PROJECT 1: Handling systematic uncertainties in background shapes

1 Introduction

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 Outline of the proposed solution and preliminary results

Let X be a continuous random variable defined over the search interval $[\mathcal{L},\mathcal{U}] \in \mathbb{R}$. Let F_b be the true background distribution with density f_b, and denote with G_b the misspecified background distribution with density g_b . The density for the signal is f_s. We assume that both g_b and f_s are in $L^2[\mathcal{L},\mathcal{U}]$ and, for simplicity, we assume that both densities are completely specified, that is, they do not depend on free parameters. If f_b was known, a suitable model for our data would specify as

$$f(x,\eta) = (1-\eta)f_b(x) + \eta f_s(x), \quad \text{with } x \in [\mathcal{L},\mathcal{U}],\ \eta \in [0,1), \tag{1}$$

where the mixture parameter, η, corresponds to the (unknown) relative intensity of the signal. Notice that, since the background can never be completely resolved, it is reasonable to assume that η is bounded away from one. Since f_b is unknown, and g_b only consists of a misspecified version of it, a naive analysis conducted by replacing f_b in (1) with g_b would lead to the problems described in Section 1. Instead, here we consider the following approximation of (1),

$$f(x,\eta,\boldsymbol{\beta}) = (1-\eta)g_b(x)\left[1+\sum_{j=1}^{m}\beta_j T_j(x)\right]+\eta f_s(x) \tag{2}$$

where the term in the square brackets plays a role of "compensator" which allows us to model the departure of g_b from f_b. Technically, it consists of an orthonormal expansion used to approximate the density ratio $\frac{f_b(x)}{g_b(x)}$. We require that the functions T_j, $j=1,\dots,m$ in (2) to be such that,

$$\langle T_j, 1\rangle_{G_b} = 0 \quad \langle T_j, T_k\rangle_{G_b} = \mathbb{1}_{\{j=k\}}, \quad \text{for all } j,k=1,\dots,m, \tag{3}$$

with $\langle h,q\rangle_{G_b} = \int h(x)q(x)\mathrm{d}G_b(x)$. Extensions to situations where g_b depends on unknown parameters are rather trivial and only require us to construct our T_j functions so that, in addition to (3), they are also orthogonal to the score function, that is, the derivative of the log-likelihood of g_b taken with respect to the unknown parameters [e.g., 7, Chapter 9]. We postpone Section **??** the discussion of the far less trivial scenario where f_s also depends on unknown parameters. The coefficients of the expansion in (2) are collected in the vector $\boldsymbol{\beta} = (\beta_1,\dots,\beta_m)$, which we assume satisfies the regularity conditions necessary to ensure that (2) is a bona-fide density [e.g., 8], that is, it is non-negative and it integrates to one.

From (2), it follows that the approximated background model is

$$\tilde{f}_b(x) = g_b(x)\left[1+\sum_{j=1}^{m}\beta_j T_j(x)\right]. \tag{4}$$

In Algeri [2], the density in (4) is estimated directly on a source-free sample and the resulting estimate, namely $\widehat{f}_b$, is used as a proxy for the true unknown background model f_b. Inference is then conducted on the physics sample by replacing f_b in (1) with $\widehat{f}_b$. When a source-free sample is not available, however, (2) can only be estimated on the physics sample. Therefore, it is necessary to ensure that, while modeling $\frac{f_b(x)}{g_b(x)}$ in (2) by means of our compensator in the square brackets, we do not model the signal as well. To circumvent this difficulty, we must introduce an additional requirement on our T_j functions.

We begin by rewriting our model in (2) as

$$f(x,\lambda,\boldsymbol{\tau}) = g_b(x)\left[1+\lambda S_1(x)+\sum_{j=1}^{m}\tau_j T_j(x)\right], \tag{5}$$

where $S_1(x) = \frac{S(x)}{||S(x)||_{G_b}}$ with $S(x) = \frac{f_s(x)}{g_b(x)}-1$, $||S(x)||_{G_b} = \sqrt{\langle S,S\rangle_{G_b}}$, $\lambda = \eta||S(x)||_{G_b}$, and $\boldsymbol{\tau}$ is the parameter vector of components $\tau_j = (1-\eta)\beta_j$, for all $j=1,\dots,m$. We proceed by constructing our T_js so that, in addition to (3), they also satisfy

$$\langle T_j, S_1\rangle_{G_b} = 0, \quad \text{for all } j=1,\dots,m. \tag{6}$$

The model in (5) enjoys an interesting geometrical interpretation. Specifically, consider the smooth manifold $\mathcal{M} = \{f(x,\eta,\boldsymbol{\beta})|(\eta,\boldsymbol{\beta})\in[0,1)\times\mathbb{R}^m\}$, and let $\mathcal{V} = \left\langle \frac{\partial}{\partial\eta}f(x,0,\mathbf{0}), \frac{\partial}{\partial\beta}f(x,0,\mathbf{0})\right\rangle$ be the tangent space induced to $\mathcal{M}$ at $(\eta,\boldsymbol{\beta}) = (0,\mathbf{0})$. The density in (5) can be seen as the first-order approximation of (2) obtained by embedding $\mathcal{M}$ in the affine space obtained by attaching $\mathcal{V}$ to $\mathcal{M}$ at $(0,\mathbf{0})$ [e.g., 9, Theorem 9]. Hence, S_1 can be interpreted as the term responsible for modeling

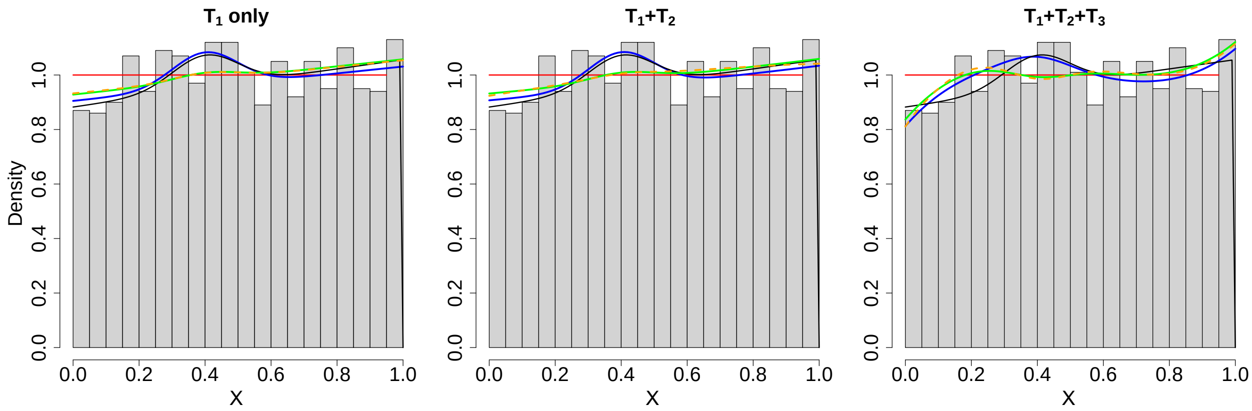

Figure 1: The solid black line is the true model. The blue solid line is the estimate obtained for it obtained by fitting (5) via MLE on $\boldsymbol{x}_{bkg+sig}$. The green solid line is the sub-model of the latter corresponding to the approximated background distribution in (4). The orange dashed line is the estimate for the background model obtained by fitting (4) on $\boldsymbol{x}_{bkg}$ via the MLE.

departure from g_b in the direction of the signal, whereas the T_j functions model departures from g_b in the direction of f_b. This can be seen more clearly through a numerical example.

Numerical example. Suppose $X \in [0,1]$ and let $g_b(x) = 1$ for all $x \in [0,1]$, that is, we assume the background to be uniform. Whereas, f_s is the density of a $N(0.4, 0.1)$ random variable truncated over $[0,1]$. The true background model is a distributed as a $N(5,5)$ random variable truncated over $[0,1]$. We generate a set of $n = 2000$ observations from f_b and we call this data set $\boldsymbol{x}_{bkg}$. We then simulate another sample of size $n = 2000$ from f_s, namely $\boldsymbol{x}_{sig}$, and we obtain our final sample $\boldsymbol{x}_{bkg+sig}$ by randomly selecting observations from $\boldsymbol{x}_{bkg}$ and $\boldsymbol{x}_{sig}$ with probability $1-\eta = 0.97$ and $\eta = 0.03$, respectively. Our goal is to approximate the latter via (5) with $g_b(x) = 1$, $S_1(x) = \frac{f_s(x)-1}{\sqrt{\int [f_s(x)-1]^2 \mathrm{d}x}}$, and we construct our T_j functions so that

$$T_1(x) = x - \langle x, S_1 \rangle S_1(x) - \langle x, 1 \rangle, \quad T_2(x) = x^2 - \langle x^2, T_1 \rangle T_1(x) - \langle x^2, S_1 \rangle S_1(x) - \langle x^2, 1 \rangle, \text{ etc.} \quad (7)$$

We proceed by estimating the parameters $(\lambda, \boldsymbol{\tau})$ in (5) on $\boldsymbol{x}_{bkg+sig}$ via Maximum Likelihood Estimation (MLE), we plug-in the estimates obtained in (5) and in (4) (after suitable transformation). That is, we first estimate the full model and we then consider the corresponding sub-model which provides an estimate of f_b. For the sake of comparison, we also estimate (4) on the background only sample $\boldsymbol{x}_{bkg}$. This is done just for the sake of illustration, that is, we want to have a sense of how similar/different the two estimators of (4) are when proceeding with a two-samples analysis (as in Algeri [2]) compared to a one-sample analysis (the method proposed here). In other words, we want to see how better/worse we would do if we had access to a background-only sample to estimate (4) directly. The results for $m = 1, 2$ and 3 are shown in Fig. 1.

Interestingly, in all three cases, the proposed procedure recovers the same estimate of the background distribution that one would obtain on a source-free sample (the green solid and orange dashed lines are very close to one anther). In our specific example, the function T_1, in addition to S_1, is sufficient to closely approximate the true distribution. No significant changes occur when adding T_2, whereas the inclusion T_3 leads to overfitting; hence the need of a robust model selection procedure. Details on the latter are given in Section 3.

3 Discussion and future developments

Identification of a suitable orthonormal basis. In our numerical example, the functions $\{T_j(x)\}_{j=1}^m$ are constructed by simply orthonormalizing a polynomial sequence with respect to S_1. Because of this, standard properties of polynomial orthonormal sequences are no longer guaranteed. For example, an orthonormal polynomial sequence on a bounded interval forms a complete basis as a result of Weierstrass approximation theorem. Due to (6), however, the same is no longer true for our system of functions $\{S_1(x), T_j(x)\}_{j=1}^m$. Therefore, in order to ensure that the expansion in (5) is indeed an orthonormal expansion for the density ratio $\frac{f(x)}{g_b(x)}$, the first step of this project will consists of proving that, when the T_j functions are constructed as in (7), $\{S_1(x), T_j(x)\}_{j=1}^m$ forms a complete basis in $L^2(G_b)$ or, alternatively, identify a suitable set of functions T_j which ensure this result, while satisfying (3) and (6).

Moreover, in Algeri [3], I have introduced a novel orthonormal tensor basis of functions which allows us to extend the two-samples analysis (that is, situations where researchers have access to both a physics and a source-free sample) to situations where the data are discrete and/or multivariate. In this project, I will exploits the results of Algeri [3] to generalize the framework outlined in Section 2 to situations where the search region is multidimensional and only the physics sample is accessible to the scientists.

Model selection. A crucial point of our analytical framework is model selection. Specifically, as highlighted through our numerical example, it is necessary to implement a suitable procedure to select which basis functions, T_j, should be discarded when estimating (5) in order to avoid overfitting. In the two-samples analysis I have proposed in Algeri [2], given a maximum number of $m_{\max}$ coefficients, the background model (4) is first estimated considering all the $m_{\max}$ terms (the full model). The final estimate of (4), however, is constructed so that it only includes the first m largest estimated coefficients which maximize either of the following criteria

$$AIC(m) = \sum_{j=1}^{m} \widehat{\tau}_{(j)}^2 - \frac{2m}{n} \quad \text{and} \quad BIC(m) = \sum_{j=1}^{m} \widehat{\tau}_{(j)}^2 - \frac{m \log n}{n} \tag{8}$$

where $\widehat{\tau}_{(j)}$ denotes the jth largest estimated coefficient, in order of magnitude, among all the $m_{\max}$ considered. Further considerations on the differences between estimates constructed via BIC versus whose obtained via AIC are discussed in Algeri [3]. In principle, one could apply these criteria also to our setting. For instance, in our numerical example, both AIC and BIC correctly select, in addition to S_1, only the first basis function, T_1. As shown in Fig. 1, such model clearly provides the best fit, while avoiding the inclusion of unnecessary extra terms. These two criteria, however, do not offer any guarantee that, while modeling the departure of F_b from G_b, we do not "overly compensate" by modeling also the signal. That is because, while each T_j is orthogonal to S_1 under g_b, the same is no longer true under under f in (1), and we may expect their correlation to increase with the number of T_j functions introduced in our model (as it may be the case when aiming to recover complex background shapes). Thus, to address this difficulty, it is necessary to identify a model selection procedure that not only provides "the best fit", but also accounts for the correlation between our T_j functions and S_1. This can be done by re-adapting the criteria in (8).

Specifically, let r_{1j}, $j = 1, \ldots, m_{\max}$ be the sample correlations between each T_j and S_1, we define the modified BIC and AIC criteria as

$$\widetilde{AIC}(m) = \sum_{j=1}^{m} \Big[\widehat{\tau}_{(j)}^2 - \frac{2}{n}\Big] \mathbb{1}_{\{|r_{1j}| \leq 2/\sqrt{n}\}} \quad \text{and} \quad \widetilde{BIC}(m) = \sum_{j=1}^{m} \Big[\widehat{\tau}_{(j)}^2 - \frac{\log n}{n}\Big] \mathbb{1}_{\{|r_{1j}| \leq 2/\sqrt{n}\}}, \tag{9}$$

where the condition $|r_{1j}| \leq 2/\sqrt{n}$ is approximately equivalent to a test for correlation at $\alpha = 0.05$ significance. In this project, I will investigate the statistical properties of estimators for (5) based on (9). This will include (but will not be limited to) a thorough assessment of the accuracy of our approach in estimating the full model and the background-only model as a function of the signal intensity and of the discrepancy between the true and the postulated background models.

(Post-selection) inference. Our ultimate goal is that of performing a test of hypothesis to assess if a signal is present or not. When considering the reparametrized model in (5), this translates into a problem of testing $H_0 : \lambda = 0$ versus $H_1 : \lambda > 0$, where λ is simply a re-scaled version of the true signal intensity η (recall that $\lambda = \eta||S_1||_{G_b}$). Such test can be easily be performed using the profile likelihood ratio test (LRT) and correcting the resulting p-value as in Chernoff [10] to account for the fact that, under H_0, λ lies on the boundary of its parameter space. It has to be noted that, while in general model selection does affect the distribution of the test statistics considered, in our setting, the selection is limited to the parameter $\boldsymbol{\tau}$. Since the latter is just a nuisance parameter, the null distribution of the LRT is unaffected by the selection process. Nonetheless, it is often of interest to assess if the initial background model postulated by the scientists, G_b, is a good approximation for the true background distribution, F_b, or not. In this case, the interest is in testing the hypotheses $H_0 : \boldsymbol{\tau} = 0$ versus $H_1 : \boldsymbol{\tau} \neq 0$, and thus the problem of adjusting inference for post-selection does arise. In the two-samples analysis, such test can be performed by considering, for instance, the so-called deviance statistic [e.g., 2, 3] which consists of the sum of the squares of the coefficients in our expansion, i.e., $D_m = \sum_{j=1}^{m} \widehat{\tau}^2_{(j)}$. When m is selected via (8), adequate post-selection adjustments for tests based on D_m have already been introduced by Algeri [3], Algeri and Zhang [11]. In this project, it will be necessary to extend these approaches to tackle situations where the selection is performed via the modified AIC and BIC criteria in (9).

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.

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

[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] Olivier Thas. *Comparing distributions*, volume 233. Springer, 2010.

[8] D.E. Barton. On neyman's smooth test of goodness of fit and its power with respect to a particular system of alternatives. *Scandinavian Actuarial Journal*, 1953(sup1):24–63, 1953.

[9] P. Marriott. On the local geometry of mixture models. *Biometrika*, 89(1):77–93, 2002.

[10] H. Chernoff. On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, pages 573–578, 1954.

[11] S. Algeri and X. Zhang. Exhaustive goodness of fit via smoothed inference and graphics. *Journal of Computational and Graphical Statistics*, pages 1–12, 2021.