Determining the Effective Sample Size of a Parametric Prior

Satoshi Morita[1], Peter F. Thall[2] and Peter Müller[2]

[1]Department of Epidemiology and Health Care Research, Kyoto University
Graduate School of Medicine, Kyoto 606-8501, Japan
satoshi_morita@pbh.med.kyoto-u.ac.jp

[2]Department of Biostatistics, The University of Texas
M. D. Anderson Cancer Center, Houston, TX 77030, U.S.A

SUMMARY. We present a definition for the effective sample size of a parametric prior distribution in a Bayesian model, and propose methods for computing the effective sample size in a variety of settings. Our approach first constructs a prior chosen to be vague in a suitable sense, and updates this prior to obtain a sequence of posteriors corresponding to each of a range of sample sizes. We then compute a distance between each posterior and the parametric prior, defined in terms of the curvature of the logarithm of each distribution, and the posterior minimizing the distance defines the effective sample size of the prior. For cases where the distance cannot be computed analytically, we provide a numerical approximation based on Monte Carlo simulation. We provide general guidelines for application, illustrate the method in several standard cases where the answer seems obvious, and then apply it to some non-standard settings.

KEY WORDS: Bayesian analysis; Effective sample size; Parametric prior distribution; ε-information prior; Computationally intensive methods.

1. Introduction

A fundamental question in any Bayesian analysis is the amount of information contained in the prior. For many commonly used models, the answer seems straightforward. For example, it can be argued that a beta(a, b) distribution has effective sample size (ESS) $a + b$. This is based on the fact that a binomial variable Y from a sample of size n with success probability θ following a beta(a, b) prior implies a beta$(a+Y, b+n-Y)$ posterior. In other words, given a sample of size n, the prior sum $a + b$ becomes the posterior sum $a+b+n$. Thus, saying that a given beta(a, b) prior has ESS $m = a+b$ requires the implicit reasoning that the beta(a, b) may be identified with a beta$(c + Y, d + m - Y)$ posterior arising from a previous beta(c, d) prior having a very small amount of information. A simple way to formalize this is to set $c + d = \varepsilon$ for an arbitrarily small value $\varepsilon > 0$ and solve for $m = a + b - (c + d) = a + b - \varepsilon$.

More generally, one may match a given prior $p(\theta)$ with the posterior $q_m(\theta \mid Y)$ arising from an earlier prior $q_0(\theta)$ that is chosen to be vague in a suitable sense and that was updated by a sample of size m, and consider m to be the ESS of $p(\theta)$. In this general formulation, $p(\theta)$, $q_0(\theta)$ and $q_m(\theta \mid Y)$ play roles analogous to those of the beta(a, b), beta(c, d) and beta$(a+Y, b+n-Y)$ distributions given above. In some cases one may find the hyperparameters of $q_m(\theta \mid Y)$ as a function of m, compare $q_m(\theta \mid Y)$ with $p(\theta)$ and solve for m analytically. For many parametric Bayesian models, however, this analytic approach does not work, and it is not obvious how to determine the ESS of the prior. A simple example is the usual normal linear regression model where the observed response variable Y for predictor X has mean $\beta_0 + \beta_1 X$ and variance σ^2, so that $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$. A traditional, technically convenient prior is that (β_0, β_1) is bivariate normal and σ^2 is inverse chi-squared, with hyperparameters chosen either for computational convenience or by elicitation. In either case, there is no obvious answer to the question of what the

ESS of the prior may be. Moreover, for many commonly used choices of $q_0(\boldsymbol{\theta})$, the joint prior $p(\boldsymbol{\theta})$ cannot be matched with $q_m(\boldsymbol{\theta} \mid Y)$ analytically.

Understanding the prior ESS is important when applying Bayesian methods in settings with a small to moderate sample size. For example, when fitting a Bayesian model to a data set of 10 observations, an *a priori* ESS of 1 is reasonable, whereas a prior ESS of 20 implies that the prior, rather than the data, dominates posterior inferences. If the prior is elicited from a domain expert, then an informative prior is desirable (Chaloner and Rhame, 2001; Garthwaite et al., 2005). In contrast, if the prior is only a technically convenient ad-hoc choice, as is often the case in practice, then understanding the ESS may prompt the investigator to reconsider the prior choice. Thus, it is important to have a good idea of the prior's ESS when interpreting one's inferences. This is especially important from the viewpoint of defending Bayesian methods against the concern that the prior may inappropriately introduce artificial information.

In this paper, we present a definition for the ESS of a prior $p(\boldsymbol{\theta})$ in a Bayesian parametric model, and we provide methods for computing the ESS in a wide variety of settings. Our approach relies on the idea of constructing an "ε-information" prior $q_0(\boldsymbol{\theta})$, considering a sample $\mathbf{Y}$ of size m and the posterior $q_m(\boldsymbol{\theta} \mid \mathbf{Y})$, and computing a distance between $q_m(\boldsymbol{\theta} \mid \mathbf{Y})$ and $p(\boldsymbol{\theta})$ in terms of the curvature (second derivatives) of $\log\{p(\boldsymbol{\theta})\}$ and $\log\{q_m(\boldsymbol{\theta} \mid \mathbf{Y})\}$. The value of m minimizing the distance is the prior ESS. For cases where the distance cannot be computed analytically, we provide a numerical approximation based on Monte Carlo simulations from $q_m(\boldsymbol{\theta} \mid \mathbf{Y})$. In cases where $\boldsymbol{\theta}$ is multivariate, one may compute multiple ESSs, one associated with each of several subvectors of $\boldsymbol{\theta}$.

Section 2 presents a motivating application and defines ε-information priors and ESS. Computational methods are presented in Section 3. Section 4 gives guidelines for using

ESS computations in specific settings. Applications are described in Sections 5 and 6. In Section 7, we discuss connections between our proposed procedures and related methods given by Hodges and Sargent (2001), Spiegelhalter, Freedman and Parmar (1992), and Spiegelhalter, Best, Carlin, and van der Linde (2002).

2. Effective Sample Size

The following example illustrates why it may be useful to determine the ESS of a prior. We consider a design for a phase I trial to determine an optimal dose combination $X = (X_1, X_2)$ of two cytotoxic agents (Thall et al., 2003). The toxicity probability at X is given by the 6-parameter model

$$\pi(X, \boldsymbol{\theta}) = \frac{\alpha_1 X_1^{\beta_1} + \alpha_2 X_2^{\beta_2} + \alpha_3 (X_1^{\beta_1} X_2^{\beta_2})^{\beta_3}}{1 + \alpha_1 X_1^{\beta_1} + \alpha_2 X_2^{\beta_2} + \alpha_3 (X_1^{\beta_1} X_2^{\beta_2})^{\beta_3}}, \tag{1}$$

where all parameters in $\boldsymbol{\theta} = (\alpha_1, \beta_1, \alpha_2, \beta_2, \alpha_3, \beta_3)$ are non-negative. Under this model, if only agent 1 is administered at dose X_1, with $X_2 = 0$, as in a single-agent phase I trial, then $\pi(X, \boldsymbol{\theta}) = \pi_1(X_1, \boldsymbol{\theta}_1) = \alpha_1 X_1^{\beta_1} / (1 + \alpha_1 X_1^{\beta_1})$ only depends on X_1 and $\boldsymbol{\theta}_1 = (\alpha_1, \beta_1)$. Similarly, if $X_1 = 0$ then $\pi(X, \boldsymbol{\theta}) = \pi_2(X_2, \boldsymbol{\theta}_2) = \alpha_2 X_2^{\beta_2} / (1 + \alpha_2 X_2^{\beta_2})$ only depends on X_2 and $\boldsymbol{\theta}_2 = (\alpha_2, \beta_2)$. The parameters $\boldsymbol{\theta}_3 = (\alpha_3, \beta_3)$ characterize interactions that may occur when the two agents are used in combination. The model parameter vector thus is partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$. Since phase I trials of combinations generally require that each agent previously has been tested alone, it is natural to obtain informative priors on $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, but assume a vague prior on $\boldsymbol{\theta}_3$. Denoting by $Ga(a, b)$ the gamma distribution with mean a/b and variance a/b^2, the elicitation process (Thall et al., 2003, section 3) yielded the priors $\alpha_1 \sim Ga(1.74, 4.07)$, $\beta_1 \sim Ga(10.24, 1.34)$ for the effects of agent 1 alone, $\alpha_2 \sim Ga(2.32, 5.42)$, $\beta_2 \sim Ga(15.24, 1.95)$ for the effects of agent 2 alone, and $\alpha_3 \sim Ga(0.33, 0.33)$, $\beta_3 \sim Ga(0.0008, 0.0167)$ for the interaction parameters.

Since doses must be selected sequentially in phase I trials based on very small amounts

of data, an important question is what ESS may be associated with the prior. Our proposed methods (Section 5, below) show that the overall ESS of this prior is $m = 1.5$. However, since informative priors on $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ were obtained and a vague prior on $\boldsymbol{\theta}_3$ was desired, it also is important to determine the prior ESS of each subvector. Applying our proposed methods yielded prior ESSs $m_1 = 547.3$ for $\boldsymbol{\theta}_1$, $m_2 = 756.8$ for $\boldsymbol{\theta}_2$, and $m_3 = 0.01$ for $\boldsymbol{\theta}_3$. The small value for m_3 confirms that the prior on $\boldsymbol{\theta}_3$ reflects little information about the interaction of the two agents. The large numerical discrepancy between $m = 1.5$ and $(m_1, m_2) = (547.3, 756.8)$ is desirable. It reflects the fact that, for each $i = 1, 2$, "$\boldsymbol{\theta}_i$" has a very different meaning in the submodel $\pi_i(X_i, \boldsymbol{\theta}_i)$ parameterized by $\boldsymbol{\theta}_i$ alone versus its meaning in the full 6-parameter model $\pi(X, \boldsymbol{\theta})$. See, for, example, Berger and Pericchi (2001). From a geometric viewpoint, if $\pi(X, \boldsymbol{\theta})$ is thought of as a response surface varying as a function of the two-dimensional dose (X_1, X_2), since the edges of the surface correspond to the submodels $\pi_1(X_1, \boldsymbol{\theta}_1)$ where $X_2 = 0$ and $\pi_2(X_2, \boldsymbol{\theta}_2)$ where $X_1 = 0$, the large values of m_1 and m_2 indicate that the locations of the edges were well known, while the small overall ESS $m = 1.5$ says that otherwise very little was known about the surface. In practice, one would report m_1, m_2, m_3 and m to the clinician from whom the priors were elicted. The clinician could then judge whether m_1 and m_2 are reasonable characterizations of his/her prior information about the single agents, and compare m to the trial's sample size. In the motivating application, a trial of gemcitabine and cyclophosphamide for advanced cancer, the large values of m_1 and m_2 were appropriate since there was substantial clinical experience with each single agent, and the small overall ESS also was appropriate since no clinical data on the two agents used together were available and a sample size of 60 patients was planned.

This example illustrates four key features of our proposed method, namely that (1) ESS is a readily interpretable index of a prior's informativeness, (2) it may be useful

to compute ESSs for both the entire parameter vector and for particular subvectors, (3) ESS values may be used as feedback in the elicitation process, and (4) even when standard distributions are used, it may not be obvious how to define a prior's ESS.

The intuitive motivation for the following construction is to mimic the rationale, given in Section 1, for why the ESS of a $beta(a, b)$ equals $a + b$. As a general Bayesian framework, let $f(Y \mid \boldsymbol{\theta})$ denote the probability distribution function (pdf) of an s-dimensional random vector Y, and let $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ be the prior on the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, where $\tilde{\boldsymbol{\theta}}$ denotes the vector of hyperparameters. The likelihood of an i.i.d. sample $\mathbf{Y}_m = (Y_1, \ldots, Y_m)$ is then given by $f_m(\mathbf{Y}_m \mid \boldsymbol{\theta}) = \prod_{i=1}^m f(Y_i \mid \boldsymbol{\theta})$.

We define an ε-information prior $q_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0)$ by requiring it to have the same mean, $\mathrm{E}_{q_0}(\boldsymbol{\theta}) = \mathrm{E}_p(\boldsymbol{\theta})$, and correlations, $\mathrm{Corr}_{q_0}(\theta_j, \theta_{j'}) = \mathrm{Corr}_p(\theta_j, \theta_{j'})$, $j \neq j'$, as $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$, while inflating the variances of the elements of $\boldsymbol{\theta}$ so that $\mathrm{Var}_{q_0}(\theta_j) \gg \mathrm{Var}_p(\theta_j)$, in such a way that $q_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0)$ has small information but $\mathrm{Var}_{q_0}(\theta_j)$ must exist for $j = 1, \ldots, d$. Table 1 illustrates how to specify $q_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0)$ for several standard parametric priors. Given the likelihood $f_m(\mathbf{Y}_m \mid \boldsymbol{\theta})$ and ε-information prior $q_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0)$, we denote the posterior by $q_m(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_m) \propto q_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0) f_m(\mathbf{Y}_m \mid \boldsymbol{\theta})$ and the marginal distribution under $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ by

$$f_m(\mathbf{Y}_m \mid \tilde{\boldsymbol{\theta}}) = \int f_m(\mathbf{Y}_m \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}) d\boldsymbol{\theta}. \tag{2}$$

When $\tilde{\boldsymbol{\theta}}$ is fixed we write $f(\mathbf{Y_m})$ for brevity. To define the ESS(s), consider the following three cases based on $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$. For implementation, we find it useful to distinguish between these cases although, formally, Cases 1 and 2 are special instances of Case 3.

Case 1: $d = 1$, with $p(\theta \mid \tilde{\boldsymbol{\theta}})$ a univariate parametric model. For this case, we will define one ESS. Examples include the beta, gamma, univariate normal with known variance, and inverse-χ^2 distributions.

Case 2: $d \geq 2$ with $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ a d-variate parametric model. For this case, we will define

one ESS. Examples include the Dirichlet and multivariate normal (MVN) distributions.

The following case deals with settings where it is scientifically appropriate to define two or more ESSs for $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$.

Case 3: $d \geq 2$ with $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ written as a product of K parametric distributions, $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}) = \prod_{k=1}^{K} p_k(\boldsymbol{\theta}_k, \mid \tilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1})$ where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ is partitioned into K subvectors, for $1 < K \leq d$. In this case, a vector of K ESSs, one for each subvector, may be meaningful. An example is a normal-inverse-χ^2 distribution where $(\theta_1, \theta_2) = (\sigma^2, \mu)$, the variance and mean of a normal sampling model, with $p(\theta_1, \theta_2) = p_1(\sigma^2)p_2(\mu \mid \sigma^2)$ and $\sigma^2 \sim Inv\text{-}\chi^2(\tilde{\nu}, \tilde{\sigma}^2)$ and $\mu \mid \sigma^2 \sim N(\tilde{\mu}, \sigma^2/\tilde{\phi})$. Here $K = d = 2$ and the two subvectors of $\boldsymbol{\theta}$ are the single parameters σ^2 and μ. We will discuss other examples in Sections 4 and 5.

To define the distance between $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ and $q_m(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_m)$ in Cases 1 and 2, the basic idea is to find the sample size, m, that would be implied by normal approximations of the prior $p(\boldsymbol{\theta})$ and the posterior $q_m(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_m)$. This led us to use the second derivatives of the log densities to define the distance. The real validation and justification of our definition, however, comes from comparing the resulting ESS values with the commonly reported ESS in standard settings. We carry out these comparisons in Section 5.

Let $\bar{\boldsymbol{\theta}} = \mathrm{E}_p(\boldsymbol{\theta})$ denote the prior mean under $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$. We define

$$D_{p,j}(\boldsymbol{\theta}) \ = -\frac{\partial^2 \log\{p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})\}}{\partial \theta_j^2},$$

and

$$D_{q,j}(m, \boldsymbol{\theta}, \mathbf{Y}_m) \ = -\frac{\partial^2 \log\{\, q_m(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_m)\}}{\partial \theta_j^2}, \qquad j = 1, \ldots, d.$$

Denote $D_{p,+}(\boldsymbol{\theta}) = \sum_{j=1}^{d} D_{p,j}(\boldsymbol{\theta})$ and $D_{q,+}(m, \boldsymbol{\theta}) = \sum_{j=1}^{d} \int D_{q,j}(m, \boldsymbol{\theta}, \mathbf{Y}_m)\ f_m(\mathbf{Y}_m) d\mathbf{Y}_m$. We define a distance between $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ and $q_m(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_m)$ for sample size m as the

difference of the trace of the two information matrices,

$$\delta(m, \bar{\boldsymbol{\theta}}, p, q_0) \;=\; \left| D_{p,+}(\bar{\boldsymbol{\theta}}) - D_{q,+}(m, \bar{\boldsymbol{\theta}}) \right| . \tag{3}$$

That is, we define the distance in terms of the trace of the information matrix (2nd derivative of the log density) of the prior $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$, and the expected information matrix of the posterior $q_m(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_m)$, where the expectation is with respect to the marginal $f_m(\mathbf{Y}_m)$. When $d = 1$, since the '+' subscript is superfluous, we write $D_p(\bar{\boldsymbol{\theta}})$ and $D_q(m, \bar{\boldsymbol{\theta}})$.

DEFINITION 1: *The ESS of* $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ *with respect to the likelihood* $f_m(\mathbf{Y}_m \mid \boldsymbol{\theta})$ *is the integer* m *that minimizes the distance* $\delta(m, \bar{\boldsymbol{\theta}}, p, q_0)$.

Algorithm 1, below, will generalize this to allow non-integer-valued m. An essential point is that the ESS is defined as a property of a prior and likelihood pair, so that, for example, a given prior might have two different ESS values in the context of two different likelihoods.

The definition of the distance (3) involves some arbitrary choices. We chose this definition after an extensive empirical investigation (not shown) of alternative formulations. Instead of evaluating the curvature at the prior mean, one could use the prior mode. Similarly, one could marginalize $\boldsymbol{\theta}$ with respect to the prior, average over $\mathbf{Y}_m$ with respect to $f_m(\mathbf{Y}_m \mid \boldsymbol{\theta})$ rather than the marginal $f_m(\mathbf{Y}_m)$, or use the determinant rather than the trace of the information matrix. One also could define $\delta(.)$ in terms of Kullback-Liebler divergence, or variances. We investigated all of these alternatives and evaluated the resulting ESS in each of several standard cases, and found that the proposed distance (3) was best at matching the results that are commonly used as ESS values.

For Case 3, a more general definition is required. A motivating example is the logistic

regression model, $\text{logit}\{\pi(X,\boldsymbol{\theta})\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ where $d = 3$, $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2)$ and $\beta_j \sim N(\tilde{\mu}_j, \tilde{\sigma}_j^2)$ independently with $\tilde{\boldsymbol{\theta}} = (\tilde{\mu}_j, \tilde{\sigma}_j^2\})$ for $j = 0, 1, 2$. In this case, the subvectors of interest are $\boldsymbol{\theta}_1 = \beta_0$ and $\boldsymbol{\theta}_2 = (\beta_1, \beta_2)$, so two ESS values, m_1 and m_2, may be computed. To accommodate Case 3, we generalize (3) by defining a set of K subvector-specific distances. Let γ_k be the set of indices of the elements of $\boldsymbol{\theta}_k$, and denote $D^k_{p,+}(\boldsymbol{\theta})$ $= \sum_{j\in\gamma_k} D_{p,j}(\boldsymbol{\theta})$ and $D^k_{q,+}(m_k,\boldsymbol{\theta}) = \sum_{j\in\gamma_k} \int D_{q,j}(m_k,\boldsymbol{\theta},\mathbf{Y}_{m_k}) f_{m_k}(\mathbf{Y}_{m_k}) d\mathbf{Y}_{m_k}$. For each $k = 1, \cdots, K$, we define the distance between $p_k(\boldsymbol{\theta}_k \mid \tilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1})$ and $q_{m_k}(\boldsymbol{\theta}_k \mid \tilde{\boldsymbol{\theta}}_{0,k}, \mathbf{Y}_{m_k}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1})$ to be

$$\delta_k(m_k, \bar{\boldsymbol{\theta}}, p, q_0) \;=\; \left| D^k_{p,+}(\bar{\boldsymbol{\theta}}) - D^k_{q,+}(m_k, \bar{\boldsymbol{\theta}}) \right|. \tag{4}$$

DEFINITION 2: *Assume $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ as in Case 3. Let m_k = arg min $\delta_k(m, \bar{\boldsymbol{\theta}}, p, q_0)$. We define $(m_1, \ldots, m_K)$ to be the ESSs for the prior $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ with respect to the model $f_m(\mathbf{Y}_m \mid \boldsymbol{\theta})$ and the partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$.*

3. Computational Methods

Let $\bar{\boldsymbol{\theta}} = (\bar{\theta}_1, \ldots, \bar{\theta}_d)$ denote the prior mean vector. With the following algorithms, we generalize Definitions 1 and 2 to allow non-integer ESS values.

Algorithm 1, for Cases 1 and 2: Let M be a positive integer chosen so that, initially, it is reasonable to assume that $m \leq M$.

Step 1. Specify $q_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0)$.

Step 2. For each $m = 0, \ldots, M$, compute $\delta(m, \bar{\boldsymbol{\theta}}, p, q_0)$.

Step 3. The ESS is the interpolated value of m minimizing $\delta(m, \bar{\boldsymbol{\theta}}, p, q_0)$.

In practice, Step 2 is carried out either analytically or using the simulation-based numerical approximation described in Section 3.2.

Algorithm 2, for Case 3: For each $k = 1, \ldots, K$, let M_k be a positive integer chosen so that, initially, it is reasonable to assume that $m_k \leq M_k$.

Step 1. Specify $q_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0) = \prod_{k=1}^{K} q_{0,k}(\boldsymbol{\theta}_k \mid \tilde{\boldsymbol{\theta}}_{0,k}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1})$.

Step 2. For each $k = 1, \ldots, K$, and $m_k = 0, \ldots, M_k$, compute $\delta_k(m_k, \bar{\boldsymbol{\theta}}, p, q_0)$.

Step 3. The ESS of $\boldsymbol{\theta}_k$ is the interpolated value of m_k minimizing $\delta_k(m_k, \bar{\boldsymbol{\theta}}, p, q_0)$.

If the hyperparameter $\tilde{\boldsymbol{\theta}}$ of $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ includes a *degree of freedom* (*d.f.*) parameter $\tilde{\nu}$, as with an inverse-χ^2, inverse-gamma, inverse-Wishart, or t distribution, then the corresponding hyperparameter of $q_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0)$ is $\tilde{\nu}_0 = \tilde{\nu}_{min} + \varepsilon$, where $\tilde{\nu}_{min}$ is the smallest integer that ensures the second moments of $\boldsymbol{\theta} \sim q_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0)$ exist and $\varepsilon > 0$ is arbitrarily small. In such cases, we add $D_{q,+}(\tilde{\nu}_{min}, \bar{\boldsymbol{\theta}}) - D_{q,+}(0, \bar{\boldsymbol{\theta}})$ to $D_{q,+}(m, \bar{\boldsymbol{\theta}})$ and add $D^k_{q,+}(\tilde{\nu}_{min}, \bar{\boldsymbol{\theta}}) - D^k_{q,+}(0, \bar{\boldsymbol{\theta}})$ to $D^k_{q,+}(m_k, \bar{\boldsymbol{\theta}})$ to ensure that ESS $> \tilde{\nu}_{min}$.

For each $m = 1, \cdots, M$, when $\int D_{q,j}(m, \bar{\boldsymbol{\theta}}, \mathbf{Y}_m) f_m(\mathbf{Y}_m) d\mathbf{Y}_m$ cannot be computed analytically, we use the following simulation-based approximation. Given $\bar{\boldsymbol{\theta}} = \mathrm{E}_p(\boldsymbol{\theta})$, we first simulate Monte Carlo sample $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}$ from $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ for large T, e.g. T = 100,000. For each $t = 1, \ldots, T$, simulate $Y_1^{(t)}, \ldots, Y_M^{(t)}$ from $f_M(\mathbf{Y}_M \mid \boldsymbol{\theta}^{(t)})$. Use the Monte Carlo average $T^{-1} \sum_{t=1}^{T} D_{q,j}(m, \bar{\boldsymbol{\theta}}, \mathbf{Y}_m^{(t)})$ in place of $\int D_{q,j}(m, \bar{\boldsymbol{\theta}}, \mathbf{Y}_m) f_m(\mathbf{Y}_m) d\mathbf{Y}_m$. For Case 3, the same method is used to evaluate $D^k_{q,+}(m_k, \bar{\boldsymbol{\theta}})$ in (4).

For regression models of Y as a function of a u-dimensional predictor X, we extend Definition 1 by augmenting the regression model with a probability distribution $g_m(\mathbf{X}_m \mid \boldsymbol{\xi})$ for the covariates and prior $r(\boldsymbol{\xi} \mid \tilde{\boldsymbol{\xi}})$, usually assuming independence, $g_m(\mathbf{X}_m \mid \boldsymbol{\xi}) = \prod_{i=1}^{m} g(X_i \mid \boldsymbol{\xi})$. Then we define

$$f_m(\mathbf{Y}_m) = \int f_m(\mathbf{Y}_m \mid \mathbf{X}_m, \boldsymbol{\theta})\, g_m(\mathbf{X}_m \mid \boldsymbol{\xi})\, f(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})\, r(\boldsymbol{\xi} \mid \tilde{\boldsymbol{\xi}})\, d\boldsymbol{\theta}\, d\boldsymbol{\xi}.$$

In this case, we simulate $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}$ from $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ and $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(T)}$ from $r(\boldsymbol{\xi} \mid \tilde{\boldsymbol{\xi}})$, then simulate each $X_1^{(t)}, \ldots, X_M^{(t)}$ from $g_M(\mathbf{X}_M \mid \boldsymbol{\xi}^{(t)})$, and $Y_i^{(t)}$ from $f(Y_i \mid \boldsymbol{\theta}^{(t)}, X_i^{(t)})$ for each

$i = 1, \ldots, M$, to obtain $(Y_1^{(t)}, X_1^{(t)}), \ldots, (Y_M^{(t)}, X_M^{(t)})$. Finally, we compute the Monte Carlo average $T^{-1} \sum_{t=1}^{T} D_{q,j}(m, \bar{\boldsymbol{\theta}}, \mathbf{Y}_m^{(t)}, \mathbf{X}_m^{(t)})$. For Case 3, the same method is used to evaluate $D_{q,+}^k(m_k, \bar{\boldsymbol{\theta}})$ in (4).

4. Guidelines for Application

Before illustrating how the above methods for computing ESS may be applied in particular cases, we provide general guidelines for using ESS values in some commonly encountered settings of Bayesian inference.

1. *Prior Elicitation.* When eliciting a prior from an area expert, ESS values may be provided as a readily interpretable form of feedback. The area expert may use this as a basis to modify his/her judgments, if desired, and this process may be iterated. For example, in the motivating example of Section 2, we would report the ESS values $m_1 = 547$ and $m_2 = 756$ to the investigator planning the study. If his/her prior were based on earlier single agent trials with around 100 patients each, (s)he would be prompted to revise the replies to the prior elicitation questions.

2. *Formalizing Uninformative Priors.* Often an investigator wishes to formalize vague prior information. The ESS can be used to confirm that the chosen prior carries little information, as desired. For example, in for the motivating example in section 2.1, the reported ESS $m_3 = 0.01$ for the interaction parameter confirms that this prior is vague.

3. *Reviewing Others' Analyses.* When interpreting or formally reviewing a Bayesian data analysis, the ESS of the analyst's prior provides a tool for evaluating the reasonableness of the analysis. In particular, if it is claimed that a vague or uninformative prior was used, the ESS provides an objective index to evaluate this claim. If appropriate, one may alert the analyst if a prior appears to be overly informative. Similarly, if an informative prior based on historical data is used in the analysis, reporting the ESS enables the reviewer to verify that the prior data is given appropriate weight.

4. *Sensitivity Analyses.* In performing a conventional Bayesian sensitivity analysis in which prior parameters are varied and corresponding posterior values of interest are computed, the ESS of each prior may be computed to enhance interpretation of this analysis. The ESS itself may be used as an index of prior informativeness in such a sensitivity analysis.

5. *Designing Outcome-Adaptive Experiments.* When formulating a prior as part of a Bayesian model to be used in a sequentially outcome-adaptive experiment, the ESS may be used to calibrate the prior to ensure that the data, rather than the prior, will dominate early decisions during the trial.

6. *Reviewing Bayesian Designs.* When interpreting or formally reviewing a Bayesian design, such as that given in a clinical trial protocol, the ESS of the prior provides a tool for determining the extent to which the prior may influence the design's decisions. Currently, an important reservation about using Bayesian inference in a regulatory environment, such as the planning of clinical trial protocols, is the difficulty of evaluating and judging the appropriateness of prior distributions in complex probability models. The ESS provides a useful tool to mitigate such concerns.

5. Validation with Standard Models

We validate the proposed definition of ESS by computing the implied sample sizes in standard models (Table 2) for which commonly reported prior equivalent sample sizes exist. Following Gelman et al. (2004), we denote $Be(\alpha, \beta)$, $Bin(n, \theta)$, $Ga(\alpha, \beta)$, $Exp(\theta)$, $N(\mu, \sigma^2)$, $Inv\text{-}\chi^2(\nu, s^2)$ $Dir(\alpha_1, \ldots, \alpha_J)$, $Mn(n, \theta_1, \ldots, \theta_J)$, and $BeBin(n, \alpha, \beta)$ for the beta, binomial, gamma, exponential, normal, scaled inverse-χ^2, Dirichlet, multinomial and beta-binomial distributions. The corresponding ε-information priors are given in Table 1. For each model in Table 2, the reported ESS matches the obvious choice.

Example 1: *Beta/Binomial Model.* $\delta(m, \bar{\boldsymbol{\theta}}, p, q_0) = \{(\tilde{\alpha} - 1)\bar{\theta}^{-2} + (\tilde{\beta} - 1)(1 - \bar{\theta})^{-2}\}$

$- \{(\tilde{\alpha}/c + \sum_{Y=0}^{m} Y f_m(\mathbf{Y}_m) - 1)\bar{\theta}^{-2} + (\tilde{\beta}/c + m - \sum_{Y=0}^{m} Y f_m(\mathbf{Y}_m) - 1)(1-\bar{\theta})^{-2}\}$, where $f_m(\mathbf{Y}_m) = BeBin(n, \tilde{\alpha}, \tilde{\beta})$ and $\bar{\theta} = \mathrm{E}_p(\theta) = \tilde{\alpha}/(\tilde{\alpha} + \tilde{\beta})$. Figure 1 shows a plot of $\delta(m, \bar{\theta}, p, q_0)$ against m in the case $\tilde{\boldsymbol{\theta}} = (\tilde{\alpha}, \tilde{\beta}) = (3,7)$. Using $\theta \sim Be(3,7)$, the computed ESS is 10, matching the commonly reported ESS in this case. Analogous plots (not shown) in all other cases examined below are very similar in appearance to Figure 1.

Example 2: *Gamma/Exponential Model.* $\delta(m, \bar{\theta}, p, q_0) = (\tilde{\alpha} - 1)\bar{\theta}^{-2} - (\tilde{\alpha}/c + m - 1)\bar{\theta}^{-2}$, where $\bar{\theta} = \tilde{\alpha}/\tilde{\beta}$, and the ESS is found analytically to be $\tilde{\alpha}$, as desired.

Example 3: *Univariate Normal With Known Variance.* For $Y \mid \theta \sim N(\theta, \sigma^2)$ with σ^2 known and prior $\theta \mid \tilde{\boldsymbol{\theta}} \sim N(\tilde{\mu}, \tilde{\sigma}^2)$, so that $\tilde{\boldsymbol{\theta}} = (\tilde{\mu}, \tilde{\sigma}^2)$, one may compute analytically $D_p(\theta) = -\partial^2 \log\{p(\theta \mid \tilde{\boldsymbol{\theta}})\}/\partial\theta^2 = 1/\tilde{\sigma}^2$ and, similarly, $D_q(m, \bar{\theta}) = m/\sigma^2$. Thus, $\delta(m, \bar{\theta}, p, q_0) = |1/\tilde{\sigma}^2 - m/\sigma^2|$, so the ESS $= \sigma^2/\tilde{\sigma}^2$, the ratio of the known variance in the likelihood to the prior variance of θ. In applying this model to a clinical trial setting where θ is the difference between two treatment effects, Spiegelhalter, Freedman and Parmar (1994, section 3.1.2) propose assuming that $\tilde{\sigma}^2 = \sigma^2/n_0$ to obtain a prior that "... is equivalent to a normalized likelihood arising from a (hypothetical) trial of n_0 patients with an observed value $\tilde{\mu}$ of the treatment difference statistic." Thus, in this case, the two methods for defining prior effective sample size agree.

Example 4: *Inverse-χ^2/Normal Model.* We find analytically that $D_p(\theta) = -(\sigma^2)^{-2}(\tilde{\nu} + 2)/2 + (\sigma^2)^{-3}\tilde{\nu}\tilde{\sigma}^2$, while $\int D_q(m, \bar{\theta}, \mathbf{Y}_m) f_m(\mathbf{Y}_m) d\mathbf{Y}_m$ is obtained by simulation. As explained in Section 3, the adjustment factor $\{D_q(4, \bar{\theta}) - D_{q,}(0, \bar{\theta})\}$ is added to $D_q(m, \bar{\theta})$. For $\tilde{\boldsymbol{\theta}} = (\tilde{\nu}, \tilde{\sigma}^2) = (20,1)$, ESS $= 20 = \tilde{\nu}$, as desired.

Example 5: *Dirichlet/Multinomial Model.* From Table 1, denote $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_J)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ and $\mathbf{S} = (S_1, \ldots, S_J)$ with $S_j = \sum_{i=1}^{m} Y_{ji}$. Compute $D_{q,j}(m, \theta)$ analytically, as with the beta−binomial. For $d = 3$ and $\tilde{\boldsymbol{\theta}} = (10,\ 15,\ 25)$, ESS $= 50 = \sum \tilde{\alpha}_j$, as desired.

Example 6: *Power Priors.* Ibrahim and Chen (2000) propose a class of "power priors" based on an initial prior $p_0(\boldsymbol{\theta} \mid c_0)$, a likelihood $L(\boldsymbol{\theta} \mid D_0)$ of historical data D_0, and a scalar prior parameter a_0. The power prior is $p(\boldsymbol{\theta} \mid D_0, a_0) \propto L(\boldsymbol{\theta} \mid D_0)^{a_0} p_0(\boldsymbol{\theta} \mid c_0)$, so that a_0 weights the historical data relative to the data that will be obtained in the future. To see how one would compute the ESS of a power prior, consider the beta/binomial model with a beta(1,1) initial prior and D_0 consisting of 3 successes in 10 historical trials. The power prior is $p(\theta \mid D_0, a_0) = p(\theta \mid (3, 10), a_0) \propto \{\theta^3(1-\theta^7)\}^{a_0}\theta(1-\theta)$, and it follows easily (Case 1) that ESS $= a_0 10 + 2$. More generally, the ESS of $p(\boldsymbol{\theta} \mid D_0, a_0)$ is a_oESS$\{L(\boldsymbol{\theta} \mid D_0)\}$ + ESS$\{p_0(\boldsymbol{\theta} \mid c_0)\}$, the weighting parameter times the ESS of the historical data likelihood treated as a function of $\boldsymbol{\theta}$ plus the ESS of the initial prior.

Hodges and Sargent (2001) derive a formula for the effective degrees of freedom (EDF) of a richly parameterized model, and illustrate this for a balanced one-way normal linear random effects model for Nn observations $\{Y_{ij}, i = 1, \ldots, N, j = 1, \ldots, n\}$, given by the likelihood $Y_{i1}, \cdots Y_{in} \mid \theta_i, \sigma^2 \sim$ iid $N(\theta_i, \sigma^2)$ for each i, and prior $\theta_1, \cdots, \theta_N \mid \tilde{\mu}, \tilde{\sigma}^2$ $\sim$ iid $N(\tilde{\mu}, \tilde{\sigma}^2)$. They show that the EDF for this model is $\rho = (nN+\phi)/(n+\phi)$, where ϕ $= \sigma^2/\tilde{\sigma}^2$, the ratio of the residual variance and the prior variance. Recall from Example 3 that ϕ is the ESS of the simple normal model with known variance. In the limiting case with $\phi \to \infty$, i.e., all θ_i are equal, $\theta_i = \mu$, we find $\rho = 1$. In other words, for large ESS and essentially only one group, Hodges and Sargent report $\rho \approx 1$. At the other extreme, for $\phi \to 0$, i.e., for small ESS and θ_i's very different from each other, they report $\rho \approx N$. However, such comparisons should not be overinterpreted. EDF and ESS are quite different summaries. Formally, the EDF is a function of the sample size n. In contrast, ESS is not a function of n. Rather it reports an equivalent sample size for the given model.

Using an information theoretic argument, Spiegelhalter, Best, Carlin and van der

Linde et al. (2002) also derive a measure for the effective number of parameters in complex models, such as generalized linear (mixed effects) models, p_D, defined as the difference between the posterior mean of the deviance and the deviance evaluated at the posterior means of the parameters of interest. But, similar to the effective degrees of freedom ρ, the nature of p_D is different from the proposed ESS. Formally, p_D is a function of the data, while the ESS is not.

6. Application to Some Non-Standard Cases

The following examples show how ESS values may be computed in settings where no commonly agreed upon ESS exists, using the numerical approximations described earlier to obtain $\delta(m, \bar{\boldsymbol{\theta}}, p, q_0)$.

Example 7: *Logistic Regression.* Thall and Lee (2003) use a logistic regression model to determine a maximum tolerable dose in a phase I clinical trial. Each patient receives one of six doses 100, 200, 300, 400, 500, 600 mg/m^2, denoted by $d_1, \ldots, d_6$, with standardized doses $X_{(z)} = \log(d_z) - 6^{-1} \sum_{l=1}^{6} \log(d_l)$. The outcome variable is the indicator $Y_i = 1$ if a patient i suffers toxicity, 0 if not. A logistic model $\pi(X_i, \boldsymbol{\theta}) = \Pr(Y_i = 1 \mid X_i, \boldsymbol{\theta}) = \text{logit}^{-1}\{\eta(X_i, \boldsymbol{\theta})\}$ with $\eta(X_i, \boldsymbol{\theta}) = \mu + \beta X_i$ is assumed, where $\text{logit}^{-1}(x) = e^x/(1 + e^x)$. Hence $d = 2$, $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\mu, \beta)$, and the likelihood for m patients is

$$f_m(\mathbf{Y}_m \mid \mathbf{X}_m, \boldsymbol{\theta}) \; = \prod_{i=1}^{m} \pi(X_i, \boldsymbol{\theta})^{\,Y_i} \{1 - \pi(X_i, \boldsymbol{\theta})\}^{1-Y_i}.$$

Thall and Lee (2003) obtained independent normal priors for μ and β, based on elicited mean $\pi(X, \boldsymbol{\theta})$ for $d = 200$ and 500, and setting $\tilde{\sigma}_\mu = \tilde{\sigma}_\beta = 2$ based on preliminary sensitivy analyses, which yielded $N(\tilde{\mu}_\mu, \tilde{\sigma}_\mu^2) = N(-0.1313, 2^2)$ and $N(\tilde{\mu}_\beta, \tilde{\sigma}_\beta^2) = N(2.3980, 2^2)$. For this application, Algorithms 1 and 2 may be applied to compute one ESS of $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ and two ESSs m_μ and m_β of the priors for μ and β, as follows. For Step 1, specify $q_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0) = N(\tilde{\mu}_\mu, c\tilde{\sigma}_\mu^2)\; N(\tilde{\mu}_\beta, c\tilde{\sigma}_\beta^2)$, with $c = 10{,}000$. Next, compute

$D_{p,1}(\boldsymbol{\theta}) = (\tilde{\sigma}_\mu^2)^{-1}$, $D_{p,2}(\boldsymbol{\theta}) = (\tilde{\sigma}_\beta^2)^{-1}$, $D_{q,1}(m, \boldsymbol{\theta}, \mathbf{Y}_m, \mathbf{X}_m) = \sum_{i=1}^m \pi(X_i, \boldsymbol{\theta})\{1 - \pi(X_i, \boldsymbol{\theta})\}$, and $D_{q,2}(m, \boldsymbol{\theta}, \mathbf{Y}_m, \mathbf{X}_m) = \sum_{i=1}^m X_i^2\, \pi(X_i, \boldsymbol{\theta})\{1 - \pi(X_i, \boldsymbol{\theta})\}$. Since $D_{q,1}(m, \boldsymbol{\theta}, \mathbf{Y}_m, \mathbf{X}_m)$ and $D_{q,2}(m, \boldsymbol{\theta}, \mathbf{Y}_m, \mathbf{X}_m)$ depend on $\mathbf{X}_m$ but not on $\mathbf{Y}_m$, this simplifies the simulation method given in Section 3.2. We assume a uniform distribution on the six doses for the probability model $g(X_i \mid \boldsymbol{\xi})$. Draw $X_1^{(t)}, \ldots, X_M^{(t)}$ independently from $\{X_{(1)}, \ldots, X_{(6)}\}$ with probability 1/6 each, for $t = 1, \ldots, 100,000$. Then, using the plug-in vector $\bar{\boldsymbol{\theta}} = (\bar{\mu}, \bar{\beta}) = (\tilde{\mu}_\mu, \tilde{\mu}_\beta)$, compute $\delta(m, \bar{\boldsymbol{\theta}}, p, q_0)$ for each $m = 0, \ldots, M$, $\delta_1(m_\mu, \bar{\boldsymbol{\theta}}, p, q_0)$ for each $m_\mu = 0, \ldots, M_1$, and $\delta_2(m_\beta, \bar{\boldsymbol{\theta}}, p, q_0)$ for each $m_\beta = 0, \ldots, M_2$. As shown in Table 3, m = 2.3, m_μ = 1.4 and m_β =6.3.

Since the standardized doses X_i were defined to be centered at 0, one may interpret m_μ as the ESS for the prior on the average effect, and m_β as the ESS for the dose effect. The prior indicates greater knowledge about the effects of the doses than about the average response. Since m = 2.3, after enrolling 3 patients, the information from the likelihood starts to dominate the prior, as desired.

As a sensitivity analysis, Table 3 summarizes corresponding results for $\tilde{\sigma}_\mu^2 = \tilde{\sigma}_\beta^2 = 0.5^2$, 1.0^2, 3.0^2, and 5.0^2. As a basis for comparison, we also include the ESS at each dose obtained by the crude method of equating the mean and variance of $\pi(X_{(z)}, \boldsymbol{\theta})$ at each dose to the corresponding values for a beta, $\mathrm{E}(\theta) = \tilde{\alpha}/(\tilde{\alpha} + \tilde{\beta})$ and $\mathrm{Var}(\theta) = \{\mathrm{E}(\theta)(\,1 - \mathrm{E}(\theta)\,)\}\,/(\tilde{\alpha} + \tilde{\beta} + 1)$, and solving for $\tilde{\alpha} + \tilde{\beta}$. We denote by $\bar{m}$ the average of the ESSs $m_{X_{(1)}}, \ldots, m_{X_{(6)}}$ at the six doses, obtained in this way. The results indicate that the crude method provides smaller estimates of the ESS for $\tilde{\sigma}^2 < 5.0^2$.

It also is useful to examine how the ESS in this example would vary with a_0 if one wished to re-weight the prior by replacing it with a power prior $\{p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})\}^{a_0}$. Identifying $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$ with $L(\boldsymbol{\theta} \mid D_0)$ in the set-up of Ibrahim and Chen (2000), and considering the additional ESS of an initial prior to be negligible, the ESS may be computed by applying

Algorithms 1 and 2 and setting the ϵ-information prior to be $\{q_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0)\}^{a_0}$. This yields the values summarized in Table 4. These values illustrate, as in Example 6 given earlier, that the power a_0 acts essentially as a multiplier in the ESS domain, aside from the additive ESS of an initial prior.

Example 8: *Two-Agent Dose-Response Model:* The next example is that described earlier in Section 2, a design to find acceptable dose combinations of two cytotoxic agents used together in a phase I trial. Recall the definition of $\pi(X, \boldsymbol{\theta})$ given in equation (1). The likelihood for m patients with toxicity indicators $\mathbf{Y_m} = (Y_1, \ldots, Y_m)$ and dose pairs $\mathbf{X_m} = (X_1, \ldots, X_m)$ is

$$f(\mathbf{Y}_m \mid \mathbf{X}_m, \boldsymbol{\theta}) \ = \prod_{i=1}^{m} \pi(X_i, \boldsymbol{\theta})^{Y_i} \{1 - \pi(X_i, \boldsymbol{\theta})\}^{1-Y_i}. \tag{5}$$

Based on (5) and the gamma priors given in Section 2, for this case, Algorithm 1 is used to compute one ESS, m, of $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$. The three ESSs m_1, m_2 and m_3 for $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ can be computed using Algorithm 2. In Step 1, with c = 10,000, $q_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0) = \prod_{k=1}^{3}$ $Ga(\tilde{a}_{k,1}/c, \tilde{a}_{k,2}/c)$ $Ga(\tilde{b}_{k,1}/c, \tilde{b}_{k,2}/c)$. In Step 2, we computed $D_{p,1}(\boldsymbol{\theta}) = (\tilde{a}_{1,1} - 1)\alpha_1^{-2}$, $\ldots$, $D_{p,6}(\boldsymbol{\theta}) = (\tilde{b}_{3,1} - 1)\beta_3^{-2}$ analytically. The numerical methods given in Section 3 give $\delta_k(m_k, \bar{\boldsymbol{\theta}}, p, q_0)$ for k = 1, 2, 3, yielding the values $m = 1.5$, $m_1 = 547.3$, $m_2 = 756.8$, and $m_3 = 0.01$, as reported earlier.

Example 9: *Linear Regression.* The last example is a linear regression model used to analyze a small data set $(Y_1, X_1), \cdots, (Y_{10}, X_{10})$ where Y_i is December rainfall and X_i is November rainfall for ten consecutive years $i = 1, \ldots, 10$ (Congdon, 2001). The sampling model is $Y_i \mid X_i, \boldsymbol{\theta} \sim N(\mu_i, 1/\tau)$ with $\mu_i = \alpha + \beta(X_i - \bar{X})$ and τ denoting the precision where $\bar{X}$ is the sample average of the original predictor, so $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) = (\alpha, \beta, \tau)$. Let $N(x;\, m, s)$ indicate that the random variable x is normally distributed with moments (m, s). In Congdon (2001), an independent prior $p(\boldsymbol{\theta}) = p_1(\theta_1, \theta_2 \mid \tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2) \cdot p_2(\theta_3 \mid \tilde{\boldsymbol{\theta}}_3)$

is assumed, with $p_1(\theta_1, \theta_2) = N(\theta_1;\, \tilde{\mu}_\alpha, \tilde{\sigma}^2_\alpha) \cdot N(\theta_2;\, \tilde{\mu}_\beta, \tilde{\sigma}^2_\beta)$ and $p_2 = Ga(\tilde{a}, \tilde{b})$. Congdon (2001) uses $\tilde{\mu}_\alpha = \tilde{\mu}_\beta = 0$, $\tilde{\sigma}^2_\alpha = \tilde{\sigma}^2_\beta = 1000$, $\tilde{a} = \tilde{b} = 0.001$. Algorithm 2 was used to compute two ESSs: m_1 for $p_1(\theta_1, \theta_2 \mid \tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2)$ and m_2 of $p_2(\theta_3 \mid \tilde{\boldsymbol{\theta}}_3)$. The plug-in vector is $\bar{\boldsymbol{\theta}} = \mathrm{E}_p(\boldsymbol{\theta}) = (\tilde{\mu}_\alpha, \tilde{\mu}_\beta, \tilde{a}/\tilde{b})$. In Step 1, specify $q_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0) = q_{0,1}(\theta_1 \mid \tilde{\boldsymbol{\theta}}_{0,1})$ $q_{0,1}(\theta_2 \mid \tilde{\boldsymbol{\theta}}_{0,2})$ $q_{0,2}(\theta_3 \mid \tilde{\boldsymbol{\theta}}_{0,3}) = N(\tilde{\mu}_\alpha, c\tilde{\sigma}^2_\alpha)$ $N(\tilde{\mu}_\beta, c\tilde{\sigma}^2_\beta)$ $Ga(\tilde{a}/c, \tilde{b}/c)$, with $c = 10{,}000$. In Step 2, compute analytically $D_{p,1}(\boldsymbol{\theta}) = (\tilde{\sigma}^2_\alpha)^{-1}$, $D_{p,2}(\boldsymbol{\theta}) = (\tilde{\sigma}^2_\beta)^{-1}$, $D_{p,3}(\boldsymbol{\theta}) = (\tilde{a} - 1)\tau^{-2}$, $D_{q,1}(m_1, \boldsymbol{\theta}, \mathbf{Y}_{m_1}, \mathbf{X}_{m_1}) = (c\tilde{\sigma}^2_\alpha)^{-1} + \tau m_1$, and $D_{q,3}(m_2, \boldsymbol{\theta}, \mathbf{Y}_{m_2}, \mathbf{X}_{m_2}) = (\tilde{a}/c - 1)\tau^{-2} + m_2\tau^{-2}/2$. For this case, only $D_{q,2}(m_1, \boldsymbol{\theta}, \mathbf{Y}_{m_1}, \mathbf{X}_{m_1}) = (c\tilde{\sigma}^2_\beta)^{-1} + \tau \sum_{i=1}^{10} X_i^2$ depends on $\mathbf{X}$. Following the methods in Section 3, we simulated $X_1^{(t)}, \ldots, X_{M_1}^{(t)} \sim$ i.i.d. $N(0, 1)$ for $t = 1, \ldots, 100{,}000$ to obtain $m_1 = 0.001$ and $m_2 = 0.002$. We interpret the reported ESSs as evidence of very vague priors. As a sensitivity analysis, we also computed the ESSs of two alternative priors $p'(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}) = N(0, 100)$ $N(0, 10)$ $Ga(1, 1)$ and $p''(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}) = N(0, 1)$ $N(0, 1)$ $Ga(2, 2)$, which gave $m_1 = 0.06$ and $m_2 = 2.0$ for $p'(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$, and $m_1 = 1.0$ and $m_2 = 4.0$ for $p''(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$.

7. Discussion

The methods proposed in this paper are useful in Bayesian analysis, particularly in settings with elicited priors or where the data consist of a relatively small number of observations. By computing ESSs, one may avoid the use of an overly informative prior in the sense that inference is dominated by the prior rather than the data. As noted in our guidelines for application, other uses of ESS values include interpreting or reviewing others' Bayesian analyses or designs, using the ESS values themselves to perform sensitivity analyses in the prior's informativeness, and calibrating the parameters of outcome-adaptive Bayesian designs.

Extension of our methods to accommodate hierarchical models is not straightforward. This is a potentially important area for future research, since it would be useful

to compute ESS values in such settings. Other potential applications involving more complicated problems include mixture priors synthesizing multiple component priors, or the class of ε-contaminated priors where ε reflects the amount of uncertainty in the prior information (Greenhouse and Wasserman, 1995).

Acknowledgments

Satoshi Morita's work was supported in part by Grant H16-TRANS-003 from the Ministry of Health, Labour, and Welfare in Japan. Peter Thall's work was partially supported by NCI grant RO1 CA 83932. Peter Muller's work was partially supported by NCI grant R01 CA 075981. We thank the associate editor and the referee for their thoughtful and constructive comments and suggestions.

References

Berger, J. and Pericchi, L. (2001) Objective Bayesian methods for model selection: introduction and comparison (with discussion). In 'Model Selection' (P. Lahiri, ed.), Inst. of Math. Stat. Lecture Notes, Monograph Series, vol. 38.

Chaloner, K. and Rhame, F. S. (2001). Quantifying and documenting prior beliefs in clinical trials. *Statistics in Medicine* **20**, 581-600.

Congdon, P. (2001). *Bayesian Statistical Modelling*. Chichester: John Wiley and Sons, pp. 94-99.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association* **100**, 680-701.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990). Illustration

of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* **85**, 972-985.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis* (*2nd Edition*). New York: Chapman and Hall/CRC.

Greenhouse, J. and Wasserman, L. (1995) Rob8ust Bayesain methods for monitoring clinical trials. *Statistics in Medicine* **14**, 1379-1391.

Hodges J.S. and Sargent, D.J. (2001) Counting degrees of freedom in hierarchical and other richly-parameteriazed models. *Biometrika* **88**, 367-379.

Ibrahim, J. G., and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science* **15**, 46-60.

Spiegelhalter, D.J., Freedman, L.S. and Parmar, M.K.B. (1994) Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society, Series A* **157**, 357-416.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* **64**, 583-639.

Thall, P. F., and Lee, S. J. (2003) Practical model-based dose-finding in phase I clinical trials: methods based on toxicity. *Int J Gynecol Cancer* **13**, 251-261.

Thall, P. F., Millikan, R.E., Mueller, P., and Lee, S. J. (2003). Dose-finding with two agents in phase I oncology trials. *Biometrics* **59**, 487-496.

Table 1. *Examples of ε-information prior distributions. The hyperparameters c, c_1, and c_2 are very large constants chosen to inflate the variances of the elements of $\boldsymbol{\theta}$ under q_0.*

d	Distribution	$p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$	$q_0(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_0)$
1	*Beta*	$Be(\tilde{\alpha}, \tilde{\beta})$	$Be(\tilde{\alpha}/c, \tilde{\beta}/c)$
1	*Gamma*	$Ga(\tilde{\alpha}, \tilde{\beta})$	$Ga(\tilde{\alpha}/c, \tilde{\beta}/c)$
1	*Univariate normal with known variance*	$N(\tilde{\mu}, \tilde{\sigma}^2)$	$N(\tilde{\mu}, c\tilde{\sigma}^2)$
1	*Scaled inverse-χ^2*	*Inv-*$\chi^2(\tilde{\nu}, \tilde{\sigma}^2)$	*Inv-*$\chi^2\left(4 + c^{-1}, \tilde{\nu}\tilde{\sigma}^2/2(\tilde{\nu} - 2)\right)$
2	*Normal-inverse-χ^2*	$N(\tilde{\mu}, \tilde{\sigma}^2/\tilde{\phi}) * Inv\text{-}\chi^2(\tilde{\nu}, \tilde{\sigma}^2)$	$N(\tilde{\mu}, c\tilde{\sigma}^2/\tilde{\phi}) *$ *Inv-*$\chi^2\left(4 + c^{-1}, \tilde{\nu}\tilde{\sigma}^2/2(\tilde{\nu} - 2)\right)$
3	*Dirichlet*	$Dir(\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3)$	$Dir(\tilde{\alpha}_1/c, \tilde{\alpha}_2/c, \tilde{\alpha}_3/c)$
3	*Multivariate normal*	$MVN\left(\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\sigma}_1^2, \tilde{\sigma}_2^2, \tilde{\sigma}_{12}\right)$	$MVN\left(\tilde{\mu}_1, \tilde{\mu}_2, c_1^2\tilde{\sigma}_1^2, c_2^2\tilde{\sigma}_2^2, c_1c_2\tilde{\sigma}_{12}\right)$

Table 2. *Prior, likelihood and corresponding posterior q_m with respect to the ε-information prior q_0, and traditionally reported prior effective sample size, ESS, for some common models. In line three, we denote $s^2 = \sum_{i=1}^{m}(Y_i - \tilde{\nu}_0)^2$.*

$p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$	$f(\mathbf{Y}_m \mid \boldsymbol{\theta})$	$q_m(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}, \mathbf{Y}_m)$	ESS
$Be(\tilde{\alpha}, \tilde{\beta})$	$Bin(n, \theta)$	$Be(c^{-1}\tilde{\alpha} + Y,\ c^{-1}\tilde{\beta} + m - Y)$	$\tilde{\alpha} + \tilde{\beta}$
$Ga(\tilde{\alpha}, \tilde{\beta})$	$Exp(\theta)$	$Ga(c^{-1}\tilde{\alpha} + m,\ c^{-1}\tilde{\beta} + \sum Y_i)$	$\tilde{\alpha}$
Inv-$\chi^2(\tilde{\nu}, \tilde{\sigma}^2)$	$N(0, \sigma^2)$	Inv-$\chi^2(\tilde{\nu}_0 + m,\ \frac{\tilde{\nu}_0\tilde{\sigma}^2 + s^2}{\tilde{\nu}_0 + m})$	$\tilde{\nu}$
$Dir(\tilde{\boldsymbol{\alpha}})$	$Mn(n, \boldsymbol{\theta})$	$Dir(c^{-1}\tilde{\boldsymbol{\alpha}} + \mathbf{S})$	$\sum \tilde{\alpha}_j$

Table 3. *Comparison of ESSs computed using the proposed method and the crude method that matches first and second moments to a beta, for the logistic regression model,* $\pi(X_i, \boldsymbol{\theta}) = Pr(Y_i = 1 \mid X_i, \boldsymbol{\theta}) = exp\,(\mu + \beta X_i)\ /\ \{1 + exp(\mu + \beta X_i)\}$*, where the priors are* $\mu \sim N(\tilde{\mu}_\mu, \tilde{\sigma}^2_\mu)$ *with* $\tilde{\mu}_\mu = -0.1313$ *and* $\beta \sim N(\tilde{\mu}_\beta, \tilde{\sigma}^2_\beta)$ *with* $\tilde{\mu}_\beta = 2.3980$.

	Proposed method			*Crude method*						
$\tilde{\sigma}^2_\mu = \tilde{\sigma}^2_\beta$	m	m_μ	m_β	$\bar{m}^*$	$m_{X_{(1)}}$	$m_{X_{(2)}}$	$m_{X_{(3)}}$	$m_{X_{(4)}}$	$m_{X_{(5)}}$	$m_{X_{(6)}}$
0.5^2	37.1	22.7	101.3	18.2	23.5	18.2	17.0	16.6	16.8	17.3
1.0^2	9.3	5.7	25.3	4.5	4.1	4.7	4.8	4.6	4.4	4.2
2.0^2	2.3	1.4	6.3	1.3	1.0	1.4	1.5	1.5	1.3	1.2
3.0^2	1.0	0.6	2.8	0.7	0.5	0.8	0.8	0.8	0.7	0.7
5.0^2	0.4	0.2	1.0	0.4	0.3	0.4	0.4	0.4	0.4	0.3

$*\ \bar{m}\ = 6^{-1} \sum_{z=1}^{6} m_{X_{(z)}}$.

Table 4. *ESSs for power priors* $\{p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})\}^{a_0}$ *based on the prior* $\{p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})\}$ *in the logistic regression example, using hyperparameter values* $\tilde{\mu}_\mu = -0.1313$ *and* $\tilde{\mu}_\beta = 2.3980$*, as in Table 3, with* $\tilde{\sigma}^2_\mu = \tilde{\sigma}^2_\beta = 4$.

a_0	m	m_μ	m_β
0.5	1.2	0.7	3.2
1	2.3	1.4	6.3
2	4.6	2.8	12.6
4	9.3	5.7	25.3

Figure Legends

Figure 1. *Plot of* $\delta(m, \bar{\theta}, p, q_0)$ *against* m *for the beta/binomial model with* $\tilde{\boldsymbol{\theta}} = (\tilde{\alpha}, \tilde{\beta}) = (3, 7)$.