Bayesian Model Choice via Mixture Distributions

With application to epidemics and population process models

Theo Kypraios
University of Nottingham
http://www.maths.nott.ac.uk/∼tk

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA
Mathematical Sciences
Mathematical Sciences Main Entrance

School of Mathematics and Statistics @ Newcastle University

Bayesian model choice via mixture distributions with application to epidemics and population process models

P. D. O'Neill, T. Kypraios
School of Mathematical Sciences, University of Nottingham, UK

Outline

This is talk is about a new and quite general method for model choice within a Bayesian framework, and in particular, how to compute Bayes Factors.

Plan of the talk:

- Motivation.
- General Theory.
- Computational matters.
- Examples.
- Discussion

Motivation

Motivation

- Our motivation comes from the setting of partially observed population processes in which missing data imputation is typically required to facilitate inference.

- For example, given a sequence of event times: Homogenous vs non-homogeneous Poisson process?

- Alternatively, given some disease outbreak data: which of two competing stochastic epidemic model is more appropriate?

- The method itself is quite general, and can also be applied to data consisting of independent and identically distributed observations, or to data arising from more complex scenarios.

Motivation

- In a Bayesian framework, questions of model choice can be addressed using Bayes factors, which quantify the relative likelihood of any two models given the data and within-model prior distributions.
- Bayes factors can suffer from two practical drawbacks, namely:
    1. they can be difficult to compute, and
    2. they can be highly sensitive to the choice of within-model prior distributions, and in particular apparently natural choices can give misleading results.
- Here our focus is towards addressing the first difficulty.
- Alternative methods of Bayesian model assessment have their own difficulties in the setting we consider.

Existing Methods

- For example, neither the Deviance Information Criterion (DIC) nor Bayesian Information Criterion (BIC) appear entirely natural for settings where the data are typically far from being independent observations, as is the case when the data are realisations of a stochastic process.

- For problems involving missing data, such as the epidemic example above, it is not even clear how suitable information criteria should best be defined Celeux et al (2006) give nine candidates, for instance).

- Finally, methods involving a comparison between the observed data and what the fitted model would predict typically involve a subjective judgement as to precisely what should be compared, and how.

Reversible Jump

- In all but the simplest cases, Bayes factors must be evaluated numerically. In principle, this can be achieved via reversible jump Markov chain Monte Carlo methods (Green 1995).

- To be precise, consider two models M_1 and M_2 with parameters θ_1 and θ_2, respectively, where $\theta_j \in \Theta_j$.

- Define $k \in \{1, 2\}$ to be a model indicator which specifies the model under consideration.

- Reversible jump methods proceed by defining a Markov chain with state space $\{1\} \times \Theta_1 \cup \{2\} \times \Theta_2$ such that the proportion of time for which $k = j$ converges to the posterior model probability $P(M_j|x)$, where x denotes the observed data.

Reversible Jump: Challenge

Given model prior probabilities $P(M_j)$, the Bayes factor in favour of model 1 is given by the expression

$$P(M_2)P(M_1|x)/P(M_1)P(M_2|x),$$

which can be estimated from the Markov chain output.

The main practical challenge in implementing reversible jump algorithms is constructing efficient between-model proposal distributions.

In other words, defining how the Markov chain jumps between the different components of the union of model parameter spaces.

Proposed Framework

Main Idea and Roadmap

- We propose a new method for evaluating Bayes factors which goes some way to removing the implementation challenges of reversible jump methods.
- The key idea is to consider a hypermodel which is itself a mixture model whose components are the two or more competing models of interest.
- A Markov chain Monte Carlo algorithm can then be defined on the product space of all model parameters and mixture probabilities.
- Our key result shows that the Bayes factors for the models can be expressed in terms of the posterior means of the mixture probabilities, and thus estimated from the MCMC output.
- Our methods allow the incorporation of missing data, and for model parameters to be shared between models.

Is that Really a New Idea?

Defining a Markov chain on a product (rather than union) of model-parameter spaces is the approach pioneered by Carlin and Chib (1995), and further developed to more general settings Green and O'Hagan (1998), Dellaportas et al. (2002) and Godsill (2001).

- This approach, as for RJMCMC methods, involves defining a probability distribution over the set of possible models, and introduces a parameter which indicates which model is chosen.
- The product-space approach also relies on defining so-called pseudo-priors for the within-model parameters, upon which algorithm efficiency is crucially dependent, and this can be difficult in practice.

In our setting:

- The possible models are combined into one mixture model.
- Our methods do not involve the need to introduce such pseudo-priors, although for some missing data problems we need to specify similar prior distributions for the missing data.

A Different Kind of Mixture

- The typical situation under consideration is one in which the data are assumed to comprise independent and identically distributed observations, each of which comes from the proposed mixture distribution(s).

- In contrast, for our methods we consider only one observation from the mixture model, but this single observation consists of all the observed data.

- In other words, rather than assuming each data point can individually come from any model in the mixture, we assume all data points come from one of the models in the mixture.

- This assumption is completely natural when the data are observations from a stochastic process and do not consist of independent points.

- However, our methods apply equally well to independent and identically distributed observations as it will be illustated later on.

Mixture Model Without Missing Data

- Suppose we observe data x, and wish to consider n competing models $M_1, \ldots, M_n$.
- For $i = 1, \ldots, n$ denote the probability density of x under model i by $\pi_i(x|\theta_i)$, where θ_i denotes the vector of within-model parameters, and set $\theta = (\theta_1, \ldots, \theta_n)$.
- We assume that all the $\pi_i(x|\theta_i)$ are densities with respect to the same common reference measure.
- Define a mixture model by

$$\pi(x|\alpha, \theta) = \sum_{i=1}^{n} \alpha_i \pi_i(x|\theta_i), \tag{1}$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)$ satisfies $\sum_{i=1}^{n} \alpha_i = 1$ and $\alpha_i \geq 0$ for $i = 1, \ldots, n$.

Computing Bayes Factors

We now show how Bayes factors can be computed directly from certain summaries of the posterior distribution of α given the data x.

We assume that α and θ are independent *a priori*, and that the prior density $\pi(\theta)$ has marginal densities $\pi_i(\theta_i)$, $i = 1, \ldots, n$, which are equal to the desired within-model prior densities.

By Bayes' Theorem,

$$\pi(\alpha|x) = \frac{\pi(x|\alpha)\pi(\alpha)}{\pi(x)} = \frac{\pi(\alpha)\sum_{i=1}^{n}\alpha_i m_i(x)}{\pi(x)},$$

where $\pi(\alpha)$ denotes the prior density of α and, for $i = 1, \ldots, n$,

$$m_i(x) = \int \pi_i(x, |\theta_i)\pi(\theta)\, d\theta = \int \pi_i(x, |\theta_i)\pi_i(\theta_i) d\theta_i,$$

integrating over all θ_j with $j \neq i$.

Computing Bayes Factors

Note also that

$$1 = \int \pi(\alpha|x)\, d\alpha = \pi(x)^{-1} \sum_{i=1}^{n} E[\alpha_i] m_i(x),$$

whence

$$\pi(x) = \sum_{i=1}^{n} E[\alpha_i] m_i(x). \tag{2}$$

Now for $i \neq j$, the Bayes factor in favour of M_i relative to M_j is defined to be $B_{ij} = B_{ij}(x) = m_i(x)/m_j(x)$.

Computing Bayes Factors

However,

$$\begin{aligned} E[\alpha_i|x] &= \int \alpha_i \pi(\alpha|x)\, d\alpha \\ &= \pi(x)^{-1} \int \alpha_i \left(\sum_{j=1}^{n} \alpha_j m_j(x) \right) \pi(\alpha)\, d\alpha \\ &= \pi(x)^{-1} \sum_{j=1}^{n} E[\alpha_i \alpha_j] m_j(x), \end{aligned}$$

which combined with (2) yields that

$$E[\alpha_i|x] = \frac{\sum_{j=1}^{n} E[\alpha_i \alpha_j] m_j(x)}{\sum_{j=1}^{n} E[\alpha_j] m_j(x)}, \quad i = 1, \ldots, n. \tag{3}$$

Computing Bayes Factors

Recall:

$$E[\alpha_i|x] = \frac{\sum_{j=1}^{n} E[\alpha_i\alpha_j]m_j(x)}{\sum_{j=1}^{n} E[\alpha_j]m_j(x)}, \quad i = 1, \ldots, n.$$

Next, fix $k \in \{1, \ldots, n\}$.

Dividing the numerator and denominator of the fraction above by $m_k(x)$ and rearranging we obtain

$$\sum_{j=1}^{n} (E[\alpha_j]E[\alpha_i|x] - E[\alpha_i\alpha_j])B_{jk}(x) = 0, \quad i = 1, \ldots, n. \qquad (4)$$

It remains to solve equations (4) to find $B_{jk}(x)$, $j = 1, \ldots, n$.

Computing Bayes Factors

Define A as the $n \times n$ matrix with elements

$$A_{ij} = E[\alpha_i|x]E[\alpha_j] - E[\alpha_i\alpha_j], \quad 1 \leq i, j \leq n.$$

Note that A depends on x, although we suppress this dependence in our notation.

Lemma 1

(a) If det $\tilde{A}_{-k} \neq 0$ then

$$B_{jk}(x) = \frac{\det \tilde{A}_{-jk}}{\det \tilde{A}_{-k}}. \tag{5}$$

where we

- define $\tilde{A}_{-k}$ as the $(n-1) \times (n-1)$ matrix formed by removing the kth row and kth column of A;
- similarly for $j \neq k$ define $\tilde{A}_{-jk}$ as the $(n-1) \times (n-1)$ matrix formed from $\tilde{A}_{-k}$ by replacing the elements A_{ij} with $-A_{ik}$, $i = 1, \ldots, n$, $i \neq k$.

Computing Bayes Factors

Lemma 1

(b) Suppose that $0 < m_i(x) < \infty$ for $i = 1, \ldots, n$. Then if either (i) $n = 2$ and $0 < E[\alpha_1] < 1$, or (ii) α has a Dirichlet prior distribution, $\mathcal{D}(p_1, \ldots, p_n)$, then

$$B_{jk}(x) = \frac{A_{jk}}{A_{kj}}.$$

Note: Dirichlet prior is both straightforward to use and flexible enough for computational purposes as described below.

An Example with Two Competing Models

We give special attention to the case $n = 2$ since this is of practical importance.

Here we have $\alpha = (\alpha_1, 1 - \alpha_1)$ and Lemma 1 yields that

$$B_{12} = \frac{E[\alpha_1] - E[\alpha_1^2] - E[\alpha_1|x](1 - E[\alpha_1])}{E[\alpha_1]E[\alpha_1|x] - E[\alpha_1^2]}.$$

It follows that

$$\frac{E[\alpha_1] - E[\alpha_1^2]}{1 - E[\alpha_1]} \leq E[\alpha_1|x] \leq \frac{E[\alpha_1^2]}{E[\alpha_1]},$$

A practical consequence is that any numerical estimate of $E[\alpha_1|x]$ lying outside these bounds must be incorrect.

An Example with Two Competing Models

Under the further assumption that $\pi(\alpha)$ is a uniform density, so that $\alpha_1 \sim U(0,1)$, we obtain

$$B_{12} = \frac{3E[\alpha_1|x] - 1}{2 - 3E[\alpha_1|x]},$$

$$E[\alpha_1|x] = \frac{2m_1(x) + m_2(x)}{3(m_1(x) + m_2(x))}$$

and

$$1/3 \leq E[\alpha_1|x] \leq 2/3$$

.

Finally, if α is assigned a Dirichlet prior distribution, bounds for $E[\alpha_i|x]$ for any value of n can be also obtained.

Mixture Model With Missing Data

- In our setting, the data x may be a partial observation of a stochastic process.

- In consequence, $\pi_i(x|\theta_i)$ in (1) may be intractable, meaning that it cannot be analytically or numerically evaluated in an efficient manner.

- We adopt data augmentation to overcome this problem, as follows:

  Let $y = (y_1, \ldots, y_N)$ be a vector comprising different kinds of 'missing data', and for $i = 1, \ldots, n$ let $\mathcal{I}(i) \subseteq \{1, \ldots, N\}$ and define $y_{\mathcal{I}(i)}$ as the vector with components y_j, $j \in \mathcal{I}(i)$.

- Thus $y_{\mathcal{I}(i)}$ denotes the missing data for model i, in practice usually chosen to make the augmented probability density $\pi_i(x, y_{\mathcal{I}(i)}|\theta_i)$ tractable.

- If model i does not require missing data, then $y_{\mathcal{I}(i)}$ is null.

- The calculation of the Bayes Factor can be derived in a similar manner as to the case without missing data.

Posterior Computation

- We employ Markov chain Monte Carlo methods (MCMC) and our objective is to sample from the target density

$$\pi(\alpha, \theta, y|x) \propto \pi(x, y|\alpha, \theta)\pi(\alpha)\pi(\theta), \tag{6}$$

- The first issue is that of assigning any missing data prior density terms in $\pi(x, y|\alpha, \theta)$.
- Recall that θ are the parameters, x is observed data and y is the missing data.

Missing data prior densities

Although the desired Bayes factors are invariant to the choice of any missing data prior densities, this choice is important in practice for computations.

This is largely a problem-specific issue, but we make two general remarks:

- if all models share the same missing data (y_1, say) then no missing data prior densities are required, and the target density becomes

$$\pi(\alpha, \theta, y|x) \propto \sum_{i=1}^{n} \alpha_i \pi_i(x, y_1|\theta_i)\pi(\alpha)\pi(\theta).$$

- It can be beneficial to assign missing data priors which mimic the marginal density of the $y_{-\mathcal{I}(i)}$ components in other models.

Sampling from the Target Density

- Sampling from the target density defined will typically be possible via a range of standard Markov chain Monte Carlo methods.
- The fact that the target density is a sum will usually make direct Gibbs sampling infeasible.
- However we can use the approach of Diebolt and Robert (1994), which relies on the introduction of allocation variables.

Markov Chain Monte Carlo

Introduce $z = (z_1, \dots, z_n)$ such that $z_i \in \{0, 1\}$ and $\sum_{i=1}^n z_i = 1$.

Thus z can take n possible values, each of which is a vector of zeroes other than a 1 at one position.

Define the augmented likelihood

$$\pi(z, x, y|\alpha, \theta) = \prod_{i=1}^n (\alpha_i \pi_i(x, y_{\mathcal{I}(i)}|\theta_i)\pi_i(y_{-\mathcal{I}(i)}|x, y_{\mathcal{I}(i)}, \theta))^{z_i},$$

- If the prior distribution on α is Dirichlet, $\mathcal{D}(p_1, \dots, p_n)$, it follows that α has full conditional distribution

$$\alpha| \cdots \sim \mathcal{D}(p_1 + z_1, \dots, p_n + z_n).$$

- The full conditional distribution of z is multinomial $\mathcal{M}(1; q_1, \dots, q_n)$, where the probabilities are given by

$$q_i \propto \alpha_i \pi_i(x, y_{\mathcal{I}(i)}|\theta_i)\pi_i(y_{-\mathcal{I}(i)}|x, y_{\mathcal{I}(i)}, \theta), \quad i = 1, \dots, n.$$

Markov Chain Monte Carlo

- For $j = 1, \ldots, n$, θ_j has full conditional distribution given by

  $$\pi(\theta_j | \cdots) \propto \pi(\theta)\pi_i(x, y_{\mathcal{I}(i)}|\theta_i)\pi_i(y_{-\mathcal{I}(i)}|x, y_{\mathcal{I}(i)}, \theta)$$

  where i denotes the current model, i.e. $z_i = 1$.

  Simplification often occurs in practice: for instance, if $\theta_1, \ldots, \theta_n$ are independent *a priori* and there are no missing data prior densities then we obtain

  $$\pi(\theta_j | \cdots) \propto \begin{cases} \pi_j(\theta_j) & z_j = 0, \\ \pi_j(x, y_{\mathcal{I}(j)}|\theta_j)\pi_j(\theta_j) & z_j = 1. \end{cases}$$

- Finally, any missing data component y_j, $j = 1, \ldots, N$, has full conditional distribution given by

  $$\pi(y_j | \cdots) \propto \begin{cases} \pi_i(x, y_{\mathcal{I}(i)}|\theta_i) & j \in \mathcal{I}(i), \\ \pi_i(y_{-\mathcal{I}(i)}|x, y_{\mathcal{I}(i)}, \theta_i) & j \notin \mathcal{I}(i), \end{cases}$$

  where i denotes the current model.

Connections with other approaches

- The framework we adopt is related to that described in Carlin and Chib (1995) and Godsill (2001), in which the target distribution of interest is defined over a product space of models and their parameters.

- Basically, our approach is not equivalent or special case of the latter.

- It is the existence of the α parameter which distinguishes our formulation from that of Carlin and Chib (1995) and Godsill (2001).

Examples

Poisson Process vs Linear Birth Process

- **Data**: Event times $x = (x_1, \ldots, x_n)$ observed during a time interval $[0, T]$, where $0 \le x_1 \le x_2 \le \ldots \le x_n \le T$.

- **Models**: a homogeneous Poisson process of rate λ (M_1) and a linear birth process $\{X(t) : t \in [0, T]\}$ with per-capita birth rate μ and $X(0) = 1$ (M_2).

- **Priors**: $\lambda \sim \text{Exp}(\theta)$ and $\mu \sim \text{Exp}(\theta)$ (independent).

- **Likelihoods**:

$$\pi_1(x|\lambda) = \lambda^n \exp\{-(\lambda - 1)T\}$$

$$\pi_2(x|\mu) = n!\mu^n \exp\{-\mu[(n+1)T - S(x)] + T\},$$

  where $S(x) = \sum_{j=1}^{n} x_j$.

  (Written densities with respect to the reference measure induced by a unit rate Poisson process on $[0, T]$)

Poisson Process vs Linear Birth Process

The Bayes factor in favour of M_1 relative to M_2 is

$$\begin{aligned} B_{12} = \frac{\int \pi_1(x|\lambda)\pi(\lambda)\, d\lambda}{\int \pi_2(x|\mu)\pi(\mu)\, d\mu} &= \frac{\int_0^\infty \theta\lambda^n \exp\{-\lambda(T+\theta)\}\ d\lambda}{\int_0^\infty \theta n!\mu^n \exp\{-\mu[(n+1)T - S(x) + \theta]\}\ d\mu} \\ &= \frac{[(n+1)T - S(x) + \theta]^{n+1}}{(T+\theta)^{n+1} n!}. \end{aligned}$$

Poisson Process vs Linear Birth Process

Assuming that $\alpha_1 \sim U(0,1)$ *a priori*, a simple Gibbs sampler for the target density consists of parameter updates as follows:

$$
\begin{aligned}
\alpha_1 | \cdots &\sim Beta(z_1 + 1, 2 - z_1), \\
z_1 | \cdots &\sim Bern\left(\frac{\alpha_1 \pi_1(x|\lambda)}{\alpha_1 \pi_1(x|\lambda) + (1-\alpha_1)\pi_2(x|\mu)} \right), \\
\lambda | \cdots &\sim \begin{cases} \Gamma(1, \theta) & z_1 = 0, \\ \Gamma(n+1,\, T + \theta) & z_1 = 1, \end{cases} \\
\mu | \cdots &\sim \begin{cases} \Gamma(n+1, (n+1)T - S(x) + \theta) & z_1 = 0, \\ \Gamma(1, \theta) & z_1 = 1, \end{cases}
\end{aligned}
$$

where $\Gamma(m, \xi)$ denotes a Gamma distribution.

Poisson Process vs Linear Birth Process

Table: Example 3: $\hat{B}_{12}$ compared to true values (B_{12}).

n	T	$S(x)$	θ	$\hat{B}_{12}$	B_{12}
5	10	36	1	1.15	1.148
5	10	36	0.01	1.58	1.587
5	10	25	1	10.25	10.239
10	20	150	1	0.18	0.181

- If we were to do RJMCMC, we would have to find a way of proposing a value of μ given λ for jumps from M_1 to M_2, and vice versa.
- In practice it is not immediately obvious how best to do this.
- An approach suggested in Green (2003) is to propose μ independently of λ, (e.g. from $\pi(\mu|x)$). This is similar to what we obtain above.

Non-nested regression models for Pines data

We consider the well-known model choice problem of assigning non-nested linear regression models to the pines data set.

The data describe the maximum compression strength parallel to the grain y_i, the density x_i, and the resin-adjusted density z_i for 42 specimens of radiata pine.

These data have been analyzed by several authors in order to compare methods for estimating Bayes factors.

Non-nested regression models for Pines data

The two competing models we consider are

$$\begin{aligned} M_1 : \quad & y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2); \\ M_2 : \quad & y_i = \gamma + \delta(z_i - \bar{z}) + \eta_i, \qquad \eta_i \sim N(0, \tau^2), \end{aligned}$$

which in vector notation we write as

$$Y = X(\alpha, \beta)^T + \epsilon$$

and

$$Y = Z(\gamma, \delta)^T + \eta$$

.

We assign identical prior distributions to the papers that have analysed the data in the literature.

We assigned a Beta$(100, 1)$ prior distribution for α_1.

Non-nested regression models for Pines data

Using the allocation variables approach, parameter updates for a Gibbs sampling algorithm are as follows, where $\theta_1 = (\alpha, \beta, \sigma^2)$ and $\theta_2 = (\gamma, \delta, \tau^2)$:

$$
\begin{aligned}
\alpha_1 | \cdots &\sim \mathit{Beta}(100 + z_1, 2 - z_1), \\
z_1 | \cdots &\sim \mathit{Bern}\left(\frac{\alpha_1 \pi_1(x|\theta_1)}{\alpha_1 \pi_1(x|\theta_1) + (1-\alpha_1)\pi_2(x|\theta_2)} \right), \\
\theta_i | \cdots &\sim \begin{cases} (N(\mu_0, v_0), IG(a_0, b_0)) & z_i = 0, \\ (N(\mu_i, v_i), IG(a_i, b_i)) & z_i = 1, \end{cases}
\end{aligned}
$$

where $v_1, v_2, \mu_1, \mu_2, a_1, a_2, b_1, b_2$ are available in closed form.

Non-nested regression models for Pines data

We carried out 100 runs of our method, this being the same as the number of runs used for the methods described in Friel and Pettitt (2008).

Pines data set: Comparison of Bayes factors from different methods.

Method	Bias	Standard Error
RJMCMC	67	2678
RJ corrected	9	124
Power posterior (serial MCMC)	10	132
Power posterior (population MCMC)	22	154
Mixture method	10	39

The bias is calculated by comparison with the estimate of 4862 obtained by numerical integration in Green and O'Hagan (1998).

Pima Indians

- The well-known Pima Indians data set consists of diabetes incidence for $n = 532$ Pima Indian women along with data on seven covariates, and can be naturally modelled using logistic regression.

- In Friel and Wyse (2012), two specific models are considered; denote them by M_2 and M_3. These models have four explanatory variables in common (number of pregnancies, plasma glucose concentration, body mass index and diabetes pedigree function) while M_3 has one additional variable (age).

- The corresponding Bayes factor is calculated using various different methods in Friel and Wyse (2012), yielding values in the range 12.83 to 13.96, the latter value coming from a lengthy reversible jump MCMC run which could plausibly be regarded as the most reliable estimate.

Pima Indians

- To illustrate our method for three models we also consider an additional model which has the same explanatory variables as M_2 other than diabetes pedigree function. This model is denoted by M_1.
- Thus the parameters of the three models under consideration are all of different dimension.
- However, we applied our method by having parameters shared between the models (ie those that they refer to the same covariates).
- We then assigned independent Gaussian prior distributions to the parameters with mean zero and variance 100.

Pima Indians

- We implemented our method using an allocation variable approach with Metropolis-Hastings updates for the θ_{3j} parameters.

- From 10^7 iterations of the resulting Markov chain we obtained estimates $B_{12} = 0.048$ and $B_{23} = 13.94$.

- The latter is in line with the values reported in Friel and Wyse (2012), especially the reversible jump MCMC estimate of 13.96...

- ... while the former is close to an estimate of 0.042 which we obtained using the Laplace approximation method.

The SIR Model

- Consider a closed population contains $N + a$ individuals of whom N are initially susceptible and a initially infective.

- Each infective remains so for a period of time distributed according to a specified random variable T_I, known as the infectious period, after which it becomes removed and plays no further part in the epidemic.

- During its infectious period an infective makes contact with each other member of the population at times given by a homogeneous Poisson process of rate β/N.

- The infectious periods of different individuals and the Poisson processes between different pairs of individuals are assumed to be mutually independent.

- The epidemic ends when there are no infectives left in the population.

Epidemic Models with Two Different Infection Periods

- We consider two competing models, namely that $T_I \sim \Gamma(1, \gamma)$ (M_1) and $T_I \sim \Gamma(m, \lambda)$ (M_2), where the shape parameter m will be assumed known.
- Observed data consist of removal times only (ie infection times are not observed) and hence both model likelihoods $\pi_1(r|\gamma)$ and $\pi_2(r|\lambda)$ are intractable.
- We introduce "missing data" such that $\pi(i, r|\ \gamma)$ is tractable and the target density is:

$$\begin{aligned}\pi(\alpha_1, \beta_1, \beta_2, \gamma, \lambda|r) &\propto [\alpha_1 \pi_1(i, r|p, i_p, \beta_1, \gamma) \\ &+ (1-\alpha_1)\pi_2(i, r|p, i_p, \beta_2, \lambda)] \\ &\times \pi(\beta_1)\pi(\beta_2)\pi(\gamma)\pi(\lambda).\end{aligned}$$

- We need no missing data priors densities because the the missing data appear in both model likelihoods.

Epidemic Models with Two Different Infection Periods

- We consider the Susceptible-Infective-Removed model with $N = 50$ initially susceptible individuals and one initial infective.
- We use the allocation variable approach.
- We assign $\Gamma(1, 1)$ and $\alpha \sim Beta(1, 1)$ *a priori*.
- For each scenario we simulated 100 data sets, and evaluated the Bayes factor using our algorithm.
- In practice, one is rarely interested in data from epidemics with few cases, so we also evaluated the Bayes factors using a subset of each of the 100 simulations in which the epidemic had clearly 'taken off (evaluated by eye), which we refer to as major epidemics.

Epidemic Models with Two Different Infection Periods

Table: Bayes factors from algorithm output.

Scenario	True model	β	M_2	$E[B_{12}](st.dev.)$ (all simulations)	$E[B_{12}](st.dev.)$ (major epidemics)
A	$\Gamma(5,5)$	2	$\Gamma(5,\lambda)$	0.06 (0.06)	0.008 (0.006)
B	$\Gamma(1,0.75)$	1	$\Gamma(1,\lambda)$	1.03 (0.17)	1.05 (0.22)
C	$\Gamma(1,1)$	3	$\Gamma(2,\lambda)$	3022 (3969)	2291 (3428)

- Clear evidence in favour of models M_2 and M_1 for scenarios A and C respectively while for scenario B the mean of B_{12} is close to the true value of 1.

- In scenario A there is a marked difference in the Bayes factors when using all simulations compared to using only major epidemics. Why?

Conclusions

- We have presented a new method for evaluating Bayes factors.
- Although motivated by epidemic models and population processes, our approach is clearly applicable in more general settings.
- The methods permit data imputation as necessary, and can cater for models which share common parameters.
- It seems likely that they are best suited to situations in which there are only a small number of competing models.
- Constructing missing data prior densities, when required, seems likely to require problem-specific insights in order to obtain reasonably efficient algorithms.

Appendix

Connections with other approaches

- The framework we adopt is related to that described in Carlin and Chib (1995) and Godsill (2001), in which the target distribution of interest is defined over a product space of models and their parameters.

- Consider two models defined by densities $\pi_1(x|\theta_1)$ and $\pi_2(x|\theta_2)$, and independent within-model prior densities $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$.

- The framework of Carlin and Chib (1995) and Godsill (2001) introduces a model indicator $k \in \{1, 2\}$ to denote the 'current' model. The target density of interest is specified by

$$\pi(k, \theta_1, \theta_2|x) \propto \pi_k(x|\theta_k)\pi_k(\theta_k|k)\pi(\theta_{3-k}|\theta_k, k)\pi(k),$$

  where it is necessary to specify $\pi(\theta_{3-k}|\theta_k, k)$, i.e. the 'prior' for the non-current model parameter.

- Assuming θ_1 and θ_2 to be independent of each other and k gives that $\pi(\theta_{3-k}|\theta_k, k) = \pi_{3-k}(\theta_{3-k})$.

Connections with other approaches

Conversely, our formulation has target density

$$\pi(\alpha, \theta_1, \theta_2 | x) \propto \pi(\alpha)\pi_1(\theta_1)\pi_2(\theta_2)[\alpha_1\pi_1(x|\theta_1) + \alpha_2\pi_2(x|\theta_2)].$$

If we adopt the allocation-variable approach, the target density becomes

$$\pi(z, \alpha, \theta_1, \theta_2 | x) \propto \pi(\alpha)\pi_1(\theta_1)\pi_2(\theta_2)[\alpha_1\pi_1(x|\theta_1)]^{z_1}[\alpha_2\pi_2(x|\theta_1)]^{z_2},$$

from which we see that it is the existence of the α parameter which distinguishes our formulation from that of Carlin and Chib (1995) and Godsill (2001).

- The general formulation in Godsill (2001) also allows each model to potentially share parameters with other models.
- Specifically, the parameters of model k will be some subset of a set of parameters $\{\theta_1, \ldots, \theta_N\}$ (similar but not technically equivalent to the way that we have dealt with shared missing data).

Connections with other approaches

- The fundamental difference is that missing data may not always require a prior, whereas model parameters always do.

- Our approach also enables each model to potentially share parameters with other models, since we allow for arbitrary prior dependency between $\theta_1, \ldots, \theta_n$.