Supervised by **Benjamin Guedj**

Graphs in Machine Learning
ABC and scoring

Marc Etheve
Master 2 MVA
ENS Paris-Saclay
marc.etheve@ens-paris-saclay.fr

Jean-Baptiste Remy
Master 2 MVA
ENS Paris-Saclay
jean-baptiste.remy@ens-paris-saclay.fr

Table des matières

1 Introduction

In complex models, such as graphical models, the likelihood $l(\theta|x)$ may not be available, either for analytical reasons or computational considerations. The ABC method, standing for *Approximate Bayesian Computation*, allows in these cases to simulate from an approximation of $\pi(\theta|x)$, using an approximative reject sampling algorithm. Two sources of noise are introduced, by allowing an error margin in the acceptance scheme and by diminishing the dimension of the data using a statistic.
The choice of this statistic being crucial for the quality of the simulation, we are here interested in the evaluation of a scoring statistic within the ABC framework.

Before addressing the subject of this project, it is worth noticing that it has little in common with the class "Graphs in ML". Actually, it turned out to be closer to classes such as "Computational Statistiscs" or again "Probabilistic Graphical Models".

2 Approximate Bayesian Computation

As previously mentioned, an ABC procedure aims at simulating from the posterior $\pi(\theta|x^0)$ where x^0 is an observation of a random variable X following a distribution p_θ, and assuming a prior π on θ (see [1] for a proper introduction to ABC).

We can simulate exactly from this posterior by the simple rejection sampling displayed in Algorithm 1.

Algorithm 1 Bayesian simulation as Accept-Reject Sampling

Given an observation x^0
for $t = 1$ to N do **do**
  repeat
    Generate θ^* from the prior $\pi(.)$
    Generate x^* from the model $f(.|\theta^*)$
    Accept θ^* if $x^0 = x^*$
  until acceptance
end for
return the N accepted values of θ^*

For diffuse distributions, this exact algorithm may not be appropriate, the exact acceptance implying a prohibitive simulation time. Hence we introduce a first source of approximation in rejection scheme, accepting all the simulations close enough to the observations. This is described in Algorithm 2, with ρ a chosen distance measure.

Algorithm 2 ABC (basic version)

Given an observation x^0
for $t = 1$ to N do **do**
  repeat
    Generate θ^* from the prior $\pi(.)$
    Generate x^* from the model $f(.|\theta^*)$
    Compute the distance $\rho(x^0, x^*)$
    Accept θ^* if $\rho(x^0, x^*) < \varepsilon$
  until acceptance
end for
return the N accepted values of θ^*

However, the curse of dimensionality may still imply a low rate of acceptance for this algorithm if the observation lay in a high dimensional space. Therefore, one may want to use a statistic of the

observation as a dimensionality reduction technique. This leads to Algorithm 3, which we will use as ABC procedure in the following.

Algorithm 3 ABC (version with summary)

Given an observation x^0
for $t = 1$ to N do **do**
  Generate $\theta^{(t)}$ from the prior $\pi(.)$
  Generate $x^{(t)}$ from the model $f(.|\theta^{(t)})$
  Compute $d_t = \rho(S(x^0), S(x^{(t)}))$
end for
Order distances $d_{(1)} \leq d_{(2)} \leq ... \leq d_{(N)}$
return the values $\theta^{(t)}$ associated with the k smallest distances.

According to this Algorithm (and not considering the noise induced by ε), we no longer simulate from $\pi(\theta|x^0)$ but from $\pi(\theta|S(x^0))$, the two distributions being equal if and only if S is a sufficient statistic. Unfortunately, such a statistic may not be easily available for models where an ABC procedure is needed. Therefore, we are looking for a low-dimensional but informative enough statistic.

3 Scoring

As defined in [2], a scoring function aims at ordering the observations $(x_i)_{i=1}^n,\ x_i \in \mathbb{R}^d$ in a consistent way with the labels $(y_i)_{i=1}^n,\ y_i \in \{0,1\}$. Such an estimate can for instance be obtained by minimizing the empirical ranking risk

$$L_n\ :\ s \mapsto \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{1}_{[(Y_i - Y_j)(s(X_i) - s(X_j)) < 0]}$$

with s a scoring function in $\mathcal{S}$ defined as

$$\mathcal{S} = \left\{ s_\beta\ :\ x \mapsto \sum_{j=1}^{d} \beta_j \phi(x_j),\ \beta \in \mathbb{R}^d \right\}.$$

Adopting a bayesian point of view, we assume a prior π_β on β and a pseudo-posterior

$$\hat{\eta}(\beta; x) \propto \exp\left[-\delta L_n(s_\beta(x))\right] \pi_\beta(\beta)$$

where $\delta > 0$ controls the importance of the empirical risk in the posterior, and the shape of π_β driving the sparsity of the distribution.

In this framework, to estimate a scoring function on a sample x, we just need to simulate from $\hat{\eta}$ and approximate β_x by a realisation or a simulated mean $\hat{\beta}_x$.

4 Experimental framework

4.1 The model

We want to assess the pertinence of the scoring function in the ABC scheme. In order to be fair, we hence need to compare it with a standard statistic in the considered framework. The scoring function being based on explaining variables, it has to be also the case for the benchmark. In addition, in order to compare the obtained results, we need to design an experiment in which the posterior distribution of θ is known.

Therefore, we propose here to tackle the linear bayesian regression inference problem. So let us define the model as

$$Y|\theta, X \sim \mathcal{N}\left(X\theta, \sigma^2 I_n\right)$$
$$\theta \sim \mathcal{N}\left(\mu, \tau^2 I_d\right)$$

with $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times d}$, $\theta, \mu \in \mathbb{R}^d$ and $\sigma^2, \tau^2 \in \mathbb{R}_+^*$.
A simple derivation leads to

$$\theta | X, Y \sim \mathcal{N}\left(\Lambda^{-1}\left(\frac{X^\top Y}{\sigma^2} + \frac{\mu}{\tau^2}\right), \Lambda^{-1}\right)$$

with

$$\Lambda = \frac{X^\top X}{\sigma^2} + \frac{I_d}{\tau^2}.$$

4.2 The statistics

The exercise of comparing two statistics in the ABC procedure needs to be tackled with caution. Indeed, the "level of convergence" of the simulated θ depends not only of the quantity of information contained in the statistic but also of the rejection rate associated with the distance measure (ε in Algorithm 2 or $\frac{k}{N}$ in Algorithm 3). The first criterion is actually what we want to compare, whereas the second is an undersirable bias in our experiment.
The chosen measure being the euclidian distance, this bias has two main sources. On one hand, the dimensions of the statistics matter. Indeed, euclidian distances in a higher dimensional space are more likely to be high given a distribution. On the other hand, the euclidian distance being a linear combination of squared errors, the variance on each dimension of the statistic is important, since it will be more difficult to meet a distance condition if a dimension has higher variance.
As a consequence, we force the statistics to have the same dimensions. Concerning the variance-based bias, as we cannot directly standardize the variance of the scoring function, we standardize X to prevent a dimension from driving excessively the distance measures.

These technical issues having been taken care of, we propose to compare in this model the benchmark statistic $\hat{\theta}_{reg}$ and the scoring statistic $\hat{\theta}_{sco}$ defined as

$$\hat{\theta}_{reg} = \left(X^\top X\right)^{-1} X^\top Y$$
$$\hat{\theta}_{sco} \sim \hat{\eta}(.; X)$$

considering the set of scoring function

$$\mathcal{S} = \left\{ s_\beta \ : \ x \mapsto \sum_{j=1}^{d} \beta_j x_j = \beta^\top x, \ \beta \in \mathbb{R}^d \right\}.$$

We precise here that we take an experiment in which the dictionary $\phi = Id$ is imposed in order to a have fair competition with the regression statistic. Besides, this naive scoring function is quite adapted to the model. We can observe this in Figure 1, where the scoring function associated with the observed sample is plotted against its actual value.

4.3 Sampling from the pseudo-posterior

We defined the scoring statistic as a random variable drawn from the pseudo-posterior $\hat{\eta}$. Unfortunately, this random variable cannot be easily simulated and thus we need to make use of approximate sampling. Constating the impossibility of performing inverse transform sampling but taking advantage of the fact that we have

$$\hat{\eta}(\theta; X) \leq \pi(\theta) \ \forall \ \theta \in \mathbb{R}^d,$$

we implemented the rejection sampling procedure described in Algorithm 4, which only requires the distribution up to a normalizing constant.

The procedure described in Algorithm 4 provides a decreasing acceptance rate as δ increases, since

$$\frac{\partial}{\partial \delta} \mathbb{P}\left(U \leq \frac{\exp\left[-\delta L_n(s_\theta(X))\right]}{\pi(\theta)} | X, \theta\right) \leq 0.$$

Actually this is quite unfortunate since δ drives the importance of the ranking in the posterior. This computational issue is the reason why we defined $\hat{\theta}_{sco} \sim \hat{\eta}(.; X)$ and not $\hat{\theta}_{sco} = \frac{1}{T}\sum_{t=1}^{T} \theta_{(t)}; \theta_{(t)} \overset{iid}{\sim} \hat{\eta}(.; X)$.

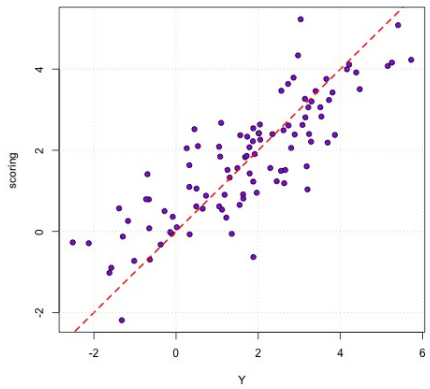

FIGURE 1 – Naive scoring on the observation sample

Algorithm 4 REJECTION SAMPLING from $\hat{\eta}(.; X)$

Generate θ from $\pi(.)$
Generate U from $\mathcal{U}_{[0,1]}$
Accept θ if $U \leq \exp\left[-\delta L_n(s_\theta(X))\right]$

5 Results

The exploration of the results have been quite painfull, due to the prohibitive computational time. It is worth noticing here that the encountered exponentially increasing time in δ (shown in minutes in Figure 2) is a direct consequence of bothe the pseudo-posterior $\hat{\eta}$ and the rejection sampling. Indeed, we could not find a better proposal than π to implement the algorithm, which time is prohibitive when δ becomes too large.

Some alternatives of rejection sampling could have been considered here. For instance, we can think of Metropolis-Hastings' procedure, or even discretizing the support of θ, performing multinomial sampling with computed probabilities proportional to $\hat{\eta}(\theta)$ and them adding continuous noise. But the sampling techniques were not the subject of the current project, so we did not investigate further these alternatives.

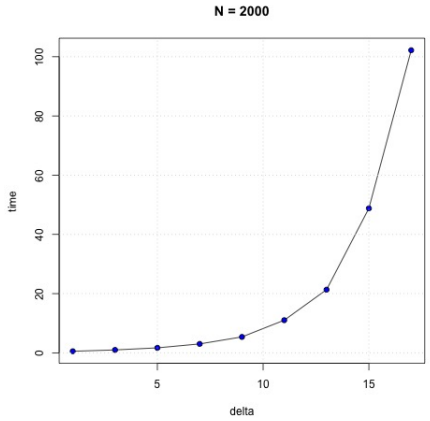

FIGURE 2 – Computational time for ABC as a function of δ

Back to the evaluation of the statistics in the ABC scheme, we need to define some metrics to compare those statistics. The goal of ABC being the simulation according to the posterior distribution of the parameter, we need to compare the obtained simulations with the true posterior, which is known in our experiment. We will then compare the posterior distributions visually and by the evolution of moments as a function of the experiment's parameters, namely the acceptance rate and δ. The optimal parameters depending on the parametrization of our dataset, the evaluation had to be made on a single dataset.

Y has been generated from $\theta_{hidden} = [-1\ 1]^\top$ and $\sigma^2 = 1$, while the prior was an unbiased conjugate prior $\mathcal{N}(\theta_{hidden}; I_d)$. We chose to center the prior around the true value to fasten the rejection sampling according to $\hat{\eta}$.

In order to dispose of as much information as possible on the observed data, we did not compute $\hat{\theta}^{obs}_{sco} \sim \hat{\eta}(.; X)$ but rather $\hat{\theta}^{obs}_{sco} = \frac{1}{100}\sum_{t=1}^{100} \theta_{(t)}$. θ_{obs} being computed a single time, we could indeed afford to increase its computation time in a negligeable way.

5.1 Comparing the posteriors

With $N = 8000$ simulations, $n = 200$ accepted θ_{sim} and $\delta = 5$, we yield the distributions showed in Figure 3.

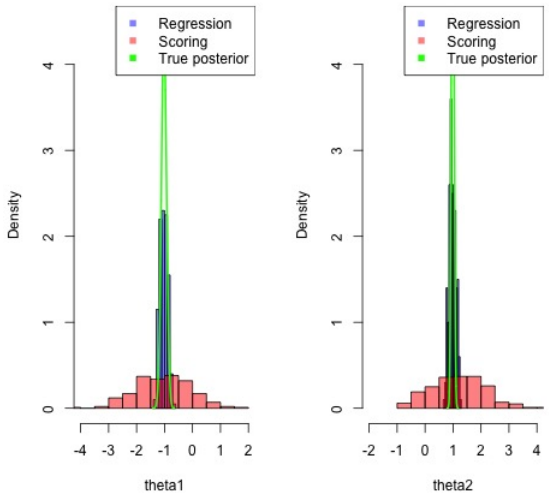

FIGURE 3 – Posterior distributions with ABC sampling

As expected, both the distributions are centered around the true expectation of the true posterior. However, we observe a much higher variance for the scoring statistic.
Nevertheless, at this stage, we cannot conclude on the quality of this statistic, the choice of $delta$ and $\varepsilon = N/n$ being arbitrary.

5.2 The acceptance rate

A way of comparing the two statistics is to estimate their convergence toward the true posterior distribution as the number of simulation goes to infinity (with a stable number of accepted samples, so with an ABC acceptance rate going to 0). We made this evaluation with a δ (=5) with relatively high acceptance rate in the rejection sampling of $\hat{\eta}$. As it can be observed in Figure 4, the regression statistic has the expected behaviour, the ABC acceptance rate driving the decrease of the posterior variance. However, we do not observe this result for the scoring statistic. As explained in the following, it may be due to the wrong scaling of the considered δ.

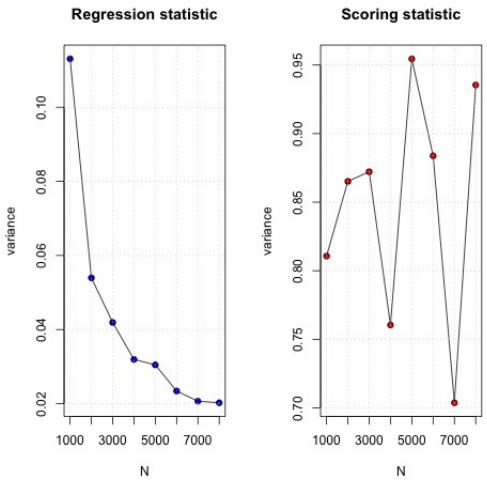

FIGURE 4 – Posterior variance as a function of N

5.3 Importance of the temperature parameter δ

Following our evaluation procedure, we now seek to evaluate the importance of the temperature parameter δ. This parameter drives the weight of the scoring function in the pseudo-posterior $\hat{\eta}$, so we expect its increase to tighten the posterior distribution around the mean, hence decreasing the variance. Reminding the computational time issue (Figure 2), we limited our experiments to a restricted grid search. Figure 5 and Figure 6 show the obtained results.

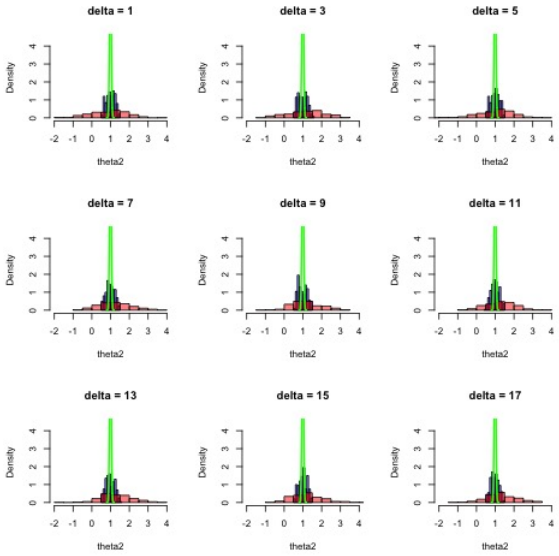

FIGURE 5 – Posterior distributions given δ

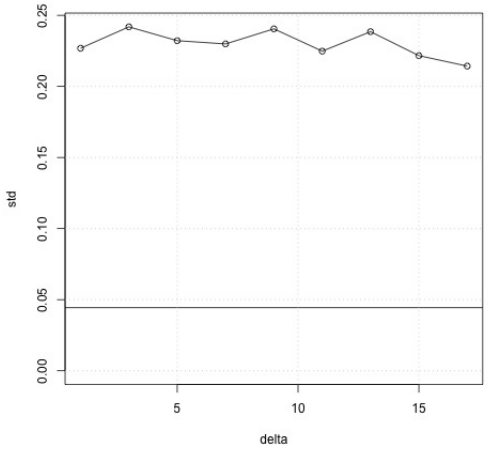

FIGURE 6 – Posterior standard error as a funtion of δ

We observe very poor improvements with the increase of δ. The low rate of decrease of the standard error let us think that for the experiment parameters, the optimal value of δ is much higher, but would require a way too important sampling time.

6 Conclusion

During this project, we have designed an experiment in which the studied scoring statistic could be compared we a standard one on a natural basis. We observed the expected behavior for the benchmark, with a converging variance and an unbiased posterior. However we could not find this result for the scoring statistic. Unfortunately, this unexpected behavior is not categorical, since the observed results tend to let us think that for (really) higher values of δ, we could observe the correct behavior.

As a consequence, we have to point out that the scoring statistic, as appealing as it may be, implies two sources of difficulty which we were not able to solve here. First, the optimal δ has to be estimated by grid search due to the nature of both the ABC and sampling schemes. The second difficulty, which is linked to the first one, is that as we get closer to the optimal value of δ (starting from 0 in the grid search), the sampling according to the pseudo-posterior $\hat{\eta}$ becomes way more time consuming. Again, this obstacle may potentially be overpassed with a better sampling procedure.

References

[1] Robert, C. P. (2016). Approximate Bayesian Computation : A Survey on Recent Results. In Monte Carlo and Quasi-Monte Carlo Methods (pp. 185-205). Springer International Publishing.

[2] Guedj, B., & Robbiano, S. (2015). PAC-Bayesian High Dimensional Bipartite Ranking. arXiv preprint arXiv :1511.02729.