Likelihood-Free Frequentist Inference for Calorimetric Muon Energy Measurement in High-Energy Physics

Luca Masserano
Department of Statistics and Data Science
Carnegie Mellon University
lmassera@andrew.cmu.edu

Tommaso Dorigo
INFN
Sezione di Padova
tommaso.dorigo@cern.ch

Rafael Izbicki
Department of Statistics
Federal University of São Carlos
rafaelizbicki@gmail.com

Mikael Kuusela
Department of Statistics and Data Science
Carnegie Mellon University
mkuusela@andrew.cmu.edu

Ann B. Lee
Department of Statistics and Data Science
Carnegie Mellon University
annlee@stat.cmu.edu

Abstract

Muons have proven to be excellent probes of new physical phenomena, but the precision of traditional curvature-based measurements of their energy degrades at high energies. Recent work has shown the feasibility of a new avenue for the precise estimation of high-energy muons by exploiting the pattern of energy losses in a dense, finely segmented calorimeter using convolutional neural networks (CNNs). However, CNN predictions of the muon energy suffered from significant bias, which hampers the reliability of traditional methods for quantifying the uncertainty of the estimates. Indeed, to date, there is no known solution to the general problem of producing reliable uncertainty estimates of internal parameters of a statistical model from point predictions. In this paper, we propose WALDO, a new method that reframes the Wald test and uses the Neyman construction to convert point predictions into valid confidence sets. We show that WALDO achieves confidence sets with correct coverage regardless of the true muon energy value, while leveraging predictions from a CNN over a high-dimensional input space. In addition, we show that despite an increasing dimensionality, WALDO is able to extract useful information from a finer segmentation of the calorimeter, yielding smaller confidence sets, and hence more precise estimates of the muon energies.

1 Introduction

Muons are essential in studies of fundamental physics because they have proven to be excellent probes of new phenomena: the precise estimation of their energy and direction has allowed the discovery of particles that produce muons in their decay, including the Higgs boson [1, 3, 4, 6, 13]. Traditionally, muon energy estimates rely only on a measurement of their momentum given the curvature radius of the particle's trajectory. At very high energies, this leads to low precision because the trajectory becomes empirically indistinguishable from a straight line within practically achievable

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

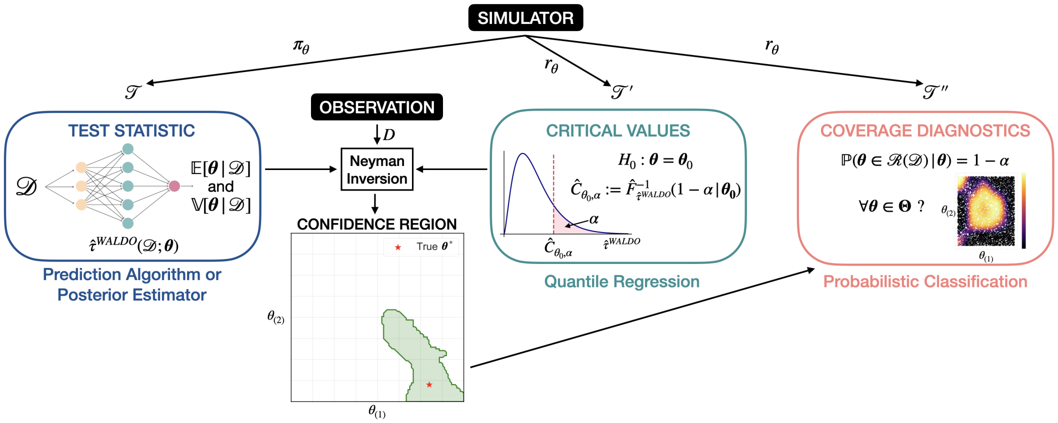

Figure 1: **Schematic diagram of WALDO.** The Neyman construction of confidence sets (*bottom*) requires estimation of a test statistic τ^{WALDO} (*left*), and of critical values $C_{\theta_0,\alpha}$ across the whole parameters space (*center*). Diagnostics (*right*) are then used to make sure that those confidence regions achieve the desired level of conditional coverage. See Section 2.1 and Algorithm 1 for details.

magnetic fields. On the other hand, the probability of energy losses along the muon path increases at high energy and within dense materials, lending itself to a complementary estimate based on the detected energy released in calorimeters. Leveraging this fact, we pose two questions: ***(i)*** Can we construct confidence sets with correct coverage of the true energy of ultra-relativistic muons using the information contained in the pattern and magnitude of radiative deposits in a dense calorimeter? ***(ii)*** Is it possible to extract additional information from finer segmentations of the calorimeter to allow for tighter constraints (i.e., smaller confidence sets with correct coverage) on muon energy estimates? If so, how large is the improvement? Quantifying the latter would allow scientists to optimize their detector designs, since manufacturing very small calorimeter cells is expensive.

Following the work of Kieseler et al. (2022) [15], we use an idealized experiment with calorimetric muon data simulated through GEANT4 [2]. This high-fidelity stochastic simulator encodes the likelihood function only implicitly, thus making standard statistical tools unsuitable. Note that the objective is to infer an unknown internal parameter θ – the muon energy – of the physical process that generates the data, using just one observation $\mathbf{x}$ (the calorimetric energy deposits) and simulations of the form $\{(\theta^{(j)}, \mathcal{D}^{(j)})\}_{j=1}^{B}$; that is, the observed sample size $n = 1$, hence $\mathcal{D}^{(j)} = \mathbf{x}^{(j)}$. This is a challenging inverse problem, which lends itself naturally to be tackled within a Likelihood-Free Inference (LFI) setting. The work done in [12] was the first to investigate how the information carried by muon-calorimeter interactions may be used to obtain estimates of muon energy in a particle collider. The authors exploited 16 custom input features derived from the raw energy deposits to predict muon energy using a gradient-boosted k-nearest-neighbor algorithm. This work was further expanded in [15], which leveraged the granularity of high-dimensional calorimeter data and the capacity of convolutional neural networks to improve the precision of muon energy estimates. Neural networks are indeed particularly suited when inputs $\mathbf{x}$ are very high-dimensional (for energy deposits in a calorimeter, $\mathbf{x} \in \mathbb{R}^{51,200}$) and the outcome variable is a scalar quantity (muon energy θ). These models have proved to be very powerful, but the above work also clearly showed that predictions of θ suffered from a strong bias, mainly due to the low signal-to-noise ratio in the calorimeter data at very high-energies. If used with conventional uncertainty quantification methods for prediction algorithms, this leads to inaccurate uncertainties that do not guarantee the desired coverage level across the parameter space, thereby hindering scientific conclusions.

To date, there is no known solution to the general problem of producing reliable uncertainty estimates of internal parameters of a statistical model from point predictions. Here we introduce WALDO, a novel method that reframes the Wald test [22] and uses the Neyman construction [18] to convert point predictions into confidence sets $\mathcal{R}(\mathcal{D})$ such that

$$\mathbb{P}(\theta \in \mathcal{R}(\mathcal{D})|\theta) = 1 - \alpha, \quad \forall \theta \in \Theta, \tag{1}$$

that is, $\mathcal{R}(\mathcal{D})$ has correct conditional coverage across the whole parameter space. It does so by leveraging prediction algorithms to compute a test statistic based on estimates of the conditional mean $\mathbb{E}[\theta|\mathcal{D}]$ and conditional variance $\mathbb{V}[\theta|\mathcal{D}]$. Figure 1 summarizes the components of WALDO and Algorithm 1 details the steps needed to construct $\mathcal{R}(\mathcal{D})$. We pair WALDO with the CNN of [15] and show that the resulting confidence sets are both *conditionally valid* and also *shorter* than standard prediction intervals in the muon energy problem. The code to reproduce our results is available at `https://github.com/anonymoussoftware/waldo`. A flexible implementation of the whole framework is available for installation on PyPI at `https://pypi.org/project/lf2i/`.

Notation We refer to the parameter of interest as $\theta \in \Theta \subset \mathbb{R}$ and to a sample of observable data of size n as $\mathcal{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, with $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$. Note that n is distinct from B, B' and B'', i.e., the number of simulations required at different steps of our method. For the muon energy problem, we have $n = 1$ and $\mathcal{D} = \mathbf{x}$ since each observation comes from a different parameter value. We distinguish between observable data and actual observations by denoting the latter as D. We refer to confidence regions as $\mathcal{R}(\mathcal{D})$. The terms "set", "region" and "interval" are used interchangeably.

2 Methodology

Algorithm 1: Confidence set for θ via WALDO

- **// Estimate conditional mean and variance**
- Simulate $\mathcal{T} = \{(\theta^{(j)}, \mathcal{D}^{(j)})\}_{j=1}^{B}$
- Estimate $\mathbb{E}[\theta|\mathcal{D}]$ on $\mathcal{T}$ under squared error loss
- Compute $(\theta^{(j)} - \mathbb{E}[\theta|\mathcal{D}^{(j)}])^2 \; \forall j = 1, \ldots, B$ and estimate $\mathbb{V}[\theta|\mathcal{D}] = \mathbb{E}[(\theta - \mathbb{E}[\theta|\mathcal{D}])^2|\mathcal{D}]$
- **// Estimate critical values**
- Simulate $\mathcal{T}' = \{(\theta^{(j)}, \mathcal{D}^{(j)})\}_{j=1}^{B'}$
- Predict $\{\widehat{\mathbb{E}}[\theta|\mathcal{D}^{(j)}], \widehat{\mathbb{V}}[\theta|\mathcal{D}^{(j)}]\}_{j=1}^{B'}$
- Compute $\{\widehat{\tau}^{\text{WALDO}}(\mathcal{D}^{(j)}; \theta^{(j)})\}_{j=1}^{B'}$
- Estimate critical values $C_{\theta,\alpha}$ via quantile regression of $\widehat{\tau}^{\text{WALDO}}(\mathcal{D}; \theta)$ on θ
- **// Confidence set via Neyman inversion**
- Predict $\widehat{\mathbb{E}}[\theta|D]$ and $\widehat{\mathbb{V}}[\theta|D]$
- Predict $\widehat{C}_{\theta_0,\alpha} \; \forall \theta_0 \in \Theta_{grid}$
- Initialize $\mathcal{R}(D) \leftarrow \emptyset$
- **for** $\theta_0 \in \Theta_{grid}$ **do**
  - **if** $\widehat{\tau}^{\text{WALDO}}(D; \theta_0) \leq \widehat{C}_{\theta_0;\alpha}$ **then**
    - $\mathcal{R}(D) \leftarrow \mathcal{R}(D) \cup \{\theta_0\}$
- **return** confidence set $\mathcal{R}(D)$

A key ingredient of WALDO is the equivalence between hypothesis tests and confidence sets, which was formalized by Neyman in 1937 [18] and widely applied in high-energy physics [8, 9]. The basic idea is to invert a series of level-α hypothesis tests of the form

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0, \tag{2}$$

for all $\theta_0 \in \Theta$. After observing a sample D, one constructs a confidence region $\mathcal{R}(D)$ for θ by taking all values of θ_0 that were not rejected by the series of tests above. By construction, the set $\mathcal{R}(D)$ satisfies Equation 1, i.e., it has the correct $1 - \alpha$ coverage level across the *entire* Θ. Albeit simple, the Neyman construction is hard to implement in practice within an LFI setting without resorting to large-n approximations like Wilks' theorem [23], since it requires estimating critical values $C_{\theta_0,\alpha}$ that define the level-α acceptance region for *every* hypothesis test that we invert. In the context of LFI, one usually either resorts to asymptotic results [8, 16] or Monte Carlo approaches [17, 21], but the latter become computationally prohibitive as the dimensionality of the parameter space increases [7].

Recently, [10, 11] proposed a fast construction of Neyman confidence sets with finite-n conditional coverage in an LFI setting; the general machinery was referred to as likelihood-free frequentist inference (LF2I). Assuming we have access to a simulator F_θ that can produce high-fidelity observable data $\mathcal{D}$ at different parameter settings θ, LF2I breaks down the construction of a confidence set (including diagnostics) into the following steps: ***(i)*** estimating a test statistic $\tau(\mathcal{D}, \theta)$ from a first simulated set $\mathcal{T}$, ***(ii)*** estimating critical values $C_{\theta,\alpha}$ from a second simulated set $\mathcal{T}'$, ***(i)*** + ***(ii)*** constructing the confidence set by retaining all θ for which the corresponding test does not reject the null, and ***(iii)*** an independent check of the empirical conditional coverage $\mathbb{P}[\theta \in \mathcal{R}(\mathcal{D})|\theta]$ of the constructed set across all $\theta \in \Theta$ using a third simulated set $\mathcal{T}''$. Dalmasso et al. (2022) [11] proved that, by using quantile regression for ***(ii)***, one can control type I error at level α for all $\theta \in \Theta$, regardless of the test statistic. In their work, they defined likelihood-based test statistics estimated via probabilistic classification, whereas we base our framework on a new test statistic that leverages point predictions from algorithms such as CNNs (see Section 2.1).

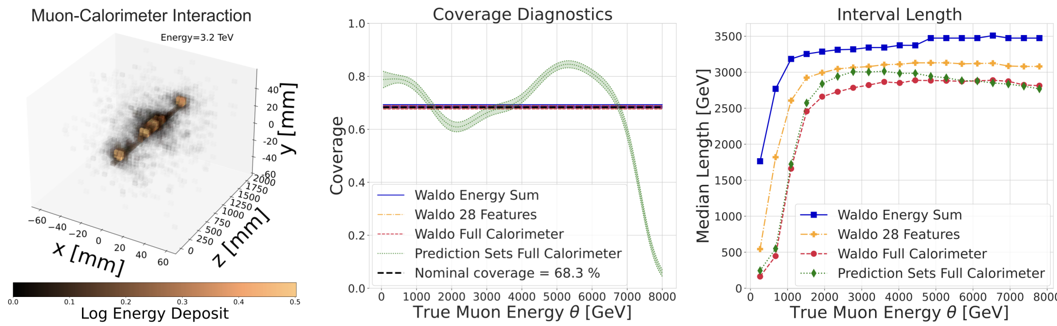

Figure 2: **Muon energy estimates via calorimetric measurements of increasing dimensionality.** WALDO (blue, orange, red) guarantees nominal coverage (68.3%), while 1σ prediction intervals (green) under- or over-cover in different regions of Θ and are wider on average than the corresponding WALDO intervals. *Left:* Energy deposited by a $\theta \approx 3.2$ TeV muon entering a homogeneous lead tungstate calorimeter with $32\times32\times50 = 51{,}200$ cells. *Center:* empirical coverage estimated via WALDO diagnostics: reliable inference requires coverage close to the nominal level. *Right:* Median lengths of intervals: lower values imply higher precision (i.e., smaller sets).

2.1 WALDO: bridging classical statistics with prediction algorithms

Since any test that controls type I error at level α may be used, we can exploit the one introduced by Wald [22], which is based on the following test statistic: $\tau^{\text{WALD}}(\mathcal{D};\theta_0) := (\widehat{\theta}^{\text{MLE}} - \theta_0)^2/\mathbb{V}(\widehat{\theta}^{\text{MLE}})$, where $\widehat{\theta}^{\text{MLE}}$ is the maximum-likelihood estimator of θ and $\mathbb{V}(\widehat{\theta}^{\text{MLE}})$ can be any consistent estimator of its variance. WALDO replaces $\widehat{\theta}^{\text{MLE}}$ and its variance in the Wald statistic with the typical output of prediction algorithms. We define the WALDO test statistic as

$$\tau^{\text{WALDO}}(\mathcal{D};\theta_0) := \frac{(\mathbb{E}[\theta|\mathcal{D}] - \theta_0)^2}{\mathbb{V}[\theta|\mathcal{D}]}, \tag{3}$$

where $\mathbb{E}[\theta|\mathcal{D}]$ and $\mathbb{V}[\theta|\mathcal{D}]$ are the conditional mean and variance of θ given the data $\mathcal{D}$, respectively. We can then replace the step ***(i)*** in the LF2I machinery in the following way: Estimate $\mathbb{E}[\theta|\mathcal{D}]$ and $\mathbb{V}[\theta|\mathcal{D}]$ by leveraging the fact that prediction algorithms approximate the conditional mean of the outcome variable given the inputs $\mathcal{D}$, when minimizing the squared error loss. The predictor is trained on a simulated set $\mathcal{T} = \{(\theta^{(j)}, \mathcal{D}^{(j)})\}_{j=1}^{B}$, where θ can be drawn from any prior distribution π_θ.

Remark. Note that the building blocks of the WALDO test statistic in Equation 3 can also be seen as the posterior mean $\mathbb{E}[\theta|\mathcal{D}]$ and posterior variance $\mathbb{V}[\theta|\mathcal{D}]$. We can then leverage modern neural posterior estimators (such as normalizing flows [19]) and approximate the above quantities via Monte Carlo sampling from the estimated posterior distribution. This variant of the approach is particularly amenable for settings with multiple parameters of interest and observed samples of size $n > 1$.

3 Measuring high-energy muons with a finely segmented calorimeter

We now return to the main goal of this work: constructing *valid* (correct conditional coverage) and *precise* (tight) confidence sets for high-energy muons using the pattern and magnitude of the radiated energy deposits in a dense calorimeter, which would then address the key questions ***(i)*** and ***(ii)*** outlined in Section 1. We have available 886,716 3D input "images" $\mathbf{x}$ and scalar true muon energies θ obtained through GEANT4 [2], a high-fidelity stochastic simulator. See Figure 2 (left panel) for an illustration of one simulated $\mathbf{x}_i$ for a particular θ_i. The data are available in [14]. As the interest is on constraining muon energies as much as possible while guaranteeing conditional coverage, we use three versions of the same dataset with increasing dimensionality: a 1D input equal to the sum over all calorimeter cells with deposited energy $E > 0.1$ GeV, for each muon; 28 custom features extracted from the spatial and energy information of the calorimeter cells as described in [15]; and the full calorimeter measurements ($\mathbf{x}_i \in \mathbb{R}^{51,200}$). For the first two datasets, we estimate $\mathbb{E}[\theta|\mathcal{D}]$ and $\mathbb{V}[\theta|\mathcal{D}]$ via Gradient Boosted Trees [5]. For the full calorimeter data, we leverage the CNN developed in [15]. We use Gradient Boosted Trees for quantile regression [20].

Answering ***(i)*** in Section 1 affirmatively, Figure 2 (center) shows that confidence sets constructed with WALDO achieve exact conditional coverage (68.3%) regardless of the dataset used. The corresponding

1σ prediction intervals ($\mathbb{E}[\theta|\mathcal{D}] \pm \sqrt{\mathbb{V}[\theta|\mathcal{D}]}$) using full calorimeter data, instead, exhibit over- or under-coverage in different regions over Θ, which in the latter case means that prediction sets contain the true value with much lower probability than anticipated. As for question ***(ii)***, we make two observations (see Figure 2; right panel): First, using the raw higher-dimensional energy deposits with WALDO allows to reduce the uncertainty around muon energies. Second, confidence sets constructed with WALDO are even shorter than the corresponding prediction intervals, while also guaranteeing conditional coverage.

Broader impact statement Our work introduces a new method, WALDO, that converts point predictions into conditionally valid confidence sets of internal parameters in an LFI setting. By leveraging its properties, we showed that ***(i)*** it is possible to construct confidence sets with correct coverage for the energy of ultra-relativistic muons using their interactions with dense calorimeters; and ***(ii)*** finer segmentations of the calorimeter carry additional information which Waldo can exploit to further constrain muon energies. Domain sciences, particularly the physical sciences, routinely seek to constrain parameters of interest using theoretical (or simulation) models together with experimental data. Assuming we have access to a high-fidelity simulator, WALDO provides reliable constraints that can be used to deduce trustworthy scientific conclusions in situations where other uncertainty quantification methods are either unavailable, unreliable or inefficient.

Acknowledgements We thank Niccolò Dalmasso for early feedback and discussions on this work, and for providing code previously written for LF2I. We are also indebted to Jan Kieseler and to Giles C. Strong for providing the muon energy data and the structure of the deep neural network employed for the studies described in Section 3, respectively. We also thank the STAMPS research group for many valuable discussions on the details of WALDO. This work is supported in part by NSF DMS-2053804, NSF PHY-2020295, and the C3.ai Digital Transformation Institute. RI is grateful for the financial support of CNPq (309607/2020-5) and FAPESP (2019/11321-9). We are also grateful to Microsoft for providing Azure computing resources for this work.

References

[1] Georges Aad, Tatevik Abajyan, B Abbott, J Abdallah, S Abdel Khalek, Ahmed Ali Abdelalim, R Aben, B Abi, M Abolins, OS AbouZeid, et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716 (1):1–29, 2012.

[2] Sea Agostinelli, John Allison, K Amako, John Apostolakis, H Araujo, Pedro Arce, Makoto Asai, D Axen, Swagato Banerjee, G Barrand, et al. GEANT4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research A*, 506(3):250–303, 2003.

[3] J-E Augustin, Adam M Boyarski, Martin Breidenbach, F Bulos, JT Dakin, GJ Feldman, GE Fischer, D Fryberger, G Hanson, B Jean-Marie, et al. Discovery of a narrow resonance in e+ e- annihilation. *Physical Review Letters*, 33(23):1406, 1974.

[4] Serguei Chatrchyan, Vardan Khachatryan, Albert M Sirunyan, Armen Tumasyan, Wolfgang Adam, Ernest Aguilo, Thomas Bergauer, M Dragicevic, J Erö, C Fabjan, et al. Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc. *Physics Letters B*, 716(1): 30–61, 2012.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[6] Cdf Collaboration et al. Observation of top quark production in Pbar-P collisions. *arXiv preprint hep-ex/9503002*, 1995.

[7] Robert D. Cousins. Lectures on statistics in theory: Prelude to statistics in practice, 2018.

[8] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71(2):1–19, 2011.

[9] Kyle Cranmer. Practical Statistics for the LHC. *arXiv e-prints*, art. arXiv:1503.07622, Mar 2015.

[10] Niccolo Dalmasso, Rafael Izbicki, and Ann Lee. Confidence sets and hypothesis testing in a likelihood-free inference setting. In *International Conference on Machine Learning*, pages 2323–2334. PMLR, 2020.

[11] Niccolo Dalmasso, Luca Masserano, David Zhao, Rafael Izbicki, and Ann B Lee. Likelihood-free frequentist inference: Confidence sets with correct conditional coverage. *arXiv preprint arXiv:2107.03920*, 2022.

[12] Tommaso Dorigo, Jan Kieseler, Lukas Layer, and Giles Strong. Muon energy measurement from radiative losses in a calorimeter for a collider detector. *arXiv preprint arXiv:2008.10958*, 2020.

[13] SW Herb, DC Hom, LM Lederman, JC Sens, HD Snyder, JK Yoh, JA Appel, BC Brown, CN Brown, WR Innes, et al. Observation of a dimuon resonance at 9.5 GeV in 400-GeV proton-nucleus collisions. *Physical Review Letters*, 39(5):252, 1977.

[14] Jan Kieseler, Giles Chatham Strong, Filippo Chiandotto, Tommaso Dorigo, and Lukas Layer. Preprocessed dataset for “Calorimetric measurement of multi-TeV muons via deep regression", August 2021. URL `https://doi.org/10.5281/zenodo.5163817`.

[15] Jan Kieseler, Giles C Strong, Filippo Chiandotto, Tommaso Dorigo, and Lukas Layer. Calorimetric measurement of multi-TeV muons via deep regression. *The European Physical Journal C*, 82(1):1–26, 2022.

[16] Louis Lyons and Nicholas Wardle. Statistical issues in searches for new phenomena in high energy physics. *Journal of Physics G: Nuclear and Particle Physics*, 45(3):033001, 2018.

[17] James G MacKinnon. Bootstrap hypothesis testing. *Handbook of computational econometrics*, 183:213, 2009.

[18] Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380, 1937.

[19] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[20] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[21] Valérie Ventura. Bootstrap tests of hypotheses. In *Analysis of parallel spike trains*, pages 383–398. Springer, 2010.

[22] Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482, 1943.

[23] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62, 03 1938. doi: 10.1214/aoms/1177732360.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Sections 2 and 3.
   (b) Did you describe the limitations of your work? [Yes] .
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 1, where we included a link to all the code used. The dataset used in Section 3 is referenced therein.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 3
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 3
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We included a link to our code in Section 1.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]