Bayesian Inference and Experimental Design for Implicit Models

Michael Gutmann

michael.gutmann@ed.ac.uk

Institute for Adaptive and Neural Computation
School of Informatics, University of Edinburgh

19 April 2019

Program

Implicit models

Learning the parameters (likelihood-free inference)

Performing experimental design

Program

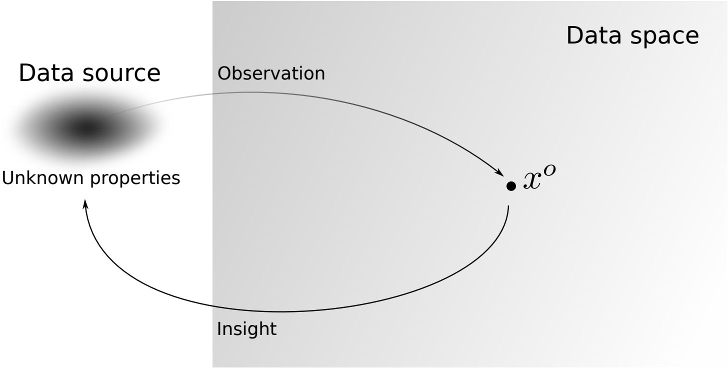
Implicit models

Learning the parameters (likelihood-free inference)

Performing experimental design

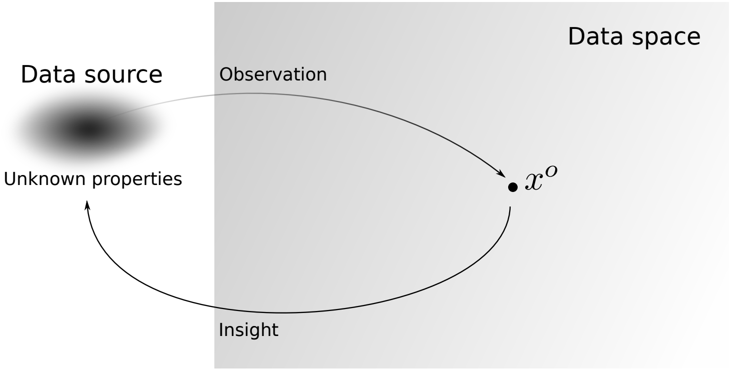
Overall goal

- Goal: Understand properties of a data source of interest
- Enables predictions, decision making under uncertainty, ...



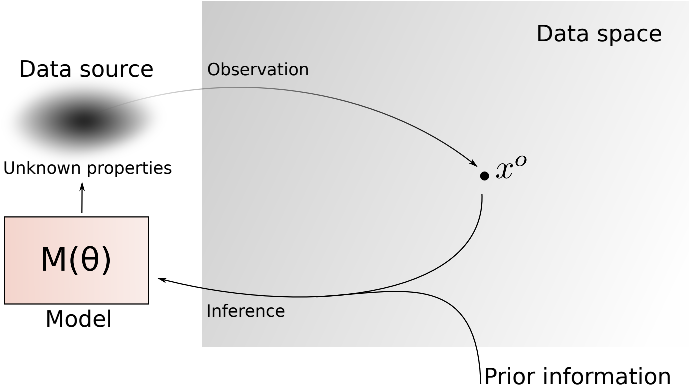
Two fundamental tasks

- Inference task : Given $\boldsymbol{x}^o$, what can we robustly say about the properties of the source?
- Experimental design task : How to obtain a $\boldsymbol{x}^o$ that is maximally useful for learning about the properties?



Using models to learn from data

- Set up a model with properties that the unknown data source might have.
- The potential properties are the parameters $\boldsymbol{\theta}$ of the model.



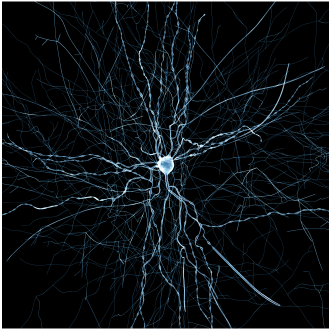
Implicit models

(Diggle and Gratton, JRSS, 1982)

- Models specified by a data generating mechanism
  - e.g. stochastic nonlinear dynamical systems
  - e.g. computer models / simulators of some complex physical or biological process
- Only assumption: sampling – simulating data – from the model is possible
- No closed form expression for probability density functions $p(\boldsymbol{x}|\boldsymbol{\theta})$.
- Different communities use different names:
  - Simulator-based models
  - Stochastic simulation models
  - Implicit models
  - Generative (latent-variable) models
  - Probabilistic programs

Implicit models are widely used

- Evolutionary biology:
  Simulating evolution
- Neuroscience:
  Simulating neural circuits
- Health science:
  Simulating the spread of an infectious disease
- Computer vision:
  Simulating naturalistic scenes
- Robotics:
  Simulating the outcome of an action
- ...



Simulated neural activity in rat somatosensory cortex
(Figure from `https://bbp.epfl.ch/nmc-portal`)

Strengths of implicit models

- Direct implementation of hypotheses of how the observed data were generated.
- Neat interface with scientific models (e.g. from physics or biology).
- Modelling by replicating the mechanisms of nature that produced the observed/measured data. ("Analysis by synthesis")
- Possibility to perform experiments in silico.

Weaknesses of implicit models

- Generally elude analytical treatment.
- Hard to assess identifiability.
- Principled inference and experimental design is difficult.

Weaknesses of implicit models

- Generally elude analytical treatment.
- Hard to assess identifiability.
- Principled inference and experimental design is difficult.

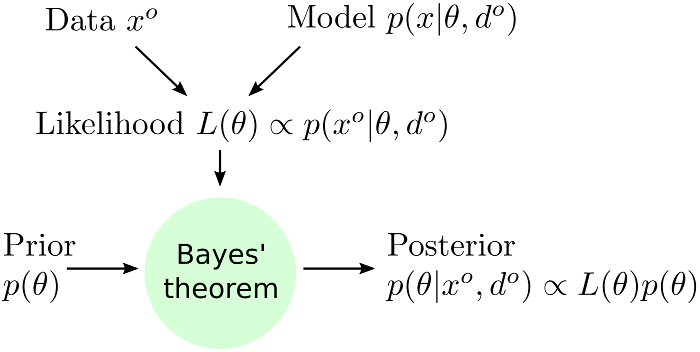
Main reason: *Likelihood function is too expensive to evaluate*

This talk considers two tasks

1. Learning the parameters of implicit models
2. Performing experimental design for implicit models

Bayesian approach to learning

- Learning $\equiv$ probabilistic inference
- Assume data $\boldsymbol{x^o}$ has been collected in an experiment with setup (design) $\boldsymbol{d^o}$.



Bayesian approach to experimental design

- Experimental design $\equiv$ utility optimisation problem
- Utility depends on the goal (parameter estimation, model comparison, prediction)
- For parameter estimation:
  maximise expected information gain (change of our belief) when an experiment with design $\boldsymbol{d}$ is performed

$$U(\boldsymbol{d}) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{d}} \left[\mathrm{KL} \left(p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{d}) \;||\; p(\boldsymbol{\theta}) \right) \right] \tag{1}$$

- Same as maximising mutual information between $\boldsymbol{x}$ and $\boldsymbol{\theta}$
- Functional of the posterior (and hence the likelihood function)

Principled but computationally hard for implicit models

- Difficulty essentially due to high-dimensional integrals
- One reason for the integrals: unobserved variables $\boldsymbol{z}$ which makes the likelihood function intractable

$$L(\boldsymbol{\theta}) \propto p(\boldsymbol{x}^o \,|\, \boldsymbol{\theta}, \mathbf{d}) \tag{2}$$

$$\propto \int p(\boldsymbol{x}^o, \boldsymbol{z} \,|\, \boldsymbol{\theta}, \mathbf{d}) \mathrm{d}\boldsymbol{z} \tag{3}$$

- Makes both Bayesian inference and experimental design computationally very difficult

Program

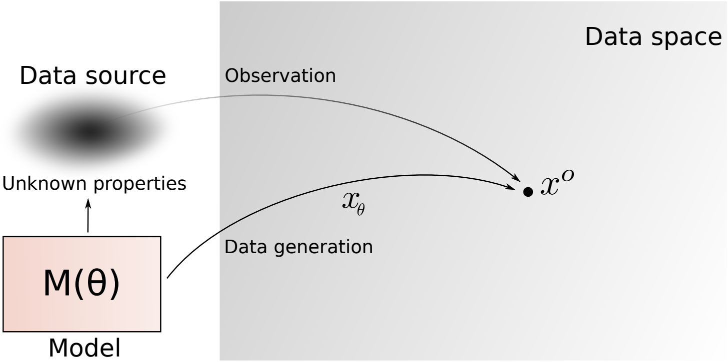
Implicit models

Learning the parameters (likelihood-free inference)

Performing experimental design

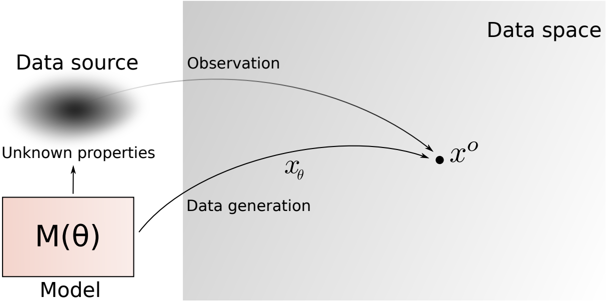
The likelihood function $L(\boldsymbol{\theta})$

- Probability that the model generates data like $\boldsymbol{x}^o$ when using parameter value $\boldsymbol{\theta}$
- Well defined but generally intractable for implicit models



Three foundational issues in likelihood-free inference (LFI)

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
2. How should we compute the probability of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
3. For which values of $\boldsymbol{\theta}$ should we compute it?



Likelihood: Probability that the model generates data like $\boldsymbol{x}^o$ for parameter value $\boldsymbol{\theta}$

LFI via synthetic likelihood

(Simon Wood, Nature, 2010)

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   - $\Rightarrow$ Compute summary statistics $\boldsymbol{t}_{\boldsymbol{\theta}} = \psi(\boldsymbol{x}_{\boldsymbol{\theta}})$
   - $\Rightarrow$ Model their distribution as a Gaussian with mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$.
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   - $\Rightarrow$ Compute likelihood function with $\psi(\boldsymbol{x}^o)$ as observed data
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   - $\Rightarrow$ Use obtained "synthetic" likelihood function as part of a Monte Carlo method

LFI via synthetic likelihood

(Simon Wood, Nature, 2010)

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   - $\Rightarrow$ Compute summary statistics $\boldsymbol{t}_{\boldsymbol{\theta}} = \psi(\boldsymbol{x}_{\boldsymbol{\theta}})$
   - $\Rightarrow$ Model their distribution as a Gaussian with mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$.
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   - $\Rightarrow$ Compute likelihood function with $\psi(\boldsymbol{x}^o)$ as observed data
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   - $\Rightarrow$ Use obtained "synthetic" likelihood function as part of a Monte Carlo method

Difficulties:

- Choice of ψ
- Gaussianity assumption may not hold
- Typically high computational cost

LFI via approximate Bayesian computation

1. How should we assess whether $\boldsymbol{x_\theta} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ Check whether $||\psi(\boldsymbol{x_\theta}) - \psi(\boldsymbol{x}^o)|| \leq \epsilon$
2. How should we compute the proba of the event $\boldsymbol{x_\theta} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ By counting
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   $\Rightarrow$ Sample from the prior (or other proposal distributions)

LFI via approximate Bayesian computation

1. How should we assess whether $\mathbf{x}_{\boldsymbol{\theta}} \equiv \mathbf{x}^o$?
   $\Rightarrow$ Check whether $||\psi(\mathbf{x}_{\boldsymbol{\theta}}) - \psi(\mathbf{x}^o)|| \leq \epsilon$
2. How should we compute the proba of the event $\mathbf{x}_{\boldsymbol{\theta}} \equiv \mathbf{x}^o$?
   $\Rightarrow$ By counting
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   $\Rightarrow$ Sample from the prior (or other proposal distributions)

Difficulties:

- Choice of $\psi()$ and ϵ
- Typically high computational cost

Recent review: Lintusaari et al (2017) "Fundamentals and recent developments in approximate Bayesian computation", Systematic Biology

Overview of some of my work

1. How should we assess whether $\boldsymbol{x_\theta} \equiv \boldsymbol{x}^o$?
   - $\Rightarrow$ Use classification (Gutmann et al, 2014, 2018)

2. How should we compute the proba of the event $\boldsymbol{x_\theta} \equiv \boldsymbol{x}^o$?
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   - $\Rightarrow$ Use Bayesian optimisation (Gutmann and Corander, 2013, 2016)
   - $\Rightarrow$ Decision making under uncertainty (Järvenpää, 2018a, 2018b)

   Compared to standard approaches: speed-up by a factor of 1000 or more

1. How should we assess whether $\boldsymbol{x_\theta} \equiv \boldsymbol{x}^o$?
2. How should we compute the proba of the event $\boldsymbol{x_\theta} \equiv \boldsymbol{x}^o$?
   - $\Rightarrow$ Use density ratio estimation (Thomas et al, 2016, Dinev and Gutmann, 2018)
   - $\Rightarrow$ Combine strengths of two classical approaches: regression ABC and sequential ABC (Chen and Gutmann, AISTATS, 2019)

Overview of some of my work

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   - $\Rightarrow$ Use classification (Gutmann et al, 2014, 2018)

2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   - $\Rightarrow$ Use Bayesian optimisation / Decision making under uncertainty (Gutmann and Corander, 2013, 2016; Järvenpää, 2018a, 2018b)
     Compared to standard approaches: speed-up by a factor of 1000 more

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   - $\Rightarrow$ Use density ratio estimation (Thomas et al, 2016, Dinev and Gutmann, 2018)
   - $\Rightarrow$ Combine strengths of two classical approaches: regression ABC and sequential ABC (Chen and Gutmann, AISTATS, 2019)

Basic idea

(Thomas et al, 2016, arXiv:1611.10242)

- Frame posterior estimation as ratio estimation problem

$$\log p(\boldsymbol{\theta}|\mathbf{x}) = \log \left[\frac{p(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})}{p(\mathbf{x})} \right] = \log p(\boldsymbol{\theta}) + h(\mathbf{x}, \boldsymbol{\theta}) \tag{4}$$

$$h(\mathbf{x}, \boldsymbol{\theta}) = \log \left[\frac{p(\mathbf{x}|\boldsymbol{\theta})}{p(\mathbf{x})} \right] \tag{5}$$

- Estimating $h(\mathbf{x}, \boldsymbol{\theta})$ is the difficult part since $p(\mathbf{x}|\boldsymbol{\theta})$ unknown.
- Estimate $\hat{h}(\mathbf{x}, \boldsymbol{\theta})$ yields estimate of the likelihood function and posterior

$$\hat{L}(\boldsymbol{\theta}) \propto \exp \left[\hat{h}(\mathbf{x}^o, \boldsymbol{\theta}) \right] \qquad \hat{p}(\boldsymbol{\theta}|\mathbf{x}^o) = p(\boldsymbol{\theta}) \exp \left[\hat{h}(\mathbf{x}^o, \boldsymbol{\theta}) \right] \tag{6}$$

- We call this approach LFIRE: Likelihood-Free Inference by Ratio Estimation

Estimating density ratios

- For implicit models, we do not know

$$p(\mathbf{x}|\boldsymbol{\theta}) \qquad\qquad p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{7}$$

  but we can draw samples from them.

- There are several methods available to estimate the log-ratio $h(\mathbf{x}, \boldsymbol{\theta})$ from the samples

$$\mathbf{x}_i^{\theta} \sim p(\mathbf{x}|\boldsymbol{\theta}) \qquad\qquad i = 1, \ldots, n_{\boldsymbol{\theta}} \tag{8}$$

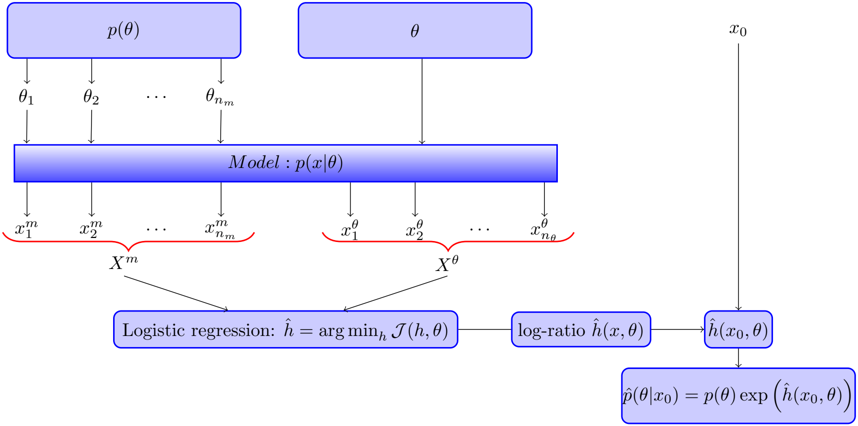
$$\mathbf{x}_i^{m} \sim p(\mathbf{x}) \qquad\qquad i = 1, \ldots, n_m \tag{9}$$

  (see e.g. textbook by Sugiyama et al, 2012)

- Bregman divergence provides general framework
  (Gutmann and Hirayama, 2011; Sugiyama et al, 2011)

- Here: density ratio estimation by logistic regression details

Estimating the posterior by LFIRE



(Thomas et al, 2016, arXiv:1611.10242)

Solving other inference tasks by ratio estimation

- Ratio estimation was used to estimate unnormalised models
  (Gutmann & Hyvärinen, 2010, 2012)
- Related to classification approach to judge whether $\boldsymbol{x}_\theta \equiv \boldsymbol{x}^o$
  (Gutmann et al, 2014, 2018)
- Can be used to train generative adversarial networks
  (see e.g. review by Mohamed and Lakshminarayanan, 2017)
- It was used to estimate likelihood ratios
  (Pham et al, 2014; Cranmer et al, 2015)

Auxiliary model

- We need to specify a model for the log-ratio h.
- For simplicity: linear model

$$h(\boldsymbol{x}) = \sum_{i=1}^{b} \beta_i \psi_i(\boldsymbol{x}) = \boldsymbol{\beta}^\top \psi(\boldsymbol{x}) \tag{10}$$

  where $\psi_i(\boldsymbol{x})$ are summary statistics (feature extractors)
- L_1 penalty on β for weighing and selecting summary statistics
- Features can be learned: e.g. convolutional neural networks for time series (Dinev and Gutmann, 2018, arXiv:1810.09899)

Key properties

1. Already the linear model generalises the synthetic likelihood approach.
2. Supports learning/selection of summary statistics
3. "Amortised inference": Learned model of the ratio can be re-used for different observed data sets $\boldsymbol{x}_k^o$ without new computations:

$$\hat{p}(\boldsymbol{\theta}|\boldsymbol{x}_k^o) = p(\boldsymbol{\theta}) \exp\left[\hat{h}(\boldsymbol{x}_k^o, \boldsymbol{\theta})\right] \tag{11}$$

Example: application to ARCH model

- Model:

$$x^{(t)} = \theta_1 x^{(t-1)} + e^{(t)} \tag{12}$$

$$e^{(t)} = \xi^{(t)} \sqrt{0.2 + \theta_2 (e^{(t-1)})^2} \tag{13}$$

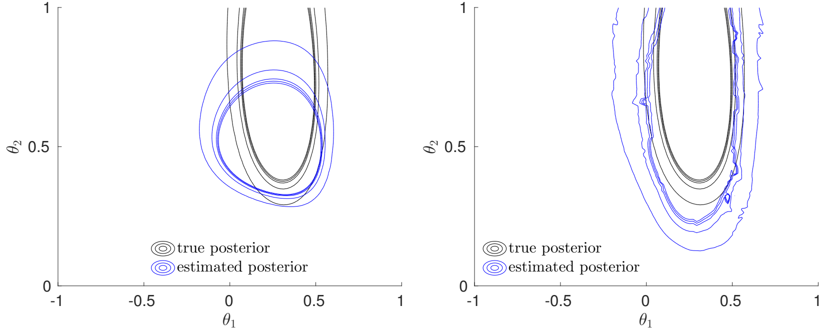
  $\xi^{(t)}$ and $e^{(0)}$ independent standard normal r.v., $x^{(0)} = 0$
- 100 time points
- Parameters: $\theta_1 \in (-1, 1), \quad \theta_2 \in (0, 1)$
- Uniform prior on θ_1, θ_2

Example: application to ARCH model

- Summary statistics:
  - auto-correlations with lag one to five
  - all (unique) pairwise combinations of them
  - a constant
- To check robustness: 50% irrelevant summary statistics (drawn from standard normal)
- Comparison with synthetic likelihood with equivalent set of summary statistics (relevant sum. stats. only)

Example: generalisation of the synthetic likelihood

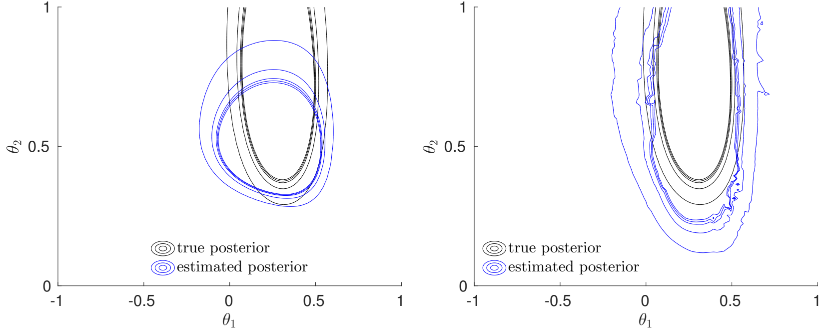
(Thomas et al, 2016, arXiv:1611.10242)



(a) synthetic likelihood (b) proposed method

Example: selection of summary statistics

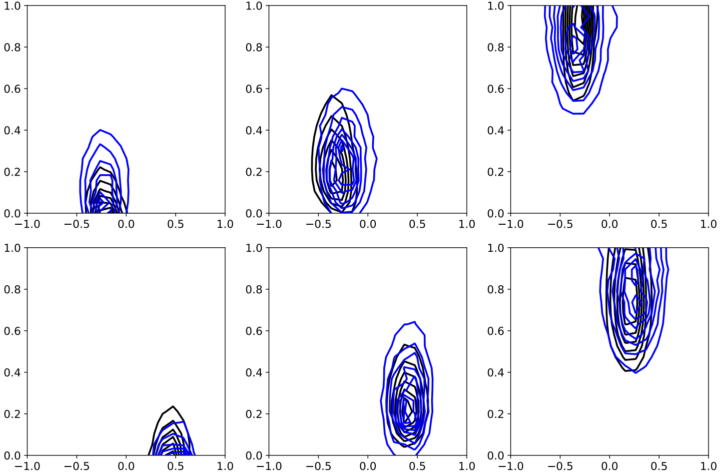
(Thomas et al, 2016, arXiv:1611.10242)



(c) synthetic likelihood (d) proposed method subject to noise

Example: "amortised inference"

$$\hat{p}(\boldsymbol{\theta}|\boldsymbol{x}_k^o) = p(\boldsymbol{\theta}) \exp\left[\hat{h}(\boldsymbol{x}_k^o, \boldsymbol{\theta})\right]$$



(Results with learned summary statistics, Dinev and Gutmann, 2018,

Program

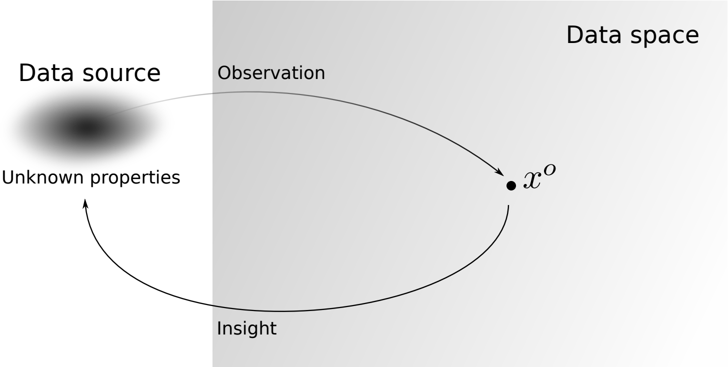
Implicit models

Learning the parameters (likelihood-free inference)

Performing experimental design

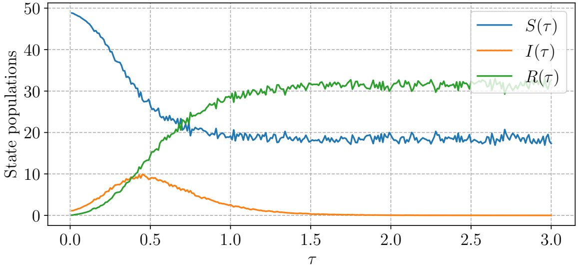
Two fundamental tasks

- Inference task: Given $\boldsymbol{x}^o$, what can we robustly say about the properties of the source?
- Experimental design task : How to obtain a $\boldsymbol{x}^o$ that is maximally useful for learning about the properties?



Example: stochastic SIR model

- Stochastic epidemiological model describing the population of susceptibles $S(\tau)$, infected $I(\tau)$ and recovered $R(\tau)$ as a function of time.
- Parameters $\boldsymbol{\theta}$: rate of infection β and the rate of recovery γ.
- Exp design task : find the optimal times at which to perform the measurements to most accurately estimate β and γ.



(Figure by Steven Kleinegesse)

Experimental design by mutual information maximisation

- Utility to maximise

$$U(\boldsymbol{d}) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{d}} \left[\mathrm{KL} \left(p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{d}) \;||\; p(\boldsymbol{\theta}) \right) \right] \tag{14}$$

  $\boldsymbol{d}$: for example a sequence of measurement times, $\boldsymbol{d} = (\tau_1, \ldots, \tau_n)$.
- Pro: Does not make a Gaussianity or unimodality assumption of the posterior
- Con: Two major difficulties:
  1. Hard to compute
  2. Hard to maximise (since typically no gradient or closed-form expression available)

Difficulty 1—approximate the mutual information

Two steps:

1. Approximate the expectation with a sample average

$$
\begin{aligned}
U(\boldsymbol{d}) &= \mathbb{E}_{\boldsymbol{x}|\boldsymbol{d}}\left[\mathrm{KL}\left(p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{d}) \,||\, p(\boldsymbol{\theta})\right)\right] && (15)\\
&= \int p(\boldsymbol{x}|\boldsymbol{d}) \int p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{d}) \log \frac{p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{d})}{p(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{x} && (16)\\
&= \int p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{d}) \log \frac{p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{d})}{p(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{x} && (17)\\
&\approx \frac{1}{N} \sum_{i=1}^{N} \log \left[\frac{p(\boldsymbol{\theta}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{d})}{p(\boldsymbol{\theta}^{(i)})}\right], && (18)
\end{aligned}
$$

   where $\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta})$ and $\boldsymbol{x}^{(i)} \sim p(\boldsymbol{x}|\boldsymbol{d}, \boldsymbol{\theta}^{(i)})$.

2. Estimate $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{d})$ using LFIRE

Difficulty 1—make use of LFIRE

- From LFIRE (for each fixed design $\boldsymbol{d}$)

$$\hat{p}(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{d}) = p(\boldsymbol{\theta}) \exp\left[\hat{h}_{\boldsymbol{d}}(\boldsymbol{x}, \boldsymbol{\theta})\right] \tag{19}$$

- Hence:

$$\log \frac{p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{d})}{p(\boldsymbol{\theta})} \approx \hat{h}_{\boldsymbol{d}}(\boldsymbol{x}, \boldsymbol{\theta}) \tag{20}$$

and

$$\hat{U}(\boldsymbol{d}) = \frac{1}{N} \sum_{i=1}^{N} \hat{h}_{\boldsymbol{d}}(\boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{(i)}) \tag{21}$$

$$\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta}) \quad \boldsymbol{x}^{(i)} \sim p(\boldsymbol{x}|\boldsymbol{d}, \boldsymbol{\theta}^{(i)})$$

- Benefit of amortisation property of LFIRE: one run of LFIRE required for each $\boldsymbol{d}$.

Difficulty 2—use BO to maximise the utility

- We can approximate the utility pointwise for each $\boldsymbol{d}$

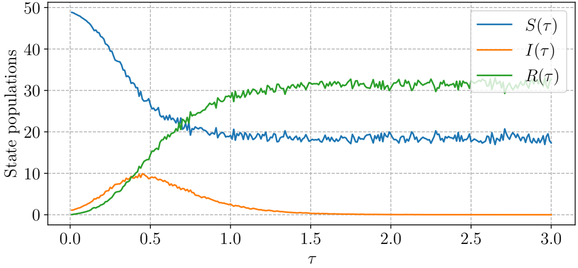
$$\hat{U}(\boldsymbol{d}) = \frac{1}{N} \sum_{i=1}^{N} \hat{h}_{\boldsymbol{d}}(\boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{(i)}) \tag{22}$$

  Deals with first difficulty.

- Second technical difficulty: How to maximise $\hat{U}(\boldsymbol{d})$?
  - Computing $\hat{U}(\boldsymbol{d})$ is relatively costly and no gradient information is available.
  - $\hat{U}(\boldsymbol{d})$ is noisy due to approximation
- Use Bayesian optimisation (BO) to determine $\mathrm{argmax}_{\boldsymbol{d}}\, \hat{U}(\boldsymbol{d})$
  - Builds a surrogate model of $\hat{U}(\boldsymbol{d})$ smoothing out noise introduced by the sample average approximation.
  - Trade-off between exploration (improve the model) and exploitation (find optimum according to the model)

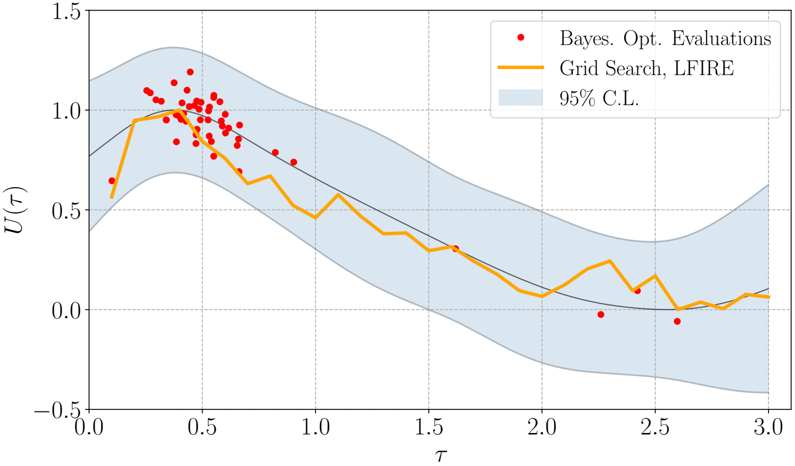
Example: stochastic SIR model

- $S(\tau)$: susceptibles; $I(\tau)$: infected: $R(\tau)$: recovered
- Parameters $\boldsymbol{\theta}$: rate of infection β and the rate of recovery γ.
- Exp design task : find the optimal times at which to perform the measurements to most accurately estimate β and γ.



(Figure by Steven Kleinegesse)

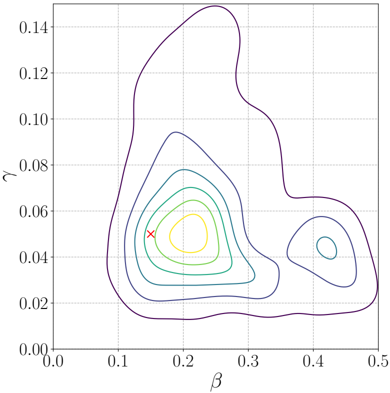
Results: one measurement



- Optimal measurement time: $\tau^* = 0.365$
- Convergence after ~ 10 evaluations

(Kleinegesse and Gutmann, AISTATS 2019)

Results: one measurement



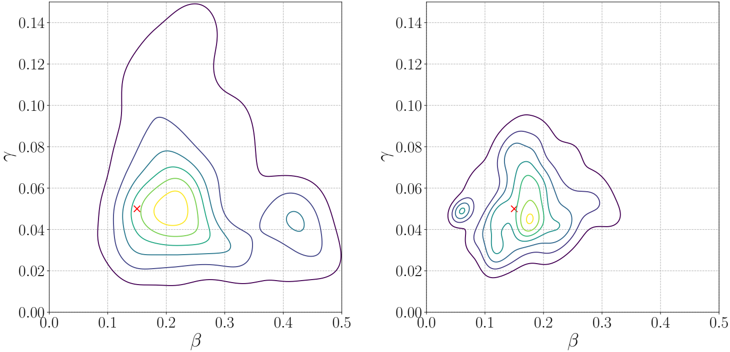
- Prior: Uniform distribution on $[0, 0.5]$ for both parameters
- One observation already provides reasonable information
- Estimation of recovery rate γ better than infection rate β for one observation

(Kleinegesse and Gutmann, AISTATS 2019)

Results: multiple measurements

- Design of multiple measurements:

$$\boldsymbol{d} = [\tau_1, \tau_2, \dots, \tau_8]^\top \text{ with } \tau_1 < \dots < \tau_8$$



- Convergence after ~ 15 evaluations for 8 dimensions

(Kleinegesse and Gutmann, AISTATS 2019)

Conclusions

- Three topics:
  1. Implicit models
  2. Inference for implicit models—likelihood-free inference (LFI)
  3. Experimental design for implicit models by mutual information maximisation
- Likelihood-free inference by ratio estimation (LFIRE)
- LFIRE to estimate both posteriors and the mutual information
- Bayesian optimisation to maximise the mutual information
- "Methods talk" with simple examples but:
  - We have applied the LFI methods in multiple domains in collaboration with domain experts (e.g. genetics, epidemiology of infectious diseases, robotics)
  - First steps towards more challenging experimental design applications.

Density ratio estimation by logistic regression

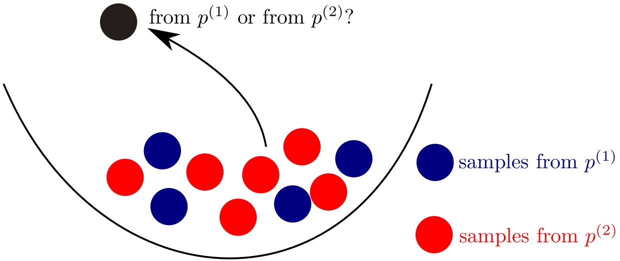
- Samples from two data sets

$$\mathbf{x}_i^{(1)} \sim p^{(1)}, \quad i = 1, \ldots, n^{(1)} \tag{23}$$

$$\mathbf{x}_i^{(2)} \sim p^{(2)}, \quad i = 1, \ldots, n^{(2)} \tag{24}$$

- Probability that a test data point $\mathbf{x}$ was sampled from $p^{(1)}$

$$\mathbb{P}(\mathbf{x} \sim p^{(1)} | \mathbf{x}, h) = \frac{1}{1 + \nu \exp(-h(\mathbf{x}))}, \qquad \nu = \frac{n^{(2)}}{n^{(1)}} \tag{25}$$



Density ratio estimation by logistic regression

- Estimate h by minimising

$$\mathcal{J}(h) = \frac{1}{n}\left\{\sum_{i=1}^{n^{(1)}} \log\left[1 + \nu \exp\left(-h_i^{(1)}\right)\right] + \sum_{i=1}^{n^{(2)}} \log\left[1 + \frac{1}{\nu}\exp\left(h_i^{(2)}\right)\right]\right\}$$

$$h_i^{(1)} = h\left(\boldsymbol{x}_i^{(1)}\right) \qquad h_i^{(2)} = h\left(\boldsymbol{x}_i^{(2)}\right)$$

$$n = n^{(1)} + n^{(2)}$$

- Objective is the re-scaled negated log-likelihood.
- For large $n^{(1)}$ and $n^{(2)}$

$$\hat{h} = \mathrm{argmin}_h \, \mathcal{J}(h) = \log \frac{p^{(1)}}{p^{(2)}}$$

without any constraints on h

back