Likelihood-Free Inference

Imperial College London

Brandon Ellis-Frew and Jasturan Padda

Supervisor: Prof. Alan Heavens, Group: Astrophysics

Background, Motivation and Project Outline

Goals:

- Find the conditional (posterior) probability distribution of a set of model parameters, given some observed data and a forward simulator model
- Finding an appropriate compression scheme to reduce dimensionality of the problem
- Develop working algorithms and validate these using a specified analytic problem
- Apply these methods to higher dimensional, real case studies

Motivations:

- Modelling complex systems, from which we have access to observed data, is a general problem applicable to a wide variety of quantitative fields
- Many fields use forward simulator models which contain parameters that require estimation
- If the processes are too complex to calculate analytic likelihood functions, we must use Likelihood-Free Inference (LFI) techniques
- Examples include the estimation of cosmological parameters within cosmology, and even the Lotka-Volterra Equations for predator-prey modelling [1]

Bayes Theorem

Consider some model with parameters, θ_j and observed data, x_i. Using Bayes Theorem, we can write [2]:

$$p(\theta_j|x_i) = \frac{p(x_i|\theta_j)p(\theta_j)}{p(x_i)} \quad (1)$$

where we refer to $p(\theta_j|x_i)$ as the posterior, $p(x_i|\theta_j)$ as the likelihood, $p(\theta_j)$ as the prior and $p(x_i)$ as the evidence.

The posterior, by definition, is the conditional probability distribution given the observed data, x_i. This term is therefore what we require since it is the updated distribution from taking the new information into account alongside the prior. This is illustrated in equation 1; it is the product of the likelihood term and the prior. We can apply this understanding to Likelihood-Free Inference techniques where, in complex scenarios, the likelihood may be unknown or intractable. This involves computing the posterior, without knowledge of the likelihood, from comparisons between simulated data and observed data.

Approximate Bayesian Computation (ABC)

- ABC is the most traditional approach to the inference problem; it involves drawing model parameters from the prior distribution, then forward simulating data with which we compare to the observed data.
- If the simulated data falls within a specified threshold of observed data, which is evaluated using a distance metric, the selected parameters are accepted or otherwise rejected.
- This enables the approximation of the posterior distribution of θ from the accepted parameter values [3]. This is illustrated in the algorithm flowchart in Fig.1.

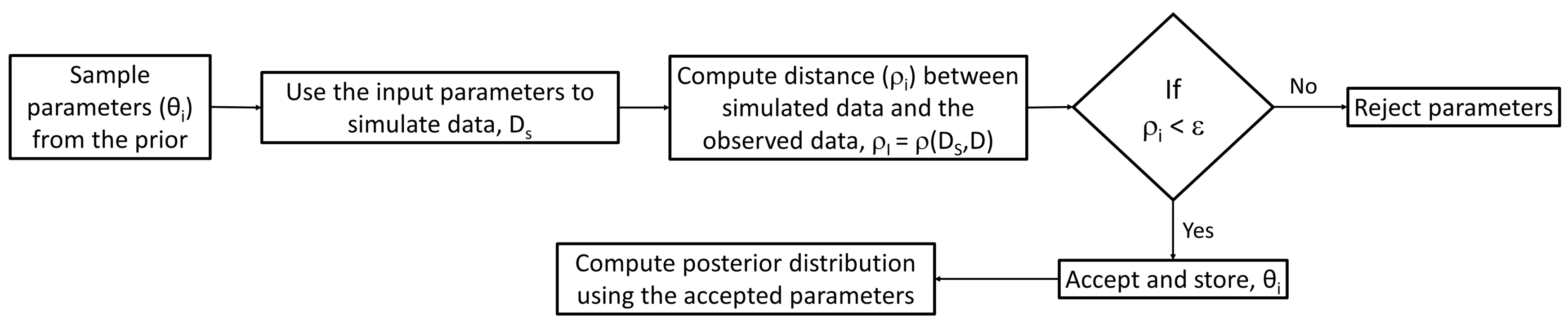

Fig. 1. A flowchart demonstrating the ABC rejection sampling algorithm

Density Estimation Likelihood Free Inference (DELFI)

- DELFI is a parametric method where we utilise Neural Density Estimators (NDEs) [4]
- NDEs tackle distribution estimation problems using various statistical methods alongside numerous network architectures [4]
- The specific method we are utilising models the conditional density, $p(d|\theta)$, and obtains the likelihood by evaluating at the observed data [4]
- The posterior can be computed by taking the product between the likelihood and the prior, $p(\theta|d) \propto p(d|\theta) \times p(\theta)$ [4]
- This allows for freedom in attaining simulations, therefore we can draw parameters from the most relevant regions in parameter space without incurring problems [4]
- This is especially effective when simulations are computationally expensive, as all simulation results are used in the training and validation process of the neural network [4]

References

[1] Weisstein, Eric W. "Lotka-Volterra Equations." From MathWorld--A Wolfram Web Resource. https://mathworld.wolfram.com/Lotka-VolterraEquations.html

[2] Dr Mark Richards (2019) Lecture 1: Introduction and properties of probabilities, Mathematics and Statistics of Measurement, Imperial College London.

[3] Leclercq, F. (2018) Bayesian optimisation for likelihood-free cosmological inference. American Physical Society. 98 (6).

[4] Justin Alsing, Tom Charnock, Stephen Feeney, Benjamin Wandelt, Fast likelihood-free cosmology with neural density estimators and active learning, https://arxiv.org/pdf/1903.00007.pdf

[5] Justin Alsing, Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology.

[6] Tom Charnock, Guilhem Lavaux, Benjamin D. Wandelt, Automatic physical inference with information maximising neural networks, https://arxiv.org/pdf/1802.03537.pdf

Density Estimation Likelihood Free Inference (cont.)

Mixture Density Networks (MDNs)

- MDNs rely on the fact that a conditional distribution can be represented via the sum over a suitable number of Gaussian components with appropriate coefficients **$\{r_k(\theta;w)\}$** ,

$$p(\boldsymbol{t}|\boldsymbol{\theta};\boldsymbol{w}) = \sum_{k=1}^{n_c} r_k(\boldsymbol{\theta};\boldsymbol{w})\mathcal{N}[\boldsymbol{t}|\boldsymbol{\mu}_k(\boldsymbol{\theta};\boldsymbol{w}), \boldsymbol{C}_k \equiv \boldsymbol{\Sigma}_k(\boldsymbol{\theta};\boldsymbol{w})\boldsymbol{\Sigma}_k^T(\boldsymbol{\theta};\boldsymbol{w})] \quad (2)$$

where **$\{\mu_k(\theta;w)\}$** are means, and **$\{\Sigma_k(\theta;w)\}$** covariances [4].

Masked Autoregressive Flows (MAFs)

- MAFs obtain the approximate posterior distribution through a series of invertible, transformations carried out by Masked Autoencoders for Distribution Estimations (MADEs)
- With this approach the posterior distribution is decomposed into a product of conditionals using the probability chain rule [4]

$$p(\boldsymbol{t}|\boldsymbol{\theta};\boldsymbol{w}) = \prod_{i=1}^{\dim(\boldsymbol{t})} p(t_i|\boldsymbol{t}_{1:i-1},\boldsymbol{\theta};\boldsymbol{w}) = \mathcal{N}[\mathbf{u}(\mathbf{t},\boldsymbol{\theta};\boldsymbol{w})|\mathbf{0},\mathbf{I}] \times \prod_{n=1}^{N_{MADEs}} \prod_{i=1}^{\dim(t)} \frac{1}{\sigma_i^n(\boldsymbol{t},\boldsymbol{\theta};\boldsymbol{w})} \quad (3)$$

Neural Network Loss Function

- Neural Networks are trained using an objective function; this is something that needs minimizing/maximizing with respect to the, w. For both MDNs and MAFs we take this to be the negative log-likelihood, [4]

$$-lnU(\boldsymbol{w}|\{\boldsymbol{\theta},\boldsymbol{t}\}) = -\sum_{i=1}^{N_{samples}} \ln p(\boldsymbol{t}_i|\boldsymbol{\theta}_i;\boldsymbol{w}) \quad (4)$$

Data Compression

- For data with high dimensionality, it may be inefficient or not possible to perform computations regarding LFI
- Data compression, therefore, becomes an essential problem whereby we must compress a high dimensional data set to a set of summary statistics of lower dimensionality
- Compressing data must be done carefully such that we are retaining the maximum amount of information
- Approximate-Score Compression is a compression method used to compress a set of N data to one compressed summary statistic per parameter of interest

$$\boldsymbol{t} = \nabla_{\boldsymbol{\theta}}\mathcal{L}_* \quad (5)$$

where L_* is the natural logarithm of a quasi-likelihood function evaluated at some chosen fiducial point in parameter space [5]

Results and Discussion

- The initial analytical model for testing is drawing 100 points from a gaussian distributed dataset with additional gaussian noise
- Here we demonstrate the success of both ABC and DELFI, whilst using sufficient summary statistics.
- For both methods we see a match between the experimental PDFs generated and the analytical forms we derived.
- Summary statistics used here were the mean and variances of the data samples.

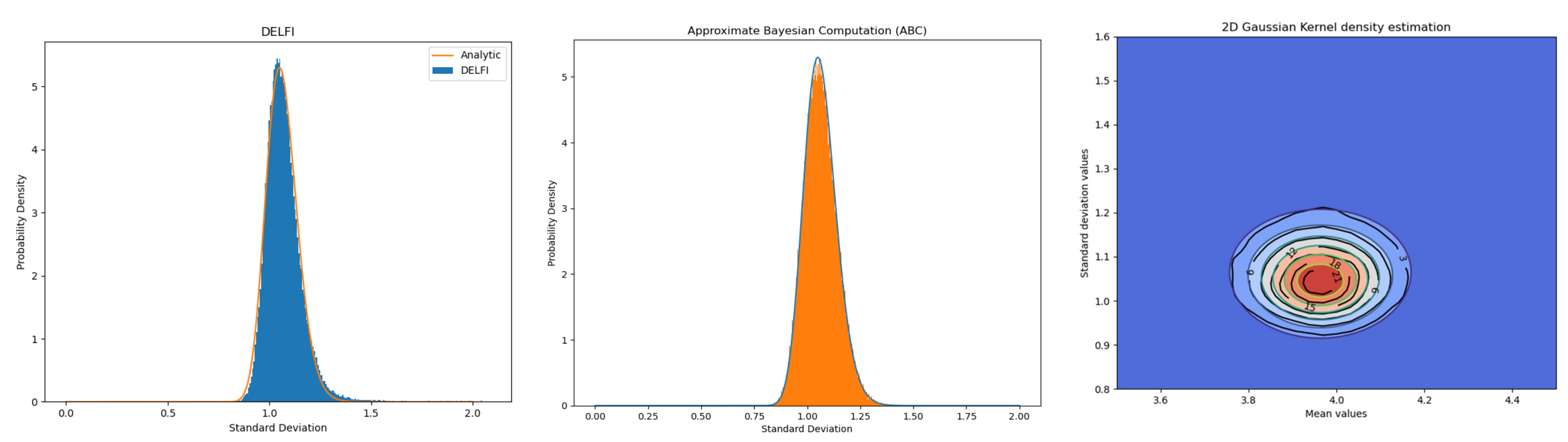

Fig. 2. (left- DELFI) marginalised PDF for standard deviation, (middle-ABC) marginalised PDF for standard deviation, (right-DELFI) joint PDF for mean and standard deviation.

Conclusion and Future Work

Conclusion

- We have explored both non-parametric (ABC) and parametric models (DELFI), whilst utilising compressed summary statistics to make the inference task easier to carry out.
- Throughout we have multiple successes in finding the desired PDFs and parameter estimation in general. Thus, highlighting the effectiveness the techniques employed.

Future for the field

- Increasing the speed at which both ABC and DELFI can be carried out.
- Applying these techniques to new types of problems.
- Further research into data compression schemes, such as recent developments in Information Maximising Neural Networks (IMNN) [6].