Model-based likelihood free inference: BOLFI and LFIRE

Based on slides by Michael Gutmann

Owen Thomas

Department of Biostatistics, University of Oslo

13th March 2019

Task

Perform Bayesian inference for models where

1. the likelihood function is too costly to compute
2. sampling – simulating data – from the model is possible

Program

Program

Background

Previous work

BOLFI

LFIRE

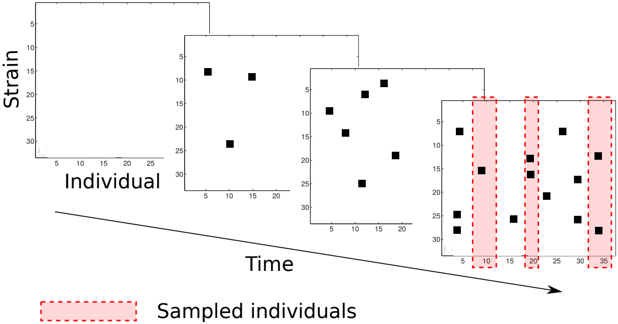
Simulator-based models

- Goal: Inference for models that are specified by a mechanism for generating data
  - e.g. stochastic dynamical systems
  - e.g. computer models / simulators of some complex physical or biological process
- Such models occur in multiple and diverse scientific fields.
- Different communities use different names:
  - Simulator-based models
  - Stochastic simulation models
  - Implicit models
  - Generative (latent-variable) models
  - Probabilistic programs

Examples

Simulator-based models are widely used:

- Evolutionary biology: Simulating evolution
- Neuroscience: Simulating neural circuits
- Ecology: Simulating species migration
- Health science: Simulating the spread of an infectious disease

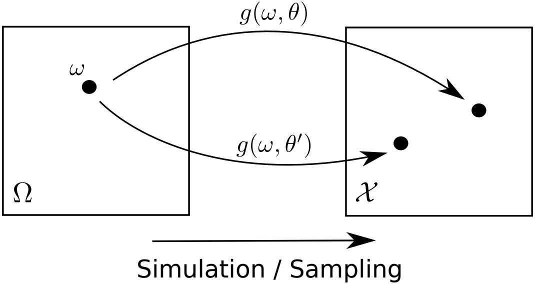


Definition of simulator-based models

- Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space.
- A simulator-based model is a collection of (measurable) functions $g(., \boldsymbol{\theta})$ parametrised by $\boldsymbol{\theta}$,

$$\boldsymbol{\omega} \in \Omega \mapsto \boldsymbol{x_\theta} = g(\boldsymbol{\omega}, \boldsymbol{\theta}) \in \mathcal{X} \tag{1}$$

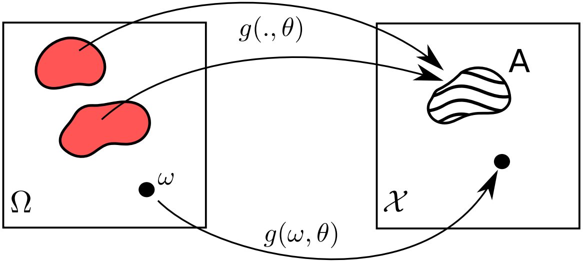
- For any fixed $\boldsymbol{\theta}$, $\boldsymbol{x_\theta} = g(., \boldsymbol{\theta})$ is a random variable.



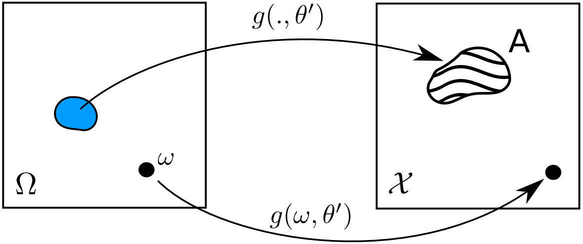
Implicit definition of the model pdfs

$$\Pr(x \in A \mid \theta) = \mathcal{P}(\{\omega : g(\omega, \theta) \in A\})$$

Parameter value θ



Parameter value θ'



Advantages of simulator-based 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.

Disadvantages of simulator-based models

- Generally elude analytical treatment.
- Can be easily made more complicated than necessary.
- Statistical inference is difficult.

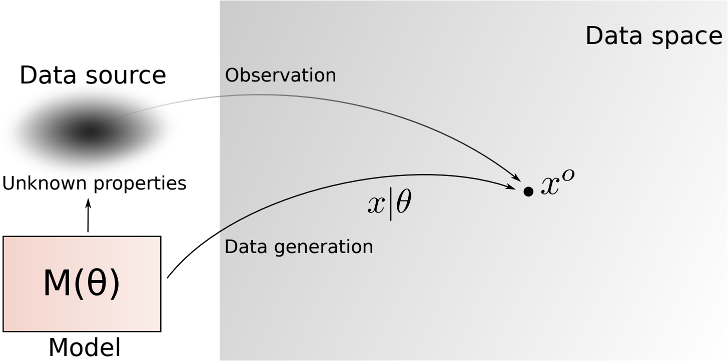
Disadvantages of simulator-based models

- Generally elude analytical treatment.
- Can be easily made more complicated than necessary.
- Statistical inference is difficult.

  Main reason: *Likelihood function is intractable*

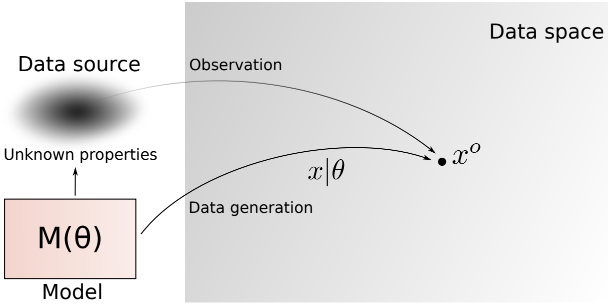
The likelihood function $L(\boldsymbol{\theta})$

- Probability that the model generates data like $\boldsymbol{x}^o$ when using parameter value $\boldsymbol{\theta}$
- Generally well defined but intractable for simulator-based / implicit models



Three foundational issues

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 $\boldsymbol{\theta}$

Program

Background

Previous work

BOLFI

LFIRE

Approximate Bayesian computation

1. How should we assess whether $\boldsymbol{x_\theta} \equiv \boldsymbol{x}^o$?
   - $\Rightarrow$ Check whether $d(\theta, \boldsymbol{x}^o) = ||T(\boldsymbol{x_\theta}) - T(\boldsymbol{x}^o)|| \leq \epsilon$
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)

Approximate Bayesian computation

1. How should we assess whether $\boldsymbol{x_\theta} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ Check whether $d(\theta, \boldsymbol{x}^o) = ||T(\boldsymbol{x_\theta}) - T(\boldsymbol{x}^o)|| \leq \epsilon$
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)

Difficulties:

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

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

Synthetic likelihood

(Simon Wood, Nature, 2010)

1. How should we assess whether $\boldsymbol{x_\theta} \equiv \boldsymbol{x}^o$?
2. How should we compute the proba of the event $\boldsymbol{x_\theta} \equiv \boldsymbol{x}^o$?
   - $\Rightarrow$ Compute summary statistics $\boldsymbol{t_\theta} = T(\boldsymbol{x_\theta})$
   - $\Rightarrow$ Model their distribution as a Gaussian
   - $\Rightarrow$ Compute likelihood function with $T(\boldsymbol{x}^o)$ 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

Synthetic likelihood

(Simon Wood, Nature, 2010)

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   - $\Rightarrow$ Compute summary statistics $\boldsymbol{t}_{\boldsymbol{\theta}} = T(\boldsymbol{x}_{\boldsymbol{\theta}})$
   - $\Rightarrow$ Model their distribution as a Gaussian
   - $\Rightarrow$ Compute likelihood function with $T(\boldsymbol{x}^o)$ 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 $T()$
- Gaussianity assumption may not hold
- Typically high computational cost

Overview of some related work

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?

   $\Rightarrow$ Use classification (Gutmann et al, 2014, 2017)

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 (Gutmann and Corander, 2013-2016)
   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, arXiv:1611.10242)

Overview of some related work

1. How should we assess whether $\boldsymbol{x_\theta} \equiv \boldsymbol{x}^o$?
   - $\Rightarrow$ Use classification (Gutmann et al, 2014, 2017)

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 BOLFI (Gutmann and Corander, 2013-2016)

     Compared to standard approaches: speed-up by a factor of 1000 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 LFIRE (Thomas et al, 2016, arXiv:1611.10242)

Program

Background

Previous work

BOLFI

LFIRE

BOLFI: Creating Proxies for the Discrepancy

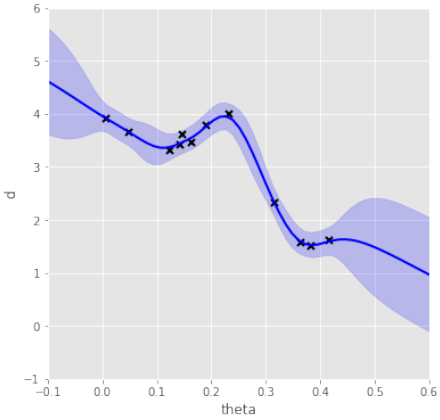
(Gutmann and Corander, 2013-2016)

- It is possible to recast likelihood free inference as characterising the discrepancy surface $d(\theta)$ as a function of θ.
- Observations of the discrepancy for a given θ can be considered an evaluation of a nonnegative response function d given covariates θ.
- It is possible to build a proxy model for the discrepancy using the well established tools of nonlinear regression.
- We can fully separate the problems of acquisition for the proxy model, and sampling from the proxy model.

- A flexible Bayesian nonparametric prior over functions is the Gaussian Process (GP), a stochastic process in which all realisations of the discrepancy $d(\theta)$ are assumed to be jointly normally distributed:

$$p(d|\theta) \sim \mathcal{N}(m(\theta), K(\theta, \theta'))$$

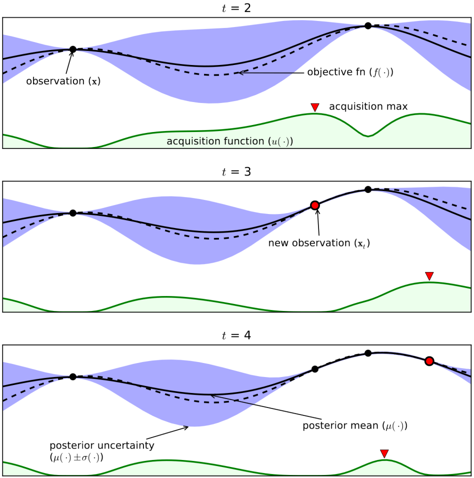
(Rasmussen and Williams, 2006)



Bayesian Optimisation I

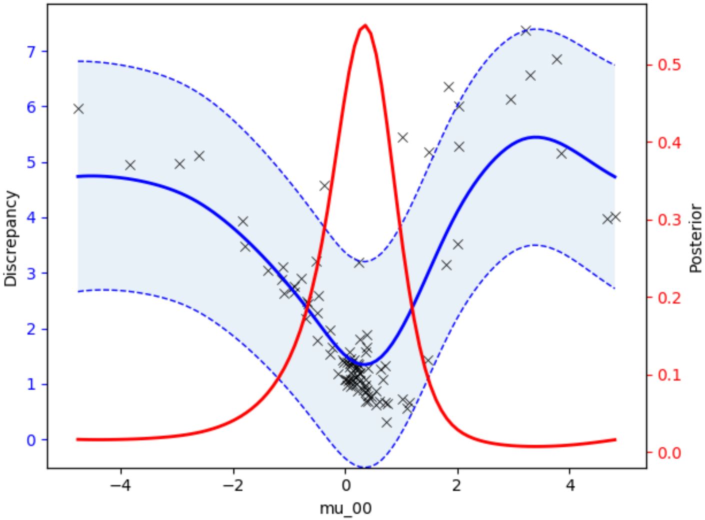
- GPs have two important properties that enable Bayesian Optimisation
  - Good posterior uncertainty characterisation
  - Flexible, non-parametric structure
- Values of θ are deterministically queried to maximise their information about the optimum of the function.
- There is no analytical solution for the optimum of a GP, but there exist various heuristics that can be useful in different circumstances.
- We define an acquisition function using the posterior mean and variance to determine where to evaluate.
  - Upper Confidence Bound
  - Expected Improvement

Bayesian Optimisation II



(https://towardsdatascience.com/shallow-understanding-on-bayesian-optimization-324b6c1f7083)

BOLFI example



Posterior Sampling

- Once we have a model for the discrepancy function $d(\theta)$, we can use it to build proxies for the likelihood $p(X|\theta)$.
- This can be achieved through Kernel Density Estimation (KDE).
- For a uniform kernel, the likelihood proxy is proportional to the probability of the discrepancy being below a given threshold, i.e.:

$$\tilde{\pi}(X|\theta) \propto p(d(\theta) < h) \tag{2}$$

- Given the likelihood proxy $\tilde{L}(\theta)$ and the prior $p(\theta)$, we can use an algorithm of our choice to sample the posterior.
- In practice, Hamiltonian Monte Carlo (HMC) is often used.

Results

- BOLFI performs extremely well in situations where simulations are expensive and acquisitions must be done intelligently.
- Smart acquisitions results in much fewer calls to the simulator, sometimes of $\mathcal{O}(10^3)$. (Gutmann and Corander, 2013-2016)
- Convergence properties are secured by the theory behind Bayesian Optimization.
- A natural extension of contemporary probabilistic numerics techniques to the likelihood free paradigm.

Program

Background

Previous work

BOLFI

LFIRE

Basic idea

(Thomas et al, 2016, arXiv:1611.10242)

- Frame posterior estimation as ratio estimation problem

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x})} = p(\boldsymbol{\theta})r(\boldsymbol{x},\boldsymbol{\theta}) \tag{3}$$

$$r(\boldsymbol{x},\boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x})} \tag{4}$$

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

$$\hat{L}(\boldsymbol{\theta}) \propto \hat{r}(\boldsymbol{x}^o,\boldsymbol{\theta}), \qquad \hat{p}(\boldsymbol{\theta}|\boldsymbol{x}^o) = p(\boldsymbol{\theta})\hat{r}(\boldsymbol{x}^o,\boldsymbol{\theta}). \tag{5}$$

Estimating density ratios in general

- Relatively well studied problem (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

Density ratio estimation by logistic regression

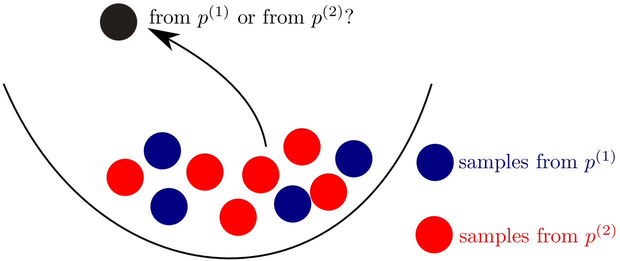
- Samples from two data sets

$$\mathbf{x}_i^{(1)} \sim p^{(1)}, \quad i = 1, \ldots, n^{(1)} \tag{6}$$

$$\mathbf{x}_i^{(2)} \sim p^{(2)}, \quad i = 1, \ldots, n^{(2)} \tag{7}$$

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

$$\mathbb{P}(\mathbf{x} \sim p^{(1)} | \mathbf{x}, h) = \frac{1}{1 + \nu \exp(-h(\mathbf{x}))}, \qquad \nu = \frac{n^{(2)}}{n^{(1)}} \tag{8}$$



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

Estimating the posterior

- Property was used to estimate unnormalised models (Gutmann & Hyvärinen, 2010, 2012)
- It was used to estimate likelihood ratios (Pham et al, 2014; Cranmer et al, 2015)
- For posterior estimation, we use
  - data generating pdf $p(\boldsymbol{x}|\boldsymbol{\theta})$ for $p^{(1)}$
  - marginal $p(\boldsymbol{x})$ for $p^{(2)}$ (Other choices for $p(\boldsymbol{x})$ possible too)
  - sample sizes entirely under our control

Estimating the posterior

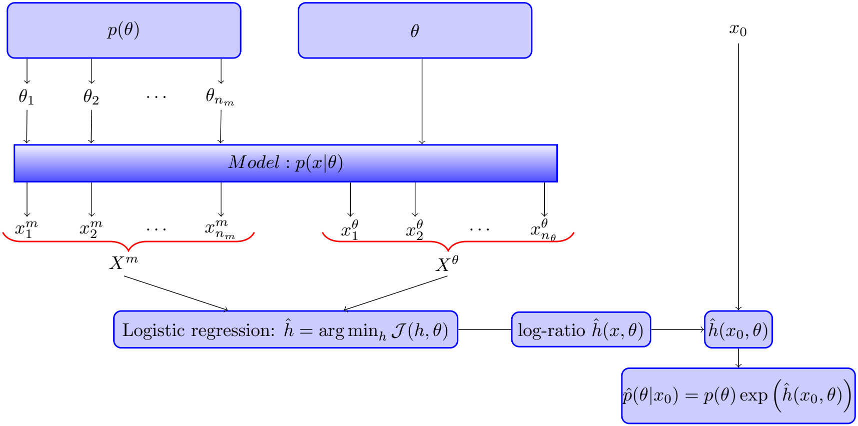
- Logistic regression (point-wise in $\boldsymbol{\theta}$)

$$\hat{h}(\boldsymbol{x}, \boldsymbol{\theta}) \rightarrow \log \frac{p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x})} = \log r(\boldsymbol{x}, \boldsymbol{\theta}) \tag{9}$$

- Estimated posterior and likelihood function:

$$\hat{p}(\boldsymbol{\theta}|\boldsymbol{x}^o) = p(\boldsymbol{\theta}) \exp(\hat{h}(\boldsymbol{x}^o, \boldsymbol{\theta})) \quad \hat{L}(\boldsymbol{\theta}) \propto \exp(\hat{h}(\boldsymbol{x}^o, \boldsymbol{\theta})) \tag{10}$$

Estimating the posterior



(Thomas et al, 2016, arXiv:1611.10242)

Auxiliary model

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

$$h(\boldsymbol{x}) = \sum_{i=1}^{b} \beta_i \psi_i(\boldsymbol{x}) = \beta^\top \psi(\boldsymbol{x}) \tag{11}$$

  where $\psi_i(\boldsymbol{x})$ are summary statistics
- More complex models possible

Exponential family approximation

- Logistic regression yields

$$\hat{h}(\boldsymbol{x}; \boldsymbol{\theta}) = \hat{\beta}(\boldsymbol{\theta})^\top \psi(\boldsymbol{x}), \quad \hat{r}(\boldsymbol{x}, \boldsymbol{\theta}) = \exp(\hat{\beta}(\boldsymbol{\theta})^\top \psi(\boldsymbol{x})) \tag{12}$$

- Resulting posterior

$$\hat{p}(\boldsymbol{\theta}|\boldsymbol{x}^o) = p(\boldsymbol{\theta}) \exp(\hat{\beta}(\boldsymbol{\theta})^\top \psi(\boldsymbol{x}^o)) \tag{13}$$

- Implicit exponential family approximation of $p(\boldsymbol{x}|\boldsymbol{\theta})$

$$\hat{r}(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\hat{p}(\boldsymbol{x}|\boldsymbol{\theta})}{\hat{p}(\boldsymbol{x})} \tag{14}$$

$$\hat{p}(\boldsymbol{x}|\boldsymbol{\theta}) = \hat{p}(\boldsymbol{x}) \exp(\hat{\beta}(\boldsymbol{\theta})^\top \psi(\boldsymbol{x})) \tag{15}$$

- Implicit because $\hat{p}(\boldsymbol{x})$ never explicitly constructed.

Remarks

- Vector of summary statistics $\psi(\boldsymbol{x})$ should include a constant for normalisation of the pdf (log partition function)
- Normalising constant is estimated via the logistic regression
- Simple linear model leads to a generalisation of synthetic likelihood
- L_1 penalty on β for weighing and selecting summary statistics

Application to ARCH model

- Model:

$$x^{(t)} = \theta_1 x^{(t-1)} + e^{(t)} \tag{16}$$

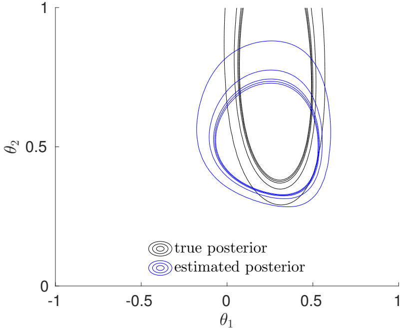
$$e^{(t)} = \xi^{(t)} \sqrt{0.2 + \theta_2 (e^{(t-1)})^2} \tag{17}$$

  $\xi^{(t)}$ 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

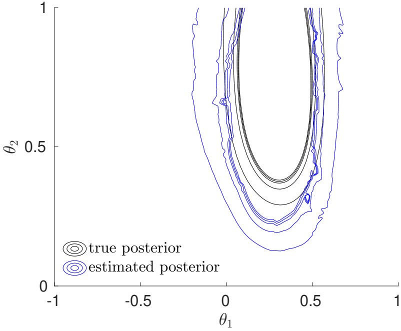
Application to ARCH model

- 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 posterior

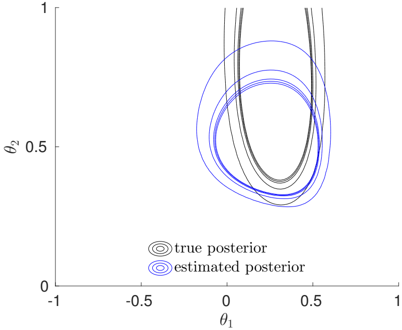


(a) synthetic likelihood

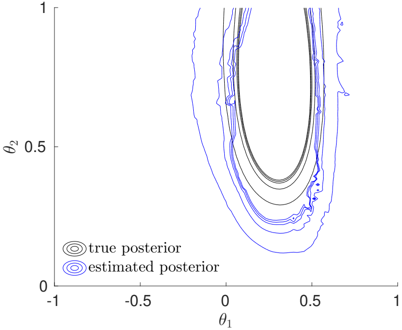


(b) proposed method

Example posterior



(c) synthetic likelihood



(d) proposed method subject to noise

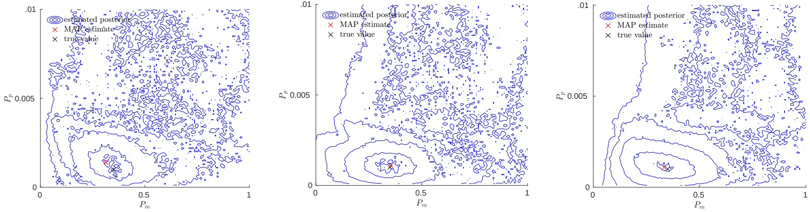
Observations

- Compared two auxiliary models: exponential vs Gaussian family
- For same summary statistics , typically more accurate inferences for the richer exponential family model
- Robustness to irrelevant summary statistics thanks to L_1 regularisation

Application to cell proliferation model

- A stochastic lattice model for generating instances of cell proliferation data.
- Two parameters P_m, P_p describing the dynamic properties.
- The summary statistics are the Hamming distances between each of the 145 time instances of the cell lattice and their squares, giving 291 in total, including a constant.
- Very high-dimensional summary statistic space.

Application to cell proliferation model



(e) $n_\theta = n_m = 50$ (f) $n_\theta = n_m = 100$ (g) $n_\theta = n_m = 150$

Figure: Cell spreading model: Contour plots of the LFIRE proliferation likelihood for the parameters P_m and P_p for the cell spreading model. Each panel corresponds to different numbers of simulated data points $n_\theta = n_m$ to train the classifier. The true values and MAP estimates of the parameters are also displayed in the plots.

Observations

- LFIRE with regularisation successfully characterises a posterior in the presence of high numbers of summary statistics.
- L_1 regularisation discretely selects greater numbers of relevant summary statistics as number of simulated data increases.
- The $n_\theta = n_m = 50$, 100 and 150 simulations select an average of 17.3, 23.9 and 30.5, respectively.

Conclusions

- Background and previous work on inference with simulator-based / implicit statistical models
- Our work on:
  - Framing the posterior estimation problem as a density ratio estimation problem
  - Estimating the ratio with logistic regression
  - Using regularisation to automatically select summary statistics
- Multitude of research possibilities:
  - Choice of the auxiliary model
  - Choice of the loss function used to estimate the density ratio
  - Combine with Bayesian optimisation framework to reduce computational cost

Conclusions

- Background and previous work on inference with simulator-based / implicit statistical models
- Our work on:
  - Framing the posterior estimation problem as a density ratio estimation problem
  - Estimating the ratio with logistic regression
  - Using regularisation to automatically select summary statistics
- Multitude of research possibilities:
  - Choice of the auxiliary model
  - Choice of the loss function used to estimate the density ratio
  - Combine with Bayesian optimisation framework to reduce computational cost

More results and details in arXiv:1611.10242v1