Partially Exchangeable Networks and Architectures for Learning Summary Statistics in Approximate Bayesian Computation

ICML 2019

Samuel Wiqvist

Centre for Mathematical Sciences, Lund University, Sweden

samuel_wiqvist

June 13, 2019

- Joint work with Pierre-Alexandre Mattei (IT University Copenhagen), Umberto Picchini (Chalmers/University of Gothenburg), and Jes Frellsen (IT University Copenhagen)

ABC: Simulation-based inference

- ABC only requires that we can simulate data from our model $p(y|\theta)$, thus ABC is very generic, and can be applied for models where the likelihood is intractable;
- ABC in a nut-shell:
  1. Generate parameter proposals $\theta^\star$ from the prior $p(\theta)$;
  2. Accept $\theta^\star$ if the generated data $y^\star \sim p(y|\theta^\star)$ is *similar* to our observed data y^{obs};
  3. Repeat Step 1-2 for a large number of times;
  4. The accepted θ's are samples from an approximation to the posterior $p(\theta|y^{\text{obs}})$.
- *Curse-of-dimensionality*: Instead of comparing $y^\star$ with y^{obs} we compare a set of summary statistics $S(y^\star)$ and $S(y^{\text{obs}})$;
- The main focus of our work is how to automatically learn summary statistics $S(\cdot)$ that are informative for θ.

ABC: Simulation-based inference

- ABC only requires that we can simulate data from our model $p(y|\theta)$, thus ABC is very generic, and can be applied for models where the likelihood is intractable;
- ABC in a nut-shell:
  1. Generate parameter proposals $\theta^\star$ from the prior $p(\theta)$;
  2. Accept $\theta^\star$ if the generated data $y^\star \sim p(y|\theta^\star)$ is *similar* to our observed data y^{obs};
  3. Repeat Step 1-2 for a large number of times;
  4. The accepted θ's are samples from an approximation to the posterior $p(\theta|y^{\text{obs}})$.
- *Curse-of-dimensionality*: Instead of comparing $y^\star$ with y^{obs} we compare a set of summary statistics $S(y^\star)$ and $S(y^{\text{obs}})$;
- The main focus of our work is how to automatically learn summary statistics $S(\cdot)$ that are informative for θ.

ABC: Simulation-based inference

- ABC only requires that we can simulate data from our model $p(y|\theta)$, thus ABC is very generic, and can be applied for models where the likelihood is intractable;
- ABC in a nut-shell:
  1. Generate parameter proposals $\theta^\star$ from the prior $p(\theta)$;
  2. Accept $\theta^\star$ if the generated data $y^\star \sim p(y|\theta^\star)$ is *similar* to our observed data y^{obs};
  3. Repeat Step 1-2 for a large number of times;
  4. The accepted θ's are samples from an approximation to the posterior $p(\theta|y^{\text{obs}})$.
- *Curse-of-dimensionality*: Instead of comparing $y^\star$ with y^{obs} we compare a set of summary statistics $S(y^\star)$ and $S(y^{\text{obs}})$;
- The main focus of our work is how to automatically learn summary statistics $S(\cdot)$ that are informative for θ.

ABC: Simulation-based inference

- ABC only requires that we can simulate data from our model $p(y|\theta)$, thus ABC is very generic, and can be applied for models where the likelihood is intractable;
- ABC in a nut-shell:
  1. Generate parameter proposals $\theta^\star$ from the prior $p(\theta)$;
  2. Accept $\theta^\star$ if the generated data $y^\star \sim p(y|\theta^\star)$ is *similar* to our observed data y^{obs};
  3. Repeat Step 1-2 for a large number of times;
  4. The accepted θ's are samples from an approximation to the posterior $p(\theta|y^{\text{obs}})$.
- *Curse-of-dimensionality*: Instead of comparing $y^\star$ with y^{obs} we compare a set of summary statistics $S(y^\star)$ and $S(y^{\text{obs}})$;
- The main focus of our work is how to automatically learn summary statistics $S(\cdot)$ that are informative for θ.

How to select/learn summary statistics

- The problem of selecting informative summary statistics is the main challenge when applying ABC in practice;
- Usually, summary statistics are ad-hoc and “handpicked” out of subject-domain expertise;
- In they show that the best summary statistics (in terms of quadratic loss for θ) is the posterior mean $E(\theta|y)$;
- Deep learning methods that learn the posterior mean as a summary statistic for ABC have already been considered.

How to select/learn summary statistics

- The problem of selecting informative summary statistics is the main challenge when applying ABC in practice;
- Usually, summary statistics are ad-hoc and "handpicked" out of subject-domain expertise;
- In they show that the best summary statistics (in terms of quadratic loss for θ) is the posterior mean $E(\theta|y)$;
- Deep learning methods that learn the posterior mean as a summary statistic for ABC have already been considered.

How to select/learn summary statistics

- The problem of selecting informative summary statistics is the main challenge when applying ABC in practice;
- Usually, summary statistics are ad-hoc and "handpicked" out of subject-domain expertise;
- In[1] they show that the best summary statistics (in terms of quadratic loss for θ) is the posterior mean $E(\theta|y)$;
- Deep learning methods that learn the posterior mean as a summary statistic for ABC have already been considered.

[1] Paul Fearnhead and Dennis Prangle. "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3 (2012), pp. 419–474.

How to select/learn summary statistics

- The problem of selecting informative summary statistics is the main challenge when applying ABC in practice;
- Usually, summary statistics are ad-hoc and "handpicked" out of subject-domain expertise;
- In[1] they show that the best summary statistics (in terms of quadratic loss for θ) is the posterior mean $E(\theta|y)$;
- Deep learning methods that learn the posterior mean as a summary statistic for ABC have already been considered[2].

[1] Paul Fearnhead and Dennis Prangle. "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3 (2012), pp. 419–474.

[2] Bai Jiang et al. "Learning summary statistic for approximate Bayesian computation via deep neural network". In: *Statistica Sinica* (2017), pp. 1595–1618.

Designing the PEN architecture

- We build on the earlier ideas and we want to target time series models;
- Thus, we construct a regression function $y \mapsto E(\theta|y)$ that is d-block-switch invariant, yielding following regression problem:

$$\theta^i = E(\theta|y^i) + \xi^i = \underbrace{\rho_{\beta_\rho}\left(y^i_{1:d}, \sum_{l=1}^{M-d} \phi_{\beta_\phi}(y^i_{l:l+d})\right)}_{\text{PEN}-d} + \xi^i.$$

- We have a universal approximation theorem for this architecture;
- DeepSets is a special case of PEN.

Designing the PEN architecture

- We build on the earlier ideas and we want to target time series models;
- Thus, we construct a regression function $y \mapsto E(\theta|y)$ that is d-block-switch invariant, yielding following regression problem:

$$\theta^i = E(\theta|y^i) + \xi^i = \underbrace{\rho_{\beta_\rho}\left(y^i_{1:d}, \sum_{l=1}^{M-d} \phi_{\beta_\phi}(y^i_{l:l+d})\right)}_{\text{PEN}-d} + \xi^i.$$

- We have a universal approximation theorem for this architecture;
- DeepSets is a special case of PEN.

Designing the PEN architecture

- We build on the earlier ideas and we want to target time series models;
- Thus, we construct a regression function $y \mapsto E(\theta|y)$ that is d-block-switch invariant, yielding following regression problem:

$$\theta^i = E(\theta|y^i) + \xi^i = \underbrace{\rho_{\beta_\rho}\left(y^i_{1:d}, \sum_{l=1}^{M-d} \phi_{\beta_\phi}(y^i_{l:l+d})\right)}_{\text{PEN}-d} + \xi^i.$$

- We have a universal approximation theorem for this architecture;
- DeepSets is a special case of PEN.

Designing the PEN architecture

- We build on the earlier ideas and we want to target time series models;
- Thus, we construct a regression function $y \mapsto E(\theta|y)$ that is d-block-switch invariant, yielding following regression problem:

$$\theta^i = E(\theta|y^i) + \xi^i = \underbrace{\rho_{\beta_\rho}\left(y^i_{1:d}, \sum_{l=1}^{M-d} \phi_{\beta_\phi}(y^i_{l:l+d})\right)}_{\text{PEN}-d} + \xi^i.$$

- We have a universal approximation theorem for this architecture;
- DeepSets[3] is a special case of PEN.

[3] Manzil Zaheer et al. "Deep sets". In: *Advances in Neural Information Processing Systems*. 2017, pp. 3391–3401.

AR(2) model

- An autoregressive time series model of order two (AR(2)) follows:

$$y_l = \theta_1 y_{l-1} + \theta_2 y_{l-2} + z_l, \qquad z_l \sim N(0, 1).$$

- The AR(2) model is a Markov model of order 2 and the requirement for PEN-d (d > 0) is therefore fulfilled;
- We use a PEN-2 network (and compare with several different other methods).

AR(2) model

- An autoregressive time series model of order two (AR(2)) follows:

$$y_l = \theta_1 y_{l-1} + \theta_2 y_{l-2} + z_l, \qquad z_l \sim N(0,1).$$

- The AR(2) model is a Markov model of order 2 and the requirement for PEN-d (d $>$ 0) is therefore fulfilled;
- We use a PEN-2 network (and compare with several different other methods).

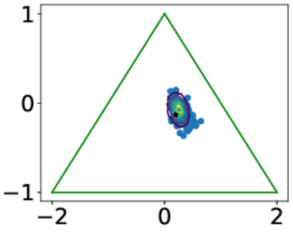
AR(2) model

- An autoregressive time series model of order two (AR(2)) follows:

$$y_l = \theta_1 y_{l-1} + \theta_2 y_{l-2} + z_l, \qquad z_l \sim N(0,1).$$

- The AR(2) model is a Markov model of order 2 and the requirement for PEN-d (d $>$ 0) is therefore fulfilled;
- We use a PEN-2 network (and compare with several different other methods).

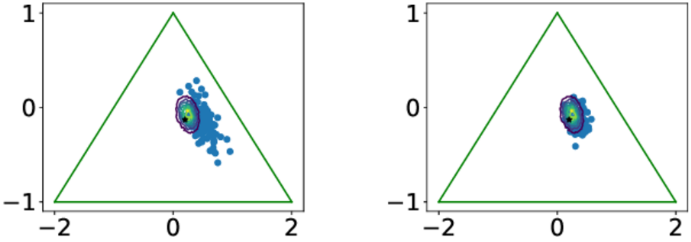
AR(2) model: Inference results with 10^6 training data points



(a) Handpicked (b) MLP (10^6) (c) PEN-2 (10^6)

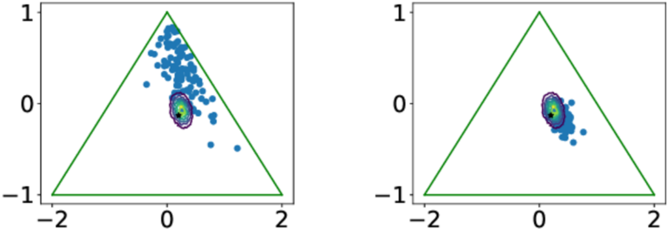
Figure: Green line: prior distribution; contour plot: exact posterior, the blue dots are 100 samples from the several ABC posteriors.

AR(2) model: Inference results with 10^5 training data points



(d) MLP (10^5) (e) PEN-2 (10^5)

Figure: Green line: prior distribution; contour plot: exact posterior, the blue dots are 100 samples from the several ABC posteriors.



(f) MLP (10^4) (g) PEN-2 (10^4)

Figure: Green line: prior distribution; contour plot: exact posterior, the blue dots are 100 samples from the several ABC posteriors.

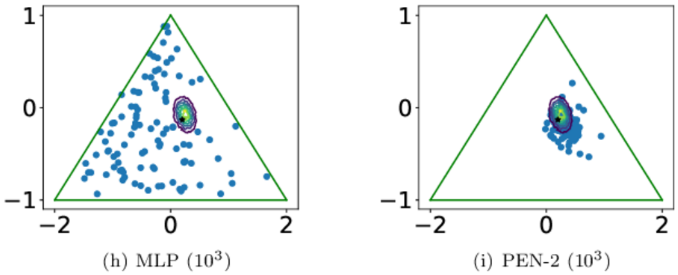


Figure: Green line: prior distribution; contour plot: exact posterior, the blue dots are 100 samples from the several ABC posteriors.

Conclusions

- PEN is more data efficient than the other methods;
- Does PEN work for time-series models that are not Markovian? Check out the paper/poster to find out!;
- Learning summary statistics for ABC is only one possible application for PEN.

Conclusions

- PEN is more data efficient than the other methods;
- Does PEN work for time-series models that are not Markovian? Check out the paper/poster to find out!;
- Learning summary statistics for ABC is only one possible application for PEN.

Conclusions

- PEN is more data efficient than the other methods;
- Does PEN work for time-series models that are not Markovian? Check out the paper/poster to find out!;
- Learning summary statistics for ABC is only one possible application for PEN.

Thank you for listening!

Find the paper at: tinyurl.com/pen-and-abc

Poster (today at 6:30PM): Pacific Ballroom #87