Approximate Bayesian Computation with Indirect Moment Conditions

Alexander Gleim
(Bonn Graduate School of Economics)

&

Christian Pigorsch
(University of Bonn)

COMPSTAT, August 24, 2010

Introduction

- Bayesian statistics regards parameters of a given model as both unknown and stochastic
- Bayesian inference makes use of prior information on the model parameter which is then updated by observing a specific data sample via the Bayes Theorem

$$p(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{\int_{\Theta} p(y|\theta)\pi(\theta)}$$

- $p(\theta|y)$ is called the posterior density of the parameter θ and Bayesian inference on θ is based on $p(\theta|y)$
- In what follows we deal with posterior sampling in the case where the likelihood function of the model is of unknown form

Approximate Bayesian Computation

- ▷ We seek draws from the posterior distribution $p(\theta|y) \propto p(y|\theta)\pi(\theta)$ where the likelihood cannot be computed exactly
  1: Generate θ^* from prior $\pi(\theta)$
  2: Simulate $\hat{y}$ from likelihood $p(y|\theta^*)$
  3: Accept θ^* if $\hat{y} = \tilde{y}$
  4: return to 1:
- ▷ Results in iid draws from $p(\theta|\tilde{y})$
- ▷ Success of ABC algorithms depends on the fact that it is easy to simulate from $p(y|\theta)$
- ▷ Problems arise in the following cases ($\rightsquigarrow$ step 3)
  - y is high–dimensional
  - y lives on a continuous state–space

  in that the acceptance rate is prohibitively small (or even exactly 0)
- ▷ Rely on approximations to the true posterior density

Approximate Bayesian Computation

- Approximate methods can be implemented as
  1: Generate θ^* from prior $\pi(\theta)$
  2: Simulate $\hat{y}$ from likelihood $p(y|\theta^*)$
  3: Accept θ^* if $d(S(\hat{y}), S(\tilde{y})) \leq \epsilon$
  4: return to 1:
- Results in iid draws from $p(\theta|d(S(\hat{y}), S(\tilde{y})) \leq \epsilon)$
- Need to specify a metric d, a tolerance level ϵ as well as summary statistics S
  - If $\epsilon = \infty$, then $\theta^* \sim \pi(\theta)$
  - If $\epsilon = 0$, then $\theta^* \sim p(\theta|S(\tilde{y}))$
- The introduction of a tolerance level ϵ allows for a discrete approximation of an originally continuous posterior density
- The problem of high–dimensional data is dealt with (sufficient) summary statistics

Why sufficient summary statistics?

- A sufficient statistic $S(y)$ contains as much information as the entire data sample y ($\rightsquigarrow$ model dependent)
- For sufficient summary statistics and ϵ small

$$p(\theta|d(S(\hat{y}), S(\tilde{y})) \leq \epsilon) \overset{a}{\sim} p(\theta|\tilde{y})$$

- Neyman factorization lemma

$$p(y|\theta) = g(S(y)|\theta)\ h(y)$$

- Verifying sufficiency for a model described by $p(y|\theta)$ is impossible when the likelihood function is unknown

Indirect approach

▷ General idea

- We cannot prove sufficiency within the structural model of interest, $p(y|\theta)$
- Find an analytically tractable auxiliary model, $f(y|\rho)$ that explains the data well
- Establish sufficient summary statistics within the auxiliary model (i.e. sufficient for ρ)
- Find conditions under which sufficiency for ρ carries over to sufficiency for θ

▷ This approach is in tradition with the Indirect Inference literature (see Gourieroux et al. (1993), Gallant and McCulloch (2009), Gallant and Tauchen (1996, 2001, 2007))

Structural model

- Our observed data $\{\tilde{y}_t, \tilde{x}_{t-1}\}_{t=1}^n$ is considered to be a sample from the structural model

$$p(x_0|\theta^\circ) \prod_{t=1}^n p(y_t|x_{t-1};\theta^\circ)$$

with θ° denoting the true structural parameter value

- We are naturally not restricted to the time invariant (i.e. stationary) case

- Only requirement

We have to be able to easily simulate from $p(\cdot|\theta)$

Auxiliary model

- Assume we have an analytically tractable auxiliary model which approximates the true data generating process to any desired degree

$$\{f(x_0|\rho), f(y_t|x_{t-1};\rho)\}_{t=1}^{n}$$

- We denote with

$$\tilde{\rho}_n = \arg\max_{\rho} \frac{1}{n}\sum_{t=1}^{n} \log f(\tilde{y}_t|\tilde{x}_{t-1};\rho)$$

its Maximum Likelihood Estimate and with

$$\tilde{\mathcal{I}}_n = \frac{1}{n}\sum_{t=1}^{n}\left[\frac{\partial}{\partial\rho}\log f(\tilde{y}_t|\tilde{x}_{t-1};\tilde{\rho}_n)\right]\left[\frac{\partial}{\partial\rho}\log f(\tilde{y}_t|\tilde{x}_{t-1};\tilde{\rho}_n)\right]^{\mathsf{T}}$$

its corresponding estimate of the Information Matrix

Indirect moment conditions

- We take the auxiliary score as a sufficient statistic for the auxiliary parameter ρ

$$S(y, x|\theta, \rho) = \sum_{t=1}^{n} \frac{\partial}{\partial \rho} \log f(y_t(\theta)|x_{t-1}; \rho)$$

- We compute the score by using a simulated sample $\{\hat{y}_t, \hat{x}_{t-1}\}_{t=1}^{n}$, replacing ρ by its MLE $\tilde{\rho}_n$, i.e.

$$\hat{S}(\hat{y}, \hat{x}|\theta, \tilde{\rho}_n) = \sum_{t=1}^{n} \frac{\partial}{\partial \rho} \log f(\hat{y}_t(\theta)|\hat{x}_{t-1}; \tilde{\rho}_n)$$

- We use $\hat{S}(\hat{y}, \hat{x}|\theta, \tilde{\rho}_n)$ as summary statistic and weight the moments by $(\tilde{\mathcal{I}}_n)^{-1}$, i.e.

$$\hat{S}(\hat{y}, \hat{x}|\theta, \tilde{\rho}_n)^{\mathsf{T}} (\tilde{\mathcal{I}}_n)^{-1} \hat{S}(\hat{y}, \hat{x}|\theta, \tilde{\rho}_n)$$

ABC with Indirect Moments

- Let us now consider how to implement indirect moment conditions within ABC
  1. Compute the ML estimate of the auxiliary model parameter $\tilde{\rho}_n$, based on observations $\{\tilde{y}_t\}_{t=1}^n$
  2. Generate θ^* from prior $\pi(\theta)$
  3. Simulate $\{\hat{y}_t, \hat{x}_{t-1}\}_{t=1}^n$ from likelihood $p(y|\theta^*)$
  4. Accept θ^* if $d(S(\hat{y}), S(\tilde{y})) \leq \epsilon$
     (a) Replace $S(\hat{y})$ by $\hat{S}(\hat{y}, \hat{x}|\theta^*, \tilde{\rho}_n) = \sum_{t=1}^n \frac{\partial}{\partial \rho} \log f(\hat{y}_t(\theta^*)|\hat{x}_{t-1}; \tilde{\rho}_n)$
     (b) Note that $S(\tilde{y}) = S(\tilde{y}, \tilde{x}|\theta, \tilde{\rho}_n) = \sum_{t=1}^n \frac{\partial}{\partial \rho} \log f(\tilde{y}_t|\tilde{x}_{t-1}; \tilde{\rho}_n) = 0$ by construction for all candidate θ
     (c) Calculate the distance d by the chi-squared criterion $\hat{S}(\hat{y}, \hat{x}|\theta^*, \tilde{\rho}_n)^{\mathsf{T}}(\tilde{\mathcal{I}}_n)^{-1}\hat{S}(\hat{y}, \hat{x}|\theta^*, \tilde{\rho}_n)$ where moments are weighted according to $(\tilde{\mathcal{I}}_n)^{-1}$
  5. Return to 2.

Sufficiency within the auxiliary model

- We use summary statistics that are based on the score of the auxiliary model, i.e.

$$s_\rho = \frac{\partial}{\partial \rho} \log f(y_t | x_{t-1}; \rho)$$

- Barndorff–Nielsen, Cox (1978) showed that the normed likelihood function $\bar{f}(\cdot) = f(\cdot) - f(\tilde{\rho})$ is indeed a minimal sufficient statistic
- More general, minimal sufficiency holds true for any statistic $T(y)$ that generates the same partition of the sample space as the mapping r: $y \mapsto f(y|\cdot)$ (see Barndorff–Nielsen, Jørgensen (1976))
- For these reasons we can regard the auxiliary score s_ρ to be minimal sufficient for the auxiliary parameter ρ

Sufficiency within the structural model

- Assumption

  There exists a map g: $\theta \mapsto \rho$ such that

  $$p(y_t|x_{t-1};\theta) = f(y_t|x_{t-1};g(\theta))$$

  for all $\theta \in \Theta$ for which our prior beliefs have positive probability mass, i.e. $\pi(\theta) > 0$

- General idea

  Given a model $f(y|\rho)$ for which a sufficient statistic $S(y)$ exists and a nested sub model $p(y|\theta)$ (i.e. the map g holds exactly) then $S(y)$ is also sufficient for $p(y|\theta)$

- Assumption can be seen in light of the indirect inference literature:
  - Compared to GSM (Gallant, McCulloch (2009)) there is no need to compute the map explicitly
  - Compared to EMM (Gallant, Tauchen (1996)) the smooth embeddedness assumption is strengthened to hold not only in an open neighborhood of the true parameter value θ°

Toy example

▷ Structural model

We consider $X_i \sim \exp(\lambda)$, i.e.

$$p_X(X|\lambda) = \lambda \exp(-\lambda X)\, \mathbb{I}_{X \geq 0}$$

▷ Auxiliary model

We consider $X_i \sim \Gamma(\alpha^{(x)}, \beta^{(x)})$, i.e.

$$f_X(X|\alpha^{(x)}, \beta^{(x)}) = \frac{(\beta^{(x)})^{\alpha^{(x)}}}{\Gamma(\alpha^{(x)})} X^{\alpha^{(x)}-1} \exp(-\beta^{(x)} X)\, \mathbb{I}_{X>0}$$

The map is thus g: $\lambda \mapsto (1, \lambda)$

Toy example

- Exact inference
  - conjugate prior: $\lambda \sim \Gamma(\alpha^{(\lambda)}, \beta^{(\lambda)})$
  - likelihood: $\mathcal{L} = \lambda^n \exp(-\lambda \sum X_i)$
  - posterior: $\lambda | X \sim \Gamma(\alpha^{(\lambda)} + n, \beta^{(\lambda)} + \sum X_i)$
- For each value of $\epsilon = (1, 0.1, 0.01)$ we run IABC until we obtain 100.000 draws from $p(\lambda | d(S(\hat{X}), S(\tilde{X})) \leq \epsilon)$
- We have a total of $n = 60$ observations $\tilde{X}_i$, iid exponentially distributed with $\lambda = 1$
- We chose the prior on λ to be $\pi(\lambda) = \Gamma(1, 1)$

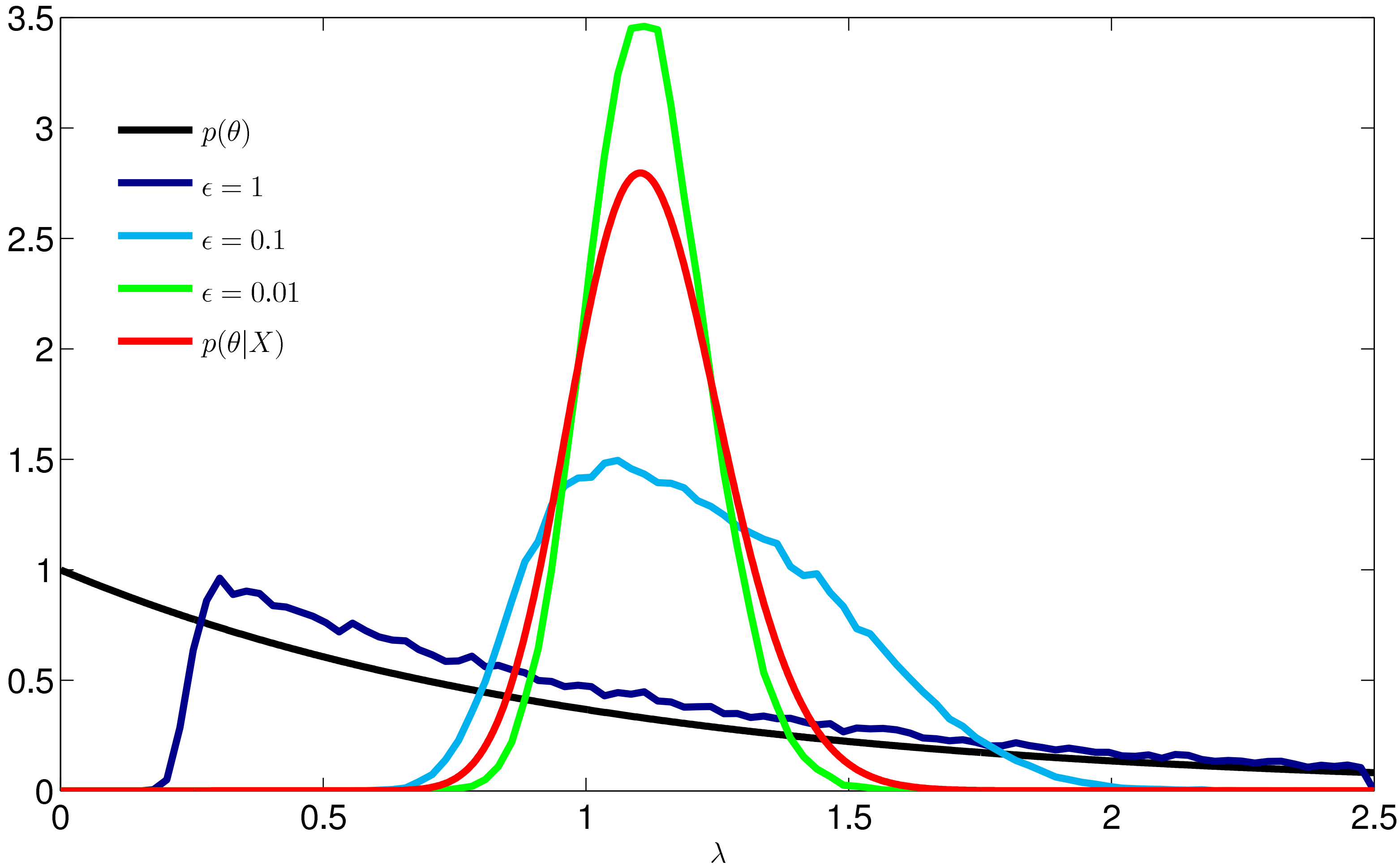

Figure 1: Histogram for posterior draws of λ for different values of ϵ

Conclusion

- Indirect moment conditions indeed provide a systematic method of choosing sufficient summary statistics

- An efficient way of weighting the different moments is presented

- A meaningful interpretation to the tolerance level ϵ is made available by normalizing the moments and using a chi–squared distance function
($\rightsquigarrow$ sensible assessment of how good the approximation to the true posterior is)

- As the results of our simulation example have shown, Indirect ABC is computationally efficient among available alternatives (e.g. GSM – Bayesian Indirect Inference)