Objective Bayesian Inference:

Estimation

James O. Berger

Institute of Statistics and Decisions Sciences

Duke University, USA

1^{st} *ISBASA Workshop, Stellenbosch, South Africa, April 2002*

Outline

- Introduction to Bayesian analysis through a medical diagnosis example.
- Objective and subjective Bayesian analysis - history.
- Objective Bayesian analysis and frequentist statistics: solving the conditioning problem.
- Ease of implementing objective Bayesian analysis.
- Comments for subjective Bayesians.
- Comments for objective Bayesians.

Bayesian Analysis

- Model the data probabilistically.
- Model unknown features of the data-model using probability distributions, called *prior* distributions.
- Use probability theory (often Bayes theorem) to find the probability distribution of quantities of interest, given the data; these are typically called *posterior* distributions.

A Medical Diagnosis Example (with Mossman)

The Medical Problem:

- Within a population, $p_0 = Pr(\text{Disease } D)$.
- A diagnosic test results in either a Positive (P) or Negative (N) reading.
- $p_1 = Pr(P \,|\, \text{patient has } D)$.
- $p_2 = Pr(P \,|\, \text{patient does not have } D)$.

It follows from Bayes theorem that

$$\theta = Pr(D|P) = \frac{p_0 p_1}{p_0 p_1 + (1 - p_0) p_2}.$$

The Statistical Problem: The p_i are unknown. Based on (independent) data $X_i \sim \text{Binomial}(n_i, p_i)$ (arising from medical studies), find a $100(1-\alpha)\%$ confidence set for θ.

Suggested Solution: Assign p_i the Jeffreys-rule prior

$$\pi(p_i) \propto p_i^{-1/2}(1-p_i)^{-1/2}.$$

By Bayes theorem, the posterior distribution of p_i given the data, x_i, is

$$\pi(p_i \mid x_i) = \frac{p_i^{-1/2}(1-p_i)^{-1/2} \times \binom{n}{x_i} p_i^{x_i}(1-p_i)^{n_i - x_i}}{\int p_i^{-1/2}(1-p_i)^{-1/2} \times \binom{n}{x_i} p_i^{x_i}(1-p_i)^{n_i - x_i}\, dp_i},$$

which is the Beta$(x_i + \frac{1}{2}, n_i - x_i + \frac{1}{2})$ distribution.

Finally, compute the desired confidence set (formally, the $100(1-\alpha)\%$ equal-tailed posterior credible set) by

- drawing random p_i from the Beta$(x_i + \frac{1}{2}, n_i - x_i + \frac{1}{2})$ distributions, $i = 0, 1, 2$;
- computing the associated $\theta = p_0 p_1 / [p_0 p_1 + (1 - p_0) p_2]$;
- repeating this process $10,000$ times;
- using the $\frac{\alpha}{2}\%$ upper and lower percentiles of these generated θ to form the desired confidence limits.

$n_0 = n_1 = n_2$	(x_0, x_1, x_2)	95% confidence interval
20	(2,18,2)	(0.107, 0.872)
20	(10,18,0)	(0.857, 1.000)
80	(20,60,20)	(0.346, 0.658)
80	(40,72,8)	(0.808, 0.952)

Table 1: The 95% equal-tailed posterior credible interval for $\theta = p_0p_1/[p_0p_1 + (1 - p_0)p_2]$, for various values of the n_i and x_i.

Subjective and Objective Bayes Analysis

- In *subjective Bayesian* analysis
  - prior distributions represent an individual's beliefs about unknowns;
  - Bayes theorem shows how prior beliefs become posterior beliefs, through learning from data.
- In *objective Bayesian* analysis
  - prior distributions are chosen to represent 'neutral' knowledge about unknowns;
  - the posterior distribution is claimed to represent evidence about unknowns arising from just data.

History of Objective and Subjective Bayes

- *Objective Bayesian* inference, using constant prior densities for unknowns, was prominent from 1775–1925, under the name *inverse probability*.
- By 1940, the prevailing statistical philosophies were either *Fisherian* (associated with Ronald Fisher) or *frequentist* (associated with Jerzy Neyman).
- *Subjective Bayesian* analysis became prominent by 1960; it is often mistakenly equated with Bayesianism
- Harold Jeffreys revived the *objective Bayesian* school from 1940-1970.

Features of Objective Bayesian Analysis

- It is as objective as anything can be in statistics.
- There are standard objective priors and techniques available for a wide array of problems (Jeffreys priors, reference priors, ...).
- Objective Bayesian analysis solves the widespread conditioning problem in statistics (examples).
- Objective Bayesian analysis is often the easiest way to obtain good frequentist procedures (examples).
- Objective Bayesian analysis is great for difficult problems (examples).

The Conditioning Problem

Basic question for a frequentist: What is the sequence of possible data for which to consider frequentist evaluations?

(Fisher: "relevant subset;" Lehmann: "frame of reference.")

Artificial example: Observe X_1 and X_2, where

$$X_i = \begin{cases} \theta + 1 \text{ with probability } 1/2 \\ \theta - 1 \text{ with probability } 1/2. \end{cases}$$

Consider the confidence set for θ

$$C(X_1, X_2) = \begin{cases} \frac{1}{2}(X_1 + X_2) & \text{if } X_1 \neq X_2 \\ X_1 - 1 & \text{if } X_1 = X_2. \end{cases}$$

Unconditional coverage:

$$P_\theta(C(X_1, X_2) \text{ contains } \theta) = 0.75.$$

This is silly: if $x_1 \neq x_2$, we know $C(x_1, x_2) = \theta$;
if $x_1 = x_2$, $C(X_1, X_2)$ equals θ only with probability 1/2.

One must use the ***conditional frequentist*** **approach:**

- Define the conditioning statistic $S = |X_1 - X_2|$, measuring the "strength of evidence" in the data (here ranging from $s = 0$ to $s = 2$);

- Compute frequentist coverage conditional on the strength of evidence S.

$$\begin{aligned} P_\theta(C(X_1, X_2) \text{ contains } \theta \mid s = 2) &= 1 \\ P_\theta(C(X_1, X_2) \text{ contains } \theta \mid s = 0) &= \tfrac{1}{2}. \end{aligned}$$

Note: The correct answer is obtained trivially by objective Bayesian analysis, using a constant prior density for θ.

Genetics Example (A. Kong): Mapping genes for complex traits based on affected half-sib data X.

Goal: A 95% confidence set for

$\theta =$ true location of a susceptibility gene.

Unconditional method: Kruglyak and Lander (1995) found $C^u(X)$ such that $P_\theta(C^u(X)$ contains $\theta) = 0.95$.

Conditional method: Conditioning on the location ancillary statistic S results in a confidence set $C^c(X)$ such that $P_\theta(C^c(X)$ contains $\theta|S) = 0.95$.

Differences can be significant:

$1 - \alpha^u(s) = P_\theta(C^u(X)$ contains $\theta|S)$ can be near zero for some s. Furthermore, $Pr(1 - \alpha_u(S) < 0.90) = 0.11$.

Using objective Bayesian analysis to easily obtain good frequentist and good conditional performance

- Bayesian analysis is the only known general method of obtaining good conditional performance.
- So simply find (objective) priors that yield Bayes procedures with good frequentist properties.
  - the constant prior often works well
  - the *Jeffreys-rule* prior is optimal in one-dimension
  - large literature on *probability matching* priors
  - *reference priors* are my favorite.

Diagnosis Example (continued): One can evaluate the frequentist properties of the objective Bayesian confidence sets. In particular, we look at the percentage of the time that the 95% equal-tailed sets miss on the left and on the right (ideal would be 2.5% each) for the indicated parameter values and when $n_0 = n_1 = n_2 = 20$.

(p_0, p_1, p_2)	O-Bayes	Log Odds	Gart-Nam	Delta
$(\frac{1}{4}, \frac{3}{4}, \frac{1}{4})$	2.86,2.71	1.53,1.55	2.77,2.57	2.68,2.45
$(\frac{1}{10}, \frac{9}{10}, \frac{1}{10})$	2.23,2.47	0.17,0.03	1.58,2.14	0.83,0.41
$(\frac{1}{2}, \frac{9}{10}, \frac{1}{10})$	2.81,2.40	0.04,4.40	2.40,2.12	1.25,1.91

My View on Foundations

- Fully subjective Bayesian analysis is philosophically compelling, but is very difficult in complex models. Also, the appearance of objectivity is often required.
- Frequentist reasoning has important roles in objective Bayesian analysis; indeed, we know that 'we have it right' when the two agree.
- Objective Bayesian analysis is much easier to implement than frequentist inference on (most) complex problems; even non-optimal objective priors (e.g., a constant) usually work fine.

A Multiple Comparisons Example

- The probability model and analysis goal for the data:
  - Suppose, for $i = 1, \ldots, m$, that observation X_i is normally distributed with mean μ_i and variance 1, to be denoted $X_i \sim N(\mu_i, 1)$.
  - Most of the μ_i are thought likely to be zero, and it is desired to detect those that are nonzero.

- The prior distribution for the unknown μ_i:
  - Let p denote the *unknown* prior probability that a given μ_i is zero; thus $1-p$ is the prior probability that it is nonzero:
    - assign p the uniform density on the interval $(0,1)$.
  - Assume that the nonzero μ_i follow a $N(0,V)$ distribution, with V unknown:
    - assign V the prior density $\pi(V) = 1/(1+V)^2$.

- Bayes theorem allows computation of the posterior distribution of unknown quantities given the data.
  - Of primary interest are the posterior probabilities that the μ_j are nonzero, given by

$$p_j = 1 - \frac{\int_0^1 \int_0^1 p \prod_{i \neq j} \left(p + (1-p)\sqrt{1-w}\, e^{w X_i{}^2/2)}\right) dp dw}{\int_0^1 \int_0^1 \prod_{i=1}^{m} \left(p + (1-p)\sqrt{1-w}\, e^{w X_i{}^2/2)}\right) dp dw}.$$

  - $(p_1, p_2, \ldots, p_m)$ can be computed numerically if m is moderate. For large m, it is most efficient to do the computation via importance sampling, with a common importance sample for all p_j.

- Note that, in the Bayesian approach, the 'penalty' for adding additional (mean zero) comparisons is that p will concentrate on smaller values. This is automatic, and one need not do any adjustments (e.g. Bonferoni).

Example: Draw ten $N(0, 4^2)$ 'signal' observations:

-8.48, -5.43, -4.81, -2.64, -2.40, 3.32, 4.07, 4.81, 5.81, 6.24

Generate $n = 10, 50, 500,$ and 5000 $N(0, 1)$ 'noise' observations.

n	Central seven 'signal' observations -5.4	-4.8	-2.6	-2.4	3.3	4.1	4.81	#noise $p_i > .6$
10	1	1	.94	.89	.99	1	1	1
50	1	1	.71	.59	.94	1	1	0
500	1	1	.26	.17	.67	.96	1	2
5000	1.0	.98	.03	.02	.16	.67	.98	1

Table 2: The posterior probabilities that the central 'signal' means are nonzero (the others always had $p_i = 1$).

Note: The penalty for multiple comparisons is automatic.

Comments for Subjective Bayesians

- An objective Bayesian analysis can provide a reference for the effect of the prior in a subjective analysis.
- It is rare that subjective elicitation can be thoroughly done for all unknowns in a problem, so that some utilization of objective priors is almost inevitable.
- An objective Bayesian analysis can be run initially, to assess if subjective priors are even needed. (Perhaps the data will 'swamp the prior'.)

- Through study of objective priors, one can obtain insight into possibly bad behavior of standard (e.g., conjugate) subjective priors.

Example: Covariance matrices

Consider the multivariate normal model $\boldsymbol{X} \sim N_k(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. A commonly used subjective prior for $\boldsymbol{\Sigma}$ is the inverse Wishart prior (for subjectively specified a and b)

$$\pi(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-a/2} \exp(-\frac{1}{2}\mathrm{tr}[b\boldsymbol{\Sigma}^{-1}]).$$

Write $\mathbf{\Sigma} = \boldsymbol{H}^t\boldsymbol{D}\boldsymbol{H}$, where $\boldsymbol{H}$ is orthogonal and $\boldsymbol{D}$ is diagonal with entries $d_1 > d_2 > \cdots > d_k$. Change of variables yields

$$\pi(\mathbf{\Sigma})\, d\mathbf{\Sigma} = \frac{1}{|\boldsymbol{D}|^{a/2}} e^{(-\frac{1}{2}\mathsf{tr}[b\boldsymbol{D}^{-1}])} \prod_{i<j}(d_i - d_j) \cdot I_{[d_1 > \cdots > d_k]}\, d\boldsymbol{D} d\boldsymbol{H}.$$

Thus, unexpectedly, one is using a prior distribution that very strongly forces the eigenvalues of the covariance matrix **apart**, contrary to typical prior beliefs.

Objective Bayes: The reference prior (Yang and Berger, 1993) is

$$\pi^R(\boldsymbol{D}, \boldsymbol{H}) = |\boldsymbol{D}|^{-1} d\boldsymbol{D} d\boldsymbol{H},$$

which has no such problem (and gives much better statistical inferences).

The Point: Proper priors that are commonly used by subjectivists can have hidden undesirable features; objective Bayesian tools can expose these features and allow for development of better subjective priors.

Comments for Objective Bayesians

- Use of improper objective priors is better than use of 'vague proper priors.'
  - Improprieties in the posterior can be detected. (In contrast, a vague proper prior will appear okay in such a situation, but will actually yield an arbitrary answer.)
  - If model selection or mixtures are involved, vague proper priors are typically terrible.
  - (Improper) objective priors (e.g., *reference priors*) often have excellent properties.

- Model selection must be treated very carefully.
- Use of data dependent priors needs formal study, as a part of objective Bayesian analysis.
- Posterior propriety is a serious concern, but is mitigated if reference priors are used.

A Spatial Example (Berger, de Oliveira and Sansó, 2001)

$Z(\mathbf{s})$ is a random process observed at $\mathbf{s}_i \in \Re^k, i = 1, \ldots, n$,

$$Z(\mathbf{s}) = \sum_{j=1}^{p} \beta_j f_j(\mathbf{s}) + \varepsilon(\mathbf{s})$$

where $f_j(\mathbf{s})$ are known regression functions, β_j are unknown coefficients and ε is a Gaussian process with zero mean and

$$cov(\varepsilon(\mathbf{s}), \varepsilon(\mathbf{u})) = cov(Z(\mathbf{s}), Z(\mathbf{u})) = \sigma^2 K_\vartheta(\|\mathbf{s} - \mathbf{u}\|)$$

Example: $K_\vartheta(\|\mathbf{s} - \mathbf{u}\|) = e^{-\|\mathbf{s}-\mathbf{u}\|/\vartheta}, \ \ \vartheta > 0.$

Common Objective Priors

Natural objective priors are of the form

$$\pi(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\vartheta}) = \frac{\pi(\boldsymbol{\vartheta})}{\sigma^2}.$$

How should $\pi(\boldsymbol{\vartheta})$ be chosen?

- Subjective information is rarely available.
- Uniform prior: $\pi(\boldsymbol{\vartheta}) = 1$.
  Problem: The resulting posterior is improper.
- Other common objective priors can also fail to give a proper posterior.

The Reference Prior

$$\pi^R(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\vartheta}) = \frac{1}{\sigma^2} [\sum_{i=1}^{n-p} (w_i(\boldsymbol{\vartheta}) - \overline{w}(\boldsymbol{\vartheta}))^2]^{\frac{1}{2}},$$

where

- $w_1, \ldots, w_{n-p}$ are eigenvalues of W_ϑ, with $\overline{w}$ their average;
- $W_\vartheta = \frac{\partial}{\partial\vartheta}\Sigma_\vartheta^{-1}\Sigma_\vartheta^{-1}(I - X(X'\Sigma_\vartheta^{-1}X)^{-1}X'\Sigma_\vartheta^{-1})$;
- X $(n \times p)$ is defined by $X_{ij} = f_j(\mathbf{s}_i)$;
- Σ_ϑ $(n \times n)$ is defined by $\Sigma_{\vartheta,ij} = K_\vartheta(\|\mathbf{s}_i - \mathbf{s}_j\|)$.

Properties and comments:

- The reference prior always yields a proper posterior for suitable $K_\vartheta(\|\mathbf{s} - \mathbf{u}\|)$.
- We are currently studying the frequentist performance of posterior credible sets from this prior.

Example: Consider the spatial data set from Davis (1973), consisting of 52 topographic elevations (in feet above sea level) over a small area on the side of a hill.

We model the unknown mean function via a second order polynomial in the spatial coordinates $\mathbf{s} = (x, y)$,

$$E\{Z(\mathbf{s})\} = \beta_1 + \beta_2 x + \beta_3 y + \beta_4 x^2 + \beta_5 xy + \beta_6 y^2.$$

Three families of isotropic covariance functions are entertained: the Matérn families

$$K^M_{\boldsymbol{\vartheta}}(d) = (2^{\theta_2 - 1}\Gamma(\theta_2))^{-1}(d/\theta_1)^{-\theta_2}\mathcal{K}_{\theta_2}(d/\theta_1)$$

(where $\mathcal{K}_{\theta_2}(.)$ is the modified Bessel function of second kind and order θ_2) with θ_2 equal to $\frac{i}{2}$, $i = 1, 2, 3$.

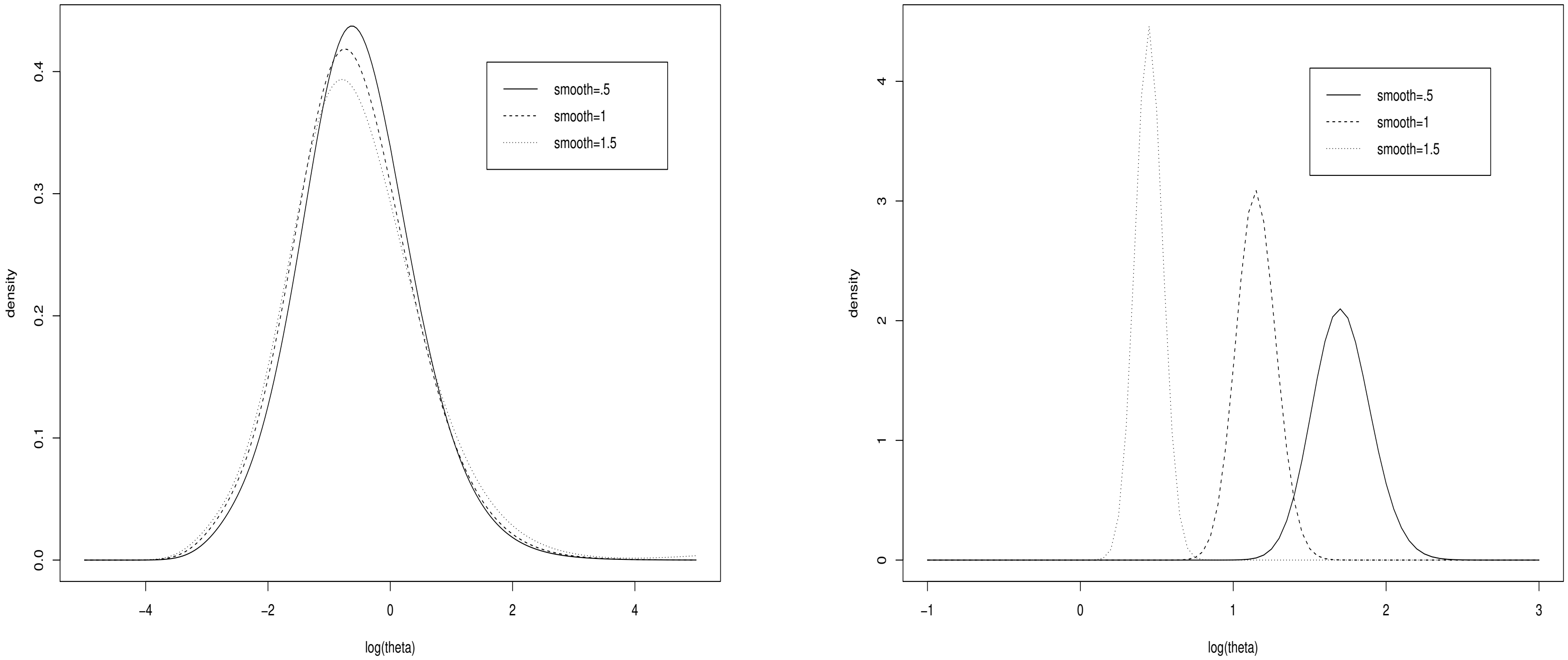
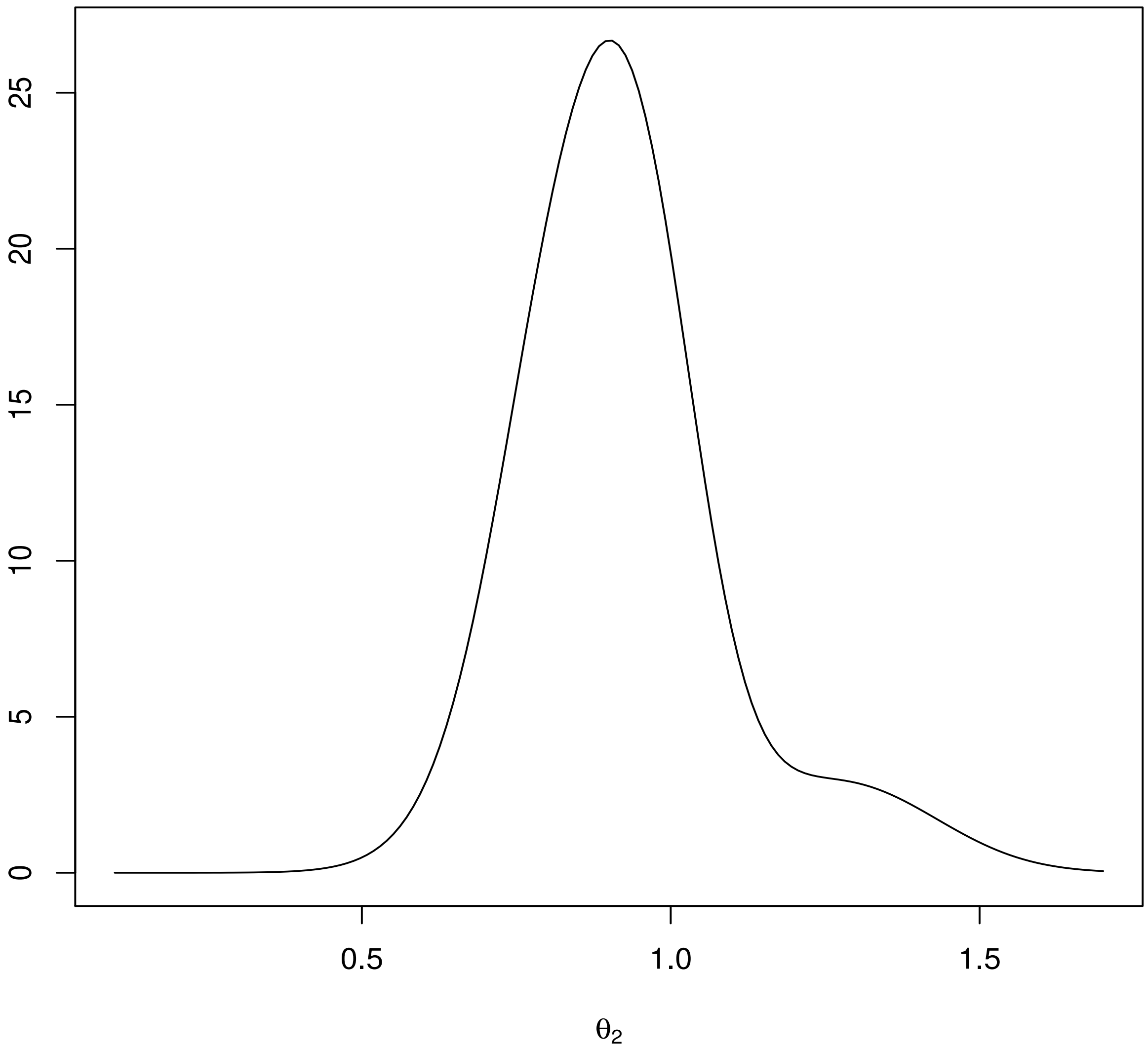


Figure 1: Marginal reference prior density (left) and marginal reference posterior density (right) of $\log(\theta_1)$ for the Davis data set, under the three Matérn families, $\theta_2 \in \{0.5, 1, 1.5\}$.

Alternatively, one could consider both parameters of the Matérn family of correlation functions as unknown, and compute the reference integrated likelihood for θ_2, $m(\mathbf{z} \mid \theta_2)$.


0
5
10
15
20
25
0.5
1.0
1.5
θ_2