Reliable Approximate Bayesian computation (ABC) model choice via random forests

Christian P. Robert

Université Paris-Dauphine, Paris & University of Warwick, Coventry

Max-Plank-Institut für Physik,

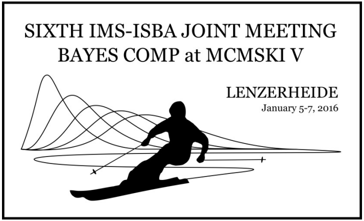
October 16, 2015

bayesianstatistics@gmail.com

Joint with J.-M. Cornuet, A. Estoup, J.-M. Marin, & P. Pudlo

The next MCMSkv meeting:

- Computational Bayes section of ISBA major meeting:
- MCMSki V in Lenzerheide, Switzerland, Jan. 5-7, 2016



- MCMC, pMCMC, SMC^2, HMC, ABC, (ultra-) high-dimensional computation, BNP, QMC, deep learning, &tc
- Plenary speakers: S. Scott, S. Fienberg, D. Dunson, K. Latuszynski, T. Lelièvre
- Call for contributed 9 sessions and tutorials opened
- "Switzerland in January, where else...?!"

Outline

Intractable likelihoods

ABC methods

ABC for model choice

ABC model choice via random forests

intractable likelihood

Case of a well-defined statistical model where the likelihood function

$$\ell(\theta|\mathbf{y}) = f(y_1, \dots, y_n|\theta)$$

- is (really!) not available in closed form
- cannot (easily!) be either completed or demarginalised
- cannot be (at all!) estimated by an unbiased estimator
- examples of latent variable models of high dimension, including combinatorial structures (trees, graphs), missing constant $f(x|\theta) = g(y,\theta)\big/Z(\theta)$ (eg. Markov random fields, exponential graphs,. . .)

© Prohibits direct implementation of a generic MCMC algorithm like Metropolis–Hastings which gets stuck exploring missing structures

intractable likelihood

Case of a well-defined statistical model where the likelihood function

$$\ell(\theta|\mathbf{y}) = f(y_1, \dots, y_n|\theta)$$

- is (really!) not available in closed form
- cannot (easily!) be either completed or demarginalised
- cannot be (at all!) estimated by an unbiased estimator

© Prohibits direct implementation of a generic MCMC algorithm like Metropolis–Hastings which gets stuck exploring missing structures

Necessity is the mother of invention

Case of a well-defined statistical model where the likelihood function

$$\ell(\theta|\mathbf{y}) = f(y_1, \ldots, y_n|\theta)$$

is out of reach

Empirical A to the original B problem

- Degrading the data precision down to tolerance level ε
- Replacing the likelihood with a non-parametric approximation based on simulations
- Summarising/replacing the data with insufficient statistics

Necessity is the mother of invention

Case of a well-defined statistical model where the likelihood function

$$\ell(\theta|\mathbf{y}) = f(y_1, \dots, y_n|\theta)$$

is out of reach

Empirical A to the original B problem

- Degrading the data precision down to tolerance level ε
- Replacing the likelihood with a non-parametric approximation based on simulations
- Summarising/replacing the data with insufficient statistics

Necessity is the mother of invention

Case of a well-defined statistical model where the likelihood function

$$\ell(\theta|\mathbf{y}) = f(y_1, \ldots, y_n|\theta)$$

is out of reach

Empirical A to the original B problem

- Degrading the data precision down to tolerance level ε
- Replacing the likelihood with a non-parametric approximation based on simulations
- Summarising/replacing the data with insufficient statistics

Approximate Bayesian computation

Intractable likelihoods

ABC methods

- Genesis of ABC
- abc of ABC
- Advances and interpretations
- Summary statistic

ABC for model choice

ABC model choice via random forests

Genetic background of ABC

▸ skip genetics

ABC is a recent computational technique that only requires being able to sample from the likelihood $f(\cdot|\theta)$

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

[Griffith & al., 1997; Tavaré & al., 1999]

Demo-genetic inference

Each model is characterized by a set of parameters θ that cover historical (time divergence, admixture time ...), demographics (population sizes, admixture rates, migration rates, ...) and genetic (mutation rate, ...) factors

The goal is to estimate these parameters from a dataset of polymorphism (DNA sample) $\mathfrak{y}$ observed at the present time

Problem:

most of the time, we cannot calculate the likelihood of the polymorphism data $f(\mathfrak{y}|\theta)$...

Demo-genetic inference

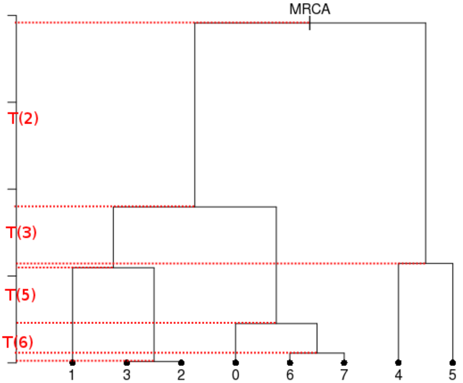
Each model is characterized by a set of parameters θ that cover historical (time divergence, admixture time ...), demographics (population sizes, admixture rates, migration rates, ...) and genetic (mutation rate, ...) factors

The goal is to estimate these parameters from a dataset of polymorphism (DNA sample) $\mathfrak{y}$ observed at the present time

Problem:

most of the time, we cannot calculate the likelihood of the polymorphism data $f(\mathfrak{y}|\theta)$...

Kingman's colaescent



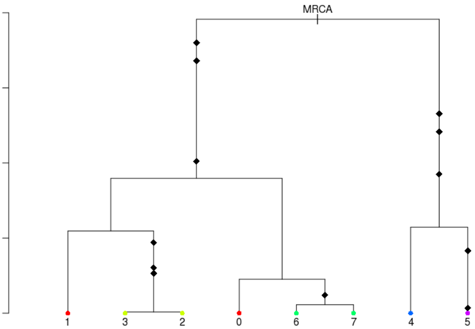
Kingman's genealogy

When time axis is normalized, $T(k) \sim \mathrm{Exp}(k(k-1)/2)$

Mutations according to the Simple stepwise Mutation Model (SMM)

- date of the mutations $\sim$ Poisson process with intensity $\theta/2$ over the branches
- MRCA $= 100$
- independent mutations: ± 1 with pr. $1/2$

Kingman's colaescent



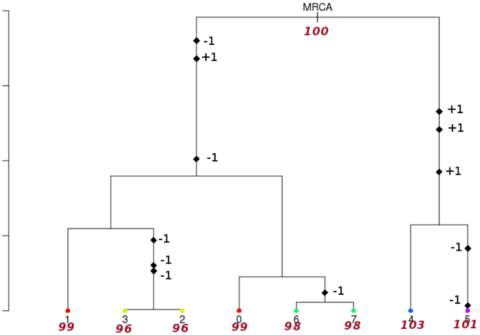
Kingman's genealogy

When time axis is normalized, $T(k) \sim \mathrm{Exp}(k(k-1)/2)$

Mutations according to the Simple stepwise Mutation Model (SMM)

- date of the mutations $\sim$ Poisson process with intensity $\theta/2$ over the branches
- MRCA $= 100$
- independent mutations: ± 1 with pr. $1/2$

Kingman's colaescent



Observations: leafs of the tree

$$\widehat{\theta} = ?$$

Kingman's genealogy

When time axis is normalized,
$T(k) \sim \mathrm{Exp}(k(k-1)/2)$

Mutations according to the Simple stepwise Mutation Model (SMM)

- date of the mutations $\sim$ Poisson process with intensity $\theta/2$ over the branches
- MRCA $= 100$
- independent mutations: ± 1 with pr. $1/2$

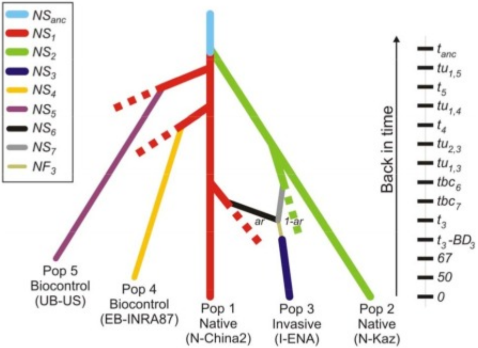
Instance of ecological questions [message in a beetle]

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

[Lombaert & al., 2010, PLoS ONE]

▸ beetles in forests

Worldwide invasion routes of *Harmonia Axyridis*



[Estoup et al., 2012, Molecular Ecology Res.]

ⓒ Intractable likelihood

Missing (too much missing!) data structure:

$$f(\mathbf{y}|\theta) = \int_{\mathcal{G}} f(\mathbf{y}|G,\theta) f(G|\theta) \mathrm{d}G$$

cannot be computed in a manageable way...

[Stephens & Donnelly, 2000]

The genealogies are considered as nuisance parameters

This modelling clearly differs from the phylogenetic perspective where the tree is the parameter of interest.

ⓒ Intractable likelihood

Missing (too much missing!) data structure:

$$f(\mathbf{y}|\theta) = \int_{\mathcal{G}} f(\mathbf{y}|G,\theta) f(G|\theta) \mathrm{d}G$$

cannot be computed in a manageable way...

[Stephens & Donnelly, 2000]

The genealogies are considered as nuisance parameters

> *This modelling clearly differs from the phylogenetic perspective where the tree is the parameter of interest.*

A?B?C?

- A stands for approximate [wrong likelihood / picture]
- B stands for Bayesian
- C stands for computation [producing a parameter sample]

ABC methodology

Bayesian setting: target is $\pi(\theta)f(x|\theta)$

When likelihood $f(x|\theta)$ not in closed form, likelihood-free rejection technique:

Foundation

For an observation $y \sim f(y|\theta)$, under the prior $\pi(\theta)$, if one keeps *jointly* simulating

$$\theta' \sim \pi(\theta)\,, z \sim f(z|\theta')\,,$$

until the auxiliary variable z is equal to the observed value, $z = y$, then the selected

$$\theta' \sim \pi(\theta|y)$$

[Rubin, 1984; Diggle & Gratton, 1984; Tavaré et al., 1997]

ABC methodology

Bayesian setting: target is $\pi(\theta)f(x|\theta)$
When likelihood $f(x|\theta)$ not in closed form, likelihood-free rejection technique:

Foundation

For an observation $\mathbf{y} \sim f(\mathbf{y}|\theta)$, under the prior $\pi(\theta)$, if one keeps *jointly* simulating

$$\theta' \sim \pi(\theta)\,, \mathbf{z} \sim f(\mathbf{z}|\theta')\,,$$

until the auxiliary variable $\mathbf{z}$ is equal to the observed value, $\mathbf{z} = \mathbf{y}$, then the selected

$$\theta' \sim \pi(\theta|\mathbf{y})$$

[Rubin, 1984; Diggle & Gratton, 1984; Tavaré et al., 1997]

A as A...pproximative

When y is a continuous random variable, strict equality $\mathbf{z} = \mathbf{y}$ is replaced with a tolerance zone

$$\rho(\mathbf{y}, \mathbf{z}) \leqslant \epsilon$$

where ρ is a distance

Output distributed from

$$\pi(\theta)\, P_\theta\{\rho(\mathbf{y}, \mathbf{z}) < \epsilon\} \overset{\text{def}}{\propto} \pi(\theta|\rho(\mathbf{y}, \mathbf{z}) < \epsilon)$$

[Pritchard et al., 1999]

A as A...pproximative

When y is a continuous random variable, strict equality $\boldsymbol{z} = \mathbf{y}$ is replaced with a tolerance zone

$$\rho(\mathbf{y}, \boldsymbol{z}) \leqslant \epsilon$$

where ρ is a distance
Output distributed from

$$\pi(\theta)\, P_\theta\{\rho(\mathbf{y}, \boldsymbol{z}) < \epsilon\} \overset{\text{def}}{\propto} \pi(\theta|\rho(\mathbf{y}, \boldsymbol{z}) < \epsilon)$$

[Pritchard et al., 1999]

ABC algorithm

In most implementations, further degree of A...pproximation:

Algorithm 1 Likelihood-free rejection sampler

```
for $i = 1$ to $N$ do
    repeat
        generate $\theta'$ from the prior distribution $\pi(\cdot)$
        generate $\mathbf{z}$ from the likelihood $f(\cdot|\theta')$
    until $\rho\{\eta(\mathbf{z}), \eta(\mathbf{y})\} \leqslant \epsilon$
    set $\theta_i = \theta'$
end for
```

where $\eta(\mathbf{y})$ defines a (not necessarily sufficient) statistic

ABC recap

Likelihood free rejection sampling

Tavaré et al. (1997) Genetics

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

Only keep θ's such that the distance between the corresponding simulated dataset and the observed dataset is small enough.

Tuning parameters

- $\epsilon > 0$: tolerance level,
- $\eta(z)$: function that summarizes datasets,
- $\rho(\eta, \eta')$: distance between vectors of summary statistics
- N: size of the output

ABC recap

Likelihood free rejection sampling

Tavaré et al. (1997) Genetics

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

Only keep θ's such that the distance between the corresponding simulated dataset and the observed dataset is small enough.

Tuning parameters

- $\epsilon > 0$: tolerance level,
- $\eta(z)$: function that summarizes datasets,
- $\rho(\eta, \eta')$: distance between vectors of summary statistics
- N: size of the output

Output

The likelihood-free algorithm samples from the marginal in z of:

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

where $A_{\epsilon,\mathbf{y}} = \{z \in \mathcal{D}|\rho(\eta(z),\eta(\mathbf{y})) < \epsilon\}$.

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

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

Output

The likelihood-free algorithm samples from the marginal in $\boldsymbol{z}$ of:

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

where $A_{\epsilon,\mathbf{y}} = \{\boldsymbol{z} \in \mathcal{D} | \rho(\eta(\boldsymbol{z}), \eta(\mathbf{y})) < \epsilon\}$.

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

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

Output

The likelihood-free algorithm samples from the marginal in z of:

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

where $A_{\epsilon,\mathbf{y}} = \{\mathbf{z} \in \mathcal{D} | \rho(\eta(\mathbf{z}), \eta(\mathbf{y})) < \epsilon\}$.

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

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

Not so good..!

Comments

- Role of distance paramount (because $\epsilon \neq 0$)
- Scaling of components of $\eta(\mathbf{y})$ is also determinant
- ϵ matters little if “small enough"
- representative of “curse of dimensionality"
- small is beautiful!
- the data as a whole may be paradoxically weakly informative for ABC

▸ how approximative is ABC?

Simulating from the prior is often poor in efficiency

Either modify the proposal distribution on θ to increase the density of x's within the vicinity of y...

[Marjoram et al, 2003; Bortot et al., 2007, Sisson et al., 2007]

...or by viewing the problem as a conditional density estimation and by developing techniques to allow for larger ϵ

[Beaumont et al., 2002]

.....or even by including ϵ in the inferential framework [ABC$_\mu$]

[Ratmann et al., 2009]

ABC (simul') advances

how approximative is ABC?

Simulating from the prior is often poor in efficiency
Either modify the proposal distribution on θ to increase the density of x's within the vicinity of y...

[Marjoram et al, 2003; Bortot et al., 2007, Sisson et al., 2007]

...or by viewing the problem as a conditional density estimation and by developing techniques to allow for larger ϵ

[Beaumont et al., 2002]

.....or even by including ϵ in the inferential framework [ABC$_\mu$]

[Ratmann et al., 2009]

ABC (simul') advances

▸ how approximative is ABC?

Simulating from the prior is often poor in efficiency
Either modify the proposal distribution on θ to increase the density of x's within the vicinity of y...

[Marjoram et al, 2003; Bortot et al., 2007, Sisson et al., 2007]

...or by viewing the problem as a conditional density estimation and by developing techniques to allow for larger ϵ

[Beaumont et al., 2002]

.....or even by including ϵ in the inferential framework [ABC$_\mu$]

[Ratmann et al., 2009]

ABC (simul') advances

▸ how approximative is ABC?

Simulating from the prior is often poor in efficiency
Either modify the proposal distribution on θ to increase the density of x's within the vicinity of y...

[Marjoram et al, 2003; Bortot et al., 2007, Sisson et al., 2007]

...or by viewing the problem as a conditional density estimation and by developing techniques to allow for larger ϵ

[Beaumont et al., 2002]

.....or even by including ϵ in the inferential framework [ABC$_\mu$]

[Ratmann et al., 2009]

ABC as knn

[Biau et al., 2013, Annales de l'IHP]

Practice of ABC: determine tolerance ϵ as a quantile on observed distances, say 10% or 1% quantile,

$$\epsilon = \epsilon_N = q_\alpha(d_1, \ldots, d_N)$$

- Interpretation of ε as nonparametric bandwidth only approximation of the actual practice
  [Blum & François, 2010]
- ABC is a k-nearest neighbour (knn) method with $k_N = N\epsilon_N$
  [Loftsgaarden & Quesenberry, 1965]

ABC as knn

[Biau et al., 2013, Annales de l'IHP]

Practice of ABC: determine tolerance ϵ as a quantile on observed distances, say 10% or 1% quantile,

$$\epsilon = \epsilon_N = q_\alpha(d_1, \ldots, d_N)$$

- Interpretation of ε as nonparametric bandwidth only approximation of the actual practice
  [Blum & François, 2010]

- ABC is a k-nearest neighbour (knn) method with $k_N = N\epsilon_N$
  [Loftsgaarden & Quesenberry, 1965]

ABC as knn

[Biau et al., 2013, Annales de l'IHP]

Practice of ABC: determine tolerance ϵ as a quantile on observed distances, say 10% or 1% quantile,

$$\epsilon = \epsilon_N = q_\alpha(d_1, \ldots, d_N)$$

- Interpretation of ε as nonparametric bandwidth only approximation of the actual practice
  [Blum & François, 2010]

- ABC is a k-nearest neighbour (knn) method with $k_N = N\epsilon_N$
  [Loftsgaarden & Quesenberry, 1965]

Which summary?

Fundamental difficulty of the choice of the summary statistic when there is no non-trivial sufficient statistics [except when done by the experimenters in the field]

- Loss of statistical information balanced against gain in data roughening
- Approximation error and information loss remain unknown
- Choice of statistics induces choice of distance function towards standardisation

- may be imposed for external/practical reasons (e.g., DIYABC)
- may gather several non-B point estimates [the more the merrier]
- can [machine-]learn about efficient combination

Which summary?

Fundamental difficulty of the choice of the summary statistic when there is no non-trivial sufficient statistics [except when done by the experimenters in the field]

- Loss of statistical information balanced against gain in data roughening
- Approximation error and information loss remain unknown
- Choice of statistics induces choice of distance function towards standardisation

- may be imposed for external/practical reasons (e.g., DIYABC)
- may gather several non-B point estimates [the more the merrier]
- can [machine-]learn about efficient combination

Which summary?

Fundamental difficulty of the choice of the summary statistic when there is no non-trivial sufficient statistics [except when done by the experimenters in the field]

- Loss of statistical information balanced against gain in data roughening
- Approximation error and information loss remain unknown
- Choice of statistics induces choice of distance function towards standardisation

- may be imposed for external/practical reasons (e.g., DIYABC)
- may gather several non-B point estimates [the more the merrier]
- can [machine-]learn about efficient combination

attempts at summaries

How to choose the set of summary statistics?

- Joyce and Marjoram (2008, SAGMB)
- Fearnhead and Prangle (2012, JRSS B)
- Ratmann et al. (2012, PLOS Comput. Biol)
- Blum et al. (2013, Statistical Science)
- LDA selection of Estoup & al. (2012, Mol. Ecol. Res.)

Semi-automatic ABC

Fearnhead and Prangle (2012) [FP] study ABC and selection of summary statistics for parameter estimation

- ABC considered as inferential method and calibrated as such
- randomised (or ‘noisy’) version of the summary statistics

$$\tilde{\eta}(\mathbf{y}) = \eta(\mathbf{y}) + \tau\epsilon$$

- *optimality* of the posterior expectation

$$\mathbb{E}[\theta|\mathbf{y}]$$

of the parameter of interest as summary statistics $\eta(\mathbf{y})$!

ABC for model choice

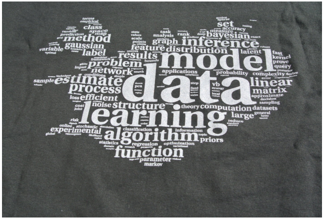
Intractable likelihoods

ABC methods

ABC for model choice

- Formalised framework

ABC model choice via random forests



Generic ABC for model choice

Algorithm 2 Likelihood-free model choice sampler (ABC-MC)

```
for $t = 1$ to $T$ do
    repeat
        Generate $m$ from the prior $\pi(\mathcal{M} = m)$
        Generate $\theta_m$ from the prior $\pi_m(\theta_m)$
        Generate $z$ from the model $f_m(z|\theta_m)$
    until $\rho\{\eta(z), \eta(\mathbf{y})\} < \epsilon$
    Set $m^{(t)} = m$ and $\theta^{(t)} = \theta_m$
end for
```

[Grelaud & al., 2009; Toni & al., 2009]

ABC model choice

ABC model choice

A) Generate large set of (m, θ, z) from the Bayesian predictive, $\pi(m)\pi_m(\theta)f_m(z|\theta)$

B) Keep particles (m, θ, z) such that $\rho(\eta(y), \eta(z)) \leqslant \epsilon$

C) For each m, return $\widehat{p_m}$ = proportion of m among remaining particles

If ε tuned towards k resulting particles, then $\widehat{p_m}$ k-**nearest neighbor** estimate of

$$\mathbb{P}\Big(\{\mathcal{M} = m\}\Big|\eta(y)\Big)$$

Approximating posterior prob's of models = regression problem where

- response is $\mathbb{1}\{\mathcal{M} = m\}$,
- covariates are summary statistics $\eta(z)$,
- loss is, e.g., L^2

Method of choice in **DIYABC** is local polytomous logistic regression

Machine learning perspective [paradigm shift]

ABC model choice

A) `Generate a large set of` (m, θ, z)`'s from Bayesian predictive,` $\pi(m)\pi_m(\theta)f_m(z|\theta)$

B) `Use machine learning tech. to infer on` $\arg\max_m \pi(m|\eta(y))$

In this perspective:

- (iid) "data set" reference table simulated during stage A)
- observed y becomes a new data point

Note that:

- predicting m is a classification problem $\iff$ select the best model based on a maximal a posteriori rule
- computing $\pi(m|\eta(y))$ is a regression problem $\iff$ confidence in each model

Ⓒ classification is much simpler than regression (e.g., dim. of objects we try to learn)

Warning

the lost of information induced by using non sufficient summary statistics is a genuine problem

Fundamental discrepancy between the genuine Bayes factors/posterior probabilities and the Bayes factors based on summary statistics. See, e.g.,

- Didelot et al. (2011, Bayesian analysis)
- X et al. (2011, PNAS)
- Marin et al. (2014, JRSS B)
- . . .

Call instead for machine learning approach able to handle with a large number of correlated summary statistics:

random forests well suited for that task

A stylised problem

Central question to the validation of ABC for model choice:

When is a Bayes factor based on an insufficient statistic $T(\mathfrak{y})$ consistent?

Note/warnin: © drawn on $T(\mathfrak{y})$ through $B_{12}^{T}(\mathfrak{y})$ necessarily differs from © drawn on $\mathfrak{y}$ through $B_{12}(\mathfrak{y})$

[Marin, Pillai, X, & Rousseau, JRSS B, 2013]

A stylised problem

Central question to the validation of ABC for model choice:

When is a Bayes factor based on an insufficient statistic $\boldsymbol{T}(\mathfrak{y})$ consistent?

Note/warnin: © drawn on $\boldsymbol{T}(\mathfrak{y})$ through $B_{12}^{\boldsymbol{T}}(\mathfrak{y})$ necessarily differs from © drawn on $\mathfrak{y}$ through $B_{12}(\mathfrak{y})$

[Marin, Pillai, X, & Rousseau, JRSS B, 2013]

A benchmark if toy example

Comparison suggested by referee of PNAS paper [thanks!]:

[X, Cornuet, Marin, & Pillai, Aug. 2011]

Model $\mathfrak{M}_1$: $\mathbf{y} \sim \mathcal{N}(\theta_1, 1)$ opposed to model $\mathfrak{M}_2$: $\mathbf{y} \sim \mathcal{L}(\theta_2, 1/\sqrt{2})$, Laplace distribution with mean θ_2 and scale parameter $1/\sqrt{2}$ (variance one).

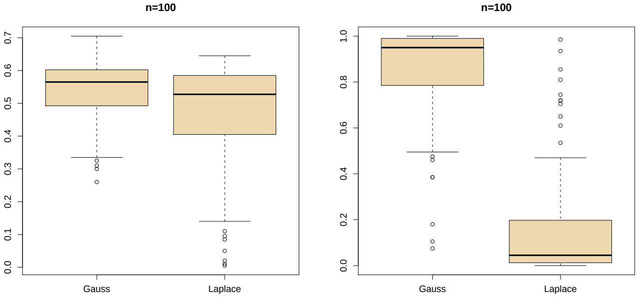
Four possible statistics

1. sample mean $\overline{\mathbf{y}}$ (sufficient for $\mathfrak{M}_1$ if not $\mathfrak{M}_2$);
2. sample median $\text{med}(\mathbf{y})$ (insufficient);
3. sample variance $\text{var}(\mathbf{y})$ (ancillary);
4. median absolute deviation $\text{mad}(\mathbf{y}) = \text{med}(|\mathbf{y} - \text{med}(\mathbf{y})|)$;

A benchmark if toy example

Comparison suggested by referee of PNAS paper [thanks!]:
[X, Cornuet, Marin, & Pillai, Aug. 2011]

Model $\mathfrak{M}_1$: $\mathbf{y} \sim \mathcal{N}(\theta_1, 1)$ opposed to model $\mathfrak{M}_2$: $\mathbf{y} \sim \mathcal{L}(\theta_2, 1/\sqrt{2})$, Laplace distribution with mean θ_2 and scale parameter $1/\sqrt{2}$ (variance one).



Framework

▸ move to random forests

Starting from sample

$$\mathfrak{y} = (y_1, \ldots, y_n)$$

the observed sample, not necessarily iid with *true* distribution

$$\mathfrak{y} \sim \mathfrak{P}^n$$

Summary statistics

$$\boldsymbol{T}(\mathfrak{y}) = \boldsymbol{T}^n = (T_1(\mathfrak{y}), T_2(\mathfrak{y}), \cdots, T_d(\mathfrak{y})) \in \mathbb{R}^d$$

with *true* distribution $\boldsymbol{T}^n \sim G_n$.

Framework

▸ move to random forests

Ⓒ Comparison of

- under $\mathfrak{M}_1$, $\mathbf{y} \sim F_{1,n}(\cdot|\theta_1)$ where $\theta_1 \in \Theta_1 \subset \mathbb{R}^{p_1}$
- under $\mathfrak{M}_2$, $\mathbf{y} \sim F_{2,n}(\cdot|\theta_2)$ where $\theta_2 \in \Theta_2 \subset \mathbb{R}^{p_2}$

turned into

- under $\mathfrak{M}_1$, $\boldsymbol{T}(\mathbf{y}) \sim G_{1,n}(\cdot|\theta_1)$, and $\theta_1|\boldsymbol{T}(\mathbf{y}) \sim \pi_1(\cdot|\boldsymbol{T}^n)$
- under $\mathfrak{M}_2$, $\boldsymbol{T}(\mathbf{y}) \sim G_{2,n}(\cdot|\theta_2)$, and $\theta_2|\boldsymbol{T}(\mathbf{y}) \sim \pi_2(\cdot|\boldsymbol{T}^n)$

Checking for adequate statistics

Run a practical check of the relevance (or non-relevance) of $\boldsymbol{T}^n$

null hypothesis that both models are compatible with the statistic $\boldsymbol{T}^n$

$$H_0 : \inf\{|\mu_2(\theta_2) - \mu_0|; \theta_2 \in \Theta_2\} = 0$$

against

$$H_1 : \inf\{|\mu_2(\theta_2) - \mu_0|; \theta_2 \in \Theta_2\} > 0$$

testing procedure provides estimates of mean of $\boldsymbol{T}^n$ under each model and checks for equality

Checking in practice

- Under each model $\mathfrak{M}_i$, generate ABC sample $\theta_{i,l}, l = 1, \cdots, L$
- For each $\theta_{i,l}$, generate $\mathbf{y}_{i,l} \sim F_{i,n}(\cdot|\psi_{i,l})$, derive $\boldsymbol{T}^n(\mathbf{y}_{i,l})$ and compute

$$\hat{\mu}_i = \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{T}^n(\mathbf{y}_{i,l}), \quad i = 1, 2\,.$$

- Conditionally on $\boldsymbol{T}^n(\mathbf{y})$,

$$\sqrt{L}\{\hat{\mu}_i - \mathbb{E}^\pi[\mu_i(\theta_i)|\boldsymbol{T}^n(\mathbf{y})]\} \rightsquigarrow \mathcal{N}(0, V_i),$$

- Test for a common mean

$$H_0 : \hat{\mu}_1 \sim \mathcal{N}(\mu_0, V_1)\,, \hat{\mu}_2 \sim \mathcal{N}(\mu_0, V_2)$$

against the alternative of different means

$$H_1 : \hat{\mu}_i \sim \mathcal{N}(\mu_i, V_i), \qquad \text{with } \mu_1 \neq \mu_2\,.$$

ABC model choice via random forests

Intractable likelihoods

ABC methods

ABC for model choice

ABC model choice via random forests

- Random forests
- ABC with random forests
- Illustrations

Leaning towards machine learning

Main notions:

- ABC-MC seen as learning about which model is most appropriate from a huge (reference) table
- exploiting a large number of summary statistics not an issue for 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, and replacing those by posterior predictive expected loss~~
- estimating posterior probabilities of the selected model 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]

Improved classification performances by averaging over classification schemes of randomly generated training sets, creating a "forest" of (CART) decision trees, inspired by Amit and Geman (1997) ensemble learning

[Breiman, 2001]

CART construction

Basic classification tree:

Algorithm 3 CART

```
start the tree with a single root
repeat
    pick a non-homogeneous tip $v$ such that $Q(v) \neq 1$
    attach to $v$ two daughter nodes $v_1$ and $v_2$
    for all covariates $X_j$ do
        find the threshold $t_j$ in the rule $X_j < t_j$ that minimizes $N(v_1)Q(v_1) + N(v_2)Q(v_2)$
    end for
    find the rule $X_j < t_j$ that minimizes $N(v_1)Q(v_1) + N(v_2)Q(v_2)$ in $j$ and set this best rule to node $v$
until all tips $v$ are homogeneous ($Q(v) = 0$)
set the labels of all tips
```

where Q is Gini's index

$$Q(v_i) = \sum_{y=1}^{M} \hat{p}(v_i, y)\{1 - \hat{p}(v, y)\} \; .$$

Growing the forest

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

- boostrap resampling of the dataset and
- random subset-ing [of size $\sqrt{t}$] of the covariates driving the classification at every node of each tree

Covariate x_τ that drives the node separation

$$x_\tau \gtrless c_\tau$$

and the separation bound c_τ chosen by minimising entropy or Gini index

Breiman and Cutler's algorithm

Algorithm 4 Random forests

```
for $t = 1$ to $T$ do
    //*$T$ is the number of trees*//
    Draw a bootstrap sample of size $n_{\text{boot}} \neq n$
    Grow an unpruned decision tree
    for $b = 1$ to $B$ do
        //*$B$ is the number of nodes*//
        Select $n_{\text{try}}$ of the predictors at random
        Determine the best split from among those predictors
    end for
end for
Predict new data by aggregating the predictions of the $T$ trees
```

ABC with random forests

Idea: Starting with

- possibly large collection of summary statistics $(s_{1i}, \ldots, s_{pi})$ (from scientific theory input to available statistical softwares, to machine-learning alternatives, to pure noise)
- ABC reference table involving model index, parameter values and summary statistics for the associated simulated pseudo-data

run R randomforest to infer $\mathfrak{M}$ from $(s_{1i}, \ldots, s_{pi})$

ABC with random forests

Idea: Starting with

- possibly large collection of summary statistics $(s_{1i}, \ldots, s_{pi})$ (from scientific theory input to available statistical softwares, to machine-learning alternatives, to pure noise)
- ABC reference table involving model index, parameter values and summary statistics for the associated simulated pseudo-data

run R randomforest to infer $\mathfrak{M}$ from $(s_{1i}, \ldots, s_{pi})$

at each step $\mathrm{O}(\sqrt{p})$ indices sampled at random and most discriminating statistic selected, by minimising ~~entropy~~ Gini loss

ABC with random forests

Idea: Starting with

- possibly large collection of summary statistics $(s_{1i}, \ldots, s_{pi})$ (from scientific theory input to available statistical softwares, to machine-learning alternatives, to pure noise)
- ABC reference table involving model index, parameter values and summary statistics for the associated simulated pseudo-data

run R randomforest to infer $\mathfrak{M}$ from $(s_{1i}, \ldots, s_{pi})$

Average of the trees is resulting summary statistics, highly non-linear predictor of the model index

Outcome of ABC-RF

Random forest predicts a (MAP) model index, from the observed dataset: The predictor provided by the forest is “sufficient" 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
- usual ABC-MC approximation equally highly variable and hard to assess
- random forests define a natural distance for ABC sample via agreement frequency

Outcome of ABC-RF

Random forest predicts a (MAP) model index, from the observed dataset: The predictor provided by the forest is “sufficient" 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
- usual ABC-MC approximation equally highly variable and hard to assess
- random forests define a natural distance for ABC sample via agreement frequency

Posterior predictive expected losses

We suggest replacing unstable approximation of

$$\mathbb{P}(\mathfrak{M} = m | x_o)$$

with x_o observed sample and m model index, by average of the selection errors across all models given the data x_o,

$$\mathbb{P}(\widehat{\mathfrak{M}}(X) \neq \mathfrak{M} | x_o)$$

where pair $(\mathfrak{M}, X)$ generated from the predictive

$$\int f(x|\theta) \pi(\theta, \mathfrak{M} | x_o) \mathrm{d}\theta$$

and $\widehat{\mathfrak{M}}(x)$ denotes the random forest model (MAP) predictor

Posterior predictive expected losses

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

Posterior probability of the selected model

Given the MAP estimate provided by the random forest, $\widehat{\mathfrak{M}}(s(X))$, consider the posterior estimation error

$$\begin{aligned}
\mathbb{E}[\mathbb{I}(\widehat{\mathfrak{M}}(\mathbf{s}_{\text{obs}}) \neq \mathfrak{M})|\mathbf{s}_{\text{obs}}] &= \sum_{i=1}^{k} \mathbb{E}[\mathbb{I}(\widehat{\mathfrak{M}}(\mathbf{s}_{\text{obs}}) \neq \mathfrak{M} = i)|\mathbf{s}_{\text{obs}}] \\
&= \sum_{i=1}^{k} \mathbb{P}[\mathfrak{M} = i)|\mathbf{s}_{\text{obs}}] \times \mathbb{I}(\widehat{\mathfrak{M}}(\mathbf{s}_{\text{obs}}) \neq i) \\
&= \mathbb{P}[\mathfrak{M} \neq \widehat{\mathfrak{M}}(\mathbf{s}_{\text{obs}})|\mathbf{s}_{\text{obs}}] \\
&= 1 - \mathbb{P}[\mathfrak{M} = \widehat{\mathfrak{M}}(\mathbf{s}_{\text{obs}})|\mathbf{s}_{\text{obs}}] \,,
\end{aligned}$$

© posterior probability that the true model is not the MAP

Posterior probability of the selected model

Given the MAP estimate provided by the random forest, $\widehat{\mathfrak{M}}(s(X))$, consider the posterior estimation error

$$
\begin{aligned}
\mathbb{E}[\mathbb{I}(\widehat{\mathfrak{M}}(\mathbf{s}_{\text{obs}}) \neq \mathfrak{M})|\mathbf{s}_{\text{obs}}] &= \sum_{i=1}^{k} \mathbb{E}[\mathbb{I}(\widehat{\mathfrak{M}}(\mathbf{s}_{\text{obs}}) \neq \mathfrak{M} = i)|\mathbf{s}_{\text{obs}}] \\
&= \sum_{i=1}^{k} \mathbb{P}[\mathfrak{M} = i)|\mathbf{s}_{\text{obs}}] \times \mathbb{I}(\widehat{\mathfrak{M}}(\mathbf{s}_{\text{obs}}) \neq i) \\
&= \mathbb{P}[\mathfrak{M} \neq \widehat{\mathfrak{M}}(\mathbf{s}_{\text{obs}})|\mathbf{s}_{\text{obs}}] \\
&= 1 - \mathbb{P}[\mathfrak{M} = \widehat{\mathfrak{M}}(\mathbf{s}_{\text{obs}})|\mathbf{s}_{\text{obs}}] \,,
\end{aligned}
$$

© posterior probability that the true model is not the MAP

Posterior probability estimated by another forest

- since

$$\mathbb{P}[\mathfrak{M} \neq \hat{\mathfrak{M}}(\mathbf{s}_{\text{obs}})|\mathbf{s}_{\text{obs}}] = \mathbb{E}[\mathbb{I}(\hat{\mathfrak{M}}(\mathbf{s}_{\text{obs}}) \neq \mathfrak{M})|\mathbf{s}_{\text{obs}}]$$

  function of $\mathbf{s}_{\text{obs}}$, $\Psi(\mathbf{s}_{\text{obs}})$, ...

- ...estimation based on the reference table simulated from prior predictive, using all simulated pairs $(\mathfrak{M}, s)$
- construction of a random forest $\widehat{\Psi}(s)$ predicting the error $\mathbb{E}[\mathbb{I}(\hat{\mathfrak{M}}(s) \neq \mathfrak{M})|s]$
- association of $\widehat{\Psi}(\mathbf{s}_{\text{obs}})$ with $\widehat{\mathfrak{M}}(\mathbf{s}_{\text{obs}})$

Algorithmic implementation

Algorithm 5 Approximation of the posterior probability

(a) Use the trained RF to predict model by $\widehat{\mathfrak{M}}(S(\boldsymbol{x}))$ for each $(m, S(\boldsymbol{x}))$ in the reference table and deduce $\iota = \mathbb{I}(\widehat{\mathfrak{M}}(s) \neq \mathfrak{M})$

(b) Train a new RF $\widehat{\Psi}(s)$ on this reference table (ι, s) predicting success $\Psi(s)$

(c) Apply to $s = \mathbf{s}_{\text{obs}}$ and deduce $\widehat{\Psi}(\mathbf{s}_{\text{obs}})$ as estimate of $\mathbb{P}[\mathfrak{M} = \widehat{\mathfrak{M}}(\mathbf{s}_{\text{obs}})|\mathbf{s}_{\text{obs}}]$

toy: MA(1) vs. MA(2)

Comparing an MA(1) and an MA(2) models:

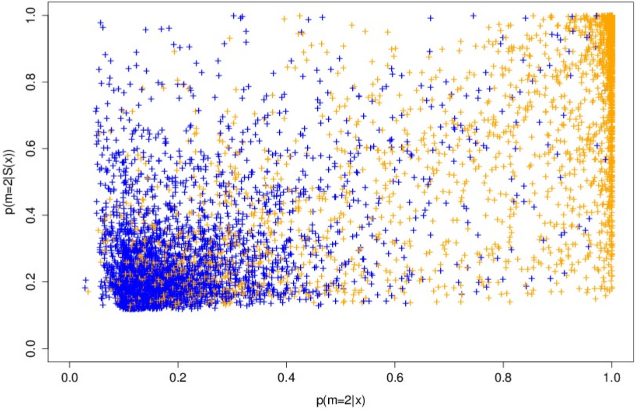
$$x_t = \epsilon_t - \vartheta_1 \epsilon_{t-1} [-\vartheta_2 \epsilon_{t-2}]$$

Earlier illustration using first two autocorrelations as $S(x)$

[Marin et al., Stat. & Comp., 2011]

Result #1: values of $p(m|x)$ [obtained by numerical integration] and $p(m|S(x))$ [obtained by mixing ABC outcome and density estimation] highly differ!

toy: MA(1) vs. MA(2)



Difference between the posterior probability of $MA(2)$ given either x or $S(x)$. Blue stands for data from $MA(1)$, orange for data from $MA(2)$

toy: MA(1) vs. MA(2)

Comparing an MA(1) and an MA(2) models:

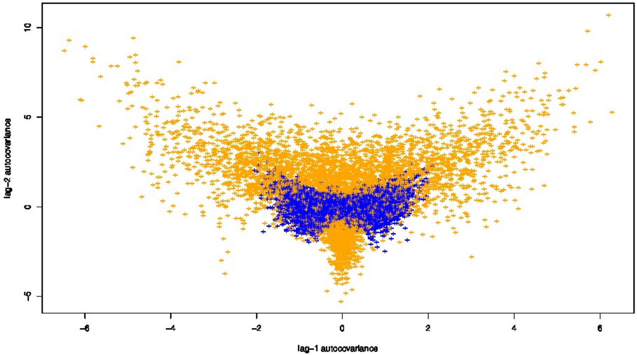
$$x_t = \epsilon_t - \vartheta_1 \epsilon_{t-1} [-\vartheta_2 \epsilon_{t-2}]$$

Earlier illustration using two autocorrelations as $S(x)$

[Marin et al., Stat. & Comp., 2011]

Result #2: Embedded models, with simulations from MA(1) within those from MA(2), hence linear classification poor

toy: MA(1) vs. MA(2)



Simulations of $S(x)$ under $MA(1)$ (blue) and $MA(2)$ (orange)

toy: MA(1) vs. MA(2)

Comparing an MA(1) and an MA(2) models:

$$x_t = \epsilon_t - \vartheta_1 \epsilon_{t-1} [-\vartheta_2 \epsilon_{t-2}]$$

Earlier illustration using two autocorrelations as $S(x)$

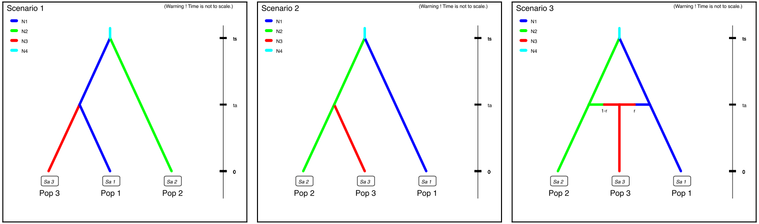
[Marin et al., Stat. & Comp., 2011]

Result #3: On such a small dimension problem, random forests should come second to k-nn ou kernel discriminant analyses

toy: MA(1) vs. MA(2)

classification **method**	**prior** **error rate** (in %)
LDA	27.43
Logist. reg.	28.34
SVM (library e1071)	17.17
“naïve” Bayes (with G marg.)	19.52
“naïve” Bayes (with NP marg.)	18.25
ABC k-nn ($k = 100$)	17.23
ABC k-nn ($k = 50$)	16.97
Local log. reg. ($k = 1000$)	16.82
Random Forest	17.04
Kernel disc. ana. (KDA)	16.95
True MAP	12.36

Evolution scenarios based on SNPs



Three scenarios for the evolution of three populations from their most common ancestor

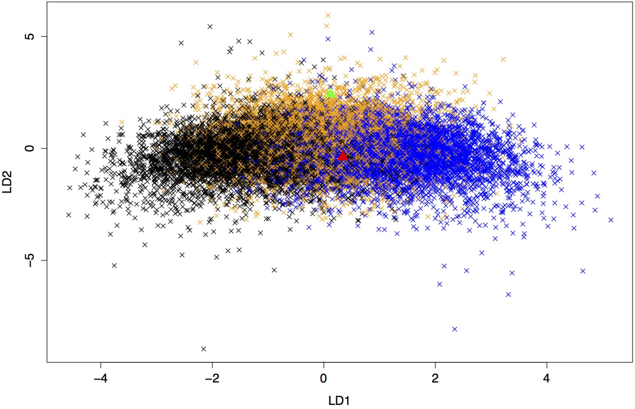
Evolution scenarios based on microsatellites

classification method	**prior error**[*] **rate** (in %)
raw LDA	35.64
“naïve” Bayes (with G marginals)	40.02
k-nn (MAD normalised sum stat)	37.47
k-nn (unormalised LDA)	35.14
RF without LDA components	35.14
RF with LDA components	33.62
RF with only LDA components	37.25

[*] estimated on pseudo-samples of 10^4 items drawn from the prior

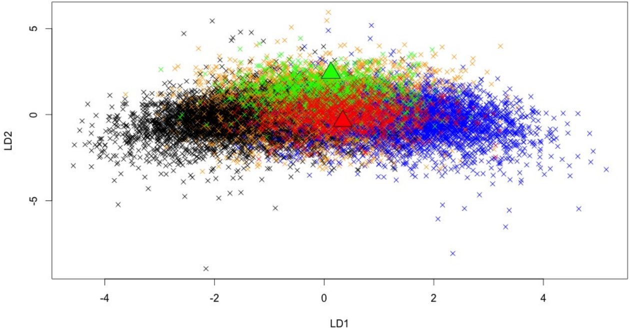
Evolution scenarios based on microsatellites

Posterior predictive error rates



Evolution scenarios based on microsatellites

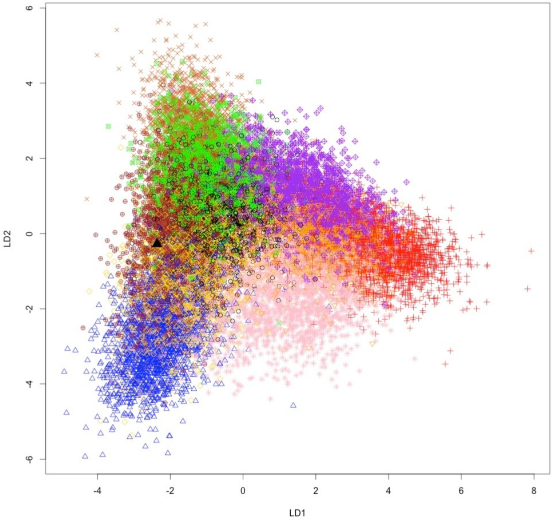
Posterior predictive error rates



favourable: 0.183 error – unfavourable: 0.435 error

Back to Asian Ladybirds [message in a beetle]

Comparing 10 scenarios of Asian beetle invasion ◂ beetle moves



Back to Asian Ladybirds [message in a beetle]

Comparing 10 scenarios of Asian beetle invasion ◂ beetle moves

classification method	**prior error**[†] **rate** (in %)
raw LDA	38.94
"naïve" Bayes (with G margins)	54.02
k-nn (MAD normalised sum stat)	58.47
RF without LDA components	38.84
RF with LDA components	35.32

[†] estimated on pseudo-samples of 10^4 items drawn from the prior

Back to Asian Ladybirds [message in a beetle]

Comparing 10 scenarios of Asian beetle invasion ◂ beetle moves

Random forest allocation frequencies

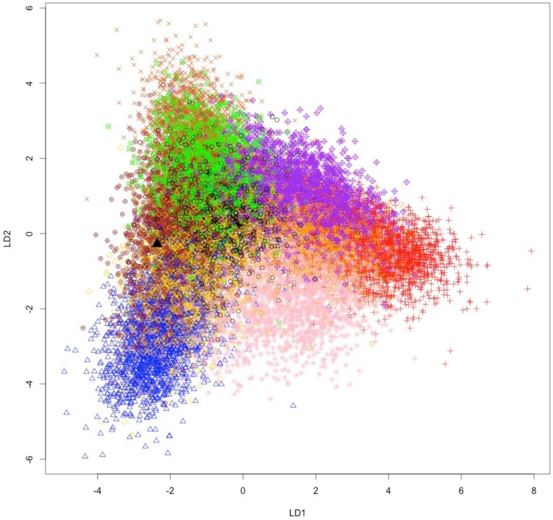
```
1     2   3     4     5     6     7     8    9     10
0.168 0.1 0.008 0.066 0.296 0.016 0.092 0.04 0.014 0.2
```

Posterior predictive error based on 20,000 prior simulations and keeping 500 neighbours (or 100 neighbours and 10 pseudo-datasets per parameter)

```
0.3682
```

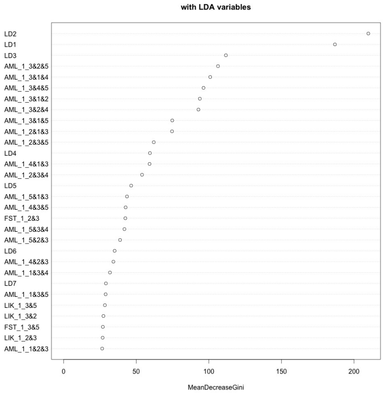
Back to Asian Ladybirds [message in a beetle]

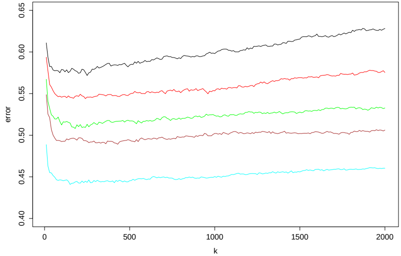
Comparing 10 scenarios of Asian beetle invasion



Back to Asian Ladybirds [message in a beetle]

Comparing 10 scenarios of Asian beetle invasion





posterior predictive error 0.368

Back to Asian Ladybirds [message in a beetle]

Harlequin ladybird data: estimated prior error rates for various classification methods and sizes of reference table.

Classification method	Prior error rates (%)		
trained on	$N_{\text{ref}} = 10,000$	$N_{\text{ref}} = 20,000$	$N_{\text{ref}} = 50,000$
linear discriminant analysis (LDA)	39.91	39.30	39.04
standard ABC (knn) on DIYABC summaries	57.46	53.76	51.03
standard ABC (knn) on LDA axes	39.18	38.46	37.91
local logistic regression on LDA axes	41.04	37.08	36.05
random forest (RF) on DIYABC summaries	40.18	38.94	37.63
RF on DIYABC summaries and LDA axes	36.86	35.62	34.44

Conclusion

Key ideas

- $\pi(m|\eta(\mathbf{y})) \neq \pi(m|\mathbf{y})$
- Rather than approximating $\pi(m|\eta(\mathbf{y}))$, **focus on selecting the best model** (classif. vs regression)
- Assess confidence in the selection via **posterior probability of MAP model**

Consequences on ABC-PopGen

- Often, RF $\gg$ k-NN (less sensible to high correlation in summaries)
- RF requires **many less prior simulations**
- RF selects automatically **relevant summaries**
- **Hence can handle much more complex models**

Conclusion

Key ideas

- $\pi(m|\eta(\mathbf{y})) \neq \pi(m|\mathbf{y})$
- Use **a seasoned machine learning technique** selecting from ABC simulations: minimise 0-1 loss mimics MAP
- Assess confidence in the selection via **RF estimate of posterior probability of MAP model**

Consequences on ABC-PopGen

- Often, **RF $\gg$ *k*-NN** (less sensible to high correlation in summaries)
- RF requires **many less prior simulations**
- RF incorporates **all available summaries**
- **Hence can handle much more complex models**

Further features

- unlimited aggregation of arbitrary summary statistics
- recovery of discriminant statistics when available
- automated implementation with reduced calibration
- self-evaluation by posterior ~~predictive error~~ probability
- soon to appear in DIYABC