THE UNIVERSITY of ADELAIDE

DarkMachines

High dimensional sampling project

Martin White, Joaquin Vanschoren

Aim of the challenge

- Find awesome techniques for sampling high dimensional parameter spaces that we have never used in particle astrophysics (with help from our machine learning experts)
- Develop realistic physics case studies, and evaluate each technique (with the help of our physics experts)
- Prepare a detailed summary that compares the techniques and presents our results

Today

- A brief reminder of the project
- Will present some tangible progress
- Will set up discussions/immediate coding problems for the working sessions

A typical particle astrophysics problem

We have a bunch of data from different experiments that might be sensitive to dark matter:

- colliders (LHC + previous)
- measurements of the magnetic moment of the muon
- electroweak precision tests
- dark matter direct detection experiments
- searches for antimatter in cosmic rays, nuclear cosmic ray ratios
- radio astronomy data
- effects of dark matter on reionisation, recombination and helioseismology
- relic density (CMB + other data)
- neutrino masses and mixings
- Indirect DM searches (e.g. FERMI-LAT, HESS, CTA, IceCube, etc)

A typical particle astrophysics problem

- We may have a particular theory of dark matter (e.g. a particular Lagrangian in particle physics)
- Which values of the parameters of that theory are preferred given the data?
- How probable or likely is the model relative to other models of dark matter?
- The likelihood of the model can be expressed as a composite likelihood assuming each set of measurements is independent:

$$\mathcal{L} = \mathcal{L}_{\mathrm{collider}}\mathcal{L}_{\mathrm{DM}}\mathcal{L}_{\mathrm{flavor}}\mathcal{L}_{\mathrm{EWPO}}\cdots$$

A typical particle astrophysics problem

$$\mathcal{L} = \mathcal{L}_{\rm collider}\mathcal{L}_{\rm DM}\mathcal{L}_{\rm flavor}\mathcal{L}_{\rm EWPO}\cdots$$

- We either want to map the shape of the multi-dimensional likelihood surface and use it to define confidence intervals (Frequentist), or use a prior and our likelihood to define a posterior, and map that (Bayesian)

- The likelihood is not known analytically, but can be mapped by sampling the function: for each parameter point, we can simulate the various experiments and compare the theoretical predictions to data to obtain a likelihood

Other problems

- These sorts of problems are ubiquitous in physics, e.g.: fitting parton density functions to experimental data to obtain the structure of the proton, extracting the neutrino sector parameters from accelerator and atmospheric data, extracting flavour physics parameters, …

- In each case, the parameters of the particle physics model are usually poorly constrained *a priori*, but there are additional nuisance parameters that are better-constrained (e.g. experimental systematics, mass measurements of SM particles, velocity of dark matter in the frame of the Earth, etc)

Slow likelihood calculations

- A particular feature of interesting problems is that the calculation of each likelihood might be very slow

- For the GAMBIT Large Hadron Collider observables, we managed to get our simulations to run in 5s, but this takes massive parallelisation

- PDF fits require a calculation that takes ~20s, and there are over 100 nuisance parameters to scan over

- Cosmological calculations that require simulating the recombination history of the universe might need minutes per point

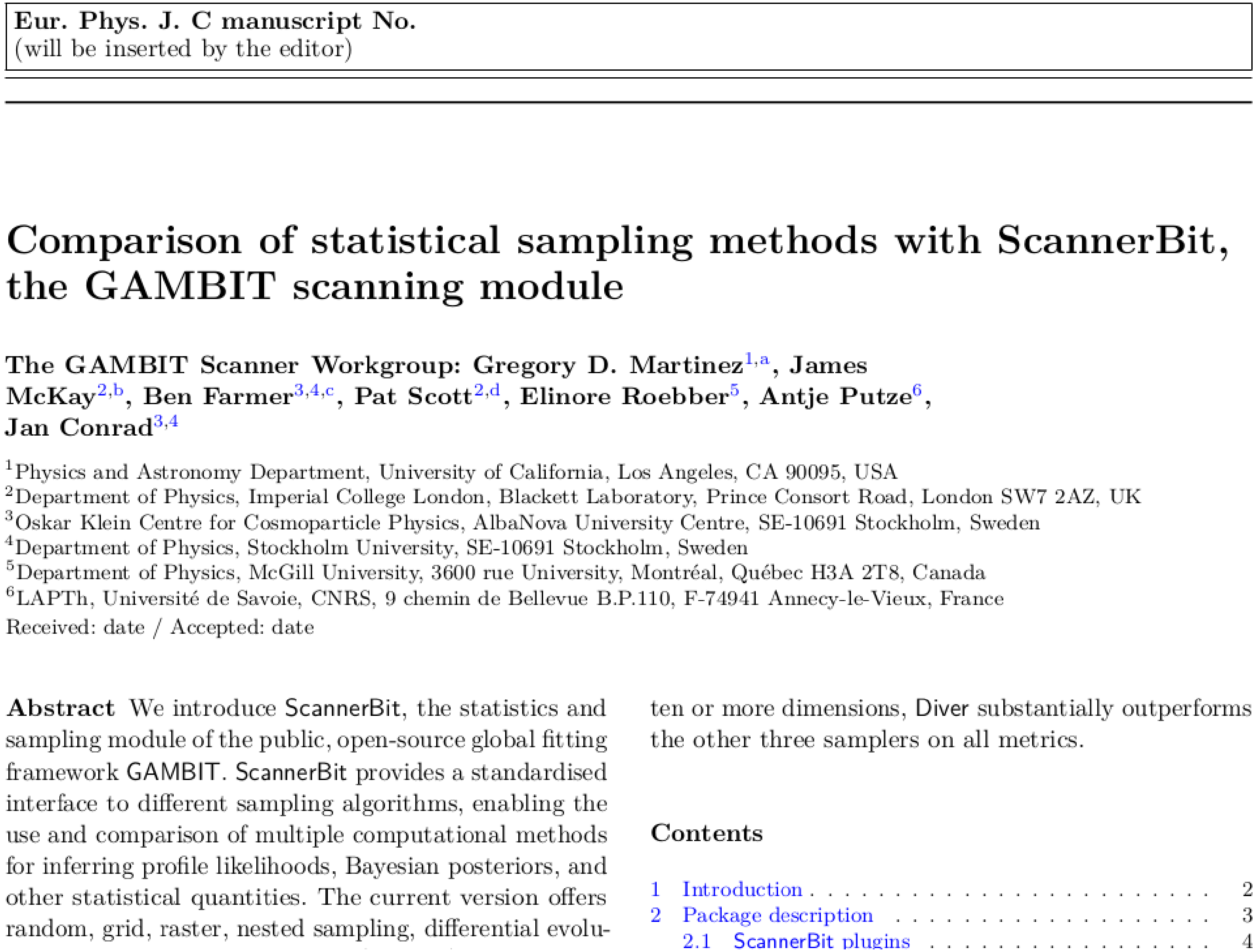
Not all problems are equally challenging

- The posterior is usually unimodal in cosmological applications (thus Markov Chain Monte Carlo techniques have remained popular), or in PDF fits

- The posterior is multimodal in, e.g., global fits of supersymmetric models, with very thin regions of interest in some cases (due to special conditions being needed to reproduce the correct dark matter relic density)

- Composite Higgs theories represent the biggest challenges I have yet seen (horrible thin sheets in the parameter space, which require delicate cancellations between sectors of the theory to get the right SM Higgs mass and quark masses)

GAMBIT sampling



Eur. Phys. J. C manuscript No.
(will be inserted by the editor)

Comparison of statistical sampling methods with ScannerBit, the GAMBIT scanning module

The GAMBIT Scanner Workgroup: Gregory D. Martinez[1,a], James McKay[2,b], Ben Farmer[3,4,c], Pat Scott[2,d], Elinore Roebber[5], Antje Putze[6], Jan Conrad[3,4]

[1]Physics and Astronomy Department, University of California, Los Angeles, CA 90095, USA
[2]Department of Physics, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK
[3]Oskar Klein Centre for Cosmoparticle Physics, AlbaNova University Centre, SE-10691 Stockholm, Sweden
[4]Department of Physics, Stockholm University, SE-10691 Stockholm, Sweden
[5]Department of Physics, McGill University, 3600 rue University, Montréal, Québec H3A 2T8, Canada
[6]LAPTh, Université de Savoie, CNRS, 9 chemin de Bellevue B.P.110, F-74941 Annecy-le-Vieux, France

Received: date / Accepted: date

Abstract We introduce ScannerBit, the statistics and sampling module of the public, open-source global fitting framework GAMBIT. ScannerBit provides a standardised interface to different sampling algorithms, enabling the use and comparison of multiple computational methods for inferring profile likelihoods, Bayesian posteriors, and other statistical quantities. The current version offers random, grid, raster, nested sampling, differential evolu-

ten or more dimensions, Diver substantially outperforms the other three samplers on all metrics.

Contents

ScannerBit algorithms

- ScannerBit contains custom code or interfaces for the following methods:

  Random

  Grid

  Markov Chain Monte Carlo (MCMC)

  Ensemble Monte Carlo

  Nested Sampling

  Differential evolution

Progress so far

- Kickoff meeting was held in October
- Initial suggestions for topics to investigate were:

1) Bayesian optimisation

2) Active learning

3) Multi-fidelity optimisation

4) Likelihood-free inference

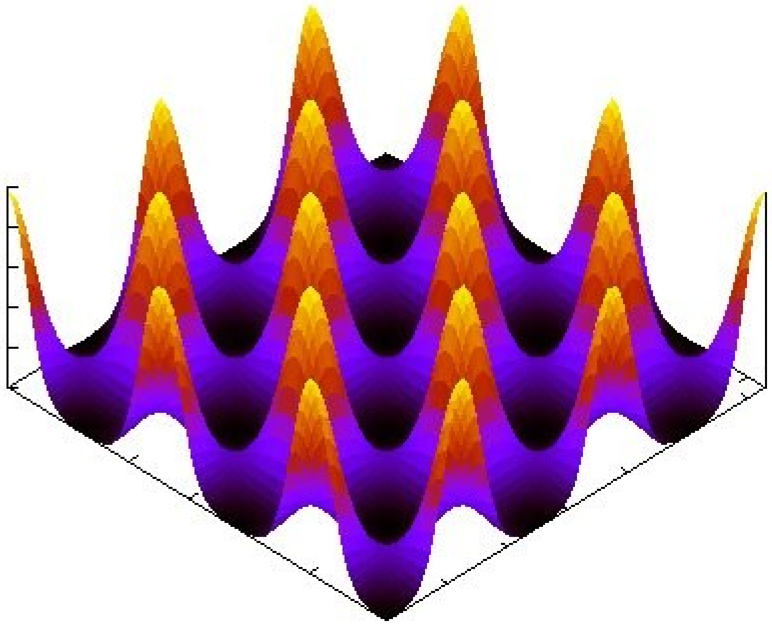
5) Multi-armed bandit, Thompson sampling.

6) Revisiting BAMBI with a better machine learning module

7) Training classifiers to get quick estimates of which models are physical for complex cases (e.g. composite Higgs, 19D SUSY).

Basic idea

- Repeat a ScannerBit style study with a wider variety of techniques, and a series of toy functions + physics cases
- Have initially settled on the MultiNest “eggbox” likelihood for testing



MNRAS, Vol. 421, Issue 1, pg. 169-180 (2012)

BAMBI

- BAMBI = MultiNest + artificial neural networks
- Basic idea: Train a neural network on the likelihood function during the sampling
- Start using the approximation to the likelihood when the estimate is deemed to be accurate enough
- The original version of BAMBI used the SkyNet neural net training algorithm (arXiv:1309.0790)

New: pyBAMBI

- The BAMBI functionality has been rewritten in python:

  https://github.com/DarkMachines/pyBAMBI
- Have replaced SkyNet with a keras-based neural net implementation
- There are still plenty of things to think about:

1) Come up with a sensible default for network architecture

2) Investigate training on the posterior, and not just the likelihood

3) Investigate different methods for assessing the accuracy of the trained net

Bayesian optimisation

- Have had a presentation from Eduardo Garrido Merchan
- Have recruited an Adelaide student to compare Bayesian optimisation with conventional global fit approach for toy problems (plus an IceCube example)
- Can we use this week to discuss particular Bayesian optimisation implementations that we want to try out/get some skeleton code?

Active learning

- Have had a presentation from Bob Stienen
- No active coding in the group so far
- Can we use this week to start the ball rolling?

Other ideas

- The bottle neck in LHC simulations is Monte Carlo event generation
- Can we use the Deep Generative Model approach outlined in arXiv:1901.00875 for signal generation?
- Suggestion: use the recent GAMBIT 4D EWMSSM study as a test case (only 4 signal parameters, events are relatively uncomplicated, thus a nice testbed)

Summary

- The sampling project is well-defined, but we should use this week to set up rapid progress over the next few months
- There are lots of places to contribute:

1) Further development of pyBAMBI

2) Commissioning Bayesian optimisation and active learning techniques in the code base

3) Detailed comparison of sampling methods

4) Novel approaches to LHC signal MC generation