Background modeling and signal estimation using Gaussian Processes in the H$\rightarrow \gamma\gamma$ channel

Simon Millerjord

Thesis submitted for the degree of
Master in Subatomic Physics
60 credits

Department of Physics
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Autumn 2019

Background modeling and signal estimation using Gaussian Processes in the H$\rightarrow\gamma\gamma$ channel

Simon Millerjord

Background modeling and signal estimation using Gaussian Processes in the H→ $\gamma\gamma$ channel

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Abstract

The signal to background ratio for Higgs decaying to 2 photons is small, therefore, powerful estimation of the background is needed to accurately measure the signal. Since the underlying physical function is unknown, various functional parameterizations are considered for the background estimation. When the number of events increases, the relative uncertainty decreases. This might give rise to some previously hidden features of the distribution, leading to the need for re-estimation of the background to avoid creating spurious signal. The current process is lengthy and awkward, and therefore, this study has focused on investigating Gaussian Processes (GP) as a new method for estimating the background and signal distributions. GP is a machine learning method that does not depend on a specific parametric function and could therefore be employed in numerous scientific areas. Herein, it has been shown that GP manages to find the underlying function for a large number of toy data set, and it proved to be independent of the integrated luminosity. From bias tests performed, it was shown that bias generated during fitting, does not scale with the luminosity relative to the expected signal for a standard model Higgs boson. It was further shown that the background component of the GP does not model a signal when this is injected into the testing distributions. Last, it was found that a signal estimation using the GP's internal parameters was unsuccessful. However, using the linearity of the GP, it proved possible to estimate the number of signal events in a distribution.

Acknowledgements

Først og fremst, må jeg takke min veileder Alex Read. Dette har vært en komplisert og spennende masterperiode hvor vi har lært masse og hatt det ganske gøy. Takk for at du tar imot alle spørsmål, uansett hvor dumme de er, og at døren din alltid er åpen! Takk for all den kunnskapen du har gitt meg og alle de fine diskusjonene vi har hatt. Du er en fantastisk motivator! En spesiell takk til Eirik S.H. for alle diskusjonene og idéene vi har hatt om GP. Vi har kommet langt! Takk til Heidi for at du er glad i rare og stygge(?) figurer og plott! En stor takk til Øyvind for støtten og kaffen nå i innspurten, og gjennomlesing av denne oppgaven. Jeg trengte et sceneskifte av og til. Jeg må også takke Pumpeklubben Simen, Andreas og Odd Petter for hjelpen med å få igang treningen jeg hadde utsatt så lenge. Slitsomt, men gøy! Jeg vil takke fattern for at du støtter meg i alt jeg vil drive med, og at du ikke vil legge på når vi snakker på telefonen. Og selvfølgelig at du kommer med mat når frysen er nesten tom. Og selvfølgelig storebror Stian, takk for samtalene over den gode ølen! Mastergjengen: Helén, Mona, Oda, og Michael. Det har vært hektisk og mye vanskelig lærdom, men også mye gøy med turer til CERN. Takk for det! HEP-gruppa må jo også få en stor takk! Takk for all støtte og kunnskap dere har delt med meg. Mye quiz, litt fysikk. DnD-gjengene mine må også takkes! Dere har underholdt meg masse de siste årene og fått fokuset litt bort fra jobb og plikt en gang i blant. Også alle de jeg startet og gikk på bachelor sammen med. Altfor mange å nevne dessverre, men dere har alle bidratt til å gjøre studietiden og livet mitt bedre!

Så må jeg jo gi en kjempetakk til samboeren min, Karoline. Uten deg hadde nok ikke dette gått. Takk for at du alltid har vært der for meg når jeg trenger noen å snakke med eller å klemme på! Jeg elsker deg!!!

Til slutt vil jeg dedikere denne oppgaven til min kjære mor. We did it!!!

Contents

Chapter 1

Introduction

In the future of the Large Hadron Collider and ATLAS detector, the amount of data that will be collected will increase by a factor of twenty or so, with a goal of reaching at least 3000 fb^{-1} by 2037. This increase of integrated luminosity does come with a price; the analysis of such event-rich data demand sophisticated analysis procedures and powerful statistical techniques. Two methods utilized in particle physics analysis for modeling the observations are Monte Carlo simulations and data-driven parametrization of functions. This thesis will focus on data-driven modeling, used for the property measurements of the Higgs boson in the two photon final state channel. The data-driven techniques utilized by ATLAS depend on parametric functions to model the backgrounds and estimate possible signals [1]. When more data is taken at ATLAS and this analysis is repeated, the choice of background model must be redone. A function considered for the background modeling must be tested and pass certain criteria imposed on the possible bias of a fit. There are a set of functions that are considered for the background, but at a point in the future these might all fail the criteria, meaning that new functions must be constructed.

This thesis proposes using the machine learning method Gaussian Processes for the data-driven analysis of the $H \to \gamma\gamma$ channel. The method is rarely used in high energy physics. Gaussian Processes are explored in an article by Frate et. al.[2], where it is claimed that background modeling using Gaussian Processes is luminosity independent. The article uses Gaussian Processes as a corrector to a functional fit. This thesis will study the approach proposed by the authors, and see if the luminosity independence can be confirmed. The background modeling and signal estimation will be done fully by Gaussian Processes (which goes slightly beyond the proposition by Frate et. al.[2]).

The scope of this thesis is the conceptual study of applying GP to the

$H \to \gamma\gamma$ channel. First, a simple introduction to the Standard Model is given, followed by a description of the LHC and ATLAS detector, then the method of parametric function fits is explained. An introduction to Gaussian Processes and its properties, including the implementation of GP into software, is described.

The results will thereafter be presented and discussed, followed by the conclusions that can be drawn from this study. The work is not yet complete to the point that the method can without reservation replace the functional parameterization currently in use by ATLAS. Therefore, additional work is needed to complete the study, which is sketched out in the outlook.

1.1 The Standard Model

When conducting experiments in physics, as in all other fields of science, theory is need to describe the phenomenon that is observed. The most successful theory in particle physics thus far is called the Standard Model (SM), an overview is given in figure 1.1. This model is a theory that has been able to both explain the known particles and their interactions, as well as predicting new physics. The last piece of the puzzle, the Higgs boson, was discovered in 2012 [3, 4]. The model separates particles by their intrinsic property called spin, where the matter particles are called fermions and have spin-1/2, and the force carrying particles are called bosons and have spin-1. Spin in this quantum mechanical sense is not the same as one would think of it in the classical sense, as for example a volleyball spinning around an axis, because the elementary particles are in the SM considered to be pointlike objects, i.e. no internal structure.

1.1.1 Fermions

The fermions are called the matter particles, as all matter is made of these. There are 12 fermions which are subdivided into quarks and leptons. Actually, all matter is made of 4 of them as these are the only stable fermions. They are called the first generation as they are the first to be discovered and are the up (u) and down (d) quarks, and they make up the protons $p(uud)$ and neutrons $n(udd)$. A nucleus consists of protons and neutrons, and, together with the electron (e^-), form atoms. The electron neutrino ν_e is not part of this matter composition, but plays an important role in the matter transformation, for example β-decay where a neutron decays into a proton, electron and anti-electron neutrino.

There are three generation of fermions in total. In the second and third

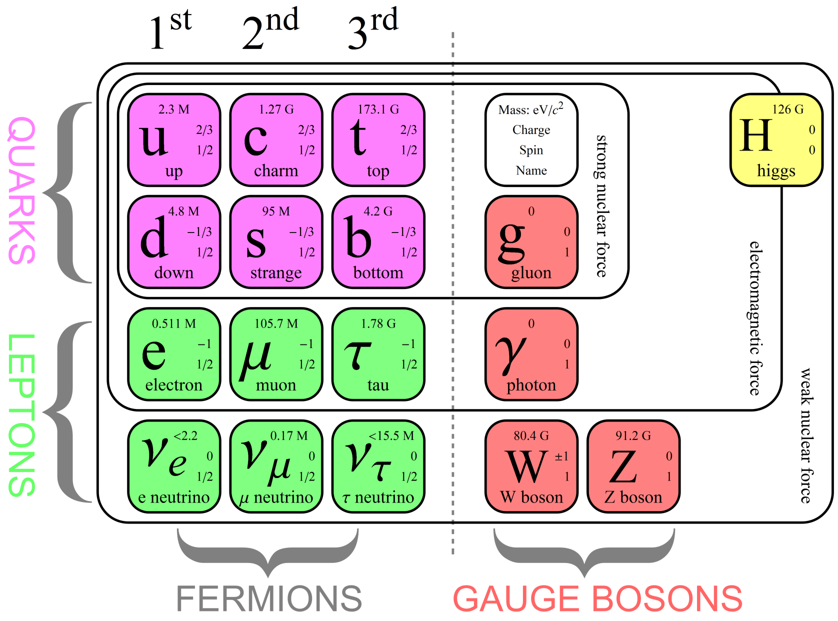


Figure 1.1: Overview over the Standard Model. The figure is taken from [5]

generation the particles have similar properties as the 1st generation, except the masses of the particles. 2nd generation are heavier than the 1st with particles such as the charm (c) and strange (s) quark, and the muon (μ^-) and muon-neutrino (ν_μ). The 3rd generation is heavier still, and contain the top (t) and bottom (b) quark, the tau (τ^-), and the tau-neutrino (ν_τ).

Almost all fermions have an electric charge, neutrinos are electrically neutral, and the charged leptons have integer charges $(e^-, \mu^- \tau^-) : -1$ while the quarks have non-integer charges $(u, c, t) : +2/3, (d, s, b) : -1/3$.

In addition to the electric charge, the quarks have another quantity: color charge. This is, like spin, an intrinsic property that must satisfy certain criteria. There are 3 values which we identify as red, blue, and green. The color charge gives the quarks the possibility to interact via the strong force, in the same way as particles with electric charge experience electromagnetic force. A property of the color charge is that it can not be free, i.e. only bound states of quarks are allowed. And these states must be colorless. There are two possible bound states; one is called *mesons* and consist of a quark-antiquark pair with the same color (antiquark has "anticolor"), and the other is called *baryons* and contain either three quarks or three anti-quarks each with a different color.

1.1.2 Bosons

Bosons are particles with 1-spin and mediate the forces in the SM. The forces are the electromagnetic force, the weak force, and the strong force. Gravity is not represented in the SM.

The *photon* is the boson mediating the electromagnetic interactions. It has not mass, is electrically neutral, and is stable. This mean that the electromagnetic range is infinite.

For the weak interactions, there are three carriers: W^-, W^+, and Z^0. These particles are heavy, only trumped by the top quark and the Higgs boson, and therefore decay quickly giving the weak force a very limited range of only $\approx 10^{-15}$m.

In the strong interaction, *gluons* are the mediator. This boson only interacts with particles having a color charge, like the quarks. It can also interact with other gluons, as they have both color *and* anticolor, and their combinations are: $b\bar{g}$, $g\bar{b}$, $r\bar{g}$, $g\bar{r}$, $r\bar{b}$, $b\bar{r}$, $\frac{1}{\sqrt{2}}(r\bar{r} - g\bar{g})$, $\frac{1}{6}(r\bar{r} + g\bar{g} - 2b\bar{b})$. When two color charged particles increase in distance, the strong force between them increases as well. This is in stark contrast to the weak and electromagnetic force which decreases for increasing distance, the weak force more rapidly than the EM force. This gives rise to a phenomenon called *asymptotic freedom*: at short distances, less than $\approx 10^{-15}$ m, the quarks and gluons can move freely. Outside this radius, the particles will be subjugated to color confinement. I.e. if two quarks are forced apart, the energy applied will create a quark-antiquark pair that then again confines the original quarks and preserves the colorless state.

One of the big questions about the SM was how the mass of the particles came to be. If the gauge bosons have mass, then the electroweak symmetry is broken and the SM would not be a consistent theory. A solution to this was the Brout-Englert-Higgs mechanism [6, 7]. This mechanism works though spontaneous symmetry breaking of gauge symmetry. The mathematical details will not be discussed in this thesis. To give an example of spontaneous symmetry breaking, consider a finite rod. The rod will be rotational symmetric around its axis. If a force is applied to the rod at one end, the rod will bend and no longer be rotationally symmetric, but the symmetry is not spontaneously broken, just broken. Spontaneously means, in this sense, that the direction of the bending can not be known beforehand. To illustrate the difference, the force are now applied at both ends towards the rod parallel with the axis. The rod will then, when a critical amount of force is applied, bend in a random direction. Then the rotational symmetry is spontaneously broken. An illustration of this for the gauge symmetry breaking is given in figure 1.2.

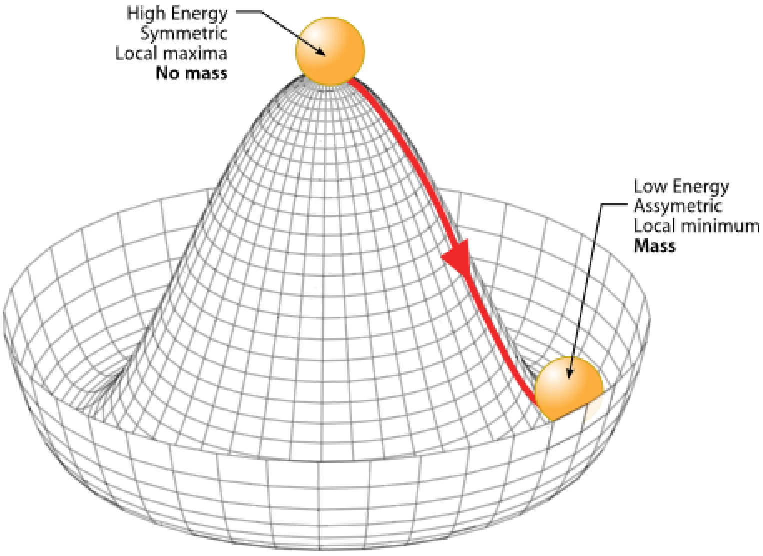


Figure 1.2: An illustration of the Higgs potential and how particles acquire mass. The figure is taken from [8]

The Brout-Englert-Higgs field has a potential as shown in figure 1.2, and when particles fall into the well, the energy-difference gives them mass. If a particle does not interact with the BEH-field it will be massless like the photon, and stronger interaction means larger mass.

An excitation of the field itself, gives rise to the Higgs boson, proposed by Peter Higgs in 1964[6]. This particle has spin-0 meaning it is a scalar particle and is electrically neutral. It was the last piece of the SM-puzzle, and was discovered by ATLAS and CMS in 2012. [3] This confirmed the existence of the BEH-field.

1.2 The LHC and the ATLAS detector

1.2.1 LHC

The accelerator

The Large Hadron Collider is the largest and most powerful particle accelerator in the world. With its 27 km circumference it can accelerate bunches of 10^{11} protons to high energies that are collided at experiment locations 40 million times per second. LHC is designed to deliver a center-of-mass energy of $\sqrt{s} = 14TeV$ at an integrated luminosity of $10^{34}cm^{-2}s^{-1}$. There are four large experiments (and a number of small ones) on the LHC ring: ATLAS, LHCb, ALICE and CMS. ATLAS and CMS are general purpose detectors able to study a wide spectrum particles, LHCb specializes in the study of the bottom-quark, and ALICE is a heavy ion collision detector.

Particle collisions

In high energy particle accelerators like the LHC, there are two quantities that are the most important: the center-of-mass energy $\sqrt{s}$ and the instantaneous luminosity $\mathcal{L}$. Each of these properties will be discussed below. To accelerate the particles to the energy desired they need to be charged and stable. There are many types of accelerators, such as hadron colliders, electron-positron colliders, and electron-proton colliders.

The center-of-mass energy in the collision is given via the invariant quantity s that is the sum of the energy and momentum of the two interacting protons:

$$s = (E_1 + E_2)^2 - (\mathbf{p}_1 + \mathbf{p}_2)^2 = m^2 \tag{1.1}$$

In the above equation 1.1 natural units are used, i.e. c = 1. The center-of-mass energy determines the heaviest particles that can be produced, as the largest mass a produced particle can have is $m = \sqrt{s}$. In the interaction the energy is usually less than $\sqrt{s}$ as the protons are composite particles of quarks and gluons. They are the interacting components in the hard scattering process and are collectively called *partons*. Only have a fraction of the total proton momentum is carried by these partons, and the probability for a parton, of a certain quark flavor, to have a certain momentum is detailed in the Parton Distribution Functions. The highest center-of-mass energy that current technologies can achieve is desired, to extend the range of particles that can be produced, discovered and studied.

Reaching high enough energies is one part of the problem of discovering new physics. Another part is the fact that all of the particles in the SM has a certain probability to be produced. Heavier particles have less probability to be produced in the collisions than the light ones. For example the Higgs boson is heavier than most particles in the SM, so it is seldom created. How can enough Higgs bosons be produced to observe them? The solution to this is given in the instantaneous luminosity, $\mathcal{L}$. This is the measure of the number of collisions per second and per cm^2 that can be produced. Assuming that the beams collide head on having a Gaussian profile, the instantaneous luminosity can be calculated as [9]

$$\mathcal{L} = f\frac{n_1 n_2}{4\pi\sigma_x\sigma_y} = 10^{34}cm^{-2}s^{-1}, \tag{1.2}$$

where n_1 and n_2 is the number of particles in each colliding bunch, f is the frequency of the beam crossings, and $\sigma_x\sigma_y$ is the root-mean-square of the beam size in the xy-plane. Unfortunately the properties of the bunches are not known exactly. To find the number of events for a process, the instantaneous luminosity is integrated with respect to time. The luminosity has the units $cm^{-2}s^{-1}$, but for the integrated luminosity the units are the inverse femtobarn fb^{-1} which is a measure of the number of particle collisions per cross-section. Multiplied with the cross-section, the probability for the interaction to occur with, for a process gives an estimate of the number of events for that process.

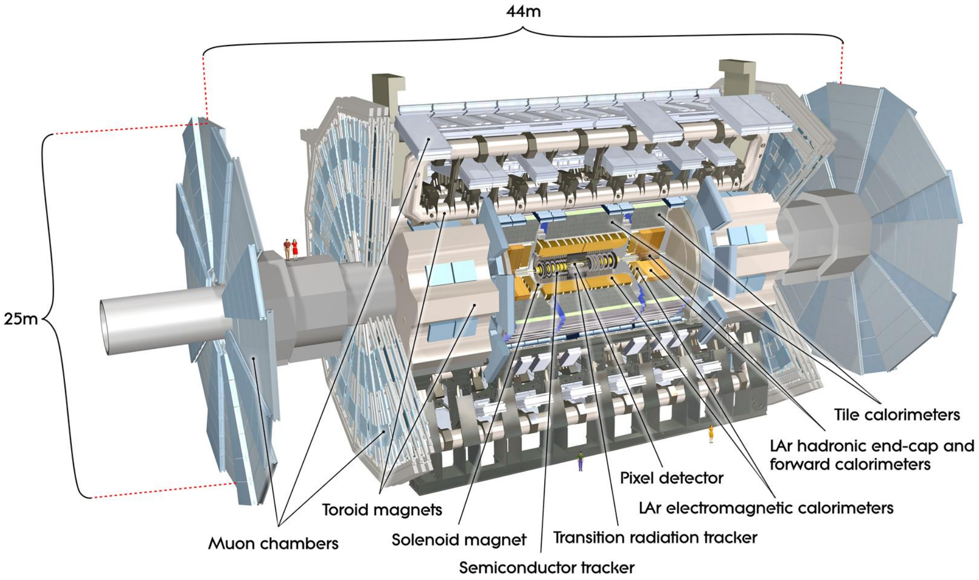
$$N = \sigma \int \mathcal{L}(t)dt \tag{1.3}$$

From equation 1.3 the importance of the luminosity is obvious: higher luminosity, more interactions. The running period for the experiment also increases the number of events accumulated. In the late fall 2018 Run 2 was completed with a total luminosity of $189.3fb^{-1}$ [10], and the LHC is now in the process of upgrading towards Run 3 beginning in 2021. The expected integrated luminosity for Run 3 is $\approx 300fb^{-1}$. During the shut down in 2022-24, LHC will be upgraded to the High Luminosity LHC for Run 4, which has an expected luminosity of $1000fb^{-1}$.

1.2.2 ATLAS

In this section, an overview of the ATLAS detector will be given, based on [11]. ATLAS is a general purpose detector constructed for studying a large variety of particles and phenomena produced in the pp collisions from the LHC. The detector is has a cylindrical shape and is forward-backward symmetric. A side-cut of the detector is given in figure 1.3. The coordinate system in the detector is a Cartesian system with origin in the collision point, the z-axis points along the beam line, the x-axis points to the center of the LHC ring, and the y-axis is pointing straight up. As the detector is cylindrical, angles are also used for geometry. ϕ is the azimuthal angle around the beam line and has the interval $(0, 2\pi)$ and the polar angle θ, giving the direction relative to the beam line, in the interval $0, \pi)$. A preferred quantity is the Pseudorapidity η which is defined from the polar angle as

$\eta = -\ln\tan(\theta/2)$. And in this ϕ–η-plane, the distance ΔR between particles is given as $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$.



44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

Figure 1.3: The ATLAS detector [11]

Inner detector

In the Inner Detector (ID) particles from the pp-collisions are tracked as they interact with the layers of the ID, and by using these hits the particle trajectories and the interaction vertices can be reconstructed. Figure 1.4 shows the setup of the ID. The combination of the track detection and a 2 T magnetic field is used to identify charged particles as these are bent in the field, and the direction of the curve gives the sign on the charge. The ID comprised of TRT, SCT, and Pixel detectors which covers the region $|\eta| < 2.5$.

The part of the ID furthest from the interaction point is the Transition Radiation Tracker (TRT), consisting of xenon-based gas filled straw tubes. These tubes gives higher electron identification capability.

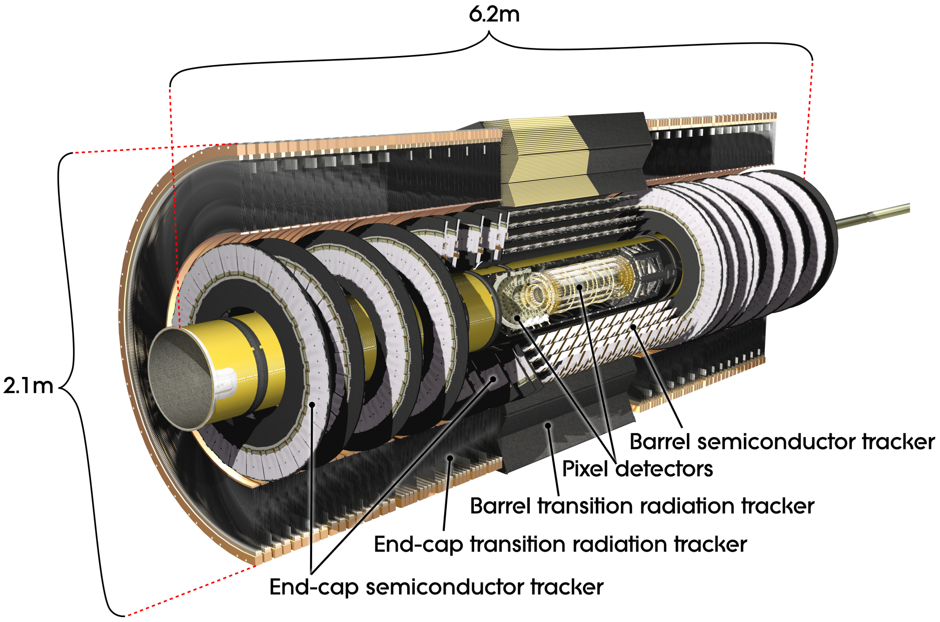

6.2m
2.1m
Barrel semiconductor tracker
Pixel detectors
Barrel transition radiation tracker
End-cap transition radiation tracker
End-cap semiconductor tracker

Figure 1.4: The Inner Detector of ATLAS [11]

Calorimeters

In the calorimeters the energy of the final state particles are measured, and the range they cover is $\eta > 4.9$. The electromagnetic calorimeter (ECal) is located after the ID and absorbs the energy of electrons, positrons, and photons. It is divided into three parts; the barrel, and the two end-caps. The barrel covers $|\eta| < 1.475$ and the end-caps cover $1.37 < |\eta| < 3.2$. The energy is measured using liquid argon and lead absorber plates. The ECal has high precision in the energy measurement, with resolution $\sigma_E/E = 10\%/\sqrt{E} \oplus 0.7\%$.

Wrapped around the ECal is the hadronic calorimeter (HadCal), which absorbs the energy of hadrons. It is made of steel for absorption and scintillating tiles as the active material. The barrel region covers $|\eta| < 1.0$ with the extended barrel covering $0.8 < |\eta| < 1.7$.

The end-caps of the HadCal consist of copper plates as absorption material and liquid Argon as the active material. The coverage of the end-caps is $1.5 < |\eta| < 3.2$.

To get a fuller coverage the forward calorimeter is integrated into the end-cap and is divided into three parts. One made from copper for electromagnetic measurements and two others built with tungsten for hadronic measurements. This region covers $3.1 < |\eta| < 4.9$.

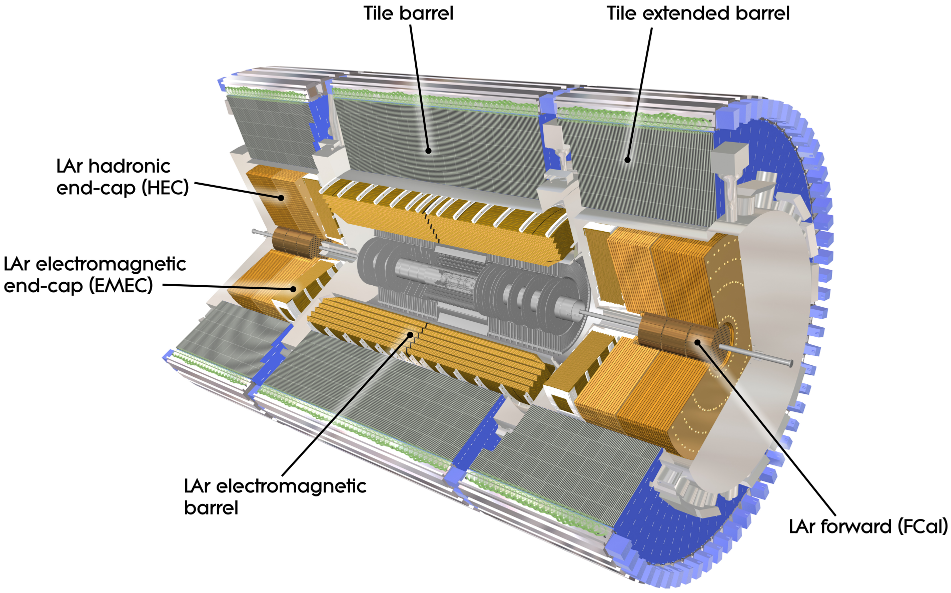

Tile barrel
Tile extended barrel
LAr hadronic end-cap (HEC)
LAr electromagnetic end-cap (EMEC)
LAr electromagnetic barrel
LAr forward (FCal)

Figure 1.5: The Calorimeters of the ATLAS detector[11]

Muon system

The muon spectrometer is located outside the HadCal. This detector is not meant to absorb the muons, just detect their passage through the muon system. The system is based on magnetic deflection of muons in the toroid magnets or in the end-cap magnets. The former has a coverage of $|\eta| < 1.4$ while the latter has a coverage of $1.6 < |\eta| < 2.7$. In the so called transition region, $1.4 < |\eta| < 1.6$, the deflection is provided by a combination of both the toroid and the end-cap magnets. The muon spectrometer provides a resolution of $\sigma_{p_T}/p_T = 10\%$ at $p_t = 1TeV$, which is achieved by three layers of precision chambers.

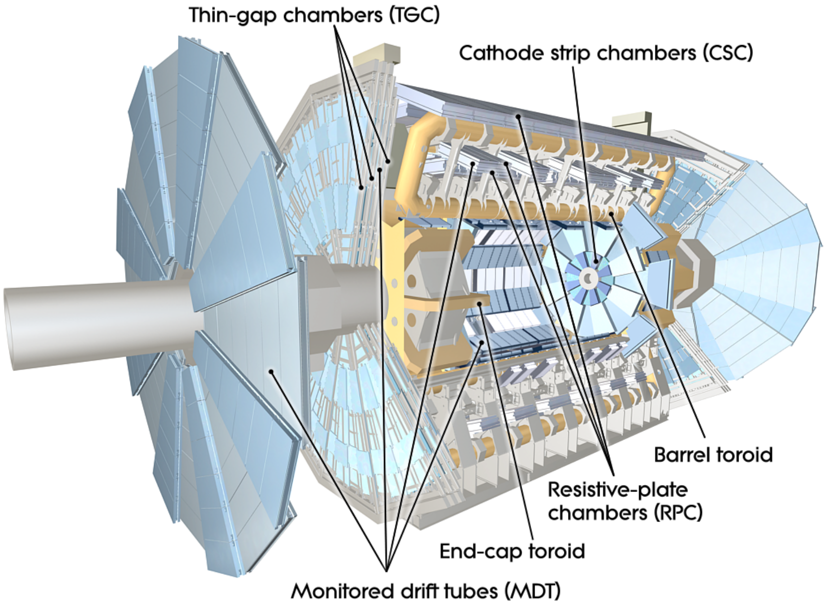

Thin-gap chambers (TGC)
Cathode strip chambers (CSC)
Barrel toroid
Resistive-plate chambers (RPC)
End-cap toroid
Monitored drift tubes (MDT)

Figure 1.6: The Muon system of the ATLAS detector[11]

Trigger system

The trigger system is required to limit the amount of data stored. This system ensures that only interesting events containing high p_T particles are selected for storage, and the Level 1 trigger is a hardware trigger that searches for such events. When L1 triggers, the information of the event is passed to the software based Level 2 trigger. At this level, the trigger use the full information to select events of interest. The last stage in the trigger system, is the event filter which use offline analysis procedures. The decision time for each trigger is less than 2.5μs for L1, $\approx$40 ms for L2, and at the order of 4 seconds for the event filter.

1.3 The Higgs production and event selection

Why is the H→ $\gamma\gamma$ channel used to study the Higgs boson? There are many other channels that the Higgs boson can decay to. Figure 1.7 shows the branching ratios of different processes, giving a indication of which decay channels are more probable. $b\bar{b}$ has the highest branching ratio at the Higgs mass, followed by WW, gg, and $\tau\tau$. But the QCD interactions give these channels huge backgrounds, and the ratio of signal to background is tiny, therefore it is difficult to separate the signal from the background.

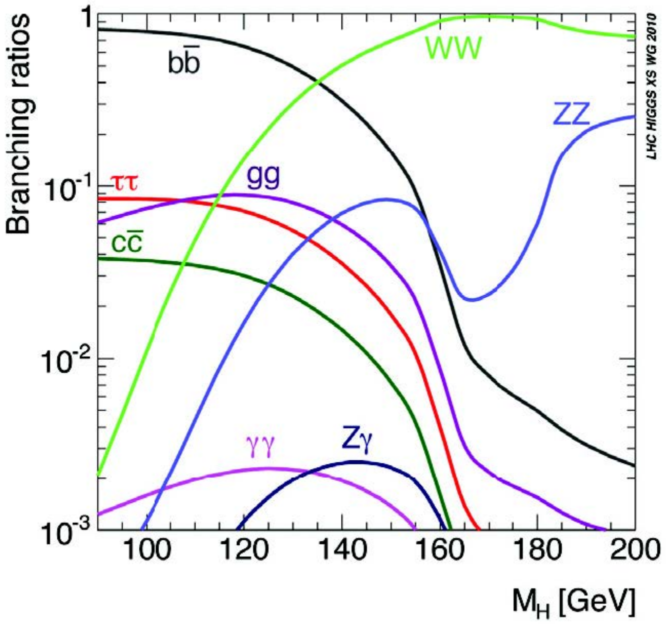


Figure 1.7: This figure show Branching ratios for different decay channels vs the Higgs mass.

In this section the photon identification in the detector is briefly discussed, and from what processes they are created. Also other factors for the background will be touched upon.

Detecting photons in the ATLAS detector, is done via the Inner Detector and the Electromagnetic Calorimeter. Particles moving through the ID are tracked, but the photons do not interact with the tracking material. They deposit all their energy in the ECal in electromagnetic cascades. In the reconstruction of the collision event, the photons are identified as a electromagnetic shower exceeding some energy level that does not have a track in the ID leading to the deposit. An example of this is seen in figure 1.8, where the dashed line indicate no interaction with the material.

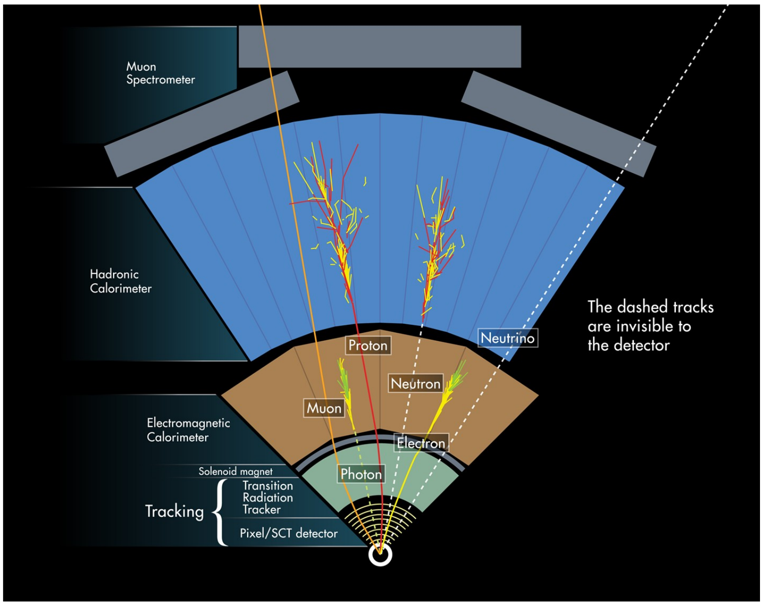


Figure 1.8: This figure show the layers of the ATLAS detector and how particles move and interact within them.

1.3.1 The background processes

The background is the name given to non-Higgs processes that either have the same final-state as the signal process, or where measurements in the detector are misidentified as photons passing the identification criteria. The former is called the irreducible background, and the latter the reducible background. As the names suggest the reducible background can be dealt with to some degree, and the irreducible part is where it can not be known if the photons comes from a Higgs boson decay or some other process. The irreducible processes are the Box process, the Born process, and the Bremsstrahlung process. The Feynman diagram for the Box process can be seen in figure 1.9, the Born diagram in figure 1.10, and Bremsstrahlung in figure 1.11. These give a "clean" spectrum, as the distribution of $m_{\gamma\gamma}$ is monotonic and decreasing.

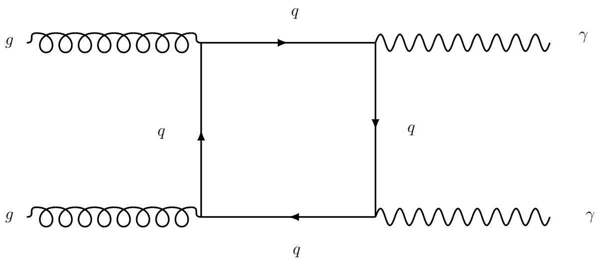


Figure 1.9: This is the Feynman diagram for the Box process. Here two gluons interact with a quark-box where the final state is two photons.

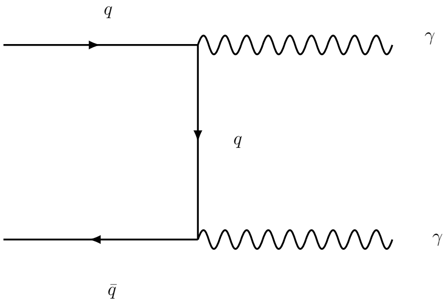


Figure 1.10: This is the Feynman diagram for the Born process of quark-antiquark pair annihilation into two photons.

The reducible background is where the reconstruction misidentifies other particles as photons, and consist mostly of jets and π^0s. They are γ-jets, jets-jets, and the Drell-Yan process. The DY process is when a $q\bar{q}$-pair annihilate

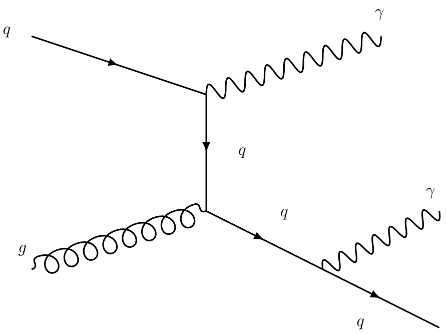


Figure 1.11: This is the Feynman diagram for the Bremsstralhung process. A quark radiate a photon by interacting with a gluon, then ratiates a photon.

to a photon that produce a e^+e^--pair. These electron are then misidentified as photons. For the γ-jet, the jet is misidentified as a photon, and the same is the situation for jet-jet. Here both jets are identified as photons.

1.3.2 The signal process

The Higgs decay process producing two photons is through a W-boson or top-quark loop. The Feynman diagram for this process is shown in figure 1.12

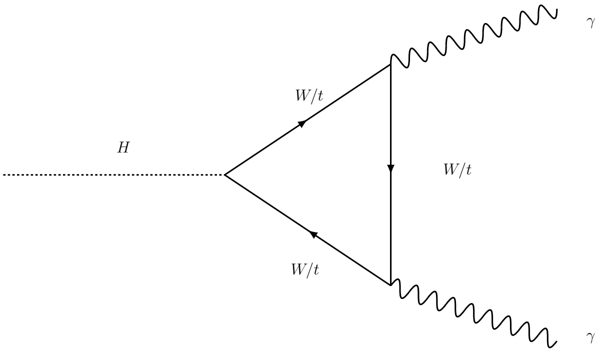


Figure 1.12: This is the Feynman diagram for the Higgs decay process. Here a produced Higgs boson decay through a W-boson loop to two photons. The loop might also be top-quarks.

Chapter 2

Analysis

2.1 Data driven methods

The most common methods for analysis utilize Monte Carlo (MC) simulations of the production and decay modes for different particles in the collisions that are based on the SM and hypothetical theories that go beyond the SM, as well as how the particles interact with the detector. One example of MC background simulation is shown in the bottom plot of figure 2.1, where the data is the black dots, the background is estimated using simulated events in red and purple and the blue is the simulated signal. The top plot of figure 2.1 is the 2012 result for the $H \to \gamma\gamma$ channel using a data driven method for estimating the background and signal. In the study of $H \to \gamma\gamma$ data driven methods are used instead of MC to estimated the background. The two photon final state is well populated, but the signal strength relative to the background is very small. This means that small fluctuations can have a large impact on the sensitivity of the signal. To model the background a suited function must be chosen. The signal model is a double sided Crystal Ball function, which is explained later. To then test which hypothesis, background-only or signal+background, that describes the data best, a test-statistic is used. The procedure will be described after model selection.

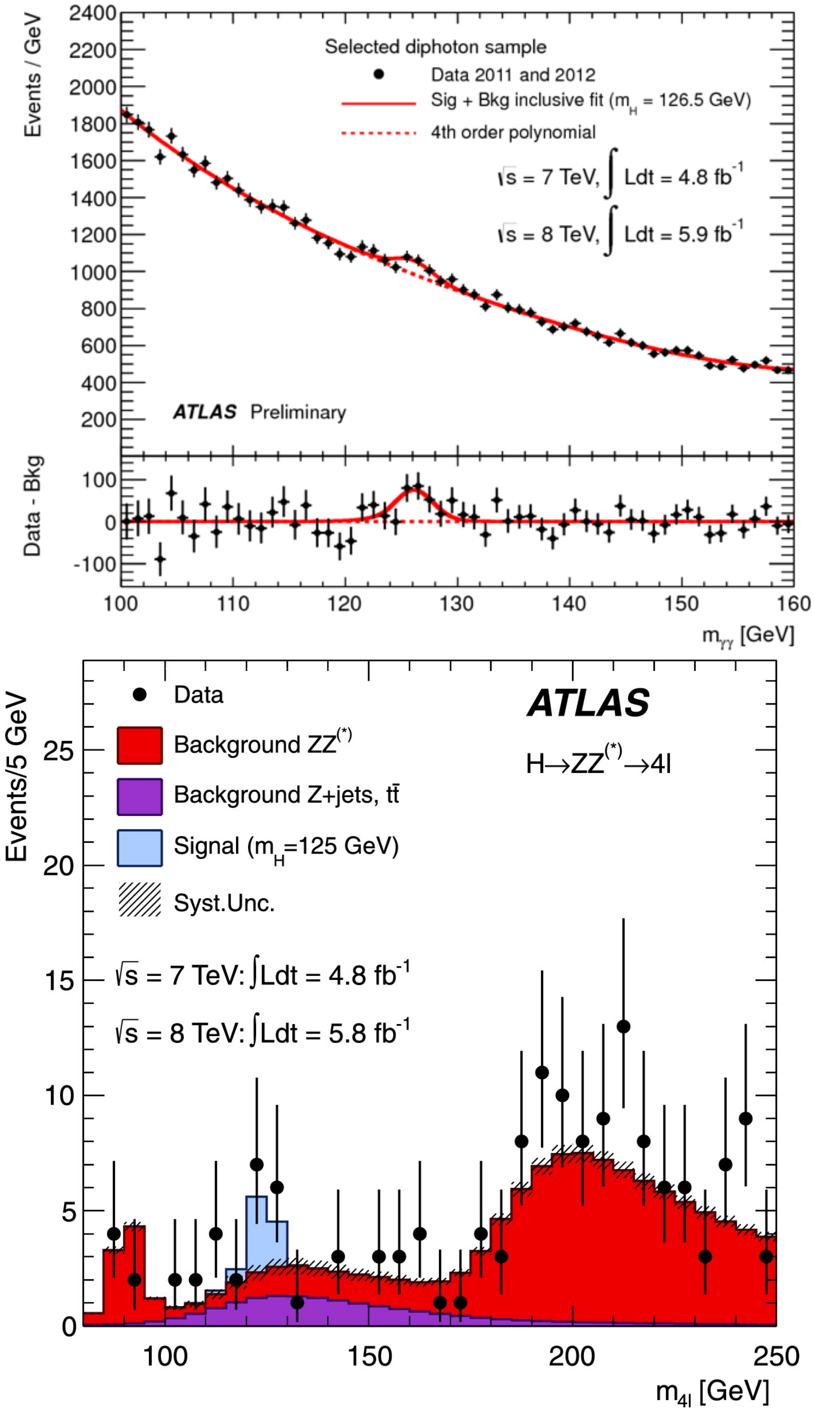


Figure 2.1: This figure display two of the main results of Higgs searches in 2012 with the ATLAS detector. The top plot is the two photon final state, and the bottom is the four leptons final state. The figures are taken from the 2012 article[3].

2.1.1 Background model

To be able to discern the signal from the background, precise modeling of the background distribution must be performed. The distribution for $H \to \gamma\gamma$ in figure 2.1 seems to be following a falling exponential or polynomial function, but the underlying functions are unknown. This presents the challenge to choose which function should be used for the background modeling. Functions from different function families will have difficulty describing one another. These differences give an indication of the systematic error one makes when choosing a particular function family. In figure 2.2 an example of this is provided. The polynomial fit function does not adequately represent the underlying exponential function. If more freedom is added to the fit function as parameters, the fit might approach the model but it will start to overfit, i.e. it will start to absorb also potential signal bumps, ceasing to be a general solution. As the amount of data increases, the ability of a previously chosen function to correctly model the background over the full relevant spectrum is compromised, and the procedure of selecting the function that describes the background must be redone. The functions must then pass some criteria, to be described below, and the one with the least degrees of freedom is chosen.

The different functions considered for the background modeling are [1]

- Exponential $= Ae^{-\xi x}$
- Exponential polynomial $= e^{\sum_{i=0}^{n} \alpha_i x^i}$
- Bernstein polynomial $= A\sum_{i=1}^{n} c_i C_n^i u^i (1-u)^{n-1}$, $u = \frac{m - m_{min}}{m_{max} - m_{min}}$ [12]

There have been considered combinations of functions as well, for example the double exponential $= A(ae^{-\xi_1 x} + (1-a)e^{-\xi_2 x}$ and functions with a turn on in the lower part of the invariant mass distribution, but none have been stable in the fitting. The fits must always converge (this is most important when studying the performances of such fits with large numbers of MC “toys”) as there can not be any doubt that the optimization of the parameters will find a minimum.

A method of finding these optimized parameters, is to use the method of Least Squares

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - f(x_i;\theta))^2}{\sigma_i^2} \tag{2.1}$$

where $f(x;\theta)$ is the predicted value in bin i given θ, and y_i is the observed value. σ_i is the uncertainty on the observed value. In this method the uncertainty for each observation is assumed to be Poisson distributed giving $\sigma_i = \sqrt{y_i}$. When the data values are large, $y_i > 10$, the Poisson uncertainty can be approximated as Gaussian. And it is recommended to use Pearsons χ^2 [13], which uses the expected number of events to specify the uncertainty. Using the observations as errors can introduce some bias as the values are Poisson-shifted, but the expected number from a fitted function is an average. The $\boldsymbol{\theta}$'s that minimizes this function, are chosen as the optimized parameters.

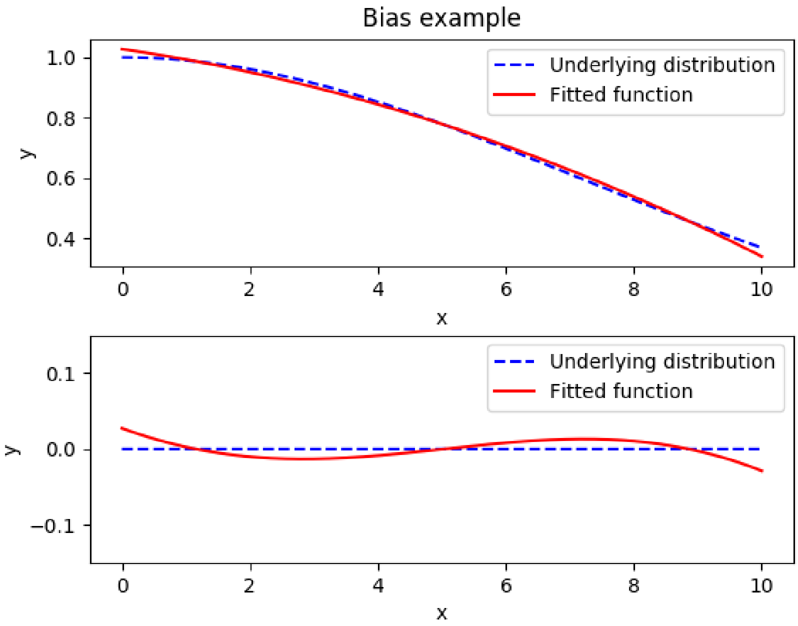


Figure 2.2: This figure show how bias arise when a fitting function (a 2-degree polynomial) does not describe the underlying distribution (an exponential of a 2-degree polynomial). The blue line is the underlying function and the red is the best fitted function. The top plot is the distribution, and the bottom is the residuals. The red line overestimates at the left edge and in the middle right, and underestimates at the right edge and in the middle left. If a signal was present at $x \sim 3$, the signal strength would be overestimated. If a signal was present at $x \sim 7$, the signal strength would be underestimated.

2.1.2 Spurious signal

An important test of the function fits, is the test of bias. If a function fitted to the data has some systematic deviation in the residuals, which is the difference between the fit and the background model, then the model is biased. This deviation must be accounted for since such residuals in the fits to real data could easily be mistaken as signal. A signal+background function is then fitted to the background-only distribution, and the signal yield from this fit is called a Spurious signal. The lower plot in figure 2.2 displays an example of a bias in the fit. Fitting a signal+background model to this example distribution and the fitted signal has a non-zero amplitude, this signal is a spurious signal. The possibility of such a deviation is implemented in the likelihood function that is used to optimize the parameters in the function. There is certain requirements [1] of this test of which the fit need only pass one:

- The spurious signal is not larger than 10% of the expected signal
- The spurious signal is not larger than 20% of the spurious signal uncertainty.

Expected signal in this context is the expected signal under the hypothesis of a SM Higgs boson with the mass being tested.

2.1.3 Signal model

The signal is typically modelled as a double-sided Crystal Ball (CB) function [1]. Modelling the signal as a Gaussian peak is a simple approximation that can be easily implemented and tested on. The CB-function on the other hand, is used to fully model the signal as it takes into account the non-Gaussian form in the low-(-) and high-mass(+) tails, parametrized by $\alpha^{\pm}_{CB,i}$ and $n^{\pm}_{CB,i}$. It consist of a Gaussian peak in the center with a power law in both tails.

$$f_i^{sig}(m_{\gamma\gamma};\Delta\mu_{CB,i},\sigma_{CB,i},\alpha^{\pm}_{CB,i},n^{\pm}_{CB,i}) = \mathcal{N}_c \begin{cases} e^{-t^2/2} \\ (\frac{n^-_{CB,i}}{|\alpha^-_{CB,i}|})^{n^-_{CB,i}} e^{-|\alpha^-_{CB,i}|^2/2} (\frac{n^-_{CB,i}}{\alpha^-_{CB,i}} - \alpha^-_{CB,i} - t)^{-n^-_{CB,i}} \\ (\frac{n^+_{CB,i}}{|\alpha^+_{CB,i}|})^{n^+_{CB,i}} e^{-|\alpha^+_{CB,i}|^2/2} (\frac{n^+_{CB,i}}{\alpha^+_{CB,i}} - \alpha^+_{CB,i} - t)^{-n^+_{CB,i}} \end{cases} \tag{2.2}$$

where $t = \frac{m_{\gamma\gamma}-m_H-\Delta\mu_{CB,i}}{\sigma_{CB,i}}$, and $\mathcal{N}_c$ is a normalization factor. When $-\alpha^-_{CB,i} \leq t \leq \alpha^+_{CB,i}$, the first function is applied; when $t < -\alpha^-_{CB,i}$, the second function is applied; last, when $t > \alpha^+_{CB,i}$, the third function is applied.

2.1.4 Test statistics and the likelihood function

To be able to claim a discovery of new physics, or to reject its existence, a statistic is used to test hypotheses. These hypotheses are usually the background-only, meaning that the background distribution is the underlying function, and the signal+background hypotheses, where an excess of events above the background is predicted by the signal being tested. The expectation value for the numbers of events is $E[n_i] = \mu s_i + b_i$, where s_i and b_i are the mean predicted number of events in bin i, and μ is the parameter indicating the signal strength, with $\mu = 0$ corresponding to the background hypothesis and $\mu = 1$ as the nominal signal hypothesis. The number of signal and background events are found from [14]

$$s_i = s_{tot} \int f_s(x; \theta_s) dx \tag{2.3}$$

$$b_i = b_{tot} \int f_b(x; \theta_b) dx \tag{2.4}$$

where $f_s(x; \theta_s)$ and $f_b(x; \theta_b)$ are the chosen model with some parameters, giving the probability distribution functions for the signal and background. To test a specific value of μ, the so-called *profile likelihood ratio* is used:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})}. \tag{2.5}$$

The numerator of equation 2.5 is the conditional likelihood function, conditional since μ is fixed and $\hat{\hat{\theta}}$ is the value of this estimator which maximizes the likelihood in this case. The denominator is the unconditional likelihood; both the parameters θ and μ are free, and $\hat{\theta}$ and $\hat{\mu}$ are the fitted estimators which maximizes the likelihood for this case.
Multiplying the Poisson probabilities for each bin, the likelihood is written as

$$L(\mu, \theta) = \prod_{i=1}^{N} \frac{(\mu s_i)^{n_i}}{n_i!} \exp(-(\mu s_i + b_i)) \prod_{j=1}^{M} \frac{u_j^{m_j}}{m_j!} \exp(-u_j) \tag{2.6}$$

u_j is here the expectation value assuming only background. In equation 2.5 it is observed that $0 \leq \lambda \leq 1$, where ~ 1 means the tested value for μ agrees well with the data given the model being tested. This is used to test different signal strengths which can be compared to the prediction from the Standard Model.

2.1.5 ROOT and Minuit

The framework for doing analysis in particle physics, is the CERN-created ROOT package [15]. This package is developed at CERN over many decades. It implements the jargon used by particle physicist to present data and analysis results in a similar manner, and makes life easier for its users as there are many elements to handle at once. ROOT utilizes MINUIT2, a package for minimization in the analysis methods. This package is also developed at CERN. Both of these packages are written in C++ which is the most common language used by the different research groups. as it is quite fast when doing iterations as well as being object-oriented. This is a vital part of a cut-and-count analysis. Here cuts are used to remove the uninteresting events from the analysis. For example, which particles are in the final-state and how their properties are, might decide if they are included. There are millions of events going through a yes/no-type algorithm, therefore speed is desired. This is what C++ excels at. PYTHON is starting to be more used as it is more user-friendly. Declaration of variables is not needed, and how graphics are handled is much better. It is much slower though, therefore a combination is often used; C++ for iteration and PYTHON for data handling.

2.2 Gaussian Processes

Gaussian Processes, abbreviated GP, is a machine learning method based on Bayesian statistics, which performs prediction of data based only on the data and the knowledge about the data. As stated in the introduction, precise measurements of the Higgs boson are still desired, but the methods used now are slow and problematic. A solution proposed is using GP for background modeling and signal estimation. In this chapter the basics of GP are first presented, before moving onto how GP is utilized for the analysis of Higgs boson decays to 2 photons. An introduction to Bayesian statistics is first made and the connection to the Gaussian distribution is presented, which is based on the more detailed derivation in the book by D.S. Sivia [16]. Then GP regression is introduced, as outlined in the book by C. E. Rasmussen and C. K. I. Williams [17]. The last part specifies the background and signal modeling.

Gaussian processes is used in many different fields, as it is diverse and brings a broad space of usage: Geostatistics, astrophysics, climate studies, and more.

2.2.1 Introduction to Bayesian Statistics

The statistics used in the method described in the parametric function section is called *frequentist*, while Gaussian Processes on the other hand, is based on *Bayesian* statistics. The fundamental difference between the two, is in the way they view uncertainty. Both use probability to describe this uncertainty, so the definition of probability must then also be different. The frequentist way handles the uncertainty probability as objective. This means that the uncertainty is fully described from the outcome of the experiment. An example is a coin flip; the probability for heads or tails are 50-50, so both outcomes are equally uncertain. If the uncertainty is handled in a Bayesian way subjective uncertainty is added, which introduces other factors: is the coin bent in any way? Is the weight shifted one way? Maybe even the person performing the flip is skilled in counting the rotations and can easier get a preferred side? This is described as prior knowledge, and will be discussed further in this chapter.

Probability

Probability is defined as how likely it is for something to occur. Looking again at the coin flip, there are two possible outcomes and one of these are chosen. Then the probability is deduced relative to this. The probability for say heads, is

$$P(H) = \frac{\text{Wanted outcomes}}{\text{Possible outcomes}} = \frac{\text{Heads}}{\text{All}} = \frac{1}{2} \tag{2.7}$$

Another important definition is $P(H) + P(\bar{H}) = 1$, where the bar above the H implies *not heads*, and the sum of probabilities for outcome H and not outcome H must be 1 as one of them must happen. Also, if the coin is flipped twice, what is the probability for one of each outcome? Assuming Heads in the first toss then Tails, this can be written as

$$P(H,T) = P(T|H)P(H), \tag{2.8}$$

where the probability for Heads and Tails are the product of the probability for Heads and the probability of Tails given Heads in the first flip. This rule holds if the outcome Tails is independent on Heads: $P(T|H) = P(T)$. Eq. 2.8 assuming Tails then Heads is $P(T,H) = P(H|T)P(T)$, and the probability for both to occur in must be the same, giving $P(T,H) = P(H,T)$ showing that the probability for both outcomes happening is independent of the order, $P(H|T)P(T) = P(T|H)P(H)$. This equality can be generalized for some variable X and some parameter θ: $P(\theta|X)P(X) = P(X|\theta)P(\theta)$. This leads to

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{2.9}$$

which is known as *Bayes' rule*, derived by Thomas Bayes [18].

A much used object in this form of statistics, are Probability Distribution Functions (PDF) describing how the probability of some quantity is distributed. Bayes' rule provides a posterior distribution function for the parameter θ given the variable X. An example where this quantity is interesting, is for a dice thrown. One question is what outcome X to expect based on θ. Another, more interesting, question is: based on the observations of the dice throws, what is the probability that the dice is fair? This is the information gained from Bayes' rule. In the rule, $P(\theta)$ is the *prior* probability for the parameter derived from the known information, e.q. if the coin from before is bent, $P(X|\theta)$ is the *likelihood*, and $P(X)$ is the *marginal likelihood.* The likelihood modifies the prior information based on the observations, and the marginal likelihood is the probability for the observations without the parameters. This is written as

$$P(X) = \int P(X|\theta)P(\theta)d\theta \tag{2.10}$$

The whole point with this derivation, is to show how some assumption for a parameter of interest can be updated by introducing observations. This is the main difference between bayesian and frequentist approaches; the usage of prior information.

The probability distribution for a parameter will, according to the Central Limit Theorem [19], converge to a Normal distribution for a large number of observations. This is utilized for the Gaussian Processes Regression. For the Normal distribution there are two parameters, the *mean* and the *variance.* The mean is the most likely value for the parameter and is written as

$$m(\theta) = \int \theta P(\theta)d\theta \tag{2.11}$$

and the variance is, given the mean value θ_0,

$$V(\theta) = \int (\theta - \theta_0)^2 P(\theta) d\theta \tag{2.12}$$

What is then the mean and variance for the Gaussian distribution of the posterior PDF? The most likely value θ_0 is the value that maximizes $P(\theta|X)$, therefore two requirements must be met:

$$\frac{dP(\theta|X)}{d\theta}\bigg|_{\theta_0} = 0 \tag{2.13}$$

$$\frac{d^2P(\theta|X)}{d\theta^2}\bigg|_{\theta_0} < 0 \tag{2.14}$$

As a Gaussian is used to approximate the value of the parameter θ, how is the mean and variance relative to that? The mean of the Gaussian is the most likely value, and the variance is needed to understand how uncertain the value is. The PDF of the posterior can be Taylor expanded around the mean value. Setting $L = \ln(P(\theta|X))$ for simplicity

$$L = L(\theta_0) + \frac{dL}{d\theta}\bigg|_{\theta_0} (\theta - \theta_0) + \frac{1}{2}\frac{d^2L}{d\theta^2}\bigg|_{\theta_0} (\theta - \theta_0)^2 + ... \tag{2.15}$$

The first term is a constant A, and the second is zero per the assumption of maximum probability. Ignoring higher order terms as they are much smaller than the third term and taking the exponential to return to the posterior function

$$f(\theta) \approx A \exp\Big(\frac{1}{2}\frac{d^2L}{d\theta^2}\bigg|_{\theta_0} (\theta - \theta_0)^2\Big) \tag{2.16}$$

This equation is recognised as having a Gaussian shape. The normalized general Gaussian distribution is written

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\Big) \tag{2.17}$$

where the parameters μ and σ^2 is the mean and variance respectively. To find the mean and variance from the Taylor expansion, a comparison is made between the two functions. The mean μ in 2.17 is identified to be θ_0 in 2.16, and the variance σ^2 must then be $\Big(-\frac{d^2L}{d\theta^2}\Big)^{-1/2}$.

It is now established that the probability distribution for a parameter or variable can be approximated as a Gaussian distribution, and that this approximation can change according to observed data. But so far this is only established for one variable. GP usually takes into account more than one covariate where one covariate can have more than one value. Each of these values corresponds to one observed value. What would the approximated Gaussian look like if another variable,ϕ, is added?

The same conditions apply, but now for both θ *and* ϕ. The Taylor expansion of the posterior PDF around θ_0 and ϕ_0 is then

$$L = L(\theta,\phi) + \frac{dL}{d\theta}\bigg|_{\theta_0,\phi_0}(\theta-\theta_0)\;\frac{dL}{d\phi}\bigg|_{\theta_0,\phi_0}(\phi-\phi_0) \tag{2.18}$$

$$+\frac{1}{2}\,\frac{d^2L}{d\theta^2}\bigg|_{\theta_0,\phi_0}(\theta-\theta_0)^2+\frac{1}{2}\,\frac{d^2L}{d\theta\phi}\bigg|_{\theta_0,\phi_0}(\theta-\theta_0)(\phi-\phi_0) \tag{2.19}$$

$$+\frac{1}{2}\,\frac{d^2L}{d\phi\theta}\bigg|_{\theta_0,\phi_0}(\phi-\phi_0)(\theta-\theta_0)+\frac{1}{2}\,\frac{d^2L}{d\phi^2}\bigg|_{\theta_0,\phi_0}(\phi-\phi_0)^2+... \tag{2.20}$$

The first derivative terms is zero per the conditions, and the cross-terms for the second derivatives can be added together since $\frac{\partial^2}{\partial x\partial y}=\frac{\partial^2}{\partial y\partial x}$. Returning to the PDF, the exponential is applied and the approximation is

$$f(\theta,\phi) \approx A\exp\Big(\frac{1}{2}\,\frac{d^2L}{d\theta^2}\bigg|_{\theta_0,\phi_0}(\theta-\theta_0)^2\Big)\cdot\exp\Big(\frac{1}{2}\,\frac{d^2L}{d\phi^2}\bigg|_{\theta_0,\phi_0}(\phi-\phi_0)^2\Big) \tag{2.21}$$

$$\cdot\exp\Big(\frac{d^2L}{d\theta d\phi}\bigg|_{\theta_0,\phi_0}(\theta-\theta_0)(\phi-\phi_0)\Big) \tag{2.22}$$

The two first exponential terms are recognized as univariate Gaussian distributions for θ and ϕ, and the third is the multivariate Gaussian distribution for both variables. This long expression can be shortened by writing the derivatives in a matrix format.

$$K = \begin{pmatrix} \frac{d^2L}{d\theta^2} & \frac{d^2L}{d\theta\phi} \\ \frac{d^2L}{d\theta\phi} & \frac{d^2L}{d\phi^2} \end{pmatrix} \tag{2.23}$$

which gives the approximation

$$f(\theta, \phi) \approx A \exp\left((\theta - \theta_0, \phi - \phi_0)K \begin{pmatrix} \theta - \theta_0 \\ \phi - \phi_0 \end{pmatrix}\right) \tag{2.24}$$

To summarize, the posterior PDF for a variable can be approximated as a Gaussian distribution $\mathcal{N}(m[\theta], \mathbb{V}[\theta])$. If the distribution are multivariate, the covariance can be calculated. This gives the covariance matrix $\text{cov}(\theta_i, \theta_j)$ which holds all the variances and covariances.

2.2.2 Regression

Now that the relation between probability and the Gaussian distribution is established, it is time to see how this is utilized in the Gaussian Processes Regression (GPR). This section will first introduce (GP) in general from a machine learning standpoint and the notation used, then exploring prior and posterior distributions, and how functions can be drawn from this. Lastly, the more specific usage for the thesis is described.

As discussed in the first part of this chapter, a posterior probability distribution for some value can be estimated as a Gaussian distribution where, based on observations, the best estimate is the mean and the variance of the distribution is σ^2. This is the basis for Gaussian processes. GP uses the Gaussian distribution to predict function values.

Consider a data set consisting of an input vector $\mathbf{x}$ and target value y, and that $y = f(x)$ describing the relation between the input and output. GP is a machine learning method, so this data is used in some way for training and multiple input vectors and target values are needed. A $n \times m$ design matrix, X, is usually made with n inputs and m covariates. Instead of covariates, there can be m data sets. When the GP performs a prediction at a new input value x_*, it is over the function value at the point.

$$y_* = \mathcal{N}(f(x_*), \mathbb{V}(x_*)). \tag{2.25}$$

As the PDF's for a data point is a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, the set of data points will have a multivariate Gaussian distribution. A prediction drawn from the PDF will be a vector containing the mean for each variable and the variance is combined into the covariance matrix. Then functions are drawn from the distribution

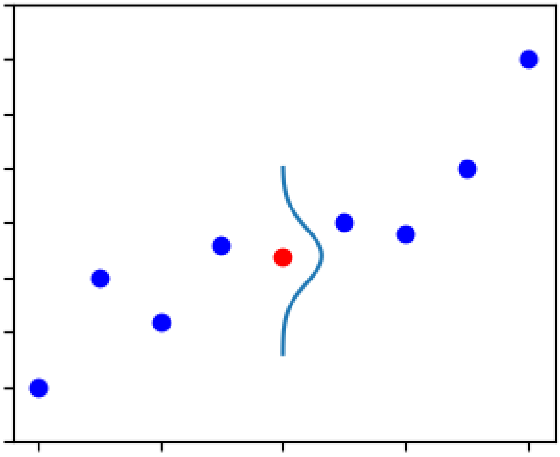


Figure 2.3: This figure shows how Gaussian Processes assumes a Gaussian distribution over a test value, here in red. The blue points are observed data that the GP would be conditioned on.

$$f \sim \mathcal{N}(m(\mathbf{f}), cov(\mathbf{f})) \tag{2.26}$$

Here the covariance matrix is calculated from the covariance function, also called *kernel*

$$cov(y_i, y_j) = k(x_i, x_j) \tag{2.27}$$

meaning that the kernel value evaluating two points, will provide the covariance between the points, i.e. how they are located relative to each other.

2.2.3 Covariance functions

Before the data is added to this machinery, the prior information is decided. This indicates the assumptions made on the data. Covariance functions, kernels, are the most important part of Gaussian Processes. As the data is unseen by the method, the covariance is unknown. But if there is some knowledge, e.g. the data varies periodically, the distance in function value space increases or decreases, or even if the data rapidly varies with a small amplitude for short distances as well as slow variations with large amplitude, this can be added to the kernel. The kernels take two input values and calculated the covariance between them, resulting in a matrix with all covariances. This matrix is symmetric, $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$. Kernels that depend only on the difference of the points, $\mathbf{x} - \mathbf{x}'$, are called *stationary.* The kernels are also called *isotropic* if it only depends on the euclidean distance between the points, $r = |\mathbf{x} - \mathbf{x}'|$.

Exponential Squared Kernel

The *squared exponential* kernel is written

$$k_{SE}(x, x') = \exp\Big(- \frac{r^2}{2l^2}\Big) \tag{2.28}$$

where l is the *characteristic length scale.* This length scale determines how smooth the function is. This can be thought of as how large the distance between points must be before the predicted function can vary greatly, i.e. the function values becomes close to uncorrelated. The SE kernel is infinitely differentiable making it very smooth. This thesis will focus mostly on this kernel.

Matérn Class of Covariance Kernels

The Matérn Class is given by

$$k_{Matern}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}\big(\frac{\sqrt{2\nu}r}{l}\big)^{\nu} K_{\nu}\big(\frac{\sqrt{2\nu}r}{l}\big) \tag{2.29}$$

where ν and l are positive values, $\Gamma(\nu)$ is the gamma function, and the K_ν is a modified Bessel function [20]. ν determines how smooth the function is, and when this goes to infinity the SE kernel is obtained. When ν is a half-integer $p + 1/2$, where p is a positive integer, the kernel becomes a product of a polynomial and an exponential function

$$k_{\nu=p+1/2}(r) = \exp\big(- \frac{\sqrt{2\nu}r}{l}\big)\frac{\Gamma(p+1)}{\Gamma(2p+1)}\sum_{i=0}^{p}\frac{(p+1)!}{i!(p-1)!}\big(\frac{\sqrt{8\nu}r}{l}\big)^{p-1} \tag{2.30}$$

The most interesting cases, and most common for machine learning, are $\nu = 3/2$ and $\nu = 5/2$,

$$k_{\nu=3/2}(r) = \big(1 + \frac{\sqrt{3}r}{l}\big) \exp\big(-\frac{\sqrt{3}r}{l}\big) \tag{2.31}$$

$$k_{\nu=5/2}(r) = \big(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\big) \exp\big(-\frac{\sqrt{5}r}{l}\big) \tag{2.32}$$

For $\nu = 1/2$ the kernel becomes a rough exponential, and for $\nu \geq 7/2$ the difference between the kernels is harder to distinguish.

Local Gaussian Kernel

The Local Gaussian kernel is a *non*-stationary kernel. It is

$$k_{LocalGauss}(x, x') = \exp\big(-\frac{(x - x_0)^2 + (x' - x_0)^2}{2\omega}\big) \tag{2.33}$$

where x_0 is the mean of the Gaussian called the *location* and ω is the squared width of the distribution. This kernel is of interest for the signal modeling of the simulated Higgs peak.

Constant kernel

The constant kernel is used when a constant number is added or multiplied to a kernel. For example, an amplitude term multiplied to the squared exponential kernel. This then makes the constant number a hyperparameter for the overall kernel.

$$k_{Constant}(x, x') = c \tag{2.34}$$

Hyperparameters

The parameters of the kernel functions are called *Hyperparameters*, to emphasize the fact that they are parameters of a non-parametric model. They can be varied to match the expected value, or be optimized for a data set. In the SE and Matérn kernels, the hyperparameters are the length scale l. In addition it is common to have a variance σ_f^2 for the intensity of the covariance, called the amplitude. Lastly, the noise variance σ_n^2 is also added to the diagonal of the covariance matrix, since noise-free processes are rare in physics. In figure 2.4, the data points are sampled from a GP prior using the SE kernel with hyperparameters $(l, \sigma_f^2, \sigma_n^2) = (1, 1, 0.1)$, then the GP is conditioned on the data points and a new posterior prediction is made, seen as the pink line in the figure. The gray band is the 2σ error. The uncertainty of the updated GP predicted mean increases in the regions with no data, but updated GP does reproduce the model that was used to generate the fake data set very well.

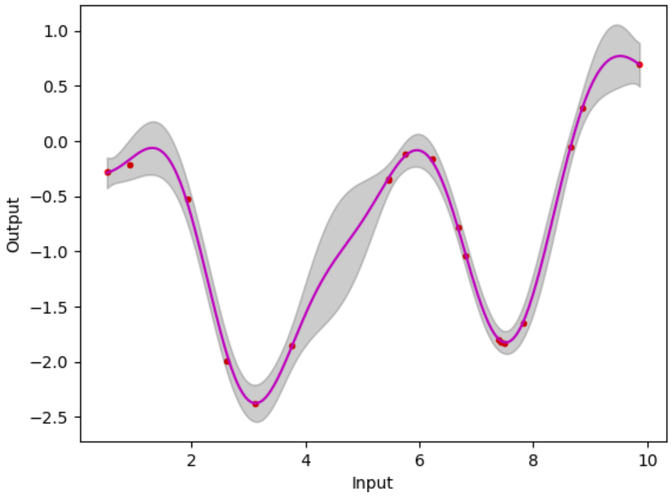


Figure 2.4: This is an example of the optimized length scale for a set of data points. The variance band increases somewhat where there is no information.

To explore how the hyperparameters affect the predictions, a short and a long length scale is used, in figure 2.5 and figure 2.6 respectively. In figure 2.5, the hyperparameters used is $(0.1, 1.08, 1e-6)$. This allows for more rapid variation of the mean function. The noise has been reduced, therefore the mean is more restricted to be closer to the data points, than in figure 2.4. This is justified from increased flexibility of the prediction. Around $x = 4.5$, the mean function moves up then down seemingly without reason. The reason for this behavior, is the lack of data between $x \approx 4$ and $x \approx 5.5$. When there is no new information for the GP, the prediction will fall back to the only knowledge at the point, the prior. The uncertainties in the data-less intervals are also much larger than in figure 2.4, which has a more even band.

In figure 2.6, the hyperparameters used are $(3, 1.16, 0.9)$. The mean follows the general trend of the data, but the prediction is a noisy and slow varying function which does not describe the data well. It is easy to see why choosing (optimizing) the hyperparameters is important and is a large part of the analysis.

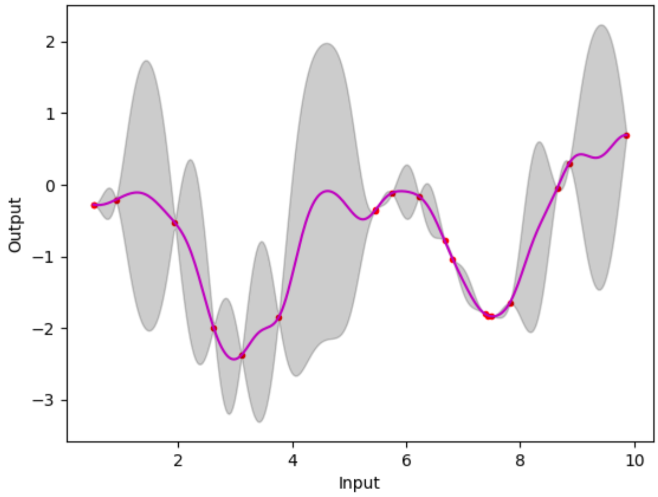


Figure 2.5: This figure shows an example of too a short length scale, where the predictive mean goes back to the prior where there is little information.

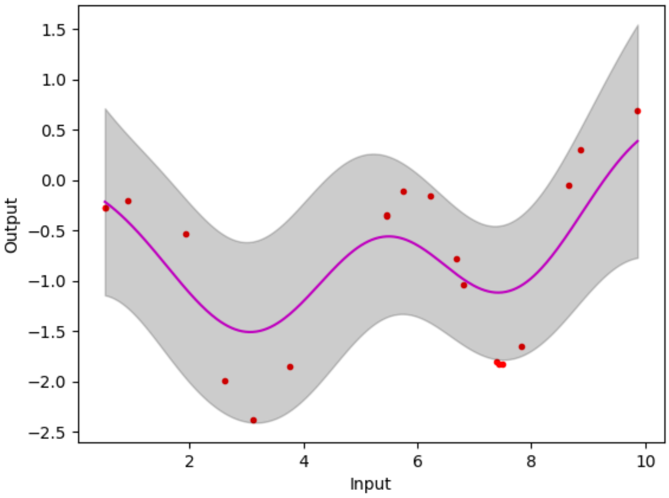


Figure 2.6: This figure shows an example of a predicted mean where the length scale is too large to describe the data points well.

Function space view

Gaussian Processes is all about doing inference in the function space, instead of the parameter space of a particular function, as the GP describes a distribution over functions. It is completely specified by a mean function and the covariance matrix given as[17]

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{2.35}$$

$$k(\mathbf{x}, \mathbf{x'}) = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x'}) - m(\mathbf{x'}))] \tag{2.36}$$

The mean $m(\mathbf{x})$ is the expectation value for the function at location $\mathbf{x}$, and the covariance is the expected product of the difference of the function value and the mean for two locations $\mathbf{x}$ and $\mathbf{x}'$. Combined these provide a Gaussian distribution for the function value with the mean and covariance given above for each input vector. This collection of Gaussian distributed function values is the Gaussian Processes, $\mathcal{GP}$, from which a random function can be drawn, in the same manner as a random value is drawn from a distribution. This is written as

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{2.37}$$

As GP is a Bayesian method, a prior is needed that will be updated on the observations. The most natural prior is a random Gaussian vector drawn from a normal distribution given as

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, K(X, X)). \tag{2.38}$$

The mean is assumed to be zero, meaning that drawing from a very large number of functions and summing them will produce an average of zero. And the covariance of the points is decided by the covariance function k that generates the covariance matrix K. Random functions can be drawn from this prior as shown in figure 2.7. These functions are drawn from the same prior information, but they are all different. The matrix X_* is constructed from the input values, of size $n^* \times m$ where n^* is the number of data points each with m features.

In figure 2.8, data points are added to the GP and a posterior mean is predicted which is the dashed black line in the plot. The other three are randomly drawn samples from this posterior. The error in the points, in order, is $(0.01, 0.01, 0.1, 0.5, 1e-6)$ and are chosen for illustration. The variance is larger where information is missing or is poor. It is observed that for the first two points, $x = 2$ and $x = 3$, the functions lies very close to the points as the noise-level is small. The same goes for the last point. The third point

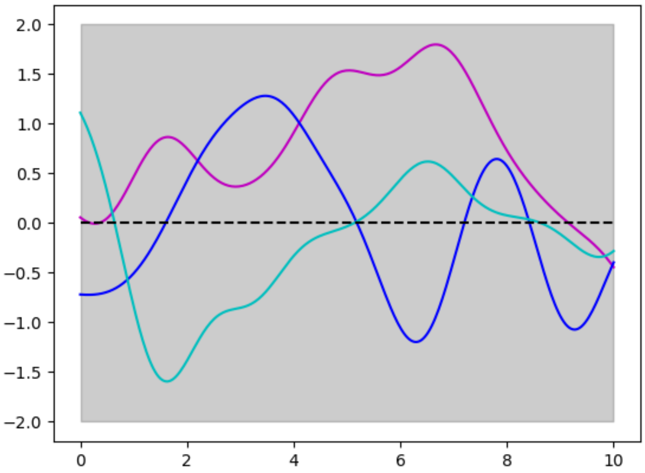


Figure 2.7: This figure shows an example of drawing samples from the prior. The same information returns three different functions. The dashed line is the prior mean, the grey band is the 2σ standard deviation from the prior variance, and the solid lines are the drawn samples.

at $x = 5$ has some noise and the functions are still close, but can now vary more. And for the fourth point at $x = 7$, the noise-level is high. Therefore the functions can vary greatly, and they tend to be far away from the point. By chance the blue function is far away from the others. This is allowed by the posterior covariance (the posterior variances are shown in the figure).

To illustrate the usage of GPR on a specific model with noise, consider the model $y = f(\mathbf{x}) + \epsilon$ where ϵ is the additive Gaussian noise with variance σ_n^2. The noisy observations $\mathbf{y}$ have the prior

$$\operatorname{cov}(y_i, y_j) = k(\mathbf{x}_i, \mathbf{x}_j) + \sigma_n^2 \delta_{ij} \tag{2.39}$$

where k is the kernel function and δ_{ij} is the Kronecker delta making sure the noise is only added to the variances, not covariances, as $\delta_{ij} = 1$ if $i = j$ and zero otherwise. This can also be written as

$$\operatorname{cov}(\mathbf{y}) = K(X, X) + \sigma_n^2 I \tag{2.40}$$

K is the covariance matrix for the matrix X of the inputs. The noise is added only to the diagonal. Introducing the test values, X_*, gives the joint distribution [17]

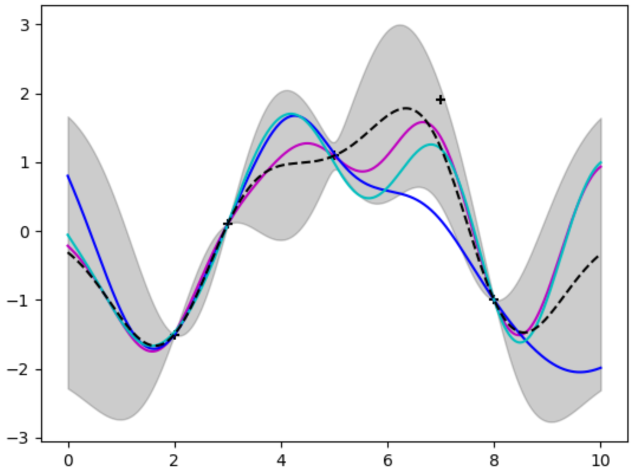


Figure 2.8: This figure shows drawn samples from the posterior function, where five data points constrain the predictive mean. The dashed line is the posterior mean, the gray band is the 2σ standard deviation from the posterior variance, and the solid lines are the drawn samples. The imporant features are commented in detail in the text

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right) \tag{2.41}$$

Conditioning this joint distribtion on the observations, $\mathbf{y}$, gives

$$f_*|X,\mathbf{y},X_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*)) \tag{2.42}$$

where $\bar{\mathbf{f}}_*$ is the predictive mean defined as

$$\bar{\mathbf{f}}_* \equiv \mathbb{E}[\mathbf{f}_*|X,\mathbf{y},X_*] = K(X_*,X)(K(X,X) + \sigma_n^2 I)^{-1}\mathbf{y} \tag{2.43}$$

and the covariance is

$$\text{cov}(\mathbf{f}_*) = K(X_*,X_*) - K(X_*,X)[K(X,X) + \sigma_n^2 I]^{-1} K(X,X_*) \tag{2.44}$$

The equations 2.43 and 2.44 are the main equations for GPR. These equations can be written simpler, by setting $K(X,X)$ to K and $K(X,X_*)$

to K_*, $K_* \to \mathbf{k}_*$ giving the covariance for one test point $\mathbf{x}_*$ and the training points. This new notation gives

$$\bar{f}_* = \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{y} \tag{2.45}$$

$$\mathbb{V}(f_*) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{k}_* \tag{2.46}$$

This is, again, the main equations for prediction in GPR. These are implemented for this thesis in the programming language PYTHON through the package GEORGE. The details of the package will be expanded upon later. The algorithm used by the package is, from [17], given in algorithm 1.

```
input: X (input), $\mathbf{y}$ (targets), k (covariance function), $\sigma_n^2$ (noise), $\mathbf{x}_*$
  (test input)
L := cholesky(K + $\sigma_n^2$)
$\boldsymbol{\alpha}$ := $L^\top$/(L/$\mathbf{y}$)
$f_*$ := $\mathbf{k}_*^\top \boldsymbol{\alpha}$
$\mathbf{v}$ := L/$\mathbf{k}_*$
$\mathbb{V}[f_*]$ := $k(\mathbf{x}_*, \mathbf{x}_*)$ - $\mathbf{v}^\top \mathbf{v}$
log p($\mathbf{y}$—X) := -$\frac{1}{2}\mathbf{y}^\top \boldsymbol{\alpha} - \Sigma_i \log L_{ii} - \frac{n}{2} \log 2\pi$
return: $f_*$ (mean), $\mathbb{V}[f_*]$ (variance), $\log p(\mathbf{y}|X)$ (log marginal
  likelihood)
```

Algorithm 1: This is the algorithm for conditioning the prior function, predict a posterior mean based on the data, and return the log marginal likelihood for the mean. This algorithm is from [17].

The algorithm 1 use Cholesky decomposition on the covariance matrix to obtain the decomposition matrix L. This matrix is then used, together with the targets $\mathbf{y}$, to find the vector $\boldsymbol{\alpha}$. This $\boldsymbol{\alpha}$ is multiplied with the test-covariance to calculate the conditional predictive mean. L is also used to find the vector $\mathbf{v}$ which, combined with the test-covariance, gives the variance of the predictive mean $\mathbb{V}[f_*]$. Even more, this algorithm returns the log marginal likelihood, $\log p(\mathbf{y}|X)$, utilized in the optimization of the hyperparameters that is described in detail in the next section.

2.2.4 Model optimization

A crucial part of the regression is the selection of kernel and the optimization of the hyperparameters. The hyperparameters are found by maximizing the *marginal likelihood.* This marginal likelihood is the integral of the likelihood multiplied with the prior

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|X)d\mathbf{f} \tag{2.47}$$

This is similar to equation 2.10, but instead of marginalizing over the parameter θ, the likelihood is marginalized over the function values $\mathbf{f}$. The prior is Gaussian, $\mathbf{f}|X \sim \mathcal{N}(\mathbf{0}, K)$, and the likelihood is also Gaussian, $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma_n^2 I)$, only factorized. Then the following can be observed

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, K + \sigma_n^2 I) = \frac{1}{\sqrt{(2\pi)\sigma^2}} \exp\Big(- \frac{\mathbf{x}^2}{2\sigma^2}\Big) \tag{2.48}$$

This leads to

$$p(\mathbf{y}|X) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)\sigma_i^2}} \exp\Big(- \frac{y_i^2}{2\sigma_i^2}\Big) \tag{2.49}$$

Taking the logarithm of equation 2.49, the log marginal likelihood is obtained [17]

$$\log p(\mathbf{y}|X, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^\top K_y^{-1}\mathbf{y} - \frac{1}{2}\log|K_y| - \frac{n}{2}\log 2\pi \tag{2.50}$$

$\boldsymbol{\theta}$ is added to imply that the marginal likelihood is conditioned on the hyperparameters of the kernel, and n is the number of data points. $K_y = K_f + \sigma_n^2 I$ is the covariance matrix for the targets $\mathbf{y}$ with noise, and K_f is the covariance of the noise-free $\mathbf{f}$. The first term in equation 2.50, $-\frac{1}{2}\mathbf{y}^\top K_y^{-1}\mathbf{y}$, is the data-fit. This is the only term containing the target points. The second term, $-\frac{1}{2}\log|K_y|$, is the complexity penalty which depends only on the kernel and the noise of the input. The last term, $-\frac{n}{2}\log 2\pi$, is a normalization. The negative complexity penalty increases for larger length scales, as the model is more complex for short length scales than for large. And if the noise is big this also increases the complexity term, to prevent too severe overfitting.

2.3 Implementation

2.3.1 Simulated data and underlying function

The template used in this thesis, is a set of simulated data made to be very similar to real data for the studied process[21]. This simulated data corresponded to an integrated luminosity luminosity of 3.6 fb^{-1} with a center of mass energy of 7 TeV, so it was scaled by a factor of 10 to give a data set with an approximate integrated luminosity of 36 fb^{-1}. The data set was not used directly, as there might be Poisson fluctuations added. Therefore, a fit was performed on the simulated data to retrieve an underlying function based on the "real" process. The chosen function was a Bernstein polynomial of 5th order. The reasoning for using such a complicated function, was to present a tough challenge to the GPR to test if the new method could find this underlying function, and if the background modeling is independent on the luminosity as claimed by Frate et. al. [2].

2.3.2 Modeling procedure

For this thesis, PYTHON is the chosen programming language, and GPR is implemented from many different sources. SCIKIT-LEARN is a well-established library for machine learning and does have a GPR part. This was used in the beginning of the thesis to explore the basics of GPR, then the package GEORGE [22] was used since it has more options, and Frate et.al. [2] also utilized this package. The Squared Exponential, Matérn and the Local Gaussian kernels (described above in section 2.2.3) are all readily implemented in GEORGE. The steps for performing the fit are seen in algorithm 2. First, the hyperparameters of the kernel must be optimized. The code snippet below shows the built-in call of GEORGE that calculates the log marginal likelihood discussed in the regression section and returns the value for some set of observations. The call initializes the kernel with the set of hyperparameters currently being tested, initiates the GP object, and computes the covariance matrix where the uncertainty is added to the diagonal of the matrix. Lastly, the log likelihood is calculated for the observations, which is returned.

Listing 2.1: This code snippet shows how the log likelihood is calculated for each set of hyperparameters

```
class log_like_gp:
    def __call__(self,Amp,length):
        kernel = Amp*ExpSquaredKernel(metric=length)
        gp = GP(kernel,solver=HODLRSolver)
        gp.compute(self.x, yerr=np.sqrt(self.y))
        return −gp.log_likelihood(self.y)
```

To scan over the hyperparameters for the optimization, another package is used. It is called iminuit[23], and is the python version of Minuit2[24]. This package utilizes the migrad algorithm, using Quasi-Newton method and DFP formula [23], to search for the maximum log marginal likelihood by scanning through parameter values. When iminuit has found the maximum, the value is returned with the hyperparameters. This is done 100 times for each fit, where the new maximum log likelihood value is compared to the last, and then the largest one is kept with the corresponding hyperparameters. The example code in listing 2.1 returns the log likelihood is for background-only fits. If the background kernel used for the modeling is changed or a signal kernel is incorporated, these changed must be done in this call-function.

```
lnprob = log_like_gp(x,y)
MLML, BFP = fit_minuit_gp(...,lnprob)
initialize kernel
initialize GP object
compute covariance matrix
call predict; constrain on observation and predict on x-values
```

Algorithm 2: This is the procedure for the fitting.

When the optimization is done, the kernel is re-initialized with the optimized hyperparameters and the covariance matrix is recalculated, as seen in algorithm 2. To include as much information as possible, a mean function is given as a part of the prior. The function is actually just the median value of the distribution being modelled, so the prior is shifted from $\mathbf{0}$ to this new value. Then GPR updates the prior function based on the observations, and provides a prediction. Usually, the prediction test point, $\mathbf{x}_*$, is some new unseen point. In this thesis, the final prediction will be on the same input values as the observations. Then GPR will return the conditional prior as the predicted posterior which is used as a fitted function.

2.3.3 Background modeling

To model the background, the right kernel must be chosen. For this thesis, the Exponential Squared and the Matérn kernels are studied. The hyperparameters (in fact the squares of the hyperparameters in iminuit) are for both kernels the length scale and amplitude. Therefore, to interpret the hyperparameters, the square root must be taken.

For the background modeling, there are tests that must be performed. Can the GPR find the underlying function used to generate the toy data sets? And if not, will the generated bias lead to a spurious signal that must be accounted for? How these issues are tested, is explained here.

Is GPR is able to find the underlying function, when a background-only fit is performed? To test this, 20000 toy data sets were generated, and fitted using GPR. The toys are created by taking the template, looping through the bins and setting each bin content as a value n from a Poisson generator using the template value as expected event count. The uncertainty in each bin is $\sqrt{n}$. To test the goodness-of-fit, Pearson's χ^2 is used as the test statistic[13]. This is given as

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - f_i)^2}{\sigma_{f_i}^2} \tag{2.51}$$

where f_i is the fitted function value in bin i, n is the number of bins, and y_i is the observed value. The uncertainty in the denominator is the uncertainty of the fitted value, giving $\sigma_{f_i}^2 = \sqrt{f_i}$. As the toy data are randomly generated, it is difficult to assume that one of these can represent the actual data distribution, it would be a very lucky strike. This is why so many toys were generated, to get a distribution over the test statistic. If the assumption that $\sigma_{f_i} \sim \mathcal{N}$ is true, the distribution of the test statistic will follow a χ^2 distribution.

A χ^2 distribution has some nice properties; the mean k of the distribution is the number of degrees of freedom for the fit and the variance is $2k$. k is given as the number of data points in the distribution to be fitted, minus the number of parameters in the function. Therefore, dividing the χ^2 value for each fit by the number of degrees of freedom, a value around 1 should be retrieved for the average. How close to one the value is, indicates how good the fit is. This procedure is done for the 20000 toys and a χ^2/ndf distribution is obtained, and the mean value of this distribution converges towards a value that characterizes how well the model explains the underlying function.

Following from the background modeling, the claim of luminosity independence from Cranmer et. al. [2] is the most logical object of interest to test. The result from the article are seen in figure 2.9. The green points are the average χ^2/ndf for the Gaussian Processes Regression and the blue for the parametric function. It is observed that the green is apparently consistently above one. This will be discussed later. Instead of many different integrated luminosities, only four were tested in this thesis: 36 fb^{-1}, 360 fb^{-1}, 1800 fb^{-1}, 3600 fb^{-1}, with 20000 toys for each.

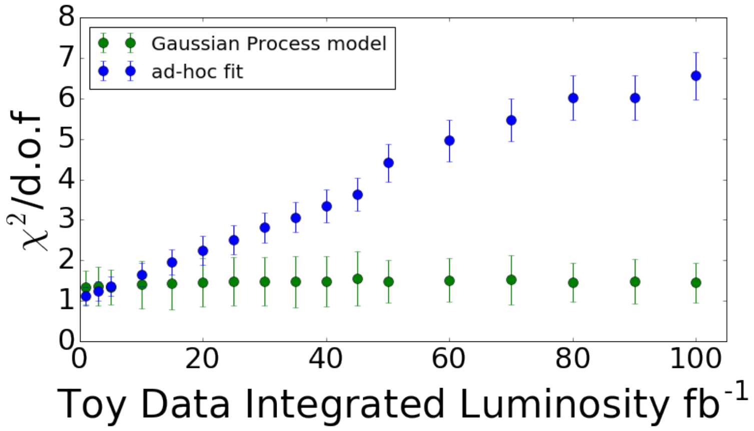


Figure 2.9: This is the evolution of the χ^2/ndf for increasing luminosity, as presented in [2].

If GPR is to be the preferred method for background modeling, then a test for bias must be performed. The values of the parameters for a signal estimation are affected by a potential bias. Usually an ensemble of simulated data is used for this test, where a fit is performed with a parametric function to each toy, and the sum of all the residuals is calculated. To save time, a representative data set is often used, which is called an Asimov data set [14]. This is the underlying function, which the average of all toys will converge to, with the Poisson error $\sqrt{n_i}$ for each bin i. One fit is performed and residuals are calculated, which gives an indication of bias for the best modeling of the underlying function.

A Gaussian signal peak was used to simulate a signal at $m_H = 125$ GeV. The width of the signal was $\sigma = 2$ GeV and the number of events was 2000. This number of events was chosen to be similar as the total number of events in table 5 in [1]. The flexibility of GPR was one concern with the fit to background-only data. If the model is too flexible, it might ”swallow” the signal, which is not desired. Therefore, a background-only fit was made to the signal+background data where the Gaussian peak is injected. The residuals of this fit can then be compared to the background-only fit on the background-only data.

2.3.4 Signal modeling

The signal modeling was performed using a signal kernel; the Local Gaussian Kernel. This kernel contain the hyperparameters location x_{loc}, the mean of the Gaussian, and the *log_width*, which is the width of the Gaussian. Both of these values are fixed during the modeling procedure. The hyperparameter *log_width* is actually the logarithmic value of the *squared* width. This is important to know when the best fit parameters are interpreted. In addition, an amplitude-term A is multiplied in the kernel, which is given from IMINUIT as squared. This gives

$$k_{LocalGaussian} = A \cdot \exp\Big(- ((x_1 - x_{loc})^2 + (x_2 - x_{loc})^2)/(2 \cdot \log\text{\_width})\Big) \tag{2.52}$$

This kernel is added to the background kernel, and fitted to the toy data including a signal peak where the location is fixed at 125 GeV and the log_width is fixed at log(4) giving a width of 2. The test here is to see if a signal amplitude can be extracted from the fit, with uncertainty. As the value of the hyperparameter is the squared of the signal amplitude, $Amp_{Hyperparameter} = A_{sig}^2$, the squareroot must be taken to get the amplitude. The uncertainty of the signal amplitude is calculated as

$$\sigma_{A_{sig}} = \frac{\sigma_{Amp_{HP}}}{2A_{sig}} \tag{2.53}$$

Alternative signal estimation method

A different method of estimating the signal over a background can be used by ... Gaussian Processes are linearly independent. A posterior mean prediction is calculated from the covariance matrix K from the kernel, noise σ^2 of the data points, and data points $\mathbf{y}$

$$\mathbf{f} = K(K + \sigma^2 I)^{-1}\mathbf{y} \tag{2.54}$$

In this thesis the predictions performed by the Gaussian Process, are on the same input values corresponding to the target values used to constrain the prior. The kernel used to calculate the matrix K for signal+background estimation, is the sum of the local gaussian signal kernel and squared exponential background kernel, $K = K_s + K_b$. Let $S = (K + \sigma^2 I)^{-1}$. Inserting into equation 2.54:

$$\mathbf{f} = (K_s + K_b)S\mathbf{y} = K_s S\mathbf{y} + K_b S\mathbf{y} \tag{2.55}$$

The two terms are recognized as two mean calculations. Therefore, the predicted mean can be written as

$$\mathbf{f} = \mathbf{f}_s + \mathbf{f}_b \tag{2.56}$$

From equation 2.56, it is clear that the calculation of the predicted signal mean and predicted background mean, can be calculated separately. GEORGE has the possibility of choosing which kernel to use for the prediction. The package will calculate the matrix S using the full kernel with added noise, then constrain the prior on the target values $\mathbf{y}$. The predicted mean is then retrieved from the package.

The prediction from GP is an estimate on the number of events in each input value, as the numbers in the bins are simulating counted events. The distribution is discrete, therefore, the sum can be taken of the predicted mean for the signal. This gives the number of estimated signal events in the distribution being modelled,

$$N_{sig} = \sum_{i=1}^{n} \mu_{i,sig}. \tag{2.57}$$

But what about the uncertainty on this number of signal events?

Assume that $X = A + B$. The variance of this sum is

$$\begin{aligned} Var(X) &= Cov(X,X) = Cov(A+B, A+B) && (2.58)\\ &= Cov(A,A) + Cov(A,B) + Cov(B,A) + Cov(B,B) && (2.59)\\ &= Var(A) + Var(B) + 2Cov(A,B) && (2.60) \end{aligned}$$

The variance of N_{sig} is then

$$Var(N_{sig}) = Cov(N_{sig}, N_{sig}) = Cov\Big(\sum_{i=1}^{n} \mu_{i,sig}, \sum_{i=1}^{n} \mu_{i,sig}\Big) \tag{2.61}$$

In equation 2.60, the variances and covariances of A and B are added together, therefore

$$Cov\Big(\sum_{i=1}^{n} \mu_{sig}, \sum_{i=1}^{n} \mu_{sig}\Big) = \sum_{i=1}^{n} Var(\mu_{i,sig}) + 2 \sum_{i,j:i<j} Cov(\mu_{i,sig}, \mu_{j,sig}) \tag{2.62}$$

It is easy to observe that the first term is the sum of the diagonal elements in the posterior covariance matrix, and the second term is two times the sum of all off-diagonal elements in one triangle of the posterior covariance matrix. The terms are added together, therefore

$$Cov\Big(\sum_{i=1}^{n}\mu_{sig},\sum_{i=1}^{n}\mu_{sig}\Big)=\sum_{i,j}^{n}Cov(\mu_{i,sig},\mu_{j,sig})=Var(N_{sig}) \qquad (2.63)$$

The variance of the number of signal events is the sum of the posterior covariance matrix for the posterior mean. Both the predicted mean and posterior covariance matrix are outputs from GEORGEs predict statement.

2.3.5 Number of degrees of freedom

The number of degrees of freedom is defined as the freedom to move. A set of data with n entries has n degrees of freedom. This is because all the data points in the set can vary on their own assuming they are independent. The data sets in this thesis will contain a number of events for each bin, and all of these values are independent on each other. Consider the example: a data set contain 10 entries. The values of these can be arbitrarily chosen, so they are free to have any value, giving this data set 10 degrees of freedom. Now, some knowledge about the data is introduced; the mean of the data set is 5. The mean for a set of values is found via the formula $mean$ = sum of values/number of values, therefore the sum of the data points must be $mean\times n$. So for the set of 10 values, the sum must be 50. This is a constraint on the degrees of freedom, which can be shown by choosing some values. The first value can be arbitrary, for example 4, and this can be done for the next seven. The set is now $[4,5,7,3,9,1,8,X,Y]$, where X and Y are not chosen yet. The sum must be $4+5+7+3+9+1+8+X+Y=50$, and the value for either X or Y can be chosen arbitrarily. But choosing one, means that the other looses its ability to vary: $X=4$ means that $Y=9$, and $X=6$ means that $Y=7$. When this information was introduced, the number of degrees of freedom was reduced by one, $10-1=9$. The data set has 9 degrees of freedom. Introducing more such information, the number of degrees of freedom will be reduced even further.

For regression problems, the concept of number of degrees of freedom becomes more cloudy and difficult. A model with p parameters can vary in all the parameters, giving the model p degrees of freedom, or *flexibility to vary.* Consider a distribution with n data points. From the last paragraph, this distribution has n number of degrees of freedom. If this distribution is fitted with a model with p parameters, the parameters are optimized using the information in the data set, which impose restrictions on the number of degrees of freedom. When the parameters are optimized, some of the data points in the distribution are not free to vary anymore, reducing the number of degrees of freedom for the fitted model by the number of parameter p. Then the number of degrees of freedom for the distribution is $n - p$.

GPR is a General Additive Model, or a *linear smoother* [17, 25], and does not contain any parametric function. The method of counting parameters does not work for GPR, therefore the number of degrees of freedom must be calculated some other way. Consider the predicted mean for the input values

$$\bar{f} = K(K + \sigma_n^2 I)^{-1}\mathbf{y}. \tag{2.64}$$

Say that K has the eigendecomposition $K = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$, where λ_i is the ith eigenvalue and $\mathbf{u}_i$ is the corresponding eigenvector. K is real and symmetric positive semidefinite, so the eigenvalues are real and ≥ 0. The eigenvectors are orthogonal. Let $\mathbf{y} = \sum_{i=1}^{n} \gamma_i \mathbf{u}_i$ where $\gamma_i = \mathbf{u}_i^\top y$. The predicted mean can then be written

$$\bar{f} = \sum_{i=1}^{n} \frac{\gamma_i \lambda_i}{\lambda_i + \sigma_n^2} \mathbf{u}_i. \tag{2.65}$$

If the fraction is much less than 1 for the ith $\mathbf{u}$, this component in $\mathbf{y}$ is practically removed. This smooths out any highly varying components in $\mathbf{y}$. So for this smoother, the *effective* number of degrees of freedom, is defined as [17, 25]

$$\sum_{i=1}^{n} \frac{\lambda_i}{\lambda_i + \sigma_n^2} = tr(K(K + \sigma_n^2 I)^{-1}) \tag{2.66}$$

where tr is the trace.

Chapter 3

Results and discussion

The results of the different tests will be presented and discussed in this chapter. It is divided into three parts; background modeling, bias test, and signal estimation.

Gaussian Processes Regression requires prior information that is updated based on data and then used for fitting. The prior used for testing was the squared exponential kernel and a mean corresponding to the median value of the data set being modelled.

To generate toy data sets, a 5-order Bernstein polynomial was used as an underlying function, with added Poisson noise. In figure 3.1 an example of such a toy data set is shown, where the red line is the underlying function and the black dots are one toy data set.

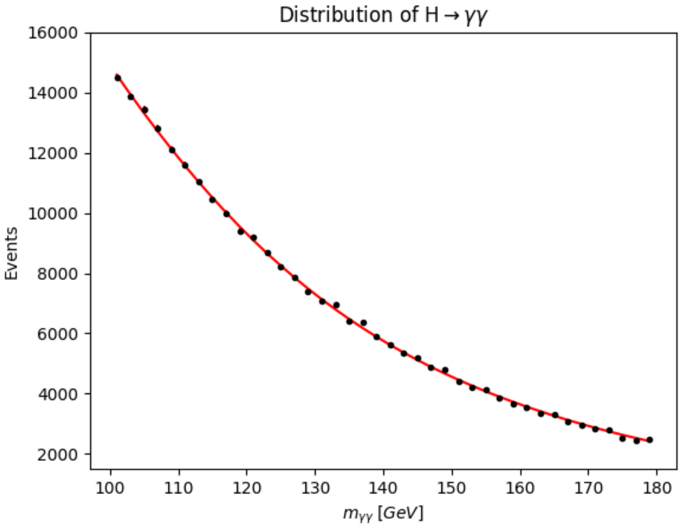


Figure 3.1: This figure shows an example of a toy distribution corresponding to an integrated luminosity of 36 fb^{-1} generated using Poisson fluctuations. The black points are the toy data set, and the red line is the underlying function.

3.1 Background modeling

The first test that was performed, was how well GPR could find the underlying function of the data sets. For this a χ^2 per number of degrees of freedom was calculated (as discussed in section 2.3.3), and the distribution of the test statistic for 20000 toy data sets is seen in figure 3.2. The shape of the distribution appears to follow a χ^2.

The distribution seems to be shifted to the left, and calculating the mean of the distribution, $\mu_{\chi^2/ndf} = 0.924$, confirms this suspicion. The immediate concern was that GPR is overfitting slightly for each toy distribution. Instead of diving into the realm of overfitting, an easier check was performed. Are the numbers of degrees of freedom used for the model correct? The number of degrees of freedom for the model is assumed to be 2. One for the amplitude and one for the length scale in the kernel. The number of degrees of freedom for the test statistic distribution is then $40-2 = 38$, where 40 is the number of bins. To confirm this assumption, the χ^2 distribution was calculated, as seen in figure 3.3. The mean of this distribution should be the number of degrees of freedom for the distribution. The mean is found to be $\mu_{\chi^2} = 35.1$, which means that the number of degrees of freedom for the model should be around 5. Two are identified, where are the missing three? The theoretical χ^2 for 35.1

numbers of degrees of freedom is also plotted in figure 3.3, confirming that the test statistic distribution follows a χ^2 and the assumption of Gaussian-like errors holds.

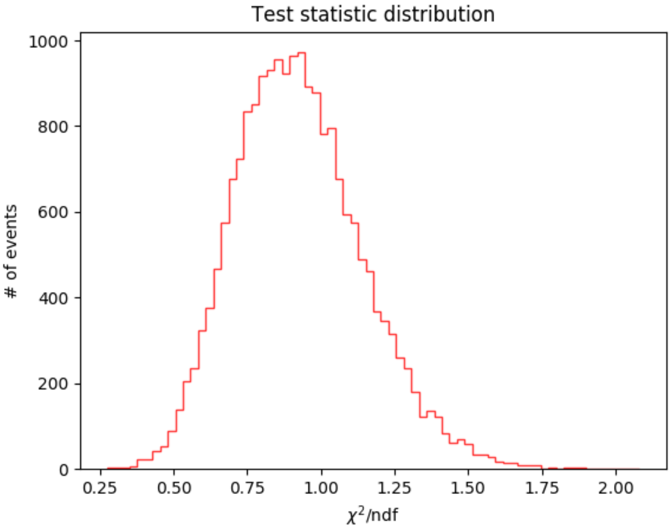


Figure 3.2: This plot show how the test statistic χ^2/ndf is distributed.

After some time, and discussions, the possibility of using the *effective* number of degrees of freedom for calculating the freedom of the model came to mind. The effective number of degrees of freedom were calculated, as described in 2.3.5 for each toy and the mean of these was found to be $\mu_{effndf} = 4.65$. This coincides with the mean μ_{χ^2} of the χ^2 distribution in figure 3.3. Using the effective number of degrees of freedom, the new distribution of χ^2 per number of degrees of freedom is shown in blue in figure 3.4, and the mean of this distribution is $\mu_{\chi^2/effndf} = 0.993$. It is easy to see the impact of using the effective vs the ”classic” (number of data points minus the number of explicit parameters) number of degrees of freedom.

The average of the test statistic χ^2 per effective number of degree of freedom for 20000 toys, are 0.993. This means that thee Gaussian Process manages to find the underlying function for an integrated luminosity of 36 fb^{-1}, but will this still be the case when the number of events increases? A χ^2 distribution for each luminosity was calculated, and is found in figure 3.5. There is no observed difference between the distributions. They overlap quite well, and are very similar to the distribution in 3.3. This indicates that the GP manages to find the underlying function for increasing event counts, meaning that GP is independent of the integrated luminosity. To

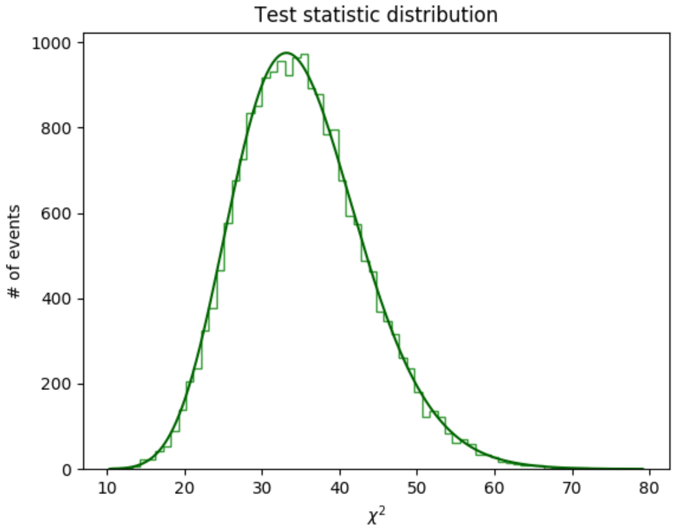


Figure 3.3: This figure shows the distribution of the χ^2 test statistic for each fit performed by Gaussian Processes. The mean of the distribution is 35.1. The smooth line is the theoretical χ^2 distribution for 35.1 numbers of degrees of freedom.

easier observe this, the means of the χ^2/ndf for the GP were calculated. To check if a parametric function could find the underlying function for increasing luminosity, a χ^2/ndf distribution was also calculated for the parametric function. The function used is the exponential of a second degree polynomial, $\exp(ax^2+bx+c)$, called Epoly2. The results for both the GP and parametric function are found in figure 3.6, where both the mean values of the GP test statistic and the test statistic for Epoly2, are plotted. The performance of the parametric function fit declines when the integrated luminosity increases. The means of the test statistic for the GP fits lies very close to 1, confirming that the Gaussian Process is independent of the luminosity. This modeling technique manages to find the underlying function when the number of events increases.

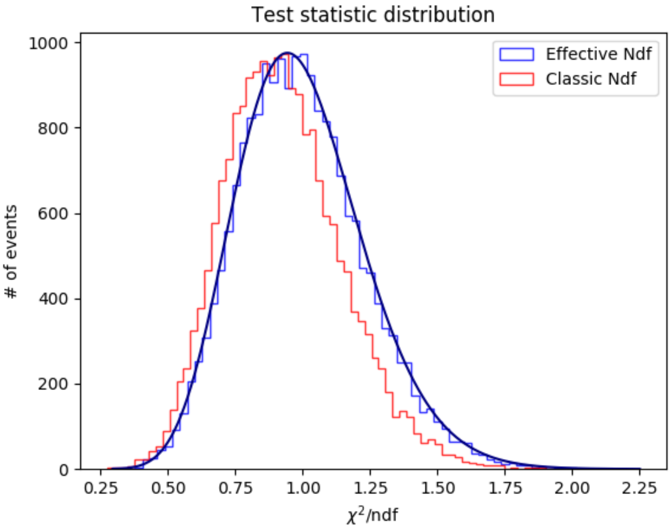


Figure 3.4: This figure shows the χ^2/ndf distribution for the fits performed with GP, using both the classical number of degrees of freedom of 2 and the effective number of degrees of freedom. The blue line is the effective ndf with a mean of 0.993, and the red is the classical method with a mean of 0.924. The solid blue line is the theoretical χ^2 distribution for 35.1 numbers of degrees of freedom.

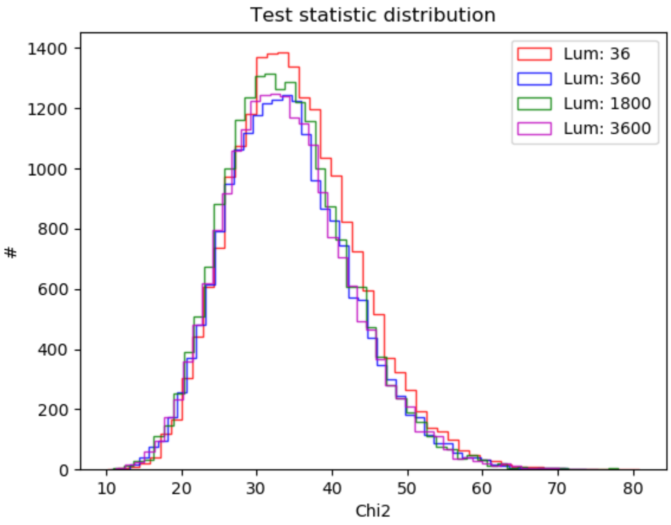


Figure 3.5: In this figure, the χ^2 distribution for each integrated luminosity value is shown.

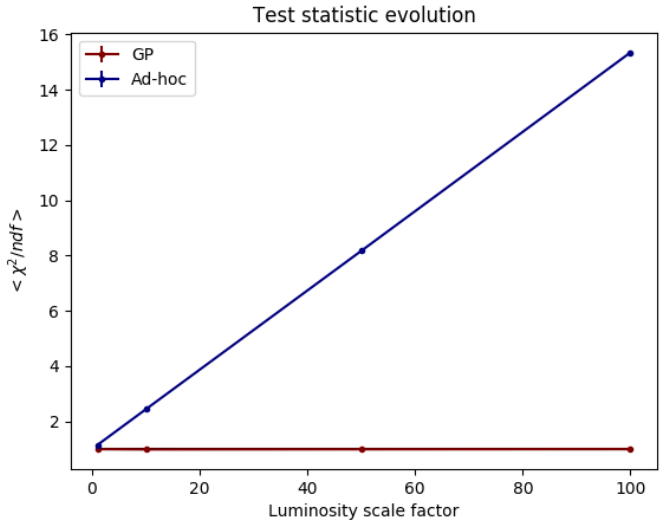


Figure 3.6: In this figure, the mean of the test statistic χ^2/ndf for both the parametric fits (blue points and line) and GP fits (red points and line) is shown. The luminosity scale factor is the value used to scale the toys to the desired integrated luminosity. The errorbars are the 1σ deviations.

3.2 Testing for bias

In this part, the results from the bias test will be presented and studied. It is important to note that the figures in this part might contain a turquoise colored dashed line: This represents an approximation of the expected signal for a SM Higgs boson at a given luminosity.

To test if the background model has a bias when modeling a distribution, a background-only model is fitted to the Asimov data set (described in section 2.3.3) of a signal-free background distribution. The distribution with an integrated luminosity of 36 fb^{-1} was fitted with a GP using only a background kernel, and the residuals are seen in figure 3.7. There is observed some wave-like bias, but the amplitude of this bias is small, less than 3.7% of the expected signal. The uncertainty on the residuals, are larger, around 25% of the expected signal. To see if this bias scales with the integrated luminosity, the modeling was redone on distributions with higher number of events. The result for 360 fb^{-1} is seen in figure 3.8, and both the bias and uncertainty is, relative to the expected signal, decreasing. And when the integrated luminosity is increased even further, figure 3.9 and 3.10 show that the bias and uncertainty decreases even more, relative to the expected signal. The actual values of the bias and uncertainty are larger for higher number of events, but they do not scale proportionally with the integrated luminosity. This mean that modeling using GP does not generate a large bias in the residuals.

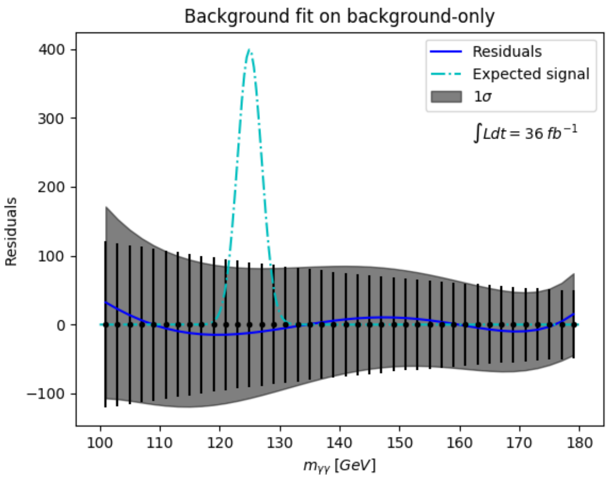


Figure 3.7: This figure shows the residuals of a background-only GP fit on a signal-free Asimov data set corresponding to an integrated luminosity of 36 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviations. The black points is the zero residual indicators.

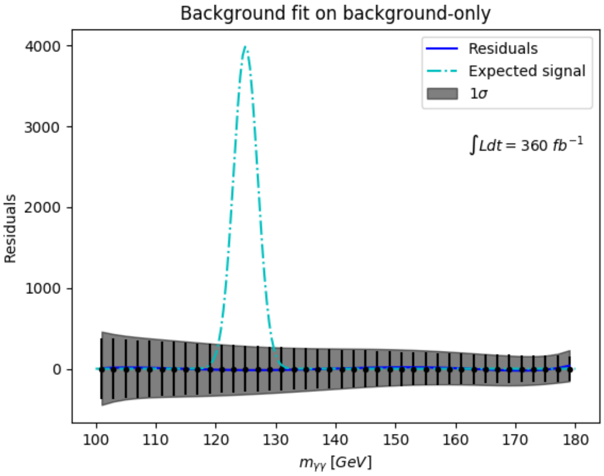


Figure 3.8: This figure shows the residuals of a background-only GP fit on a signal-free Asimov data set corresponding to an integrated luminosity of 360 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviations. The black points is the zero residual indicators.

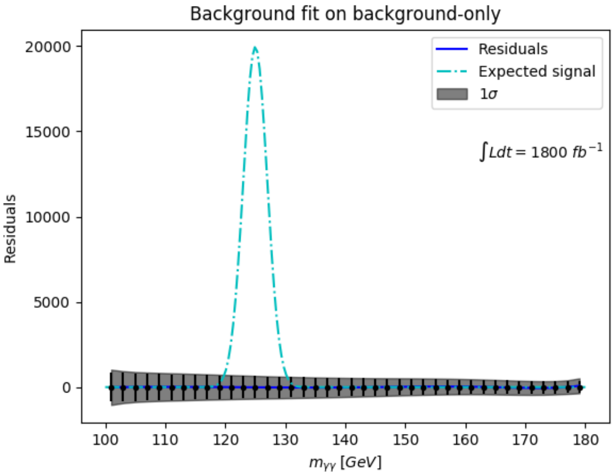


Figure 3.9: This figure shows the residuals of a background-only GP fit on a signal-free Asimov data set corresponding to an integrated luminosity of 1800 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviation. The black points is the zero residual indicators.

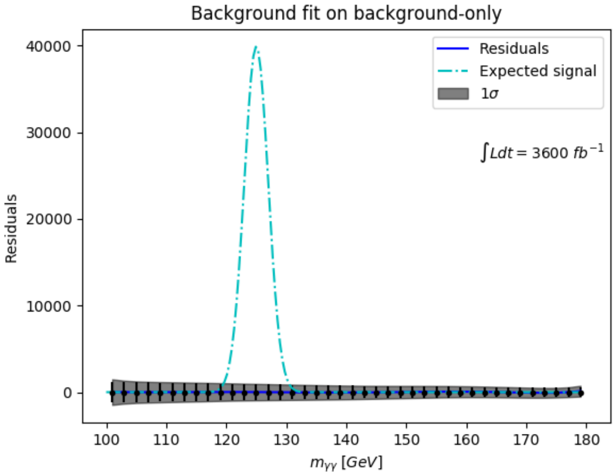


Figure 3.10: This figure shows the residuals of a background-only GP fit on a signal-free Asimov data set corresponding to an integrated luminosity of 3600 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviation. The black points is the zero residual indicators.

Gaussian Processes may be very flexible, therefore it is important to check if a signal present in the distribution is swallowed by the model, or if the model only focuses on the background. A fit was performed using GP with only a background kernel to a distribution containing a signal, corresponding to an integrated luminosity of 36 fb^{-1}. The residuals are seen in figure 3.11. The signal peak is observed in the residuals, which means that the background model does *not* swallow the signal. The residuals are calculated by subtracting the predicted mean from the Asimov data set, and the blue line in figure 3.11 does show increased bias, especially around the signal region. The residuals on the signal do seem to follow the shape of the expected signal with the same amplitude, only shifted down by the bias.

Will the bias seen in figure 3.11 decrease relative to the expected signal, as in the previous section, or will it scale with the integrated luminosity? Figure 3.13 shows the residuals for a fit on a distribution with 360 fb^{-1}, and the signal is still observed, but so is the bias. Comparing the lowest point in the plots in figure 3.11 and 3.13, the residual is increased by a factor 10, just like the integrated luminosity. So it seems this bias does scale. Inspecting figure 3.14 and 3.15, the residual of the signal is still observed, with a shape comparable to the expected signal. The bias of the background residuals persists, and the shape of these has changed a bit; there is more swinging, and also the lowest point has decreased further somewhat. This decrease might explain why the signal residuals in the two lower luminosity plots seems to have a larger amplitude than the ones for higher luminosities.

The reason for this bias, comes from the fact that the background model tries to accommodate the signal without rapidly changing shape. This can be seen in figure 3.12, where the blue is the background modeling on a distribution with a signal injected. This accommodation also gives a more wiggly shape on the residuals for the higher luminosities.

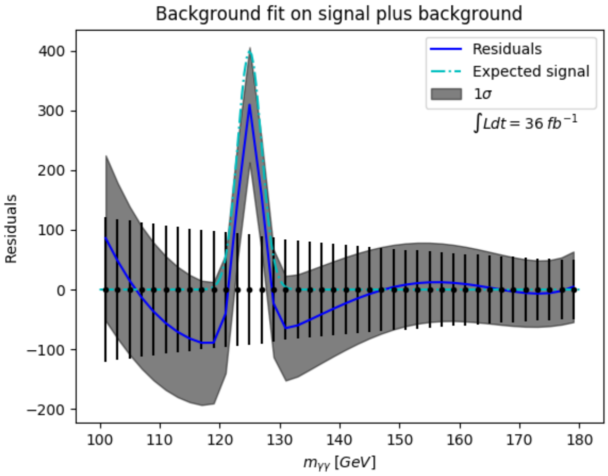


Figure 3.11: This figure shows the residuals of a background-only GP fit on an Asimov data set where a signal is present, corresponding to an integrated luminosity of 36 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviations. The black points is the zero residual indicators.

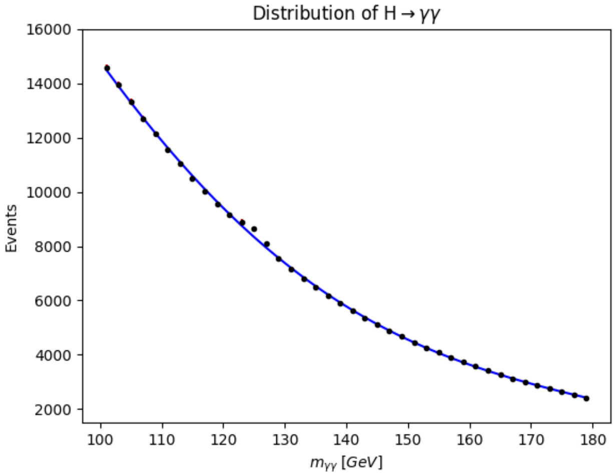


Figure 3.12: This figure shows the distribution of an Asimov data set with a signal present in the distribution corresponding to an integrated luminosity of 36 fb^{-1}, and the distribution has been fitted with a background only GP as the blue line. The errorbars are given in red for the Asimov data set, but they are small and covered by the point.

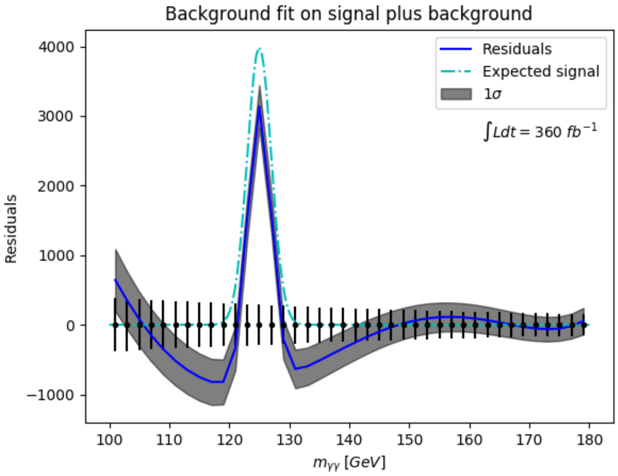


Figure 3.13: This figure shows the residuals of a background-only GP fit on an Asimov data set where a signal is present, corresponding to an integrated luminosity of 360 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviations. The black points and red line is the zero residual indicators.

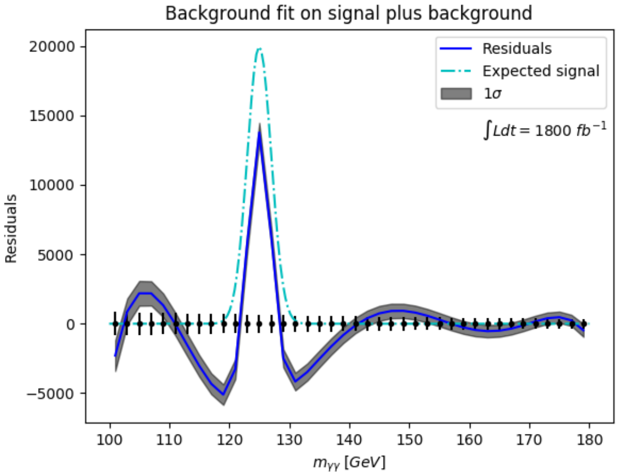


Figure 3.14: This figure shows the residuals of a background-only GP fit on an Asimov data set where a signal is present, corresponding to an integrated luminosity of 1800 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviations. The black points is the zero residual indicators.

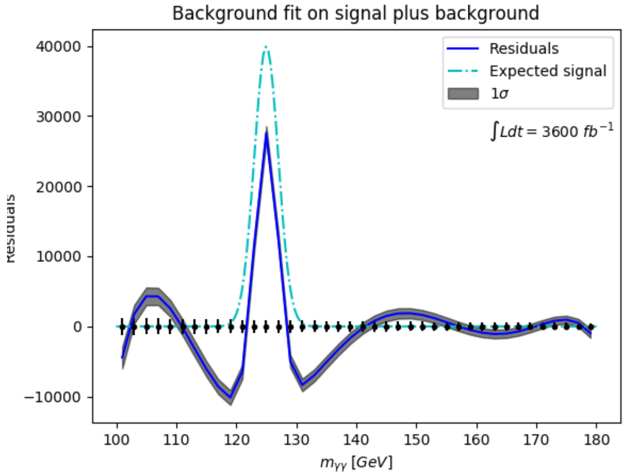


Figure 3.15: This figure shows the residuals of a background-only GP fit on an Asimov data set where a signal is present, corresponding to an integrated luminosity of 3600 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviations. The black points is the zero residual indicators.

As previously observed, there is some bias in the background modeling, so to estimate the impact on a signal extraction, a spurious signal test was performed. A model with both background and signal kernel is fitted to the signal-free distribution, to see if the signal amplitudes were different from zero. In table 3.1, the results for the different luminosities are given. The estimated amplitudes are zero. However, the plot in figure 3.16 does have a small bias in the form of a downward fluctuation. As per the construction of the amplitude term, negative values are not taken into account, and so this negative deviation is not estimated. Fortunately, the bias follows the same behavior as in figure 3.7 so adding the signal kernel did not visibly affect the residuals, and the bias is less than 3.7 % of the expected signal. Figure 3.17 shows that the bias in the residuals decrease with increasing integrated luminosity, as seen before in figure 3.10.

Table 3.1: In this table, the result of a spurious signal test is shown, with a fixed mean at the mass point 125 GeV. The amplitudes of the signal are zero or close to zero.

Luminosity [fb^{-1}]	**Estimated Signal**
36	0
360	0
1800	0.06
3600	0.03

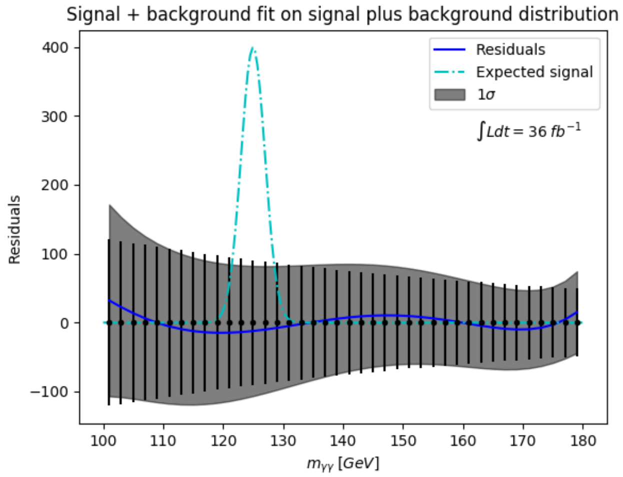


Figure 3.16: This figure shows the residuals of a GP fit using a signal and background kernel on a signal-free Asimov data set, corresponding to an integrated luminosity of 36 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviations. The black points is the zero residual indicators.

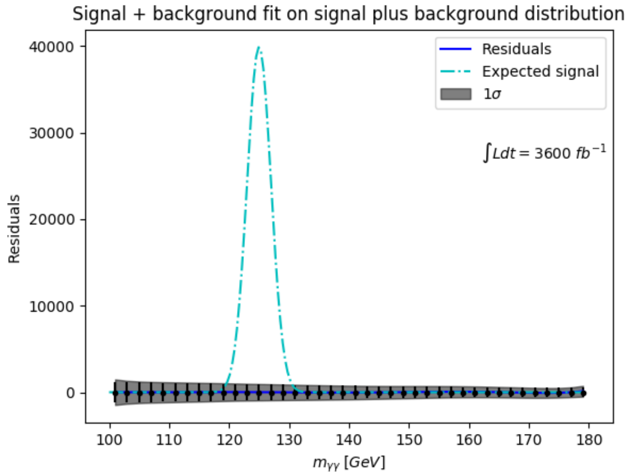


Figure 3.17: This figure shows the residuals of a GP fit using a signal and background kernel on a signal-free Asimov data set, corresponding to an integrated luminosity of 3600 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviation. The black points and red line, is the zero residual indicators.

3.2.1 Different underlying function

The underlying function used so far in the thesis, has been a Bernstein polynomial of 5th order. What if some other function is used instead, will GP still be able to model the background with a low bias or will there be some radical difference? The Asimov data set is now represented by a Epoly2-function fitted to the simulated data, in the same manner as the Bernstein polynomial.

The residuals of a background-only kernel on a signal-free background distribution with an integrated luminosity of 36 fb^{-1}, is seen in figure 3.18. Comparing these residuals with the ones in figure 3.7, the trend is very similar. There is bias observed in the residuals, and there is a wave-structure present. The largest non-edge deviation is shifted to the left in the plot and is somewhat larger than in figure 3.7, while the deviations on the right side are smaller, and the uncertainty is also of the same magnitude. The bias in the signal region is 2.9% of the expected signal, and the spurious signal is estimated to 0.

The integrated luminosity is increased to 360 fb^{-1}, to see if the same evolution is observed for Epoly2 as for Bernstein of 5th order. Figure 3.19 shows the residuals, and the same trend is seen; the bias and uncertainties decrease relative to the expected signal. And increasing even further, the relative bias and uncertainties continue to decrease. There is one difference that must be noted; when the integrated luminosity is 3600 fb^{-1}, the residuals in figure 3.10 is practically zero, but in figure 3.20 there are still some deviation in the signal region. This indicates that Gaussian Processes are not totally independent on the underlying function, but within the framework presented in this thesis, this dependence is negligible.

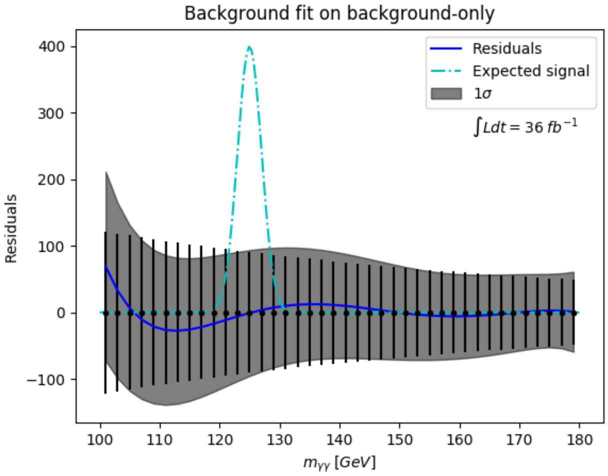


Figure 3.18: This figure shows the residuals of a background-only GP fit on a signal-free Asimov data set where the underlying function is the Epoly2, corresponding to an integrated luminosity of 36 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviation. The black points is the zero residual indicators.

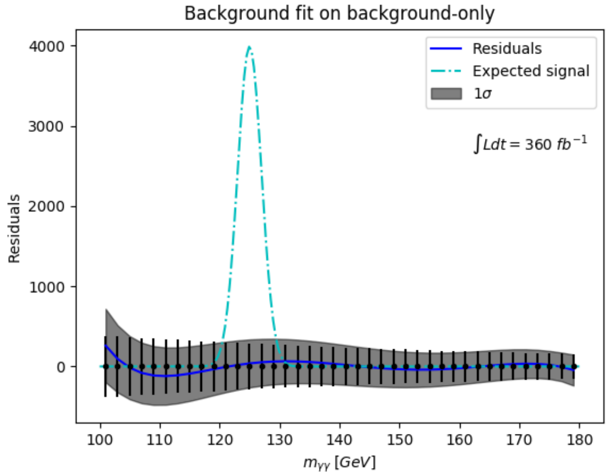


Figure 3.19: This figure shows the residuals of a background-only GP fit on a signal-free Asimov data set where the underlying function is the Epoly2, corresponding to an integrated luminosity of 360 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviation. The black points is the zero residual indicators.

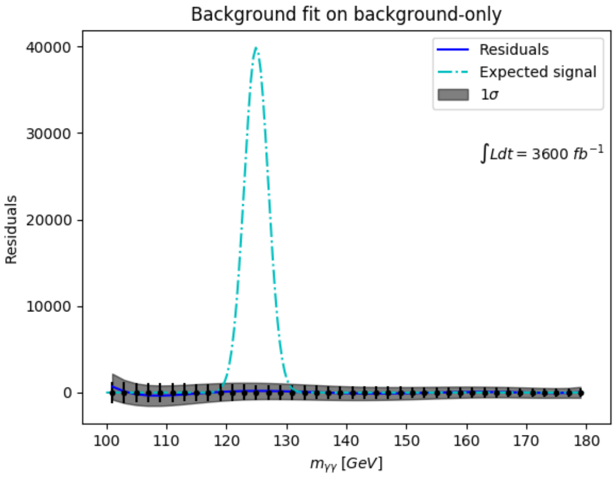


Figure 3.20: This figure shows the residuals of a background-only GP fit on a signal-free Asimov data set where the underlying function is the Epoly2, corresponding to an integrated luminosity of 3600 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is the expected signal, and the grey area are the 1σ standard deviation. The black points is the zero residual indicators.

3.2.2 The Matérn kernel

There are other kernels that may be used for background estimation instead of the squared exponential. The Matérn kernel (described in section 2.2.3) was also tested for the background modeling to see if it would perform better than the more commonly used squared exponential. This kernel has a parameter ν which controls the complexity of the kernel, and the hyperparameter length scale l. The parameter ν was fixed to 3/2, and l was free for optimization. A constant kernel (as described in 2.2.3) is multiplied with the Matérn kernel to represent the amplitude A. The result from a fit using only the background kernel on an Asimov data set with a signal injected is shown in figure 3.21 and figure 3.22. The signal is partly absorbed for 36 fb^{-1} and almost completely absorbed for 3600 fb^{-1}. Therefore, the Matérn class of kernels are not suited for background estimation in this problem, as the signal is swallowed by the fit due to the flexibility of the kernel. These kernels are actually products of an exponential function and a polynomial, making them more adaptable for rapid deviations.

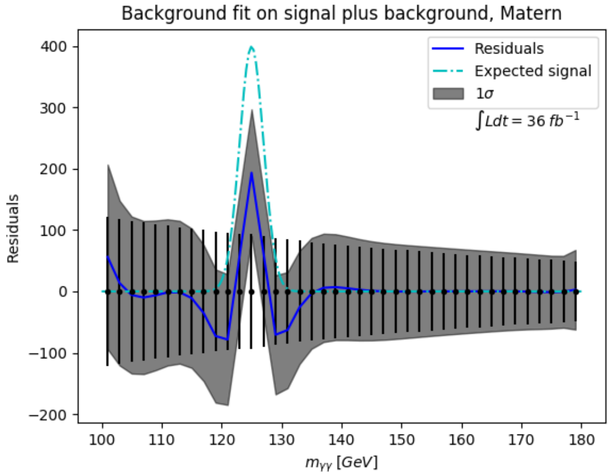


Figure 3.21: This figure shows the residuals of a GP fit using a Matérn kernel as the background kernel on an Asimov data set with a signal present in the distribution, corresponding to an integrated luminosity of 36 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviation. The black points is the zero residual indicators.

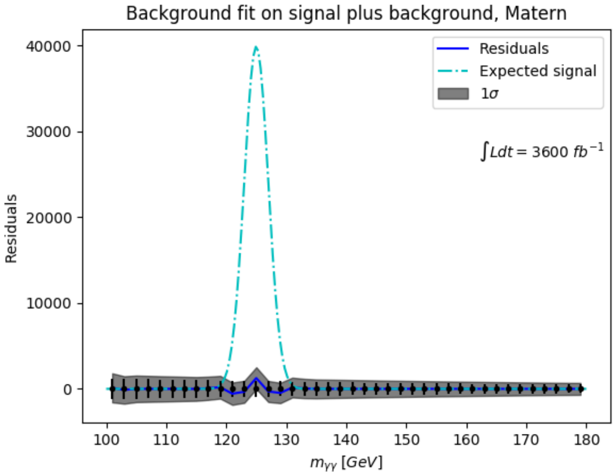


Figure 3.22: This figure shows the residuals of a GP fit using a Matérn kernel as the background kernel on an Asimov data set with a signal present in the distribution, corresponding to an integrated luminosity of 3600 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviation. The black points is the zero residual indicators.

3.3 Signal estimation and modeling

To study how well Gaussian Processes can model a distribution with a signal present, a distribution with a signal injected was fitted using GP with both a signal and background kernel. The function used to simulate a signal distribution was a Gaussian distribution as described in section 2.3.3, and the signal kernel for the GP is given in section 2.3.4.

3.3.1 Estimating a signal amplitude

When doing precise measurements of the Higgs boson properties, extracting the estimated parameters of the fitted model with as small as possible uncertainty is important, as these provide information about the Higgs boson production. So, even though GP does fit the distribution very well, can the hyperparameters serve this purpose? To study this, the signal amplitude was retrieved from IMINUIT with the calculated uncertainty. The standard deviation of the extracted hyperparameters are calulated as given in section 2.3.4. The results are displayed in table 3.2.

Table 3.2: This table shows the estimated signal amplitudes, with uncertainties, for the different integrated luminosities. The fit was performed with a GP using a signal and background kernel on an Asimov data set with a signal present. The expected signal is the known amplitude of the injected signal.

Luminosity [fb^{-1}]	**Estimated Signal**	**Expected Signal**
36	373.2 ±271.0	398.9
360	3955.0 ±2778.6	3989.4
1800	19927.8 ±14096.5	19947.1
3600	39901.5 ±28218.5	39894.2

The *estimated signal* in table 3.2 is the signal amplitude from IMINUIT, and expected is the normalized amplitude of the signal injected into the distribution. The estimated signal amplitude for a distribution at 36 fb^{-1}, is quite close to the actual value of the amplitude. The uncertainty of this parameter is large, about 70 % of the estimation. Such a large uncertainty is not desired as the measurement should be estimated with great precision. It is observed that for increasing integrated luminosity, the estimated signal amplitude continues to be close to the actual amplitude, but the large uncertainty persists. Therefore, this method of using a signal kernel combined with a background kernel for precisely estimating the parameters of the signal distribution, is not the optimal one. In the literature, there is no mention of errors on the hyperparameters themselves, only errors related to the predictions and target values. Either the kernels that are used must be changed or tweaked, or other ways to estimate the parameters of interest must be used. Also, the possibility of taking negative signals into account in the Spurious Signal fit, should be implemented.

Alternative method for signal estimation

The method for extracting the estimated number of signal events from the sum of the posterior mean, as described in section 2.3.4, was tested using an Asimov data set with an injected signal. The sum of the posterior covariance matrix is the variance of the estimated number of signal events. To retrieve the error on the estimated number of signal events, the square root of the covariance-sum was taken. Table 3.3 shows the results of the modeling for all integrated luminosities. The *estimated signal* is the calculated signal, and the *expected signal* is the signal injected into the distribution with a bin width of 2.

Table 3.3: This table shows the estimated number of signal events, with uncertainties, for the different integrated luminosities. The fit was performed with a GP using a signal and background kernel on an Asimov data set with a signal present. The expected signal is the known number of events of the injected signal.

Luminosity [fb^{-1}]	**Est. nr of sig. events**	**Exp. nr of sig. events**
36	895.5 ± 198.9	1000
360	9867.3 ± 655.0	10000
1800	49907.4 ± 1482.4	50000
3600	99989.0 ± 2106.2	100000

The estimated number of signal events for a distribution at 36 fb^{-1} is 89.6% of the expected signal yield with an uncertainty of 22% on the estimation. This result is not optimal. However, increasing the integrated luminosity to 360 fb^{-1}, improved the estimation to 98.7% of the expected signal with an uncertainty of only 6.6% of the estimated number of events. When increasing the luminosity even further, the estimated signal yield was 99.8% of the expected signal count with an uncertainty of 2.9% on the estimation. The last test was for an integrated luminosity of 3600 fb^{-1}. This enhanced the signal estimation to 99.989% of the expected number of events with an uncertainty of 2.1%. Therefore, this method of using the linearity of GP, and the sum of the predicted signal mean and covariance, is a very promising technique for extracting the signal yield from a distribution.

3.3.2 Modeling a signal over a background distribution

The residuals from modeling a distribution corresponding to an integrated luminosity of 36 fb^{-1}, is seen in figure 3.23. Some bias is seen in the background residuals, but this is not worse than in figure 3.7. There are a small spike in the residuals in the signal region, which is likely due to the fact that in addition to fitting the signal, the signal kernel absorbs some of the remaining bias in the background model. The background-bias 1.6% larger than in figure 3.7, meaning that a small part of the background tries to accommodate some of the signal. The uncertainty is comparable in size as in figure 3.7, but now there is an increase in the signal region, which is added to the residual uncertainty from the signal kernel. Figure 3.24 shows the distribution of an Asimov data set with as signal injected corresponding to an integrated luminosity of 36 fb^{-1}, fitted with a GP using both a signal kernel and background kernel.

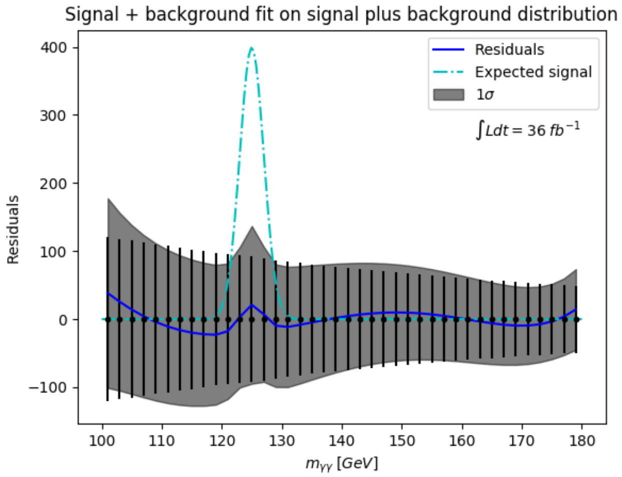


Figure 3.23: This figure shows the residuals of a GP fit using a signal and background kernel on an Asimov data set with a signal present in the distribution, corresponding to an integrated luminosity of 36 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviation. The black points is the zero residual indicators.

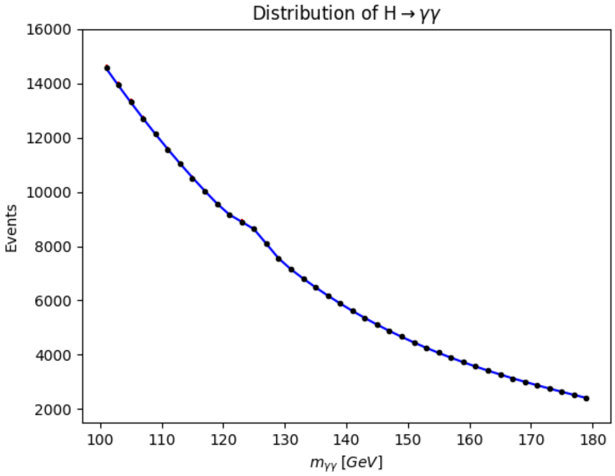


Figure 3.24: This figure shows the distribution of an Asimov data set with a signal present in the distribution corresponding to an integrated luminosity of 36 fb^{-1}, and the distribution has been fitted with a signal+background GP as the blue line. The errorbars are given in red for the Asimov data set, but they are small and covered by the point.

Increasing the integrated luminosity to 360 fb^{-1}, the residuals in figure 3.25 have the same behavior as in 3.8. The bias in the residuals, and their uncertainties, decrease relative to the signal. Continuing to increase the integrated luminosity to 1800 fb^{-1} and further to 3600 fb^{-1} decreases the bias and uncertainties relative to the signal even more, as seen in figures 3.26 and 3.27. GP does a very good job of fitting the representative data sets without overfitting, and as the integrated luminosity will only increase in the future, it seems that GP is very suitable for modeling the distribution with and without an injected signal.

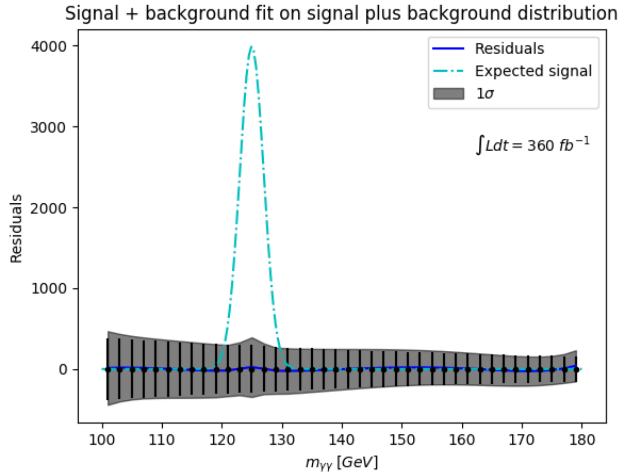


Figure 3.25: This figure shows the residuals of a GP fit using a signal and background kernel on an Asimov data set with a signal present in the distribution, corresponding to an integrated luminosity of 360 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line isan approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviation. The black points is the zero residual indicators.

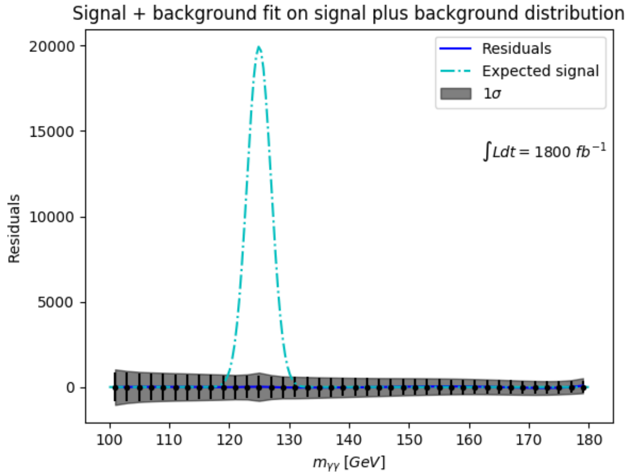


Figure 3.26: This figure shows the residuals of a GP fit using a signal and background kernel on an Asimov data set with a signal present in the distribution, corresponding to an integrated luminosity of 1800 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviation. The black points is the zero residual indicators.

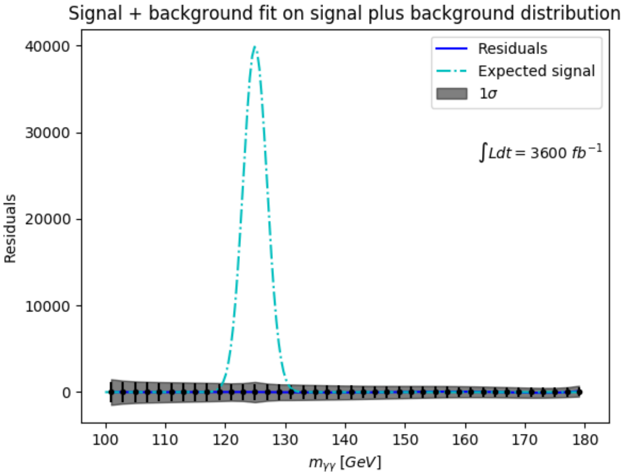


Figure 3.27: This figure shows the residuals of a GP fit using a signal and background kernel on an Asimov data set with a signal present in the distribution, corresponding to an integrated luminosity of 3600 fb^{-1}. The dark blue line is the residuals, the turquoise dashed line is an approximation of the expected signal for a SM Higgs boson, and the grey area are the 1σ standard deviation. The black points is the zero residual indicators.

Chapter 4

Conclusion and outlook

In this thesis, the performance of the machine learning method Gaussian Processes used to model a background distribution similar to the background to the $H \to \gamma\gamma$ signal at LHC with and without an injected Gaussian signal at 125 GeV, has been studied. The goal for this study, was to find out if GP is a viable alternative to parametric background modelling (specifically for the $H \to \gamma\gamma$ channel) and if it can be applied without modification for ever-increasing luminosity of the data.

Gaussian Processes manages to find the underlying function used to generate the toy distributions resembling data with an integrated luminosity of 36 fb^{-1}, with a $\chi^2/ndf = 0.993$. When the integrated luminosity was increased, the Gaussian Process continued to perform well, managing to find the underlying function. The mean of the χ^2/ndf was shifted from 0.92 to 0.993 by using the effective number of degrees of freedom, instead of counting the number of hyperparameters in the kernel. This shows the importance of using the correct calculation method for finding the number of degrees of freedom.

The modeling of a signal-free distribution using Gaussian Processes with only a background kernel, does give rise to some bias for the 36 fb^{-1} distribution. But this bias is less than 3.7% of the expected signal. The uncertainty of the residuals is about 25% of this expected signal, which is too large. When the integrated luminosity increases, however, this bias and uncertainty diminishes. When the signal-free distribution is modelled by Gaussian Processes using both a signal and background kernel, the spurious signal is estimated to be zero, however, the signal kernel does not take into account negative amplitudes.

The residuals do contain the injected signal, when modeling a distribution with a signal injected with Gaussian Processes with only a background kernel, meaning that the Gaussian Process tested with a squared exponential as kernel with a length scale l above 70, does not "swallow" the signal. However, a larger bias is now present in the residuals. This is because GP tries to accommodate the signal without varying rapidly. When a signal kernel is used together with the background kernel for the modeling of a distribution with a signal present, the Gaussian Process manages to fit the distribution resulting in small residuals, which decreased with increasing integrated luminosity. For a distribution with an integrated luminosity of 36 fb^{-1}, the bias is 5.3% of the expected signal, but this decreases relative to the expected signal distribution for increasing luminosities. The bias seen when the background kernel tries to accommodate for the signal is reduced by 18.7%.

Extracting an estimate of the signal amplitude using the optimized hyperparameter representing the signal amplitude of the signal kernel, did not prove successful. The values of the extracted estimation did correspond well with the known amplitudes, but the uncertainties of the estimated signal amplitudes were large. Therefore, precise measurements of the parameters of a signal using a signal kernel, is not the optimal approach.

Using the linearity of the GP to estimate the signal yield, proved successful. The estimation, relative to the expected number of events, improved from 89.6% to 99.989% when the integrated luminosity was increased from 36 fb^{-1} to 3600 fb^{-1}. The uncertainties were 22% and 2.1%, respectively. These results indicate that this approach is extremely promising for estimating the signal parameters.

The Matérn kernel did not perform well as a background kernel, as it swallowed the signal injected into the distribution.

Comparing two fits on a signal-free distribution using only a background kernel, where one fit is on the Bernstein polynomial of 5th order and the other is on the Epoly2 function, proved that while there are differences, these are negligible. Gaussian Processes, for the purpose of this analysis, is independent of the underlying function.

This finding is in agreement with the result found by Frate et. al. [2], who performed used the same method on a very different background distribution.

Outlook

The results from this study looks very promising, although there are still many unresolved questions. The first approach would be to apply the method to actual data. Then, exploring how a negative signal amplitude could be allowed in the kernel-only method might be useful. It is also important to continue to explore the estimation of the number of signal events (and it's uncertainty) in the GP-framework, using the predicted signal mean and covariance. If these methods does not prove fruitful, another strategy could be to combine a background modeling GP and a parametric fit to extract signal parameters. Last, it could be interesting to create and study different kernels that can implement more information from the knowledge about the $H \to \gamma\gamma$ channel.

Bibliography

[1] M. Aaboud, G. Aad, B. Abbott, O. Abdinov, B. Abeloos, S. H. Abidi, O. S. AbouZeid, N. L. Abraham, H. Abramowicz, H. Abreu, and et al. Measurements of Higgs boson properties in the diphoton decay channel with 36 fb-1 of pp collision data at s=13 TeV with the ATLAS detector. *Physical Review D*, 98(5), Sep 2018. ISSN 2470-0029. doi: 10.1103/physrevd.98.052005. URL `http://dx.doi.org/10.1103/PhysRevD.98.052005`.

[2] Meghan Frate, Kyle Cranmer, Saarik Kalia, Alexander Vandenberg-Rodes, and Daniel Whiteson. Modeling smooth backgrounds and generic localized signals with gaussian processes, 2017.

[3] G. Aad, T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, A.A. Abdelalim, O. Abdinov, R. Aben, B. Abi, M. Abolins, and et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716(1): 1–29, Sep 2012. ISSN 0370-2693. doi: 10.1016/j.physletb.2012.08.020. URL `http://dx.doi.org/10.1016/j.physletb.2012.08.020`.

[4] S. Chatrchyan, V. Khachatryan, A.M. Sirunyan, A. Tumasyan, W. Adam, E. Aguilo, T. Bergauer, M. Dragicevic, J. Erö, C. Fabjan, and et al. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1):30–61, Sep 2012. ISSN 0370-2693. doi: 10.1016/j.physletb.2012.08.021. URL `http://dx.doi.org/10.1016/j.physletb.2012.08.021`.

[5] University of Zurich. Standard model. URL `https://www.physik.uzh.ch/groups/serra/StandardModel.html`. Accessed: 2019-12-11.

[6] Peter W. Higgs. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.*, 13:508–509, Oct 1964. doi: 10.1103/PhysRevLett.13.508. URL `https://link.aps.org/doi/10.1103/PhysRevLett.13.508`.

[7] F. Englert and R. Brout. Broken Symmetry and the Mass of Gauge Vector Mesons. *Phys. Rev. Lett.*, 13:321–323, 1964. doi: 10.1103/PhysRevLett.13.321. [,157(1964)].

[8] Fred Bellaiche. What is the higgs boson anyway? URL `http://www.quantum-bits.org/?p=233`.

[9] Mark Thomson. *Modern Particle Physics*. Cambridge University Press, 2013. doi: 10.1017/CBO9781139525367.

[10] Rende Steerenberg. LHC report: Protons: mission accomplished.

[11] G. Aad et al. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3:S08003, 2008. doi: 10.1088/1748-0221/3/08/S08003.

[12] N. S. Bernstein. Démonstration du théoréme de weierstrass fondée sur le calcul des probabilités. *Comm. Kharkov Math. Soc. 13*, 1912.

[13] Karl Pearson F.R.S. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900. doi: 10.1080/14786440009463897. URL `https://doi.org/10.1080/14786440009463897`.

[14] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71(2), Feb 2011. ISSN 1434-6052. doi: 10.1140/epjc/s10052-011-1554-0. URL `http://dx.doi.org/10.1140/epjc/s10052-011-1554-0`.

[15] Fons Rademakers Rene Brun. Root - an object oriented data analysis framework. *Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81-86*, 1996.

[16] D. S. Sivia and J. Skilling. *Data Analysis - A Bayesian Tutorial*. Oxford Science Publications. Oxford University Press, 2nd edition, 2006.

[17] Chris K. I. Williams Carl Edward Rasmussen. *Gaussian Processes for Machine Learning*. MIT Press books, 2006.

[18] Mr. Bayes and Mr. Price. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions (1683-1775)*, 53:370–418, 1763. ISSN 02607085. URL `http://www.jstor.org/stable/105741`.

[19] P.S. Laplace. *Théorie analytique des probabilités.* Courcier, 1820. URL https://books.google.no/books?id=J7gl92QRdZMC.

[20] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* 1965.

[21] Lilian Smedstad. *The Search for the Standard Model Higgs Boson in $H\rightarrow\gamma\gamma$ Decays with the ATLAS Detector in 4.9 fb^{-1} of 2011 Data at $\sqrt{s}=7$ TeV.* PhD thesis, University of Oslo, 2013.

[22] Sivaram Ambikasaran, Daniel Foreman-Mackey, Leslie Greengard, David W. Hogg, and Michael O'Neil. Fast direct methods for gaussian processes, 2014.

[23] iminuit team. iminuit – a python interface to minuit. https://github.com/scikit-hep/iminuit. Accessed: 2019-12-05.

[24] F. James and M. Roos. Minuit – a system for function minimization and analysis of the parameter errors and correlations. *Computer Physics Communications*, 10:343–367, December 1975. doi: 10.1016/0010-4655(75)90039-9.

[25] Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–510, 1989. ISSN 00905364. URL http://www.jstor.org/stable/2241560.