Overall Objective Priors

Jim Berger, Jose Bernardo and Dongchu Sun

Duke University, University of Valencia and University of Missouri

Recent advances in statistical inference: theory and case studies

University of Padova

March 21-23, 2013

Outline

- Background
- Previous approaches to development of an overall prior
- New approaches to development of an overall prior
- Specifics of the reference distance approach
- Specifics of the prior modeling approach
- Summary

Background

- Objective Bayesian methods have priors defined by the model (or model structure).
- In models with a single unknown parameter, the acclaimed objective prior is the *Jeffreys-rule prior* (more generally, the *reference prior*).
- In multiparameter models, the optimal objective (e.g., reference or matching) prior depends on the quantity of interest, e.g., the parameter concerning which inference is being performed.
- But often one needs a single overall prior
  - for prediction
  - for decision analysis
  - when the user might consider non-standard quantities of interest
  - for computational simplicity
  - for sociological reasons

Example: *Bivariate Normal Distribution*, with mean parameters μ_1 and μ_2, standard deviations σ_1 and σ_2, and correlation ρ.

Berger and Sun (AOS2008) studied priors that had been considered for 21 quantities of interest (original parameters and derived ones such as μ_1/σ_1).

- An optimal prior for each quantity of interest was suggested.
- An overall prior was also suggested:
  - The primary criterion used to judge candidate overall priors was reasonable frequentist coverage properties of resulting credible intervals for the most important quantities of interest.
  - The prior (from Lindley and Bayarri)

$$\pi^O(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_1\sigma_2(1-\rho^2)}$$

was the suggested overall prior.

Previous Approaches to Development of an Overall Prior

- I. Group-invariance priors
- II. Constant or vague proper priors
- III. The Jeffreys-rule prior

Notation:

Data: $\boldsymbol{x}$

Unknown model parameters: $\boldsymbol{\theta}$

Data density: $p(\boldsymbol{x} \mid \boldsymbol{\theta})$

Prior density: $\pi(\boldsymbol{\theta})$

Marginal (predictive) density: $p(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$

Posterior density: $\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{\theta})\ \pi(\boldsymbol{\theta})/p(\boldsymbol{x})$

I. Group-invariance priors: If $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ has a group invariance structure, then the recommended objective prior is typically the right-Haar prior.

- Often works well for all parameters that define the invariance structure.
  Example: If the sampling model is $\mathrm{N}(x_i \mid \mu, \sigma)$, the right-Haar prior is $\pi(\mu, \sigma) = 1/\sigma$, and this is fine for either μ or σ (yielding the usual objective posteriors).

- But it may be poor for other parameters.
  Example: For the bivariate normal problem, one right-Haar prior is $\pi_1(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = 1/[\sigma_1^2(1-\rho^2)]$, which is fine for μ_1, σ_1 and ρ, but leads to problematical posteriors for μ_2 and σ_2 (Berger and Sun, 2008).

- And it may not be unique.
  Example: For the bivariate normal problem, another right-Haar prior is $\pi_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = 1/[\sigma_2^2(1-\rho^2)]$.

- The situation can be even worse if the right-Haar prior is used for derived parameters.

  Example: *Multi-normal means:* Let x_i be independent normal with mean μ_i and variance 1, for $i = 1 \cdots, m$.

  - The right-Haar (actually Haar) prior for $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ is $\pi(\boldsymbol{\mu}) = 1$.
  - It results in a sensible $\mathrm{N}(\mu_i \mid x_i, 1)$ posterior for each individual μ_i.
  - But it is terrible for $\theta = \frac{1}{m}|\boldsymbol{\mu}|^2 = \frac{1}{m}\sum_{i=1}^m \mu_i^2$ (Stein).
    - ∗ The posterior mean of θ is $[1 + \frac{1}{m}\sum_{i=1}^m x_i^2]$;
    - ∗ this converges to $[\theta + 2]$ as $m \to \infty$;
    - ∗ indeed, the posterior concentrates sharply around $[\theta + 2]$ and so is badly *inconsistent*.

II. Constant or vague proper priors are often used as the overall prior.

- The problems of a constant prior are well-documented, including
  - lack of invariance to transformation (the original problem with Laplace's 'inverse probability'),
  - frequent posterior impropriety (as in the first full Bayesian analyses of Gaussian spatial models with an exponential correlation structure, when constant priors were used for the range parameter),
  - and possible terrible performance (as in the previous example).
- Vague proper priors (such as a constant prior over a large compact set)
  - are at best equivalent to use of a constant prior (and so inherit the flaws of a constant prior);
  - can be worse, in that they can hide problems such as a lack of posterior propriety.

III. The Jeffreys-rule prior: If the data model density is $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ the Jeffeys-rule prior for the unknown $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\}$ has the form

$$|I(\boldsymbol{\theta})|^{1/2} d\theta_1 \ldots d\theta_k$$

where $I(\boldsymbol{\theta})$ is the $k \times k$ Fisher information matrix with (i, j) element

$$I(\boldsymbol{\theta})_{ij} = \mathrm{E}_{\boldsymbol{x} \mid \boldsymbol{\theta}} \left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) \right].$$

This is the optimal objective prior (from many perspectives) for (regular) one-parameter models, but has problems for multi-parameter models:

- The right-Haar prior in the earlier multi-normal mean problem is also the Jeffreys-rule prior there, and yielded inconsistent estimators. (It also yields inconsistent estimators in the Neyman-Scott problem.)
- For the $\mathrm{N}(x_i \mid \mu, \sigma)$ model, the Jeffreys-rule prior is $\pi(\mu, \sigma) = 1/\sigma^2$, which results in posterior inferences for μ and σ that have 'degrees of freedom' equal to n, not the correct $n - 1$.

- For the bivariate normal example, the Jeffreys-rule prior is $1/[\sigma_1^2\sigma_2^2(1-\rho^2)^2]$;
  - it yields the natural marginal posteriors for the means and standard deviations,
  - but results in quite inferior objective posteriors for ρ and various derived parameters (Berger and Sun, 2008)).
- in p-variate normal problems, the Jeffreys-rule prior for a covariance matrix can be very bad (Stein, Yang and Berger, 1992).
- It can overwhelm the data:

Example: *Multinomial distribution:* Suppose $\boldsymbol{x} = (x_1, \ldots, x_m)$ is multinomial $\mathrm{Mu}(\boldsymbol{x} \mid n; \theta_1, \ldots, \theta_m)$, where $\sum_{i=1}^m \theta_i = 1$. If the sample size n is small relative to the number of classes m, we have a large sparse table. The Jeffreys-rule prior, $\pi(\theta_1, \ldots, \theta_m) \propto \prod_{i=1}^m \theta_i^{-1/2}$ is a proper prior that can overwhelm the data.

- Suppose $n = 3$ and $m = 1000$, with $x_{240} = 2$, $x_{876} = 1$, other $x_i = 0$.
- The posterior means resulting from the Jeffreys prior are

$$\mathrm{E}[\theta_i \,|\, \boldsymbol{x}] = \frac{x_i + 1/2}{\sum_{i=1}^{m}(x_i + 1/2)} = \frac{x_i + 1/2}{n + m/2} = \frac{x_i + 1/2}{503} \, ,$$

  so $\mathrm{E}[\theta_{240} \,|\, \boldsymbol{x}] = \frac{2.5}{503}$, $\mathrm{E}[\theta_{876} \,|\, \boldsymbol{x}] = \frac{1.5}{503}$, $\mathrm{E}[\theta_i \,|\, \boldsymbol{x}] = \frac{0.5}{503}$ otherwise.
- Thus cells 240 and 876 only have total posterior probability $\frac{4}{503} = 0.008$, even though all 3 observations are in these cells.
- The problem is that the Jeffreys-rule prior added 1/2 to all the zero cells, making them much more important than the cells with data!
- Note that the uniform prior on the simplex is even worse, since it adds 1 to each cell. The prior $\prod_i \theta_i^{-1}$ adds zero to each cell, but the posterior is improper unless all cells have nonzero entries.

For specific problems there have been improvements such as the "independence Jeffreys-rule prior," but such prescriptions have been adhoc and have not lead to a general alternative definition.

New Approaches to Development of an Overall Prior

- **A.** The *reference distance approach*
- **B.** The *hierarchical approach*
  - **B1.** *Prior averaging*
  - **B2.** *Prior modeling approach*

A. The Reference Distance Approach: Choose a prior that yields marginal posteriors for all parameters that are close to the reference posteriors for the parameters in an average distance sense (to be specified).

Example: *Multinomial example (continued):*

- The reference prior, when θ_i is of interest, differs for each θ_i.
- It results in a Beta reference posterior $\text{Be}(\theta_i \,|\, x_i + \frac{1}{2}, n - x_i + \frac{1}{2})$.
- Goal: identify a single joint prior for $\boldsymbol{\theta}$ whose marginal posteriors could be expected to be close to each of the reference posteriors just described, in some average sense.
- Consider, as an overall prior, the Dirichlet $\text{Di}(\boldsymbol{\theta} \,|\, a, \ldots, a)$ distribution, having density proportional to $\prod_i \theta_i^{(a-1)}$.
  - The marginal posterior for θ_i is then $\text{Be}(\theta_i \,|\, x_i + a, n - x_i + (m-1)a)$.
  - The goal is to choose a so these are, in any average sense, close to the reference posteriors $\text{Be}(\theta_i \,|\, x_i + \frac{1}{2}, n - x_i + \frac{1}{2})$.

- The recommended choice is (approximately) $a = 1/m$:
  - ∗ This prior adds only $1/m = 0.001$ to each cell in the earlier example;
  - ∗ Thus

$$\mathrm{E}[\theta_i \,|\, \boldsymbol{x}] = \frac{x_i + 1/m}{\sum_{i=1}^{m}(x_i + 1/m)} = \frac{x_i + 1/m}{n+1} = \frac{x_i + 0.001}{4}\,,$$

so that $\mathrm{E}[\theta_{240} \,|\, \boldsymbol{x}] \approx 0.5$, $\mathrm{E}[\theta_{876} \,|\, \boldsymbol{x}] \approx 0.25$, and $\mathrm{E}[\theta_i \,|\, \boldsymbol{x}] \approx \frac{1}{4000}$ otherwise, all sensible (recall $x_{240} = 2$, $x_{876} = 1$, other $x_i = 0$).

A. The Hierarchical approach: Utilize hierarchical modeling to transfer the reference prior problem to a ‘higher level’.

A1. Prior Averaging: Starting with a collection of reference (or other) priors $\{\pi_i(\boldsymbol{\theta}), i = 1, \ldots, k\}$ for differing parameters or quantities of interest, use the average prior, such as

$$\pi(\boldsymbol{\theta}) = \sum_{i=1}^{k} \pi_i(\boldsymbol{\theta}) \, .$$

This is hierarchical as it coincides with giving each prior an equal prior probability of being correct, and averaging out over this hyperprior.

Example: *Bivariate Normal example (continued):* Faced with the two right-Haar priors, a natural prior to consider is their average, given by

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\sigma_1^2(1-\rho^2)} + \frac{1}{2\sigma_2^2(1-\rho^2)} \, .$$

- It is shown in Sun and Berger (2007) that this prior is *worse* than either right-Haar prior alone, suggesting that averaging improper priors is not a good idea.

- Interestingly, the geometric average of these two priors is the recommended overall prior for the bivariate normal $\pi^O(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = 1/[\sigma_1\sigma_2(1-\rho^2)]$, but justification for geometric averaging is currently lacking.

Another problem with prior averaging is that there can be too many reference priors to average.

Example: *Multinomial example (continued):* The reference prior $\pi_i(\theta)$, when θ_i is the parameter of interest, depends on the parameter ordering chosen in the derivation (e.g. $\{\theta_i, \theta_1, \theta_2, \ldots, \theta_m\}$).

- All choices lead to the same marginal reference posterior $\text{Be}(\theta_i \,|\, x_i + \frac{1}{2}, n - x_i + \frac{1}{2})$.
- In constructing an overall prior by prior averaging, each of the orderings would have to be considered.
- There are $m!$ reference priors to be averaged.

Conclusion: For the reasons indicated above, we do not recommend the prior averaging approach.

A2. Prior Modeling Approach: In this approach one

- Chooses a class of *proper* priors $\pi(\boldsymbol{\theta} \,|\, a)$ that reflects the desired structure of the problem.
- Forms the marginal likelihood $p(\boldsymbol{x} \mid a) = \int p(\boldsymbol{x} \mid a)\pi(\boldsymbol{\theta} \,|\, a)\; d\boldsymbol{\theta}$.
- Finds the reference prior, $\pi^R(a)$, for a in this marginal model.
- Thus the overall prior becomes

$$\pi^O(\boldsymbol{\theta}) = \int \pi(\boldsymbol{\theta} \mid a)\pi^R(a)da\,,$$

although computation is typically easier in the hierarchical formulation.

Example: *Multinomial (continued):*

- The Dirichlet $\mathrm{Di}(\boldsymbol{\theta} \,|\, a, \ldots, a)$ class of priors is natural, reflecting the desire to treat all the θ_i similarly.

- The marginal model is then

$$\begin{aligned} p(\boldsymbol{x} \,|\, a) &= \int \binom{n}{x_1 \ldots x_m} \left(\prod_{i=1}^{m} \theta_i^{x_i} \right) \frac{\Gamma(m\, a)}{\Gamma(a)^m} \prod_{i=1}^{m} \theta_i^{a-1} d\boldsymbol{\theta} \\ &= \binom{n}{x_1 \ldots x_m} \frac{\Gamma(m\, a)}{\Gamma(a)^m} \frac{\prod_{i=1}^{m} \Gamma(x_i + a)}{\Gamma(n + m\, a)} \,. \end{aligned}$$

- The reference prior for $\pi^R(a)$ would just be the Jeffreys-rule prior for this marginal model, and is given later.

- The overall prior for $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{\theta}) = \int \mathrm{Di}(\boldsymbol{\theta} \,|\, a, \ldots, a)\, \pi^R(a) da \,.$$

Specifics of the Reference Distance Approach

Defining a distance (divergence): *Intrinsic discrepancy* (Bernardo and Rueda, 2002; Bernardo, 2005, 2001)

Definition 1 *The* **intrinsic discrepancy** $\delta\{p_1, p_2\}$ *between two probability distributions for the random vector* $\boldsymbol{\psi}$ *with densities* $p_1(\boldsymbol{\psi}) \in \boldsymbol{\Psi}_1$ *and* $p_2(\boldsymbol{\psi}) \in \boldsymbol{\Psi}_2$ *is*

$$\delta\{p_1, p_2\} = \min\left\{\int_{\boldsymbol{\Psi}_1} p_1(\boldsymbol{\psi}) \log \frac{p_1(\boldsymbol{\psi})}{p_2(\boldsymbol{\psi})}\, d\boldsymbol{\psi}, \int_{\boldsymbol{\Psi}_2} p_2(\boldsymbol{x}) \log \frac{p_2(\boldsymbol{\psi})}{p_1(\boldsymbol{\psi})}\, d\boldsymbol{\psi}\right\}$$

assuming that at least one of the integrals exists.

The (non-symmetric) (Kullback-Leibler) logarithmic divergence, in scenarios where there is a 'true' distribution $p_2(\boldsymbol{\psi})$,

$$\kappa\{p_1 \,|\, p_2\} = \int_{\boldsymbol{\Psi}_2} p_2(\boldsymbol{x}) \log \frac{p_2(\boldsymbol{\psi})}{p_1(\boldsymbol{\psi})}\, d\boldsymbol{\psi},$$

is another reasonable choice (and is usually equivalent to the intrinsic discrepancy).

The exact solution scenario: If a prior $\pi^O(\boldsymbol{\theta})$ yields marginal posteriors that are equal to the reference posteriors for each of the quantities of interest, then the resulting intrinsic discrepancies are zero and $\pi^O(\boldsymbol{\theta})$ is a natural choice for the overall prior.

Example: *Univariate normal distribution:* For the $\mathrm{N}(x_i \mid \mu, \sigma)$ distribution,

- suppose μ and σ are the quantities of interest;
- $\pi^O(\mu, \sigma) = \sigma^{-1}$ is the reference prior when either μ or σ is the quantity of interest;
- hence π^O is an optimal overall prior.

Suppose, in addition to μ and σ, the centrality parameter $\theta = \mu/\sigma$ is also a quantity of interest.

- The reference prior for θ is (Bernardo, 1979) $\pi_\theta(\theta, \sigma) = (1 + \frac{1}{2}\theta^2)^{-1/2}\sigma^{-1}$;
- this yields different marginal posteriors than does $\pi^O(\mu, \sigma) = \sigma^{-1}$;
- hence we would not have an exact solution.

General (Proper) Situation:

- Suppose the model is $p(\boldsymbol{x} \mid \boldsymbol{\omega})$ and the quantities of interest are $\{\theta_1, \ldots, \theta_m\}$, with *proper* reference priors $\{\pi_i^R(\boldsymbol{\omega})\}_{i=1}^m$.
  - $\{\pi_i^R(\theta_i \,|\, \boldsymbol{x})\}_{i=1}^m$ are the corresponding marginal reference posteriors.
  - $p_i^R(\boldsymbol{x}) = \int_{\boldsymbol{\Omega}} p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, \pi_i^R(\boldsymbol{\omega})\, d\boldsymbol{\omega}$ are the corresponding (proper) marginal densities or prior predictives.
- $\{w_i\}_{i=1}^m$ are weights giving the importance of each quantity of interest.
- A family of priors $\mathcal{F} = \{\pi(\boldsymbol{\omega} \,|\, \boldsymbol{a}), \boldsymbol{a} \in \mathcal{A}\}$ is considered.

The best overall prior within $\mathcal{F}$ is defined to be that which minimizes, over $\boldsymbol{a} \in \mathcal{A}$, the **average expected intrinsic loss**

$$d(\boldsymbol{a}) = \sum_{i=1}^{m} w_i \int_{\mathcal{X}} \delta\{\pi_i^R(\cdot \,|\, \boldsymbol{x}),\, \pi_i(\cdot \,|\, \boldsymbol{x}, \boldsymbol{a})\}\, p_i^R(\boldsymbol{x})\, d\boldsymbol{x}\,.$$

Big Issue: When the reference priors are not proper (the usual case), there is no assurance that $d(\boldsymbol{a})$ is finite. There is no clear way to proceed otherwise, so we are studying if $d(\boldsymbol{a})$ is often finite in the improper case.

Example: *Multinomial model:* Consider the multinomial model with m cells and parameters $\{\theta_1, \ldots, \theta_m\}$, with $\sum_{i=1}^m \theta_i = 1$. We seek to find the $\text{Di}(\boldsymbol{\theta} \,|\, a, \ldots, a)$ prior that minimizes the average expected intrinsic loss.

- The reference posterior for each of the θ_i's is $\text{Be}(\theta_i \,|\, x_i + \frac{1}{2}, n - x_i + \frac{1}{2})$.
- The marginal posterior of θ_i for the Dirchlet prior is $\text{Be}(\theta_i \,|\, x_i + a, n - x_i + (m-1)a)$.
- The intrinsic discrepancy between these marginal posteriors is

$$\delta_i\{a \,|\, \boldsymbol{x}, m, n\} = \delta_{\text{Be}}\{x_i + \tfrac{1}{2}, n - x_i + \tfrac{1}{2}, x_i + a, n - x_i + (m-1)a\}\,,$$

$$\delta_{\text{Be}}\{a_1, \beta_1, a_2, \beta_2\} = \min[\kappa_{\text{Be}}\{a_2, \beta_2 \,|\, a_1, \beta_1\}, \kappa_{\text{Be}}\{a_1, \beta_1 \,|\, a_2, \beta_2\}\,,]$$

$$\kappa_{\text{Be}}\{a_2, \beta_2 \,|\, a_1, \beta_1\} = \int_0^1 \text{Be}(\theta_i \,|\, a_1, \beta_1) \log\left[\frac{\text{Be}(\theta_i \,|\, a_1, \beta_1)}{\text{Be}(\theta_i \,|\, a_2, \beta_2)}\right] d\theta_i$$

$$= \log\left[\frac{\Gamma(a_1 + \beta_1)}{\Gamma(a_2 + \beta_2)} \, \frac{\Gamma(a_2)}{\Gamma(a_1)} \, \frac{\Gamma(\beta_2)}{\Gamma(\beta_1)}\right]$$
$$+ (a_1 - a_2)\psi(a_1) + (\beta_1 - \beta_2)\psi(\beta_1) - ((a_1 + \beta_1) - (a_2 + \beta_2))\psi(a_1 + \beta_1),$$

and $\psi(\cdot)$ is the digamma function.

- The discrepancy $\delta_i\{a \,|\, x_i, m, n\}$ between the two posteriors of θ_i only depends on the data through x_i and the reference predictive for x_i is

$$p(x_i \,|\, n) = \int_0^1 \mathrm{Bi}(x_i \,|\, n, \theta_i)\, \mathrm{Be}(\theta_i \,|\, 1/2, 1/2)\, d\theta_i = \frac{1}{\pi}\, \frac{\Gamma(x_i + \frac{1}{2})\, \Gamma(n - x_i + \frac{1}{2})}{\Gamma(x_i + 1)\, \Gamma(n - x_i + 1)}\,,$$

  – because the sampling distribution of x_i is $\mathrm{Bi}(x_i \,|\, n, \theta_i)$,

  – and the marginal reference prior for θ_i is $\pi_i(\theta_i) = \mathrm{Be}(\theta_i \,|\, 1/2, 1/2)$.

- Noting that each θ_i yields the same expected loss, the average expected intrinsic loss is

$$d(a \,|\, m, n) = \sum_{x=0}^{n} \delta\{a \,|\, x, m, n\}\, p(x \,|\, n)\,.$$

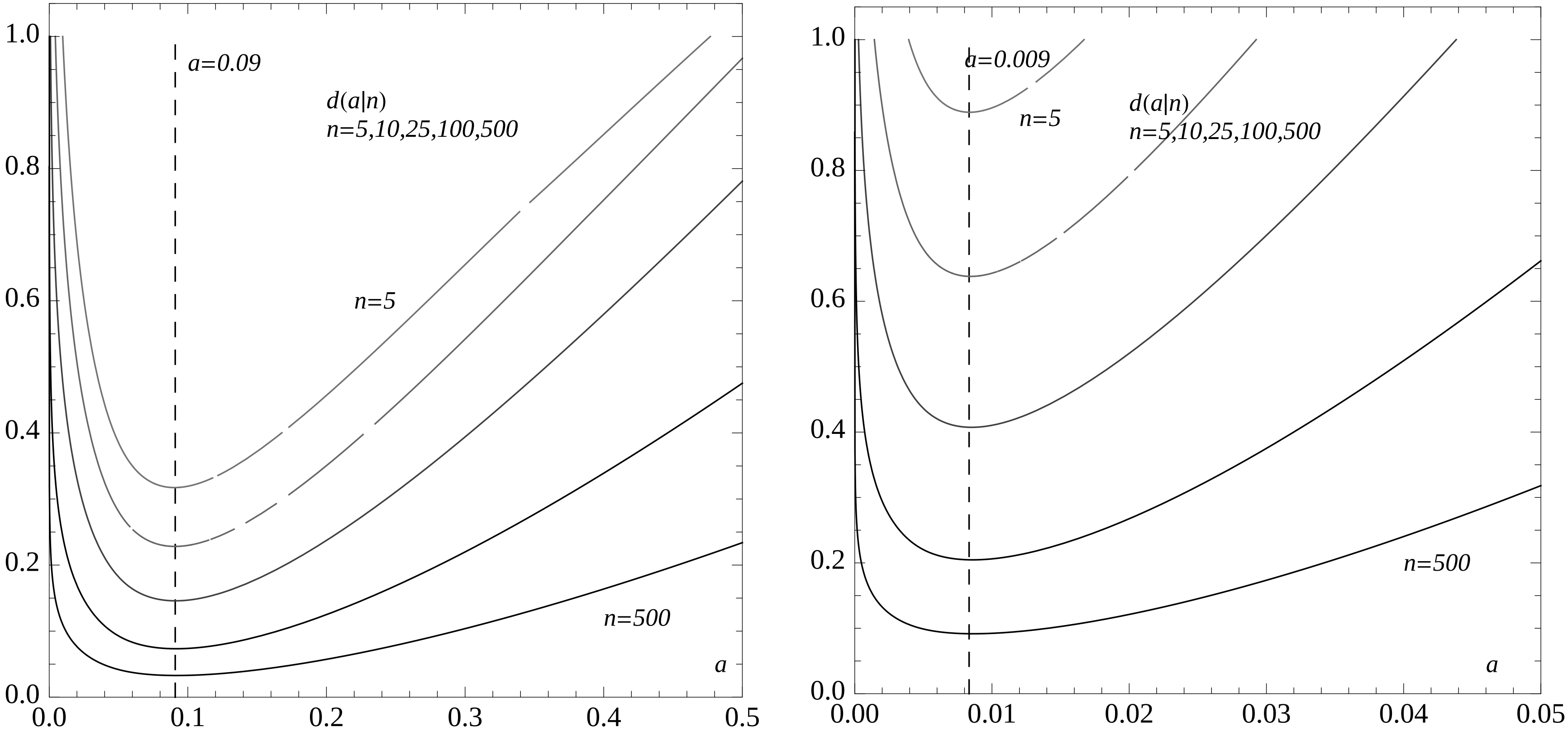

Figure 1: Expected intrinsic losses, of using a Dirichlet prior with parameter $\{a, \ldots, a\}$ in a multinomial model with m cells, for sample sizes $5, 10, 25, 100$ and 500. Left panel, $m = 10$; right panel, $m = 100$. In both cases, the optimal value for all sample sizes is $a^* \approx 1/m$. Exact values for $n = 25$ are 0.091 and 0.0085.)

Specifics of the Prior Modeling Approach

- Multinomial Example
- Bivariate Normal Example

Example: *Multinomial (continued):*

- The Dirichlet $\mathrm{Di}(\boldsymbol{\theta}\,|\,a,\ldots,a)$ class of priors is natural, reflecting the desire to treat all the θ_i similarly.

- The marginal model is then

$$\begin{aligned} p(\boldsymbol{x}\,|\,a) &= \int \binom{n}{x_1 \ldots x_m} \left(\prod_{i=1}^{m} \theta_i^{x_i}\right) \frac{\Gamma(m\,a)}{\Gamma(a)^m} \prod_{i=1}^{m} \theta_i^{a-1} d\boldsymbol{\theta} \\ &= \binom{n}{x_1 \ldots x_m} \frac{\Gamma(m\,a)}{\Gamma(a)^m} \frac{\prod_{i=1}^{m} \Gamma(x_i + a)}{\Gamma(n + m\,a)} \,. \end{aligned}$$

- The reference prior for $\pi^R(a)$ would just be the Jeffreys-rule prior for this marginal model, and is given later.

- The overall prior for $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{\theta}) = \int \mathrm{Di}(\boldsymbol{\theta}\,|\,a,\ldots,a)\,\pi^R(a)da\,.$$

Derivation of $\pi^R(a)$: $p(\boldsymbol{x} \,|\, a)$ is a regular one-parameter model, so the reference prior is the Jeffreys-rule prior.

- The marginal (predictive) density of any of the x_i's is

$$p_1(x_i \,|\, a, m, n) = \binom{n}{x_i} \frac{\Gamma(x_i + a)\,\Gamma(n - x_i + (m-1)a)\,\Gamma(m\,a)}{\Gamma(a)\,\Gamma((m-1)a)\,\Gamma(n + m\,a)}\,.$$

- Computation yields

$$\pi^R(a \,|\, m, n) \propto \left[\sum_{j=0}^{n-1} \left(\frac{Q(j \,|\, a, m, n)}{(a+j)^2} - \frac{m}{(m\,a+j)^2} \right)\right]^{1/2},$$

where $Q(j \,|\, a, m, n) = \sum_{l=j+1}^{n} p_1(l \,|\, a, m, n), \quad j = 0, \ldots, n-1.$

- $\pi^R(a)$ can be shown to be a proper prior. Why did that happen?

  It can be shown that

  $$p(\boldsymbol{x} \,|\, a) = \begin{cases} O(a^{r-1}), & \text{as } a \to 0, \\ \binom{n}{\boldsymbol{x}} m^{-n}, & \text{as } a \to \infty, \end{cases}$$

  where r is the number of nonzero x_i. Thus the likelihood is constant at ∞, so the prior must be proper at infinity for the posterior to exist.

- It can be shown that, for sparse tables, where m/n is relatively large, the reference prior is well approximated by the proper prior

  $$\pi^*(a \,|\, m, n) = \frac{1}{2} \frac{n}{m} \, a^{-1/2} \Big(a + \frac{n}{m} \Big)^{-3/2} .$$

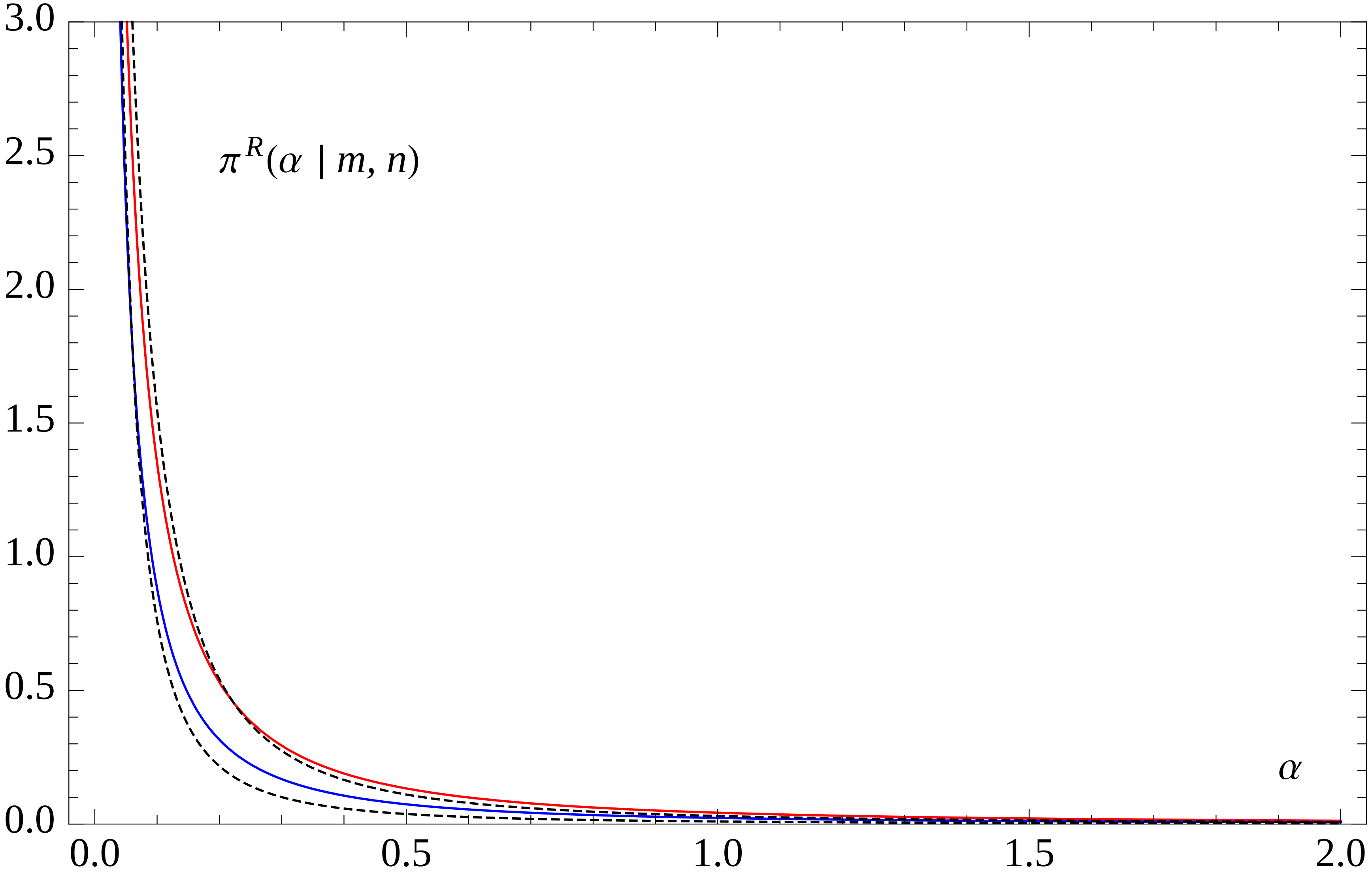

Figure 2: Reference priors $\pi^R(a \,|\, m, n)$ (solid lines) and its approximations (dotted lines) for $(m = 150, n = 10)$ (upper curve) and for $(m = 500, n = 10)$ (lower curve)

Computation with the hierarchical reference prior:

1. The obvious *MCMC sampler* is:

Step 1. Use a Metropolis Hastings move to sample from the marginal posterior $\pi^R(a \mid \boldsymbol{x}) \propto \pi^R(a)\, p(\boldsymbol{x} \mid a)$.

Step 2. Given a, sample from the usual beta posterior $\pi(\theta \mid a, \boldsymbol{x})$.

2. The *empirical Bayes approximation* is to fix a at it's posterior mode $\widehat{a}^R$, which exists and is nonzero if $r \geq 2$.

Using the ordinary empirical Bayes estimate from maximizing $p(\boldsymbol{x} \mid a)$ is problematical, since the likelihood does not go to zero at ∞. For instance, if all $x_i = 1$, $p(\boldsymbol{x} \mid a)$ has a likelihood increasing in a.

Asymptotic posterior mode as m and n go to ∞, but $n/m \to 0$:

$$\widehat{a} = \begin{cases} \frac{(r-1.5)}{m \log n} & \text{if } \frac{r}{n} \to 0, \\ \frac{c^* n}{m} & \text{if } \frac{r}{n} \to c < 1, \\ \frac{n^2}{2m(n-r)} & \text{if } \frac{r}{n} \to \ 1 \text{ and } \frac{(n-r)^2}{n} \to \infty. \end{cases},$$

where r is the number of nonzero x_i and c^* is the solution to $c^* \log(1 + \frac{1}{c^*}) = c$.

- While $\widehat{a}$ is of $O(\frac{1}{m})$, it also depends on r and n.
- For instance, suppose $r = n/2$ (*i.e.*, there are $n/2$ nonzero entries); then $\widehat{a} = 0.40n/m$.

Example: *Bivariate Normal (continued):* There are actually a continuum of right-Haar priors given as follows.

- For the orthogonal matrix $\mathbf{\Gamma} = \begin{pmatrix} \cos(\beta) & -\sin(\beta) \\ \sin(\beta) & \cos(\beta) \end{pmatrix}$, $-\pi/2 < \beta \leq \pi/2$,
- the right-Haar prior based on the transformed data $\mathbf{\Gamma}\boldsymbol{X}$ is

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho \mid \beta) = \frac{\sin^2(\beta)\sigma_1^2 + \cos^2(\beta)\sigma_2^2 + 2\sin(\beta)\cos(\beta)\rho\sigma_1\sigma_2}{\sigma_1^2\sigma_2^2(1-\rho^2)}.$$

- We thus have a class of priors indexed by a hyperparameter β.
- The natural prior distribution on β is the (proper) uniform distribution (being uniform over the set of rotations is natural.)
- The resulting prior is

$$\pi^O(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} \pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho \mid \beta) d\beta \propto \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) \frac{1}{(1-\rho^2)}$$

the same bad prior as the average of the original two right-Haar priors.

Empirical hierarchical approach: Find the empirical Bayes estimate $\hat{\beta}$ and use $\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho \mid \hat{\beta})$ as the overall prior.

This was shown in Sun and Berger (2007) to result in a terrible overall prior, much worse than either the individual reference priors or even the bad prior average.

Summary

- There is an important need for overall objective priors for models.
- The reference distance approach is natural, and seems to work well when reference priors are proper.
- It is unclear if the reference distance approach can be used when the reference priors are improper.
- The prior averaging approach is not recommended when the reference priors are improper and can be computationally difficult even when they are proper.
- The prior modeling approach seems excellent (as usual), and is recommended if one can find a natural class of proper priors to initiate the hierarchical analysis.
- The failure of the hierarchical approach for the right-Haar priors in the bivariate normal example was dramatic, suggesting that using improper priors are the bottom level of a hierarchy is a bad idea.

Thanks!