Estimating the evidence for statistical models

Nial Friel

University College Dublin
nial.friel@ucd.ie

March, 2011

Bayesian model choice

Given data y and competing models: $m_1, \ldots, m_l$, each with parameters $\theta_1, \ldots, \theta_l$, respectively.

Bayesian inference:

$$\pi(\theta_k, m_k|y) \propto \pi(y|\theta_k, m_k)\pi(\theta_k|m_k)p(m_k)$$

Model evidence

Within model m_k:

$$\pi(\theta_k|y, m_k) \propto \pi(y|\theta_k, m_k)\pi(\theta_k|m_k)$$

Constant of proportionality is

$$\pi(y|m_k) = \int_{\theta_k} \pi(y|\theta_k, m_k)\pi(\theta_k|m_k)d\theta_k.$$

This is often called the marginal likelihood, integrated likelihood or evidence and is difficult to compute in general.

Posterior model probabilities

Suppose we could compute $\pi(y|m_k)$. Then, using Bayes theorem we get,

$$\pi(m_k|y) = \frac{\pi(y|m_k)\pi(m_k)}{\sum_1^I \pi(y|m_k)\pi(m_k)}.$$

Bayes factors

If we have two competing models:

$$\frac{\pi(m_1|y)}{\pi(m_2|y)} = \frac{\pi(y|m_1)}{\pi(y|m_2)} \times \frac{\pi(m_1)}{\pi(m_2)}$$

posterior odds = Bayes factor $\times$ prior odds

The Bayes factor, $B_{12} = \frac{\pi(y|m_1)}{\pi(y|m_2)}$.

The larger B_{12} is, the greater the evidence in favour of M_1 compared to M_2.

Bayesian model averaging

Predictions can be made by averaging over all models, weighted proportional to the posterior model probability, thereby incorporating model uncertainty.

$$\pi(y^*|y) = \sum_{k=1}^{I} \pi(y^*|m_k, y)\pi(m_k|y)$$

This is the average of the posterior distribution for y^* under each model weighted by the corresponding posterior model probabilities.

Why estimating the model evidence is a challenge

- $\pi(y|m_k)$ is an integral of a (usually) highly variable function over a high-dimensional parameter space.
- Analytic tractability is sometimes possible, often where conjugate priors are used. This is quite rare.
- Consequently, sophisticated Monte Carlo methods are needed.

Within model search or across model search?

Within model search:

Inference for $\pi(\theta_k|y)$ separately for every m_k. This is used to estimate $\pi(y|m_k)$, for all k.

There are many approaches under this heading.

Across model search:

Here inference is carried out over the joint model and parameter space, $\pi(\theta_k, m_k|y)$. In an MCMC setting, only one chain is needed!

Reversible jump Markov chain Monte Carlo developed by Green (1995) is the dominant approach. ($> 1,400$ citations to date...)

Laplace's method (eg Tierney and Kadane 1986)

Assume that $\pi(\theta_k|y)$ is highly peaked around the posterior mode $\tilde{\theta}_k$ eg if sample size is large enough.
Define

$$I(\theta_k) = \log\{\pi(y|\theta_k)\pi(\theta_k)\}$$

- Expand $I(\theta_k)$ as a quadratic about $\tilde{\theta}_k$ and then exponentiate.
- Result gives an approximation to $\pi(y|\theta_k)\pi(\theta_k)$ as a Gaussian with mean $\tilde{\theta}_k$ and covariance $\tilde{\Sigma} = (-D^2 I(\tilde{\theta}_k))^{-1}$, where $D^2 I(\tilde{\theta}_k)$ is the Hessian matrix of second derivatives.
- Integrating this approximation yields

$$\pi(y) \approx (2\pi)^{d/2} |\tilde{\Sigma}|^{1/2} \pi(y|\tilde{\theta}_k)\pi(\tilde{\theta}_k).$$

Harmonic mean estimator (Newtown and Raftery (1994))

$$\pi(y) = 1/\left(\frac{1}{n}\sum_{1}^{n}\pi(y|\theta_i)\right), \quad \theta_i \sim \pi(\theta|y).$$

Why does this hold?

$$\mathbf{E}\left\{\frac{1}{\pi(\theta|y)}\pi(\theta|y)\right\} = \int \frac{\pi(y|\theta)\pi(\theta)}{\pi(\theta|y)\pi(y)}d\theta = \frac{1}{\pi(y)}\int \pi(\theta)\, d\theta = \frac{1}{\pi(y)}.$$

The bad news?

Harmonic mean estimator (Newtown and Raftery (1994))

$$\pi(y) = 1/\left(\frac{1}{n}\sum_{1}^{n}\pi(y|\theta_i)\right), \quad \theta_i \sim \pi(\theta|y).$$

Why does this hold?

$$\mathbf{E}\left\{\frac{1}{\pi(\theta|y)}\pi(\theta|y)\right\} = \int \frac{\pi(y|\theta)\pi(\theta)}{\pi(\theta|y)\pi(y)}d\theta = \frac{1}{\pi(y)}\int \pi(\theta)\, d\theta = \frac{1}{\pi(y)}.$$

The bad news?

Harmonic mean estimator (Newtown and Raftery (1994))

$$\pi(y) = 1/\left(\frac{1}{n}\sum_{1}^{n}\pi(y|\theta_i)\right), \quad \theta_i \sim \pi(\theta|y).$$

Why does this hold?

$$\mathbf{E}\left\{\frac{1}{\pi(\theta|y)}\pi(\theta|y)\right\} = \int \frac{\pi(y|\theta)\pi(\theta)}{\pi(\theta|y)\pi(y)}d\theta = \frac{1}{\pi(y)}\int \pi(\theta)\, d\theta = \frac{1}{\pi(y)}.$$

The bad news?

Harmonic mean estimator (Newtown and Raftery (1994))

$$\pi(y) = 1/\left(\frac{1}{n}\sum_{1}^{n}\pi(y|\theta_i)\right), \quad \theta_i \sim \pi(\theta|y).$$

This estimator is based solely on draws from the posterior. But the posterior is typically much more peaked than the prior, eg, when the posterior is insensitive to the prior. Hence in such situations, the harmonic mean estimator will not change much as the prior changes.

But $\pi(y)$ is *very sensitive* to changes in the prior.

This drawback is very well documented. See Radford Neal's blog, for example.

Chib's method (Chib 1995)

Chib (1995) presented a generic method which can be applied to output from the Gibbs sampler.

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

Re-writing this,

$$\pi(y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(\theta|y)}.$$

So we could estimate $\log \pi(y)$ as

$$\log \pi(y) = \log \pi(y|\theta^*) + \log \pi(\theta^*) - \log \hat{\pi}(\theta^*|y)$$

where $\hat{\pi}(\theta^*|y)$ is an estimate of the posterior density at a point θ^* of high posterior prob.

Chib's method (Chib 1995)

Chib's method relies on estimating $\pi(\theta^*|y)$.

Suppose the vector θ can be partitioned as $(\theta_1, \theta_2, \theta_3)$, where the full-conditional distribution of each θ_i is standard.

$$\pi(\theta^*|y) = \pi(\theta_1^*|\theta_2^*, \theta_3^*, y)\pi(\theta_2^*|\theta_3^*, y)\pi(\theta_3^*|y)$$

Gibbs sampling can be used to estimate each factor on the LHS:

$$\begin{aligned}
\pi(\theta_2^*|\theta_3^*, y) &= \frac{1}{N}\sum_j \pi(\theta_2^*|\theta_1^{(j)}, \theta_3^*). \\
\pi(\theta_3^*|y) &= \frac{1}{N}\sum_j \pi(\theta_3^*|\theta_1^{(j)}, \theta_2^{(j)}).
\end{aligned}$$

Chib's method (Chib 1995)

In general, Chib's method can be applied when θ is partitioned into an arbitrary number of blocks.

The only requirement is that the full-conditional sampling of each block is possible.

Annealed Importance Sampling (Neal 2001)

AIS is a very clever algorithm which shows how tempering can be used to define an importance samping function to sample from complex distributions.

Aside: Importance sampling to sample from a target $f(x)$ using an importance function $g(x)$:

$$x^{(1)}, \ldots, x^{(N)} \sim g(x)$$

$$\mathbf{E}_f a(x) = \frac{\sum w^{(i)} a(x^{(i)})}{\sum w^{(i)}}, \quad \text{where} \quad w^{(i)} = \frac{f(x^{(i)})}{g(x^{(i)})}$$

Further,

$$\frac{1}{N} \sum w^{(i)} \to \frac{z_f}{z_g} \quad \text{as} \quad N \to \infty,$$

where $z_f = \int_x f(x)dx$ and $z_g = \int_x g(x)dx$.

Annealed Importance Sampling (Neal 2001)

Define

$$\pi_i(\theta|y) = \pi(\theta)^{1-t_i}\pi(\theta|y)^{t_i}, \text{ where } 1 = t_0 > \cdots > t_n = 0.$$

Thus π_{t_0} and π_{t_n} corresponds to posterior and prior, respectively.

Let T_i denote a Markov transition kernel with invariant π_{t_i}.

For $j = 1, \ldots, N$

- Sample θ_{n-1} from π_{t_n}
- Sample θ_{n-2} from θ_{n-1} using T_{n-1}
- $\ldots$
- Sample θ_0 from θ_1 using T_1.
- Set

$$\theta^{(j)} = \theta_0 \text{ and } w^{(j)} = \frac{\pi_{n-1}(\theta_{n-1})}{\pi_n(\theta_{n-1})}\frac{\pi_{n-2}(\theta_{n-2})}{\pi_{n-1}(\theta_{n-2})}\cdots\frac{\pi_0(\theta_0)}{\pi_1(\theta_0)}.$$

Annealed Importance Sampling

AIS yields:

1. An independent sample $\{\theta^{(i)}\}$ from $\pi(\theta|y)$.
2. An estimator of the evidence

$$\pi(y) \approx \frac{1}{n} \sum_{i=1}^{n} w^{(i)}.$$

Evidence estimation via power posteriors (NF and Pettitt (2008))

Consider the Power posterior:

$$\pi(\theta|y,t) \propto \{\pi(y|\theta)\}^{T(t)} p(\theta)$$

where $T : [0,1] \to [0,1]$ is defined st $T(0) = 0$ and $T(1) = 1$.
Its normalising constant is

$$z(y|t) = \int_\theta \{\pi(y|\theta)\}^t p(\theta)\, d\theta.$$

$z(y|t=1)$: Posterior model evidence.
$z(y|t=0)$: Integral of the prior for θ, which equals 1.

Evidence via power posteriors

The evidence follows the identity:

$$\log \pi(y) = \log \left\{ \frac{z(y|t=1)}{z(y|t=0)} \right\} = \int_0^1 \mathbf{E}_{\theta|t} \log \pi(y|\theta) dt.$$

Proof:

$$\begin{aligned}
\frac{d}{dt} \log(z(y|t)) &= \frac{1}{z(y|t)} z'(y|t) \\
&= \frac{1}{z(y|t)} \int \frac{d}{dt} \log(\pi(y|\theta))^t \pi(\theta) d\theta \\
&= \int \log(\pi(y|\theta)) \frac{\pi(y|\theta)^t \pi(\theta) d\theta}{z(y|t)} \\
&= \mathbf{E}_{\theta|t} \log(\pi(y|\theta)).
\end{aligned}$$

Evidence via power posteriors

$$\frac{d}{dt}\log z(y|t) = \mathbf{E}_{\theta|t}\log(\pi(y|\theta))$$

This is the mean deviance wrt to $(\theta|y,t)$ - the power posterior. Integrating wrt t yields,

$$\log \pi(y) = \log\left\{\frac{z(y|t=1)}{z(y|t=0)}\right\} = \int_0^1 \mathbf{E}_{\theta|t}\log\pi(y|\theta)dt.$$

This is essentially an application of thermodynamic integration, which was first developed in the statistical physics community, and outlined in Gelman and Meng (1998).

In practice: Discretise $t \in [0, 1]$, $0 = t_0 < t_1, \ldots, t_n = 1$.

For each t_i: Sample $\theta \sim \pi(\theta|y, t)$ and estimate $E_i = \mathbf{E}_{\theta|t_i} \log \pi(y|\theta)$.

$$\pi(y) = \sum_{i=1}^{n} (t_i - t_{i-1}) \left(\frac{(E_{i-1} + E_i)}{2} \right)$$

Sensitivity of $p(y)$ to the prior - toy example

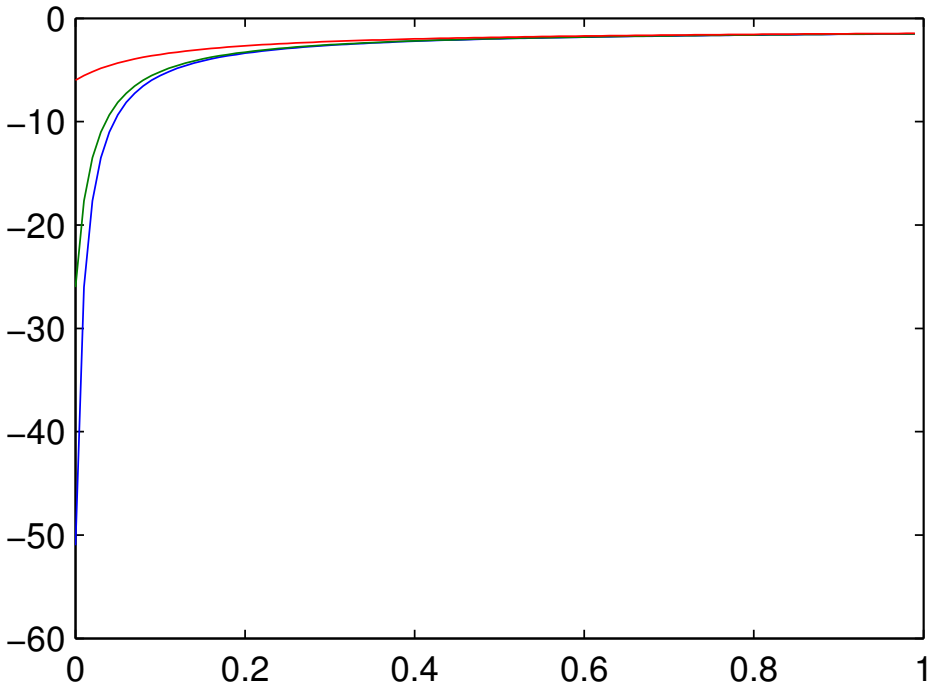
How does sensitivity to the prior impact on this method?
Suppose $y = \{y_i\}$ iid $N(\theta, 1)$. A priori, $\theta \sim N(m, v)$. Then the power posterior $\theta|y, t \sim N(m_t, v_t)$, where

$$m_t = \frac{nt\bar{y} + m/v}{nt + 1/v} \quad \text{and} \quad v_t = \frac{1}{nt + 1/v}$$

and

$$\mathbf{E}_{\theta|y,t} \log \pi(y|\theta) = -\frac{\log 2\pi}{2} - \frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{n}{2}\frac{(m-\bar{y})^2}{(vmt+1)^2} - \frac{n}{2}\frac{1}{(nt+1/v)}$$

When $t = 0$ final term is $-nv/2$. As $v \to \infty$ so too does $\mathbf{E}_{\theta|y,t}$.



Expected deviance, under the distribution $\theta|y, t$ plotted against t for prior variance equal to $10, 5, 1$.

As v increases, so too does the rate at which the mean deviance changes with t

Connection to Fractional Bayes estimator

The fraction $z(y|t=1)/z(y|t=a)$ where a is close to 0, is precisely the estimate of the marginal likelihood used in the 'Fractional Bayes' estimate of the Bayes factor (O'Hagan 95).

$$\pi(y) \approx \frac{z(y|t=1)}{z(y|t=a)} = \frac{\int_\theta \pi(y|\theta)\pi(\theta)\, d\theta}{\int_\theta \{\pi(y|\theta)\}^a \pi(\theta)\, d\theta}$$
$$= \int_a^1 \mathbf{E}_{\theta|t} \log \pi(y|\theta) dt$$

This method was proposed to compute Bayes factor with un-informative priors. Impropriety in $\pi(\theta)$ cancels above and below. Essentially a fraction a of the data is borrowed for the prior.

Power posterior approach

- It is realitively straightforward to code/implement.
- It is a generic method. In some cases it can be implemented in WinBUGS.
- Choosing the temperature schedule is vital – this is the weakness of this approach. Behrens, NF, Hurn (2011) offer some possibility in this direction.

Nested sampling (Skilling, 2006)

(For the moment (for ease of notation), let $L(\theta) = \pi(y|\theta)$.)

$$\pi(y) = \int L(\theta)\pi(\theta)\, d\theta = \int L(\theta)\, dX,$$

where $dX = \pi(\theta)\, d\theta$ is an element of prior mass.

Define

$$X(\lambda) = \int_{L(\theta)>\lambda} \pi(\theta)\, d\theta$$

as a cumulant prior mass.

Write the inverse function as $L(X)$, ie $L(X(\lambda)) = \lambda$. This then allows us to express the evidence as a $1-$dimensional integral:

$$\pi(y) = \int_0^1 L(X)\, dX.$$

Nested sampling (Skilling, 2006)

(For the moment (for ease of notation), let $L(\theta) = \pi(y|\theta)$.)

$$\pi(y) = \int L(\theta)\pi(\theta)\, d\theta = \int L(\theta)\, dX,$$

where $dX = \pi(\theta)\, d\theta$ is an element of prior mass.
Define

$$X(\lambda) = \int_{L(\theta)>\lambda} \pi(\theta)\, d\theta$$

as a cumulant prior mass.
Write the inverse function as $L(X)$, ie $L(X(\lambda)) = \lambda$. This then allows us to express the evidence as a $1-$dimensional integral:

$$\pi(y) = \int_0^1 L(X)\, dX.$$

Nested sampling (Skilling, 2006)

(For the moment (for ease of notation), let $L(\theta) = \pi(y|\theta)$.)

$$\pi(y) = \int L(\theta)\pi(\theta)\, d\theta = \int L(\theta)\, dX,$$

where $dX = \pi(\theta)\, d\theta$ is an element of prior mass.
Define

$$X(\lambda) = \int_{L(\theta)>\lambda} \pi(\theta)\, d\theta$$

as a cumulant prior mass.
Write the inverse function as $L(X)$, ie $L(X(\lambda)) = \lambda$. This then allows us to express the evidence as a $1-$dimensional integral:

$$\pi(y) = \int_0^1 L(X)\, dX.$$

Nested sampling

The main computational burden is the requirement to sample θ from the prior subject to the constraint that $L(\theta) > l$.
This is roughly similar to the computational effort of slice sampling (Neal, 2003).

The evidence is estimated by sorting draws from the prior according to their likelihood.

$$\pi(y) = Z = \sum_{i=1}^{l-1} (X_i - X_{i+1}) L_i.$$

Sketch of algorithm

Sample $\theta_1, \ldots, \theta_N$ from the prior.

Repeat for $i = 1, \ldots, I$:

- Find the point θ_k with the smallest likelihood, l_i, among the N current θ_i's.

  Set $X_i = exp(i/N)$ and $w_i = X_{i-1} - X_i$.

  Increment Z by $L_i w_i$.
- Replace θ_k with a point sampled from the prior subject to $L(\theta) > l_i$.

Doubly intractable distributions

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

Here we assume that the likelihood, $\pi(y|\theta)$, is impossible to evaluate.

Doubly intractable distributions

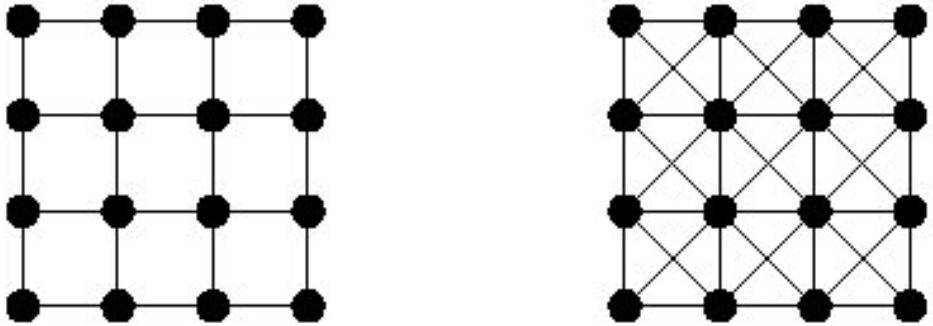
Gibbs random fields, which find use in spatial statistics and statistical network analysis, involves intractable likelihood models.

Ising model

- Defined on a lattice $y = \{y_1, \ldots, y_n\}$.
- Lattice points y_i take values $\{-1, 1\}$.
- Full conditional $\pi(y_i|y_{-i}, \theta) = \pi(y_i|\text{neighbours of } i, \theta)$.

$$\pi(y|\theta) \propto q(y|\theta) = \exp\left\{\frac{1}{2}\theta_1 \sum_{i\sim j} y_i y_j\right\}.$$

Here $\sim$ means "is a neighbour of".



1st order and 2nd order Ising models.

$$\pi(y|\theta) = \frac{\exp(\theta^T s(y))}{z(\theta)}$$

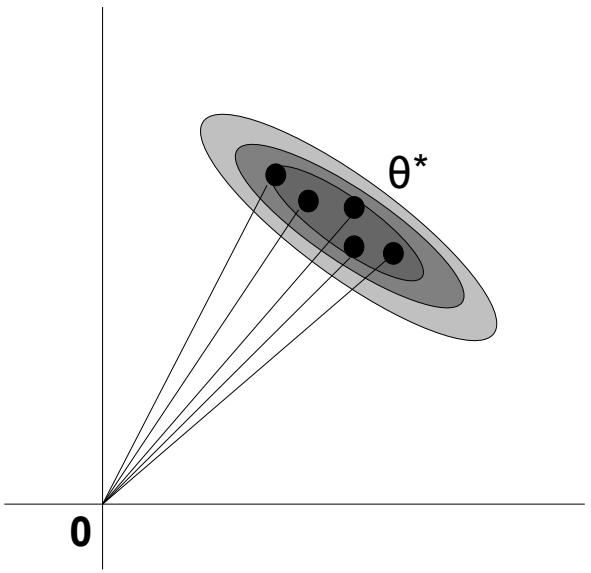
$s(y)$ is a sufficient statistics and counts the number of 'like' neighbours.

$$z(\theta) = \sum_{x_1} \cdots \sum_{x_n} q(y|\theta).$$

Model evidence for MRFs – our approach

$$\pi(y) = \frac{q(y|\theta)\pi(\theta)}{z(\theta)\pi(\theta|y)} \quad \forall\theta.$$

- Draw from the posterior, and estimate $\pi(\theta^*|y)$ for a high probability θ^*.
- Estimate $z(\theta)$ using thermodynamic integration.



Auxiliary variable method (Møller *et al.*, 2006)

Introduce an auxiliary variable y' on the same space as the data y and extend the target distribution

$$\pi(\theta, y'|y) \propto \pi(y|\theta)\pi(\theta)\pi(y'|\theta_0),$$

for some fixed θ_0.
Joint update (θ^*, y'^*) with proposal:

$$h(\theta^*, y'^*|\theta, y') = h_1(y'^*|\theta^*)h_2(\theta^*|\theta, y'^*)$$

where

$$h_1(y'^*|\theta^*) = \pi(y'^*|\theta^*) = \frac{q(y'^*|\theta^*)}{z(\theta^*)}.$$

$$\alpha(\theta^*, y'^*|\theta, y') = \frac{\pi(y|\theta^*)\pi(\theta^*)\pi(y'^*|\theta_0)\pi(y'|\theta)h_2(\theta|\theta^*)}{\pi(y|\theta)\pi(\theta)\pi(y'|\theta_0)\pi(y'^*|\theta^*)h_2(\theta^*|\theta)}$$

$z(\theta^*)$ appears in $\pi(y|\theta^*)$ above and in $\pi(y'^*|\theta^*)$ below, and therefore cancels. Similarly $z(\theta)$ cancels above and below.

The choice of θ_0 is important. eg the maximum pseudolikelihood estimate based on y.

Exchange algorithm (Murray, Ghahramani & MacKay 2006)

Sample from an augmented distribution

$$\pi(\theta', y', \theta|y) \propto \pi(y|\theta)\pi(\theta)h(\theta'|\theta)\pi(y'|\theta')$$

whose marginal distribution for θ is the posterior of interest

- $\pi(y'|\theta')$ is the same likelihood model on which y is defined.
- $h(\theta'|\theta)$ arbitrary distribution for the augmented variable θ' which might depend on θ (eg random walk distribution centred at θ)

Exchange algorithm – How it works

1. Gibbs update of (θ', y')
   - *i* Draw $\theta' \sim h(\cdot|\theta)$
   - *ii* Draw $y' \sim \pi(\cdot|\theta')$
2. Exchange move from (θ, y), (θ', y') to (θ', y), (θ, y') with probability

$$\alpha = \min\left(1, \underbrace{\frac{q(y'|\theta)}{q(y|\theta)}}_{*} \frac{\pi(\theta')}{\pi(\theta)} \frac{h(\theta|\theta')}{h(\theta'|\theta)} \underbrace{\frac{q(y|\theta')}{q(y'|\theta')}}_{**} \times \underbrace{\frac{z(\theta)z(\theta')}{z(\theta)z(\theta')}}_{1}\right)$$

- Exchange move proposes to "offer" the data y the auxiliary θ' and similarly to "offer" the auxiliary data y' the parameter θ
- The affinity between θ' and y is measured by (**) and the affinity between θ and y' by (*)

Exchange algorithm – How it works

1 Gibbs update of (θ', y')
   - *i* Draw $\theta' \sim h(\cdot|\theta)$
   - *ii* Draw $y' \sim \pi(\cdot|\theta')$

2 Exchange move from (θ, y), (θ', y') to (θ', y), (θ, y') with probability

$$\alpha = \min\left(1, \underbrace{\frac{q(y'|\theta)}{q(y|\theta)}}_{*} \frac{\pi(\theta')}{\pi(\theta)} \frac{h(\theta|\theta')}{h(\theta'|\theta)} \underbrace{\frac{q(y|\theta')}{q(y'|\theta')}}_{**} \times \underbrace{\frac{z(\theta)z(\theta')}{z(\theta)z(\theta')}}_{1}\right)$$

- Exchange move proposes to "offer" the data y the auxiliary θ' and similarly to "offer" the auxiliary data y' the parameter θ
- The affinity between θ' and y is measured by (**) and the affinity between θ and y' by (*)

Exchange algorithm for the Ising model

$$\alpha = \min\left(1, \frac{\pi(\theta')}{\pi(\theta)} \exp\left\{(\theta - \theta')^t(s(y') - s(y))\right\}\right)$$

The term

$$\exp\left\{(\theta - \theta')^t(s(y') - s(y))\right\}$$

can be viewed as a measure of distance between the observed data y and the auxiliary data y'.
It is somewhat similar to the accept/reject step in ABC (approximate Bayesian computation).
Note: If $\theta \approx \theta'$, then $\alpha \approx 1$. This does not necessarily happen with ABC.

Exchange algorithm for the Ising model

- The main difficulty is the need to draw an exact sample $y' \sim \pi(\cdot|\theta')$
- Perfect sampling is an obvious approach.
- A pragmatic alternative is to take a realisation from a long MCMC run with stationary distribution $\pi(y'|\theta')$ as an approximate draw.

Simulation study: Ising model

Data y simulated from an Ising model defined on a 16×16 lattice, with a single interaction parameter θ.

Two competing models: 4 and 8 nearest neighbours.

Here the lattices are sufficently small to allow a very accurate estimate of the Bayes factor:

The normalising constant $z(\theta)$ can be calculated exactly for a grid of $\{\theta_i\}$ values, which can then be plugged into the right hand side of:

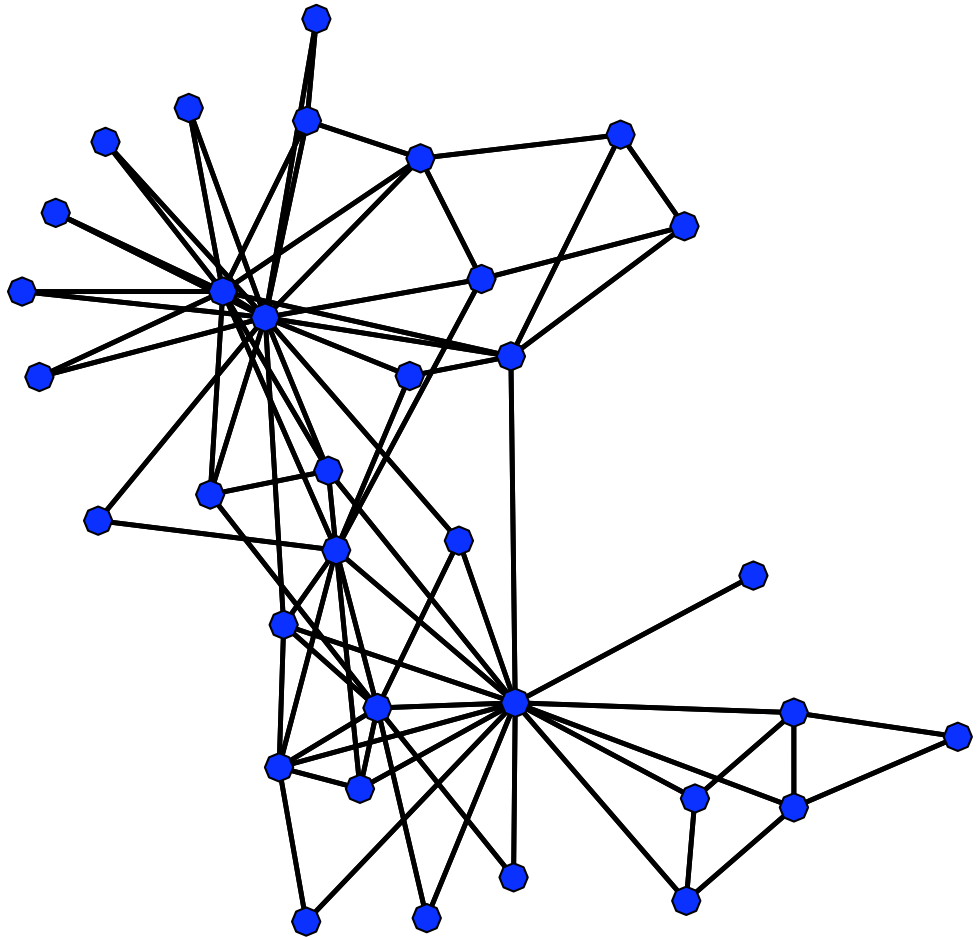
$$\pi(\theta_i|y) \propto \frac{q(y|\theta_i)}{z(\theta_i)}\pi(\theta_i), \quad i = 1, \ldots, n.$$

Summing up the right hand side yields an estimate of $\pi(y)$. This serves as a groundtruth to compare with the corresponding MCMC-based estimate of the model evidence.

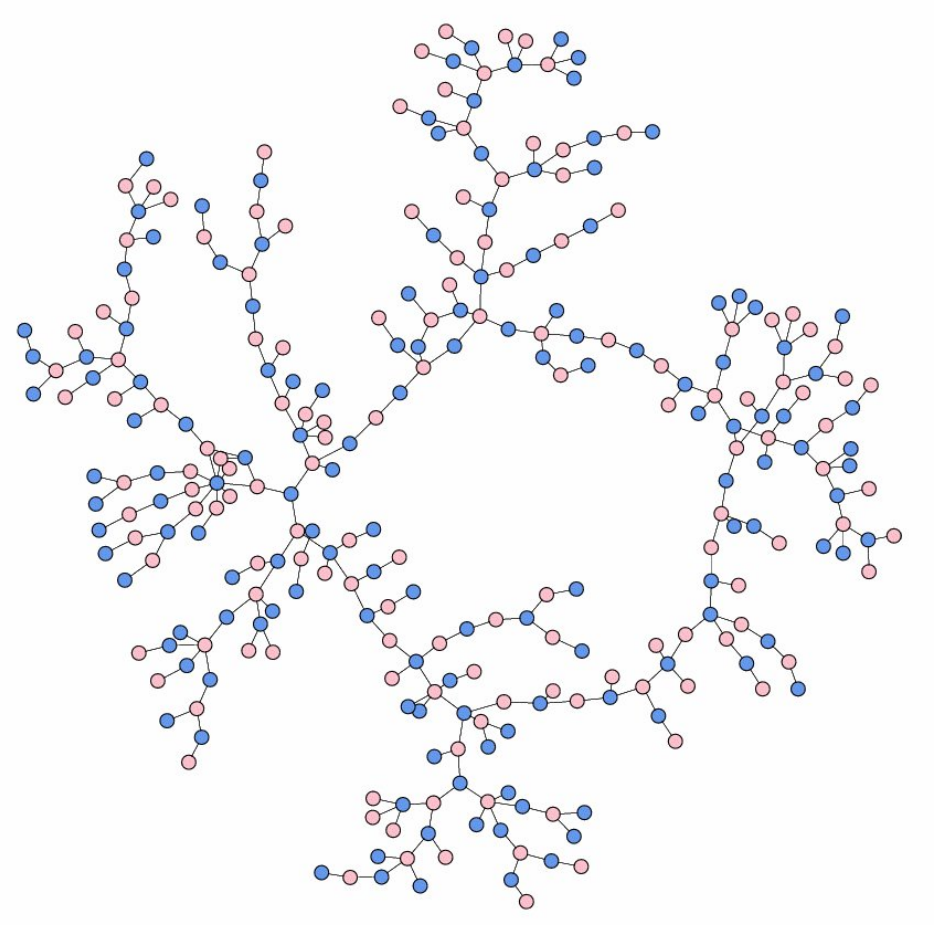
Results: Ising model

θ	$\hat{BF}$	BF
0.1	2.51	1.88
0.2	13.48	13.57
0.3	9.135	6.95
0.4	3.35	2.05

Friendships in a karate club in a US university.



High school dating



The exponential random graph (or p^*) model

First proposed by Frank and Strauss (JASA, 1986).
Let $y_{ij} = 1$ denote an edge connecting nodes i and j, and 0, otherwise.

Data y is an adjacency matrix indicating nodes which are connected by an edge.

1. Edges y_{ij} and y_{kl} are neighbours of one another, if they share a common node.
2. If y_{ij} and y_{kl} are not neighbours, then y_{ij} and y_{ij} are conditionally independent, given the rest of the graph.

The exponential random graph (or p^*) model

First proposed by Frank and Strauss (JASA, 1986).
Let $y_{ij} = 1$ denote an edge connecting nodes i and j, and 0, otherwise.

Data y is an adjacency matrix indicating nodes which are connected by an edge.

1. Edges y_{ij} and y_{kl} are neighbours of one another, if they share a common node.
2. If y_{ij} and y_{kl} are not neighbours, then y_{ij} and y_{ij} are conditionally independent, given the rest of the graph.

The p^* model

$$\pi(y|\theta) = \frac{\exp\{\theta^t s(y)\}}{z(\theta)} = \frac{q(y|\theta)}{z(\theta)}$$

- y observed graph
- $s(y)$ known vector of sufficient statistics
- θ vector of parameters
- $z(\theta)$ normalizing constant

$$z(\theta) = \sum_{\text{all possible graphs}} \exp\{\theta^t s(y)\}$$

- $2^{\binom{n}{2}}$ possible undirected graphs of n nodes
- Calculation of $z(\theta)$ is infeasible for non-trivially small graphs

The p^* model

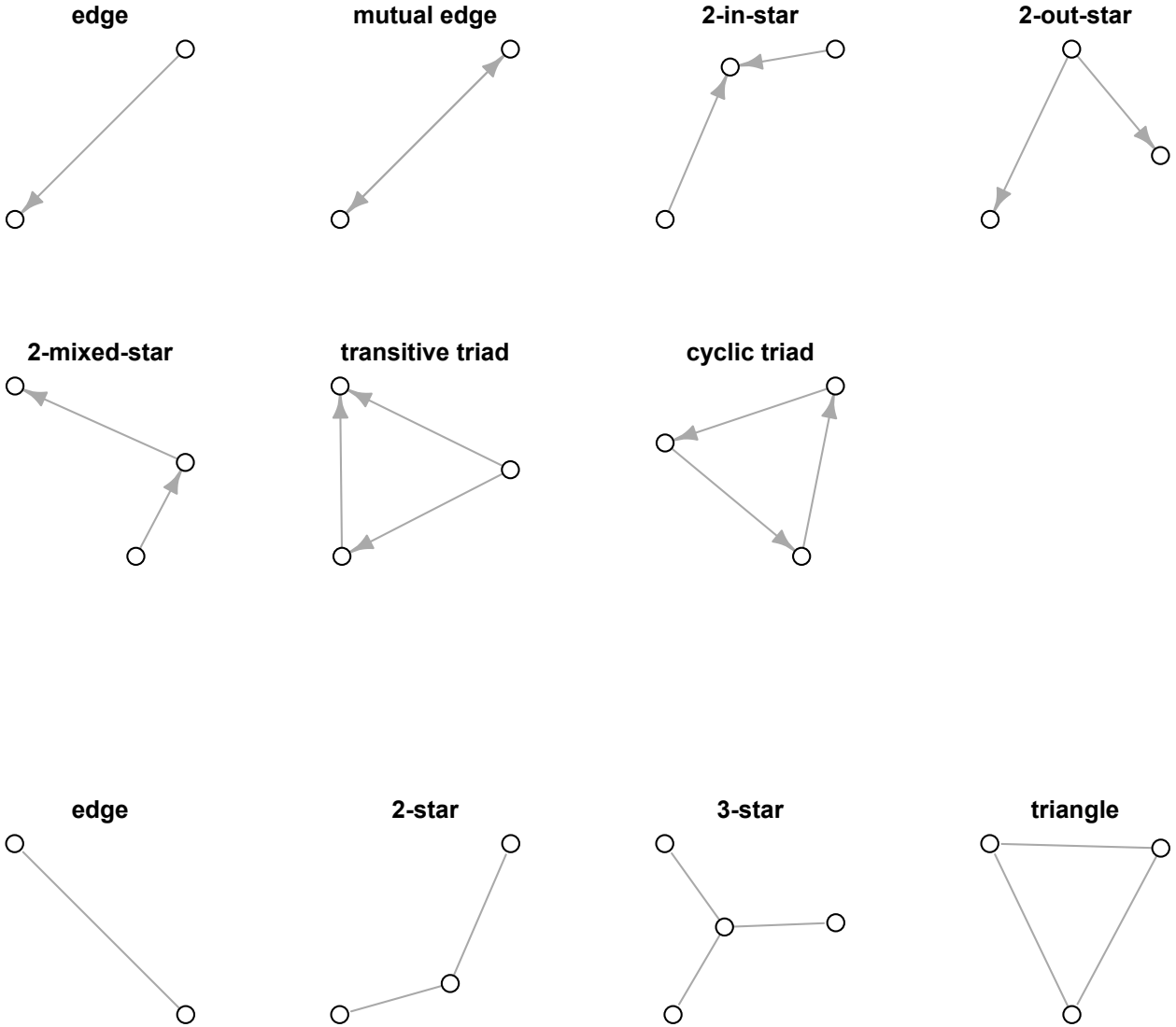
$$\pi(y|\theta) = \frac{\exp\{\theta^t s(y)\}}{z(\theta)} = \frac{q(y|\theta)}{z(\theta)}$$

- y observed graph
- $s(y)$ known vector of sufficient statistics
- θ vector of parameters
- $z(\theta)$ normalizing constant

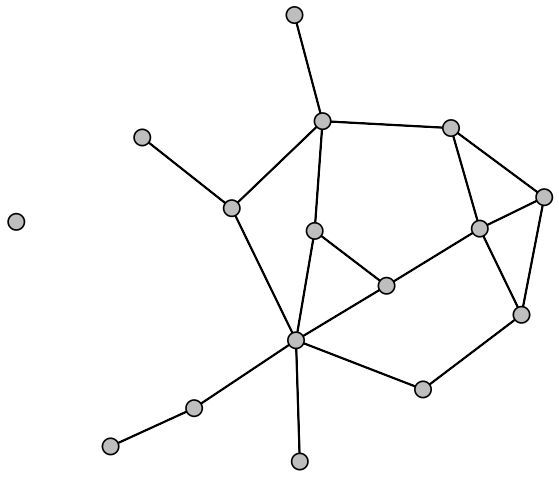
$$z(\theta) = \sum_{\textit{all possible graphs}} \exp\{\theta^t s(y)\}$$

- $2^{\binom{n}{2}}$ possible undirected graphs of n nodes
- Calculation of $z(\theta)$ is infeasible for non-trivially small graphs

Model Specification: Network Statistics



ERGM: Florentine network



Model 1:	$y \sim$ edges $+$ 3-star
Model 2:	$y \sim$ edges $+$ 2-star
Model 3:	$y \sim$ edges $+$ 2-star $+$ 3-star

ERGM: Florentine network

Here it is difficult to establish a groundtruth. For this purpose, we ran an 'independence' RJMCMC sampler:

1. Sample from each model, separately, using the exchange algorithm. (Here used the `Bergm` package of Caimo and NF (2011)).
2. RJMCMC: Use the posterior mean and variance for model k, as proposal parameters when proposing to jump to model k.

This works well, since the model space is small, but also because each posterior model is unimodal.

Acceptance rates for the jump proposals were around 40%, suggesting that the proposal distributions were a good fit to each posterior model.

This is essentially the `AutoRJ` approach outlined in Chapter 6 of Green (2003).

ERGM: Florentine network

Here it is difficult to establish a groundtruth. For this purpose, we ran an 'independence' RJMCMC sampler:

1. Sample from each model, separately, using the exchange algorithm. (Here used the `Bergm` package of Caimo and NF (2011)).
2. RJMCMC: Use the posterior mean and variance for model k, as proposal parameters when proposing to jump to model k.

This works well, since the model space is small, but also because each posterior model is unimodal.

Acceptance rates for the jump proposals were around 40%, suggesting that the proposal distributions were a good fit to each posterior model.

This is essentially the `AutoRJ` approach outlined in Chapter 6 of Green (2003).

ERGM: Florentine network

Here estimates of posterior model probabilities based on `AutoRJ` are compared to those based on estimates of the model evidence for each model.

	$\pi(m_1\|y)$	$\pi(m_2\|y)$	$\pi(m_3\|y)$
AutoRJ	0.29	0.69	0.02
"Model evidence" based	0.35	0.59	0.06

Concluding remarks

- Model evidence is difficult to compute!
- Often complex Monte Carlo methods are needed. There are plenty of methods in the Bayesian toolbox.
- A quick solution is not necessarily the best one!

References

- Chib, S. (1995) *Marginal likelihood using Gibbs output*. Journal of the American Statistical Association, 90, 1313 – 1321.
- Friel, N and Pettitt, AN (2008) *Marginal likelihood via power posteriors*. Journal of the Royal Statistical Society, Series B, 70, 589 – 607.
- Newton MA and Raftery, AE (1994) *Approximate Bayesian inference by the weighted likelihood bootstrap (with Discussion).* Journal of the Royal Statistical Society, Series B, 56, 3 – 48.
- Neal, R (2001) *Annealed importance sampling*. Statistics and Computing, 11, 125 – 139.
- Murray I., Ghahramani, Z., and MacKay, D. (2006) *MCMC for doubly-intractable distributions*. In Proceedings of the 22nd annual conference on uncertainty in artificial intelligence
- Ciamo A., Friel N. (2011) *Bayesian inference for the exponential random graph model*. Social Networks, 33, 41 – 55.
- Skilling, J. (2006) *Nested sampling for general Bayesian computation* Bayesian Analysis, 1, 833 – 860.