The 'ABC' of Data Science on HPC Scale

Antonietta Mira
joint work with Ritabrata Dutta

SOS21, Davos, March 21, 2017

What is "big data"?

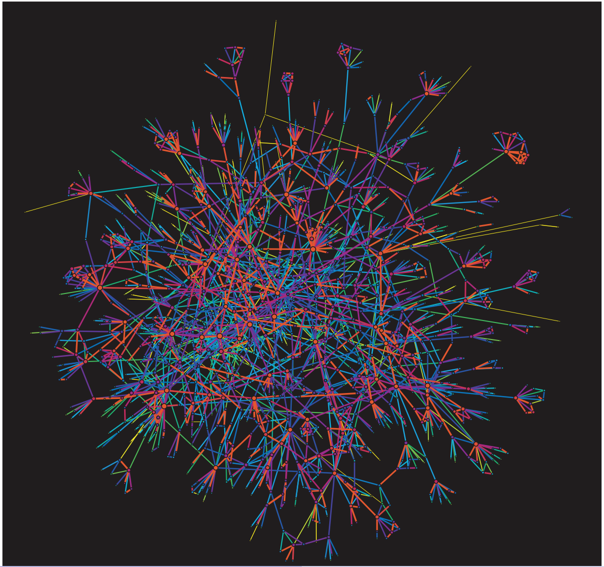
BIG = E-NORMI = EX (out of) NORMA (norm)
= EXTRA-ORDINARY LARGE

DATA = DATUS = GIVEN =Immediate, straightforward
DATA SCIENTISTS mediate big data and extract information

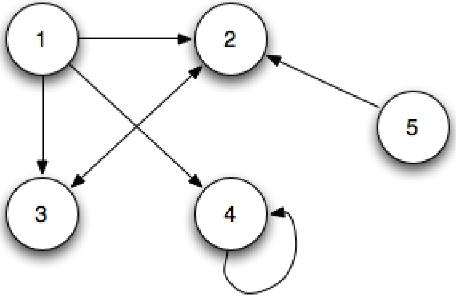
- big data can be small or fat (small n, large p)
- but typically is complex: not i.i.d. - not Gaussian - not linear
- unstructured, distributed
- smart data
- value chain: information - knowledge - decisions - actions

Terry Speed, 2014:
"... big data refers to things one can do at a large scale, that cannot be done at a smaller one, to extract new insights, or create new forms and value, in ways that change markets, organizations, the relationships between governments, citizens an more."

Communication network of 7 million nodes + 23 million ties



Relationships $\Longrightarrow$ Adjacency matrix $\Longrightarrow$ Model



	1	2	3	4	5
1	0	1	1	1	0
2	0	0	1	0	0
3	0	1	0	0	0
4	0	0	0	1	0
5	0	1	0	0	0

+ edge weights + covariates + time + spatial coordinates + . . .
= complex relational/network data

Network data

- social network: friendships
- financial network: EU overnight interbank money transfer
- economy: EU countries import-export
- biology/ genetics networks: protein-protein interaction
- communication network: CDR
- information / knowledge network: patents
- citation / collaboration network: wikipedia

What is the role of statistics in the Big Data era?

David Dunson, 2015:
"I would describe statistics as the science of variability, meaning that the main goal of statistics is to develop methods and algorithms for the mathematical exploration, elicitation and control of variability, and the uncertainty it generates. Inference and uncertainty quantification are at the core of statistics and they have generated correlated siblings like prediction, testing, controlling for dependence, confounding, randomization."

Model-driven approach

Data

Data-driven approach

Data-driven model approach

What is the role of statistics in the Big Data era?

Statistics challenges in the big/complex data era?

- multi-resolution: separate signal from noise
- dive for perceived signals in what would have been discarded as noise a decade ago
- multi-phase: data arriving at my desk are almost never the original raw data
- too dirty, too confidential, too large
- pre-processing with different goals/assumptions
- a single model is too simple to handle heterogeneity
- multiplicity of models capture multiplicity of incompatible assumptions
- multi-source
- different sources and some not collected for inference purposes
- sampling bias of observational / self-reported data

Statistics challenges in the big/complex data era?

- dimension reduction / summary / compression
- error rate control
- uncertainty quantification
- assure coherence among different scales of time/space
- support real-time decision making
- complex data - complex models
- big data - big errors
- big methodological and computational challenges

Why Bayesian statistics?

"A model-based revolution"
Sir Adrian Smith, DG Knowledge & Innovation, U. of London

Bayesian methods allow us to:

- Think differently about estimating and interpreting unknowns
  "what are possible values of this parameter?"
- Combine prior information with the data
  "what else do I know about this parameter and model?"
- Regularize the LHD and average the posterior
- Describe many sources of uncertainty in the model
  "how sure am I about the inputs and outputs of my model?"
- Analyze complex systems with hierarchical / multi-level models "Divide and conquer strategy"
- Perform model comparison and model averaging
- Bayesian non-parametric
- Bayesian computation (MCMC, Variational, INLA, ABC)

Big picture of statistical inference

GIVEN:

- Data $= y = (y_1, \ldots, y_n)$
- Statistical model which describes data, $p_{y|\theta}(y|\boldsymbol{\theta})$, indexed by Parameters $= \boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$
- Observed data $= y_{obs}$
- Prior probability density function for $\boldsymbol{\theta}$, $p_{\boldsymbol{\theta}}$

WANTED:

- Some probabilistic statement about $\boldsymbol{\theta}$
  - which value of $\boldsymbol{\theta}$ has, most likely, generated y_{obs} ?
  - what is the mean value of $\boldsymbol{\theta}$ given y_{obs}?
  - which interval contains θ_1 with probability 0.95 ?
  - $\ldots$

Big picture of statistical inference

GIVEN:

- Data $= y = (y_1, \ldots, y_n)$
- Statistical model which describes data, $p_{y|\theta}(y|\boldsymbol{\theta})$, indexed by Parameters $= \boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$
- Observed data $= y_{obs}$
- Prior probability density function for $\boldsymbol{\theta}$, $p_{\boldsymbol{\theta}}$

WANTED:

- Some probabilistic statement about $\boldsymbol{\theta}$
  - which value of $\boldsymbol{\theta}$ has, most likely, generated y_{obs} ?
  - what is the mean value of $\boldsymbol{\theta}$ given y_{obs}?
  - which interval contains θ_1 with probability 0.95 ?
  - ...

Different types of statistical models

1. Statistical model as family of pdfs, e.g.

$$p_{y|\theta}(y|\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i-\mu)^2\right), \quad \boldsymbol{\theta} = (\mu, \sigma)$$

2. Unnormalized statistical model
   (the partition function, of $p_{y|\theta}$ is not known)

$$p^0_{y|\theta}(y|\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(y_i-\mu)^2\right)$$

3. Simulator-based (generative/mechanistic) model
   (shape and scale of $p_{y|\theta}$ are not known but sampling is possible if parameters are given)

$$y \sim p_{y|\theta}(y|\boldsymbol{\theta}), \qquad y_i = \mu + \sigma z_i \quad z_i \sim \mathcal{N}(0,1)$$

Big picture of statistical inference

GIVEN:

- Data $= y = (y_1, \ldots, y_n)$
- Statistical model which describes data, $p_{y|\theta}(y|\boldsymbol{\theta})$, indexed by Parameters $= \boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$
- Observed data $= y_{obs}$
- Prior probability density function for $\boldsymbol{\theta}$, $p_{\boldsymbol{\theta}}$

WANTED:

- Some probabilistic statement about $\boldsymbol{\theta}$
  - which value of $\boldsymbol{\theta}$ has, most likely, generated y_{obs} ?
  - what is the mean value of $\boldsymbol{\theta}$ given y_{obs}?
  - which interval contains θ_1 with probability 0.95 ?
  - ...

Likelihood-Based Inference

- Likelihood function: pdf of the observed data y_{obs} as a function of the model parameters

$$L(\boldsymbol{\theta}) \propto p_{y|\boldsymbol{\theta}}(y_{obs}|\boldsymbol{\theta})$$

- Plays a central role in statistical inference
  - Maximum likelihood estimation:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, L(\boldsymbol{\theta})$$

  - Bayesian inference:

$$p_{\boldsymbol{\theta}|y}(\boldsymbol{\theta}|y_{obs}) \propto L(\boldsymbol{\theta}) p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$$

- *Not available for unnormalized and simulator-based models*

Why simulator-based models?

- Allow to use knowledge domain on how the data were generated without having to make excessive compromises in the modeling
- Neat interface with physical, social, medical, biological . . . models of data
- Scale well with big data
- No limits on the number of unobserved/latent variables
- Easier to study the effect of interventions on simulator-based (mechanistic) models rather than statistical models

Examples

- Astrophysics:
  Simulating the formation of galaxies, stars, or planets
- Evolutionary biology:
  Simulating species evolution
- Ecology:
  Simulating species migration over time
- Neuroscience:
  Simulating neural circuits
- Health science:
  Simulating the spread of an infectious disease
- Meteorology :
  Simulating weather prediction

Approximate Bayesian Computation (ABC) references

- ABC in population genetics, Beaumont, Zhang, Balding - Genetics, 2002
- Comparative evaluation of a new effective population size estimator based on approximate Bayesian computation Tallmon, Luikart, Beaumont - Genetics, 2004
- Inferring population history with DIY ABC: a user-friendly approach to ABC, Cornuet, Santos, Beaumont, Robert, Marin, . . . - Bioinformatics, 2008
- COMPUTER PROGRAMS: onesamp: a program to estimate effective population size using ABC, Tallmon, Koyuk, Luikart, Beaumont - Molecular Ecology Resources, 2008
- Adaptive ABC, Beaumont, Cornuet, Marin, Robert - Biometrika, 2009
- Approximate Bayesian computation without summary statistics: the case of admixture, Sousa, Fritz, Beaumont, Chikhi - Genetics, 2009
- Review: Marin, Statistics and Computing, 2012

Principle behind ABC

- Replace LHD, $p_y(y|\theta)$, by SUMMARY LHD $p_S(S(y)|\theta)$ where $S(y)$ = summary statistics
- But $p_S(S(y)|\theta)$ is also unknown
- Use an APPROXIMATE SUMMARY LHD: $\tilde{p}_S(S(y)|\theta)$, based on pseudo data y^* generated from the model
- The POSTERIOR is also approximate and summarized: $\tilde{p}_S(\theta|S(y)) \propto p(\theta)\tilde{p}_S(S(y)|\theta)$

Approximate Bayesian Computation (ABC)

- ABC algorithms are iterative

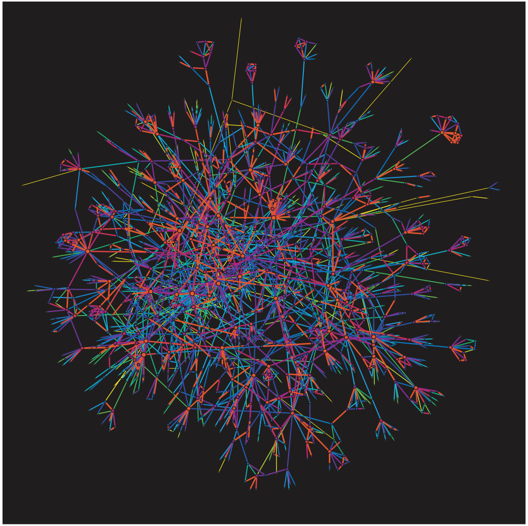
  The basic steps at each iteration are:

  1. proposing a parameter θ^*,
  2. simulating pseudo data y^*
  3. accepting or rejecting the proposed θ^* based on a comparison of y^* with the real observed data y_{obs}

- How to actually measure the discrepancy between the observed and the simulated pseudo data is a major difficulty in these methods

Motivating example

Technology generates new types of data and new modeling challenges



Motivation

- Systems of scientific and societal interest have large numbers of interacting components
- Representation as networks:
  node = component, edge = interaction
- E.G.: Friendship/Advisory network, Citation network, Webpage link network, Protein-Protein interactions
- Distinction between models of two things:
  - Models of network structure (e.g, Erdös-Rényi)
  - Models of dynamical processes on networks (e.g., SI model)
- Why care about network structure?
  - Interplay between network structure and the behavior of dynamical processes on networks (e.g., hubs in epidemics)

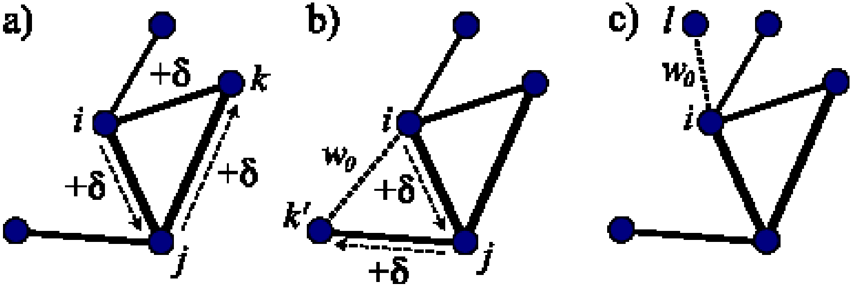
Network Models

Distinction between two types of models of network structure:

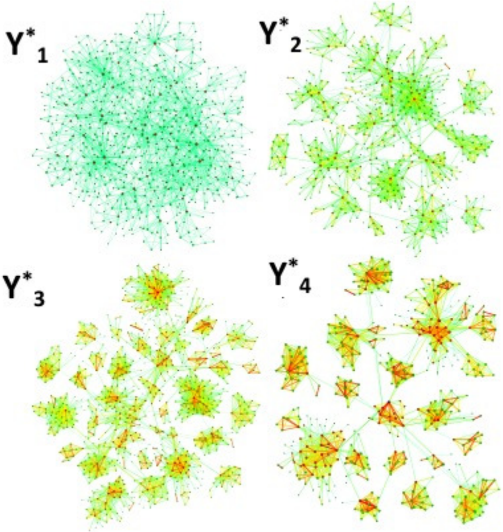
- **Statistical models** (e.g. ERGM,Goyal-Blitzstein-DeGruttola)
  DATA DRIVEN
  - **Pros**: inference on model parameters; hypothesis testing; model selection
  - **Cons**: scalability; hard to incorporate domain knowledge
- **Mechanistic models** (e.g. Price model)
  KNOWLEDGE DRIVEN
  assume that microscopic mechanisms that govern network formation and evolution are known, ask what happens if we apply these mechanisms repeatedly
  - **Pros**: easy to incorporate domain knowledge, scalability
  - **Cons**: no inferential tools; no model comparison

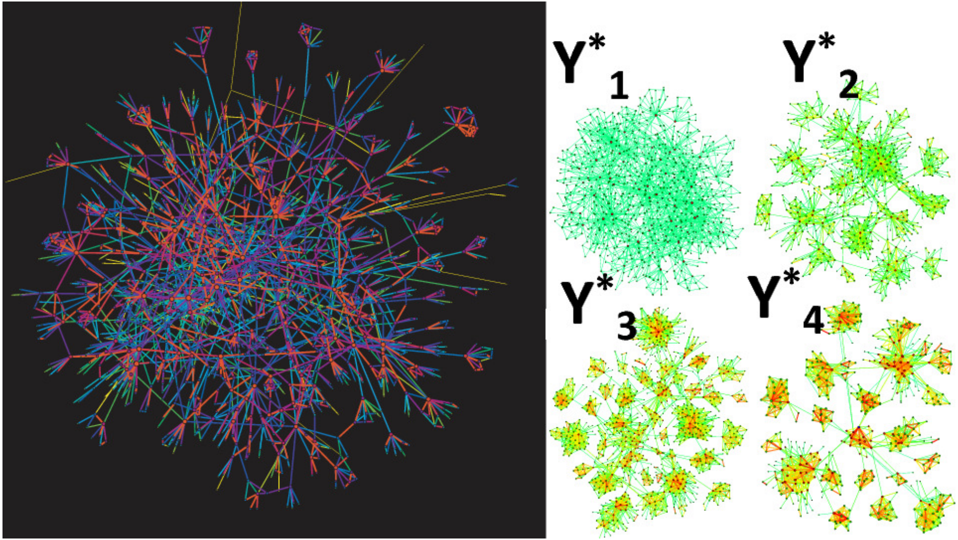
Mechanistic Model of Social and Contact Networks

- From the perspective of time expenditure of subject i:
  - spend time with existing friends (a)
  - become friend of a friend (b)
  - make totally new friends (c)



Mechanistic Model of Social and Contact Networks




Y*1
Y*2
Y*3
Y*4

Approximate Bayesian Computation (ABC)

- ABC rejection sampler is the simplest form of ABC

ABC rejection sampler

- Sample parameter θ^* from the prior $p(\theta)$
- Simulate dataset y^* under the given model specified by θ^*: $y^* \sim p(\cdot|\theta^*)$
- Accept θ^* if $\rho(y^*, y) \leq \epsilon$

- Distance measure $\rho(y^*, y)$ determines the level of discrepancy between the simulated data y^* and the observed data y

- The accepted θ^* are approximately distributed according to the desired posterior and, crucially, obtained without the need of explicitly evaluating the LHD

Approximate Bayesian Computation (ABC)

- It may be unfeasible to compute the distance $\rho(y^*, y)$ for high-dimensional data
- Lower dimensional summary statistic $S(y)$ to capture the relevant information in y
- Comparison is done between $S(y^*)$ and $S(y)$: accept θ^* if $\rho(S(y^*), S(y)) \leq \epsilon$
- If S is sufficient wrt θ, then it contains all information in y about θ (by definition), and using $S(y)$ in place of the full dataset does not introduce any error
- For most models it may be impossible to find sufficient statistics S, in which case application relevant summary statistics need to be used
- Use of non-sufficient summary statistics introduces a further level of approximation

ABC for Mechanistic Network Models

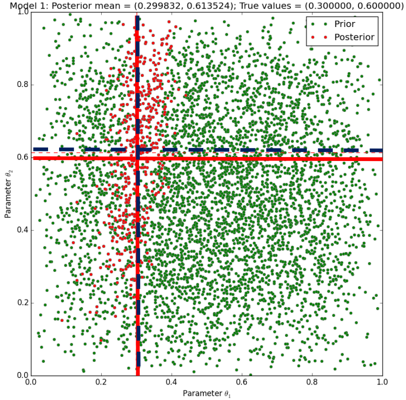
- ABC + mechanistic network models = generic + sound inferential framework

ABC rejection sampler for mechanistic network models

- Observe an empirical graph G
- Set up mechanistic network model M
- Sample parameter θ^* from the prior $p(\theta)$
- Simulate graph G^* from the mechanistic network model M using parameter θ^*
- Accept θ^* if $\rho(S(G^*), S(G)) \leq \epsilon$ using application relevant summaries S

- Some simple network summaries: degree sequence, k-stars, subgraph counts, centrality measures (betweenness, eigenvector, random walk, etc.), etc.
- Can use KNN to identify points in the space of summary statistics close to $S(G)$
- From Rejection-ABC to SMC-ABC by Drovandi + Pettitt (2015)

Inference on δ and w



ABC + generative network model

Prior and posterior draws

True parameter values: $\delta = 0.3$ and $w = 0.6$ (solid lines)

Posterior means: $\delta = 0.299$ and $w = 0.613$ (dashed lines)

$H_0 : \delta > \delta^*$ VS $H_1 : \delta \leq \delta^*$ for some arbitrary $\delta^* = 0.35$

Bayesians compute $P(H_0|y) = \int_{\theta^*}^{\infty} p(\theta|y)\, d\theta$
The integral can be estimated by summing over a finite set of samples θ_t from the posterior resulting in the estimator $\hat{P}(H_0|y) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{\delta_t > \delta^*}$
The posterior odds are defined as

$$\frac{P(H_0|y)}{P(H_1|y)} = \frac{P(H_0|G)}{P(H_1|G)} = \frac{P(H_0|G)}{1 - P(H_0|G)} \approx \frac{0.032}{0.968} \approx 0.033,$$

suggesting that H_1 is $1/0.033 = 30.25$ i.e. over 30 times more likely than H_0
We can confidently reject the null HP

ABC for model comparison (Part I)

- Observe an empirical graph G
- Identify alternative possible mechanistic network models M_1 and M_2
- Draw model index from the model prior: $\tau_1 = P(\mathcal{M} = 1) = P(\mathcal{M} = 2) = \tau_2$
- Draw parameter θ^* from the prior $p(\theta|\mathcal{M})$
- Simulate graph G^* from the given mechanistic network model using parameter θ^*
- Accept θ^* if $\rho(S(G^*), S(G)) \leq \epsilon$ using any summaries S

ABC for model comparison (Part II)

- Draw from the ABC approximation of the joint posterior $p(\theta, \mathcal{M}|y)$
- Generate n independent pseudo-data sets for each such draw (ABC approximation of the posterior predictive distribution)
- Compute posterior error rate using the random forest classifier i.e., how frequently it returns the true model index

ABCpy: A parallelized python library for ABC

ABCpy: efficient library to automatically parallelize ABC algorithms with a modular structure that allows

- no-HPC experts and no-ABC experts from different domains to run ABC in parallel
  USER-FRIENDLY
- ABC experts to develop parallel versions of different algorithms
  MODULAR
- HPC experts to develop different parallelization frameworks for ABCpy
  EXTENSIBLE
- researchers to compare efficiency parallelized ABC algorithms
  BENCHMARK

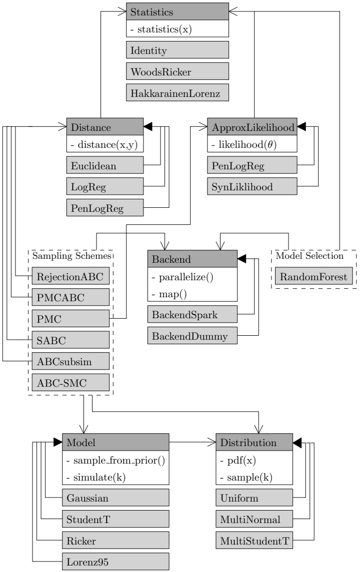
Abstract classes
dark grey

Derived classes
light grey

Filled arrows
inheritance

No filled arrows
association

map-reduce framework
with master/worker
architecture



Ricker Model: Stochastic population growth

- The unobservable population size $N(t)$ is

$$N(t) = rN(t-1)\exp\left(\frac{\sigma e(t)}{N(t-1)}\right)$$

- The observable population size $y(t)$ over discrete time $t = 0, \ldots, T$ is

$$y(t) \sim Poisson(\phi N(t))$$

- r is the growth rate, σ is the deviation of the innovation rate, and ϕ is a scaling parameter

- Goal: estimate the parameters r, σ, ϕ given the observed data

Lorenz Model for Numerical Weather Prediction

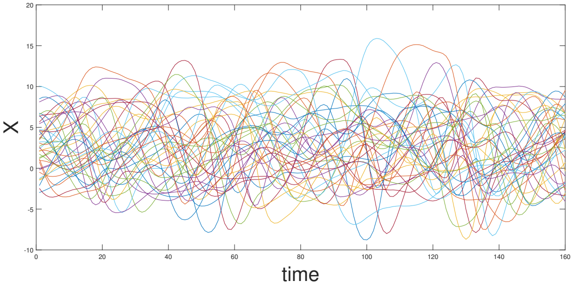
- Modification of weather prediction model of Lorenz (1995) when fast weather variables are unobserved (Wilks, 2005)
- $(Y_1^t, \ldots, Y_{40}^t)$: slow weather variables observed at time t
- Known: Initial value $(Y_1^0, \ldots, Y_{40}^0)$
- Goal: Simulate weather variables in future for numerical weather prediction with $t \in [0, 4]$ corresponding to 20 days

Model: SDEs of Weather Variables

- The weather variables follow the coupled Stochastic DE:
$\frac{dy_k^{(t)}}{dt} = -y_{k-1}^{(t)}(y_{k-2}^{(t)} - y_{k+1}^{(t)}) - y_k^{(t)} + 10 - g(y_k^{(t)}, \theta) + \eta_k^{(t)}$
- $g(y_k^{(t)}, \theta)$ = **deterministic** parametrization of the net effect of the unobserved (fast) variables on the observable ones
$g(y_k^{(t)}, \theta) = \sum_{i=1}^{2} \theta_i \left(y_k^{(t)}\right)^{i-1}$
- $\eta_k^{(t)}$ = **stochastic** forcing term representing the uncertainty due to forcing the fast variables, updated for an interval Δt
$\eta_k^{(t+\Delta t)} = \phi\eta_k^{(t)} + (1-\phi^2)^{\frac{1}{2}} e^{(t)}, t \in \{0, \Delta t, \ldots, T\Delta t\}$ $\eta^{(0)} = (1-\phi^2)^{\frac{1}{2}} e^{(0)}$ and $e^{(t)}$ are indep. standard normal r.v.
- we discretize the 20 days time interval in 5760 steps and use an **SDE 4th order solver**

Inference: Unknown coupling Parameters

- Unknown parameters: (θ_1, θ_2)
- Observed weather variables for $t \in [0, 4]$ in T = 160 equal intervals simulated using $(\theta_1, \theta_2) = (2.1, 0.1)$



ABCpy: Linear scaling up

We consider:

- The *speedup* $\mathcal{S}_{\mathcal{A}}(n)$ of a parallel algorithm $\mathcal{A}$ on n cores with respect to a baseline (number of cores) m, $m \leq n$, is the ratio of the algorithms running time $t(m)$ on m cores and the running time $t(n)$ on n cores:
  $\mathcal{S}_{\mathcal{A}}(n) = t(m)/t(n)$

- The *efficiency* $\mathcal{E}_{\mathcal{A}}(n)$ of an algorithm $\mathcal{A}$ on n cores is the speedup normalized by the numbers of cores:
  $\mathcal{E}_{\mathcal{A}}(n) = \mathcal{S}_{\mathcal{A}}(n)/n$

PMCABC: Linear scaling up

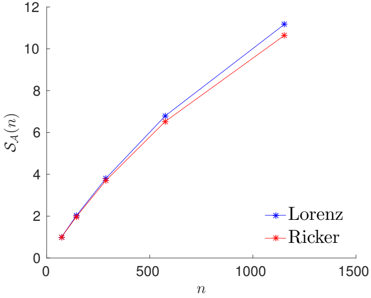


Figure 1: Speedup for PMCABC

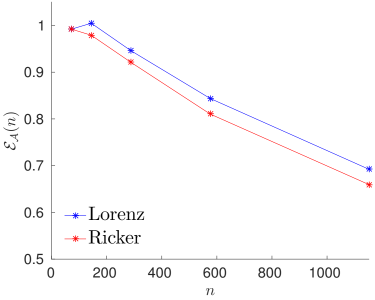


Figure 2: Efficiency for PMCABC

Data simulation from Ricker model: milliseconds
Data simulation from Lorenz model: seconds
CSCS Piz Daint with Apache Spark: 1 master, 2 to 32 workers, 72 to 1152 cores
10 min at most (32 workers)

PMC: Linear scaling up

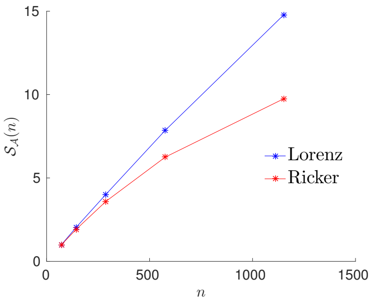


Figure 3: Speedup for PMC

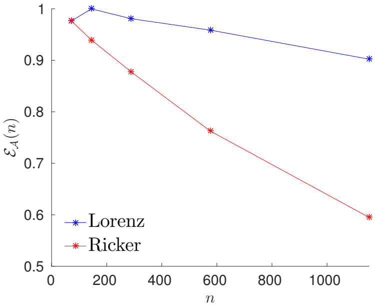


Figure 4: Efficiency for PMC

Amdahl's law

Data simulation from Ricker model: milliseconds

Data simulation from Lorenz model: seconds