STAT 206

Mar 21

DD AM (5 Mar 21 discussion section)

9 Mar: lecture

A Laplace Approximation

In the context of Bayes factors as an approach to Bayesian model comparison

Remember Bayes factors from week 1?

Suppose that we have uncovered an ensemble $\mathcal{M} = \{M_1, M_2, \ldots, M_m\}$ of worthwhile-to-consider models (e.g., using the cheating approach)

$\underline{\underline{P}} = (\underline{\underline{Q}}, \underline{\underline{L}}) \to (\underline{\theta}, D, B)$ ← unknown quantities of principal interest

$\to M_j$ ($j = 1, \ldots, m$):

$$\begin{cases} (\underline{\theta}_j \mid [PM_j] B) \sim p(\underline{\theta}_j \mid [PM_j] B) \\ (\mathcal{I}_i \mid \underline{\theta}_j [SM_j] B) \overset{IID}{\sim} p(y_i \mid \underline{\theta}_j [SM_j] B) \quad (i = 1, \ldots, n) \end{cases}$$

k_j = # free parameters in $\underline{\theta}_j$

$$D = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

ex. NB10 case study $\mathcal{M} = \{\text{Gaussian}, t\}$ ($m = 2$)

M_1:
$$\begin{cases} \underline{\theta}_1 = (\mu, \sigma) \\ (\underline{\theta}_1 \mid [PM_1] B) \sim p(\underline{\theta}_1 \mid [PM_1] B) \\ (\mathcal{I}_i \mid \underline{\theta}_1 [SM_1 : N] B) \overset{IID}{\sim} p(y_i \mid \underline{\theta}_1 [SM_1 : NB) \end{cases}$$

M_2: $\left\{ \begin{array}{l} (\underset{\sim}{\theta}_2 \mid [p M_2] B) \sim p(\underset{\sim}{\theta}_2 \mid [p M_2] B) \quad (2) \\ (\mathcal{I}_i \mid \underset{\sim}{\theta}_2 [s M_2 : \mathbb{T}] B) \sim p(y_i \mid \underset{\sim}{\theta}_2 [s M_2 : \mathbb{F}] B) \end{array} \right\}$

$(i = 1, \ldots, n)$ $\quad \underset{\sim}{\theta}_2 = (\mu, \sigma_t, r)$ $\quad$ not the same as σ_N in M_1

~~have~~

$\theta = \mu$ has the same meaning in both models

Since $|\mathcal{M}| = (m) < \infty$ (i.e., the number of models in $\mathcal{M}$ is finite) it's sufficient to have a method that compares models 2 at a time: M_1 vs. M_2; better of (M_1, M_2) vs. M_3; best of M_1, M_2, M_3 vs. M_4; ...

The Bayes factor story makes the following provisional (& heroic) assumptions: ① There is a "true" data-generating model M_{DG}; ② $M_{DG} \in \underline{\mathcal{M}}$.

(this is called the M-closed viewpoint)

Then first pretend that $\mathcal{M} = \{M_1, M_2\}$; then

$$\left[\frac{p(M_2 \mid D\,B)}{p(M_1 \mid D\,B)}\right] = \left[\frac{p(M_2 \mid B)}{p(M_1 \mid B)}\right]\left[\frac{p(D \mid M_2\,B)}{p(D \mid M_1\,B)}\right] \quad (*)$$

$$\left(\begin{array}{l}\text{posterior odds}\\ \text{in favor of } M_2\\ \text{over } M_1 \text{ given}\\ D \;\&\; B\end{array}\right) = \left(\begin{array}{l}\text{prior odds}\\ \text{in favor}\\ \text{of } M_2 \text{ over}\\ M_1 \text{ given}\\ B\end{array}\right) \cdot \left(\begin{array}{l}\text{Bayes factor}\\ \text{in favor of}\\ M_2 \text{ over } M_1\\ \text{given } D \text{ and}\\ B\end{array}\right)$$

Q: How specify prior model probabilities $p(M_j \mid B)$?

A: "Please ask me an easier question."

Seriously: many people try to "avoid" this issue by assigning a uniform prior across models:

$$p(M_j \mid B) = \begin{cases} \frac{1}{m} & \text{for } j = 1, \ldots, m \text{ in } \mathcal{M} = \{M_1, \ldots, M_m\} \\ 0 & \text{else} \end{cases}$$

Another idea is to penalize more complex models in the prior (details later if time).

(Occam's Razor)

This task is difficult: what were your prior odds in favor of the t model over the Gaussian model in the NBIO case study before you saw the data? One possible (DR) decision rule in Bayesian model comparison:

DR$_1$: Choose M_{j^*} such that $P(M_{j^*} | DB)$ is the largest posterior model probability across the models in $\mathcal{M} = \{M_1, \ldots, M_m\}$. This requires specifying the prior model (PMPs) probabilities in ⊛ above. One idea: use sensitivity analysis over a variety of PMPs;

if all reasonable PMP choices lead to the same best model, breathe a sigh of relief; if not, "think harder".

Many people punt on this issue by using just the Bayes factors to decide: DR_2: choose the model with the highest Bayes factor in favor of it (which is equivalent to DR_1 with a uniform $(\frac{1}{m})$ prior on $\mathcal{M} = \{M_1, \ldots, M_m\}$)

The terms in the Bayes factor are of the form $p(D \mid M_j; B)$; how compute or approximate?

this is called the marginal likelihood or model evidence

It's actually the prior predictive distribution for the dataset D before it's been observed, given only that $(M_j = M_{DG})$ & B.

$p(D|M_j B)$ is hard to think about without knowing $\underset{\sim}{\theta}_j$, so let's bring $\underset{\sim}{\theta}_j$ into the calculations in the usual way:

$$p(D|M_j B) = \int_{(H)_j} p(D|\underset{\sim}{\theta}_j, M_j B)\, p(\underset{\sim}{\theta}_j|M_j B)\, d\underset{\sim}{\theta}_j$$

But $p(D|\underset{\sim}{\theta}_j, M_j B)$ is just the sampling distribution for the dataset D under model M_j; so this says to take a weighted average of the sampling distribution in M_j, weighted by the prior distribution for the parameter vector $\underset{\sim}{\theta}_j$ of M_j.

Serious problem: (LI) if we try to use a low-information prior for $\underset{\sim}{\theta}_j$, $p(D|M_j B)$ can be extremely & uncomfortably sensitive to the LI details.

[simple example]

$(\mu \mid [PM:LI]\, B) \sim N(0, 1000^2)$ (1)

$(Y_i \mid \mu\, [SM:N]\, B) \overset{IID}{\sim} p(y_i \mid \mu\, [SM:N]\, B)$

NBIO data

Call this model M^*

$(i = 1, \ldots, n)$

$(k=1) \quad \theta = (\mu)$

$N(\mu, \sigma^2)$

$p(y \mid \mu\, [SM:N]\, B) =$

(known: set $\sigma = 5$ (sample sd in NBIO data))

$$\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right]$$

$$= \sigma^{-n} (2\pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right]$$

consider this a function of μ for fixed (known) $y = (y_1, \ldots, y_n)$ and σ

(y: NBIO data vector; σ: known)

LI prior for inference: $N(0, 1000^2)$ (LI_2: $\mu_0 = 0$, $\sigma_0 = 1000$)

evidence for M^*

$$p(y \mid M^* B) = \int_{-\infty}^{\infty} \sigma^{-n} (2\pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \cdot p(\mu \mid [PM:LI_2]\, B)\, d\mu$$

what if we tried to approximate this integral with (IID) MC?

The basic (IID) MC algorithm for approximating an integral of the form

$$(**) = \int_{\Theta} g(\theta) \cdot \boxed{p(\theta)}\, d\theta$$

PDF

is simply as follows:

① draw θ_j^* from $p(\theta)$ for $j = 1, \ldots, m$ (IID; large number of MC draws)

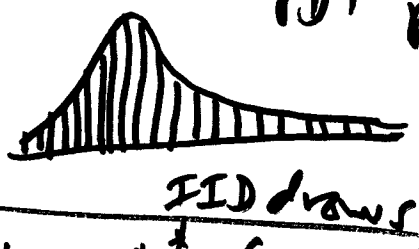
② approximate (**) by

$$\int_{\Theta} g(\theta) \cdot p(\theta)\, d\theta \approx \frac{1}{m} \sum_{j=1}^{m} g(\theta_j^*)$$

that's all there is to it.

To see why,

imagine approximating $p(\theta)$ with many equally-wide histogram bars



IID draws

sampling θ_j^* from $p(\theta)$ will cause each histogram bar to have a # of θ_j^* values in it that's (approx.) proportional to $p(\text{hist. bar center})$

(**) is a weighted average of $g(\theta)$ values, weighted by the $p(\theta)$ values.

(R code) Applying the MC integration idea to marginal likelihoods is now straightforward

$$p(D \mid M_j, \beta) \doteq \frac{1}{M} \sum_{i=1}^{M} p(D \mid \theta_{ij}^*, M_j, \beta),$$

in which $\theta_{ij}^* \overset{IID}{\sim} p(\theta_j \mid \hat{\beta}^{M_j})$ for $i = 1, \ldots, M$

(R code) But if we want $p(\theta_j \mid \hat{\beta}^{M_j})$ to be a LI prior, we must be careful in how we specify such a prior (i.e., LI prior ⊛ specification in Bayesian model comparison is different from LI ⊛⊛ prior specification in Bayesian inference & predictions) ← (easier)

& border: using MC to approximate a marginal likelihood with a LI prior → next page (⊛⊛⊛)

many LI priors in inference lead to the same posterior

(k++) can (as mentioned above) be extremely sensitive to the LI prior details — I've found that the following works well: ① do enough of a likelihood analysis to identify the interval (99.99%) for each parameter: suppose that this interval is (say) $\left[(\hat{\theta}_{ij})_{left}, (\hat{\theta}_{ij})_{right}\right]$ for parameter θ_{ij} (& similarly for the other parameters) ② specify the LI prior as follows: $(\theta_{ij} | M_k B) \sim U\left[(\hat{\theta}_{ij})_{left}, (\hat{\theta}_{ij})_{right}\right]$ (and so on (independently) for the other parameters) (R code)

the Bayes factor in favor of M_2 over M_1 is of the form $BF(M_2 \| M_1 | D B) =$

$$\frac{p(D|M_2 B)}{p(D|M_1 B)}$$

marginal likelihood values $p(D|M_j B)$ tend to be

really small (e.g., 10^{-100}), so the log function comes to the rescue again:

Choose M_2 over M_1 if $BF(M_2 \| M_1 | DB) > 1$ $\leftrightarrow$ (3 equivalent decision rules)

if $p(D|M_2 B) > p(D|M_1, B)$ $\leftrightarrow$

if $\log p(D|M_2 B) > \log p(D|M_1 B)$

so attention shifts to $\log p(D|M_j B)$

here's where Mr. Laplace's approximation idea comes in:

here's Mr. Laplace's idea: ~~[illegible]~~ un-normalized

$p(D|\underset{\sim}{\theta}_j M_j B)$ = likelihood function for $\underset{\sim}{\theta}_j$ in M_j,

$$\log p(D|M_j B) = \log \int_{\Theta_j} \underbrace{p(D|\underset{\sim}{\theta}_j M_j B)\, p(\underset{\sim}{\theta}_j|M_j B)}_{\text{integrand}}\, d\underset{\sim}{\theta}_j$$

and $p(\underset{\sim}{\theta}_j | M_j, B)$ is the prior for $\underset{\sim}{\theta}_j$ in model M_j, so the integrand is the un-normalized posterior for $\underset{\sim}{\theta}_j$ under model M_j

for even moderate n when $D = \underset{\sim}{x} = (y_1, \ldots, y_n)$, this posterior should be approximately multivariate Normal with mean vector $\hat{\underset{\sim}{\theta}}_j$ (MLE) and covariance matrix $\hat{\Sigma}_{\underset{\sim}{\theta}_j} = \hat{I}^{-1}_{\underset{\sim}{\theta}_j}$, the inverse of the observed information matrix

Let's integrate the integrand only over that subset of Θ_j where the integrand is positive (the other parts of Θ_j contribute zero to the integral); then we can write the integrand as $p(D | \underset{\sim}{\theta}_j, M_j, B) p(\underset{\sim}{\theta}_j | M_j, B) = \exp\{\log[\,]\}$ (nice trick)

$x > 0 \rightarrow x = \exp[\log x]$

As we did in computing the deviance, let's choose the likelihood function $\ell(\underset{\sim}{\theta}_j | D M_j B)$ so that it exactly equals the sampling dist. $p(D | \underset{\sim}{\theta}_j M_j B)$ (this uniquely defines what would otherwise be the arbitrary multiplicative constant in the definition of the likelihood) i.e.

$$g(\theta) \Longleftrightarrow$$

then integrand $= \exp\{ \underbrace{ll(\underset{\sim}{\theta}_j | M_j B) + \log p(\underset{\sim}{\theta}_j | M_j B)}_{\hat{g}(\underset{\sim}{\theta}_j)} \}$

Mr. Laplace now expands this function $\hat{g}(\underset{\sim}{\theta}_j)$ in a 3-term Taylor series approximation around the point $\underset{\sim}{\theta}_j = \underset{\sim}{\bar{\theta}}_j$

For simplicity let's look at the special case in which $k_j = \left(\begin{array}{c}\text{\# of free} \\ \text{parameters} \\ \text{in } M_j\end{array}\right) = 1$ a scalar

i.e., $\underset{\sim}{\theta}_j = \theta_j$

The Taylor series approximation

then looks like this:

$$g(\theta_j) \doteq g(\hat{\theta}_j) + \left[\frac{d}{d\theta_j} g(\theta_j)\right]_{\theta_j=\hat{\theta}_j} \cdot (\theta_j - \hat{\theta}_j) + \frac{1}{2}\left[\frac{d^2}{d\theta_j^2} g(\theta_j)\right]_{\theta_j=\hat{\theta}_j} \cdot (\theta_j - \hat{\theta}_j)^2$$

substituting in $g(\theta_j) = ll(\theta_j | D m_j \cdot B) + \log p(\theta_j | m_j \cdot B)$ gives the following:

$$ll(\theta_j | D m_j B) + \log p(\theta_j | m_j \cdot B) \doteq ll(\hat{\theta}_j | D m_j B) + \log p(\hat{\theta}_j | m_j B) + \left\{\left[\frac{d}{d\theta_j} ll(\theta | D m_j B)\right]_{\theta=\hat{\theta}} + \left[\frac{d}{d\theta_j} \log p(\theta_j | m_j B)\right]_{\theta=\hat{\theta}_{MLE}}\right\} \cdot (\theta - \hat{\theta}_j) + \frac{1}{2}\left\{\left[\frac{d^2}{d\theta_j^2} ll(\theta | D m_j \cdot B)\right]_{\theta_j=\hat{\theta}_j} + \left[\frac{d^2}{d\theta_j^2} \log p(\theta_j | m_j \cdot B)\right]_{\theta=\hat{\theta}_j}\right\} \cdot (\theta_j - \hat{\theta}_j)^2$$

looks like a real mess but major simplification is imminent

Simplification: ① $\hat{\theta}_j$ maximizes $\log p(\theta_j | D M_j B)$

so $\left[\frac{d}{d\theta_j} \log p(\theta_j | D M_j B)\right]_{\theta_j = \hat{\theta}_j} = 0$

② Assume now that the prior is LI: then it will be close to flat in the region around $\theta_j = \hat{\theta}_j$, so $\left[\frac{d}{d\theta_j} \log p(\theta_j | M_j B)\right]_{\theta_j = \hat{\theta}_j}$ is also $\doteq 0$

Therefore the entire linear term

$\left[\quad\right](\theta_j - \hat{\theta}_j) \doteq 0$

③ By the same reasoning as in ②,

$\left[\frac{d^2}{d\theta_j^2} \log p(\theta_j | M_j B)\right]_{\theta_j = \hat{\theta}_j}$ is also $\doteq 0$

$= -\hat{I}(\hat{\theta}_j)$

and $\left[\frac{d^2}{d\theta_j^2} \log p(\theta_j | D M_j B)\right]_{\theta = \hat{\theta}_j} = -$ (observed information) (!)

So Mr. Laplace's approximation becomes

$$(\text{integrand}) = p(D|\underset{\sim}{\theta}_j M_j B) \cdot p(\underset{\sim}{\theta}_j | M_j B)$$

$$\doteq \exp\left[ll(\hat{\theta}_j | D M_j B) + \log p(\hat{\theta}_j | M_j B) - \frac{1}{2} \hat{I}(\hat{\theta}_j) \sum_i \left[(\theta_j - \hat{\theta}_j)^2 \right] \right]$$ where

integrating over θ_j, so the first two terms are constant in θ_j & can be brought out in front of the integral

This yields (with $k_j = 1$)

$$\begin{pmatrix} \text{evidence} \\ \text{in favor} \\ \text{of } M_j \end{pmatrix} = p(D|M_j B) \bullet$$

$$= \int_R p(D|\theta_j M_j B)\, p(\theta_j | M_j B)\, d\theta_j$$

$$= \exp\left(ll(\hat{\theta}_j | D M_j B) \right) \cdot \exp\left[\log p(\hat{\theta}_j | M_j B) \right] \cdot (*)$$

in which $\textcircled{*} = \int_R \exp\left[-\frac{1}{2}\hat{I}(\hat{\theta}_j)(\theta_j - \hat{\theta}_j)^2\right] d\theta_j$

But this is the ~~just integral of an~~ un-normalized Normal distribution (!) As Wolfram will tell you after some algebra

$$\int_{-\infty}^{\infty} \exp\left[-c_1(\theta - c_2)^2\right] d\theta = \sqrt{\frac{\pi}{c_1}} \quad (c_1 > 0)$$

so

$$\begin{pmatrix} \text{evidence} \\ \text{in favor} \\ \text{of } M_j \end{pmatrix} = p(D \mid M_j, B)$$

$$\doteq \exp\left[\mathcal{L}(\hat{\theta}_j \mid D M_j, B)\right] \cdot \exp\left[\log p(\hat{\theta}_j \mid M_j, B)\right] \cdot \sqrt{\frac{\pi}{\frac{1}{2}\hat{I}(\hat{\theta}_j)}}$$

and therefore

$$\log\begin{pmatrix} \text{evidence} \\ \text{in favor} \\ \text{of } M_j \end{pmatrix} = \log p(D \mid M_j, B) \qquad (**)$$

$$\doteq \boxed{\mathcal{L}(\hat{\theta}_j \mid D M_j)} + \log p(\hat{\theta}_j \mid M_j, B) + \frac{1}{2}\log(2\pi) - \frac{1}{2}\log|\hat{I}(\hat{\theta}_j)|$$

an old friend

Looking back at the Taylor series approximation

on p. (14), $\log(\text{integrand}) \doteq$

$$\ell\ell(\tilde{\theta}_j \mid D, M_j, B) + \log p(\tilde{\theta} \mid M_j, B)$$ ← $\underset{=}{O}(1)$ as $n \to \infty$

$$+ \left\{ \text{complicated constant} \right\} (\theta - \tilde{\theta}_j)$$ ← $\underset{=}{O}\left(\frac{1}{\sqrt{n}}\right)$ (think of $SE(\tilde{\theta}_j) = \frac{c}{\sqrt{n}}$)

$$+ \left\{ \text{complicated constant} \right\} (\theta - \tilde{\theta}_j)^2$$ ← $\underset{=}{O}\left(\frac{1}{n}\right)$ (think of $\hat{V}(\tilde{\theta}_j)$)

The generalization of (**) on p. (17) turns out to be (including the $\underset{=}{O}(\cdot)$ story):

Laplace approximation to log marginal likelihood

$$\log\left(\begin{matrix}\text{evidence} \\ \text{in favor} \\ \text{of } M_j\end{matrix}\right) = \log p(D \mid M_j, B) \doteq \ell\ell(\tilde{\theta}_j \mid D, M_j, B)$$

(***)

$$+ \log p(\tilde{\theta}_j \mid M_j, B) + \frac{k_j}{2} \log(2\pi) - \frac{1}{2} \log \left| \hat{I}(\tilde{\theta}_j) \right| + \underset{=}{O}\left(\frac{1}{n}\right)$$

(k_j = # of free parameters in M_j)

in which $|\hat{I}(\tilde{\theta}_j)|$ is the determinant of the observed information matrix and $\tilde{\theta}_j$ is the (k_j-dimensional) MLE vector.

Since the Laplace approximation is accurate to $O(\frac{1}{n})$, it's quite accurate; after all, the CLT is only accurate to $O(\frac{1}{\sqrt{n}})$ (R code)

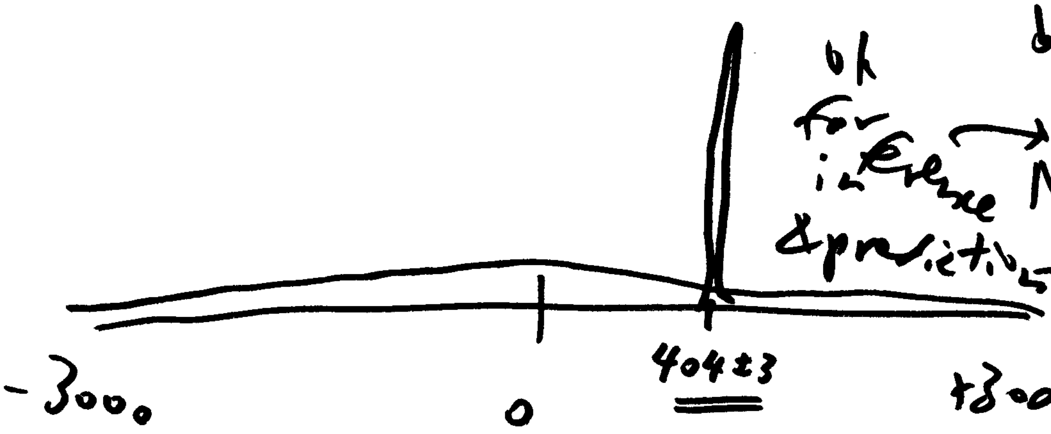
You can compute (***) using only output from 'optim' applied to the log likelihood function with the 'hessian = T' option

---

One more wrinkle in this story:

One way to obtain useful LI priors for Bayesian model comparison with Bayes factors

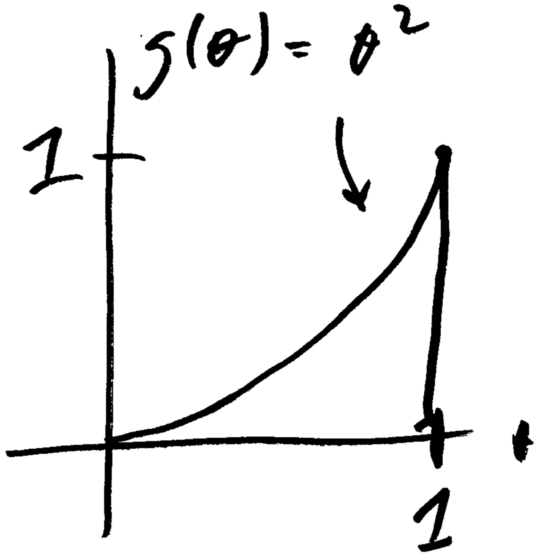
(to be continued)



ok for inference & prediction → $N(0, \sigma = 1000)$ (crossed out)

but crop for model comparison

$\sim U(404, 404)$

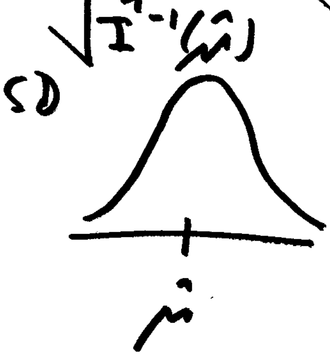


$$\int_0^1 g(\theta)\,d\theta = \int_0^1 \theta^2\,d\theta = \frac{\theta^3}{3}\Big|_0^1 = \frac{1}{3}$$

$$\int_0^1 g(\theta)\,\underbrace{p(\theta)}_{\text{PDF}}\,d\theta \qquad p(\theta) = 1 \leftarrow U(0,1)$$

$$\int_0^1 \overline{g(\theta)}^{\,\theta^2}\,d\theta = \frac{1}{M}\sum_{i=1}^{M} \boxed{g(\theta_i^*)}^{(\theta_i^*)^2},$$

$$\theta_i^* \overset{IID}{\sim} U(0,1)$$



$p(\mu \mid D\, M_j \ldots B),$ (n large)

$c_1 \exp\left[-c_2 (\mu - \hat{\mu})^2\right]$