Reference Posteriors From a Mixture of Reference Priors

Michael D. Sonksen
sonksen@stat.unm.edu
Department of Mathematics and Statistics
The University of New Mexico
Albuquerque, NM 87131, USA

December 20, 2018

Abstract

In multi-parameter settings, the reference prior for a given likelihood is not uniquely defined. Instead, the reference prior algorithm of Berger and Bernardo (1992) produces different reference priors depending on a ordering and grouping of the model parameters. In some instances an explicit formula for all possible reference prior can be found. We define a reference posterior using a mixture of all possible reference priors. We utilize the Dirichlet process to define the prior distribution for the grouping and ordering of the parameters. The discrete nature of Dirichlet process priors makes it an ideal mixing distribution. Examination of the posterior probabilities for the various models provides insight into the viability of specific orderings and groupings. We illustrate this methodology and consider the associated computational issues with a multinomial model and a constrained Poisson rate model.

1 Introduction

Reference priors, first conceived by Bernardo (1979), are among the most popular non-informative priors used in practice. In multi-parameter problems, there is not one unique reference prior, instead a reference prior is defined as function a specific group and ordering of the parameters. Because of the large number of possible reference priors we consider an approach to define the reference posterior as a mixture of all reference posteriors obtained from a reference prior. This is implemented by modeling the grouping/ordering as a realization from a distribution following a Dirichlet Process.

The rest of the paper is organized as follows. In Section 2, we review the theory of reference priors and the dirichlet process. In Section 3, we describe our proposed reference posterior as a mixture of reference priors. In Section 4, we examine theoretical properties of this reference posterior. In Section 5, we consider two examples. A discussion of the proposed model and

1.1 Reference Priors

Reference priors are a class of non-informative priors which are an extension of Jeffreys prior. The theory and derivation of reference priors was developed in a string of papers, namely Bernardo (1979) Berger and Bernardo (1992b), Berger and Bernardo (1992a), and Berger et al. (2009).

Bernardo (1979) defined the reference prior as a minimizer of expected distance for one parameter or groups of parameters of interest and nuisence parameters. Bernardo's idea, is that the prior distribution should impact the posterior distribution as little as possible. One way to define

this is to define the reference prior as the distribution for which the Expected Kullbach-Liebler divergence (Equation (1)) between the posterior distribution (f) and prior distribution (g) is as large as possible (where the expectation is taken with respect to the marginal distribution of the data).

$$KL(f, g) = \mathrm{E}_f \left[\frac{f(\theta|X)}{g(\theta)} \right] \tag{1}$$

Unfortunately, the prior which minimizes Equation (1) is often discrete even for continuous parameters. Berger and Bernardo (1992b) modified the definition of the reference prior to be the minimizer of the expected KL divergence where the expectation is taken with respect to repeated experiments $Z_1, Z_2, \ldots, Z_s$. The minimizer for a fixed s is found and then passed to the limit. The justification for this is that the reference prior should be useful in repeated experiments.

When the support of Θ is not compact, the expected K-L divergence may be infinity. To overcome this problem, Berger and Bernardo (1992a) recommend considereing a compact parameter space Θ_l such that

$$\lim_{l \to \infty} \Theta_l = \Theta.$$

The minimizing function is then found for an arbitrary l and a the reference prior prior is taken as the limit as $l \to \infty$.

Berger et al. (2009) showed that a reference prior may also be represented as a the maximizer of missing information. They provide an explicit form of the reference prior when there is only one parameter. Unfortunately, this work has yet to be extended to multiple parameters.

When more than one parameter exists, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, the reference prior is simply Jeffreys prior. However, it is well known that in multi-parameter likelihoods, Jeffreys prior can lead to poor posterior distributions (strong inconsistency, impropriety). To alleviate this problem, Berger and Bernardo (1992a) and Berger and Bernardo (1992b) defined the reference prior as conditional on some grouping and ordering of the parameters. This sequential reference prior is defined by first grouping the parameters into m groups ordered groups $\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}, \ldots, \boldsymbol{\theta}_{(m)}$. Where group i has n_i number of parameters. Further define $N_j = \sum_{i=1}^{j} n_i$. Let S bet the inverse of the Fisher Information matrix (orderd by the groups) and S_j is the upper left $N_j \times N_j$ upper left corner of S. If $H_j = S_j^{-1}$ and h_j= lower right $n_j \times n_j$ corner of H_j then the reference prior for this ordering and grouping is

$$\pi^l(\boldsymbol{\theta}) = \left(\prod_{i=1}^{m} \frac{|h_i(\boldsymbol{\theta})|^{1/2}}{\int_{\Theta^l(\boldsymbol{\theta}_{[i-1]})} |h_i(\boldsymbol{\theta})|^{1/2} d\boldsymbol{\theta}_{(i)}} \right) 1_{\Theta^l}(\boldsymbol{\theta})$$

$$\pi(\boldsymbol{\theta}) = \lim_{l \to \infty} \frac{\pi^l(\boldsymbol{\theta})}{\pi^l(\boldsymbol{\theta}^*)}$$

if $|h_j|$ only depends on parameters from $\theta_{(1)}, \ldots, \theta_{(j)}$. Where $\boldsymbol{\theta}^*$ is any fixed point in Θ.

Obviously, this sequential defintion of the reference priors means that there can be many different reference priors to use in a problem. Pratically, users often try several different reference priors based on intuative grouping/orderings and exaimine either theoretical properties (prior moments, posterior coverage probabilities, etc) or criterior based on a given sample (DIC, p-value diagnostics, Bayes factors) to select a prior distribution for use.

In this work, we recommend averaging over all possible reference priors. To average over all possible reference priors, we will utized the Nonparametric Bayes framework. Specifically, we will consider a Dirichlet Process mixture to model the ordering/grouping of our parameters.

1.2 Dirichlet Processes

The Dirichlet process (**??**) is a commonly used tool in the non-parametric Bayesian literature. **?** described the dirichlet process as a probability distribution with support on discrete probability functions. Common descriptions of the Diriclet process include stick breaking processes and the chineses restaurant process. We will not describe the Dirichlet process, and it's properties in detail here, but instead refer the reader to **?**, **?**, and the referenences therein.

The DP is characterized by it's base measure G_0 and mass parameter α. We will represent such a process by $DP(\alpha, G_0)$. The base measure G_0 defines the possible values and shape of realizations from the DP. The mass parameter, α, affects the variance of π in the stick breaking representation. A realization from a DP process, $G|\alpha, G_0 \sim DP(\alpha, G_0)$, is a discrete probability distribution.

The realizations from a Dirichlet process are rarely used to model the data directly. Instead, parameters are often assumed to follow realizations from a dirichlet process. A classic example is the use of the dirichlet process in hierarchical models. Parameters, representing latent group means, can be assumed to follow a DP.

The estimation of summaries of the posterior distribution of models with a DP is often done with Markov chain Monte Carlo (MCMC) methods. In particular, the stick-breaking representation can be used to define a Metropolis-Hastings algorithm with stationary distibution approximately that of the posterior distribution. In this work, we use Algorithm 7 of **?** for all computation. For other algorithms, see **?**, **?**, **?**.

2 Description of Methods

2.1 Ordering and Grouping as Latent Parameters

Assume that all reference priors are proper or lead to a proper posterior distribution. The ordering and grouping of k parameters can be represented through a latent parameter vector $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_k)$. Each element of this vector is tied to one of the parameters. A larger ϕ_i implies that the parameter is of more inferential importance and when $\phi_i = \phi_j$ the two parameters are of the same importance. For example, $\boldsymbol{\phi} = (1, 1, 4, 2, 2)$ would imply the grouping / ordering: $\boldsymbol{\theta}_1 = \{\theta_3\}$, $\boldsymbol{\theta}_{(2)} = \{\theta_4, \theta_5\}$ and $\boldsymbol{\theta}_{(3)} = \{\theta_1, \theta_2\}$. This represenation allows us to model the grouping/ordering by modeling $\boldsymbol{\phi}$. The choice of having larger ϕ denote a more important parameter is arbitrary, but works well in our examples.

2.2 Application to MDP

To model $\boldsymbol{\phi}$, we assume that each component independently follows a realization from the Dirichlet process with base measure F_0 and mass parameter α. Because we do not care about the distribution of ϕ, outside of the ordering of realizations, the choice of a continuous F_0 is arbitrary and can be chosen for computational convience (in this work we use the standard normal). The discrete nature of realizations from a Dirichlet process naturally will allow the elements of $\boldsymbol{\phi}$ to be equal with positive probability. The mass parameter α, which in essence controls the number of groups on average, can be fixed or assumed to follow a distribution. We consider both of these cases in our examples in Section 4.

Defining the reference prior as a mixture overall possible reference priors allows the observed data to select the prior distribution which provides a better fit.

2.3 Improper Reference Priors

In the examples and derivations we consider in this work all of the grouping and orderings yield prior distributions which correspond the proper posterior distributions.

When the parameter space is not compact, the reference priors could be improper and there is a danger that the resulting posterior distribution will be improper.

In this case, we propose two options to avoid this disasterous problem.

If we can identify which grouping/orderings induce a reference prior which yields an improper posterior, those grouping/orderings can be assigned a prior probability of zero.

The latter situation may not be feasible, especially if a large number of grouping/orderings exist.

Instead, one can utilize a common assumption of the reference prior derivation for non-compact parameter spaces.

In the reference prior formula, the reference prior, $\pi_l(\boldsymbol{\theta})$ is derived for a compact subspace Θ_l such that $\lim_{l\to\infty} \Theta_l = \Theta$. As an approximation, we can average over all reference priors for a given l and use the posterior probabilities of the ordering/groupings to identify potential reference priors for use. Of course one would have to check if the recommended reference priors induce a proper posterior distribution.

Another potential concern is that the reference priors may not be of closed form or easily found. In these cases the reference priors can be approximated using the algorithm in Section 3.2 of **?**.

2.4 Computation

The model may be fit using the algorithms of **?**.

3 Properties

-Minimizing KL Divergence (long shot)

When considering non-parameteric mixtures of priors, the is often an interest in showing that the resulting posterior distribution is consistent.

-Posterior Consistency (prove!!!!)

4 Examples

We will consider three examples to illustrate the proposed methodology. The first utilizes a simulations study around the multinomial distribution to invistigate small sample properties of the proposed models. We also examine a constrained Poisson rate model which is used to estimate mortality rates from a well known actuarial dataset.

4.1 Multinomial Model

A likelihood for which the form of all reference priors is known is the multinomial model (Berger and Bernardo, 1992b). If we assume that there are k parameters, the likelihood for $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_k)$

when $N \leq \sum_{i=1}^{k} Y_i$ is known is

$$f(\boldsymbol{Y}|\boldsymbol{\theta}) = \binom{N}{Y_1, Y_1, \ldots, Y_n} \prod_{i=1}^{k} \left(1 - \sum_{i=1}^{k} \theta_i\right)^{N - \sum_{i=1}^{k} Y_i} \prod_{i=1}^{k} \theta_i^{Y_i}.$$

Berger and Bernardo (1992b) showed that the ordered-group reference prior has the form:

$$\pi_{\texttt{multinom}}(\boldsymbol{\theta}) \propto \prod_{i=1}^{k}$$

We will average over all possible reference priors by allowing the parameters ($\boldsymbol{\phi}$) indexing the ordered groups to follow realizations of the Dirichlet Process with a standard normal base measure. For the mass parameter of the Dirichlet Process, we will consider four different values (0.01,0.1,1.0,10.0). This makes our model for a given α:

$$\begin{aligned} F_0 &= N(0,1), \\ F &\sim DP(F_0, \alpha), \\ \phi_i|F &\overset{iid}{\sim} F, \\ \boldsymbol{\theta}|\boldsymbol{\phi} &\sim \pi_{\text{multinom}}(\boldsymbol{\theta}|\boldsymbol{\phi}), \\ Y_i|\theta_i, N_i &\overset{ind}{\sim} Multinomial(N, \boldsymbol{\theta}). \end{aligned}$$

To investigate properties of this mixture of reference priors model, we performed the following simulation study. We examine different values of k and fixed $\boldsymbol{\theta}$. To investigate the small sample properties, we consider the frequentist coverage probabilities of the resulting 95% posterior intervals with 1000 intervals compared for each set of parameters. The posterior intervals were estimated from quantiles of posterior samples generated the Metropolis-Hasting algorithm described in Section **??**.

The estimated frequestist coverage probabilities are displayed in Table **??**.

N	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6

4.2 Constrained Poisson Rate Model

Sonksen and Peruggia (2012) provides another model where the reference prior for any grouping/ordering is known. They model the mortality data of Broffitt (1988) by assuming that the number of observed deaths (Y_i) in N_i total people of age i follows a Poisson distribution.

$$Y_i|\psi_i, N_i \overset{ind}{\sim} Poisson(N_i\psi_i) \tag{2}$$

Where ψ_i is the mortality rate of age group i for $i = 35, 36, \ldots, 64$. The observed mortality rates (Y_i/N_i) are displayed in Table 1.

Sonksen and Peruggia (2012) showed that the reference prior, for a given ordered group $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(m)}$, of the model in Expression (2) satisfies.

$$\pi_{\text{poi\_ref}}(\boldsymbol{\theta}|\boldsymbol{\phi}) \propto \frac{1}{\prod_{i=1}^{k} \sqrt{\theta_i}} \times \frac{1}{\prod_{j=2}^{m} (\sqrt{\gamma_j} - \sqrt{\eta_j})^{n_j}} \times I_{\Theta_{\text{Incr}}}(\boldsymbol{\theta}), \tag{3}$$

Table 1: Insurance records on mortality data for male individuals in 30 age groups (Broffitt, 1988). For each age group, the table reports the total number of individuals and the observed number of deaths. If an individual joined or left the insurance policy during a given year, he is counted as 0.5 of a person.

age	size	deaths	age	size	deaths	age	size	deaths
35	1771.5	3	45	1931.0	8	55	1204.5	11
36	2126.5	1	46	1746.5	13	56	1113.5	13
37	2743.5	3	47	1580.0	8	57	1048.0	12
38	2766.0	2	48	1580.0	2	58	1155.0	12
39	2426.0	2	49	1467.5	7	59	1018.5	19
40	2368.0	4	50	1516.0	4	60	945	12
41	2310.0	4	51	1371.5	7	61	853	16
42	2306.5	7	52	1343.0	4	62	750	12
43	2059.5	5	53	1304.0	4	63	693	6
44	1917.0	2	54	1232.5	1	64	594	10

where,

$$\gamma_{j+1} = \begin{cases} \min[\boldsymbol{\theta}_{(1:j)} : \boldsymbol{\theta}_{(1:j)} > \max(\boldsymbol{\theta}_{(j+1)})], & \text{if } \max[\boldsymbol{\theta}_{(1:\mathrm{j})}] > \max[\boldsymbol{\theta}_{(\mathrm{j}+1)}], \\ u, & \text{if } \max[\boldsymbol{\theta}_{(1:\mathrm{j})}] < \max[\boldsymbol{\theta}_{(\mathrm{j}+1)}], \end{cases}$$

and

$$\eta_{j+1} = \begin{cases} \max[\boldsymbol{\theta}_{(1:j)} : \boldsymbol{\theta}_{(1:j)} < \min(\boldsymbol{\theta}_{(j+1)})], & \text{if } \min[\boldsymbol{\theta}_{(1:\mathrm{j})}] < \min[\boldsymbol{\theta}_{(\mathrm{j}+1)}], \\ 0, & \text{if } \min[\boldsymbol{\theta}_{(1:\mathrm{j})}] > \min[\boldsymbol{\theta}_{(\mathrm{j}+1)}]. \end{cases}$$

To complete the model specification, we assume that the measure F_0 is a standard normal and that the mass parameter α follows an exponential distribution with mean of 1. Making our final model:

$$\begin{aligned} \alpha &\sim Exp(1), \\ F_0 &= N(0,1), \\ F &\sim DP(F_0, \alpha), \\ \phi_i | F &\overset{iid}{\sim} F, \\ \boldsymbol{\theta}|\boldsymbol{\phi} &\sim \pi_{\text{poi\_ref}}(\boldsymbol{\theta}|\boldsymbol{\phi}), \\ Y_i|\theta_i, N_i &\overset{ind}{\sim} Poisson(N_i\theta_i). \end{aligned}$$

We utilized the algorithm described in Section 2.4 to obtain samples from the posterior distribution of all parameters. Figure **??** displays in red the estimated posterior means (solid lines), with 95% posterior intervals (dashed line), for all of the mortality rates. The posterior means and 95% intervals for the recommended reference prior in Sonksen and Peruggia (2012) are displayed in green. We see that, at least visually, the DP mixture of reference priors provides a better fit. In terms of DIC, the DP mixture of reference priors also outperforms the recommended model.

An interesting exercise is to examine which grouping and orderings have the highest posterior probability. This is allows us to see which reference priors the data prefers and can give us some insight into what the grouping and ordering mean in the reference prior algorithm. Table 2 displays the ten groupings and orderings which have the highest posterior probability. It is of note that the

$\boldsymbol{\theta}_{(1)}$	$\boldsymbol{\theta}_{(2)}$	$\boldsymbol{\theta}_{(3)}$	$\boldsymbol{\theta}_{(4)}$	Rank
$\{\theta_1\}$	$\{\theta_2, \ldots, \theta_{29}\}$	$\{\theta_{30}\}$		1
$\{\theta_1\}$	$\{\theta_{30}\}$	$\{\theta_2, \ldots, \theta_{29}\}$		2
$\{\theta_{30}\}$	$\{\theta_1\}$	$\{\theta_2, \ldots, \theta_{29}\}$		3
$\{\theta_1, \theta_{30}\}$	$\{\theta_2, \ldots, \theta_{29}\}$			4
$\{\theta_{30}\}$	$\{\theta_2, \ldots, \theta_{29}\}$	$\{\theta_1\}$		5
$\{\theta_2, \ldots, \theta_{29}\}$	$\{\theta_1\}$	$\{\theta_{30}\}$		6
$\{\theta_{30}\}$	$\{\theta_1, \ldots, \theta_{29}\}$			7
$\{\theta_2\}$	$\{\theta_1\}$	$\{\theta_3, \ldots, \theta_{29}\}$	$\{\theta_{30}\}$	8
$\{\theta_1, \ldots, \theta_{29}\}$	$\{\theta_{30}\}$			9
$\{\theta_9\}$	$\{\theta_1\}$	$\{\theta_2, \ldots, \theta_8, \theta_{10}, \ldots, \theta_{29}\}$	$\{\theta_{30}\}$	10

Table 2: Posterior Rankings

the third most probable model is the one Sonksen and Peruggia (2012) selected through intuition and DIC.

4.3 One-Way Balanced Random Effects Model

As an example, consider a standard One-Way, balanced, random effects model:

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + \epsilon_{ij} \\ \alpha_i|\tau^2 &\overset{iid}{\sim} N(0, \tau^2) \\ \epsilon_{ij}|\sigma^2 &\overset{iid}{\sim} N(0, \sigma^2) \end{aligned}$$

for $i = 1, 2, \ldots n$, $j = 1, 2, \ldots, k$. Since the random effects ($\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ are by design normally distributed, there are only 3 parameters in need of a prior. The group mean μ, the error variance σ^2, and the random effect variance τ^2.

In this model, there is only 13 possible groupings and orderings. All of which yield a proper posterior distribution. Berger and Bernardo (1992b) gives the form of the reference priors for each possible ordering and grouping.

$$\pi(\mu, \sigma^2, \tau^2|\phi) \propto \begin{cases} \sigma^{-2}\left(n\tau^2 + \sigma^2\right)^{-3/2} & \text{if } \phi = (1,1,1) \\ \sigma^{-5/2}\left(n\tau^2 + \sigma^2\right)^{-1} & \text{if } \phi = (1,1,2) \\ \tau^{-3C_n/2}\sigma^{-2}\psi(\tau^2/\sigma^2) & \text{if } \phi = (1,2,1) \\ \sigma^{-1}\left(n\tau^2 + \sigma^2\right)^{-3/2} & \text{if } \phi = (2,1,2) \\ \tau^{-1}\sigma^{-2}\left(n\tau^2 + \sigma^2\right)^{-1/2}\psi(\tau^2/\sigma^2) & \text{if } \phi = (2,2,1) \\ \sigma^{-2}\left(n\tau^2 + \sigma^2\right)^{-1} & \text{if } \phi = (1,2,2), (2,1,1), (1,2,3), (2,1,3), (3,1,2) \\ \tau^{-C_n}\sigma^{-2}\phi(\tau^2/\sigma^2) & \text{if } \phi = (1,3,2), (2,3,1), (3,2,1) \end{cases}$$

With $C_n = 1 - \sqrt{n-1}\left(\sqrt{n} + \sqrt{n-1}\right)^{-3}$, $\psi(\tau^2/\sigma^2) = \left((n+1) + \left(1 + n\tau^2/\sigma^2\right)^{-2}\right)^{1/2}$ for $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, n$. See **?** for an alternative reference prior based on a reparameterization.

For data, we set $n = 10$ and $m = 3$ and generated α and y from the model in Equation (4) setting $\mu = 2$, $\sigma^2 = 16$ and $\tau^2 = 4$.

5 Discussion

References

Berger, J. O., Bernardo, J., and Sun, D. (2009). The formal definition of reference priors. *Annals of Statistics*, 37:905–938.

Berger, J. O. and Bernardo, J. M. (1992a). On the development of reference priors. *Bayesian Statistics 4*, 4:35–60.

Berger, J. O. and Bernardo, J. M. (1992b). Ordered group reference priors with application to the multinomial problem. *Biometrika*, 79(1):25–37.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 41:113–147.

Broffitt, J. D. (1988). Increasing and increasing convex Bayesian graduation. *Transactions of the Society of Actuaries*, 40:115–148.

Sonksen, M. D. and Peruggia, M. (2012). Reference priors for constrained rate models of count data. *Journal of Statistical Planning and Inference*, To appear:1–25.