Statistical methods applied to the search of sterile neutrinos

Matteo Agostini\textsuperscript{1,2,a}, Birgit Neumair\textsuperscript{1,b}

\textsuperscript{1} Physik Department E15, Technische Universität München, James-Franck-Straße 1, 85748 Garching, Germany
\textsuperscript{2} Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK

Received: 15 January 2020 / Accepted: 23 July 2020 / Published online: 18 August 2020
© The Author(s) 2020

Abstract  The frequentist statistical methods applied to search for short-baseline neutrino oscillations induced by a sterile neutrino with mass at the eV scale are reviewed and compared. The comparison is performed under limit setting and signal discovery scenarios, considering both when an oscillation would enhance the neutrino interaction rate in the detector and when it would reduce it. The sensitivity of the experiments and the confidence regions extracted for specific data sets change considerably according to which test statistic is used and the assumptions on its probability distribution. A standardized analysis approach based on the most general kind of hypothesis test is proposed.

1 Introduction

A vast experimental program has been mounted in the last decade to search for a new elementary particle named sterile neutrino [1]. The sterile neutrino is a particular type of neutrino that does not interact through the weak force. Since Pontecorvo postulated its existence in 1967 [2], the sterile neutrino has become increasingly popular and its existence is nowadays often invoked to explain the mysterious origin of neutrino masses and dark matter [3]. The discovery of sterile neutrinos would hence be a milestone towards the development of new theories beyond the Standard Model, with deep repercussions in particle physics and cosmology.

The phenomenology of sterile neutrinos depends on their hypothetical mass value. The main target of the ongoing experimental efforts is sterile neutrinos with a mass of the order of the eV, whose existence has been hinted at by various experiments [4–6] and is still under debate [1,7]. The statistical data treatment for this kind of searches presents various challenges and has not been standardized yet. Currently, different statistical methods are used in the field, each addressing a different statistical question, and thus providing different results. This situation prevents a direct comparison of the performance of the experiments and of their outcome. In addition, approximations often adopted in the statistical analysis can lead to significantly inaccurate results.

In this article we review the statistical methods used in the search for sterile neutrinos at the eV mass scale, expanding the discussion of Refs. [8–10] and performing a comprehensive comparison of the analysis approaches in scenarios with and without a signal. Section 2 describes the phenomenology of eV-mass sterile neutrinos, the signature sought after by the experiments, and the features of two toy experiments that are used in this article to compare the analysis techniques. Section 3 reviews the statistical methods and concepts used in the field. The performance of the different methods are discussed in Sects. 4 and 5. Finally, in Sect. 6, the methods are compared and a standardized analysis is proposed.

2 Phenomenology and experiments

Neutrinos of three different flavours have been observed: the electron ($\nu_e$), the muon ($\nu_\mu$) and the tau neutrino ($\nu_\tau$) [11]. These standard neutrinos can be detected by experiments because they interact through the weak force. Neutrinos can change flavor as they move through space. This phenomenon, called neutrino flavour oscillation, is possible because neutrinos of different flavours do not have a fixed mass but are rather a quantum-mechanical superposition of different mass eigenstates (i.e. $v_1$, $v_2$, and $v_3$), each associated to a distinct mass eigenvalue ($m_1$, $m_2$ and $m_3$).

A sterile neutrino ($\nu_4$) would not interact through the weak force and cannot be directly detected. However its existence would affect the standard neutrino oscillations in two ways. Firstly, a standard neutrino could oscillate into an undetectable sterile neutrino, leading to a reduction of the observed event rate within the detector. Secondly, the mass eigenstate ($\nu_4$ with mass $m_4$) primarily associated to
the sterile neutrino would enhance the transformation probability between standard neutrinos, leading to the detection of a neutrino flavor that is not emitted by the source. The experiments looking for a reduction of the interaction rate are called “disappearance” experiments while the ones seeking for an enhanced neutrino conversion are called “appearance” experiments. In principle, more than one sterile neutrino with mass at the eV scale could exist. In this work we will focus on the scenario in which there is only one eV-mass sterile neutrino.

The current-generation sterile-neutrino experiments are designed to search for oscillations among standard neutrinos at a short distance from the neutrino source, where the effect of neutrino oscillations is expected to be negligible unless eV-mass sterile neutrinos exist. The oscillation probability expected by these so-called short-baseline experiments can be approximated by:

\[ P(\nu_\alpha \rightarrow \nu_\alpha) = 1 - \sin^2(2\theta_{\alpha\alpha}) \sin^2\left( k \cdot \Delta m^2 \cdot \frac{L}{E} \right) \]  
\[ P(\nu_\alpha \rightarrow \nu_\beta) = \sin^2(2\theta_{\alpha\beta}) \sin^2\left( k \cdot \Delta m^2 \cdot \frac{L}{E} \right) \]  

where \( P(\nu_\alpha \rightarrow \nu_\alpha) \) is the survival probability for a specific neutrino of flavor \( \alpha \) and \( P(\nu_\alpha \rightarrow \nu_\beta) \) is the probability for a neutrino of flavor \( \alpha \) to transform into the flavor \( \beta \) (\( \nu_\alpha \) and \( \nu_\beta \) indicate any of the standard neutrino flavors, i.e.: \( \nu_e, \nu_\mu \) and \( \nu_\tau \)). The mixing angles (i.e. \( \theta_{\alpha\alpha} \) and \( \theta_{\alpha\beta} \)) and the difference between the squared mass eigenvalues (i.e. \( \Delta m^2 \)) are physical constants.\(^1\) The experiments aim at extracting these constants from the measurement of the oscillation probability as a function of the distance travelled by the neutrino before its interaction (\( L \)) and its initial energy (\( E \)). The maximum value of the oscillation probability is proportional to \( \sin^2(2\theta) \) that acts as a scaling factor, while the modulation of the probability is determined by \( \Delta m^2 \). The constant \( k = 1.27 \text{MeV/eV}^2/\text{m} \) applies when \( \Delta m^2 \) is expressed in eV\(^2\), \( L \) in meters and \( E \) in MeV. The modulation of the oscillation as a function of \( L/E \) is shown in Fig. 1a for a selection of \( \Delta m^2 \) values.

The features of various short-baseline experiments are summarized in Table 1. Different kinds of neutrino sources and detection technologies are used. The most common neutrino sources are nuclear reactors (producing electron anti-neutrinos up to 10 MeV), radioactive sources (electron neutrinos and anti-neutrinos up to a few MeV), and particle accelerators (muon neutrinos and anti-neutrinos up to several GeV). The detector designs are very different, but they mostly rely on scintillating materials and light detectors, or on liquid-argon time-projection chambers (LAr TPC).\(^1\)

In order to extract the sterile neutrino parameters, both \( L \) and \( E \) must be reconstructed for each detected neutrino. The oscillation baseline \( L \) is well defined as either it is much larger than the dimensions of the source and the detector – as in accelerator-based experiments – or the source is relatively compact and the detector is capable of reconstructing the position of an event – as in experiments with radioactive isotopes or reactors. The reconstruction of the event position is achieved with the physical segmentation of the detector.

\(^1\) Within the expanded oscillation phenomenology, sterile neutrinos are described through additional non-interacting flavors, which are connected to additional mass states via an extended PMNS matrix. The sterile mixing angles can be expressed as a function of the elements of the extended matrix: \( \sin^2(\theta_{\alpha\beta}) = 4|U_{\alpha4}|^2|\bar{U}_{\beta4}|^2 - |U_{\beta4}|^2 \). The mass squared difference is typically defined as \( \Delta m^2 = m_3^2 - m_1^2 \) under the approximation that \( m_1, m_2, m_3 \ll m_4 \). More details can be found in Ref. [7].
Table 1 Features and parameters of a selection of short-baseline experiments grouped according to the source of neutrinos. The kind of neutrinos emitted by the source and those detected by the experiment are shown in the third column. For each experiment the accessible range of $L_{\text{rec}}$, $E_{\text{rec}}$, and $L_{\text{rec}}/E_{\text{rec}}$ are given along with the binning used to analyze the data and the expected number of neutrino and background events. The number of neutrino events is given assuming an oscillation probability of one: $P(\nu_{\alpha} \rightarrow \nu_{\alpha}) = 1$ in the case of disappearance experiments and $P(\nu_{\alpha} \rightarrow \nu_{\beta}) = 1$ in the case of appearance experiments. Absolute resolutions on the reconstructed baseline and energy ($\sigma_L$ and $\sigma_E$) are quoted for the mean value of $L_{\text{rec}}$ and $E_{\text{rec}}$. The parameters quoted in this table are sometimes effective or approximated quantities and are intended to give an idea of the signal expected in each experiment. The last two rows show the parameters of two toy experiments used in this work to compare statistical methods. The toy experiments have parameters typical of disappearance searches based on nuclear reactors and appearance searches based on accelerators.

| Detection technology | Sought-after oscillation | $L_{\text{rec}}$ [m] range | $E_{\text{rec}}$ [MeV] range | $L_{\text{rec}}/E_{\text{rec}}$ [m/MeV] range | Binning $L_{\text{rec}} \times E_{\text{rec}}$ | $\bar{N}^e$ | $\bar{N}^b$ |
|----------------------|--------------------------|-----------------------------|-------------------------------|--------------------------------|-----------------------------|-------|-------|
| **Experiments with neutrinos from nuclear reactors:** |
| DANSS [12] | Gd-coated plastic scintillator | $\bar{\nu}_e \rightarrow \bar{\nu}_e$ | 11–13 | 0.2 | 1–7 | 0.6 | 1.5–13 | 3×24 | 10^6 | 10^4 |
| NEOS [13] | Gd-loaded liquid scintillator | $\bar{\nu}_e \rightarrow \bar{\nu}_e$ | 24 | 1 | 1–7 | 0.1 | 3.5–24 | 1×60 | 10^5 | 10^4 |
| NEUTRINO-4 [14] | Gd-loaded liquid scintillator | $\bar{\nu}_e \rightarrow \bar{\nu}_e$ | 6–12 | 0.2 | 1–6 | 0.3 | 1–12 | 24×9 | 10^5 | 10^4 |
| PROSPECT [15] | Li-loaded liquid scintillator | $\bar{\nu}_e \rightarrow \bar{\nu}_e$ | 7–9 | 0.15 | 1–7 | 0.1 | 1–9 | 6×16 | 10^4 | 10^4 |
| SoLiD [16] | Li-coated PVT scintillator | $\bar{\nu}_e \rightarrow \bar{\nu}_e$ | 6–9 | 0.05 | 1–7 | 0.2 | 1–9 | 10^5 | 10^5 |
| STEREO [17] | Gd-loaded liquid scintillator | $\bar{\nu}_e \rightarrow \bar{\nu}_e$ | 9–11 | 0.3 | 2–7 | 0.1 | 1.3–5.5 | 6×11 | 10^4 | 10^4 |
| **Experiments with neutrinos from radioactive sources:** |
| BEST [18] | Ga radiochemical | $\nu_e \rightarrow \nu_e$ | 0.1–1 | 0.6 | 0.4–1.4 | – | 0.1–2.5 | | 10^4 | 10^2 |
| SOX [19, 20] | liquid scintillator | $\nu_e \rightarrow \nu_e$ | 4–12 | 0.15 | 2–3 | 0.3 | 1.3–6.5 | | 10^4 | 10^2 |
| **Experiments with neutrinos from particle accelerators:** |
| JSNS [21] | Gd-loaded liquid scintillator | $\bar{\nu}_e \rightarrow \bar{\nu}_e$ | 24 | 5 | 10–50 | 5 | 0.5–2.5 | | 10^5 | 10^2 |
| LSND [4] | liquid scintillator | $\bar{\nu}_e \rightarrow \bar{\nu}_e$ | 30 | 10 | 20–60 | 3 | 0.5–1.5 | | 1×10 | 10^4 | 30 |
| MiniBooNE [22] | mineral oil | $\nu_\mu \rightarrow \nu_e/\bar{\nu}_e \rightarrow \bar{\nu}_e$ | 500 | 50 | 200–3000 | 20 | 0.2–2.5 | | 1×11 | 10^5/10^4 | 10^3/10^2 |
| SBN@FNAL [23] | liquid-Ar TPC | $\nu_\mu \rightarrow \nu_e$ | 110–600 | 50 | 200–3000 | 15 | 0.1–3 | | 10^5 | 10^3 |
| **Toy experiments:** |
| Disappearance | $\bar{\nu}_e \rightarrow \bar{\nu}_e$ | 7–10 | 0.5 | 2–7 | 0.1 | 1–5 | | 6×10 | 10^5 | 10^4 |
| Appearance | $\nu_\mu \rightarrow \nu_e$ | 500–550 | 50 | 200–1200 | 10 | 0.4–2.4 | | 1×20 | 10^5 | 10^3 |
and/or with advanced analysis techniques based on the properties of the scintillation or ionization signal.

The strategy used to reconstruct $E$ varies according to the primary channel through which neutrinos interact in the detector. Experiments using low energy anti-neutrinos can measure $E$ through a calorimetric approach thanks to the fact that anti-neutrinos interact via inverse beta decay and their energy is entirely absorbed within the detector. In experiments with high energy neutrinos interacting through charged-current quasi-elastic reactions, $E$ is estimated from the kinematic of the particles produced in the interaction. Some experiments measure neutrinos that interact through electron scattering and release only a random fraction of their energy inside the detector. In this cases the energy cannot be accurately reconstructed and monoenergetic neutrino sources are typically used. In the following we will use $L_{\text{rec}}$ and $E_{\text{rec}}$ to refer to the reconstructed value of baseline and energy.

Table 1 shows for each experiment the range and resolution of $L_{\text{rec}}$ and $E_{\text{rec}}$. To maximize the sensitivity to sterile neutrino masses at the eV-scale, the experiments are designed to be sensitive to $L_{\text{rec}}/E_{\text{rec}}$ values of the order of 1 m/MeV. The experiments can thus observe multiple oscillations within the detector for $\Delta m^2$ values at the eV scale. As the sought-after signal is similar among the experiments, the issues and challenges related to the statistical data treatment are the same.

The strategy used to reconstruct $E$ varies according to the primary channel through which neutrinos interact in the detector. Experiments using low energy anti-neutrinos can measure $E$ through a calorimetric approach thanks to the fact that anti-neutrinos interact via inverse beta decay and their energy is entirely absorbed within the detector. In experiments with high energy neutrinos interacting through charged-current quasi-elastic reactions, $E$ is estimated from the kinematic of the particles produced in the interaction. Some experiments measure neutrinos that interact through electron scattering and release only a random fraction of their energy inside the detector. In this cases the energy cannot be accurately reconstructed and monoenergetic neutrino sources are typically used. In the following we will use $L_{\text{rec}}$ and $E_{\text{rec}}$ to refer to the reconstructed value of baseline and energy.

Table 1 shows for each experiment the range and resolution of $L_{\text{rec}}$ and $E_{\text{rec}}$. To maximize the sensitivity to sterile neutrino masses at the eV-scale, the experiments are designed to be sensitive to $L_{\text{rec}}/E_{\text{rec}}$ values of the order of 1 m/MeV. The experiments can thus observe multiple oscillations within the detector for $\Delta m^2$ values at the eV scale. As the sought-after signal is similar among the experiments, the issues and challenges related to the statistical data treatment are the same.

The analysis of an experiment can exploit two complementary pieces of information. When the neutrino energy spectrum, flux and cross section are accurately known, the integral number of neutrino interactions expected within the detector can be computed for a given oscillation hypothesis and compared with the observed one. This approach is often called “rate” analysis. Alternatively, the relative change of rate as a function of the interaction position and neutrino energy can be compared with the expectations under different oscillation hypotheses, leaving unconstrained the integral number of events. This second approach is known as “shape” analysis. Rate and shape analysis are used simultaneously to maximize the experimental sensitivity, however they are affected by different systematic uncertainties and for a specific experiment only one of the two might be relevant. In the following we will discuss these two analyses separately. Results for a specific experiment can be estimated by interpolating between these two extreme cases. Experiments based on nuclear reactors sometimes use the so-called “ratio” method [17], in which the energy spectrum measured in a given part of the detector is normalized against what observed in a reference section. The ratio method has features similar to the shape analysis and it is not explicitly considered in the following.

Two toy experiments are used in this work to compare different analysis techniques. The first one is an example of disappearance experiment representative of the projects using nuclear reactors or radioactive isotopes as anti-neutrino source. In these experiments, the electron anti-neutrinos partially convert into sterile neutrinos with different probabilities as a function of $L_{\text{rec}}$ and $E_{\text{rec}}$. The anti-neutrino energy spectrum is considered between 2 and 7 MeV and the range of oscillation baselines accessible by the experiment is from 7 to 10 m ($L_{\text{rec}}/E_{\text{rec}}=1–5$ m/MeV with a resolution varying between 5 and 10%). The second toy experiment is an example of appearance experiment in which muon neutrinos transform into electron neutrinos with a probability enhanced by the existence of sterile neutrinos. In this case, typical for experiments based on particle accelerators, the neutrino energy varies between 200 and 1200 MeV and the oscillation baseline between 500 and 550 m ($L_{\text{rec}}/E_{\text{rec}}=0.4–2.4$m/MeV with a resolution varying between 10 and 25%).

The toy disappearance experiment can observe an oscillatory pattern in the event rate as a function of both energy and baseline, whereas the appearance experiment can observe it only in energy as the baseline can be regarded as fixed. The energy distribution of both the neutrinos emitted by the source and the background events is assumed to be flat. This is not the case in real experiments where the initial neutrino energy is peaked at some value and the background spectrum depends on the background sources. However this approximation does not affect the study reported in this work. The experimental parameters of our toy experiments, including the number of signal and background events, are summarized in the last two rows of Table 1. The uncertainties on the signal and background rates are assumed to have the realistic values of 10% and 5% for the appearance experiment, and of 2% both for the signal and background rate in the disappearance experiment.

An example of the oscillation probability reconstructed from our toy experiments is shown in Fig. 1b, c as a function of $L_{\text{rec}}/E_{\text{rec}}$. The probability is reconstructed from a set of pseudo-data generated with Monte Carlo simulations from a model with no sterile neutrinos. The oscillation probability expected assuming the existence of sterile neutrinos is shown for four mass values ($\Delta m^2=0.1, 0.5, 2$ and $10$ eV$^2$). Given the $L_{\text{rec}}/E_{\text{rec}}$ resolution of our toy experiments, an oscillatory pattern can be observed only for $\Delta m^2$ of 0.5 and 2 eV$^2$. For higher $\Delta m^2$ values, the frequency of the oscillation becomes too high and only an integrated change of the rate is visible. For smaller $\Delta m^2$ values, the oscillation length approaches the full $L_{\text{rec}}/E_{\text{rec}}$ range to which the experiment is sensitive, resulting in a loss of sensitivity. In the appearance experiments the discrimination power among different oscillatory patterns relies only on $E_{\text{rec}}$ since $L_{\text{rec}}$ is fixed.

In this work we focus on short-baseline experiments. We do not consider other oscillation experiments (e.g. Daya Bay [24], Double Chooz [25], RENO [26], MINOS [27], NOvA [28] and Ice Cube [29]) for which the oscillation probability cannot be approximated by Eqs. (1) and (2) as it is
either complicated by the overlap between oscillations driven by multiple mass eigenstates or by matter effects [7]. We also do not consider approaches that are not based on oscillations such as the study of cosmological structures [30], the high-precision spectroscopy of beta-decays (e.g. KATRIN [31]), or electron captures (e.g. ECHO [32]). The statistical issues of these searches are different from those of the short-baseline experiments and would require a specific discussion.

3 Statistical methods

The goal of short-baseline experiments is to search for a signal due to a sterile neutrino with mass at the eV-scale by measuring the oscillation probability at different $L$ and $E$ values. The parameters of interest associated to the sterile neutrino are the mixing angle and its mass eigenvalue. However, because of the functional form of Eqs. (1) and (2), the observables of the experiments are a function of the angle and mass, i.e.: $\sin^2(2\theta)$ and $\Delta m^2$. In the following we will refer to $\sin^2(2\theta)$ and $\Delta m^2$ as the parameters of interest of the analysis.

The role of statistical inference applied to the data from sterile neutrino searches can be divided into four tasks:

1. point estimation: the computation of the most plausible value for $\sin^2(2\theta)$ and $\Delta m^2$;
2. hypothesis testing: given a hypothesis on the value of $\sin^2(2\theta)$ and $\Delta m^2$, decide whether to accept or reject it in favor of an alternative hypothesis. Among the different tests that can be performed, testing the hypothesis that there is no sterile neutrino signal (i.e. $\sin^2(2\theta) = 0$ or $\Delta m^2 = 0$) is of primary interest for an experiment aiming at a discovery;
3. interval estimation: construct a set of $\sin^2(2\theta)$ and $\Delta m^2$ values that includes the true parameter values at some predefined confidence level;
4. goodness of fit: estimate if the data can be described by the model.

The statistical methods used by sterile neutrino experiments are based on the likelihood function. The point estimation is carried out using maximum likelihood estimators, i.e. by finding the values of $\sin^2(2\theta)$ and $\Delta m^2$ that correspond to the maximum of the likelihood function. The hypothesis testing is based on the ratio of likelihoods. The interval estimation is carried out by inverting a set of likelihood-ratio based hypotheses tests, and grouping the hypotheses that are accepted. The goodness-of-fit test can be carried out assuming the most plausible value of the parameters of the model (i.e. the maximum likelihood estimator for $\sin^2(2\theta)$ and $\Delta m^2$) and using for instance a Pearson $\chi^2$ or a “likelihood ratio” test [33,34].

While the procedures for point estimation and goodness of fit are not controversial, the hypothesis testing differs significantly among the experiments since multiple definitions of the hypotheses are possible. Changing the hypothesis definition does not only affect the outcome of the hypothesis test but also of the interval estimation, which is performed by running a set of hypothesis tests. The comparison of tests based on different hypothesis definitions is the subject of Sects. 4, 5 and 6.

In this section we review the ingredients needed to build the tests and the statistical concepts that will be used in the following. Firstly, we consider the likelihood function and derive a general form that can be applied to all experiments (Sect. 3.1). Then we discuss the possible hypothesis definitions and the resulting test statistics (Sect. 3.2). The properties of the test statistic probability distributions are described in Sect. 3.3. Finally, in Sect. 3.4 we examine the construction of confidence regions and in Sect. 3.5 the concept of power of a test and sensitivity.

Bayesian methods have not been applied in the search for sterile neutrinos so far. Even if their usage could be advantageous, we will not consider them in the following and keep the focus on the methods that are currently in use.

3.1 The likelihood function

Short-baseline experiments measure the oscillation baseline and the energy of neutrinos, i.e. a pair of $(L_{\text{rec}}, E_{\text{rec}})$ values for each event. $L_{\text{rec}}$ and $E_{\text{rec}}$ are random variables whose probability distributions depend on the true value of $L$ and $E$. Monte Carlo simulations are used to construct the probability distributions of $L_{\text{rec}}$ and $E_{\text{rec}}$ for a neutrino event given a $\sin^2(2\theta)$ and $\Delta m^2$ value, $p_e(L, E | \sin^2(2\theta), \Delta m^2)$, and for a background event, $p_b(L, E)$. Additional quantities are sometimes measured, however they are ultimately used to constrain the background or the systematic uncertainties and can be neglected in this work.

To our knowledge, all the experiments organize the data in histograms and base their analysis on a binned likelihood function. The use of histograms is motivated by the fact that the number of neutrino events is large, between $10^4$ and $10^6$ as shown in Table 1. Binning the data leads to a new set of random variables that are the numbers of observed events in each bin: $N_{\text{obs}}^{i,j} = \{N_{11}^{\text{obs}}, N_{12}^{\text{obs}}, \ldots, N_{ij}^{\text{obs}} \}$ where $i$ runs over the $L_{\text{rec}}$ bins and $j$ over the $E_{\text{rec}}$ bins. Consistently, we indicate with $h_{ij}^{e}$ and $h_{ij}^{b}$ the integral of the probability distribution function for neutrino and background events over each bin:

$$h_{ij}^e = \int_{L,E \in \text{bin}_{ij}} p_e(L, E | \sin^2(2\theta), \Delta m^2) \, dL \, dE \quad (3)$$

$$h_{ij}^b = \int_{L,E \in \text{bin}_{ij}} p_b(L, E) \, dL \, dE$$
on the number of neutrino events. Conversely, a pure rate analysis will be emulated by enlarging the size of the bins in $h^{c}$ and $h^{b}$ up to the point at which there is a single bin and any information on the number of events as a function of $L_{\text{rec}}$ or $E_{\text{rec}}$ is lost.

### 3.2 Hypothesis testing and test statistics

The hypothesis testing used nowadays in particle physics is based on the approach proposed by Neyman and Pearson in which the reference hypothesis $H_0$ (i.e. the null hypothesis) is compared against an alternative hypothesis $H_1$ [35]. The test is a procedure that specifies for which data sets the decision is made to accept $H_0$ or, alternatively, to reject $H_0$ and accept $H_1$. Usually a hypothesis test is specified in terms of a test statistic $T$ and a critical region for it. The test statistic is a function of the data that returns a real number. The critical region is the range of test statistic values for which the null hypothesis is rejected.

The critical region is chosen prior the analysis such that the test rejects $H_0$ when $H_0$ is actually true with a desired probability. This probability is denoted with $\alpha$ and called the “size” of the test. In the physics community, it is more common to quote $1-\alpha$ and refer to it as the “confidence level” of the test. For instance, if $H_0$ is rejected with $\alpha = 5\%$ probability when it is true, the test is said to have 95% confidence level (CL). In order to compute the critical thresholds, the probability distribution of the test statistic must be known. In our work, the distributions are constructed from large ensembles of pseudo-data sets generated via Monte Carlo techniques.

In sterile neutrino searches a hypothesis is defined by a set of allowed values for $\sin^2(2\theta)$ and $\Delta m^2$. The null hypothesis is defined as:

$$H_0: \{\sin^2(2\theta), \Delta m^2 : \sin^2(2\theta) = X, \Delta m^2 = Y\} \quad \text{(7)}$$

where $X$ and $Y$ are two particular values. Since the mixing angle and the mass eigenvalue are defined as non-negative numbers by the theory and $m_3 \geq m_1$, the most general version of the alternative hypothesis is

$$H_1: \{\sin^2(2\theta), \Delta m^2 : 0 \leq \sin^2(2\theta) \leq 1, \Delta m^2 \geq 0\}. \quad \text{(8)}$$

A test based on these two hypotheses leads to a generalized likelihood-ratio test statistic of the form [35]:

$$T = -2 \ln \sup_{\eta} \frac{\sup_{\sin^2(2\theta), \Delta m^2, \eta} \mathcal{L} (\sin^2(2\theta), \Delta m^2, \eta | N_{\text{obs}})}{\mathcal{L} (\sin^2(2\theta), \Delta m^2, \eta | N_{\text{obs}})} \quad \text{(9)}$$

where the denominator is the maximum of the likelihood for the observed data set over the parameter space allowed for the parameters of interest ($\{\sin^2(2\theta), \Delta m^2\} \in H_1$) and the
nuisance parameters. The numerator is instead the maximum of the likelihood in the restricted space in which \( \sin^2(2\theta) \) and \( \Delta m^2 \) are equal to the value specified by \( H_0 \).

If the value of \( \Delta m^2 \) or \( \sin^2(2\theta) \) are considered to be known because of theoretical predictions or of a measurement, then the parameter space of the alternative hypothesis can be restricted. Restricting the parameter space is conceptually equivalent to folding into the analysis new assumptions and changes the question addressed by the hypothesis test. The smaller is the parameter space allowed by the alternative hypothesis, the greater the power of the test will be.

Three tests have been used in the context of sterile neutrino searches and are summarized in Table 2. The most general test is the one that we just described and that leads to the test statistic given in Eq. (9). We will indicate this test statistic in the analysis new assumptions and changes the question addressed by the hypothesis test. The smaller is the parameter space allowed by the alternative hypothesis, the greater the power of the test will be.

The second test statistic used in the field can be traced back to the situation in which the mass squared difference is considered to be perfectly known and is equal to the value of the null hypothesis (\( \Delta m^2 = Y \)). In this case the alternative hypothesis and its related test statistic are:

\[
\begin{align*}
H_1 &= \{ \sin^2(2\theta), \Delta m^2 : 0 \leq \sin^2(2\theta) \leq 1, \Delta m^2 = Y \} \\
T_1 &= -2 \ln \sup_{\sin^2(2\theta), \eta} \frac{L(\sin^2(2\theta), \Delta m^2 = Y, \eta|N_{\text{obs}})}{L(\sin^2(2\theta), \eta|N_{\text{obs}})}. \quad (10)
\end{align*}
\]

While the numerator of \( T_1 \) is the same of \( T_2 \), the maximum of the likelihood at the denominator is now computed over a narrower parameter space, restricted by the condition \( \Delta m^2 = Y \).

The third test corresponds to the simplest kind of hypothesis test that can be performed. Both the null and alternative hypothesis have the parameters of interest fully defined. The alternative hypothesis is now the no-signal hypothesis and this leads to a test of the form:

\[
\begin{align*}
H_1 &= \{ \sin^2(2\theta), \Delta m^2 : \sin^2(2\theta) = 0, \Delta m^2 = 0 \} \\
T_0 &= -2 \ln \sup_{\eta} \frac{L(\sin^2(2\theta) = 0, \Delta m^2 = 0, \eta|N_{\text{obs}})}{L(\sin^2(2\theta), \eta|N_{\text{obs}})}. \quad (13)
\end{align*}
\]

The numerator and denominator are the maximum likelihoods for fixed values of the parameters of interest, where the maximum is computed over the parameter space allowed for the nuisance parameters. By construction, the no-signal hypothesis is always accepted when it is used as \( H_0 \), since the test statistic becomes identically equal to zero.

Nowadays, the value of \( \sin^2(2\theta) \) or \( \Delta m^2 \) is still considered to be unknown and all the parameter space accessible by the experiment is probed in search for a signal. This situation should naturally lead to the usage of \( T_2 \). However the maximization of the likelihood required by \( T_2 \) is challenging from the computational point of view. Reducing the dimensionality of the parameter space over which the likelihood is maximized can enormously simplify the analysis and, for such a practical reason, \( T_1 \) and \( T_0 \) are used even if the restriction of the parameter space is not intended.

In the neutrino community, the analysis based on \( T_2 \) has been called “2D scan” or “global scan” while the analysis based on \( T_1 \) is known as “raster scan” [8,9]. In the absence of nuisance parameters, the definition of these test statistics reduce to those discussed in Ref. [8]. \( T_0 \) has been used in the framework of a method called “Gaussian CLs” [10].

The search for new particles at accelerators presents many similarities with the search for sterile neutrinos. For instance, in the search for the Higgs boson, the sought-after signal is a peak over some background. The two parameters of interest are the mass of the Higgs boson, which defines the position of the peak, and its coupling with other particles, which defines the amplitude. Similarly, in the search for sterile neutrinos \( \Delta m^2 \) defines the shape of the signal and \( \sin^2(2\theta) \) its strength. When the Higgs boson is searched without assumptions on its mass and coupling, a test similar to \( T_2 \) is performed (i.e. a “global p-value” analysis). When the mass is assumed to be known, a test similar to \( T_1 \) is used (i.e. a “local p-value” analysis) [36]. Procedures for converting a local into a global p-value analysis have been developed in the last years [37,38] and are nowadays used to avoid the direct usage of \( T_2 \) that is computationally demanding. This procedure is known as a correction for the “look-elsewhere effect”. We have studied the correction described in Ref. [37] and found that it does not provide accurate results for sterile neutrino experiments because of the oscillatory nature of the sought-after signature. Our studies are discussed in Appendix E.

3.3 Test statistic probability distributions

The test statistic \( T_2 \) and \( T_1 \) can assume any non-negative value. If the absolute maximum of the likelihood corresponds to \( H_0 \), these test statistics are identically zero. The farther the absolute maximum is from the parameter space of the null hypothesis, the larger the test statistic value becomes. If the null hypothesis is true, the probability distribution of this kind of test statistic is expected to converge to a chi-square function in the large sample limit, but only if the regularity conditions required by Wilks’ theorem are met [39]. In particular, given the ratio between the dimensionality of the parameter space for the null and alternative hypothesis (i.e. the number of free parameters of interest in the likelihood maximization), \( T_2 \) would converge to a chi-square with two
degrees of freedom and $T_1$ to a chi-square with one degree of freedom. As discussed in Sect. 4.3, the conditions required by Wilks’ theorem are not always valid in sterile neutrino experiments and the assumption that the test statistic follows a chi-square distribution can lead to significantly inaccurate results.

The probability distributions of $T_0$ are qualitatively different from those of $T_2$ and $T_1$. $T_0$ is negative when the tested signal hypothesis is more likely than the no-signal hypothesis, positive in the opposite case. The larger is the test statistic value, the more the tested hypothesis is disfavoured. Under mild conditions, the probability distribution of $T_0$ converges to a Gaussian function [10].

Our results are based on test statistic probability distributions constructed from ensembles of pseudo-data. Firstly a grid in the $\sin^2(2\theta)$ vs. $\Delta m^2$ space is defined. Secondly, for each point on the grid an ensemble of pseudo-data is generated. The probability distributions are hence constructed by computing the test statistic for the pseudo-data in the ensemble. The pseudo-data are generated for a fixed value of the nuisance parameters. More details on our procedure are described in Appendix B.

### 3.4 Interval estimation and confidence regions

The results of a neutrino oscillation search are generally summarized by a two-dimensional confidence region in the $\sin^2(2\theta)$ vs. $\Delta m^2$ space. The confidence region defines a set of parameter values that are compatible with the data given a certain confidence level. The construction of a confidence region is formally referred to as an interval estimation.

One of the most popular statistical techniques to construct a confidence region is through the inversion of a set of hypothesis tests [35,40]. This is also the technique used by experiments searching for sterile neutrinos. The construction starts from the selection of a specific test and its resulting test statistic. The parameter space considered for $H_1$ is naturally the space in which the region will be defined. Usually a grid is fixed over this space and a test is run for each point. The tests in this set have the same $H_1$ but $H_0$ is changed to the value of the parameters at each point. This standard construction guarantees that the properties of the test statistic carry over to the confidence region and the confidence level of the region is equal to that of the test [35].

Since the confidence region is constructed in the parameter space considered by the alternative hypothesis, tests based on $T_2$ would naturally lead to two-dimensional confidence regions in the $\sin^2(2\theta)$ vs. $\Delta m^2$ space, tests based on $T_1$ to one-dimensional regions in the $\sin^2(2\theta)$ space, and tests based on $T_0$ to point-like regions. As we already mentioned, $T_0$ and $T_1$ are used even if the restriction of the parameter space is not intended. To build two-dimensional regions for $T_0$ and $T_1$ a non-standard procedure is used, i.e. the final region is created as union of one-dimensional or point-like regions. For instance, while $T_1$ would require the value of $\Delta m^2$ to be known, one can technically construct a one-dimensional $\sin^2(2\theta)$ region for a scan of “known” $\Delta m^2$ values and then take the union of these regions. Assuming that the true $\Delta m^2$ is among the scanned values, the confidence level of the union will be the same of the test.

The standard procedure for constructing confidence regions ensures that inverting uniformly most powerful tests provides uniformly most accurate confidence regions, i.e. regions with minimal probability of false coverage [35]. This is not true for the non-standard procedure described above, which indeed produces regions with peculiar features and pathologies as discussed in Sects. 4 and 5.
3.5 Power of the test and sensitivity

The performance of the different kinds of hypothesis tests can be studied by comparing their expected outcome under the assumption that an hypothesis is true. The idea of expected outcome is captured by the statistical concept of “power” of a test. The power is defined as the probability that the test rejects the null hypothesis when the alternative hypothesis is true [35].

In high-energy physics the concept of power is replaced by the idea of sensitivity of an experiment. The sensitivity is defined as the set of hypotheses for which the test would have a 50% power. The focus is thus shifted from the expected outcome of a test to a set of hypotheses. Two kinds of sensitivities are commonly used. The exclusion sensitivity is the set of hypotheses that, if true, would allow in 50% of the experiments to reject the no-signal hypothesis (H0 is now the no-signal hypothesis). More details on how the sensitivities are defined and computed can be found in Appendix C.

Both sensitivities will be displayed as contours in the $\sin^2(2\theta)$ vs. $\Delta m^2$ parameter space and for a 95% CL test. A larger confidence level is typically required for a discovery, however we prefer to use the same value for the exclusion and discovery sensitivity to ease their comparison.

4 2D and raster scan

In this section we compare the confidence regions built using $T_2$ and $T_1$. The comparison is done using the toy experiments introduced in Sect. 2: a disappearance experiment representative of searches based on reactor neutrinos and an appearance experiment representative of the accelerator-based experiments. First we focus on the sensitivity of the toy experiments (Sect. 4.1) and then consider the results extracted for specific sets of data (Sect. 4.2). Finally, in Sect. 4.3, we study the impact of approximating the test statistic distributions with chi-square functions. Our results and conclusions fully agree with previous works [8, 9].

4.1 Sensitivity

The exclusion and discovery sensitivities of our toy disappearance experiment based on the statistic $T_2$ are shown in Fig. 2a. The exclusion sensitivity (black lines) delimits the parameter space that has a 50% chance to be rejected by a 95%-CL test under the assumption that sterile neutrinos do not exist. The discovery sensitivity (red lines) delimits the set of hypotheses which, assuming those to be true, have a 50% chance that the no-signal hypothesis is rejected by a 95%-CL test. The figure shows separately the sensitivity for a rate and shape analysis (dotted lines) that are useful to illustrate which kind of information contributes most to the overall sensitivity as a function of $\Delta m^2$. Three $\Delta m^2$ regions can be identified in Fig. 2a:

- $\Delta m^2 > 10$ eV$^2$: the oscillation length is smaller than the detector resolution on $L_{\text{rec}}$ and/or $E_{\text{rec}}$, making the experiment sensitive only to an overall reduction of the integral rate (sensitivity dominated by the rate analysis);
- $0.1$ eV$^2 < \Delta m^2 < 10$ eV$^2$: the oscillation length is larger than the experimental resolution and smaller than the range of $L_{\text{rec}}$ and/or $E_{\text{rec}}$ values accessible by the detector, making the experimental sensitivity dominated by the shape analysis;
- $\Delta m^2 < 0.1$ eV$^2$: the oscillation length becomes larger than the detector dimensions. The experimental sensitivity decreases with the $\Delta m^2$ value (i.e. with increasing oscillation length) and larger $\sin^2(2\theta)$ values are needed in order to observe a signal. The sensitivity is approximately proportional to the product $\sin^2(2\theta) \times \Delta m^2$.

Example of the expected oscillations in these three regions are shown in Fig. 1b, c.

The total sensitivity is given by a non-trivial combination of the sensitivity of the rate and shape analysis. The rate and shape analysis are emulated by considering only parts of the likelihood function (see Sect. 3.1) that would otherwise have common parameters. The sensitivity for the rate analysis is higher than the total one for high $\Delta m^2$ values. This feature is related to the fact that $\sin^2(2\theta)$ and $\Delta m^2$ are fully degenerate parameters in a rate analysis based on $T_2$. A rate analysis uses a single piece of information and cannot fix two correlated parameters. Given an observed number of events, the global maximum of the likelihood function can be obtained for infinite combinations of $\sin^2(2\theta)$ and $\Delta m^2$ values. The degeneracy is however broken when the rate information is combined with the shape one. The number of effective degrees of freedom of the problem changes, and this results in a reduction of sensitivity.

The sensitivities of our toy appearance experiment computed for $T_2$ is shown in Fig. 2b. The same $\Delta m^2$ regions discussed for the disappearance experiment can be identified, even if the relative weight of the shape and rate information is different. In particular, in the appearance searches a shape analysis can provide information on $\Delta m^2$ but not on $\sin^2(2\theta)$. The number of expected $\nu_e$ events is indeed proportional to the product of the oscillation amplitude $\sin^2(2\theta)$ and flux of $\nu_\mu$ neutrinos. If the flux is left unconstrained in the fit, no statement can be made about the oscillation amplitude. This is the reason why the shape analysis contribution is not displayed. The rate analysis accounts for the bulk of the sensitivity and adding the shape information does not result...
Discovery and exclusion sensitivities at 95% CL for toy disappearance and appearance experiments. The sensitivities are shown for the test statistics $T_2$ and $T_1$. Whenever possible, the contribution of the shape and rate analysis are displayed separately. The discovery sensitivity is typically associated to a larger confidence level compared to the exclusion sensitivity, however here we use the same value to highlight their differences.

Fig. 2
Exclusion and discovery sensitivity at 95% CL for a toy disappearance and appearance experiment. The sensitivities are shown for the test statistics $T_2$ and $T_1$. Whenever possible, the contribution of the shape and rate analysis are displayed separately. The discovery sensitivity is typically associated to a larger confidence level compared to the exclusion sensitivity, however here we use the same value to highlight their differences.

in a net improvement, in the sense that the slight improvement is compensated by the reduction of sensitivity due to the increased number of effective degrees of freedom discussed above. Having a sensitivity dominated by the rate analysis is typical for experiments using accelerators as neutrino source.

The exclusion and discovery sensitivities are similar to each other for both the disappearance and appearance experiments. Some differences are however present. When computing the discovery sensitivity, the hypothesis tested is the no-signal hypothesis. Since $\sin^2(2\theta) = 0$ or $\Delta m^2 = 0$ are points at the edge of the allowed parameter space, the number of degrees of freedom of the problem decreases when testing them and the power of the test increases. The exclusion sensitivity is instead computed for values of the parameters far from the edges. This is the reason why the discovery sensitivity is in general expected to be stronger than that the exclusion one. However the situation is reversed in the shape analysis because of a peculiar feature of the sterile neutrino signature. Statistical fluctuations between bins mimic an oscillation signal and the maximum of the likelihood is always found far from the no-signal hypothesis. This decreases significantly the power of the test for a discovery while it does not affect much the exclusion case. If the shape analysis dominates the overall sensitivity, as in our disappearance experiment, its features propagate also to the combined sensitivities.

Figure 2c, d show the sensitivities for our toy disappearance and appearance experiments computed for $T_1$. The overall features are similar to those of $T_2$ and the weight of the rate and shape information in the three $\Delta m^2$ regions are also consistent. However, since the parameter space of the alternative hypothesis is now restricted, $T_1$ has greater power than $T_2$ for a given $\Delta m^2$ value. This leads to sensitivities that are stronger by up to a factor 2 in terms of $\sin^2(2\theta)$. This is particular evident for high $\Delta m^2$ values where, differently from $T_2$, now the number of effective degrees of freedom in the alternative hypothesis is always one (only $\sin^2(2\theta)$ is free) and the total sensitivity is equal to the one of the rate analysis. The restriction of the parameter space is also the reason why
the maximum of the likelihood can now correspond to the no-signal hypothesis and the discovery sensitivity is stronger than the exclusion sensitivity even in a shape analysis.

It should be emphasized that the mixing angle in the plot for the disappearance experiment is different from that of the appearance experiment. The minimal value of $\sin^2(2\theta)$ accessible by an experiment cannot be used as a figure of merit to compare disappearance and appearance experiments. A comparison can however be done assuming a specific theoretical model that connects the value of $\sin^2(2\theta_{ee})$ and $\sin^2(2\theta_{\mu e})$ [7].

4.2 Results from observed data sets

The confidence region derived from an observed data set can significantly differ from the expectations because of statistical fluctuations on the number of signal and background events. This issue is particularly relevant when no signal is observed and an upper limit on a parameter is reported. Frequentists upper limits can indeed become extremely strong in case of background fluctuations.

In sterile neutrino searches, when no signal is observed, the confidence region extends down to $\sin^2(2\theta) = 0$ for most of the $\Delta m^2$ values and it is bounded by an upper limit on $\sin^2(2\theta)$ that plays the role of the maximum signal strength. It is hence informative to report the observed upper limit along with its expected value and variance under the no-signal hypothesis. This has been first proposed in Ref. [8] and it is nowadays common practice.

Figure 3a shows the confidence region derived with T2 from a pseudo-data set generated for the toy disappearance experiment under the no-signal hypothesis. In addition to the confidence region, the expected distribution of the upper limit is displayed in terms of its median value and 68%/95% central intervals. The median is exactly the exclusion sensitivity plotted in Fig. 2a. The observed upper limit fluctuates around the median expectation. This is true for all possible realizations of the data as the likelihood is maximized for a specific phase of the oscillatory pattern that matches the statistical fluctuations between the bins of the data set. This phase is reproduced at regularly spaced values of $\Delta m^2$ over the full parameter space. The limit gets weaker when the phase helps describing the data, stronger when it does not. The overall shift of the observed limit with respect to the median value is instead due to the fact that the random number of events injected in this particular data set is slightly above its median expectation. The width of the green and yellow bands gives an idea of the magnitude of the fluctuations at a given $\Delta m^2$, as they contain the upper limit on $\sin^2(2\theta)$ with a probability of 68% and 95% respectively.

The results and expectations based on T1 are shown in Fig. 3b. For a given $\Delta m^2$ value, T1 has greater power than T2 as the parameter space allowed under the alternative hypothesis is smaller. This leads to stronger limits in terms of $\sin^2(2\theta)$. On the other hand, the non-standard construction of the confidence region can lead to accept the no-signal hypothesis (i.e. $\sin^2(2\theta) = 0$) at some $\Delta m^2$ value and reject it at others (see for instance $\Delta m^2 \sim 1$ eV$^2$). This can happen because the tests performed by T1 at different $\Delta m^2$ values are independent by each other.

The difference between T2 and T1 is more evident when a signal is present in the data. Figure 4 shows the reconstructed confidence regions for a pseudo-data set generated assuming a sterile neutrino with $\sin^2(2\theta) = 0.04$ and $\Delta m^2 = 1$ eV$^2$. The confidence regions are shown for 68% and 95% CL along with the discovery sensitivity. The analysis based on T2 is able to properly pin down the signal and it returns a two-dimensional confidence region surrounding the true parameter values. T1 returns a $\sin^2(2\theta)$ region that is similar to that of T2 for $\Delta m^2$ values close to the true one. However it returns an allowed region for any $\Delta m^2$ value. This is again due to the non-standard construction of the confidence region used in combination with T1. As tests performed at different $\Delta m^2$
conditions. The parameter \( \sin^2 \Delta \) originates from an hypothesis test in which effective number of free parameters of interest \([42]\). Sterile values are independent from each other, an allowed \( \sin^2 \theta \) of a sterile neutrino with \( \sin^2 \Delta \). Confidence regions at 68% and 95% CL for pseudo-data generated by the toy disappearance experiment assuming the existence of a sterile neutrino with \( \sin^2 (2\theta) = 0.04 \) and \( \Delta m^2 = 1 \text{eV}^2 \). The 95% CL discovery sensitivity is shown for both the statistic \( T_2 \) (red line) and \( T_1 \) (grey line). The right panel shows the minimal size of a test based on \( T_1 \) that is required to reject the no-signal hypothesis as a function of the \( \Delta m^2 \) value (grey line). The minimal test size for \( T_2 \) (red line) is independent by \( \Delta m^2 \) as all values are tested simultaneously. Differently from \( T_2, T_1 \) would reject the no-signal hypothesis even for a test size of \( 10^{-7} \) (corresponding to a 4\( \sigma \) two-sided Gaussian deviation). This discrepancy is due to the look-elsewhere effect.

values are independent from each other, an allowed \( \sin^2 (2\theta) \) interval is always found.

In summary, the greater power of \( T_1 \) in terms of \( \sin^2 (2\theta) \) comes at the cost of losing any capability in constraining \( \Delta m^2 \). This is consistent with the fact that this test statistic originates from an hypothesis test in which \( \Delta m^2 \) is considered to be a known and fixed parameter.

4.3 Validity of Wilks’ theorem

Constructing the probability distributions of the test statistic through Monte Carlo methods can be computationally challenging and often it is avoided by invoking the asymptotic properties of the generalized likelihood ratio test. If the regularity conditions required by Wilks’ theorem are met \([39, 41]\), \( T_2 \) and \( T_1 \) are indeed expected to follow a chi-square distribution with a number of degrees of freedom equal to the effective number of free parameters of interest \([42]\). Sterile neutrino experiments do not typically fulfill these regularity conditions. The parameter \( \sin^2 (2\theta) \) is often reconstructed next to the border of its allowed range (only positive values are physically allowed) or far from its true value when the statistical fluctuations mimic a signal. This induces a bias in its maximum likelihood estimator. In addition, for \( T_2 \), the alternative hypothesis becomes independent by \( \Delta m^2 \) when \( \sin^2 (2\theta) \) is equal to zero.

The impact of assuming Wilks’ asymptotic formulas has been evaluated by studying the coverage probability, i.e. the probability that the confidence region covers the true value of the parameters of interest \([35]\). If the asymptotic formulas are a good approximation of the actual test statistic distribution, the coverage should be equal to the confidence level of the test used to create the confidence region. A direct comparison of the test statistic distributions is discussed in Appendix D.

The coverage probability computed for \( T_2 \) assuming the validity of Wilks’ theorem is shown in Fig. 5 for both our toy disappearance and appearance experiments, considering separately a rate and shape analysis. The test statistic distributions have been approximated by a chi-square with one or two degrees of freedom, according to the number of non-degenerate parameters of interest in the alternative hypothesis (see insets in the figure). The coverage is generally correct in the parameter space where the experiment is sensitive to a signal. The rate analysis shows just a slight overcoverage where the experiment is not sensitive. This is expected as \( \sin^2 (2\theta) \) is bounded to positive values, causing an effective reduction of the degrees of freedom of the test when the signal is reconstructed close to the edge of the allowed parameter space \([43]\).

The shape analysis has instead a severe undercoverage for \( \sin^2 (2\theta) \) values below the sensitivity of the experiment and the coverage can be as low as 60%, while its nominal value should be 95%. The undercoverage is connected to the fact that when a binned analysis is performed, it is always possible to find a sterile neutrino hypothesis whose oscillatory pattern helps reproducing the statistical fluctuations between bins. As a result, even if no signal is present in the data, the maximum of the likelihood always corresponds to some oscillation hypothesis. This is conceptually equivalent to overfitting and it artificially increases the degrees of freedom of the test and the test statistic values (see discussion in Appendix D). A region of overcoverage is present also in the parameter space within the sensitivity of the experiment at low \( \Delta m^2 \) values, where the oscillation length becomes close to the dimension of the detector, creating a degeneracy between the parameters of interest. Together with the restriction of the parameter space \( \sin^2 (2\theta) \leq 1 \), the number of effective degrees of freedom changes.

When the analysis includes both the rate and shape information, the coverage shows a combination of the features discussed above. In particular, in the parameter space beyond the sensitivity of the experiment, the overcoverage of the rate analysis partially compensates for the undercoverage of the shape analysis. Severe undercoverage regions are however still present, consistently with the results obtained in Ref. \([8]\).

The difference between the outcome of a test based on probability distributions constructed with Monte Carlo techniques and their chi-square approximation is shown in Fig. 6. Both the sensitivities and the confidence regions reconstructed from pseudo data are significantly different, up to 70% in terms of \( \sin^2 (2\theta) \). For experiments with a sensitivity dominated by the shape analysis, the confidence region...
can even switch from an upper limit to an island, leading to an unjustified claim for a discovery. The probability for this event to occur can be significant, up to 40% in our toy experiment for the considered hypothesis. More details on the probability distributions of $T_2$ and the option to compute exclusion sensitivities based on the Asimov data set are discussed in Appendix D.

While the asymptotic approximation is not satisfactory for tests based on $T_2$, it is instead very good for tests based on $T_1$. The coverage of $T_1$ has exactly the same features of Fig. 5a, d and therefore it is not shown here. The coverage is correct in the region in which the experiment is sensitive and is slightly higher ($97.5\%$) in the parameter space beyond the experimental sensitivity. The possibility of avoiding a Monte Carlo construction of the probability distributions of $T_1$ is a significant advantage and contributed to make $T_1$ popular in the sterile neutrino community.

5 Testing of simple hypotheses

The exclusion sensitivity based on $T_0$ for our toy disappearance experiment is shown in the first row of Fig. 6 separately for the rate, shape and combined analysis. This test provides a sensitivity significantly stronger than what obtained with $T_2$ and $T_1$. This is expected as the test involves “simple” hypotheses in which the parameters of interest are fixed. The parameter space of the alternative hypothesis is now even more restricted than for $T_1$ and the test has maximum
Fig. 6  Comparison of the sensitivity and confidence regions at 95% CL for the toy disappearance experiment obtained with T2, T1 and T0. The exclusion and discovery sensitivities (first and third row respectively) have been computed using 10,000 pseudo-data sets. The confidence regions for a concrete data-set have instead been calculated for a pseudo-data set generated under the no-signal hypothesis (second row) and one generated assuming a sterile neutrino signal with $\sin^2(2\theta) = 0.04$ and $\Delta m^2 = 1\text{eV}^2$ (fourth row). The rate (left) and shape analysis (middle) are shown independently and combined together (right column). The probability distribution of the test statistics are computed using Monte Carlo techniques. Results obtained when approximating the distributions with chi-square functions are shown for T2.
power. The discovery sensitivity cannot be calculated as the no-signal hypothesis is always accepted when used as $H_0$.

The confidence regions extracted for specific pseudo-data sets not containing a signal is shown in the second row of Fig. 6. $T_0$ can provide extremely stringent constraints on $\sin^2(2\theta)$ that are orders of magnitudes beyond the sensitivity. To mitigate this behaviour this test statistic is used in combination with the CLs method [44] that penalizes constraints stronger than the sensitivity by introducing an overcoverage in the test. The combination of $T_0$ and the CLs method is known as “Gaussian CLs” [10].

The plots in the fourth row of Fig. 6 show the confidence regions extracted for a pseudo-data set with an injected signal. These regions have two peculiarities. Similarly to $T_1$, the non-standard construction of the confidence region produces an allowed $\sin^2(2\theta)$ interval for each $\Delta m^2$ value. However, differently from $T_1$, the $\sin^2(2\theta)$ intervals are now always connected to the no-signal hypothesis $\sin^2(2\theta)=0$, even for the true $\Delta m^2$ value, as the alternative hypothesis in the test is now fixed to the no-signal hypothesis.

In conclusion, while $T_0$ has a greater power and can produce the strongest limits in terms of $\sin^2(2\theta)$, it produces confidence regions that cannot constrain either of the parameters of interest. We confirm that the probability distribution of $T_0$ converges to a normal distribution for our toy appearance and disappearance experiment as reported in Ref. [10].

6 Comparison and discussion

The main difference among the statistical methods applied to the search for sterile neutrinos has been traced back to the definition of the alternative hypothesis in the hypothesis testing procedure. The considered definitions lead to three different test statistics that are used to construct confidence regions in the $\sin^2(2\theta)$ vs. $\Delta m^2$ parameter space. The sensitivities and confidence regions constructed for each test are compared in Fig. 6.

In $T_2$, the parameter space of the alternative hypothesis covers all possible values of $\sin^2(2\theta)$ and $\Delta m^2$. This test is the natural choice when the values of the parameters of interest are unknown and a generic search over the full parameter space is intended. Using this test for an interval estimation procedure provides naturally two-dimensional confidence regions in the $\sin^2(2\theta)$ vs. $\Delta m^2$ space. The probability distributions of this test statistic are not well approximated by chi-square functions in the analysis of sterile neutrino experiments, and such an approximation can lead to very inaccurate confidence regions and even to erroneously reject the no-signal hypothesis.

In $T_1$, the value of $\Delta m^2$ is assumed to be known prior to the experiment and the parameter space of the alternative hypothesis is restricted to a unique $\Delta m^2$ value. $T_1$ naturally generates one-dimensional confidence regions in the $\sin^2(2\theta)$ space. Two-dimensional confidence regions can be technically created as an union of $\sin^2(2\theta)$ intervals, each computed for a different fixed value of $\Delta m^2$. Such confidence regions have proper coverage but also some pathologies. In particular, while the constraints on $\sin^2(2\theta)$ are more stringent than for $T_2$, the test has no capability to constrain $\Delta m^2$ and the confidence region extends over any $\Delta m^2$ value. The conditions of Wilks’ theorem are almost fulfilled and its probability distribution follows accurately a chi-square function in the parameter space close to the physical border where the probability distribution becomes half a chi-square function and half a delta-Dirac function at zero [43].

The test statistic $T_0$ compares two simple hypotheses with a fixed value of the parameters of interest. The alternative hypothesis is defined as the no-signal hypothesis. Thus, the no-signal hypothesis is accepted by construction when used as $H_0$. The natural confidence regions constructed using this test are point-like. Two-dimensional regions in the $\sin^2(2\theta)$ vs $\Delta m^2$ can be obtained as union of point-like confidence regions, but this non-standard construction produces regions that do not constrain the parameters of interest and only set upper limits on $\sin^2(2\theta)$. These limits are consistently stronger than for $T_1$ and $T_2$ as this test has maximum power. The asymptotic formulas, namely Gaussian distributions, seem to describe well the probability distribution of this test statistic in a large set of conditions.

In summary, all the test statistics are conceptually correct and have a natural scope of application. The more information is available, the more we can restrict the parameter space of the alternative hypotheses and the greater the power of the test becomes. However, some issues arise when these tests are used – regardless of what is their natural scope – to build two-dimensional confidence regions in the $\Delta m^2$ vs. $\sin^2(2\theta)$ space. The regions produced by $T_1$ and $T_0$ do not constrain $\Delta m^2$ as they have an allowed $\sin^2(2\theta)$ interval for any $\Delta m^2$ value. In addition, the regions produced by $T_0$ do not even constrain $\sin^2(2\theta)$ as the $\sin^2(2\theta)$ interval is always connected to $\sin^2(2\theta)=0$, even in the presence of a strong signal. Since the primary goal of the current experiments is to find a signal at unknown $\sin^2(2\theta)$ and $\Delta m^2$ values, we find natural to adopt an analysis that is able to pin down simultaneously both oscillation parameters and recommend the usage of $T_2$.

To ease the comparison of the performance and results from different experiments, it would be convenient for the field to adopt a standardized analysis procedure. Based on the results presented in this article and expanding the proposal of Ref. [8], such a standard analysis could follow these steps:

1. identification of the most likely value for $\sin^2(2\theta)$ and $\Delta m^2$ defined as the value corresponding to the maximum of the likelihood function over the space
\[ \sin^2(2\theta), \Delta m^2 : 0 \leq \sin^2(2\theta) \leq 1, \Delta m^2 \geq 0 \] (i.e., maximum likelihood estimators); 

2. check that the data is compatible with the model corresponding to the most likely value of the parameters of interest by using a “likelihood ratio” goodness-of-fit test [34] whose probability distribution is verified or constructed with Monte Carlo techniques; 

3. construct the two-dimensional confidence region based on $T_2$. If the no-signal hypothesis is accepted, the confidence region will extend down to vanishing $\sin^2(2\theta)$ values and its upper limit can be plotted along with its median value (i.e., the exclusion sensitivity) and 68/95/100% central intervals expected under the no-signal hypothesis (as in Fig. 3). If the no-signal hypothesis is instead rejected, the confidence region can be plotted for different confidence levels along with the discovery sensitivity as in Fig. 4.

When the number of events observed by an experiment is large and each bin of the data set contains tens of counts or more, the Poisson probability in the likelihood function can be approximated by a Gaussian probability. In this case, the likelihood function can be converted into a chi-square function. This treatment can be regarded as a sub-case of what is discussed in the previous sections. However, independently by the number of events, Wilks’ theorem is not valid for $T_2$ because of the presence of physical borders and of the statistical fluctuations in the data sample that mimic a sterile neutrino signature. In addition, the alternative hypothesis becomes independent by $\Delta m^2$ when $\sin^2(2\theta)$ tends to zero. The construction of the test statistic probability distribution through Monte Carlo techniques is hence mandatory in order to ensure accurate results. The Monte Carlo construction is computationally demanding, but it is feasible as proved by the experiments that are already performing it. Indeed, the proposed analysis based on $T_2$ is similar to the one used by e.g. MiniBooNE and PROSPECT.

The inapplicability of Wilks’ theorem and the non-trivial interplay between the rate and shape analysis have repercussions also on the global fits for which the likelihood of each experiment must be combined and the probability distribution of $T_2$ must be computed by generating simultaneously pseudo-data for all the experiments considered. For this reason, it would be useful if the experiments would release in addition to their likelihood fit function and their data, also the probability distributions of the individual signal and background components used in the fit and for the pseudo-data generation.

7 Conclusions

The statistical methods used to search for short-baseline neutrino oscillations induced by an hypothetical sterile neutrino with mass at the eV scale have been reviewed and compared. Three hypothesis testing procedures are used in the field to create confidence intervals. Each procedure is based on a specific test statistic. We identified how two out of the three tests make implicit assumptions on the value of the parameters of interest for sterile neutrinos, i.e. $\Delta m^2$ and $\sin^2(2\theta)$. Making different assumptions changes the question addressed by the test and, consequently, changes the result of the analysis.

For the first time, the performance of the three tests have been compared in a coherent way over a comprehensive set of limit setting and signal discovery scenarios. In particular, we considered both disappearance and appearance experiments as well as rate- and shape-based analyses. For each scenario we constructed the probability distributions of the test statistic using Monte Carlo techniques and found that they can differ significantly from the usual asymptotic approximations. The confidence regions reconstructed by the three tests can be significantly different, making hard to compare the results from experiments that adopt different analyses. Our results are consistent with those obtained in Refs. [8–10] for specific scenarios.

The current generation sterile-neutrino searches aim at finding a signal anywhere in the parameter space available to the experiment. This should naturally lead to an analysis based on a test without any assumption on the value of the oscillation parameters. In this way, the analysis will be able to constrain both the value of $\Delta m^2$ and $\sin^2(2\theta)$ when a signal is observed. Thus, we recommend the use of $T_2$ and the construction of its probability distributions through Monte Carlo techniques.

Acknowledgements The authors are greatly thankful to Hans Niederhauser for the valuable suggestions on how to interpret the statistical methods and structure the comparison. The authors would also like to thank T. Lasserre, P. Litchfield, L. Oberauer, G. Ranucci, M. Wurm and S. Schoenert for many helpful discussions and suggestions during the work and the preparation of the manuscript. This work has been supported by the Deutsche Forschungsgemeinschaft (SFB1258).

Data Availability Statement This manuscript has no associated data or the data will not be deposited. [Authors’ comment: This manuscript has not associated data].

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes
were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Funded by SCOAP3.

Appendix A: The likelihood

The general form of the likelihood used in the analysis of sterile neutrino experiments is given in Sect. 3.1. The computational task of finding the maximum of the likelihood is typically performed by minimizing the negative logarithm of the likelihood (NLL). Moving to the logarithm space is convenient from the computational and numerical point of view. This is one of the reason why the test statistics are defined as the logarithm of the likelihood.

Figure 7 shows the NLL as a function of $\Delta m^2$ and $\sin^2(2\theta)$ for two sets of pseudo-data of the toy disappearance experiment. The two sets of pseudo-data are respectively a realization of the no-signal hypothesis and a realization of a hypothesis $\{\sin^2(2\theta) = 0.04$ and $\Delta m^2 = 1\text{ eV}^2\}$. They are the same data sets used for the comparison of the performance of the statistical methods in the previous sections. Local minima at regularly-spaced $\Delta m^2$ values are present for both data sets. This feature is due to the oscillatory nature of the sought-after signal and appears in any realizations of the data.

The presence of multiple minima makes it difficult for a minimization algorithm to converge to the absolute minimum, in particular for those algorithms relying on the derivative of the function (e.g. the algorithms known as SIMPLEX and MIGRAD in the MINUIT software package [45]). To reliably find the absolute minimum we adopt a scanning approach in which $\Delta m^2$ is increased progressively with uniform steps in the logarithmic space, each step having a length of $\log(\Delta m^2/\text{eV}^2) \lesssim 0.01$. At each $\Delta m^2$ value a minimization against $\sin^2(2\theta)$ is performed. This minimization is not problematic because, when the value of $\Delta m^2$ is fixed, the likelihood function along $\sin^2(2\theta)$ is a smooth function with a unique minimum.

The NLL for the pseudo-data generated under the no-signal hypothesis shows another important feature: the absolute minimum does not correspond to the no-signal hypothesis. This is the case for all the realizations of the data of the toy disappearance experiment. The sought-after oscillatory signature is indeed mimicked by the statistical fluctuations between adjacent bins and the data are described always better by an oscillatory hypothesis than by the no-signal hypothesis. As discussed later in Appendix D, this leads to a deformation of the test statistic probability distribution.

Appendix B: Generation of pseudo-data

The generation of pseudo-data is performed with Monte Carlo techniques. The experimental parameters of the two toy experiments used in this work are quoted in Sect. 2 and Table 1. Pseudo-data for a specific hypothesis $H(X,Y) : \{\sin^2(2\theta), \Delta m^2 : \sin^2(2\theta) = X, \Delta m^2 = Y\}$ are generated according to the probability distribution of signal ($h_{ij}^s(X,Y)$) and background ($h_{ij}^b$) events as a function of the $i$-th bin in $L_{\text{MC}}$ and $j$-th bin $E_{\text{MC}}$. To construct a set of pseudo-data, a Poisson random number is generated for each $ij$ bin using as expectation

$$\lambda_{ij} = \tilde{N}^s(X,Y) \cdot h_{ij}^s(X,Y) + \tilde{N}^b \cdot h_{ij}^b$$

where $\tilde{N}^s(X,Y)$ is the expected number of neutrino interactions under the hypothesis $H(X,Y)$ and $\tilde{N}^b$ is the expected number of background events. For each set of pseudo-data also the external constraints $\tilde{N}^s(X,Y)$ and $\tilde{N}^b$ (see Eq. 6) are
Appendix C: Hypothesis testing and sensitivity

The definition of the set of accepted/rejected hypotheses in this article is based on a Neyman construction in which the ordering principle is defined by the test statistic. The construction follows the steps outlined in this section.

Let \( T \) be any of the three statistics in Table 2. \( T \) is used to test a specific hypothesis \( H(X, Y) \): \{sin\(^2\)(2\(\theta\)), \( \Delta m \)\(^2\) : sin\(^2\)(2\(\theta\)) = \( X \), \( \Delta m \)\(^2\) = \( Y \)\} given a set of data. We will use the symbol \( T_{X,Y} \) to make explicit which hypothesis is being tested. The set of hypotheses accepted or rejected by the test statistic is defined according to how the observed value of the test statistic \( T_{X,Y}^{\text{obs}} \) compares to its critical value \( T_{X,Y}^{\text{crit}} \). To this purpose, the critical value is determined for a predefined test size \( \alpha \) as:

\[
\alpha = \int_{T_{X,Y}^{\text{crit}}}^{\infty} f(T_{X,Y}|H(X,Y)) \, dT_{X,Y}
\]

The probability distributions \( f(T_{X,Y}|H(X,Y)) \) are constructed by computing the test statistic for each data set in ensembles of 10\(^5\) pseudo-data sets produced under the true hypothesis \( H(X,Y) \).

For each hypothesis to be tested the following steps are carried out:

1. compute \( T_{X,Y}^{\text{crit}} \) given the specific probability distribution for \( H(X,Y) \) and a test size of 0.05;
2. \( T_{X,Y}^{\text{obs}} \) is evaluated for the considered data set;
3. the hypothesis is accepted if \( T_{X,Y}^{\text{obs}} \leq T_{X,Y}^{\text{crit}} \), otherwise the hypothesis is rejected.

Two definitions of sensitivity have been used in this article: the exclusion sensitivity that defines the set of hypotheses that would be rejected assuming the no-signal hypothesis, and the discovery sensitivity that defines the set of hypotheses for which we expect to reject the no-signal hypothesis. The formal definition of these two quantities is based on the median expected value of the test statistic under certain hypotheses that can be calculated using an ensemble of pseudo-data.

The exclusion sensitivity is defined as the set of hypotheses for which the test is expected to provide a median test statistic value exactly of \( T_{X,Y}^{\text{crit}} \) assuming the data were generated from the no-signal hypothesis. To find the set of hypotheses fulfilling this requirement, the following steps are carried out for each \( H(X,Y) \):

1. compute \( T_{X,Y}^{\text{crit}} \) given the specific probability distribution for \( H(X,Y) \) and a test size of 0.05;
2. compute \( \text{med}[T_{X,Y}|H(0,0)] \), i.e. the median value of \( f(T_{X,Y}|H(0,0)) \);

Since the set of hypotheses considered is usually discrete, finding a hypothesis that meets perfectly the condition above is not possible. The requirement is hence softened: for a given \( \Delta m \)\(^2\) value, the hypotheses are tested for increasing values of sin\(^2\)(2\(\theta\)) and the first one fulfilling the condition \( \text{med}[T_{X,Y}|H(0,0)] > T_{X,Y}^{\text{crit}} \) is added to the set.

The discovery sensitivity is defined as the set of hypotheses that, if true, would result in a rejection of the no-signal hypothesis with a probability of 50%. To find the set of hypotheses fulfilling this requirement, the following steps are carried out for each \( H(X,Y) \):

1. \( T_{0,0}^{\text{crit}} \) is computed given the specific probability distribution for \( H(0,0) \) and a test size of 0.05;
2. compute \( \text{med}[T_{0,0}|H(X,Y)] \), i.e. the median value of \( f(T_{0,0}|H(X,Y)) \);

Similarly to the previous case, a softer requirement is typically applied when considering a discrete set of hypotheses.

Appendix D: Probability distributions of \( T_2 \)

Figure 8 shows the probability distributions of \( T_2 \) computed from an ensemble of pseudo-data generated assuming the no-signal hypothesis and their expected asymptotic formulas. The distributions are shown for the toy disappearance and appearance experiments, assuming a rate analysis, a shape analysis, and their combination.

The top panels show the distributions obtained when testing the no-signal hypothesis (i.e. testing the same hypothesis used to generate the pseudo-data). In this case, according to
Fig. 8 Distributions of $T_2$ computed from an ensemble of pseudo-data generated assuming the no-signal hypothesis and their asymptotic formulas. The distributions are shown for a test of the no-signal hypothesis (panel a and b using $10^7$ sets of pseudo-data) and of an oscillation hypothesis (panel c and d using $10^4$ sets of pseudo-data). The oscillation hypothesis has $\Delta m^2=1\,\text{eV}^2$ and $\sin^2(2\theta)=0.002/0.02/0.07$, respectively for the appearance/disappearance(shape&combined)/disappearance(rate) experiment. The value of the angle is chosen such that the hypothesis is very close to the experimental sensitivity for each kind of analysis.

Wilks’ theorem, the distributions should tend asymptotically to a chi-square function. However, considering that $\sin^2(2\theta)$ is bounded to positive values, the distributions are expected to be described by a half chi-square [43]. The bottom panels show the distributions obtained when testing an oscillation hypothesis (i.e. testing an hypothesis different from the one used to generate the pseudo-data). In this case, according to Wald [46], the distributions should tend asymptotically to a non-central chi-square. In both cases the number of degrees of freedom of the chi-square function is given by the difference between the number of free parameters in the alternative and the null hypothesis. The non-centrality parameter is defined through the Asimov data set as discussed in Ref. [42].

The distributions for the rate analysis are to a first approximation described by the asymptotic formulas. On the contrary, the distributions for the shape analysis and the combination of rate and shape differ significantly and are deformed towards higher values of the test statistic. This is due to the fact that the sought-after oscillatory signature is often preferred to the no-oscillation hypothesis and the best fit does not correspond to the true hypothesis. In other words, the extra flexibility of a model with oscillations can always be used to better describe the data and the statistical fluctuations between bins.

The exclusion sensitivities based on the asymptotic formulas are compared to those based on the distributions constructed through Monte Carlo techniques in Fig. 9. The approximated sensitivities calculated using the asymptotic formulas are accurate within 10%. This might seem inconsistent with the fact that the test statistic distributions are not well described by the asymptotic formulas. However, the distributions are consistently shifted at higher values for all possible hypotheses and this coherent bias preserves the relationship between the distribution quantiles. This might not be valid if the sensitivity is computed for an other confidence level (i.e. an other quantile of the distribution) or if the experimental parameters are changed.
Appendix E: Look-elsewhere effect

The statistical significance associated to the observation of a sterile neutrino signal can be expressed through the p-value computed for the no-signal hypothesis given the observed data:

\[
p = \int_{T_{0,0}}^{\infty} f(T_{0,0}|H(0,0)) \, dT_{0,0}
\]

While \( T_2 \) provides a unique p-value for the no-signal hypothesis, \( T_1 \) provides a p-value for each tested value of \( \Delta m^2 \) (see Fig. 4). For this reason the p-value provided by \( T_1 \) is often called local p-value, while the p-value provided by \( T_2 \) is called global p-value. A procedure to estimate the global p-value using the local estimation has been proposed in Ref. [37] based on previous results from Davis [47]. The procedure is based on a linear correction of the minimum local p-value found:

\[
P_{\text{global}} \approx \min_{\Delta m^2} P_{\text{local}} + \langle N_u \rangle
\]

where \( \langle N_u \rangle \) is the mean number of “upcrossings” above the level \( u \) in the range of considered \( \Delta m^2 \) values for a test of the no-signal hypothesis based on \( T_1 \). Each upcrossing corresponds to a \( \Delta m^2 \) value for which the signal hypothesis is preferred over the no-signal hypothesis at a certain level \( u \). The mean number of upcrossings above the level \( u \) and lower level \( u_0 \) is connected by the relationship \( \langle N_u \rangle = \langle N_{u_0} \rangle e^{-(u-u_0)/2} \). \( \langle N_{u_0} \rangle \) can be estimated from a small ensemble of pseudo-data. The possibility of using a small ensemble is convenient because if \( u_0 \) is small the number of upcrossings per data set becomes large. The approximation becomes valid for \( u \to \infty \), an upper limit on the p-value is given otherwise.

Computing a global p-value from a local estimation could be interesting for sterile neutrino searches as the local p-value construction can be performed assuming the asymptotic formula for the probability distributions of \( T_1 \). The global p-value estimated through the correction discussed above is shown in Fig. 10 for the toy disappearance experiment. The p-value computed from \( T_2 \) using an ensemble of pseudo-
data is also shown. The distributions have the same shape for large values of the test statistic but a different offset. The offset implies that the p-value extrapolated from $T_1$ would be overestimated by a factor of 1.4. This factor is not constant and depends on the specific experiment.

While a correction can provide accurate results in problems such as a peak search, the oscillatory nature of the signal sought-after by sterile neutrino experiments induces a correlation in the number of upcrossings. Such a correlation is due to the harmonics in Eqs. (1) and (2) occurring at different values of $\Delta m^2$. It might be possible to correct the number of upcrossings to account for the spurious occurrences but this requires additional studies.

References

1. C. Giunti, T. Lasserre, Rev. Nucl. Part. Sci. 69, 163–190 (2019). https://doi.org/10.1146/annurev-nucl-101918-023755Ann
2. B. Pontecorvo, Neutrino experiments and the problem of conservation of leptonic charge. Sov. Phys. JETP 26, 984–988 (1968)
3. A. Boyarsky, M. Drewes, T. Lasserre, S. Mertens, O. Ruchayskiy, Sterile neutrino dark matter. Prog. Part. Nucl. Phys. 104, 1–45 (2019). https://doi.org/10.1016/j.ppnp.2018.07.004
4. A. Aguilar-Arevalo et al., (LSND Collaboration), Evidence for neutrino oscillations from the observation of $\nu_e$ appearance in a $\nu_x$ beam. Phys. Rev. D 64, 112007 (2001). https://doi.org/10.1103/PhysRevD.64.112007
5. G. Mention, M. Fechner, T. Lasserre, T. Mueller, D. Lhuillier, M. Cribier et al., The reactor antineutrino anomaly. Phys. Rev. D 83, 073006 (2011). https://doi.org/10.1103/PhysRevD.83.073006
6. C. Giunti, M. Laveder, Y. Li, Q. Liu, H. Long, Update of short-baseline electron neutrino and antineutrino disappearance. Phys. Rev. D 86, 113014 (2012). https://doi.org/10.1103/PhysRevD.86.113014
7. M. Dentler, A. Hernandez-Cabezudo, J. Kopp, P.A. Machado, M. Maltoni, I. Martinez-Soler et al., Updated global analysis of neutrino oscillations in the presence of eV-scale sterile neutrinos. JHEP 08, 010 (2018). https://doi.org/10.1007/JHEP08(2018)010
8. G.J. Feldman, R.D. Cousins, A unified approach to the classical statistical analysis of small signals. Phys. Rev. D 57, 3873–3889 (1998). https://doi.org/10.1103/PhysRevD.57.3873
9. L. Lyons, Raster scan or 2-D approach? arXiv:1404.7395
10. X. Qian, A. Tan, J.J. Ling, Y. Nakajima, C. Zhang, The Gaussian CL$_{s}$ method for searches of new physics. Nucl. Instrum. Methods A 827, 63–76 (2016). https://doi.org/10.1016/j.nima.2016.04.089
11. M. Tanabashi et al., (Particle Data Group Collaboration), Rev. Part. Phys. Rev. D 98, 030001 (2018). https://doi.org/10.1103/PhysRevD.98.030001
12. I. Alekseev et al., (DANSS Collaboration), Search for sterile neutrinos at the DANSS experiment. Phys. Lett. B 787, 56–63 (2018). https://doi.org/10.1016/j.physletb.2018.10.038
13. Y. Ko et al., (NEOS Collaboration), Sterile neutrino search at the NEOS experiment. Phys. Rev. Lett. 118, 121802 (2017). https://doi.org/10.1103/PhysRevLett.118.121802
14. A. Serebrov et al., (NEUTRINO-4 Collaboration), First observation of the oscillation effect in the neutrino-4 experiment on the search for the sterile neutrino. Pisma Zh. Eksp. Teor. Fiz. 109, 209–218 (2019). https://doi.org/10.1134/S0232134619040040
15. J. Ashenfelter et al., (PROSPECT Collaboration), First search for short-baseline neutrino oscillations at HFIR with PROSPECT. Phys. Rev. Lett. 121, 251802 (2018). https://doi.org/10.1103/PhysRevLett.121.251802
16. Y. Abreu et al. (SoLid Collaboration), SoLid: A short baseline reactor neutrino experiment. arXiv:2002.05914
17. H. Almazan et al., (STEREO Collaboration), Sterile neutrino constraints from the STEREO experiment with 66 days of reactor-on data. Phys. Rev. Lett. 121, 161801 (2018). https://doi.org/10.1103/PhysRevLett.121.161801
18. V. Barinov, V. Gavrin, V. Gorbatchev, D. Gorbunov, T. Ibragimova, BEST potential in testing the eV-scale sterile neutrino explanation of reactor antineutrino anomalies. Phys. Rev. D 99, 111702 (2019). https://doi.org/10.1103/PhysRevD.99.111702
19. G. Bellini et al., (Borexino Collaboration), SOX: Short distance neutrino oscillations with Borexino. JHEP 08, 038 (2013). https://doi.org/10.1007/JHEP08(2013)038
20. J. Gaffiot et al., Experimental parameters for a cerium 144 based intense electron antineutrino generator experiment at very short baselines. Phys. Rev. D 91, 072005 (2015). https://doi.org/10.1103/PhysRevD.91.072005
21. S. Ajimura et al., Technical Design Report (TDR): Searching for a sterile neutrino at J-PARC MLF (E56, JSNS2). arXiv:1705.08629
22. A. Aguilar-Arevalo et al., (MiniBooNE Collaboration), Significant excess of electronlike events in the MiniBooNE short-baseline neutrino experiment. Phys. Rev. Lett. 121, 221801 (2018). arXiv:1805.12028
23. M. Antonello et al. (MINICORE Collaboration), A proposal for a three detector short-baseline neutrino oscillation program in the Fermilab booster neutrino beam. arXiv:1503.01520
24. F. P. An et al., (Daya Bay Collaboration), Improved search for a light sterile neutrino with the full configuration of the Daya Bay experiment. Phys. Rev. Lett. 117, 151802 (2016). https://doi.org/10.1103/PhysRevLett.117.151802
25. D. Hellwig and T. Matsubara (DOUBLE CHOOZ Collaboration), Sterile neutrino search with the double Chooz experiment. J. Phys. Conf. Ser. 888 (2017) 012133. https://doi.org/10.1088/1742-6596/888/1/012133
26. I. S. Yeo (RENO Collaboration), Search for sterile neutrinos at RENO. J. Phys. Conf. Ser. 888 (2017) 012139. https://doi.org/10.1088/1742-6596/888/1/012139
27. P. Adamson et al., (MINOS+ Collaboration), Search for sterile neutrinos in MINOS and MINOS+ using a two-detector fit. Phys. Rev. Lett. 122, 091803 (2019)
28. P. Adamson et al. (NOvA Collaboration), Search for active-sterile neutrino mixing using neutral-current interactions in NOvA. Phys. Rev. D 96, 072006 (2017). https://doi.org/10.1103/PhysRevD.96.072006
29. M. Aartsen et al., (IceCube Collaboration), Search for sterile neutrino mixing using three years of IceCube DeepCore data. Phys. Rev. D 95, 112002 (2017). https://doi.org/10.1103/PhysRevD.95.112002
30. N. Aghanim et al. (Planck Collaboration), Planck 2018 results. VI. Cosmological parameters. arXiv:1807.06209
31. J. Angrik et al. (KATRIN Collaboration), KATRIN design report 2004, FZKA-7090 (2005)
32. L. Gastaldo, C. Giunti, E. Zavanin, Light sterile neutrino sensitivity baseline neutrino experiment. Phys. Rev. Lett. 113014 (2012). https://doi.org/10.1103/PhysRevLett.113014
35. G. Casella, R.L. Berger, *Statistical Inference*, 2nd edn. (Thomson Learning, Pacific Grove, 2002)
36. ATLAS, CMS, LHC HIGGS COMBINATION GROUP, Procedure for the LHC Higgs boson search combination in summer 2011, ATL-PHYS-PUB-2011-011, CMS-NOTE-2011-005 (2011)
37. E. Gross, O. Vitells, Trial factors for the look elsewhere effect in high energy physics. Eur. Phys. J. C **70**, 525–530 (2010). https://doi.org/10.1140/epjc/s10052-010-1470-8
38. G. Ranucci, The profile likelihood ratio and the look elsewhere effect in high energy physics. Nucl. Instrum. Methods A **661**, 77–85 (2012). https://doi.org/10.1016/j.nima.2011.09.047
39. S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. **9**, 60–62 (1938). https://doi.org/10.1214/aoms/1177732360
40. A. Stuart, K. Ord, *Kendall’s Advanced Theory of Statistics: Distribution Theory*, vol. 1 (Wiley, New York, 2009)
41. S. Algeri, J. Aalbers, K. Dundas Mora, J. Conrad, Searching for new physics with profile likelihoods: Wilks and beyond. arXiv:1911.10237
42. G. Cowan, K. Cranmer, E. Gross, O. Vitells, Asymptotic formulae for likelihood-based tests of new physics. Eur. Phys. J. C **71**, 1554 (2011). https://doi.org/10.1140/epjc/s10052-011-1554-0
43. H. Chernoff, On the distribution of the likelihood ratio. Ann. Math. Stat. **25**, 573–578 (1954). https://doi.org/10.1214/aoms/1177728725
44. A. L. Read, Modified frequentist analysis of search results (The CL(s) method), in Workshop on confidence limits, CERN, Geneva, Switzerland, 17–18 Jan 2000: Proceedings, pp. 81–101, 8, (2000)
45. F. James, M. Roos, Minuit: A system for function minimization and analysis of the parameter errors and correlations. Comput. Phys. Commun. **10**, 343–367 (1975). https://doi.org/10.1016/0010-4655(75)90039-9
46. A. Wald, Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Am. Math. Soc. **54**, 426–482 (1943)
47. R.B. Davies, Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika **74**, 33–43 (1987). https://doi.org/10.1093/biomet/74.1.33