Computing Bayes: Bayesian computation from 1763 to 2017!

Gael Martin

Department of Econometrics and Business Statistics

Monash University, Melbourne

Bayes on the Beach, November, 2017

In the Beginning.....1763

- The Royal Society, London, December 23, 1763:
- **Richard Price** read:

  An Essay Towards Solving a Problem in the Doctrine of Chances

  by

  Reverend Thomas Bayes

- Three years after Bayes' death
- '**Bayesian inference**' has its birth......

In the Beginning.....1763

- The **question posed**?
- If perform n Bernoulli trials, with $\theta =$ probability of 'success'
  - Rolling a ball across a 'billiard' table n times
  - $\Rightarrow$ 'success' if ball lands within a particular distance from the edge
- And record: $\mathbf{y} = (1, 1, 0, 1, 0, ...0....)'$
- What is:

$$\Pr ob(a < \theta < b|\mathbf{y})?$$

In the Beginning.....1763

- The **answer offered?**

$$\Pr ob(a < \theta < b|\mathbf{y}) = \int_a^b p(\theta|\mathbf{y})d\theta$$

- where:

$$p(\theta|\mathbf{y}) = \frac{L(\theta|\mathbf{y})p(\theta)}{p(\mathbf{y})} = \text{posterior pdf}$$

- with:

$$\begin{aligned} L(\theta|\mathbf{y}) &= \text{Bernoulli likelihood function} \\ p(\theta) &= \text{a uniform prior on } (0,1) \\ p(\mathbf{y}) &= \text{the marginal likelihood} \end{aligned}$$

In the Beginning.....1763

- First application (we think...) of '**inverse probability**'
- **Given** a set of *observations* ($\mathbf{y}$)
- Produced according to an assumed probability distribution
  - (Bernoulli here....)
- Can we **invert** the problem to make a **probability statement** about the *unknown and unobservable* θ?
- $\equiv$ '**Bayesian inference**' in our modern language....
- $\vdots$
- Computational challenge??

In the Beginning.....1763

- **Closed-form** solution for $p(\theta|\mathbf{y})$...(**beta** density)
- **However:**

$$\Pr ob(a < \theta < b|\mathbf{y}) = \int_a^b p(\theta|\mathbf{y})d\theta = \text{'}\textbf{incomplete beta function}\text{'}$$

- does **not** have a closed form!
- (and was not yet numerically tabulated!)
- And the fact that Bayes could not find an accurate numerical solution
- Has been proposed as a possible reason for his not publishing the work! (**Stigler (1983) 'The History of Statistics'**)
- $\Rightarrow$ **computational issues a feature of 'Bayesian inference' from its birth!!**

Reverend Thomas Bayes: 1701-1761:

- Why is a Presbyterian clergyman in the mid-1700's playing around with billiard balls and mathematics??

Protestant Reformation: 1517+

- October 31st 1517: Castle Church, Wittenberg, Germany
- Martin Luther (a monk) nails to the door: 95 'theses' or 'objections' to the workings of the Roman Catholic Church
- And so begins (the most publicized) break from the established Church of Rome
- The Swiss follow: Ulrich Zwingli, John Calvin (mid-1500s.....)
- All 'reformers' or protesters'....creating the new **Protestant movement**
- Stepping outside of the authority of the Pope
- Advocating a more personal connection with God
- Including ordinary people appointing their own pastors

Protestant Reformation: 1517+

- Across the English Channel?
- Tumultuous time....
- Henry the 8th/Mary 1st/Elizabeth 1st
- 'Protestants' (Church of England variety...) have ascendancy under Elizabeth
- Simultaneously, in Scotland, Calvin's brand of Protestantism spreads
- $\Rightarrow$ Presbyterians
- By Bayes time (1701-1761): 'Non-conformist' (e.g. Presbyterians) and Church of England clergy dotted throughout the British Isles
- $\Rightarrow$ Reverend Thomas Bayes preaching in Tunbridge Wells (England) 1734 +

The Scottish Enlightenment (1700s/1800s)

- An 'easy' gig! (**Bryson (2010) 'At Home: a Short History of Private Life'!!)**
- The odd sermon on Sunday...
- A fair bit of spare time!
- Time to explore ideas
- 'Gentleman' scholars
- (Bayes had studied both theology **and mathematics** at the University of **Edinburgh**)
- Ideas; discovery; questioning; scientific experimentation valued in the time of the **Enlightenment**
- .....so what we see with **Bayes** all makes sense......

Pierre-Simon Laplace: 1749–1827

- But **Bayes** dies early
- Work eventually publicized by Price....but appears to have disappeared from view thereafter
- Then along comes **Pierre**.....

Pierre-Simon Laplace: 1749–1827

- Appears to have discovered '**Bayes Theorem**' **independently** (1770 +)
- Applied method of **inverse probability** to several problems, with priors determined via more abstract reasoning
- Along the way introduced the **Laplace (analytical) approximation** to (Bayesian) integrals!
- $\Rightarrow$ **first computational solution to intractable Bayesian problems!!**
- The method of **inverse probability** remained dominant in the 1800s
    - (**Feinberg (2006), 'When did Bayesian Inference become "Bayesian"'**)

Pierre to Arnold......

- Somewhat usurped in the 1900s by ('frequentist') notions of:
  - **Maximum likelihood estimation** and associated 'sampling properties' (**Fisher, 1922**)
  - **Hypothesis testing**/p-values/confidence intervals (**Neyman/Pearson, 1930+**)
- Despite works on '**Bayesian inference**' by:
  - **De Finetti (1930, 1937)**
  - **Jeffreys (1939)**
  - **Savage (1954)**
  - **Lindley (1965, 1971)**
  - **Arnold Zellner (1971)**

State of Play in 'Bayesian Inference' in 1970s?

- **Zellner, 1971: 'Bayesian Inference in Econometrics'**
- Key aspects of coverage?
  - Gaussian (and associated) distributions dominate
  - natural conjugate priors
  - $+$ non-informative (Jeffreys) priors
  - $\Rightarrow$ analytical solutions for **posterior moments**
  - $\Rightarrow$ analytical solutions for **marginal posteriors**
  - $\Rightarrow$ analytical solutions for **marginal likelihoods**
  - $\Rightarrow$ analytical solutions for **predictives**

State of Play in 'Bayesian Inference' in 1970s?

- Some use of **low-dimensional (deterministic) numerical integration**
- (+ use of numerical tabulations of common integrals)
- Some use of **analytical approximations**
- **No** mention of **simulation-based computation**.....
- However.....

State of Play in Bayesian Computation in 1980s?

- Assumed DGPs (**models**) are becoming more **complex** and **high-dimensional**; e.g:
  - full models of the economy
  - more complex time series (e.g. unit root/cointegration) models
  - latent variable (including state space) models
  - ⋮
- Neither Bayes with **deterministic numerical integration**
- Nor Bayes with **analytical approximations**
- was viable as a **general** inferential method
- Plus, computers speeding up!
- Enter stage left: **simulation-based computation........**

What IS the computational challenge in Bayes?

- Virtually all quantities of interest in Bayesian statistics can be expressed as:

$$E(g(\boldsymbol{\theta})|\mathbf{y}) = \int_{\boldsymbol{\theta}} g(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

- for some $g(\boldsymbol{\theta})$:

$$\begin{aligned}
E(\boldsymbol{\theta}|\mathbf{y}) &= \int_{\boldsymbol{\theta}} \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \\
p(\theta_1|\mathbf{y}) &= \int_{\boldsymbol{\theta}} p(\theta_1|\boldsymbol{\theta}_{-1}, \mathbf{y})p(\boldsymbol{\theta}_{-1}|\mathbf{y})d\boldsymbol{\theta}_{-1} \\
\Pr ob(a &< \boldsymbol{\theta} < b|\mathbf{y}) = \int_{\boldsymbol{\theta}} \mathbf{I}_{(a<\boldsymbol{\theta}<b)} p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \\
p(y_{T+1}|\mathbf{y}) &= \int_{\boldsymbol{\theta}} p(y_{T+1}|\boldsymbol{\theta}, \mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}
\end{aligned}$$

- **all** $\equiv E(g(\boldsymbol{\theta})|\mathbf{y})$ for **some** $g(\boldsymbol{\theta})$

What IS the computational challenge in Bayes?

- i.e **implementing Bayes is all about evaluating integrals!!!**
- $\equiv E(g(\boldsymbol{\theta})|\mathbf{y})$ for **some** $g(\boldsymbol{\theta})$
- Only when assuming **simple** models
  - and **standard** - including natural conjugate - priors
- will such integrals ($\equiv$ expectations) be available in closed form!
- For most empirically realistic models
- The integrals need to be **estimated** in some way......
- **Three** main options:

Bayesian Numerical Methods

1. **Deterministic** numerical integration methods:

$$\int_{\boldsymbol{\theta}} g(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = \int_{\theta_1}\int_{\theta_2}..\int_{\theta_p} g(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \approx \sum^{G}\sum^{G}..\sum^{G}......$$

- Computational burden $= G^p$
- **'curse' of dimensionality**
- $\Rightarrow$ **no good in high-dimensional case**!

Bayesian Numerical Methods

2. **Analytical approximation** of the **integrand**: $\Rightarrow$ closed-form integrals

    - 'Laplace' method
    - Integrated Nested Laplace (INLA) method
    - Variational Bayes

- All feasible, but only ever produce **approximate** results

Bayesian Numerical Methods

3. **Stochastic simulation (or sampling)** methods

- With modern computing power: 'exact' solutions are attainable
- Plus: a very natural way of thinking about **the estimation of an expectation**
- $\Rightarrow$ the dominant approach in the literature.....

Bayesian Simulation Methods

Overview

- Given:

$$E(g(\boldsymbol{\theta})|\mathbf{y}) = \int_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$$

  for some $g(\boldsymbol{\theta})$

- **All simulation** methods involve:
  - **sampling** from $p(\boldsymbol{\theta}|\mathbf{y})$
  - and using that sample to estimate $E(g(\boldsymbol{\theta})|\mathbf{y})$
- From Statistics 101: we estimate a **population mean** with a **sample mean!!**
- So, at the end of the day we will (usually) do two **simple** things:

1. Construct a **sample mean** of some function of M posterior draws:
$$\overline{g(\boldsymbol{\theta})} = \frac{1}{M}\sum_{j=1}^{M} g(\boldsymbol{\theta}^{(j)})$$
2. (Legitimately) use frequentist concepts to:
   - Construct a **standard error** that measures the **accuracy** of $\overline{g(\boldsymbol{\theta})}$ as an estimate of $E(g(\boldsymbol{\theta})|\mathbf{y})$
   - **WLLN** $\Rightarrow$ **consistency** of $\overline{g(\boldsymbol{\theta})}$ as an estimate of $E(g(\boldsymbol{\theta})|\mathbf{y})$ (as $M \to \infty$)
   - **CLT** $\Rightarrow$ **asymptotic normality** of $\overline{g(\boldsymbol{\theta})}$ (as $M \to \infty$)

- The **hard** part? Getting the draws from $p(\boldsymbol{\theta}|\mathbf{y})$!

Bayesian Simulation Methods

Independent sampling: Monte Carlo sampling

- An **independent** sample from $p(\boldsymbol{\theta}|\mathbf{y})$ is ideal: each new draw brings 'fresh' information about $p(\boldsymbol{\theta}|\mathbf{y})$
  - $\Rightarrow$ high accuracy $\equiv$ small (simulation) standard error
- **Monte Carlo** sampling produces an independent sample from $p(\boldsymbol{\theta}|\mathbf{y})$ **directly**
  - Great when $p(\boldsymbol{\theta}|\mathbf{y})$ is of a standard form but $E(g(\boldsymbol{\theta})|\mathbf{y})$ is not!
  - Think of **Bayes** and his **beta probability**!
- But complex model $\Rightarrow$ complex $L(\boldsymbol{\theta}|\mathbf{y}) \Rightarrow p(\boldsymbol{\theta}|\mathbf{y})$ non-standard
- $\Rightarrow$ for **realistic** models:

$$\boxed{p(\boldsymbol{\theta}|\mathbf{y}) \textbf{ cannot be simulated from directly}}$$

- Enter **importance sampling......**

- **Kloek and (Herman) van Dijk (1978)**
- **Dutch** econometricians. Why?
- Back to the Protestant reformation!!
- **1568** - the (mainly) Protestant Dutch threw off the their imperial overlord: the Catholic Spanish
- Struck out independently.....invented the powerful mercentile state
  - $\Rightarrow$ a strong tradition in **economics/econometrics**
  - $\Rightarrow$ *Econometric Institute of the Erasmus University Rotterdam*
  - **Kloek** and **van Dijk**
- But back to the integrals!!!

- **Importance sampling:** Simple idea!
- Say have $q(\boldsymbol{\theta}|\mathbf{y}) \approx p(\boldsymbol{\theta}|\mathbf{y})$, and from which we **can sample**
- Estimate $E(g(\boldsymbol{\theta})|\mathbf{y})$ as

$$\overline{g(\boldsymbol{\theta})}^{IS} = \sum_{j=1}^{M} \left(g(\boldsymbol{\theta}^{(j)}) w(\boldsymbol{\theta}^{(j)}) \right) / \sum_{j=1}^{M} w(\boldsymbol{\theta}^{(j)})$$

  - Using draws of $\boldsymbol{\theta}$ from the **importance density** $q(\boldsymbol{\theta}|\mathbf{y})$

- where: $w(\boldsymbol{\theta}^{(j)}) = p^*(\boldsymbol{\theta}^{(j)}|\mathbf{y}) / q(\boldsymbol{\theta}^{(j)}|\mathbf{y})$
- for some **kernel** p^* of p:

$$p(\boldsymbol{\theta}|\mathbf{y}) = c \times p^*(\boldsymbol{\theta}|\mathbf{y}) \propto L(\boldsymbol{\theta}|\mathbf{y}) \times p(\boldsymbol{\theta})$$

- **Need to be able to evaluate $L(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})$**

- Great!! Problem solved??
- As long as we can write down the assumed **DGP** we are in business?
- Ummm.....how to choose $q(\boldsymbol{\theta}|\mathbf{y})$ to 'match' $p(\boldsymbol{\theta}|\mathbf{y})$ when $\boldsymbol{\theta}$ is of high dimension???
- **Light bulb moment!**
- Why not break a **high**-dimensional problem down into a sequence of **lower**-dimensional problems??

Bayesian Simulation Methods

Gibbs sampling

- **Geman and Geman (1984), Gelfand and Smith (1990)**
- Simple (and **revolutionary**!) idea:
- **Hard** to sample from a (complex) **joint** posterior
- **Easier** to sample from (lower dimensional; simpler) **conditional** posteriors
- Why?
- **Conditioning** always makes life easier
- Something that is **unknown** is treated (temporarily....) as **known**
- + **Low-dimensional** problems easier to deal with

- E.g., say have $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)'$

$$p(\boldsymbol{\theta}|\mathbf{y}) = p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})$$

- Draw $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ **iteratively** from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})$ and $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})$
- Under regularity $\Rightarrow$ yields draws from the **joint**: $p(\boldsymbol{\theta}|\mathbf{y})$
- Cost??
- Drawing sequentially via the conditionals creates **dependence** in the sample
- $\Rightarrow$ a **Markov chain** with **invariant distribution** equal to $p(\boldsymbol{\theta}|\mathbf{y})$
- **Gibbs** an example of a **Markov chain Monte Carlo (MCMC)** algorithm

Bayesian Simulation Methods

Gibbs sampling

- $\Rightarrow$ Need to verify conditions for convergence to $p(\boldsymbol{\theta}|\mathbf{y})$
- $\Rightarrow$ Need to monitor convergence (and 'burn-in') in practice.....
- $\Rightarrow$ Need more draws to produce the same level of accuracy as an independent sample
- All that done though.....once we have the draws we do the usual simple things with them
- (Standard error formulae simply reflect the dependence in the draws)
- Gibbs sampling a good starting point in many complicated models
- Exploits the simplicity that comes from conditioning

- Take, for e.g. a **state space** model
  - with '**static**' parameters $\boldsymbol{\theta}_1$ and **random** parameters $\boldsymbol{\theta}_2$ $(\dim(\boldsymbol{\theta}_2) \geq n!)$
- $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y})$ will not be amenable to analytical treatment
- But:
  - $p(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2, \mathbf{y})$ is often simple (reflecting a **linear regression** structure)
  - $p(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1, \mathbf{y})$ exploits **filtering techniques**
- Can also introduce **auxiliary** latent variables in order to produce simple conditionals
- $\Rightarrow$ integrated out via the Gibbs procedure.....
- $\Rightarrow$ draws on parameters of interest retained

- Introduced by **Tanner and Wong (1987)** **as 'data augmentation'**
- **Note:**
- **For** $p(\theta_1|\theta_2, y)$ **and** $p(\theta_2|\theta_1, y)$
- **to be standard enough to be simulated from directly**
- $L(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)$ **needs to be available**

Bayesian Simulation Methods

Gibbs sampling

- **Important:** even when DGP **is** available
- Typically, not **all** conditionals are standard and hence **can** be drawn from!
- (e.g. $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})$ in a **non-linear state space model**)
- Trick? Draw from it **indirectly**
- By inserting another MCMC chain **within Gibbs:**
- **Metropolis-Hastings (MH)**
    - **Metropolis (1953)** - Los Alamos (US)....nuclear physicists....inventors of the atomic bomb.....
- **Magic!** Insertion produces a **hybrid** chain with $p(\boldsymbol{\theta}|\mathbf{y})$ still the **invariant** distribution.....

- The thrust of **MH within Gibbs** (applied to $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})$ say)
  - Draw from $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})$ via a **candidate** $q(\boldsymbol{\theta}_2) \approx p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})$
  - (Note the **dimension reduction** via **Gibbs**.....)
  - Accept **candidate** draw of $\boldsymbol{\theta}_2$ with a **probability** that depends on the ratio:

$$\frac{p^*(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})}{q(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})}$$

- $$\boxed{\textbf{Need to be able to evaluate } p^*(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y}) \Leftrightarrow}$$

- $$\boxed{\textbf{Need to be able to evaluate } L(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})}$$

Bayesian Simulation Methods

Pseudo-marginal MCMC

- **In summary......all methods so far:**
-

  $$\boxed{\textbf{Require the evaluation of: } L(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})!!}$$

- What if this is not the case? E.g. continuous time models with **unknown** transitions ($p(y_t|y_{t-1}, \boldsymbol{\theta})$?
- Or, if $\dim(\mathbf{y})$ is so large in dimension that evaluation of $p(\mathbf{y}|\boldsymbol{\theta})$ (product of n terms....) is essentially infeasible??
- For *some* problems rescue comes via the magic of an **unbiased estimate** of $p(\mathbf{y}|\boldsymbol{\theta})$!
- and the use of so-called **pseudo-marginal MCMC** methods

- Say have some:

$$\widehat{p(\mathbf{y}|\boldsymbol{\theta})}$$

- where:

$$E_{\mathbf{u}}\left[\widehat{p(\mathbf{y}|\boldsymbol{\theta})}\right] = p(\mathbf{y}|\boldsymbol{\theta})$$

- $\mathbf{u} =$ the auxiliary **random variables** underpinning the estimate
- and are intimately related to model-specific latent random variables
- Make this additional source of uncertainty explicit:

$$\widehat{p(\mathbf{y}|\boldsymbol{\theta})} = g(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})$$

Bayesian Simulation Methods

Pseudo-marginal MCMC

- Apply usual trick $\Rightarrow$ augment the 'unknowns' of the problem with $\mathbf{u}$:

$$g(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) \propto g(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})g(\mathbf{u})p(\boldsymbol{\theta})$$

- $\Rightarrow$ apply MCMC to the augmented space $(\boldsymbol{\theta}, \mathbf{u})$
- $\Rightarrow$ produce **marginal** inferences about $\boldsymbol{\theta}$
- ('**pseudo**' due to the true likelihood not being used....)
- What do we get?
- Is $p(\boldsymbol{\theta}|\mathbf{y})$ the **invariant distribution** of an MCMC algorithm applied to $(\boldsymbol{\theta}, \mathbf{u})$?
- **Yes! Due to the unbiasedness of likelihood estimate**!
    - **Beaumont, 2003**, **Andrieu and Roberts, 2009**

- **Pseudo-marginal** applied in a **state space model**?
  - $\Rightarrow$ **particle filtering**-based estimate of $p(\mathbf{y}|\boldsymbol{\theta}_1)$
  - $\Rightarrow$ $\mathbf{u}$ = vector of uniforms driving the **particle filter**
  - $\Rightarrow$ **Particle MCMC (PMCMC)** **Andrieu et al. 2010**
-

Releases the burden of having to

evaluate <u>all</u> components of $p(\mathbf{y}|\theta)$

- E.g. some filtering methods require only **simulation** from the *transition* densities
- But *measurement* densities still need to be evaluated (in the particle weights)

- Finally, **pseudo-marginal MCMC** has been applied specifically to reduce **computational load** associated with evaluating $p(\mathbf{y}|\boldsymbol{\theta})$ when $\dim(\mathbf{y})$ is large
- ‘**Big data**’
- **Quiroz, Villani, Kohn and Tran 2017** **subsample** the data to produce an **unbiased** estimate of $p(\mathbf{y}|\boldsymbol{\theta})$
- $\Rightarrow \widehat{p(\boldsymbol{\theta}|\mathbf{y})}$

'Exact' Bayesian Inference

- All done??
- Have access to **multiple** simulation-based methods: **MC/IS/MCMC/PM-MCMC**
- to produce $\widehat{p(\boldsymbol{\theta}|\mathbf{y})}$
- i.e. **exact Bayesian inference** (up to simulation error)
- But.....how to conduct posterior inference on $\boldsymbol{\theta}$ when:
  - The DGP $p(\mathbf{y}|\boldsymbol{\theta})$ is **intractable** in a way that precludes use of **exact** (including **pseudo-marginal**) methods?
  - Or the dimension of $\boldsymbol{\theta}$ so large that exploration/marginalization **infeasible** via **exact** methods?
  - Or when the expertise to produce a finely-tuned efficient **exact** algorithm is not available?
- Can/must resort to **approximate Bayesian inference**

'Approximate' Bayesian Inference

- **Goal** then is to produce **an approximation to** $p(\boldsymbol{\theta}|\mathbf{y})$:

  (i) Approximate Bayesian computation (**ABC**)

  (ii) Bayesian Synthetic likelihood

  - (i) and (ii) nested under **3. Simulation methods**

  (iii) Variational Bayes

  (iv) Integrated nested Laplace (INLA)

  - (iii) and (iv) nested under **2. Analytical approximation methods**

(i) Approximate Bayesian Computation

- Aim is to produce **draws** from an **approximation** to $p(\boldsymbol{\theta}|\mathbf{y})$
- and use draws to **estimate** that **approximation**
- The simplest (accept/reject) form of the algorithm:
  1. Simulate $(\boldsymbol{\theta}^i)$, $i = 1, 2, ..., N$, from $p(\boldsymbol{\theta})$
  2. Simulate pseudo-data $\mathbf{z}^i$, $i = 1, 2, ..., N$, from $p(\mathbf{z}|\boldsymbol{\theta}^i)$
  3. Select $(\boldsymbol{\theta}^i)$ such that:

$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}^i)\} \leq \varepsilon$$

    - $\boldsymbol{\eta}(.)$ is a (vector) **summary statistic**
    - $d\{.\}$ is a distance criterion
    - the tolerance ε is arbitrarily small
- (Recent reviews: **Marin, Publo, Robert and Ryder, 2011; Sisson and Fan, 2011; Drovandi, 2017**)

(i) Approximate Bayesian Computation

- **Note:**

  $$\boxed{\textbf{Evaluation of } L(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta) \textbf{ is not required}}$$

- $$\boxed{\textbf{Only simulation of } p(\mathbf{y}|\theta) \textbf{ is required}}$$

- In practice: $\boldsymbol{\eta}(.)$ is never **sufficient** $\Rightarrow$

- i.e. $\boldsymbol{\eta}(.)$ does not reproduce information content of $\mathbf{y}$

- Selected draws (as $\varepsilon \longrightarrow 0$) estimate $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ (**not** $p(\boldsymbol{\theta}|\mathbf{y})$)

- **Selection** of $\boldsymbol{\eta}(.)$

- And hence, **proximity** of $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ to $p(\boldsymbol{\theta}|\mathbf{y})$ still an open and hot topic!

- More after tea!!!!

(ii) Bayesian Synthetic Likelihood

- **ABC** attempts to estimate $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ via simulation
- Given:
$$p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y})) \propto p(\boldsymbol{\eta}(\mathbf{y})|\boldsymbol{\theta})p(\boldsymbol{\theta})$$
- in essence ABC approximates $p(\boldsymbol{\eta}(\mathbf{y})|\boldsymbol{\theta})$ via simulation, as:
$$p(\boldsymbol{\eta}(\mathbf{y})|\boldsymbol{\theta}) \approx \frac{1}{N}\sum_{i=1}^{N}\mathcal{I}\left(d\{\boldsymbol{\eta}(\mathbf{y}),\boldsymbol{\eta}(\mathbf{z}^i)\} \leq \varepsilon\right)$$

for the accept/reject version

- **BSL** (Price, Drovandi, Lee and Nott, 2017):
- Approximate $p(\boldsymbol{\eta}(\mathbf{y})|\boldsymbol{\theta})$ as:

$$p_S(\boldsymbol{\eta}(\mathbf{y})|\boldsymbol{\theta}) \approx N(\boldsymbol{\mu}_N(\boldsymbol{\theta}), \boldsymbol{\Sigma}_N(\boldsymbol{\theta}))$$

  where $\boldsymbol{\mu}_N(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_N(\boldsymbol{\theta})$ are computed from N simulated draws of $\boldsymbol{\eta}(\mathbf{z})$ from $p(\mathbf{z}|\boldsymbol{\theta}^i)$, **for a given $\boldsymbol{\theta}$**
- Draws from $p_S(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ are obtained by embedding $p_S(\boldsymbol{\eta}(\mathbf{y})|\boldsymbol{\theta})$ within (say) an MCMC algorithm
- **Both ABC** and **BSL** can thus be seen as versions of **pseudo-marginal** methods!
  - although inference is only ever conditional on $\boldsymbol{\eta}(\mathbf{y})$ (not $\mathbf{y}$)
  - and hence is only ever approximate......
- **Note, again:**

$$\boxed{\textbf{Only simulation of } p(\mathbf{y}|\boldsymbol{\theta}) \textbf{ is required}}$$

(iii) Variational Bayes

- Simultaneous with the development of new (**simulation-based**) **approximation** methods by **statisticians/econometricians**
- **Computer science/machine learning** community have been developing their own (**deterministic) approximation** tool:
- **Variational inference/Variational Bayes**
- In the spirit of **calculus of variations** $\Rightarrow$
- **Approximate** $p(\boldsymbol{\theta}|\mathbf{y})$ by some $q^*(\boldsymbol{\theta}) \in Q$ s.t:

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta}) \in Q}{\arg\min} KL\left(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta}|\mathbf{y})\right) = E_{q(\boldsymbol{\theta})}\left[\log\left(\frac{p(\boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta})}\right)\right]$$

  - Nice reviews by **Ormerod and Wand, 2010 and Blei, Kucukelbir and McAuliffe, 2017**

- Approximating $p(\boldsymbol{\theta}|\mathbf{y})$ via **simulation** replaced by
- Approximating $p(\boldsymbol{\theta}|\mathbf{y})$ via **optimization**

- Trade-off between:
  - Choosing q to be flexible enough to capture features of $p(\boldsymbol{\theta}|\mathbf{y})$
  - Choosing q to be tractable enough to enable efficient optimization

- **Problem?** If don't know $p(\boldsymbol{\theta}|\mathbf{y})$ how can we approximate it via:

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta})\in Q}{\arg\min} KL\left(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})\right)???$$

- $\Rightarrow$ **minimizing** $KL \equiv$ **maximizing:**

$$E_{q(\boldsymbol{\theta})}\left[\log\left(\frac{p(\mathbf{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right)\right]$$

  - where $p(\mathbf{y},\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ is (assumed to be) available!

(iii) Variational Bayes

- Further:

$$E_{q(\boldsymbol{\theta})}\left[\log\left(\frac{p(\mathbf{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right)\right] \leq \log p(\mathbf{y}) = \int_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

- $\Rightarrow$ a lower bound for the **marginal likelihood** (or **'evidence'**)
- used as an approximation to $p(\mathbf{y})$
  - which would typically be approximated as an **additional step** in simulation (e.g. MCMC) settings
- Critically: to implement **VB**:

$$\boxed{\textbf{Evaluation of } p(\mathbf{y}|\theta) \textbf{ is required!}}$$

(iii) Variational Bayes

- Note however!

1. **Barthelme and Chopin, 2014, 'Expectation-Propagation for Likelihood-Free Inference'**
   - Use of **VB** principles to implement **ABC**
2. **Tran, Nott and Kohn, 2106, 'Variational Bayes with Intractable Likelihood'**
   - Use an **unbiased estimate** of $p(\mathbf{y}|\boldsymbol{\theta})$ within **VB**
3. **Ong, Nott, Tran, Sisson and Drovandi, 2106, 'Variational Bayes with Synthetic Likelihood'**
   - Use a **synthetic likelihood** estimate of $p(\mathbf{y}|\boldsymbol{\theta})$ within **VB**
   -

All loosen the requirements on the tractability of $p(\mathbf{y}|\boldsymbol{\theta})$

(iv) INLA

Remember Pierre?

- Yet *another* stream of **approximate inference** builds on **Pierre's** simple idea for approximating an integral:

$$\int_x e^{\{nf(x)\}} dx \approx e^{\{nf(\widehat{x})\}} \int_x e^{\left\{ \frac{-n|f''(\widehat{x})|}{2}(x-\widehat{x})^2 \right\}} dx$$
$$= e^{\{nf(\widehat{x})\}} \sqrt{\frac{2\pi}{n|f''(\widehat{x})|}}$$

- **Optimization** needed to obtain $\widehat{x}$
- Building on **Laplace (1774) and Tierney and Kadane, 1986**
- **Rue, Martino and Chopin, 2009**
- apply this idea to a very broad class of models:
- '**latent Gaussian models**' (or '**Gaussian process models**')

- $\Rightarrow$ **Integrated Nested Laplace (INLA)** approx. of $p(\boldsymbol{\theta}|\mathbf{y})$
    - A combination of (nested) **Laplace (LA)** approximations
    - Plus a (low-dimensional) numerical integration (**IN**) step
- Again: **INLA** (like **VB**) amounts to replacing **simulation** by **optimization**
    - $\Rightarrow$ much attention given to the matter of **numerical opt.** in the given model class
    - The **optimization** in **INLA** being over a high dimensional vector of latent states.....
- Critically, the application of **INLA:**

$$\boxed{\textbf{Requires the evaluation of } p(\mathbf{y}|\boldsymbol{\theta})!}$$

    - (Augmentation with other methods for dealing with the case where $p(\mathbf{y}|\boldsymbol{\theta})$ is **intractable** is surely possible.......)

The 21st Century and Beyond?

- So......where are we heading now?
- What does this wealth of computational developments mean: for the future of statistical inference?
- Back in 2008 I had just finished reading: *'The Story of French'*
- An historical perspective on the language and its place in the world
- Coincidentally, I was asked to name and chair a debate between Christian Robert (**Bayesian**) and Russell Davidson (**frequentist**), entitled:

The 21st Century Belongs to Bayes

- Certain analogies between language and statistical paradigm became clear!

The 21st Century and Beyond?

-
$$\begin{aligned}\text{French} &= \text{Linga franca until 20th century}\\ &= \text{characterized by clear, coherent rules of grammar}\\ &= \text{characterized by a strong sense of correct usage}\\ &= \text{Bayesian inferential paradigm!}\end{aligned}$$

-
$$\begin{aligned}\text{English} &= \text{Linga franca in 20th century} +\\ &= \text{evolved quite differently}\\ &= \text{freely borrowing from many other languages}\\ &= \text{an amalgam of different approaches and structures}\\ &= \text{Classical/frequentist inferential paradigm!}\end{aligned}$$

The 21st Century and Beyond?

- According to the last chapter in *The Story of French'*,
- the authors bravely assert that **in the 21st century** the elegant, logical and coherent language of French may regain its preeminence!
- Is it the same with **Bayes** ???
- **In particular** - now armed as it is with this immense array of new computational tools!

The 21st Century and Beyond?

- So elegance and coherence in approach:
- Quantifying uncertainty about what is **unknown** conditional on what is **known** using the language of probability: $p(\boldsymbol{\theta}|\mathbf{y})$
- Underpinned by the ability to compute $p(\boldsymbol{\theta}|\mathbf{y})$
- Whether '**exactly**' or in some '**approximate**' fashion
- in almost every imaginable situation......
- Surely, our man in 1700's England with the billiard balls and the time to explore ideas.....
- Is now our man for the 21st century and beyond.....