Statistical Science
2002, Vol. 17, No. 2, 193–208

A Simulation-based Approach to Bayesian Sample Size Determination for Performance under a Given Model and for Separating Models

Fei Wang and Alan E. Gelfand

Abstract. Sample size determination (SSD) is a crucial aspect of experimental design. Two SSD problems are considered here. The first concerns how to select a sample size to achieve specified performance with regard to one or more features of a model. Adopting a Bayesian perspective, we move the Bayesian SSD problem from the rather elementary models addressed in the literature to date in the direction of the wide range of hierarchical models which dominate the current Bayesian landscape. Our approach is generic and thus, in principle, broadly applicable. However, it requires full model specification and computationally intensive simulation, perhaps limiting it practically to simple instances of such models. Still, insight from such cases is of useful design value. In addition, we present some theoretical tools for studying performance as a function of sample size, with a variety of illustrative results. Such results provide guidance with regard to what is achievable. We also offer two examples, a survival model with censoring and a logistic regression model.

The second problem concerns how to select a sample size to achieve specified separation of two models. We approach this problem by adopting a screening criterion which in turn forms a model choice criterion. This criterion is set up to choose model 1 when the value is large, model 2 when the value is small. The SSD problem then requires choosing n_1 to make the probability of selecting model 1 when model 1 is true sufficiently large and choosing n_2 to make the probability of selecting model 2 when model 2 is true sufficiently large. The required n is $\max(n_1, n_2)$. Here, we again provide two illustrations. One considers separating normal errors from t errors, the other separating a common growth curve model from a model with individual growth curves.

Key words and phrases: Average posterior variance criterion, fitting and sampling priors, linear and generalized linear models, random effects models, Bayes factor, likelihood and penalized likelihood, screening criterion.

1. INTRODUCTION

Experimental design is a multifaceted activity but, undeniably, sample size determination (SSD) is a crucial aspect. There is by now a substantial literature in the classical case which is summarized roughly through the 1980s in books such as Kraemer and Thiemann (1987), Cohen (1988) and Desu and Raghavarao (1990). More recent work has moved toward general

Fei Wang is Assistant Professor in the Health Services Department of the School of Public Health, Boston University, Boston, Massachusetts 02118 (e-mail: feiwang@bu.edu). Alan E. Gelfand is Professor, Department of Statistics at the University of Connecticut, Storrs, Connecticut 06269 (e-mail: alan@stat.duke.edu).

regression settings and generalized linear models, for example, Self and Mauritsen (1988), Self, Mauritsen and O'Hara (1992) and Muller, LaVange, Ramey and Ramey (1992). A recent sophisticated illustration in the context of longitudinal data studies appears in Liu and Liang (1997).

EXAMPLE 1. Consider the usual normal theory linear regression setting. That is, assume $\mathbf{y} = \mathbf{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{X}_n$ is subscripted to denote sample size and is $n \times p$, $\boldsymbol{\beta}$ is $p \times 1$ and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$. SSD may be sought to address specified inferential performance. However, the classical SSD approach described below restricts inferential performance to hypothesis testing regarding linear transformation of $\boldsymbol{\beta}$, requiring a null and alternative hypothetical value. In practice, for $\phi = h(\boldsymbol{\beta})$ we may seek to control the performance of an interval estimate for ϕ or, if ϕ is viewed as an effect size, to assess something about the chance that $\phi > \phi^*$, that is, of detecting an effect of size at least ϕ^*. We return to this example in Section 5.

A generic strategy when interest is confined to a single parameter, say θ, assumes an estimator $\hat{\theta}$ of θ which is consistent and approximately normal. Then, with a specified null and alternative along with an estimated standard error for $\hat{\theta}$, we have a routine "one sample normal" calculation for sample size n which should perform reasonably well if the resulting n is reasonably large. This approach is advocated by Shuster (1993).

The general classical strategy in the situation of more complicated parametric modeling structure is as follows. In the case of Gaussian data, for an estimable parameter of interest, the usual associated F statistic will have a null distribution which is a central F. Typically, at an alternative value for the parameter, the resulting noncentrality parameter is increasing in sample size n. Then, a sample size is determined such that, at this n, the probability of rejecting the null achieves a prescribed level. For generalized linear models, under conditions, likelihood ratio tests and score tests have approximate chi-square distributions which are central under the null and, under alternatives, approximately noncentral. The noncentrality parameter can be approximated using the sample or estimated expected information matrix along with the familiar delta method. Again, if this parameter is increasing in n we can proceed as above.

The limitations of the classical approach are evident. In simpler cases, for example, the binomial, one needs an estimate of the quantity for which performance is desired in order to obtain the required sample size. In other cases, one needs an estimate of the variability in the data or else of the standard error of the parameter estimate. In the general case, one needs a value of the parameter vector in order to calculate the noncentrality parameter as a function of sample size. Where does one obtain these numbers from? Moreover, should these numbers merely be *inserted* into an SSD formula without some recognition of their uncertainty or variability? How comfortable are we with the various approximations which are implicit in the resulting SSD formula? Perhaps, most importantly, in the general case is one interested exclusively in performance which is measured through power at an alternative value of a noncentrality parameter?

In this sense, a Bayesian approach may be more attractive, but the literature on Bayesian SSD is more recent and more narrow, focusing primarily on normal and binomial one- and two-sample problems. A recent issue of *The Statistician* (**46** 2, 1997) summarizes this work. In particular, see the articles by Lindley (1997), Pham-Gia (1997), Adcock (1997), Joseph and Belisle (1997) and Joseph and Wolfson (1997). An earlier issue of *The Statistician* (**44** 2, 1995) is also of interest, with papers by Joseph, Wolfson and Du Berger (1995a, b) and discussions by Pham-Gia (1995) and Adcock (1995). The most recent work is that of Rahme, Joseph, and Gyorkos (2000) and Inoue, Berry and Parmigiani (2000).

This effort illuminates two primary issues. The first is the distinction between a formal utility approach which provides SSD through a maximization of expected utility [Lindley (1997)] and a performance based approach which chooses SSD to control inference for a parameter of interest to a specified degree of error. The other issue is an elaboration of a variety of performance measures and their respective advantages and disadvantages.

Our contribution here is not to join the debate over the first issue nor is it to argue for a particular class of measures for the second issue. Rather, it is to move the Bayesian SSD problem forward to handle the range of models that classical SSD work has been addressing over the past decade (as referenced above). Our simulation-based approach sacrifices explicit SSD formulas and is computationally intensive but is feasible for at least a portion of the wide range of hierarchical models which dominate the current Bayesian landscape. It was anticipated in Joseph, Wolfson and Du

Berger (1995b, pages 169–170), and was illustrated for a binomial model without covariates in Zou and Normand (2001).

We also address a second SSD problem. Often before collecting data, we can consider several competing explanatory models. It might be useful to be able to determine a sample size such that, after the data is collected, using a particular model choice criterion, our chance of choosing a particular model given that model is correct is sufficiently high. We refer to this problem as SSD to separate models and confine ourselves to the separation of two models. A natural initial question is whether, for a given pair of models and a given model choice criterion, separation through sample size is achievable? As a simple example, if both models have the same likelihood, differing only in the prior, then with increasing sample size, since the data overwhelms the prior, the posteriors will become indistinguishable so model separation is not achievable. In other words, prior sensitivity is a characteristic of a given sample size and evaporates with increasing sample size.

EXAMPLE 2. A frequent concern when modeling continuous data is whether the assumption of Gaussian errors is acceptable. Perhaps a heavier-tailed distribution, say a t with small degrees of freedom, is appropriate. Consider the simplest version where model 1 asserts $y_1, \ldots, y_n \sim N(\mu, \sigma^2)$ and model 2 asserts $y_1, \ldots, y_n \sim t_2(\mu, \sigma)$. With μ and σ unknown, how large must n be so that, for an adopted model comparison criterion, the chance of selecting the normal model, given it is true, achieves a specified level and the chance of selecting the t model given it is true also achieves a specified level. We return to this example in Sections 8 and 9.

The classical literature unfortunately is limited to the case of nested models under usual regularity conditions. Then the likelihood ratio statistic will have an asymptotic chi-square distribution under the reduced model. As a result, the likelihood ratio is inconsistent. As sample size increases we can not guarantee that the chance that we select the reduced model when it is true is arbitrarily large. Under special asymptotics, under the full model, the likelihood ratio statistic will have an approximate noncentral chi-square distribution with noncentrality parameter which increases in sample size. Now suppose a full model parameter value is selected. In addition, suppose a probability of choosing the full model when the full model is true is specified, thus defining the model separation criterion in this setting. Then, for sample size large enough, the noncentrality parameter, resulting from the full model parameter value, will be large enough to achieve the desired power thus determining the required sample size. Unfortunately, in practice the functional connection of the noncentrality parameter to the sample size through the design matrix may be vague. Also, it may not be apparent which full model parameter value to adopt.

The Bayesian literature is scant. Weiss (1997) confines himself to the case of hypothesis testing considering nonnested composite null and alternative hypotheses. He uses the Bayes factor, illustrating with some very elementary examples. Rubin and Stern (1998) raise the important point that the failure of the available data to reject a simpler model in favor of a more complex one may be due to having insufficient data to criticize the simpler model. Evidently, the context is nested models but hypothesis testing is not proposed. Rather, posterior predictive distributions of so-called discrepancy variables are employed to diagnose particular failures of the simpler model. Sample size is determined to make these distributions sufficiently concentrated, implying a high chance of revealing failure of the simpler model when it is not operating.

Hence the format of the paper is as follows. In Sections 2–6 we focus on SSD for model performance; in Sections 7–9 we consider SSD for model separation. In Section 2 we detail a range of performance criteria. This suggests Section 3 which introduces the roles of the sampling prior and the fitting prior. In Section 4 we lay out the simulation-based approach including a flow chart to elaborate the computational development. In Section 5 we provide a collection of useful theoretical results with regard to performance as a function of sample size. Such results are needed to provide guidance with regard to what is achievable. In Section 6 we give two nonstandard illustrations. The first considers a survival model with censoring, the second a logistic regression. Section 7 formalizes the model separation problem. The use of sampling and fitting priors and the general simulation-based approach follow that of Sections 3 and 4, respectively. Section 8 presents analytical results for several separation problems. Finally, Section 9 presents two examples. One considers separation of a normal error model from a t-error model. The second seeks, in the context of longitudinal data, to separate a common growth curve model from a hierarchical model with individual growth curves.

2. MODEL PERFORMANCE CRITERIA

A Bayesian model specifies a likelihood and a prior. Denote the data associated with a sample size of n by $\mathbf{y}^{(n)}$ and let $\boldsymbol{\theta}$ be the vector of all model parameters. Then we write the model as the joint distribution of $\mathbf{y}^{(n)}$ and $\boldsymbol{\theta}$, that is,

$$f(\mathbf{y}^{(n)} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}). \tag{1}$$

The posterior for $\boldsymbol{\theta}$ is proportional to (1). Here $f(\boldsymbol{\theta})$ can be a hierarchical specification, for example, $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$ with $\boldsymbol{\theta}^{(1)}$, say, the first-stage parameters and $\boldsymbol{\theta}^{(2)}$ the hyperparameters. Interest focuses on inference regarding the scalar $\varphi = h(\boldsymbol{\theta})$.

We consider exclusively the problem of choosing n to achieve specified expected behavior with regard to the posterior $f(\varphi \mid \mathbf{y}^{(n)})$. We do not consider that there is a cost to sampling and thus do not formulate a utility function which reflects the trade-off between performance and cost [Lindley (1997)]. Therefore we do not have an expected utility maximization problem. We do note that the utility maximization approach can be handled within our general computational strategy following the ideas in Müller and Parmigiani (1995) but no further details are presented here.

The literature mentioned in Section 1 discusses a variety of performance criteria with regard to $f(\varphi \mid \mathbf{y}^{(n)})$ and associated c.d.f. $F(\varphi \mid \mathbf{y}^{(n)})$. These include:

1. The average posterior variance criterion (APVC). Given $\varepsilon \geq 0$, the APVC seeks n such that

$$E \operatorname{var}(\varphi \mid \mathbf{y}^{(n)}) \leq \varepsilon. \tag{2}$$

   Of course other measures of dispersion might be appropriate, for example, the average posterior interquantile range.

2. The average coverage criterion (ACC). Suppose $A(\mathbf{y}^{(n)})$ is a set on R^1 determined by $\mathbf{y}^{(n)}$ of fixed length l. For instance, $A(\mathbf{y}^{(n)})$ could be a symmetric set of the form $(\hat{\varphi}_n - l/2, \hat{\varphi}_n + l/2)$ where $\hat{\varphi}_n$ is an estimate of φ such as the posterior mean or median. Alternatively, with a highest posterior density (HPD) set, $A(\mathbf{y}^{(n)}) = \{\varphi : f(\varphi \mid \mathbf{y}^{(n)}) \geq c_n(l)\}$ where $c_n(l)$ is chosen such that the Lebesgue measure of $A(\mathbf{y}^{(n)})$ is l. Then given $\alpha \geq 0$, ACC seeks n such that

$$E \Pr\bigl(\varphi \in A(\mathbf{y}^{(n)}) \mid \mathbf{y}^{(n)}\bigr) \geq 1 - \alpha. \tag{3}$$

   Of course, for a fixed l, the HPD choice will receive a smaller n than the symmetric choice. Thus, the latter is conservative, but simpler to compute. See Joseph, Wolfson and Du Berger (1995b) in this regard.

3. The average length criterion (ALC). Consider the interval $A(\mathbf{y}^{(n)}) = (F^{-1}_{\varphi|\mathbf{y}^{(n)}}(\alpha/2),\ F^{-1}_{\varphi|\mathbf{y}^{(n)}}(1-\alpha/2))$. That is, $A(\mathbf{y}^{(n)})$ is an equal (in probability) tail, $1-\alpha$ posterior interval estimate for φ. Given $l \geq 0$, the ALC criterion seeks n such that

$$E\bigl(F^{-1}_{\varphi|\mathbf{y}^{(n)}}(1-\alpha/2) - F^{-1}_{\varphi|\mathbf{y}^{(n)}}(\alpha/2)\bigr) \leq l. \tag{4}$$

   Again, a $1-\alpha$ HPD interval can replace the equal tail interval.

4. In certain applications φ will be interpreted as an *effect size* in which case $\Pr(\varphi > 0 \mid \mathbf{y}^{(n)})$ (the posterior probability of detecting an effect) or, more generally, $\Pr(\varphi > \varphi^* \mid \mathbf{y}^{(n)})$ (the posterior probability of detecting an effect of size at least φ^*) may be of interest. Then, given $\alpha \geq 0$, we seek n such that

$$E \Pr(\varphi > \varphi^* \mid \mathbf{y}^{(n)}) \geq 1 - \alpha. \tag{5}$$

Note that in (2)–(5), the expectation is calculated with respect to the marginal distribution of $\mathbf{y}^{(n)}$ which therefore must be proper.

Implicit in (2)–(5) is the existence of a limit for the left side as $n \to \infty$ and that this limit permits the specified inequality. In fact, rough monotonicity of the left sides in n would seem to be implicit as well. These theoretical matters are taken up in the next section. Not surprisingly, establishing the existence of limits is much easier than demonstrating monotonicity so we generally depend upon the results of our simulation-based approach to clarify the rough monotonicity.

To unify notation we remark that each of (2) through (5), for an appropriate nonnegative function $T(\mathbf{y}^{(n)})$, can be written as

$$E\bigl(T(\mathbf{y}^{(n)})\bigr) \leq \varepsilon. \tag{6}$$

Also, in the above, if it is of interest, we can replace $f(\varphi \mid \mathbf{y}^{(n)})$ with a posterior predictive distribution, $f(y_{\text{new}} \mid \mathbf{y}^{(n)})$, and again obtain a form like (6). Lastly, if we have multiple performance objectives, this will result in a set of expectations each like (6). If the required n is computed for each one, the maximum of these n's achieves all objectives.

3. FITTING AND SAMPLING PRIORS

Bayesian sample size determination is a form of "preposterior" analysis. It is done in the absence of data. Nonetheless, given a proper model and a sample size, we can certainly simulate data from the model to learn how the resultant posterior will behave.

An important aspect of our approach is that we distinguish between a "sampling" prior and a "fitting" prior. In particular, while we often have useful prior information, it is generally preferable to let the data drive the inference. Priors which are relatively noninformative are encouraged. An analysis which more closely resembles a likelihood analysis (but avoids worrisome asymptotics) results. Hence in (1) we think of $f(\boldsymbol{\theta})$ as vague, perhaps improper as long as $f(\boldsymbol{\theta} \mid \mathbf{y}^{(n)})$ is proper. We refer to this prior as the fitting prior and denote it by $f^{(f)}(\boldsymbol{\theta})$. This is the prior that we would anticipate using to fit the model once the data is obtained.

By contrast, in practical sample size determination we are usually interested in achieving a certain level of performance if $\boldsymbol{\theta}$ is likely to be in some specified portion of the parameter space. We capture this through a sampling prior for $\boldsymbol{\theta}$, denoted by $f^{(s)}(\boldsymbol{\theta})$. The sampling prior arises in a "what if" spirit. Drawing upon expertise, we may speculate upon a variety of informative scenarios regarding $\boldsymbol{\theta}$ and capture each with a suitable sampling prior. Moreover, if $\phi = h(\boldsymbol{\theta})$ is of interest we may create $f^{(s)}(\boldsymbol{\theta})$ from $f^{(f)}(\boldsymbol{\theta})$ by insuring that $f^{(s)}$ is proper and very informative with regard to ϕ. We take "very informative" as a uniform prior over some suitable bounded interval. In terms of capturing variability appropriately, allowing uncertainty in ϕ and, in fact, in $\boldsymbol{\theta}$, seems preferable to fixing $\boldsymbol{\theta}$, as classical approaches require. The sampling prior is necessarily proper and generates the $\boldsymbol{\theta}$'s in the Bayesian model. Given such a $\boldsymbol{\theta}$, say, $\boldsymbol{\theta}^*$, $\mathbf{y}^{(n)*}$ is generated from $f(\mathbf{y}^{(n)} \mid \boldsymbol{\theta}^*)$. Then $\mathbf{y}^{(n)*}$ is subjected to the fitting model to ascertain what sort of posterior analysis ensues. Also, the sampling prior, in conjunction with $f(\mathbf{y}^{(n)} \mid \theta)$ provides the (proper) marginal distribution $f^{(s)}(\mathbf{y}^{(n)})$ under which (6) is computed. $\mathbf{y}^{(n)*}$ is a realization from this distribution. With regard to (6), $T(\mathbf{y}^{(n)})$ is calculated under $f^{(f)}(\mathbf{y}^{(n)})$, its expectation under $f^{(s)}(\mathbf{y}^{(n)})$. Neither $f^{(s)}(\boldsymbol{\theta} \mid \mathbf{y}^{(n)})$ nor $f^{(f)}(\mathbf{y}^{(n)})$ is of interest.

4. THE SIMULATION-BASED SSD APPROACH

The simulation-based SSD approach requires exploring an appropriate set of n's such that, at each one, $E(T(\mathbf{y}^{(n)}) \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)}))$ is calculated. With multiple T_i's, this expectation must be obtained for each one. The set of n's need not be known in advance but may evolve according to (6). If the right-hand side of (6) is inflexible, we would attempt to find a pair n_1 and n_2 such that the associated expectations bracket ε. Then, perhaps a bisection or more refined search would be used to find the required n. If the right-hand side of (6) is flexible, then we may propose a grid of n values, obtain the left-hand side of (6) for each one and, using simple interpolation, develop a plot of $E(T(\mathbf{y}^{(n)}) \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)}))$ versus n.

Generically, $T(\mathbf{y}^{(n)})$ is a functional of $f^{(f)}(\boldsymbol{\theta} \mid \mathbf{y}^{(n)})$. More broadly, it may be a functional of $f^{(f)}(y_{\text{new}}, \boldsymbol{\theta} \mid \mathbf{y}^{(n)}) = f(y_{\text{new}} \mid \boldsymbol{\theta}, \mathbf{y}^{(n)}) f^{(f)}(\boldsymbol{\theta} \mid \mathbf{y}^{(n)})$. Only in the simplest cases will $f^{(f)}(\boldsymbol{\theta} \mid \mathbf{y}^{(n)})$ be available explicitly. For these, few choices of $T(\mathbf{y}^{(n)})$ will be available explicitly and even fewer of these will admit an explicit expectation with regard to $f^{(s)}(\mathbf{y}^{(n)})$ [especially since $f^{(s)}(\mathbf{y}^{(n)})$ itself will rarely be available explicitly].

Fortunately, we can compute $E(T(\mathbf{y}^{(n)}) \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)}))$ using simulation. In particular, following the end of Section 3, we can obtain arbitrarily many realizations $\mathbf{y}_l^{(n)*}$, $l = 1, \ldots, L$, from $f^{(s)}(\mathbf{y}^{(n)})$. Hence, a Monte Carlo integration for $E(T(\mathbf{y}^{(n)}) \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)}))$ is $L^{-1} \sum_{l=1}^{L} T(\mathbf{y}_l^{(n)*})$ so we only need to compute $T(\mathbf{y}_l^{(n)*})$. However, given $\mathbf{y}_l^{(n)*}$, we can use either direct or iterative simulation to sample either $f^{(f)}(\boldsymbol{\theta} \mid \mathbf{y}_l^{(n)*})$ or $f^{(f)}(y_{\text{new}}, \boldsymbol{\theta} \mid \mathbf{y}_l^{(n)*})$ and hence, to obtain the corresponding functional of the sample as an arbitrarily accurate approximation to the functional $T(\mathbf{y}_l^{(n)*})$.

It is apparent that the required computation is intensive. Figure 1 provides a flow chart to summarize the steps.

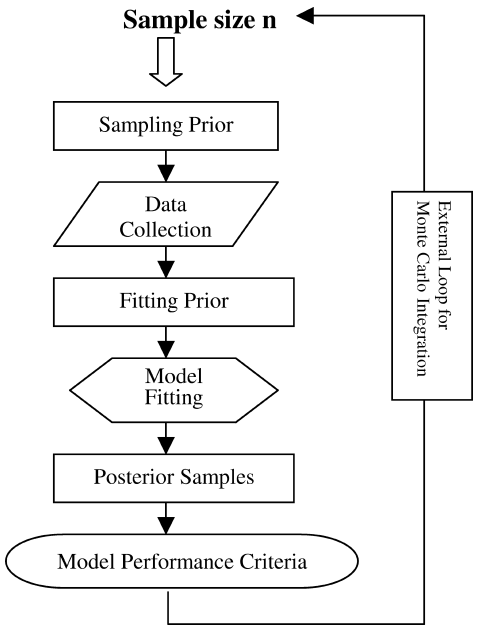


FIG. 1. *Flow chart.*

5. ANALYTICAL RESULTS FOR MODEL PERFORMANCE SSD

It is insightful to start with a simple example where calculations can be done explicitly. Suppose y_i, $i = 1, \ldots, n \sim N(\theta, \sigma^2)$ with σ^2 known and let $f^{(f)}(\theta) = N(\mu_f, \tau_f^2)$, $f^{(s)}(\theta) = N(\mu_s, \tau_s^2)$. Then, $f^{(f)}(\theta \mid \mathbf{y}^{(n)}) = N\big(\frac{n\tau_f^2 y_n + \sigma^2\mu_f}{n\tau_f^2+\sigma^2}, \frac{\sigma^2\tau_f^2}{\sigma^2+n\tau_f^2}\big)$. Also, by sufficiency, $f^{(s)}(\mathbf{y}^{(n)})$ can be replaced with $f^{(s)}(\bar{y}_n)$ where $f^{(s)}(\bar{y}_n) = N(\mu_s, \sigma^2/n + \tau_s^2)$.

With the APVC criterion, $T(\mathbf{y}^{(n)}) = \sigma^2\tau_f^2/(\sigma^2 + n\tau_f^2)$ and thus $E(T(\mathbf{y}^{(n)}) \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)}))$ decreases strictly to 0 as $n \to \infty$. With, say, an effect size criterion, $T(\mathbf{y}^{(n)}) = \Pr(\theta > \theta^* \mid \theta \sim f^{(f)}(\theta \mid \mathbf{y}^{(n)}))$, we obtain

$$T(\mathbf{y}^{(n)}) = \Phi\Bigg(\frac{n\tau_f^2(\bar{y}_n - \theta^*) + \sigma^2(\mu_f - \theta^*)}{n\tau_f^2 + \sigma^2} \Bigg/ \sqrt{\sigma^2\tau_f^2/(\sigma^2 + n\tau_f^2)}\Bigg)$$

and, after some calculation,

$$E\big(T(\mathbf{y}^{(n)}) \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)})\big) = \Phi\Bigg((\mu_s - \theta^* + \sigma^2\mu_f/n\tau_f^2) \Bigg/ \sqrt{\frac{\sigma^2}{n} + \tau_s^2 + \frac{\sigma^2(\sigma^2 + n\tau_f^2)}{n^2\tau_f^2}}\Bigg). \tag{7}$$

The limit of (7) as $n \to \infty$ is $\Pr(\theta > \theta^* \mid \theta \sim f^{(s)}(\theta))$. The fitting prior is overwhelmed in the posterior by the increasing amount of data arising under the sampling model and in the limit we obtain the probability under the sampling prior. Also the need to obtain this limit in order to determine what α's are achievable in (5) is demonstrated (or alternatively, to modify $f^{(s)}$ so that a desired α is achievable).

Our objective is analytical assessment of $\lim_{n\to\infty} b_n$ where $b_n = E(T(\mathbf{y}^{(n)}) \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)}))$. Theorem 1 provides two limiting results.

THEOREM 1. (i) *Let* $h_n(\boldsymbol{\theta}) = \int T(\mathbf{y}^{(n)}) f(\mathbf{y}^{(n)} \mid \boldsymbol{\theta})\, d\mathbf{y}^{(n)}$. *If* $h_n(\boldsymbol{\theta}) \to h(\boldsymbol{\theta})$ *then, provided interchange of limit and integration is valid,* $\lim h_n = \int h(\boldsymbol{\theta}) f^{(s)}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$. *Alternatively, if, given* $\boldsymbol{\theta}$, $T(\mathbf{y}^{(n)}) \xrightarrow{P} h(\boldsymbol{\theta})$ *and* $\sup_n b_n(\boldsymbol{\theta}) \le M < \infty$, *then* $\lim_{n\to\infty} b_n = \int h(\boldsymbol{\theta}) \cdot f^{(s)}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$.

(ii) *Suppose* $T(\mathbf{y}^{(n)})$ *is a linear functional, that is,* $T(\mathbf{y}^{(n)}) = \int \gamma(\boldsymbol{\theta}) f^{(f)}(\boldsymbol{\theta} \mid \mathbf{y}^{(n)})\, d\boldsymbol{\theta}$. *Define* $T^{(s)}(\mathbf{y}^{(n)}) = \int b(\boldsymbol{\theta}) f^{(s)}(\boldsymbol{\theta} \mid \mathbf{y}^{(n)})\, d\boldsymbol{\theta}$. *If* $T(\mathbf{y}^{(n)}) - T^{(s)}(\mathbf{y}^{(n)}) \xrightarrow{P} 0$ *and* $\sup_n \int |T(\mathbf{y}^{(n)}) - T^{(s)}(\mathbf{y}^{(n)})| f^{(s)}(\mathbf{y}^{(n)})\, d\mathbf{y}^{(n)} \le M < \infty$, *then* $\lim b_n = \int \gamma(\boldsymbol{\theta}) f^{(s)}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$.

The straightforward proof is given in Appendix A.

We note that the alternative conditions in (i) may be easier to check since often the convergence requirement is available through standard asymptotics and the boundedness requirement can be established without computing $h_n(\boldsymbol{\theta})$ explicitly.

For (ii), the boundedness condition will be achieved if either $b(\boldsymbol{\theta})$ is bounded or $f^{(f)}(\boldsymbol{\theta})$, hence $f^{(f)}(\boldsymbol{\theta} \mid \mathbf{y}^{(n)})$ has bounded support. The convergence in probability condition will hold when usual regularity conditions hold. That is, $T(\mathbf{y}^{(n)}) - b(\hat{\boldsymbol{\theta}}_n) \xrightarrow{P} 0$, where $\hat{\boldsymbol{\theta}}_n$ is the maximum likelihood estimator of $\boldsymbol{\theta}$ based on $\mathbf{y}^{(n)}$, implies $T(\mathbf{y}^{(n)}) - T^{(s)}(\mathbf{y}^{(n)}) \xrightarrow{P} 0$. Intuitively, as n grows large, the likelihood (the data) overwhelms either prior. Lastly, we see the importance of distinguishing a sampling prior from a fitting prior. If they are the same then, in the above case, $E(T(\mathbf{y}^{(n)}) \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)})) = \int b(\boldsymbol{\theta}) f^{(s)}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$, free of n. There is no SSD problem.

We conclude this section with some examples where we work primarily with APVC. However, note that controlling APVC typically ensures that ACC and ALC can be controlled. For instance, using Chebyshev's inequality $\Pr(\varphi \in (E(\varphi \mid \mathbf{y}^{(n)}) - d, E(\varphi \mid \mathbf{y}^{(n)}) + d) \mid \mathbf{y}^{(n)}) \ge 1 - \mathrm{var}(\phi \mid \mathbf{y}^{(n)})/4d^2$. Taking expectations with respect to $\mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)})$ yields

$$E_{f^{(s)}(\mathbf{y}^{(n)})} \Pr\big(\phi \in \big(E(\phi \mid \mathbf{y}^{(n)}) - d, E(\phi \mid \mathbf{y}^{(n)}) + d\big) \mid \mathbf{y}^{(n)}\big) \ge 1 - \mathrm{APVC}/4d^2. \tag{8}$$

From (8), for fixed length d, if APVC $\to 0$, ACC $\to 1$. But also, if we fix posterior coverage to $1 - \alpha$, then (8) yields $1 - \alpha \ge 1 - \mathrm{APVC}/d^2$, that is, $d^2 \le \mathrm{APVC}/\alpha$ so, again, if APVC $\to 0$, ALC $\to 0$.

One and two sample problems. Consider first the one parameter natural exponential family (NEF) with density $f(y \mid \theta) = c(y)\exp(\theta y - \chi(\theta))$, the so-called canonical parametrization. It is evident that, for a sample of size n, y_n is sufficient. Also, the form of the conjugate prior is well known, $f(\theta) = k(m, \mu_0) \cdot \exp(m(\mu_\theta\theta - \chi(\theta))$. In the case where $f(y \mid \theta)$ has a quadratic variance function (QVF), $V(\mu) = \nu_0 + \nu_1\mu + \nu_2\mu^2$ where $\mu = E(y \mid \theta) = \chi'(\theta)$, then, as in Morris (1983, Theorem 5.4), the posterior variance of μ is $\mathrm{var}(\mu \mid y_n) = V(y_0)/(N - \nu_2)$ where

$N = n+m$ and $y_0 = (n\bar{y}_n + m\mu_0)/N$. Also, since marginally, $E(\bar{y}_n) = \mu_0$ and $\text{var}(\bar{y}_n) = NV(\mu_0)/n(m-\nu_2)$, provided $m > \nu_2$ [again, Morris (1983)], we can compute APVC explicitly. Indeed, after some calculation, $\text{APVC} = V(\mu_0)(1 + \frac{\nu_2 n}{N(m-\nu_2)})/(N-2)$ where $\text{APVC} \to 0$ as $n \to \infty$. In fact, it is easy to see that if n is sufficiently large, APVC decreases strictly to 0.

The normal case with σ^2 known is a special case of the above. When σ^2 is unknown and $f^{(f)}(\mu, \sigma^2) \propto 1/\sigma$, $\sqrt{n}(\mu - \bar{y}_n)/s_n \sim t_{n-1}$ where s_n^2 is the usual sample variance [Box and Tiao (1973), Section 2.4]. Hence, $\text{var}(\mu \mid \mathbf{y}^{(n)}) = \frac{n-1}{n-3}\frac{s_n^2}{n}$ and $\text{APVC} = \frac{(n-1)}{n(n-3)} \cdot E(s_n^2 \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)})) = \frac{(n-1)}{n(n-3)} E(\sigma^2 \mid \sigma^2 \sim f^{(s)}(\sigma^2))$. So, if the mean of $f^{(s)}(\sigma^2)$ exists, APVC $\to 0$ as $n \to \infty$ and, in fact, monotonically.

The above results directly extend to the usual two sample problem. If $x_1, \ldots, x_n$ i.i.d. $f(x \mid \theta_1)$ and $y_1, \ldots, y_n$ i.i.d. $f(y \mid \theta_2)$ and θ_1 and θ_2 are, a priori, independent, $\text{var}(\theta_1 - \theta_2 \mid \mathbf{x}^{(n_1)}, \mathbf{y}^{(n_2)}) = \text{var}(\theta_1 \mid \mathbf{x}^{(n_1)}) + \text{var}(\theta_2 \mid \mathbf{y}^{(n_2)})$. So, APVC for $\theta_1 - \theta_2$ approaches 0 if APVC approaches 0 for each of θ_1 and θ_2. We can apply the foregoing results choosing n_1 and n_2 to make each of these less than $\varepsilon/2$, ensuring that the APVC for $\theta_1 - \theta_2$ is less than ε.

Linear regression and generalized linear models. Consider first the usual normal theory linear regression setting. That is, we assume $\mathbf{y}^{(n)} = X_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}^{(n)}$ where X_n is subscripted to denote sample size and is $n \times p$ with rank p, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients and $\boldsymbol{\varepsilon}^{(n)} \sim N(\mathbf{0}, \sigma^2 I_n)$.

The classical SSD approach notes that, in testing $H_0: \boldsymbol{\beta} = \mathbf{0}$, the noncentrality parameter at a given $\boldsymbol{\beta}$ and n is $\lambda_n = \boldsymbol{\beta}^T X_n^T X_n \boldsymbol{\beta}/\sigma^2$. Hence, if λ_n is sufficiently large, say λ^*, the probability that the usual F test rejects H_0 at this λ^* can be made as large as desired. But since $X_n^T X_n = O(n)$, for any $\boldsymbol{\beta} \neq \mathbf{0}$, $\lambda_n \to \infty$ as $n \to \infty$. Therefore, if n is large enough, λ_n will reach λ^* and this n becomes the required sample size. There is a bit of circularity in this argument because λ^* in fact depends upon n since the denominator d.f. of the F distribution does. For large n, since $F_{p,n-p} \approx \chi_p^2$, presumably this is not a serious problem. What is problematic is that, without explicit assumptions regarding X_n, the connection of λ_n to n through X_n is vague. Even if this is made precise, a value of $\boldsymbol{\beta}$ is required to calculate λ_n. Where does this $\boldsymbol{\beta}$ come from? Moreover, such a performance measure is appropriate only if the sole objective of the experiment is to reject $H_0: \boldsymbol{\beta} = \mathbf{0}$. Note that so-called *local* asymptotics which make $\boldsymbol{\beta} = \boldsymbol{\beta}_n = O(n^{-1/2})$ hence $\lambda_n = O(1)$ can not achieve arbitrary power.

For a generalized linear model, the F statistic is replaced by a score or likelihood ratio statistic. Under suitable conditions, such a statistic has an approximate chi-square distribution with an approximated noncentrality parameter λ_n which, as above, tends to ∞ as $n \to \infty$. Hence, the foregoing approach can be applied.

The Bayesian SSD approach presumes broader inferential interest, in particular about some (or all) of the β's individually. We continue to illustrate with the APVC. Hence, we have the multiple criteria, APVC for $\beta_j \leq \varepsilon_j$, $j = 1, 2, \ldots, p$ (or perhaps a subset of these).

In a preposterior mode, we assume that each vector $\mathbf{X_i}$ is random and, in fact, that the $\mathbf{X_i}$ are i.i.d. Hence, in our simulation-based approach, given n, we generate $\mathbf{X_i}^*$, $i = 1, 2, \ldots, n$, collecting them into X_n^*. We draw $(\boldsymbol{\beta}^*, \sigma^{2*})$ from $f^{(s)}(\boldsymbol{\beta}, \sigma^2)$, then $\mathbf{y}^{(n)*}$ from $N(X_n^*\boldsymbol{\beta}^*, \sigma^{2*} I_n)$ and proceed as in Section 4. In particular, if $f^{(f)}(\boldsymbol{\beta}, \sigma^2) = N((\boldsymbol{\beta} \mid (\boldsymbol{\beta}_0, c\sigma^2) IG(\sigma^2 \mid a, b)$ then we can directly simulate the fitting posterior; that is, $f^{(f)}(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}^{(n)}) = f^{(f)}(\boldsymbol{\beta}, \mid \mathbf{y}^{(n)}) f^{(f)}(\sigma^2 \mid \mathbf{y}^{(n)})$ where the latter two distributions are multivariate normal and inverse gamma, respectively. For a general $f^{(f)}(\boldsymbol{\beta}, \sigma^2)$ we would use a Gibbs sampler as in Gelfand and Smith (1990). The choice of distribution for $\mathbf{X_i}$ depends upon the application. For convenience we might assume that the components of $\mathbf{X_i}$ are independent uniforms over ranges determined by expertise.

Hence, we consider when APVC for $\beta_j \to 0$ as $n \to \infty$. In fact, with X_n random, we only ask when APVC for $\beta_j \xrightarrow{P} 0$. In particular with fitting prior $f^{(f)}(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma$, the posterior covariance matrix for $\boldsymbol{\beta}$ is $(n-p)(n-p-2)^{-1}\hat{\sigma}^2(X_n^T X_n)^{-1}$ where $\hat{\sigma}^2$ is the usual mean square error [Box and Tiao (1973), Section 2.7]. Standard calculation shows that $(X_n^T X_n)^{-1}$ is $O_p(n^{-1})$. Hence, as with the one sample normal case above, if $f^{(s)}(\sigma^2)$ has a finite mean, APVC for $\beta_j \xrightarrow{P} 0$.

For the generalized linear model we offer an approximate argument, omitting details. Assuming a NEF model with canonical link, under a flat prior for $\boldsymbol{\beta}$, provided the posterior exists, $\boldsymbol{\beta} \mid \mathbf{y}^{(n)} \dot{\sim} N(\hat{\boldsymbol{\beta}}_n, (X_n^T \cdot M_n X_n)^{-1})$ where $\hat{\boldsymbol{\beta}}_n$ is the MLE for $\boldsymbol{\beta}$ based on $\mathbf{y}^{(n)}$ and M_n is a diagonal matrix with $(M_n)_{ii} = \chi''(X_i^T\hat{\boldsymbol{\beta}})$. Hence, analogous to the Gaussian case, $(X_n^T M_n X_n)_{jj}^{-1}$ will be $O_p(n^{-1})$ under appropriate constraints on $f^{(s)}(\boldsymbol{\beta})$.

A random effects model. We consider the simplest balanced random effects model, $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$ with ε_{ij} i.i.d. $N(0, \sigma_e^2)$, α_i i.i.d. $N(0, \sigma_\alpha^2)$. In this setting, we take Jeffreys' prior as the fitting prior, that is, $f^{(f)}(\mu, \sigma_\alpha^2, \sigma_e^2) \propto \sigma_e^{-2}(\sigma_e^2 + J\sigma_\alpha^2)^{-1}$. The resulting posterior is straightforward to sample using Markov chain Monte Carlo methods but, analytically, we are limited to approximation. Suppose interest lies in APVC for σ_e^2 and for σ_α^2. Intuitively, the former is controlled by letting J grow large, the latter by letting I grow large as well. Appendix B provides formal details.

For an unbalanced setting with, for population i, $j = 1, 2, \ldots, J_i$, the same conclusions will obviously hold if $\min J_i \to \infty$. In particular, this means that if we choose $I = I_0$ and $J = J_0$ to achieve specified APVC performance, we will achieve at least this performance for any design with $\min J_i \geq J_0$.

6. TWO EXAMPLES OF SSD FOR PERFORMANCE

We illustrate the foregoing development in two standard settings, a survival model with censoring and a logistic regression. Neither has previously been considered with regard to Bayesian SSD.

6.1 A Single Event Survival Data Problem

We consider a survival model incorporating a Weibull hazard, that is, $h(t, \boldsymbol{\theta}) = \lambda\gamma t^{\gamma-1}$, $\lambda > 0, \gamma > 0$, $\boldsymbol{\theta} = (\lambda, \gamma)$. We introduce right censoring which could be imposed at random but which we treat as fixed, that is, T is censored if $T > T_c$. Random censoring would simply require sampling (T, T_c) pairs. In either case we can design the censoring mechanism to roughly achieve some predetermined expected proportion of censoring. For convenience we take the fitting prior to be a product of vague gamma distributions. Such choices provide a full conditional distribution for λ which is an updated gamma and a log-concave full conditional distribution for γ. Hence a Gibbs sampler is routine, sampling γ through adaptive rejection sampling using the approach of Gilks and Wild (1992). For the sampling prior we take $\lambda \sim U(\underline{\lambda}, \bar{\lambda})$ and, independently, $\gamma \sim U(\underline{\gamma}, \bar{\gamma})$. Possible choices for γ might make the support greater than 1 (less than 1) insuring a decreasing (increasing) hazard. With this choice and a choice of some median survival time of interest [median $= (\log 2/\lambda)^\gamma$], rough choices for $\underline{\lambda}$ and $\bar{\lambda}$ can be made. Sample size determination to control inference regarding λ and γ is not likely as useful as that for say a survival time quantile or the value of the survival function at a given time.

As a concrete illustration we set the fitting prior for both λ and γ to be Ga(0.01, 0.01), that is, a gamma distribution with mean 1. The sampling prior is $\lambda \sim U(0.01, 0.02)$, $\gamma \sim U(3.0, 3.5)$ and $T_c = 5.0$. (The expected censoring rate is roughly 10%–20%.) In Figure 2 we consider APVC and ALC for the posterior median survival time and for the survival function at $t = 4.0$. We illustrate with $n = 10, 11, \ldots, 20$, though, of course, we could look at any range of n with any appropriate spacing. The Monte Carlo integration for expectations with respect to $f^{(s)}(\mathbf{y}^{(n)})$ used $L = 2000$.

6.2 A Logistic Regression Problem

We consider a logistic regression setting where the objective is to model p_{ij}, the probability of an occurrence for the jth individual in the ith group, $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$. We assume $\log \frac{p_{ij}}{1-p_{ij}} = \beta_0 + \beta_1 X_i + \beta_2 Z_{ij}$ and seek inference regarding β_1, the coefficient of the population level covariate and β_2, the coefficient of the individual level covariate, each to specified precision. In particular, we set $I = 2$ and let $X_i = 0, 1$ indicating which of the two groups was sampled.

With binary response, the fitting prior cannot be flat or else an improper posterior results. Hence we take independent normal priors for β_0, β_1 and β_2 with mean 0 and variance large relative to the scale of the data. Under the fitting prior, models are fitted using a Gibbs sampler with adaptive rejection sampling. The sampling prior can match the fitting prior for β_0 but is very informative for β_1 and β_2, that is, $\beta_1 \sim U(\underline{\beta}_1, \bar{\beta}_1)$, $\beta_2 \sim U(\underline{\beta}_2, \bar{\beta}_2)$. With I fixed at 2, sample size determination involves letting J increase, sampling J individuals from group 1 and J from group 2.

As a concrete illustration we let $Z_{ij} \sim U(0, 1)$ with, as fitting prior, $\beta_0, \beta_1, \beta_2$ all $N(0,10)$. The sampling prior is $\beta_0 \sim N(0, 10)$, $\beta_1 \sim U(1, 1.5)$, $\beta_2 \sim U(1, 2)$. In Figure 3 we obtain APVC and ALC for both β_1 and β_2 for $J = 10, 15, 20, \ldots, 50$. The Monte Carlo integrations for expectation with respect to $f^{(s)}(\mathbf{y}^{(n)})$ used $L = 2000$. If we sought APVC for β_1 equal to 0.2 we would interpolate to $J = 32$. If in addition we sought APVC for β_2 equal to 0.05 then we would take $J = 36$ and this choice of J achieves both performance specifications.

7. FORMALIZING SSD FOR THE MODEL SEPARATION PROBLEM

We approach the model separation problem through model choice criteria. We do not enter the debate as to

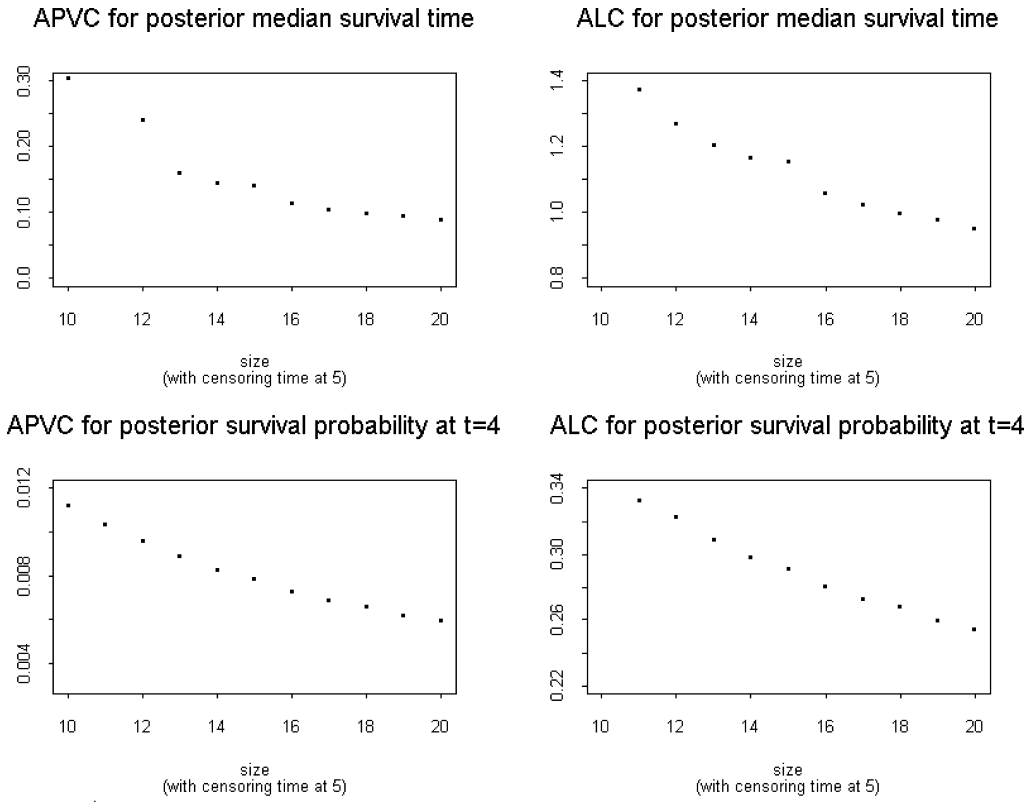


FIG. 2. *Sample size determination under a Weibull model with censoring.*

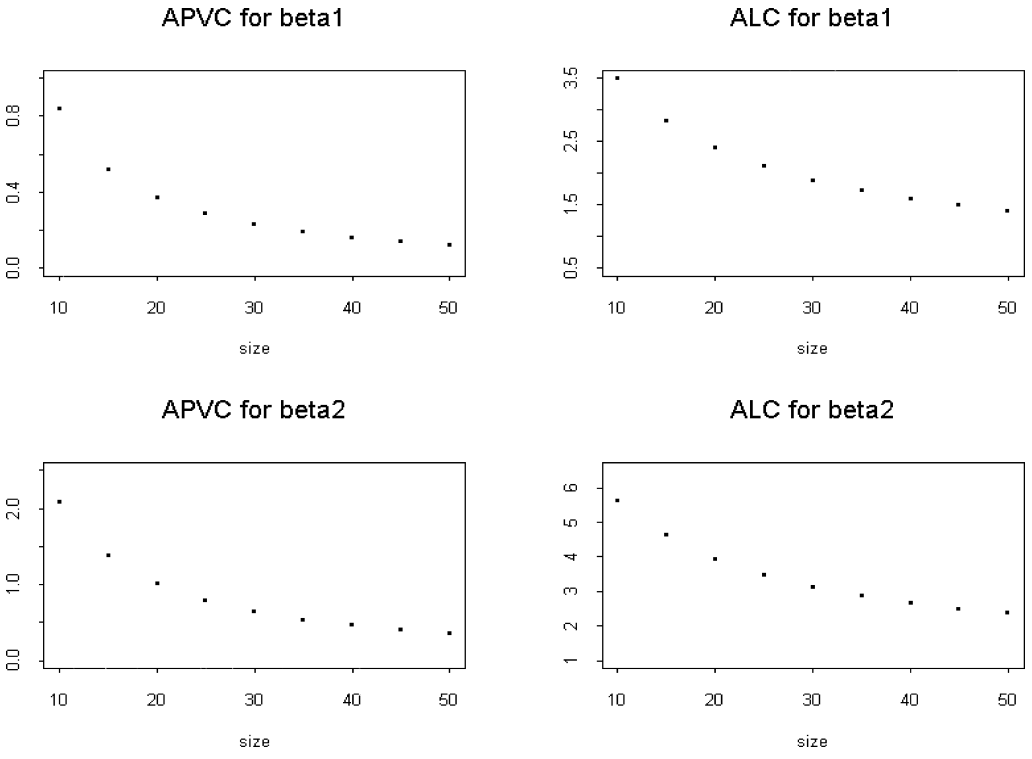


FIG. 3. *Sample size determination under a logistic regression model.*

whether one criterion is preferable to another. Rather, we simplify to the adoption of a model screening criterion [Kass and Raftery (1995)], leaving the selection to the user. In particular, denote the data associated with a sample of size n by $\mathbf{y}^{(n)}$ and let $\boldsymbol{\theta}_i$ denote the vector of model parameters under model M_i, $i = 1, 2$. Then the two Bayesian models to be separated are

$$f(\mathbf{y}^{(n)} \mid \boldsymbol{\theta}_i, M_i) f(\boldsymbol{\theta}_i \mid M_i), \quad i = 1, 2, \tag{9}$$

where the first term in (9) is the likelihood under model M_i and the second term is the prior under M_i.

We denote the screening criterion by S and its value for data $\mathbf{y}^{(n)}$ and model M_i by $S_i(\mathbf{y}^{(n)})$. Such a criterion might be the marginal density ordinate at $\mathbf{y}^{(n)}$, that is, $S_i(\mathbf{y}^{(n)}) = \int f(\mathbf{y}^{(n)} \mid \boldsymbol{\theta}_i, M_i) f(\boldsymbol{\theta}_i \mid M_i)\, d\boldsymbol{\theta}_i$. It might be a pseudodensity ordinate arising under cross validation [Geisser and Eddy (1979)], that is, $S_i(\mathbf{y}^{(n)}) = \prod_i f(y_i^{(n)} \mid \mathbf{y}_{(i)}^{(n)})$ where $y_i^{(n)}$ is the ith component of $\mathbf{y}^{(n)}$ and $\mathbf{y}_{(i)}^{(n)}$ denotes the vector of components of $\mathbf{y}^{(n)}$ with $\mathbf{y}_{(i)}^{(n)}$ deleted. It might be a posterior predictive density ordinate [Aitkin (1991)], that is, $S_i(\mathbf{y}^{(n)}) = f(\mathbf{y}_{\text{new}}^{(n)} \mid \mathbf{y}^{(n)}, M_i)$ evaluated at $\mathbf{y}_{\text{new}}^{(n)} = \mathbf{y}^{(n)}$. It might be a functional of the posterior $f(\boldsymbol{\theta}_i \mid \mathbf{y}^{(n)})$, for example, $E \log f(\boldsymbol{\theta}_i \mid \mathbf{y}^{(n)}) \mid \boldsymbol{\theta}_i \sim f(\boldsymbol{\theta}_i \mid \mathbf{y}^{(n)})$ [Spiegelhalter, Best, Carlin and van der Linde (2002)]. It might be a functional of $f(\mathbf{y}_{\text{new}}^{(n)} \mid \mathbf{y}^{(n)}, M_i)$ as in Gelfand and Ghosh (1998). It might be a functional of $f(\mathbf{y}_{\text{new}}^{(n)}, \boldsymbol{\theta}_i \mid \mathbf{y}^{(n)}, M_i)$ as in Gelman, Meng and Stern (1996). In fact, it could be a non-Bayesian criterion such as a penalized likelihood [see, e.g., Gelfand and Dey (1994)].

Then, without loss of generality, we assume that the screening criterion has been formulated such that it is positive and, the larger the value, the more support for the model. Defining $T(\mathbf{y}^{(n)}) = \ln S_1(\mathbf{y}^{(n)})/S_2(\mathbf{y}^{(n)})$, we "choose" M_1 when $T(\mathbf{y}^{(n)}) > 0$, "choose" M_2 when $T(\mathbf{y}^{(n)}) < 0$. Then the SSD problem to separate two models becomes

$$\text{Choose } n_1 \text{ such that } \Pr\bigl(T(\mathbf{y}^{(n)}) > a_1 \mid M_1\bigr) = 1 - \alpha_1 \tag{10a}$$

$$\text{Choose } n_2 \text{ such that } \Pr\bigl(T(\mathbf{y}^{(n)}) < a_2 \mid M_2\bigr) = 1 - \alpha_2. \tag{10b}$$

Set $n = \max(n_1, n_2)$. Here $a_2 \le 0 \le a_1$ reflect the strength of relative model support one wishes to detect under each model. Obviously, α_1 and α_2 indicate the confidence in such detection. To clarify (10), $\Pr(T(\mathbf{y}^{(n)}) > a_1 \mid M_1) = \int f(\boldsymbol{\theta}_1 \mid M_1) \int 1(T(\mathbf{y}^{(n)}) > a_1) f(\mathbf{y}^{(n)} \mid \boldsymbol{\theta}_1, M_1)\, d\mathbf{y}^{(n)}\, d\boldsymbol{\theta}_1$. A similar expression arises for $\Pr(T(\mathbf{y}^{(n)}) < a_2 \mid M_2)$.

Weiss (1997) proposes (10) informally where $T(\mathbf{y}^{(n)})$ is the log of the Bayes factor and the setting is hypothesis testing. Note that one may elect to choose n to satisfy only one of the criteria in (10). For instance, if $M_1 \subset M_2$ we may seek a sample size to reject M_1 in favor of M_2, that is, use (10b). This is the approach of Rubin and Stern (1998).

Following Section 3 we introduce fitting and sampling priors. For model M_i the fitting prior, denoted by $f^{(f)}(\boldsymbol{\theta}_i \mid M_i)$, $i = 1, 2$, is the rather vague prior we would anticipate using to fit M_i once the data is obtained. By contrast, the sampling prior for model M_i, denoted by $f^{(s)}(\boldsymbol{\theta}_i \mid M_i)$, $i = 1, 2$, is the prior under which we seek to achieve model separation with regard to (10). These priors induce the marginal distributions $f^{(s)}(\mathbf{y}^{(n)} \mid M_i)$ under which the probabilities in (10) are calculated.

The simulation-based SSD approach requires computing the left-hand side of (10) for each n. As in Section 4 for a given n under model M_i, we can obtain arbitrarily many realizations $\mathbf{y}_l^{(n)*}$, $l = 1, 2, \ldots, L$, from $f^{(s)}(\mathbf{y}^{(n)} \mid M)$. When $i = 1$, $L^{-1} \sum_l 1(T(\mathbf{y}_l^{(n)*}) > a_1)$ is a Monte Carlo integration for $\Pr(T(\mathbf{y}^{(n)}) > a_1 \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_1))$. When $i = 2$, $L^{-1} \cdot \sum_l 1(T(\mathbf{y}_l^{(n)*}) < a_2)$ is a Monte Carlo integration for $\Pr(T(\mathbf{y}^{(n)}) < a_2 \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_2))$. It only remains to compute $T(\mathbf{y}_l^{(n)*})$ for a given $\mathbf{y}_l^{(n)*}$ which, in turn, requires $S_1(\mathbf{y}_l^{(n)*})$ and $S_2(\mathbf{y}_l^{(n)*})$. All the foregoing screening criteria can be obtained through either direct or iterative simulation.

We next turn to the question of when (10) can be achieved. The answer depends upon the behavior of the sequence of distributions $f^{(s)}(T(\mathbf{y}^{(n)}) \mid M_i)$, $n = 1, 2, \ldots$, $i = 1, 2$, which is conveniently studied through the behavior of the sequence of conditional distributions $f(\mathbf{y}^{(n)} \mid \boldsymbol{\theta}_i, M_i)$, $n = 1, 2, \ldots$. For all customary screening criteria, under M_1 given say $\boldsymbol{\theta}_1$ we may discern three behaviors for $T(\mathbf{y}^{(n)})$:

$$T(\mathbf{y}^{(n)}) \xrightarrow{P} c(\boldsymbol{\theta}_1) > 0, \quad T(\mathbf{y}^{(n)}) \xrightarrow{P} \infty \quad \text{or} \quad T(\mathbf{y}^{(n)}) \xrightarrow{d} T_0 \sim f_0(t_0 \mid \boldsymbol{\theta}_1). \tag{11}$$

In certain cases f_0 does not depend upon $\boldsymbol{\theta}_1$. Similarly, under M_2, given $\boldsymbol{\theta}_2$, with $\boldsymbol{\theta}_2, c(\boldsymbol{\theta}_2) < 0$, and $-\infty$ replacing $\boldsymbol{\theta}_1, c(\boldsymbol{\theta}_1), \boldsymbol{\theta}_1 > 0$ and ∞ in (11). Of course, for a pair M_1, M_2, the limiting behavior under M_1 need not be the same as that under M_2.

Theorem 2 summarizes what can be said regarding (10a) under each of the limits in (11). Analogous results can be developed for (10b). All three limiting cases are illustrated through the analytical results of the next section.

THEOREM 2. *With regard to the sequence of conditional distributions* $f(\mathbf{y}^{(n)} \mid \boldsymbol{\theta}_1, M_1), n = 1, 2, \ldots,$ *at a given* $\boldsymbol{\theta}_1$:

(i) *If* $T(\mathbf{y}^{(n)}) \xrightarrow{p} \infty$ *then* (10a) *is achievable*.

(ii) *If* $T(\mathbf{y}^{(n)}) \xrightarrow{p} c(\boldsymbol{\theta}_1) > 0$, *then* (10a) *need not be achievable*.

(iii) *If* $T(\mathbf{y}^{(n)}) \xrightarrow{d} T_0 \sim f_0(t_0 \mid \boldsymbol{\theta}_1)$ *then* (10a) *need not be achievable*.

The proof is straightforward and given in Appendix C.

8. ANALYTICAL RESULTS FOR MODEL SEPARATION SSD

The preceding development has reduced the model separation problem to the examination of the distribution of the statistic $T(\mathbf{y}^{(n)})$ when $\mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_1)$ and when $\mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_2)$. Here, we are interested in analytical examination of $\lim_{n\to\infty} P(T(\mathbf{y}^{(n)}) > a_1 \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_1))$ and $\lim_{n\to\infty} P(T(\mathbf{y}^{(n)}) < a_2 \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_2))$. Below, we present results for three problems: (i) hypothesis testing where both H_0 and H_A have positive Lebesgue measure in the space of $\boldsymbol{\theta}$ (the nonsingular null case), (ii) the nested (generalized) linear models case, and (iii) the choice between error distributions.

Hypothesis testing with a composite null hypothesis. Suppose $H_0 : \boldsymbol{\theta} \in \Theta_0$ and $H_A : \boldsymbol{\theta} \in \Theta_0^c$. Let $f^{(f)}(\boldsymbol{\theta} \mid M_1) = f^{(f)}(\boldsymbol{\theta} \mid M_2) = f^{(f)}(\boldsymbol{\theta})$ be a common proper prior fitting density for $\boldsymbol{\theta}$ and let $P_{H_0} = \int_{\Theta_0} f^{(f)}(\boldsymbol{\theta}) > 0$, $P_{H_A} = \int_{\Theta_0^c} f^{(f)}(\boldsymbol{\theta}) > 0$. The posterior probabilites $P_{H_0}(\mathbf{y}^{(n)}) \equiv \Pr(\boldsymbol{\theta} \in \Theta_0 \mid \boldsymbol{\theta} \sim f^{(f)}(\boldsymbol{\theta} \mid \mathbf{y}^{(n)}))$ and $P_{H_A}(\mathbf{y}^{(n)}) \equiv \Pr(\boldsymbol{\theta} \in \Theta_0^c \mid \boldsymbol{\theta} \sim f^{(f)}(\boldsymbol{\theta} \mid \mathbf{y}^{(n)}))$ are of interest.

Letting M_1 be H_0 and M_2 be H_A, we recall that

$$\frac{P_{H_0}(\mathbf{y}^{(n)})}{P_{H_A}(\mathbf{y}^{(n)})} = BF_{12}^{(n)} \frac{P_{H_0}}{P_{H_A}}, \tag{12}$$

where $BF_{12}^{(n)}$ is the Bayes factor for model M_1 given data $\mathbf{y}^{(n)}$. From (12), $P_{H_0}(\mathbf{y}^{(n)}) \xrightarrow{P} 1 \Leftrightarrow \ln BF_{12}^{(n)} \xrightarrow{P} \infty$ and $P_{H_A}(\mathbf{y}^{(n)}) \xrightarrow{P} 1 \Leftrightarrow \ln BF_{12}^{(n)} \xrightarrow{P} -\infty$. Next, we introduce a sampling prior $f^{(s)}(\boldsymbol{\theta} \mid M_1)$ restricted to H_0 which induces $f^{(s)}(\mathbf{y}^{(n)} \mid M_1)$ and a sampling prior $f^{(s)}(\boldsymbol{\theta} \mid M_2)$ restricted to H_A which induces $f^{(s)}(\mathbf{y}^{(n)} \mid M_2)$. Then, if $T(\mathbf{y}^{(n)}) = BF_{12}^{(n)}$, it suffices to study the behavior as n grows large of $P_{H_0}(\mathbf{y}^{(n)})$ under $f^{(s)}(\mathbf{y}^{(n)} \mid M_1)$ and of $P_{H_A}(\mathbf{y}^{(n)})$ under $f^{(s)}(\mathbf{y}^{(n)} \mid M_2)$.

Under usual regularity conditions $P_{H_0}(\mathbf{y}^{(n)}) - 1(\hat{\boldsymbol{\theta}}^{(n)} \in \Theta_0) \xrightarrow{P} 0$ where $\hat{\boldsymbol{\theta}}^{(n)}$ is the MLE of $\boldsymbol{\theta}$ based upon $\mathbf{y}^{(n)}$. So, given $\boldsymbol{\theta}$, when $\hat{\boldsymbol{\theta}}^{(n)} \xrightarrow{P} \boldsymbol{\theta}$, $P_{H_0}(\mathbf{y}^{(n)}) \xrightarrow{P} 1(\boldsymbol{\theta} \in \Theta_0)$. Thus

$$\int P_{H_0}(\mathbf{y}^{(n)}) f^{(s)}(\mathbf{y}^{(n)} \mid M_1) \to \Pr\bigl(\boldsymbol{\theta} \in \Theta_0 \mid \boldsymbol{\theta} \sim f^{(s)}(\boldsymbol{\theta} \mid M_1)\bigr). \tag{13}$$

Hence, if the support of $f^{(s)}(\boldsymbol{\theta} \mid M_1)$ is contained in Θ_0, (13) is equal to 1, that is,

$$\lim_{n\to\infty} \Pr(P_{H_0}(\mathbf{y}^{(n)}) > 1 - \varepsilon \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_1)) = 1$$

for all $\varepsilon > 0$. Similarly, if the support of $f^{(s)}(\boldsymbol{\theta} \mid M_2)$ is contained in Θ_0^c, $\lim_{n\to\infty} \Pr(P_{H_A}(\mathbf{y}^{(n)}) > 1 - \varepsilon \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_2)) = 1$ for all $\varepsilon > 0$. Hence, with sampling priors chosen in this fashion we can choose n_1 large enough so that $\Pr(BF_{12}^{(n)} > a_1 \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_1))$ is arbitrarily close to 1 and n_2 large enough so that $\Pr(BF_{12}^{(n)} < a_2 \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_2))$ is arbitrarily close to 1.

Nested linear models. In relating a response variable Y to an explanatory variable X, the nature of the regression function is of interest. A possible design question is choice of sample size to distinguish a lower order polynomial (e.g., a linear relationship) from a higher order polynomial (e.g., a quadratic relationship), that is, to separate nested linear models.

In the literature, in the case of regular models (models where dimension remains fixed as sample size increases), many screening criteria, $S_i(\mathbf{y}^{(n)})$ can be expressed in the form

$$S_i(\mathbf{y}^{(n)}) = \ln f(\mathbf{y}^{(n)} \mid \hat{\boldsymbol{\theta}}_i^{(n)}, M_i) - k(n, p_i) + O_p(1) \tag{14}$$

[Gelfand and Dey (1994)]. In (14), $\hat{\boldsymbol{\theta}}_i^{(n)}$ is the MLE of $\boldsymbol{\theta}_i$ under model M_i and $k(n, p_i)$ is a penalty function which is increasing in n and in p_i where p_i is the dimension of model M_i. In other words, $S_i(\mathbf{y}^{(n)})$ is approximately a penalized log likelihood.

Then, from (14), for $T(\mathbf{y}^{(n)}) = \ln S_1(\mathbf{y}^{(n)})/S_2(\mathbf{y}^{(n)})$, $T(\mathbf{y}^{(n)}) = \ln \lambda_n + (k(n, p_2) - k(n, p_1)) + O_p(1)$ where $\lambda_n = f(\mathbf{y}^{(n)} \mid \hat{\boldsymbol{\theta}}_1^{(n)}, M_1)/f(\mathbf{y}^{(n)} \mid \hat{\boldsymbol{\theta}}_2^{(n)}, M_2)$. If $M_1 \subset M_2$,

under customary conditions, $-2\ln\lambda_n \xrightarrow{d} \chi^2_{p_2-p_1}$ under M_1. Hence, the behavior as $n\to\infty$ of $T(\mathbf{y}^{(n)})$ given $\boldsymbol{\theta}_1$ is readily examined.

In particular, if $k(n,p)$ does depend upon n and $\lim_{n\to\infty} k(n,p)=\infty$ then $\lim_{n\to\infty}\Pr(T(\mathbf{y}^{(n)}) > a_1 \mid \boldsymbol{\theta}_1, M_1)=1$. If $\lim_{n\to\infty} k(n,p)$ is finite, then $\lim_{n\to\infty}\Pr(T(\mathbf{y}^{(n)}) > a_1 \mid \boldsymbol{\theta}_1, M_1)=b_1<1$. (This is merely an elaboration of the well-known inconsistency of the likelihood ratio test for nested models.) Since, in both cases the limit does not depend upon $\boldsymbol{\theta}_1$, these limits hold marginally for $T(\mathbf{y}^{(n)})$ regardless of the choice of $f^{(s)}(\boldsymbol{\theta}_1 \mid M_1)$. So, provided $k(n,p)\to\infty$ as $n\to\infty$, we can choose n_1 to make $\Pr(T(\mathbf{y}^{(n)}) > a_1 \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_1))$ arbitrarily close to 1.

The Bayes factor, the intrinsic Bayes factor [Berger and Pericchi (1996)] and the fractional Bayes factor [O'Hagan (1995)] all provide such a $k(n,p)$. On the other hand, the familiar AIC [Akaike (1973)], the pseudo-Bayes factor [Geisser and Eddy (1979)] and the posterior Bayes factor [Aitkin (1991)] do not. See Gelfand and Dey (1994) for details. Recently proposed criteria such as the DIC [Spiegelhalter, Best, Carlin and van der Linde (2002)] and the posterior predictive loss approach [Gelfand and Ghosh (1998)], at least for regular Gaussian linear models, do not as well.

What happens under M_2? Under suitable conditions, $-2\ln\lambda_n$ will have an approximate noncentral χ^2 distribution. In fact, for a Gaussian linear model, if $(E\mathbf{y}^{(n)}) = X_{1n}\boldsymbol{\beta}_1$ under M_1, $= X_{1n}\boldsymbol{\beta}_1 + X_{2n}\boldsymbol{\beta}_2$ under M_2 then $-2\log\lambda_n \dot{\sim} \chi^2_{p_2-p_1,\gamma_n/2\sigma^2}$ where $\gamma_n = \boldsymbol{\beta}_2^T X_{2,n}^T X_{2,n}\boldsymbol{\beta}_2$. Hence, since customarily $X_{2,n}^T X_{2,n} = O(n)$, $\boldsymbol{\beta}_2$ must be $O(n^{-1/2})$ in order to obtain a finite nonzero limit for γ_n. If $\boldsymbol{\beta}_2 = O_p(1)$ then $\gamma_n \xrightarrow{P}\infty$ and given $\boldsymbol{\beta}_2$, $\ln\lambda_n \xrightarrow{P} -\infty$ and, in fact, $\ln\lambda_n = -O_p(n)$. Similar results apply to generalized linear models [see, e.g., the Appendix of McCullagh and Nelder (1989)]. Hence, if $k(n,p)$ is of order less than n, given $\boldsymbol{\beta}_2$, $T(\mathbf{y}^{(n)}) \xrightarrow{P} -\infty$ and hence marginally $T(\mathbf{y}^{(n)}) \xrightarrow{P} -\infty$ for $\mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_2)$. This order condition holds for all $k(n,p)$ in the literature so for all of the familiar $T(\mathbf{y}^{(n)})$ we can choose n_2 large enough so that $\Pr(T(\mathbf{y}^{(n)}) < a_2 \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_2))$ is arbitrarily close to 1.

Separating error distributions. A frequent concern when modeling continuous data is whether the assumption of Gaussian errors is acceptable. Perhaps a heavier-tailed error distribution (typically a t with small degrees of freedom) is appropriate. Here we consider, at the design stage, whether we can choose a sample size large enough to separate two such error distributions. The formal analysis may be simplified to the separation of two scale parameter families. That is, we presume $y_1,\ldots,y_n$ i.i.d. $\sigma^{-1}f_i(y/\sigma)$, under M_i, $i=1,2$, for example, M_1 is $N(0,\sigma^2)$, M_2 is σt_2 where t_2 denotes a t distribution with 2 degrees of freedom.

TABLE 1
The constants c_1 and c_2 for separating normal (f_1) from t_ν (f_2)

ν	$\mathbf{c}_1$	$\mathbf{c}_2$
1	0.26	$-\infty$
2	0.12	−2.94
5	0.03	−0.11
10	0.01	−0.02
20	0.003	−0.005

Again working with the Bayes factor, we show in Appendix B that, under mild conditions, $\lim_{n\to\infty}\Pr(T(\mathbf{y}^{(n)}) > 0 \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_1)) = 1$ and $\lim_{n\to\infty}\Pr(T(\mathbf{y}^{(n)}) < 0 \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_2)) = 1$ where $T(\mathbf{y}^{(n)}) = \ln BF_{12}^{(n)}$. Hence, we can choose n_1 and n_2 to make these probabilities arbitrarily large. The argument in Appendix B is applicable if f_2 arises as a location mixture of f_1. It is also applicable for separating scale parameter models which are not error distributions such as an exponential from a Weibull.

Recalling the definition of c_1 and c_2 from Appendix D as Kullback–Leibler distances at $\sigma=\sigma_0$, suppose, for example, that $|c_1|<|c_2|$. The above argument suggests that $-\bar{V}_n$ will tend to be larger under f_2 than $-\bar{V}_n$ will tend to be under f_1, that is, that $-\ln BF_{12}^{(n)}$ will tend to be larger under f_2 than $\ln BF_{12}^{(n)}$ will tend to be under f_1. This suggests that for a given $1-\alpha$, n_2 will be smaller than n_1, that is, separation is *easier* under M_2. In fact, this is illustrated in one of our examples in the next section. Here we conclude with Table 1 which provides c_1 and c_2 where f_1 is normal and f_2 is a t with several choices of d.f.

9. TWO EXAMPLES

We provide two illustrations of SSD for model separation. The first addresses separation of a normal model for the data from a heavier-tailed model, in particular a t model. The second considers separation of a common growth curve model from a model with individual growth curves.

9.1 Normal Distribution versus t_2 Distribution

Under $M_1, y_1, \ldots, y_n \sim N(\mu, \sigma^2)$, under M_2, $y_1, \ldots, y_n \sim t_v(\mu, \sigma)$. For simplicity we set $\mu = 0$, taking $f^{(f)}(\sigma \mid M_1) = f^{(f)}(\sigma) = IG(2, 5)$, that is, an inverse gamma distribution with mean 5 and infinite variance. We also set $f^{(s)}(\sigma \mid M_1) = f^{(s)}(\sigma \mid M_2 = U(4, 6))$. Illustration in the previous section ensures that separation is achievable.

Figure 4 show plots of $\Pr(\ln BF_{12}^{(n)} > 0 \mid M_1)$ and $\Pr(\ln BF_{12}^{(n)} < 0 \mid M_2)$ versus n for $v = 1$ (Cauchy), 2 and 5. The jaggedness of the curves is attributable to random error arising from the Monte Carlo integrations to obtain these probabilities. Note that when $v = 1$ we can separate Cauchy from normal with confidence exceeding 0.9 under either model by taking n as small as 15. Of course, with v larger, separation must become more difficult (since the t becomes closer to the normal). This is seen in panel (1c) under M_1, even with v as small as 5. Table 1 and the associated discussion suggest an asymmetry in this separation problem, that separation under M_2 is easier than under M_1 for $v = 2$ and 5. Panels (1b) and (1c) provide detailed support for this conclusion.

9.2 Common Linear Growth Curve versus Individual Linear Growth Curves

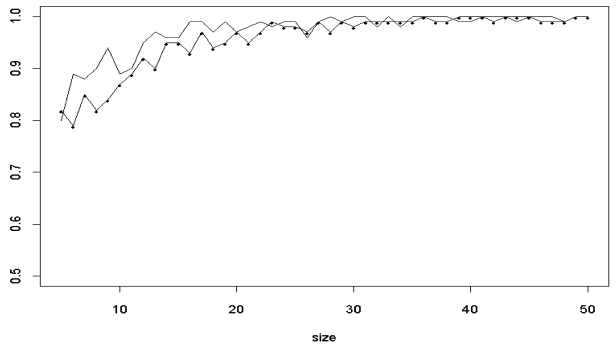
Under M_1, $y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij}$, under M_2, $y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \varepsilon_{ij}$ where $i = 1, \ldots, n$, $j = 1, 2, \ldots, J$. In fact, we set $J = 4$ and, for simplicity, $t_{ij} = j$. With J fixed, the SSD problem is to determine the number of subjects n to separate M_1 from M_2. Because the Bayes factor becomes increasingly difficult to compute as n increases, we illustrate with the screening criterion proposed in Gelfand and Ghosh (1998). This criterion is routine to compute under Markov chain Monte Carlo model fitting, which will be required under M_2. It also allows improper priors. In the present situation, for a given model M, if μ_{ij} is the posterior predictive mean of y_{ij} and σ_{ij}^2 is the posterior predictive variance of y_{ij}, then $S(\mathbf{y}^{(n)}) = (\sum_i \sum_j \sigma_{ij}^2 + \sum_i \sum_j (y_{ij} - \mu_{ij})^2)^{-1}$ (since model support is to increase in S) whence

$$T(\mathbf{y}^{(n)}) = \ln \frac{\sum_i \sum_j \sigma_{ij}^{2(2)} + \sum_i \sum_j (y_{ij} - \mu_{ij}^{(2)})^2}{\sum_i \sum_j \sigma_{ij}^{2(1)} + \sum_i \sum_j (y_{ij} - \mu_{ij}^{(1)})^2}.$$

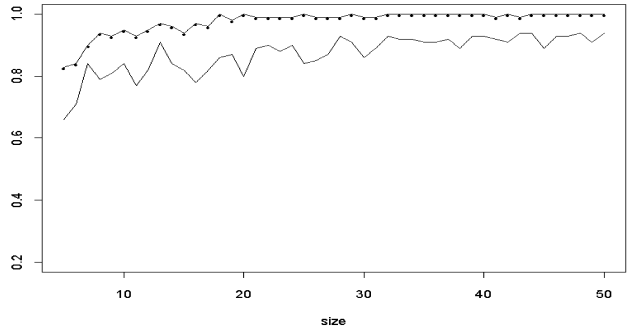
The fitting prior under M_1 is

$$f^{(f)}(\beta_0, \beta_1, \sigma^2 \mid M_1) = f_1^{(f)}(\beta_0) f_1^{(f)}(\beta_1) f_1^{(f)}(\sigma^2)$$

(1a)



(1b)



(1c)

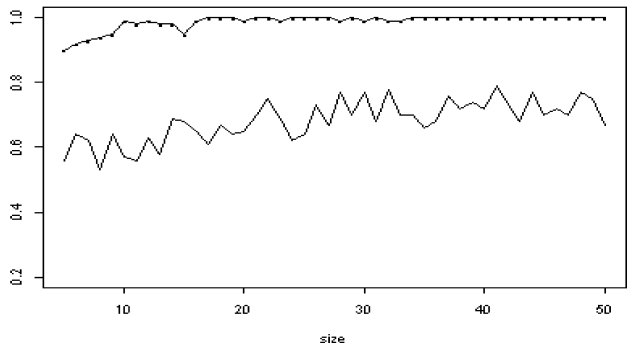


FIG. 4. *Sample size for separating normal* (M_1) *from* $t_v(M_2)$. *In* (1a) $v = 1$, *in* (1b) $v = 2$, *in* (1c) $v = 5$. *The "solid" curve is* $\Pr(\ln BF_{12} > 0 \mid M_1)$, *the "dot-connected" curve is* $\Pr(\ln BF_{12} < 0 \mid M_2)$.

where $f_1^{(f)}(\beta_0) = f_1^{(f)}(\beta_1) = 1$ and $f_1^{(f)}(\sigma^2) = IG(2, 5)$ (as in the previous example). The illustrative sampling prior is $f^{(s)}(\beta_0, \beta_1, \sigma^2 \mid M_1) = f_1^{(s)}(\beta_0) f_1^{(s)}(\beta_1) \cdot f_1^{(s)}(\sigma^2)$ where $f_1^{(s)}(\beta_0) = f_1^{(s)}(\beta_1) = U(0, 1)$ and

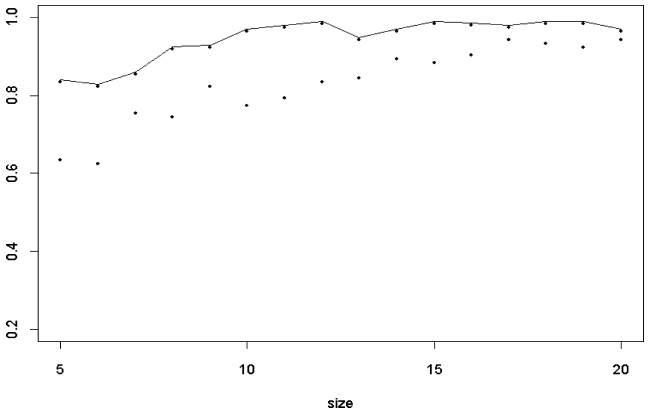


FIG. 5. *Sample size determination to separate an individual linear growth curve model (M_2) from a common growth curve model (M_1). The criterion is that of Gelfand and Ghosh. Solid curve is under M_1, unconnected dots are under M_2.*

$f_1^{(s)}(\sigma^2) = U(4,6)$. The fitting prior under M_2 follows the customary normal-Wishart-inverse gamma form, $\binom{\beta_{0i}}{\beta_{1i}} \sim N(\binom{\mu_0}{\mu_1}, \Sigma)$, $f_2^{(f)}(\mu_0)$ and $f_2^{(f)}(\mu_1)$ are flat, $f_2^{(f)}(\Sigma) \sim IW((\rho R)^{-1}, \rho)$ where $\rho = 2$, $R = \begin{pmatrix} 100 & 0 \\ 0 & 0.1 \end{pmatrix}$ and again $f_2^{(f)}(\sigma^2) = IG(2,5)$. As shown in Gelfand, Hills, Racine-Poon and Smith (1990), in this case all full conditional distributions are standard and a Gibbs sampler for model fitting is routine to implement. The sampling prior under M_2 is again illustrative. Given n, $f_2^{(s)}(\beta_{0i}) = U(i/n, (i+1)/n)$, $f_2^{(s)}(\beta_{1i}) = U(2i/n, (2i+1)/n)$ and again $f_2^{(s)}(\sigma^2) = U(4,6)$.

Figure 5 plots $\Pr(T(\mathbf{y}^{(n)}) > 0 \mid M_1)$ and $\Pr(T(\mathbf{y}^{(n)}) < 0 \mid M_2)$ using $T(\mathbf{y}^{(n)})$ above. Again jaggedness is attributable to error in the Monte Carlo integrations contributing to the calculation of $T(\mathbf{y}^{(n)})$. Model separation is quite good even for n as small as 20.

APPENDIX A

Here we provide a proof of Theorem 1. The following standard probability result [see, e.g., Chung (1974), page 95] is useful.

LEMMA 1. *If $Z_n \xrightarrow{d} Z$ and $\sup_n E|Z_n| \le M$, then $\lim_{n\to\infty} E|Z_n| = E|Z| < \infty$.*

For (i) of the theorem, usual application of the Lebesgue dominated convergence theorem or monotone convergence theorem enables the interchange of limit and integration needed for the first part of (i). For the second part of (i), if we work conditionally, given $\boldsymbol{\theta}$, the lemma provides the result immediately since $T(\mathbf{y}^{(n)}) \ge 0$.

For (ii) of the theorem, the lemma is immediately applicable with $Z_n = T(\mathbf{y}^{(n)}) - T^{(s)}(\mathbf{y}^{(n)})$.

APPENDIX B

For the random effects model of Section 5 we show that, as $J \to \infty$, APVC for $\sigma_e^2 \to 0$ and if, in addition, $I \to \infty$ APVC for $\sigma_\alpha^2 \to 0$. We modify notation slightly, replacing $\mathbf{y}^{(n)}$ with $\mathbf{y}^{(I,J)}$ and define $Q_1 = \sum\sum(y_{ij} - y_{i\cdot})^2$, $Q_2 = J\sum_i(y_{i\cdot} - y_{\cdot\cdot})^2$, $v_1 = I(J-1)$, $v_2 = I-1$ and $r = Q_2/(Q_1+Q_2)$. From Box and Tiao (1973), expression (5.2.37), $f^{(f)}(\sigma_e^2 \mid \mathbf{y}^{(I,J)}) \approx a^{-1}Q_1 IG(b/2, 1/2)$ where $a = (\frac{v_1}{2}+1)\frac{I_r(v_2/2, v_1/2+2)}{I_r(v_2/2, v_1/2+1)} - \frac{v_1}{2}\frac{I_r(v_2/2, v_1/2+1)}{I_r(v_2/2, v_1/2)}$ and $b = \frac{v_1}{a}\frac{I_r(v_2/2, v_1/2+1)}{I_r(v_2/2, v_1/2)}$. Also, from their expression (5.2.61), $f^{(f)}((\sigma_e^2 + J\sigma_\alpha^2) \mid \mathbf{y}^{(I,J)}) \approx c^{-1}Q_2 IG(d/2, 1/2)$ where $c = (v_2/2+1)\cdot\frac{I_r(v_2+2, v_1/2)}{I_r(v_2/2+1, v_1/2)} - v_2/2\frac{I_r(v_2/2+1, v_1/2)}{I_r(v_2/2, v_1/2)}$ and $d = \frac{v_2}{c}\cdot\frac{I_r(v_2/2+1, v_1/2)}{I_r(v_2/2, v_1/2)}$. Hence, after routine calculation, the posterior variance for σ_e^2,

$$T_e(\mathbf{y}^{(I,J)}) \approx 2Q_1^2/a^2(b-2)^2(b-4). \tag{B.1}$$

Instead of working with σ_α^2 directly, we work with $(\sigma_e^2/J) + \sigma_\alpha^2$. Again routine calculation yields its posterior variance for

$$T_\alpha(\mathbf{y}^{(I,J)}) \approx 2Q_2^2/J^2c^2(d-2)^2(d-4). \tag{B.2}$$

Next, as anticipated above, we claim that $T_e \xrightarrow{P} 0$ as $J \to \infty$ and that $T_\alpha \xrightarrow{P} 0$ if, in addition, $I \to \infty$. First note that $I_w(s,t) \to 1$ as $t \to \infty$, which implies $\frac{tI_w(s,t+1)}{I_w(s,t)}$ behaves like t as $t \to \infty$. Next, recall the familiar identity for incomplete Beta integrals [see, e.g., Abramowitz and Stegun (1965), page 944], $I_w(s,t) = s^{-1}B(s,t)w^s(1-w)^t + I_w(s+1,t) = t^{-1}B(s,t)w^s(1-w^t + I_w(s,t+1)$ where $B(s,t) = \Gamma(s+t)/\Gamma(s)\Gamma(t)$. This implies directly that $\lim_{t\to\infty}((t+1)\frac{I_w(s,t+2)}{I_w(s,t+1)} - t\frac{I_w(s,t+1)}{I_w(s,t)})$ exists and equals 1.

Applying these results to (B.1) and (B.2) we find that as $J \to \infty$, $a \xrightarrow{P} 1$, $b \xrightarrow{P} \infty$, $(Q_1/v_1) \xrightarrow{P} (\sigma_e^2)^2$, hence $(Q_1/b)^2 \to (\sigma_e^2)^2$ and thus $T_e \xrightarrow{P} 0$ as $J \to \infty$. Also, $c \xrightarrow{P} 1$ and $d \xrightarrow{P} v_2$ and $Q_2/J = \frac{Q_2}{\sigma_e^2 + J\sigma_\alpha^2}\frac{\sigma_e^2 + J\sigma_\alpha^2}{J} \xrightarrow{P} \sigma_\alpha^2\chi_{v_1}^2$. Hence if in addition $v_2 \to \infty$, that is $I \to \infty$, $T_\alpha \xrightarrow{P} 0$. It is also apparent from (B.1) and (B.2) and the foregoing discussion that the boundedness of

the expectation condition holds so that finally, as $J \to \infty$, APVC for $\sigma_e^2 \to 0$ and if, in addition, $I \to \infty$, APVC for $\sigma_\alpha^2 + \sigma_e^2/J$, hence for $\sigma_\alpha^2 \to 0$.

APPENDIX C

Here we provide a proof of Theorem 2. Let $b(a_1) = \lim_{n\to\infty} \Pr(T(\mathbf{y}^{(n)}) > a_1 \mid M_1)$. Recall that $\Pr(T(\mathbf{y}^{(n)}) > a_1 \mid M_1) = \int f^{(s)}(\boldsymbol{\theta}_1 \mid M_1) \int 1(T(\mathbf{y}^{(n)}) > a_1) f(\mathbf{y}^{(n)} \mid \boldsymbol{\theta}_1, M_1)\, d\mathbf{y}^{(n)}\, d\boldsymbol{\theta}_1$. Hence we can use the Lebesgue dominated convergence theorem to study $b(a_1)$ and hence to argue (i), (ii) and (iii).

For (i), by definition $\lim_{n\to\infty} \Pr(T(\mathbf{y}^{(n)}) > a_1 \mid \boldsymbol{\theta}_1, M_1) = 1$ for all a_1 and $\boldsymbol{\theta}_1$. Hence $b(a_1) = 1$ so (10a) is achievable.

For (ii), we have $b(a_1) = \Pr(c(\boldsymbol{\theta}_1) > a_1 \mid \boldsymbol{\theta}_1 \sim f^{(s)}(\boldsymbol{\theta}_1 \mid M_1))$. So, if $a_1 = 0$, (10a) holds but for $a_1 > 0$, depending upon $f^{(s)}(\boldsymbol{\theta}_1 \mid M_1)$, (10a) may not be achievable.

For (iii), $\lim_{n\to\infty} \Pr(T(\mathbf{y}^{(n)}) > a_1 \mid \boldsymbol{\theta}_1, M_1) = 1 - F_0(a_1 \mid \boldsymbol{\theta}_1)$ so $b(a_1) = 1 - \int F_0(a_1 \mid \boldsymbol{\theta}_1) f^{(s)}(\boldsymbol{\theta}_1 \mid M_1)\, d\boldsymbol{\theta}_1$. If the support for f_0 puts positive mass on $(-\infty, a_1)$ then (10a) need not be achievable.

APPENDIX D

For the problem of separating two scale parameter families in Section 8, we argue that under usual regularity conditions, $\lim_{n\to\infty} \Pr(T(\mathbf{y}^{(n)}) > 0 \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_1)) = 1$ and $\lim_{n\to\infty} \Pr(T(\mathbf{y}^{(n)}) < 0 \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_2)) = 1$ where $T(\mathbf{y}^{(n)}) = \ln BF_{12}^{(n)}$. Since $\Pr(T(\mathbf{y}^{(n)}) > 0 \mid \mathbf{y}^{(n)} \sim f^{(s)}(\mathbf{y}^{(n)} \mid M_1)) = \int f_1^{(s)}(\sigma) \Pr(T(\mathbf{y}^{(n)}) > 0 \mid \mathbf{y}^{(n)} \sim \sigma^{-n} \prod_{i=1}^n f_1(y_i/\sigma))$, if $\lim_{n\to\infty} \Pr(BF_{12}^{(n)} > 1 \mid \mathbf{y}^{(n)} \sim \sigma^{-n} \cdot \prod_{i=1}^n f_1(y_i/\sigma)) = 1$ for each σ, we have demonstrated the first limit. Let $f^{(f)}(\sigma)$ be a common proper fitting prior for σ. Then

$$
\begin{aligned}
BF_{12}^{(n)} &= \frac{\int \sigma^{-n} \prod_{i=1}^n f_1(y_i/\sigma) f^{(f)}(\sigma)}{\int \sigma^{-n} \prod_{i=1}^n f_2(y_i/\sigma) f^{(f)}(\sigma)} \\
&= \frac{\int \lambda(\sigma; \mathbf{y}^{(n)}) \sigma^{-n} \prod_{i=1}^n f_2(y_i/\sigma) f^{(f)}(\sigma)}{\int \sigma^{-n} \prod_{i=1}^n f_2(y_i/\sigma) f^{(f)}(\sigma)} \qquad \text{(D.1)} \\
&= E\big(\lambda(\sigma; \mathbf{y}^{(n)}) \mid \sigma \sim f(\sigma \mid \mathbf{y}^{(n)}, M_2)\big),
\end{aligned}
$$

where $\lambda(\sigma; \mathbf{y}^{(n)}) = \prod_{i=1}^n f_1(y_i/\sigma)/f_2(y_i/\sigma)$.

Hence given $\sigma = \sigma_0$, under usual regularity conditions, $BF_{12}^{(n)} - \lambda(\sigma_0; \mathbf{y}^{(n)}) \xrightarrow{P} 0$. Therefore, $\lim_{n\to\infty} \Pr(BF_{12}^{(n)} > 1 \mid \mathbf{y}^{(n)} \sim \sigma_0^{-n} \prod_{i=1}^n f_1(y_i/\sigma_0)) = \lim_{n\to\infty} \Pr(\lambda(\sigma_0; \mathbf{y}^{(n)}) > 1 \mid \mathbf{y}^{(n)} \sim \sigma_0^{-n} \cdot \prod_{i=1}^n f_1(y_i/\sigma_0))$. But $\log \lambda(\sigma_0; \mathbf{y}^{(n)}) = \sum_{i=1}^n V_i$ where the V_i are i.i.d., $V_i = f_1(y_i/\sigma_0)/f_2(y_i/\sigma_0)$. So, if $y_i \sim \sigma_0^{-1} f_1(y_i/\sigma_0)$, $\bar{V}_n \xrightarrow{\text{a.s.}} \int \sigma_0^{-1} f_1(y/\sigma_0) \log(f_1(y/\sigma_0)/f_2(y/\sigma_0))\, dy \equiv c_1 > 0$ where it is clear that c_1 does not depend upon σ_0. As a result, $\lim_{n\to\infty} P(\sum V_i > 0) = 1$ and we are done.

Similarly, $\lim_{n\to\infty} \Pr(BF_{12}^{(n)} < 1 \mid \mathbf{y}^{(n)} \sim \sigma_0^{-n} \cdot \prod_{i=1}^n f_2(y_i/\sigma_0)) = \lim_{n\to\infty} \Pr(\lambda(\sigma_0; \mathbf{y}^{(n)}) < 1 \mid \mathbf{y}^{(n)} \sim \sigma_0^{-n} \prod_{i=1}^n f_2(y_i/\sigma_0))$. Now $\bar{V}_n \xrightarrow{\text{a.s.}} \int \sigma_0^{-1} \cdot f_2(y/\sigma_0) \log(f_1(y/\sigma_0)/f_2(y/\sigma_0))\, dy \equiv c_2 < 0$ so $\lim_{n\to\infty} P(\sum V_i < 0) = 1$ enabling the second limit in the foregoing paragraph.

ACKNOWLEDGMENTS

A portion of this work is contained in the first author's Ph.D. dissertation completed at the University of Connecticut. The work of the second author was supported in part by NSF Grant DMS-99-71206.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A. (1965). *Handbook of Mathematical Functions*. Dover, New York.

ADCOCK, C. J. (1995). The Bayesian approach to determination of sample sizes: Some comments on the paper by Joseph, Wolfson and du Berger. *The Statistician* **44** 155–161.

ADCOCK, C. J. (1997). Sample size determination: A review. *The Statistician* **46** 261–283.

AITKIN, M. (1991). Posterior Bayes factors (with discussion). *J. Roy. Statist. Soc. Ser. B* **53** 111–142.

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory* (B. N. Petrov and F. Csáki, eds.) 267–281. Akadémiai Kiadó, Budapest.

BERGER, J. O. and PERICCHI, L. R. (1996). The intrinsic Bayes factor for linear models (with discussion). In *Bayesian Statistics V* (J. O. Berger, J. M. Bernardo, A. P. David, D. V. Lindley and A. F. M. Smith, eds.) 25–44. Oxford Univ. Press.

BOX, G. E. P. and TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA.

CHUNG, K. L. (1974). *A Course in Probability Theory*, 2nd ed. Academic Press, New York.

COHEN, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Erlbaum, Hillsdale, NJ.

DESU, M. M. and RAGHAVARAO, D. (1990). *Sample Size Methodology*. Academic Press, New York.

GEISSER, S. and EDDY, W. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74** 153–160.

GELFAND, A. E. and DEY, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. Roy. Statist. Soc. Ser. B* **56** 501–514.

GELFAND, A. E. and GHOSH, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85** 1–11.

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. *J. Amer. Statist. Assoc.* **85** 972–985.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

GELMAN, A., MENG, X.-L. and STERN, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sinica* **6** 733–807.

GILKS, W. R. and WILD, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41** 337–348.

INOUE, L., BERRY, D. A. and PARMIGIANI, G. (2000). A Bayesian view of the classical sample size determination. Technical report, Dept. Biostatistics, MD Anderson Cancer Center, Houston.

JOSEPH, L. and BELISLE P. (1997). Bayesian sample size determination for normal means and differences between normal means. *The Statistician* **46** 209–226.

JOSEPH, L. and WOLFSON, D. B. (1997). Interval-based versus decision theoretic criteria for the choice of sample size. *The Statistician* **46** 145–149.

JOSEPH, L., WOLFSON, D. B. and DU BERGER, R. (1995a). Sample size calculations for binomial proportions via highest posterior density intervals. *The Statistician* **44** 143–154.

JOSEPH, L., WOLFSON, D. B. and DU BERGER, R. (1995b). Some comments on Bayesian sample size determination. *The Statistician* **44** 167–171.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

KRAEMER, H. C. and THIEMANN, S. (1987). *How Many Subjects? Statistical Power Analysis in Research*. Sage, Newbury Park, CA.

LINDLEY, D. V. (1997). The choice of sample size. *The Statistician* **46** 129–138.

LIU, G. and LIANG, K.-Y. (1997). Sample size calculations for studies with correlated observations. *Biometrics* **53** 937–947.

MCCULLAGH, P. and NELDER J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

MORRIS, C. N. (1983). Natural exponential families with quadratic variance functions: Statistical theory. *Ann. Statist.* **11** 515–529.

MULLER, K. E., LAVANGE, L. M., RAMEY, S. L. and RAMEY, C. T. (1992). Power calculations for general linear multivariate models including repeated measures applications. *J. Amer. Statist. Assoc.* **87** 1209–1226.

MÜLLER, P. and PARMIGIANI, G. (1995). Optimal design via curve fitting of Monte Carlo experiments. *J. Amer. Statist. Assoc.* **90** 1322–1330.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison. *J. Roy. Statist. Soc. Ser. B* **57** 99–138.

PHAM-GIA, T. (1995). Sample size determination in Bayesian statistics: A commentary. *The Statistician* **44** 163–166.

PHAM-GIA, T. (1997). On Bayesian analysis, Bayesian decision theory and the sample size problem. *The Statistician* **46** 139–144.

RAHME, E., JOSEPH, L. and GYORKOS, T. W. (2000). Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *Appl. Statist.* **49** 119–128.

RUBIN, D. B. and STERN, H. S. (1998). Sample size determination using posterior predicitive distributions. *Sankhyā Ser. B* **60** 161–175.

SELF, S. G. and MAURITSEN, R. H. (1988). Power/sample size calculations for generalized linear models. *Biometrics* **44** 79–86.

SELF, S. G., MAURITSEN, R. H. and O'HARA, J. (1992). Power calculations for likelihood ratio tests in generalized linear models. *Biometrics* **48** 31–39.

SHUSTER, J. J. (1993). *Practical Handbook of Sample Size Guidelines for Clinical Trials*. CRC Press, Boca Raton, FL.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. Roy. Statist. Soc. Ser. B*. To appear.

WEISS, R. (1997). Bayesian sample size calculations for hypothesis testing. *The Statistician* **46** 185–191.

ZOU, K. H. and NORMAND, S. L. (2001). On determination of sample size in hierarchical binomial models. *Statistics in Medicine* **20** 2163–2182.