ARTICLE IN PRESS

ELSEVIER

Available at WWW.MATHEMATICSWEB.ORG POWERED BY SCIENCE DIRECT®

Journal of Statistical Planning and Inference ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Decomposing posterior variance☆

Paul Gustafson*, Bertrand Clarke

Department of Statistics, University of British Columbia, 333-6356 Agricultural Road, Vancouver, Canada BC V6T 1Z2

Received 10 January 2001; accepted 8 September 2002

Abstract

We propose a decomposition of posterior variance somewhat in the spirit of an ANOVA decomposition. Terms in this decomposition come in pairs. Given a single parametric model, for instance, one term describes uncertainty arising because the parameter value is unknown while the other describes uncertainty propagated via uncertainty about which prior distribution is appropriate for the parameter. In the context of multiple candidate models and model-averaged estimates, two additional terms emerge resulting in a four-term decomposition. In the context of multiple spaces of models, six terms result. The value of the decomposition is twofold. First, it yields a fuller accounting of uncertainty than methods which condition on data-driven choices of models or model spaces. Second, it constitutes a novel approach to the study of prior influence in Bayesian analysis.

Keywords: Bayesian robustness; Model averaging; Prior sensitivity; Standard error

1. Introduction

Say a Bayesian analysis is to be undertaken in the face of uncertainty about the correct parameter value within a parametric model, the correct model within a collection or space of models, and the correct space within a collection of spaces. For instance, in the context of estimating an unknown function the different spaces might correspond to different types of basis functions, the different models might correspond to different subsets of basis functions of a given type, and the different parameter values might correspond to different coefficients for the subset of basis functions. In addition to uncertainty about the space, model and parameter themselves, there may be

☆ This research was supported by the Natural Sciences and Engineering Research Council of Canada.
* Corresponding author. Tel.: +1-604-822-1300; fax: +1-604-822-6960.
E-mail address: gustaf@stat.ubc.ca (P. Gustafson).

0378-3758/02/$ - see front matter
doi:10.1016/S0378-3758(02)00491-3

uncertainty about how best to weight, or assign priors to, these quantities. Fortunately Bayes theorem provides a principled way to construct point estimates in the face of these multiple uncertainties. Moreover, to assess the precision of such an estimate the posterior standard deviation of the estimand can be reported as a *standard error* (SE).

This article focusses on a decomposition for such a standard error in the face of multiple sources of uncertainty. Formally, we examine $\mathrm{Var}(\Psi|D)$, the posterior variance of a scalar estimand Ψ given data D. There is no impediment to considering vector estimands, but for ease of exposition we develop our ideas in the context of a scalar estimand. The central idea is to give an ANOVA-like decomposition for this quantity, so that the total SE, which we take to be the square-root of the posterior variance, can be expressed as a root sum of squares. In particular, each individual component in this decomposition is viewed as the SE due to a particular source of uncertainty. To be more specific, our decomposition takes the form

$$\begin{aligned}\mathrm{SE}^2[\mathrm{tot}] &= \mathrm{Var}(\Psi|D)\\ &= \mathrm{SE}^2[\mathrm{par}] + \mathrm{SE}_*^2[\mathrm{par}] + \mathrm{SE}^2[\mathrm{mod}] + \mathrm{SE}_*^2[\mathrm{mod}]\\ &\quad + \mathrm{SE}^2[\mathrm{spc}] + \mathrm{SE}_*^2[\mathrm{spc}], \end{aligned} \tag{1}$$

where par, mod, and spc refer to parameter, model and model space, respectively.

The operational meaning of the decomposition (1) is quite simple. With x being either par, mod, or spc, $\mathrm{SE}[x]$ describes the a posteriori uncertainty associated with the estimate $\hat{\psi} = E(\Psi|D)$ that results from the value of x being unknown for a given prior on x, while $\mathrm{SE}_*[x]$ describes the additional a posteriori uncertainty resulting from uncertainty about which prior to assign to x. In this sense, $\mathrm{SE}_*[x]$ measures prior influence, and can be regarded as being a "higher-order" term in relation to $\mathrm{SE}[x]$. Of course all six terms in (1) are not always needed. In the face of a single space of models only the first four terms are manifested, and with a single model only the first two terms are manifested.

There is considerable literature advocating *Bayesian model averaging* as a technique for estimation in the face of competing models. Some key references include Draper (1995), Kass and Raftery (1995), Clyde (1999), Hoeting et al. (1999), and Fernandez et al. (2001). A theme in this literature is that Bayesian model averaging leads to more realistic uncertainty assessments than methods which use the data to arrive at a single model and then make uncertainty assessments as if that model were known to be correct. Indeed, Draper, Kass and Raftery, and Hoeting et. al. give decompositions which can be viewed as corresponding to *some* of the terms in (1). They do not, however, include terms which reflect uncertainty about appropriate priors. That is, the SE_* terms in (1) are entirely absent. Thus a key contribution of this article is to develop these terms via a hierarchical prior specification and appropriate use of conditional means and expectations.

There are two major and transparent benefits arising from the decomposition. First, SE[tot] better reflects the uncertainty associated with a point estimate by avoiding unwarranted conditioning on a model or model space selected in a data-driven manner. Moreover, the constituent terms give insight into which a priori uncertainties are more or less responsible for the a posteriori uncertainty in the estimate.

Second, the decomposition yields a novel approach to the assessment of prior influence in Bayesian analysis. Most formal schemes for such assessment are based on extremes, examining, for instance, the range of a posterior quantity as the prior distribution varies in a large class of distributions (see, for example, the volume edited by Rios Insua and Ruggeri, 2000, and the references therein). This sort of approach, however, has not become popular in practice. This lack of appeal may derive in part from incompatibility with the mainstream view that uncertainty assessments should be based on typical or average error rather than worst-case error. In this regard our decomposition of the posterior variance under a single model into SE[par] and SE_*[par] terms may prove more appealing. Moreover, our approach is less computationally demanding than an 'extremal' analysis of prior influence.

The remainder of the article is organized as follows. In Section 2 we lay out the formal structure that defines the terms in (1), and then in Section 3 we suggest routes to the hierarchical prior specification demanded by this structure. Section 4 focusses on computational issues. Sections 5–7 illustrate the use of the proposed partitioned standard error in three practical examples. Finally, in Section 8 we revisit the main motivating ideas behind the decomposition.

2. The basic decomposition

2.1. Forms of the terms

The decomposition of the form (1) that we propose is based on the standard identity relating a variance to a conditional mean and a conditional variance. That is,

$$\mathrm{Var}(V) = E\,\mathrm{Var}(V|W) + \mathrm{Var}\,E(V|W). \tag{2}$$

2.2. Parameter uncertainty

Say that a single parametric model is under consideration, with Θ and D denoting the parameter vector and the observable data respectively. The prior distribution for Θ is taken to depend on a hyperparameter Ω. We are interested in representing uncertainty about which prior is appropriate, and therefore Ω itself is assigned a prior distribution rather than a fixed value. Thus we are adopting the common hierarchical Bayes approach to prior specification. If we let $\Psi = \Psi(\Theta)$ denote the scalar estimand and apply (2) conditional on D we obtain

$$\mathrm{Var}(\Psi|D) = E_{\Omega|D}\mathrm{Var}(\Psi|\Omega, D) + \mathrm{Var}_{\Omega|D}E(\Psi|\Omega, D) \tag{3}$$

as a decomposition of the posterior variance for the estimand. The terms on the right in (3) now include the uncertainty about the prior explicitly. In the first term, the inner variance summarizes the a posteriori uncertainty about Ψ for a given prior. The outer expectation then averages this uncertainty across priors, with a weighting determined by the posterior distribution of the hyperparameter Ω given the data. Thus the first term reflects the usual sense of statistical uncertainty that results from not knowing

the correct parameter value, hence we set

$$\mathrm{SE}^2[\mathrm{par}] = E_{\Omega|D}\mathrm{Var}(\Psi|\Omega, D). \tag{4}$$

In contrast the second term reflects the across-prior variation in the estimator $E(\Psi|\Omega, D)$, the posterior mean for a given prior. Thus we set

$$\mathrm{SE}_*^2[\mathrm{par}] = \mathrm{Var}_{\Omega|D}E(\Psi|\Omega, D) \tag{5}$$

to represent the a posteriori uncertainty in the estimator $E(\Psi|D)$ that results from being uncertain about the appropriate choice of prior. Thus in the face of certainty about the appropriate model, (4) and (5) provide a two-term partition for the posterior variance of the estimand, corresponding to the first two terms in (1). To identify the other terms in (1) in the presence of additional sources of uncertainty, we will apply (2) repeatedly.

2.3. Parameter and model uncertainty

Now say that a priori there are multiple candidates for the appropriate model M. For simplicity we will generically refer to the parameter vector within a given model as Θ, although of course the notation Θ_M would be more precise, as the parameter vectors for different models are logically distinct entities. Also, the estimand Ψ should be regarded as a function of both Θ and M, and must have a common interpretation across models.

We assume the candidate models are assigned prior probabilities, encapsulated as a hyperparameter Λ. In analogy to the treatment of a prior within a model, Λ is assigned a prior distribution rather than a fixed value. Application of (2) then yields

$$\mathrm{Var}(\Psi|D) = E_{M|D}\mathrm{Var}(\Psi|M, D) + \mathrm{Var}_{M|D}E(\Psi|M, D). \tag{6}$$

The inner variance in the first term of (6) is simply the posterior variance of the parameter of interest given a single model, and so can be decomposed exactly as per (3). This yields

$$\begin{aligned} E_{M|D}\mathrm{Var}(\Psi|M, D) = {} & E_{M|D}E_{\Omega|M,D}\mathrm{Var}(\Psi|\Omega, M, D) \\ & + E_{M|D}\mathrm{Var}_{\Omega|M,D}E(\Psi|\Omega, M, D) \end{aligned}$$

and so we identify

$$\begin{aligned} \mathrm{SE}^2[\mathrm{par}] &= E_{M|D}E_{\Omega|M,D}\mathrm{Var}(\Psi|\Omega, M, D) \\ &= E_{M,\Omega|D}\mathrm{Var}(\Psi|\Omega, M, D) \end{aligned} \tag{7}$$

and

$$\mathrm{SE}_*^2[\mathrm{par}] = E_{M|D}\mathrm{Var}_{\Omega|M,D}E(\Psi|\Omega, M, D). \tag{8}$$

That is, these terms differ from the analogous terms in Section 2.2 only via the additional outer expectation to average across competing models.

To identify the third and fourth terms in (1), we decompose the second term in (6) according to (2) to obtain

$$\mathrm{Var}_{M|D}E(\Psi|M,D) = E_{\Lambda|D}\mathrm{Var}_{M|\Lambda,D}E(\Psi|M,D) + \mathrm{Var}_{\Lambda|D}E_{M|\Lambda,D}E(\Psi|M,D).$$

The first term on the right represents the uncertainty arising from not knowing the correct model for a given prior distribution across models, while the second term represents the variation in the model-averaged estimator $E_{M|\Lambda,D}E(\Psi|M,D)$ as the across-model prior determined by Λ varies. Thus we identify

$$\mathrm{SE}^2[\mathrm{mod}] = E_{\Lambda|D}\mathrm{Var}_{M|\Lambda,D}E(\Psi|M,D) \tag{9}$$

and

$$\mathrm{SE}_*^2[\mathrm{mod}] = \mathrm{Var}_{\Lambda|D}E_{M|\Lambda,D}E(\Psi|M,D) = \mathrm{Var}_{\Lambda|D}E(\Psi|\Lambda,D). \tag{10}$$

Consequently (7)–(10) yield a four-term decomposition for the posterior variance of the parameter of interest, corresponding to the first four terms in (1).

We emphasize that $\mathrm{SE}^2[\mathrm{par}]$ and $\mathrm{SE}^2[\mathrm{mod}]$ are slight extensions of the usual two terms given for a posterior variance in the presence of model uncertainty but fixed prior distributions (Draper, 1995; Kass and Raftery, 1995; Hoeting et al., 1999). In particular, we can write (7) as

$$\mathrm{SE}^2[\mathrm{par}] = E_{M|D}E_{\Omega|M,D}\mathrm{Var}(\Psi|\Omega,M,D).$$

For fixed within-model priors Ω is known so that no actual averaging takes place in the inner expectation and we reduce back to the usual term. Similarly, for a fixed across-model prior Ψ is known, no actual averaging takes place in the expectation in (9), and we reduce back to the usual term. Thus the present extension involves extra averaging with respect to plausible values of the hyperparameters governing the priors (Ω and Λ) in light of the data. Typically we expect this averaging to yield similar values for SE[par] and SE[mod] as would be obtained by fixing the hyperparameters.

2.4. Parameter, model and model space uncertainty

It remains to see how the last two terms in (1) can arise. Operationally they are manifested when we consider different possibilities for the model space, which we denote as S. Again for the sake of simplicity we retain M to denote a particular model within a model space, although M_S would be a more precise notation. We let Δ be a hyperparameter which indexes the prior distribution on S, and as usual consider Δ to itself have a prior distribution rather than a fixed value.

Using (2) we obtain

$$\mathrm{Var}(\Psi|D) = E_{S|D}\mathrm{Var}(\Psi|S,D) + \mathrm{Var}_{S|D}E(\Psi|S,D), \tag{11}$$

in which the variance inside the first term is the same as the left-hand side of (6) for a particular model space S. Since we already have a decomposition of (6) into the sum

of (7)–(10), we simply average each term with respect to $S|D$ in order to obtain the analogous terms in the more general scenario. That is, we can determine the first four terms in the general six-term decomposition (1).

To get the last two terms in (1), note that the argument of the variance in the second term of (11) is $E(\Psi|S,D)$, the model-averaged estimator for a given model space S. Applying (2) to this variance we obtain

$$\begin{aligned}\mathrm{Var}_{S|D}E(\Psi|S,D) &= E_{\Delta|D}\mathrm{Var}\{E(\Psi|S,D)|\Delta,D\} + \mathrm{Var}_{\Delta|D}E\{E(\Psi|S,D)|\Delta,D\}\\ &= E_{\Delta|D}\mathrm{Var}\{E(\Psi|S,D)|\Delta,D\} + \mathrm{Var}_{\Delta|D}E(\Psi|\Delta,D). \qquad (12)\end{aligned}$$

The first term on the right in (12) represents the uncertainty that results from not knowing the correct model space for a given prior across the model spaces, and so is taken to be SE[spc]. The second term examines the variation in the model and model-space averaged estimator $E(\Psi|\Delta,D)$ as the prior across models spaces varies, and so is taken to be SE_*[spc]. Altogether then we have the six terms listed in (1), applicable when there is uncertainty about both the values and priors for the parameter, model and model space.

Note that our development is intensely hierarchical; from lowest level to highest level we deal with parameter nested inside model nested inside model space. Point estimates and SE components are computed via E and Var operators applied iteratively, with inner terms corresponding to lower levels. In principle one could extend the approach *ad infinitum*, with an extra two terms arising at each level of nesting. In practical terms, however, even one more iteration to "spaces of spaces" would likely be of little interest.

3. Specifying extended priors

Here we describe relatively general approaches to the hierarchical specification of priors as demanded by the development in the previous section. In principle, there are many ways to approach this specification, and so we emphasize that other schemes could be envisioned without straying from the framework of Section 2.

3.1. Prior within a model

In the case of the prior for a continuous parameter θ within a given model, we presume there is some fixed or "baseline" prior density $f_0(\theta)$ around which we want to construct a hierarchical prior. We start by considering a one-to-one parameterization from θ to $(\phi,\mu_1,\ldots,\mu_s)$, so that the components of μ have independent Uniform(0,1) distributions under the baseline prior. Typically, the probability integral transform can be used to obtain a suitable reparameterization.

With the reparameterization in hand, we take the conditional distribution of $\phi|\mu_1,\ldots,\mu_s$ under the extended prior to be the same as under the baseline prior, but specify the prior distribution of each μ_i in a hierarchical fashion. In particular we take

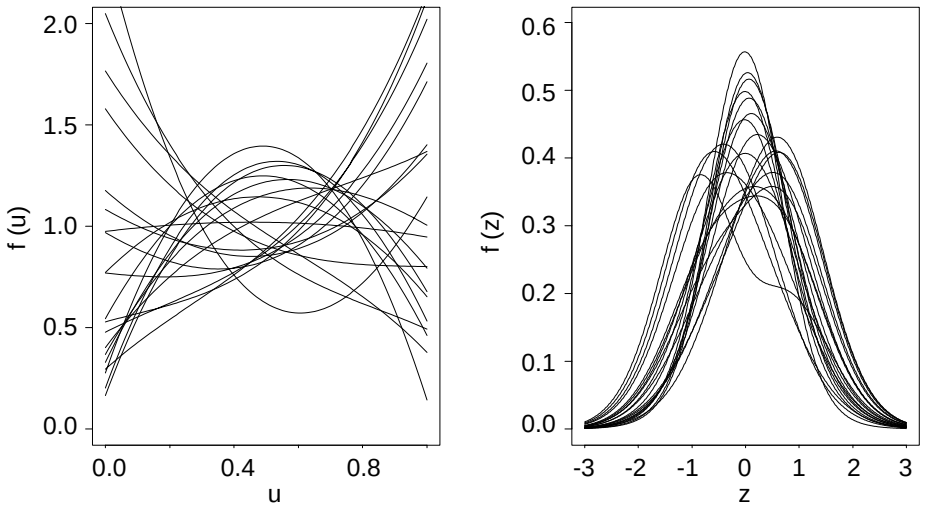

Fig. 1. Default Prior Uncertainty. The left panel gives densities $f(\mu|w)$ for a sample of w values using the default specification. The right panel gives corresponding densities after transformation to a standard normal baseline.

$f(\mu|\omega) = \prod_{i=1}^{s} f(\mu_i|\omega_i)$, where

$$f(\mu_i|\omega_i) = \sum_{j=1}^{r} \omega_{ij} b(\mu_i|\alpha_j, \beta_j) \tag{13}$$

with $b(\cdot|\alpha, \beta)$ denoting a beta density with parameters α and β.

Since each ω_i is an r-dimensional probability vector, it is simple to complete the hierarchical specification by assigning independent Dirichlet$(c_1\bar{\omega})$ distributions to each ω_i, where $\bar{\omega}$ is a specified probability vector. Thus, we can view the distribution of priors on μ_i as being centred at $f(\mu_i|\omega_i = \bar{\omega})$, with c_1 controlling the degree of concentration around this centre.

While there is much flexibility with this specification, we have found the choices of $r = 4$, $(\alpha_1, \beta_1) = (1, 1)$, $(\alpha_2, \beta_2) = (2, 2)$, $(\alpha_3, \beta_3) = (1, 4)$, $(\alpha_4, \beta_4) = (4, 1)$, $\bar{w} = (2/6, 2/6, 1/6, 1/6)$, and $c_1 = 4$ to provide reasonable default settings. Note, in particular, that $f(\mu_i|\omega_i = \bar{\omega})$ is the Uniform$(0, 1)$ density. Thus the extended prior for each μ_i is centred around the uniform distribution, and consequently the extended prior for θ is centred around the corresponding baseline prior. Fig. 1 gives plots of $f(\mu_i|\omega_i)$ for a sample of ω_i values, on the uniform scale and also transformed to the standard normal scale. We see that the chosen settings induce substantial variation in the prior, without giving weight to prior densities which are artificially rough. On the other hand, of course, completely different specifications can be used as the analyst sees fit. For instance, one intuitive hierarchical extension of a baseline normal distribution involves placing a prior over the degrees of freedom in the Student's t distribution.

3.2. Priors across models and model spaces

In analogy to the treatment of the parameter within a model, we must also extend the prior distribution over models within a model space. There is a simple way to accomplish this when there are a finite number of models in the space. In particular, let the k-dimensional probability vector λ denote the prior probabilities for the k models in the space, and say $\lambda = \bar{\lambda}$ corresponds to the baseline prior probabilities. We extend the prior by assigning a Dirichlet($c_2\bar{\lambda}$) distribution to λ. As a default we suggest $c_2 = 4$, to be consistent with the default specification for the magnitude of uncertainty about the prior within a model.

Similarly, if δ denotes the prior probabilities assigned to competing model spaces, then we replace the baseline prior $\delta = \bar{\delta}$ with a Dirichlet($c_3\bar{\delta}$) distribution for δ. Again for the sake of consistency we take $c_3 = 4$ as a default value.

4. Computing the terms

Next we describe the three key algorithms used to compute the partitioned standard errors. We emphasize that these algorithms are relatively simple additions to whatever MCMC analysis is used under the baseline priors.

4.1. Posterior sampling under the extended prior

Algorithm 1 enables MCMC sampling from the posterior distribution of $(\theta, \omega|d)$, the posterior distribution under the extended prior in the context of a single model. We use a standard "trick" for posterior analysis of mixture models: a vector of latent variables ξ is introduced, where ξ_i indicates which component of the mixture gives rise to μ_i as defined in the extended prior. Thus we are interested in the joint posterior distribution of $(\theta, \xi, \omega|d)$, and we apply MCMC updates to $(\theta|\xi, \omega, d)$, $(\xi|\theta, \omega, d)$ and $(\omega|\theta, \xi, d)$ in turn.

To update $(\theta|\xi, \omega, d)$ we simply tweak whatever updating schemes are used for θ under the baseline prior. In particular, we do not alter the scheme used to propose a candidate value θ^* given a current value θ. We simply modify the Metropolis-Hasting acceptance probability computed under the baseline prior by the multiplicative factor

$$\prod_{i=1}^{s} \frac{b(\mu_i(\theta^*)|\alpha_{\xi_i}, \beta_{\xi_i})}{b(\mu_i(\theta)|\alpha_{\xi_i}, \beta_{\xi_i})},$$

to account for the replacement of the uniform prior for μ_i with the appropriate mixture component from (13).

The next updating step involves sampling from the discrete distribution of $(\xi|\theta, \omega, d)$. The conditional probability that ξ_i takes the value j is proportional to $b(\mu_i(\theta)|\alpha_j, \beta_j) \times \omega_{ij}$, and so upon normalization independent updates to each component of ξ are easily implemented.

The final update involves sampling from $(\omega|\theta,\xi,d)$. This is also very simple, since the conditional distribution of $(\omega_i|\theta,\xi=j,d)$ is Dirichlet$(c_1\bar{\omega}+v_j)$, where v_j is the unit vector with one as the jth entry and zero elsewhere.

As a general comment we find this algorithm to be quite efficient under the default mixture specification. In particular, the four beta densities comprising the mixture (13) are relatively flat with respect to one another, and this facilitates good MCMC mixing. We also emphasize that there is very little effort required to implement this algorithm beyond what is required for MCMC sampling under the baseline prior.

4.2. Computing the decomposition for a given model

Algorithm 2 is designed to compute $E_{\Omega|D}\mathrm{Var}(\Psi|\Omega,D)$ and $\mathrm{Var}_{\Omega|D}E(\Psi|\Omega,D)$ in the context of a single model. We do this by drawing two MCMC samples. The first, $\theta^{(1,j)}$, $j=1,\ldots,t$, is drawn from the posterior distribution resulting from the baseline prior. The second, $(\theta^{(2,j)},\omega^{(2,j)})$, $j=1,\ldots,t$ is drawn from the extended prior using Algorithm 1. For a given ω we can use importance sampling to obtain

$$\hat{E}(\omega)=\frac{\sum_{j=1}^{t}\psi^{(1,j)}\{f(\theta^{(1,j)}|\omega)/f_0(\theta^{(1,j)})\}}{\sum_{j=1}^{t}\{f(\theta^{(1,j)}|\omega)/f_0(\theta^{(1,j)})\}}$$

as an estimate of $E(\Psi|\Omega{=}\omega,D)$. Thus, the sample variance of $\hat{E}(\omega^{(2,1)}),\ldots,\hat{E}(\omega^{(2,t)})$ is a Monte Carlo estimate of $\mathrm{Var}_{\Omega|D}E(\Psi|\Omega,D)$. Moreover, by interchanging the mean and variance operations we can similarly obtain a Monte Carlo estimate of $E_{\Omega|D}\mathrm{Var}(\Psi|\Omega,D)$. Again, the relative flatness of the component densities in (13) under the default specification makes this algorithm quite efficient.

4.3. Computing the across-model terms

While the first two algorithms yield the requisite within-model computations, Algorithm 3 yields the across-model quantities for a given space. Given the marginal density of the data under the various models, $f(d|M=i)$, $i=1,\ldots,k$, it is trivial to sample from $M|\lambda,d$ since

$$\Pr(M=i|\lambda,d)=\frac{f(d|M=i)\lambda_i}{\sum_{j=1}^{k}f(d|M=j)\lambda_j}.$$

It is also simple to sample from $\Lambda|M=i,d\sim$ Dirichlet$(c_2\bar{\lambda}+v_i)$, and thus a MCMC sample from $M,\lambda|D$ can be obtained. Moreover, since the quantities of interest $\mathrm{Var}_{M|\Lambda,D}E(\Psi|M,D)$ and $E_{M|\Lambda,D}\mathrm{Var}(\Psi|M,D)$ are readily computed for a given value of Λ, this MCMC sample yields Monte Carlo estimates of the terms (9) and (10) without the added complexity of the second algorithm.

In this algorithm we assume that $f(d|m)$ can be computed. In some problems this is a difficult task (see, for instance, DiCiccio et al., 1997, Han and Carlin, 2001), although the challenge may be the same for both the baseline and extended priors within models. In the examples of Sections 6 and 7, however, we use the approximation

$f(y|m) \approx f_0(y|m)$, as there are closed-form expressions for the marginal density of the data under the baseline prior but not under the extended prior. We expect this approximation to be quite reasonable, as a slight "flattening" of the prior within each model will not have a great influence on the relative a posteriori weighting of the models.

We also note that Algorithm 3 adapts trivially to compute the requisite across-space quantities if needed.

5. Example: case-control analysis with imprecise exposure assessment

In this section we apply our techniques in a problem investigated by Gustafson et al. (2001). Even though it involves only a single model, our approach to uncertainty analysis may be particularly germane in this example as the estimand is a function of both identifiable and nonidentifiable parameters. Consequently the extent to which inferences will depend on the prior is not clear.

Consider case-control analysis with a single binary exposure variable. Let r_0 and r_1 be the prevalences of this exposure amongst the control and case populations, and say that the log odds-ratio $\psi = \log[\{r_1/(1-r_1)\}/\{r_0/(1-r_0)\}]$ is the parameter of interest. To allow for possible nondifferential misclassification in the exposure assessment, define the sensitivity p to be the probability of correct classification for a subject who is actually exposed, and the sensitivity q to be the probability of correct classification for a subject who is actually unexposed. Thus $\theta = (r_0, r_1, p, q)$ comprises the parameter vector in this problem. Gustafson et al. (2001) discuss Bayesian inference about ψ in this setting at length.

In realistic applications the classification probabilities p and q may be known roughly but not exactly. For instance, consider the following scenario. The investigator believes the exposure assessment scheme is very good, but is not entirely convinced that it is perfect. Consequently he assigns independent Beta$(29, 1)$ distributions to p and q as a baseline prior. This distribution has a monotone increasing density with finite mode at 1, corresponding to the investigator's best guess that the classification scheme is perfect. However, the distribution assigns roughly 0.95 probability to the interval (0.9,1), and so the possibility of slight misclassification is not ruled out. On the other hand, little information about the prevalences r_0 and r_1 is available, so these parameters are assigned uniform priors.

In the context of this scenario, consider two fictitious data sets. In Scenario A, 10/200 controls and 20/200 are classified as exposed. In Scenario B 250/5000 controls and 500/5000 cases are classified as exposed. Thus Scenario B involves the same proportions of apparent exposure as Scenario A, but with 25 times more data.

We consider uncertainty about the priors on p and q by applying the probability integral transform and taking $\mu_1 = F_0(p)$ and $\mu_2 = F_0(q)$ in the formation of an extended prior. The MCMC analysis for a baseline prior suggested by Gustafson et al. (2001) is easy adapted to the present context via Algorithms 1 and 2 as described in Section 4. A posterior sample size of $t = 5000$ after 1000 burn-in iterations suffices to give stable Monte Carlo calculation of the point estimate $\hat{\Psi} = E(\Psi|D)$ and the corresponding partitioned SE, as reported in Table 1.

Table 1
Point estimates and partitioned standard errors in Example 1

	$\hat{\Psi}$	SE[tot]	SE[par]	SE_*[par]
A	1.11	0.81	0.81	0.06
B	1.21	0.65	0.64	0.10

It is seen that SE[tot] and SE[par] decrease only slightly in Scenario B relative to Scenario A, whereas in regular models we would expect to see about a fivefold decrease in SE due to a 25-fold increase in the amount of data. Moreover, SE_*[par] is actually *larger*, and appreciable relative to SE[par], at the larger sample size. In contrast, with regular models we expect the prior to become less important as the sample size grows. Both of these quantitative findings agree with the qualitative findings of Gustafson et al. (2001) concerning the effect of nonidentifiability in this problem. These authors, however, did not have tools with which to quantify the relative role of the prior as in Table 1. Given that the nonidentifiability in this problem is unavoidable, the partition of SE[tot] gives valuable insight into the relative import of the data and prior.

6. Example: linear regression with uncertainty about predictors and interactions

Ein-Dor and Feldmesser (1987) provide data on the characteristics and benchmark performance of $n = 209$ central processing units (CPUs). Following these authors, we consider four predictors of performance: A, the machine cycle time (in nanoseconds), B, the average main memory size (in kilobytes), C, the cache memory size (in kilobytes), and D, the average number of input channels. To reduce skewness, square-root transformations are applied to the response and the four predictors. Then the response and the four predictors are linearly rescaled to have mean zero and variance one, to aid in model interpretation.

We consider the $k = 2^{10} = 1024$ models obtained by either including or excluding each of the four predictors and each of the six possible pairwise interaction terms. For a particular model m involving p_m main effect and interaction terms, our baseline assumptions are

$$Y|\beta_m, \sigma^2 \sim \mathrm{N}(X_m\beta, \sigma^2 I),$$

$$\beta_m|\sigma^2 \sim \mathrm{N}(0, \sigma^2 I),$$

$$p(\sigma^2) \propto \sigma^{-2},$$

where X_m is the $n \times (p_m + 1)$ design matrix for model m. The choice of prior for β_m can be viewed as a 'unit-information' prior as it leads to $(X_m'X_m + I)^{-1}X_m'y$ as the model-specific posterior mean for β_m, where the predictor standardization implies that each diagonal element of $X_m'X_m$ is $n - 1$. The baseline prior for m is taken to be uniform over the k different models. Note that standard linear model analysis yields a closed-form expression for $f_0(d|m)$, the marginal density of the data for a given model.

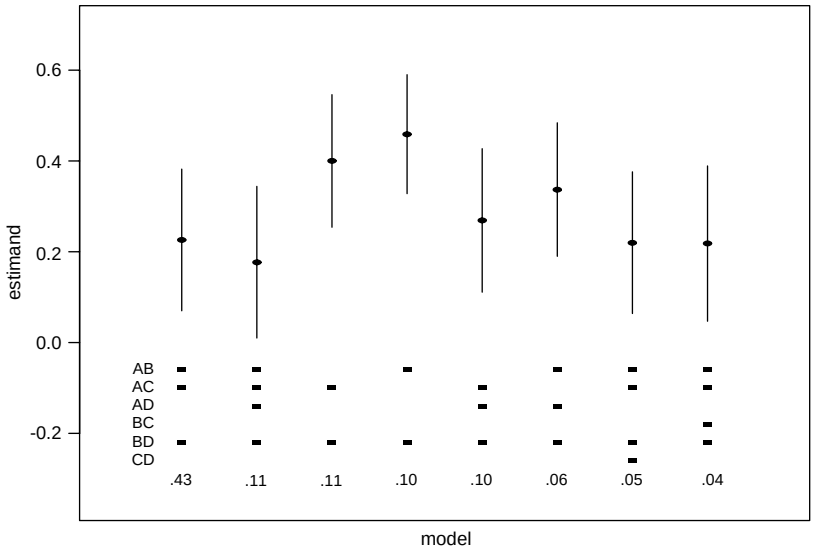

Fig. 2. The eight models with appreciable posterior probability in the regression example of Section 6. For each model $\hat{\psi}$ plus/minus one SE[par] is indicated. Each model contains all the main effects, and is therefore identified by which of the interaction terms (AB, AC, AD, BC, BD, CD) are present. The posterior probability of each model is given at the bottom of the figure.

To apply our analysis we use the default settings of Section 3, with $\mu_i = \Phi(\beta_i/\sigma)$ for $i = 1, \ldots, (p_m + 1)$ under model m. Thus we are considering uncertainty about the prior for the regression coefficients but ignoring uncertainty about the improper prior for the variance term σ^2. The estimand is taken as $\psi = \beta_1 + \cdots + \beta_{p_m+1}$, which can be interpreted as the estimated performance, in terms of SD above or below average, of a CPU which is one SD above average in each characteristic.

Following the idea of Occam's window as advocated by Hoeting et al. (1999), we restrict attention to the eight models having baseline posterior probability of at least $\exp(-3) \approx 1/20$ times that of the highest posterior probability model. After re-normalizing, the posterior model probabilities range from 0.43 down to 0.04. Using Algorithms 1 and 2, each of these models is subjected to the extended prior analysis using Monte Carlo sample sizes of $t = 5000$ after 1000 burn-in iterations. Again, this was found to be sufficient to yield stable values for the model-specific estimates and partitioned standard errors. Fig. 2 gives the posterior mean $\hat{\psi}$ and SE[par] under each of the eight models. Note that $\hat{\psi}$ does exhibit considerable variation across the competing models. The SE_*[par] terms are small in relation to SE[par], with the ratio of the former to the latter ranging from 0.068 to 0.083 across the eight models. Given that the decomposition is additive on the squared scale, this implies that SE[tot] $\approx$ SE[par] for each model. Next we implement the across-model analysis, again using a Monte Carlo sample size of $t = 5000$. Aggregating the model-specific SE[par] and SE_*[par] terms as dictated in Section 2 yields the following overall uncertainty analysis. The point estimate is $E(\Psi|D) = 0.27$, with SE[tot] $= 0.18$. The components of the SE are

SE[par] = 0.15, SE_*[par] = 0.01, SE[mod] = 0.08, and SE_*[mod] = 0.04. Thus while the within-model uncertainty about the parameter is the largest contributor to the overall uncertainty, both SE[mod] and SE_*[mod], which might typically be ignored, are of appreciable size in relation to SE[par].

7. Example: multiple uncertainties in Bayesian curve-fitting

As an example involving a full six-term partitioned SE, consider Bayesian curve-fitting for the data from Silverman (1985) on acceleration versus time in a simulated motorcycle crash. The data, collected at $n = 133$ irregularly-spaced timepoints, appear in Fig. 3. We focus on estimating three quantities: the minimum acceleration, the maximum acceleration, and the acceleration at the last timepoint.

We follow the Bayesian approach to smoothing advocated by Smith and Kohn (1996). For a fixed set of knots $z_1, \ldots, z_m$, consider the $d + m + 1$ functions $1, x, \ldots, x^d$, $(x - z_1)_+^d, \ldots, (x - z_m)_+^d$. These constitute a basis for the regression splines of order d on the knots. Smith and Kohn suggest the use of variable selection methods, so that some of the basis functions (i.e. some of the knots, given the choice of basis) may be removed on the grounds that they are not contributing sufficiently to the fit. Thus for a given choice of d we have a model space containing 2^{d+m+1} models.

In the present example we consider both quadratic (d=2) and cubic (d=3) regression splines, and thus have two competing model spaces. For illustrative purposes we use

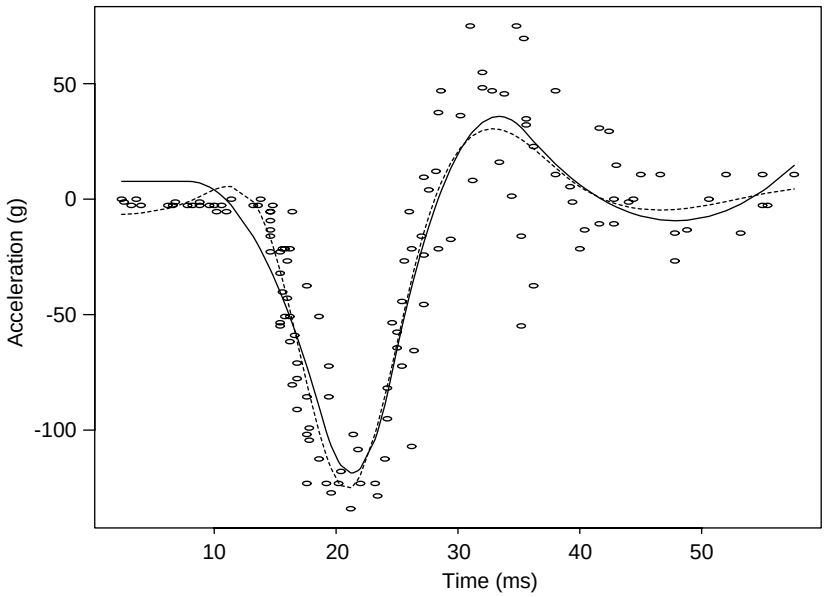

Fig. 3. The motorcycle data of Example 3, with the best quadratic-model fit (solid line) and the best cubic-model fit (dashed line) superimposed.

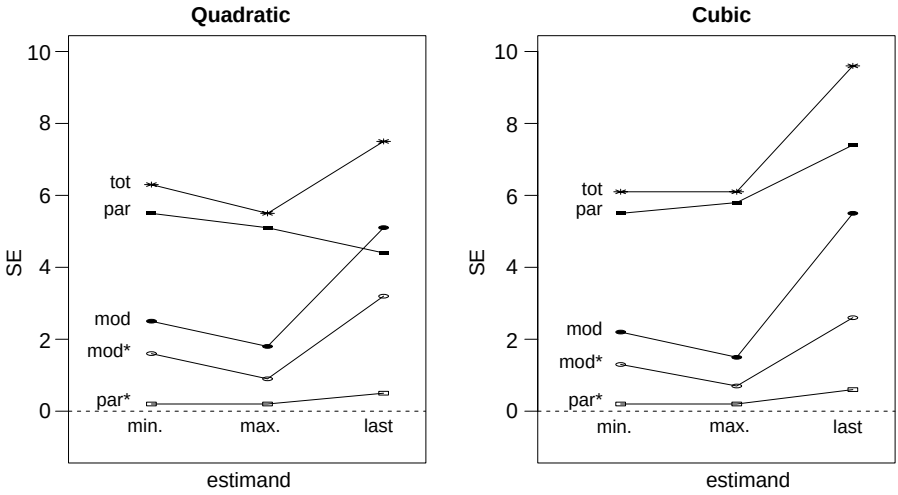

Fig. 4. Partitioned standard errors for the two model spaces considered in the curve-fitting example of Section 7. The left panel gives SE[tot], SE[par], SE_*[par], SE[mod] and SE_*[mod] for the three estimands under the quadratic spline model space. The right panel gives the same quantities for the cubic spline model space.

$m = 9$ equally spaced knots when $d = 2$ and $m = 8$ equally-spaced knots when $d = 3$, in order to have a manageably sized space of 2^{12} models in each case.

For a given model m we use the baseline specification

$$Y|\beta_m, \sigma^2 \sim \mathrm{N}(X_m\beta_m, \sigma^2 I),$$

$$\beta_m|\sigma^2 \sim \mathrm{N}(0, \sigma^2 T_m),$$

$$p(\sigma^2) \propto \sigma^{-2},$$

using $T_m = n(X_m'X_m)^{-1}$ which corresponds to a "unit-information" version of Zellner's g-prior (Zellner, 1986, Fernandez et al., 2001). Fits from the highest posterior probability model in each space are superimposed on Fig. 3.

Again to keep the analysis manageable we restrict to models with baseline posterior probability no less than $\exp(-3) \approx 1/20$ times that of the highest posterior probability model. This yields 50 models, 23 of which are from the quadratic model space and 27 of which are from the cubic model space.

To perform the within-model uncertainty analysis we obtain the components of μ by applying the standard normal distribution function to the components of $\sigma^{-1}T_m^{-1/2}\beta_m$, and then use the default settings from Section 3. Applying this to the models under consideration via Algorithms 1 and 2 and then averaging over models within each space via Algorithm 3 leads to posterior means of $\hat{\psi}_{\min} = -122.3$, $\hat{\psi}_{\max} = 36.7$, $\hat{\psi}_{\text{last}} = 10.9$ for the quadratic spline model space, and $\hat{\psi}_{\min} = -123.4$, $\hat{\psi}_{\max} = 31.0$, $\hat{\psi}_{\text{last}} = 7.0$ for the cubic spline model space. For each space SE[tot] and the four constituent terms are displayed in Fig. 4. Aggregating across spaces, again with the help of Algorithm 3,

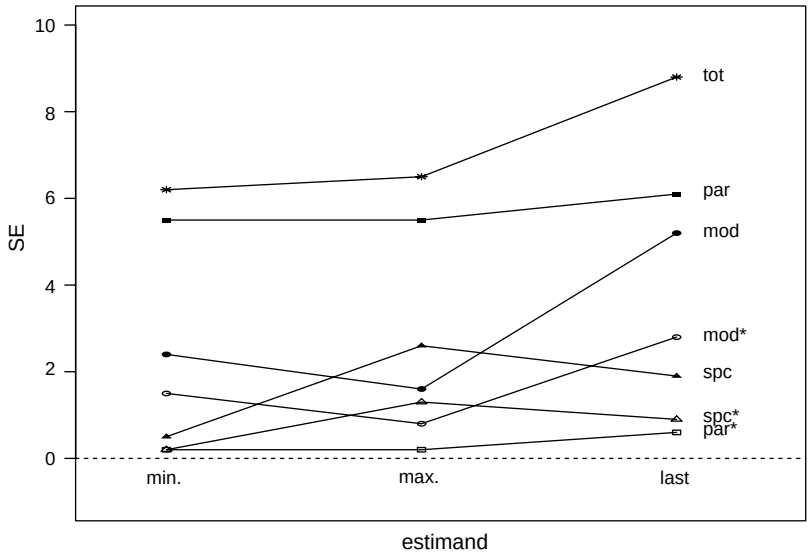

Fig. 5. Full partitioned standard errors for the curve-fitting example of Section 7. Each of SE[tot], SE[par], SE_*[par], SE[mod], SE_*[mod], SE[spc] and SE_*[spc] is given for each estimand.

yields posterior means $\hat{\psi}_{\min} = -122.9$, $\hat{\psi}_{\max} = 33.6$, $\hat{\psi}_{\text{last}} = 8.8$. The corresponding SE terms are given in Fig. 5. Perhaps the overarching comment is that while SE[par] is the largest component in the partitioned SE for each estimand, it does not dwarf the other components which might typically be ignored. For instance, in estimating the acceleration at the last timepoint, uncertainty about the model within the space (i.e. which knots should be used) is almost as large as uncertainty about the parameter within the model (i.e. what coefficients should be used). And in estimating the maximum acceleration, SE[spc] is quite large in relation to SE[par], indicating substantial uncertainty about whether quadratic or cubic splines are most appropriate. Some of the uncertainties about priors are also non-negligible, with SE_*[mod] when estimating the acceleration at the last timepoint and SE_*[spc] when estimating the maximum acceleration being cases in point.

8. Discussion

Throughout the examples of the previous three section, each SE[x] term is larger than its corresponding SE_*[x] term. Of course this is to be expected; the uncertainty that derives from not knowing x is larger than the uncertainty that derives from not knowing how to weight x a priori. Moreover, in Example 3 which involves the full six-term decomposition, SE[par] > SE[mod] > SE[spc]. We suggest that this ordering would be typical, as the progression from parameter to model to model space moves us away from the "heart" of the inferential process, that is, the exact relationship

between possible outcomes and parameter values. Nonetheless, the examples do caution us against ignoring model or model space uncertainty.

We view simplicity, both computational and conceptual, as a strength of our approach. As discussed in Section 4, the computational burden over and above that of the baseline analysis is quite modest. On the other hand, the conceptual simplicity of our decomposition (1) derives from the existence of a joint distribution on all relevant quantities, with respect to which uncertainties can be assessed. We could not use conditional expectation and variance to our advantage without this underlying structure. In this regard, the Bayes paradigm and the desire to decompose or partition uncertainty go hand in hand.

A point which deserves elaboration is the remark in Section 2.3 that an estimand must have a common interpretation across models, or more generally across models and spaces of models. One could apply the decomposition to estimands without such an interpretation, but this would be quite meaningless. As an example, say that the different models correspond to different transformations of the response variable in a regresssion context. For instance, different powers in the Box-Cox transformation could be considered. One could apply the decomposition with the estimand taken to be the regression coefficient for one of the regressors. However, estimates of this coefficient would tend to vary greatly across models, simply because the coefficient has a different interpretation under each model. Thus a high value of SE[mod] would be almost guaranteed. In the examples of Sections 6 and 7 the estimands are fitted values at particular values of the X variables. Since a fitted value has a common interpretation across models as a predicted Y value for the given X values, the decomposition is readily interpreted.

Of course the goal of accounting for various sources of uncertainty as completely as possible when making inferential statements is not unique to Bayesian analysis. It is well known in general that standard errors and confidence intervals determined conditionally on a model that has been chosen from a model space on empirical grounds are too small. Surprisingly, however, the literature on solutions to this problem is somewhat sparse (see, for instance, Aitkin, 1974; Buckland et al., 1997; Regal and Hook, 1991; Zhang, 1992).

In fact, to some extent we may be able to ascribe non-Bayesian interpretations and analogs to some of the terms in our decomposition. In particular, SE[mod] is essentially a weighted sum of squared deviations between model-specific estimates and a model-averaged estimate, and hence variants of it using weights derived in non-Bayesian ways are readily obtained. For instance, Buckland et al. (1997) use AIC values to obtain model weights which differ from Bayes weights as used here. While this may not have the conceptual appeal of the Bayes approach, it would typically involve easier computations. Moreover, we can make a rough analogy between changing the prior distribution over competing models within the Bayes paradigm and changing the model selection principle (MSP) or paradigm used to choose between models. That is to say a non-Bayes approach to weighting the models can be "back-transformed" to a prior distribution (albeit data-dependent) over the models. Thus SE_*[mod] can be viewed as a rough surrogate for capturing uncertainty due to choice of MSP. There is a considerable literature to suggest that depending on one's statistical goals and

settings different MSPs are optimal, and combining several model selection principle can outperform any fixed choice, much as averaging models may outperform usage of a fixed model.

One interesting extension to our decomposition would be an examination of the uncertainty due to imperfections in the data. For instance, many analyses of data must deal with censored responses, missing covariates, or imprecisely measured covariates. While most reported standard errors incorporate the extra uncertainty caused by these imperfections, it is not always easy to understand how much of the uncertainty is attributable to these sources. In principle a decomposition of posterior variance in the spirit of (1) could determine how much of the posterior variability would remain if the data were measured perfectly. Since the typical route to Bayes analysis in such problems involves MCMC on an extended parameter space which includes the true but unobserved measurements, such a development may indeed be practical.

References

Aitkin, M., 1974. Simultaneous inference and the choice of variable subsets in multiple regression. Technometrics 16, 221–227.

Buckland, S.T., Burnham, K.P., Augustin, N.H., 1997. Model selection: an integral part of inference. Biometrics 53, 603–618.

Clyde, M.A., 1999. Bayesian model averaging and model search strategies. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 6, Oxford Univ. Press, pp. 157–185.

DiCiccio, T.J., Kass, R.E., Raftery, A., Wasserman, L., 1997. Computing Bayes factors by combining simulation and asymptotic approximations. J. Amer. Statist. Assoc. 92, 903–915.

Draper, D., 1995. Assessment and propagation of model uncertainty (with discussion). J. Roy. Statist. Soc. Ser. B 57, 45–97.

Ein-Dor, P., Feldmesser, J., 1987. Attributes of performance of central processing units: a relative performance prediction model. Comm. Assoc. Comput. Mach. 30, 308–317.

Fernandez, C., Ley, E., Steel, M.F.J., 2001. Benchmark priors for Bayesian model averaging. J. Econometrics 100, 381–427.

Gustafson, P., Le, N.D., Saskin, R., 2001. Case-control analysis with partial knowledge of exposure misclassification probabilities. Biometrics 57, 598–609.

Han, C., Carlin, B.P., 2001. MCMC methods for computing Bayes factors: a comparative review. J. Amer. Statist. Assoc. 96, 1122–1132.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Statist. Sci. 14, 382–417. Corrected version available at http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90, 773–795.

Regal, R.R., Hook, E.B., 1991. The effects of model selection on confidence intervals for the size of a closed population. Statist. Med. 10, 717–721.

Rios Insua, D., Ruggeri, F., 2000. Robust Bayesian Analysis. Lecture Notes in Statistics, Vol. 152. New York, Springer, Berlin.

Silverman, B.W., 1985. Some aspects of the spline-smoothing approach to non-parametric regression curve fitting (with discussion). J. Roy. Statist. Soc. Ser. B 47, 1–52.

Smith, M., Kohn, R., 1996. A Bayesian approach to nonparametric bivariate regression. J. Econometrics 75, 317–343.

Zellner, A., 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Goel, P.K., Zellner, A. (Eds.), Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti. North-Holland, Amsterdam, pp. 233–243.

Zhang, P., 1992. Inference after variable selection in linear regression models. Biometrika 79, 741–746.