PROJECTS Overview

1 Handling systematic uncertainties in background shapes

A common difficulty arising in searches of new astrophysical phenomena is the impossibility of correctly specifying the distribution of the background. In physics and astronomy, the "background" refers to all the astronomical objects or particles which are not those we aim to discover. In other words, one can think of it as the collection of all the elements of the universe whose behavior has already been studied, it is well known, and therefore, they are of no interest when attempting to detect new phenomena that have never been observed before.

Model-uncertainty arising from background mismodeling can dramatically compromise the sensitivity of the experiment. Specifically, overestimating the background distribution in the signal region increases the chance of missing the detection of new phenomena. Conversely, underestimating the background outside the signal region leads to an artificially enhanced sensitivity and a higher likelihood of a false discovery.

Several methods have been proposed in literature to address this problem. For instance, Yellin's optimum interval method [1] allows us to construct suitable upper limits under the assumption that the data available is either generated from the signal or from background sources mimicking the latter. In Yellin's framework, observations generated from background sources which behave differently from the expected signal are assumed to be completely resolved (that is, excluded from the analysis when pre-processing the data) or negligible. When the background cannot be completely resolved, the methods I have proposed in Algeri [2, 3] and those introduced by Priel et al. [4] exploit the information contained in a source-free sample (i.e., a sample which is known not to contain any signal) to provide a suitable data-driven correction/estimate for the postulated/unknown background model. The resulting estimator is then employed to perform inference on the physics sample (i.e., a sample which may or may not contain the signal of interest). Model-independent solutions have been recently proposed by Chakravarti et al. [5] in the context of a two-samples analysis. Finally, when a source-free sample is not available, the discrete profiling method of Dauncey et al. [6] allows us to perform valid inference given a set of plausible candidate background models.

None of these methods, however, deals with the rather common scenario where the background cannot be completely resolved, its distribution is either unknown or known only partially, and we do not have access to a source-free sample. In this project, I will address this problem by extending the methods I have introduced in Algeri [2, 3] to the case where only one sample (the physics sample) is available.

2 Handling systematic uncertainties due to nuisance parameters

A wide variety of problems arising in astroparticle physics can be adequately described by means of mixture models. In practical applications, however, even if the functional form of both the background and the signal are known, they typically depend on unknown parameters to be estimated. While classical inferential results, such as Wilks' theorem [7], are often valid when testing hypotheses in presence of nuisance parameters, they are not applicable in many situations relevant to astrophysical searches. That is because the classical regularity conditions required by the theory of maximum likelihood fail to hold [e.g., 8].

In searches for the Higgs boson, for instance, scientists focus on the analysis of proton-to-proton collisions observed on a given mass range [e.g., 9]. From a statistical perspective, the ultimate goal is that of assessing if a bump with unknown location is present on top of a smooth background. The main issue arising in this classical "bump-hunting" problem is that, given that location of the signal is unknown, the corresponding parameter is not identifiable under the hypothesis of background

only. In high energy physics, this statistical problem is often referred to as *look-elsewhere effect* [e.g., 10]. Such nomenclature is used to emphasize that, while claiming a discovery, we want to account for the probability of potentially measuring a random fluctuation as strong as the signal observed anywhere over the search region considered.

To ensure an adequate statistical treatment of this class of problems, we consider the model

$$f(\boldsymbol{x}, \eta, \boldsymbol{\beta}, \boldsymbol{\theta}) = (1-\eta) f_b(\boldsymbol{x}, \boldsymbol{\beta}) + \eta f_s(\boldsymbol{x}, \boldsymbol{\theta}), \quad \text{with } \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^D,\ \eta \in [0, 1), \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^m$ is a nuisance parameter characterizing the background distribution and $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ is a nuisance the parameter characterizing the distribution of the signal.

To assess if the data provides sufficient statistical evidence in support of the hypothesis that the signal of interest is present, we proceed by testing

$$H_0 : \eta = 0 \quad \text{versus} \quad H_1 : \eta > 0. \tag{2}$$

This implies that, since $\eta \in [0, 1)$, the parameter $\boldsymbol{\beta}$ is identifiable, whereas $\boldsymbol{\theta}$ is not identifiable under H_0. In addition to bump-hunting, classical situations where nuisance parameters are present only under the alternative include regression models in which structural changes, such as break-points and threshold-effects, occur [11, 12, 13, 14]. The general problem has long been studied, starting at least from the seminal work of Hotelling [15] and Davies [16, 17], and further investigated in the econometrics literature by Andrews and Ploberger [18] and Hansen [19, 12, 20]. Unfortunately, these methods often require case-by-case mathematical computations [e.g., 16], estimation of the covariance structure of the underlying process [e.g., 19], or choosing weighting functions [e.g., 18], which strongly limit their applicability when dealing with complex models, such as those typically involved in astrophysical searches. Moreover, simulating the null distribution of classical test statistics, such as the profile LRT or the score statistic, may be unfeasible when dealing with stringent significance requirements [e.g., 21].

More recently, Gross and Vitells [22] have introduced a numerical solution that drastically reduces the computational complexity of a full simulation of the profile LRT statistic. In Algeri and van Dyk [23, 24], we have exploited classical results in extreme value theory and modern developments in random fields theory [e.g., 25] to extend the approach of Gross and Vitells [22] to a more general class of test statistics and to the multidimensional setting. Despite these methods offer a good trade-off between analytical calculations and numerical simulations, the resulting inference is not distribution-free, it is only approximate, and the underlying regularity conditions are often difficult to verify in practice.

In this project, we will rely on recent developments in the theory of empirical processes to introduce an entire new family of tests to deal with the look-elsewhere effect. The methods proposed are designed to overcome the computational issues arising when the models under study are too complex to simulate from them or even to evaluate over several Monte Carlo or bootstrap replicates.

References

[1] S. Yellin. Finding an upper limit in the presence of an unknown background. *Physical Review D*, 66(3):032005, 2002.

[2] S. Algeri. Detecting new signals under background mismodeling. *Physical Review D (a top physics journal)*, 101(1):015003, 2020. Solution proposed for Project 1 when 2 samples are given

[3] S. Algeri. Informative goodness-of-fit for multivariate distributions. *Electronic Journal of Statistics*, 15(2):5570–5597, 2021. Extension of the above in multi D

[4] N. Priel, L. Rauch, H. Landsman, A. Manfredini, and R. Budnik. A model independent safeguard against background mismodeling for statistical inference. *Journal of Cosmology and Astroparticle Physics*, 2017(05):013, 2017.

[5] P. Chakravarti, M. Kuusela, J. Lei, and L. Wasserman. Model-independent detection of new physics signals using interpretable semi-supervised classifier tests. *arXiv:2102.07679*, 2021.

[6] P.D. Dauncey, M. Kenzie, N. Wardle, and G.J. Davies. Handling uncertainties in background shapes: the discrete profiling method. *Journal of Instrumentation*, 10(04):P04015, 2015.

[7] S.S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

[8] S. Algeri, J. Aalbers, K.D. Morå, and J. Conrad. Searching for new phenomena with profile likelihood ratio tests. *Nature Reviews Physics*, 2(5):245–252, 2020.

[9] D.A. van Dyk. The role of statistics in the discovery of a higgs boson. *Annual Review of Statistics and Its Application*, 1:41–59, 2014.

[10] S. Algeri, D.A. van Dyk, J. Conrad, and B. Anderson. On methods for correcting for the look-elsewhere effect in searches for new physics. *Journal of Instrumentation*, 11(12):P12010, 2016.

[11] D.W.K. Andrews. Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, pages 821–856, 1993.

[12] B.E. Hansen. The likelihood ratio test under nonstandard conditions: testing the markov switching model of gnp. *Journal of applied Econometrics*, 7(S1), 1992.

[13] B.E. Hansen. Threshold effects in non-dynamic panels: Estimation, testing, and inference. *Journal of econometrics*, 93(2):345–368, 1999.

[14] R.B. Davies. Hypothesis testing when a nuisance parameter is present only under the alternative: linear model case. *Biometrika*, pages 484–489, 2002.

[15] H. Hotelling. Tubes and spheres in n-spaces, and a class of statistical problems. *American Journal of Mathematics*, 61(2):440–460, 1939.

[16] R.B. Davies. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 64(2):247–254, 1977.

[17] R.B. Davies. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 74(1):33–43, 1987.

[18] D.W.K. Andrews and W. Ploberger. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica: Journal of the Econometric Society*, pages 1383–1414, 1994.

[19] B.E. Hansen. Inference when a nuisance parameter is not identified under the null hypothesis. *Rochester Center for Economic Research Working Paper No. 296*, 1991.

[20] B.E. Hansen. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica: Journal of the econometric society*, pages 413–430, 1996.

[21] L. Lyons. Discovering the significance of 5 sigma. *arXiv:1310.1284*, 2013.

[22] E. Gross and O. Vitells. Trial factors for the look elsewhere effect in high energy physics. *The European Physical Journal C*, 70(1-2):525–530, 2010. Main reference for the look-elsewhere effect problem in Project 2

[23] S. Algeri and D.A. van Dyk. Testing one hypothesis multiple times. *Statistica Sinica*, 31: 959–979, 2021.

[24] S. Algeri and D.A. van Dyk. Testing one hypothesis multiple times: The multidimensional case. *Journal of Computational and Graphical Statistics*, pages 1–37, 2019.

Solution proposed so far for the look-elsewhere problem in multi-D

[25] Jonathan E Taylor and Keith J Worsley. Random fields of multivariate test statistics, with applications to shape analysis. *The Annals of Statistics*, 36(1):1–27, 2008.

The solution I have in mind to address problem 2 will require to learn the method described here: *https://journals.aps.org/prd/abstract/10.1103/PhysRevD.105.035030* and that I will teach you at the end of the course.