UNIVERSITY OF HELSINKI

Machine Learning Accelerated Likelihood-Free Event Reconstruction in Dark Matter Direct Detection

[a]U. Simola, [b]B. Pelssers, [b]D. Barge, [b]J. Conrad, [a]J. Corander
[a]University of Helsinki, Department of Mathematics and Statistics, Helsinki, Finland
[b] Stockholm University, Department of Physics, Stockholm, Sweden

Likelihood-Free Inference Workshop

18-22 March 2019 @ Flatiron Institute, NYC

Outline

UNIVERSITY OF HELSINKI

What is Dark Matter?

According to most recent studies, we cannot explain **85% of matter in the Universe**. Why so?

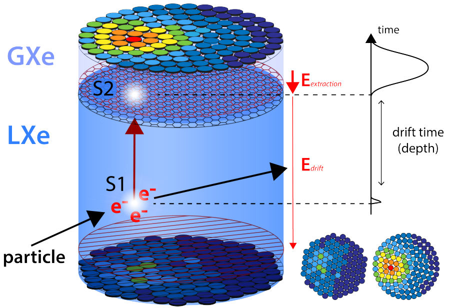
1. Looking for **something unseen** for redeeming the **Baryonic matter** theory (i.e. Dark Matter is made of ordinary matter after all):
   - Massive Astrophysical Compact Halo Object (M.A.C.H.O.)
   - Modified Newtonian Dynamics (Mo.N.D.)
   - Tensor–Vector–Scalar Gravity (Te.Ve.S.)
   - ...
2. **Move towards a Non-baryonic matter theory**
   - Supersymmetric particles
   - Gravitationally-Interacting Massive Particles (G.I.M.Ps)
   - **Weakly Interacting Massive Particles (W.I.M.Ps)**
   - ...

UNIVERSITY OF HELSINKI

Are we living in a WIMP–World?

- WIMP is a **hypothetical massive particle**, that are thought to constitute dark matter, **which interacts only via gravity and the weak nuclear force**.
- **Searching for WIMPs is done in 3 ways**:
  1. **Production**: in particle colliders such as the LHC
  2. **Annihilation**: indirect detection
  3. **Scattering**: direct detection

XENON1T detector



Dual–phase Time Projection Chambers use both the **scintillation** and the **ionization** signals to detecting particles scattering on atoms in the detector

- **Distinguish** between **nuclear and electronic recoil events**
- **Properly reconstruct the spatial position of the recoil events**, to discard background events at the edge of the TPC
- We use the **S2 hit-pattern** to reconstruct the **2–D (x, y) spatial position** of the recoil events

Statistical Model

- **Top and bottom S2 hit-patterns** consist of respectively on **127** and **121 photomultiplier tubes (PMTs)**
- The **likelihood function usually defined** assumes that **the number of photoelectrons counted in a certain PMT follows a Poisson** distribution with parameter:

$$\lambda_i = N_{obs} \frac{LCE_i(x,y)}{\sum_{j \in PMTs} LCE_j(x,y)}, \tag{1}$$

with N_{obs} being the total number of observed photoelectrons in the S2 hit pattern and $LCE_i(x,y)$ is the **light collection efficiency (LCE)** function of PMT i for photons produced at position (x,y).

Drawbacks of needing the LCE maps

- The **statistical model** does not include any other processes beyond the Poisson process. We know that there are **PMT afterpulses and detector systematics** which are very hard to include in the statistical model
- **The LCE functions are not analytically known** but are rather numerically estimated using optical photon Monte Carlo simulations
- Those simulations take into account both **the geometry of the detector** and **the optical and reflective properties of the employed materials**
- **The LCE maps are not defined on the continuum** but rather simulated on a grid, after which an interpolation is used

UNIVERSITY OF HELSINKI

Approximate Bayesian Computation

- Approximate Bayesian Computation (ABC) is a framework for inference for situations in which the **likelihood function is intractable** ("likelihood-free" approach)
- Issues with writing down a likelihood:
  1. The statistical model is too complex
  2. No general accepted theory is available
  3. Strong dependency in the data
  4. Observational limitations (i.e. truncations and censures)
- A **forward process / simulator** is available
- The goal is to retrieve **a suitable approximation of the posterior distribution**

Basic ABC algorithm

Algorithm 1 Basic ABC algorithm by *Pitchard et al., (1999)*

1: Sample θ_{prop} from the prior $\pi(\theta)$
2: Produce y_{prop} from the forward model $f(y \mid \theta_{\text{prop}})$
3: Define a **summary statistics** $s(\cdot)$, a **distance metric** $\rho(s(y_{\text{obs}}), s(y_{\text{prop}}))$ and a **tolerance** ϵ
4: Accept θ_{prop} if $\rho(s(y_{\text{obs}}), s(y_{\text{prop}})) < \epsilon$. Repeat until the desired particle sample size N is achieved

Any ABC procedure relies on:

- **Determining the tolerance** ϵ
- **Defining highly informative summary statistics** and **suitable distance functions** in order to **compare the observed and the simulated samples**

Bayesian Optimization for Likelihood–Free Inference

- One of the **major obstacles to likelihood–free inference is the computational cost** of the method, since **most of the parameters proposed result in large distances between observed and the simulated samples**
- The basic idea behind the Bayesian Optimization for Likelihood–Free Inference (BOLFI) is to find, avoiding unnecessary computations, **relevant regions of the parameter space** where the distance between observed and the simulated samples is small
- **The problem becomes to infer the stochastic relation between the parameter estimates and the distances**, and BOLFI addresses this task by using Gaussian processes

2–D (x, y) Position Reconstruction

Given the observed **S2 top hit pattern**, r_{obs}, and a simulated one, r_{prop}, 2 different BOLFI analyses are performed:

1. by using as distance function the **Euclidean distance**

$$\rho(r_{\text{obs}}, r_{\text{prop}})_{\text{Euclidean}} = \sqrt{\sum_{i=1}^{n} (r^i_{\text{obs}} - r^i_{\text{prop}})^2} \tag{2}$$

2. by using as distance function the **Bray–Curtis dissimilarity**

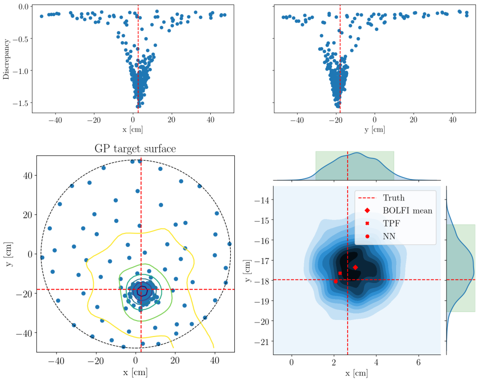
$$\rho(r_{\text{obs}}, r_{\text{prop}})_{\text{Bray–Curtis}} = \frac{\sum_{i=1}^{n} |r^i_{\text{obs}} - r^i_{\text{prop}}|}{\sum_{i=1}^{n} |r^i_{\text{obs}} + r^i_{\text{prop}}|}, \tag{3}$$

   where n is the total number of PMTs

3. Both priors x_{prop} and y_{prop} are Normally distributed

BOLFI output

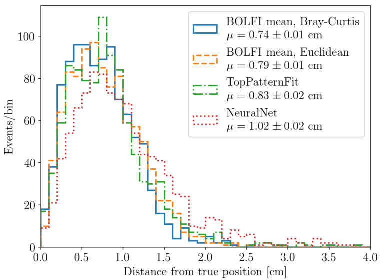
A typical output provided by BOLFI, where the input coordinates are ($x_{\text{input}} = 2.63$ cm, $y_{\text{input}} = -17.96$ cm)

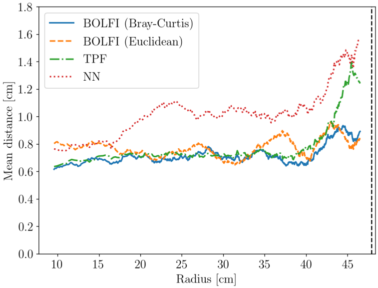


Comparison with the commonly employed methods

Let $d_{\text{euc}} = \sqrt{(x_{\text{input}} - x_{\text{rec}})^2 + (y_{\text{input}} - y_{\text{rec}})^2}$ as the Euclidean distance, a comparison with the commonly employed methods is done once 1000 events have been reconstructed

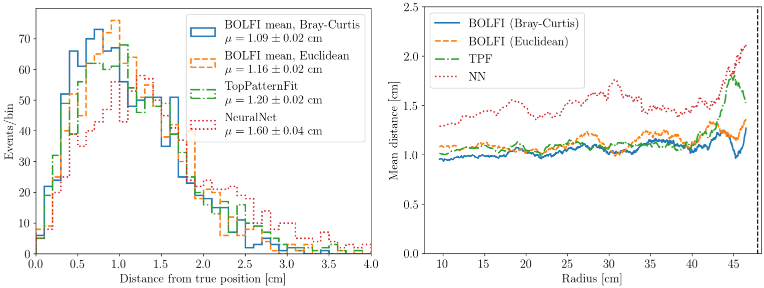
- Size of the charge signal = 25





Comparison with the commonly employed methods

- Size of the charge signal = 10



2–D (x, y) Position and Energy (e) Reconstruction

- Beyond the 2–D (x, y) Position, in this last example also **the number of ionization electrons e is unknown**
- Both the **S2 top and bottom hit patterns** are used
- The **energy distance is combined with the Bray–Curtis dissimilarity** to improve the quality of the comparison between observed and simulated hit patterns:

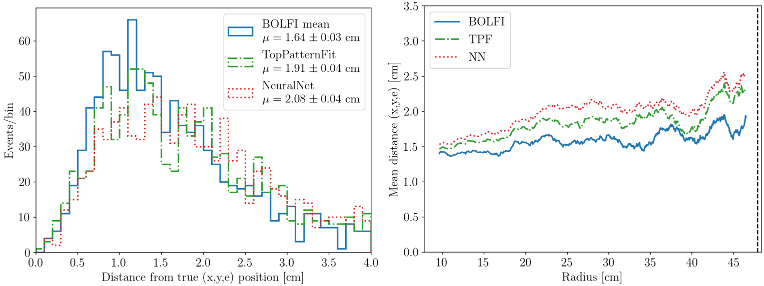
$$\rho(r_{\text{obs}}, r_{\text{prop}})_{\text{energy}} = \int_{-\infty}^{+\infty} \left(\hat{F}(r_{\text{obs}}) - \hat{F}(r_{\text{prop}})\right)^2 d\hat{F}(r_{\text{prop}}),$$

  where $\hat{F}(r_{\text{obs}})$ and $\hat{F}(r_{\text{prop}})$ are the densities estimated respectively using r_{obs} and r_{prop}.
- The prior for the energy is $e_{\text{prop}} = \text{logNormal}(\text{PAX e}, 5)$

Comparison with the commonly employed methods

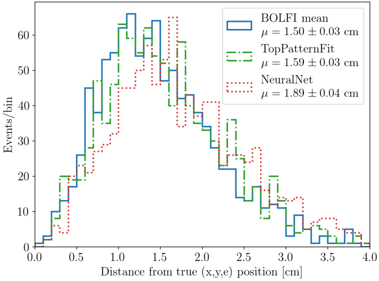
Let $d_{\text{euc}} = \sqrt{(x_{\text{input}} - x_{\text{rec}})^2 + (y_{\text{input}} - y_{\text{rec}})^2 + (e_{\text{input}} - e_{\text{rec}})^2}$ as the Euclidean distance, a comparison with the commonly employed methods is done once 1000 events have been reconstructed

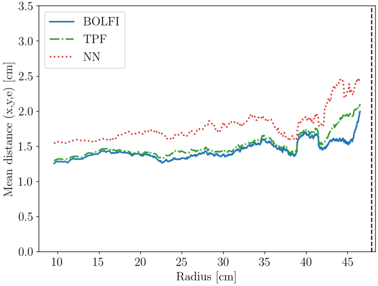
- Size of the charge signal = 25



Comparison with the commonly employed methods

- Size of the charge signal = 10





UNIVERSITY OF HELSINKI

Final Remarks

- When focusing on the **3–D (x, y, e) reconstruction, BOLFI improves the accuracy of the reconstruction over TPF by 14% and 5%**, respectively when the size of the charge signal is equal to 25 electrons and 10 electrons
- **BOLFI can reconstruct background events (radius $R >$ 30 cm) more precisely with respect to TPF**
- **The uncertainties associated to the parameters of interest retrieved by BOLFI are always the smallest among all the tested methods**
- **No LCE maps are needed**; the LCE information is used in the simulator but it **simulates extra processes that are not in the LCE maps**

UNIVERSITY OF HELSINKI

Essential Bibliography

- Simola, U., et al. Machine Learning Accelerated Likelihood-Free Event Reconstruction in Dark Matter Direct Detection. JINST, 14, P03004 (2019)
- XENON collaboration, The XENON1T Dark Matter Experiment, Eur. Phys. J. C 77 (2017) 881 [arXiv:1708.07051].
- M.U. Gutmann and J. Corander, Bayesian optimization for likelihood-free inference of simulator-based statistical models, J. Mach. Learn. Res. 17 (2016) 1.
- J. Lintusaari et al., Elfi: Engine for likelihood-free inference, J. Mach. Learn. Res. 19 (2018)
- Rubin, Vera C., W. Kent Ford Jr, and Norbert Thonnard. Extended rotation curves of high-luminosity spiral galaxies. IV-Systematic dynamical properties, SA through SC. The Astrophysical Journal 225 (1978): L107-L111. (Fig.1)
- XENON collaboration, The pax data processor v6.8.0, March (2018).

Thank You