Expanding limits of statistical methods for high energy physics

Viktor Ananiev

Main supervisor:
Prof. Alexander L. Read

Department of Physics

The Faculty of Mathematics and Natural Sciences

Abstract

The Standard Model serves as the fundamental framework for describing elementary particles and their interactions. During my doctoral studies as a member of the ATLAS collaboration, my primary focus centered on statistical methods, particularly their application in searches searches for new particles. Even though all particles within the Standard Model have been discovered, particle searches remain relevant, especially in anticipation of future runs of the Large Hadron Collider, such as Run 3 and the High Luminosity LHC, spanning several decades from today. These upcoming experiments aim to uncover new particles that may provide explanations for missing components of the Standard Model, such as Dark Matter, Neutrino masses, or Gravitational forces.

A key contribution is a novel method for calculating the global significance in the search for new resonances within particle physics. This method employs a Gaussian Process to model the significance in the search area, utilizing an approximate covariance matrix derived from a carefully designed set of Asimov datasets. This innovative approach enables precise and computationally efficient calculation of the trials factor.

Notably, this research has yielded the SigCorr framework, a versatile tool for constructing analytical pipelines in particle physics. While the published works delve into specific methodologies, SigCorr holds the promise of enabling exploration across a broad spectrum of approaches to trials factor calculations.

I had an opportunity to contribute to a range of significant open-source projects through collaborations facilitated by my membership in the Marie Curie Innovative Training Network "INSIGHTS". In addition to the aforementioned SigCorr, I also developed Cicliminds for climate data visualization, and contributed enhancements to the Boost.Math C++ library, the Sherpa Monte Carlo generator, the MGVI.jl Bayesian inference framework, and the Statsmodels Python library for descriptive statistics.

In summary, this doctoral thesis presents groundbreaking statistical methods and software for high-energy physics, aligning with the mission of INSIGHTS. Its contributions could greatly improve precision and efficiency in particle searches and statistical analysis in physics and other fields.

Sammendrag

Standardmodellen fungerer som det grunnleggende rammeverket for å beskrive elementære partikler og deres interaksjoner. Under doktorgradsstudiene mine som medlem av ATLAS-kollaborasjonen, var mitt primære fokus på statistiske metoder, spesielt søket etter nye partikler. Selv om alle partikler i standardmodellen er oppdaget, forblir søket etter nye partikler relevante, spesielt i påvente av fremtidige "runs" av Large Hadron Collider, som Run 3 og den planlagte High Luminosity LHC, som spenner over flere tiår fra i dag. Disse kommende eksperimentene tar sikte på å avdekke nye partikler som kan gi forklaringer på manglende komponenter i standardmodellen, slik som mørk materie, nøytrinomasser og gravitasjonskrefter.

Et nøkkelbidrag er en ny metode for å beregne den globale signifikansen i jakten på nye partikler innen partikkelfysikk. Denne metoden bruker en Gaussisk prosess for å modellere signifikansen i søkeområdet, ved å bruke en omtrentlig kovariansmatrise avledet fra et nøye designet sett av Asimov datasett. Denne innovative tilnærmingen muliggjør presis og effektiv beregning av prøvefaktoren.

Spesielt har forskningen gitt opphav til Sigcorr-rammeverket, et allsidig verktøy for å konstruere analytiske arbeidsflytkjeder for partikkelfysikk. Selv om de publiserte verkene fokuserer på spesifikke metoder, har Sigcorr potensialet til å muliggjøre utforskninger som spenner over et bredt spekter av tilnærminger i beregningene av prøvefaktorer.

Jeg hadde mulighet til å bidra til en rekke betydelige *open source*-prosjekter gjennom samarbeid som er tilrettelagt av mitt medlemskap i Marie Curie Innovative Training Network "INSIGHTS". I tillegg til den nevnte Sigcorr, utviklet jeg også Cicliminds for klimadatavisualisering, og bidro med forbedringer for Boost.Math C ++ biblioteket, Sherpa Monte Carlo Generator, MGVI.JL Bayesian Inference Framework, og Statsmodels Python Library for beskrivende statistikk.

Oppsummert presenterer denne doktorgradsavhandlingen banebrytende statistiske metoder for høyenergifysikk, og samsvarer med INSIGHTS' oppdrag. Bidragene kan forbedre presisjonen og effektiviteten i partikkelsøk og statistisk analyse innen fysikk og andre felt.

Acknowledgements

This thesis would not have come to existence without my supervisor. Alex Read, thank you for giving me the opportunity to pursue this challenge. You were always available for me, patiently answering my many questions, engaging into fruitful discussions and brainstorming, correcting the articles, and I mean both: articles as publications and the ones used before nouns. Thank you for supporting me all the way!

To all my friends and colleagues who I met at the HEPP and THEP sections at the Department of Physics. Thank you for your warm company, lunch quizzes, and for switching to English each time I appeared on the horizon. Erik Adli and Farid Ould-Saada, thank you for being supportive and proactive, making my work in HEPP and ATLAS stress-free. Heidi Sandaker and Are Raklev, thank you for your guidance, thoughtful feedback and consistent support throughout these years. The soul of our group, Eli, thanks to you I felt included in the community from my first day at the UiO.

My friends and colleagues from the corporate world of Borshch, Genomics and News Media, time with you has never felt like work to me *(it was meant to be a good thing)*. These occasional excursions into the tech world kept me motivated and set the pace for my academic advancements. Thank you to all the kind people I work with in the world of Krill today, for cheering me on and encouraging me to grind through!

To my teachers and mentors in Ukraine, from school and university, who taught me invaluable lessons in science and in life.

I would like to thank my closest friends from Ukraine, including the Boardgames gang, the Dutch gang, the Gang of 22 and the Core gang. You have been my mental support through many years, and form the corner of my emotional stability today. Vlad, thank you for our regular Rustathons, which have evolved into regular writing sessions, but only temporarily. A special thank you goes to Mykola. Our numerous chats on life, science, tech and various other topics not only kept me afloat but also facilitated some scientific contributions to this thesis.

To the friends who contributed to my life during and beyond the period of my PhD studies: thank you. To Niv and Yakir for inspiring and mentoring; Stacy for good memes, both digital and real life; Katya for our studdy-buddies; Pawel for showing me the world outside my laptop; and Thay for all the parties and chats we had.

Dear Kristian and Hannah, through these years we have become close friends, moreover, our families have grown close as well. Thank you for being open,

welcoming, and supportive, especially during the hard times. I am grateful for the holidays we have spent together, the boardgame nights, cultural events, and all the good memories we have created and those to come.

My family, thank you for supporting me in all my decisions, for paying attention, and always finding ways to help. I have come this far thanks to you.

Nastya, thank you for always being around, even though, sometimes, it can be difficult between countries; for understanding me when it seems impossible; for being my stand-up partner and making my life a good show, and for your patience and support every day.

I would like to extend my thanks to Norway, and in particular to Oslo Commune, for providing exceptional support to my family during the ongoing war in our home country.

I am grateful to the Military Forces of Ukraine for keeping my family, friends and all the people safe during the Russian invasion into Ukraine. To all the organizations, Ukrainian and International, to volunteers, and to everyone who supports Ukraine and stands for the justice and peace: *Дякую!*

Contents

List of Figures

Chapter 1

Introduction

Paper geckos bend.
Fairy tales become math —
the Standard Model.

The Standard Model (SM) of particle physics is the commonly accepted model of the Universe at the microscopic level. It assumes that the world is built of elementary particles, and the Standard Model is the essence of our knowledge about the properties of these particles and their interactions (Figure 1.1). Statistical data analysis plays an important role in extracting this knowledge from complex measurements probing the intricate fabric of the Universe.

Advancement of statistical methods for data analysis was one of the major goals of the International Training Network of Statistics for High Energy Physics and Society (INSIGHTS ITN) [1], which I am proud to have been a member of. The network was focusing on multivariate analysis, parametric modeling, and Bayesian computation, to address the challenges posed by high-energy physics experiments, particularly at the Large Hadron Collider, training a new generation of physicists and fostering collaborations with interdisciplinary partners for broader societal impact. As a consequence, a significant part of my research achievements is associated with statistical methods in high energy physics.

Standard Model of Elementary Particles

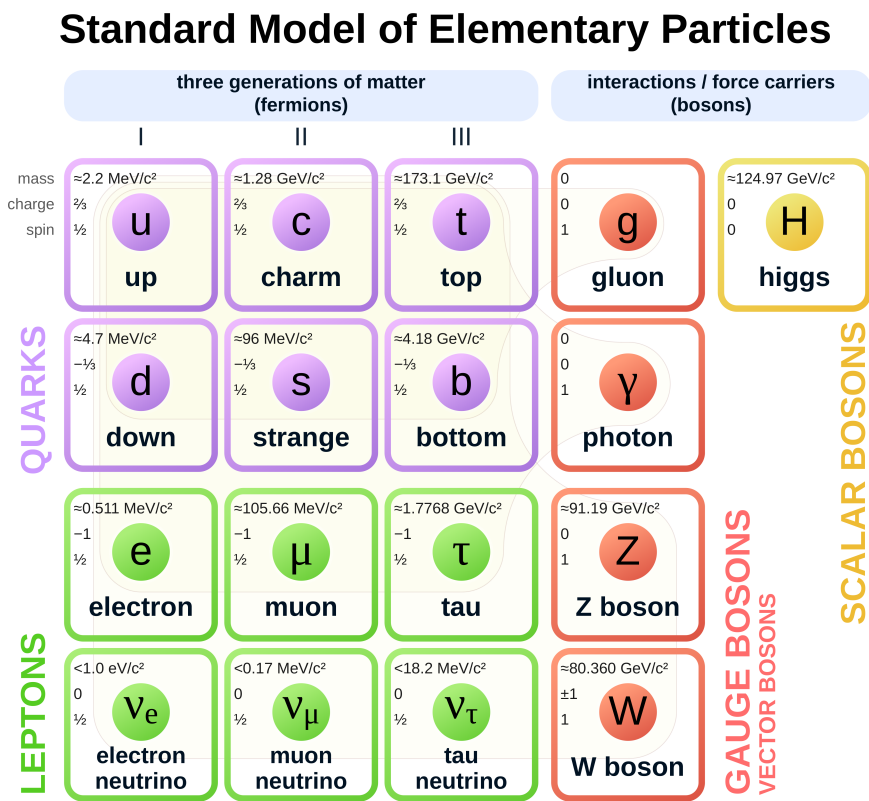Figure 1.1: Elementary particles of the Standard Model divided into two groups: *fermions* which constitute all known matter, and *bosons* which mediate the interactions. The 6 anti-quarks and 6 anti-leptons are not shown. (Taken from Ref. [2].)

Development of the Standard Model. Historically, the first experimentally confirmed particles predicted by the SM were $W^{\pm}$ and Z^0, so called *weak bosons*.[1]

They were observed at the European Organization for Nuclear Research (CERN) in 1983 using the Super Proton Synchrotron [3, 4]. This was a major success of CERN and the SM, which encouraged the scientific community to continue searching for other SM particles not yet observed at the time.

By the time of discovery of the weak bosons most of the fermions had been already discovered. The only ones missing were the top quark, which got confirmed in 1995 [5] by the CDF and DØ experiments [6], and the tau neutrino, first directly observed in 2000 [7]. It was already known, therefore, that most of the fermions in the SM have a non-zero mass.

It turns out that naive mass terms in the fermionic equations make quantum calculations in the Standard Model (SM) divergent at high energies [8]. However, there is a procedure called *renormalization*, which suggests to conceal these infinite terms in the fundamental parameters of the theory. Since experimental observables would include the quantum corrections, the divergent raw values will not make any harm. According to 't Hooft and Veltman [9], the renormalization procedure can always be implemented for theories that are symmetric in a particular way, respecting the so-called Gauge symmetry.

The famous Higgs mechanism [10–12] suggests a way to include the fermionic mass terms into the theory while respecting the Gauge symmetries required for renormalizability. In the big picture, the electro-weak sector consists of a complex vector weak triplet and a complex singlet B^0, with a scalar Higgs doublet constituting the Higgs sector. Consequently, the Higgs mechanism assumes that the world we observe can be described in terms of small excitations of the Higgs field around some local minimum of its potential. The approximate model including these excitations can, again, be reformulated as a particle model but less symmetric. In other words, the Higgs mechanism is a symmetry-breaking procedure.

As a result of the symmetry breaking [13], the weak triplet, B^0 and the Higgs doublet are replaced by massive $W^\pm$, Z, Higgs, and massless photon fields. Yukawa couplings, initially describing the interactions between fermions and the Higgs field, generate fermion masses and the couplings of the fermions to the Higgs boson, which, in turn, constitute the intimate connection between the masses of fermions and the decay rates of the Higgs boson to the fermions.

The search for the Higgs boson, the direct evidence of the existence of the Higgs field needed for the SM to be consistent, was one of the major reasons for building the Large Hadron Collider (LHC) (Figure 1.2). It was built by CERN between 1998 and 2008 and lies in a 27 kilometer long circular tunnel, around 100 meters underground at the France-Switzerland border near Geneva.

The LHC is a machine designed for accelerating and colliding protons and heavy nuclei, such as lead, at high energies. The data produced as a result of these collisions is recorded with the detectors installed across the LHC ring. The Compact Muon Solenoid (CMS) and A Toroidal LHS Apparatus (ATLAS) are two large multi-purpose detectors installed at two of the points where the LHC beam

---

[1]Technically, most of the leptons and quarks were the first, but the discovery of the weak bosons confirmed the existence of the electroweak sector, which was a major step verifying the unified theory of interactions, i.e. the Standard Model we know today.
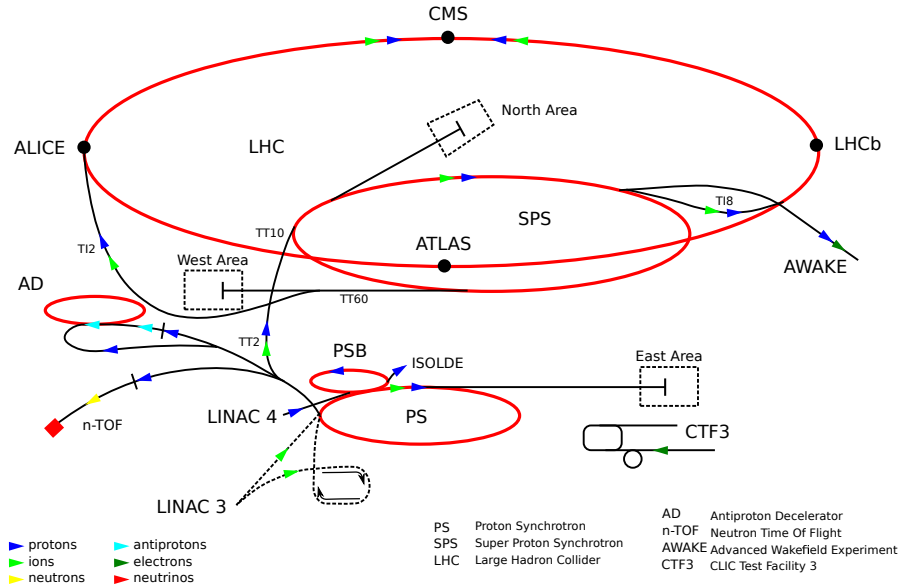
Figure 1.2: CERN Accelerator Complex. (Taken from Ref. [14]).

crossings produce collisions. The corresponding collaborations were the ones that discovered the long-sought Higgs boson in 2012 [15, 16].

Scientific accomplishments presented in this thesis are directly related to the ATLAS Collaboration, as I was an ATLAS member while doing the underlying research, and, therefore, will focus on the ATLAS detector in more detail.

The search for the Standard Model Higgs boson was used as a benchmark during the design of the ATLAS detector. The design was supposed to cover a wide range of the Higgs decay channels, sensitivity of the measurements of which depended on the then unknown mass of the boson. To illustrate how strongly the search is affected by the Higgs mass, we refer to Figure 1.3, which shows the dependency of the branching ratios of the Higgs decay channels on the Higgs boson mass. The mass range on the plot is constrained to 100-200 GeV, where the lower bound of 114.4 GeV was set by the Large Electron-Positron Collider [17], the predecessor of the LHC, and the approximate upper bound was estimated from the electroweak processes [18].

Although the branching ratio of a channel does not necessarily reflect the sensitivity of this channel, it, at least, can help us grasp the variability the experiment should be prepared for. This and other new physics searches motivated such aspects of the detector construction as high acceptance in pseudorapidity and azimuthal angle coverage, very good calorimetry and charged particle momentum resolution with a special focus on high p_T muons [19].

Progressing through the Run 1 (2009–2013) and Run 2 (2015–2018) phases, the LHC has developed towards the higher collision energies and gained enough resilience to operate stably in the 13-14 TeV regime. After the second long shutdown (2018–2022) with the main focus on improving the luminosity, today, we are in the Run 3 phase, which began in July 2022 and is an intermediate step before the High Luminosity LHC, which is planned for 2029-2041. With the precision measurements of the Standard Model particles being of major interest [21] for the
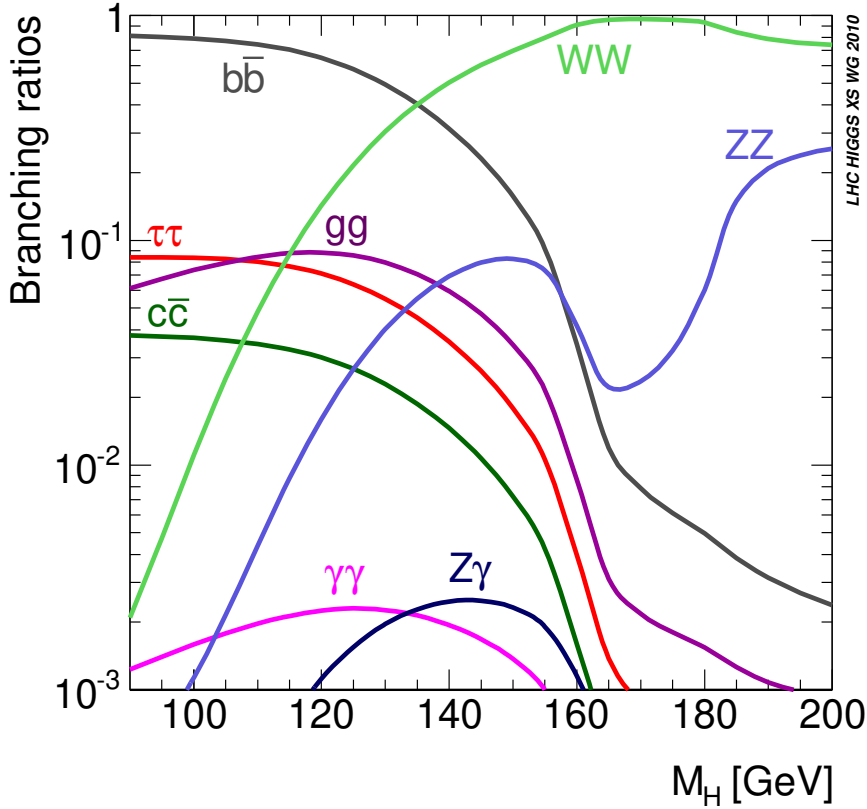
Figure 1.3: Higgs decay branching ratios as a function of its mass. (Taken from Ref. [20].)

Run 3, a special attention is dedicated to the precision measurements of the Higgs boson interactions, a potential proxy to the new physics.

For the analysis part, as the most sensitive, at least, for the Higgs searches, the observable based on the invariant mass of the decay products was used. Such searches, also known as *resonance searches*, are still relevant in the context of the Run 3 and the High Luminosity LHC, as we are looking for new particles that could explain such missing parts of the SM as Dark matter, Neutrino masses or Gravitational forces. In the next paragraph I will proceed by introducing resonance searches in more detail, explaining the arising challenges and how a Gaussian process can help tackle them.

Gaussian process approach to Look-elsewhere effect. In a resonance search, one typically scans a range of masses, aiming to find a significant peak at one or another mass. A quantitative measure of incompatibility between the observed data and the background model is typically determined by applying Wilks' theorem [22] to the likelihood-ratio test statistic (LRT) [23]. The test is conducted for a set of mass points in the search region with the resulting probability, in each point, usually being expressed in terms of local p-value (local p_0), as the probability for the background to give an excess larger than the one observed in the search data. Alternatively, one can associate the p-value to the number of standard deviations (Z) for a Gaussian distribution, and, when applied

to the search region, the procedure results in a field of varying significance levels across the region, shown in Figure 1.4.

Although the resonance could appear at any point in the search region, the search typically focuses on the area with the highest (local) significance. This situation presents a common challenge: accurately determining the global significance of the observation. It is crucial to take into account that random background fluctuations might produce a peak of significance anywhere in the search area, generating a "fake" point of highest significance. This concept is often referred to as the look-elsewhere effect (LEE). The probability for a background fluctuation anywhere in the search region to be larger than the largest excess observed in the data defines the global p-value (global p_0), and is the major target of the research presented in this thesis.

If every mass point in the search region were tested, the collection of measured significances would form a field of normally distributed random variables. Importantly, these variables are not independent; they are interconnected due to the non-zero width of the signal model and, as explored in Paper I, the rigidity of the background model. Similar to the multivariate Gaussian distribution, this group of variables, known as a *Gaussian process*, can be defined by specifying a mean for each point and a two-point correlation function, the Gaussian process kernel.

A number of studies have explored the look-elsewhere effect (LEE), and in my research, I specifically focus on those that model the significance field using a Gaussian process.

The direct method for estimating the trials factor involves analyzing a large number of Monte Carlo datasets — toy datasets created from the background model, each simulating one experimental measurement — and counting how often the significance curve surpasses a certain local significance level. Even the analysis of these Monte Carlo toy simulations is often computationally intensive, which is why the toys are commonly referred to as "brute force toys".

Typically, analyses set only an upper limit on the trials factor (TF) [24–26], by converting a local p-value into a global one, employing a Gaussian process indirectly to connect areas of low and high significance.

Gross and Vitells explored the look-elsewhere effect in their research [24, 25]. They proposed a method to approximate the upper bound on the trials factor, which also tends to approach the true value for local significances $\gtrsim 3\sigma$. Their technique involves setting this upper limit based on the average Euler characteristic of the χ^2 field, which arises from scanning the signal nuisance parameter. In one dimensional cases, this average Euler characteristic can be roughly calculated using the mean count of so-called up-crossings (Figure 1.4) at a certain local significance level and then extrapolated to higher levels. However, this approach tends to be overly conservative for lower significances, and also requires a relatively large amount of toys to get a precise estimate for the average number of up-crossings.

In Paper I, it was proposed that a small set of purposefully crafted background samples could be used efficiently for more precise estimations of the trials factor, and also for better predictions of low-significance up-crossings. Notably, the described method required explicit values for the parameters of the Gaussian
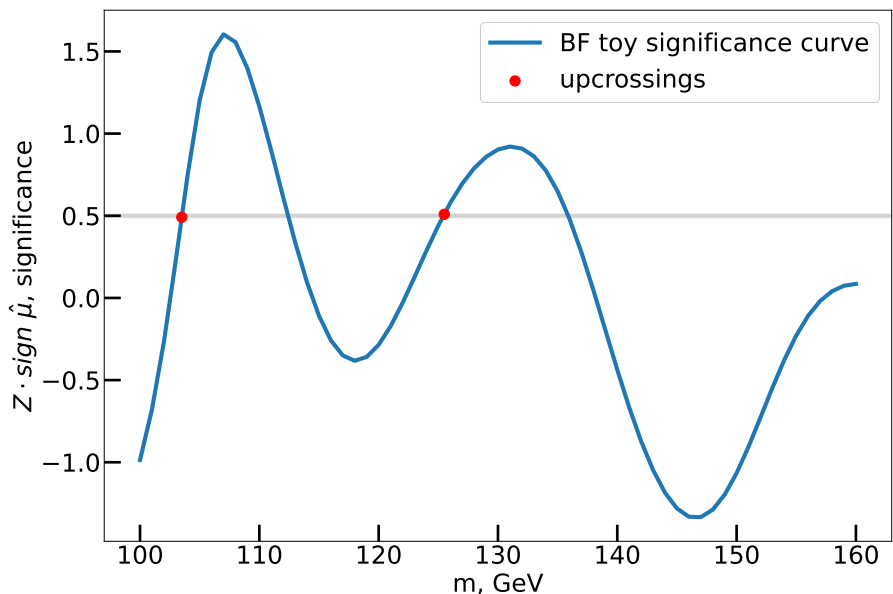
Figure 1.4: An example of a significance field (blue curve) for a likelihood ratio scan of the invariant mass of the Higgs boson in a study of brute force (BF) toys. Red dots denote locations of the up-crossings of threshold 0.5 by this significance curve. (Taken from Paper I.)

process, unlike the well-established implicit approach. The latter utilizes the up-crossings count at some reference level to compensate for the lack of information about the covariance and, by doing this, it sacrifices the precision of the trials factor estimate in exchange for avoiding the need to estimate the full covariance.

In works presented in this thesis the focus is on models where the test statistic is based on a χ^2 distribution with 1 degree of freedom, a fairly common case in High Energy Physics. Models that incorporate the signal through an extra signal strength parameter usually fit this requirement. This lets the χ^2 field to be changed into a field of significances (Figure 1.4). In Paper I, a Gaussian Process (GP) was suggested to be used for approximation of this field. Similarly to how a single Asimov dataset [23] is usually used to estimate expected significance, a special set of Asimov-like background samples (Section 2) was introduced to calculate a covariance matrix that approximates the GP kernel.

This GP facilitates making approximate Monte Carlo simulations (called *GP toys*), which represent the approximate significance[2] levels of the fitted Monte Carlo background samples straightforwardly, skipping the usual fitting step. This method is a fast way to estimate the trials factor using an approximate light-weight brute force approach.

It is important to note that the concept of applying the GP to hypothesis testing in scenarios where a nuisance parameter exists only under the alternative hypothesis, is not new and has been previously explored by Davies [26]. However, the contribution elaborated in Paper I, is an enhancement of this concept, in which a practical method for estimating the GP kernel was proposed.

---

[2]Later in the text we do not specify again that the constructed toy significances are approximate, as, we assume, it is implied that any use of Monte Carlo toys is an approximation.

Paper II presents a model which linearly approximates the fluctuations around the Asimov dataset. As also detailed in Section 2 of this thesis, this procedure enables an analytical approach to optimizing the log-likelihood, resulting in a closed form expression for significances. It, consequently, allows for analytical calculation of the significance covariance matrix, an integral part of the proposed GP-based method.

This linear model is beneficial as it eliminates the need for complex "toy fits" previously necessary in Paper I, replacing them with the requirement for calculations of derivatives from the best-fit background model, which are often available through fitting software.

Importantly, the foundational assumptions in Paper II are in alignment with those in the empirical study of Paper I. Consequently, Paper II not only lays the theoretical basis for the applications of set of Asimov background samples introduced in Paper I, but also benefits from its extensive validations.

Scientific software. It may already be evident that in this research field, up-crossings, the Euler characteristic, and Gaussian processes are frequently utilized. When combined, they offer numerous advantageous properties for estimating trials factors. This is how SigCorr (Paper III), a pioneering project that integrates all these elements into a single Python package, was born. This package was instrumental in generating, fitting, and visualizing the datasets used for the studies contributing to this thesis.

SigCorr is equipped with tools that enable parallel analysis of both MC toys and GP toys in a cohesive manner. In conjunction with the fitting framework, it facilitates the path from statistical model definition to estimating a trials factor. Moreover, SigCorr incorporates validated approximations and asymptotics, derived from seminal papers in the field, for rapid computation of the trials factor and its upper limit.

The modular yet coherent nature of SigCorr sets it apart in this domain, offering users the flexibility to experiment with various methods, conduct cross-validations, and select or develop the most appropriate approach for their analysis.

Furthermore, the data and Jupyter notebooks [27] employed for creating figures in Paper I are designed to be compatible with the SigCorr package, facilitating their reuse in future studies. For instance, this same data was utilized to validate the findings of Paper II.

Apart from SigCorr, more open-source contributions are presented in Chapter B. The list of contributions includes improvements of already existing software, like Monte Carlo generator Sherpa [28] and Boost C++ Library [29], as well as new projects: Cicliminds [30], a climate visualization toolkit developed in collaboration with the Center for International Climate Research CICERO, and MGVI.jl [31], which implements an optimization method for Bayesian statistical analysis and was developed in collaboration with Max Planck Institute for Physics.

Chapter 2

Set of Asimov background samples

In resonance searches, typically, there is a unique nuisance parameter M in the signal model, which is absent in the background-only hypothesis. Values of this parameter, M, represent the points that define the search region. The dimensionality of the search region can vary: it might be one-dimensional, as in the case where the mass of a hypothetical particle is unknown, or it can extend to multiple dimensions. For instance, a two-dimensional search region is considered when, also, the decay width is not specified uniquely. The details of a two-dimensional search region are discussed in Section 3.3 of Paper I. Additionally, the fundamental structure of a statistical model, typically employed in high-energy physics experiments for new particle or resonance discovery, is thoroughly explained in Paper I.

First, the proposed method using the model described by Gross and Vitells in their study of trials factor and up-crossings [24] was successfully verified. Subsequently, to showcase the efficacy of the set of Asimov background samples, a more complex example, drawing inspiration from the Higgs searches in the $H \to \gamma\gamma$ channel, was created.

In this $H \to \gamma\gamma$-inspired binned model, where i denotes the bins and $\boldsymbol{\theta}$ denotes the vector of nuisance parameters, the expected background $b_i(\boldsymbol{\theta})$ has an exponential shape and serves as the null hypothesis H_0. Both the normalization and width parameters of the exponential function are treated as nuisance parameters. The expected signal $\mu s_i(\boldsymbol{\theta})$ has a Gaussian form, and in conjunction with the background b_i, it constitutes the alternative hypothesis H_1, an estimate of the expected signal + background:

$$n_i(\mu, \boldsymbol{\theta}, M) = \mu s_i(\boldsymbol{\theta}, M) + b_i(\boldsymbol{\theta}), \tag{2.1}$$

where μ is the signal strength parameter.

The search utilizes the likelihood ratio test statistic, as detailed in [22], which is calculated for each point M within the search grid $\mathcal{M}$.

To illustrate the model inspired by $H \to \gamma\gamma$, Figure 2.5 displays the predicted background-only hypothesis b_i, a dataset sampled from it, and the resulting significance curve obtained from applying the LRT at each point in the search grid $\mathcal{M}$. This figure shows how noticeable fluctuations in the data correspond to peaks in the significance curve. The significance derived from the LRT appears

negative in places where there is a deficit in the data compared to the predicted background, and positive when there are data excesses.
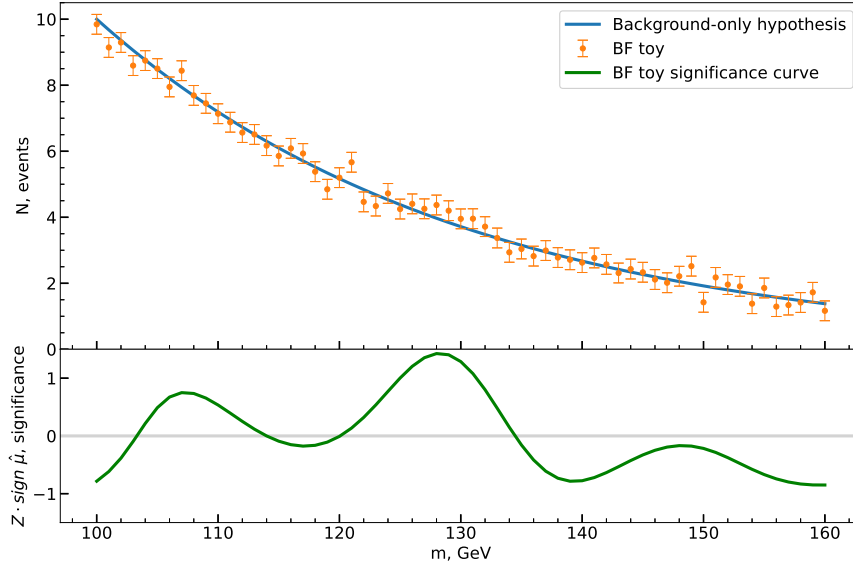

Figure 2.5: One brute force toy (orange) sampled from a background-only hypothesis (blue) of the $H \to \gamma\gamma$ inspired model. The signal significance curve is plotted below (green), where each point of the curve corresponds to a different choice of the signal hypothesis. (Taken from Paper I.)

Set of Asimov background samples. Paper I proposes to approximate the ensemble of curves $Z(M)$ with a family of samples from some Gaussian Process. Such an approximation increases the performance of the sampling without introducing significant bias. Properties of Z-curves (significance curves) are reflected in the GP mean, that should be set to 0, and the kernel, which must have 1 on the diagonal.[1]

Similarly to the concept of the representative Asimov dataset that may be used to estimate the expected local significance [23], we propose to use a special set of background Asimov datasets to estimate the GP covariance between the points of the significance curves (see Figure 2.5), and subsequently estimate the global significance based on a few generic *assumptions for the set of background Asimov datasets*:

1. the background model should be well approximated by its linear expansion around the best fit parameters,

2. the data can be binned and fluctuations in different bins of the dataset are independent,

---

[1]The diagonal elements, in the limit of infinite statistics, are expected to converge to 1.0, because the test statistic for each individual mass hypothesis should be distributed as χ^2 with 1 degree of freedom, which consequently makes the significance normally distributed with unit variance.

3. the fluctuations in each bin follow a Gaussian distribution.

(a)

(b)
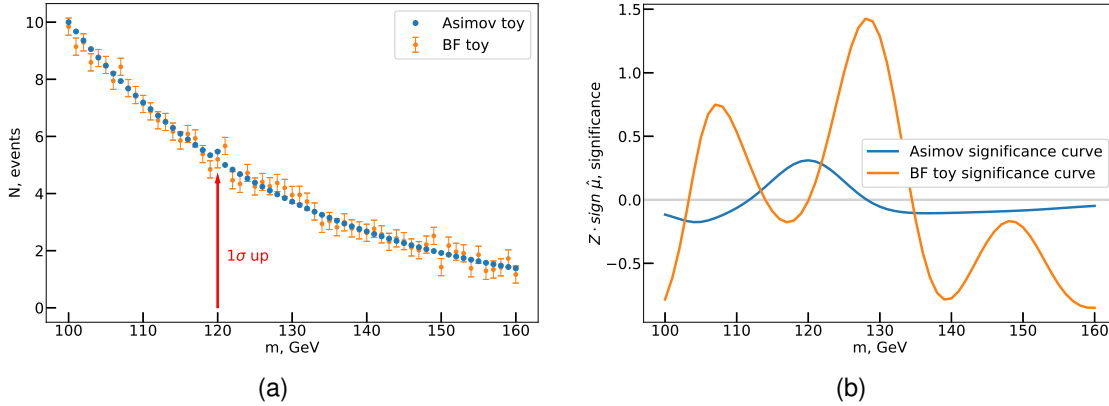
Figure 2.6: A background sample, an Asimov toy and the corresponding significance curves from the $H \to \gamma\gamma$ inspired model. In (a), as an example, a brute force toy sampled from the background template (orange dots) and one data set from the set of Asimov data sets (blue dots) are shown. The red arrow specifies the bin that was distorted in this particular Asimov data set. In (b) the significance curves from the signal plus background fits for the toy sample (orange line) and the Asimov data set (blue line) are shown. (Taken from Paper I.)

When the requirements are met, the *set of Asimov datasets*, according to Paper I, is constructed as follows:

1. Specify the binning of the data (for example invariant mass m) and the grid of scan points for the signal hypothesis (i.e. values for M).

2. Construct the Asimov dataset for the background model.

3. For each chosen data bin produce a new dataset, where a 1σ upward fluctuation[2] is introduced in that bin of the background Asimov dataset (see an example of such a dataset in Figure 2.6a).

Figure 2.6b shows a significance curve for one dataset of the set of Asimov datasets. Interestingly, although an excess of data was introduced in the middle of the mass range, the significance undershoots zero on the sides of the search region, which later results in anti-correlations visible in the significance covariance matrix.

The covariance matrix of the GP can now be determined by calculating the sample covariance using the assembled set of Asimov samples. However, it is important to note that this dataset incorporates only upward fluctuations, based on the assumption that errors will become symmetric in the asymptotic regime. As a result, the calculation of the sample covariance must be approached with caution.

---

[2]The precise magnitude and direction of the introduced fluctuation are not important and are accounted for by additionally normalizing the estimated covariance, as part of the method described in Paper I.

It is worth mentioning that the approximate linearity assumption was not foreseen at the time of writing Paper I, but in the following work (Paper II) it was identified as required.

If the assumption holds, the relation between significance and signal strength becomes linear, resulting in the covariance contribution of each dataset from the set of Asimov datasets to become independent of the sign and magnitude of the artificial 1σ fluctuation. It also allows the partial contributions of each Asimov sample to be assembled into a single estimate for the GP covariance, becoming a basis for the superposition principle assumed in Paper I.

Linear approximation to the covariance matrix. In Paper II, a series of simple steps were developed with a goal to derive the trials factor from a linear approximation of a non-linear parametric statistical model by estimating the covariance matrix of the significance field. Notably, the process only involves a single background-only fit to the data. Demortier briefly touched upon a method for calculating the covariance matrix from a linear model in his work [32, p. 23-33]. The contribution of this thesis extends Demortier's method by providing a rigorous mathematical framework for the method and showcasing its application to non-linear background shapes. In Paper II, the estimated covariance matrix acts as a proxy for the calculation of the trials factor estimate, either via the analytic calculation of average up-crossings or using the GP toys.

This method was studied in application to the same statistical model structure used to introduce the set of Asimov samples (see Equation 2.1). Also, it was assumed that the best fit of the background model b_i to the data d_i, represented as $b_i(\hat{\boldsymbol{\theta}}) = \hat{b}_i$, was known for the analysis. To simplify the explanation, the reference point for the model parameters $\boldsymbol{\theta}$ was chosen to be $\hat{\boldsymbol{\theta}} = \mathbf{0}$.

Following the approach in Paper II, it became possible to calculate the autocovariance matrix Σ_{MN} of the significance field between two signal mass hypotheses M and N. This calculation used the signal model s_M and derivatives of the background model $\dfrac{\partial b}{\partial \theta}$:

$$\Sigma_{MN} = \frac{\langle s_M|}{\sqrt{\langle s_M|\,P\,|s_M\rangle}} P \frac{|s_N\rangle}{\sqrt{\langle s_N|\,P\,|s_N\rangle}}, \quad M, N \in \mathcal{M}$$

$$P = \mathbb{1} - \Delta(\Delta^\intercal \Delta)^{-1}\Delta^\intercal, \tag{2.2}$$

$$\Delta_{i\alpha} = \frac{\partial b_i(\boldsymbol{\theta})}{\partial \theta^\alpha}\Big|_{\boldsymbol{\theta}=\mathbf{0}},$$

where bra-ket notation was used to reduce the number of indices. The elements of the ket-vector $|s_M\rangle$ are associated with different dataset bins, and the index M marks the signal locations on the search grid $\mathcal{M}$, each representing a unique signal hypothesis.

This result makes the direct method for calculating up-crossings, as well as the trials factor, detailed in Chapter 2.1 of Paper I, even more practical, and helps avoid the need for sampling to estimate the average number of up-crossings. Although, the calculation does require knowing the covariance matrix, it is now easier to calculate with this new linear approximation method.

When presenting the method of Asimov datasets in application to estimation of the trials factor [33], along with the excitement of the audience, numerous questions suggesting to go beyond binned models were brought up. As a first step towards unbinned analysis, in Appendix A, the steps from Paper II were repeated, but the starting point was set to a more general model. The calculation resulted in a closed form expression for the significance for a general quadratic log-likelihood (Equation A.8), with the Equation 2.17 from Paper II being its partial case.

Taking this farther towards closed form covariance matrix estimates will open opportunities for analyses with low bin counts that are usually not suitable for a binned approach.

Chapter 3

Conclusion

This doctoral thesis has addressed several crucial aspects of statistical methods in the field of high-energy physics and beyond, particularly focusing on searches for new resonances involving the precise estimation of the trials factor, a task for which existing methods have a tendency to become excessively conservative, especially when dealing with lower local significance regions. To address this issue, in Paper I, a novel approach based on Gaussian processes was introduced.

The newly proposed method does not entirely eliminate the need for Monte Carlo simulations but significantly enhances the computational efficiency of the process. Extensive testing of the suggested method was conducted on various models, even those with multiple nuisance parameters, designed intentionally to exhibit diverse statistical characteristics. The results demonstrated remarkable agreement between the trials factors calculated using our new approach and those derived from large samples of brute force Monte Carlo simulations. As part of a comprehensive cross-check, the results from the Gross and Vitells paper [24], which introduced the high-significance approximation based on up-crossings, were successfully reproduced .[1]

The practical realization of the method and the implementation of test models suggested in Paper I have been made available for the research community through our published work (Paper III) and the SigCorr package itself, shared on GitLab [34]. The extensive documentation provided with SigCorr is intended to lower the threshold for the use and further developments of the package. For example, one can consider the possible integration of SigCorr into the ScikitHEP [35] environment with particular focus on pyhf [36] and HepStats [37] packages.

Another major contribution of this thesis is the development of an efficient method for estimating the covariance matrix of statistical significance in new particle searches (Paper II). This method relies on a linear expansion of the statistical model around its background-only best fit to the data. We not only derived a closed-form expression for the linear approximation of the significance covariance matrix but also provided elegant expressions for the best-fitted signal strength and statistical significance within this approximation. To judge the

---

[1]The data sets and the Jupyter notebooks that have been used for the cross-checks are available on Zenodo [27].

validity of the linear approximation for the search in question, one can compare our approximate results to the results of a full fit to the data.

Crucially, it was demonstrated that the suggested covariance matrix adheres to the superposition principle with regard to the data fluctuations, making it a suitable proxy for the covariance matrix constructed using the set of Asimov background samples. Comparing these two approaches, particularly in the context of a Higgs boson decay model, the differences between the values of the covariance matrices constructed using the two methods were found to be small. Consequently, all validations conducted as part of the empirical study, i.e. trials factor estimation with a set of Asimov background samples, can be applicable to the linear approximation approach, which itself may serve as a theoretical foundation for the empirical method of the set of Asimov background samples.

In addition to the above, the research results of this thesis also delivered a practical expression approximating the mode of the non-central χ^2 distribution in Paper IV. This expression, based on an asymptotic expansion, has a practical utility as an initial guess for iterative procedures approximating the mode of the non-central χ^2 distribution numerically. As a result, the Boost.Math C++ library received a contribution for this small improvement [38]

Beyond the primary research in statistical methods for high-energy physics, through collaborative work contributions to the advancement of research tools across diverse technologies and scientific domains were published. The list includes such famous software as Sherpa for which support for custom form factors for models defined with UFO format [39] was enabled, and no less known Statsmodels, which now allows omnibus survival tests for data with any number of groups [40].

The close collaboration with CICERO resulted in the development of Cicliminds [30], a versatile tool for comprehensive climate data analysis, and Max-Planck Institute for Physics, together with which MGVI.jl [31], a Julia package offering efficient estimation for a posterior distribution, was introduced.

In summary, this doctoral thesis has introduced innovative statistical methods, addressed challenging issues in high-energy physics, and made contributions to the field's ongoing research efforts that fit precisely the mission statement of the INSIGHTS ITN [1]. These findings have the potential to significantly enhance the precision and efficiency of new particle searches and statistical analysis in physics and beyond.

Papers

Paper I

Gaussian Process-based calculation of look-elsewhere trials factor

PAPER • OPEN ACCESS

Gaussian Process-based calculation of look-elsewhere trials factor

To cite this article: V. Ananiev and A.L. Read 2023 *JINST* **18** P05041

View the article online for updates and enhancements.

Gaussian Process-based calculation of look-elsewhere trials factor

V. Ananiev and A.L. Read[*]

Department of Physics, University of Oslo,
Boks 1072 Blindern, Oslo, NO-0316, Norway

E-mail: a.l.read@fys.uio.no

Abstract: In high-energy physics it is a recurring challenge to efficiently and precisely (enough) calculate the global significance of, e.g., a potential new resonance. We propose a new method that models the significance in the search region as a Gaussian Process. The kernel of the Gaussian Process is approximated with a covariance matrix and is calculated with a carefully designed set of background-only data sets, comparable in number to the random background-only data sets used in a typical analysis that relies on the average upcrossings of the significance. The trials factor for both low and moderate significances can subsequently be calculated to the desired precision with a computationally inexpensive random sampling of the Gaussian Process. In addition, once the covariance of the Gaussian Process is determined, the average number of upcrossings can be computed analytically. In our work we give some highlights of the analytic calculation and also discuss some peculiarities of the trials factor estimation on a finite grid. We illustrate the method with studies of three complementary statistical models.

---

[*]Corresponding author.

Contents

1 Introduction

In a typical search for a new particle or resonance in a high-energy physics experiment, the statistical model consists of a relatively narrow signal peak on top of a broad background in the invariant mass spectrum. In general, the location of the signal is not known a priori so the likelihood is maximized for signal masses and amplitudes in the "search region" of mass. In some models the width of the signal is not uniquely specified, leading to a 2-dimensional search region. The background model has typically one or more nuisance parameters that describe the form and normalization of the background distribution. The signal parameters such as width and mass, however, are not defined under the background hypothesis. In this case, the significance at the location of maximum likelihood of the signal plus background hypothesis (local significance) should be reduced to account for random fluctuations of the background that can occur anywhere in the search region. The corrected significance is called "global significance", and the ratio between the corresponding global and local p-values or tail probabilities, which is a function of the observed local significance, is called the "trials factor" (TF). The whole situation is often referred to in high energy physics as the "look-elsewhere effect".

The most straightforward way to estimate the trials factor is to analyse a large ensemble of Monte Carlo data sets (toy data sets or "toys") generated from the background model and to count how many times the significance exceeds a local significance level. However, Monte Carlo toys are usually expensive to analyse in terms of computational resources.

Gross and Vitells, in their studies of the look-elsewhere effect [1, 2], suggested a way to estimate the upper bound on the trials factor, which for large local significances $\gtrsim 3\sigma$ approaches the true value. Their method sets the upper bound to the trials factor based on the average Euler characteristic

of the χ^2 field, a field that emerges from the signal nuisance parameter scan. In the one-dimensional case the average Euler characteristic can be estimated from the average number of upcrossings (figure 1) counted at some reference level of local significance and then propagated to higher levels. This method, however, can be very conservative when applied to lower significances. In this work, we argue that a relatively small number of carefully designed background toys can be used efficiently to provide more precise and less conservative estimates for the trials factor, as well as more precise estimates for low-significance upcrossings.

We focus on models with a test statistic that follows a χ^2 distribution with 1 degree of freedom. This allows us to turn the χ^2 field into a significance field (figure 1) that we then propose in section 2 to approximate with a Gaussian Process (GP) whose kernel is approximated with a covariance matrix calculated with the carefully designed toys. The resulting GP allows us to efficiently sample approximate Monte Carlo toys and, therefore, to quickly evaluate the trials factor via an approximate brute force method.

It is worth mentioning that the idea of using the GP in relation to hypothesis testing when a nuisance parameter is present only under the alternative is not new and was studied by Davies [3]. We, however, build on top of this idea with a suggestion of a practical way to estimate the GP kernel.

As described in section 2.1, once the covariance of the GP is determined, the average number of upcrossings can, in fact, be computed analytically. We also discuss some peculiarities of the trials factor estimation on a finite grid in section 2.2.

We demonstrate our method applied to the background template model used by Gross and Vitells (GV) in their study (section 3.1) as well as exponential background models with 1-dimensional (section 3.2) and 2-dimensional (section 3.3) search regions inspired by searches for a resonance due to $H \rightarrow \gamma\gamma$ decays in proton-proton collisions at the Large Hadron Collider [4, 5].
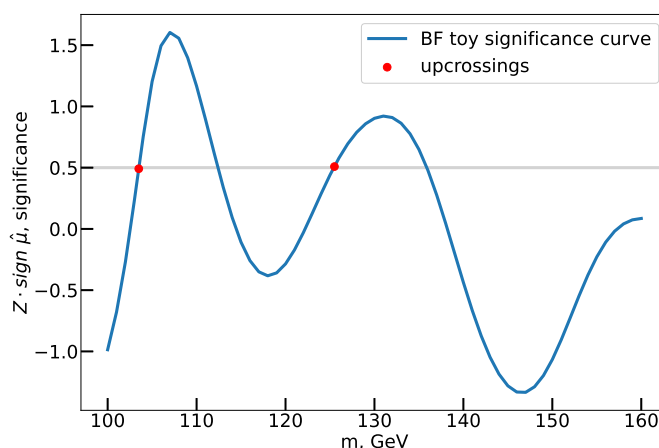
Figure 1. An example of a significance field (blue curve) for a likelihood ratio scan of the invariant mass of the Higgs boson in a study of brute force (BF) toys. Red dots denote locations of the upcrossings of threshold 0.5 by this significance curve.

The Jupyter notebook and the data that were used to produce figures for this paper are publicly available at Zenodo [6]. The Python package that we developed and used to produce and fit the data sets, which is also extensively used in the notebook for visualizations, is also publicly available [7] and includes the documentation [8] with recipes of how to reproduce the results we present in this work.

2 Method

In resonance searches the signal model has a nuisance parameter M that is not present under the background-only hypothesis and that denotes the set of points that constitute the search region. The search region itself may be one-dimensional, for a case when the mass of a hypothetical particle is unknown, or many dimensional, for example 2-dimensional, when the decay width is also not specified uniquely. We elaborate on the 2-D search region in section 3.3.

We focus on the case where the signal model has 1 free parameter (e.g. signal amplitude or strength μ, that we assume is fitted unconstrained[1]) for each point in M, therefore the profile likelihood test statistic

$$t(M) = -2 \ln \lambda(M), \ \lambda(M) = \frac{\mathcal{L}(0, \hat{\hat{\theta}})}{\mathcal{L}(\hat{\mu}, M, \hat{\theta})} \tag{2.1}$$

follows a χ^2 distribution with 1 degree of freedom in the large sample limit [9]. Here θ is one or more additional nuisance parameters that maximize the likelihood under the background ($\hat{\hat{\theta}}$) or signal plus background ($\hat{\theta}$) hypotheses.

We want to construct a standard normal random variable from $t(M)$. For large samples:

$$\sqrt{t(M)} \sim \sqrt{\chi_1^2} = |N(0, 1)| . \tag{2.2}$$

To restore the sign of the standard normal distribution we use the sign of the signal strength $\hat{\mu}$, which is known to be normally distributed around 0 in the absence of signal, in the large sample limit as a maximum likelihood estimator.

We then construct the normally distributed significance $Z(M)$ as follows:

$$Z(M) = \sqrt{t(M)} \operatorname{sign} \hat{\mu}. \tag{2.3}$$

To compute the trials factor we need to determine, under the background hypothesis $H_0 : \mu = 0$, the probability that $Z(M)$ exceeds the local significance observed in the data for any M in the search region.

We propose to approximate the ensemble of curves $Z(M)$ with a family of samples from some Gaussian Process. Such an approximation increases the performance of the sampling without introducing significant bias. Properties of Z-curves (significance curves) are reflected in the GP mean, that should be set to 0, and the kernel K with $K(M_i, M_i) = 1$, i.e., the kernel must have 1 on the diagonal.

The crucial step of our proposal is to estimate $\hat{\Sigma}$, which is the covariance matrix, an approximation of the GP kernel evaluated on the grid specific to the problem. Properties of the kernel K, and subsequently of the covariance matrix $\hat{\Sigma}$, can be formally expressed as follows:

$$\hat{\Sigma}_{ij} = \operatorname{cov} \left[Z(M_i), Z(M_j) \right], \tag{2.4}$$
$$\hat{\Sigma}_{ii} = 1.$$

---

[1]We need both positive and negative values of the best-fit signal strength $\hat{\mu}$ to be able to reconstruct the sign of the significance and, therefore, both sides of its distribution. This is a crucial step required to treat the significance as a GP, which is at the core of the method suggested in this paper. We, however, argue that unconstrained fits generally are not an issue for analyses of large samples, while analyses of small samples can benefit from brute force calculations of the TF with no significant penalty in performance.

The significance Z evaluated on the grid becomes a multivariate Gaussian random vector $\hat{Z}$ with covariance $\hat{\Sigma}$:

$$\hat{Z} \sim \mathcal{N}(\vec{0}, \hat{\Sigma}). \tag{2.5}$$

Similarly to the concept of the Asimov data set that may be used to estimate the expected local significance [10], we propose to use a special set of background Asimov data sets to estimate the GP covariance, and subsequently estimate the global significance.

The *set of Asimov data sets* is constructed as follows:

1. Specify the binning of the data (for example invariant mass m) and the grid of scan points for the signal hypothesis (i.e. values for M).

2. Construct the Asimov data set for the background model.

3. For each chosen data bin produce a new data set, where a 1σ upward fluctuation is introduced in that bin of the background Asimov data set (see section 3 for concrete examples).

We assume that the data in different bins are uncorrelated.[2] To construct the approximate covariance, using eq. (2.4), we also use the fact that the covariance matrix of a sample of independent measurements can be decomposed into a sum of the partial covariances, in our case calculated from the set of Asimov data sets:

$$\hat{C}_{ij}^a = Z^a(M_i) \cdot Z^a(M_j), \tag{2.6}$$

$$\hat{C} = \sum_a \hat{C}^a, \tag{2.7}$$

where index a enumerates Asimov data sets and $\hat{C}$ is a sum of partial covariances. Note, however, the expression for the partial covariance $\hat{C}^a$ does not include the sample mean. It was intentionally set to 0, because we know the mean of the significance should be 0. We also know that variance of the significances is 1, so the final touch is to impose 1 on the diagonal of $\hat{\Sigma}$ by rescaling the $\hat{C}$ as follows:

$$\hat{\Sigma}_{ij} = \frac{\hat{C}_{ij}}{\sqrt{\hat{C}_{ii}\hat{C}_{jj}}}. \tag{2.8}$$

By normalizing this way, we again use the assumption that data in the bins are uncorrelated and we also assume that the covariance structure is approximately independent of the scale of a fluctuation.

2.1 Number of upcrossings directly from the covariance matrix

Given a differentiable kernel for a Gaussian Process, it is possible to calculate the average number of upcrossings at any level without recourse to any random sampling [11, 12]. This represents a small advancement with respect to the Gross and Vitells extrapolation method.

To motivate the formula in eq. (2.14) for the density of upcrossings that we used directly in the form presented by Lutes et al. [11, Example 7.2], we first cite the generic expression for the density

---

[2]There are situations when this assumption is not valid. This happens, for example, in analyses that allow each event to be reconstructed in multiple ways, and these reconstructed configurations contribute to the distribution of the observable simultaneously. One can also think of events recorded very closely in time becoming correlated due to out of time pile-up, when the detector does not have time to relax to its initial state.

$\nu_X^+(u, t)$ of the average number of upcrossings of the level u by any stochastic process $X(t)$ [11, eq. (7.3)]:

$$\nu_X^+(u, t) = p_{X(t)}(u) \int_0^\infty \nu p_{\dot{X}(t)}[\nu \mid X(t) = u] \mathrm{d}\nu. \tag{2.9}$$

When $X(t)$ is a Gaussian process, $p_{X(t)}$ is the corresponding Gaussian density. The derivative $\dot{X}(t)$, when conditioned on $X(t) = u$, follows a Gaussian distribution with mean $\mu_\star$ and standard deviation $\sigma_\star$:

$$p_{\dot{X}(t)}[\nu \mid X(t) = u] = \mathcal{N}(\mu_\star(t), \sigma_\star(t))[\nu], \tag{2.10}$$

$$\mu_\star(t) = \frac{K_\bullet(t, t)}{K(t, t)} u = \rho_\bullet(t, t) \frac{\sigma_\bullet(t)}{\sigma(t)} u, \tag{2.11}$$

$$\sigma_\star(t) = \sigma_\bullet \sqrt{1 - \rho_\bullet^2(t, t)},$$

where $\sigma(t) = \sqrt{K(t, t)}$ is the standard deviation of the GP and $\mathcal{N}[\nu]$ is a Gaussian pdf evaluated at ν. The above expressions use a dot notation for derivatives along a single or both axes:

$$K_\bullet(x, y) = \frac{\partial K(x, y)}{\partial y}, \ K_{\bullet\bullet}(x, y) = \frac{\partial^2 K(x, y)}{\partial x \partial y}. \tag{2.12}$$

Consequently, the effective derivatives of the standard deviation and correlation are:

$$\sigma_\bullet(t) = \sqrt{K_{\bullet\bullet}(t, t)}, \ \rho_\bullet(x, y) = \frac{K_\bullet(x, y)}{\sigma(x)\sigma_\bullet(y)}. \tag{2.13}$$

The expression for the density $\nu_X^+(u, t)$ of the average number of upcrossings at the level u by the non-stationary GP with a covariance kernel K and zero mean was derived by Lutes et al. [11, Example 7.2] by integrating explicitly eq. (2.9) with the conditional Gaussian distribution $p_{\dot{X}}$ substituted from eq. (2.10) and is:

$$\nu_X^+(u, t) = \frac{\mathrm{e}^{\frac{-u^2}{2\sigma^2(t)}}}{\sqrt{2\pi}\sigma(t)} \left(\mu_\star(t) \Phi \left[\frac{\mu_\star(t)}{\sigma_\star(t)} \right] + \frac{\sigma_\star(t)}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\mu_\star^2(t)}{2\sigma_\star^2(t)}} \right), \tag{2.14}$$

where Φ is the cumulative density function (CDF) of the standard normal distribution.

The analytic results above are valid for a continuous differentiable GP kernel. To evaluate these expressions numerically it proved useful to apply 2D spline interpolation to our covariance matrices before computing the derivatives. In this way we managed to reduce the uncertainty introduced by the coarse grid without sacrificing performance much, because spline interpolation provides exact spline derivatives straightaway. We then used Simpson quadrature to integrate the computed density over the mass range.

## 2.2 Effect of coarse binning on upcrossings

The Gross and Vitells upper bound is based on the average number of upcrossings of some level u by the significance curve $Z(M)$. When computing this number numerically, from MC toys, some finite resolution grid is used to approximate the curve. It turns out that the resolution of the grid affects

the resulting number of upcrossings. In this section we investigate this effect, which is relevant for any Gross and Vitells procedure, whether it is based on brute force toys or on the suggested GP toys.

We illustrate the effect on an example of an analytically defined Gaussian process with a squared exponential kernel:

$$K(x, y) = e^{-\frac{(x-y)^2}{\alpha^2}}, \quad \alpha^2 = 10. \quad (2.15)$$

We generated 10^6 samples from this GP and counted the average number of upcrossings above the level $Z = 0.2$ on a set of grids with different resolutions. In figure 2 we show how the accuracy decreases with the resolution, starting when the bin size exceeds about one third of the correlation length.
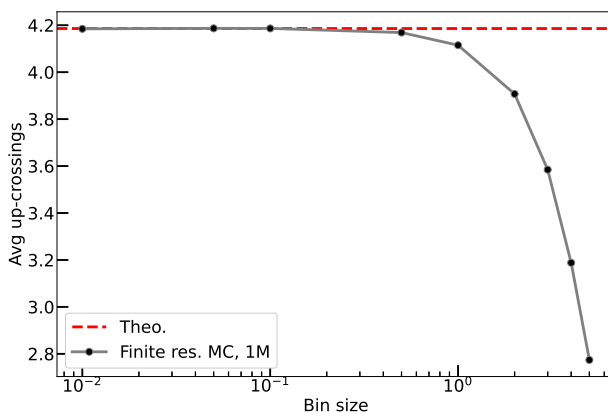
Figure 2. The average number of upcrossings as a function of bin size. The analytical estimation for the average number of upcrossings (red dashed line) was calculated following section 2.1 for a squared exponential kernel (eq. (2.15)). The average number of upcrossings for various bin sizes was estimated from 10^6 GP toys (black dots connected by a gray line). The statistical errors are smaller than the size of the dots.

The underestimation of the number of upcrossings affects the upper bound for the global significance, i.e. the numerator of the trials factor, subsequently the trials factor itself will be underestimated. The local significance, however, which is the denominator, is also affected by the finite grid resolution. The tip of the peak at high significance most probably will appear between grid points, therefore the local significance will also be underestimated.

We investigated the dependency of the effective peak width on its height. We sampled 10^6 toys from the GP (eq. (2.15)). For each sample we detected the highest point (x_{max}, y_{max}), chose a window of 40 around it to select one peak and fitted it with a squared exponential shape $y = A^2 e^{(x-B)^2/C^2} + D$. For each sample we recorded a pair (y_{max}, C), where C is the effective peak width. We then removed the failed fit outliers by rejecting the 1% lowest and 1% highest heights. We observed that the average width of the peak decreases with its height (figure 3), therefore, the finite grid effect on the trials factor gains a slight compensation.

With the Asimov set of samples we are not trying to improve this particular aspect of the Gross and Vitells approach, yet, GP toys are also affected by the choice of the grid so we wanted to shed more light on this caveat. Although we are not very concerned by the choice of a grid for the signal nuisance parameters, of course it makes good sense to choose a grid that allows peaks to be described by at least a handful of points.
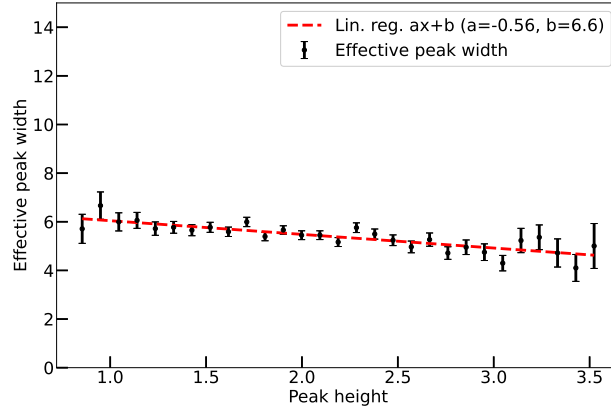
Figure 3. The average effective width of the highest peak of GP samples for a squared exponential kernel (eq. (2.15)) as a function of its height for 10^6 GP toys (black dots) fitted to a linear trend (red dashed line).

3 Examples

The models we use as examples are inspired by searches for new resonances. The invariant masses of signal candidates are binned. Two of the search models have signals with specified widths whereas the third includes the signal width as a free parameter (on a grid) in the likelihood maximization.

3.1 Background template model

Gross and Vitells used this model in their study [1]. It consists of a fixed background shape modeled by the probability density function of the Rayleigh distribution with scale parameter σ_b, and a Gaussian signal with resolution $\sigma_s(M)$ that grows linearly with mass. Background counts D_i are sampled from the Poisson distribution with the rate $b_0 \cdot b_i$ in each bin i of the data grid m_i defined as follows:[3]

$$b_i = N \frac{m_i}{\sigma_b} e^{-\frac{m_i^2}{2\sigma_b^2}},$$

$$D_i \sim \text{Poisson}[b_0 \cdot b_i],$$
(3.1)

where N was chosen such that $\sum b_i = 1$ and, therefore, b_0 is the total number of events we expect to observe.

To test the precision of our GP-based approximation of the trials factor we compute the baseline via 10^6 brute force MC simulations of the experiment. We scan each background simulation by testing signal hypotheses for M in a range that is narrower, in this example, than the data grid m:

$$s_i(M) = \frac{1}{\sqrt{2\pi}\sigma_s(M)} e^{-\frac{(m_i - M)^2}{2\sigma_s^2(M)}},$$

$$\sigma_s(M) = A \left(1 + \frac{M}{B}\right),$$
(3.2)

---

[3]$\sigma_b = 40\,\text{GeV}$, $b_0 = 2000$, $A = 2.5\,\text{GeV}$, $B = 50\,\text{GeV}$, $m_i = 0.5 - 154.5\,\text{GeV}$ with a step of 1 GeV, $M_i = 5 - 120\,\text{GeV}$ with a step of 0.25 GeV.

and maximize the likelihood with Poisson statistics in the bins of data:

$$N_i = \mu s_i(M) + \beta b_i,$$

$$-\log \mathcal{L}(\mu, M, \beta) = -\sum_i D_i \log(N_i) + \sum N_i,$$

(3.3)

where μ and β correspond to the numbers of inferred signal and background events that we vary to maximize the likelihood $\mathcal{L}$ with respect to data D_i, and M is the nuisance parameter that defines the location of the hypothetical signal peak (the constant term of $\log \mathcal{L}$, which plays no role in the maximization, is not shown).

An example of a background simulation and the corresponding significance curve are shown in figure 4. As a cross-check we compute the average number of upcrossings for the significance threshold of $\sqrt{0.5}$, the same threshold Gross and Vitells used in their work. We get 4.3071 ± 0.0016 upcrossings, which is consistent with their result of 4.34 ± 0.11.

Next, we generate the set of Asimov background data sets. In this case there are 155 of them, one for each bin of data. One of the Asimov data sets and the corresponding significance curve are shown in figure 4 for comparison.


| (a) | (b) |
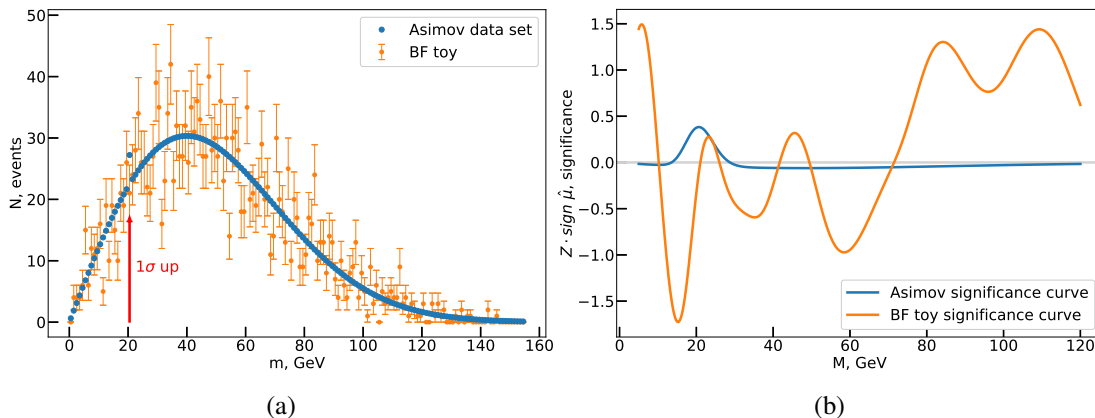
Figure 4. A background sample, an Asimov toy and the corresponding significance curves from the Background template model. In (a) we show an example of a brute force toy sampled from the background template (orange dots) and one data set from the set of Asimov data sets (blue dots). The red arrow specifies the bin that was distorted in this particular Asimov data set. In (b) we show the significance curves from the signal plus background fits for the toy sample (orange line) and the Asimov data set (blue line).

For each Asimov data set we compute the partial covariance (see figure 5(a)), then we sum them and normalize according to eq. (2.7) to produce the Asimov GP covariance, shown in figure 5(b). As expected, the width of the ridge on the diagonal of the covariance matrix is seen to have the same trend as the signal resolution that grows linearly with the mass hypothesis (eq. (3.2)).

Before using the covariance matrix to sample the significance curves, let us first evaluate how well we approximate the covariance matrix estimated from the brute force toys. For this we subtract the Asimov covariance, shown in figure 5(b), from the covariance matrix constructed from 10^6 brute force toys, shown in figure 6(a), and plot the difference between the two covariances in figure 6(b). The differences between Asimov and brute force covariance are 100 times smaller than the values
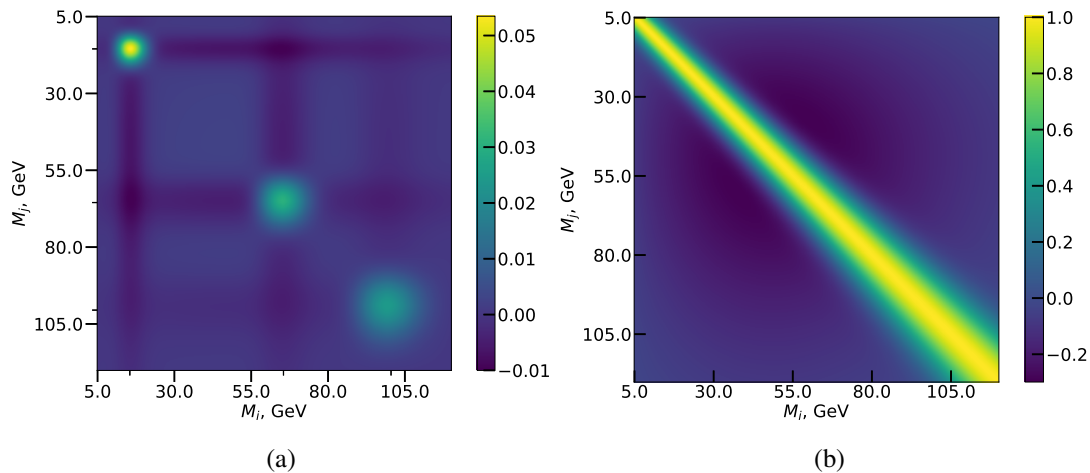
Figure 5. Partial contributions to the covariance $\hat{\Sigma}(M_i, M_j)$ (a) from 15.5, 65.5 and 100.5 GeV Asimov data sets and the full Asimov GP covariance (b) for the Background template model.
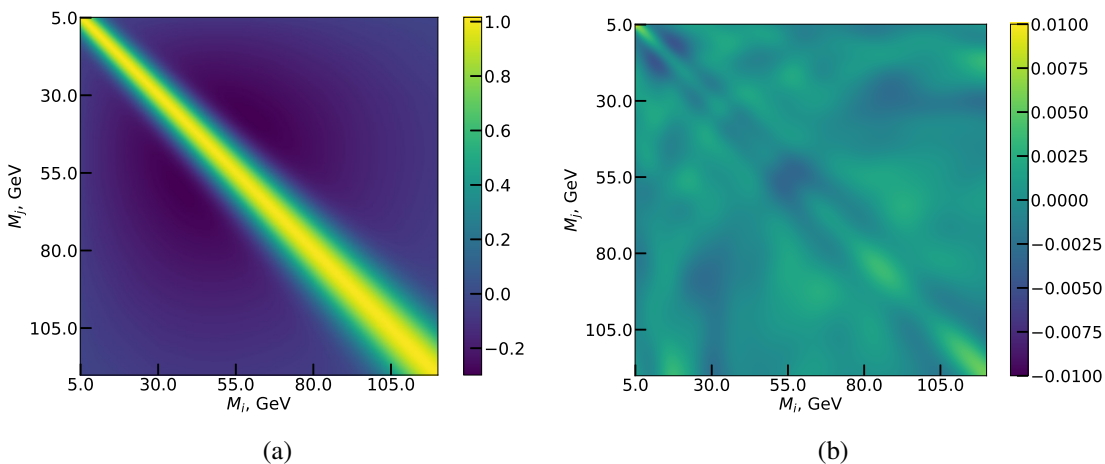

Figure 6. The difference (b) between the brute force (a) and Asimov (figure 5(b)) covariance matrices calculated for the Background template model.

in the covariance itself. We obtain similar patterns of differences for statistically independent simulations, so we conclude the pattern is systematic.

In our experience, the trials factor appears to be the most sensitive metric to evaluate approximations of the global significance. We estimate the trials factor from $360 \cdot 10^6$ significance curves sampled from the Asimov GP. In figure 7 we compare the Asimov GP trials factor, as a function of the local significance, to the brute force approach and approximations that include the Gross and Vitells upper bound and the rule of thumb estimate. The latter was calculated as a ratio of the mass range to the signal resolution, that we average, afterwards, over the mass range, since the resolution changes with the energy scale. We explore the sensitivity of the TF to the Asimov GP deviations by amplifying the difference between Asimov GP covariance and the brute force covariance by factors of 10 upwards and downwards. From this simple sensitivity analysis we see that visible deviations only appear in the very high significance region (here $\sim 6\sigma$), in which also the Asimov GP approach

is affected by coarse binning effects, and in which the Gross and Vitells upper bound serves as a good approximation to the true TF value. In the same figure we also show that the analytic method described in section 2.1 is in excellent agreement with the Gross and Vitells extrapolated upper bound.

We observe a significant deviation of the Asimov GP from the brute force trials factor estimate at high local significance levels. Qualitatively, we attribute this to the non-gaussianity of the errors introduced by Poissonian statistics that we approximated with a Gaussian distribution. Since the Poisson density is not symmetric around the mean, and is more pronounced for large upward fluctuations, the approximating Gaussian distribution gains upward bias. This effect is smaller at low local significances because the Poisson distribution can be well approximated by a quadratic shape around the mean. We confirmed this by studying the trials factors for larger integrated background rates and observing that the differences at high significance were reduced.
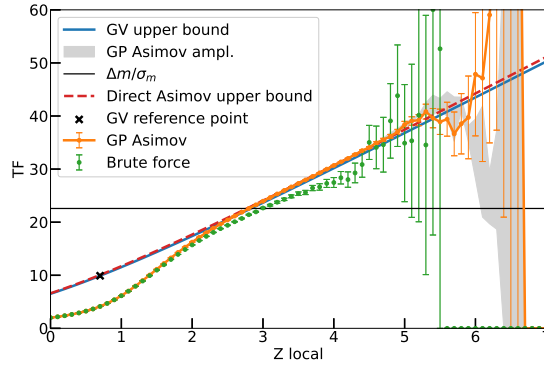
Figure 7. Trials Factor TF as a function of local significance Z_{local} for the Background template model. We estimated TF from 10^6 brute force toys (green dots) and compare it to the Asimov GP results (orange line). We color in gray the area enclosed between the two amplified lines from the TF sensitivity study for the Asimov covariance (section 3.1). The solid blue line shows the Gross and Vitells upper bound [1, eq. (3)] extrapolated from the reference significance threshold $\sqrt{0.5}$ (marked by the black cross on the plot). For comparison we show the semi-analytical estimate (section 2.1) of the Gross and Vitells upper bound (red dashed line) and the rule of thumb (black line).

3.2 $H \rightarrow \gamma\gamma$ inspired model

This model is inspired by the ATLAS and CMS searches for the Higgs boson decaying to 2 photons [4, 5]. In this case the background is modeled as an exponential distribution with Gaussian errors σ_b in each bin of data. In order to challenge our method with a more realistic background scenario, in addition to the normalization b_0, the rate parameter α_b is also unconstrained in the maximum likelihood fits:[4]

$$b_i(\alpha_b) = e^{-(m_i - 100)\alpha_b},$$
(3.4)
$$D_i \sim \mathcal{N}(b_0 \cdot b_i, \sigma_b).$$

We also simplify the signal model by setting its resolution σ_s to a constant:

$$s_i(M) = \frac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{(m_i - M)^2}{2\sigma_s^2}}.$$
(3.5)

---

[4]$\alpha_b = 0.033\,\text{GeV}^{-1}$, $b_0 = 10$, $\sigma_b = 0.3$, $\sigma_s = 5\,\text{GeV}$, $m_i = M_i = 100 - 160\,\text{GeV}$ with a step of $1\,\text{GeV}$.

We unify the grids and use the same bins for both sampling of the MC toys (m_i) and signal location scan (M_i). The structure of the likelihood function in this case is:

$$N_i = \mu s_i(M) + \beta b_i(\alpha_b),$$ (3.6)

$$-\log \mathcal{L}(\mu, M, \beta, \alpha_b) = \sum_i (N_i - D_i)^2.$$

Following the same algorithm as for the Gross and Vitells model, we first show an example of the brute force and Asimov background samples (figure 8(a)) together with the corresponding significance curves (figure 8(b)).
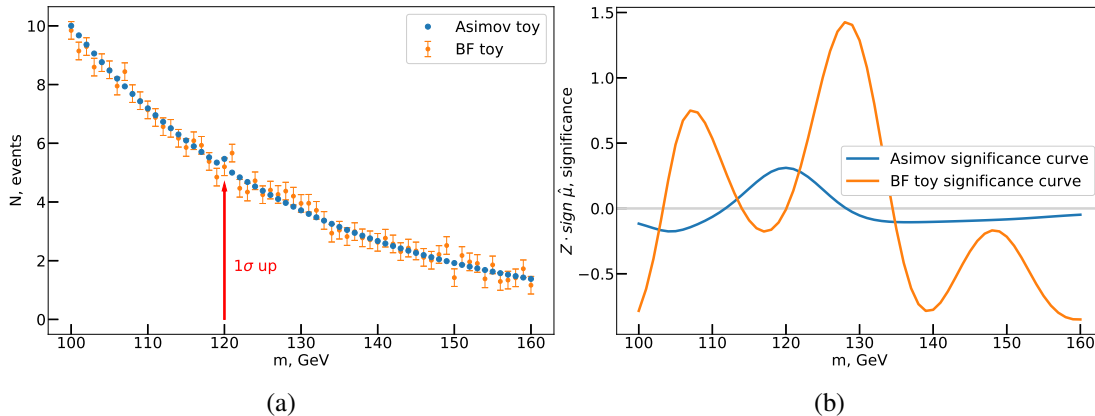
Figure 8. A background sample, an Asimov toy (a) and the corresponding significance curves (b) from the $H \to \gamma\gamma$ inspired model. See the detailed explanation of the two plots in figure 4.

Already from the background fit to the Asimov sample (figure 8(b)), we can see that the increased flexibility of the background model introduces more sophisticated long-range correlations. Notice the asymmetry between the left and right sides of the peak. When the signal hypothesis is near the Asimov fluctuation at 120 GeV, the signal part of the model accommodates it, but well below and above this the rate parameter of the background model tries to compensate for the local excess, which distorts the background model and leads to weak but clear regions of long-range anti-correlation and correlation. We confirm this observation from the plots of the GP covariance that we estimate from the Asimov samples (figure 9(b)).

We then compare the Asimov covariance to the one calculated with brute force (figure 10). It shows the same degree of accuracy as we have observed for the Background template model: the differences between the Asimov and brute force covariances are 100 times smaller than the values of the covariance itself.

The various computations of the trials factor for the $H \to \gamma\gamma$ inspired model are shown in figure 11. Since the likelihood and the generation of Monte Carlo data sets both use a Gaussian pdf for the data in each bin, there is no significant difference between the brute force and GP results at high significance, in contrast to the results for the Background template model.

### 3.3 $H \to \gamma\gamma$ model with 2-dimensional search region

We keep the background model unchanged from the 1-D example (eq. (3.4)), however, we want to study a signal model with more than one nuisance parameter. In this case a 1-dimensional
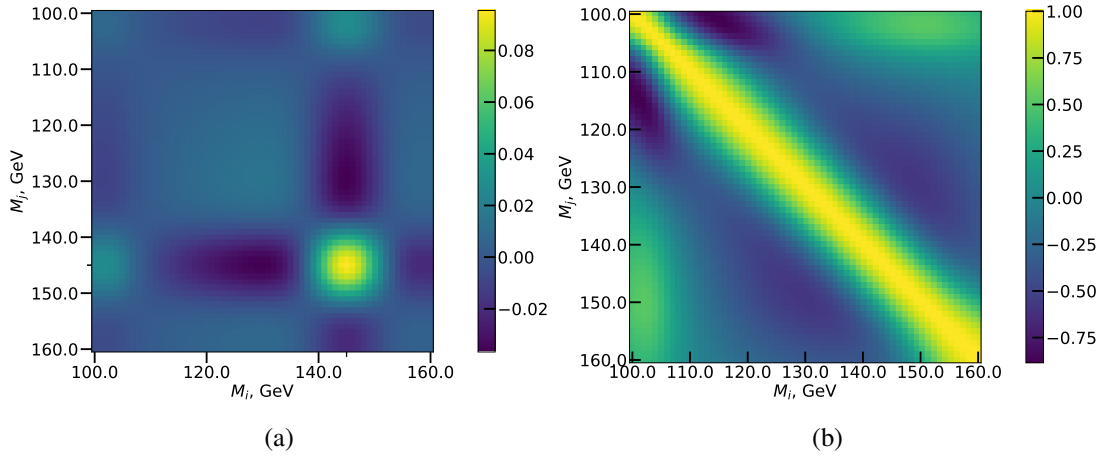
Figure 9. The partial contribution to the covariance $\hat{\Sigma}(M_i, M_j)$ (a) from the 145.5 GeV Asimov data set and the full Asimov GP covariance (b) for the $H \to \gamma\gamma$ inspired model.
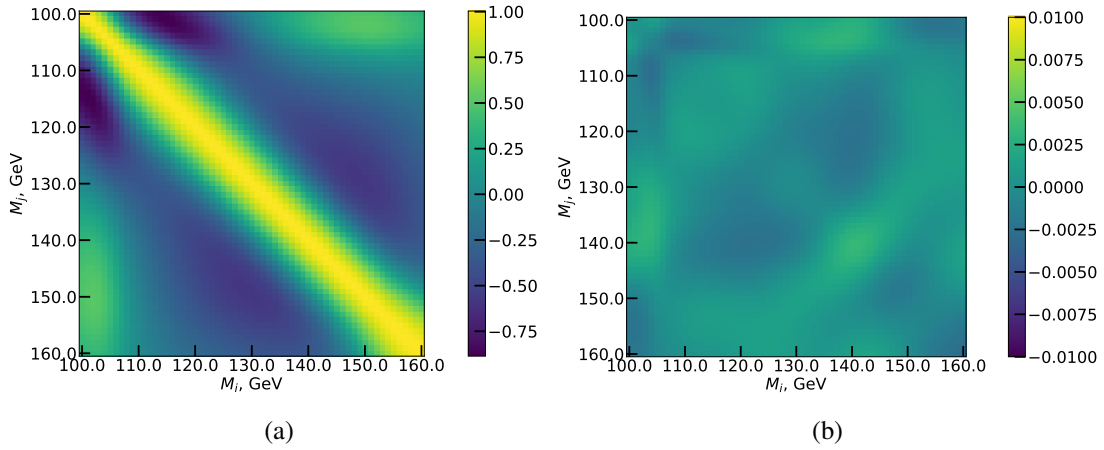
Figure 10. The difference (b) between the brute force (a) and Asimov (figure 9(b)) covariance matrices calculated for the $H \to \gamma\gamma$ inspired model.
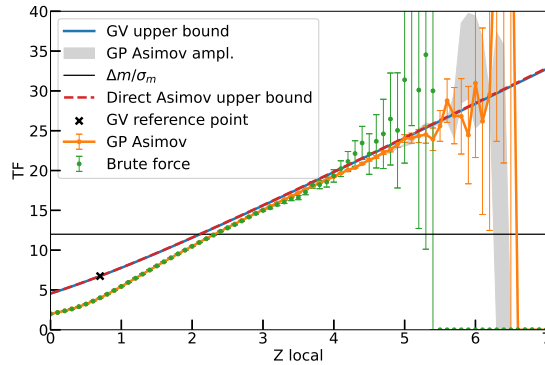
Figure 11. Trials Factor TF as a function of local significance Z_{local} for the $H \to \gamma\gamma$ inspired model. See the detailed explanation of the plot in figure 7.

significance curve becomes a multidimensional surface. However, each point on this surface still behaves like a standard normal random variable. As defined in section 2, M denotes the set of points that constitute the search region.

For example, in the 2-D case, when in addition to *mass m_s* we also add *width σ_s* as a second nuisance parameter, M gains a second axis, consequently, the significance surface becomes 2-dimensional. The signal model for this example is[5]

$$s_i(M) = s_i(m_s, \sigma_s) = \frac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{(m_i - m_s)^2}{2\sigma_s^2}} . \tag{3.7}$$

The likelihood function is the same as for the 1-D case (see eq. (3.6)).

For the covariance calculation the situation does not change if we think in terms of M, i.e., the expressions in eq. (2.4) still hold. For visualization, on the contrary, it is convenient to distinguish between *mass* and *width* dimensions. The transition between M_{mn}, that preserves *mass* and *width* correspondingly, and M_i can be constructed as follows:

$$\{M_1, M_2, \ldots\} \tag{3.8}$$
$$\Downarrow$$
$$\{M_{11}, \ldots, M_{1k}, M_{21}, \ldots, M_{2k}, M_{31}, \ldots\}.$$

In figure 12 we visualize the Asimov covariance matrix in the 2-dimensional case, which we unwrapped according to eq. (3.8). We also show different projections of this covariance and compare the Asimov covariance to the brute force covariance computed with $3 \cdot 10^6$ toys in figure 13. We again observe that the differences between covariances (figure 13(e)) are 100 times smaller than the values in the covariance matrix in figure 12.


(a) (b)
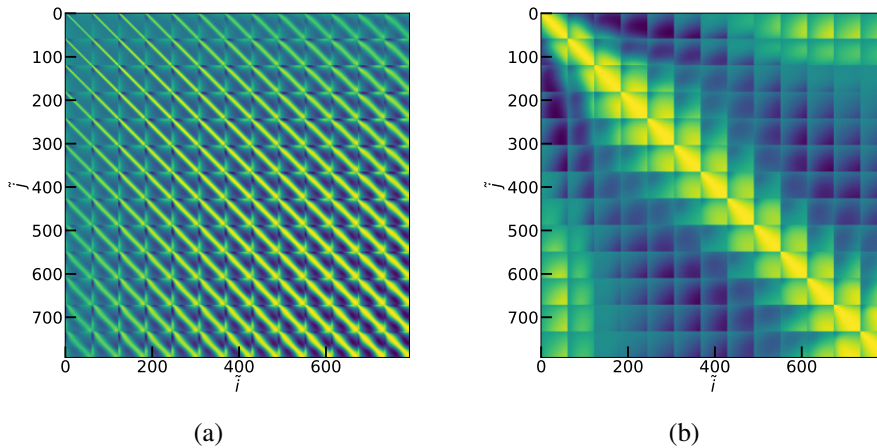
Figure 12. 4D covariance matrix $M_{\tilde{i}\tilde{j}}$ unwrapped into 2D following eq. (3.8) (a) and its transposed version (b), for the $H \rightarrow \gamma\gamma$ inspired model with a 2D search region. For the purpose of illustration, the less frequently changing dimension on each plot has a reduced number of points, consequently, the plots have a reduced number of "squares". The notation $M_{\tilde{i}\tilde{j}}$ emphasises that the grid was artificially downsampled.

---

[5]$\alpha_b = 0.033\,\text{GeV}$, $b_0 = 10$, $\sigma_b = 0.3\,\text{GeV}$, $M_{ij} = (a_i, b_j)$, $m_i = 100 - 160\,\text{GeV}$ with a step of $1\,\text{GeV}$; $a_i = 100 - 160\,\text{GeV}$ and $b_i = 1 - 10\,\text{GeV}$ with 61 points on the grid each.
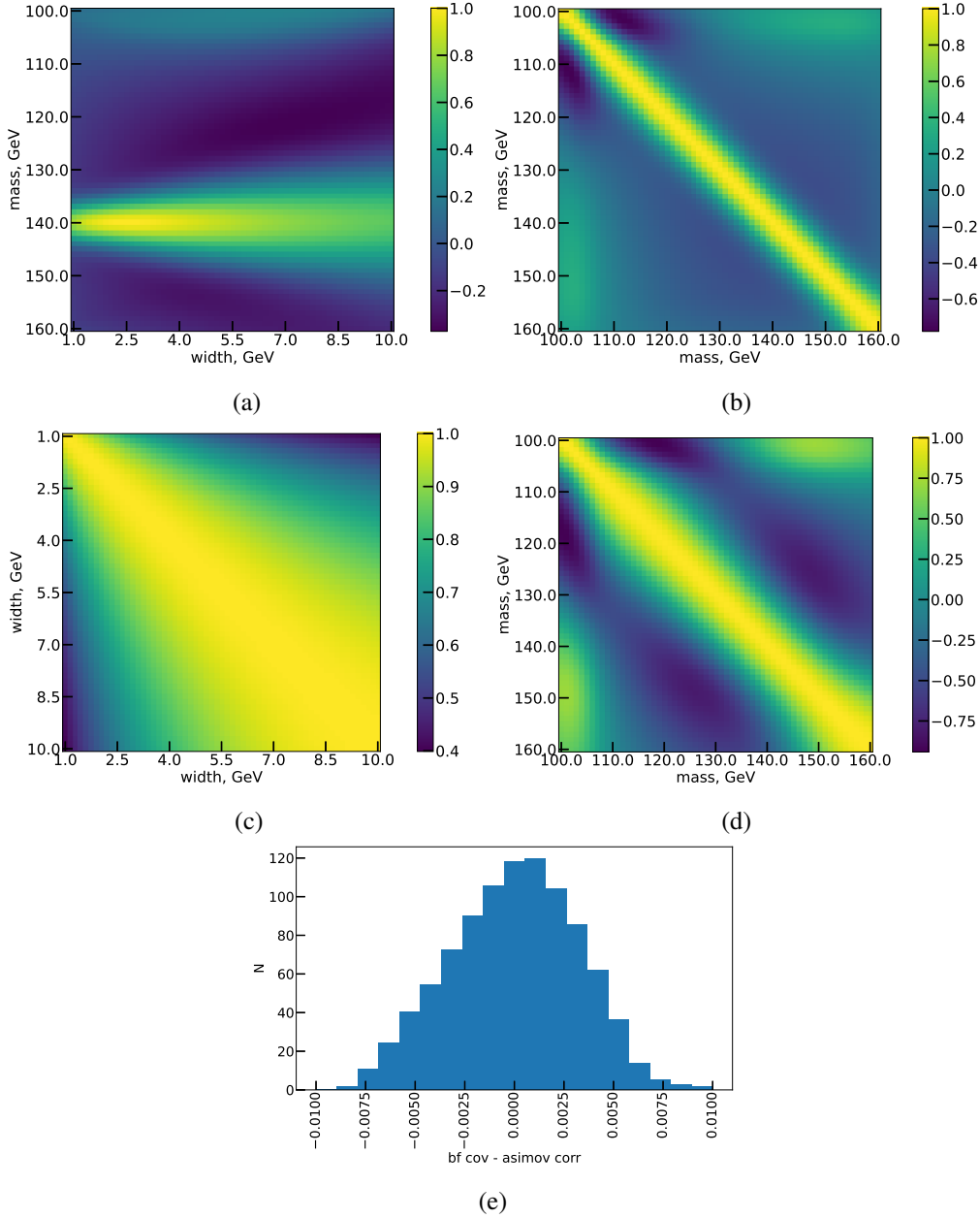
Figure 13. Covariance for the 2D $H \rightarrow \gamma\gamma$ model (in the 2D case a covariance between two significance surfaces). The covariance between the point on the first surface and 1 chosen point of mass and width ($m_s = 140 \, \text{GeV}$, $\sigma_s = 2.5 \, \text{GeV}$) from the second surface is shown in (a). The covariance between the slices of surfaces, where σ_s was fixed for both surfaces, is shown in (b) for $\sigma_s = 3.25 \, \text{GeV}$ and in (d) for $\sigma_s = 7 \, \text{GeV}$. Similarly, (c) shows the pattern for the slices where m_s is set to 130 GeV for both surfaces. The histogram in (e) shows the distribution of the values in the cells of the matrix of differences between the brute force and Asimov covariances, where N denotes the absolute bin counts.

The trials factors we obtained for this model are shown in figure 14. In this case the Vitells and Gross upper bound [2] is based on the Euler characteristic of the 2-dimensional significance surface. The upper bound calculation requires the Euler characteristic to be known at two significance levels
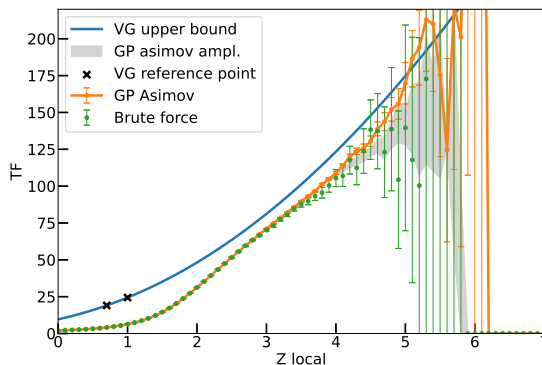
Figure 14. Trials Factor TF as a function of local significance Z_{local} for the $H \to \gamma\gamma$ inspired model with 2D search region. The green dots, orange line, and gray area are explained in figure 7. The solid blue line shows the Vitells and Gross upper bound [2] extrapolated from the two reference significance thresholds ($\sqrt{0.5}$ and 1) (black crosses).

(black crosses) before extrapolation. From the qualitative comparison of figure 14 to figure 7 and figure 11 it seems that for a higher dimensional search region, the Vitells and Gross upper bound becomes even more conservative.

4 Performance analysis

Several times in our work we mentioned that GP toys are more efficient in comparison to brute force toys. In this section we will quantify this statement.

We chose the $H\gamma\gamma$ inspired model as a benchmark. There were two reasons that guided our choice:

- The $H\gamma\gamma$ inspired model is more complex than the background template model,

- Adding more dimensions, according to section 3.3, only increases the number of fits per MC sample, but does not change the relative fitting performance.

Fitting on a regular laptop with 4 cores and one Python process, it took us 9 m 30 sec to produce 10^3 significance curves, which is 62 000 fits for the 61-bin grid, or about 2 significance curves per second.

To estimate the covariance matrix with our proposed set of Asimov samples requires $(61 + 1) \times 61 = 3782$ fits to obtain 61 significance curves, which took 23 seconds, or also about 2 significance curves per second. Having the covariance, we then managed to produce $10 \cdot 10^6$ GP toys in 27 sec, or roughly 4×10^5 GP toys per brute force toy.

In summary, for little additional effort and CPU time, we can create much larger samples to estimate global p-values or average up-crossings.

5 Conclusion

A precise estimation of the look-elsewhere effect in searches for new resonances in high energy physics is challenging. Existing methods are overly conservative unless the observed local significance is high enough. In this paper we proposed and evaluated a new approach based on Gaussian

processes. The new approach doesn't eliminate the production of Monte Carlo "toys", however, the computational process becomes much more efficient. We tested the approach on several models, including models with more than 1 nuisance parameter, that intentionally were constructed to have different statistical features. The trials factors calculated with our new approach and with large samples of brute force Monte Carlo simulations showed excellent agreement. As a cross-check, we reproduced the result of the Gross and Vitells paper [1] that introduced the high-significance approximation based on upcrossings. An actual realization of the method and implementation of the test models was published and is available for use out of the box [7].

Acknowledgments

References

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

[2] O. Vitells and E. Gross, *Estimating the significance of a signal in a multi-dimensional search*, *Astropart. Phys.* **35** (2011) 230 [arXiv:1105.4355].

[3] R.B. Davies, *Hypothesis testing when a nuisance parameter is present only under the alternative*, *Biometrika* **74** (1987) 33.

[4] ATLAS collaboration, *Search for the Standard Model Higgs boson in the diphoton decay channel with $4.9\,fb^{-1}$ of pp collisions at $\sqrt{s} = 7\,TeV$ with ATLAS*, *Phys. Rev. Lett.* **108** (2012) 111803 [arXiv:1202.1414].

[5] CMS collaboration, *Search for the standard model Higgs boson decaying into two photons in pp collisions at $\sqrt{s} = 7\,TeV$*, *Phys. Lett. B* **710** (2012) 403 [arXiv:1202.1487].

[6] V. Ananiev and A.L. Read, *Datasets of fitted toy monte carlo samples for "gaussian process-based calculation of look-elsewhere trials factor"*, 2023, https://doi.org/10.5281/ZENODO.7793309.

[7] https://gitlab.com/sigcorr/sigcorr.

[8] V. Ananyev and A. Read, *Documentation and examples for SigCorr*, https://sigcorr.gitlab.io/sigcorr/dev/.

[9] S.S. Wilks, *The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses*, *Ann. Math. Statist.* **9** (1938) 60.

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

[11] L.D. Lutes and S. Sarkani, *Random Vibrations. Analysis of Structural and Mechanical Systems*, Butterworth Heinemann, Boston (2004).

[12] H. Cramér and M.R. Leadbetter, *Stationary and related stochastic processes: sample function properties and their applications*, Wiley (1967).

Paper II

Linear approximation to the statistical significance autocovariance matrix in the asymptotic regime

SISSA

PAPER • OPEN ACCESS

Linear approximation to the statistical significance autocovariance matrix in the asymptotic regime

To cite this article: V. Ananiev and A.L. Read 2023 *JINST* **18** P10018

View the article online for updates and enhancements.

41

Linear approximation to the statistical significance autocovariance matrix in the asymptotic regime

V. Ananiev and A.L. Read[*]

Department of Physics, University of Oslo,
Boks 1048 Blindern, Oslo, NO-0316, Norway

E-mail: a.l.read@fys.uio.no

Abstract: Approximating significance scans of searches for new particles in high-energy physics experiments as Gaussian fields is a well-established way to estimate the trials factors required to quantify global significances. We propose a novel, highly efficient method to estimate the covariance matrix of such a Gaussian field. The method is based on the linear approximation of statistical fluctuations of the signal amplitude. For one-dimensional searches the upper bound on the trials factor can then be calculated directly from the covariance matrix. For higher dimensions, the Gaussian process described by this covariance matrix may be sampled to calculate the trials factor directly. This method also serves as the theoretical basis for a recent study of the trials factor with an empirically constructed set of Asmiov-like background datasets. We illustrate the method with studies of a $H \to \gamma\gamma$ inspired model that was used in the empirical paper.

---

[*]Corresponding author.

Contents

1 Introduction

In high energy physics searches for new particles that appear in the data as resonances [1, 2], one usually scans a mass region and hopes to find a peak of high significance at some mass. The significance at each mass of the scan is generally found by applying Wilks' theorem [3] to the likelihood-ratio test statistic (LRT) [4] for each point, and results in a field of significances measured across the search region.

While the resonance may appear anywhere in the search region, the analysis usually targets the highest (local) significance, which leads to the recurring challenge of estimating the global significance of this observation. The necessity of calculating the probability for a background fluctuation to give such a peak of significance anywhere in the search region, and not simply where the significance is maximal, is commonly referred to as the look-elsewhere effect (LEE).

There have been a number of studies investigating the LEE, and in our work we pay particular attention to those describing the significance field with a Gaussian process. While some studies [5–7] set the upper bound on the trials factor, which converts a local p-value into a global one, and only use a Gaussian process implicitly to link the low and high significance regions, other studies [8] require explicit values for the Gaussian process parameters.

In this paper we establish a chain of lightweight steps from a non-linear parametric statistical model to the trials factor by estimating the covariance matrix of the significance field. To construct the estimate involving only one background only fit to the data, we apply linear expansion to the non-linear background shape. The way to calculate the covariance matrix starting from a linear model was briefly discussed by Demortier ([9], pg. 23–33). As part of our work, we give a strict mathematical formulation of the method and demonstrate a practical application of it to non-linear background shapes, with the estimated covariance matrix serving as a proxy for the straightforward trials factor estimate.

A common input for the methods that quantify the LEE is a set of maximum likelihood fits to some number of Monte Carlo generated data realizations. They may be used to estimate the trials factor in the lower significance region, or the covariance matrix of the Gaussian process itself (the

significance autocovariance). The challenge, then, is to fit enough datasets to estimate the trials factor with a satisfactory precision, while keeping the number of fits as small as possible.

In high-energy physics searches for a new particle or a resonance, typically, the likelihood-ratio test statistic is used to construct the p-value for each point on a search grid. In the asymptotic regime, the test statistic follows a χ^2 distribution.

For analyses that use a Gaussian process to model the significance, the number of degrees of freedom of the test statistic distribution is, typically, 1. For this case, in section 2, we suggest a method to estimate the significance covariance matrix that makes use of a single background-only fit to the data.

We replace the set of fits that were required in our previous work, with derivatives of the best-fit-to-the-data background model. Fortunately, the derivatives can often be extracted from the fit software.

Core assumptions. In section 3 we show that three quite generic requirements:

1. the background model should be well approximated by its linear expansion around the best fit parameters,

2. the assumption that the data can be binned and fluctuations in different bins of the dataset are independent,

3. the fluctuations in each bin follow a Gaussian distribution,

together, are consistent with the assumptions made in the empirical study by Ananiev & Read [8], which relied on the additivity (superposition) principle for the fluctuations to empirically estimate the covariance matrix of the significances. We argue, therefore, that this work serves as a theoretical basis for the method of the set of Asimov background samples introduced in the study, and at the same time may rely on its validations.

1.1 Statistical model

The basic structure of a statistical model commonly used in high-energy physics experiments that search for a new particle or a resonance was described in detail in the empirical study [8]. For the present study, we chose the $H \rightarrow \gamma\gamma$ inspired model as a benchmark, because it satisfies without approximation the second and third requirements above.

The search is conducted with the likelihood ratio test statistic evaluated for each point M of the search grid $\mathcal{M}$.

In this binned model, the expected background $b_i(\boldsymbol{\theta})$ has an exponential shape and is used as the null-hypothesis H_0. The shape of the expected signal $\mu s_i(\boldsymbol{\theta})$ is Gaussian and together with background b_i forms the alternative H_1, expected signal + background estimate:

$$n_i(\mu, \boldsymbol{\theta}, M) = \mu s_i(\boldsymbol{\theta}, M) + b_i(\boldsymbol{\theta}), \tag{1.1}$$

where i enumerates bins, $\boldsymbol{\theta}$ denotes the vector of nuisance parameters and μ is the signal strength parameter.

Generally, in the asymptotic regime (e.g. large sample), and neglecting constant terms, log-likelihoods for H_0 and H_1 may be approximated as follows:[1]

$$-2 \ln \mathcal{L}_0(\mu = 0, \boldsymbol{\theta}) = \sum_i \left(\frac{d_i - b_i(\boldsymbol{\theta})}{\sigma_i} \right)^2, \tag{1.2}$$

$$-2 \ln \mathcal{L}_1(\mu, \boldsymbol{\theta}, M) = \sum_i \left(\frac{d_i - b_i(\boldsymbol{\theta}) - \mu s_i(M, \boldsymbol{\theta})}{\sigma_i} \right)^2,$$

where i enumerates bins, $M \in \mathcal{M}$ denotes the point in the search region $\mathcal{M}$ of parameters which are not present under the background-only hypothesis, $\boldsymbol{\theta}$ are the nuisance parameters, and d_i corresponds to the binned data with errors σ_i.[2]

Our goal is to estimate the covariance matrix Σ_{MN} of the statistical significances Z_M and Z_N evaluated at two different points of the search region $\mathcal{M}$:

$$\Sigma_{MN} = \langle Z_M Z_N \rangle_d, \quad M, N \in \mathcal{M}, \tag{1.3}$$

$$Z_M = \text{sign}(\hat{\mu}) \sqrt{t_\mu(M)} \sim \mathcal{N}[0, 1], \tag{1.4}$$

$$t_\mu(M) = -2 \ln \frac{\mathcal{L}_0(\mu = 0, \boldsymbol{\theta}_0)}{\mathcal{L}_1(\hat{\mu}, \boldsymbol{\theta}_0 + \boldsymbol{\theta}_1, M)} \sim \chi^2_{\text{d.o.f=1}}, \tag{1.5}$$

where $t_\mu(M)$ is the likelihood-ratio test statistic (LRT), Z_M is the so-called signed-root LRT, $\boldsymbol{\theta}_0$ are the nuisance parameters that maximize the background-only likelihood $\mathcal{L}_0$, $\boldsymbol{\theta}_0 + \boldsymbol{\theta}_1$ together with the signal strength $\hat{\mu}$ maximize the signal+background likelihood $\mathcal{L}_1$, and $\mathcal{N}[0, 1]$ denotes the standard normal distribution.

To give a feeling of the $H\gamma\gamma$ inspired model, in figure 1 we plotted the shape of the background-only hypothesis b_i, one sample of data sampled from it, and a corresponding significance curve (eq. (1.4)). Notice how clearly visible bumps in the data are reflected in peaks of the significance curve.

We would like to remark that for the signal+background model we are fitting $\boldsymbol{\theta}$ as a deviation from $\boldsymbol{\theta}_0$. This is essential for the proper separation of variables in the subsequent calculations.

We assume that the best fit of the background model b_i to the data d_i is available for the study as $b_i(\hat{\boldsymbol{\theta}}) = \hat{b}_i$. In order to simplify the notation, we make use of the freedom to choose the reference point for the model parameters $\boldsymbol{\theta}$ and define the best fit parameters to be $\hat{\boldsymbol{\theta}} = \mathbf{0}$.

2 Method

To simplify the notation, we redefine d_i, s_i and b_i to include σ_i:

$$\frac{d_i}{\sigma_i} \mapsto d_i, \quad \frac{s_i}{\sigma_i} \mapsto s_i, \quad \frac{b_i}{\sigma_i} \mapsto b_i. \tag{2.1}$$

---

[1]The $H\gamma\gamma$ inspired model assumes Gaussian errors in its definition [8]. The expressions for log-likelihoods (eq. (1.2)) in case of this model are, therefore, exact.

[2]We have assumed that the errors σ_i are independent of the nuisance parameters $\boldsymbol{\theta}$. With a linear correction to σ_i it is still possible to get a closed form expression for the test statistic and significance. The calculation of the covariance would require sampling toys to average out the fluctuations. No additional fits would be required, however, so this may be a potential option for more sophisticated analyses.
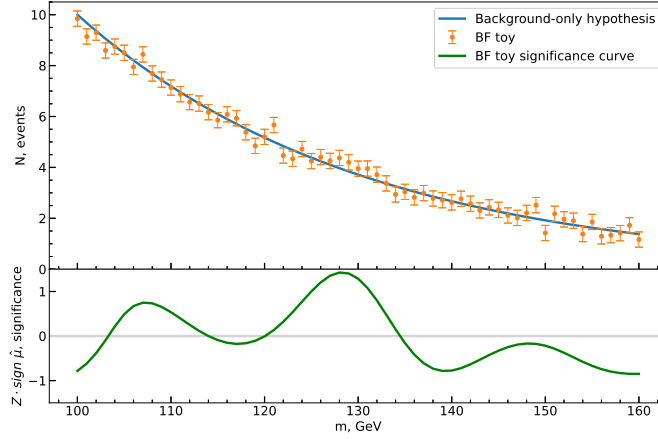
Figure 1. One brute force toy (orange) sampled from a background-only hypothesis (blue) of the $H \to \gamma\gamma$ inspired model. The signal significance curve is plotted below (green), where each point of the curve corresponds to a different choice of the signal hypothesis.

The log-likelihoods then become:

$$-2\ln \mathcal{L}_0 = \sum_i (d_i - b_i(\boldsymbol{\theta}))^2, \tag{2.2}$$

$$-2\ln \mathcal{L}_1 = \sum_i (d_i - b_i(\boldsymbol{\theta}) - \mu s_i(\boldsymbol{\theta}))^2.$$

For every realization of the data, we expect the deviations of the fit parameters μ and $\boldsymbol{\theta}$ from 0 to be small (in the absence of a signal), and therefore the first-order expansion of $b_i(\boldsymbol{\theta})$ and $s_i(\boldsymbol{\theta})$ around $\mathbf{0}$ to be accurate enough.

The log-likelihoods then are:

$$-2\ln \mathcal{L}_0 = \sum_i \left(d_i - \hat{b}_i - \Delta_{i\beta}\theta^\beta\right)^2, \tag{2.3}$$

$$-2\ln \mathcal{L}_1 = \sum_i \left(d_i - \hat{b}_i - \Delta_{i\beta}\theta^\beta - \mu s_i(\mathbf{0})\right)^2,$$

where $\Delta_{i\alpha} = \frac{\partial b_i(\boldsymbol{\theta})}{\partial \theta^\alpha}\big|_{\boldsymbol{\theta}=\mathbf{0}}$ is the Jacobian of the best-fit background model and the Einstein summation rule applies to the indices β. Since the signal model s_i contributes to the log-likelihoods eq. (2.3) only at lowest order, thus is constant, we simplify $s_i(\mathbf{0})$ to s_i from now on.

The equations that define optimal values of $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_1$, and μ then are:

$$\frac{\partial \mathcal{L}_0}{\partial \theta_\alpha}\big|_{\boldsymbol{\theta}_0} \propto \sum_i (d_i - \hat{b}_i - \Delta_{i\beta}\theta_0{}^\beta) \cdot \Delta_{i\alpha} = 0, \tag{2.4}$$

$$\frac{\partial \mathcal{L}_1}{\partial \theta_\alpha}\big|_{\boldsymbol{\theta}_1, \hat{\mu}} \propto \sum_i (d_i - \hat{b}_i - \Delta_{i\beta}(\theta_0{}^\beta + \theta_1{}^\beta) - \hat{\mu} s_i) \cdot \Delta_{i\alpha} = 0, \tag{2.5}$$

$$\frac{\partial \mathcal{L}_1}{\partial \mu}\big|_{\boldsymbol{\theta}_1, \hat{\mu}} \propto \sum_i (d_i - \hat{b}_i - \Delta_{i\beta}(\theta_0{}^\beta + \theta_1{}^\beta) - \hat{\mu} s_i) \cdot s_i = 0. \tag{2.6}$$

To reduce the number of indices, we rewrite the expressions above with bra-ket notation:

$$\langle d - \hat{b} | \, \Delta = \langle \theta_0 | \, \Delta^\mathsf{T} \Delta, \tag{2.7}$$

$$\mathbf{0} = \langle \theta_1 | \, \Delta^\mathsf{T} \Delta + \hat{\mu} \, \langle s | \, \Delta, \tag{2.8}$$

$$\langle d - \hat{b} | s \rangle = \langle \theta_0 + \theta_1 | \, \Delta^\mathsf{T} \, | s \rangle + \hat{\mu} \, \langle s | s \rangle, \tag{2.9}$$

where in eq. (2.8) we used eq. (2.7) to cancel the θ_0 contribution. We can solve eq. (2.7) and eq. (2.8) for θ_0 and θ_1 correspondingly:

$$\langle \theta_0 | = \langle d - \hat{b} | \, \Delta (\Delta^\mathsf{T} \Delta)^{-1}, \tag{2.10}$$

$$\langle \theta_1 | = -\hat{\mu} \, \langle s | \, \Delta (\Delta^\mathsf{T} \Delta)^{-1}. \tag{2.11}$$

It is important to mention that, although Δ itself is generally singular, the product $\Delta^\mathsf{T} \Delta$ appears to be a Hessian of $-2 \ln \mathcal{L}_1$ with respect to θ_1. For the background model best-fit point $\theta = \mathbf{0}$ to be a minimum, it is required that the Hessian be positive definite, thus $\Delta^\mathsf{T} \Delta$ is invertible.

We substitute eq. (2.10) and eq. (2.11) into eq. (2.9) and solve for $\hat{\mu}$:

$$\hat{\mu}(M) = \frac{\langle d - \hat{b} | \, P \, | s_M \rangle}{\langle s_M | \, P \, | s_M \rangle}, \tag{2.12}$$

$$P = \mathbb{1} - \Delta (\Delta^\mathsf{T} \Delta)^{-1} \Delta^\mathsf{T}.$$

An interesting and important fact is that P is a projector and it is symmetric:

$$P^2 = P, \quad P = P^\mathsf{T}. \tag{2.13}$$

A projector is always positive semi-definite, which means that the product below is non-negative for any non-zero $\mathbf{s}$:

$$\langle s | \, P \, | s \rangle = \langle s | \, P^2 \, | s \rangle = (P \, | s \rangle)^2 \geq 0, \quad \forall \mathbf{s} \neq \mathbf{0}. \tag{2.14}$$

Let us estimate the test statistic t_M:

$$\begin{aligned} t_M &= (-2 \ln \mathcal{L}_0) - (-2 \ln \mathcal{L}_1) \\ &= 2 \langle d - \hat{b} - \Delta \theta_0 | \Delta \theta_1 + \hat{\mu} s \rangle + \langle \Delta \theta_1 + \hat{\mu} s | \Delta \theta_1 + \hat{\mu} s \rangle. \end{aligned} \tag{2.15}$$

We again use eq. (2.7) to cancel the θ_0 contribution and eq. (2.11) to substitute the solution for θ_1:

$$t_M = \hat{\mu} \, \langle d - \hat{b} | \, P \, | s_M \rangle = \hat{\mu}^2 \, \langle s_M | \, P \, | s_M \rangle. \tag{2.16}$$

The significance Z_M, as defined in eq. (1.4), is:

$$Z_M = \hat{\mu} \sqrt{\langle s_M | \, P \, | s_M \rangle} = \frac{\langle d - \hat{b} | \, P \, | s_M \rangle}{\sqrt{\langle s_M | \, P \, | s_M \rangle}}. \tag{2.17}$$

The square root in eq. (2.17) is always defined, as the product under the square root is always positive (eq. (2.14)).

For the covariance matrix estimation, we would need to average over data. We are looking for a solution with uncorrelated fluctuations in each bin (section 1), and we recall that we normalized the errors to 1 in eq. (2.1), therefore, the following is true:

$$E_d \left\{ |d - \hat{b}\rangle \langle d - \hat{b}| \right\} = \mathbb{1}, \tag{2.18}$$

where E_d denotes the expectation value calculated across samples of the dataset.

The covariance matrix, then, is:[3]

$$
\begin{aligned}
\Sigma_{MN} &= E_d \left\{ Z_M Z_N \right\} \\
&= E_d \left\{ \frac{\langle s_M| P |d - \hat{b}\rangle \langle d - \hat{b}| P |s_N\rangle}{\sqrt{\langle s_M| P |s_M\rangle} \sqrt{\langle s_N| P |s_N\rangle}} \right\} \\
&= \frac{\langle s_M| P}{\sqrt{\langle s_M| P |s_M\rangle}} E_d \left\{ |d - \hat{b}\rangle \langle d - \hat{b}| \right\} \frac{P |s_N\rangle}{\sqrt{\langle s_N| P |s_N\rangle}} \\
&= \frac{\langle s_M|}{\sqrt{\langle s_M| P |s_M\rangle}} P \frac{|s_N\rangle}{\sqrt{\langle s_N| P |s_N\rangle}},
\end{aligned}
\tag{2.19}
$$

where we used the symmetry and projector properties of P.

It should be noted that from the data fluctuations $\mathbf{d} - \hat{\mathbf{b}}$ contributing to the covariance matrix in the form

$$\text{Fluct.} \propto E_d \left\{ |d - \hat{b}\rangle \langle d - \hat{b}| \right\}, \tag{2.20}$$

a superposition principle, relied on in ref. [8], can be derived:

$$\Sigma_{MN} = \sum_f \Sigma_{MN}^f, \tag{2.21}$$

where f enumerates independent fluctuations in different bins.

In summary, we can estimate the autocovariance matrix of the significance field from the signal model and derivatives of the background model:

$$
\begin{aligned}
\Sigma_{MN} &= \frac{\langle s_M|}{\sqrt{\langle s_M| P |s_M\rangle}} P \frac{|s_N\rangle}{\sqrt{\langle s_N| P |s_N\rangle}}, \quad M, N \in \mathcal{M} \\
P &= \mathbb{1} - \Delta (\Delta^\intercal \Delta)^{-1} \Delta^\intercal, \\
\Delta_{i\alpha} &= \frac{\partial b_i(\boldsymbol{\theta})}{\partial \theta^\alpha} |_{\boldsymbol{\theta}=\mathbf{0}}.
\end{aligned}
\tag{2.22}
$$

---

[3]To see the parallel with Demortier [9], one needs to think of the background model as a linear combination of vectors in Δ. Then eq. (2.8) defines a vector $|v_M\rangle = \frac{P|s_M\rangle}{\sqrt{\langle s_M|P|s_M\rangle}}$, which was introduced by Demortier and is orthogonal to each of the vectors constituting the background shape. The test statistic, then, can be rewritten as $t_M = \left(\langle d - \hat{b}|v_M\rangle \right)^2$, and the covariance can be expressed as $\Sigma_{MN} = \langle v_M|v_N\rangle$.

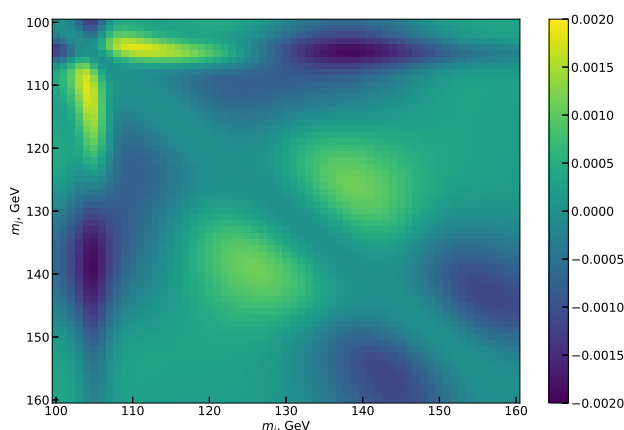# 3 Justification of the set of Asimov background samples

In this section we would like to compare the derived expression eq. (2.22) for the linear approximation of the significance covariance matrix to the empirical study [8] and the $H \rightarrow \gamma\gamma$ inspired model introduced there. To carry out the calculations we used the SigCorr package that we developed specifically for trials factor studies, which now includes functionality for the linear approximation [10].

We estimate the linear approximation using eq. (2.22) with the true parameters of the model, which were predefined in the paper. The resulting matrix shown in figure 2 is visually indistinguishable from the one presented in the empirical study.
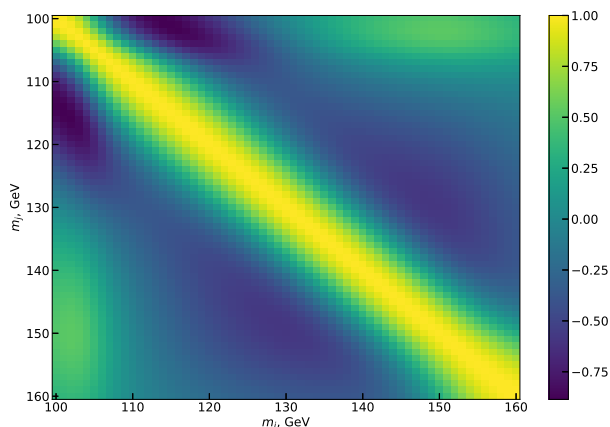
Figure 2. A linear approximation of the significance covariance matrix which was computed on the true parameters of the $H \rightarrow \gamma\gamma$ inspired model.

We also show, in figure 3, the difference between the linear approximation computed on the model's true parameters (figure 2) and the empirical estimate. We confirm that the empirical

Figure 3. The difference between the linear approximation of the significance covariance matrix computed with the true parameters of the $H \rightarrow \gamma\gamma$ inspired model (figure 2) and the covariance matrix estimated with the set of Asimov background samples [8].

– 7 –

covariance matrix is compatible with the linear approximation suggested in this paper within the accuracy of the empirical estimate.

On the one hand, the compatibility of the linear approximation and the empirical study allows us to refer to the validations conducted in the empirical study, including those regarding trials factor estimation, and to re-apply them to the method suggested in this paper. The direct calculation of the up-crossings from the covariance matrix, described in [8], becomes particularly appealing now, since it requires only a single fit of the statistical model to the data.

The linear approximation, on the other hand, serves as the theoretical basis for the empirical set of Asimov background samples used to estimate the covariance matrix in the aforementioned work.

4   Conclusion

In this work we proposed a novel method for the estimation of the covariance matrix of statistical significance in new particle searches using a linear expansion of the statistical model around its background-only best fit to the data. In addition to the closed form expression for the linear approximation of the significance covariance matrix, we also presented elegant expressions for the best fitted signal strength and statistical significance in this approximation.

We proved that the suggested covariance matrix satisfies the superposition principle with regard to the fluctuations of the data, which makes it a good proxy to the covariance matrix constructed with the set of Asimov background samples [8].

Finally, we compared these two approaches with the example of a $H \to \gamma\gamma$ inspired model and showed that the deviations are compatible with the error of the set of Asimov background samples.

We, therefore, claim that all the validations conducted in the empirical study, including those regarding trials factor estimation, hold for the linear approximation suggested in this paper, and the linear approximation serves as a theoretical basis for the empirical set of Asimov background samples construction.

Acknowledgments

References

[1] CMS collaboration, *Search for the Standard Model Higgs Boson Decaying into Two Photons in pp Collisions at $\sqrt{s}$ = 7 TeV*, *Phys. Lett. B* **710** (2012) 403 [`arXiv:1202.1487`].

[2] ATLAS collaboration, *Search for the Standard Model Higgs boson in the diphoton decay channel with 4.9 fb^{-1} of pp collisions at $\sqrt{s}$ = 7 TeV with ATLAS*, *Phys. Rev. Lett.* **108** (2012) 111803 [`arXiv:1202.1414`].

[3] S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, *Ann. Math. Stat.* **9** (1938) 60.

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

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

[6] O. Vitells and E. Gross, *Estimating the significance of a signal in a multi-dimensional search*, *Astropart. Phys.* **35** (2011) 230 [arXiv:1105.4355].

[7] R.B. Davies, *Hypothesis testing when a nuisance parameter is present only under the alternative*, *Biometrika* **74** (1987) 33.

[8] V. Ananiev and A.L. Read, *Gaussian Process-based calculation of look-elsewhere trials factor*, 2023 *JINST* **18** P05041 [arXiv:2206.12328].

[9] L. Lyons, H.B. Prosper and A. De Roeck, eds., *Proceedings of PHYSTAT-LHC Workshop on Statistical issues for LHC physics*, Geneva, Switzerland (2007), CERN-2008-001 (2008).

[10] V. Ananiev and A.L. Read, *SigCorr: A python package for studies of trials factors*, *J. Open Source Softw.* **8** (2023) 4989.

Paper III

SigCorr: A Python package for studies of trials factors

SigCorr: A Python package for studies of trials factors

V. Ananiev ●[1*] **and A. L. Read** ●[1*¶]

1 Department of Physics, University of Oslo, Boks 1072 Blindern, Oslo, NO-0316, Norway ¶ Corresponding author * These authors contributed equally.

Summary

The Trials factor (TF) is a generalization of the Bonferroni correction for the significance of repeated experiments. In a likelihood ratio scan that searches for the most significant peak of signal amplitude as a function of some parameter, e.g. the mass of a hypothetical particle, repeated experiments emerge effectively due to the localized nature of the hypothetical signal. The trials factor in such a search is the ratio between the probability of observing such an excess anywhere in the scanned search region to the probability of observing the excess at the point of maximum observed significance. For a more detailed introduction of the TF, we refer the reader to the work of Gross and Vitells (Gross & Vitells, 2010).

There are several ways to estimate the trials factor or an upper bound for it. Most of the approaches require a number of samples from the background distribution (Monte Carlo (MC) toys) to be generated and fitted with a statistical model. This procedure is separated in the upper block of Figure 1, mainly because it is the most time-consuming part of the analysis.

The lower block in Figure 1 shows various ways to estimate the trials factor from the fitted MC toys. When expressed as a pipeline, each way can be assembled from some simple building blocks, the implementation of which is at the core of SigCorr.

The communication between the upper (fitting) and the lower (analysis) components is conducted via HDF5 files with a well defined structure, which can be referred to in the documentation of SigCorr (Ananiev & Read, 2023a). Such a weak coupling allows to easily replace parts of the pipeline with more efficient tools if needed.
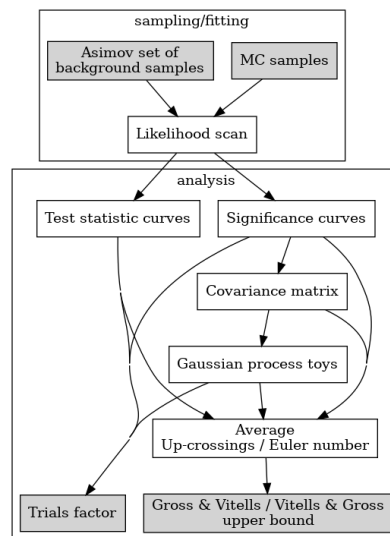
Figure 1: Various ways to estimate the trials factor.

SigCorr is a framework that implements a wide range of tools and was developed to simplify the process of construction of the pipelines shown above, and, therefore, to make studies of the global significance or a trials factor a straightforward exercise. It is worth mentioning, though, that the figure above is not exhaustive. We believe there are more ways to approach the trials factor, and some of them may also be possible to implement with the tools provided by SigCorr!

The project has a hybrid structure:

- From the perspective of fitting the Monte Carlo toys, it is a framework. Users will have to implement their own statistical models following the guides and examples in the documentation.

- From the perspective of analysis of the fitted toys, however, SigCorr is a Swiss Army knife that assembles into one tool the knowledge from the frequently cited HEP papers that studied statistical hypothesis testing and the trials factor (Cowan et al., 2011; Gross & Vitells, 2010; Vitells & Gross, 2011).

Statement of need

In high-energy physics it is a recurring challenge to efficiently and precisely (enough) calculate the global significance of, e.g., a potential new resonance. The Gross and Vitells trials factor approximation (Gross & Vitells, 2010) and (Vitells & Gross, 2011) is based on the average "up-crossings" of the significance in the search region, or generally on the average Euler characteristic of the set of significance measurements that exceed the threshold of the local significance. It has revolutionized the trials factor estimation for significances above 3 standard deviations, but the challenges of actually calculating the average up-crossings and the validity of the approximation for smaller significances remain.

In Ananiev & Read (2023b) a new method was proposed. It models the significance in the search region as a Gaussian process (GP). The method was developed to overcome the limitations of the Gross and Vitells approach via replacing expensive MC fits with lightweight GP toys.

Up-crossings, Euler characteristic and Gaussian processes are commonly relied on in this field of research. When studied together, they have many useful properties for the estimation of

trials factors. SigCorr is the first project that assembles all of them into one Python package, that also includes the tools for parallel analysis of the MC toys and GP toys in a unified fashion. Together with the fitting framework it allows the path from the statistical model definition to a TF estimate to be travelled.

SigCorr is a framework developed with a goal to cover a wide range of use cases for TF estimates. There are, however, popular and well-maintained packages which cover some aspects of this process. For example, RooFit (Verkerke & Kirkby, 2006) and PyHF (Heinrich et al., n.d.), which are widely used in particle physics, may significantly replace or augment the part of SigCorr that is responsible for optimizing the likelihoods of the statistical models. The pyBumpHunter package (Vaslin & Donini, 2023) may help with brute force estimates of the TF.

Although SigCorr covers just simplified versions of all the above, it also implements validated approximations and asymptotics for fast computations of the TF and its upper bound based on classical papers in the field. The possibility of doing this in a modular but consistent fashion makes SigCorr unique in the field, and will allow users to try different methods (Figure 1), cross-validate and choose or construct the most suitable approach to their analysis.

Examples of use

Average Euler characteristic propagation

For the Gaussian process field, for any chosen threshold on the field values, one can define a set of points above this threshold. The Euler characteristic of this set gives an estimation of the number of peaks above the threshold, which is a very important characteristic of the Gaussian process for the Trials factor studies.

A Gaussian process has a property that the average Euler characteristic at any threshold is a simple function of a few reference Euler characteristic values and the threshold (Vitells & Gross, 2011). The number of reference values required depends on the dimensionality of the Gaussian field.

Here is how to estimate the average Euler characteristic for arbitrary thresholds with SigCorr, knowing 2 reference values for a 2-dimensional Gaussian field:

```python
import numpy as np
from sigcorr.tools.stats.gp.euler_number import GPEulerNumberPropagator

# 2D Gaussian process requires 2 reference values
ref_thresholds = np.array([1, 2.])
ref_euler_nums = np.array([3.2, 1.2])
target_thresholds = np.array([0.5, 1.5, 2.5])
euler_number_propagator = GPEulerNumberPropagator(ref_thresholds,
                                                  ref_euler_nums)
target_euler_numbers = euler_number_propagator.calc(target_thresholds)
# array([3.10791656, 2.29280228, 0.46934814])
```

Batched statistics

When the number of samples is known in advance but not all samples are available at the moment (e.g. they don't fit into RAM), it is still possible to estimate mean, variance, covariance and statistical errors by processing the samples in batches.

```python
import numpy as np
from sigcorr.tools.stats.batch_stats import BatchStats2

n_samples = 3
one_sample_shape = 2
samples = np.array([[1, 2], [2, 3], [3., 4.]])
```

```
bs = BatchStats2(n_samples, one_sample_shape)
bs.push(samples[:1])  # push first batch
bs.push(samples[1:])  # push second batch

# sample mean
mean = bs.get_mean()
# array([2., 3.])

# sample variance
variance = bs.get_var()
# array([0.66666667, 0.66666667])

# covariance matrix between observables
covariance = bs.get_cov()
# array([[0.66666667, 0.66666667],
#        [0.66666667, 0.66666667]])

# variance of the covariance matrix between observables,
# can be used to estimate statistical errors
# on the sample covariance estimate
covariance_variance = bs.get_cov_var()
# array([[0.22222222, 0.22222222],
#        [0.22222222, 0.22222222]])
```

Statistical analysis of GP samples (parallel)

A set of standard normal random variables when squared follow a chi-squared distribution with 1 degree of freedom. One can see the analogy with a test statistic curve emerging from a likelihood scan. If we ask the question, what is the fraction of curves that exceed some threshold, this would resemble the procedure used to estimate the trials factor via brute force. Below we estimate the fraction of test statistic curves that exceed the threshold 1.2.

```python
import numpy as np
from sigcorr.mapreduce.gp import gp_batch_mapreduce
from sigcorr.tools.utils import get_last_from_iter
from sigcorr.mapreduce.map_reducers import ChainCalc
from sigcorr.mapreduce.map_reducers import MathCalc
from sigcorr.mapreduce.map_reducers import OverflowsCalc
from sigcorr.mapreduce.map_reducers import BatchStats1Reduce

cov = np.eye(3) + 0.1
result_iterator = gp_batch_mapreduce(
    cov,  # covariance matrix of the GP
    100,  # number of GP samples
    10,   # batch size
    (3, ),  # sample shape
    ChainCalc([MathCalc(np.square), OverflowsCalc(1.2)]),  # apply to every batch
    BatchStats1Reduce())  # aggregate
resulting_bs, num_processed = get_last_from_iter(result_iterator)
resulting_bs.get_mean()
```

Here, `ChainCalc`, `MathCalc`, `OverflowsCalc` are building blocks of the pipeline used to square the GP sample and then to set the value 0 or 1 per curve in the batch depending on whether the curve exceeds the threshold 1.2. There are other building blocks that, for example, help to estimate the significance from the maximum likelihood values (`SigsCalc`), or to compute the Euler characteristic of the batch of samples (`EulerNumberCalc`), etc.

Acknowledgements

References

Ananiev, V., & Read, A. L. (2023a). *Documentation and examples for SigCorr*. https://sigcorr.gitlab.io/sigcorr/latest/

Ananiev, V., & Read, A. L. (2023b). Gaussian process-based calculation of look-elsewhere trials factor. *Journal of Instrumentation*, *18*(05), P05041. https://doi.org/10.1088/1748-0221/18/05/P05041

Cowan, G., Cranmer, K., Gross, E., & Vitells, O. (2011). Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, *71*(2). https://doi.org/10.1140/epjc/s10052-011-1554-0

Gross, E., & Vitells, O. (2010). Trial factors for the look elsewhere effect in high energy physics. *The European Physical Journal C*, *70*(1-2), 525–530. https://doi.org/10.1140/epjc/s10052-010-1470-8

Heinrich, L., Feickert, M., & Stark, G. (n.d.). *Pyhf*. https://doi.org/10.5281/zenodo.1169739

Vaslin, L., & Donini, J. (2023). pyBumpHunter. In *GitHub repository*. https://github.com/scikit-hep/pyBumpHunter; GitHub.

Verkerke, W., & Kirkby, D. (2006, May). The RooFit Toolkit for data modeling. *Statistical Problems in Particle Physics, Astrophysics and Cosmology*. https://doi.org/10.1142/9781860948985_0039

Vitells, O., & Gross, E. (2011). Estimating the significance of a signal in a multi-dimensional search. *Astroparticle Physics*, *35*(5), 230–234. https://doi.org/10.1016/j.astropartphys.2011.08.005

Paper IV

Approximating the Mode of the Non-Central Chi-Squared Distribution

Approximating the Mode of the Non-Central Chi-Squared Distribution

V. Ananiev*, A. L. Read

Department of Physics, University of Oslo, Norway

*Corresponding author: victor.ananyev@gmail.com

Abstract. In this paper we consider the probability density function (pdf) of the non-central χ^2 distribution with arbitrary number of degrees of freedom and non-centrality. For this function we find the approximate location of the maximum and discuss related edge cases of 1 and 2 degrees of freedom. We also use this expression to demonstrate the improved performance of the C++ Boost's implementation of the non-central χ^2 and extend the domain of its applicability.

1. Introduction

Properties of the non-central χ^2 distribution were described before in literature [6–8]. However, the topic of the mode of the non-central χ^2 was significantly underrepresented. We would like to focus on the mode specifically in this paper.

Let $X_1, X_2, ..., X_n$ be normally distributed random variables with unit variance and means $\mu_1, \mu_2, ..., \mu_n$. The sum $X_1^2 + X_2^2 + ... + X_n^2$ follows the non-central χ^2 distribution with $k = n$ degrees of freedom and non-centrality $\lambda = \mu_1^2 + \mu_2^2 + ... + \mu_n^2$. The probability density function of this distribution has a closed form expression:

$$f_{k,\lambda}(x) = \frac{1}{2} \exp^{-\frac{x+\lambda}{2}} \left(\frac{x}{\lambda}\right)^{\frac{k-2}{4}} I_{\frac{k-2}{2}}(\sqrt{\lambda x}), \tag{1.1}$$
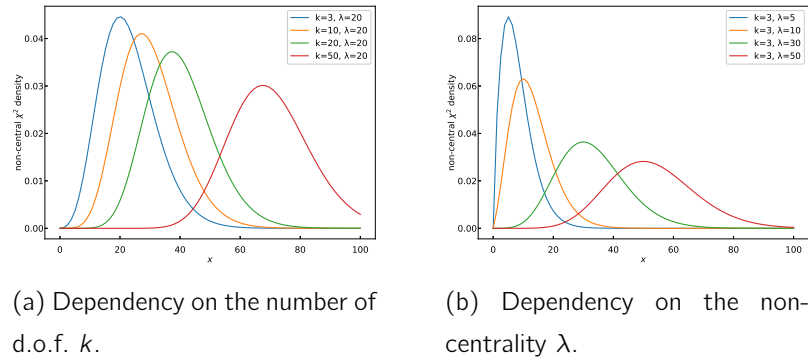
where $I_\nu(x)$ is a modified Bessel function of the first kind.

We are interested in the value of x_{mode} that maximizes $f_{k,\lambda}(x)$. Typical shapes of the pdf of the non-central χ^2 distribution are shown in Fig. 1.

(a) Dependency on the number of d.o.f. k.

(b) Dependency on the non-centrality λ.

Figure 1. Non-central χ^2 distributions and behavior of the mode.

When the number of degrees of freedom k is fixed, we can plot the dependency of the maximum of the pdf as a function of the non-centrality parameter λ, see Fig. 2.
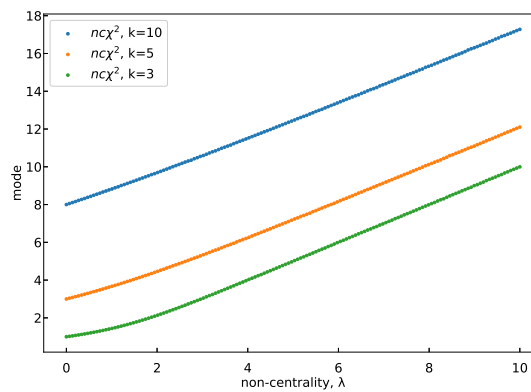

Figure 2. Mode of the non-central χ^2 as a function of the non-centrality parameter λ.

We observe that the bigger λ is the better the mode appears to be approximated with a straight line. The derivation of the line parameters together with the analysis of the edge cases of small number of degrees of freedom, where the mode does not exist, constitute the main results of the paper.

2. Derivation

2.1. **Master equation.** In this section we obtain the transcendental equation (Eq. 2.2) that determines the mode of the non-central χ^2 distribution. We reduce it to the ordinary differential equation (Eq. 2.5), where the non-centrality parameter λ is the argument, and the number of degrees of freedom k is a parameter. Finally, we solve the ODE approximately with a Taylor expansion (Eq. 2.10) and investigate edge cases of 1 and 2 degrees of freedom (Sec. 2.3).

We start by setting the derivative of the density of the non-central χ^2 (Eq. 2.1) to zero. This leads us to the transcendental equation (Eq. 2.2) that determines the mode of the distribution:

$$\frac{d}{dx}\chi^2_{k,\lambda}(x) = \frac{1}{2}\chi^2_{d,\lambda}(x) \cdot \left[-1 + \frac{k-2}{2x} + \sqrt{\frac{\lambda}{x}}\frac{I'_{\frac{k-2}{2}}(\sqrt{\lambda x})}{I_{\frac{k-2}{2}}(\sqrt{\lambda x})}\right] , \tag{2.1}$$

$$\frac{d}{dx}\chi^2_{k,\lambda}(x) = 0 \Rightarrow \sqrt{\lambda x}I'_{\frac{k-2}{2}}(\sqrt{\lambda x}) = (x - \frac{k-2}{2})I_{\frac{k-2}{2}}(\sqrt{\lambda x}) . \tag{2.2}$$

We can eliminate the derivative in Eq. 2.2 by using the differential equation for the modified Bessel function [1, Eq. 10.25.1]:

$$t^2\frac{d^2}{dt^2}I_\nu(t) + t\frac{d}{dt}I_\nu(t) - (t^2 + \nu^2)I_{d,\lambda}(t) = 0 . \tag{2.3}$$

To make use of Eq. 2.3, we need the expression for $I''_{\frac{k-2}{2}}$, therefore, we differentiate Eq. 2.2 by λ. Since the mode depends on the non-centrality λ, we should remember that $x = x(\lambda)$, thus $\frac{dx}{d\lambda} = x'$. The resulting expression for $I''_{\frac{k-2}{2}}$ is as follows:

$$\sqrt{\lambda x}I''_{\frac{k-2}{2}}(\sqrt{\lambda x}) = (x - \frac{k}{2})I'_{\frac{k-2}{2}}(\sqrt{\lambda x}) + \frac{2\sqrt{\lambda x}x'}{x + \lambda x'}I_{\frac{k-2}{2}}(\sqrt{\lambda x}) . \tag{2.4}$$

We substitute $I'_{\frac{k-2}{2}}$ (Eq. 2.2) and $I''_{\frac{k-2}{2}}$ (Eq. 2.4) into the differential equation for the modified Bessel function (Eq. 2.3). We then use the property [1, Eq. 10.29.4] to decrease the order of the derivatives of the modified Bessel functions. Assuming that the Bessel function itself is non-zero at the mode, we arrive to the following differential equation for the mode as a function of the non-centrality parameter λ:

$$\lambda x'(x - k - \lambda + 4) + x(x - k - \lambda + 2) = 0 . \tag{2.5}$$

2.2. **Approximate solution.** We observed that the linear approximation works better with growing λ, thus we introduce the asymptotic parameter $t = \frac{k}{\lambda} << 1$ to build the expansion. We expect the solution to be linear in λ, however the asymptotic expansion of $x(t) = C_0 + C_1 t + ...$ won't provide us with a solution linear in λ. Therefore, we reparametrize $x(t)$ with a new function $y(t) = tx(t)$:

$$t = \frac{k}{\lambda} , \tag{2.6}$$

$$y(t) = tx(t) . \tag{2.7}$$

We obtain the following equation after the reparametrization:

$$-(y't - y)(y - kt - k + 4t) + y(y - kt - k + 2t) = 0 . \tag{2.8}$$

To solve Eq. 2.8, we expand $y(t)$ into the Taylor series by the scale parameter $t = \frac{k}{\lambda}$. We would like to find the linear solution and one extra term that estimates the error. Thus, we cut the series at

the third power of t in order to account for the derivative. After solving algebraic equations for the coefficients near each power of t, we arrive to the resulting approximate expression for the mode:

$$y(t) = C_0 + C_1 t + C_2 t^2 + C_3 t^3 + O(t^4) \,, \tag{2.9}$$

$$C_0 = k, \quad C_1 = k - 3, \quad C_2 = \frac{k-3}{2k} \,, \tag{2.10}$$

$$\boxed{x_{mode} = \lambda + k - 3 + \frac{k-3}{2\lambda} + O\left(\frac{k^2}{\lambda^2}\right)} \,. \tag{2.11}$$

We plot the linear approximation Eq. 2.11 together with the precise numerical solution Fig. 2 in order to verify the approximation is correct, see Fig. 3.
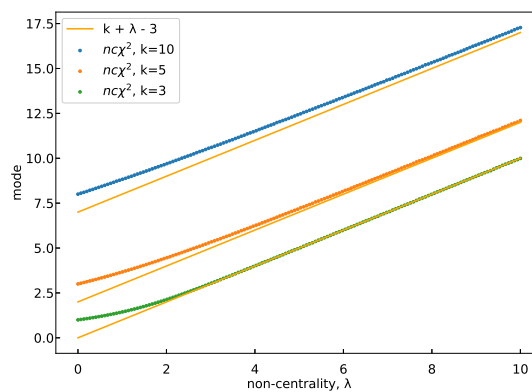

Figure 3. Linear approximation to the mode of the non-central χ^2 compared to the more precise numerical solution as a function of the non-centrality parameter λ.

2.3. **Small number of degrees of freedom.**

2.3.1. *Case $k < 2$.* The asymptotic behavior of the modified Bessel function at $x \to 0$ [1, Eq. 10.30.1] shows that the pdf of the non-central χ^2 diverges, thus it doesn't have a mode:

$$\chi^2_{k,\lambda}(x) \to \frac{1}{2\Gamma(\frac{k}{2})} \frac{1}{(2\lambda)^{\frac{k-2}{2}}} e^{-\frac{\lambda}{2}} \left(\sqrt{\lambda x}\right)^{k-2}, \quad x \to 0 \,. \tag{2.12}$$

2.3.2. *Case $k = 2$.* In this case, the pdf at $x = 0$ is finite. If the derivative at $x = 0$ is positive, then the maximum is not there. The expression for the derivative (Eq. 2.13) and its asymptotic behavior at $x \to 0$ (Eq. 2.14) are shown below:

$$\frac{d}{dx}\chi^2_{k,\lambda}(x) = \frac{1}{2}\chi^2_{d,\lambda}(x) \cdot \left[-1 + \sqrt{\frac{\lambda}{x}}\frac{I_{-1}(\sqrt{\lambda x})}{I_0(\sqrt{\lambda x})}\right] \,, \tag{2.13}$$

$$\frac{d}{dx}\chi^2_{k,\lambda}(x) \to \frac{1}{2}\chi^2_{d,\lambda}(x) \cdot \left[-1 + \frac{\lambda}{2}\right], \quad x \to 0 \,. \tag{2.14}$$

We observe that when $\lambda > 2$, the pdf of the non-central χ^2 doesn't have its maximum at $x = 0$. In the region $\lambda < 2$, the asymptotic scale $t = \frac{k}{\lambda} > 1$, hence our approximation is inapplicable in this region and we refrain from analysing it.

3. Application

There exist a number of numerical procedures for finding the mode of a distribution [2, Ch. 10]. Some of them require the search region to be specified.

For example, the widely used C++ library Boost [3] identifies the search region based on an initial guess for the mode x_0. Boost iteratively checks regions of the form $[x_0/2, 2x_0]$, $[x_0/2^2, 2^2 x_0]$, etc. When the value of the pdf at both ends of the region becomes smaller than the value at the initial guess point x_0, the algorithm initiates the search for the maximum inside of the region.

At the time of writing, Boost used $x_0 = k + 1$ as the initial guess. We already know, based on the approximate solution (Eq. 2.11), that the chosen guess will undershoot at large non-centrality values λ. Let's estimate λ above which the method will require the second iteration for the region to cover the mode. For this we compare the linear estimate for the location of the mode (Eq. 2.11) to the initial guess x_0 used by Boost:

$$k + \lambda - 3 > 2 \cdot (k + 1) ,\tag{3.1}$$

$$\lambda > k + 5 .\tag{3.2}$$

With Eq. 3.2, for any number of d.o.f. k we are able to specify the threshold α, defined by $\lambda = \alpha k$, at which the original initial guess starts undershooting:

$$\alpha > \frac{5}{k} + 1 .\tag{3.3}$$

We see that large k corresponds to small thresholds α. The most conservative estimate for the threshold would be at the smallest k possible: $k = 2$. Thus, $\alpha = 3.5$ is the threshold that approximately works for $k = 2$ and is the overestimated threshold for bigger values of k.

The threshold α (Eq. 3.3) is closely related to the asymptotic scale t (Eq. 2.6) that we used for finding the approximate solution, specifically: $t = \frac{k}{\lambda} = \frac{1}{\alpha}$. For example, the conservative threshold $\alpha \approx 3.5$ corresponds to the asymptotic scale $t \approx 0.25 < 1$. It means that the region where the original guess of Boost undershoots, is, at the same time, the region where our approximate solution for the mode becomes applicable and can be used as a corrected initial guess. However, the fact that we use the conservative threshold may lead to the situation where the original method has already started undershooting but λ is not yet big enough to turn on the corrected regime.

3.1. **Dependency on λ.** We fix the threshold to the conservative value $\frac{k}{\lambda} = 0.25$. We then plot the dependency of the run time on the non-centrality λ for a set of d.o.f. k: 2, 15, 50, see Fig. 4. For benchmarking we use the Google benchmark library [4]. The benchmarking script itself became a part of the Boost.Math [5]. Using this script we measure the run time 100 times and use the mean as a central value. The error bar is computed as a standard deviation. We add noise with standard deviation $\sigma = 10^{-6}$ to parameters k and λ to avoid caching effects. The vertical line on the plots shows the threshold where the original initial guess for smaller λ is switched to the corrected value at

bigger λ. Therefore, we expect that both lines coincide below the threshold and the improved solution would lie lower above the threshold. One can notice missing values on the curve representing the original initial guess. The reason for this is the numerical instability of the algorithm in Boost, that has been resolved after we corrected the initial guess.
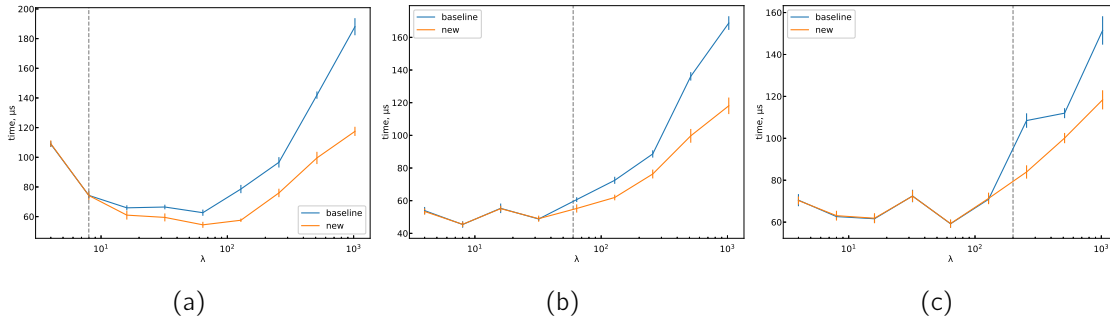

(a)   (b)   (c)

Figure 4. Run time as a function of the non-centrality λ for d.o.f. $k = 2$ (4a), $k = 15$ (4b), $k = 50$ (4c). Vertical line shows the threshold at which the corrected expression replaces the original initial guess.

3.2. **Dependency on d.o.f. (k).** In the set of plots in Fig. 5, we fix the asymptotic scale to values $\frac{k}{\lambda} = 0.25, 0.15, 0.05$ and investigate the dependency of the run time on the number of d.o.f. Since the threshold is fixed, the difference in the run time is caused by the actual position where the original initial guess starts to undershoot, the non-conservative threshold. The farther the fixed threshold is from the non-conservative threshold, the more significant the effect of undershooting at the test point will be. Therefore, we expect the difference in the run time to grow with number of d.o.f, as follows from Eq. 3.3. For each value of the asymptotic scale, in addition to the full plot, we also show a zoomed version that shows the region where both original and improved methods were able to converge (Fig. 5).

4. Conclusion

In this paper we present an approximate expression for the mode of the non-central χ^2 distribution: $x_{mode} \approx k + \lambda - 3$, where k is the number of degrees of freedom and λ is the non-centrality parameter. The approximation is based on an asymptotic expansion and is valid in the region where the scale parameter $\frac{k}{\lambda} << 1$ and where the mode exists $k > 2$. The approximate formula can be used as the initial guess for iterative procedures searching for a precise solution. Run time performance and the domain of applicability of the Boost implementation of the mode search was improved using the presented approximate expression. The improvement became a part of the Boost.Math library [5].
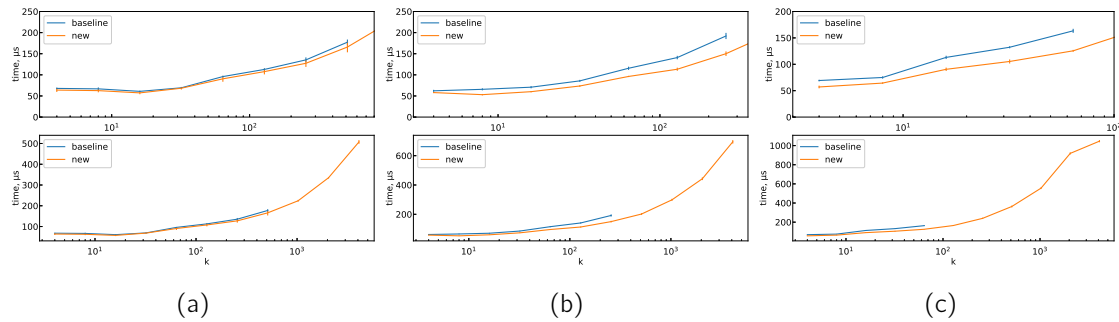
Figure 5. Run time as a function of the number of d.o.f. (k) for the asymptotic scale values $\frac{k}{\lambda} = 0.25$ (5a), $\frac{k}{\lambda} = 0.15$ (5b), $\frac{k}{\lambda} = 0.05$ (5c). The upper plot in each pair shows the zoomed version, focused on the region where both original and improved methods were able to converge.

5. Acknowledgements

References

[1] F.W.J. Olver et al., NIST Digital Library of Mathematical Functions, (2021). http://dlmf.nist.gov/.

[2] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes 3rd Edition, (2007). http://numerical.recipes/.

[3] Boost C++ Libraries, v1.76.0, (2021). https://www.boost.org/.

[4] Google benchmark, (2021). https://github.com/google/benchmark.

[5] Boost Math, (2021). https://github.com/boostorg/math/pull/645.

[6] S. András, Á. Baricz, Properties of the Probability Density Function of the Non-Central Chi-Squared Distribution, J. Math. Anal. Appl. 346 (2008), 395–402. https://doi.org/10.1016/j.jmaa.2008.05.074.

[7] D. Horgan, C.C. Murphy, On the Convergence of the Chi Square and Noncentral Chi Square Distributions to the Normal Distribution, IEEE Commun. Lett. 17 (2013), 2233–2236. https://doi.org/10.1109/LCOMM.2013.111113.131879.

[8] L. Saulis, Asymptotic Expansion for the Distribution and Density Functions of the Quadratic Form of a Stationary Gaussian Process in the Large Deviation Cramer Zone, Nonlinear Anal.: Model. Control. 6 (2001), 87–101. https://doi.org/10.15388/NA.2001.6.1.15218.

[9] S.S. Sawant, D.A. Levin, V. Theofilis, Analytical Prediction of Low-Frequency Fluctuations Inside a One-Dimensional Shock, arXiv:physics.flu-dyn (2021). https://arxiv.org/abs/2101.00664.

[10] V. Pereyra, Iterated Deferred Corrections for Nonlinear Operator Equations, Numer. Math. 10 (1967), 316–323. https://doi.org/10.1007/BF02162030.

Appendix

A. Likelihood-ratio test statistic for quadratic log-likelihood with 1 signal d.o.f.

The derivation starts by defining the log-likelihood for both signal, H_0, and background, H_1, hypotheses.

$$-2\log\mathcal{L}_0 = g + \langle j|\theta\rangle + \langle\theta|\,\hat{h}\,|\theta\rangle\,, \tag{A.1}$$
$$-2\log\mathcal{L}_1 = g + \langle j, s'|\theta,\mu\rangle + \langle\theta,\mu|\,\hat{H}\,|\theta,\mu\rangle\,,$$

where g is a scalar, $|j\rangle$ is the first derivative and $\hat{h}$ is a hessian of the background only log-likelihood $-2\mathcal{L}_0$ with respect to θ. They are considered constant parameters of the model but may depend on the data.

Vector $|\theta,\mu\rangle$ is the concatenation of the vector θ and the one-dimensional vector containing μ. The same holds for $|j, s'\rangle$.

Additionally, s' and $\hat{H}$ are parameters of the signal+background log-likelihood $-2\mathcal{L}_1$, with s' representing its gardient with respect to the nuisance parameter μ only present under H_1. The structure of $\hat{H}$, however, should be described in more detail:

$$\hat{H} = \begin{bmatrix} \hat{h} & \mathbf{q} \\ \mathbf{q} & s'' \end{bmatrix}\,, \tag{A.2}$$

where $\hat{h}$ is the same background-only hessian, the second order mixed derivative of the $-2\mathcal{L}_1$ with respect to $|\theta\rangle$ and μ is a vector of $\mathbf{q} = |q\rangle$, and the second derivative with respect to μ is s''.

Log-likelihoods can be rewritten, taking the detailed structure of $\hat{H}$ into account:

$$-2\log\mathcal{L}_0 = g + \langle j|\theta\rangle + \langle\theta|\,\hat{h}\,|\theta\rangle\,, \tag{A.3}$$
$$-2\log\mathcal{L}_1 = g + \langle j|\theta\rangle + \langle\theta|\,\hat{h}\,|\theta\rangle + \mu s' + 2\mu\langle q|\theta\rangle + \mu^2 s''\,.$$

It is assumed that $\mathcal{L}_0$ is in its maximum when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and the maximum of $\mathcal{L}_1$ is in $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \boldsymbol{\theta}_1$ constructed as a deviation from $\boldsymbol{\theta}_0$. The following equations

on the derivatives of the log-likelihoods then take place:

$$
\begin{cases}
-2\dfrac{\partial \mathcal{L}_0}{\partial \boldsymbol{\theta}_0} = |j\rangle + 2\hat{h}\,|\theta_0\rangle = 0, \\[2ex]
-2\dfrac{\partial \mathcal{L}_1}{\partial \boldsymbol{\theta}_1} = |j\rangle + 2\hat{h}\,|\theta_0 + \theta_1\rangle + 2\mu\,|q\rangle = 0, \\[2ex]
-2\dfrac{\partial \mathcal{L}_1}{\partial \mu} = 2\,\langle q|\theta_0 + \theta_1\rangle + s' + 2s''\mu = 0.
\end{cases}
\tag{A.4}
$$

Firstly, Equation A.4 is solved for $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ using the first two equations in the system:

$$
|\theta_0\rangle = -\frac{1}{2}\hat{h}^{-1}\,|j\rangle,
\tag{A.5}
$$
$$
|\theta_1\rangle = -\frac{1}{2}\mu\hat{h}^{-1}\,|q\rangle.
$$

Then this solution (Equation A.5) is substituted into the third equation in A.4 and is solved for μ:

$$
\mu = \frac{1}{2}\frac{\langle q|\,\hat{h}^{-1}\,|j\rangle - s'}{s'' - \langle q|\,\hat{h}^{-1}\,|q\rangle}.
\tag{A.6}
$$

Finally, the solutions A.5 and A.6 are used to compute the test statistic t:

$$
t = (-2\log \mathcal{L}_0) - (-2\log \mathcal{L}_1) = \mu^2(s'' - \langle q|\,\hat{h}^{-1}\,|q\rangle)
\tag{A.7}
$$
$$
= \frac{1}{4}\frac{\left(\langle q|\,\hat{h}^{-1}\,|j\rangle - s'\right)^2}{s'' - \langle q|\,\hat{h}^{-1}\,|q\rangle}.
$$

The signal model brings only 1 extra degree of freedom (μ), which means the test statistic t, in the asymptotic regime, follows the χ^2 distribution with 1 degree of freedom. The significance, therefore, can be derived from t:

$$
\boxed{Z = \sqrt{t} = \frac{1}{2}\frac{\langle q|\,\hat{h}^{-1}\,|j\rangle - s'}{\sqrt{s'' - \langle q|\,\hat{h}^{-1}\,|q\rangle}}}\quad.
\tag{A.8}
$$

B. Open source contributions

In the course of my research and academic journey, I was actively contributing to open source software, leveraging my expertise to make meaningful contributions to various statistical methods in physics. While some projects were developed internally, others resulted from collaborations with various institutions. Below I only list significant open source contributions to statistical software.

Custom form factors in Sherpa. Sherpa is a Monte Carlo event generator for the simulation of high-energy reactions of particles [28]. It is multipurpose, which means it can be applied to different models of physics including those beyond the Standard Model of particle physics. Simulation in Sherpa starts from the definition of a model, which includes listing particles, their properties and properties of interactions between these particles.

Sherpa developed a language to define these parameters, but also can ingest Universal FeynRules Output (UFO) [41], model definitions compatible with other software packages like MadGraph [42]. Since support for external definitions is an additional feature, these capabilities may be incomplete.

In the beginning of my journey as a PhD student, we investigated an opportunity to measure the CP properties of the Higgs boson by analysing the electron-positron pairs created when both of the photons from the $H \to \gamma\gamma$ decay are converted by interactions with the silicon layers of the ATLAS inner detector (or the future ITk all-silicon tracker). Such events are expected to have little background, with the additional advantage of the spin-0 silicon nucleus allowing the spin properties of the converted pairs to be analysed undisturbed. The rate of production of the electron-positron pairs can be described with the Bethe-Heitler process [43], with possible CP-sensitive observables discussed by Bishara et al [44].

The interaction of a converting photon with a silicon nucleus can be summarized with a form factor, which is not constant, but depends non-trivially on the momentum of the silicon atom [44]:

$$G_2^{\text{el}}(q^2) = \frac{M^2 a^4 q^4}{(1 - a^2 q^2)^2},$$ (B.1)

where q is the 4-momentum transfer between the nucleus with mass M and the converting photon, and a is an effective scale.

With my contribution [39] I made Sherpa custom form factors for MadGraph models even more flexible, which also allowed me to explore the possibility of measuring the CP of the Higgs boson with converted photons in $H \to \gamma\gamma$ decays.
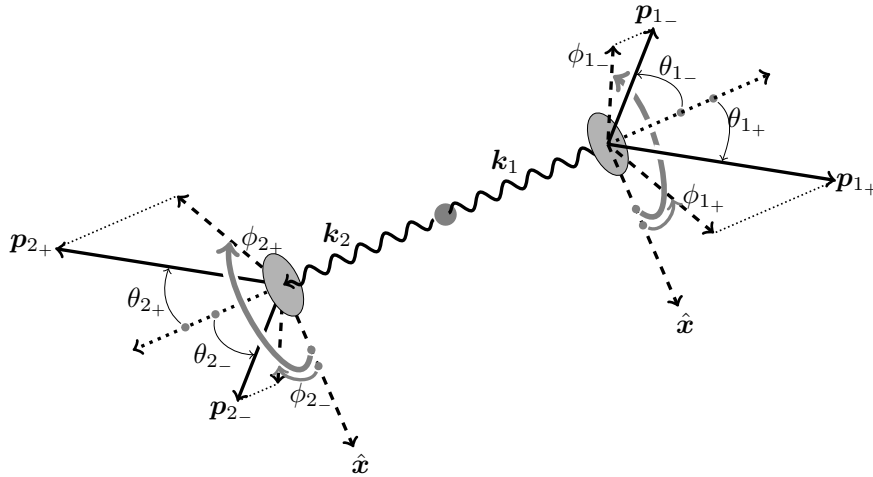
Figure B.1: Conversion of a photon pair into two electron-positron pairs in matter. (Taken from Ref. [44].)

Omnibus tests for survival analysis. Statsmodels [45] is a Python package for statistical computations that provides a complement to the well known SciPy [46], and includes tools for descriptive statistics, estimation, and inference for statistical models. To emphasize the importance of the Statsmodels package, it is worth mentioning that it has $> 9K$ stars on GitHub, which is comparable to $\approx 12K$ stars[1] that SciPy has.

Many solutions in Statsmodels are based on existing packages in the R ecosystem. The functionality of one such R package, the SurvDiff module of the Survival package [47], is used for survival analysis generally applied to clinical trials data.

At the statistics course I was attending during my PhD curriculum, there was a homework assignment involving this R package, which I chose to do in Python until I discovered that the functionality of SurvDiff in Python was limited. The test provided by Statsmodels at that time supported only 2 different groups in a trial. During the work on my assignment, I enabled support for omnibus tests [40] allowing any number of groups in the data for the SurvDiff implementation in Statsmodels.

Mode of the non-central chi-squared distribution. Independently from the Gaussian process study, we encountered a beautiful asymptotic behavior of the non-central chi-squared distribution. This distribution also appears in statistical tests in physics, when data is sampled from a distribution that is incompatible with the null-hypothesis [23].

The non-central χ^2 distribution follows the sum $X_1^2 + X_2^2 + \cdots + X_k^2$ of normally distributed random variables $X_1, X_2, \ldots, X_k$ with unit variance and means $\mu_1, \mu_2, \ldots, \mu_k$. The distribution is said to have k degrees of freedom and

---

[1]On GitHub, "stars" are a way to bookmark and express appreciation for interesting repositories. Starring a repository helps users track and revisit projects, while also contributing to the repository's recognition.

non-centrality $\lambda = \mu_1^2 + \mu_2^2 + \cdots + \mu_k^2$. The probability density function of this distribution has a closed form expression:

$$f_{k,\lambda}(x) = \frac{1}{2} \exp^{-\frac{x+\lambda}{2}} \left(\frac{x}{\lambda}\right)^{\frac{k-2}{4}} I_{\frac{k-2}{2}}(\sqrt{\lambda x}) , \qquad \text{(B.2)}$$

where $I_\nu(x)$ is a modified Bessel function of the first kind.

The way the number of degrees of freedom and the non-centrality parameter affect the shape of the distribution is illustrated in Figure B.2.
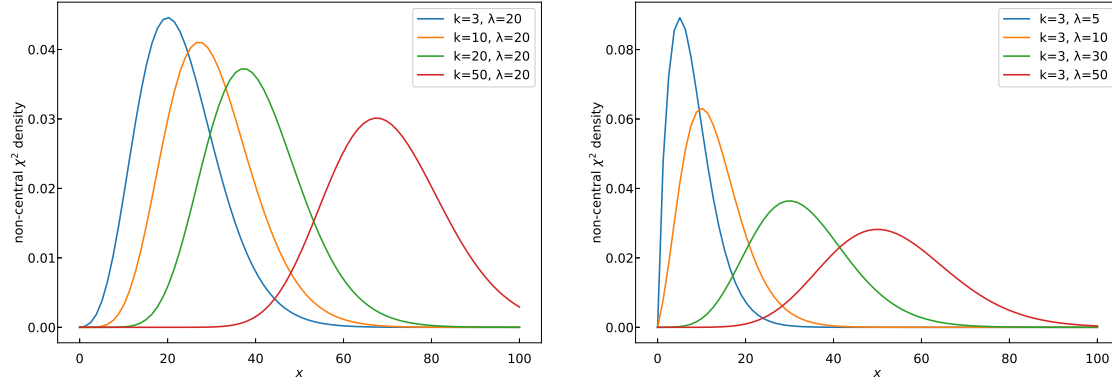
(a) Dependency on the number of d.o.f. k for non-centrality $\lambda = 20$.

(b) Dependency on the non-centrality λ for number of d.o.f. $k = 3$.

Figure B.2: Non-central χ^2 distributions for various values of non-centrality λ and degrees of freedom k. (Taken from Paper IV.)

The result of Paper IV is an estimate of the mode of the non-central χ^2 distribution with number of degrees of freedom k and non-centrality λ:

$$x_{mode} = \lambda + k - 3 + \frac{k-3}{2\lambda} + O\left(\frac{k^2}{\lambda^2}\right). \qquad \text{(B.3)}$$

In Figure B.3 the more precisely numerically estimated mode is compared to the estimate of Equation B.3. The more precise calculation was conducted with the iterative optimization routine implemented in Boost [29], which is a suit of free peer-reviewed portable C++ source libraries.

I used this result to speed up the calculation and even widen the domain of applicability of the algorithm used in Boost. The Boost.Math module of the library implements a solution for the estimation of the mode of the non-central χ^2 distribution, and my contribution became a part of it [38].

Cicliminds with CICERO. One of the big challenges in the field of climate research is to get a good qualitative overview of climate models, and also be able to dig deeper when needed. The Expert Team on Climate Change Detection and Indices (ETCCDI) introduced a set of summary statistics of climate data, so-called ETCCDI indices, which were extensively studied by Sillmann et al. [48] to provide a quantitative summary of climate change as well as extreme events in the climate data.
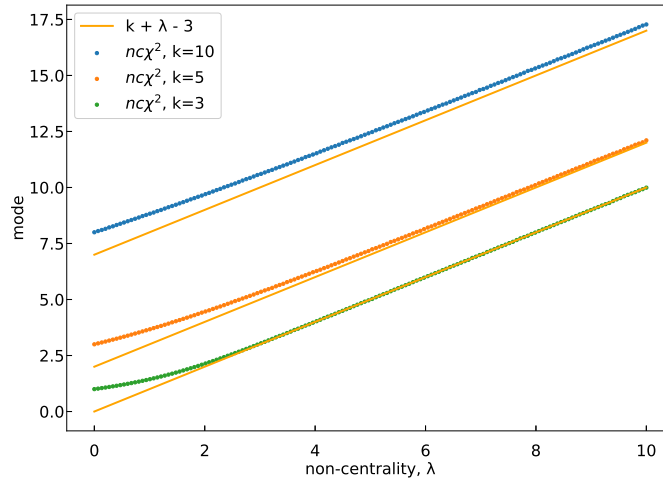
Figure B.3: Linear approximation to the mode of the non-central χ^2 compared to the more precise numerical solution as a function of the non-centrality parameter λ. We plot the simplified approximation, linear in λ and k, because specifically this form was used in Boost C++ [38]. (Taken from Paper IV.)

Many climate models have been developed and used to simulate the change of the climate, both for historical time ranges and extrapolating to the future. Extrapolations are made based on sets of assumptions called *scenarios*, and can be simulated several times with different random seeds, which results in an *ensemble* of simulations. When combined together with the number of observable indices of interest, the number of datasets may reach thousands, which requires a proper way to navigate all of them.

In collaboration with the Center for International Climate Research (CICERO) I developed Cicliminds [30], a tool for quick and comprehensive insight into the ETCCDI climate index datasets.

The graphical interface of the tool consists of two main control panels used for filtering the data (Figure B.4a) and for configuring the visualization (Figure B.4b). The joint dataset is then used to build various visualizations predefined in the tool, some examples of which we demonstrate in Figure B.5.

The example visualizations are based on one of the indices taken as a benchmark, and were presented in the package documentation [30]. In Figure B.5a the histograms of measurements of this index were averaged across the globe. Each curve represents a summary across the models and the range of years, with the aggregation parameters selected in the interface shown in Figure B.4b.

A comparison of the future scenarios is shown in Figure B.5b. The baseline value, estimated from the reference window, was, optionally, subtracted from each of the time series.
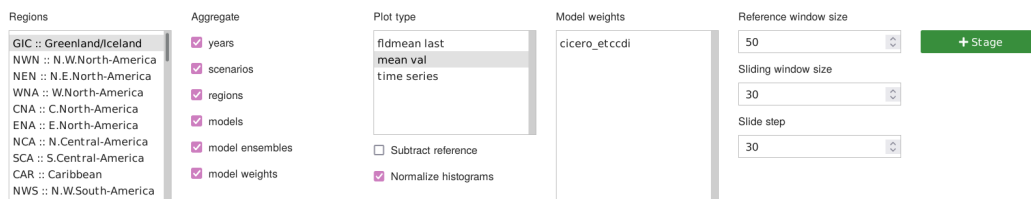
Figure B.5c demonstrates the distribution of the index values across the globe, aggregated across the rest of the dimensions.

This is only a subset of the functionality, and there is a large potential in expanding the tool with more visualization types and features, which Cicliminds is well designed for.
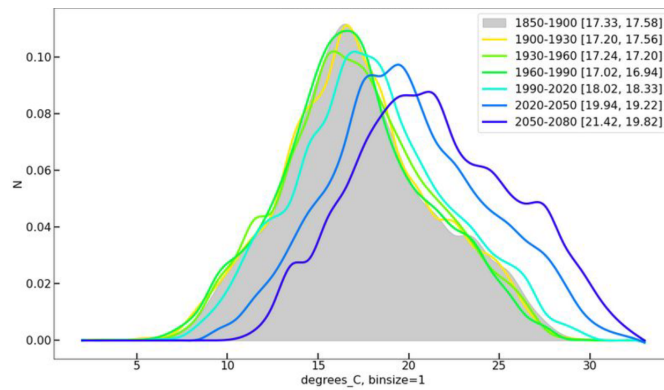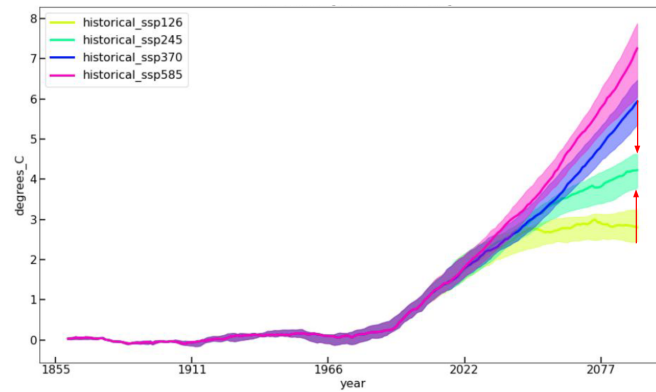
Figure B.4: Graphical user interface of Cicliminds: dataset filters (a) and visualization control (b). (Taken from the documentation for the Cicliminds package available on GitHub [30].)
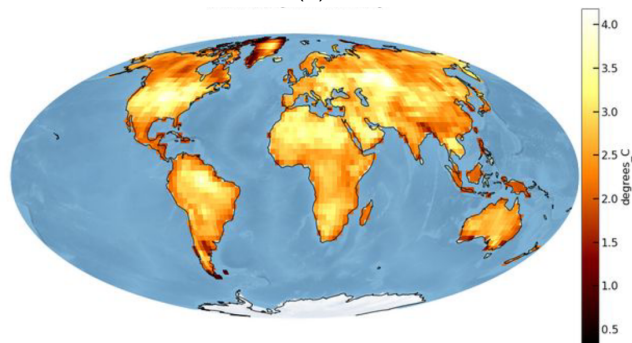
(a)

(b)

(c)

Figure B.5: Examples of the visualizations provided by Cicliminds. The plotted index ($tnxETCCDI$) defines the annual maximum of the daily minumum temperature aggregated across the models selected in Figure B.4. In (a), an example of histogram evolution is shown. The functionality of the time series comparison is demonstrated in (b), which also visualizes the variability of each series. In (c) we show a heatmap of the measurements of the index, aggregated across the rest of dimensions. (Taken from the documentation for the Cicliminds package available on GitHub [30].)

MGVI.jl with Max-Planck Institute for Physics. Metric Gaussian Variational Inference (MGVI) is an iterative method developed by Knollmüller et al [49] that performs a series of Gaussian approximations to the posterior distribution. In collaboration with the Max-Planck Institute for Physics I developed the first version of the Julia [50] implementation of MGVI, called MGVI.jl [31].

The new package is fresh but powerful. It combines the efficient MGVI procedure for estimating posterior distributions with the benefits of Julia's highly performant ecosystem of scientific software.

To demonstrate the functionality of the package in Figure B.6, which was also provided in the documentation for the package [31], the coal mining disaster dataset [51] was chosen. The histogram of the disaster counts is assumed to be Poisson-distributed, with the Poisson rates across bins being modeled as a continuous curve. The curve is derived from a realization of a Gaussian process with zero mean and a finite width covariance kernel, selected with the help of the MGVI optimization procedure. In addition to the parameters selecting the particular realization of the Gaussian process, the parameters of the GP kernel itself are also considered free.

The model described in the previous paragraph was inspired by the recurrent problem of smoothing sparse data distributions in High Energy Physics. This class of problems is particularly important for modelling background distributions, and we believe the MGVI method is worth further exploration in this regard.
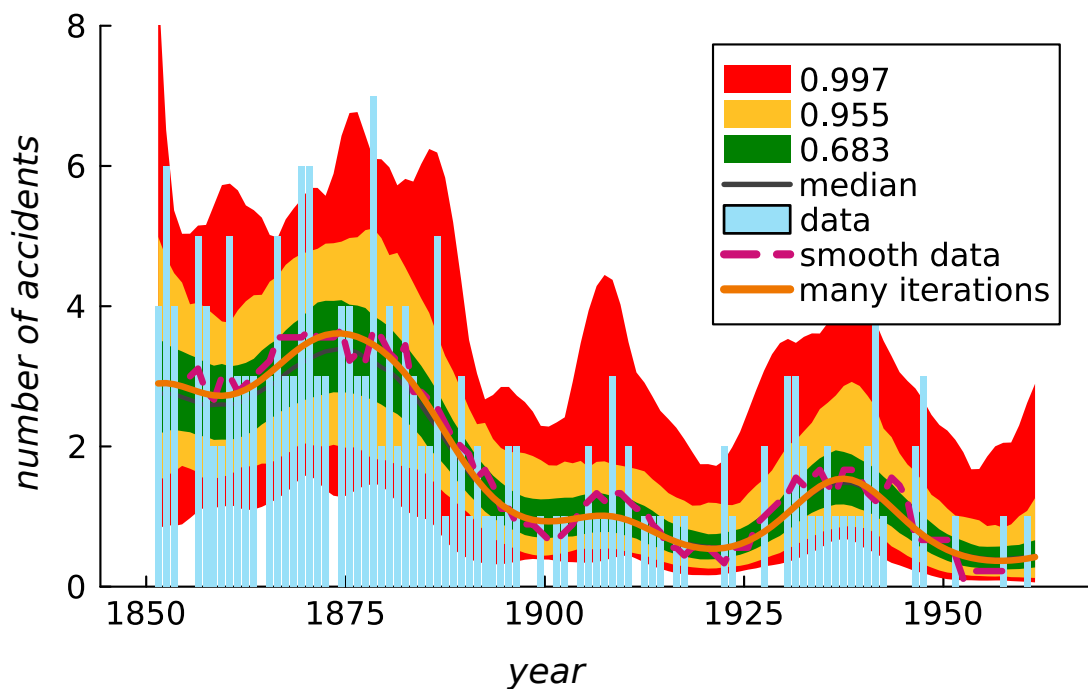

Figure B.6: Result of the posterior fit (orange) of a Poisson distribution to the data (blue), with its rate density modeled with a Gaussian process. The colored regions are spreads of the samples from the fitted model, selected from 1σ (green), 2σ (yellow) and 3σ (red) credibility intervals. (Taken from the documentation for the MGVI.jl package available on GitHub [31].)

Appendix B. Open source contributions

Bibliography

[1] *International Training Network of Statistics for High Energy Physics and Society, INSIGHTS*. URL: https://insights-itn.eu/.

[2] Wikimedia Commons. *Standard Model of Elementary Particles, (accessed on 18.11.2023)*. 2023. URL: https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg.

[3] UA1 Collaboration. "Experimental observation of lepton pairs of invariant mass around 95 GeV/c2 at the CERN SPS collider". In: *Physics Letters B* 126.5 (July 1983), pp. 398–410. ISSN: 0370-2693. DOI: 10.1016/0370-2693(83)90188-0. URL: http://dx.doi.org/10.1016/0370-2693(83)90188-0.

[4] UA2 Collaboration. "Evidence for Z0 -> e+e- at the CERN p collider". In: *Physics Letters B* 129.1–2 (Sept. 1983), pp. 130–140. ISSN: 0370-2693. DOI: 10.1016/0370-2693(83)90744-x. URL: http://dx.doi.org/10.1016/0370-2693(83)90744-X.

[5] Fermilab. "Observation of Top Quark Production in $\bar{p}p$ Collisions with the Collider Detector at Fermilab". In: *Physical Review Letters* 74.14 (Apr. 1995), pp. 2626–2631. ISSN: 1079-7114. DOI: 10.1103/physrevlett.74.2626. URL: http://dx.doi.org/10.1103/PhysRevLett.74.2626.

[6] DØ Collaboration. "Observation of the Top Quark". In: *Physical Review Letters* 74.14 (Apr. 1995), pp. 2632–2637. ISSN: 1079-7114. DOI: 10.1103/physrevlett.74.2632. URL: http://dx.doi.org/10.1103/PhysRevLett.74.2632.

[7] K. Kodama et al. "Observation of tau neutrino interactions". In: *Physics Letters B* 504.3 (Apr. 2001), pp. 218–224. ISSN: 0370-2693. DOI: 10.1016/s0370-2693(01)00307-0. URL: http://dx.doi.org/10.1016/S0370-2693(01)00307-0.

[8] W Buchmüller and C Lüdeling. "Field Theory and Standard Model". In: (2006). DOI: 10.5170/CERN-2006-014.1. URL: https://cds.cern.ch/record/984122.

[9] Gerardus 't Hooft and Martinus J G Veltman. *Diagrammar*. CERN Yellow Reports: Monographs. Geneva: CERN, 1973. DOI: 10.5170/CERN-1973-009. URL: https://cds.cern.ch/record/186259.

[10] Peter W Higgs. "Broken symmetries and the masses of gauge bosons". In: *Physical review letters* 13.16 (1964), p. 508.

[11] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". In: *Phys. Rev. Lett.* 13 (9 Aug. 1964), pp. 321–323. DOI: 10.1103/PhysRevLett.13.321. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.321.

[12] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. "Global Conservation Laws and Massless Particles". In: *Phys. Rev. Lett.* 13 (20 Nov. 1964), pp. 585–587. DOI: 10.1103/PhysRevLett.13.585. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.585.

[13] Jeffrey Goldstone, Abdus Salam, and Steven Weinberg. "Broken symmetries". In: *Physical Review* 127.3 (1962), p. 965.

[14] Forthommel. *CERN Accelerator Complex, (accessed on 19.11.2023).* CC BY-SA 3.0 https://creativecommons.org/licenses/by-sa/3.0, via Wikimedia Commons. 2021. URL: https://commons.wikimedia.org/wiki/File:Cern-accelerator-complex.svg.

[15] CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physics Letters B* 716.1 (Sept. 2012), pp. 30–61. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.021. URL: http://dx.doi.org/10.1016/j.physletb.2012.08.021.

[16] The ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* 716.1 (Sept. 2012), pp. 1–29. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.020. URL: http://dx.doi.org/10.1016/j.physletb.2012.08.020.

[17] CERN Collaboration. "Search for the Standard Model Higgs Boson at LEP. Large Electron Positron". In: *Phys. Lett. B* 565 (2003). 22 pages, 10 figures Report-no: CERN-EP/2003-011, pp. 61–75. DOI: 10.1016/S0370-2693(03)00614-2. URL: https://cds.cern.ch/record/610122.

[18] Giovanni Ridolfi. "Search for the Higgs boson: theoretical perspectives". In: *Frascati Phys. Ser.* 22 (2001), pp. 291–304. URL: https://cds.cern.ch/record/506136.

[19] The ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08003–S08003. ISSN: 1748-0221. DOI: 10.1088/1748-0221/3/08/s08003. URL: http://dx.doi.org/10.1088/1748-0221/3/08/S08003.

[20] LHC Higgs Cross Section Working Group CERN. *Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables.* 2011. DOI: 10.5170/CERN-2011-002. URL: http://cds.cern.ch/record/1318996.

[21] CERN. *Run 3 of the Large Hadron Collider.* https://home.cern/press/2022/run-3. 2022.

[22] S. S. Wilks. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". In: *The Annals of Mathematical Statistics* 9.1 (1938), pp. 60–62. ISSN: 00034851. DOI: http://dx.doi.org/10.1214/aoms/1177732360. URL: https://www.jstor.org/stable/2957648.

[23] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *The European Physical Journal C* 71.2 (Feb. 2011). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-011-1554-0.

[24] Eilam Gross and Ofer Vitells. "Trial factors for the look elsewhere effect in high energy physics". In: *The European Physical Journal C* 70.1-2 (Oct. 2010), pp. 525–530. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-010-1470-8.

[25] Ofer Vitells and Eilam Gross. "Estimating the significance of a signal in a multi-dimensional search". In: *Astroparticle Physics* 35.5 (Dec. 2011), pp. 230–234. ISSN: 0927-6505. DOI: 10.1016/j.astropartphys.2011.08.005.

[26] R. B. Davies. "Hypothesis testing when a nuisance parameter is present only under the alternative". In: *Biometrika* 64.2 (Aug. 1977), pp. 247–254. ISSN: 0006-3444. DOI: 10.1093/biomet/64.2.247. URL: https://doi.org/10.1093/biomet/64.2.247.

[27] Victor Ananyev and Alex Read. *Datasets of fitted toy Monte Carlo samples for "Gaussian Process-based calculation of look-elsewhere trials factor": https://doi.org/10.5281/zenodo.7861345.* URL: https://doi.org/10.5281/zenodo.7861345.

[28] Enrico Bothmann et al. "Event generation with Sherpa 2.2". In: *SciPost Phys.* 7 (2019), p. 034. DOI: 10.21468/SciPostPhys.7.3.034. URL: https://scipost.org/10.21468/SciPostPhys.7.3.034.

[29] *Boost C++ Libraries, v1.76.0.* 2021.

[30] Victor Ananyev and CICERO. *Cicliminds: interactive visualization of ETCCDI indices.* https://github.com/vindex10/cicliminds. 2022.

[31] Victor Ananyev et al. *MGVI.jl.* https://github.com/bat/MGVI.jl. 2021.

[32] "PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics". In: *CERN CDS.* CERN Yellow Reports: Conference Proceedings (June 2008). Ed. by Louis Lyons, Harrison Prosper, and Albert de Roeck. Report number: CERN-2008-001, p. 219. DOI: 10.5170/CERN-2008-001. URL: https://dx.doi.org/10.5170/CERN-2008-001.

[33] "Look-elsewhere effect with Gaussian Process". In: *HDBS Workshop.* (Internal) https://indico.cern.ch/event/1132691/contributions/4994710/. ATLAS Collaboration, CERN. Uppsala University, Sweden, 2022.

[34] V. Ananiev and A. L. Read. "SigCorr: A Python package for studies of trials factors". In: *Journal of Open Source Software* 8.87 (2023), p. 4989. DOI: 10.21105/joss.04989. URL: https://doi.org/10.21105/joss.04989.

[35] *SciKit-HEP Project.* URL: https://github.com/scikit-hep/.

[36] Lukas Heinrich, Matthew Feickert, and Giordon Stark. *pyhf.* DOI: 10.5281/zenodo.1169739. URL: https://github.com/scikit-hep/pyhf/.

[37] SciKit-HEP Project. *hepstats.* URL: https://github.com/scikit-hep/hepstats/.

[38] Victor Ananyev. *Improve mode of the non-central chi-squared algorithm. Boost C++ Libraries.* https://github.com/boostorg/math/pull/645. 2021.

[39]   Victor Ananyev. *UFO model improvements. Sherpa.* https://gitlab.com/sherpa-team/sherpa/-/merge_requests?scope=all&state=merged&author_username=vindex10. 2020.

[40]   Victor Ananyev. *SurvDiff improvements to StatsModels.* https://github.com/statsmodels/statsmodels/pull/6626. 2020.

[41]   Céline Degrande et al. "UFO – The Universal FeynRules Output". In: *Computer Physics Communications* 183.6 (June 2012), pp. 1201–1214. DOI: 10.1016/j.cpc.2012.01.022. URL: https://doi.org/10.1016/j.cpc.2012.01.022.

[42]   J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: *JHEP* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].

[43]   "On the stopping of fast particles and on the creation of positive electrons". In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 146.856 (Aug. 1934), pp. 83–112. ISSN: 2053-9150. DOI: 10.1098/rspa.1934.0140. URL: http://dx.doi.org/10.1098/rspa.1934.0140.

[44]   Fady Bishara et al. "Probing CP violation in h → yy with converted photons". In: *Journal of High Energy Physics* 2014.4 (Apr. 2014). ISSN: 1029-8479. DOI: 10.1007/jhep04(2014)084. URL: http://dx.doi.org/10.1007/JHEP04(2014)084.

[45]   Skipper Seabold and Josef Perktold. *statsmodels: Econometric and statistical modeling with python.* https://github.com/statsmodels/statsmodels/. 2010.

[46]   Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[47]   Terry M Therneau. *A Package for Survival Analysis in R.* R package version 3.2-7. 2020. URL: https://CRAN.R-project.org/package=survival.

[48]   J. Sillmann et al. "Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate". In: *Journal of Geophysical Research: Atmospheres* 118.4 (Feb. 2013), pp. 1716–1733. DOI: 10.1002/jgrd.50203. URL: https://doi.org/10.1002/jgrd.50203.

[49]   Jakob Knollmüller and Torsten A. Enßlin. "Metric Gaussian Variational Inference". In: (2020). arXiv: 1901.11033 [stat.ML].

[50]   Jeff Bezanson et al. "Julia: A fresh approach to numerical computing". In: *SIAM Review* 59.1 (2017), pp. 65–98. DOI: 10.1137/141000671. URL: https://epubs.siam.org/doi/10.1137/141000671.

[51]   R. G. JARRETT. "A note on the intervals between coal-mining disasters". In: *Biometrika* 66.1 (1979), pp. 191–193. ISSN: 1464-3510. DOI: 10.1093/biomet/66.1.191. URL: http://dx.doi.org/10.1093/biomet/66.1.191.