On An Empirical Likelihood based Solution to Approximate Bayesian Computation Problem

Sanjay Chaudhuri
University of Nebraska-Lincoln

Shubhroshekhar Ghosh and Kim Cuc Pham, National University of Singapore

Approximate Bayesian Computation (ABC)

$$\theta \to \boxed{\text{BLACKBOX}} \to data.$$

- We have an observation $X_o = (X_{o1}, \ldots, X_{on})$ obtained from the "black-box" for some unknown value of θ say θ_o.

- The goal is to make inference about θ_o.

- 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: Too many trees.

- Example: Non-linear differential equations: Complicated model.

- Based on the values generated from the black box ABC methods make inferences about the parameter without requiring an user to specify a model for the data generating process.

Basic ABC

- Suppose $\theta \in \Theta$ and $\pi(\theta)$ is a prior defined on Θ.

- 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.

Direct Estimation of the True Posterior

- An approximate method is used. The steps are as follows:

  1. Choose a small tolerance $\epsilon > 0$, a distance function d, and a vector of summary statistics g.

  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 $d(g(X_o), g(X_1)) < \epsilon$, and return to Step 2.

- The acceptance rate and the accuracy of the posterior depend crucially on ϵ. See (Allingham et. al. [2009]).

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

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

- How the tolerance ϵ affects the accuracy of the estimated posterior is not well studied.

ESTIMATING THE LIKELIHOOD OF $g(X_o)$ GIVEN θ

- In view of the requirements of MCMC sampling, it is actually sufficient to estimate the likelihood of $g(X_o)$ for every value of θ from the generated replications X_i, $i = 1, 2, \ldots, m$.

- Wood [2010] asymptotic normality of g for all $\theta \in \Theta$. He uses a multivariate Gaussian Likelihood for $g(X_o)$ where the mean and the covariance matrix at each θ are estimated using the generated replications.

- An et.al. [2020] take a more semi-parametric approach, where the marginal density for each component in the summary vector is estimated using kernel density estimator. The joint density of the summary vector is the estimated using a Copula.

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

- Random forest based classification techniques which directly estimate the test ratio in a MCMC step 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:
$$\hat{w} = argmax_{w\in\mathcal{W}_\theta}\prod_{i=1}^{n} w_i, \text{ where } \mathcal{W}_\theta = \left\{w \ : \ \sum_{i=1}^{n} w_i h(X_{oi},\theta) = 0\right\} \cap \Delta_{n-1}.$$

- 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.

Data Dependent EL based Direct Estimator of the Posterior

- The proposed estimator hinges crucially on the following observation.

- Suppose we could generate m i.i.d. data sets of length n, ie. X_1, ..., X_m from the black-box putting same parameter value θ_1.

- If $\theta_0 = \theta_1$, for each i, For any summary g, $g(X_i)$ and $g(X_o)$ are identically distributed. So it clearly follows that:

$$E_{\theta_0}\left[g(X_i) - g(X_o)\right] = 0,\ \forall\ i = 1, 2, \ldots, m.$$

- We now build an EL using the above relationship and estimate the posterior.

- Define:

$$\mathcal{W}_{\theta_1} = \left\{w\ :\ \sum_{i=1}^{m} w_i \left\{g\left(X_i\right) - g\left(X_o\right)\right\} = 0\right\} \cap \Delta_{m-1}.$$

$$l(\theta_1) = \frac{1}{m} \max_{w \in \mathcal{W}_{\theta_1}} \sum_{i=1}^{m} \log m w_i + \widehat{H}^0_{g(x_1)|\theta}(\theta_1),$$

EL based Direct Estimator of the Posterior (contd.)

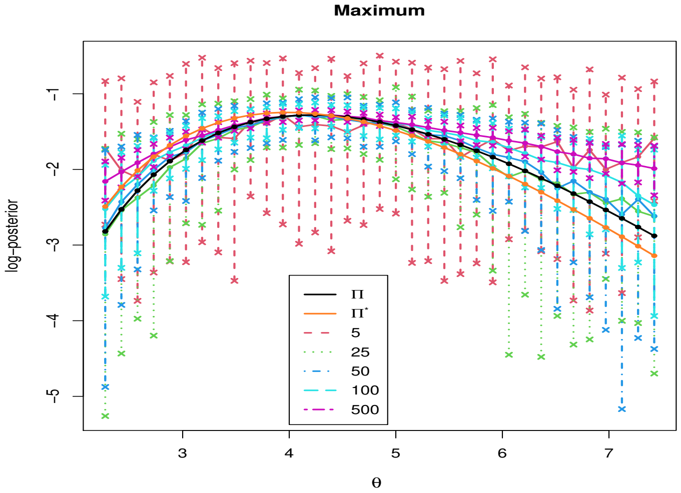
- Here $\hat{H}^0_{g(x)_i|\theta}(\theta_1)$ is the estimate of the true differential entropy of the conditional density of $g(x_1)$ given θ at θ_1.

- We define an estimator of the true posterior at θ_1 as:

$$\hat{\Pi}(\theta_1 \mid g(X_o)) = \frac{e^{l(\theta_1)}\pi(\theta_1)}{\int e^{l(t)}\pi(t)dt} \propto e^{\left\{\frac{1}{m}\max_{w\in\mathcal{W}_{\theta_1}}\sum_{i=1}^m \log m w_i + \hat{H}^0_{g(x_1)|\theta}(\theta_1)\right\}}\pi(\theta_1).$$

- If the maximisation problem is infeasible, we define $\hat{\Pi}(\theta_1 \mid g(X_o)) = 0$.

- Example: Estimate θ from $X \sim N(0,\theta)$.
- $n = 100$, $\theta_o = 4$, $\pi(\theta) = U(0,10)$.
- $g(X_i) = max_j(X_{ij})$.
- The true log-posterior is in black.
- The coloured lines are the estimated posterior for $m = 5$, 25, 50, 100, 500.
- Vertical lines are point-wise 95% confidence bands.



Direct Information Projection

- For a $\theta \in \Theta$, assume X_1 was generated using θ and for convenience, denote, $g_o = g(X_o)$, $g_1 = g(X_1)$.
- By construction for each θ, g_1 is conditionally independent of g_o given θ.
- We focus on the conditional density of g_1 and θ given g_o.
- By construction it can be shown that the *true* conditional is given by:

$$f_0(\theta, g_1 \mid g_o) = f_0(g_1 \mid \theta)\Pi(\theta \mid g_o).$$

- That is the true density if of the form $q'(\theta)f_0(g_1 \mid \theta)$, where $q'(\theta)$ is a density defined on θ.
- We take a density $f(\theta, g_1 \mid g_o)$ and project it on the above set of densities by minimising the KL-divergence.
- It can be shown that the projection is given by:

$$\frac{e^{E^0_{g_1|\theta}[\log f(\theta,g_1,g_o)]+H^0_{g_1|\theta}(\theta)}}{\int e^{E^0_{g_1|t}[\log f(t,g_1,g_o)]+H^0_{g_1|t}(t)}dt} f_0(g_1 \mid \theta) = f'(\theta \mid g_o) f_0(g_1 \mid \theta).$$

- That is for a user chosen density f, $f'(\theta \mid g_o)$ is an estimate of $\Pi(\theta \mid g_o)$.

Connection to the Proposed EL

- In the EL formulation, we estimate the expectation by the mean of the log EL weights.

- Furthermore, in order to uniquely find an optimal density f, we need to ensure that,

  1. the corresponding conditional density of g_1 given θ is the same as the corresponding conditional density of g_o given θ, and
  2. the corresponding marginal density of θ is the prior π.

- The first constraint is approximately ensured by the particular choice of the constraints.

- It can be shown that the best estimate of the underlying density can be obtained from a reverse information projection.

- In short, The proposed EL-based method is well-justified.

Asymptotic Properties

- Under mild conditions, as $n \to \infty$, the approximate posterior converges to the degenerate distribution supported on the true parameter value θ_0.

- Even if the number of replications, i.e. m, grow much faster than sample size n (even exponentially if the underlying distribution is normal), the posterior consistency would be retained.

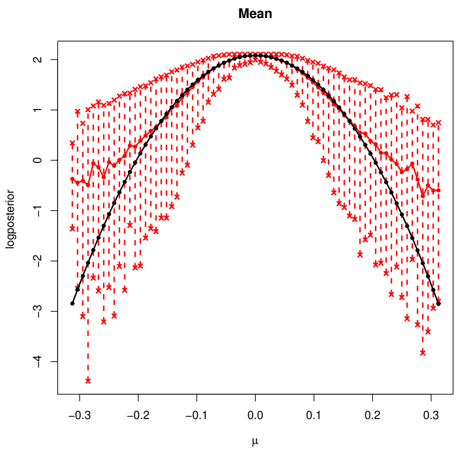
- The number of replications growing to infinity by itself ensures that the probability of $exp(\frac{1}{m}\sum_{i=1}^{m} \log(\widehat{w}(\theta_o)))$ does not collapse to zero grows to one.

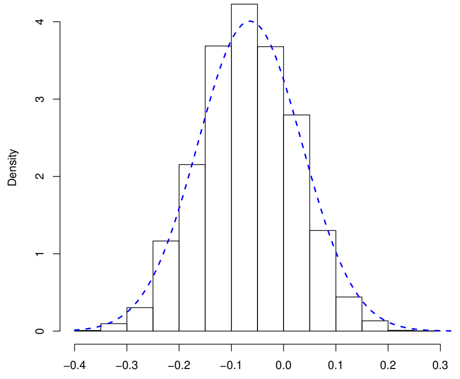
- We can show that for fixed n, and fixed number of summaries, the posterior will be more flat as m grows.

- In order to maintain the frequentist coverage, larger m would require larger number of summaries.

Results for a Gaussian mean

- Estimation of mean from a standard normal distribution.
- We assume a standard normal prior on the mean.
- Take $n = 100$, $m = 25$.
- For mean as the summary, the histogram of the observations drawn via MCMC matches with the true posterior, i.e. $N\left(\sum_{i=1}^{n} X_{oi}/(n+1), (n+1)^{-1}\right)$.





- We present the coverages and average lengths of the 95% confidence intervals for various choices of the summary statistics:

Choice of g	rep	Cov.	Ave. Len.
$1st$ moment (mean)	25	0.95	0.360
Median	25	0.95	0.446
$1st$ and $2nd$ central moments	40	0.94	0.330
Mean and median	40	0.94	0.330
$1st$, $2nd$ and $3rd$ central moments	70	0.91	0.307
3 quartiles	75	0.93	0.329

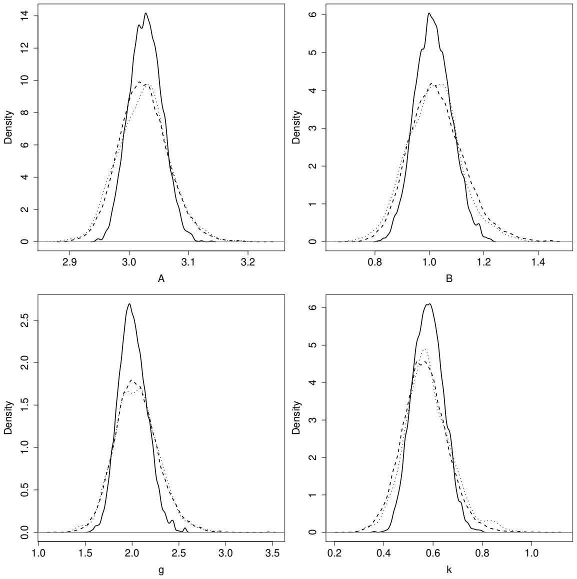
RESULTS: $g - and - k$ DISTRIBUTION

- 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.

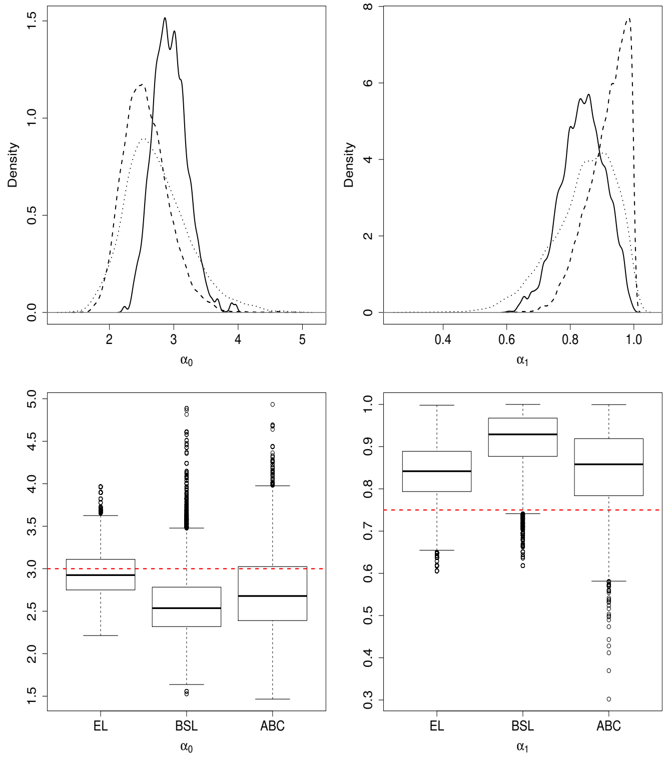


- $\theta_o = (A_o, B_o, g_o, k_o) = (3, 1, 2, 0.5)$, $n = 1000$, $m = 40$, $\theta \sim U(0, 10)^4$.
- Summary statistics used: mean and three quartiles.
- EL: $bold - line$, BSL: $dashed - line$, ABC: $dotted - line$.
- The usual summaries based on octiles results in slightly inferior performance in estimating k.
- The synthetic likelihood is expected to work well here.

ARCH(1) MODEL

- The ARCH(1) model is defined as

$$X_{ij} = \sigma_{ij}\epsilon_{ij}, \quad \sigma_{ij}^2 = \alpha_0 + \alpha_1 X_{i(j-1)}^2, \text{with } \epsilon_j \sim \ N(0,1),\ \alpha_0 > 0,\ \alpha_1 \in (0,1).$$



- True parameters $(3, 0.75)$, $n = 1000$, $m = 50$, $\pi = U((0,5) \times (0,1))$.
- Summary: Three quartiles of $|X_i|$. The fourth one is the following:

$$Y_{ij} = X_{ij}^2 - \sum_{j=1}^{n} X_{ij}^2/n,$$

$$g_4(X_i) = \frac{1}{n}\sum_{j=2}^{n}\left[1_{\{(Y_{ij} \cdot Y_{i(j-1)}) \geq 0\}} - 1_{\{(Y_{ij} \cdot Y_{i(j-1)}) \leq 0\}}\right].$$

- This summary controls the $lag-1$ dependence in the data.
- Summaries are not asymptotically normal.

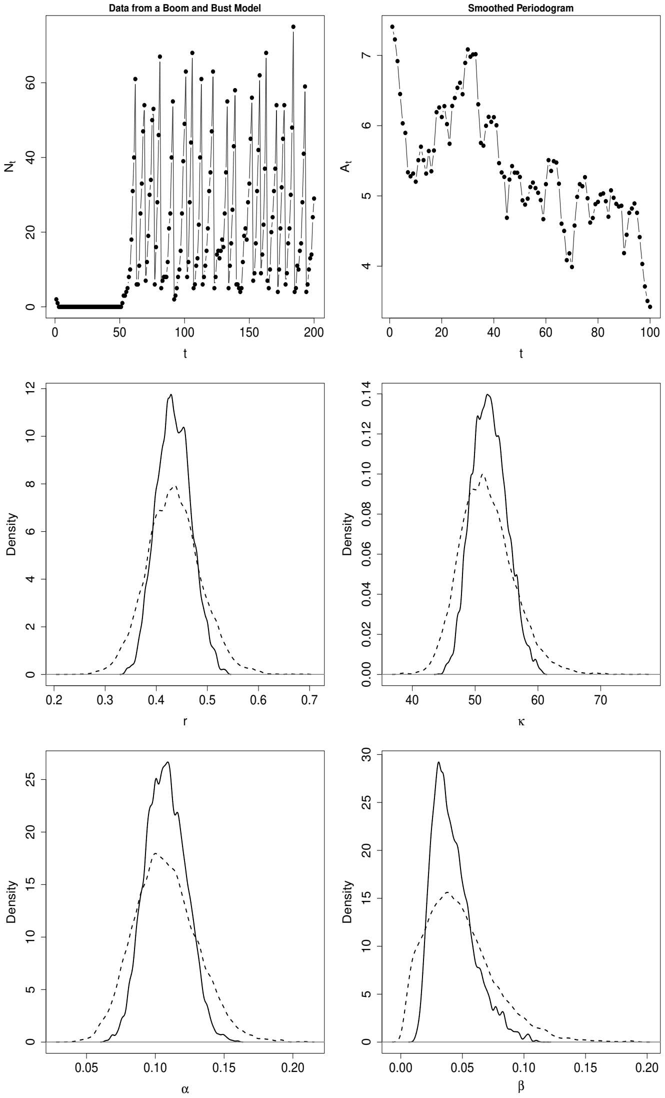
- Synthetic likelihood estimates are quite different from those obtained by the rejection ABC. This is specially true for α_1.

Boom and Bust Model

- This a four parameter (r,κ,α,β) stochastic model taking values in the set of non-negative integers.
- Given N_t and the parameters:

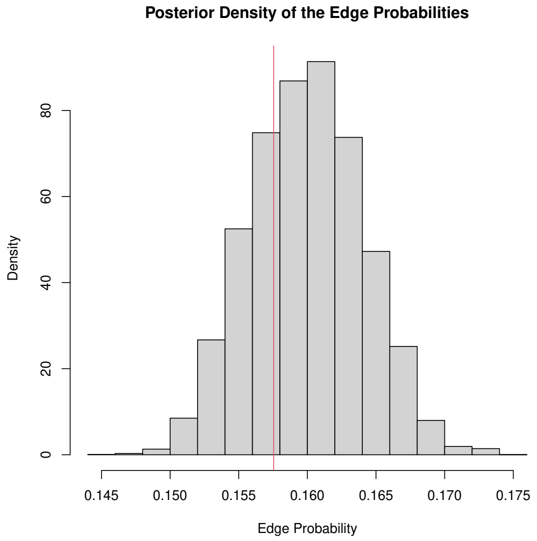
$$N_{t+1} \sim \begin{cases} Poisson(N_t(1+r)+\epsilon_t, & \text{if } N_t \leq \kappa \\ Binom(N_t, \alpha)+\epsilon_t, & \text{if } N_t > \kappa \end{cases},$$

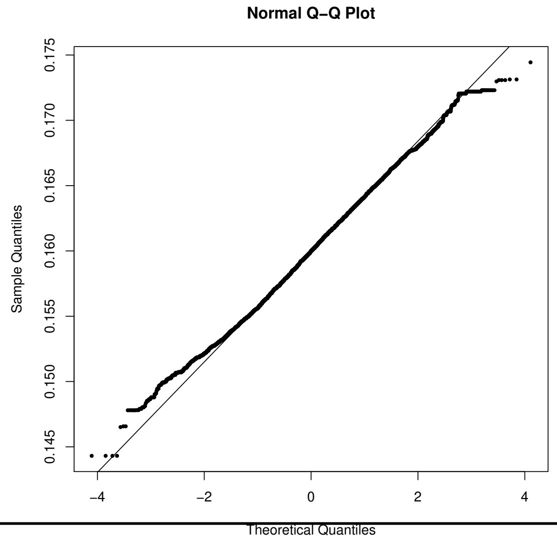
  where $\epsilon_t \sim Poisson(\beta)$.
- $n = 200$, (burn-in 50), $m = 40$.
- True values $(0.4, 50, 0.09, 0.05)$.
- We used the following types summaries:
  1. Proportion of observations in the interval $(0, 15)$. (Upcrossings !)
  2. Proportion of differences $N_{t+1} - N_t$ strictly larger than 2. (Errors !)
  3. We computed the smoothed periodograms and used the proportion of log-amplitudes in $(5.120, 6.278)$. (Autocovariances !)
  4. The numbers chosen judiciously from the observed data.



EDGE PROBABILITY OF AN ERDÖS-RENYI RANDOM GRAPH

- We estimate the edge probability p of an Erdös-Renyi random graph with number of vertex $n = 100$.
- We put a $Beta(1.5, 1.5)$ prior on p.
- The number of edges and the number of triangles were used as summary statistics.
- We used $m = 25$ replications.
- Even though we are estimating parameters, this method does not show any weight degeneracy issues.
- In fact, the posterior is quite close to normal.
- This method can be generalised to more complicated models for p.





STEREOLOGICAL DATA

- Inclusions are microscopic particles introduced in the steel production process. The size of the largest inclusion in a block is thought to be important for steel strength.

- The data considered here were first analysed by Anderson and Coles [2002], and consist of measurements on inclusions from planar cross-sections.

- We try to model the diameters of inclusions in a block of steel.

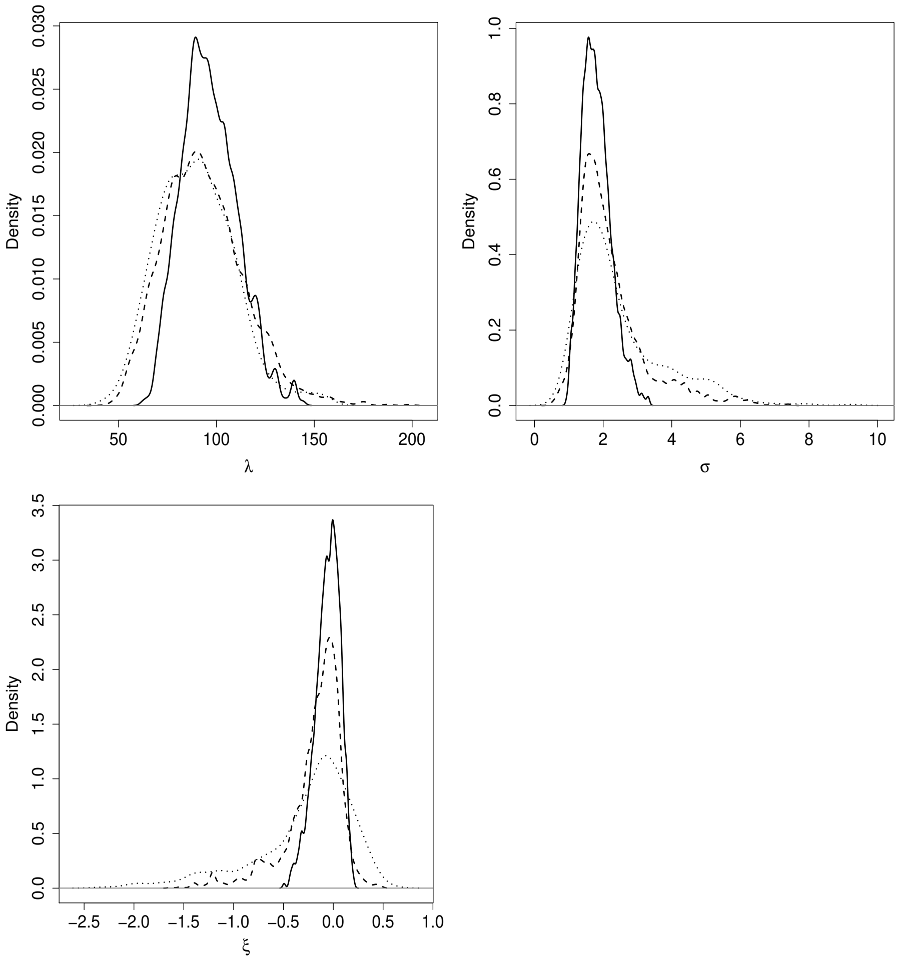
- We assume that the inclusion centres follow a homogeneous Poisson process with rate λ.

- Conditional on exceeding a threshold value v_0, the largest inclusion diameter V, is assumed to follow a generalised Pareto distribution:

$$\text{pr}(V \leq v | V > v_0) = 1 - \left\{ 1 + \frac{\xi(v - v_0)}{\sigma} \right\}_+^{-\frac{1}{\xi}}.$$

- Given V, the other two principal diameters are determined by multiplying V with two independent $U[0,1]$ random variables.

RESULTS: STEREOLOGICAL DATA



- Solid line: proposed EL based method with $m = 25$.
- Dashed line: Synthetic likelihood with the same summary statistics.
- Dotted line: Rejection ABC with $\epsilon = 0.00005$ based on $10,000,000$ datasets.
- Here $n = 112$.
- L is the number of inclusions. The summary statistics used are:
  $a)$ $(L - 112)/100$,
  $b)$ the mean of the observed planar measurements,
  $c)$ the median of the of the observed planar measurements, and
  $d)$ the proportion of planar measurement less than or equal to 6.

- We assume independent uniform priors for λ, σ and ξ with ranges $(1, 200)$, $(0, 10)$ and $(-5, 5)$ respectively.

CONCLUSION

- We provide a solution to the ABC problem using empirical likelihood based method.

- The proposed method estimates the posterior directly and is based on an interpretable likelihood 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.

- The parameter estimates have interpretable and favourable properties.

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

- 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.