Can Cross-Validation help to tune an ABC Algorithm?

Johanna Bertl

Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) is a simulation method to find the posterior distribution of a parameter θ of a model $\mathcal{M}$, in cases where the likelihood $p(x|\theta)$ cannot be calculated analytically.

A standard ABC algorithm consists of the following steps, which are repeated m times:

1. Simulate θ^* from the prior distibution $\pi(\theta)$.
2. Simulate data X from $\mathcal{M}(\theta^*)$.
3. Compute the summary statistic S^* from the simulated data and compare it to S, the summary statistic from the real data. If $d(S^*, S) \leq \epsilon$, keep θ^*, else, reject it.

From the θ^*, which were accepted, the posterior density can be estimated, e.g. with a kernel density estimator.

This algorithm raises two questions:

- Which summary statistic(s) S are appropriate? If no sufficient statistics exist, it is not trivial to identify the statistics that carry the most information about a parameter. High dimensional summary statistics may be very informative, but at the same time they make the proportion of accepted θ^*'s very small.
- How shall ϵ be chosen? The smaller ϵ is, the better the estimated posterior density approximates the true posterior density. In practice, however, a small ϵ means that only very few θ^* are accepted, so the stochastic noise in the estimation of the density is very high.

The Coalescent

The coalescent is a widely used stochastic process to analyse genetic variation in population genetics. It is an important area of application of ABC, because the likelihood function of the parameters, that determine the shape of the genealogical tree of n individuals in the last generation, can only be worked out with a tremendous

computational effort. Even for the usually small sample sizes of $n \leq 50$ the likelihood cannot be computed in a reasonable time.

A further aspect that makes the coalescent an interesting field for ABC is that for the parameters of a coalescent model, there usually are no sufficient statistics.

The Role of Cross-Validation

By now, as ABC is quite a "young" method, not many approaches exist to answer the open questions. My talk will give an overview of the methods that have yet been proposed.

Furthermore I will present the results of a simulation study in my master thesis, where I am currently investigating, if cross-validation can be of help to choose ϵ in a simple coalescent model. As cross-validation is a method to estimate the risk of a statistical algorithm, it may be a useful tool to choose among ABC algorithms with different values of ϵ.

Results

The application of 5-fold cross-validation and 100 times repeated 5-fold cross-validation on data which was simulated from a coalescent model showed that cross-validation might be an appropriate tool to choose ϵ. In 5-fold cross-validation the influence of random noise is slightly too strong, but the results of the 100 times repeated 5-fold cross-validation are promising and encourage further research.