Statistica Neerlandica (2005) Vol. 59, nr. 1, pp. 57–69

Bayesian model selection using encompassing priors

Irene Klugkist*, Bernet Kato and Herbert Hoijtink

Department of Methodology and Statistics, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands

This paper deals with Bayesian selection of models that can be specified using inequality constraints among the model parameters. The concept of encompassing priors is introduced, that is, a prior distribution for an unconstrained model from which the prior distributions of the constrained models can be derived. It is shown that the Bayes factor for the encompassing and a constrained model has a very nice interpretation: it is the ratio of the proportion of the prior and posterior distribution of the encompassing model in agreement with the constrained model. It is also shown that, for a specific class of models, selection based on encompassing priors will render a virtually objective selection procedure. The paper concludes with three illustrative examples: an analysis of variance with ordered means; a contingency table analysis with ordered odds-ratios; and a multilevel model with ordered slopes.

Key Words and Phrases: Bayes factors, inequality constraints, objective Bayesian inference, posterior probability.

1 Inequality constrained statistical models

Researchers often have one or more (competing) theories about their field of research. Consider, for example, theories about the effect of behavioral therapy versus medication for children with an attention deficit disorder (ADD). Some researchers in this area believe medication is the only effective treatment for ADD, some believe strongly in behavioral therapy, and others may expect an additive effect of both therapies. To test or compare the plausibility of these theories they need to be translated into statistical models. Subsequently, empirical data can be used to determine which model is best. Inequality constraints on model parameters can be useful in the specification of statistical models.

This paper deals with competing models that have the same parameter vector, but in one or more of the models parameters are subjected to inequality constraints. To continue the example, consider an experiment where children with ADD are randomly assigned to one of four conditions: no treatment (1), behavioral therapy (2), medication (3), and behavioral therapy plus medication (4). Let the outcome

*I.Klugkist@fss.uu.nl

variable of interest be the score on an attention test, and μ_j be the average in group j ($j = 1,\ldots,4$). The three theories presented above can be translated into the following models: $(\mu_3, \mu_4) > (\mu_1, \mu_2)$ (positive effect of medication), $(\mu_2, \mu_4) > (\mu_1, \mu_3)$ (positive effect of behavioral therapy), $\mu_4 > (\mu_2, \mu_3) > \mu_1$ (additive effect of therapy and medication).

The unconstrained model is called the encompassing model and plays a central role in the model selection procedure described in this paper. Note that all the constrained models are nested in the encompassing model. Two goals are distinguished: select the best theory of a set of competing theories (i.e. constrained models) or find out if a theory (i.e. a constrained model) is better than the unconstrained, encompassing model. Illustrations of both situations will be provided in the examples of this paper. The notation used for the model parameters of interest, that is, the vector of parameters subjected to inequality constraints in one or more of the nested models, is $\boldsymbol{\theta}$. Parameters that are unconstrained in all (encompassing and nested) models, i.e. the nuisance parameters, are denoted by $\boldsymbol{\omega}$. Assuming that the ADD data of the example are normally distributed, $\boldsymbol{\theta} = \{\mu_1, \mu_2, \mu_3, \mu_4\}$ and $\boldsymbol{\omega}$ contains the nuisance parameter σ^2.

The model selection procedure is based on the Bayes factor. For data $\boldsymbol{D}$ and models M_q and $M_{q'}$, the Bayes factor is

$$\mathrm{BF}_{q'q} = \frac{P(\boldsymbol{D}|M_{q'})}{P(\boldsymbol{D}|M_q)} = \frac{\int L(\boldsymbol{D}|\boldsymbol{\theta}, \boldsymbol{\omega}, M_{q'}) g(\boldsymbol{\theta}, \boldsymbol{\omega}|M_{q'}) \mathrm{d}\boldsymbol{\theta}, \boldsymbol{\omega}}{\int L(\boldsymbol{D}|\boldsymbol{\theta}, \boldsymbol{\omega}, M_q) g(\boldsymbol{\theta}, \boldsymbol{\omega}|M_q) \mathrm{d}\boldsymbol{\theta}, \boldsymbol{\omega}}, \tag{1}$$

that is, the ratio of the marginal likelihoods of $M_{q'}$ and M_q (see for instance, KASS and RAFTERY (1995)). As can be seen in (1), Bayes factors are sensitive to the prior distribution of the parameters of each model. However, in this paper it will be shown that for sets of models where the constrained models are nested in an unconstrained, encompassing model, only one prior distribution needs to be specified, the so-called *encompassing prior*. The prior distributions for the parameters of the nested models follow from the encompassing prior by restriction of the parameter space according to the constraints. Furthermore, it will be shown that, for specific classes of models, model selection based on encompassing priors is virtually objective. Estimation of the Bayes factor traditionally involves the calculation of marginal likelihoods, which often involves computational problems. However, using the encompassing prior approach leads to a straightforward estimate of the Bayes factor, without requiring the computation of marginal likelihoods.

In Section 2 the new approach to estimating Bayes factors in the context of encompassing priors is introduced. In Section 3 specification of the encompassing prior is outlined, and subsequently the sensitivity of Bayes factors to the encompassing prior is examined. In Sections 4, 5 and 6, three illustrations are provided, dealing with respectively normal linear data with constraints on independent means, a three-way contingency table with ordered odds ratios, and a multilevel analysis with inequality constraints on the slopes. The paper will be concluded with a short discussion in Section 7.

2 Estimation of posterior probabilities using encompassing priors

In the encompassing model no constraints are put on the model parameters and all other models are nested in this model. Furthermore, the prior distribution of the parameters in the encompassing model will be called the encompassing prior. In what follows the encompassing model will be denoted M_1 and the encompassing prior distribution will be denoted $g(\boldsymbol{\theta}, \boldsymbol{\omega}|M_1)$. The prior distribution of any model M_q, $q = 2,\ldots,Q$ that is nested in the encompassing model M_1 can be obtained from $g(\boldsymbol{\theta}, \boldsymbol{\omega}|M_1)$ by restricting the parameter space in accordance with the constraints imposed by a model M_q and is given by

$$g(\boldsymbol{\theta}, \boldsymbol{\omega}|M_q) = \frac{g(\boldsymbol{\theta}, \boldsymbol{\omega}|M_1) I_{M_q}(\boldsymbol{\theta}, \boldsymbol{\omega})}{\int g(\boldsymbol{\theta}, \boldsymbol{\omega}|M_1) I_{M_q}(\boldsymbol{\theta}, \boldsymbol{\omega}) d(\boldsymbol{\theta}, \boldsymbol{\omega})}. \tag{2}$$

In (2), $I_{M_q}(\boldsymbol{\theta}, \boldsymbol{\omega})$ is the indicator function for model M_q, such that $I_{M_q}(\boldsymbol{\theta}, \boldsymbol{\omega})$ equals 1 if the parameter values are in accordance with the constraints imposed by model M_q, and equals 0 otherwise. It should be noted that only the prior distribution of the parameters of the encompassing model has to be specified and that the prior distributions of the other models can be derived using (2).

Let $\{M_1, M_2,\ldots,M_Q\}$ denote the finite set of all models under consideration, that is the set of competing models, and $\boldsymbol{D}$ denote the observed data. In the sequel a method which will be used to compute posterior probabilities is introduced. The method is based on the principle of encompassing priors and it works as follows: consider two models, the encompassing model M_1 and another model M_q (nested in M_1). In general, for any model M_q, the marginal likelihood can be written as

$$P(\boldsymbol{D}|M_q) = \frac{L(\boldsymbol{D}|\boldsymbol{\theta}, \boldsymbol{\omega}) g(\boldsymbol{\theta}, \boldsymbol{\omega}|M_q)}{P(\boldsymbol{\theta}, \boldsymbol{\omega}|\boldsymbol{D}, M_q)},$$

where the numerator is a product of the likelihood function of the data and the prior distribution of $(\boldsymbol{\theta}, \boldsymbol{\omega})$ under model M_q, while the denominator is the posterior density of $(\boldsymbol{\theta}, \boldsymbol{\omega})$ under model M_q (see Chib (1995), Section 2). Consequently the Bayes factor of M_q to M_1 is given by:

$$\mathrm{BF}_{q1} = \frac{P(\boldsymbol{D}|M_q)}{P(\boldsymbol{D}|M_1)} = \frac{L(\boldsymbol{D}|\boldsymbol{\theta}, \boldsymbol{\omega}) g(\boldsymbol{\theta}, \boldsymbol{\omega}|M_q)/P(\boldsymbol{\theta}, \boldsymbol{\omega}|\boldsymbol{D}, M_q)}{L(\boldsymbol{D}|\boldsymbol{\theta}, \boldsymbol{\omega}) g(\boldsymbol{\theta}, \boldsymbol{\omega}|M_1)/P(\boldsymbol{\theta}, \boldsymbol{\omega}|\boldsymbol{D}, M_1)}. \tag{3}$$

Suppose $\boldsymbol{\theta}^*$ is a value of $\boldsymbol{\theta}$ that is allowed in the constrained model. Then substituting $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ in (3) renders

$$\mathrm{BF}_{q1} = \frac{g(\boldsymbol{\theta}^*, \boldsymbol{\omega}|M_q) P(\boldsymbol{\theta}^*, \boldsymbol{\omega}|\boldsymbol{D}, M_1)}{g(\boldsymbol{\theta}^*, \boldsymbol{\omega}|M_1) P(\boldsymbol{\theta}^*, \boldsymbol{\omega}|\boldsymbol{D}, M_q)}.$$

Since M_q is nested in M_1, it follows that the densities $g(\boldsymbol{\theta}^*, \boldsymbol{\omega}|M_q)$ and $P(\boldsymbol{\theta}^*, \boldsymbol{\omega}|\boldsymbol{D}, M_q)$ can be rewritten as $c_q \times g(\boldsymbol{\theta}^*, \boldsymbol{\omega}|M_1)$ and $d_q \times P(\boldsymbol{\theta}^*, \boldsymbol{\omega}|\boldsymbol{D}, M_1)$ respectively, where d_q and c_q are constants. In other words, if model M_q is nested in model M_1 then the prior and posterior densities of M_q can be rewritten in terms of the prior and posterior densities of M_1. Effectively, $1/c_q$ is the proportion of the

prior distribution of M_1 in agreement with M_q and $1/d_q$ is the proportion of the posterior distribution of M_1 in agreement with M_q. This procedure can be applied to any number of competing models as long as an encompassing model within which each of them is nested can be specified. A computational advantage of this method is that a researcher only needs to specify the prior distribution and subsequently the posterior distribution of the encompassing model and then sample from each of them to determine the proportion of parameter vectors (i.e., $1/c_q$ and $1/d_q$ respectively) from each in agreement with any model nested in the encompassing model. Subsequently, $\mathrm{BF}_{q1} = P(M_q|\boldsymbol{D})/P(M_1|\boldsymbol{D}) = c_q/d_q$, for $q = 1,\ldots,Q$.

Assuming that the models are *a priori* equally probable (i.e., $P(M_q) = 1/Q$ for $q = 1,\ldots,Q$), posterior model probabilities can be derived from Bayes factors using

$$P(M_q|\boldsymbol{D}) = \frac{\mathrm{BF}_{q1}}{\mathrm{BF}_{11} + \mathrm{BF}_{21} + \cdots + \mathrm{BF}_{Q1}}, \text{ for } q = 1, \ldots, Q, \tag{4}$$

where $\mathrm{BF}_{11} = 1$.

3 Sensitivity of posterior probabilities

As explained in the previous section, only $g(\boldsymbol{\theta}, \boldsymbol{\omega}|M_1)$, that is, the prior distribution of the parameters of the encompassing model, has to be specified. For notational convenience, in the sequel the notation $g(\boldsymbol{\theta}, \boldsymbol{\omega})$ is used. Specification of the encompassing prior is based on the following four principles.

1. The encompassing prior should not favour the unconstrained or any of the constrained models. This is achieved using similar and independent prior distributions for each element of $\boldsymbol{\theta}$, i.e.

   $$g(\boldsymbol{\theta}, \boldsymbol{\omega}) = g(\theta) \ldots g(\theta) g(\boldsymbol{\omega}). \tag{5}$$

   Stated otherwise, $g(\theta_k) = g(\theta)$.
2. The prior element $g(\theta)$ is continuous on the real line or a subsection of the real line.
3. The prior element $g(\theta)$ should be vague, that is, it should not exclude regions of the parameter space with substantial posterior probability. For example, if a variable is measured on a scale from 1 to 10, the prior distribution for the mean of this variable could be a uniform distribution on the interval 1–10. If θ is a probability, the interval is 'naturally' bounded by 0 and 1. If natural bounds are not available, priors can also be data-based. For example, if at least 99% of the posterior distribution of each element of $\boldsymbol{\theta}$ is within the interval l–u, $g(\theta)$ could have a mean and variance computed such that it corresponds to a normal distribution with 0.05th and 0.95th percentile l and u.

4. The encompassing prior should be vague with respect to the nuisance parameters $\boldsymbol{\omega}$. This will be achieved using scaled inverse chi-square and inverse Wishart distributions with one degree of freedom and data based scale factors, that is, sample variance and sample covariance matrix, respectively.

The sensitivity of posterior probabilities with respect to the actual specification of the encompassing prior is an issue that deserves further attention. For a large class of models posterior probabilities are virtually objective, that is, independent of the actual specification of a vague encompassing prior. This will be illustrated using $1/c_q$ and $1/d_q$, the proportion of the encompassing prior and posterior in agreement with a constrained model, respectively. The sensitivity with respect to $g(\boldsymbol{\omega})$ and $g(\boldsymbol{\theta})$ will be discussed separately.

From (5) it is immediately clear that $g(\boldsymbol{\omega})$ does not influence $1/c_q$. It is clear that $g(\boldsymbol{\omega})$ does influence the posterior distribution, and thus $1/d_q$. However, it is well-known (see, for example, GELMAN, CARLIN, STERN and RUBIN (2000), p.101) that this influence is small if the sample size is large compared with the degrees of freedom of the scaled inverse chi-square and inverse Wishart distributions.

Similarly, if $g(\theta)$ is vague, its influence on the posterior distribution is negligible, that is, its influence on $1/d_q$ is negligible. For a specific class of models the choice of $g(\theta)$ does not influence $1/c_q$. Let $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}\}$. The class uses one or more constraints of the form

$$\sum_{p=1}^{P} \theta_p^{(1)} > \sum_{s=1}^{S} \theta_s^{(2)}, \tag{6}$$

with the restriction $P = S$. For encompassing priors it holds that $\theta_p^{(1)}, \theta_s^{(2)} \overset{i.i.d.}{\sim} g(\theta)$. The consequence is that $P(\sum_{p=1}^{P} \theta_p^{(1)} > \sum_{s=1}^{S} \theta_s^{(2)})$ is 0.50, and, consequently, independent of the actual choice of $g(\theta)$. The same holds for combinations of constraints of the form (6). For example, $P(\theta_1 > \theta_2 > \theta_3) = 1/3! = 1/6$, independent of the choice of $g(\theta)$. Similar considerations holds for models using one or more constraints of the form

$$\prod_{p=1}^{P} \theta_p^{(1)} > \prod_{s=1}^{S} \theta_s^{(2)}, \tag{7}$$

with the restriction $P = S$.

Besides (6) and (7), there are other constraints for which model selection based on encompassing priors is virtually objective. However, these constraints are not used in the examples and will not be discussed in this paper. There are also constraints for which the model selection is strongly affected by the encompassing prior. An example is a model with approximate equality constraints between model parameters. The evidence in favour of the constrained model increases with the amount of vagueness of the encompassing prior, a phenomenon known as Bartlett's or Lindley's paradox (e.g. LINDLEY (1957); BERNARDO and SMITH (1994)).

4 One-way analysis of variance with ordered means

The data for this illustration deal with sales numbers of breakfast cereals. A food company wants to test four different package designs: design 1 is a short, thick box with a cartoon on it; design 2 is a tall, slim box with a cartoon on it; design 3 is a short, thick box with a photo of an athlete; and design 4 is a tall, slim box with a photo of an athlete. Each store included in the experiment was randomly assigned one of the package designs. The stores were chosen to be comparable in location and sales volume. Other relevant conditions that could affect sales, such as price, amount and location of shelf space, and special promotion efforts were kept the same for all stores. Sales, in number of boxes sold, were recorded and are presented in Table 1 (NETER, KUTNER, NACHTSHEIM and WASSERMAN, 1996).

The data $\boldsymbol{D} = \{\boldsymbol{y}, \boldsymbol{d}\}$ are assumed to be normally distributed:

$$y_i = \sum_{j=1}^{4} \mu_j d_{ji} + \varepsilon_i, \text{ with } \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

where y_i denotes the sales for the ith shop ($i = 1,\ldots,19$), and, $d_{ji} = 1$ if the ith shop has package design j, and zero otherwise. Consequently, the regression coefficient μ_j represents the mean sales for the jth package design.

The semi-conjugate encompassing prior distribution with $\boldsymbol{\theta} = \{\mu_1, \mu_2, \mu_3, \mu_4\}$ and $\boldsymbol{\omega} = \{\sigma^2\}$, is

$$g(\boldsymbol{\theta}, \boldsymbol{\omega}) = \text{Inv-}\chi^2(\sigma^2|v, \varphi^2) \times \prod_{j=1}^{4} N(\mu_j|\eta, \tau^2). \tag{8}$$

In (8), $N(\mu_j|\eta, \tau^2)$ denotes a Normal distribution with mean η and variance τ^2 and Inv-$\chi^2(\sigma^2|v, \varphi^2)$ denotes a scaled inverse Chi-square distribution with degrees of freedom v and scale φ.

Different prior specifications, i.e. values for η, τ^2 and φ^2, will be used in the analysis of this example. The first encompassing prior is data-based, using the method as described in the example of principle 3 (for μ_j) and principle 4 (for σ^2) in Section 3. This leads to the values $\eta = 20.1$, $\tau^2 = 11.8$, $v = 1$ and $\varphi^2 = 12.6$. Three other encompassing prior distributions are specified by varying the values for η, τ^2 and φ^2 (see Table 2). These values are chosen such that, going from the first to the fourth encompassing prior, the distributions become more and more diffuse (increasing τ^2). In addition, η and φ^2 are varied rather randomly.

Table 1. Breakfast cereal sales data.

	Design 1	Design 2	Design 3	Design 4
Sales (No. of boxes):	11	12	23	27
	17	10	20	33
	16	15	18	22
	14	19	17	26
	15	11		28

Table 2. Prior sensitivity of posterior model probabilities for sales data.

Prior specification		Theory		
$g(\theta)$	$g(\omega)$	1	2	3
$N(20.1; 11.8)$	Inv-$\chi^2(1; 12.6)$	0.9718	0.0282	0.0000
$N(20; 25)$	Inv-$\chi^2(1; 25)$	0.9775	0.0225	0.0000
$N(0; 500)$	Inv-$\chi^2(1; 25)$	0.9823	0.0177	0.0000
$N(100; 1000)$	Inv-$\chi^2(1; 50)$	0.9772	0.0228	0.0000

Three competing theories about the effect of the package design on the sales numbers exist. The first theory states that packages with a photo of an athlete sell better than packages with a cartoon, and, to a lesser extent, that slim boxes sell better than thick boxes, that is $\mu_1 < \mu_2 < \mu_3 < \mu_4$. Theory 2 states that slim boxes sell better than thick boxes, and, to a lesser extent, that the athlete photo sells better than the cartoon, i.e. $\mu_1 < \mu_3 < \mu_2 < \mu_4$. The last theory states that the cartoon sells better than the athlete, and, to a lesser extent, that slim boxes sell better than thick boxes, that is $\mu_3 < \mu_4 < \mu_1 < \mu_2$. Note that these constraints are of the form (6). Also note that the encompassing model is *not* included in the set of models of interest.

Posterior probabilities are computed using samples of 100,000 parameter vectors from each encompassing prior and the corresponding posterior distribution. For each encompassing prior and each of the three theories, the quantities $1/c_q$ and $1/d_q$ are estimated using the proportion of the 100,000 parameter vectors from prior and posterior in agreement with the constraints specified by the model at hand. The resulting posterior probabilities can be found in Table 2. As can be seen, the posterior model probabilities are virtually independent of the specification of the encompassing prior. Stated otherwise, the influence on $1/d_q$ is indeed negligible. Theory 1 has the largest posterior probability (97–98%) and therefore it can be concluded that the first theory about the effect of the package designs on sales numbers has the strongest support from the data.

5 Contingency table analysis with ordered odds ratios

The data (obtained from AGRESTI (2002), p. 322) to be analysed in this section are presented in Table 3. It is a three-way contingency table containing counts of high school seniors using (combinations of) alcohol (a), cigarettes (c) and marijuana (m).

Table 3. Alcohol, cigarette and marijuana use for high school seniors.

		Marijuana use	
Alcohol use	Cigarette use	Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

Note, that this model has no nuisance parameters $\boldsymbol{\omega}$, that $\boldsymbol{\theta}$ is a vector of probabilities and $\boldsymbol{D} = \boldsymbol{x}$.

Analysis of these data is based on a multinomial likelihood

$$L(\boldsymbol{x} \mid \boldsymbol{\theta}) \propto \prod_{a=0}^{1} \prod_{c=0}^{1} \prod_{m=0}^{1} \theta_{acm}^{x_{acm}},$$

where 0 denotes the response 'No' and 1 the response 'Yes', and a conjugate Dirichlet encompassing prior distribution

$$g(\boldsymbol{\theta}) \propto \prod_{a=0}^{1} \prod_{c=0}^{1} \prod_{m=0}^{1} \theta_{acm}^{x_0-1},$$

where x_0 denotes the prior sample size in each cell of the contingency table. In the sequel analyses will be presented for x_0 equal to 1, 0.5 and 0.001. If a contingency table contains only two cells, these correspond to uniform uninformative prior distributions for θ, $\sin^{-1}(\sqrt{\theta})$ and logit(θ), respectively (see, for example, GELMAN, CARLIN, STERN and RUBIN (2000), pp. 55–56; LEE (1997), pp. 83–85). For eight cells (like in the example at hand) the marginal distribution of each θ is Beta(1, 7), Beta(0.5, 3.5) and Beta(0.001, 0.007), respectively. In Figure 1 these marginal distributions are displayed. As can be seen, the prior information strongly depends on the choice of x_0: the smaller x_0, the larger the prior density of very small and very large values of θ. Note that Figure 1 does *not* contain a line for the vertical and horizontal axis. The virtually horizontal and vertical lines displayed in Figure 1 constitute the prior distribution for $x_0 = 0.001$.

AGRESTI and COULL (2002) present an overview of methods for the analysis of contingency tables under inequality constraints. Apparently, there are no classical

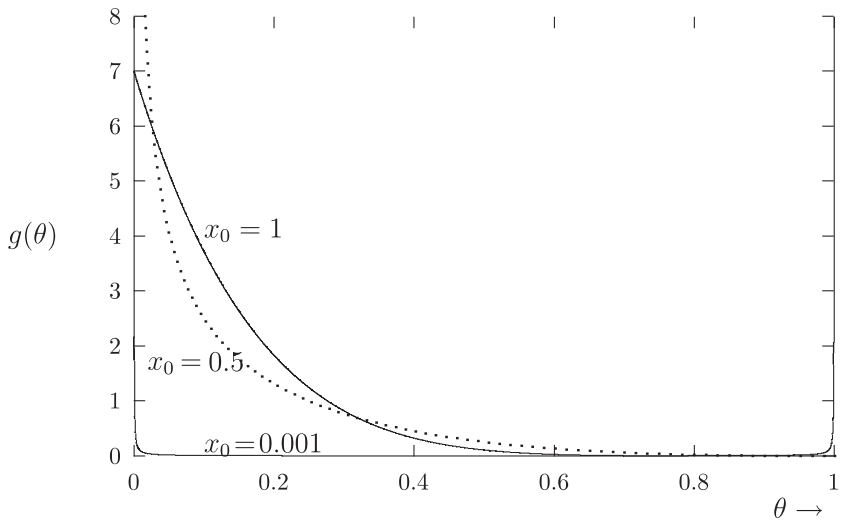

Fig. 1. Marginal prior density for three uninformative encompassing priors.

alternatives for the Bayesian analysis executed in the sequel. Posterior probabilities will be computed for five different models. Let γ_{ac} denote the marginal odds-ratio for alcohol and cigarettes, that is,

$$\frac{(\theta_{111}+\theta_{110})(\theta_{001}+\theta_{000})}{(\theta_{101}+\theta_{100})(\theta_{011}+\theta_{010})},$$

and let γ_{am} and γ_{cm} have corresponding definitions. Odds-ratios larger than 1 indicate that both substances involved are used together. Odds-ratios smaller than 1 indicate that both substances are not used together.

The first model is the unconstrained model. The second model reflects the theory that cigarettes and marijuana are often used together, and that cigarette use and alcohol consumption are often combined, that is, $\gamma_{cm} > 1$ and $\gamma_{ac} > 1$. Note that these constraints can be rewritten in the form (7). The third model is an extension of the second model: $\gamma_{cm} > 1$, $\gamma_{ac} > 1$ and $\gamma_{am} > 1$. The fourth model adds to the third model the theory that the association between alcohol and marijuana use is weaker than the other two associations, that is, $\gamma_{am} < \gamma_{cm}$ and $\gamma_{am} < \gamma_{ac}$. Note that these constraints can also be rewritten in the form (7). The fifth model adds to the third model the theory that the association between alcohol and marijuana use is stronger than the other two associations, that is, $\gamma_{am} > \gamma_{cm}$ and $\gamma_{am} > \gamma_{ac}$.

In Table 4 the posterior probabilities resulting from the use of different encompassing priors are displayed. The posterior probabilities are computed using a sample of 100,000 parameter vectors $\boldsymbol{\theta}$ from each encompassing prior and the corresponding unconstrained posterior distribution. For each encompassing prior the quantities $1/c_q$ and $1/d_q$ are estimated for Models 2–5 using the proportion of the 100,000 parameter vectors from prior and posterior in agreement with the constraints specified by the model at hand. Subsequently (4) is used to obtain the posterior probabilities.

As can be seen in Table 4, the posterior probabilities are almost independent of the choice of the encompassing prior. This provides some support for our claim that $1/d_q$ is almost independent of the encompassing prior. The encompassing priors chosen are quite different and even probabilities associated with very small cell counts in Table 3 appear to be unaffected. It can furthermore be concluded that model five has the largest posterior probability. Stated otherwise, it can be concluded that the marginal association between alcohol and marijuana use is stronger than the other two marginal associations, and that all marginal associations are positive.

Table 4. Prior sensitivity of posterior probabilities for substance use.

	Model				
x_0	1	2	3	4	5
0.001	0.0364	0.1459	0.1867	0.0001	0.6300
0.5	0.0347	0.1385	0.2086	0.0004	0.6178
1	0.0300	0.1363	0.2123	0.0009	0.6165

6 Multilevel model with ordered slopes

The data to be used in this section have been introduced and analysed by VERBEKE and LESAFFRE (1999). The data consist of craniofacial growth measurements of 50 male Wistar rats. The rats were randomized to either a control group or one of the two treatment groups where the treatment consisted of a low or high dose of the drug Decapeptyl. This drug is an inhibitor for testosterone production in rats. The primary aim of the experiment was to investigate the effect of the inhibition of the production of testosterone on the craniofacial growth in male Wistar rats. The responses of interest are distances (in pixels) between well-defined points on X-ray pictures of the skull of each rat, taken after the rat had been anaesthetized.

In the same spirit as VERBEKE and LESAFFRE (1999), in this paper we will consider one of the measurements that can be used to characterize the height of the skull. The treatment started at the age of 45 days, and measurements taken every 10 days until the age of 110 days, with the first measurement taken at the age of 50 days. This would give seven measurements for each rat. For each treatment group the individual profiles are shown in Figure 2. As can be seen in the figure, not all rats have up to seven measurements. This is because some rats did not survive the anaesthesia and therefore dropped out before the end of the study. In their analyses, VERBEKE and LESAFFRE (1999) use these data to investigate the effect of drop-out on the efficiency of longitudinal experiments.

To analyse these data the following model proposed by VERBEKE and MOLENBERGHS (2000, Chapter 3) will be used:

$$y_{jk} = (\beta_1 + u_{1j}) + (\beta_2 C_j + \beta_3 L_j + \beta_4 H_j + u_{2j}) t_{jk} + \varepsilon_{jk}, \tag{9}$$

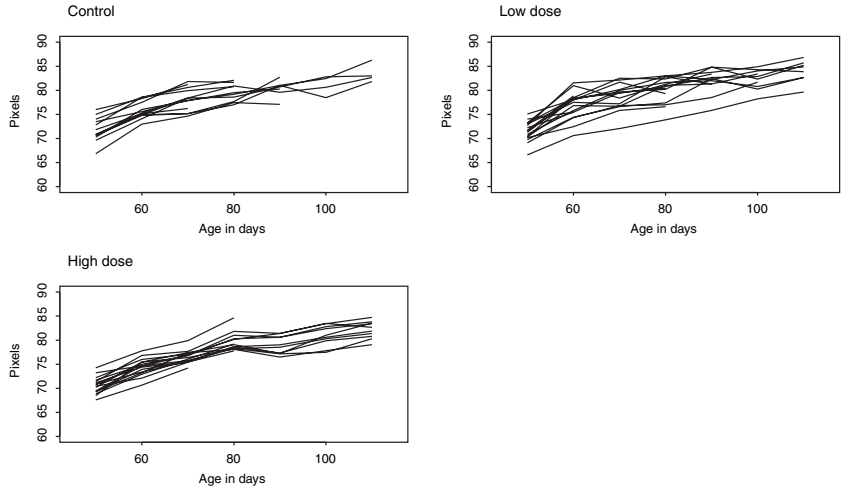

Fig. 2. Profiles for each of the treatment groups.

$$\boldsymbol{u}_j = (u_{1j}, u_{2j})^T \sim N(\mathbf{0}, \boldsymbol{V}), \varepsilon_{jk} \sim N(0, \sigma^2),$$

in which y_{jk} denotes the $k = 1,\ldots,K_j$th measurement for the $j = 1,\ldots,50$th rat and $t_{jk} = ln[1+(Age_{jk}-45)/10)]$. In (9), C_j, L_j, and H_j are indicator variables defined to be one if the subject belongs to the control, low-dose group or the high-dose group respectively, and zero otherwise. So, here $\boldsymbol{D} = \{\boldsymbol{y}, \boldsymbol{C}, \boldsymbol{L}, \boldsymbol{H}, \boldsymbol{t}\}$. Further, $\boldsymbol{V}$ is the covariance matrix of the random effects $\boldsymbol{u}_j$ and σ^2 is the variance of the level 1 residuals ε_{jk}. The transformation of the original time (age in days) implies that $t = 0$ corresponds to the start of the treatment. Note that the randomization in combination with this transformation of the original time scale allows one to assume that the subject-specific intercepts β_{1j} $(=\beta_1+u_{1j})$ do not depend on treatment. Consequently the parameter β_1 represents the average response at the start of treatment. The parameters β_2, β_3 and β_4 represent the average slopes for the control, low dose and high dose groups respectively. In this paper, interest lies in investigating some theory about the treatment effect. This theory will be translated into a model by putting constraints on the average slopes parameters since these directly measure the effect of treatment on the craniofacial growth. In this example $\boldsymbol{\theta} = \{\beta_2, \beta_3, \beta_4\}$ and $\boldsymbol{\omega} = \{\boldsymbol{V}, \sigma^2, \beta_1\}$.

Two models 1 and 2 will be compared. Model 1 is the unconstrained model. Model 2 reflects the theory that the higher the dose, the lower the growth rate. This renders the constraint $\beta_2 > \beta_3 > \beta_4$. Note that this constraint can be rewritten in the form (6). Consequently we have two competing models, $M_1 : \beta_1, \beta_2, \beta_3$ and $M_2 : \beta_2 > \beta_3 > \beta_4$.

From (9), the likelihood function of the data is

$$L(.|.) = \prod_{j=1}^{50} \int_{\boldsymbol{u}_j} \left\{ \prod_{k=1}^{K_j} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_{jk} - \boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{z}_{jk}^{\mathrm{T}}\boldsymbol{u}_j)}{2\sigma^2} \right) \right\} N(\boldsymbol{u}_j \mid \mathbf{0}, \boldsymbol{V}) \mathrm{d}\boldsymbol{u}_j,$$

where $\boldsymbol{x}_{jk}^{\mathrm{T}} = (1, C_j t_{jk}, L_j t_{jk}, H_j t_{jk})$ and $\boldsymbol{z}_{jk}^{\mathrm{T}} = (1, t_{jk})$. Using conjugate prior specifications and assuming independence between the model parameters, the encompassing prior to be used is

$$g(\boldsymbol{\theta}, \boldsymbol{\omega}) = \text{Inv-W}(\boldsymbol{V}|\lambda, \boldsymbol{S}) \times \text{Inv-}\chi^2(\sigma^2|v, \varphi^2) \times \prod_{p=1}^{4} N(\beta_p|\eta_p, \tau_p^2). \tag{10}$$

In (10), Inv-$\chi^2(\sigma^2|v, \varphi^2)$ denotes a scaled inverse Chi-square distribution with degrees of freedom v and scale φ, Inv-W$(\boldsymbol{V}|\lambda, \boldsymbol{S})$ denotes an inverse Wishart distribution with degrees of freedom λ and scale matrix $\boldsymbol{S}$ and $N(\beta_p|\eta_p, \tau_p^2)$ denotes a Normal distribution with mean η_p and standard deviation τ_p.

For the analysis, the first encompassing prior is data-based, using the method as described in principles 3 and 4 in Section 3. This leads to a $N(68.6, 0.3)$ for β_1 and a $N(7.2, 0.7)$ for each of β_2, β_3 and β_4 respectively. Turning to the prior distribution on $\boldsymbol{V}$, we take $\lambda = 1$ and $\boldsymbol{S} = \begin{pmatrix} 3.5 & -0.06 \\ -0.06 & 0.3 \end{pmatrix}$. Finally for the prior on σ^2 we set

Table 5. Prior sensitivity of posterior probabilities for rat data.

	Model	
	1	2
Prior 1	0.3724	0.6276
Prior 2	0.3701	0.6299
Prior 3	0.3697	0.6303

$\upsilon_0 = 1$ and $\sigma_0 = 1.5$. In the sequel the above mentioned specification will be referred to as Prior 1.

For a sensitivity analysis, two other specifications for the encompassing prior will be used and will be referred to as Prior 2 and Prior 3 respectively. These are obtained by increasing the variance of the (common) normal prior for the parameters β_2, β_3 and β_4 by factors of 4 and 9 respectively. Consequently under Prior 2, each of the parameters β_2, β_3 and β_4 will have a $N(7.2, 2.8)$ distribution and under Prior 3 they will each have a $N(7.2, 6.3)$ distribution. Note that the prior specifications of the other parameters (β_1, $\boldsymbol{V}$ and σ^2) remain as they were specified under Prior 1.

Subsequently 200,000 samples are drawn from each encompassing prior and the corresponding unconstrained posterior distribution. For each encompassing prior and posterior, these samples are used to estimate the quantities $1/c_2$ and $1/d_2$ respectively. Next, Bayes factors and posterior model probabilities are estimated using the procedure presented in Section 2. The results are displayed in Table 5.

From the table it is evident that the posterior probabilities are virtually independent of the choice of encompassing prior. These findings provide more support for our claim that for models with constraints of the form (6), the quantity $1/d_2$ is almost independent of the choice of encompassing prior. Further, Model 2 has the highest posterior probability. This result somewhat favours the theory that inhibiting testosterone production in rats slows down their cranofacial growth. In particular the higher the dose of the drug Decapeptyl, the lower the growth rate.

7 Discussion

In this paper we showed that to select the best model of a set of inequality constrained models, Bayesian model selection can be virtually objective for specific classes of constraints. Our approach is based on the so-called encompassing prior, which is the prior for the unconstrained model. The prior distributions for the constrained models can be derived from the encompassing prior. Using this set up, we showed that the Bayes factor for the encompassing model and any constrained model is the ratio of two proportions: the proportion of the encompassing prior ($1/c_q$), respectively posterior ($1/d_q$), in agreement with the constraints of the model at hand. We derived that $1/c_q$ is independent of the encompassing prior. Furthermore, we claimed that $1/d_q$ is virtually independent of the encompassing prior. The results

of three examples supported this claim. This means that for the presented classes of inequality constrained models a virtual objective Bayesian model selection procedure is obtained.

References

AGRESTI, A. and B. A. COULL (2002), The analysis of contingency tables under inequality constraints, *Journal of Statistical Planning and Inference* **1–2**, 45–73.

AGRESTI, A. (2002), *Categorical data analysis*, Wiley, Hoboken, New Jersey.

BERNARDO, J. M. and A. F. M. SMITH (1994), *Bayesian theory*, John Wiley & Sons, Chichester.

CHIB, S. (1995), Marginal likelihood from the Gibbs output, *Journal of the American Statistical Association* **90**, 1313–1321.

GELMAN, A., J. B. CARLIN, H. S. STERN and D. B. RUBIN (2000), *Bayesian data analysis*, Chapman and Hall, New York.

KASS, R. E. and A. E. RAFTERY (1995), Bayes factors, *Journal of the American Statistical Association* **90**, 773–795.

LEE, P. M. (1997), *Bayesian statistics: an introduction*, Arnold, London.

LINDLEY, D. V. (1957), A statistical paradox, *Biometrika* **44**, 187–192.

NETER, J., M. H. KUTNER, C. J. NACHTSHEIM and W. WASSERMAN (1996), *Applied linear statistical models*, 4th edn, McGraw-Hill/Irwin, Boston, 676–677.

VERBEKE, G. and E. LESAFFRE (1999), The effect of drop-out on the efficiency of longitudinal experiments, *Applied Statistics* **48**, 363–375.

VERBEKE, G. and G. MOLENBERGHS (2000), *Linear mixed models for longitudinal data*, Springer, New York.

Received: June 2004. Revised: December 2004.