

Bayesian and frequentist inference

Nancy Reid

March 26, 2007



Don Fraser, Ana-Maria Staicu

Overview

Methods of inference

Asymptotic theory

Approximate posteriors
matching priors

Examples

Logistic regression
normal circle

Constructing priors

Conclusions

A Basic Structure

- data $y = (y_1, \dots, y_n)$ R^n
- model $f(y; \theta)$ probability density
- parameters $\theta = (\theta_1, \dots, \theta_k)$ R^k
- inference about θ (or components of θ)
- an interval of values of θ consistent with the data
- a probability distribution for θ , conditional on the data

JAMA[®]

The Journal of the American Medical Association

ORIGINAL CONTRIBUTION

Comparison of the Atkins, Zone, Ornish, and LEARN Diets for Change in Weight and Related Risk Factors Among Overweight Premenopausal Women

The A TO Z Weight Loss Study: A Randomized Trial

Christopher D. Gardner, PhD

Alexandre Kiazand, MD

Sofiya Alhassan, PhD

Soowon Kim, PhD

Randall S. Stafford, MD, PhD

Raymond R. Balise, PhD

Helena C. Kraemer, PhD

Context Popular diets, particularly those low in carbohydrates, have challenged current recommendations advising a low-fat, high-carbohydrate diet for weight loss. Potential benefits and risks have not been tested adequately.

Objective To compare 4 weight-loss diets representing a spectrum of low to high carbohydrate intake for effects on weight loss and related metabolic variables.

Design, Setting, and Participants Twelve-month randomized trial conducted in the United States from February 2003 to October 2005 among 311 free-living, overweight/obese (body mass index, 27-40) nondiabetic, premenopausal women.

- “The average 12 month weight loss on the Atkins diet was 10.3 lbs, compared to 3.5, 5.7 and 4.8 lbs on the Zone, Learn and Ornish diets, respectively.”¹
- “mean 12 month weight loss on the Atkins diet was 4.7 kg (95% confidence interval 3.1 to 6.3 kg)”
- “mean 12 month weight loss was significantly different between the Atkins and the Zone diets ($p < 0.05$)”
- The probability that the average weight loss on the Atkins diet is greater than that on the Zone diet is 0.983

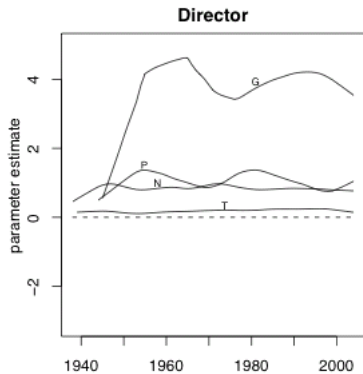
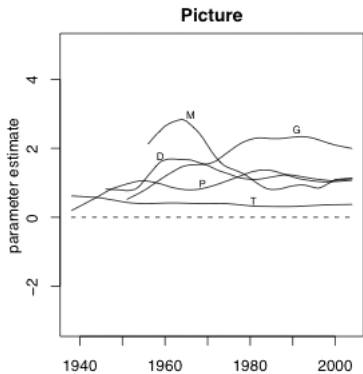
¹ JAMA, Mar 7

Iain's 2006 Oscar Predictions

The following table provides predictions for the four main categories for the 2006 Academy Awards. The nominees are presented in order of expected probability to win, with the probabilities (in percentages) given in the final column.

In a nutshell, the predictions are based on a statistical analysis of past Oscar history. In particular, I identify the influence of factors associated with past winners (such as other Oscar nominations, total number of Oscar nominations, previous Oscar nominations/wins, Golden Globe wins, Guild wins). I then calculate a score for each current nominee based on these factors - nominees with a higher score have more chance of winning (see the notes column in the table). For further details of how the predictions were made click [here](#).

Category	Nominee	Probability (%)	Notes
Picture	Babel	69.5	Although The Departed seems to be the bookmaker's favorite, Babel gets a big score for its Golden Globe (drama) win and having the most overall nominations (seven) of the five Best Picture contenders. Interestingly, the model gives The Departed and The Queen roughly equal chances of winning, with Little Miss Sunshine not far behind with its Producer's Guild win. Overall, this category seems the hardest to call of the four major races this year.
	The Departed	9.7	
	The Queen	9.5	
	Little Miss Sunshine	7.4	
	Letters from Iwo Jima	3.9	
Director	Martin Scorsese for The Departed	97.0	Martin Scorsese benefits from his Director's Guild win and from five previous director nominations, making him a clear favorite here. First-time nominee, Alejandro González Iñárritu for Babel
	Alejandro González Iñárritu for Babel	1.2	
	Clint Eastwood for Letters from Iwo Jima	0.9	



Bayesian inference

- $\pi(\theta | y) \propto f(y; \theta)\pi(\theta)$
- $\pi(\theta | y) = \frac{f(y; \theta)\pi(\theta)}{\int f(y; \theta)\pi(\theta)d\theta}$
- parameter of interest $\psi(\theta)$
- $\pi_m(\psi | y) = \int_{\{\psi(\theta)=\psi\}} \pi(\theta | y)d\theta$
- integrals often computed using simulation methods
- marginal posterior density summarizes all information about ψ available from the data

nonBayesian inference

- find a function of the data $t(y)$, say, with
 1. distribution depending only on ψ
 2. distribution that is known
- usually based on the log-likelihood function
$$\ell(\theta; y) = \log f(y; \theta) \quad \hat{\theta} \equiv \arg \sup \ell(\theta; y)$$
$$\hat{\theta}_\psi \equiv \arg \sup_{\psi(\theta)=\psi} \ell(\theta; y)$$
- likelihood ratio statistic: $2\{\ell(\hat{\theta}; y) - \ell(\hat{\theta}_\psi; y)\} \xrightarrow{d} \chi_{d_1}^2$
- Wald statistic $\hat{\Sigma}^{-1/2}(\hat{\psi} - \psi) \xrightarrow{d} N(0, I)$
- if ψ is scalar: $r^*(\psi) \sim N(0, 1)$
- statistic is derived from the model (via the likelihood function)
- distribution comes from asymptotic theory
- likelihood is not (usually) integrated but is (usually) maximized

... Bayesian inference

- well defined basis for inference
- internally consistent
- leads to optimal results from one point of view
- requires a probability distribution for θ (a **prior** distribution)
- not necessarily calibrated
- not always clear how much answers depend on the choice of prior
- requires (high-) dimensional integration, or simulation, or **approximation**
- useful in applications

... nonBayesian inference

- properties well understood
- calibrated: properties correspond to long-run frequency
- choice of function $t(y)$ may be non-optimal
- approximation to distribution of $t(Y)$ may be poor
- approximation to distribution of $t(Y)$ may be excellent
- useful in applications

- Fisher (1920s, frequentist): “I know only one case in mathematics of a doctrine which has been accepted and developed by the most eminent men of their time, and is now perhaps accepted by men now living, which at the same time has appeared to a succession of sound writers to be fundamentally false and devoid of foundation”
- Lindley (1950s, Bayesian): “Every statistician would be a Bayesian if he took the trouble to read the literature thoroughly and was honest enough to admit that he might have been wrong”

Approximate posteriors

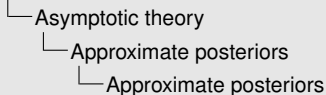
$$\int^{\psi} \pi_m(\psi | y) d\psi = \Phi(r_B^*) \{1 + O(n^{-3/2})\}$$

$$r_B^* = r + \frac{1}{r} \log \frac{q_B}{r}$$

$$r = [2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\}]^{1/2}$$

$$q_B = -\ell'_p(\psi) j_p^{-1/2}(\hat{\psi}) \frac{|j_{\lambda\lambda}(\hat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}} \frac{\pi(\hat{\theta}_\psi)}{\pi(\hat{\theta})}$$

$$\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi) = \ell(\hat{\theta}_\psi)$$



$$\int^{\infty} \pi_{\alpha}(\psi | y) d\psi = \mathbf{e}(\hat{r}_B) \{1 + O(n^{-3/2})\}$$

$$\hat{r}_B = r + \frac{1}{r} \log \frac{qr}{r}$$

$$r = [2\{f_p(\hat{\psi}) - f_p(\psi)\}]^{1/2}$$

$$q_B = -f_p''(\psi)^{-1/2}(\psi) \frac{|j_{\lambda\lambda}(\hat{\theta}_{\psi})|^{1/2} \pi(\hat{\theta}_{\psi})}{|j_{\lambda\lambda}(\hat{\theta}_{\psi})|^{1/2} \pi(\hat{\theta})}$$

$$f_p(\psi) = f(\psi, \hat{\lambda}_{\psi}) = f(\hat{\theta}_{\psi})$$

marginal posterior density:

$$\begin{aligned} \pi_m(\psi | y) &= \int_{\{\psi(\theta)=\psi\}} \exp\{\ell(\theta; y)\} \pi(\theta) d\theta / \int \exp\{\ell(\theta; y)\} \pi(\theta) d\theta \\ &= c \frac{\exp\{\ell(\hat{\theta}_{\psi})\} |j_{\lambda\lambda}(\hat{\theta}_{\psi})|^{-1/2} \pi(\hat{\theta}_{\psi})}{\exp\{\ell(\hat{\theta})\} |j(\hat{\theta})|^{-1/2} \pi(\hat{\theta})} \{1 + O(n^{-3/2})\} \\ &\doteq c \exp\{\ell(\hat{\theta}_{\psi}) - \ell(\hat{\theta})\} \frac{|j_{\lambda\lambda}(\hat{\theta}_{\psi})|^{-1/2}}{j_p(\hat{\psi})^{-1/2} |j_{\lambda\lambda}(\hat{\theta}_{\psi})|^{-1/2}} \frac{\pi(\hat{\theta}_{\psi})}{\pi(\hat{\theta})} \\ \pi_m(r | y) &= c \exp(-\frac{1}{2} r^2) \frac{r}{q_B} \\ &= c \exp(-\frac{1}{2} r^2 - \log \frac{r}{q_B}) \end{aligned}$$

... approximation to the posterior is really good

- Example:

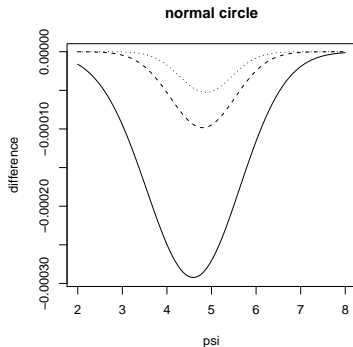
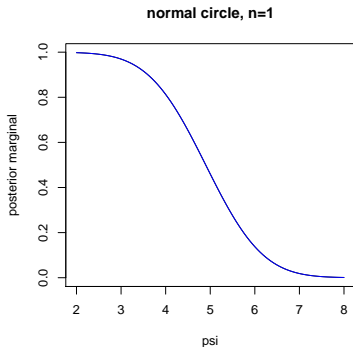
$$y_i \sim N(\theta, I), \quad y_i = (y_{i1}, \dots, y_{ik}), \quad \theta = (\theta_1, \dots, \theta_k)$$

- $\psi = \sqrt{(\theta_1^2 + \dots + \theta_k^2)}$
- flat prior $\pi(\theta)d\theta \propto d\theta$

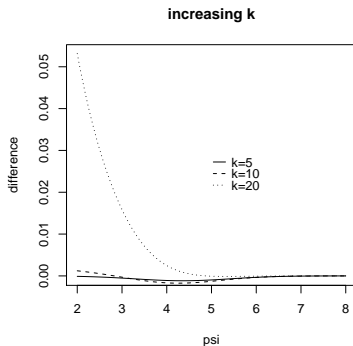
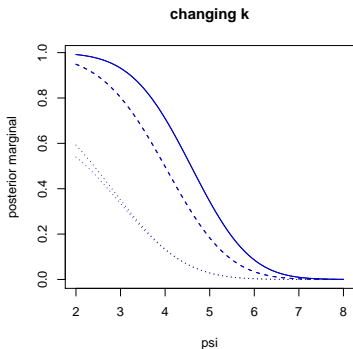
- posterior is proper, in fact is $\chi_k^{2'}(n\|y\|^2)$

- $r_B^* = \sqrt{n}(\hat{\psi} - \psi) + \frac{1}{\sqrt{n}(\hat{\psi} - \psi)} \log \left\{ \left(\frac{\psi}{\hat{\psi}} \right)^{(k-1)/2} \right\}$

... normal circle, $k=2$



... normal circle, increasing k



Are Bayesian and nonBayesian solutions really that different?

- posterior probability limit
- $\Pr_{\theta|Y}\{\theta \leq \theta^{(1-\alpha)}|y\} = 1 - \alpha$
- upper confidence bound
- $\Pr_{Y|\theta}\{\theta \leq \theta^{(1-\alpha)}(\pi, Y)\} = 1 - \alpha$
- first depends on the prior, second is evaluated under the model
- is there a prior for which both limits are equal?
- no², but there might be a prior for which they are approximately equal:

$$\Pr_{Y|\theta}\{\theta \leq \theta^{(1-\alpha)}(\pi, Y)\} = 1 - \alpha + O(1/n)$$
- matching prior

²except in a location model

... matching priors

- if $\theta \in R$, then $\pi(\theta) \propto i^{1/2}(\theta)$ (Welch & Peers, 1963)
- if $\theta = (\psi, \underline{\lambda})$ then

$$\pi(\theta) \propto i_{\psi\psi}^{1/2}(\theta)g(\lambda)$$

(Peers, 1965; Tibshirani, 1989)

- where we reparametrize if necessary so that $i_{\psi\lambda}(\theta) = 0$
- matching priors are one version of **objective** priors
- they depend on the model
- reference priors of Berger and Bernardo are a different version of objective priors
- with a slightly different goal: maximize the ‘distance’ from prior to posterior
- good news: Bayesian inference with matching priors is calibrated
- bad news: there is a whole family of such priors, even in relatively simple situations

... good news: matching priors are unique

when used in the r^* approximation (DiCiccio & Martin 93, Staicu 07)

$$\int^{\psi} \pi_m(\psi | y) d\psi = \Phi(r_B^*) \{1 + O(n^{-3/2})\}$$

$$r_B^* = r + \frac{1}{r} \log \frac{q_B}{r}$$

$$r = [2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\}]^{1/2}$$

$$q_B = -\ell'_p(\psi) j_p^{-1/2}(\hat{\psi}) \frac{|j_{\lambda\lambda}(\hat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}} \frac{\pi(\psi, \hat{\lambda}_\psi)}{\pi(\hat{\psi}, \hat{\lambda})}$$

$$= \quad \quad \quad \frac{i_{\psi\psi}^{1/2}(\psi, \hat{\lambda}_\psi) g(\hat{\lambda}_\psi)}{i_{\psi\psi}^{1/2}(\hat{\psi}, \hat{\lambda}) g(\hat{\lambda})}$$

when ψ orthogonal to λ , $\hat{\lambda}_\psi = \hat{\lambda} \{1 + O_p(n^{-1})\}$

Logistic regression

The first ten out of 79 sets of observations on the physical characteristics of urine. Presence/absence of calcium oxalate crystals is indicated by 1/0. Two cases with missing values.³

Case	Crystals	Specific gravity	pH	Osmolarity	Conductivity	Urea	Calcium
1	0	1.021	4.91	725	—	443	2.45
2	0	1.017	5.74	577	20.0	296	4.49
3	0	1.008	7.20	321	14.9	101	2.36
4	0	1.011	5.51	408	12.6	224	2.15
5	0	1.005	6.52	187	7.5	91	1.16
6	0	1.020	5.27	668	25.3	252	3.34
7	0	1.012	5.62	461	17.4	195	1.40
8	0	1.029	5.67	1107	35.9	550	8.48
9	0	1.015	5.41	543	21.9	170	1.16
10	0	1.021	6.13	779	25.7	382	2.21
⋮							⋮
⋮							⋮

³Andrews & Herzberg, 1985

... logistic regression

Model: Independent binary responses Y_1, \dots, Y_n with

$$\Pr(Y_i = 1) = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}$$

linear exponential family:

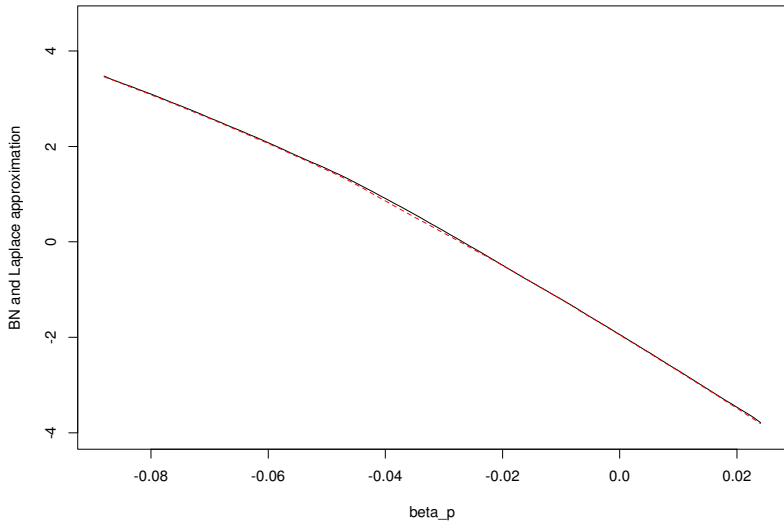
$$\ell(\beta; y) = \exp\{\beta_0 \sum y_i + \beta_1 \sum x_{1i} y_i + \dots + \beta_6 \sum x_{6i} y_i - c(\beta) - d(y)\}$$

parameter of interest $\psi = \beta_6$, say

orthogonal nuisance parameter $\lambda_j = E(yx_j)$

$\hat{\lambda}_\psi \equiv \hat{\lambda}$ (special to exponential families)

$\pi(\psi, \lambda) \propto i_{\psi\psi}^{1/2}(\psi, \lambda)$, unique



... logistic regression

Confidence/posterior interval for parameter of interest

method	lower bound	upper bound	p -value for 0
$\Phi(r^*)$	-0.058	0.00029	0.052
$\Phi(r_B^*)$	-0.058	0.00028	0.052
1st order	-0.063	-0.00062	0.047

First line uses a (very accurate) approximation to the conditional distribution of $\sum y_i x_{6i}$, given sufficient statistics for nuisance parameters (exponential family)

Fitted using `cond` library of package `hoa`

2007-03-27

Bayesian and frequentist inference

Examples

Logistic regression

... logistic regression

... logistic regression

Confidence/posterior interval for parameter of interest

method	lower bound	upper bound	p-value for 0
$\Phi(r^*)$	-0.058	0.00029	0.052
$\Phi(\hat{r}_0^*)$	-0.058	0.00028	0.052
1st order	-0.063	-0.00062	0.047

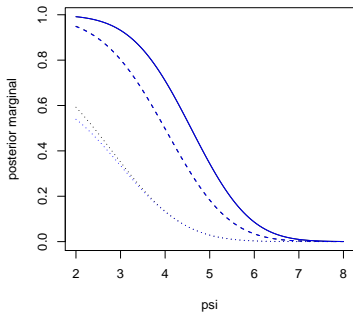
First line uses a (very accurate) approximation to the conditional distribution of $\Sigma_j X_{ij}$, given sufficient statistics for nuisance parameters (exponential family)
Fitted using `cond` library of package `hbook`

$n = 77, d = 7$, but the information in 77 binary observations seems to be comparable to the information in about 10 continuous observations

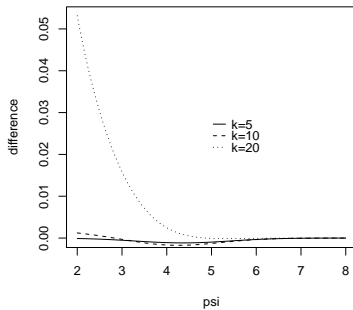
Matching priors need to be targetted on the parameter of interest

- as do all objective priors
- 'flat' priors for vector parameters don't lead to calibrated inferences
- Example: $y_i \sim N(\theta_i, 1), i = 1, \dots, k$
- exact and approximate posterior for $\sum \theta_i^2$ nearly identical

changing k



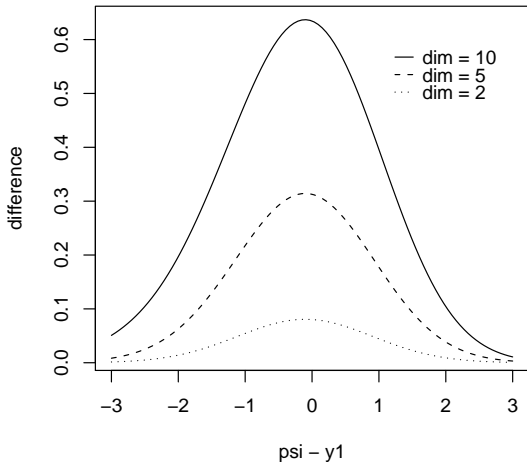
increasing k

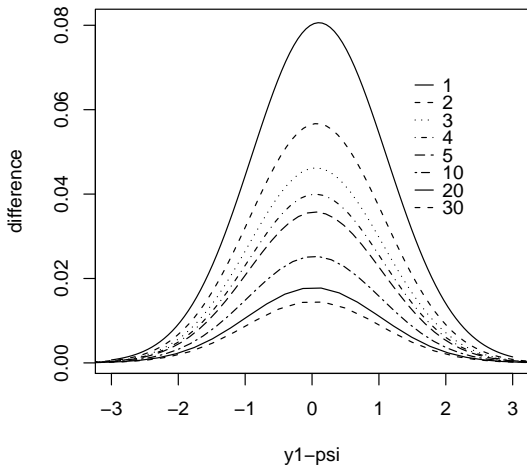


Matching priors need to be targetted on the parameter of interest

- as do all objective priors
- 'flat' priors for vector parameters don't lead to calibrated inferences
- Example: $y_i \sim N(\theta_i, 1), i = 1, \dots, k$
- exact and approximate posterior for $\sum \theta_i^2$ nearly identical
- posterior used prior $\pi(\underline{\theta}) \propto 1$, marginalize to $\psi = \sum \theta_i^2$
- posterior limits are not probability limits
- there is an exact nonBayesian solution, based on marginal density of $\sum y_i^2$ (noncentral χ^2) Dawid, Stone, Zidek, 1974

dependence on dimension



dependence on n 

nonBayesian asymptotics

- find a function of the data with known distribution depending on ψ
- our candidate: $r_F^* = r_F^*(\psi; y)$ with $N(0, 1)$ distribution

$$r_F^* = r + \frac{1}{r} \log \frac{q_F}{r}$$

-
-
- a maximum likelihood type departure
- $\chi(\theta) =$ scalar function of $\varphi(\theta)$
- $\theta \rightarrow \varphi(\theta) = \varphi(\theta; y^0)$: a re-parametrization of the model

... nonBayesian asymptotics

- where does this come from and why does it work
- using φ we can construct an approximate exponential family model
- this is the route to accurate density approximation using saddlepoint-type arguments
- to get $\varphi(\theta)$ differentiate the log-likelihood on the sample space
- $\varphi_j(\theta) = \left. \frac{d}{dt} \ell(\theta; \underline{y}^0 + t\underline{v}_j) \right|_{t=0}$
- jiggle the observed data in directions \underline{v}_j , and record the change in the log likelihood
- gives a way to implement approximate conditioning
- leads to quite accurate frequentist solutions

... does this help with finding priors?

1. strong matching priors (FR, 2002)

- posterior marginal distribution $s(\psi) = \Phi(r_B^*)$
- nonBayesian p -value function $p(\psi) = \Phi(r_F^*)$
-

$$r_F^* = r + \frac{1}{r} \log \frac{q_F}{r}$$

$$r_B^* = r + \frac{1}{r} \log \frac{q_B}{r}$$

- Set $q_F = q_B$, and solve for prior π (if possible)
- guarantees matching to 3rd order
- leads to a data dependent prior
- automatically targetted on parameter of interest
- only a function of the parameter of interest

└ Constructing priors

└ ... does this help with finding priors?

... does this help with finding priors?

- strong matching priors (FR, 2002)
 - posterior marginal distribution $s(v) = \Phi(r_{\psi}^*)$
 - nonBayesian p -value function $p(v) = \Phi(r_{\psi}^*)$

$$r_{\psi}^* = r + \frac{1}{r} \log \frac{q_{\psi}}{r}$$

$$r_{\psi}^* = r + \frac{1}{r} \log \frac{q_{\psi}}{r}$$

- Set $q_{\psi} = q_{\psi}$, and solve for prior π (if possible)
- guarantees matching to 3rd order
- leads to a data dependent prior
- automatically targetted on parameter of interest
- only a function of the parameter of interest

$$q_B = -\ell'_{\psi}(\psi) j_{\psi}^{-1/2}(\hat{\psi}) \quad \text{blue prior}$$

$$q_F = \{\chi(\hat{\theta}) - \chi(\hat{\theta}_{\psi})\} / \hat{\sigma}_{\chi}$$

... does this help

2. approximate location models

- every model $f(y; \theta)$ can be approximated by a location model (locally)
- location models have automatic prior for vector θ : flat
- change at θ that generates an increment $d\hat{\theta}$ at observed data, $W(\theta)$, say

- flat prior is $\pi(\theta) \propto |W(\theta)| = |AV(\theta)| = \left| \frac{d\hat{\theta}}{d\theta} \right|$
- V is a $n \times k$ matrix with i th row

$$V_i(\theta) = - \frac{F_{y;\theta'}(y_i^0; \theta)}{F_y(y_i^0; \theta)}$$

- gives a natural default prior for vector parameters
- still need to deal with marginalization to target parameters of interest

└ Constructing priors

└ ... does this help

... does this help

2. approximate location models

- every model $f(y; \theta)$ can be approximated by a location model (locally)
- location models have automatic prior for vector θ : flat
- change at θ that generates an increment $d\theta$ at observed data, $W(\theta)$, say
- flat prior is $\pi(\theta) \propto |W(\theta)|^{-1} = |AV(\theta)|^{-1} = \frac{d\theta}{|A|}$
- V is a $n \times k$ matrix with i th row

$$V_i(\theta) = -\frac{F_{y_i}(y_i^0; \theta)}{F_y(y^0; \theta)}$$

- gives a natural default prior for vector parameters
- still need to deal with marginalization to target parameters of interest

- approximate location model $f(y - \beta(\theta))$, say
- factors as $f_1(a)f_2(\hat{\theta} - \beta(\theta))$
- if $f_2'(\cdot) = 0$ then

•

$$d\hat{\theta} = \beta_{\theta}(\theta)d\theta$$

- flat prior is $|\beta_{\theta}(\theta)|$

Summarizing

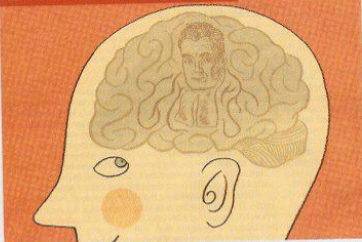
- asymptotic theory is useful for obtaining (very) good approximations to distributions
- can be used in Bayesian or nonBayesian contexts
- can be used to derive posterior intervals that are calibrated
- can be used to simplify the choice of matching priors
- finding priors is not straightforward
- likelihood plus first derivative change in the sample space is ‘intrinsic’ to inference, at least asymptotically to $O(n^{-3/2})$
- flat priors for a location parameter will not lead to calibrated posteriors for curved parameters

We're all Bayesians really

Economist Jan 14, 2006

70 Science and technology

The Economist, January 7th 2006



Psychology

Bayes rules

A once-neglected statistical technique may help to explain how the mind works

SCIENCE, being a human activity, is not immune to fashion. For example, one of the first mathematicians to study the subject of probability theory was an English clergyman called Thomas Bayes, who was born in 1702 and died in 1761. His ideas about the prediction of future events from one or two examples were popular for a while, and have never been fundamentally challenged. But they were eventually overwhelmed by those of the "frequentist" school, which developed the

they are on the right track, but only recently have they begun to look at whether the brain copes with everyday judgments in the real world in a Bayesian manner. In research to be published later this year in *Psychological Science*, Thomas Griffiths of Brown University in Rhode Island and Joshua Tenenbaum of the Massachusetts Institute of Technology put the idea of a Bayesian brain to a quotidian test. They found that it passes with flying colours.

Prior assumptions

Also in this section

- 71 Fraud and cloning
- 72 Fisheries and conservation areas
- 72 Banning trade in caviar

consequence of simple mathematical equations (or, at least, of equations that mathematicians regard as simple).

With the correct prior, even a single piece of data can be used to make meaningful Bayesian predictions. By contrast frequentists, though they deal with the same probability distributions as Bayesians, make fewer prior assumptions about the distribution that applies in any particular situation. Frequentism is thus a more robust approach, but one that is not well suited to making decisions on the basis of limited information—which is something that people have to do all the time.

Dr Griffiths and Dr Tenenbaum conducted their experiment by giving individual nuggets of information to each of the participants in their study (of which they had, in an ironically frequentist way of doing things, a total of 350), and asking them to draw a general conclusion. For example, many of the participants were told the

... Bayesian

Griffiths and Tenenbaum *Psychological Science*

- asked subjects to make predictions based on limited evidence
- concluded that the predictions were consistent with Bayesian methods of inference
- “the results of our experiment reveal a far closer correspondence between optimal statistical inference and everyday cognition than that suggested by previous research”

... Bayesian

- Goldstein (2006): “we argue first that the subjectivist Bayes approach is the only feasible method for tackling many scientific problems”
- Neal (1998): “Using ”technological” or ”reference” priors chosen solely for convenience, or out of a mis-guided desire for pseudo-objectivity ... [or] using priors that vary with the amount of data that you have collected ... have no real Bayesian justification, and since they are usually offered with no other justification either, I consider them to be highly dubious.”
- Wasserman (2006): “I think it would be hard to defend a sequence of subjective analyses that yield poor frequency coverage”

References

1. `ba.stat.cmu` Volume 2 # 3 Berger 2006, Goldstein 2006, Wasserman 2006
2. DiCiccio and Martin 1993
3. Fraser Reid 2002
4. Little 2006
5. Tibshirani 1989
6. Welch and Peers 1963