Lecture 8: Computation and Search

Jim Berger

Duke University

CBMS Conference on Model Uncertainty and Multiplicity
July 23-28, 2012

Outline

- Marginal likelihood computation via importance sampling
- Rui Paulo's slides on other methods
- Adaptive importance sampling and exoplanets
- Search in large model spaces and the inference challenge

I. Marginal Likelihood Computation via Importance Sampling

Goal: Computation of $m(\boldsymbol{x}) = \int f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$.

Importance sampling: Choose a proper distribution $q(\boldsymbol{\theta})$, easy to generate from, and such that $q(\boldsymbol{\theta})$ is roughly proportional to $f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$. Then

$$\begin{aligned} m(\boldsymbol{x}) &= \int \frac{f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\, q(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &\approx \frac{1}{L}\sum_{i=1}^{L} \frac{f(\boldsymbol{x} \mid \boldsymbol{\theta}^{(i)})\pi(\boldsymbol{\theta}^{(i)})}{q(\boldsymbol{\theta}^{(i)})} \quad \text{with} \quad \boldsymbol{\theta}^{(i)} \sim q(\boldsymbol{\theta})\,. \end{aligned}$$

$q(\boldsymbol{\theta})$ is called the importance function.

Choice of q is crucial: It should

- be easy to generate from;
- be roughly proportional to $f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$;
- have tails that are somewhat heavier than those of $f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

The reason $q(\boldsymbol{\theta})$ can't have too sharp tails is that the variance of the estimate is V/L,

$$V = \left(\int \frac{[f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})]^2}{q(\boldsymbol{\theta})} \, d\boldsymbol{\theta} - m^2(\boldsymbol{x}) \right) ,$$

and this may not exist if $q(\boldsymbol{\theta})$ has tails that are too light, which can result in an extremely slow converging algorithm.

Note: Assuming this variance is finite, it can be estimated by $\hat{V}/L$,

$$\hat{V} = \left(\frac{1}{L} \sum_{i=1}^{L} \left[\frac{f(\boldsymbol{x} \mid \boldsymbol{\theta}^{(i)})\pi(\boldsymbol{\theta}^{(i)})}{q(\boldsymbol{\theta}^{(i)})} \right]^2 - \left[\frac{1}{L} \sum_{i=1}^{L} \frac{f(\boldsymbol{x} \mid \boldsymbol{\theta}^{(i)})\pi(\boldsymbol{\theta}^{(i)})}{q(\boldsymbol{\theta}^{(i)})} \right]^2 \right) .$$

- One of the great advantages of importance sampling is the ease with which one can estimate accuracy.
- Some care is needed: monitor $\hat{V}$ as L increases to make sure it is not increasing.

Common choices of the importance function:

- If there is little data, choosing $q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$ is okay, and then $m(\boldsymbol{x}) \approx \frac{1}{L} \sum_{i=1}^{L} f(\boldsymbol{x} \mid \boldsymbol{\theta}^{(i)})$.
- If there is a lot of data and $\pi(\boldsymbol{\theta})$ does not have sharp tails (e.g. is Cauchy) choosing $q(\boldsymbol{\theta}) \propto f(\boldsymbol{x} \mid \boldsymbol{\theta})$ is okay, if the likelihood is easy to generate from (e.g., is normal).
- A common choice is $q(\boldsymbol{\theta}) = N(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{I}})$ where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{I}}$ are the mle and observed Fisher information matrix for $\boldsymbol{\theta}$.
  - But the normal distribution has too sharp tails, so a much better choice is a t-distribution with four degrees of freedom.
  - Better yet is $q(\boldsymbol{\theta}) = T_4(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}, c\hat{\boldsymbol{I}})$, and try difference $c > 1$ until convergence is fast.
  - Better yet is $q(\boldsymbol{\theta}) = T_4(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}_\pi, c\hat{\boldsymbol{I}}_\pi)$, where $\hat{\boldsymbol{\theta}}_\pi$ and $\hat{\boldsymbol{I}}_\pi$ are the maximizer and Hessian of $f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

II. Rui Paulo's Slides on Other Methods

Bayes Factors and Marginal Likelihoods

Rui Paulo
NISS/SAMSI
June 26, 2003

Problem Description

- The posterior distribution on the model space

$$\pi(\gamma \mid \boldsymbol{y}) = \frac{\pi(\gamma)\ m(\boldsymbol{y} \mid \gamma)}{\sum_{\gamma'} \pi(\gamma')\ m(\boldsymbol{y} \mid \gamma')}$$
$$= \left[1 + \sum_{\gamma' \neq \gamma} \pi_{\gamma':\gamma}\ B_{\gamma':\gamma}\right]^{-1}$$

  where

  - $\pi(\gamma)$ is the prior probability of model $\mathcal{M}_\gamma$;
  - $m(\boldsymbol{y} \mid \gamma) = \int f(\boldsymbol{y} \mid \theta_\gamma, \gamma)\ \pi(\theta_\gamma \mid \gamma)\ d\theta_\gamma$ is the marginal likelihood under model $\mathcal{M}_\gamma$;
  - $\pi_{\gamma':\gamma} = \pi_{\gamma'}/\pi_\gamma$;
  - $B_{\gamma':\gamma} = m(\boldsymbol{y} \mid \gamma')/m(\boldsymbol{y} \mid \gamma)$.

Strategies

The basic goal is to characterize the posterior distribution $\pi(\gamma \mid \boldsymbol{y})$, and for that several methods are available

I– Single-chain methods — Markov chain that moves in the model space producing a sample $\{\gamma^{(i)},\ i = 1, \ldots, M\}$ from $\pi(\gamma \mid \boldsymbol{y})$.

The Monte Carlo frequencies

$$\frac{\text{number of } \gamma^{(i)} = \gamma}{M}$$

are possible estimates for $\pi(\gamma \mid \boldsymbol{y})$.

II– Methods that require one chain per model or that compute either the marginal likelihoods or the Bayes factors one at a time, making use of the formulas on the previous slide to compute the posterior model probabilities.

I — Single-Chain Methods

- Methods that sample over the model space alone, which requires integration of the model-specific parameters θ_γ. Examples are Madigan and York (1995), Raftery et al. (1997), George and McCulloch (1997).
- Methods that sample over the model space and parameter space jointly
  - Reversible Jump of Green (1995)
  - Product space search of Carlin and Chib (1995)
  - Metropolized Carlin and Chib of Dellaportas et al. (1998)
  - Composite model space approach of Godsill (2001).

II — One Chain per Model Methods

- Chib's methods estimate the marginals under each model using a chain from that model along with details of the sampling mechanism used to produce it; (Chib, 1995 and Chib and Jeliazkov, 2001)
- RIS-IWMDE of Ibrahim, Chen, and MacEachern (1999) and Chen, Ibrahim and Yiannoutsos (1999). Uses only one sample from the full model to estimate all Bayes factors, no details of the sampling mechanism are used;
- Recent method by Ming-Hui Chen, which again uses only one sample from the full model, but in this case estimates all marginal likelihoods. No details of the sampling method are used either.
- Importance Sampling can be used to estimate the marginals under each model. Requires tuning of the importance function, which may be done using only one sample from the full model;

Aside — the IWMDE of Chen (1994)

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(j)}, \boldsymbol{\theta}_{(-j)})$ be a parameter and consider its posterior distribution,

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\ f(\boldsymbol{y} \mid \boldsymbol{\theta})\ .$$

The goal is to estimate the marginal posterior at a particular point, i.e., to estimate

$$\pi_j(\boldsymbol{\theta}^\star_{(j)} \mid \boldsymbol{y}) = \int \pi(\boldsymbol{\theta}^\star_{(j)}, \boldsymbol{\theta}_{(-j)} \mid \boldsymbol{y})\ d\boldsymbol{\theta}_{(-j)}\ .$$

Chen showed that, if $\omega(\boldsymbol{\theta}_{(j)} \mid \boldsymbol{\theta}_{(-j)})$ is a conditional density, we have that

$$\pi_j(\boldsymbol{\theta}^\star_{(j)} \mid \boldsymbol{y}) = \mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}} \frac{\pi(\boldsymbol{\theta}^\star_{(j)}, \boldsymbol{\theta}_{(-j)} \mid \boldsymbol{y})}{\pi(\boldsymbol{\theta}_{(j)}, \boldsymbol{\theta}_{(-j)} \mid \boldsymbol{y})}\ \omega(\boldsymbol{\theta}_{(j)} \mid \boldsymbol{\theta}_{(-j)})\ .$$

IWMDE of Chen (1994)

- This estimator as been named Importance-Weighted Marginal Density Estimator (IWMDE) by Chen (1994).
- The choice $\omega(\boldsymbol{\theta}_{(j)} \mid \boldsymbol{\theta}_{(-j)}) = \pi(\boldsymbol{\theta}_{(j)} \mid \boldsymbol{\theta}_{(-j)}, \boldsymbol{y})$ results in the conditional marginal density estimator (CMDE) of Gelfand et al. (1992).
- Chen notes that the CMDE is optimal among all IWMDE.
- There is a default choice for ω based on approximating the posterior by a multivariate normal and using for ω the induced conditional.

Chib's Methods

- These methods are aimed at computing marginal likelihoods and are based on the trivial but fundamental identity

$$m(\boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta})\ \pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta} \mid \boldsymbol{y})}\ , \quad \forall \boldsymbol{\theta}\ .$$

- Find a way of estimating $\pi(\boldsymbol{\theta}^\star \mid \boldsymbol{y})$ and you will have an estimate of the marginal likelihood.
- $\boldsymbol{\theta}^\star$ is usually chosen to be a point of high posterior density.

Following are several ways of using this idea in different contexts.

Latent Variable Model

- Suppose that

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) = \int \pi(\boldsymbol{\theta}, \boldsymbol{u} \mid \boldsymbol{y})\, d\boldsymbol{u}$$

  where $\boldsymbol{u}$ might be a vector of imputed latent variables.

- The posterior distribution of $\boldsymbol{\theta}$ can be viewed as the marginal posterior density of $(\boldsymbol{\theta}, \boldsymbol{u})$, and IWMDE can be used to estimate $\pi(\boldsymbol{\theta}^\star \mid \boldsymbol{y})$.

- If $[\boldsymbol{\theta} \mid \boldsymbol{u}, \boldsymbol{y}]$ is available in closed form, CMDE can be used to estimate $\pi(\boldsymbol{\theta}^\star \mid \boldsymbol{y})$, giving rise to the estimate

$$\hat{m}(\boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}^\star)\, \pi(\boldsymbol{\theta}^\star)}{\frac{1}{M} \sum_{j=1}^{M} \pi(\boldsymbol{\theta}^\star \mid \boldsymbol{u}^{(j)}, \boldsymbol{y})}\ ,$$

  where $\{\boldsymbol{u}^{(j)}, j = 1, \ldots, M\}$ is a sample from the posterior distribution of $\boldsymbol{u}$.

Two-Block Gibbs Sampler

- Suppose $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and that both full conditionals are available in closed form, i.e., including the normalizing constant.
- It is clear that

$$\pi(\boldsymbol{\theta}^\star \mid \boldsymbol{y}) = \underbrace{\pi(\boldsymbol{\theta}_1^\star \mid \boldsymbol{y})}_{\text{CMDE}} \underbrace{\pi(\boldsymbol{\theta}_2^\star \mid \boldsymbol{\theta}_1^\star, \boldsymbol{y})}_{\text{known}} .$$

- In effect,

$$\hat{\pi}(\boldsymbol{\theta}_1^\star \mid \boldsymbol{y}) = \frac{1}{M} \sum_{j=1}^{M} \pi(\boldsymbol{\theta}_1^\star \mid \boldsymbol{\theta}_2^{(j)}, \boldsymbol{y}) .$$

  Above, $\{\boldsymbol{\theta}_2^{(j)}\}$ is a sample from the posterior distribution of $\boldsymbol{\theta}_2$.

Two-Block Gibbs Sampler

Remarks:

1. If $\pi(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2^\star, \boldsymbol{y})$ is not available in closed form, in which case Metropolis-Hastings would in principle be used to sample from this full conditional, the idea can still be used but $\pi(\boldsymbol{\theta}_1^\star \mid \boldsymbol{y})$ can be estimated using IWMDE.

2. The method can be extended to the situation where you have a k-block Gibbs sampler, but we would have to introduce the idea of a "reduced run" of a Gibbs sampler.

The two previous examples were explored in Chib (1995). The next idea was introduced in Chib and Jeliazkov (2001).

Metropolis-Hastings step

- Suppose $\boldsymbol{\theta}$ is sampled in one block using the Metropolis-Hastings algorithm.

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}' \mid \boldsymbol{y}) = 1 \wedge \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}')\ \pi(\boldsymbol{\theta}')}{f(\boldsymbol{y} \mid \boldsymbol{\theta})\ \pi(\boldsymbol{\theta})}\ \frac{q(\boldsymbol{\theta}', \boldsymbol{\theta} \mid \boldsymbol{y})}{q(\boldsymbol{\theta}, \boldsymbol{\theta}' \mid \boldsymbol{y})}$$

- Detailed balance

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\star \mid \boldsymbol{y})\ q(\boldsymbol{\theta}, \boldsymbol{\theta}^\star \mid \boldsymbol{y})\ \pi(\boldsymbol{\theta} \mid \boldsymbol{y}) = \pi(\boldsymbol{\theta}^\star \mid \boldsymbol{y})\ \alpha(\boldsymbol{\theta}^\star, \boldsymbol{\theta} \mid \boldsymbol{y})\ q(\boldsymbol{\theta}^\star, \boldsymbol{\theta} \mid \boldsymbol{y})\ .$$

  Manipulating the above formula and integrating over $\boldsymbol{\theta}$ shows that

$$\pi(\boldsymbol{\theta}^\star \mid \boldsymbol{y}) = \frac{\mathsf{E}_1\ \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\star \mid \boldsymbol{y})\ q(\boldsymbol{\theta}, \boldsymbol{\theta}^\star \mid \boldsymbol{y})}{\mathsf{E}_2\ \alpha(\boldsymbol{\theta}^\star, \boldsymbol{\theta} \mid \boldsymbol{y})}\ ,$$

  where

  E_1 — expectation with respect to $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$,

  E_2 — expectation with respect to $q(\boldsymbol{\theta}^\star, \boldsymbol{\theta} \mid \boldsymbol{y})$.

Metropolis-Hastings step

Remarks:

1. The method can be extended to multiple parameter-blocks and to a latent variable framework. Check Chib and Jeliazkov (2001) for details.
2. Chib and Jeliazkov (2001) together with Chib (1995) provide a framework under which virtually any model that can be fit using MCMC techniques can have its marginal likelihood estimated.
3. One must know the details of the sampling mechanism in order to apply the methods.
4. Note the need for additional sampling from the proposal.
5. Chib and Jeliazkov (2001) conclude that if a sampling scheme is efficient in sampling from the posterior, it will give rise to an efficient method to compute the marginal likelihood.

Aside — Ratio Importance Sampling

- Suppose we want to compute

$$\frac{m_1}{m_2}$$

where m_i is the unknown normalizing constant of the un-normalized density p_i, i.e.

$$\pi_i(\boldsymbol{\theta}_i) = \frac{p_i(\boldsymbol{\theta}_i)}{m_i} \ , \quad i = 1, 2,$$

are probability densities.

The Ratio Importance Sampling identity of Chen and Shao (1997) is

$$\frac{m_1}{m_2} = \frac{\mathsf{E}_q \ p_1(\boldsymbol{\theta}_1)/q(\boldsymbol{\theta}_1)}{\mathsf{E}_q \ p_2(\boldsymbol{\theta}_2)/q(\boldsymbol{\theta}_2)} \ ,$$

where q is defined over $\Omega_1 \cup \Omega_2$, the union of the supports of each density.

Ratio Importance Sampling

As noted in the paper, if π_1 and π_2 have different dimensions, this formula is not directly applicable. Instead, this paper suggests the following idea.

Consider the case where

$$m_1 = \int p_1(\boldsymbol{\theta})\ d\boldsymbol{\theta}$$

$$m_2 = \int p_2(\boldsymbol{\theta}, \boldsymbol{\psi})\ d\boldsymbol{\theta}\ d\boldsymbol{\psi}$$

and again the goal is to estimate m_1/m_2.

Let

$$p_1^\star(\boldsymbol{\theta}, \boldsymbol{\psi}) = p_1(\boldsymbol{\theta})\ \omega(\boldsymbol{\psi} \mid \boldsymbol{\theta})$$

where ω is a completely known conditional density. Then, it is clear that

$$m_1^\star = \int p_1^\star(\boldsymbol{\theta}, \boldsymbol{\psi})\ d\boldsymbol{\theta}\ d\boldsymbol{\psi} = m_1$$

Ratio Importance Sampling

As a consequence, using RIS,

$$\frac{m_1}{m_2} = \frac{m_1^\star}{m_2} = \frac{\mathsf{E}_q \; p_1(\boldsymbol{\theta}) \; \omega(\boldsymbol{\psi} \mid \boldsymbol{\theta}) / q(\boldsymbol{\theta}, \boldsymbol{\psi})}{\mathsf{E}_q \; p_2(\boldsymbol{\theta}, \boldsymbol{\psi}) / q(\boldsymbol{\theta}, \boldsymbol{\psi})}$$

The choice $q = \pi_2 = p_2/m_2$ is particularly interesting in that it simplifies to

$$\frac{m_1}{m_2} = \mathsf{E}_{\pi_2} \frac{p_1(\boldsymbol{\theta}) \; \omega(\boldsymbol{\psi} \mid \boldsymbol{\theta})}{p_2(\boldsymbol{\theta}, \boldsymbol{\psi})} \; .$$

Note that you only need a sample from π_2 in order to estimate the ratio of the normalizing constants. This idea is explored next in the context of model selection.

The RIS-IWMDE

- Model $\mathcal{M}_\gamma$ has parameter vector $\boldsymbol{\beta}_{(\gamma)}$
- The full model parameter vector is $\boldsymbol{\beta} = (\boldsymbol{\beta}_{(\gamma)}, \boldsymbol{\beta}_{(-\gamma)})$.
- We will index the full model by $\gamma = 1$.
- Direct application of the RIS formula

$$\frac{m(\boldsymbol{y} \mid \gamma)}{m(\boldsymbol{y} \mid 1)} = \mathsf{E}_{\boldsymbol{\beta}|\boldsymbol{y}} \, \frac{f_\gamma(\boldsymbol{y} \mid \boldsymbol{\beta}_{(\gamma)}) \; \pi_\gamma(\boldsymbol{\beta}_{(\gamma)})}{f_1(\boldsymbol{y} \mid \boldsymbol{\beta}) \; \pi_1(\boldsymbol{\beta})} \; \omega(\boldsymbol{\beta}_{(-\gamma)} \mid \boldsymbol{\beta}_{(\gamma)})$$

  where the expectation is taken with respect to the full model.
- Sample from the full model allows for estimating all Bayes factors!
- This idea has been proposed and derived in Ibrahim, Chen, and MacEachern (1999) and Chen, Ibrahim and Yiannoutsos (1999).

The RIS-IWMDE

- The optimal choice for ω is $\pi(\boldsymbol{\beta}_{(-\gamma)} \mid \boldsymbol{\beta}_{(\gamma)}, \boldsymbol{y})$, which is typically not available, and so the empirical method suggested by Chen (1994) can be used instead.

- If the priors are compatible by conditioning, i.e. if

$$\pi_\gamma(\boldsymbol{\beta}_{(\gamma)}) = \pi_1(\boldsymbol{\beta}_{(\gamma)} \mid \boldsymbol{\beta}_{(-\gamma)} = \mathbf{0}) \ ,$$

  then we have

$$\begin{aligned} \frac{m(\boldsymbol{y} \mid \gamma)}{m(\boldsymbol{y} \mid 1)} &= \frac{1}{\pi_1(\boldsymbol{\beta}_{(-\gamma)} = \mathbf{0})} \, \mathsf{E}_{\boldsymbol{\beta}|\boldsymbol{y}} \, \frac{\pi_1(\boldsymbol{\beta}_{(\gamma)}, \boldsymbol{\beta}_{(-\gamma)} = \mathbf{0} \mid \boldsymbol{y})}{\pi_1(\boldsymbol{\beta}_{(\gamma)}, \boldsymbol{\beta}_{(-\gamma)} \mid \boldsymbol{y})} \, \omega(\boldsymbol{\beta}_{(-\gamma)} \mid \boldsymbol{\beta}_{(\gamma)}) \\ &= \frac{\pi_1(\boldsymbol{\beta}_{(-\gamma)} = \mathbf{0} \mid \boldsymbol{y})}{\pi_1(\boldsymbol{\beta}_{(-\gamma)} = \mathbf{0})} \qquad \text{by IWMDE.} \end{aligned}$$

Recent Method by Chen

Latent variable model:

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) = \int \pi(\boldsymbol{\theta}, \boldsymbol{u} \mid \boldsymbol{y}) \; d \; \boldsymbol{u} \; .$$

A direct application of IWMDE leads to

$$\begin{aligned} \pi(\boldsymbol{\theta}^\star \mid \boldsymbol{y}) &= \mathsf{E}_{\boldsymbol{\theta},\boldsymbol{u}|\boldsymbol{y}} \; \frac{f(\boldsymbol{y}, \boldsymbol{u} \mid \boldsymbol{\theta}^\star) \; \pi(\boldsymbol{\theta}^\star)}{f(\boldsymbol{y}, \boldsymbol{u} \mid \boldsymbol{\theta}) \; \pi(\boldsymbol{\theta})} \; \omega(\boldsymbol{\theta} \mid \boldsymbol{u}) \\ &= \pi(\boldsymbol{\theta}^\star) \; \mathsf{E}_{\boldsymbol{\theta},\boldsymbol{u}|\boldsymbol{y}} \; \frac{\omega(\boldsymbol{\theta} \mid \boldsymbol{u})}{\pi(\boldsymbol{\theta})} \; \frac{f(\boldsymbol{y}, \boldsymbol{u} \mid \boldsymbol{\theta}^\star)}{f(\boldsymbol{y}, \boldsymbol{u} \mid \boldsymbol{\theta})} \end{aligned}$$

Recent Method by Chen

- Application of this equality in the variable selection problem.
- If the priors are compatible by conditioning then one has the identity

$$\pi_\gamma(\boldsymbol{\beta}_{(\gamma)}, \boldsymbol{u} \mid \boldsymbol{y}) = \pi_1(\boldsymbol{\beta}_{(-\gamma)} = \mathbf{0} \mid \boldsymbol{\beta}_{(\gamma)}, \boldsymbol{u}, \boldsymbol{y})\, \pi_1(\boldsymbol{\beta}_{(\gamma)}, \boldsymbol{u} \mid \boldsymbol{y}) / \pi_1(\boldsymbol{\beta}_{(\gamma)} = \mathbf{0} \mid \boldsymbol{y})$$

- Substitute in the formula previously derived

$$\pi_\gamma(\boldsymbol{\beta}^\star_{(\gamma)} \mid \boldsymbol{y}) = \frac{\pi_\gamma(\boldsymbol{\beta}^\star_{(\gamma)})}{\pi_1(\boldsymbol{\beta}_{(-\gamma)} = \mathbf{0} \mid \boldsymbol{y})} \times$$

$$\times\, \mathsf{E}_{\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{y}} \frac{\omega(\boldsymbol{\beta}^\star_{(\gamma)} \mid \boldsymbol{u})}{\pi_\gamma(\boldsymbol{\beta}^\star_{(\gamma)})} \frac{f(\boldsymbol{y}, \boldsymbol{u} \mid \boldsymbol{\beta}^\star_{(\gamma)})}{f(\boldsymbol{y}, \boldsymbol{u} \mid \boldsymbol{\beta}_{(\gamma)})} \pi_1(\boldsymbol{\beta}_{(-\gamma)} = \mathbf{0} \mid \boldsymbol{\beta}_{(\gamma)}, \boldsymbol{u}, \boldsymbol{y}) \ ,$$

where the expectation is taken with respect to the posterior of the parameter and latent variable vectors under the full model.

Remarks

- Like RIS-IWMDE, the above formula allows for estimation of all posterior model probabilities using only a sample from the posterior of the full model.
- One needs to be able to evaluate analytically the conditional

$$\pi_1(\boldsymbol{\beta}_{(-\gamma)} = \mathbf{0} \mid \boldsymbol{\beta}_{(\gamma)})$$

  which is a considerable constraint in terms of applicability.
- One also needs to estimate

$$\pi_1(\boldsymbol{\beta}_{(-\gamma)} = \mathbf{0} \mid \boldsymbol{y})$$

  which can be done using IWMDE but is certainly prone to instability if $\boldsymbol{\beta}_{(-\gamma)} = \mathbf{0}$ is not a point of reasonable density under the full model posterior. Recall that RIS-IWMDE suffers from a similar drawback.
- If the optimal ω is not available, the choice for g is tricky if $\boldsymbol{u}$ is high dimensional, and repeatedly evaluating $f(\boldsymbol{y}, \boldsymbol{u} \mid \boldsymbol{\beta}_{(\gamma)})$ can be computationally intensive in the same setting.

Importance Sampling

- The idea behind importance sampling and how it applies to marginal likelihood estimation is quite easy to convey:

$$\begin{aligned} m(\boldsymbol{y}) &= \int \pi(\boldsymbol{\theta})\ f(\boldsymbol{y} \mid \boldsymbol{\theta})\ d\boldsymbol{\theta} \\ &= \int \frac{\pi(\boldsymbol{\theta})\ f(\boldsymbol{y} \mid \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\ q(\boldsymbol{\theta})\ d\boldsymbol{\theta} \\ &\approx \frac{1}{M} \sum_{j=1}^{M} \frac{\pi(\boldsymbol{\theta}^{(j)})\ f(\boldsymbol{y} \mid \boldsymbol{\theta}^{(j)})}{q(\boldsymbol{\theta}^{(j)})} \end{aligned}$$

  where $\{\boldsymbol{\theta}^{(j)},\ j = 1, \ldots, M\}$ is a sample from $q(\cdot)$.

- It is well-known that the method will be as good as the importance function chosen, and that it is hard to find good importance functions as the dimensionality of the problem grows.

Importance Sampling

- Default choices:

$$t_\nu(\hat{\boldsymbol{\theta}}, [I(\hat{\boldsymbol{\theta}})]^{-1})$$
$$t_\nu(\mathsf{E}[\boldsymbol{\theta} \mid \boldsymbol{y}], \mathsf{Var}[\boldsymbol{\theta} \mid \boldsymbol{y}])$$

- In the case of variable selection, one can actually tune the importance function using information from the full model only; one can use the induced conditional distributions as importance functions for the submodels.
- Note that contrary to the other methods that only require a sample from the full model, here additional random variables have to be generated.

Example: Probit Regression Model

- This is an example where most techniques are applicable.
- The model:

$$y_i \mid p_i \sim \text{Ber}(p_i),\ i = 1, \ldots, N,\ \text{independently}$$
$$p_i = \Phi(\boldsymbol{x}_i' \boldsymbol{\beta})$$
$$\boldsymbol{\beta} \sim \text{N}(\mathbf{0}, g\ (\boldsymbol{X}'\boldsymbol{X})^{-1}) \text{ where we take } g = N.$$

- Latent variable formulation (Albert and Chib, 1993)

$$z_i \mid \boldsymbol{\beta} \sim \text{N}(\boldsymbol{x}_i' \boldsymbol{\beta}, 1)$$
$$y_i = I\{z_i > 0\}\ .$$

Gibbs Sampler

Albert and Chib (1993)

- $\boldsymbol{z} \mid \boldsymbol{\beta}, \boldsymbol{y} \sim \prod_{i=1}^{N} [I\{y_i = 1\}\ I\{z_i \leq 0\} + I\{y_i = 0\}\ I\{z_i > 0\}]\ \phi(z_i \mid \boldsymbol{x}_i'\boldsymbol{\beta}, 1)$

- $\boldsymbol{\beta} \mid \boldsymbol{z} \sim \mathsf{N}(\frac{g}{1+g}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}, \frac{g}{1+g}\ (\boldsymbol{X}'\boldsymbol{X})^{-1})$

This fits very nicely into Chib's method, latent variable variant, and furthermore CMDE can be used since $[\boldsymbol{\beta} \mid \boldsymbol{z}, \boldsymbol{y}]$ is known in closed form

$$\hat{m}(\boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \boldsymbol{\beta}^\star)\ \pi(\boldsymbol{\beta}^\star)}{\frac{1}{M}\sum_{j=1}^{M} \pi(\boldsymbol{\beta}^\star \mid \boldsymbol{z}^{(j)}, \boldsymbol{y})}\ ,$$

where we have set $\boldsymbol{\beta}^\star$ equal to its MLE.

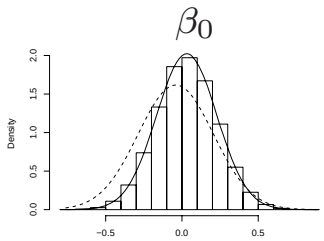
Metropolis-Hastings Step

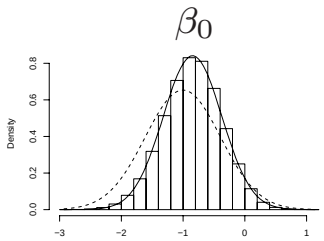
Without introducing the latent variable, it is easy to do a one-step Metropolis update with proposal

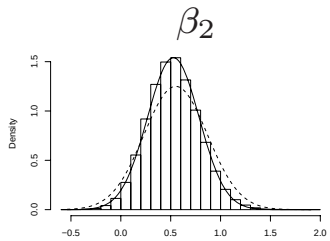
$$q(\boldsymbol{\beta}, \boldsymbol{\beta}' \mid \boldsymbol{y}) \sim t_\nu(\hat{\boldsymbol{\beta}}, c\, [I(\hat{\boldsymbol{\beta}})]^{-1})$$

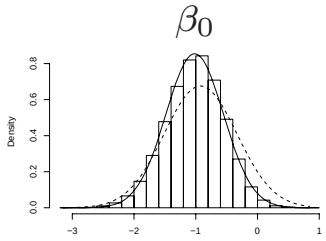
where c can be easily tuned ($c = 2.4/\sqrt{p}$ is a good guess.)

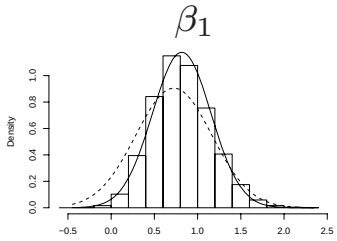
This sampling scheme fits nicely into Chib and Jeliazkov's method, becoming very simple if the point of high density is chosen to be the MLE.

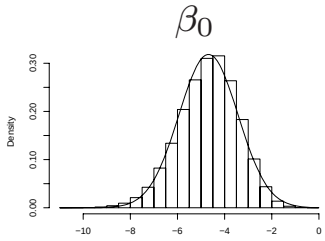

β_0
Density

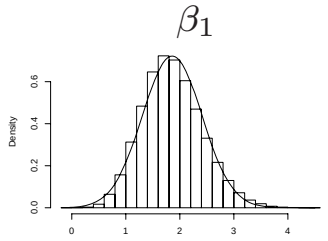

β_0
Density

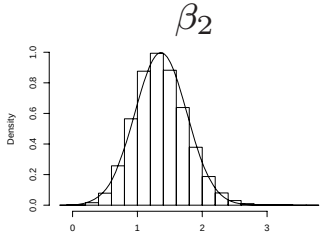

β_2
Density

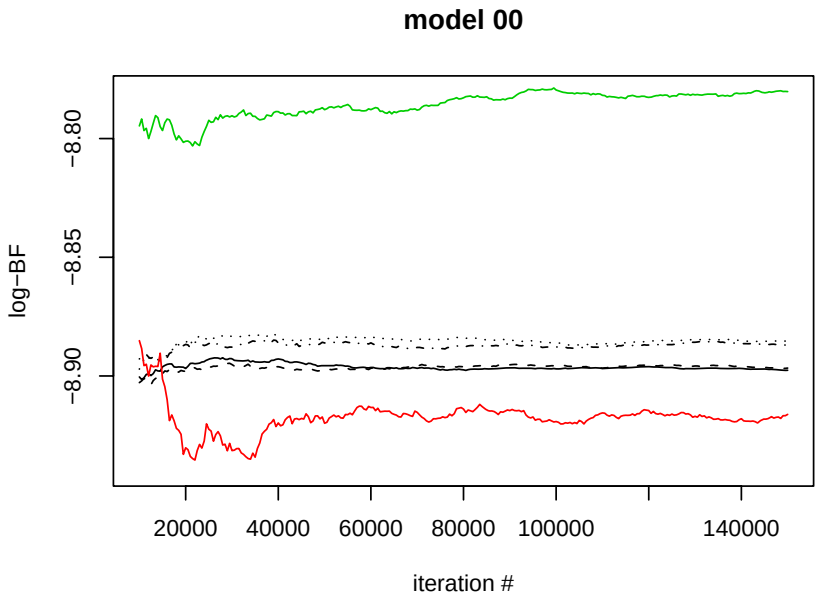

β_0
Density

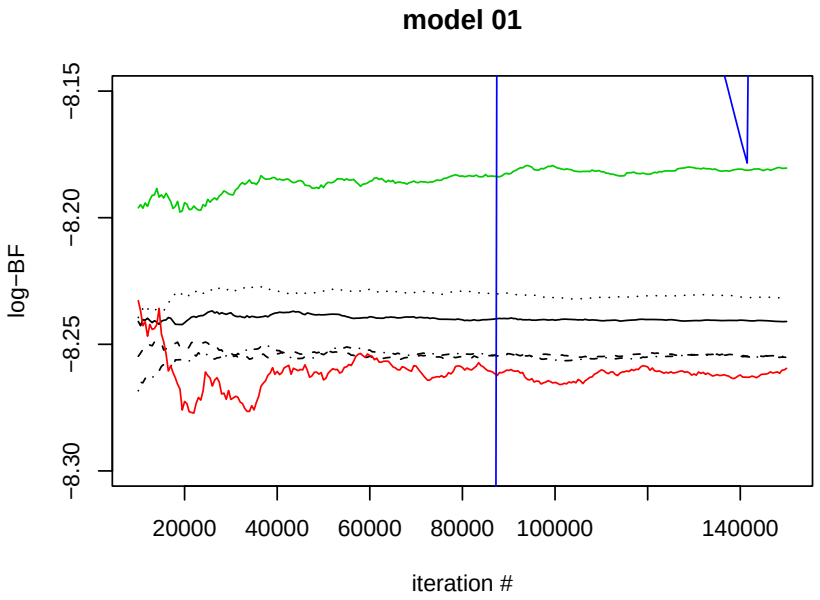

β_1
Density

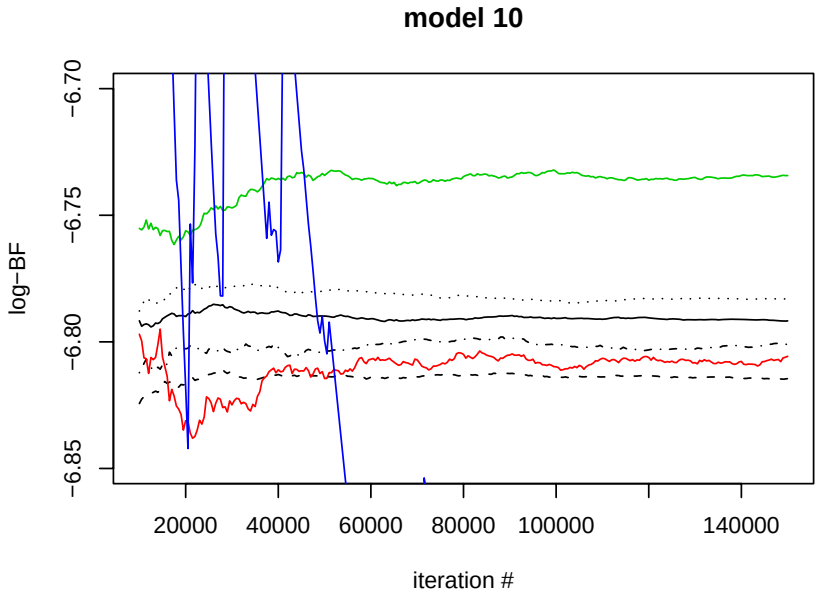

β_0
Density

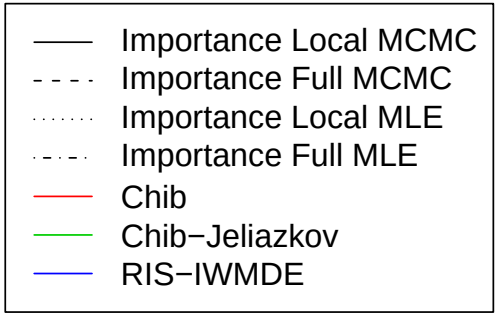

β_1
Density


β_2
Density


model 00
log−BF
−8.80
−8.85
−8.90
20000
40000
60000
80000
100000
140000
iteration #


model 01
log−BF
−8.15
−8.20
−8.25
−8.30
20000
40000
60000
80000
100000
140000
iteration #


model 10
log−BF
−6.70
−6.75
−6.80
−6.85
20000
40000
60000
80000
100000
140000
iteration #


Importance Local MCMC
Importance Full MCMC
Importance Local MLE
Importance Full MLE
Chib
Chib−Jeliazkov
RIS−IWMDE

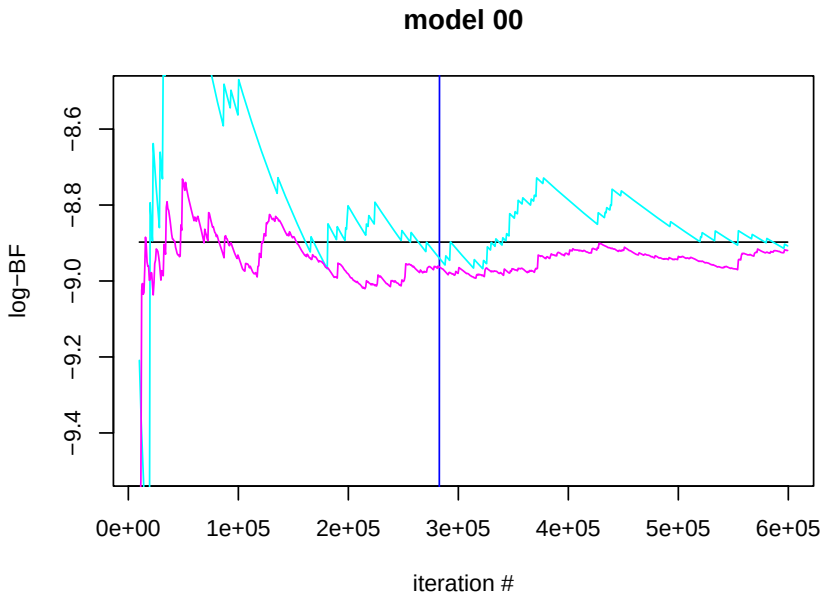
Reversible Jump

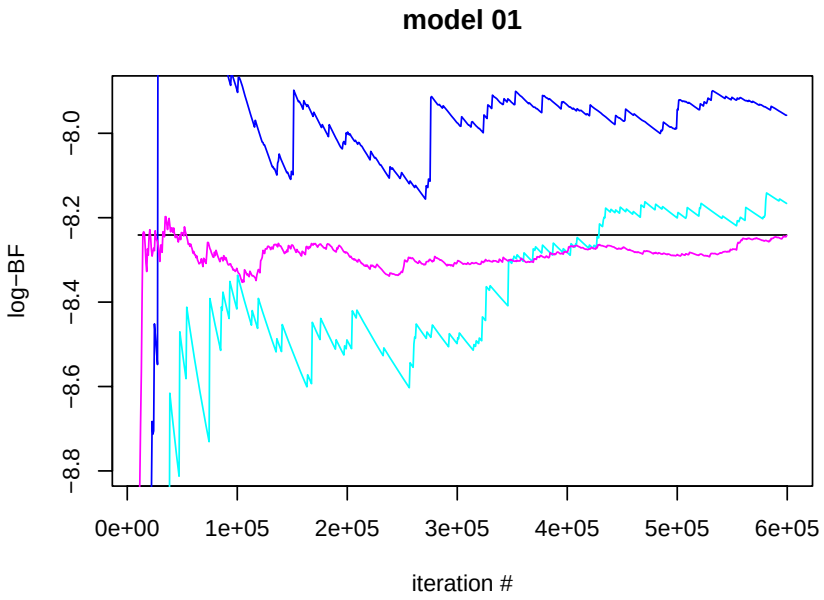
- If we propose to jump from model γ to model γ', we match the dimensions by setting $u = \boldsymbol{\beta}_{\gamma'}$, so that

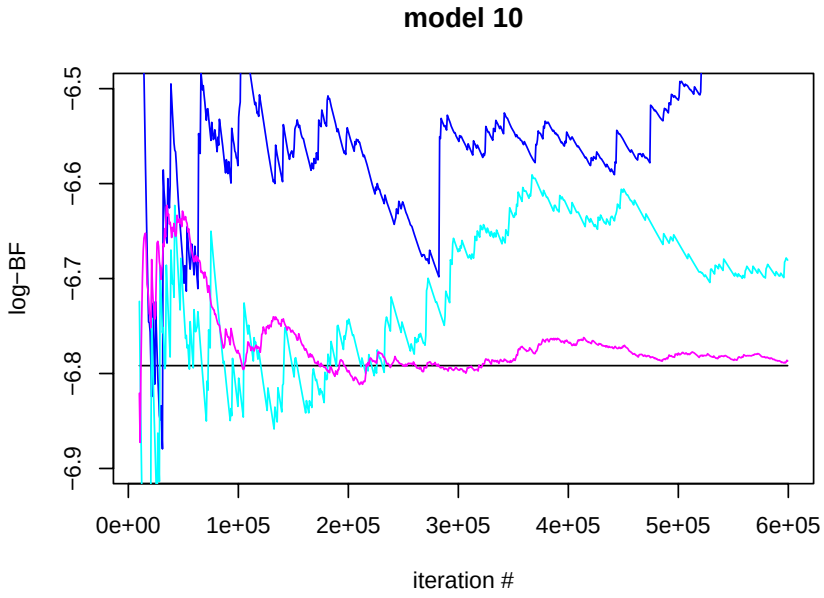
$$(\boldsymbol{\beta}_{\gamma'}, u') = g_{\gamma,\gamma'}(\boldsymbol{\beta}_{\gamma}, u)$$

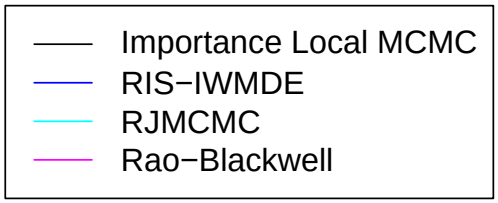
  with $g_{\gamma,\gamma'}(a, b) = (b, a)$.

- The proposal $q(\cdot \mid \boldsymbol{\beta}_{\gamma}, \gamma, \gamma')$ is simply, like before, a t density centered at the MLE and with a scale matrix proportional to the inverse of the Fisher information at the MLE.


model 00
log−BF
−8.6
−8.8
−9.0
−9.2
−9.4
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
iteration #


model 01
log−BF
−8.0
−8.2
−8.4
−8.6
−8.8
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
iteration #


model 10
log−BF
−6.5
−6.6
−6.7
−6.8
−6.9
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
iteration #


Importance Local MCMC
RIS−IWMDE
RJMCMC
Rao−Blackwell

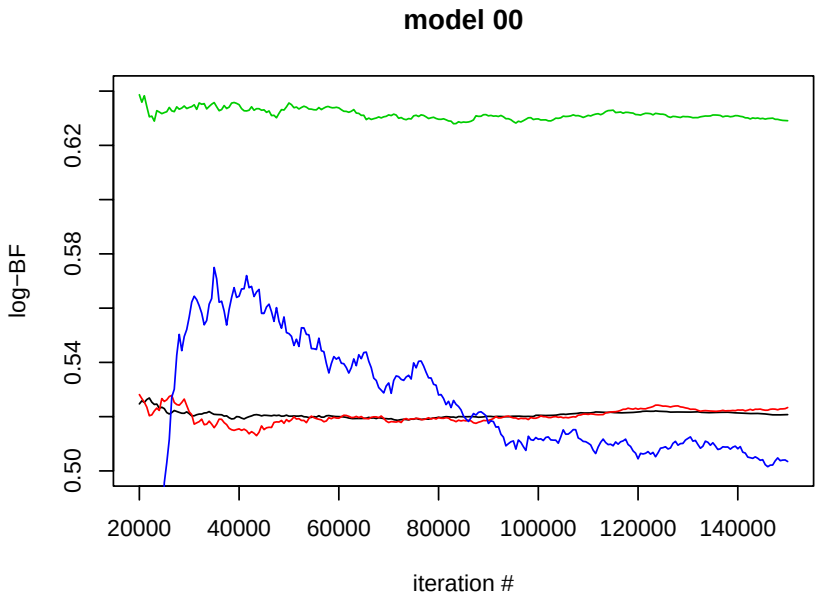
Rao-Blackwellized Estimates

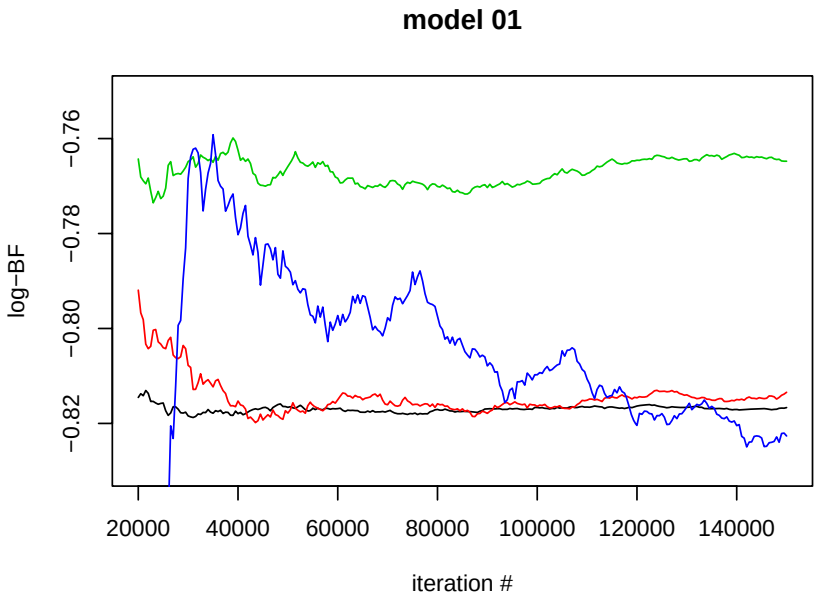
Introducing the latent variables $\boldsymbol{z}$, we can implement the following algorithm to sample jointly from the model and parameter spaces.

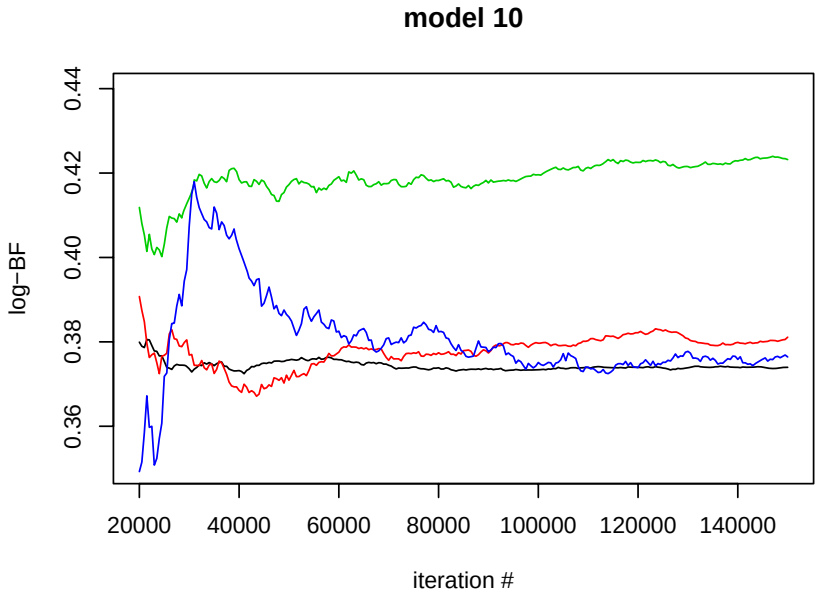
- $\gamma \mid \boldsymbol{y}, \boldsymbol{z} \propto (1+g)^{-k_\gamma/2} \ \exp(\boldsymbol{z}'\boldsymbol{P}_\gamma \boldsymbol{z}/2)$
- $\boldsymbol{\beta} \mid \gamma, \boldsymbol{y}, \boldsymbol{z} \sim \mathsf{N}(\frac{g}{1+g}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}, \frac{g}{1+g}\ (\boldsymbol{X}'\boldsymbol{X})^{-1})$
- $\boldsymbol{z} \mid \boldsymbol{\beta}, \boldsymbol{y}, \gamma \sim \prod_{i=1}^{N} [I\{y_i = 1\}\ I\{z_i > 0\} + I\{y_i = 0\}\ I\{z_i \leq 0\}]\ \phi(z_i \mid \boldsymbol{x}_i'\boldsymbol{\beta}, 1)$

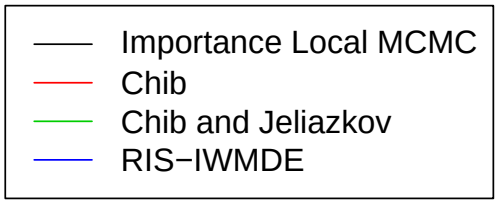
And then estimate the posterior model probabilities by

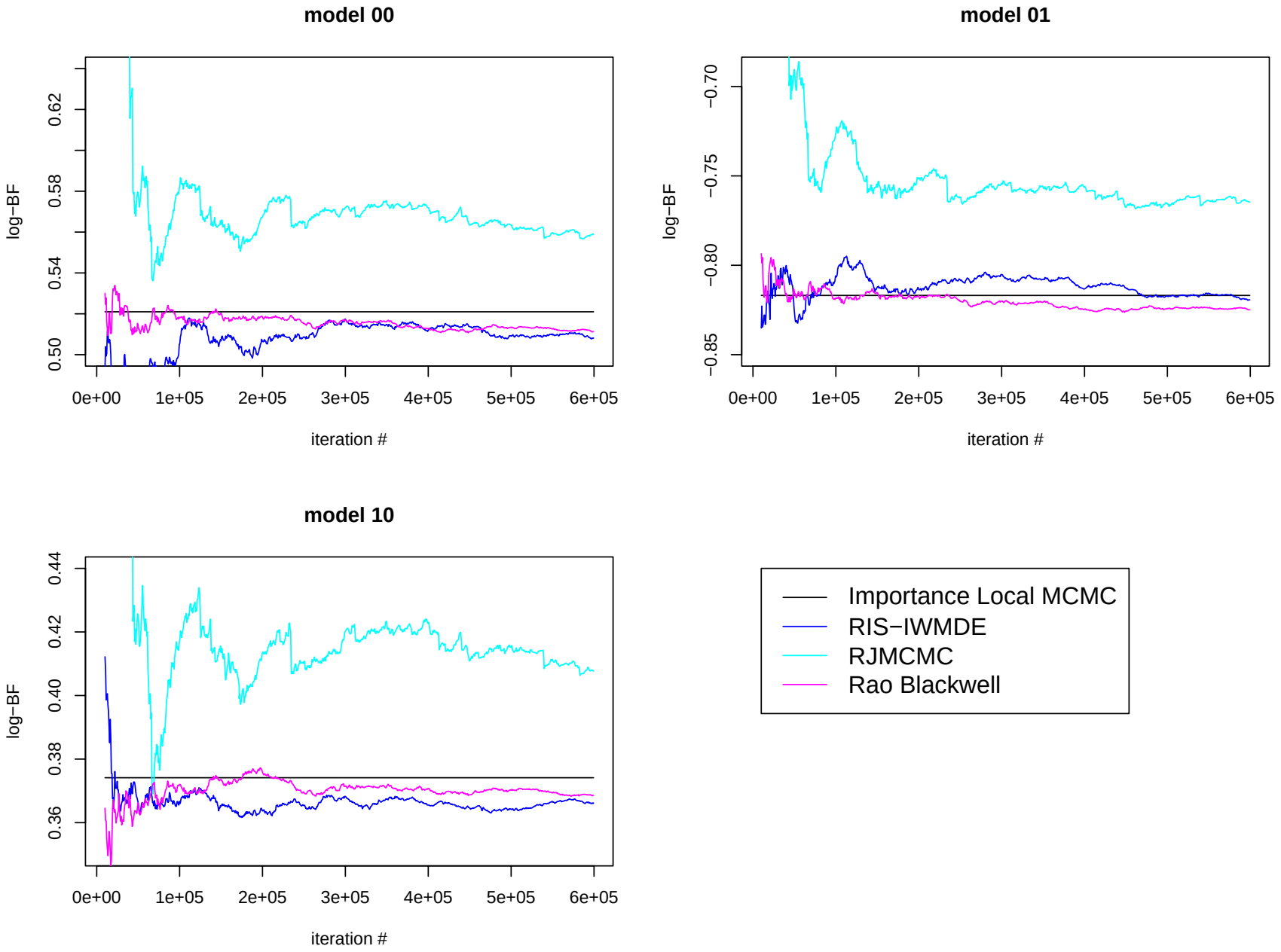
$$\begin{aligned} \pi(\gamma \mid \boldsymbol{y}) &= \mathsf{E}[\pi(\gamma \mid \boldsymbol{y}, \boldsymbol{z})] \\ &\approx \frac{1}{M} \sum_{j=1}^{M} \pi(\gamma \mid \boldsymbol{y}, \boldsymbol{z}^{(j)}) \end{aligned}$$


model 00
log−BF
0.50
0.54
0.58
0.62
20000
40000
60000
80000
100000
120000
140000
iteration #


model 01
log−BF
−0.82
−0.80
−0.78
−0.76
20000
40000
60000
80000
100000
120000
140000
iteration #


model 10
log−BF
0.36
0.38
0.40
0.42
0.44
20000
40000
60000
80000
100000
120000
140000
iteration #


Importance Local MCMC
Chib
Chib and Jeliazkov
RIS−IWMDE


model 00
log−BF
0.50
0.54
0.58
0.62
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
iteration #
model 01
log−BF
−0.85
−0.80
−0.75
−0.70
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
iteration #
model 10
log−BF
0.36
0.38
0.40
0.42
0.44
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
iteration #
Importance Local MCMC
RIS−IWMDE
RJMCMC
Rao Blackwell

Conclusions/Recommendations

- Whenever possible, use more than one method;
- When using methods based on the output of MCMC, be prepared to run your chain for much longer than needed for parameter estimation;
- The RIS-IWMDE estimates the marginal posterior under the full model evaluated at zero; that can result in considerable instability when that point has little density under the full;
- Chib's method and variants perform reasonably well, but if the sampling mechanism is elaborate they will not be easy to implement;

Conclusions/Recommendations

- Reversible jump followed by Monte Carlo frequencies can give rise to poor estimates of the posterior model probabilities;
- This was alleviated in a particular example using Rao-Blackwellization, but in general?
- Importance sampling performs amazingly well, and should not be dismissed as a viable alternative. We are working on examples where, although the parameter vector is high dimensional, we can find a very good importance function for part of that vector, being left with a relatively low-dimensional vector that is easily dealt with.

III. Adaptive Importance Sampling and Exoplanets

Long literature on adaptive importance sampling: OH and BERGER (1993), CAPPE, GUILLIN, MARIN, and ROBERT (2004), ARDIA, HOOGERHEIDE, and VAN DIJK (2008), CAPP´E, DOUC, GUILLIN, MARIN, and ROBERT (2008), CORNEBISE, MOULINES, and OLSSON (2008).

The Adaptive Importance Sampler

- It has an overall annealing layer, wherein one, as temperature $t \to 0$, is attempting to target $[f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})]^{1-t}$. (This is to try to find the modes of the integrand.)
- The importance function $q_t(\boldsymbol{\theta})$ tries to mimic the target with a mixture of T_4 densities, $q_t(\boldsymbol{\theta}) = \sum_{j=1}^{k} w_j T_4(\boldsymbol{\theta} \mid \boldsymbol{\mu}_j, \Sigma_j)$.
- Samples $\boldsymbol{\theta}^{(i)}$ are drawn from $q_t(\boldsymbol{\theta})$, and examined for high ratios of

$$\frac{[f(\boldsymbol{x} \mid \boldsymbol{\theta}^{(i)})\pi(\boldsymbol{\theta}^{(i)})]^{1-t}}{q_t(\boldsymbol{\theta}^{(i)})};$$

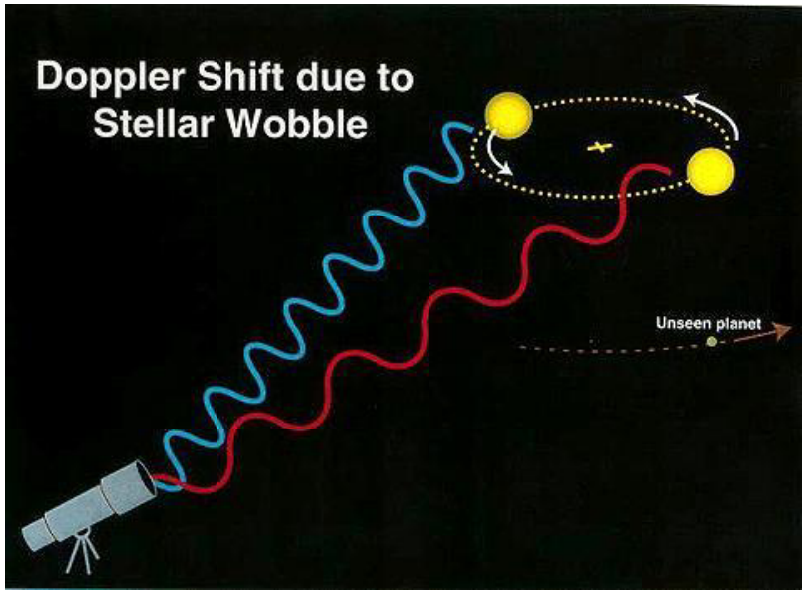
  new components of the mixture may be added at those points.
- If a weight of a component in the mixture becomes too small, the component is dropped.
- The weights of the mixture and the mean and covariance matrices are chosen by an analytic fit to the annealing target, using K-L divergence estimated from the previous draws from $q_t(\boldsymbol{\theta})$

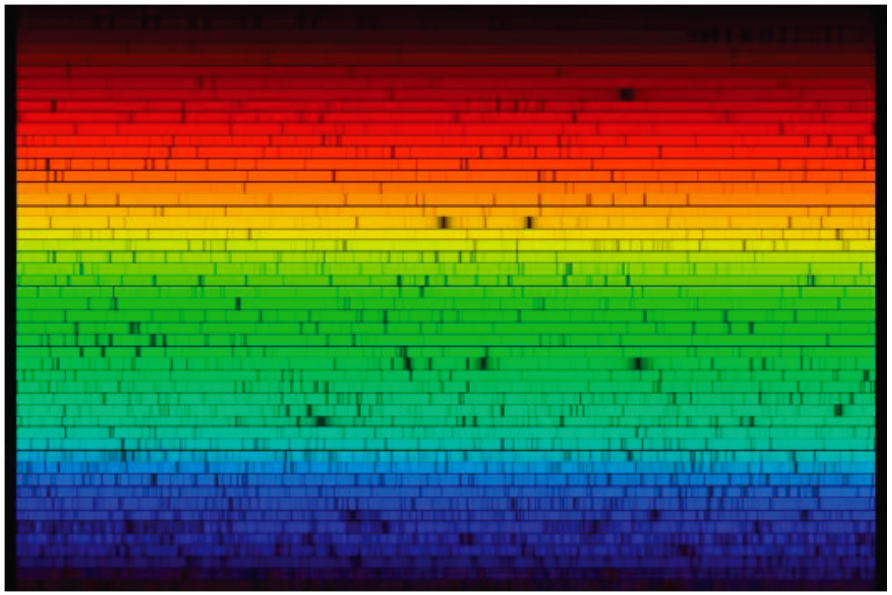
An example: *Exoplanet Detection*

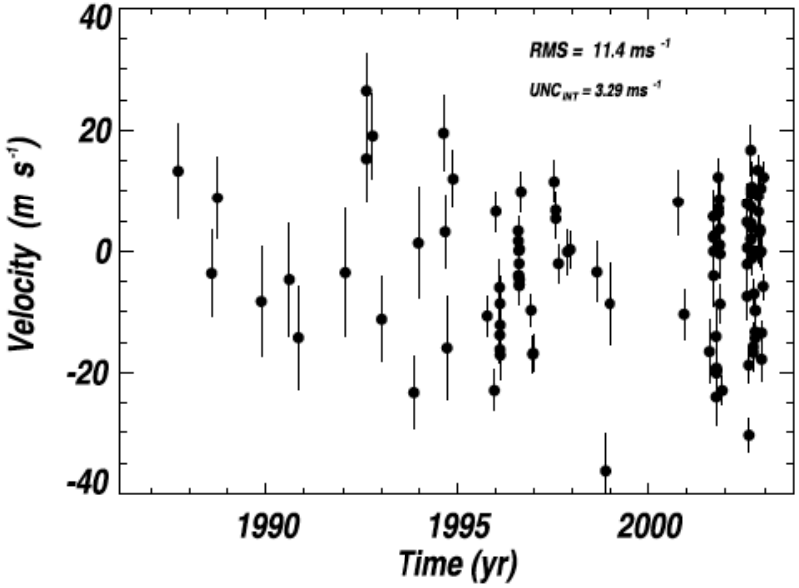
With Tom Loredo and David Chernoff (Cornell Astronomy)

Bin Liu and Merlise Clyde (Duke Statistics)

Common Detection Method: Use of Radial Velocity







Keplerian Radial Velocity Model

Parameters for single planet

- τ = orbital period (days)
- e = orbital eccentricity
- K = velocity amplitude (m/s)
- Argument of pericenter ω
- Mean anomaly at $t = 0$, M_0
- System center-of-mass velocity v_0
- Stellar jitter σ_J^2

Keplerian reflex velocity vs. time

$$v(t) = v_0 + K\left(e\cos\omega + \cos[\omega + \upsilon(t)]\right)$$

True anomaly $\upsilon(t)$ found via Kepler's equation for eccentric anomaly:

$$E(t) - e\sin E(t) = \frac{2\pi t}{\tau} - M_0; \qquad \tan\frac{\upsilon}{2} = \left(\frac{1+e}{1-e}\right)^{1/2}\tan\frac{E}{2}\,.$$

The Likelihood Function

Keplerian velocity model with parameters $\theta = \{K, \tau, e, M_0, \omega\}$:

$$d_i = v(t_i; \theta) + \varepsilon_i$$

For measurement errors with standard deviation σ_i,

$$\begin{aligned} L(\theta, v_0, \sigma_J^2) &\equiv p(\{d_i\}|\theta, v_0, \sigma_J^2) \\ &= \prod_{i=1}^{N} \frac{1}{2\pi\sqrt{\sigma_i^2 + \sigma_J^2}} \exp\left[-\frac{1}{2}\frac{[d_i - v(t_i;\theta)]^2}{\sigma_i^2 + \sigma_J^2}\right] \\ &\propto \left[\prod_i \frac{1}{2\pi\sqrt{\sigma_i^2 + \sigma_J^2}}\right] \exp\left[-\frac{1}{2}\chi^2(\theta)\right] . \end{aligned}$$

$$\text{where} \quad \chi^2(\theta, \sigma_J^2) \equiv \sum_i \frac{[d_i - v(t_i;\theta)]^2}{\sigma_i^2 + \sigma_J^2}$$

This likelihood has extreme multimodality in τ; challenging multimodality in M_0; and is smooth (but often vague) in e.

Bayesian Approach to Exoplanet Detection

- Let M_i denote the model that there are i planets ($2 + 5i$ parameters).
- Determine prior distributions $\pi(\theta_i, v_0, \sigma_J^2)$ for the parameters (semi-standard, as the result of a SAMSI program, except for σ_J^2).
- Compute the marginal likelihood of model M_i,

$$\int L_i(\theta_i, v_0, \sigma_J^2)\pi(\theta_i, v_0, \sigma_J^2)\ d\theta_i\ dv_0\ d\sigma_J^2\ .$$

  We have been working on an adaptive importance sampling algorithm for carrying out the computation.
- Typically look at Bayes factors (the ratio of marginal likelihoods) to determine the number of planets.

Example Exoplanet Results

I. HD73526, 18 observations

	Marginal Likelihood	ESS/N
$\mathcal{M}_0$	$5.9013 \times 10^{-50} \pm 5.1325 \times 10^{-52}$	0.9320
$\mathcal{M}_1$	$4.4886 \times 10^{-41} \pm 3.2093 \times 10^{-42}$	0.5698
$\mathcal{M}_2$	$1.5511 \times 10^{-42} \pm 3.2878 \times 10^{-43}$	0.3458

BayesFactor$\{\mathcal{M}_1 : \mathcal{M}_0\}$	BayesFactor$\{\mathcal{M}_2 : \mathcal{M}_1\}$
7.606×10^8	0.03456

II. HD73526, 30 observations

BayesFactor$\{\mathcal{M}_1 : \mathcal{M}_0\}$	BayesFactor$\{\mathcal{M}_2 : \mathcal{M}_1\}$
6.534×10^6	8.233×10^4

III. 47 Ursae Majoris

	Marginal Likelihood	ESS/N
$\mathcal{M}_0$	$2.0198 \times 10^{-1004} \pm 9.2572 \times 10^{-1006}$	0.1002
$\mathcal{M}_1$	$3.4400 \times 10^{-896} \pm 3.10 \times 10^{-897}$	0.5643
$\mathcal{M}_2$	$1.3500 \times 10^{-816} \pm 1.77 \times 10^{-817}$	0.3324
$\mathcal{M}_3$	$2.8970 \times 10^{-825} \pm 9.1623 \times 10^{-825}$	0.0089

BayesFactor$\{\mathcal{M}_1 : \mathcal{M}_0\}$	BayesFactor$\{\mathcal{M}_2 : \mathcal{M}_1\}$	BayesFactor$\{\mathcal{M}_3 : \mathcal{M}_2\}$
1.703×10^{108}	3.924×10^{79}	?

IV. Search in Large Model Spaces and the Inference Challenge

Bayesian Analysis in Large Model Spaces: a Search Strategy and Conceptual Issues

Notation: Data $\mathbf{x}$, arising from model M having density $f(\mathbf{x} \mid \boldsymbol{\theta})$, with unknown parameter $\boldsymbol{\theta}$ that has prior density $\pi(\boldsymbol{\theta})$.

Search Strategy: Start at any model. Upon visiting model $\mathcal{M}_l$, having unknown parameter $\boldsymbol{\theta}_l$ consisting of some subset of the variables $\theta_1, \ldots, \theta_p$, determine its marginal likelihood $m_l(\mathbf{x}) = \int f_l(\mathbf{x} \mid \boldsymbol{\theta}_l)\pi(\boldsymbol{\theta}_l)d\boldsymbol{\theta}_l$.

At any stage, with $\{\mathcal{M}_1, \ldots, \mathcal{M}_k\}$ denoting the models previously visited,

- define the current estimated posterior probabilities of models (assuming equal prior probabilities - bad choice) as

$$\widehat{P}(\mathcal{M}_l \mid \mathbf{x}) = \frac{m_l(\mathbf{x})}{\sum_{j=1}^{k} m_j(\mathbf{x})} ;$$

- define the current estimated variable inclusion probabilities (estimated overall posterior probabilities that variables are in the model) as $\hat{q}_i = \hat{P}(\theta_i \in \text{model} \mid \mathbf{x}) = \sum_{j=1}^{k} \widehat{P}(\mathcal{M}_j \mid \mathbf{x}) 1_{\{\theta_i \in \mathcal{M}_j\}}$.

1. At iteration k, compute the current posterior model and inclusion probability estimates, $\widehat{P}(\mathcal{M}_j \mid \mathbf{x})$ and $\hat{q}_i$.
2. Return to one of the $k-1$ distinct models already visited, in proportion to their estimated probabilities $\widehat{P}(\mathcal{M}_j \mid \mathbf{x})$.
3. Add/remove a variable with probability 0.5.
   - If adding a variable, choose i with probability $\propto \frac{\hat{q}_i+\mathbb{C}}{1-\hat{q}_i+\mathbb{C}}$
   - If removing, choose i with probability $\propto \frac{1-\hat{q}_i+\mathbb{C}}{\hat{q}_i+\mathbb{C}}$

   $\mathbb{C}$ is a tuning constant introduced to keep the $\hat{q}_i$ away from zero or one; $\mathbb{C} = 0.01$ is a reasonable default value.
4. If the model obtained in Step 3
   - has already been visited, return to step 2.
   - is new, update $\hat{P}(\mathcal{M}_j \mid \mathbf{x})$ and $\hat{q}_i$ and go to step 2.

Conventional Priors for Normal Linear Models

The full model for the data $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is (with σ^2 unknown)

$$\mathcal{M}_F : \mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})\,,$$

- $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}^*)$ is a $p \times 1$ vector of unknown coefficients, with β_1 being the intercept;
- $\mathbf{X} = (\mathbf{1}\ \mathbf{X}^*)$ is the $n \times p$ $(p < n)$ full rank design matrix of covariates, with the columns of $\mathbf{X}^*$ being orthogonal to $\mathbf{1} = (1, \ldots, 1)'$.

We consider selection from among the submodels $\mathcal{M}_l,\ l = 1, \ldots, 2^p$,

$$\mathcal{M}_i : \mathbf{Y} = \mathbf{X}_i\,\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I}),$$

- $\boldsymbol{\beta}_i = (\beta_1, \boldsymbol{\beta}_i^*)$ is a $d(i)$-dimensional subvector of $\boldsymbol{\beta}$ (note that we only consider submodels that contain the intercept);
- $\mathbf{X}_i = (\mathbf{1}\ \mathbf{X}_i^*)$ is the corresponding $n \times d(i)$ matrix of covariates.
- Let $f_i(\mathbf{y} \mid \boldsymbol{\beta}_i, \sigma^2)$ denote this density.

Prior density under $\mathcal{M}_i$: standard choices include

- g-Priors: $\pi_i^g(\beta_1, \boldsymbol{\beta}_i^*, \sigma^2 \mid c) = \frac{1}{\sigma^2} \times \mathrm{N}_{(d(i)-1)}\left(\boldsymbol{\beta}_i^* \mid \mathbf{0}, cn\sigma^2(\mathbf{X}_i^{*\prime}\mathbf{X}_i^*)^{-1}\right)$, where c is fixed, and is typically set equal to 1 or estimated in an empirical Bayesian fashion; *these are not information consistent.*

- Zellner-Siow priors (ZSN): the intercept model has the prior $\pi(\beta_1, \sigma^2) = 1/\sigma^2$, while the the prior for other $\mathcal{M}_i$ is

$$\begin{aligned} \pi_i^{ZSN}(\beta_1, \boldsymbol{\beta}_i^*, \sigma^2) &= \pi_i^g(\beta_1, \boldsymbol{\beta}_i^*, \sigma^2 \mid c)\,, \\ \pi_i^{ZSN}(c) &\sim \mathrm{InverseGamma}(c \mid 0.5, 0.5)\,. \end{aligned}$$

Marginal density under $\mathcal{M}_i$: $m_i(\mathbf{Y}) = \int f_i(\mathbf{y} \mid \boldsymbol{\beta}_i, \sigma^2)\pi_i(\boldsymbol{\beta}_i, \sigma^2)\, d\boldsymbol{\beta}_i d\sigma^2$

Posterior probability of $\mathcal{M}_i$, when the prior probability is $P(\mathcal{M}_i)$ (for multiplicity adjustment, choose $P(\mathcal{M}_i) = Beta(d(i), p - d(i) + 1)$), is

$$P(\mathcal{M}_i \mid \mathbf{Y}) = \frac{P(\mathcal{M}_i)m_i(\mathbf{Y})}{\sum_k P(\mathcal{M}_k)m_k(\mathbf{Y})}\,.$$

Example: An Ozone data set was considered in Breiman and Friedman (1985).

- It consists of 178 observations, each having 10 covariates (in addition to the intercept).
- We consider a linear model with all linear main effects along with all quadratic terms and second order interactions, yielding a total of 65 covariates and $2^{65} \approx 3.6 \times 10^{19}$ models.
- g-prior and Zellner-Siow priors were considered; these result in easily computable marginal model probabilities $m_i(\mathbf{y})$.
- The algorithm was implemented at different starting points with essentially no difference in results, and run through 5,000,000 iterations, saving only the top 65,536 models.
- No model had appreciable posterior probability; 0.0025 was the largest.

variables	g-prior	ZSN
x1	.860	.943
x2	.052	.060
x3	.030	.033
x4	.985	.995
x5	.195	.306
x6	.186	.353
x7	.200	.215
x8	.960	.977
x9	.029	.054
x10	.999	.999
x1.x1	.999	.999
x9.x9	.999	.999
x1.x2	.577	.732
x1.x7	.076	.142
x3.x7	.021	.022
x4.x7	.330	.459
x6.x8	.776	.859
x7.x8	.266	.296
x7.x10	.975	.952

Table 1: Posterior inclusion probabilities of variables in the ozone problem, after 5,000,000 iterations of the algorithm

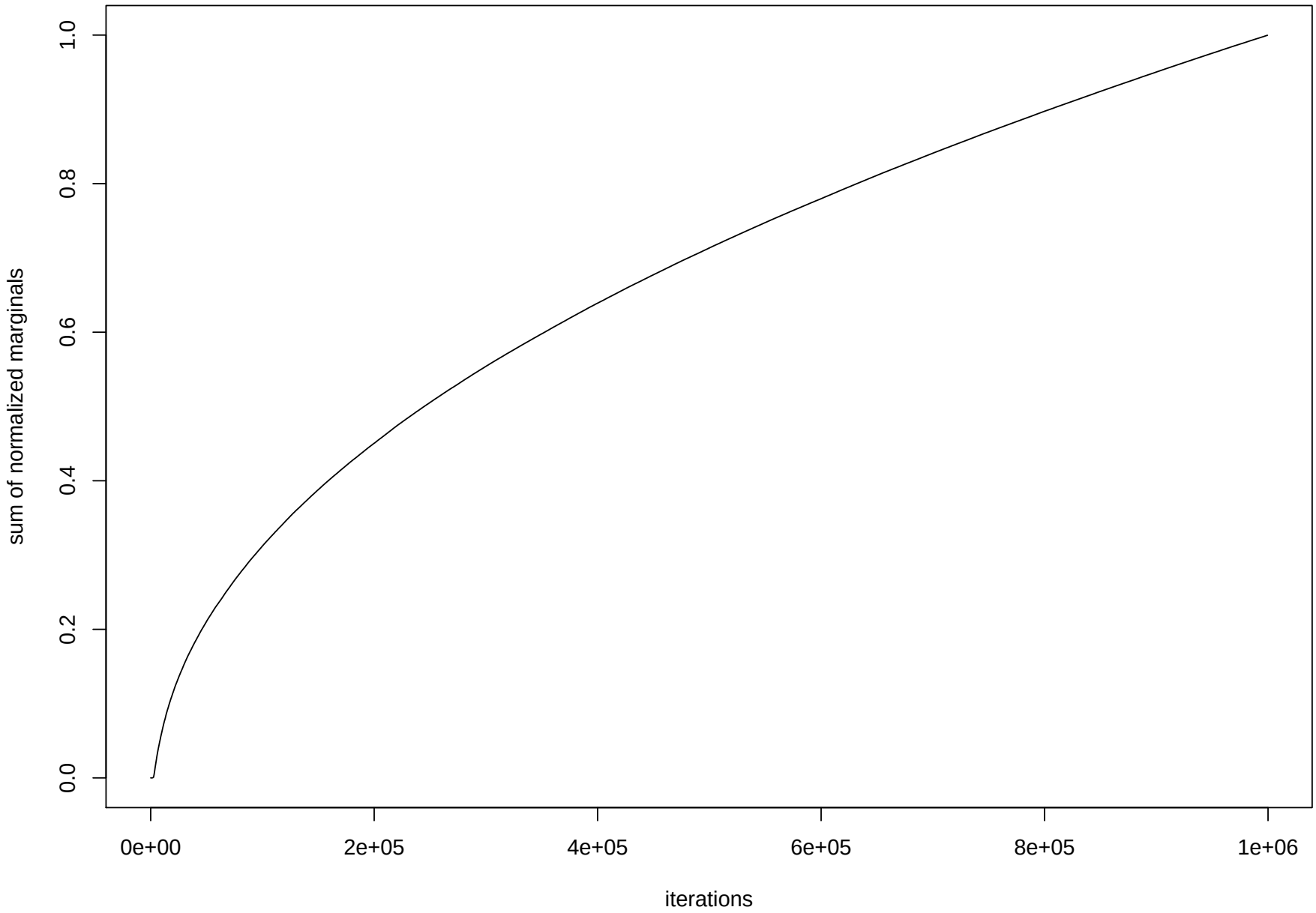


Figure 1: Retrospective cumulative posterior probability of models visited in the ozone problem.

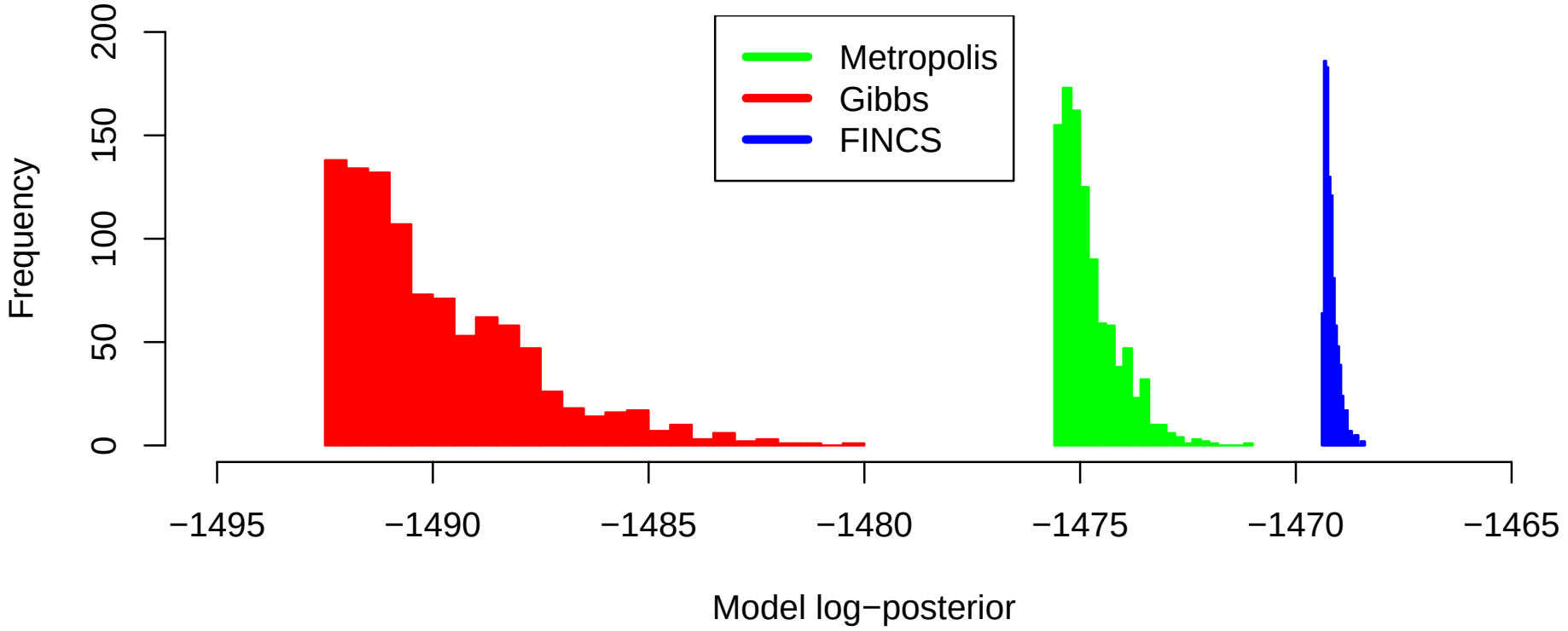


Figure 2: Carvalho and Scott (2008): graphical model selection, 25 node example. Models found by Gibbs (SVSS), Metropolis, and Feature INClusion Search. Equal computation time for searches; probability renormalization across all three model sets.

Example 2. Posterior Model Probabilities for Variable Selection in Probit Regression

(from CMU Case Studies VII, Viele et. al. case study)

Motivating example: Prosopagnosia (face blindness), is a condition (usually developing after brain trauma) under which the individual cannot easily distinguish between faces. A psychological study was conducted by Tarr, Behrmann and Gauthier to address whether this extended to a difficulty of distinguishing other objects, or was particular to faces.

The study considered 30 control subjects (C) and 2 subjects (S) diagnosed as having prosopagnosia, and had them try to differentiate between similar faces, similar 'Greebles' and similar 'objects' at varying levels of difficulty and varying comparison time.

Data:

$$\left.\begin{array}{l} C = Subject\ C \\ S = Subject\ S \\ G = Greebles \\ O = Object \\ D = Difficulty \\ B = Brief\ time \\ A = images\ match\ or\ not) \end{array}\right\} \Rightarrow \quad R = Response\ (answer\ correct\ or\ not).$$

All variables are binary. Sample size was $n = 20,083$. {C=S=1} and {G=O=1} are not possible combinations, so there are $3 \times 3 \times 2 \times 2 \times 2 = 72$ possible covariates.

Statistical modeling: For a specified covariate vector $\boldsymbol{X}_i$, let y_i and $n_i - y_i$ be the numbers of successes and failures among the responses with that covariate vector, with probability of success p_i assumed to follow the probit regression model

$$p_i = \Phi(\beta_1 + \textstyle\sum_{j=2}^{72} \boldsymbol{X}_{ij}\beta_j).$$

The full model likelihood (up to a fixed proportionality constant) is then

$$f(\boldsymbol{y} \mid \boldsymbol{\beta}) = \textstyle\prod_{i=1}^{72} p_i^{y_i}(1-p_i)^{n_i - y_i}.$$

Goal: Select from among the 2^{72} submodels which have some of the β_j set equal to zero. (Actually, only models with graphical structure were considered, i.e., if an interaction term is in the model, all the lower order effects must also be there.)

Prior Choice: conditionally induce submodel priors from a full-model prior

If all models are nested in a 'Full Model' M_l, having parameter $\boldsymbol{\beta}$, one possibility is to

- choose a prior $\pi_l(\boldsymbol{\beta})$;
- Define a prior on the submodel M_i by $\pi_i(\boldsymbol{\beta}_i) = \pi_l(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_{-i} = 0)$, where $\boldsymbol{\beta}_{-i}$ consists of the other coordinates of $\boldsymbol{\beta}$.

Example. Suppose $\pi_l(\boldsymbol{\beta})$ is $\mathcal{N}_{k_l}(0, V^{-1})$, where V is a given precision matrix. Then $\pi_i(\boldsymbol{\beta}_i)$ is $\mathcal{N}_{k_i}(0, V_i^{-1})$, where V_i is the corresponding submatrix of V.

Application to the Case Study

The full model likelihood is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{72} p_i^{y_i}(1-p_i)^{n_i-y_i},$$

where y_i and $n_i - y_i$ are the numbers of hits and failures for the specified vector of covariates $\boldsymbol{X}_i$, and

$$p_i = \Phi(\beta_1 + \sum_{j=2}^{72} \boldsymbol{X}_{ij}\beta_j).$$

A standard noninformative prior for $\boldsymbol{p} = (p_1, p_2, \ldots, p_{72})$ is the uniform prior, usable here since it is proper.
Change of variables yields $\pi_l(\boldsymbol{\beta})$ is $\mathcal{N}_{72}(0, (\boldsymbol{X}'\boldsymbol{X})^{-1})$, where
$\boldsymbol{X}' = (\boldsymbol{X}'_1, \boldsymbol{X}'_2, \ldots, \boldsymbol{X}'_{72})$.

Two Modifications of the Priors

I. Induce priors from $\pi_l(\boldsymbol{\beta}) = \mathcal{N}_{72}(0, c(\boldsymbol{X}'\boldsymbol{X})^{-1})$.

- g-priors using $c = 72$? (why not $c = n = 20,083$??)
- Also tried $c = 18$.

II. For M_j, let $\boldsymbol{H}_j$ be the matrix such that $\boldsymbol{\beta}_j = \boldsymbol{H}_j\boldsymbol{\beta}$. Then use $\pi_j(\boldsymbol{\beta}_j) = \mathcal{N}_{k_j}(0, ((\boldsymbol{H}_j\boldsymbol{X}\boldsymbol{H}_j')'\boldsymbol{H}_j\boldsymbol{X}\boldsymbol{H}_j')^{-1})$.

- Arises from applying the "uniform $\boldsymbol{p}$" argument separately to each model; equivalent to marginalizing out in the prior for the full model.

Computation of the marginal probability of a visited model: Use the modified Laplace approximation

$$\begin{aligned} m_j(\boldsymbol{y}) &= \int f_j(\boldsymbol{y} \mid \boldsymbol{\beta}_j)\pi_j(\boldsymbol{\beta}_j)\, d\boldsymbol{\beta}_j \\ &\approx f_j(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}_j)|\boldsymbol{I} + \boldsymbol{V}_j^{-1}\hat{\boldsymbol{I}}_j|^{-1/2} e^{-\frac{1}{2}\hat{\boldsymbol{\beta}}_j'\left(\hat{\boldsymbol{I}}_j^{-1}+\boldsymbol{V}_j^{-1}\right)^{-1}\hat{\boldsymbol{\beta}}_j}. \end{aligned}$$

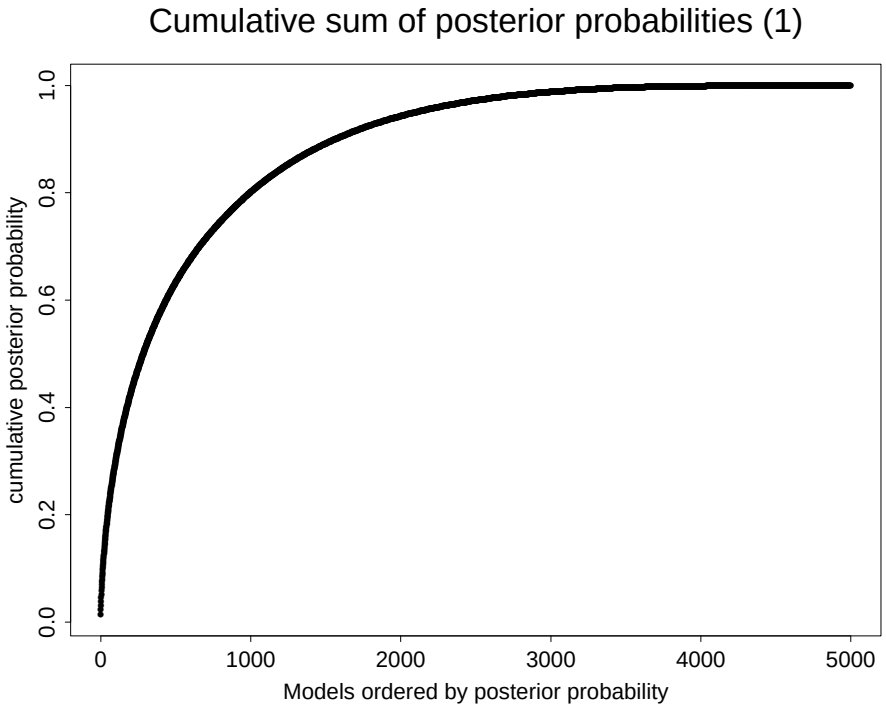
where $\hat{\boldsymbol{I}}_j$ is the observed information matrix.

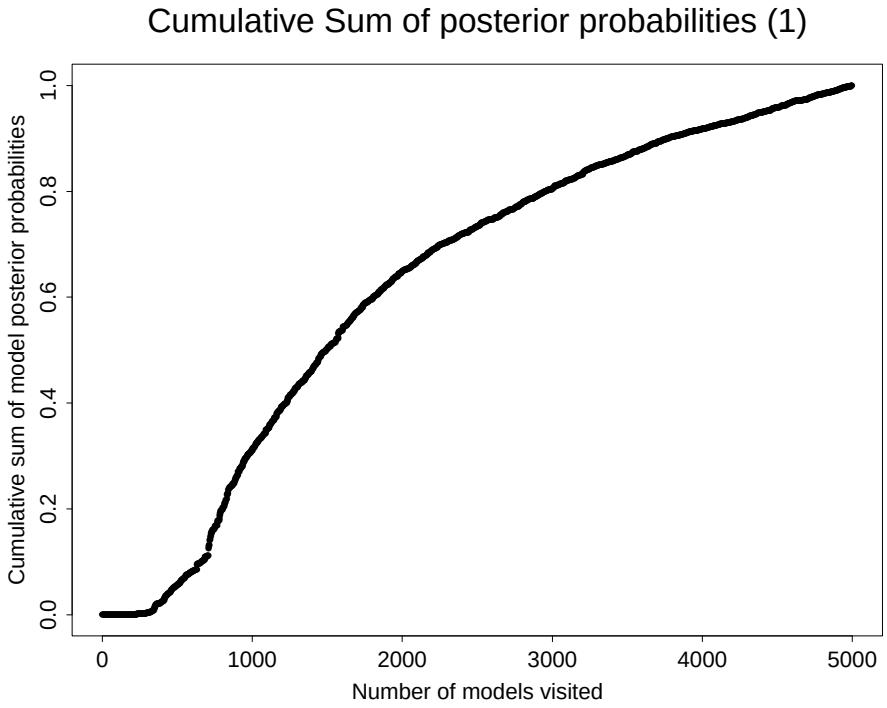
- This allows use of standard probit packages to obtain $\hat{\boldsymbol{\beta}}_j$, $f_j(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}_j)$, and $\hat{\boldsymbol{I}}_j$.
- Note that (approximate) posterior means, $\boldsymbol{\mu}_j$, and covariance matrices, $\boldsymbol{\Sigma}_j$, are also available in closed form.

Similarities Among Selected Models

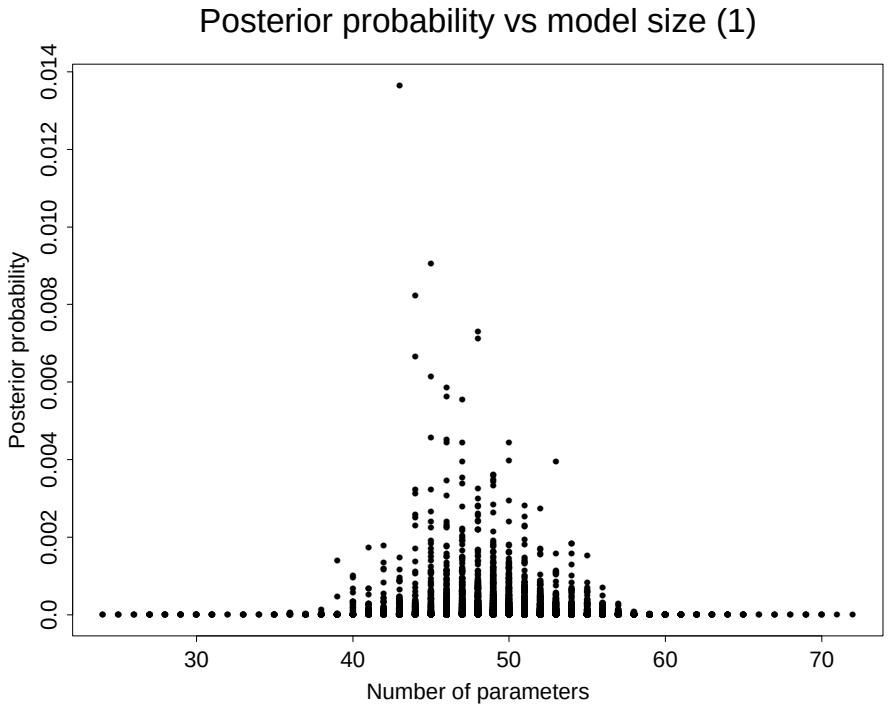
- The search found much better models than the search strategy (a combination of backward elimination and MCMC) used in the original paper. Indeed the top two models in the paper for their choice of $n = 50$ for BIC ranked #7 and #1, respectively, under Prior 1 (the uniform induced conditional prior). No other models in the paper ranked in the top ten for any prior.
- The top model for Prior 4 (the multiple uniform induced prior) ranked #5 under Prior 1.
- The models for Prior 2 ($c = 72$) and Prior 3 ($c = 18$) were considerably smaller (# variables in the mid-30's) than those for Prior 1 and Prior 4 (# variables in the mid-40's).

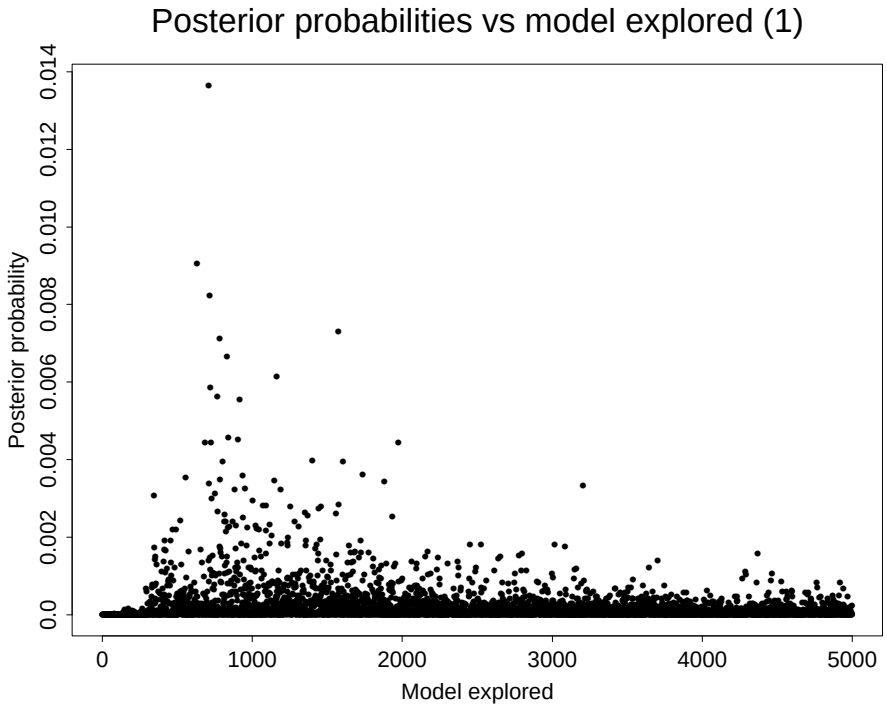
Uniform-induced conditional prior



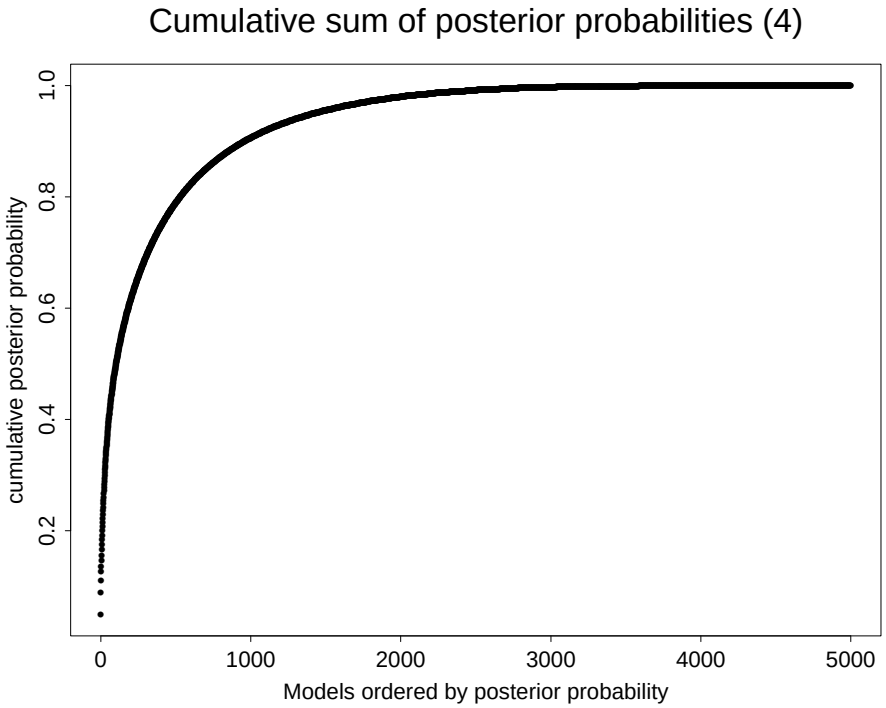


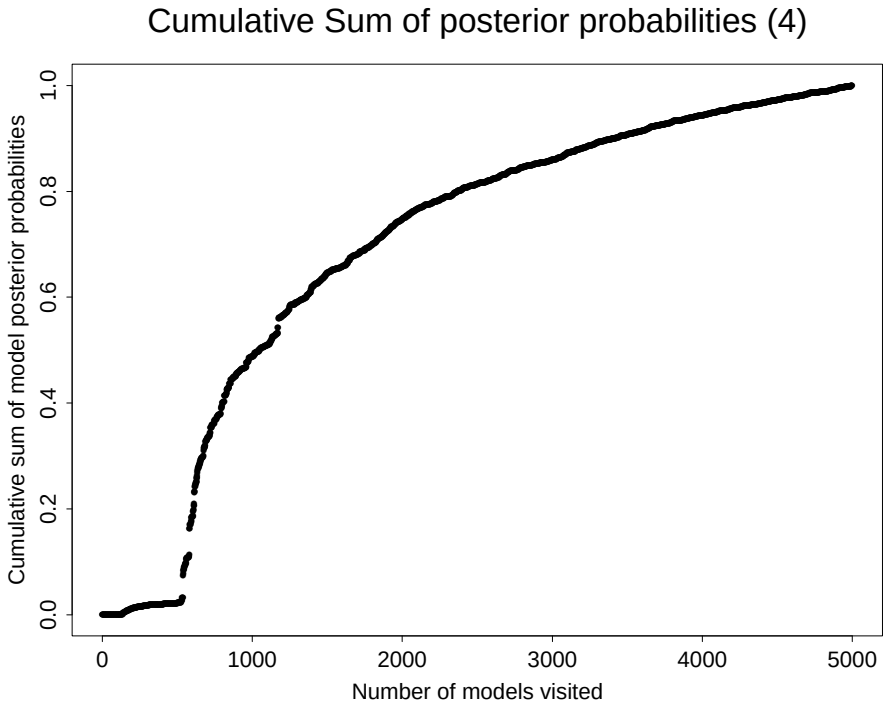
Uniform-induced conditional prior



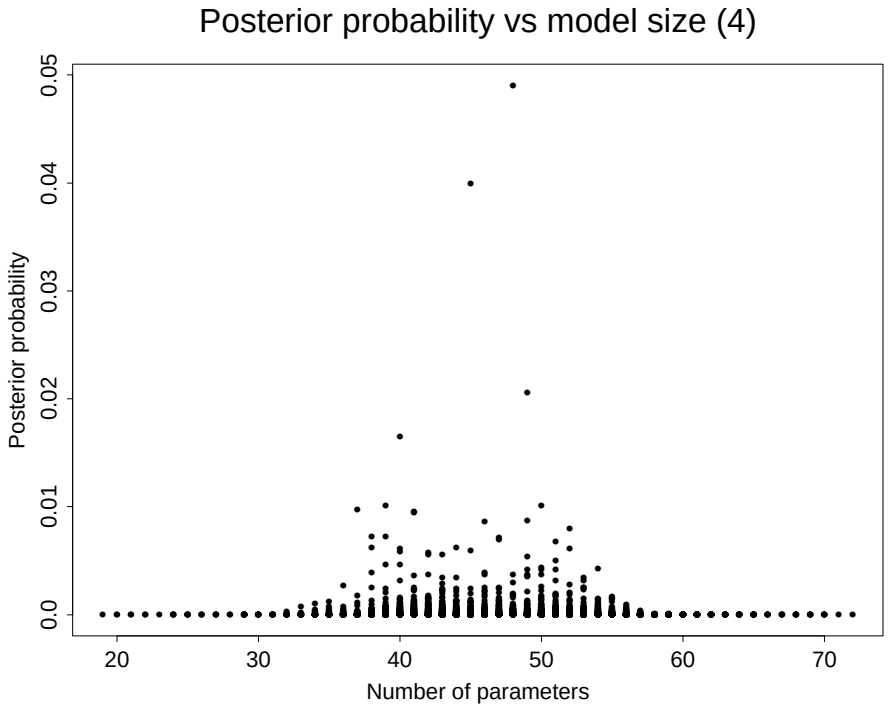


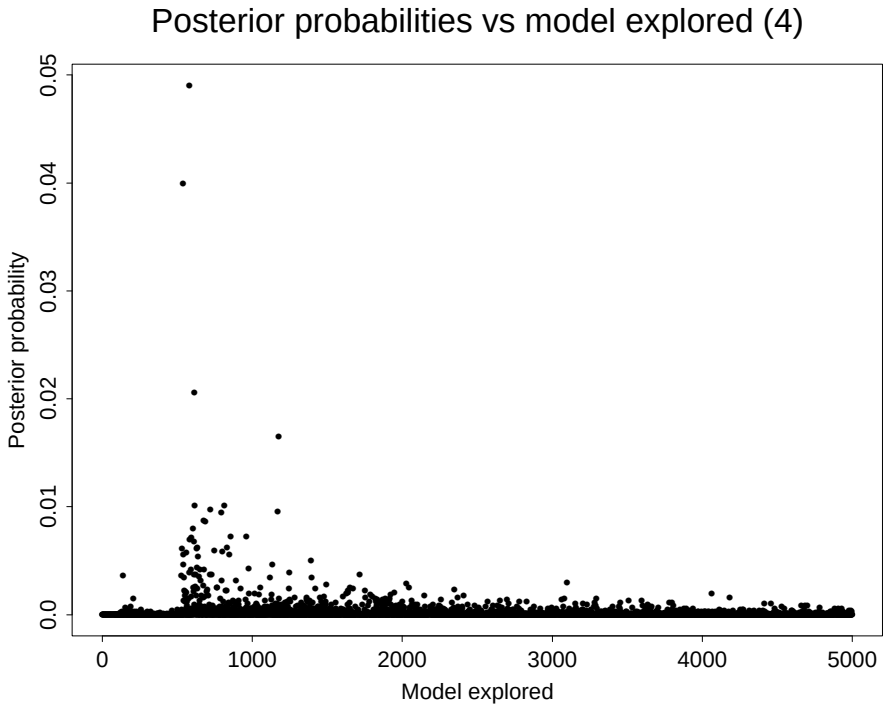
Multiple-uniform prior





Multiple-uniform prior





Utilization of the Posterior Model Probabilities

One usually must decide between

- Model Averaging: perform inferences, averaged over all models.
- Model Selection: choose a specific model, e.g.
  - the maximum posterior probability model;
  - the median probability model (that which includes only those β_i for which $P(\beta_i \in \text{model} \mid \boldsymbol{y}) \geq 0.5$).

Posterior inclusion probabilities for the 72 parameters

Var.	prior1	prior2	prior3	prior4
Int.	1	1	1	1
C	1	1	1	1
S	1	1	1	1
G	1	1	1	1
O	1	1	1	1
D	1	1	1	1
B	1	1	1	1
A	1	1	1	1
CG	0.966	0.082	0.154	0.955
CO	0.519	0.033	0.063	0.155
CD	0.780	0.035	0.109	0.454
CB	1	0.999	0.999	1
CA	1	1	1	1
SG	1	1	1	1
SO	0.999	0.877	0.969	0.995
SD	1	0.999	0.999	1
SB	1	1	1	1
SA	1	1	1	1

Var.	prior1	prior2	prior3	prior4
GD	0.998	0.358	0.519	0.833
GB	1	1	1	1
GA	1	1	1	1
OD	0.999	0.418	0.700	0.744
OB	1	1	1	1
OA	1	1	1	1
DB	1	0.999	0.999	0.999
DA	1	1	1	1
BA	1	1	1	1
CGD	0.264	0.000	0.000	0.134
CGB	0.938	0.002	0.020	0.943
CGA	0.936	0.002	0.018	0.944
COD	0.074	0.000	0.000	0.003
COB	0.417	0.001	0.003	0.080
COA	0.275	0.003	0.013	0.094
CDB	0.684	0.012	0.063	0.357
CDA	0.323	0.001	0.011	0.235
CBA	0.999	0.639	0.833	0.999

Posterior inclusion probabilities for the 72 parameters

Var.	P1	P2	P3	P4
SGD	0.997	0.079	0.335	0.661
SGB	1	0.999	0.999	1
SGA	1	1	1	1
SOD	0.990	0.165	0.434	0.640
SOB	0.091	0.016	0.034	0.116
SOA	0.999	0.843	0.956	0.993
SDB	0.999	0.821	0.957	0.998
SDA	0.999	0.998	0.999	0.999
SBA	1	0.999	0.999	1
GDB	0.996	0.096	0.349	0.673
GDA	0.442	0.012	0.025	0.616
GBA	1	1	1	1
ODB	0.128	0.004	0.016	0.037
ODA	0.999	0.407	0.689	0.737
OBA	0.329	0.065	0.139	0.045
DBA	0.999	0.770	0.920	0.997
CGDB	0.111	0.000	0.000	0.073
CGDA	0.004	0.000	0.000	0.018

Var.	P1	P2	P3	P4
CGBA	0.932	0.000	0.013	0.941
CODB	0.002	0.000	0.000	0.000
CODA	0.006	0.000	0.000	0.001
COBA	0.013	0.000	0.000	0.001
CDBA	0.088	0.000	0.001	0.104
SGDB	0.996	0.072	0.324	0.652
SGDA	0.379	0.000	0.006	0.579
SGBA	1	0.999	0.999	1
SODB	0.004	0.000	0.000	0.004
SODA	0.176	0.013	0.055	0.250
SOBA	0.004	0.000	0.000	0.002
SDBA	0.271	0.031	0.065	0.654
GDBA	0.206	0.000	0.001	0.515
ODBA	0.005	0.000	0.000	0.000
CGDBA	0.000	0.000	0.000	0.003
CODBA	0.000	0.000	0.000	0.000
SGDBA	0.144	0.000	0.000	0.473
SODBA	0.000	0.000	0.000	0.000

Here model averaging seems best (individual models do not seem to be of particular interest). Thus one estimates contrasts $d = \mathbf{x}^* \boldsymbol{\beta}$ by

$$\hat{d} = \mathbf{x}^* \bar{\boldsymbol{\beta}} \equiv \mathbf{x}^* \sum_j P(M_j \mid \boldsymbol{y}) \boldsymbol{H}'_j \boldsymbol{\mu}_j,$$

having estimated variance

$$\sigma^2 = \sum_j P(M_j \mid \boldsymbol{y}) \mathbf{x}^* \boldsymbol{H}'_j (\boldsymbol{\Sigma}_j + \boldsymbol{\mu}_j \boldsymbol{\mu}'_j) \boldsymbol{H}_j \mathbf{x}^{*\prime} - (\mathbf{x}^* \bar{\boldsymbol{\beta}})^2.$$

Contrast	Prior 1	Prior 2	Prior 3	Prior 4
SM/GR/FA	.683(0.240)	.780(0.239)	.829(0.240)	.306(0.282)
SM/OB/FA	.619(0.161)	.606(0.161)	.596(0.161)	.640(0.162)
SM/OB/GR	-.063(0.242)	-.173(0.244)	-.232(0.244)	.334(0.286)
CR/GR/FA	-.761(0.279)	.000(0.000)	.000(0.000)	-.872(0.282)
CR/OB/FA	.000(0.000)	.000(0.000)	.000(0.000)	.000(0.000)
CR/OB/GR	.761(0.279)	.000(0.000)	.000(0.000)	.872(0.279)

Median models (Posterior Probabilities) at iteration 5000:

Prior 1: [CO OB CDB SOD SOA SDA ODA DBA CGBA SGDB SGBA] (0.001)

Prior 2: [GD OB CBA SOA SDB SDA ODA DBA SGBA] (0.067)

Prior 3: [OB CBA SOA SDB SDA DBA SGBA] (0.016)

Prior 4: [OB SOD SOA ODA CGBA SGDB SGDA SGBA SDBA GDBA] (0.001)

Non-Bayesian search in model space: The same principle can be applied for any search criterion and for any model structure:

- Choose the model features (e.g. variables, graphical nodes, links in graph structures) you wish to drive the search.
- Choose the criterion for defining a good model (e.g. AIC)
- Convert the criterion to pseudo-probabilities for the models, e.g.

$$\widehat{P}(\mathcal{M}_j \mid \mathbf{x}) \propto e^{\mathrm{AIC}_j/2}\,.$$

- Define the feature inclusion probabilities as

$$\hat{q}_i = \hat{P}(\text{feature}_i \in \text{model} \mid \mathbf{x}) = \sum_{j=1}^{k} \widehat{P}(\mathcal{M}_j \mid \mathbf{x}) 1_{\{\text{feature}_i \in \mathcal{M}_j\}}\,.$$

- Apply the search algorithm.

Summary of Search

For (non-orthogonal) large model spaces,

- Effective search can be done by revisiting high probability models and changing variables (features) according to their estimated posterior inclusion probabilities. The technique is being used to great effect in many fields:
  - Statistics: Berger and Molina, 2004 CMU Case Studies; 2005 Statist. Neerland.
  - Economics: Sala-i-Martin, 2004 American Economic Review; different variant
  - CS: Schmidt, Niculescu-Mizil, and Murphy, 2007 AAAI; different variant
  - Graphical Models: Carvalho and Scott, 2007 tech report
- There may be no model with significant posterior probability (in the ozone example the largest model posterior probability was 0.0025) and thousands or millions of roughly comparable probability.
- Of most importance is thus overall features of the model space, such as the posterior inclusion probabilities of each variable, and possibly bivariate inclusion probabilities or other structural features (see, e.g., Ley and Steel, 2007 J. Macroeconomics).