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

Pierre-Alexandre Mattei and Samuel Wiqvist
(joint work with U. Picchini & J. Frellsen)

Introduction

- Paper: Wiqvist, Mattei, Picchini & Frellsen, *Partially Exchangeable Networks and Architectures for Learning Summary Statistics in Approximate Bayesian Computation*, **ICML 2019**
- A deep learning method for learning summary statistics in ABC

Umberto Picchini
(Chalmers & U. of Gothenburg)

Jes Frellsen
(Technical University of Denmark)

Quick recap of likelihood-free inference

- Observed data set $y^{\text{obs}} \in y^M$ with M units.
- Associated Bayesian model:

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

- The likelihood $p(y|\theta)$ is intractable, but we assume that we can **simulate** data from the model.

Quick recap of ABC - rejection sampling

1. Sample proposal $\theta^\star \sim p(\theta)$
2. Generate data $y^\star \sim p(y|\theta^\star)$
3. Accept proposal if $S(y^\star) \approx S(y^{\mathrm{obs}})$

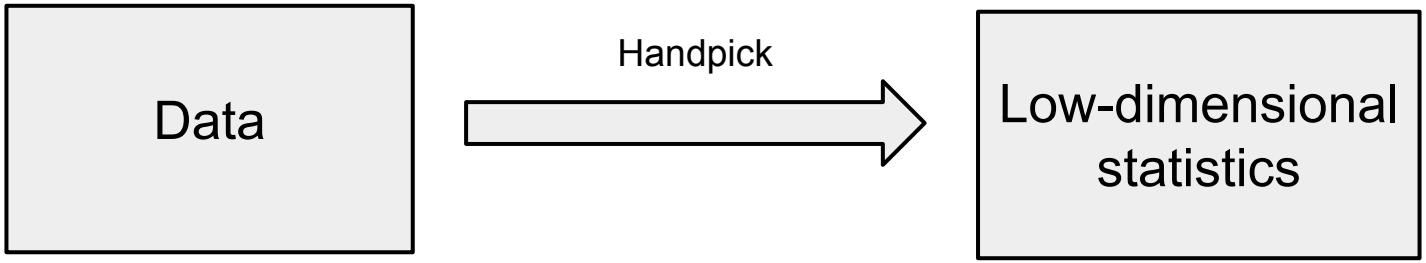
The function $S(\cdot)$ computes the summary statistics. **Main goal in our work: learning $S(\cdot)$ automatically and leveraging probabilistic symmetries in the data.**

The summary statistics function

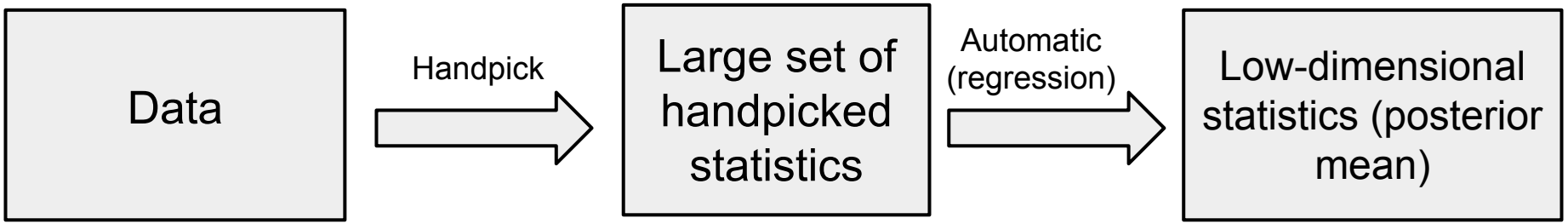
$$S : y^M \rightarrow s, \text{where}, \dim(s) < \dim(y^M).$$

- Summary statistics are usually necessary, due to curse-of-dimensionality.
- The summary statistics should be low-dimensional and informative for the parameters.
- Hand-picked summary statistics are often used (e.g. mean, quantiles, correlation, etc).
- Several methods for learning/selecting summary statistics have been developed (Handbook of ABC, chapter 5).

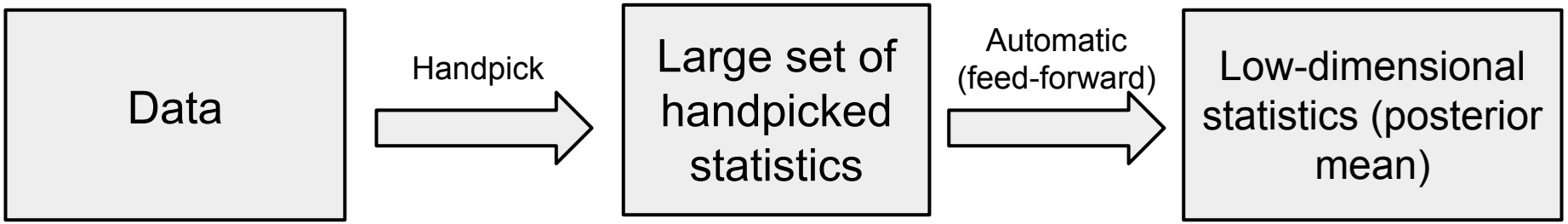
Handpicked summary statistics



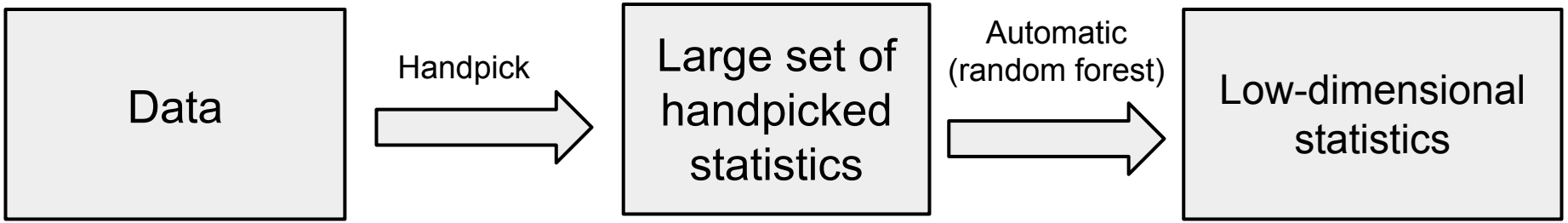
Semi-automatic (Fearnhead & Prangle (JRSSB 2012))



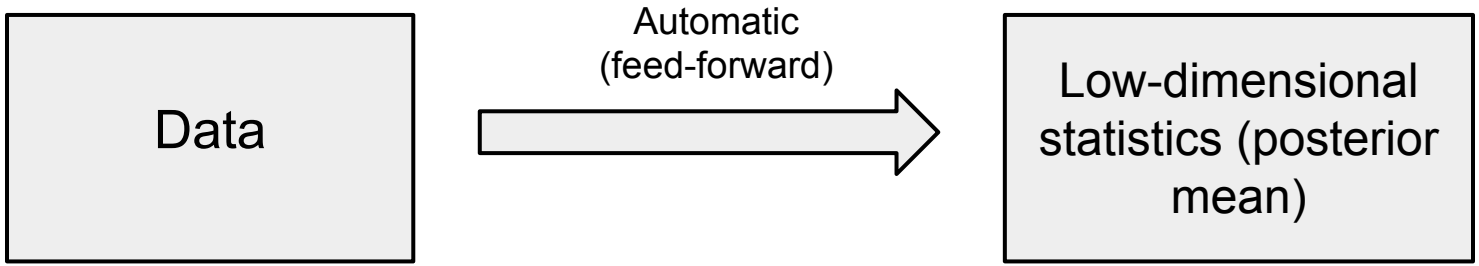
Network with pre-processing (Creel (Econ. Stat 2017))



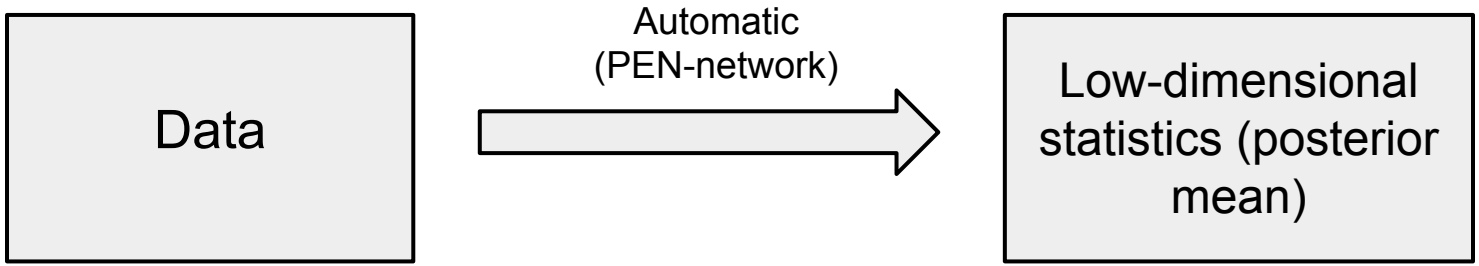
Random forest with pre-processing (Rayna et al. (Bioinformatics 2018))



Automatic network (Jiang et al. (Stat. Sin. 2017))



Automatic invariant network (Wiqvist et al. (ICML 2019))



Posterior mean as the summary statistics

- Fearnhead & Prangle (JRSSB 2012) showed that the posterior mean is the best summary statistics in terms of quadratic loss of the parameters.
- Thus, they let $S_\beta(y) \approx E(\theta|y)$ and they learn this function via simulations.
- The posterior mean is modelled with linear regression:

$$\theta_j = E(\theta_j|y) + \xi = \beta_{0j} + \beta_j h(y) + \xi \qquad (1)$$

Where j = 1:dim(θ), and $h(y)$ a vector of (non)-linear transformations.

Learning the posterior mean via simulations

1. Simulate parameter-data pairs $(\theta^i, y^i)_{i=1:N}$ from the prior predictive distribution
2. Fit the linear regression model in (1) to the parameter-data pairs, i.e:

$$\min_{\beta_j} \frac{1}{N} \sum_{i=1}^{N} ||S_{j,\beta_j}(y^i) - \theta_j^i||_2^2$$

Where $S_{j,\beta_j}(y^i)$ is the regression model in (1).

- *Once we have fitted the regression model we run ABC as usual, using $S_{j,\hat{\beta}_j}(y^\star)$ to compute the j:th summary statistic.*

Leveraging probabilistic symmetries

- Our goal is to find **a principled functional space for the summary function.**

- **Principled based on what?**

- We may not know the likelihood, but **we often know some symmetries of the model!**

From exchangeability to permutation invariance

- A first example of symmetry: **exchangeability**. A model with *M* units is exchangeable when the order of observations does not matter:

$$\forall \sigma \in S_M, p(y) = p(y_{\sigma(1)}, \cdots, y_{\sigma(M)})$$

- This is for example the case of a (conditionally) i.i.d. model. De Finetti's celebrated theorem says that the converse is also true for infinite data.
- If the data are exchangeable then the posterior is **permutation invariant:**

$$\forall \sigma \in S_M, p(\theta|y) = p(\theta|y_{\sigma(1)}, \cdots y_{\sigma(M)})$$

From exchangeability to permutation invariance

- Since the posterior is permutation invariant, it makes sense to look for **permutation invariant summary functions:**

$$\forall \sigma \in S_M, S(y) = S(y_{\sigma(1)}, \ldots, y_{\sigma(M)})$$

- Is it possible to design rich spaces of permutation invariant functions?

- **Yes, using permutation invariant neural networks!**

Permutation invariant neural networks

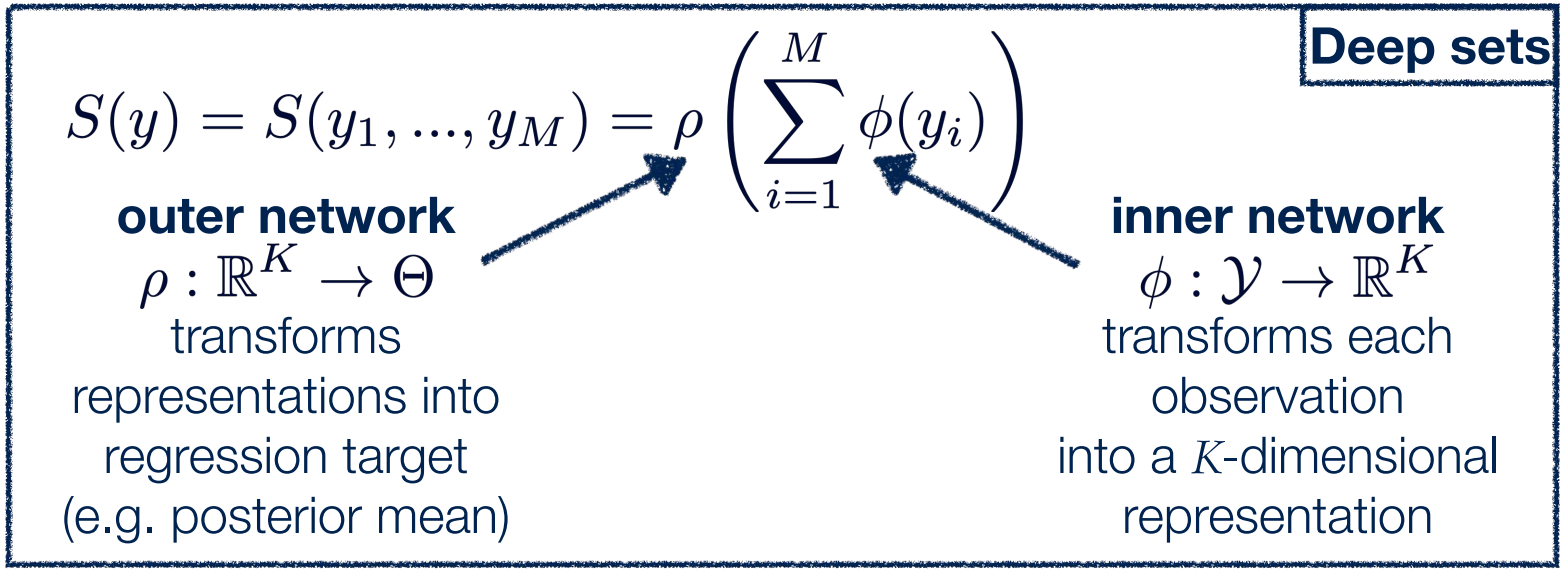
- Studied since the 1980s! (Shawe-Taylor, 1989)
- Renewed interest recently, notably with the **“Deep Sets”** architecture of Zaheer et al. (NeurIPS 2017), composed of **two neural networks.**
- A nice overview just came out in JLMR (Bloem-Reddy and Teh, 2020)

Deep Sets

Manzil Zaheer[1,2], **Satwik Kottur**[1], **Siamak Ravanbhakhsh**[1],
Barnabás Póczos[1], **Ruslan Salakhutdinov**[1], **Alexander J Smola**[1,2]
[1] Carnegie Mellon University [2] Amazon Web Services
{manzilz,skottur,mravanba,bapoczos,rsalakhu,smola}@cs.cmu.edu

Permutation invariant neural networks: Deep Sets

- The Deep Sets architecture **can approximate any permutation invariant function**, given enough capacity (Zaheer et al., 2017).
- It was used for likelihood-free inference by Chan et al. (NeurIPS 2018)



Beyond exchangeability

- **Problem:** likelihood-free models are often non-exchangeable!
  - State-space models
  - Stochastic differential equations
  - Any model with temporal structure...
- Can we find some **weaker version of exchangeability** suitable for more general models, e.g. time-series?
- **This is possible for Markov chains!** Using the notion of **partial exchangeability** (aka Markov exchangeability) of Diaconis and Freedman

The Annals of Probability
1980, Vol. 8, No. 1, 115–130

DE FINETTI'S THEOREM FOR MARKOV CHAINS

BY P. DIACONIS[1] AND D. FREEDMAN[2]

Partial exchangeability, block-switch transformations

- Rather than being invariant to all permutations, we will restrict this to a particular kind call **block-switch transformations**, which are **permutations that do not change the distribution of Markov chains.** As argued by Diaconis and Freedman (1980), **this is the "right" kind of symmetry to invoke when dealing with Markov chains** (they prove an analogue of De Finetti's theorem using these transformations).

- A ***d*-block-switch transformation interchanges two disjoint blocks** of a chain **when these two blocks start with the same *d* symbols and end with the same d symbols**.

Formal definition of the block-switch transformation

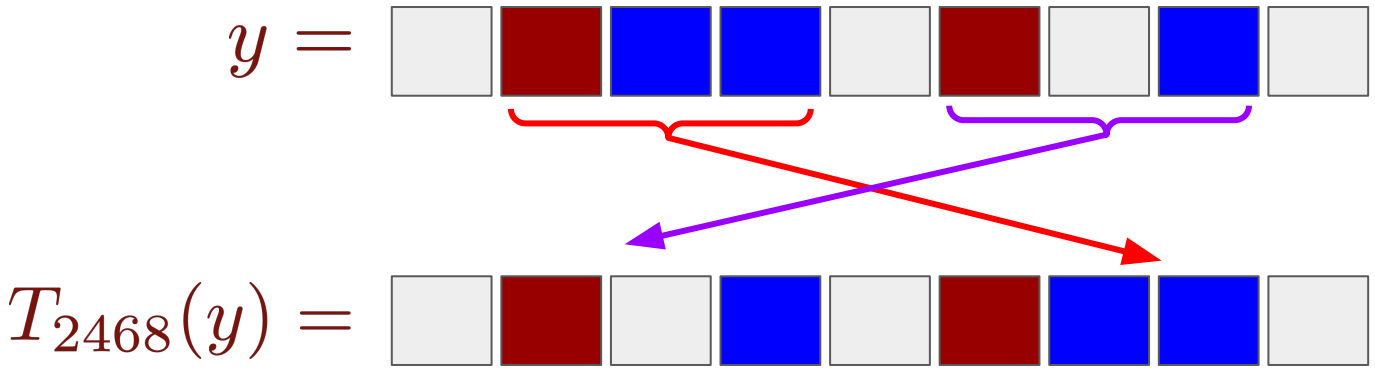
For increasing indices $b = (i, j, k, l) \in \{0, \dots, M\}^4$ such that $j - i \geq d$ and $l - k \geq d$, the d-**block-switch transformation** $T_b^{(d)}$ is defined as follows: if $y_{i:(i+d)} = y_{k:(k+d)}$ and $y_{(j-d):j} = y_{(l-d):l}$ then

$$y = y_{1:i-1}\, y_{i:j}\, y_{(j+1):(k-1)}\, y_{k:l}\, y_{(l+1):M}$$

$$T_b^{(d)}(y) = y_{1:i-1}\, y_{k:l}\, y_{(j+1):(k-1)}\, y_{i:j}\, y_{(l+1):M}\,.$$

If $y_{i:(i+d)} \neq y_{k:(k+d)}$ or $y_{(j-d):j} \neq y_{(l-d):l}$ then the block-switch transformation leaves y unchanged: $T_b^{(d)}(y) = y$.

A simple example with a chain of order 1



$$p(T_{2468}(y)) = p(y)$$

Towards a block-switch invariant summary function

- A function invariant to all block-switch transformations of order *d* is d-block-switch invariant. Similarly a model is said to be *d*-partially exchangeable.
- It is easy to see that

$$p(y|\theta) \;\; d-\text{Markov} \implies (y \mapsto E[\theta|y]) \;\; d-\text{block-switch-invariant}$$

therefore, **our summary function should be *d*-block-switch invariant!**

- Can we design neural nets that are *d*-block-switch invariant?

Partially exchangeable networks (PENs)

We proposed the following **PEN architecture**, that generalises deep sets to handle *d*-partially exchangeable data

$$\forall y \in \mathcal{Y}^M,\ S(y) = \rho\left(y_{1:d}, \sum_{i=1}^{M-d} \phi\left(y_{i:(i+d)}\right)\right)$$

outer network

$\rho : \mathcal{Y}^d \times \mathbb{R}^K \to \Theta$

transforms the representations and the beginning of the chain into regression target (e.g. posterior mean)

inner network

$\phi : \mathcal{Y}^{d+1} \to \mathbb{R}^K$

transforms each *d*-block into a K-dimensional representation

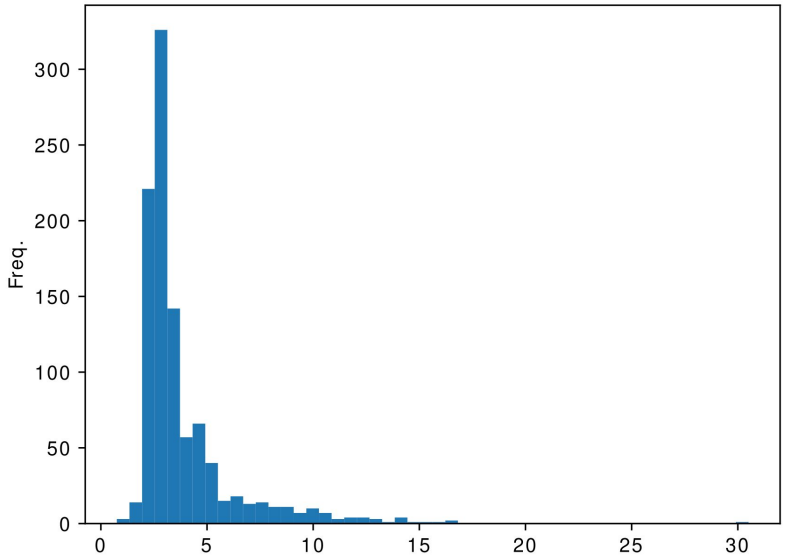
What are the properties of PENs?

$$\forall y \in \mathcal{Y}^M,\ S(y) = \rho\left(y_{1:d}, \sum_{i=1}^{M-d} \phi\left(y_{i:(i+d)}\right)\right)$$

- When *d*=0, this is exactly Deep Sets!
- When the data space is countable, we show that any block-switch invariant function admits a PEN representation.
- Related to the fact that the initial *d*-block and all (*d*+1)-blocks are sufficient statistics of a Markov chain of order *d*

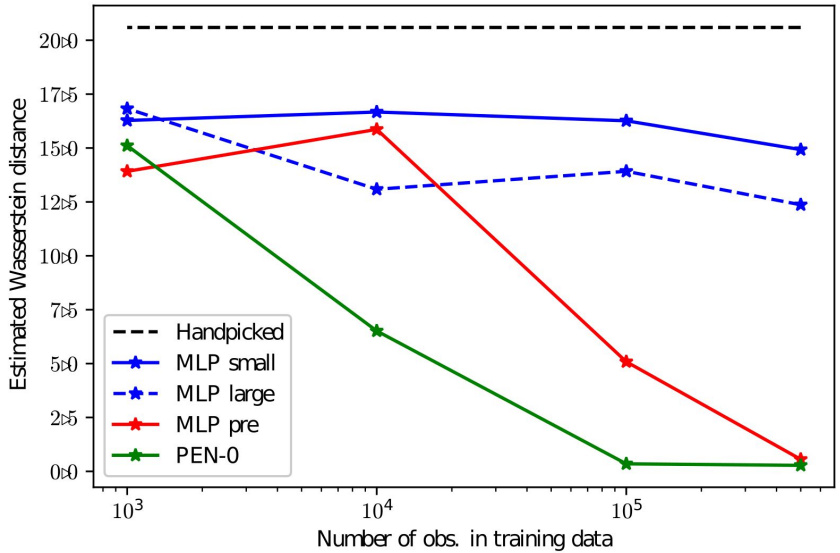
Case study - G-and-k distribution

- Task: inferring the parameters A,B,g,k (c is assumed to be known).
- The set-up was similar to other papers.
- Since the data is i.i.d we used PEN-0 (i.e. DeepSets).



Data for the g-and-k distribution.

Case study: G-and-k distribution



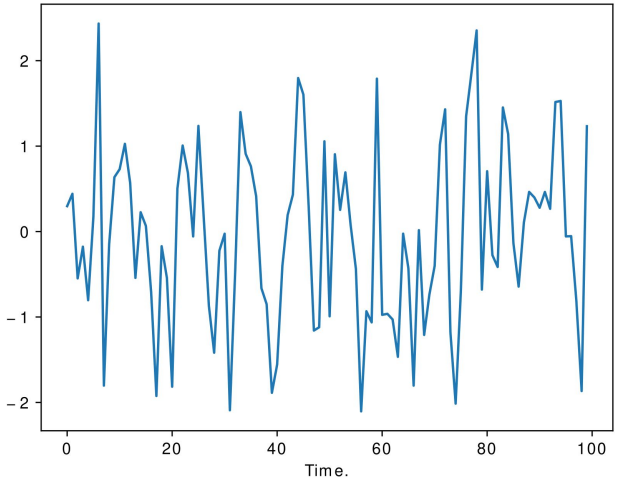
The estimated Wasserstein distances (mean over 100 repetitions) when comparing the MCMC posterior with ABC posteriors.

Case study - AR(2) process

- The AR(2) model follows:

$$y_l = \theta_1 y_{l-1} + \theta_2 y_{l-2} + z_l$$

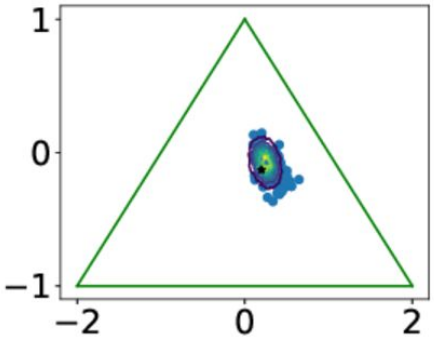
- The AR(2) model is a Markov model of order 2, thus we used PEN-2.



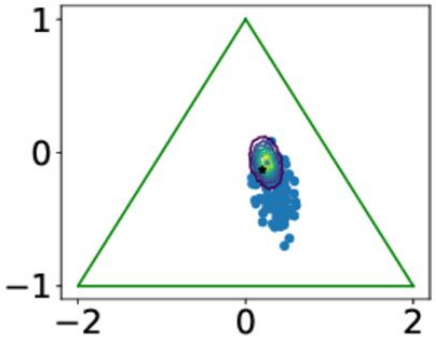
Data for the AR(2) model.

Case study AR(2)

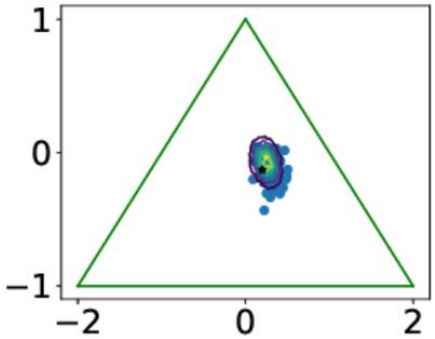
- Green line: prior
- Contour plot: true posterior



(a) Handpicked



(b) MLP (10^6)

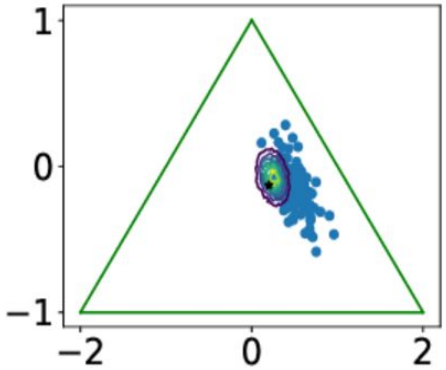


(c) PEN-2 (10^6)

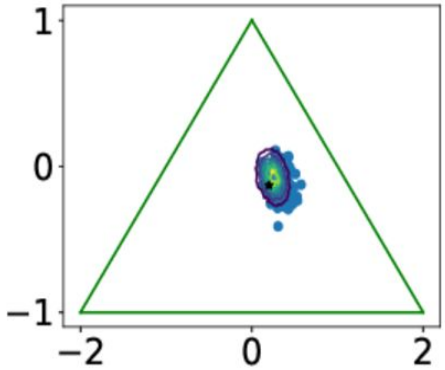
Posterior distributions (10^6 training data points).

Case study AR(2)

- Green line: prior
- Contour plot: true posterior



(d) MLP (10^5)

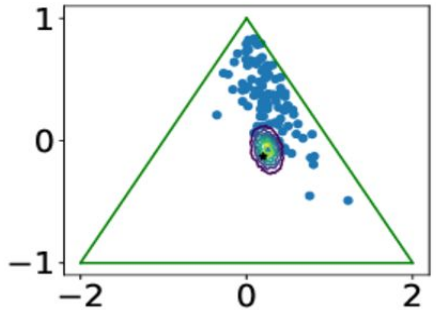


(e) PEN-2 (10^5)

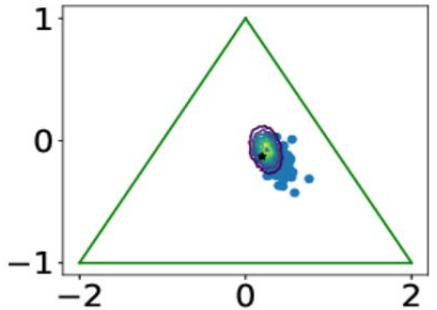
Posterior distribution (10^5 training data points).

Case study AR(2)

- Green line: prior
- Contour plot: true posterior



(f) MLP (10^4)

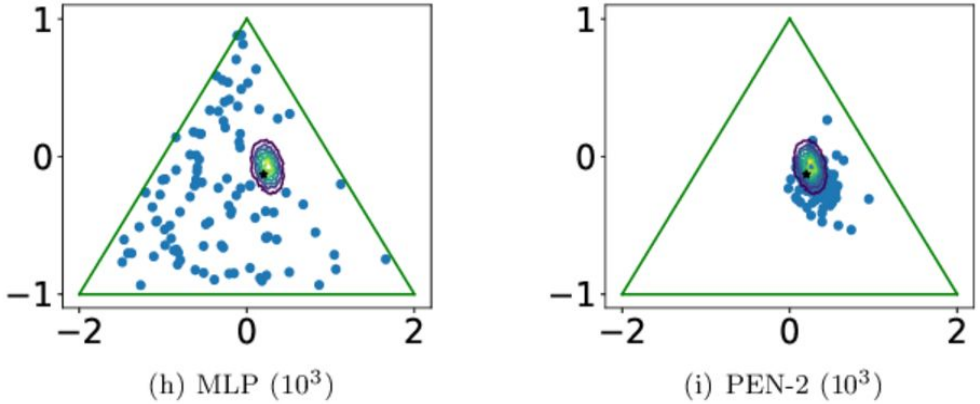


(g) PEN-2 (10^4)

Posterior distribution (10^4 training data points).

Case study AR(2)

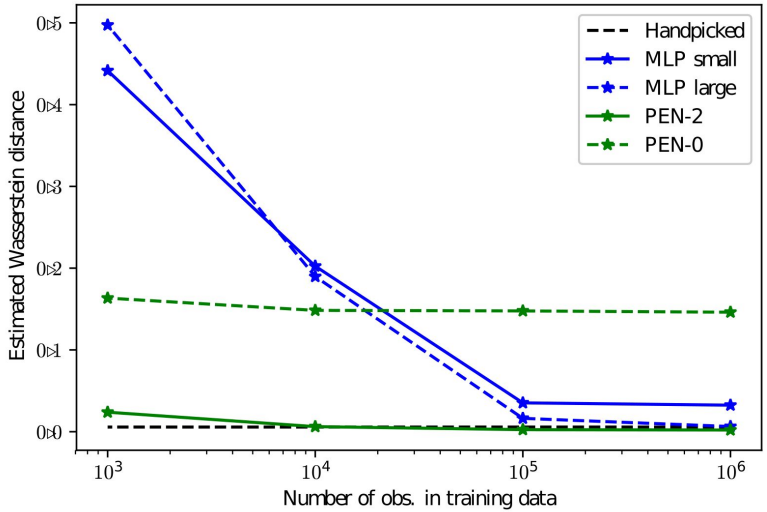
- Green line: prior
- Contour plot: true posterior



(h) MLP (10^3) (i) PEN-2 (10^3)

Posterior distribution (10^3 training data points).

Case study AR(2)



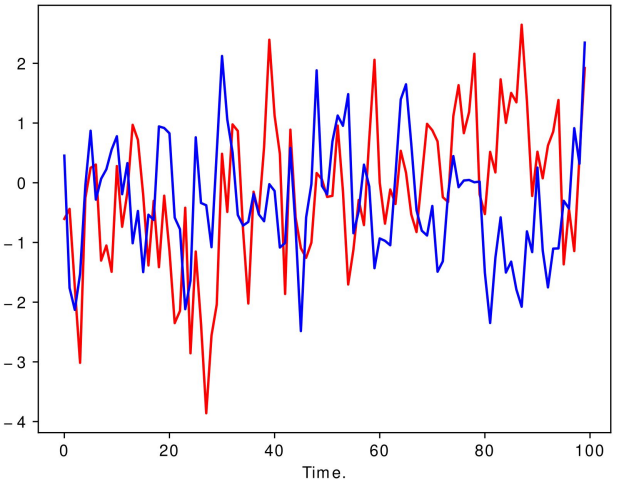
Estimated Wasserstein distances (mean over 100 data sets) when comparing the true posterior with ABC posteriors, for varying sizes of training data when using DNN models.

Case study - MA(2) process with observation noise

- The MA(2) model:

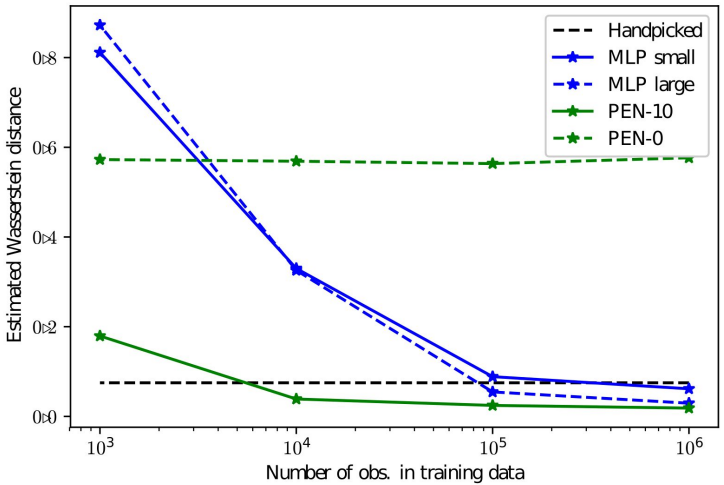
$$x_l = z_l + \theta_1 z_{l-1} + \theta_2 z_{l-2}$$
$$y_l = x_l + e_l, e_l \sim N(0, \sigma_\epsilon)$$

- Non-Markovian time series model, i.e. not partial exchangeable.
- We found PEN-10 to work the best.



Data for the MA(2). Blue without noise, red with noise.

Case study - MA(2) process with observation noise



Estimated Wasserstein distances (mean over 100 data sets) when comparing the true posterior with ABC posteriors.

Can we learn other properties of the posterior?

- In our work, we use PEN to learn the posterior mean from the prior predictive distribution.
- Radev et al. (arXiv 2020): Utilizes structures of the data (e.g. exchangeability) and learns the global posterior distribution from the prior predictive distribution.
- Chan et al. (NeurIPS 2018): Utilizes a exchangeable network to learn the parameter posterior distribution for population genetics models.

Some further ideas

- We can conclude that PEN is advantageous since it leverages symmetries.
- Can PEN be extended to other more complex symmetries?
- PEN can also be used in more advanced algorithms (e.g. ABC-MCMC, ABC-PMC).