

Research Master: Methodology and Statistics

Utrecht University, The Netherlands

MSc Thesis Jeroen C.L. Ooms

Title: The Highest Posterior Density Posterior Prior for Bayesian Model Selection

May 2009

Supervisors:

Dr. Irene Klugkist

Prof. Dr. Herbert Hoijtink

Preferred Journal of Publication: Journal of Mathematical Psychology

Word count: 5968

The Highest Posterior Density Posterior Prior for Bayesian Model Selection

Jeroen C.L. Ooms — Utrecht University

Abstract

In this paper a new type of prior is proposed that could be suitable in the context of model selection using Bayes factors. The Highest Posterior Density Posterior Prior (HPDPP) consists of a uniform distribution over the highest posterior density area, and basically results in truncation of low-density parameter space. The behavior and properties of the new prior are illustrated using constrained analysis of variance models. Both theoretical justification and simulations are used to argue that this prior has attractive properties for model selection. Because the HPDPP only uses relevant parameter space to determine the size of a model, results do not heavily depend on sample size or number of parameters. To avoid complications for interpretation it is recommended only to test exclusive models when using the HPDPP.

1 Introduction

The quality of a model is often defined as a trade-off between fit and complexity. More complex models perform better in describing the data, however they become harder to use, and might actually result in worse predictions (Myung, 2000). Model selection is about finding a balance between the two, in order to build models that are useful to the researcher and the people that have to use them. The trade-off between fit and complexity is a central theme in statistics. Many ideas and philosophies have been developed around this problem, resulting in different approaches to judge the quality of a

model. In classical statistics, usually a null hypothesis is formulated. The philosophy behind the null hypothesis significance test (NHST) is that an observed effect should be larger than what one could expect to happen by chance because of random fluctuation. Another approach is taken by information criteria, like the AIC (Akaike, 1973, 1981) or BIC (Schwartz, 1978). These methods more explicitly illustrate the balance between fit and complexity by defining a 'penalty function' based on the number of parameters in the model (Burnham and Anderson, 2002, 2004). Therefore in order for a more complex model to be preferred, a certain minimal increase in model fit is required. Although information criteria and NHST are founded in different theory, they are similar in the fact that they define an optimal balance between the fit and the number of parameters in a model.

In classical statistics, models are constrained by fixing parameters. The Bayesian framework on the other hand allows a different type of model constraints. Here, model parameters are considered to be random quantities, which enables us to put constraints on the *parameter space*. This can be used for evaluating models representing *informative hypotheses* like $\theta_1 > \theta_2 > \theta_3$, where θ_k ($k = 1, 2, 3$) denote certain model parameters. Informative hypotheses can then be modeled as subspaces of the total \mathfrak{R}^k parameter space (Hojtink et al., 2008). In this approach the focus is not on a null hypothesis or a specific parameter, but rather on comparing several candidate models. Just like in NHST or model selection based on information criteria, bigger models will usually have a better fit, but this does not mean they should always be preferred. To select the 'best' model rather than the biggest model, involving some measure of complexity is required. Complexity in this context is not defined by the number of parameters, but by the *size of the parameter space* covered by the model. However, what exactly makes up the parameter space, and how the parameter space is related to complexity is, just like in classical statistics, a matter of philosophy and debate.

It is known that results in Bayesian model selection are often highly sensitive to the specification of the prior distribution, which is usually considered undesired. Research has focused on finding priors that are objective to avoid prior sensitivity (Vaurio, 1992; Gelman et al., 2004). However, in this paper we take a different approach, and argue

that the prior distribution actually has an essential role in the fit/complexity trade-off. This idea leads to the proposal of a new type of prior that uses the *relevant parameter space*. The new prior is implemented and tested using the encompassing prior approach (EPA) as developed by Klugkist et al. (2005). The EPA provides an easy and straightforward method to evaluate support for a set of competing informative hypotheses. The method calculates Bayes factors for every constrained model compared to the unconstrained 'encompassing model' and provides an intuitive, yet powerful tool to gain insight in which of the models are supported by the data.

For simplicity in this paper we limit ourselves to analysis of variance (anova) models. Section 2 starts with a general introduction to Bayesian anova in the context of constrained models, and introduces the EPA. Furthermore the issue of prior sensitivity is discussed, and we show by example why an objective prior might not be appropriate for this application. In Section 3 the new prior is introduced and we show how to calculate the Bayes factor for anova models using this prior. In Section 4 some of the properties of the old and the new prior are compared using simulations. We end with a discussion that emphasizes the importance of using exclusive models.

2 Evaluating Informative Hypotheses

2.1 Bayesian Anova and Bayes Factors

The anova model $y \sim N(\mu, \sigma^2)$ is a generalization of the univariate normal model with multiple group means (hence μ denotes a vector) and one common residual variance σ^2 . It is the appropriate model to test whether a set of means come from one and the same population, or have different population means. In the frequentist approach, a hypothetical sampling distribution of parameter estimates $p(\hat{\mu}, \hat{\sigma}^2 | \mu, \sigma^2)$ is used, with μ, σ^2 being the 'real population values' under the null hypothesis. The Bayesian approach on the other hand allows the parameters itself to be random, and considers the derived distribution as the distribution of the parameters of interest conditional on the observed data, i.e. $p(\mu, \sigma^2 | y)$. This distribution is called the posterior distribution, and is the main ingredient for Bayesian analysis.

Let θ denote the set of parameters in the model, i.e. for the anova model: $\theta = \{\mu, \sigma^2\}$. The posterior distribution $p(\theta|y)$ is proportional to the product of the prior distribution $p(\theta)$ and the likelihood function $p(y|\theta)$ derived from the data, sometimes also denoted as $L(\theta|y)$. Note that the prior distribution $p(\theta)$ has to be specified by the user, and this choice might affect the results. The random distribution of parameters allows us to derive the marginal likelihood of a model. A marginal likelihood of a model M_i with parameters θ over data y is defined as the probability of finding data y , integrated over the total parameter space allowed by the model:

$$m_i(y) \equiv p(y|M_i) = \int p(\theta|M_i) p(y|\theta, M_i) d\theta \quad (1)$$

The Bayes factor, which is a common model selection tool in Bayesian analysis, is the ratio of the marginal likelihood of two competing models:

$$B_{ji}(y) \equiv \frac{p(y|M_j)}{p(y|M_i)} = \frac{\int p(\theta|M_j) p(y|\theta, M_j) d\theta}{\int p(\theta|M_i) p(y|\theta, M_i) d\theta} \quad (2)$$

Bayes factors can be used to make inferences about relative support, analogous to the likelihood ratio statistic in classical statistics, which uses the *maximum likelihood* (ML) rather than the *marginal likelihood* for comparing two models. A common way to interpret Bayes factors is by converting them to posterior model probabilities (PMP). For a finite set of competing models M , the PMP's are proportional to the product of the prior model probabilities, and the marginal likelihoods of the models:

$$p(M|y) \propto p(M) \times p(y|M) \quad (3)$$

Prior model probabilities represent a priori beliefs about model support, which are usually undesired in objective model selection. From (2) and (3) follows that when equal prior model probabilities are assumed, the posterior model probabilities in M are proportional to their Bayes factors with respect to a common model M_t :

$$p(M_i|y) = \frac{B_{it}(y)}{\sum_{M_j \in M} B_{jt}(y)} \quad (4)$$

The model M_t can be, but does not have to be one of the competing models. The next section shows an application that calculates Bayes factors for a set of constrained models with respect to the unconstrained model.

2.2 Constraining Models and the Encompassing Prior Approach

In this paper the focus is on comparing anova models with different constraints on the same parameter space. One way to perform this is by treating the constraints as prior information, and applying them in the prior distributions. This principle forms the basis of the encompassing prior approach (EPA). The EPA is a method for deriving Bayes factors for models with parameter space constraints like equality constraints (e.g. $\theta_1 = \theta_2$) and inequality constraints (e.g. $\theta_1 < \theta_2$). In the EPA, one prior is defined for the unconstrained model, which is called the encompassing prior. In this paper we will use a semi conjugate prior, hence assume a priori independence of parameters (Gelman et al., 2004, p. 81). From the encompassing prior, the prior for every constrained model is derived by setting the prior density to zero in parameter space not allowed by the constraints. Formally this is noted using the indicator function I , which has value 1 in the parameter space allowed by a model, and zero elsewhere. Hence the prior for a constrained anova model M_i equals:

$$p(\theta|M_i) = p(\mu, \sigma^2|M_i) = c_i \times \prod_{j=1}^k p(\mu_j) \times I_{\mu \in M_i} \times p(\sigma^2) \quad (5)$$

where $p(\mu_j)$ and $p(\sigma^2)$ are unconstrained prior distributions for the mean and variance parameters. Note that in our application, model constraints are only applied to μ parameters, not the σ^2 parameter. Because a constrained prior only contains a part of the original, unconstrained prior density, every prior distribution has to be multiplied with a marginalizing constant c_i , to make it a proper density function. The value for c_i will become important later on, and is used to measure the size of a model. The remainder of the analysis continues with the usual steps as in described Section 2.1: the prior distributions for the models are updated with the likelihood function derived from the data and the marginal likelihood for every model is derived by integrating the model parameters out of the posterior distributions. The marginal likelihoods can then be used to calculate Bayes factors and PMP's.

Analytically deriving marginal likelihoods for all constrained models can be very tedious and is probably not suitable for most applied researchers. However, Klugkist et al. (2005) explain that sampling from the unconstrained prior and unconstrained

posterior is sufficient to derive the size (c_i^{-1}) and fit (d_i^{-1}) for every of the constrained models, where:

$$p(\theta|M_i) = \frac{p(\theta|M_0)I_{M_i}}{\int p(\theta|M_0)I_{M_i}d\theta} = c_i \cdot p(\theta|M_0)I_{M_i} \quad (6)$$

and

$$p(\theta|y, M_i) = \frac{p(\theta|y, M_0)I_{M_i}}{\int p(\theta|y, M_0)I_{M_i}d\theta} = d_i \cdot p(\theta|y, M_0)I_{M_i} \quad (7)$$

Here M_0 represents the unconstrained model, and c_i and d_i are marginalizing constants that are needed to make respectively the prior and posterior of a constrained model M_i into proper densities. Model size is defined as the inverse of the marginalizing constant for the prior: c_i^{-1} . As a general rule, the more constraints a model has, the smaller the parameter space of the model, and the higher the value for c_i . The fit of a model is defined as the inverse of the marginalizing constant for the posterior: d_i^{-1} . Hence a model has a high fit when it covers a large proportion of the unconstrained posterior density, and therefore has a low marginalizing constant d_i . Klugkist and Hoijtink (2007) have shown that for constrained models, the ratio of its fit and size d_i^{-1}/c_i^{-1} equals the Bayes factor for constrained model M_i with respect to the unconstrained model M_0 . This result is used in the EPA to derive the Bayes factors for all constrained models under investigation, using only a sample from the unconstrained prior and posterior.

2.3 Prior Sensitivity

Like most Bayesian analyses, the EPA requires specification of a prior distribution. Most people would like the choice of prior, or any other subjective choice, not to influence the results too much if there is no substantial argumentation to do so. Therefore often uninformative priors are used that have practically no effect on the posterior distribution. However, ironically, uninformative priors can be highly influential in the context of Bayesian model selection. The reason for this is that the prior distribution does not just represent prior information. It also defines the parameter space over which the models are evaluated, and consequently defines the size of a model. As a result, both the location and size of the prior distribution can have a big impact on the results. The issues are commonly referred to as *prior sensitivity*, and have caused a

lot of criticism on Bayesian statistics in general (Mayo, 1996; Howson, 2002; Sober, 2002).

In the original EPA from Klugkist et al. (2005), symmetric priors are used that attribute an equal a priori probability to every ordering of parameters. Therefore the prior is always centered at the point where all model parameters are equal. In the context of inequality-constrained models, Klugkist et al. showed that using this type of prior distribution, the model selection is hardly affected by the prior specification and therefore is considered objective. However it was recognized that in the context of equality constrained models, results are still affected by the 'vagueness' of the prior. For uninformative priors, models with one or more equality constraints will be attributed an extremely small size, simply because they strike a very small proportion of the total parameter space. As a result, vague priors can lead to nearly certain acceptance of the null model (Jeffreys, 1998). The problem is commonly referred to as Lindley's Paradox (Lindley, 1957). For a detailed outline of the problem in the context of EPA we refer to (Hoijsink et al., 2008, chap. 4).

To deal with the problem of prior sensitivity, research has been done to find an optimal balance between priors that are objective, yet not too vague to cause Lindley's paradox. The use of a small part of the data, i.e. a training sample to obtain a proper 'posterior prior' is a popular method to resolve issues of prior sensitivity (Berger and Pericchi, 1996; Perez and Berger, 2002; Berger and Pericchi, 2004). In the context of the EPA, Mulder et al. (2009) have developed the conjugate expected constrained posterior prior method (CECPP). The CECPP uses a minimal training sample to obtain a posterior prior that results in c_i values that comply with the strict definition as formulated in Mulder et al. (2009). This definition ensures equal a priori probabilities for every ordering of parameters, as in the original EPA from Klugkist. In the CECPP the size of the prior is not subjective, however it is usually very big because it is based on the minimal amount of data needed to derive a proper posterior.

2.4 Issues with the CECPP

As a direct consequence of its design, the CECPP has 2 properties that are important for its behavior. The first property is that it has a static size. Because the CECPP is based on a *minimal training sample*, the size of the prior does not change when the sample size n changes. Stated otherwise: conditional on μ and σ^2 the expected size $E(c_i^{-1}|\mu, \sigma^2)$ of a model is identical for all values of n . The second property of interest is its symmetry: the definition of complexity from Mulder et al. (2009) requires that every ordering of means is equally likely in the prior. This implies that the size of an inequality constrained model is the same for all populations and all sample sizes. For example, the model $M_i : \mu_1 < \mu_2 < \mu_3$ always has a size of $(3!)^{-1} = 1/6$. Klugkist and Hoijtink (2007) and Mulder et al. (2009) argue that these properties ensure objective model selection.

However, these properties might not always work out as intended. To illustrate this point, we used EPA to calculate Bayes factors on several generated datasets. These datasets were generated to be perfect samples, meaning that the estimated mean and variance for every dataset are identical to the population they were sampled from. Because these samples technically do not include sampling error, we consider the Bayes factors for these samples to be the expected Bayes factors conditional on the specified population properties and sample sizes. The left diagram in Figure 1 shows the expected Bayes factor for the model $M1 : \mu_1 < \mu_2 < \mu_3 < \mu_4 < \mu_5 < \mu_6 < \mu_7$ versus the unconstrained model $M0$, when using the CECPP method. The expected Bayes factors were derived for varying sample sizes, with the actual population values fixed at $\mu_1 = 0, \mu_2 = 3, \mu_3 = 6, \mu_4 = 9, \mu_5 = 12, \mu_6 = 15, \mu_7 = 14, \sigma^2 = 10$. Hence, this represents a realistic situation in which most model-constraints are fulfilled, but one of them is not. One can argue about the quality of $M1$ for this population, however Figure 1 shows an unexpected property of using static priors: results can become extremely dependent on the sample size n in an unpredictable way. In this case, model $M1$ quickly gains support as n increases and at $n = 50$ it has an expected Bayes factor around 200, resulting in a PMP of 0.995 in favor of $M1$. However, for bigger samples the expected Bayes factor decreases and at $n = 300$ it is actually 0, resulting in a PMP

of 1 in favor of M_0 . Note that these differences have nothing to do with sampling error. These are Bayes factors for the model as expected for different sample sizes of this population.

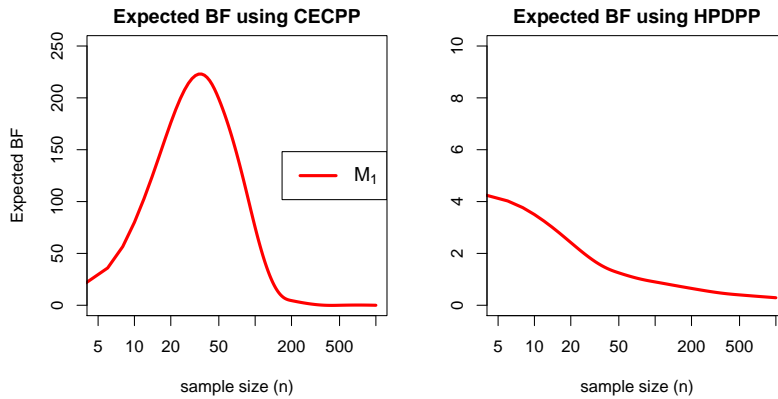


Figure 1: The expected Bayes Factor of several sample sizes for a fixed model and population, using two different priors. Notice the different scales on the y axis.

The reason for this remarkable behavior is that under the static and symmetric CECPP prior, the size of M_1 is fixed at $(7!)^{-1} = 1/5200$. This means that if at least $1/5200$ part of the posterior density is in the parameter space of M_1 , it will be the preferred model. Because the fit of M_1 depends on the size of the posterior, results become highly sensitive to sample size n . In every dataset ML estimates are identical and not completely in agreement with M_1 . However the Bayes factor for M_1 initially grows for bigger sample sizes, which seems counter intuitive. Only for sample sizes above $n = 50$ the posterior becomes so small that it has almost completely disappeared from the parameter space of M_1 . As a result the Bayes factor for M_1 collapses and quickly approximates zero.

Although the CECPP prior has been designed to be objective, its behavior does not seem unbiased at all. In fact, the static c_i^{-1} values that result from the CECPP function very similar to a prior model probability with a fixed preference for small models. For this reason we argue that a static, symmetric prior might not be very appropriate for model selection, and propose an alternative. The prior proposed in this

paper and presented in the next section results in expected Bayes factors as plotted in the right diagram of Figure 1. For the example above, this prior also results in an initial preference for model $M1$, however it gradually decreases support for bigger sample sizes and eventually approximates a Bayes factor of zero just like the CECPP prior.

3 The HPD Posterior Prior

Substantively it is not common to attach great value to big or infinite parameter spaces that represent unrealistic theory. For example in research comparing age of mortality for smokers and non-smokers, parameter space that represents ages above 120 seems relatively unimportant. Therefore we argue that it makes more sense to use the data to determine the *relevant parameter space*, and then let the model selection happen within this area. Usually only parameter space with considerable density represents realistic theory. This is why we propose to use the data to derive the highest density region for the μ parameters and evaluate the model support over this subspace of the parameter space. This can be accomplished by using an uninformative prior over the highest density area and setting the prior density to zero outside of this area. Because this area is no longer infinitely large, a completely uniform, proper prior can be used. The resulting posterior then is a truncated version of the posterior that we would have obtained with an improper uninformative prior. For this reason we refer to the area as the Highest Posterior Density (HPD) area. The uniform prior over this area will be referred to as the HPD posterior prior (HPDPP), and the posterior as the HPD posterior (HPDP). Models that do not cover any of the HPD area are attributed a Bayes factor of zero per definition.

The only arbitrary choice that is left in this application is the size of the HPD area. In this paper we will use the 95% HPD area as the relevant parameter space, 5% being a common and widely accepted error margin in many applied fields. Figure 2 illustrates the process for an example with two μ parameters. Note that although the plot only shows μ parameters, the actual distributions also contain a σ^2 parameter. However, the σ^2 parameter is of no substantial interest in our application, and is therefore not included in the calculation of the HPD region. What will happen when using the

HPDPP is that model support is evaluated *within the relevant parameter space*. Note that the size of the HPD prior is dynamic: the more data, the smaller it gets. It is also asymmetric: it centers around the ML estimates of the parameters. These properties make the HPDPP fundamentally different from most current approaches.

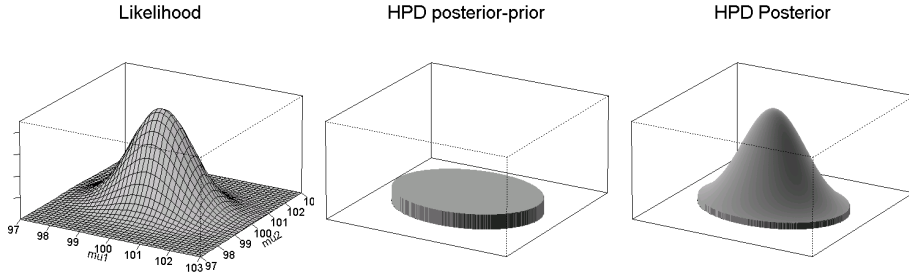


Figure 2: Likelihood, HPDPP and HPDP

To derive the HPDPP and HPDP, we use the fact that the likelihood is equal to the posterior that is obtained with a completely uninformative prior distribution. To avoid confusion, this prior will be called the *pre-prior*, and the resulting posterior the *pre-posterior*. In this application improper pre-priors are used for all parameters as described in Gelman et al. (2004). The parameters are assumed to be a priori independent, therefore the pre-prior for the anova model equals:

$$p_{pre}(\mu, \sigma^2) \propto 1/\sigma^2 \quad (8)$$

The pre-prior is updated by the likelihood distribution $p(y|\mu, \sigma^2)$, resulting in the pre-posterior $p_{pre}(\mu, \sigma^2|y)$. Note that the distribution of the pre-posterior is identical to the distribution of the likelihood. From the pre-posterior $p_{pre}(\mu, \sigma^2|y)$, the 95% HPD region over the μ -parameters is calculated. The density outside this area is then set to zero, which is formally noted using the indicator function I_{hpd} , which has value 1 within the HPD area, and 0 elsewhere. The HPDPP then equals:

$$p_{hpd}(\mu, \sigma^2) \propto p_{pre}(\mu, \sigma^2) \times I_{hpd} \quad (9)$$

The HPD posterior can be obtained by updating the HPDPP with the likelihood. Because the likelihood equals the pre-posterior, this distribution is proportional to the

truncated pre-posterior:

$$p_{hpd}(\mu, \sigma^2|y) \propto p_{hpd}(\mu, \sigma^2) \times p(y|\mu, \sigma^2) \propto p_{pre}(\mu, \sigma^2|y) \times I_{hpd} \quad (10)$$

Just like in the original EPA, a random sample from the prior (9) and posterior (10) distribution can be used to derive the Bayes factors for the candidate models. The next section will explain how samples can be obtained from the HPDPP (9) and HPDP (10).

3.1 Sampling from the Prior and Posterior

Because in this application the posterior is needed to determine the prior, we have the unusual order in which we start with determining the posterior distribution, and then derive the prior distribution. However, this is no problem since the prior in this application does not really contain any prior information, its only purpose is to estimate the size of the models.

3.1.1 Obtaining A sample from the HPD Posterior

The easiest way to obtain a sample from the HPDP is by sampling from the pre-posterior, and removing the draws outside of the HPD region. A Gibbs sampler can be used to sample from the pre-posterior like described in standard literature (Gelman et al., 2004; Gill, 2008; Lynch, 2007). The Gibbs sampler uses the conditional univariate distributions of all of the model parameters and iteratively samples a value conditional on the values of the previous iteration. For the pre-posterior of the anova model the conditional distributions equal

$$\begin{aligned} p(\mu_k|\sigma^2, y) &\propto N(\bar{y}_k, \frac{\sigma^2}{n_k}) \\ p(\sigma^{-2}|\mu, y) &\propto \Gamma(\frac{N}{2}, \frac{1}{2} \sum_i^n (y_i - \mu_k)^2) \end{aligned} \quad (11)$$

where μ_k equals the μ parameter for every group, n_k equals the sample size per group, and \bar{y}_k the group's sample mean. Note that although we are actually only interested in the μ parameters, the σ^2 parameter has to be incorporated in the Gibbs sampler to obtain the correct conditional distributions. Once a sufficiently large sample from

the pre-posterior is obtained, the draws outside of the HPD area are to be removed from the sample.

Finding the highest density area of a multivariate distribution can be very complicated when the distribution has no clear shape. However, in this application, likelihood functions for the mean parameters are independent, symmetric and monotonic increasing towards the center. The only difficult property is that they have different means and variances, but this can be solved by temporarily standardizing the draws in the sample. The distribution then becomes a multivariate standard normal density, and the 95% HPD region can easily be derived by calculating the Euclidean distance towards the center of the distribution for every draw, and remove the 5% draws with the highest distance.

The truncation is illustrated for a model with 2 μ -parameters in the four diagrams of Figure 3. The first diagram shows a plot of a sample of μ parameters from the pre-posterior, obtained from the Gibbs sampler of (11). The pre-posterior has the recognizable bivariate normal distribution for independent μ parameters. In this example the posterior mean for μ_1 equals 100 with a standard deviation of 1.3, and the posterior mean for μ_2 equals 102 with a standard deviation of 2. To be able to truncate this distribution, the standardized version of both variables is created, which is shown in the second diagram. In this standardized bivariate normal distribution, it is easy to calculate the Euclidean distance towards the center of the distribution for every draw. When this is done, we calculate the 95th percentile for these distances, and call this value r (as in radius). Next, all draws with a bigger distance to the center than r are removed from the sample. These are the draws outside the circle in diagram 2. When the same cases are removed in the non-standardized data, this results in the typical ellipse shape, covering the 95% HPD of the original distribution. The remaining draws can be considered a random sample from the HPD posterior (the third picture in Figure 2), and will be used as such for the computation of the Bayes factors.

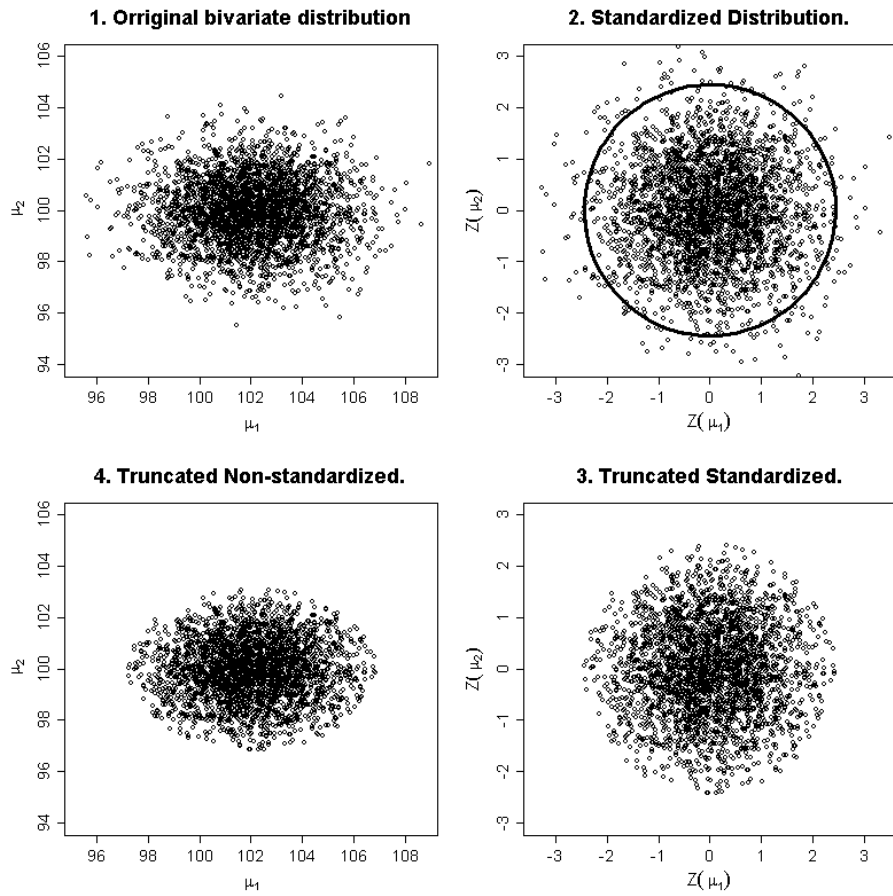


Figure 3: An example of truncating a bivariate normal distribution. Note that diagram 3 is the lower right box.

3.1.2 Obtaining a Sample from the HPD Posterior Prior

To obtain a sample from the HPDPP, a uniform distribution has to be truncated at exactly the same HPD area as was done for the posterior. Therefore the value r and the means and standard deviations for the μ distributions that were calculated from the pre-posterior are used again. The process is illustrated in Figure 4. An easy way is to start with a random sample from the uniform distribution $U(-r, r)$ for every μ parameter, as illustrated in the first diagram. Again, for all these draws the Euclidean distance towards the center is calculated, and draws with a distance above the maximum value of r are removed, which results in the second diagram of Figure 4. Because the same value for r is used, this area is exactly identical to the area within the circle of diagram 2 in Figure 3. The difference between the two samples is that these draws are uniformly distributed, whereas the sample in Figure 3 has a truncated bivariate normal distribution. The only thing left to do is transform the sample to the scaling of the original distribution. To realize this, the draws are multiplied with the standard deviation and added to the mean using the hyperparameters from the pre-posterior. This brings us to the third diagram in Figure 4, which is a uniform sample over the same HPD area as was used for the HPD posterior, and represents the second picture from Figure 2.

In the example above a model with only two μ -parameters was used to visualize the process. However, the method of sampling and using Euclidean distances easily generalizes to models with more parameters. To explore the behavior of the HPDPP, the techniques described in this section were implemented in the EPA and tested in numerous simulations. Different types of constrained anova models were tested for datasets varying in size, number of parameters and parameter values. In the next section two simulations are presented which illustrate the different properties of the CECPP and HPDPP in the context of constrained model selection.

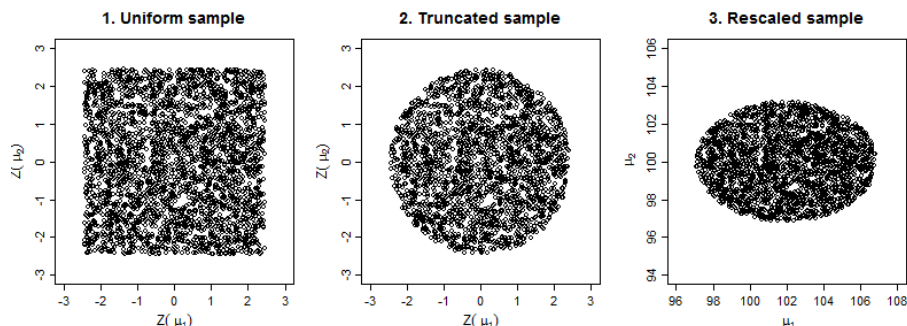


Figure 4: An example of generating a HPDPP sample

4 Comparing Behavior of the CECPP and HPDPP

Theoretically, the Bayes factor can be calculated over any subspace of the parameter space. Within the context of informative hypotheses there are a few popular type of constraints for defining a model. We have already seen the inequality constraint (e.g. $\mu_1 < \mu_2$), representing parameter space where μ_1 is smaller than μ_2 . Another type of constraint that is introduced in Hoijtink et al. (2008) is the approximate equality constraint, e.g. $\mu_1 \approx \mu_2$, that requires specification of a *minimal relevant effect size*. For example for 2 parameters we could specify a minimal effect size δ_{min} , where $\mu_1 \approx \mu_2 \equiv |\mu_1 - \mu_2| < \delta_{min}$. This constraint models the parameter space where the difference between μ_1 and μ_2 is smaller than what is considered relevant by the user. A straightforward generalization of the approximate equality for multiple parameters could be introduced by specifying a minimal between group variance (σ_μ^2) or a minimal proportion of explained variance (η^2).

4.1 Simulations

In this section two small simulations are presented that illustrate and compare the behavior of the CECPP and HPDPP. As before, perfect samples were used to derive the expected Bayes factors of the models under consideration for various sample sizes. In this section the Bayes factors are converted to posterior model probabilities (PMP) because this is how Bayes factors are often used to make inferences about model support.

In both upcoming simulations an exclusive and exhaustive set of models is evaluated, i.e. every point in the parameter space is covered by one and only one model. This is a requirement for the HPDPP prior method, but not for the CECPP method. However, using exclusive models overcomes the problem of bounded model support as described in Hoijsink et al. (2008). In the discussion we will get back to the issue of overlapping models.

4.1.1 Simulation 1

In the first simulation, the support for $M0 : \mu_1 \approx \mu_2$ versus $M1 : \mu_1 > \mu_2$ versus $M2 : \mu_1 < \mu_2$ is compared, with the minimal relevant effect size of $\delta = 3$. Because an approximate equality constrained model is included, and we require exclusive models, the models are defined by: $M0 : |\mu_1 - \mu_2| < 3$, $M1 : \mu_1 - \mu_2 > 3$, $M2 : \mu_2 - \mu_1 > 3$.

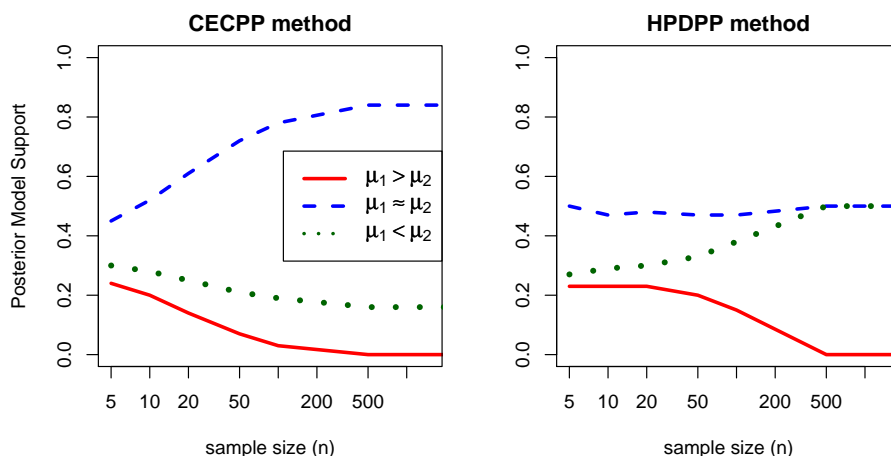


Figure 5: Expected posterior models probabilities for varying sample sizes in simulation 1.

For the purpose of this simulation, the models were tested in a population where $\mu_1 = 10$ and $\mu_2 = 13$ and $\sigma^2 = 10$. Hence, the actual population is exactly on the border of the parameter space of the models $M0$ and $M2$. As can be seen in Figure 5, under these conditions, the CECPP converges to support the model which is considered the smallest model, $M0$ in this case. The HPDPP on the other hand converges to

a fifty/fifty support: when the data are exactly on the boundary of two models, the HPDPP has asymptotically no preference for either model. This example illustrates an important difference between the CECPP and HPDPP. In the CECPP every model has a fixed size, whereas in the HPDPP, model size becomes less influential for larger sample sizes. In this example, there are only 2 parameters, and the size of the models do not take extreme values. However, for multiple parameters a model's size can become very small, and the preference for small models is magnified, as will be shown in the next simulation.

4.1.2 Simulation 2

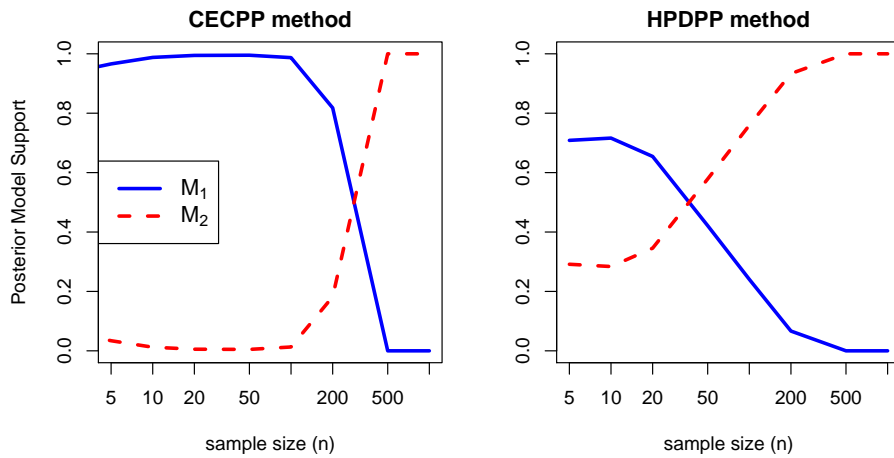


Figure 6: Expected posterior models probabilities for varying sample sizes in simulation 2.

In the second simulation the support for the model $M1 : \mu_1 < \mu_2 < \mu_3 < \mu_4 < \mu_5 < \mu_6 < \mu_7$ versus the complementary model $M2$ (i.e. all parameter space not occupied by $M1$) was investigated for the population $\mu_1 = 0, \mu_2 = 3, \mu_3 = 6, \mu_4 = 9, \mu_5 = 12, \mu_6 = 15, \mu_7 = 12, \sigma^2 = 10$. The situation is similar to the example from section 3, however, instead of comparing $M1$ to an unconstrained model, it is compared to the complementary model. As can be seen in Figure 6, both methods have a preference for $M1$ for small sample sizes, but eventually reject the model and

show a very strong preference for the complementary $M2$. This can be explained by the fact that model $M1$ is in fact very close, yet not in complete agreement with the data. The difference between the CECPP and HPDPP exists in the amount of support that is initially given to the models. The CECPP very quickly develops a very strong preference for the smaller model $M1$, and the PMP for $M1$ grows to a very convincing .995 for samples of $n=50$, but then drops quite unexpectedly. The HPDPP on the other hand seems more careful with its support, and monotonically converges to support $M2$. It recognizes that although $M1$ seems to be quite close to the data, the final constraint might be incorrect which becomes more obvious for larger sample sizes.

5 Discussion

In this paper we have introduced the HPDPP and shown some of its unique properties. The property that is most appealing for model selection is that the HPDPP is asymptotically unbiased. This means that the HPDPP does not result in a fixed preference for small models, like in the case of static priors like the CECPP. The HPDPP does take the size of models into account to prevent big models from easily gaining major support. However when more data are collected and the posterior is more peaked, the prior also becomes smaller and the size of a model becomes by design less influential on its Bayes factor. As a consequence, Bayes factors are less dependent on sample size and the number of parameters of a model, and results seem to be more stable than with the CECPP method.

Because of the asymptotic unbiasedness, it is highly recommended to test a set of exclusive models when using the HPDPP. The HPDPP does not have a strong preference for small models, hence the evidence in favor of a model will be bounded when there is a larger model overlapping the same parameter space. This behavior is often undesired in the context of model selection, and complicates interpretation. The use of exclusive models avoids this problem and is therefore in many cases more powerful.

Although not discussed in this paper, it is also possible to derive Bayes factors for point models, i.e. models of a lower dimension than the total parameter space. A typical example of a point model is an equality constrained model, e.g. $M_i : \mu_1 = \mu_2$.

Models like M_i do not cover a subspace of the parameter space, but are represented by a line within this space. Because of the difference in dimensionality, estimation of these Bayes factors is more complicated and requires alternative methods that are beyond the scope of this paper. A second problem with point models is that they do not cover any parameter space, which complicates our advice of using exclusive models. Behavior of the HPDPP for point models was shortly explored, suggesting that the problem of bounded Bayes factors also exists when a point model is overlapped or immediately adjacent to a model of a higher dimension. However, further research is needed to examine the properties of the HPDPP for point models and define guidelines for optimal use.

As a final thought, we would like to say a word about objectivity in model selection. Sceptics might feel that results can hardly be reliable when they are very dependent on the choice of prior distribution. However model selection always involves some degree of subjective judgment regarding the quality of a model. There is no single answer to what makes a good model. Statistics can quite easily quantify the fit of a model to the data, however we also want a model to be parsimonious, which requires some judgment about the complexity of a model. There are many ways to define complexity and the optimal balance between fit and complexity. Different methods implement different philosophies and result in different evaluations of model quality. In Bayesian model selection the prior plays an important role in this definition, which gives the user some degree of control over the process. The choice of prior is more explicit, but not more subjective than any other model selection criterion. It is the responsibility of the user to make sensible choices and interpret results accordingly.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csake, F., editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest: Akademiai Kiado. Akademiai Kiado.

- Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, 16:3 – 14.
- Berger, J. and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal Of The American Statistical Association*, 91(433):109–122.
- Berger, J. O. and Pericchi, L. R. (2004). Training samples in objective bayesian model selection. *The Annals of Statistics*, 32:841–869.
- Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York, NY, 2 edition.
- Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: Understanding aic and bic in model selection. *Sociological Methods Research*, 33:261 – 304.
- Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis, 2nd ed.* Chapman & Hall/CRC, London.
- Gill, J. (2008). *Bayesian Methods, A social and Behavioral Sciences Approach, 2nd ed.* Chapman & Hall/CRC, London.
- Hoijtink, H., Klugkist, I., and Boelen, A. (2008). *Bayesian Evaluation of Informative Hypotheses*. Springer New York.
- Howson, C. (2002). *Bayesianism in statistics.*, pages 39–69. Oxford: Oxford University Press.
- Jeffreys, H. (1998). *Theory of Probability*. Oxford University Press: Oxford, 3rd edition.
- Klugkist, I. and Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, 51(12):6367–6379.
- Klugkist, I., Laudy, O., and Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, 10(4):477–493.
- Lindley, D. (1957). A statistical paradox. *Biometrika*, 44:187192.

- Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer New York.
- Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: The University of Chicago Press.
- Mulder, J., Hoijtink, H., and Klugkist, I. (2009). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Under review*.
- Myung, I. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, 44(1):190–204.
- Perez, J. and Berger, J. (2002). Expected-posterior prior distributions for model selection. *Biometrika*, 89(3):491–511.
- Schwartz (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.
- Sober, E. (2002). *Bayesianism, its scope and limits.*, pages 21–38. Oxford: Oxford University Press.
- Vaurio, J. (1992). Objective prior distributions and bayesian updating. *Reliability Engineering & system safety*, 35(1):55–59.