Bayesian Synthetic Likelihood – Asymptotics and Misspecification

Professor Chris Drovandi

School of Mathematical Sciences and Centre for Data Science
ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS)
Queensland University of Technology

Collaborators: David Frazier, David Nott and Robert Kohn

8th September 2022

Bayesian Statistics

Let $y = (y_1, \ldots, y_n)^\top$ denote the observed data and define $P_0^{(n)}$ as the true distribution generating y.

The model $P_0^{(n)}$ is approximated using a parametric family of models $\{P_\theta^{(n)} : \theta \in \Theta \subseteq \mathbb{R}^{d_\theta}\}$

We wish to estimate θ via the posterior distribution:

$$\pi(\theta|y) \propto p_n(y|\theta)\pi(\theta).$$

where $\pi(\theta)$ is the prior density and $p_n(y|\theta)$ is the likelihood.

Markov Chain Monte Carlo

Construct ergodic Markov chain with invariant distribution $\pi(\theta|y)$

A common MCMC algorithm is Metropolis Hastings (MH) MCMC, where proposals θ^* are accepted with probability

$$\min\left(1, \frac{p_n(y|\theta^*)\pi(\theta^*)q(\theta|\theta^*)}{p_n(y|\theta)\pi(\theta)q(\theta^*|\theta)}\right),$$

where $q(\cdot)$ is the proposal density.

For complex models, $p_n(y|\theta)$ may be intractable.

QUT Centre for Data Science

Likelihood-free Inference

Likelihood-free methods approximate the posterior via model simulations.

Two common methods are approximate Bayesian computation (ABC) and Bayesian synthetic likelihood (BSL).

ABC and BSL often perform inference using summaries via the function: $S_n : \mathbb{R}^n \to \mathbb{R}^d$, $d \geq d_\theta$. We denote $S_n = S_n(y)$ when clear.

$$\pi(\theta|S_n) \propto g_n(S_n|\theta)\pi(\theta).$$

ABC and BSL approximate $g_n(S_n|y)$ in different ways.

Approximate Bayesian Computation

ABC[1] approximates the likelihood by drawing m mock datasets $z^1, \ldots, z^m \sim P_\theta^{(n)}$:

$$\hat{g}_\epsilon(S_n|\theta) = \frac{1}{m} \sum_{i=1}^{m} K_\epsilon[\rho\{S_n(y), S_n(z^i)\}].$$

where $\rho\{S_n(y), S_n(z_i)\}$ is the distance function, $K_\epsilon[\cdot]$ is a kernel weighting function and ϵ is the bandwidth or ABC tolerance.

Common choices are $m = 1$ and $K_\epsilon[\rho\{S_n(y), S_n(z)\}] = \mathbb{I}[\rho\{S_n(y), S_n(z) \leq \epsilon]$.

[1] Sisson et al (2018). Handbook of ABC.

QUT Centre for Data Science

Regression Adjustment

Regression adjustment (see [1] for a review) aims to improve the ABC approximation via the adjustment[2]:

$$\tilde{\theta}_i = \theta_i - (S_n(z_i) - S_n(y))^\top \hat{\beta},$$

where $\hat{\beta}$ is estimated from the regression:

$$\theta_i = \alpha + (S_n(z_i) - S_n(y)^\top \beta + e_i.$$

More sophisticated regressions are possible[1] (e.g. local linear regression and neural networks).

[1] Blum (2018). Chapter in Handbook of ABC.

[2] Li and Fearnhead (2018). Biometrika.

Limitations of ABC

Inferences can be sensitive to ϵ and ρ.

Scales poorly with summary statistic dimension[1].

Can be computationally inefficient.

Regression-adjusted ABC can perform poorly under misspecification[2].

[1] Blum (2010). JASA.
[2] Frazier et al (2020). JRSS B.

QUT Centre for Data Science

Theoretical Properties of ABC

Under correct model specification and $\epsilon \to 0$ at a "fast" rate ABC delivers[1][2]:

- Asymptotically normal posterior concentrating on θ_0 with correct coverage.
- Asymptotically normal posterior mean with correct coverage.
- Acceptance rate goes to 0.

Under correct model specification ABC regression adjustment[3] does not require $\epsilon \to 0$.

[1] Frazier et al (2018). Biometrika.
[2] Li and Fearnhead (2018a). Biometrika.
[3] Li and Fearnhead (2018b). Biometrika.

Bayesian Synthetic Likelihood

BSL[1,2] approximates $g_n(\cdot|\theta)$ using a Gaussian likelihood

$$g_A(S_n|\theta) = N\{S_n; b_n(\theta), \Sigma_n(\theta)\},$$

where $b_n(\theta)$ and $\Sigma_n(\theta)$ denote the mean and variance of the model summary statistic at θ.

We estimate $b_n(\theta)$ and $\Sigma_n(\theta)$ from m independent model simulations:

$$\overline{b}_n(\theta) = \frac{1}{m}\sum_{i=1}^{m} S_n(z^i), \quad \overline{\Sigma}_n(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[S_n(z^i) - \overline{b}_n(\theta)\right]\left[S_n(z^i) - \overline{b}_n(\theta)\right]^\top$$

[1] Wood (2010). Nature.

[2] Price et al (2018). JCGS.

Bayesian Synthetic Likelihood

The synthetic likelihood is then approximated as $N\{S_n; \overline{b}_n(\theta), \overline{\Sigma}_n(\theta)\}$.

MH-MCMC BSL targets the following posterior:

$$\overline{\pi}(\theta \mid S_n) \propto \pi(\theta)\overline{g}_n(S_n \mid \theta),$$

$$\overline{g}_n(S_n \mid \theta) = \int N\{S_n; \overline{b}_n(\theta), \overline{\Sigma}_n(\theta)\} \prod_{i=1}^{m} \mathrm{d}P_\theta^{(n)}\{S_n(z^i)\}\, \mathrm{d}S_n(z^1) \ldots \mathrm{d}S_n(z^m).$$

$\overline{g}_n(S_n \mid \theta)$ is the expectation of the estimated synthetic likelihood.

Asymptotic Properties of BSL[1] – Assumptions

Observed summaries satisfy a central limit theorem.

Define $b_n(\theta) = E[S_n(z)|\theta]$, require $b_n(\theta)$ to be continuous over Θ, able to identify θ_0 and whose derivative has full column rank at θ_0

In addition to the continuity of $\pi(\theta)$, we require the existence of a certain prior moment.

Model summaries are sub-Gaussian (can be weakened).

[1] Frazier et al (2022). JASA.

Asymptotic Properties of BSL[1] – Assumptions

Let $\Delta_n(\theta)$ be some covariance matrix estimator of summaries $S_n(z)$. Not necessarily the standard one.

We require that $\Delta_n(\theta)$ for large n is positive definite for θ sufficiently close to θ_0.

We require $v_n^2 \Delta_n(\theta)$ converges uniformly to $\Delta(\theta)$, which needs to be continuous and positive definite for θ sufficiently close to θ_0.

[1] Frazier et al (2022). JASA.

Asymptotic Properties of BSL – Results

BSL posterior is asymptotically normal.

BSL posterior mean is asymptotically normal.

Above results depend on choice of covariance matrix (BSL = ABC when using standard covariance estimator).

Above results satisfied for any $m = C\lfloor n^{\gamma} \rfloor$, with $C > 0$, $\gamma > 0$. Demonstrates that m does not strongly impact inferences.

Asymptotic Properties of BSL – Computational Efficiency

Based on a rejection sampler with a ‘good’ proposal.

BSL has a non-negligible acceptance rate asymptotically.

Similar computational efficient to regression adjusted ABC.

Misspecification of Covariance Matrix

Using a misspecified covariance matrix $\Delta_n(\theta) \neq \overline{\Sigma}_n(\theta)$ (e.g. shrinkage estimator) may lead to computational gains.

But it can lead to invalid uncertainty quantification.

We develop an adjustment method[1]: Consider the adjusted sample

$$\theta^{A,q} = \widehat{\theta} + \widehat{\Gamma}\widetilde{\Omega}^{1/2}\widehat{\Gamma}^{-1/2}(\theta^q - \widehat{\theta}), \quad q = 1, \dots, Q$$

θ^q is a sample based on misspecified covariance, $(\widehat{\theta}, \widehat{\Gamma})$ is estimated posterior mean/covariance and $\widetilde{\Omega}$ is an estimate of $\text{var}\left\{\nabla_\theta \log g_n(S_n|\widehat{\theta})\right\}$.

[1] Frazier et al (2022). JASA.

Misspecification of Covariance Matrix

Estimating $\widetilde{\Omega}$ when the model for y is correct.

1. For $j = 1, \ldots, J$, draw $S^{(j)} \sim P_{\hat{\theta}}^{(n)}$.
2. Approximate $g^{(j)} = \nabla_\theta \log g_n(S^{(j)}|\widehat{\theta})$. We estimated this based on a Gaussian process approximation of synthetic likelihood, trained on samples around $\widehat{\theta}$.
3. With $\bar{g} = J^{-1} \sum_{j=1}^{J} g^{(j)}$, return $\widetilde{\Omega} = \frac{1}{J-1} \sum_{j=1}^{J} (g^{(j)} - \bar{g})(g^{(j)} - \bar{g})^\intercal$

Gives asymptotically valid frequentist inference about θ_0.

MA(2) Example

We consider the second order moving average model (MA(2)):

$$y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2},$$

for $t = 1, \ldots, n$, where $e_t \sim N(0, 1)$, $t = -1, \ldots, n$, and n is the length of time series. Prior is uniform over $-1 < \theta_2 < 1, \theta_1 + \theta_2 > -1, \theta_1 - \theta_2 < 1$.

First 20 autocovariances as summaries.

We consider 100 datasets of size $n = 10^4$ simulated with true parameters $\theta_1 = 0.6$ and $\theta_2 = 0.2$

MA(2) Example – Results

method	m	mean ESS	min ESS	90%	tv dist	time (hrs)
BSL	200	3000	240	91/89/88	0.22	1.4
BSL diag	200	5400	1500	98/88/85	0.60	1.4
BSL adj	200	-	-	91/90/91	0.53	1.5
ABC	-	-	-	96/99/96	0.34	5.9
ABC reg	-	-	-	93/96/90	0.22	5.9

Table: Estimated coverage for credible intervals having nominal 90% credibility for standard BSL, BSL with a diagonal covariance (BSL diag), BSL diag with an adjustment (BSL adj), ABC and regression adjustment ABC (ABC reg) for $\theta_1/\theta_2/(\theta_1, \theta_2)$.

QUT Centre for Data Science

Misspecification in Synthetic Likelihood

What does misspecification mean in the context of synthetic likelihood?

Define $b_n(\theta) = E[S_n(z)|\theta]$. There does not exist any $\theta \in \Theta$ such that $b_n(\theta) = b_0$ where $b_0 = E[S_n(y)]$.

That is, there is no θ that recovers the observed summary. Referred to as *incompatibility*[1].

For synthetic likelihood, we say that the model is incompatible if

$$\lim_{n\to\infty} \inf_{\theta\in\Theta} \{b_n(\theta) - b_0\}^{\top} \{n\Sigma_n(\theta)\}^{-1} \{b_n(\theta) - b_0\} > 0. \tag{1}$$

[1] Marin et al (2014). JRSS B.

Consequences of Misspecification: Simple Example

Researcher believes the observed data $y_{1:n} = (y_1, \ldots, y_n)^\intercal$ is generated according to MA(1) model

$$y_t = e_t + \theta e_{t-1}, \quad t = 1, \ldots, n, \tag{2}$$

Actual data generating process (DGP) evolves according to the stochastic volatility (SV) model

$$y_t = \exp(h_t/2)u_t, \;\; h_t = \omega + \rho h_{t-1} + v_t \sigma_v. \tag{3}$$

Two summaries: variance and first autocovariance.

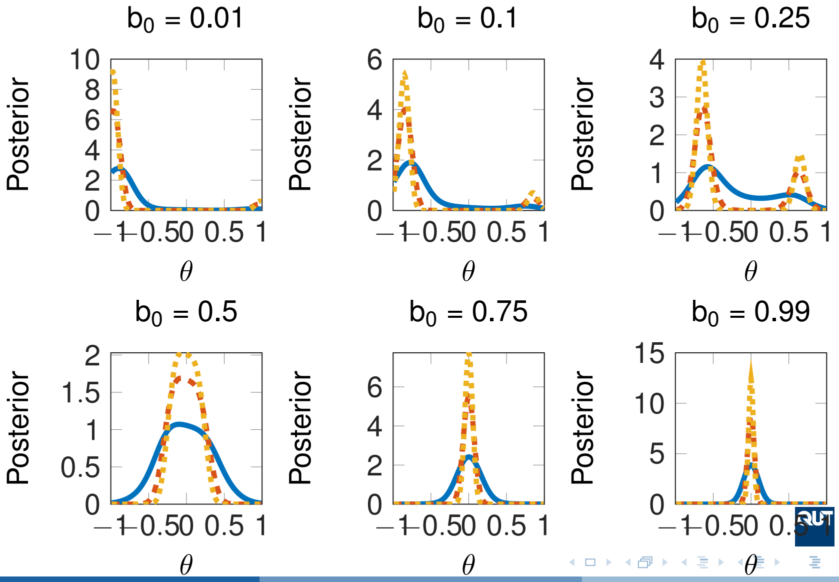
Simple Example

Under the MA(1) model we have $b(\theta) = (1 + \theta^2, \theta)^\top$.

Under the DGP we have $b_0 = \left(\exp\left(\frac{\omega}{1-\rho} + \frac{1}{2}\frac{\sigma_v^2}{1-\rho^2}\right), \quad 0\right)^\top$.

The unique minimum of $\|b(\theta) - b_0\|$ is achieved at $\theta = 0$, and we hope our inferences concentrate on this value.

Results under Misspecification



Tempering does not help

Tempering, i.e. raising the likelihood to power $\alpha < 1$ has often been to robustify Bayesian inference to misspecification.

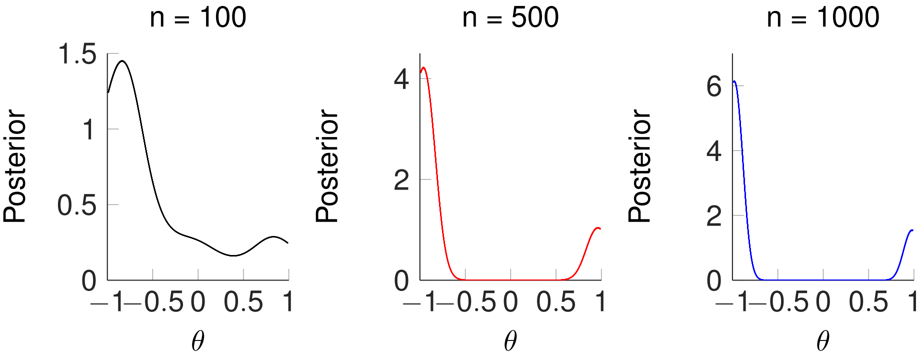
For $\alpha \geq 0$, the tempered synthetic likelihood:

$$\overline{g}_n^{\alpha}(S_n \mid \theta) = \int N\{S_n; \overline{b}_n(\theta), \overline{\Sigma}_n(\theta)\}^{\alpha} \prod_{i=1}^{m} \mathrm{d}P_{\theta}^{(n)}\{S_n(z^i)\}\, \mathrm{d}S_n(z^1) \ldots \mathrm{d}S_n(z^m)$$

which yields the posterior distribution $\overline{\pi}_{\alpha}(\theta \mid S_n) \propto \overline{g}_n^{\alpha}(S_n \mid \theta)\pi(\theta)$.

This doesn't help BSL...

Tempering does not help



Consequences of Misspecification[1]: Theoretical Results

Define:

$$M_n(\theta) = n^{-1}\partial \log g_n(S_n \mid \theta)/\partial\theta.$$

Incompatibility can result in multiple roots of $M_n(\theta) = 0$, producing multimodality in the BSL posterior.

In this case, the BSL posterior is not asymptotically Gaussian, but resembles a Gaussian density near the modes.

[1] Frazier et al (2022). https://arxiv.org/abs/2104.03436.

Consequences of Misspecification: Theoretical Results

Even under misspecification, it is possible $M_n(\theta) = 0$ has only a single solution.

This results in Gaussian-like posterior concentration around the pseudo-true value.

Overall, BSL and ABC have very different asymptotic behaviours under misspecification.

BSL Adjustments for Misspecification

Two adjustments[1] have been proposed for making BSL robust to misspecification: mean adjustment and variance inflation.

The main idea is to introduce auxiliary variables to make an incompatible model compatible again.

Here we focus on *variance inflation*.

[1] Frazier and Drovandi (2021). JCGS.

Variance Inflation

Variance inflation alters the covariance matrix of synthetic likelihood.

For $\Gamma = (\gamma_1, \ldots, \gamma_d)'$, define the regularized variance matrix $\overline{\Sigma}_n(\theta, \Gamma) = \overline{\Sigma}_n(\theta) + \overline{\Sigma}_n^{1/2}(\theta)\text{diag}\{\gamma_1, \ldots, \gamma_d\}\overline{\Sigma}_n^{1/2}(\theta)$.

Let

$$\overline{g}_n(S_n \mid \theta, \Gamma) = \int N\{S_n; \overline{b}_n(\theta), \overline{\Sigma}_n(\theta, \Gamma)\} \prod_{i=1}^{m} \text{d}P_\theta^{(n)} \left\{S_n(z^i)\right\} \text{d}S_n(z^1) \ldots \text{d}S_n$$

denote the synthetic likelihood based on $\overline{\Sigma}_n(\theta, \Gamma)$.

Variance Inflation

Γ ensures that $\|\overline{\Sigma}_n(\theta, \Gamma)^{-1/2}\{\overline{b}_n(\theta) - S_n\}\|$ can be made small even if there is no value in Θ where $\|\overline{b}_n(\theta) - S_n\|$ is small.

To regularise the new model, we impose a prior on Γ, $\pi(\Gamma)$, consisting of independent exponential densities.

The joint posterior is $\overline{\pi}(\theta, \Gamma \mid S_n) \propto \pi(\theta)\pi(\Gamma)\overline{g}_n(S_n \mid \theta, \Gamma)$, which we sample with MCMC.

Variance Inflation – MCMC Sampling

Component-wise MCMC scheme[1].

Update $\theta|\gamma$ using standard BSL Metropolis-Hastings.

Update each component of $\gamma|\theta$ using a slice sampler. Acceptance rate of 1, no additional tuning.

We don't seem to incur an additional penalty of sampling over a higher dimensional space.

[1] Frazier and Drovandi (2021). JCGS.

Variance Inflation – Theoretical Properties

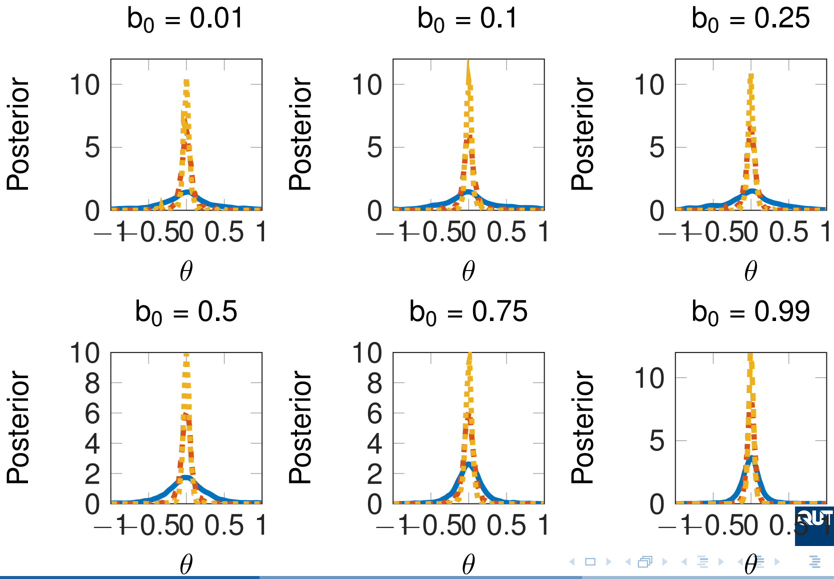
If the model is compatible, what behavior should we expect from the R-BSL approach?

Under compatibility and other mild assumptions, we show[1] that the posterior for Γ converges to the prior.

Thus incompatibility can be detected by departures from the prior.

[1] Frazier and Drovandi (2021). JCGS.

Variance Inflation – Results



Other BSL Extensions

Semi-parametric BSL – more robust to Gaussian assumption. An et al (2020) STCO.

Whitening BSL – decorrelate summaries to make shrinkage covariance more effective/accurate. Priddle et al (2022) JCGS.

Variational Bayes BSL – reduced model simulation at expensive of parametric posterior assumption. Ong et al (2018) CSDA.

Software packages – **BSL** package in R, BSL functionality in **ELFI** package in Python (coming soon). An et al (2022) Journal of Statistical Software.

Summary

BSL is an attractive method for likelihood-free inference:

- Has good asymptotic properties under correct specification.
- More computationally efficient than ABC and requires little tuning.
- Extensions have been developed to handle model misspecification.

Main limitations:

- Requires stronger condition on distribution of summaries compared to ABC.
- Generally remains model simulation intensive.

References

Frazier et al (2022). Bayesian inference using synthetic likelihood: asymptotics and adjustments. JASA.

Frazier et al (2021). Synthetic Likelihood in Misspecified Models. https://arxiv.org/abs/2104.03436.

Li and Fearnhead (2018a). On the asymptotic efficiency of approximate Bayesian computation estimators. Biometrika

Li and Fearnhead (2018b). Convergence of regression-adjusted approximate Bayesian computation. Biometrika

Frazier et al (2020). Model misspecification in approximate Bayesian computation: consequences and diagnostics. JRSS B.

Frazier and Drovandi (2021). Robust Approximate Bayesian Inference with Synthetic Likelihood. JCGS

Price et al (2018). Bayesian synthetic likelihood. JCGS.

Contact Details

Email: c.drovandi@qut.edu.au Web: chrisdrovandi.weebly.com Twitter: @chris_drovandi

QUT Centre for Data Science