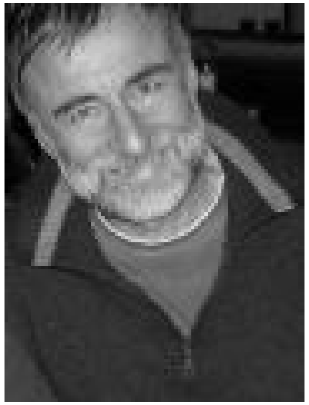
ABC model choice via random forests

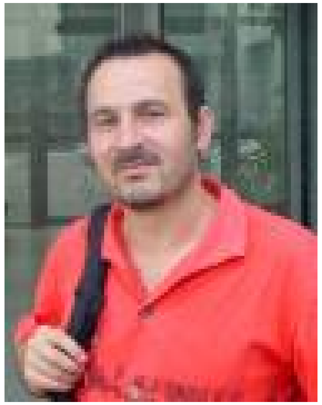
Jean-Michel Marin

Université de Montpellier
Institut de Mathématiques et Modélisation de Montpellier (I3M)
Institut de Biologie Computationnelle (IBC)
Labex Numev

Montpellier, 28/11/2014, Journée AppliBUGS

Mathieu Gautier, Pierre Pudlo & Christian Robert





Jean-Marie Cornuet
Apis mellifera specialist
The western honeybee

Arnaud Estoup
Harmonia axyridis specialist
The Asian ladybird beetle

Bayesian model choice

J models in competition

A model is characterized by a likelihood function $f_k(\mathbf{y}|\boldsymbol{\theta}_k)$ and a prior distribution on the parameter $\boldsymbol{\theta}_k \in \Theta_k$.

Prior probabilities in the model space are defined.

The posterior distribution in the model space is such that

$$\mathbb{P}^\pi\left(\mathcal{M} = k|\mathbf{y}\right) \propto \mathbb{P}(\mathcal{M} = k) \int_{\Theta_k} f_k(\mathbf{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k)\, \mathbf{d}\boldsymbol{\theta}_k\,.$$

Some computational difficulties:

- How to approximate the evidences?
- When the number of models in consideration is huge, how to explore the models's space?
- How to proceed when the calculation of the likelihood in intractable?

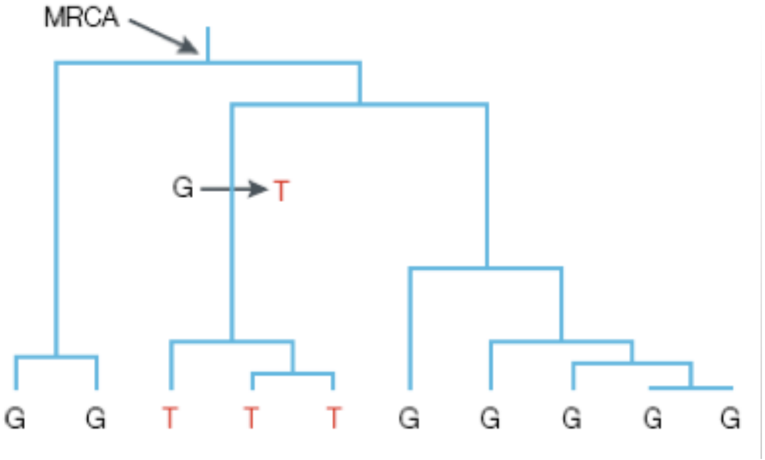
Genesis of Approximate Bayesian Computation methods

If, with Christian, we work on ABC methods, we can be very grateful to our biologist colleagues!

Arnaud and Jean-Marie typical questions:

- How the Asian ladybird beetle arrived in Europe?
- Why do they swarm right now?
- What are the routes of invasion?
- How to get rid of them?

Answer using molecular data, Kingman's coalescent process and ABC inference procedures!

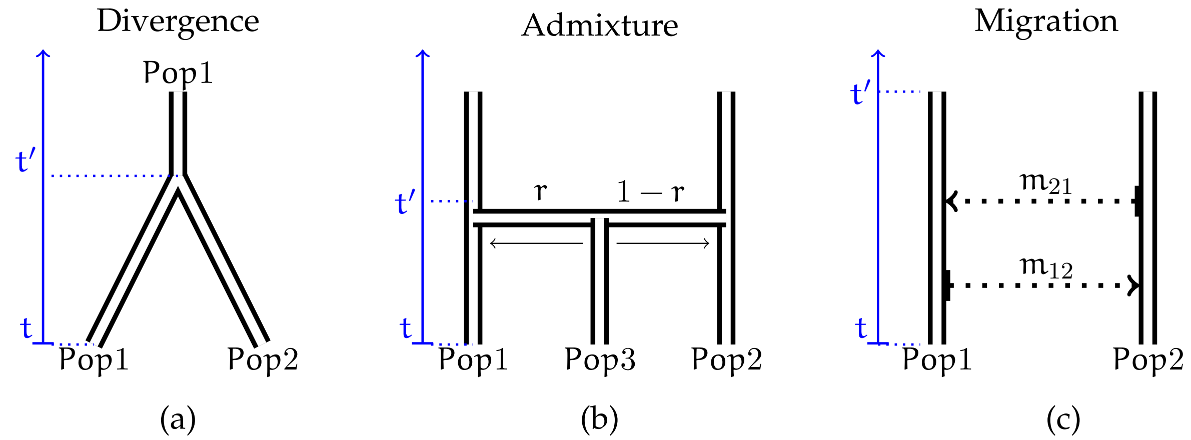


The Kingman's coalescent process is concerned with the genealogy of a sample of genes back in time to the common ancestor of the sample.

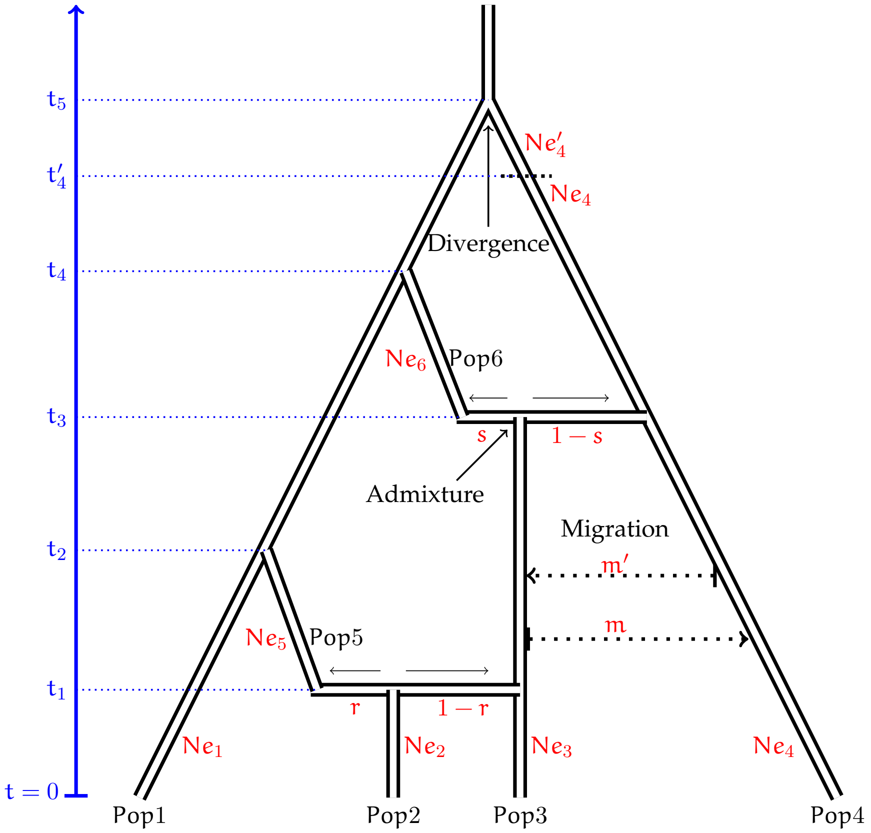
Within population model: Kingman's coalescent

Between populations: three types of events, backward in time

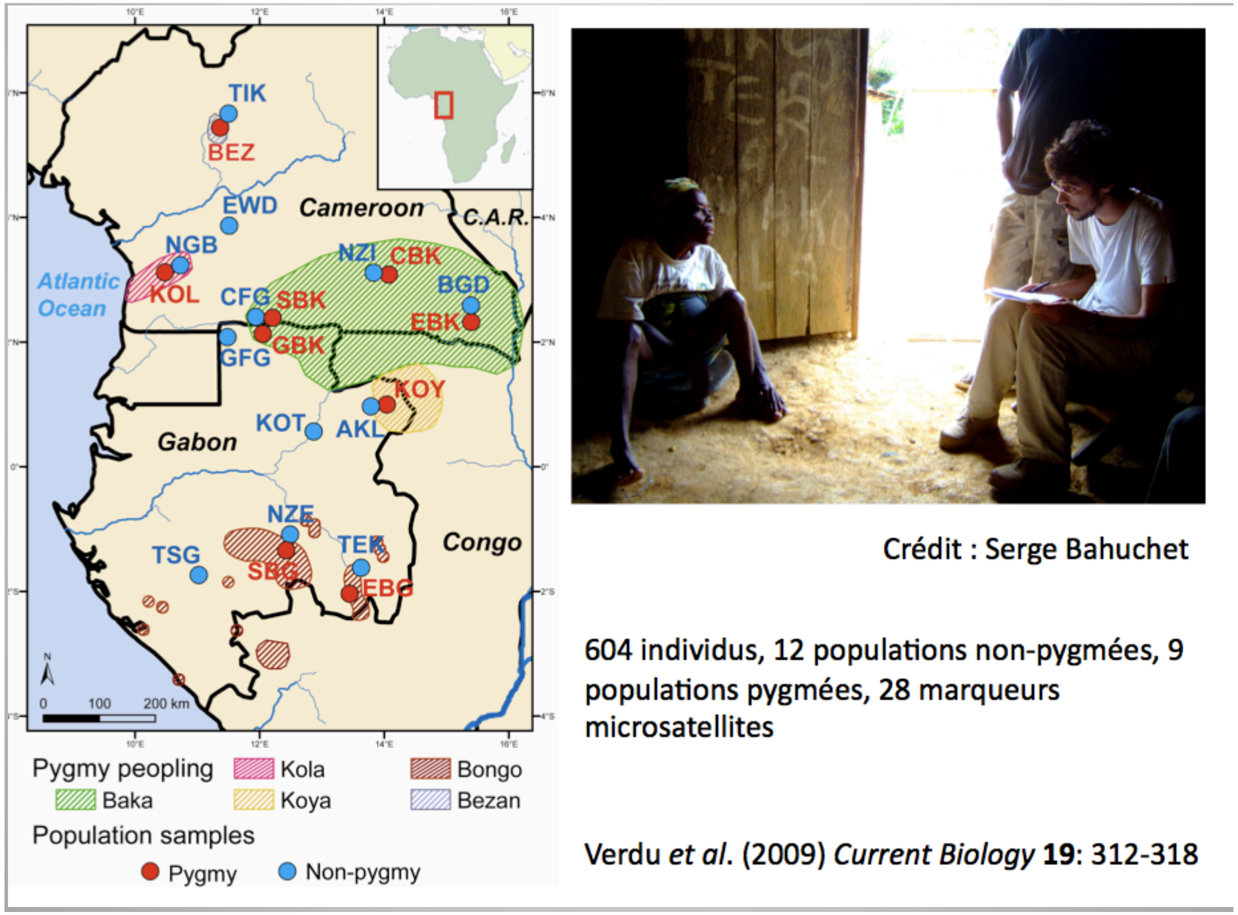
- **the divergence** is the fusion between two populations,
- **the admixture** is the split of a population into two parts,
- **the migration** allows the move of some lineages of a population to another.



The goal is to discriminate between different population scenarios from a dataset of polymorphism (DNA sample) **y** observed at the present time.



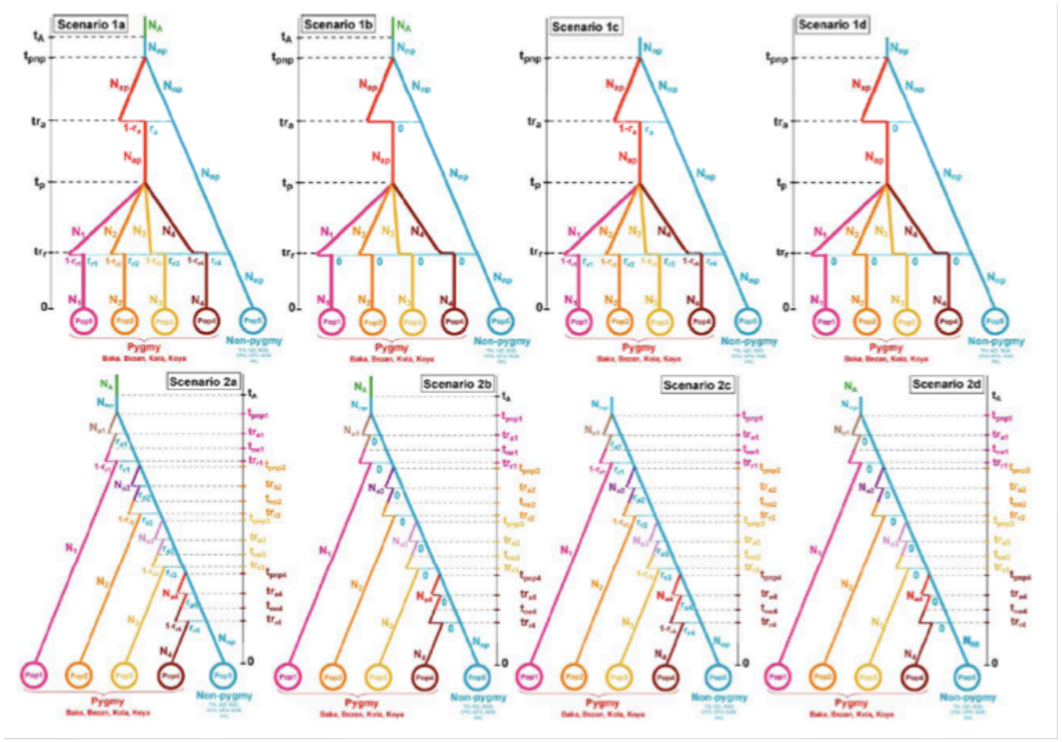
Example in human population genetics: do pygmies share a common ancestral populations?



Crédit : Serge Bahuchet

604 individus, 12 populations non-pygmées, 9 populations pygmées, 28 marqueurs microsatellites

Verdu *et al.* (2009) *Current Biology* **19**: 312-318



Kingman's coalescent is constrainted to live in the above "pipes"

For one population, the joint likelihood of the genealogy and of the polymorphism observed at the present time is available.

Data: polymorphism observed at the present time.

The calculation of the likelihood for the parameters involves integrating out the unknown tree.

Intractable likelihood!

Choosing a population scenario is a model choice problem.

When the likelihood function $f(\mathbf{y}|\boldsymbol{\theta})$ is expensive or impossible to calculate, it is extremely difficult to sample from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$. Two typical situations:

$f(\mathbf{y}|\boldsymbol{\theta}) = \displaystyle\int f(\mathbf{y}, \mathbf{u}|\boldsymbol{\theta})\mu(\mathrm{d}\mathbf{u})$, the calculation of this integral is intractable and the latent vector $\mathbf{u}$ takes values in a high dimensional space (e.g. population genetics models).

$f(\mathbf{y}|\boldsymbol{\theta}) = g(\mathbf{y}, \boldsymbol{\theta})/Z(\boldsymbol{\theta})$ and the calculation of $Z(\boldsymbol{\theta})$ is intractable (e.g. for Markov random fields).

ABC is a technique that only requires being able to sample from the likelihood $f(\cdot|\boldsymbol{\theta})$.

This technique stemmed from population genetics models, about 15 years ago, and population geneticists still significantly contribute to methodological developments of ABC.

ABC recap

Likelihood free rejection sampling

Rubin (1984) The Annals of Statistics

Tavaré et al. (1997) Genetics

Pritchard et al. (1999) Mol. Biol. Evol.

```
1) Set $i = 1$,
2) Generate $\boldsymbol{\theta}'$ from the prior distribution $\pi(\cdot)$,
3) Generate $\mathbf{z}$ from the likelihood $f(\cdot|\boldsymbol{\theta}')$,
4) If $\rho(\eta(\mathbf{z}), \eta(\mathbf{y})) \leq \epsilon$, set $\boldsymbol{\theta}_i = \boldsymbol{\theta}'$ and $i = i + 1$,
5) If $i \leq N$, return to 2).
```

We keep the $\boldsymbol{\theta}$'s values such that the distance between the corresponding simulated dataset and the observed dataset is small enough.

The likelihood free algorithm sample from the marginal in $\mathbf{z}$ of:

$$\pi_\epsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta}) \mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z})}{\int_{A_{\epsilon,\mathbf{y}} \times \Theta} \pi(\boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta}) \mathrm{d}\mathbf{z} \mathrm{d}\boldsymbol{\theta}},$$

- $\epsilon > 0$ a tolarance level,
- $\mathbb{I}_B(\cdot)$ the indicator function of a given set B,
- $A_{\epsilon,\mathbf{y}} = \{\mathbf{z} \in \mathcal{D} | \rho(\eta(\mathbf{z}), \eta(\mathbf{y})) \leq \epsilon\}$.

The idea behind ABC is that the summary statistics coupled with a small tolerance should provide a good approximation of the posterior distribution:

$$\pi_\epsilon(\boldsymbol{\theta}|\mathbf{y}) = \int \pi_\epsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) \mathrm{d}\mathbf{z} \approx \pi(\boldsymbol{\theta}|\mathbf{y}).$$

Curse of dimensionality

We have to summarize! If not the distance's values are too noisy. In the exponential family case, we use sufficient statistics, outside...

Toy example: the simulated summary statistics $\eta(\mathbf{z}_1), \ldots, \eta(\mathbf{z}_N)$ and the observed one $\eta(\mathbf{y})$ are iid with uniform distribution on $[0,1]^d$

Let $d_\infty(d, N) = \mathbb{E}\left[\min_{i=1,\ldots,N} \left\|\eta(\mathbf{y}_{\text{obs}}) - \eta(\mathbf{y}_i)\right\|_\infty\right]$

	$N = 100$	$N = 1,000$	$N = 10,000$	$N = 100,000$
$\delta_\infty(1, N)$	0.0025	0.00025	0.000025	0.0000025
$\delta_\infty(2, N)$	0.033	0.01	0.0033	0.001
$\delta_\infty(10, N)$	0.28	0.22	0.18	0.14
$\delta_\infty(200, N)$	0.48	0.48	0.47	0.46

Likelihood free MCMC sampler

(Majoram et al. (2003) PNAS)

Adaptive samplers

(Sisson et al. (2007) PNAS)
(Beaumont, Cornuet, Marin and Robert (2009) Biometrika)
(Del Moral et al. (2012) Statistics and Computing)
(Filippi et al. (2013) Statistical Applications in Genetics and Molecular Biology
(Sedki, Cornuet, Marin, Pudlo and Robert (2014) Preprint)

Regression adjustments

One central question: how to choose the set of summary statistics?

- Parameter estimation

  Joyce and Marjoram (2008) SAGMB

  Nunes and Balding (2010) SAGMB

  Fearnhead and Prangle (2012) JRSS B

  Blum et al. (2013) Statistical science

- Model choice

  Barnes et al. (2012) Statistics and Computing

  Fearnhead et al. (2014) SAGMB

ABC and model choice

Infering population history with DIY ABC: a user-friedly approach Approximate Bayesian Computation

Cornuet, Santos, Beaumont, Robert, Marin, Balding, Guillemaud, Estoup (2008) Bioinformatics

DIYABC v2.0: a software to make Approximate Bayesian Computation inferences about population history using Single Nucleotide Polymorphism, DNA sequence and microsatellite data

Cornuet, Pudlo, Veyssier, Dehne-Garcia, Gautier, Leblois, Marin, Estoup (2014) Bioinformatics

ABC algorithm for model choice

```
1) Set $i = 1$,
2) Generate $m'$ from the prior $\pi(\mathcal{M} = m)$,
3) Generate $\boldsymbol{\theta}'_{m'}$ from the prior $\pi_{m'}(\cdot)$,
4) Generate $\mathbf{z}$ from the model $f_{m'}(\cdot|\boldsymbol{\theta}'_{m'})$,
5) If $\rho(\eta(\mathbf{z}), \eta(\mathbf{y})) \leq \epsilon$, set $m^i = m'$, $\boldsymbol{\theta}^i_{m^i} = \boldsymbol{\theta}'_{m'}$ and $i = i + 1$,
6) If $i \leq N$, return to 2).
```

If $\eta(\mathbf{y})$ is a sufficient statistics for the model choice problem, this can work pretty well.

ABC likelihood-free methods for model choice in Gibbs random fields Grelaud, Robert, Marin, Rodolphe and Taly (2009) Bayesian Analysis

If not...

Lack of confidence in approximate Bayesian computation model choice Robert, Cornuet, Marin, Pillai (2011) PNAS

Relevant statistics for Bayesian model choice Marin, Pillai, Robert, Rousseau (2014) JRSS B

The use of random forests

Learning towards machine learning

In practice, people do not have a fixed tolerance level.

The tolerance level is a random variable and the number of simulations is fixed!

Real ABC algorithm for model choice

Let $N = \lfloor \alpha M \rfloor$.

```
1) For $i = 1, \ldots, M$:
    a) Generate $m_i$ from the prior $\pi(\mathcal{M} = m)$,
    b) Generate $\boldsymbol{\theta}'_{m_i}$ from the prior $\pi_{m_i}(\cdot)$,
    c) Generate $\mathbf{z}$ from the model $f_{m_i}(\cdot|\boldsymbol{\theta}'_{m_i})$,
    d) Calculate $d_i = \rho(\eta(\mathbf{z}), \eta(\mathbf{y}_{obs}))$,
6) Order the distances $d_{(1)}, \ldots, d_{(M)}$,
7) Return the $m_i$'s that correspond to the $N$-smallest distances.
```

We get $\epsilon = d_{\lfloor \alpha M \rfloor}$.

That is a knn approximation of the posterior probabilities!

We investigate some ABC model choice techniques that use others machine learning procedures.

Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics

Estoup, Lombaert, Marin, Guillemaud, Pudlo, Robert, Cornuet (2012) Molecular Ecology

Key points

- ABC model choice seen as learning about which model is most appropriate from a huge (reference) table
- exploiting a large number of summary statistics is not an issue for some machine learning methods intended to estimate efficient combinations
- abandoning (temporarily?) the idea of estimating posterior probabilities of the models, poorly approximated by machine learning methods

Random Forests

Technique that stemmed from Leo Breiman's bagging (or bootstrap aggregating) machine learning algorithm for both classification and regression
Breiman (1996) Machine Learning

Improved classification performances by averaging over classification schemes of randomly generated training sets, creating a forest of CART decision trees
Breiman (2001) Machine Learning

Breiman's solution for inducing random features in the trees of the forest:

- boostrap resampling of the dataset and
- random subseting of the covariates driving the classification at every node of each tree

Input ABC reference table involving model index, parameter values and summary statistics for the associated simulated pseudo-data [possibly large collection of summary statistics (from scientific theory input to available statistical softwares, to machine-learning alternatives)]

Output a random forest classifier to infer model indexes

Random forest predicts a MAP model index, from the observed dataset: the predictor provided by the forest is good enough to select the most likely model but not to derive associated posterior probability

- exploit entire forest by computing how many trees lead to picking each of the models under comparison but variability too high to be trusted
- frequency of trees associated with majority model is no proper substitute to the true posterior probability
- and usual ABC model choice approximation equally highly variable and hard to assess

$$\mathbb{P}^{\pi}\left(\mathcal{M}=k|\mathbf{y}\right) \propto \mathbb{P}(\mathcal{M}=k)\int_{\Theta_k} f_k(\mathbf{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k)\,\mathrm{d}\boldsymbol{\theta}_k$$

difficult to estimate

We propose to use

$$\mathbb{P}^{\pi}\left(\widehat{\mathcal{M}}(\mathbf{z})=\mathcal{M}|\mathbf{y}\right)$$

where $(\mathcal{M},\mathbf{z})$ generated from the predictive and $\widehat{\mathcal{M}}(\mathbf{z})$ denotes the random forest MAP predictor

Arguments

- Bayesian estimate of the posterior error
- integrates error over most likely part of the parameter space
- gives an averaged error rather than the posterior probability of the null hypothesis
- easily computed: given ABC subsample of parameters from reference table, simulate pseudo-samples associated with those and derive error frequency

On populations genetics examples,

random forests require many less prior simulations

random forests can deal with correlated summary statistics

assess confidence in the selection with posterior predictive expected losses

Thank you