AN USER FRIENDLY EMPIRICAL LIKELIHOOD ABC METHOD

Sanjay Chaudhuri
National University of Singapore

Shubhroshekhar Ghosh, David Nott and Kim Cuc Pham, National University of Singapore

APPROXIMATE BAYESIAN COMPUTATION (ABC)

$$\theta \rightarrow \text{BLACKBOX} \rightarrow data.$$

- We have an observation $X_o = (X_{o1}, \ldots, X_{on})$ obtained from the "black-box" for some unknown value of θ. Suppose $\pi(\theta)$ is our prior on θ.
- The goal is to make inference about the value of the parameter from which X_o was generated.
- What happens in the black box is a mystery. We assume that it is not easy to specify a data generating model. Such a model may have many components, even may not be analytically expressible.
- Example: Phylogenetic Trees.
- Example: Non-linear differential equations.
- Based on the values generated from the black box ABC methods makes inference about the parameter without requiring an user to specify a model for the data generating process.

BASIC ABC

- The basic ABC algorithm goes through the following three steps.
  1. Generate θ from $\pi(\theta)$.
  2. Simulate $X_1 = (X_{11}, \ldots, X_{1n})$ from the black box with parameter θ.
  3. Accept θ if $X_o = X_1$, and return to Step 1.
- Clearly if X is a continuous random variable the probability that $X_o = X_1$ is zero. So the above algorithm does not work well.
- Thus an approximate method is used. The steps are as follows
  1. Choose a small $\epsilon > 0$ and a distance function ρ.
  2. Generate θ from $\pi(\theta)$.
  3. Simulate $X_1 = (X_{11}, \ldots, X_{1n})$ from the black box with parameter θ.
  4. Accept θ if $\rho(X_o, X_1) < \epsilon$, and return to Step 2.
- The acceptance rate and the accuracy of the posterior depend crucially on ϵ. See (Allingham et. al. [2009]).

SOME VARIATIONS AND RECENT DEVELOPMENTS

- In many cases instead of comparing the full data, a set of summary statistics is compared. This reduces the dimension of the problem. However, if the statistics used are not sufficient for the unknown data-generation model, their would be some loss in the resulting posterior.

- Marjoram et. al. [2003] develop a MCMC ABC method by targeting a stationary distribution of the form $\Pi^{\epsilon}(\theta, X_1 \mid X_o)$.

- An SMC version of ABC with each chain was considered by Sisson et. al. [2007].

- Wood [2010] postulate a synthetic likelihood which requires asymptotic normality of the summary statistics used.

- In recent times Drovandi et. al. [2013] among others use parametric auxiliary models in indirect inference.

- More recently random forest based classification techniques have been successfully used in an ABC setup. We refer to Pham et. al. [2014] for an example.

EMPIRICAL LIKELIHOOD (EL) IN ABC

- Mengersen, Pudlo and Robert [2013] were the first to consider the use of empirical likelihood in ABC setting.

- They assumed that X_{o1}, ..., X_{on} are i.i.d and a set of constraints of the form

$$E\left[h(X_{oi}, \theta)\right] = 0 \quad \forall\ i = 1, \ldots, n$$

  are available. Here the expectation is taken w.r.t. the unknown true distribution.

- An empirical likelihood can then be calculated by re-weighting the data by weights given by:

$$\widehat{w} = argmax_{w \in \mathcal{W}_\theta} \prod_{i=1}^{n} w_i, \text{ where } \mathcal{W}_\theta = \left\{ w \ : \ \sum_{i=1}^{n} w_i h(X_{oi}, \theta) = 0 \right\} \cap \Delta_{n-1}.$$

- They use a fast importance sampling algorithm to sample from the posterior.

- However, the method requires one specify the function h, which is not easy.

A FULLY DATA DEPENDENT APPROACH.

- Let θ_o be the value of θ corresponding to X_o.
- Suppose we could generate m data sets of length n, ie. $X_1, \ldots, X_m$ from the black-box putting same parameter value θ_1.
- If $\theta_0 = \theta_1$, for each i, $\bar{X}_{i\cdot}$ and $\bar{X}_{o\cdot}$ are identically distributed. So it clearly follows that:

$$E_{\theta_0}\left[\bar{X}_{o\cdot} - \bar{X}_{i\cdot}\right] = 0, \ \forall \ i = 1, 2, \ldots, m.$$

- If we can build an EL using the above relationship, then we have a procedure which would not require us to specify any relationship between the data and the parameter directly.
- Furthermore, such a procedure would inherit good properties of EL as indicated in Mengersen et. al. [2013].
- That is what we are going to do.

THE BUILT UP

- Suppose we have pairs of independent random variables (Y_i, X_i), $i = 1, 2, \ldots, m$, with same marginal distribution and generated from the same parameter value θ.

- Let $f_\theta^{\otimes 2}(Y_i, X_i)$ be the joint density of (Y_i, X_i). Clearly,

$$f_\theta^{\otimes 2}(Y_i, X_i) = f_\theta(Y_i) f_\theta(X_i).$$

- We construct a composite likelihood for (Y_1, X_1), (Y_2, X_2), ..., (Y_m, X_m) as

$$l^{(m)}(\theta) = \prod_{i=1}^{m} f_\theta^{\otimes 2}(Y_i, X_i).$$

- In the event that $Y_i = X_o$, for all $i = 1, 2, \ldots, m$, the above likelihood becomes:

$$l^{(m)}(\theta) = \{f_\theta(X_o)\}^m \prod_{i=1}^{m} f_\theta(X_i).$$

- This is similar to the idea of data cloning (Doucet et. al. [2002], Lele [2007]).

THE BUILT UP

- We use a scaled version of the corresponding log-likelihood which is given by:

$$\frac{1}{m}\log\left(l^{(m)}(\theta)\right) = \log(f_\theta(X_o)) + \frac{1}{m}\sum_{i=1}^{m}\log\left(f_\theta(X_i)\right).$$

- Clearly as $m \to \infty$, the second term converges almost everywhere to the differential entropy defined as, $E_\theta\left[\log\left(f_\theta(X_1)\right)\right] = \log C(\theta)$. Thus for large m, the generated samples X_1, X_2, ..., X_m have little influence on the likelihood. Major contribution comes from the first term.

- Moreover, $C(\theta)$ is usually a slowly varying function of θ. For a location model $C(\theta)$ is a constant independent of θ. That is, with a pre-specified prior $\pi(\theta)$, if we define a posterior as:

$$\frac{e^{\frac{1}{m}l^{(m)}(\theta)}\pi(\theta)}{\int_{t\in\Theta} e^{\frac{1}{m}l^{(m)}(t)}\pi(t)dt}$$

  for the location models, under mild assumptions, in the limit as $m \to \infty$ the term involving $C(\theta)$ cancels out and we get the posterior only conditional on the observation X_o.

APPLICATION OF THE EMPIRICAL LIKELIHOOD

- We need to estimate $\log\left(l^{(m)}(\theta)\right)/m$ from the observation X_o an the generated data X_1, X_2, ..., X_m.
- We use empirical likelihood for this.
- In particular, we estimate the joint distribution of (X_o, X_i), $i = 1, 2 \ldots, m$.
- Consider a collection of functions $g_1, \ldots, g_r$ depending only on the observations and not on the parameter θ.
- For a θ generated from $\pi(\theta)$ and X_1, ..., X_m identically and independently generated from the model with input θ, suppose we define,

$$\mathcal{W}_\theta = \bigcap_{k=1}^{r} \left\{ w \ : \ \sum_{i=1}^{m} w_i \left\{ g_k\left(X_i\right) - g_k\left(X_o\right) \right\} = 0 \right\} \cap \Delta_{m-1}.$$

- Let $\widehat{w}(\theta) := \widehat{w}(X_1, \ldots, X_m, X_o) = \text{argmax}_{w \in \mathcal{W}} \left(\prod_{i=1}^{m} m w_i \right)$.
- When the problem is infeasible, we define $\widehat{w}(\theta) = 0$.

THE TARGET POSTERIOR

- We first estimate:

$$\frac{1}{m}\log\widehat{\left(l^{(m)}(\theta)\right)} = \frac{1}{m}\sum_{i=1}^{m}\log(\widehat{w}_i(\theta)).$$

- Now plugging this estimate in the expression of the posterior above we get our target posterior:

$$\Pi(\theta \mid X_o) := \frac{e^{\frac{1}{m}\sum_{i=1}^{m}\log(\widehat{w}_i(\theta))}\pi(\theta)}{\int_{t\in\Theta} e^{\frac{1}{m}\sum_{i=1}^{m}\log(\widehat{w}_i(t))}\pi(t)dt} \propto e^{\frac{1}{m}\sum_{i=1}^{m}\log(\widehat{w}_i(\theta))}\pi(\theta).$$

- When $\widehat{w}_i(\theta) = 0$, $\Pi(\theta \mid X_o) := 0$.

- The samples are drawn from this posterior using MCMC procedures.

- In general, computation of empirical likelihood is fast.

- The MCMC could sometimes be a bit tricky to run.

CHOICE OF SUMMARIES

- The proposed empirical likelihood estimates joint weights by matching the moments of $g(X_1)$, ..., $g(X_m)$ with that of $g(X_o)$, without requiring a direct relationship with the parameter.
- We assume that for $i \in \{o, 1, \ldots, m\}$, $X_i \in \mathbb{R}^n$.
- For some k and some positive deterministic γ_k, for each i we may define, $g_k(X_i) = \frac{1}{n}\sum_{j=1}^{n} X_{ij}^{\gamma_k}$.
- The summary g_k is the γ_kth raw sample moment and would constrain the underlying distribution through its moments.
- The γ_k sample quantile of X_i may be used for any $\gamma_k \in [0, 1]$, which would directly put a constraint on the distribution through its quantiles.
- The proportion of times X_i is larger than γ_k, i.e. $g_k(X_i) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\{X_{ij} \geq \gamma_k\}}$.
- Joint moments, up-crossings etc. Other transformations eg. spectral domain statistics can be used.

ASSUMPTIONS FOR ASYMPTOTIC PROPERTIES

- We consider limits as n and $m = m(n)$ grow unbounded.
- Suppose that, $E[g(X_i^{(n)}(\theta))]$ is finite so that we can write

$$g\left(X_i^{(n)}(\theta)\right) = E\left[g\left(X_i^{(n)}(\theta)\right)\right] + \xi_i^{(n)}(\theta) = \mathfrak{g}^{(n)}(\theta) + \xi_i^{(n)}(\theta),$$

  where $E[\xi_i^{(n)}(\theta)] = 0$ for all i, n and θ.

- We make the following assumptions.

(A1) (Indentifiability and convergence) There is a sequence of positive increasing real numbers $b_n \to \infty$, such that:

$$\mathfrak{g}^{(n)}(\theta) = b_n \{\mathfrak{g}(\theta) + o(1)\},$$

where $\mathfrak{g}(\theta)$ is a one to one function of θ that does not depend on n. Furthermore, $\mathfrak{g}(\theta)$ is continuous at θ_o and for each $\epsilon > 0$, and for all $\theta \in \Theta$, there exists $\delta > 0$, such that whenever $|| \theta - \theta_o ||> \epsilon$, $|| \mathfrak{g}(\theta) - \mathfrak{g}(\theta_o) ||> \delta$.

ASSUMPTIONS

(A2) (Feasibility) For each θ, n and $i = o, 1, \ldots, m(n)$, the vectors $\xi_i^{(n)}(\theta)$ are identically distributed, supported over the whole space, and their distribution puts positive mass on every orthant, $\mathcal{O}_s$ of $\mathbb{R}^r$, $s = 1, 2, \ldots, 2^r$. Furthermore, for every orthant $\mathcal{O}_s$, as $n \to \infty$,

$$\sup_{\{i \,:\, \xi_i^{(n)}(\theta) \in \mathcal{O}_s\}} || \, \xi_i^{(n)}(\theta) \, || \longrightarrow \infty$$

in probability, uniformly in θ.

(A3) (Growth of extrema of Errors) As $n \to \infty$,

$$\sup_{i \in \{o,1,2,\ldots,m(n)\}} \frac{|| \, \xi_i^{(n)}(\theta) \, ||}{b_n} \to 0$$

in probability, uniformly in $\theta \in \Theta$.

ASYMPTOTIC RESULTS

- Let $l_n(\theta) := exp(\sum_{i=1}^{m(n)} \log(\widehat{w}_i(\theta))/m(n))$ and for each n, we define:

$$\Theta_n = \left\{\theta \ : \ || \ \mathfrak{g}(\theta) - \mathfrak{g}(\theta_o) \ || \leq b_n^{-1}\right\}.$$

  By continuity of $\mathfrak{g}$ at θ_0, Θ_n is nonempty for each n. Furthermore, since b_n is increasing in n, Θ_n is a decreasing sequence of sets in n.

- **Lemma 1** *Under assumptions (A1) to (A3), with high probability, the likelihood* $l_n(\theta) > 0$, *for all* $\theta \in \Theta_n$.

- For $\theta \in \Theta$ and $\epsilon > 0$, by $B(\theta, \epsilon)$ we denote the ball of radius ϵ around θ.

- **Lemma 2** *Under assumptions (A1) - (A3), for every* $\epsilon > 0$, *with high probability, the empirical likelihood is zero outside* $B(\theta_0, \epsilon)$.

- From Lemma 1. and 2. it follows that:

$$\Pi_n(\theta \mid X_o(\theta_o)) = \frac{l_n(\theta)\pi(\theta)}{\int_{t\in\Theta} l_n(t)\pi(t)dt}$$

  is a valid probability measure with high probability.

- **Theorem 1** *As* $n \to \infty$, $\Pi_n(\theta \mid X_o(\theta_o))$ *converges in probability to* δ_{θ_o}, *where* δ_{θ_0} *is the degenerate probability measure supported at* θ_0.

RESULTS FOR A GAUSSIAN MEAN

- As a test case we consider estimation of mean from a standard normal distribution. We assume a standard normal prior on the mean.

- We assume that the variance is known to be 1. The coverage and the length of the 95% credible intervals were checked for different choices of the functions g_1, ..., g_r.

- We took $n = 100$, $m = 25$. The coverages are based on 100 repetitions. For each repetition, last $50,000$ of the $100,000$ samples drawn from the posterior by adaptive MCMC was used.

Choice of g	Coverage	Average Length
$1st$ moment (mean)	0.93	0.3406
Median	0.93	0.4259
$1st$ and $2nd$ moments	0.88	0.3047
$1st$, $2nd$ and $3rd$ moments	0.85	0.2710
3 quartiles	0.76	0.2807
Mean and median	0.76	0.2392
$1st$, $2nd$, $3rd$ and $4th$ moments	0.72	0.2183

EXAMPLE: $g - and - k$ DISTRIBUTION

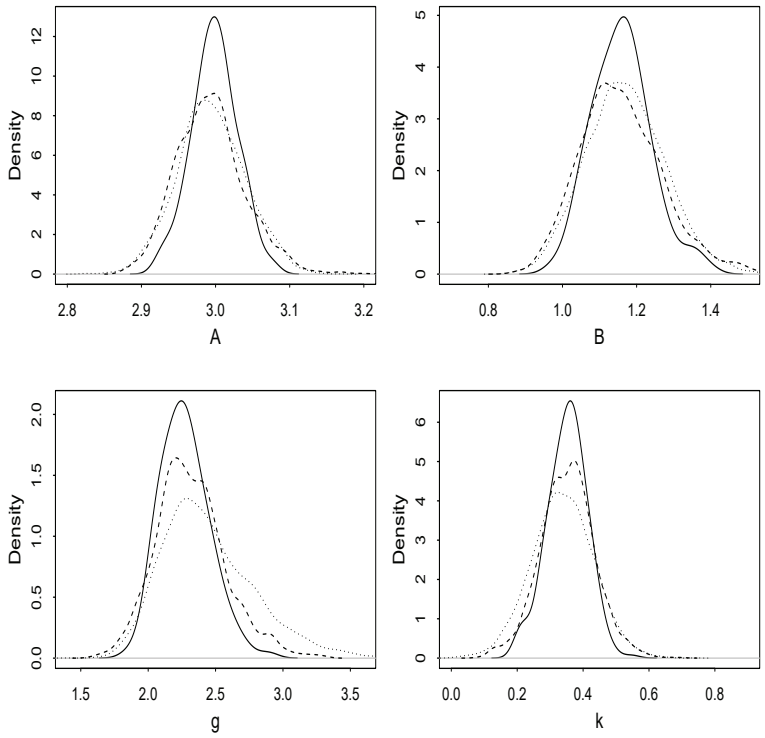
- Next we consider the $g - and - k$ distribution.
- The distribution is expressed by its quantile function

$$Q\left(r; A, B, g, k\right) = A + B\left(1 + .8\frac{1 - e^{-gz(r)}}{1 + e^{-gz(r)}}\right)\left(1 + z(r)^2\right)^k z(r),$$

  where $z(r)$ is the rth standard normal quantile. The parameters A, B, $g$and k respectively represent location, scale, skewness and kurtosis of the distribution.

- We generated sample from a $g - and - k$ with $A = 3$, $B = 1$, $g = 2$ and $k = 0.5$. The parameters were assumed to follow a $U(0, 10)$ prior.
- In our simulation we chose $n = 1,000$ and $m = 40$.
- The results are based on last $25,000$ out of $50,000$ samples drawn from the posterior using an adaptive random walk MCMC procedure.

RESULTS: $g - and - k$ DISTRIBUTION



- Summary statistics used: the mean and three quartiles.
- Solid line: proposed EL based method.
- Dashed line: Synthetic likelihood with the same summary statistics.
- Dotted line: Rejection ABC with regression adjustment, based on $5,000,000$ datasets, tolerance chosen so that 2000 samples were kept. This is the current gold standard.
- The usual summaries based on octiles results in slightly inferior performance in estimating k.

- The synthetic likelihood is expected to work well here. Both methods produce posterior similar to the rejection ABC.

EXAMPLE: ARCH(1)

- The ARCH(1) model is defined as

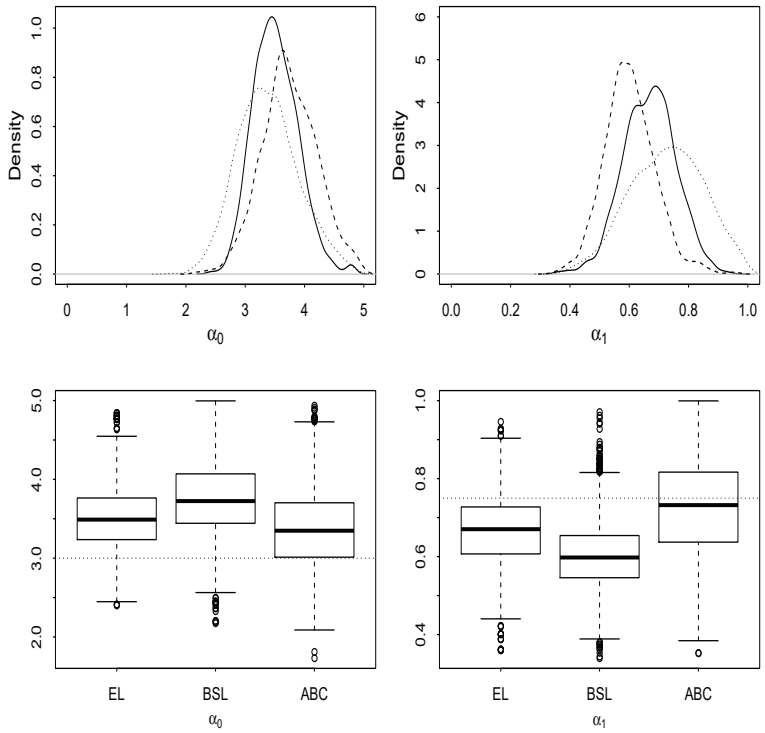
$$y_j = \sigma_j \epsilon_j, \quad \sigma_j^2 = \alpha_0 + \alpha_1 y_{j-1}^2, \text{with } \epsilon_j \sim N(0,1),\ j = 1, \ldots, n.$$

- We generate vector of size $n = 1000$ from $\alpha_0 = 3.00$ and $\alpha_1 = 0.75$. The priors on α_0 and α_1 is taken to be $U(0,5)$ and $U(0,1)$ respectively. we take $m = 20$.

- Four summaries were used. They were the lag 1 auto-covariance of y^2 and the three quartiles of $|y|$.

- The first order auto-covariance is highly non-Gaussian.

RESULTS: ARCH(1) MODEL

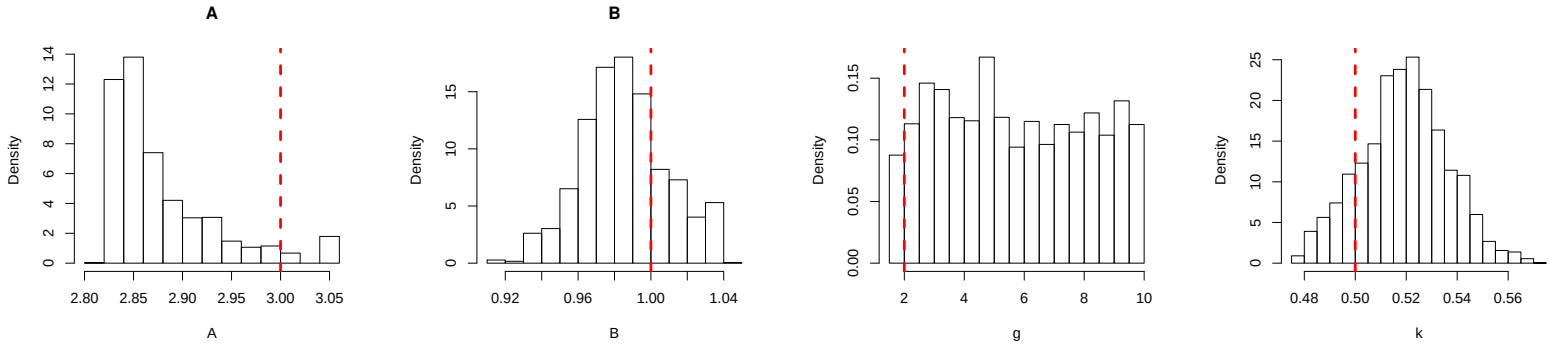


- Solid line: proposed EL based method.
- Dashed line: Synthetic likelihood with the same summary statistics.
- Dotted line: Rejection ABC with $\epsilon = .0025$ based on $1,000,000$ datasets.

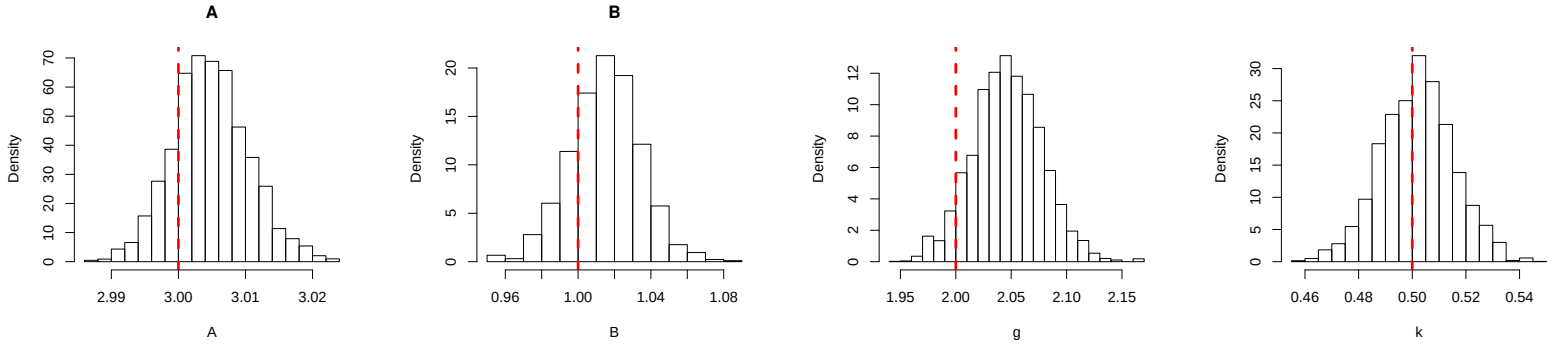
- Synthetic likelihood estimates are quite different from those obtained by the rejection ABC. This is specially true for α_1.

CHOICE OF FUNCTIONS: $g - and - k$ DISTRIBUTION

- Choice of functions g_1, g_2, ..., g_r are crucial.
- First four raw moments.



- Median and first four raw moments.



CONCLUSION

- We introduce an easy-to-use empirical likelihood based ABC method, where the only required inputs are a choice of summary statistic, it's observed value, and the ability to simulate that particular statistic under the model for any parameter value.

- We show that the proposed method is asymptotically consistent and has better performance than the synthetic likelihood when the summary statistics are not normally distributed.

- It seems the method slightly under-estimates the posterior uncertainty.

- The choice of summaries are important. Bad summary statistics may lead to slow mixing of the MCMC. However, the proposed method is no worse than the Synthetic likelihood for such summaries.

- As implemented, the proposed method cannot choose constraints. That is, unimportant summaries are given equal weight as the important ones. No guidance for choosing summary statistics.

- Good adaptive MCMC procedures are required to draw samples from the posterior.