Prior selection and model choice

Christian P. Robert

Université Paris Dauphine and CREST-INSEE
http://www.ceremade.dauphine.fr/~xian

Mathematischen Forschungsinstitut Oberwolfach
October 18, 2005

Outline

1 Bayesian Model Choice

Setup

Choice of models

Several models available for the same observation

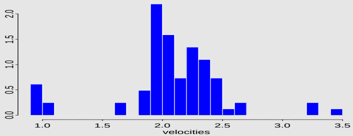
$$\mathfrak{M}_i : x \sim f_i(x|\theta_i), \qquad i \in \mathfrak{I}$$

where $\mathfrak{I}$ can be finite or infinite

Example (Galaxy normal mixture)

Set of observations of radial speeds of 82 galaxies possibly modelled as a mixture of normal distributions

$$\mathfrak{M}_i : x_j \sim \sum_{\ell=1}^{i} p_{\ell i} \mathcal{N}(\mu_{\ell i}, \sigma^2_{\ell i})$$



Bayesian resolution

B Framework

Probabilises the entire model/parameter space
This means:

- allocating probabilities p_i to all models $\mathfrak{M}_i$
- defining priors $\pi_i(\theta_i)$ for each parameter space Θ_i

Formal solutions

Resolution

1. Compute

$$p(\mathfrak{M}_i|x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)\mathrm{d}\theta_j}$$

2. Take largest $p(\mathfrak{M}_i|x)$ to determine ''best'' model,

or use averaged predictive

$$\sum_j p(\mathfrak{M}_j|x) \int_{\Theta_j} f_j(x'|\theta_j)\pi_j(\theta_j|x)\mathrm{d}\theta_j$$

Several types of problems

- Concentrate on selection perspective:
  - averaging = estimation = non-parsimonious = no-decision
  - how to integrate loss function/decision/consequences
  - representation of parsimony/sparcity (Ockham's rule)
  - how to fight overfitting for nested models

Which loss ?

Several types of problems (2)

- Choice of prior structures
  - adequate weights p_i:
    if $\mathfrak{M}_1 = \mathfrak{M}_2 \cup \mathfrak{M}_3$, $p(\mathfrak{M}_1) = p(\mathfrak{M}_2) + p(\mathfrak{M}_3)$?
  - priors distributions
    - $\pi_i(\theta_i)$ defined for every $i \in \mathfrak{I}$
    - $\pi_i(\theta_i)$ *proper* (Jeffreys)
    - $\pi_i(\theta_i)$ coherent (?) for nested models

Warning

Parameters common to several models must be treated as separate entities!

Several types of problems (3)

- Computation of predictives and marginals
  - infinite dimensional spaces
  - integration over parameter spaces
  - integration over different spaces
  - summation over many models (2^k)

[MCMC resolution = another talk]

A function of posterior probabilities

Definition (Bayes factors)

Models $\mathfrak{M}_1$ vs. $\mathfrak{M}_2$

$$B_{12} = \frac{\Pr(\mathcal{M}_1|x)}{\Pr(\mathcal{M}_2|x)} \bigg/ \frac{\Pr(\mathcal{M}_1)}{\Pr(\mathcal{M}_2)}$$

$$= \frac{\int f_1(x|\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1}{\int f_2(x|\theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2}$$

[Good, 1958 & Jeffreys, 1961]

Goto Poisson example

Self-contained concept

- eliminates choice of $\Pr(\mathfrak{M}_i)$
- but depends on the choice of $\pi_i(\theta_i)$
- Bayesian/marginal likelihood ratio
- Jeffreys' scale of evidence

A battery of difficulties

Improper priors not allowed here

If

$$\int_{\Theta_1} \pi_1(d\theta_1) = \infty \quad \text{or} \quad \int_{\Theta_2} \pi_2(d\theta_2) = \infty$$

then either π_1 or π_2 cannot be normalised uniquely but the normalisation matters in the Bayes factor Recall Bayes factor

Constants matter

Example (Poisson versus Negative binomial)

If $\mathfrak{M}_1$ is a $\mathscr{P}(\lambda)$ distribution and $\mathfrak{M}_2$ is a $\mathscr{N}\mathscr{B}(m,p)$ distribution, we can take

$$\begin{aligned} \pi_1(\lambda) &= 1/\lambda \\ \pi_2(m,p) &= \tfrac{1}{M}\,\mathbb{I}_{\{1,\cdots,M\}}(m)\,\mathbb{I}_{[0,1]}(p) \end{aligned}$$

Constants matter (cont'd)

Example (Poisson versus Negative binomial (2))

then

$$\begin{aligned} B_{12} &= \frac{\displaystyle\int_0^\infty \frac{\lambda^{x-1}}{x!} e^{-\lambda} d\lambda}{\displaystyle\frac{1}{M}\sum_{m=1}^{M}\int_0^\infty \binom{m}{x-1} p^x (1-p)^{m-x} dp} \\ &= 1 \Big/ \frac{1}{M}\sum_{m=x}^{M} \binom{m}{x-1} \frac{x!(m-x)!}{m!} \\ &= 1 \Big/ \frac{1}{M}\sum_{m=x}^{M} x/(m-x+1) \end{aligned}$$

Constants matter (cont'd)

Example (Poisson versus Negative binomial (3))

- does not make sense because $\pi_1(\lambda) = 10/\lambda$ leads to a different answer, **ten times larger!**
- same thing when both priors are improper

Improper priors on common (nuisance) parameters do not matter (so much)

Vague proper priors are not the solution

Taking a proper prior and take a "very large" variance (e.g., BUGS) will most often result in an undefined or ill-defined limit

Example (Lindley's paradox)

If testing $H_0 : \theta = 0$ when observing $x \sim \mathcal{N}(\theta, 1)$, under a normal $\mathcal{N}(0, \alpha)$ prior $\pi_1(\theta)$,

$$B_{01}(x) \overset{\alpha \longrightarrow \infty}{\longrightarrow} 0$$

Vague proper priors are not the solution (cont'd)

Example (Poisson versus Negative binomial (4))

$$\begin{aligned} B_{12} &= \frac{\displaystyle\int_0^1 \frac{\lambda^{\alpha+x-1}}{x!} e^{-\lambda\beta} \mathrm{d}\lambda}{\displaystyle\frac{1}{M} \sum_m \frac{x}{m-x+1} \frac{\beta^\alpha}{\Gamma(\alpha)}} \quad \text{if } \lambda \sim \mathcal{G}a(\alpha, \beta) \\ &= \frac{\Gamma(\alpha+x)}{x!\,\Gamma(\alpha)} \beta^{-x} \Big/ \frac{1}{M} \sum_m \frac{x}{m-x+1} \\ &= \frac{(x+\alpha-1)\cdots\alpha}{x(x-1)\cdots 1} \beta^{-x} \Big/ \frac{1}{M} \sum_m \frac{x}{m-x+1} \end{aligned}$$

depends on choice of $\alpha(\beta)$ or $\beta(\alpha) \longrightarrow 0$

Pseudo-Bayes factors

Idea

Use one part $x_{[i]}$ of the data x to make the prior proper:

- π_i improper but $\pi_i(\cdot|x_{[i]})$ proper
- and
$$\frac{\int f_i(x_{[n/i]}|\theta_i)\, \pi_i(\theta_i|x_{[i]}) \mathrm{d}\theta_i}{\int f_j(x_{[n/i]}|\theta_j)\, \pi_j(\theta_j|x_{[i]}) \mathrm{d}\theta_j}$$
independent of normalizing constant
- Use remaining $x_{[n/i]}$ to run test as if...

Motivation

- Working principle for improper priors
- Gather enough information from data to gain properness
- and use this properness to run the test on remaining data
- does not use x twice as in Aitkin's (1991)

Details

Since $\pi_1(\theta_1|x_{[i]}) = \dfrac{\pi_1(\theta_1) f^1_{[i]}(x_{[i]}|\theta_1)}{\int \pi_1(\theta_1) f^1_{[i]}(x_{[i]}|\theta_1) \mathrm{d}\theta_1}$

then

$$
\begin{aligned}
B_{12}(x_{[n/i]}) &= \frac{\int f^1_{[n/i]}(x_{[n/i]}|\theta_1)\pi_1(\theta_1|x_{[i]})\mathrm{d}\theta_1}{\int f^2_{[n/i]}(x_{[n/i]}|\theta_2)\pi_2(\theta_2|x_{[i]})\mathrm{d}\theta_2} \\
&= \frac{\int f_1(x|\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1}{\int f_2(x|\theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2} \frac{\int \pi_2(\theta_2) f^2_{[i]}(x_{[i]}|\theta_2)\mathrm{d}\theta_2}{\int \pi_1(\theta_1) f^1_{[i]}(x_{[i]}|\theta_1)\mathrm{d}\theta_1} \\
&= B^N_{12}(x) B_{21}(x_{[i]})
\end{aligned}
$$

ⓒ **Independent of scaling factor!**

More problems

- depends on the choice of $x_{[i]}$
- many ways of combining pseudo-Bayes factors
  - AIBF $= B^N_{ji} \frac{1}{L} \sum_\ell B_{ij}(x_{[\ell]})$
  - MIBF $= B^N_{ji}$ med$[B_{ij}(x_{[\ell]})]$
  - GIBF $= B^N_{ji}$ exp $\frac{1}{L} \sum_\ell \log B_{ij}(x_{[\ell]})$
- not often exact Bayes

[Berger & Pericchi, 1996]

More problems (cont'd)

Example (Mixtures)

There is no sample size that proper-ises improper priors, except if a training sample is allocated to *each* component

Reason If

$$x_1, \ldots, x_n \sim \sum_{i=1}^k p_i f(x|\theta_i)$$

and

$$\pi(\theta) = \prod_i \pi_i(\theta_i) \text{ with } \int \pi_i(\theta_i)\mathrm{d}\theta_i = +\infty\,,$$

the posterior is never defined, because

$$\Pr(\text{“no observation from } f(\cdot|\theta_i)\text{”}) = (1-p_i)^n$$

Intrinsic priors

There may exist a true prior that provides the same Bayes factor

Example (Normal mean)

Take $x \sim \mathcal{N}(\theta, 1)$ with either $\theta = 0$ ($\mathfrak{M}_1$) or $\theta \neq 0$ ($\mathfrak{M}_2$) and $\pi_2(\theta) = 1$.

Then

$$
\begin{aligned}
B^{AIBF}_{21} &= B_{21} \tfrac{1}{\sqrt{2\pi}} \tfrac{1}{n} \textstyle\sum_{i=1}^n e^{-x_1^2/2} && \approx B_{21} && \text{for } \mathcal{N}(0,2) \\
B^{MIBF}_{21} &= B_{21} \tfrac{1}{\sqrt{2\pi}} e^{-\text{med}(x_1^2)/2} && \approx 0.93 B_{21} && \text{for } \mathcal{N}(0,1.2)
\end{aligned}
$$

[Berger and Pericchi, 1998]

When such a prior exists, it is called an **intrinsic prior**

Intrinsic priors (cont'd)

Example (Exponential scale)

Take $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \exp(\theta - x)\mathbb{I}_{x \geq \theta}$
and $H_0 : \theta = \theta_0,\ H_1 : \theta > \theta_0$, with $\pi_1(\theta) = 1$
Then

$$B_{10}^A = B_{10}(x) \frac{1}{n} \sum_{i=1}^{n} \left[e^{x_i - \theta_0} - 1 \right]^{-1}$$

is the Bayes factor for

$$\pi_2(\theta) = e^{\theta_0 - \theta} \left\{ 1 - \log\left(1 - e^{\theta_0 - \theta}\right) \right\}$$

Most often, however, the pseudo-Bayes factors do not correspond to any true Bayes factor

2 Compatible priors

[Joint work with C. Celeux, G. Consonni and J.M. Marin]

Principle

Difficulty of finding simultaneously priors on a collection of models $\mathfrak{M}_i$ $(i \in \mathfrak{I})$
Easier to start from a single prior on a "big" model and to derive the others from a coherence principle

[Dawid & Lauritzen, 2000]

Projection approach

For $\mathfrak{M}_2$ submodel of $\mathfrak{M}_1$, π_2 can be derived as the distribution of $\theta_2^\perp(\theta_1)$ when $\theta_1 \sim \pi_1(\theta_1)$ and $\theta_2^\perp(\theta_1)$ is a projection of θ_1 on $\mathfrak{M}_2$, e.g.

$$d(f(\cdot\,|\theta_1), f(\cdot\,|{\theta_1}^\perp)) = \inf_{\theta_2 \in \Theta_2} d(f(\cdot\,|\theta_1)\,, f(\cdot\,|\theta_2))\,.$$

where d is a divergence measure

[McCulloch & Rossi, 1992]

Or we can look instead at the posterior distribution of

$$d(f(\cdot\,|\theta_1), f(\cdot\,|{\theta_1}^\perp))$$

[Goutis & Robert, 1998]

Operational principle for variable selection

Selection rule

Among all subsets $\mathcal{A}$ of covariates such that

$$d(\mathfrak{M}_g, \mathfrak{M}_{\mathcal{A}}) = \mathbb{E}_x[d(f_g(\cdot|x,\alpha), f_{\mathcal{A}}(\cdot|x_{\mathcal{A}}, \alpha^{\perp}))] < \epsilon$$

select the submodel with the smallest number of variables.

[Dupuis & Robert, 2001]

Kullback proximity

Alternative

Definition (Compatible prior)

Given a prior π_1 on a model $\mathfrak{M}_1$ and a submodel $\mathfrak{M}_2$, a prior π_2 on $\mathfrak{M}_2$ is *compatible* with π_1 when it achieves the minimum Kullback divergence between the corresponding marginals:
$m_1(x;\pi_1) = \int_{\Theta_1} f_1(x|\theta)\pi_1(\theta)\mathrm{d}\theta$ and
$m_2(x);\pi_2 = \int_{\Theta_2} f_2(x|\theta)\pi_2(\theta)\mathrm{d}\theta$,

$$\pi_2 = \arg\min_{\pi_2} \int \log\left(\frac{m_1(x;\pi_1)}{m_2(x;\pi_2)}\right) m_1(x;\pi_1)\,\mathrm{d}x$$

Difficulties

- Does not give a working principle when $\mathfrak{M}_2$ is not a submodel $\mathfrak{M}_1$
- Depends on the choice of π_1
- Prohibits the use of improper priors
- Worse: useless in unconstrained settings...

Case of exponential families

Models

$$\mathfrak{M}_1 : \{f_1(x|\theta), \theta \in \Theta\}$$

and

$$\mathfrak{M}_2 : \{f_2(x|\lambda), \lambda \in \Lambda\}$$

sub-model of $\mathcal{M}_1$,

$$\forall \lambda \in \Lambda, \exists\, \theta(\lambda) \in \Theta, f_2(x|\lambda) = f_1(x|\theta(\lambda))$$

Both $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are natural exponential families

$$\begin{aligned} f_1(x|\theta) &= h_1(x)\exp(\theta^{\mathsf{T}} t_1(x) - M_1(\theta)) \\ f_2(x|\lambda) &= h_2(x)\exp(\lambda^{\mathsf{T}} t_2(x) - M_2(\lambda)) \end{aligned}$$

Conjugate priors

Parameterised (conjugate) priors

$$\begin{aligned}\pi_1(\theta; s_1, n_1) &= C_1(s_1, n_1) \exp(s_1^{\mathsf T}\theta - n_1 M_1(\theta))\\ \pi_2(\lambda; s_2, n_2) &= C_2(s_2, n_2) \exp(s_2^{\mathsf T}\lambda - n_2 M_2(\lambda))\end{aligned}$$

with closed form marginals ($i = 1, 2$)

$$m_i(x; s_i, n_i) = \int f_i(x|u)\pi_i(u)du = \frac{h_i(x)C_i(s_i, n_i)}{C_i(s_i + t_i(x), n_i + 1)}$$

Conjugate compatible priors

(Q.) Existence and unicity of Kullback-Leibler projection

$$\begin{aligned}(s_2^*, n_2^*) &= \arg\min_{(s_2,n_2)} \mathfrak{KL}(m_1(\cdot; s_1, n_1), m_2(\cdot; s_2, n_2))\\ &= \arg\min_{(s_2,n_2)} \int \log\left(\frac{m_1(x; s_1, n_1)}{m_2(x; s_2, n_2)}\right) m_1(x; s_1, n_1)\mathrm{d}x\end{aligned}$$

A sufficient condition

Sufficient statistic $\psi = (\lambda, -M_2(\lambda))$

Theorem (Existence)

If, for all (s_2, n_2), the matrix

$$\mathbb{V}^{\pi_2}_{s_2,n_2}[\psi] - \mathbb{E}^{m_1}_{s_1,n_1}\left[\mathbb{V}^{\pi_2}_{s_2,n_2}(\psi|x)\right]$$

is semi-definite negative, the conjugate compatible prior exists, is unique and satisfies

$$\begin{aligned}\mathbb{E}^{\pi_2}_{s_2^*,n_2^*}[\lambda] - \mathbb{E}^{m_1}_{s_1,n_1}[\mathbb{E}^{\pi_2}_{s_2^*,n_2^*}(\lambda|x)] &= 0\\ \mathbb{E}^{\pi_2}_{s_2^*,n_2^*}(M_2(\lambda)) - \mathbb{E}^{m_1}_{s_1,n_1}[\mathbb{E}^{\pi_2}_{s_2^*,n_2^*}(M_2(\lambda)|x)] &= 0.\end{aligned}$$

Application to linear regression

$\mathfrak{M}_1$ and $\mathfrak{M}_2$ are two nested Gaussian linear regression models with Zellner's g-priors and the same variance $\sigma^2 \sim \pi(\sigma^2)$:

1. $\mathfrak{M}_1$:

   $$y|\beta_1, \sigma^2 \sim \mathcal{N}(X_1\beta_1, \sigma^2), \quad \beta_1|\sigma^2 \sim \mathcal{N}\left(s_1, \sigma^2 n_1 (X_1^{\mathsf T} X_1)^{-1}\right)$$

   where X_1 is a $(n \times k_1)$ matrix of rank $k_1 \leq n$
2. $\mathfrak{M}_2$:

   $$y|\beta_2, \sigma^2 \sim \mathcal{N}(X_2\beta_2, \sigma^2), \quad \beta_2|\sigma^2 \sim \mathcal{N}\left(s_2, \sigma^2 n_2 (X_2^{\mathsf T} X_2)^{-1}\right),$$

   where X_2 is a $(n \times k_2)$ matrix with $\text{span}(X_2) \subseteq \text{span}(X_1)$

For a fixed (s_1, n_1), we need the projection $(s_2, n_2) = (s_1, n_1)^{\perp}$

Compatible g-priors

Since σ^2 is a nuisance parameter, we can minimize the Kullback-Leibler divergence between the two marginal distributions conditional on σ^2: $m_1(y|\sigma^2; s_1, n_1)$ and $m_2(y|\sigma^2; s_2, n_2)$

Theorem

Conditional on σ^2, the conjugate compatible prior of $\mathfrak{M}_2$ wrt $\mathfrak{M}_1$ is

$$\beta_2 | X_2, \sigma^2 \sim \mathcal{N}\left(s_2^*, \sigma^2 n_2^* (X_2^T X_2)^{-1}\right)$$

with

$$\begin{aligned} s_2^* &= (X_2^T X_2)^{-1} X_2^T X_1 s_1 \\ n_2^* &= n_1 \end{aligned}$$

Variable selection

Regression setup where y regressed on a set $\{x_1, \ldots, x_p\}$ of p **potential explanatory** regressors (plus intercept)

Corresponding 2^p submodels $\mathfrak{M}_\gamma$, where $\gamma \in \Gamma = \{0,1\}^p$ indicates inclusion/exclusion of variables by a binary representation

Notations

For model $\mathfrak{M}_\gamma$,

- q_γ variables are included
- $t_1(\gamma) = \{t_{1,1}(\gamma), \ldots, t_{1,q_\gamma}(\gamma)\}$ are the indices of those variables and $t_0(\gamma)$ the indices of the variables *not* included
- For $\beta \in \mathbb{R}^{p+1}$,

$$\begin{aligned} \beta_{t_1(\gamma)} &= \left[\beta_0, \beta_{t_{1,1}(\gamma)}, \ldots, \beta_{t_{1,q_\gamma}(\gamma)}\right] \\ \beta_{t_0(\gamma)} &= \left[\beta_{t_{0,1}(\gamma)}, \ldots, \beta_{t_{0,p-q_\gamma}(\gamma)}\right] \\ X_{t_1(\gamma)} &= \left[1_n | x_{t_{1,1}(\gamma)} | \ldots | x_{t_{1,q_\gamma}(\gamma)}\right]. \end{aligned}$$

Submodel $\mathfrak{M}_\gamma$ is thus

$$y | \beta, \gamma, \sigma^2 \sim \mathcal{N}\left(X_{t_1(\gamma)} \beta_{t_1(\gamma)}, \sigma^2 I_n\right)$$

Global and compatible priors

Use Zellner's g-prior, i.e. a normal prior for β conditional on σ^2,

$$\beta | \sigma^2 \sim \mathcal{N}(\tilde{\beta}, c\sigma^2 (X^\mathsf{T} X)^{-1})$$

and a Jeffreys prior for σ^2,

$$\pi(\sigma^2) \propto \sigma^{-2}$$

▸ Noninformative g

Resulting compatible prior

$$\mathcal{N}\left(\left(X_{t_1(\gamma)}^\mathsf{T} X_{t_1(\gamma)}\right)^{-1} X_{t_1(\gamma)}^\mathsf{T} X \tilde{\beta}, c\sigma^2 \left(X_{t_1(\gamma)}^\mathsf{T} X_{t_1(\gamma)}\right)^{-1}\right)$$

[Surprise!]

Model index

For the hierarchical parameter γ, we use

$$\pi(\gamma) = \prod_{i=1}^{p} \tau_i^{\gamma_i}(1-\tau_i)^{1-\gamma_i},$$

where τ_i corresponds to the prior probability that variable i is present in the model.
Typically, when no prior information is available, $\tau_1 = \ldots = \tau_p = 1/2$, ie a uniform prior

$$\pi(\gamma) = 2^{-p}$$

Posterior model probability

Can be obtained in closed form:

$$\pi(\gamma|y) \propto (c+1)^{-(q_\gamma+1)/2}\left[y^\mathsf{T}y - \frac{c}{c+1}y^\mathsf{T}P_1 y + \frac{1}{c+1}\tilde\beta^\mathsf{T}X^\mathsf{T}P_1X\tilde\beta - \frac{2}{c+1}y^\mathsf{T}P_1X\tilde\beta\right]^{-n/2}.$$

Conditionally on γ, posterior distributions of β and σ^2:

$$\begin{aligned}
\beta_{t_0(\gamma)}|\sigma^2,y,\gamma &\sim \delta(0_{p-q_\gamma}),\\
\beta_{t_1(\gamma)}|\sigma^2,y,\gamma &\sim \mathcal{N}\left[\frac{c}{c+1}(U_1y+U_1X\tilde\beta/c), \frac{\sigma^2 c}{c+1}\left(X_{t_1(\gamma)}^\mathsf{T}X_{t_1(\gamma)}\right)^{-1}\right],\\
\sigma^2|y,\gamma &\sim \mathcal{IG}\left[\frac{n}{2}, \frac{y^\mathsf{T}y}{2} - \frac{c}{2(c+1)}y^\mathsf{T}P_1y + \frac{\tilde\beta^\mathsf{T}X^\mathsf{T}P_1X\tilde\beta}{2(c+1)} - \frac{1}{c+1}y^\mathsf{T}P_1X\tilde\beta\right].
\end{aligned}$$

Noninformative case

Use the same compatible informative g-prior distribution with $\tilde\beta = 0_{p+1}$ and a hierarchical diffuse prior distribution on c,

$$\pi(c) \propto c^{-1}\mathbb{I}_{\mathbb{N}^*}(c)$$

‣ Recall g-prior

The choice of this hierarchical diffuse prior distribution on c is due to the model posterior sensitivity to large values of c:

Taking $\tilde\beta = 0_{p+1}$ and c large does not work

Influence of c

Consider the 10-predictor full model

$$y|\beta,\sigma^2 \sim \mathcal{N}\left(\beta_0 + \sum_{i=1}^{3}\beta_i x_i + \sum_{i=1}^{3}\beta_{i+3}x_i^2 + \beta_7x_1x_2 + \beta_8x_1x3 + \beta_9x_2x_3 + \beta_{10}x_1x_2x_3, \sigma^2 I_n\right)$$

where the x_is are iid $\mathscr{U}(0,10)$

[Casella & Moreno, 2004]

True model: two predictors x_1 and x_2, i.e. $\gamma^* = (1,1,0,\ldots,0)$, and $(\beta_0,\beta_1,\beta_2) = (5,1,3)$, and $\sigma^2 = 4$.

Influence of c^2

γ	$c = 10$	$c = 100$	$c = 10^3$	$c = 10^4$	$c = 10^6$
0,1,2	0.04062	0.35368	0.65858	0.85895	0.98222
0,1,2,7	0.01326	0.06142	0.08395	0.04434	0.00524
0,1,2,4	0.01299	0.05310	0.05805	0.02868	0.00336
0,2,4	0.02927	0.03962	0.00409	0.00246	0.00254
0,1,2,8	0.01240	0.03833	0.01100	0.00126	0.00126

Noninformative case (cont'd)

In the noninformative setting,

$$\pi(\gamma|y,c) \propto (c+1)^{-(q_\gamma+1)/2} \left[y^\mathsf{T} y - \frac{c}{c+1} y^\mathsf{T} P_1 y \right]^{-n/2}$$

and

$$\pi(\gamma|y) \propto \sum_{c=1}^{\infty} c^{-1} (c+1)^{-(q_\gamma+1)/2} \left[y^\mathsf{T} y - \frac{c}{c+1} y^\mathsf{T} P_1 y \right]^{-n/2}$$

which converges for all y's

Casella & Moreno's example

γ	$\sum_{i=1}^{10^5} \pi(\gamma|y,c)\pi(c)$	$\sum_{i=1}^{10^6} \pi(\gamma|y,c)\pi(c)$
0,1,2	0.77969	0.78071
0,1,2,7	0.06229	0.06201
0,1,2,4	0.04138	0.04119
0,1,2,8	0.01684	0.01676
0,1,2,5	0.01611	0.01604

Gibbs approximation

When p large, impossible to compute the posterior probabilities of all of the 2^p models.
Use of a simulation approximation of $\pi(\gamma|y)$

Gibbs sampling

- At $t = 0$, draw γ^0 from the uniform distribution on Γ;
- At t, for $i = 1, \ldots, p$, draw $\gamma_i^t \sim \pi(\gamma_i|y, \gamma_1^t, \ldots, \gamma_{i-1}^t, \ldots, \gamma_{i+1}^{t-1}, \ldots, \gamma_p^{t-1})$

Gibbs approximation (cont'd)

Example (Simulated data)

Severe multicolinearities among predictors for a 20-predictor full model

$$y|\beta, \sigma^2 \sim \mathcal{N}\left(\beta_0 + \sum_{i=1}^{20} \beta_i x_i, \sigma^2 I_n\right)$$

where $x_i = z_i + 3z$, the z_i's and z are iid $\mathcal{N}_n(0_n, I_n)$.
True model with $n = 180$, $\sigma^2 = 4$ and seven predictor variables

$$x_1, x_3, x_5, x_6, x_{12}, x_{18}, x_{20},$$

$$(\beta_0, \beta_1, \beta_3, \beta_5, \beta_6, \beta_{12}, \beta_{18}, \beta_{20}) = (3, 4, 1, -3, 12, -1, 5, -6)$$

Gibbs approximation (cont'd)

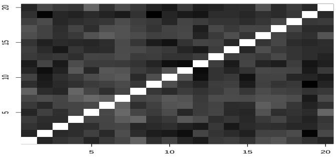


Figure: Correlations between the 20 predictors (white=1, black=0)

Gibbs approximation (cont'd)

Example (Simulated data (2))

Results

γ	$\pi(\gamma\|y)$	$\widehat{\pi(\gamma\|y)}^{GIBBS}$	$\widehat{\pi(\gamma\|y)}^{PMC}$
0,1,3,5,6,12,18,20	0.1893	0.1822	0.1891
0,1,3,5,6,18,20	0.0588	0.0598	0.0596
0,1,3,5,6,9,12,18,20	0.0223	0.0236	0.0335
0,1,3,5,6,12,14,18,20	0.0220	0.0193	0.0248
0,1,2,3,5,6,12,18,20	0.0216	0.0222	0.0212
0,1,3,5,6,7,12,18,20	0.0212	0.0233	0.0282
0,1,3,5,6,10,12,18,20	0.0199	0.0222	0.0129
0,1,3,4,5,6,12,18,20	0.0197	0.0182	0.0200
0,1,3,5,6,12,15,18,20	0.0196	0.0196	0.0168
0,1,3,5,6,8,12,18,20	0.0193	0.0197	0.0142

Gibbs ($T = 100,000$ and $T_0 = 10,000$) and PMC ($N = 10,000$, $T = 10$ and $D = 20$) results for $\tilde{\beta} = 0_{21}$ and $c = 100$

Extension

When models $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are not embedded, difficult choice of $\mathfrak{M}_1$ versus $\mathfrak{M}_2$ in above principle.

Idea of an iterative prior determination by successive replacements of π_1 and π_2 by their respective compatible priors...

Should get to the two sets of hyperparameters closest to one another.

3 Symmetrised compatible priors

[Joint work with J.A. Cano and D. Salmerón]

Postulate

Previous principle requires embedded models (or an encompassing model) and proper priors, while being hard to implement outside exponential families
Now we determine prior measures on two models $\mathfrak{M}_1$ and $\mathfrak{M}_2$, π_1 and π_2, directly by a compatibility principle.

Generalised expected posterior priors

[Perez & Berger, 2000]

EPP Principle

Starting from reference priors π_1^N and π_2^N, substitute by prior distributions π_1 and π_2 that solve the system of integral equations

$$\pi_1(\theta_1) = \int_{\mathscr{X}} \pi_1^N(\theta_1 \,|\, x) m_2(x) \mathrm{d}x$$

and

$$\pi_2(\theta_2) = \int_{\mathscr{X}} \pi_2^N(\theta_2 \,|\, x) m_1(x) \mathrm{d}x,$$

where x is an imaginary minimal training sample and m_1, m_2 are the marginals associated with π_1 and π_2 respectively.

Motivation

Eliminates the "imaginary observation" device and proper-isation through part of the data by integration under the "truth"

Assumes that both models are *equally* valid and equipped with ideal unknown priors

$$\pi_i, \quad i = 1, 2,$$

that yield "true" marginals balancing each model wrt the other

For a *given* π_1, π_2 is an **expected posterior prior**
Using both equations introduces symmetry into the game

Dual properness

Theorem (Proper distributions)

If π_1 is a probability density then π_2 solution to

$$\pi_2(\theta_2) = \int_{\mathscr{X}} \pi_2^N(\theta_2 \mid x) m_1(x) dx$$

is a probability density

ⓒ Both EPPs are either proper or improper.

Bayesian coherence

Theorem (True Bayes factor)

If π_1 and π_2 are the EPPs and if their marginals are finite, then the corresponding Bayes factor

$$B_{1,2}(\mathbf{x})$$

is either a (true) Bayes factor or a limit of (true) Bayes factors.

Obviously only interesting when both π_1 and π_2 are improper.

Existence/Unicity

Theorem (Recurrence condition)

When both the observations and the parameters in both models are continuous, if the Markov chain with transition

$$Q(\theta_1' \mid \theta_1) = \int g(\theta_1, \theta_1', \theta_2, x, x') \, dx dx' d\theta_2$$

where

$$g(\theta_1, \theta_1', \theta_2, x, x') = \pi_1^N(\theta_1' \mid x) \, f_2(x \mid \theta_2) \, \pi_2^N(\theta_2 \mid x') \, f_1(x' \mid \theta_1),$$

is recurrent, *then there exists a solution to the integral equations, unique up to a multiplicative constant.*

Consequences

- If the M chain is positive recurrent, there exists a unique pair of proper EPPS.
- The transition density $Q(\theta_1' \mid \theta_1)$ has a dual transition density on Θ_2.
- There exists a parallel M chain on Θ_2 with identical properties; if one is (Harris) recurrent, so is the other.
- **Duality property** found both in the MCMC literature and in decision theory

  [Diebolt & Robert, 1992; Eaton, 1992]
- When Harris recurrence holds but the EPPs cannot be found, the Bayes factor can be approximated by MCMC simulation

Point null hypothesis testing

Testing $H_0 : \theta = \theta^*$ versus $H_1 : \theta \neq \theta^*$, i.e.

$$\begin{aligned} \mathfrak{M}_1 &: f(x\,|\,\theta^*)\,, \\ \mathfrak{M}_2 &: f(x\,|\,\theta)\,, \theta \in \Theta. \end{aligned}$$

Default priors

$$\pi_1^N(\theta) = \delta_{\theta^*}(\theta) \text{ and } \pi_2^N(\theta) = \pi^N(\theta)$$

For x minimal training sample, consider the proper priors

$$\pi_1(\theta) = \delta_{\theta^*}(\theta) \text{ and } \pi_2(\theta) = \int \pi^N(\theta\,|\,x)\, f(x\,|\,\theta^*)\, dx$$

Point null hypothesis testing (cont'd)

Then

$$\int \pi_1^N(\theta\,|\,x)\, m_2(x)\, dx = \delta_{\theta^*}(\theta) \int m_2(x)\, dx = \delta_{\theta^*}(\theta) = \pi_1(\theta)$$

and

$$\int \pi_2^N(\theta\,|\,x)\, m_1(x)\, dx = \int \pi^N(\theta\,|\,x)\, f(x\,|\,\theta^*)\, dx = \pi_2(\theta)$$

© **$\pi_1(\theta)$ and $\pi_2(\theta)$ are integral priors**

Note

Uniqueness of the Bayes factor
Integral priors and intrinsic priors coincide
[Moreno, Bertolino and Racugno, 1998]

Location models

Two location models

$$\begin{aligned} \mathfrak{M}_1 &: f_1(x\,|\,\theta_1) = f_1(x - \theta_1) \\ \mathfrak{M}_2 &: f_2(x\,|\,\theta_2) = f_2(x - \theta_2) \end{aligned}$$

Default priors

$$\pi_i^N(\theta_i) = c_i, \quad i = 1, 2$$

with minimal training sample size **one**
Marginal densities

$$m_i^N(x) = c_i, \quad i = 1, 2$$

Location models (cont'd)

In that case, $\pi_1^N(\theta_1)$ and $\pi_2^N(\theta_2)$ are integral priors **when $c_1 = c_2$**:

$$\begin{aligned} \int \pi_1^N(\theta_1\,|\,x)\, m_2^N(x)\, dx &= \int c_2 f_1(x - \theta_1)\, dx = c_2 \\ \int \pi_2^N(\theta_2\,|\,x)\, m_1^N(x)\, dx &= \int c_1 f_2(x - \theta_2)\, dx = c_1. \end{aligned}$$

© If the associated Markov chain is recurrent,

$$\pi_1^N(\theta_1) = \pi_2^N(\theta_2) = c$$

are the unique integral priors and they are intrinsic priors
[Cano, Kessler & Moreno, 2004]

Location models (cont'd)

Example (Normal versus double exponential)

$$\mathfrak{M}_1 \; : \; \mathcal{N}(\theta, 1), \quad \pi_1^N(\theta) = c_1,$$
$$\mathfrak{M}_2 \; : \; \mathcal{DE}(\lambda, 1), \quad \pi_2^N(\lambda) = c_2.$$

Minimal training sample size one and posterior densities

$$\pi_1^N(\theta \,|\, x) = \mathcal{N}(x, 1) \text{ and } \pi_2^N(\lambda \,|\, x) = \mathcal{DE}(x, 1)$$

Location models (cont'd)

Example (Normal versus double exponential (2))

Transition $\theta \to \theta'$ of the Markov chain made of steps :

1. $x' = \theta + \varepsilon_1$, $\varepsilon_1 \sim \mathcal{N}(0, 1)$
2. $\lambda = x' + \varepsilon_2$, $\varepsilon_2 \sim \mathcal{DE}(0, 1)$
3. $x = \lambda + \varepsilon_3$, $\varepsilon_3 \sim \mathcal{DE}(0, 1)$
4. $\theta' = x + \varepsilon_4$, $\varepsilon_4 \sim \mathcal{N}(0, 1)$

$$\text{i.e.} \qquad \theta' = \theta + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4$$

random walk in θ with finite second moment, null recurrent
© **Resulting Lebesgue measures $\pi_1(\theta) = 1 = \pi_2(\lambda)$ invariant and unique solutions to integral equations**