FERMILAB-PUB-22-476-ND

The Profiled Feldman–Cousins technique for confidence interval construction in the presence of nuisance parameters

M. A. Acero,[2] B. Acharya,[31] P. Adamson,[12] L. Aliaga,[12] N. Anfimov,[25] A. Antoshkin,[25] E. Arrieta-Diaz,[27] L. Asquith,[39] A. Aurisano,[6] A. Back,[19, 23] C. Backhouse,[43] M. Baird,[44] N. Balashov,[25] P. Baldi,[24] B. A. Bambah,[16] S. Bashar,[42] A. Bat,[11] K. Bays,[4, 18] R. Bernstein,[12] V. Bhatnagar,[33] D. Bhattarai,[31] B. Bhuyan,[14] J. Bian,[24, 30] A. C. Booth,[35, 39] R. Bowles,[19] B. Brahma,[17] C. Bromberg,[28] N. Buchanan,[8] A. Butkevich,[21] S. Calvez,[8] T. J. Carroll,[41, 47] E. Catano-Mur,[46] A. Chatla,[16] R. Chirco,[18] B. C. Choudhary,[10] S. Choudhary,[14] A. Christensen,[8] T. E. Coan,[38] M. Colo,[46] L. Cremonesi,[35] G. S. Davies,[31, 19] P. F. Derwent,[12] P. Ding,[12] Z. Djurcic,[1] M. Dolce,[42] D. Doyle,[8] D. Dueñas Tonguino,[6] E. C. Dukes,[44] A. Dye,[31] R. Ehrlich,[44] M. Elkins,[23] E. Ewart,[19] G. J. Feldman,[48] P. Filip,[22] J. Franc,[9] M. J. Frank,[36] H. R. Gallagher,[42] R. Gandrajula,[28, 44] F. Gao,[34] A. Giri,[17] R. A. Gomes,[13] M. C. Goodman,[1] V. Grichine,[26] M. Groh,[8, 19] R. Group,[44] B. Guo,[37] A. Habig,[29] F. Hakl,[20] A. Hall,[44] J. Hartnell,[39] R. Hatcher,[12] H. Hausner,[47] M. He,[15] K. Heller,[30] V Hewes,[6] A. Himmel,[12] B. Jargowsky,[24] J. Jarosz,[8] F. Jediny,[9] C. Johnson,[8] M. Judah,[8, 34] I. Kakorin,[25] D. M. Kaplan,[18] A. Kalitkina,[25] J. Kleykamp,[31] O. Klimov,[25] L. W. Koerner,[15] L. Kolupaeva,[25] S. Kotelnikov,[26] R. Kralik,[39] Ch. Kullenberg,[25] M. Kubu,[9] A. Kumar,[33] C. D. Kuruppu,[37] V. Kus,[9] T. Lackey,[12, 19] K. Lang,[41] P. Lasorak,[39] J. Lesmeister,[15] S. Lin,[8] A. Lister,[47] J. Liu,[24] M. Lokajicek,[22] J. M. C. Lopez,[43] R. Mahji,[16] S. Magill,[1] M. Manrique Plata,[19] W. A. Mann,[42] M. T. Manoharan,[7] M. L. Marshak,[30] M. Martinez-Casales,[23] V. Matveev,[21] B. Mayes,[39] B. Mehta,[33] M. D. Messier,[19] H. Meyer,[45] T. Miao,[12] V. Mikola,[43] W. H. Miller,[30] S. Mishra,[3] S. R. Mishra,[37] A. Mislivec,[30] R. Mohanta,[16] A. Moren,[29] A. Morozova,[25] W. Mu,[12] L. Mualem,[4] M. Muether,[45] K. Mulder,[43] D. Naples,[34] A. Nath,[14] N. Nayak,[24] S. Nelleri,[7] J. K. Nelson,[46] R. Nichol,[43] E. Niner,[12] A. Norman,[12] A. Norrick,[12] T. Nosek,[5] H. Oh,[6] A. Olshevskiy,[25] T. Olson,[42] J. Ott,[24] A. Pal,[32] J. Paley,[12] L. Panda,[32] R. B. Patterson,[4] G. Pawloski,[30] D. Pershey,[4] O. Petrova,[25] R. Petti,[37] D. D. Phan,[41, 43] R. K. Plunkett,[12] A. Pobedimov,[25] J. C. C. Porter,[39] A. Rafique,[1] L. R. Prais,[31] V. Raj,[4] M. Rajaoalisoa,[6] B. Ramson,[12] B. Rebel,[12, 47] P. Rojas,[8] P. Roy,[45] V. Ryabov,[26] O. Samoylov,[25] M. C. Sanchez,[23] S. Sánchez Falero,[23] P. Shanahan,[12] P. Sharma,[33] S. Shukla,[3] A. Sheshukov,[25] I. Singh,[10] P. Singh,[35, 10] V. Singh,[3] E. Smith,[19] J. Smolik,[9] P. Snopok,[18] N. Solomey,[45] A. Sousa,[6] K. Soustruznik,[5] M. Strait,[30] L. Suter,[12] A. Sutton,[44] S. Swain,[32] C. Sweeney,[43] A. Sztuc,[43] B. Tapia Oregui,[41] P. Tas,[5] B. N. Temizel,[18] T. Thakore,[6] R. B. Thayyullathil,[7] J. Thomas,[43, 47] E. Tiras,[11, 23] J. Tripathi,[33] J. Trokan-Tenorio,[46] Y. Torun,[18] J. Urheim,[19] P. Vahle,[46] Z. Vallari,[4] J. Vasel,[19] T. Vrba,[9] M. Wallbank,[6] T. K. Warburton,[23] M. Wetstein,[23] D. Whittington,[40, 19] D. A. Wickremasinghe,[12] T. Wieber,[30] J. Wolcott,[42] M. Wrobel,[8] W. Wu,[24] Y. Xiao,[24] B. Yaeggy,[6] A. Yallappa Dombara,[40] A. Yankelevich,[24] K. Yonehara,[12] S. Yu,[1, 18] Y. Yu,[18] S. Zadorozhnyy,[21] J. Zalesak,[22] Y. Zhang,[39] and R. Zwaska[12]

(The NOvA Collaboration)

[1] *Argonne National Laboratory, Argonne, Illinois 60439, USA*
[2] *Universidad del Atlantico, Carrera 30 No. 8-49, Puerto Colombia, Atlantico, Colombia*
[3] *Department of Physics, Institute of Science, Banaras Hindu University, Varanasi, 221 005, India*
[4] *California Institute of Technology, Pasadena, California 91125, USA*
[5] *Charles University, Faculty of Mathematics and Physics, Institute of Particle and Nuclear Physics, Prague, Czech Republic*
[6] *Department of Physics, University of Cincinnati, Cincinnati, Ohio 45221, USA*
[7] *Department of Physics, Cochin University of Science and Technology, Kochi 682 022, India*
[8] *Department of Physics, Colorado State University, Fort Collins, CO 80523-1875, USA*
[9] *Czech Technical University in Prague, Brehova 7, 115 19 Prague 1, Czech Republic*
[10] *Department of Physics and Astrophysics, University of Delhi, Delhi 110007, India*
[11] *Department of Physics, Erciyes University, Kayseri 38030, Turkey*
[12] *Fermi National Accelerator Laboratory, Batavia, Illinois 60510, USA*
[13] *Instituto de Física, Universidade Federal de Goiás, Goiânia, Goiás, 74690-900, Brazil*
[14] *Department of Physics, IIT Guwahati, Guwahati, 781 039, India*
[15] *Department of Physics, University of Houston, Houston, Texas 77204, USA*
[16] *School of Physics, University of Hyderabad, Hyderabad, 500 046, India*
[17] *Department of Physics, IIT Hyderabad, Hyderabad, 502 205, India*
[18] *Illinois Institute of Technology, Chicago IL 60616, USA*
[19] *Indiana University, Bloomington, Indiana 47405, USA*
[20] *Institute of Computer Science, The Czech Academy of Sciences, 182 07 Prague, Czech Republic*
[21] *Institute for Nuclear Research of Russia, Academy of Sciences 7a, 60th October Anniversary prospect, Moscow 117312, Russia*
[22] *Institute of Physics, The Czech Academy of Sciences, 182 21 Prague, Czech Republic*
[23] *Department of Physics and Astronomy, Iowa State University, Ames, Iowa 50011, USA*
[24] *Department of Physics and Astronomy, University of California at Irvine, Irvine, California 92697, USA*

This document was prepared by NOvA Collaboration using the resources of the Fermi National Accelerator Laboratory (Fermilab), a U.S. Department of Energy, Office of Science, HEP User Facility. Fermilab is managed by Fermi Research Alliance, LLC (FRA), acting under Contract No. DE-AC02-07CH11359.

[25] *Joint Institute for Nuclear Research, Dubna, Moscow region 141980, Russia*
[26] *Nuclear Physics and Astrophysics Division, Lebedev Physical Institute, Leninsky Prospect 53, 119991 Moscow, Russia*
[27] *Universidad del Magdalena, Carrera 32 No 22-08 Santa Marta, Colombia*
[28] *Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824, USA*
[29] *Department of Physics and Astronomy, University of Minnesota Duluth, Duluth, Minnesota 55812, USA*
[30] *School of Physics and Astronomy, University of Minnesota Twin Cities, Minneapolis, Minnesota 55455, USA*
[31] *University of Mississippi, University, Mississippi 38677, USA*
[32] *National Institute of Science Education and Research, Khurda, 752050, Odisha, India*
[33] *Department of Physics, Panjab University, Chandigarh, 160 014, India*
[34] *Department of Physics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA*
[35] *Particle Physics Research Centre, Department of Physics and Astronomy,*
Queen Mary University of London, London E1 4NS, United Kingdom
[36] *Department of Physics, University of South Alabama, Mobile, Alabama 36688, USA*
[37] *Department of Physics and Astronomy, University of South Carolina, Columbia, South Carolina 29208, USA*
[38] *Department of Physics, Southern Methodist University, Dallas, Texas 75275, USA*
[39] *Department of Physics and Astronomy, University of Sussex, Falmer, Brighton BN1 9QH, United Kingdom*
[40] *Department of Physics, Syracuse University, Syracuse NY 13210, USA*
[41] *Department of Physics, University of Texas at Austin, Austin, Texas 78712, USA*
[42] *Department of Physics and Astronomy, Tufts University, Medford, Massachusetts 02155, USA*
[43] *Physics and Astronomy Department, University College London, Gower Street, London WC1E 6BT, United Kingdom*
[44] *Department of Physics, University of Virginia, Charlottesville, Virginia 22904, USA*
[45] *Department of Mathematics, Statistics, and Physics, Wichita State University, Wichita, Kansas 67206, USA*
[46] *Department of Physics, William & Mary, Williamsburg, Virginia 23187, USA*
[47] *Department of Physics, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA*
[48] *Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*

(Dated: June 21, 2022)

Measuring observables to constrain models using maximum-likelihood estimation is fundamental to many physics experiments. The Profiled Feldman–Cousins method described here is a potential solution to common challenges faced in constructing accurate confidence intervals: small datasets, bounded parameters, and the need to properly handle nuisance parameters. This method achieves more accurate frequentist coverage than other methods in use, and is generally applicable to the problem of parameter estimation in neutrino oscillations and similar measurements. We describe an implementation of this method in the context of the NOvA experiment.

I. INTRODUCTION

The main goal of many physics experiments is to make measurements of the properties of Nature in the form of parameters of a model. Often, those parameters cannot be observed directly, and must instead be inferred from a likelihood function, $\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta})$, which describes the probability of the observed data, $\boldsymbol{x}$, for a given set of parameter values, $\boldsymbol{\theta}$. In frequentist analyses, the best estimate for the model parameters is determined using maximum likelihood estimation. Results are usually [1] presented as one- or two-dimensional Neyman–constructed confidence intervals [2], and Wilks' theorem [3] is used to determine the confidence level which corresponds to a given likelihood value. However, Wilks' theorem is only valid if certain conditions are met, so some experimental measurements that depend on Wilks' theorem may fail to produce confidence intervals with proper frequentist 'coverage,' meaning that confidence intervals determined in the same way in many repeated experiments would not contain the true value with the reported frequency. In other words, the confidence intervals would have an actual significance different from what is reported. The Unified Approach, or more commonly in particle physics the 'Feldman–Cousins' (FC) method[1], defines a nonparametric ordering procedure for determining the critical values that define the extent of the confidence intervals. It is especially useful in situations where Wilks' theorem does not apply [4]. However, it does not give guidance on how to handle additional nuisance parameters beyond those being measured. Ensuring proper coverage in the presence of nuisance parameters is a challenge. No method can guarantee correct coverage for all possible values of the nuisance parameters, but various approaches can give more or less accurate coverage. This paper presents a technique, based on [10], that extends the Feldman–Cousins method to produce confidence intervals with accurate coverage in the presence of nuisance parameters, hereinafter referred to as 'Profiled Feldman–Cousins' or 'Profiled FC.' While deployed in the context of particular measurements made by the NOvA experiment [5–8], this method

[1] The method is named after the authors who introduced it to high energy physics, though it was previously described in [10].

is sufficiently general to apply to a range of measurements that fail to satisfy the assumptions of Wilks' theorem in similar ways.

This paper is divided into two main sections. Section II briefly introduces the Feldman–Cousins method and describes the challenge posed by nuisance parameters, defines the Profiled FC method, and compares its performance to alternative methods in a toy model inspired by neutrino oscillations. Section III takes the NOvA neutrino oscillation measurement as an example to demonstrate the implementation of this method in practice, including some methods used to validate coverage, and important features of the confidence intervals produced in this way.

II. THE PROFILED FELDMAN–COUSINS METHOD

A. The Original Feldman–Cousins Method

The most commonly used method for drawing frequentist confidence intervals is the Neyman construction [2]. Likelihood–ratio tests are performed between each point in parameter space and the best fit point, with test statistic λ defined as:

$$\lambda_i = -2\ln\frac{\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}_i)}{\mathcal{L}(\boldsymbol{x}|\hat{\boldsymbol{\theta}})} = \ell(\boldsymbol{x}|\boldsymbol{\theta}_i) - \ell(\boldsymbol{x}|\hat{\boldsymbol{\theta}}), \tag{1}$$

where $\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta})$ is the likelihood function of data $\boldsymbol{x}$ given parameter values $\boldsymbol{\theta}$, ℓ is $-2\ln\mathcal{L}$, $\boldsymbol{\theta}_i$ is the i^{th} set of fixed values of the parameters being tested for potential inclusion in the confidence interval, and $\hat{\boldsymbol{\theta}}$ is the overall maximum likelihood estimate, hereinafter referred to as 'best fit,' of all parameters to the data. Point i is included in the α-level confidence interval if the p-value from the likelihood ratio test is less than $1-\alpha$, or equivalently, if λ_i is less than a 'critical value,' c_α, given by:

$$\int_0^{c_\alpha} P(\lambda_i)d\lambda_i = \alpha, \tag{2}$$

where P is the expected distribution of the λ_i statistic assuming the true $\boldsymbol{\theta} = \boldsymbol{\theta}_i$. As can be seen from Equation 2, calculating the critical value requires knowledge of the distribution of the likelihood–ratio test statistic.

If the conditions of Wilks' theorem [3] are met, then the distribution $P(\lambda)$ asymptotically approaches a χ^2 distribution with a number of degrees of freedom equal to the number of parameters of interest[2] with deviations expected at the $\mathcal{O}(1/\sqrt{N})$ level, where N refers to the size of the data sample, $\boldsymbol{x}$. This asymptotic behavior means $P(\lambda)$ is the same for any point, i. Since the χ^2 distributions are well known, fixed critical values for drawing confidence intervals at common significance levels are tabulated and readily available.

The conditions required for Wilks' theorem to apply are: (1) the maximum likelihood estimators of the parameters have ellipsoidal distributions, and (2) the null hypothesis is 'nested' within the range of alternative hypotheses. The most common way to violate assumption (1) is a physical boundary on the allowed values of a parameter applied externally (e.g., probabilities must be between 0 and 1), but it can also be violated by an effective boundary introduced by a function with a limited range such as sin(), or degeneracies that add additional allowed regions to the estimator[3]. For the theorem to be useful in practice we also require (3) the size of the data sample, $\boldsymbol{x}$, is sufficiently large that neglecting $\mathcal{O}(1/\sqrt{N})$ deviations from the χ^2 distribution is an acceptable approximation. Many experiments of interest, including the NOvA oscillation measurement, as explained in more detail in Section III A, violate these assumptions in several ways. In such cases, another method must be used to determine suitable critical values. When the assumptions of Wilks' theorem are not satisfied, the significance of the hypothesis tests cannot be reliably determined using the χ^2 distribution, meaning the associated confidence intervals will not have the correct coverage for their reported significance. However, the likelihood-ratio test itself remains valid and optimal per the Neyman–Pearson lemma [9].

The Feldman–Cousins (FC) method [4] provides a nonparametric approach to defining confidence intervals with correct coverage and is commonly used in particle physics. A large number, N, of FC pseudoexperiments are simulated at points sampling the range of parameter values where confidence intervals will be reported. A 'Feldman–Cousins pseudoexperiment' represents a possible experimental observation at a given set of parameters, $\boldsymbol{\theta}$. Each pseudoexperiment is constructed by drawing a Poisson-distributed random number for each bin of our analysis samples, with

[2] The number of parameters of interest is equivalent to the difference in number of degrees-of-freedom between the two likelihoods in the likelihood ratio.

[3] Degeneracies act like 'inverse' boundaries since they add freedom to the estimator rather than constraining it.

the mean of those Poisson distributions being the predicted number of events in that bin given $\boldsymbol{\theta}$. For each FC pseudoexperiment, $\boldsymbol{x}_j$, the best fit of the parameter(s), $\hat{\boldsymbol{\theta}}_j$, is also found through Maximum Likelihood Estimation. The FC pseudoexperiments are then ordered by the difference in ℓ between the 'true' value used to generate the FC pseudoexperiments and the best fit,

$$\lambda_{ij} = \ell(\boldsymbol{x}_j|\boldsymbol{\theta}_i) - \ell(\boldsymbol{x}_j|\hat{\boldsymbol{\theta}}_j), \tag{3}$$

to form a distribution $P(\lambda_i)$ that differs for every $\boldsymbol{\theta}_i$. This procedure is called 'nonparametric' since the ordering of the pseudoexperiments creates a distribution for the test statistic, λ_i, without knowing in advance how it should be distributed. Then, the α-significance-level critical value for this set of true parameters, $c_\alpha(\boldsymbol{\theta}_i)$ as defined in Equation 2, is the value which is larger than the first αN of the λ_{ij} values. This procedure is then repeated for each point being tested, and the confidence interval at level α is made up of the points where $\lambda_i < c_\alpha(\boldsymbol{\theta}_i)$. If the FC pseudoexperiments are a fair representation of the data, it is straightforward to see that this procedure will give correct coverage, α, since we have empirically determined for each point in parameter space the critical value $c_\alpha(\boldsymbol{\theta}_i)$ which will cover α fraction of the pseudoexperiments generated with values $\boldsymbol{\theta}_i$.

B. The Challenge of Nuisance Parameters

While the above procedure is straightforward, it does not provide guidance on a key question when applying it in practice: how to handle nuisance parameters. We use the term 'nuisance parameters' (hereinafter referred to by $\boldsymbol{\phi}$ to distinguish them from the parameters of interest, $\boldsymbol{\theta}$) to refer to any model parameter that we do not wish to include in the specification of our final confidence intervals. These can be parameters the experiment is measuring, but whose constraints are not reported, other parameters of the model which are constrained by external experiments, or parameters representing systematic uncertainties, whose exact values are uninteresting.

The usual frequentist prescription for handling nuisance parameters is to 'profile' over them [10]. That is, at each point in the parameter space, $\boldsymbol{\theta}_i$, at which the likelihood is to be evaluated, a search is performed over all values of the nuisance parameters, and the combination of nuisance parameters that yield the maximum likelihood (minimum ℓ),

$$\hat{\hat{\boldsymbol{\phi}}}_i = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\, \ell(\boldsymbol{\theta}_i, \boldsymbol{\phi}), \tag{4}$$

is adopted. $\hat{\hat{\boldsymbol{\phi}}}_i$, which corresponds to point $\boldsymbol{\theta}_i$, is marked with two hats to distinguish it from the globally optimal nuisance parameters, $\hat{\boldsymbol{\phi}}$, which correspond to the best estimate of the parameters of interest, $\hat{\boldsymbol{\theta}}$. With these parameters defined, the likelihood ratio from Equation 1 becomes:

$$\lambda_i = \ell(\boldsymbol{x}|\boldsymbol{\theta}_i, \hat{\hat{\boldsymbol{\phi}}}_i) - \ell(\boldsymbol{x}|\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}). \tag{5}$$

In the frequentist statistical philosophy each nuisance parameter possesses an (unknown) true value. The intuition is that, absent any further information, we adopt the nuisance parameter values most compatible with the data. This procedure contrasts with the Bayesian 'marginalization' procedure, where the likelihood is taken to be the likelihood integrated over all values of the nuisance parameters, weighted by a prior probability distribution.

The coverage guarantees of the Feldman–Cousins procedure rely on our access to a collection of FC pseudoexperiments to inspect, which have been generated at the precise points we wish to include/exclude at a certain significance. In the presence of nuisance parameters, however, we no longer have access to such an ensemble since the values of the nuisance parameters are not defined a priori by the point in parameter space being tested. Nevertheless, some values must be chosen in order to generate FC pseudoexperiments. We could ensure correct coverage by defining our allowed regions in a high-dimensional space containing all the nuisance parameters, but this is impractical, both computationally and because it cannot be easily visualized. When defining a lower-dimensional allowed region, the values we choose for the nuisance parameters may differ from the true values, potentially yielding incorrect coverage.

C. Existing Methods

Several plausible approaches exist for generating the FC pseudoexperiments for point $\boldsymbol{\theta}_i$ in the presence of nuisance parameters; the methods differ both in how practical they are to use and in the accuracy of the coverage they achieve. We discuss the methods below, and point out those which are impractical to apply to real-world problems. The coverage properties of the methods that are practical to implement will be explored in Section II E.

A priori estimate: Hold the nuisance parameters fixed at their a priori assumed values in the generation of all FC pseudoexperiments, $\boldsymbol{\phi}_i = \boldsymbol{\phi}_0$. While straightforward, in the plausible case that the true values of the nuisance parameters differ from their a priori values, the a priori estimate solution ignores the information available from the data about their values and thus can easily under- or over-cover. While not expected to perform well, this method is straightforward to implement so we will examine its coverage properties in Section II E.

Conservative: At each point in the parameter space, $\boldsymbol{\theta}_i$, select the values of the nuisance parameters that yield the most conservative (largest) critical value based on FC pseudoexperiments, and thus the largest confidence interval, $\boldsymbol{\phi}_i = \mathrm{argmax}_{\boldsymbol{\phi}}\, c_{\alpha,i}(\boldsymbol{\phi})$. By taking the most conservative critical values, this method is guaranteed not to under-cover. However, because even nuisance parameters highly inconsistent with the data are considered, it is likely to substantially over-cover. Additionally, unless a closed-form estimate of the $c_{\alpha,i}(\boldsymbol{\phi})$ is available, this can be computationally infeasible for unbounded parameters or a large number of parameters.

Berger–Boos: This method is philosophically similar to the conservative method, but introduces a limiting principle for which values of nuisance parameter to consider. At each point in parameter space, $\boldsymbol{\theta}_i$, determine the range of nuisance parameters consistent with the data at significance level β, and then calculate p-values empirically (i.e. using pseudoexperiments) for all values of the nuisance parameters within that range.

The overall p-value for point $\boldsymbol{\theta}_i$ is based on the largest p-value within that set, $p = \max_{\boldsymbol{\phi}} p(\boldsymbol{\theta}_i, \boldsymbol{\phi}) + \beta$. This method is named after its proposers [11]. Since the nuisance parameters in the likelihood and the pseudoexperiments are moved together, this method does not have the same problem of over-coverage as the Conservative method, but it is still computationally infeasible for making confidence intervals or for a large number of nuisance parameters. Appendix B shows the use of this method to cross-check the significance in a single hypothesis test, which is the context in which it was originally proposed.

Highland–Cousins: When generating FC pseudoexperiments, generate the nuisance parameters from their a priori probability distributions, $\boldsymbol{\phi}_i \sim P_r(\boldsymbol{\phi}_0)$. This method is commonly called the Highland–Cousins method after its proposers [12]. The Highland–Cousins approach guarantees coverage in the sense that an ensemble of experiments in which the true values of the nuisance parameters are distributed according to the assumed a priori will have correct coverage overall, analogous to the usual frequentist requirement to have correct coverage when aggregated over repeated statistical samples. However, in a frequentist analysis, nuisance parameters do in fact have true values, and the goal is to ensure correct coverage for those true values. In the same fashion as with the a priori estimate approach, information about the nuisance parameters garnered from the experiment is here discarded. The Highland–Cousins method has also been shown to over-cover in circumstances where the nuisance parameter has a true fixed value but an estimated value that can vary experiment-to-experiment [13, 14]. Since this method requires the generation of a single set of FC pseudoexperiments, it is practical to use and its coverage properties will be investigated in Section II E.

A posteriori Highland–Cousins: At each point in parameter space, generate the FC pseudoexperiments with parameters drawn from the post-fit, or a posteriori, likelihood distribution derived from the observed data, $\boldsymbol{\phi}_i \sim P(\hat{\boldsymbol{\phi}}|\boldsymbol{\theta}_i)$. This variant has the same issue as the regular Highland–Cousins method, where the coverage is ensured for an ensemble of experiments with nuisance parameter values drawn from the a posteriori distribution rather than considering their true values. This procedure can also be impractical to apply in frequentist analyses, which do not naturally produce these a posteriori distributions. Nonetheless, by constraining the nuisance parameter values to those most consistent with the data, the coverage for the unknown true values is likely to be more accurate. This method will be investigated in Section II E.

D. The Profiled Feldman–Cousins Method

We propose an alternative procedure addressing some of the shortcomings of the existing methods:

Profiled Feldman–Cousins: At each point in parameter space, $\boldsymbol{\theta}_i$, generate the FC pseudoexperiments assuming the best-fit values of the nuisance parameters, given these parameters and the observed data, $\boldsymbol{\phi}_i = \hat{\hat{\boldsymbol{\phi}}}_i$, as defined in Equation 4.

This follows the same intuition that motivates the frequentist profiling procedure. While the best-fit nuisance parameters are certainly not exactly the true values, they are the best estimate available to us, and we expect FC pseudoexperiments generated from our best estimate of the true parameters to yield better coverage than experiments not so informed. The Profiled FC method takes the definition of the critical value from Equation 2 literally, meaning that the distribution, $P(\lambda_i)$, should be calculated for λ_i with nuisance parameters fixed at $\hat{\hat{\boldsymbol{\phi}}}_i$ as defined in Equation 5. We note that this method is a generalization of the procedure in Chapter 22 of [10] for likelihood–ratio tests, and is consistent with the best-practices recommendations from the PhyStat-DM workshop [15]. The examples in [10] focus on simple cases where $P(\lambda_i)$ does not depend on the value of the nuisance parameters[4], or where the distribution can be derived or approximated analytically. Since we cannot rely on these assumptions, we instead use FC pseudoexperiments to determine $P(\lambda_i)$ empirically for each point being tested, $\boldsymbol{\theta}_i$, along with its associated nuisance parameters, $\hat{\hat{\boldsymbol{\phi}}}_i$[5].

Note that in this procedure, the critical values depend on the observed data, which has an important practical consequence: unlike with the standard Feldman–Cousins method, it is no longer possible to generate the FC pseudoexperiments before having determined the best fit nuisance parameters profiled from the data. Some additional features and limitations are described later in Section III F.

E. Toy Model

We can gain intuition and illustrate many of the key features of the aforementioned methods using a toy model to evaluate their coverage properties for a wide range of scenarios. This model is chosen to resemble a situation that occurs in practice in the analysis of neutrino oscillation experiments, while remaining as simple and generic as possible.

The toy consists of the measurement of a single number – the number of events observed. We take the expected number to be given by

$$N_{\text{exp}} = A - B\sin\delta \pm C, \tag{6}$$

where A, B, and C are fixed constants and the expectation, N_{exp} depends on a 2 unknown parameters: a continuous, cyclic parameter, δ, and a binary parameter corresponding to a positive or negative sign for the C term. We choose values for the constants:

$$A = 80,$$
$$B = 15,$$
$$C = 10,$$

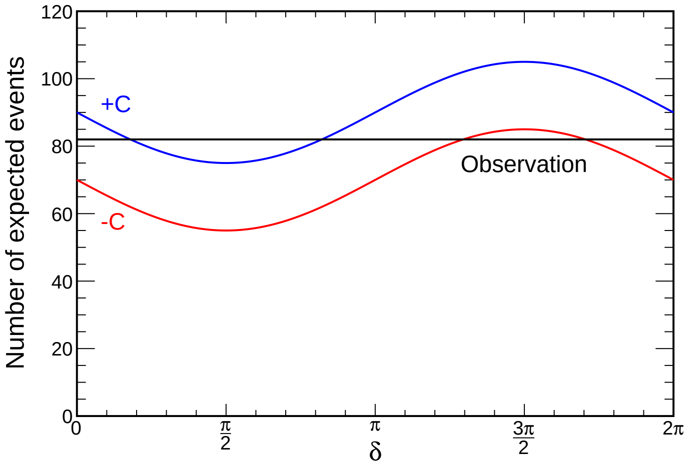


FIG. 1. The number of events expected in the toy model as a function of the continuous δ parameter (x-axis) and sign of C term (positive sign blue, negative sign red). A hypothetical observation of a particular number of events is shown in black.

[4] The prescription $P(\lambda_i) = \chi^2_k$ from Wilks' theorem is an example of such a case where P depends only on the number of degrees-of-freedom in the likelihood, k, not which point i is being tested.

[5] Determining the distribution empirically is not suggested as a solution in [10]. We speculate that this possibility is omitted because it is only practical to do with access to a detailed simulation of the likelihood and extensive computing resources not available at the time.

so that the toy model has event counts similar to current rates from the NOvA experiment [8]. Figure 1 illustrates this function, along with a hypothetical measurement that we would want to interpret. The experiment consists of making a single measurement of the number of events observed, N_{obs}, comparing to the expected number of events N_{exp}, and using that to generate confidence regions in δ or determine the sign of the C term.

Constraining ourselves for the moment to the case where the sign of C is already known (we have external information telling us for certain which sign to pick) one derives a confidence interval by first finding the value $\hat{\delta}$ that provides the best match to the observed data (the best fit given N_{obs}), and then computing:

$$\lambda(\delta) = \ell(\delta) - \ell(\hat{\delta}) \tag{7}$$

for each value of δ under consideration.

For the purposes of keeping this toy minimal, and to avoid discontinuities arising from discrete event counts[6], we will assume N_{obs} is normally distributed with mean N_{exp} and standard deviation $\sqrt{N_{\text{exp}}}$, and thus:

$$\ell(\delta) = \frac{\Big(N_{\text{exp}}(\delta) - N_{\text{obs}}\Big)^2}{N_{\text{exp}}(\delta)}. \tag{8}$$

To determine confidence intervals, one then compares $\lambda(\delta)$ to c_α and accepts all values of δ having a lower λ. According to Wilks' theorem, $\lambda \sim \chi^2_{k=1}$, and one should therefore use $c_\alpha = 1$ to achieve 68.27% coverage.

This procedure over-covers significantly, even when the sign of C is known in advance. First, most observed event counts are compatible with two values of δ, due to the periodic nature of the N_{exp} function. Second, in cases where a statistical fluctuation in the data leads to observations outside the expected range ($A - B + C < N_{\text{exp}} < A + B + C$, if C is known to be positive), no good 'fit' to the data will be available. The best available fit will be at the extreme of the function range, making $\ell(\hat{\delta})$ larger than it would be without constraints, and causing a larger region of the δ space to have a value of λ below 1. This 'physical boundary' effect is expected to be largest when the true value of δ is near $\pi/2$ or $3\pi/2$, where such a fluctuation is expected to occur 50% of the time. Figure 2 shows this over-coverage vs. the true value of δ. We evaluate coverage by generating a series of statistically fluctuated toy experiments at each true value of δ, determining the best fit and confidence interval that would be obtained for each, using $c_{68\%} = 1$, and counting the fraction of these toy experiments in which the true δ value is included in the confidence interval.

In this circumstance where the sign of C is known, the Feldman–Cousins procedure can be followed to produce perfect coverage for any value of δ. Figure 3 shows how the critical value, $c_{68\%}$, varies as a function of δ, with substantially lower critical values in the regions nearest the physical boundary to account for the effect described above. Using these critical values to evaluate the coverage of an independent set of mock experiments yields ideal coverage, as would be expected in this case since the FC pseudoexperiments were generated in exactly the same way.

In the full experiment, we do not know the true sign of C. The standard frequentist procedure in this case is to profile over the sign parameter,

$$\ell(\delta) = \min\Big(\ell^+(\delta), \ell^-(\delta)\Big), \tag{9}$$

where ℓ^+ is evaluated using the values of N_{exp} based on the positive sign for C, and similarly for ℓ^-. We can replicate this procedure in the fits performed on the FC pseudoexperiments, but we are still left with the question of how to generate the FC pseudoexperiments. We will obtain different critical values if we generate all the FC pseudoexperiments with positive vs. negative sign, as shown by the solid and dashed lines in Figure 4, because the boundaries on allowed values of N_{exp} are now wider ($A - B - C < N_{\text{exp}} < A + B + C$), and FC pseudoexperiments generated assuming a particular sign will only run up against one boundary. The previous example where the sign was known (Figure 3) showed large downward deviations in the critical value at both $\pi/2$ and $3\pi/2$ since both were boundaries on N_{exp}, but now there is only a large deviation at $3\pi/2$ for the positive sign, where it runs into the high-side boundary on N_{exp}, and at $\pi/2$ for the negative sign where it runs into the low-side boundary. In the intermediate regions around 0, π, and 2π, where the event counts in the pseudoexperiments will typically be far from the overall upper and lower limits no matter which sign we assume when generating them, the critical values closely follow each other.

[6] Typical physics analyses have many bins and continuous parameters. But the first NOvA electron neutrino appearance data, with only a handful of events in each bin, caused discontinuities to appear. An example of this type of discontinuity caused by integer event counts can be seen in Fig. 4 of [16].

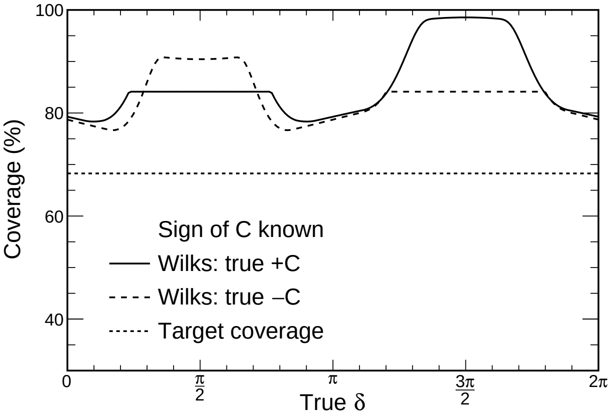


FIG. 2. Coverage for the toy experiments using Wilks' theorem in the case where the true sign of C is positive and this fact is known to the fitter (solid) and likewise true $-C$ known to the fitter (dashed). The short-dashed line indicates the desired coverage. Since there are no nuisance parameters, all other discussed techniques are equivalent to Feldman-Cousins. Since they would all perfectly match the target coverage, they are not shown in this figure.

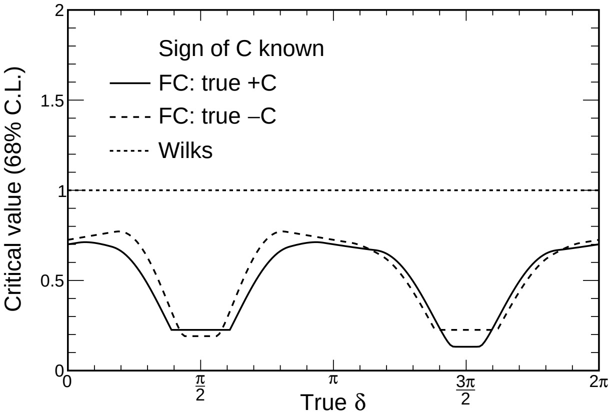


FIG. 3. Critical values evaluated for the toy experiments using the Feldman–Cousins procedure in the case where the true sign of C is positive and this fact is known to the fitter (solid) and likewise true $-C$ known to the fitter (dashed). The critical value shows substantial deviations from the expectation of Wilks' theorem (short-dashed) in those regions where the Wilks' critical value most over-covered.

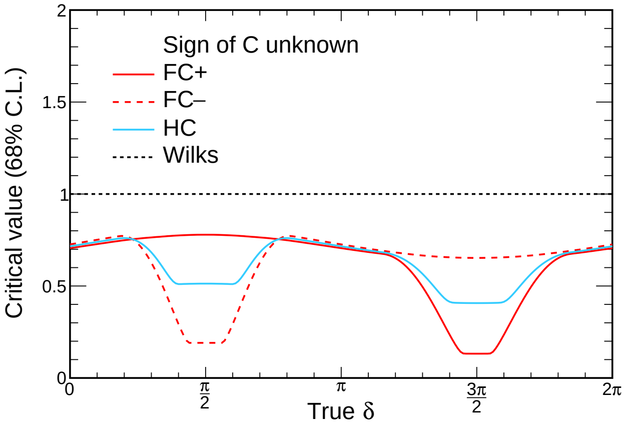


FIG. 4. Critical values for the 68% C.L. from Wilks' theorem (the horizontal black line at 1), the Feldman–Cousins procedure (red) and Highland–Cousins (light blue), where the true sign of C is positive. The Feldman–Cousins critical values are shown for two cases – generating the FC pseudoexperiments assuming positive C (solid) and assuming negative C (dashed). In our toy model, the Highland–Cousins procedure consists of generating the FC pseudoexperiments with an equal mixture of the two signs, and the blue curve splits the difference between the red curves as expected. The profiled FC procedure cannot be displayed on this plot; it amounts to choosing one or other of the Feldman–Cousins curves at each value of δ depending on the observed data.

The consequences of this behavior for the coverage of confidence intervals are shown in Figure 5, which compares the coverage vs. true values of δ and sign of C (solid/dashed for positive/negative) from Wilks' theorem (black) and from the Feldman–Cousins procedure where we arbitrarily choose to generate FC pseudoexperiments assuming the positive sign. As in Figure 2, Wilks' theorem shows over-coverage everywhere, but it is substantially worse when the true values lie near the boundaries on N_{exp} ($+C$, $\delta = 3\pi/2$ or $-C$, $\delta = \pi/2$). The Feldman–Cousins method yields ideal coverage in the $+C$ case, but large deviations in the case of true $-C$, where the FC pseudoexperiments have incorrectly encountered a physical boundary (at $3\pi/2$) or missed one (at $\pi/2$). The results for experiments generated assuming negative sign show the same qualitative behaviour, but with the roles of $\delta = \frac{\pi}{2}$ and $\delta = \frac{3\pi}{2}$ reversed.

For the present toy experiment, the Highland–Cousins procedure consists of splitting the difference by generating the FC pseudoexperiments equally from each sign (assuming a 50:50 prior expectation). This has the predictable effect

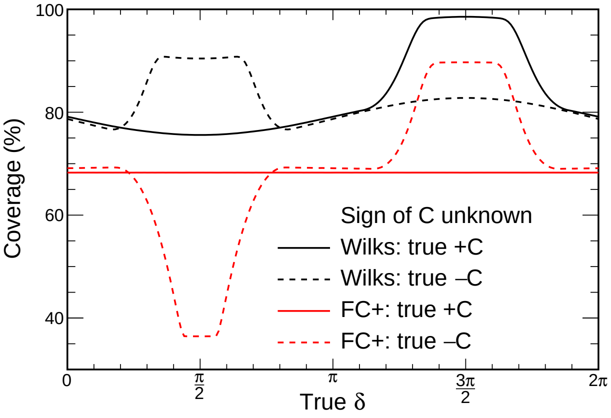


FIG. 5. The coverage obtained for our toy experiments using critical values from Wilks' theorem (black) and the Feldman–Cousins procedure, which here assumes a positive sign for C for the FC pseudoexperiments (red). Coverage is shown vs. true δ and true sign (solid/dashed for positive/negative). The true sign is *not* known at fit time and is profiled over. The Wilks' theorem critical values lead to substantial over-coverage in all cases. Since the FC pseudoexperiments have been generated assuming positive sign, the procedure produces exactly the target coverage of 68% for toy experiments with true positive sign, but for true negative sign the coverage properties are particularly poor.

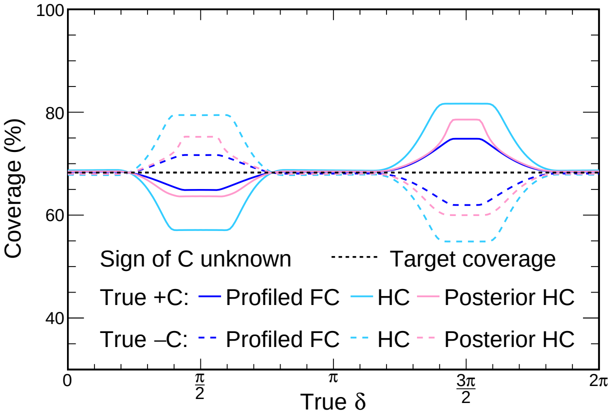


FIG. 6. The coverage obtained using critical values using the Highland–Cousins procedure (light blue) and our proposed profiling procedure (dark blue) for our toy model, where the true sign of C is unknown at fit time, and profiled over, evaluated for true positive sign (solid) and true negative sign (dashed). In both cases the coverage averaged over δ and sign is correct, but the profiling procedure exhibits substantially smaller deviations from correct coverage where these occur. Also shown is the a posteriori Highland-Cousins method (here labeled 'Posterior HC' and drawn in pink) which can be considered as an intermediate option between Highland–Cousins and our profiling method, and yields an intermediate performance.

of yielding critical values intermediate between the FC expectations from the two signs (light blue line in Figure 4) and coverage (light blue lines in Figure 6) intermediate between the 'right' and 'wrong' FC coverage (red lines, solid and dashed respectively, in Figure 5). This is certainly an improvement from the FC$^+$ (or FC$^-$) case – the 'average' coverage is correct, and there is no longer a large difference in behaviour depending on the true sign.

The procedure we propose in the present work achieves better results than any of these methods by using information from the observed data itself. If we observe a large number of events, say $\gtrsim 85$, we know it is more likely that the critical value evaluated under the $+C$ hypothesis will provide the right coverage, and similarly a small number of observed events, $\lesssim 70$, suggests the $-C$ hypothesis is more likely to provide correct coverage. If we observe an intermediate number of events (values close to 80), then we have gained no information about the true sign of C, but in that case the critical values are very similar either way.

In this case, for each toy experiment contributing to the coverage evaluation, for each value of δ whose membership in the confidence interval we need to determine, we evaluate which sign gives the best match (lowest ℓ) to the data, and generate the FC pseudoexperiments from which the critical value will be derived assuming that sign. For a continuous nuisance parameter, we would generate experiments assuming the best-fit value.

The blue lines in Figure 6 show the coverage obtained by this procedure. Deviations still occur in the regions where the two critical values differ, but the magnitude is substantially reduced compared to Highland–Cousins. The remaining mis-coverage is due to those cases where a statistical fluctuation produces a number of events more compatible with positive sign, despite the true sign being negative, or vice versa.

The Posterior Highland–Cousins approach – generating the FC pseudoexperiments distributed between the two signs based on the posterior distribution – represents an intermediate point between Highland–Cousins (generating pseudoexperiments equally from the two signs) and our profiling method (generating pseudoexperiments from the best-fit sign). Unsurprisingly, for these toy experiments it yields intermediate coverage properties – better than Highland–Cousins but not as good as our proposed method.

Source code reproducing the analysis of this toy model is publicly available [17].

III. IMPLEMENTATION IN THE NOVA ANALYSIS

The primary goal of a neutrino oscillation experiment like NOvA is to measure the parameters which govern neutrino oscillations, namely the mixing angles and phase from the PMNS mixing matrix as well as the differences between the neutrino masses [8]. Additionally, certain 'binary' questions can be addressed, for example whether the ordering of the neutrino masses is 'normal' or 'inverted,' i.e., whether m_3 is larger or smaller than m_1. These parameters, as described above, cannot be observed directly. Instead, the experiment uses a beam of muon (anti)neutrinos [18] and measures the rate of disappearance of muon (anti)neutrinos and the rate of appearance of electron (anti)neutrinos as a function of their estimated energy. Since the parameters of interest govern these disappearance and appearance rates, they can be estimated from the observed energy spectra via Maximum Likelihood Estimation [1]. The confidence intervals describing the uncertainty on these parameters are then determined using the methods described here.

After some concrete illustrations of how Wilks' conditions are not satisfied, this section describes some key technical details in the implementation of the Profiled FC method in the NOvA oscillation analysis. Substantially more details on the optimization of this method to run on High Performance Computing platforms will be available in an upcoming paper.

A. Violations of Wilks' theorem assumptions in NOvA's neutrino oscillation analysis

Feldman and Cousins first introduced the FC method in the context of a neutrino experiment [19] where the conditions for Wilks' theorem, described in Section II A were not met. The NOvA 3-flavor oscillation analysis violates these three conditions as follows:

(1) Effective boundaries: Many of the parameters of the oscillation model have effective boundaries of some kind. One example can be seen with the 2-flavor approximation of the survival probability for neutrino flavor ν_α:

$$P(\nu_\alpha \to \nu_\alpha) = 1 - \sin^2(2\theta)\sin^2\left(\frac{\Delta m^2 L}{4E}\right), \tag{10}$$

where L is the constant distance, E is the neutrino energy, and Δm^2 and θ are the independent parameters being measured. While the angle θ is unconstrained, the impact it has on the observable (the survival probability) is constrained by unitarity: if $\theta = \pi/4$, either increasing or decreasing θ will lead to a reduction in the oscillation probability. Similarly, the $\mathcal{CP}$-violating phase δ_{CP} is cyclic and not well constrained, so it also easily runs up against effective 'boundaries' in its possible impact.

(2) Nested hypotheses: The nested hypothesis assumption is not violated for all measurements, but it is clearly violated for binary questions. When there are only 2 possible disjoint outcomes (e.g., mass ordering is normal or inverted), whichever is chosen as the null cannot be a special case of the alternate.

(3) Sample size: Long-baseline neutrino experiments generally have small sample sizes because of the small neutrino interaction cross-section and large physical distances required for oscillations to occur. The most recent measurement had 82 electron neutrino candidates and 33 electron antineutrino candidates, in neutrino and antineutrino beam modes respectively [8].

The procedure followed by NOvA is presented next.

B. Fitting the data

NOvA measures the energy spectra of disappearing muon (anti)neutrinos and appearing electron (anti)neutrinos in order to constrain parameters of the neutrino oscillation model: the mixing angle θ_{23}, the mass splitting Δm^2_{32}, in particular its sign, equivalent to determining the neutrino mass ordering, and the CP–violating phase δ_{CP}. The candidate neutrino interactions are divided into different categories (based on energy resolution and particle identification criteria) to optimize the measurement's sensitivity. The compatibility between a model prediction given a set of parameter values and some data is quantified with a likelihood function $\mathcal{L}$. The best fit is found by maximizing $\mathcal{L}$, or minimizing $\ell = -2\ln\mathcal{L}$. Since the data is structured as a histogram (meaning a set of counts of independent events), the likelihood function for Poisson–distributed data [1] is used[7]:

$$\ell_{stat} = 2\sum_i \left(e_i(\boldsymbol{\theta}) - o_i + o_i \ln\frac{o_i}{e_i(\boldsymbol{\theta})}\right), \tag{11}$$

[7] Or more accurately $\ell = -2\ln\mathcal{L}/\mathcal{L}_0$, where $\mathcal{L}_0$ is the likelihood when $o_i = e_i$

where $e_i(\boldsymbol{\theta})$ is the expected number of events in bin i given parameter values $\boldsymbol{\theta}$, and o_i is the observed number of events in that same bin. The $e_i(\boldsymbol{\theta})$'s are calculated by extrapolating the muon (anti)neutrino energy spectrum measured in NOvA's near detector to its far detector assuming a set of neutrino oscillation parameters, taking into account known differences in flux and acceptance between the detectors. In addition to the oscillation parameters, around 50 systematic uncertainties are included in the fit as nuisance parameters, with penalty terms added to the likelihood in Equation 11:

$$\ell = \ell_{stat} + \sum_k \frac{\phi_k^2}{\sigma_k^2}, \tag{12}$$

where σ_k is the prior uncertainty on the k^{th} nuisance parameter ϕ_k. The sources of uncertainty vary from parameter to parameter. For example, some uncertainties are based on the uncertainties quoted by external measurements, some are based on the level of agreement between data and simulation within the experiment, and some are based on comparisons between alternative theoretical models. The values of $\sin^2\theta_{23}$, Δm^2_{32}, and δ_{CP} which minimize ℓ (i.e., the Maximum Likelihood Estimate or best fit point) are found using the Minuit2 minimizer [20]. This best fit point is the basis from which the confidence intervals and significances, the main topic of this paper and main results of the oscillation analysis, are constructed.

C. Building 1-dimensional and 2-dimensional confidence intervals

To build 1-dimensional or 2-dimensional maps of the significance, we need to sample the oscillation parameter space finely enough to catch possible local features, while also being limited by the computational costs the Profiled Feldman–Cousins approach entails. In practice, this means that the significance is evaluated at 60 points evenly distributed across the range of parameter values when building 1-dimensional significance maps. These one-dimensional plots can be constructed with the parameters constrained in one mass ordering, one θ_{23} octant[8], or a combination of both. In two dimensions, we report confidence intervals (i.e., contours) for $\sin^2\theta_{23}$ vs. δ_{CP} (estimated in a 30×30 grid) and Δm^2_{32} vs. $\sin^2\theta_{23}$ (in a 20×20 grid), for both orderings.

As explained earlier, we chose to profile the nuisance parameters. The first step is therefore to fit the data with the parameters of interest fixed at each grid point, $\boldsymbol{\theta}_i$, and find $\hat{\hat{\boldsymbol{\phi}}}_i$, the set of nuisance parameters minimizing ℓ per Equation 4. This process can be conveniently run on standard distributed computing resources and serves as an input to the more computationally intensive generation and fitting of millions of Feldman–Cousins pseudoexperiments in a High Performance Computing environment. From that first step, we can already obtain a good approximation of the significance maps. The Feldman–Cousins procedure then modifies those maps, increasing or decreasing the significance depending on the distribution of the underlying test statistic, which is why this procedure is often perceived as a correction. We can also take advantage of those approximated significances to estimate the number of FC pseudoexperiments that need to be generated at each point of the parameter space, $\boldsymbol{\theta}_i$, to reach a desired statistical accuracy when measuring the p-values from the empirical λ distributions. For each $\boldsymbol{\theta}_i$, the FC pseudoexperiments are constructed by generating Poisson–fluctuated neutrino energy spectra from the predictions made at $(\boldsymbol{\theta}_i, \hat{\hat{\boldsymbol{\phi}}}_i)$ determined above. For each FC pseudoexperiment, j, generated at point i, a likelihood ratio is estimated:

$$\begin{aligned}\lambda_{ij} &= \ell_{\text{constrained}} - \ell_{\text{unconstrained}} \\ &= \ell(\boldsymbol{x}_j|\boldsymbol{\theta}_i, \hat{\hat{\boldsymbol{\phi}}}_{ij}) - \ell(\boldsymbol{x}_j|\hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{\phi}}_j).\end{aligned} \tag{13}$$

Both likelihoods are evaluated on the FC pseudoexperiment spectrum, $\boldsymbol{x}_j$, at parameter values which minimize the likelihood function, ℓ, but they differ in which parameters are allowed to vary in the minimization. The first likelihood is evaluated after a constrained fit where the parameters of interest are fixed to the values used to generate the pseudoexperiment, $\boldsymbol{\theta} = \boldsymbol{\theta}_i$, and only the nuisance parameters are varied, denoted by $\boldsymbol{\phi} = \hat{\hat{\boldsymbol{\phi}}}_{ij}$, analogous to how $\hat{\hat{\boldsymbol{\phi}}}_i$ is determined in the fit to the real data. The second likelihood is evaluated after an unconstrained fit in which both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are varied in order to find the global minimum of $\ell(\boldsymbol{x}_j)$, denoted, $(\hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{\phi}}_j)$.

The neutrino oscillation parameter space can be degenerate, in particular for δ_{CP} and nuisance parameters like θ_{13}, or for values of θ_{23} mirrored around the value which produces maximal ν_μ disappearance. In order to avoid biases towards a particular region of parameter space, we run multiple fits with different seed values for each FC pseudoexperiment and then take the result with the lowest ℓ.

[8] $\theta_{23} < 45°$ is commonly referred to as the lower octant, while $\theta_{23} > 45°$ is the upper octant.

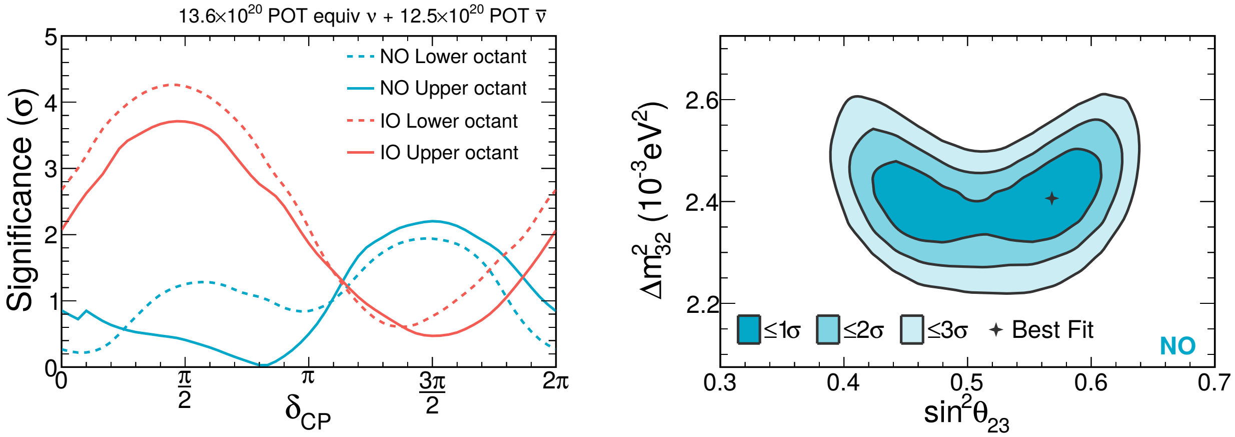


FIG. 7. Left: Significance of the data for different values of $\delta_{\rm CP}$. Right: Contour plot showing the 1-σ, 2-σ, and 3-σ domains of isosignificance in the normal ordering for Δm^2_{32} vs. $\sin^2\theta_{32}$ [8].

Between 1000 and 5000 FC pseudoexperiments are generated at each $\boldsymbol{\theta}_i$, where more FC pseudoexperiments are required for the most extreme p-values. Furthermore, given the very large number of FC pseudoexperiments that are required in the 3-sigma (and above) regions in order to accurately measure the corresponding small p-values, we choose to only perform the profiled FC procedure in regions where $\sqrt{\lambda_{\rm Wilks}} < 20$ for 1-dimensional constraints and $\sqrt{\lambda_{\rm Wilks}} < 12$ for 2-dimensional constraints.

The λ_{ij} distributions are then used to build empirical test statistic distributions for each $\boldsymbol{\theta}_i$. For 1-dimensional significance plots, a p-value is first determined at each grid point by counting the fraction of FC pseudoexperiments with a λ_{ij} larger than that of the data at that same $\boldsymbol{\theta}_i$. The p-value is then converted to a significance via $\sigma = \sqrt{2}\,\mathrm{erfc}^{-1}(p)$. The resulting collection of significances is then interpolated and smoothed taking care to preserve real discontinuous features (discussed more in Section III F). Figure 7 illustrates how significances for one or two parameters of interest can be represented. For most regions of the parameter space, we expect the underlying likelihood surface to be well–behaved but the existence of boundaries and local, nearly degenerate minima can skew the test statistic distributions, resulting in jump of significances between neighboring grid points, as illustrated in Section III F.

The procedure to establish 2-dimensional contours of isosignificance is slightly different. We first start by evaluating the standard likelihood of the data at each point $\boldsymbol{\theta}_i$ of the grid used to sample the parameter space. We then evaluate the critical likelihood corresponding to each of the significance levels of interest, namely 1σ, 2σ, and 3σ, from the set of Feldman–Cousins pseudoexperiments, again, at each grid point. Each map of critical profiled FC values is then subtracted from the map of standard likelihood obtained from the data. The intersection of the resulting surfaces with the plane 0 (or, for the inverted ordering, with the plane λ_{IH}, which is the difference between the likelihoods of the best fit point in the Inverted Ordering and the overall best fit point) represents the contours of isosignificance. A kernel smoothing procedure is finally applied to the 2-dimensional contours, taking care to consider points near $\delta_{\rm CP} = 0$ and $\delta_{\rm CP} = 2\pi$ as neighbors (due to its cyclical nature) in the $\sin^2\theta_{23}$ vs. $\delta_{\rm CP}$ contours.

D. Hypothesis tests

In addition to 1-dimensional and 2-dimensional constraints on oscillation parameters, we can perform hypothesis tests for the mass ordering, the θ_{23} octant, or a combination of both. A key benefit of the Profiled FC Method is that the procedure can naturally address these binary tests (or discrete choices in general): the FC pseudoexperiments are generated with the parameter being tested held fixed and all other parameters set to their profiled values given that constraint. For example, if the overall best fit is in the normal ordering, the test would be for rejecting the inverted ordering, so the FC pseudoexperiments would be generated in the inverted ordering with all other parameters set to the best fit to the data in that ordering. Since this procedure is only done at one point of the parameter space for each hypothesis test, we can afford to generate more FC pseudoexperiments (tens of thousands) and reach more accurate measurements of the p-values and significances than for 1D and 2D confidence intervals. The result of the procedure is, again, an empirical collection of $\lambda = \ell_{\rm constrained}$-$\ell_{\rm unconstrained}$ which can be used to determine the fraction of FC pseudoexperiments that yield a λ less compatible with the null hypothesis than the data, equating to a p-value. This likelihood–ratio test statistic slightly differs from the one defined in Equation 5: all parameters are still free to vary in the unconstrained fit, but in the constrained fit, the parameters of interest are allowed to take values within the

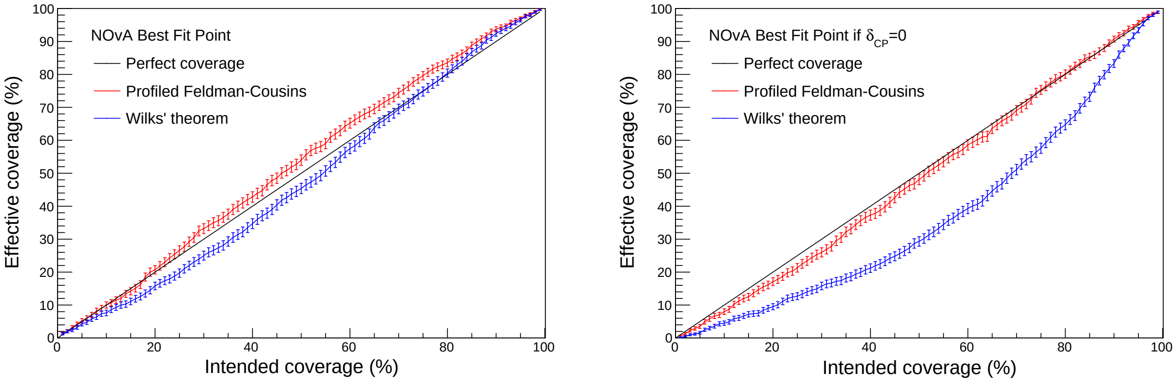


FIG. 8. The left figure shows the coverages obtained with Wilks' theorem (blue) and the Profiled Feldman–Cousins approach (red) at our overall best fit point, while the right figure shows those coverages at our best fit if $\delta_{\mathrm{CP}} = 0$. On the left, Wilks' theorem shows a good approximate coverage, while on the right, it produces a significant under-coverage, which would have the effect to artificially disfavor $\delta_{\mathrm{CP}} = 0$. The coverage obtained with the Profiled Feldman–Cousins approach is consistently more accurate. The error bars represent the statistical uncertainty on the binomial confidence interval obtained from 1000 fake experiments.

limits defined by the hypothesis being tested. This procedure is the only correct one for the estimation of our level of preference (or rejection) for a given hypothesis; it cannot be done by reading the minima of the 1-dimensional or 2-dimensional confidence intervals, as explained in more detail in Section III F. The profiled FC procedure can also be extended in a straightforward way to also calculate a CLs significance, see Appendix A for details.

E. Validation

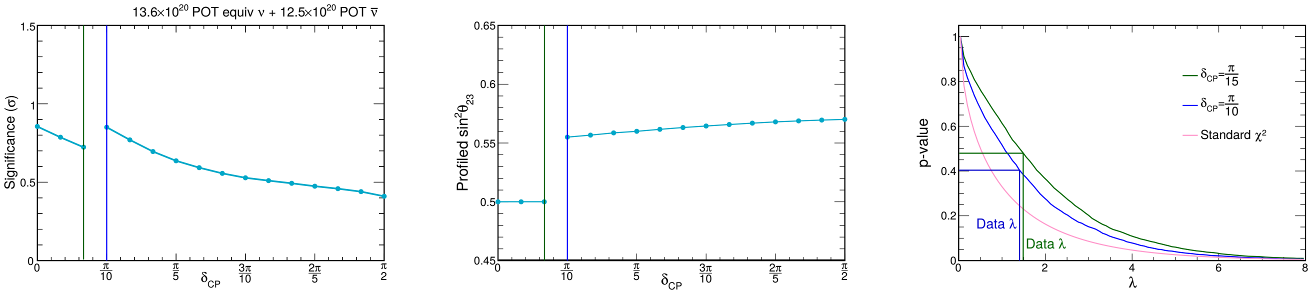
When considering any frequentist statistical procedure, a key step is to evaluate the coverage properties of that procedure for the problem at hand. The goal of the profiled FC procedure is to produce confidence intervals with coverage as close as possible to the stated level α. The examples in Section II E show that none of the procedures considered produce perfect coverage when certain truth quantities are unknown, but in those examples, the procedure we use comes the closest.

Here we give an in-situ demonstration of achieving these coverage properties with NOvA simulation. We have chosen two points of interest: our overall best fit point from [8], which is far from boundaries, leading to little impact from the Profiled FC procedure on the significance, and our preferred point if the CP–violating phase was $\delta_{\mathrm{CP}} = 0$[9]. This parameter region is degenerate which can cause the underlying test statistic distribution to deviate from a standard χ^2-distribution. For those two points, we want to estimate how the effective coverage varies for different levels of intended coverage. To do so, we repeated the Profiled FC procedure for 1000 validation pseudoexperiments generated at each of those two points, and measured how frequently the true point was actually contained in a given confidence interval. For instance, in the ideal case, we would expect the 50% confidence interval to cover the true point in 50% of the validation pseudoexperiments.

In practice, we first fit each validation pseudoexperiment, i, to determine its best fit point, $(\hat{\boldsymbol{\theta}}_i,\ \hat{\boldsymbol{\phi}}_i)$, as well as the preferred set of nuisance parameters when $\boldsymbol{\theta}$ is constrained to the value the validation pseudoexperiments were generated at, $(\boldsymbol{\theta}_0, \hat{\hat{\boldsymbol{\phi}}}_{0i})$[10]. We then generate 1000 regular FC pseudoexperiments for each *validation* pseudoexperiment (1,000,000 in total) based on each one's best fit nuisance parameters at the value being tested, $(\boldsymbol{\theta}_0, \hat{\hat{\boldsymbol{\phi}}}_{0i})$. These FC pseudoexperiments are then used to determine critical values, $c_{\alpha,i}$, for each validation pseudoexperiment, i, at a range of significances, α. The Wilks' theorem critical values are derived analytically and are, of course, the same for every validation pseudoexperiment. The effective coverage with both methods can then be measured by counting how often $\boldsymbol{\theta}_0$ falls inside the confidence intervals, or

$$\ell(\boldsymbol{\theta}_0, \hat{\hat{\boldsymbol{\phi}}}_{0i}) - \ell(\hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\phi}}_i) < c_{\alpha,i}. \tag{14}$$

[9] While this test could be done at any points, these points from the fit to NOvA data were chosen to give concrete, relevant examples.
[10] In this study the nuisance parameters just include other oscillation parameters; we did not include systematic uncertainties.



(a) Significance vs. δ_{CP}

(b) Profiled $\sin^2\theta_{23}$ vs. δ_{CP}

(c) p-value vs. critical λ

FIG. 9. (a) The quoted significance vs. δ_{CP} is discontinuous around $\delta_{\mathrm{CP}} = \frac{\pi}{10}$. This is due to the discontinuity in the profiled value of $\sin^2\theta_{23}$ as a function of δ_{CP}. (b) $\sin^2\theta_{23}$ transitions from maximal mixing to upper octant at this point. The FC pseudoexperiments are therefore generated at different points in parameter space. (c) The very similar values of λ in the data are assigned different p-values due to being compared to different empirical distributions. The p-value is obtained by integrating the empirical test-statistic distribution, $P(\lambda)$, from a lower bound, shown here on the x-axis, to $+\infty$.

The above calculation is done per *validation* pseudoexperiment, and the FC pseudoexperiments are just used to determine the critical values, $c_{\alpha,i}$.

Note that without nuisance parameters, this test would be tautological: the validation pseudoexperiments and the FC pseudoexperiments being used to determine if the test point would be inside the profiled FC confidence interval would all be drawn based solely on $\boldsymbol{\theta}_0$, and so the coverage must be correct. In the presence of nuisance parameters, however, the validation pseudoexperiments are drawn based on $(\boldsymbol{\theta}_0, \boldsymbol{\phi}_0)$ while the FC pseudoexperiments are drawn from $(\boldsymbol{\theta}_0, \hat{\hat{\boldsymbol{\phi}}}_{0i})$. Figure 8 shows how the coverages obtained under Wilks' theorem and the Profiled Feldman–Cousins approach vary for different intended coverages at the two points of parameter space considered above. Wilks' theorem generates widely different results depending on the region of the parameter space and can significantly deviate from the ideal coverage. The Profiled Feldman–Cousins method provides us with a more consistently accurate estimation of the desired coverage. Figure 8 hints that the magnitude of the corrections might decrease in the most extreme significance levels. This is not a general property and is further investigated in Section III F. We also performed a cross-check of the significance of our mass ordering determination using an alternative (and more conservative) method of handling nuisance parameters developed by Berger and Boos [11]. That procedure did not uncover a larger p-value than the one reported from the Profiled FC method, and so is consistent with that result. The details of this cross-check can be found in Appendix B.

F. Limitations and Features

The nominal output of the Feldman–Cousins method is a single confidence interval or region with proper coverage. However, it is straightforward and convenient to apply a Feldman–Cousins correction to a whole likelihood surface: each point has a likelihood, from that likelihood a p-value can be determined based on the distribution of FC pseudoexperiments at that point, and then from that p-value work backwards to an equivalent likelihood. This pseudo-likelihood surface is quite practical to work with since contours at any significance can be drawn using the Wilks' critical values. However, while the pseudo-likelihood superficially resembles an actual likelihood, it does not have the properties of a likelihood. Notably, it cannot be 'profiled' to reduce its dimensionality: a two-dimensional likelihood surface and its associated FC pseudoexperiments cannot be used to find one-dimensional confidence intervals.

The determination of the mass ordering in the most recent NOvA results provides a clear demonstration of this phenomenon [8]. The lowest significance for the Inverted Ordering has several different values in different projections of the significance: 0.6σ vs. $\sin^2\theta_{23}$ and 0.5σ vs. Δm^2_{32} or δ_{CP}. Mechanically, these differ since each projection is determined with different sets of experiments generated at different assumed true values. They are not expected to correspond in principle because assigning the likelihood of the Inverted Ordering as a whole to the lowest value of the likelihood when projected against another variable is an example of profiling, which is not a valid operation on these pseudo-likelihoods. The correct procedure is to generate FC pseudoexperiments specific to each question being asked, in this case a hypothesis test to determine the ordering. A benefit of the FC approach is that it can naturally accommodate binary questions like the neutrino mass ordering where the number of degrees of freedom for the Wilks' theorem approach is not well-defined, typically producing stronger constraints than applying Wilks' theorem with 1 degree of freedom. In this case, the significance calculated for this question directly is 1.0σ.

With this method, it is also possible for discontinuities in the corrected significance plot to emerge even if the

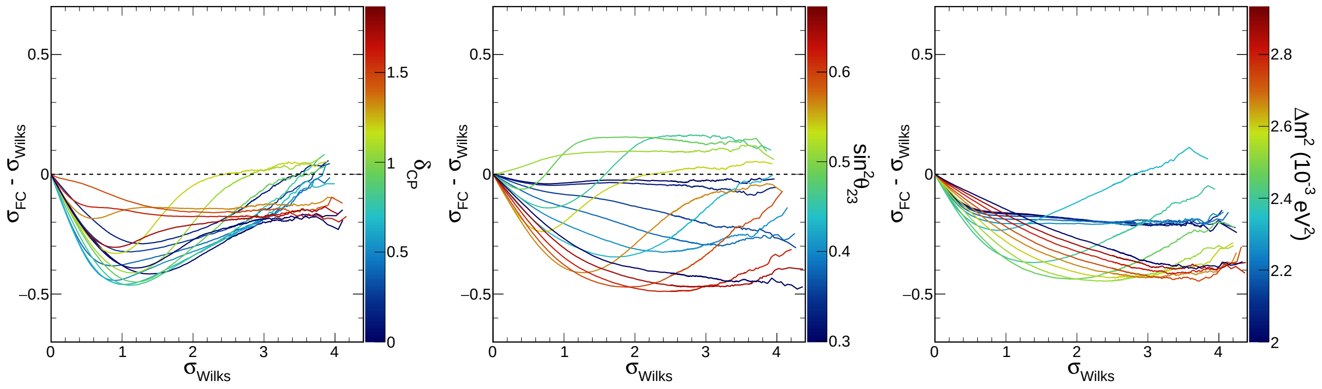


FIG. 10. The change in significance vs. the significance level at which the correction occurs for different values of, from left to right, δ_{CP}, $\sin^2\theta_{23}$, and Δm^2_{32}. The colors represent different true values of the parameter in question being tested.

underlying likelihood surface is smooth. An example of one of such a discontinuity can be found in Figure 9a around 0.1π in the plot of significance vs. δ_{CP} in the normal ordering, upper octant. This occurs because of a discontinuity in the profiled FC corrections, caused by a discontinuous change in the value of the nuisance parameters[11]. In this particular case, the global minimum moves from maximal mixing to the upper octant at this particular value of δ_{CP}, as shown in Figure 9b, leading to a change in the underlying λ distributions on either side of the discontinuity which then translates to different p-values for a given critical value, shown in Figure 9c.

A drawback of this method is its computation cost. We explored how the size of profiled FC corrections depends on the significance for which the correction is being computed. It would be convenient if the size of corrections became smaller as significance increases since corrections require more FC pseudoexperiments and get progressively more expensive to calculate at higher significance. We explored this question using the three plots which tested significance for different true values of δ_{CP}, $\sin^2\theta_{23}$, and Δm^2_{32}, and the results are shown in Figure 10. While the sizes of corrections clearly change as a function of significance, and for some true values the corrections converge towards zero, this is not true in general: the sizes of corrections at 4σ can be as large as the corrections at 2σ. In these examples, the *relative* size of the correction does decrease as the absolute significance gets larger, but we leave it to the reader to decide if the difference between 3.5σ and 4σ is more or less important than the difference between 1.5σ and 2σ.

Another limitation is that it is not possible to combine the corrected likelihoods from two separate experiments to produce a combined likelihood surface from a joint analysis. While it is possible to combine experiments using FC corrections, doing so requires more detailed information than is captured in just the likelihood and corrections [21].

IV. CONCLUSIONS

Statistical analysis is the window through which the results of experiments are viewed. Nowhere is this more true than when the parameters of interest cannot be observed directly but must be inferred using a model to interpret the observed data. The properties of that model, as well as the setup of the experiment itself, can distort the apparent power of the experiment, causing results to look more significant, or less significant, than they actually are, sometimes substantially. The Feldman–Cousins method serves a crucial role, providing a robust method for handling the common challenges that experiments encounter when Wilks' theorem cannot be relied upon, but the lack of a prescription for handling nuisance parameters complicates its adoption in practice. The Profiled FC method presented in this paper offers a straightforward prescription for handling nuisance parameters. Toy studies show the method achieves more accurate coverage when the true parameters of the underlying model are unknown compared to other plausible methods. In-situ tests in the NOvA analysis further validate the accuracy of the reported confidence intervals and significances. The Profiled FC method has been used in several NOvA oscillation analyses, including the most recent [5–8]. Given the strong basis in the literature, it is likely optimal in a wide variety of experimental contexts facing

[11] Discontinuous changes in the nuisance parameters when testing a continuous set of values of a parameter of interest are not a particular problem, and are quite common. Without FC corrections, these changes can cause a discontinuous change in the derivative of the likelihood, but do not make the value of the likelihood discontinuous

similar challenges with bounded parameters and small numbers of events. The most significant challenge to making use of Profiled FC (and Feldman–Cousins in general) is the large computational cost associated with generating and fitting the required FC pseudoexperiments. Our approach takes advantage of available High Performance Computing resources, but other approaches to improve the efficiency of this method are also being explored [22].

V. ACKNOWLEDGMENTS

This document was prepared by the NOvA collaboration using the resources of the Fermi National Accelerator Laboratory (Fermilab), a U.S. Department of Energy, Office of Science, HEP User Facility. Fermilab is managed by Fermi Research Alliance, LLC (FRA), acting under Contract No. DE-AC02-07CH11359. This work was supported by the U.S. Department of Energy; the U.S. National Science Foundation; the Department of Science and Technology, India; the European Research Council; the MSMT CR, GA UK, Czech Republic; the RAS, MSHE, and RFBR, Russia; CNPq and FAPEG, Brazil; UKRI, STFC and the Royal Society, United Kingdom; and the state and University of Minnesota. We are grateful for the contributions of the staffs of the University of Minnesota at the Ash River Laboratory, and of Fermilab. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising.

Appendix A: CLs Mass Ordering Significance

The CL_S method [23–25] was introduced as an alternative to traditional p-value calculations to address situations where an experiment might potentially make a claim of 'discovery' well beyond its sensitivity. In a nutshell, the method takes a ratio between the p-value for the null hypothesis, $\mathcal{H}^0$, and the potential discovery hypothesis, $\mathcal{H}^1$. In a true discovery, $p(\mathcal{H}^0) \ll p(\mathcal{H}^1)$, and the CL_S value will be small, while in a spurious claim, the data will be a poor fit to both hypotheses, so even though $p(\mathcal{H}^0)$ might be small, CL_S will be of order 1.

In the particular case of binary questions, the Profiled FC procedure can be naturally extended so the same FC pseudoexperiments can be re-used for the CL_S method . A mass ordering test is presented here, but the method is generic. Two modifications are needed. First, rather than evaluating $\ell_\text{constrained}$ and $\ell_\text{unconstrained}$, ℓ_NO and ℓ_IO are evaluated, but they can be readily re-interpreted: $\ell_\text{constrained}$ corresponds to the ℓ for the hypothesis being tested and $\ell_\text{unconstrained}$ corresponds to whichever ℓ is lower[12]. Second, FC pseudoexperiments need to be generated for both possible hypotheses, but given the relatively low computational cost of this test, this is a minor overall additional cost. Where the Profiled FC only reports the fraction of FC pseudoexperiments in the hypothesis being tested with λ larger than that observed in data, CL_S also requires the 'inverse': the fraction of FC pseudoexperiments generated under the hypothesis favored by the data with λ *lower* than that observed in the data, as shown in Figure 11. A small overlap of the two distributions would signify a strong discrimination power towards the mass ordering. Our data suggests a slight preference for the Normal Ordering.

Appendix B: Validation of Significance in Mass Ordering Determination

In the case of binary questions, like the choice of ordering, the situation is better thought of as a hypothesis test than a confidence interval, though they are closely related as described in Section II. For these cases, there is an alternative approach to handling nuisance parameters developed by Berger and Boos [11]. In this procedure, the p-value of a set of parameter values being tested, $\boldsymbol{\theta}$, is redefined as:

$$p_\text{BB}(\boldsymbol{\theta}) = \max_{\boldsymbol{\phi}} p(\boldsymbol{\theta}, \boldsymbol{\phi}) + \beta, \tag{B1}$$

where the max represents the largest p-value over all values of the nuisance parameters, $\boldsymbol{\phi}$, allowed at the β confidence level based on a fit to the data. By contrast, the Profiled Feldman–Cousins approach simply uses the p-value at $\hat{\hat{\boldsymbol{\phi}}}$, the maximum likelihood estimate of the nuisance parameters given $\boldsymbol{\theta}$:

$$p_\text{FC}(\boldsymbol{\theta}) = p(\boldsymbol{\theta}, \hat{\hat{\boldsymbol{\phi}}}), \tag{B2}$$

[12] Since FC pseudoexperiments generated in the Normal Ordering may have a better fit in the Inverted Ordering, and vice versa, these two ℓ's may be the same or not.

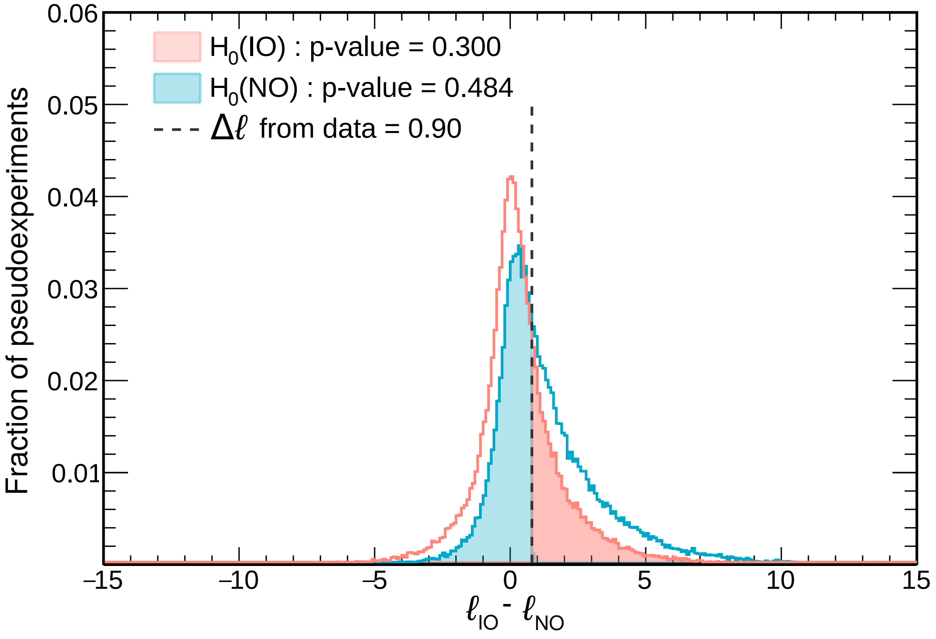


FIG. 11. Distribution of the likelihood ratio $\lambda = \ell_{\rm IO} - \ell_{\rm NO}$ for FC pseudoexperiments generated at the best fit points in the IO (red) and the NO (blue). The fraction of FC pseudoexperiments with a likelihood ratio more compatible with the null hypothesis than the data is smaller in the case of the NO, which suggests a preference for the latter. The resulting $\mathrm{CL_S}$ factor is 0.620.

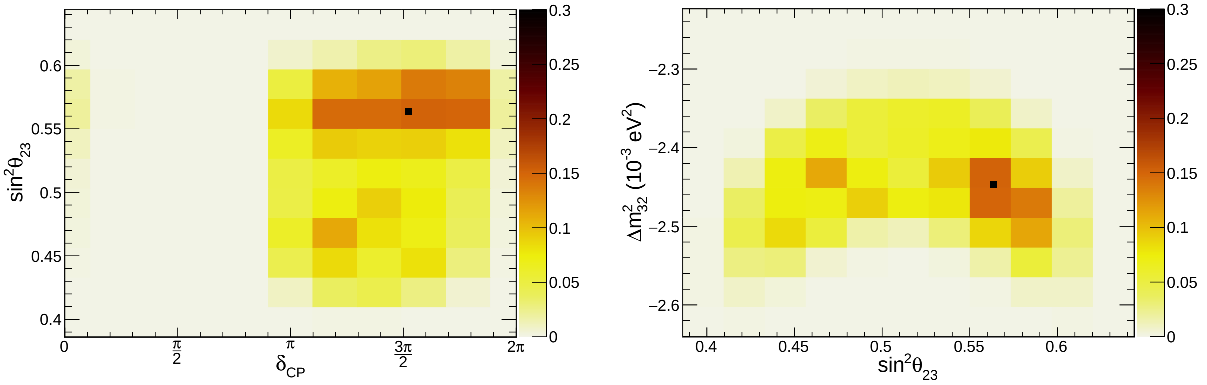


FIG. 12. The maximum p-values for the tested choices of nuisance parameters in the Berger–Boos test. All points in the full 3-dimensional space were tested, but only the largest p-value for each pair of values of the nuisance parameters is shown. All values are below $p = 0.30$, the maximum of the color scale and the significance of rejecting the inverted ordering at the best fit point, shown with a small square.

effectively assuming that the nuisance parameters which give the largest likelihood value (and thus the largest p-value under Wilks' theorem) will also have the largest p-value with the pseudoexperiment–calculated critical values. The Berger–Boos method is more conservative since it allows for the possibility that a seemingly non-optimal set of nuisance parameters will produce a 'favorable' change in the critical value and thus produce a larger effective p-value, but it is commensurately more costly to calculate since pseudoexperiments must be produced for a range of nuisance parameters.

In practice, it is not possible to test 'all' values in a multi-dimensional parameter space without an analytic form, so the possible choices of nuisance parameters must be sampled in a fashion which covers the possible space, and for each sampled set of nuisance parameters, a set of FC pseudoexperiments must be generated and used to calculate a new p-value. In this case, we are testing the p-value for rejecting the IO from the fit to data, $p = 0.30$ [8], so are taking a β of 0.005 which would not qualitatively alter the interpretation of the original p-value. This value of β then defines the ranges over which values of the nuisance parameters need to be sampled: a range in Δm^2_{32} of $[-2.623, -2.241] \times 10^{-3}\text{eV}^2$, a range in $\sin^2\theta_{23}$ of $[0.397, 0.633]$ and all values of δ_{CP}. Then, 1331 choices of nuisance parameters were tested (11 values in each dimension), sampled uniformly from the allowed space, and p-values were calculated for those choices. In order to save computational costs, pseudoexperiments were only generated for points where Feldman–Cousins corrections could plausibly raise it above the original p-value. The threshold chosen was $\lambda < 2.8$, which corresponds to $p_{\text{Wilks}} > 0.094$ assuming one degree-of-freedom. A total of 54 points fell below that threshold.

The largest p-value found was $p = 0.151$ at $\Delta m^2_{32} = -2.43 \times 10^{-3}\text{eV}^2$, $\sin^2\theta_{23} = 0.562$, and $\delta_{\text{CP}} = 1.64\pi$, which is below the $p = 0.30$ at the best fit point, so the original p-value is still the largest. This point had a $\lambda = 1.10$, which would give $p_{\text{Wilks}} = 0.295$ assuming one degree-of-freedom. This behavior was typical of most points for which FC pseudoexperiments were generated: p-values decreased (i.e., significances increased) since a binary question effectively has fewer degrees of freedom than one continuous parameter. Only 2 of the 54 points tested had $p > p_{\text{Wilks}}$, namely $p = 0.150$ and $p = 0.134$. The plots in Figure 12 show the largest p-values for rejecting the inverted ordering for different choices of the nuisance parameters.

[1] P.A. Zyla et al. (Particle Data Group), *Review of Particle Physics*, Chapter 40: Statistics. Prog. Theor. Exp. Phys. 2020, 083C01 (2020).

[2] J. Neyman. “Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability.” Philos. Trans. R. Soc. Lond. A, 236 (767): 333–380 (1937)

[3] S. S. Wilks, “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses.” Ann. Math. Statist. 9 (1) 60 - 62, March (1938)

[4] G. Feldman and R. Cousins. “Unified approach to the classical statistical analysis of small signals.” Phys. Rev. **D57** 3873-3889 (1998)

[5] P. Adamson *et al.* (NOvA Collaboration), “Constraints on Oscillation Parameters from ν_e Appearance and ν_μ Disappearance in NOvA,” Phys. Rev. Lett. **118**, no.23, 231801 (2017) arXiv:1703.03328 [hep-ex].

[6] M. A. Acero *et al.* (NOvA Collaboration), “New constraints on oscillation parameters from ν_e appearance and ν_μ disappearance in the NOvA experiment,” Phys. Rev. D **98**, 032012 (2018) arXiv:1806.00096 [hep-ex].

[7] M. A. Acero *et al.* (NOvA Collaboration), “First Measurement of Neutrino Oscillation Parameters using Neutrinos and Antineutrinos by NOvA,” Phys. Rev. Lett. **123**, no.15, 151803 (2019) arXiv:1906.04907 [hep-ex].

[8] M. A. Acero *et al.* (NOvA Collaboration), “Improved measurement of neutrino oscillation parameters by the NOvA experiment”, to appear in Phys. Rev. D. (2022) arXiv:2108.08219 [hepex].

[9] J. Neyman, E. S. Pearson. “On the problem of the most efficient tests of statistical hypotheses.” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 231 (694–706): 289–337 (1933)

[10] A. Stuart, K. Ord, and S. Arnold. *Kendall’s Advanced Theory of Statistics, Vol. 2A*, Chapter 22: Likelihood Ratio Tests and Test Efficiency (1999)

[11] R. L. Berger and D. D. Boos. “P values maximized over a confidence set for the nuisance parameter.” Journal of the American Statistical Association, Vol. 89, No. 427, pp. 1012-1016 (1994)

[12] R. Cousins and V. Highland. “Incorporating systematic uncertainties into an upper limit.” Nucl. Instr. and Meth. A320, 331 (1992).

[13] J. Conrad, O. Botner, A. Hallgren and C. P. de los Heros, “Coverage of confidence intervals for Poisson statistics in presence of systematic uncertainties,” Contribution to *Conference on Advanced Statistical Techniques in Particle Physics*, 58-63 (2002) arXiv:hep-ex/0206034 [hep-ex]

[14] F. Tegenfeldt and J. Conrad, “On Bayesian treatement of systematic uncertainties in confidence interval calculations,” Nucl. Instrum. Meth. A **539**, 407-413 (2005) arXiv:physics/0408039 [physics]

[15] D. Baxter *et al.* “Recommended conventions for reporting results from direct dark matter searches,” (2021) arXiv:2105.00599 [hep-ex]

[16] P. Adamson *et al.* (NOvA Collaboration). “First Measurement of Electron Neutrino Appearance in NOvA.” Phys. Rev. Lett. 116, 151806 (2016)

[17] `github.com/novaexperiment/fcplots/tree/main/ToyModel`

[18] P. Adamson, et al. (NOvA Collaboration). “The NuMI Neutrino Beam.” Nucl. Instr. and Meth. A806, 279 (2016)

[19] J. Altegoer *et al.* (NOMAD Collaboration), “A Search for $\nu_\mu \to \nu_\tau$ oscillations using the NOMAD detector,” Phys. Lett. B **431**, 219-236 (1998)

[20] F. James. “MINUIT Function Minimization and Error Analysis: Reference Manual Version 94.1.” CERN-D-506 (1994)

[21] A. V. Waldron, M. D. Haigh, and A. Weber. “Combining neutrino oscillation experiments with the Feldman–Cousins method.” New J. Phys. **14** 063037 (2012)

[22] L. Li, N. Nayak, J. Bian, and P. Baldi. “Efficient neutrino oscillation parameter inference using Gaussian processes.” Phys. Rev. **D101**, 012001 (2020)

[23] T. Junk. “Confidence Level Computation for Combining Searches with Small Statistics” Nucl. Instrum. Methods A434,

435 (1999).

[24] A. L. Read, “Modified frequentist analysis of search results (the CL_S method)”, in F. James, L. Lyons, and Y. Perrin (eds.), Workshop on Confidence Limits, CERN Yellow Report 2000-005, available through `cdsweb.cern.ch`

[25] A. L. Read. “Presentation of search results: the CL_S technique” J. Phys. G: Nucl. Part. Phys. 28 2693 (2002)