distributions. Although this procedure is of interest in many areas of science, including astrophysics and crystallography, here we have focused on extracting the Higgs boson signal strength from an ATLAS open dataset which consists of binned event counts [26]. This is a challenging task because the signal is masked by the background whose functional form is unknown. Thus, the inference procedure and the background model affect the statistical significance of the signal extracted by fitting a computational model to experimental observations. Traditionally, the background distribution is modeled using a polynomial fit onto which a Gaussian signal is superimposed (eq. (2.13)) [26]. Here we propose an alternative framework in which GP regression is used to model the background+signal data, with the signal added via the mean function of the Gaussian process. As with the functional fits, the mean function is represented by a Gaussian with three free parameters (eq. (2.4)), one of which, A, is especially relevant since it represents the total number of signal events found in the dataset.

JHEP02(2023)230

The GP framework is more flexible than standard fitting approaches that employ a fixed set of basis functions such as polynomials or Gaussians [7, 14]. This flexibility comes from the focus on the correlations between datapoints, which are modeled using kernel functions. Although GP methods are not limited by the prior choice of the finite basis (indeed, some popular kernels formally correspond to infinite basis sets [7]), most kernel functions depend on one or several hyperparameters, such as the characteristic length scale in the RBF kernel. A fully Bayesian treatment the dependence of the model evidence (marginal likelihood) on the hyperparameters is usually impractical; however, simply maximizing the evidence with respect to the hyperparameters may lead to overfitting for more complex kernels. In order to provide a more principled approach to the selection of the kernel type, we have considered two independent methodologies.

One approach, BIC, is based on evaluating the model evidence under the Laplace approximation and the assumption that the effects of hyperparameter priors are negligible (eq. (2.6)). With several additional approximations, notably the assumption that the Hessian matrix has full rank, BIC yields a simple and widely used correction which penalizes model complexity (eq. (2.5)) [7]. The other approach, AIC, is non-Bayesian. Instead of concentrating on the model evidence, it focuses on the degree of smoothing that results from applying a given kernel to the dataset [31, 32]. Thus, the AIC and BIC approaches to model complexity are complementary to each other and reflect different kernel properties (the amount of data smoothing vs. the shape of the log-likelihood landscape as a function of hyperparameters). Using both criteria holistically in a ranking scheme, we have chosen the RBF kernel for our GP regression models, although the results with the Matern kernel are of nearly the same quality.

We note that AIC yields approximately equal scores for the GP RBF fit and the traditional fit, which models the background using a fourth-order polynomial (table 1). The results of the two fits are also very similar visually, and both approaches are close to the Higgs simulations predictions (figures 1, 2). However, the total area A under the signal bump is somewhat higher with the GP RBF fit compared to the functional fit, with 473 and 443 Higgs boson events, respectively. Moreover, Hessian-based error analysis reveals that the standard deviation is much smaller with the GP prediction, 123 vs. 199 in the functional fit. Thus, the GP approach is preferable since it leads to a more significant prediction of the signal strength.

After investigating the relative importance of the bias in our signal extraction procedure (figure 3, tables 2, 3), we have proceeded to explore the posterior probabilities of model parameters by MCMC sampling (figure 4). The MCMC approach is necessary since it can be used to investigate the Poisson log-likelihood (eq. (2.9)), which is more appropriate for modeling integer event counts. The Poisson log-likelihood depends on the kernel hyperparameters, which were kept fixed to their values $\hat{\theta}$ obtained by fitting to the background-only data (figure 1), and on the signal strength, mean and width, which were sampled from prior distributions. The prior for signal strength A was uninformative, assigning equal weights to any non-negative value. The priors for the mean and the width were informative, modeled by Gaussians whose parameters were constrained by Higgs simulations (figure 2) and by the previous studies of instrumental errors in the ATLAS detector [5]. The resulting posterior

JHEP02(2023)230

probability for signal strength shows a clear Higgs boson signature, with 485 ± 121 Higgs boson-related events (figure 4b). These numbers are consistent with the previous estimate obtained by Hessian-based error analysis of the signal parameters in GP regression, which yielded 473 ± 123 Higgs boson-related events.

When the MCMC sampling procedure is applied to synthetic datasets where no contributions from the signal are expected, the posterior distribution for signal strength is peaked at zero and the typical predicted values are much smaller (figure 4a). The latter is clearly seen by combining the data from 5000 independently generated background-only synthetic datasets into a histogram of 95% confidence levels for signal strength A (figure 5). The corresponding confidence level obtained from the real dataset is larger than 99.84% of the histogram values and corresponds to 3.15 $\tilde{\sigma}$, where $\tilde{\sigma}$ is the distance between the median and the 84% quantile. Thus our signal strength prediction is also highly significant within the MCMC framework.

In summary, we have developed a novel GP regression framework for extracting localized signals from smooth background distributions. This problem appears in many areas of science where a weak signal of interest is masked by background events due to light scattering, extraneous emission sources, and other interfering processes. The location and the width of the signal can sometimes be guessed based on physical considerations; in other cases, scanning over multiple putative signal windows is necessary, as in LHC anomaly detection searches [45–50]. In both scenarios, only rough estimates of the position and the width of the signal window are required. Data outside of the signal window is assumed to belong to the background and a GP model without the signal contribution is fitted to it. With multiple windows, a key consideration is the "look elsewhere effect" — an enhancement of the fake signal rate and the resulting decrease of the statistical significance of the true signal due to multiple testing [51]. The interplay between the fake signal rate and our procedure for modeling the background distribution will be discussed in a future study.

We carry out model selection using both BIC and AIC considerations, including an in-depth analysis of the BIC assumptions. The extrapolation of the model across the signal window then provides an estimate of the background, from which the signal can now be separated in a second GP fit where only the signal parameters are allowed to vary, while all the background parameters remain fixed. This two-step procedure allows us to measure the signal and background parameters in a robust and reproducible manner. An application of our approach to the open Higgs boson dataset from the ATLAS detector (known as the ATLAS open dataset) yields a highly significant prediction of the Higgs boson signature, outperforming the traditional approach based on fitting a polynomial function to the background distribution.

Acknowledgments

We thank John Paul Chou, Igor Volobouev, David Shih and Yuri Gershtein for their valuable comments on the manuscript while it was in preparation and revision. This work was supported by the National Science Foundation through grants NSF-PHY-1607096 (to A.L.) and NSF-MCB-1920914 (to A.V.M.). This manuscript has been authored by

JHEP02(2023)230

Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

Open Access. This article is distributed under the terms of the Creative Commons Attribution License (CC-BY 4.0), which permits any use, distribution and reproduction in any medium, provided the original author(s) and source are credited. SCOAP^3 supports the goals of the International Year of Basic Sciences for Sustainable Development.

References

[1] D.S. Sivia and W.I.F. David, *A Bayesian approach to extracting structure-factor amplitudes from powder diffraction data*, *Acta Crystallographica A* **50** (1994) 703.

[2] W.I.F. David and D.S. Sivia, *Background estimation using a robust Bayesian analysis*, *J. Appl. Crystallography* **34** (2001) 318.

[3] T.A. Gordon, E. Agol and D. Foreman-Mackey, *A fast, two-dimensional Gaussian process method based on celerite: applications to transiting exoplanet discovery and characterization*, *Astron. J.* **160** (2020) 240.

[4] D. Foreman-Mackey, E. Agol, S. Ambikasaran and R. Angus, *Fast and scalable Gaussian process modeling with applications to astronomical time series*, *Astron. J.* **154** (2017) 220.

[5] ATLAS collaboration, *Observation of a new particle in the search for the standard model Higgs boson with the ATLAS detector at the LHC*, *Phys. Lett. B* **716** (2012) 1 [arXiv:1207.7214] [INSPIRE].

[6] Particle Data Group collaboration, *Review of particle physics*, *PTEP* **2020** (2020) 083C01 [INSPIRE].

[7] C. Bishop, *Pattern recognition and machine learning*, Springer (2006).

[8] P. Mehta et al., *A high-bias, low-variance introduction to machine learning for physicists*, *Phys. Rept.* **810** (2019) 1 [arXiv:1803.08823] [INSPIRE].

[9] J.W. Rocks and P. Mehta, *Memorizing without overfitting: bias, variance, and interpolation in over-parameterized models*, arXiv:2010.13933.

[10] S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, *Annals Math. Statist.* **9** (1938) 60 [INSPIRE].

[11] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, *Eur. Phys. J. C* **71** (2011) 1554 [*Erratum ibid.* **73** (2013) 2501] [arXiv:1007.1727] [INSPIRE].

[12] ATLAS collaboration, *Search for new phenomena in dijet mass and angular distributions from pp collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, *Phys. Lett. B* **754** (2016) 302 [arXiv:1512.01530] [INSPIRE].

[13] M. Titsias, *Variational learning of inducing variables in sparse Gaussian processes*, in *Proceedings of the twelth international conference on artificial intelligence and statistics*, D. van Dyk and M. Welling eds., *Proc. Machine Learning Research* **5**, Hilton Clearwater Beach Resort, Clearwater Beach, FL, U.S.A., 16–18 April 2009, p. 567.

[14] C.E. Rasmussen and C.K.I. Williams, *Gaussian processes for machine learning*, The MIT Press (2005).

JHEP02(2023)230

[15] K. Kersting, C. Plagemann, P. Pfaff and W. Burgard, *Most likely heteroscedastic Gaussian process regression*, in *Proceedings of the* 24th *international conference on machine learning*, ACM (2007).

[16] A. O'Hagan, *Curve fitting and optimal design for prediction*, *J. Roy. Statist. Soc. B* **40** (1978) 1.

[17] K.G. Iyer et al., *Nonparametric star formation history reconstruction with Gaussian processes. I. Counting major episodes of star formation*, *Astrophys. J.* **879** (2019) 116.

[18] C.J. Moore, C.P.L. Berry, A.J.K. Chua and J.R. Gair, *Improving gravitational-wave parameter estimation using Gaussian process regression*, *Phys. Rev. D* **93** (2016) 064001 [arXiv:1509.04066] [INSPIRE].

[19] S. Golchi and R. Lockhart, *A Bayesian search for the Higgs particle*, arXiv:1501.02226.

[20] M. Frate, K. Cranmer, S. Kalia, A. Vandenberg-Rodes and D. Whiteson, *Modeling smooth backgrounds and generic localized signals with Gaussian processes*, arXiv:1709.05681 [INSPIRE].

[21] L. Diosan, A. Rogozan and J. Pécuchet, *Evolving kernel functions for SVMs by genetic programming*, in *Sixth International Conference on Machine Learning and Applications (ICMLA* 2007*)*, IEEE (2007), p. 19.

[22] W. Bing, Z. Wen-qiong, C. Ling and L. Jia-hong, *A GP-based kernel construction and optimization method for RVM*, in 2010 *the* 2nd *International Conference on Computer and Automation Engineering (ICCAE)*, IEEE (2010), p. 419.

[23] D. Duvenaud, J.R. Lloyd, R. Grosse, J.B. Tenenbaum and Z. Ghahramani, *Structure discovery in nonparametric regression through compositional kernel search*, arXiv:1302.4922.

[24] A.G. Wilson and R. Prescott Adams, *Gaussian process kernels for pattern discovery and extrapolation*, in *Proceedings of the* 30th *international conference on machine learning*, S. Dasgupta and D. McAllester eds., *Proc. Machine Learn. Res.* **28**, PMLR, Atlanta, GA, U.S.A., 17–19 June 2013, p. 1067 [arXiv:1302.4245].

[25] D. Ruppert, M.P. Wand and R.J. Carroll, *Semiparametric regression*, Cambridge University Press (2003).

[26] ATLAS collaboration, *Review of the* 13 *TeV ATLAS open data release*, Tech. Rep. ATL-OREACH-PUB-2020-001, CERN, Geneva, Switzerland (2020).

[27] P.W. Goldberg, K.I. Williams and C.M. Bishop, *Regression with input-dependent noise: a Gaussian process treatment*, *Adv. Neural Inf. Proc. Syst.* **10** (1998) 493.

[28] F. Garwood, *Fiducial limits for the Poisson distribution*, *Biometrika* **28** (1936) 437.

[29] L. Demortier, *Interval estimation*, in *Data analysis in high energy physics: a practical guide to statistical methods*, O. Behnke, K. Kroninger, G. Schott and T. Schorner-Sadenius eds., Wiley-VCH, Berlin, Germany (2013), p. 107.

[30] G. Schwarz, *Estimating the dimension of a model*, *Annals Statist.* **6** (1978) 461 [INSPIRE].

[31] H. Akaike, *A new look at the statistical model identification*, *IEEE Trans. Automat. Contr.* **19** (1974) 716.

[32] T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer (2001).

JHEP02(2023)230

[33] T. Hastie and R. Tibshirani, *Generalized additive models*, *Statist. Sci.* **1** (1986) 297.

[34] R. Brun and F. Rademakers, *ROOT: an object oriented data analysis framework*, *Nucl. Instrum. Meth. A* **389** (1997) 81 [INSPIRE].

[35] F. James and M. Roos, *Minuit: a system for function minimization and analysis of the parameter errors and correlations*, *Comput. Phys. Commun.* **10** (1975) 343 [INSPIRE].

[36] M. Hiabu, E. Mammen and J.T. Meyer, *Local linear smoothing in additive models as data projection*, *Ann. Statist.* **17** (1989) 453 [arXiv:2201.10930].

[37] ATLAS collaboration, *Recommendations for the modeling of smooth backgrounds*, Tech. Rep. ATL-PHYS-PUB-2020-028, CERN, Geneva, Switzerland (2020) [INSPIRE].

[38] W.K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, *Biometrika* **57** (1970) 97 [INSPIRE].

[39] G.J. Feldman and R.D. Cousins, *A unified approach to the classical statistical analysis of small signals*, *Phys. Rev. D* **57** (1998) 3873 [physics/9711021] [INSPIRE].

[40] A.L. Read, *Presentation of search results: the CL_s technique*, *J. Phys. G* **28** (2002) 2693 [INSPIRE].

[41] D. Foreman-Mackey, D.W. Hogg, D. Lang and J. Goodman, *emcee: the MCMC hammer*, *Publ. Astron. Soc. Pac.* **125** (2013) 306 [arXiv:1202.3665] [INSPIRE].

[42] D. Foreman-Mackey, *corner.py: scatterplot matrices in python*, *J. Open Source Softw.* **1** (2016) 24.

[43] CMS collaboration, *Search for pair-produced three-jet resonances in proton-proton collisions at $\sqrt{s} = 13$ TeV*, *Phys. Rev. D* **99** (2019) 012010 [arXiv:1810.10092] [INSPIRE].

[44] CMS collaboration, *Search for pair-produced resonances each decaying into at least four quarks in proton-proton collisions at $\sqrt{s} = 13$ TeV*, *Phys. Rev. Lett.* **121** (2018) 141802 [arXiv:1806.01058] [INSPIRE].

[45] J.H. Collins, K. Howe and B. Nachman, *Anomaly detection for resonant new physics with machine learning*, *Phys. Rev. Lett.* **121** (2018) 241803 [arXiv:1805.02664] [INSPIRE].

[46] T. Heimel, G. Kasieczka, T. Plehn and J.M. Thompson, *QCD or what?*, *SciPost Phys.* **6** (2019) 030 [arXiv:1808.08979] [INSPIRE].

[47] P. Jawahar et al., *Improving variational autoencoders for new physics detection at the LHC with normalizing flows*, *Front. Big Data* **5** (2022) 803685 [arXiv:2110.08508] [INSPIRE].

[48] O. Amram and C.M. Suarez, *Tag N' Train: a technique to train improved classifiers on unlabeled data*, *JHEP* **01** (2021) 153 [arXiv:2002.12376] [INSPIRE].

[49] A. Hallin et al., *Classifying anomalies through outer density estimation*, *Phys. Rev. D* **106** (2022) 055006 [arXiv:2109.00546] [INSPIRE].

[50] M. Farina, Y. Nakai and D. Shih, *Searching for new physics with deep autoencoders*, *Phys. Rev. D* **101** (2020) 075021 [arXiv:1808.08992] [INSPIRE].

[51] E. Gross and O. Vitells, *Trial factors for the look elsewhere effect in high energy physics*, *Eur. Phys. J. C* **70** (2010) 525 [arXiv:1005.1891] [INSPIRE].

JHEP02(2023)230