# Selección de modelos bayesianos: Discrepancia $\chi^2$ con la distribución uniforme y posibles extensiones

Julián de la Horra

Departamento de Matemáticas
Universidad Autónoma de Madrid
E-mail: julian.delahorra@uam.es

# The problem

Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from a continuous random variable $X$.

We have to choose among $m$ different Bayesian models, $M_i$, $i = 1, ..., m$. Each Bayesian model consists of two components: a sampling density, $f_i(x|\theta)$ (where $\theta \in \Theta$), and a prior density, $\pi_i(\theta)$.

The cumulative distribution function corresponding to $f_i(x|\theta)$ will be denoted by $F_i(x|\theta)$.

In short:

$$M_i = \{f_i(x|\theta), \pi_i(\theta)\}, \; i = 1, ..., m.$$

# The problem

The problem of Bayesian model selection has been studied in many papers. For instance:

Bayesian model selection based on Bayes factors:
O'Hagan (1995; J. Roy. Statist. Soc. B)
Berger and Pericchi (1996; J. Amer. Statist. Assoc.)

Bayesian model selection based on a predictive approach:
Gelfand, Dey, and Chang (1992; Bayesian Statistics 4)
Gutierrez-Peña and Walker (2001; J. Statist. Plan. Infer.)
De la Horra and Rodríguez-Bernal (2005; Statist. Probab. Lett.)
De la Horra and Rodríguez-Bernal (2006; Commun. Statist. Theor. Meth.)

# The method

Next, a different and easier method is proposed and analyzed [De la Horra (2008; Commun. Statist. Theor. Meth.)]. The method is based on the following idea:

Let us assume that $\mathbf{X} = (X_1, ..., X_n)$ is a random sample from a continuous random variable $X$ with density function $f(x)$ and cumulative distribution function $F(x)$.

Then, $(F(X_1), ..., F(X_n))$ can be considered as a random sample from a $U(0, 1)$, because of $F(X)$ follows a $U(0, 1)$ distribution.

As a consequence, we hope that $(F(X_1), ..., F(X_n))$ will be well fitted to the $U(0, 1)$ distribution.

# First step of the method

**(1)** First of all, we measure the discrepancy between the sample we have obtained, $\mathbf{x} = (x_1, ..., x_n)$, and the distribution function $F_i(x|\theta)$ (for a fixed $\theta$), by measuring the $\chi^2$ discrepancy between $(F_i(x_1|\theta), ..., F_i(x_n|\theta))$ and the $U(0,1)$ distribution; for doing that, we partition the interval (0, 1) in $k$ subintervals, (0, 1/k), (1/k, 2/k),...,((k-1)/k, 1). So, the $\chi^2$ discrepancy will be measured as usual:

$$D_i(\mathbf{x}, \theta) = \sum_{j=1}^{k} \frac{[O_{ij}(\theta) - n(1/k)]^2}{n(1/k)} = \sum_{j=1}^{k} \frac{[O_{ij}(\theta) - (n/k)]^2}{n/k},$$

where $O_{ij}(\theta)$ is the number of elements of $(F_i(x_1|\theta), ..., F_i(x_n|\theta))$ we have obtained in each subinterval.

The idea behind this discrepancy is simple: if $F_i(x|\theta)$ (for a fixed $\theta$) is a good model, $D_i(\mathbf{x}, \theta)$ will be close to zero; if $F_i(x|\theta)$ (for a fixed $\theta$) is not a good model, $D_i(\mathbf{x}, \theta)$ will be far from zero.

# Second step of the method

**(2)** Of course, we are interested in evaluating the discrepancy between the sample we have obtained, $\mathbf{x} = (x_1, ..., x_n)$, and the whole Bayesian model, $M_i$, given by (1). The Bayesian solution is easy; first of all, we compute the usual posterior density of the parameter,

$$\begin{aligned} \pi_i(\theta|\mathbf{x}) &= \pi_i(\theta|x_1, ..., x_n) = \frac{f_i(x_1, ..., x_n|\theta)\pi_i(\theta)}{\int_\Theta f_i(x_1, ..., x_n|\theta)\pi_i(\theta)d\theta} \\ &= \frac{f_i(x_1|\theta)...f_i(x_n|\theta)\pi_i(\theta)}{\int_\Theta f_i(x_1|\theta)...f_i(x_n|\theta)\pi_i(\theta)d\theta} \,, \end{aligned}$$

and then we evaluate the posterior expected discrepancy between the sample $\mathbf{x}$ and the model $M_i$:

$$D_i(\mathbf{x}) = \int_\Theta D_i(\mathbf{x}, \theta)\pi_i(\theta|\mathbf{x})d\theta.$$

# Third step of the method

**(3)** Finally, we only have to compare $D_1(\mathbf{x})$,..., $D_m(\mathbf{x})$, and choose the Bayesian model having the smallest posterior expected discrepancy.

# Asymptotical properties of the method

It is obviously interesting to know something about the asymptotic distribution of the posterior expected discrepancy, under the "true" model $M = \{f(x|\theta), \pi(\theta)\}$.

In this case, we want to study:

$$\lim_{n} \int_{\Theta} D(x_1, ..., x_n, \theta)\pi(\theta|x_1, ..., x_n)d\theta,$$

and so, we are looking for a result in the style of Helly's theorems.

First problem we find: in Helly's theorems, the function we integrate is fixed, and here, $D(x_1, ..., x_n, \theta)$ depends on $n$. For solving this problem we replace $D(x_1, ..., x_n, \theta)$ by $D(x_1, ..., x_N, \theta)$, where $N$ is fixed and large enough. For fixing ideas and showing that this is a reasonable approximation, we give the following lemma:

# Asymptotical properties of the method

**Lemma 1.-** Let us assume that, for any fixed $\varepsilon > 0$, there exists $N$ such that for all $n \geq N$:
(a) $|D(x_1, ..., x_n, \theta) - D(x_1, ..., x_N, \theta)| < \varepsilon$, for all $\theta$ in a subset $\Theta_0 \subset \Theta$ such that $Pr(\Theta_0 | x_1, ..., x_n) \geq 1 - \varepsilon$,
b) $|D(x_1, ..., x_n, \theta) - D(x_1, ..., x_N, \theta)| < M$ for all $\theta$ in $\Theta - \Theta_0$.

Then:

$$\begin{aligned} & \left| \limsup \int_\Theta D(x_1, ..., x_n, \theta) \pi(\theta | x_1, ..., x_n) d\theta \right. \\ - & \left. \limsup \int_\Theta D(x_1, ..., x_N, \theta) \pi(\theta | x_1, ..., x_n) \right| \\ \leq & \ \varepsilon(M + 1). \end{aligned}$$

# Asymptotical properties of the method

Second problem we find: in Helly's theorems, the function we integrate is continuous, and here, $D(x_1, ..., x_N, \theta)$ is not continuous. We next obtain a result, in the style of Helly's theorems, making use of two specific conditions we have in this problem:

(a) It is very reasonable to assume that the posterior distribution converges to the degenerate distribution giving all the mass to the true value of the parameter, $\theta_0$.

(b) The $\chi^2$ discrepancy, $D(x_1, ..., x_N, \theta)$, for a fixed $N$, is a simple function of $\theta$: the number of values it can take is, perhaps huge, but finite. Therefore, we will denote

$$D(x_1, ..., x_N, \theta) = \sum_{i=0}^{r} a_i I_{\Theta_i}(\theta),$$

where $I_A(\theta)$ is the usual indicator function of the set $A$, and $\Theta_0$ is the subset of $\Theta$ containing $\theta_0$, the true value of the parameter.

# Asymptotical properties of the method

Now, we can give the main result in this section.

**Theorem.-** Let us assume that, $a.s. - \theta_0$, the posterior distribution converges to the degenerate distribution giving all the mass to the true value of the parameter, $\theta_0$, in the precise sense that, for all $\varepsilon > 0$, there exists $n_0$ such that for all $n \geq n_0$,
$Pr(\Theta_0|x_1, ..., x_n) \geq 1 - \varepsilon$.
Then:

$$\lim_n \int_\Theta D(x_1, ..., x_N, \theta)\pi(\theta|x_1, ..., x_n)d\theta = D(x_1, ..., x_N, \theta_0),\ a.s - \theta_0.$$

# Asymptotical properties of the method

**Remarks on the theorem:**
**(a)** The theorem we have just proved may be seen like a Helly theorem in which:
¶ The space of integration do not need to be the real line.
¶ The function we integrate needs to be a simple function.
¶ The distribution converges to a degenerate distribution.

**(b)** The result has been obtained for the $\chi^2$ discrepancy, but it is equally well suited for any discrepancy taking a finite number of values.

**(c)** It is well known that, under $\theta_0$, $D(x_1, ..., x_N, \theta_0)$ has (approximately) a $\chi^2_{k-1}$ distribution. Therefore, in practice, the posterior expected $\chi^2$ discrepancy will have (approximately) a $\chi^2_{k-1}$ distribution.

# Evaluating the method

Now, we have to evaluate the method by simulation, because a method may seems very natural and sensible (as this is, in my opinion) but may have a poor performance.
We consider the three following Bayesian models:

$$M_1 = \{f_1(x|\theta) \sim N(\theta, \sigma = 1); \pi_1(\theta) \propto 1\}$$

$$M_2 = \{f_2(x|\theta) \sim N(\theta, \sigma = 2); \pi_2(\theta) \propto 1\}$$

$$M_3 = \{f_3(x|\theta) \sim N(\theta, \sigma = 3); \pi_3(\theta) \propto 1\}$$

We will generate random samples with 50 elements each.

# Evaluating the method

100 random samples (with 50 elements each) are generated from a N(0, 1) distribution (model $M_1$).

For each random sample, we evaluate discrepancies $D_1(\mathbf{x})$, $D_2(\mathbf{x})$ and $D_3(\mathbf{x})$ with the models $M_1$, $M_2$ and $M_3$, respectively (these discrepancies will be evaluated by partitioning the interval (0,1) in $k = 4$ subintervals).

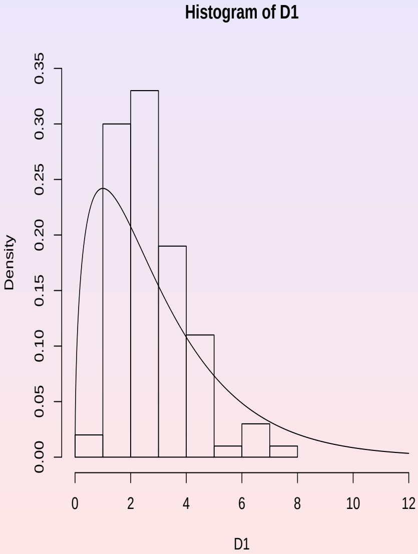
These discrepancies are evaluated (in an approximated way) by simulation.

The performance of the method is excellent: the model $M_1$ (the true model) is chosen in the 100% of the cases.

# Evaluating the method

On the other hand, theoretical results of Section 3 tell that $D_1(\mathbf{x})$ (posterior expected discrepancies under the true model) will have (approximately) a $\chi^2_{k-1} = \chi^2_3$ distribution (because we partitioned the interval (0,1) in k=4 subintervals). The histogram of the values $D_1(\mathbf{x})$ is shown in the next figure; the histogram compares quite well to the density of $\chi^2_3$ distribution. The differences we can observe in the figure are easy to justify because theoretical results of Section 3 are asymptotic.

# Evaluating the method



# Possible extensions: other discrepancy measures

In the previous transparencies, the $\chi^2$ discrepancy has been used for measuring the discrepancy between $(F_i(x_1|\theta), ..., F_i(x_n|\theta))$ and the $U(0,1)$ distribution. Of course, this is a possible discrepancy measure, but not the only one:

**(1) $\chi^2$ discrepancy**

$$D_i^1(\mathbf{x}, \theta) = \sum_{j=1}^{k} \frac{[O_{ij}(\theta) - n(1/k)]^2}{n(1/k)} = \sum_{j=1}^{k} \frac{[O_{ij}(\theta) - (n/k)]^2}{n/k}$$

**(2) Kolmogorov-Smirnov discrepancy**
Let $G_0(y)$ denote the cumulative distribution function of the $U(0,1)$, and let $G_i(y|\theta)$ denote the empirical cumulative distribution function corresponding to the sample $(F_i(x_1|\theta), ..., F_i(x_n|\theta))$. The Kolmogorov-Smirnov discrepancy is defined as usual:

$$D_i^2(\mathbf{x}, \theta) = \sup_{y \in (0,1)} |G_i(y|\theta) - G_0(y)|.$$

# Possible extensions: other discrepancy measures

**(3) $L^1$ discrepancy**
Let $g_0(y)$ denote the density function of the $U(0,1)$, and let $g_i(y|\theta)$ denote some density estimator obtained from the sample $(F_i(x_1|\theta), ..., F_i(x_n|\theta))$. The $L^1$ discrepancy is defined as usual:

$$D_i^3(\mathbf{x},\theta) = \int_0^1 |g_i(y|\theta) - g_0(y)| dy.$$

**(4) Intrinsic discrepancy**
Let us consider again $g_0(y)$ (defined over $\mathcal{X}_0 = (0,1)$) and $g_i(y|\theta)$ (defined over $\mathcal{X}_i \subset (0,1)$). The intrinsic discrepancy is defined as follows (see Bernardo (2005; Test)):

$$D_i^4(\mathbf{x},\theta) \quad = \quad \min\left\{ \int_{\mathcal{X}_i} g_i(y|\theta) \log \frac{g_i(y|\theta)}{g_0(y)} dy \ , \ \int_{\mathcal{X}_0} g_0(y) \log \frac{g_0(y)}{g_i(y|\theta)} dy \right\}$$

# Possible extensions: comparing discrepancy measures

First of all, we must choose a discrepancy measure. For doing that, we can proceed by simulation as follows:

**(1)** Fix $m$ Bayesian models (for instance, the Bayesian models used before; these Bayesian models are similar and it is more difficult to choose the correct model).

**(2)** Simulate many samples (100, 1000, ...) from the Bayesian model $M_i$. Apply to these samples the method described before, for the four discrepancy measures we have just defined, recording the percentage of correct classification with each discrepancy.

**(3)** Repeat Step (2) for each model $M_i$, $i = 1, ..., m$. Construct a table of double entry with the percentages of correct classification with each discrepancy measure and each model.

**(4)** Finally, look for the discrepancy measure having the best performance.

# Possible extensions: assessing the discrepancy

Once we have chosen the discrepancy measure we are going to use, we can apply the method described in Section 2 for choosing the Bayesian model we thing is more suitable for our data. It is important to remark that the model we choose must not be understood as the true model (nobody knows the true model), but as the best representation we can find among several Bayesian models.

Therefore, suppose we have decided to use a discrepancy measure and, then, we have chosen a Bayesian model as the best representation for our data. The posterior expected discrepancy between our data $\mathbf{x}$ and the model $M_i$ is just a number, $D_i^j(\mathbf{x})$. It is important to decide if this number indicates a small discrepancy or a large discrepancy:

If the discrepancy is small, the model we have chosen is a good representation for our data.

If the discrepancy is large, the model we have chosen is not a very good representation for our data.

# Possible extensions: assessing the discrepancy

For deciding if the discrepancy between our data $\mathbf{x}$ and the model $M_i$ is large or small we may proceed as follows:

**(1)** Simulate many samples (100, 1000, ...) from the Bayesian model we have chosen.
**(2)** Compute the posterior expected discrepancies between these samples and the Bayesian model.
**(3)** Construct the histogram corresponding to these posterior expected discrepancies.
**(4)** Compare $D_i^j(\mathbf{x})$ to this histogram.