A Bayesian test for the intraclass correlation coefficient. Computational implementation

Verónica A. González-López

Departamento de Estatística, Universidade Estadual de Campinas, Brazil

Julio M. Singer

Nelson I. Tanaka

Antonio C. Pedroso-de-Lima

Departamento de Estatística, Universidade de São Paulo, Brazil

1 Introduction

We illustrate the use of a code written in the R language. Previously we indicate the computational requirements, and we give a general description of the R-code subroutines. The computational requirements for the use of the code are

1. the sofware "R" available in http://www.r-project.org/
   with the R-package "R2WinBUGS" (install dependencies: "coda" package and "lattice" package);

2. the sofware "WinBugs" http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

Our R-code is available in http: www.ime.unicamp.br/∼veronica/intracorr/
It is composed by three main R-functions:

1. rhobetamodel;

2. posteriorprob;

3. rhohierarchicalmodel.

The three functions are in the same script, "rhobetacreatemodelv9.txt"

2 Code

In this section we describe each function,

1. "rhobetamodel": build the model and save it as "rhobetamodel.bug" in a file located in the working directory.

USAGE: run the script: "rhobetacreatemodelv9.txt"

MODEL: * General Description: Let be the model

$$y_{ij} = \mu + a_i + e_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, m, \tag{2.1}$$

where $a_i \sim \mathcal{N}(0, \sigma_a^2)$ and $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$ are independent. i represents each experimental unit, j represents each replication in the experimental unit. The dependence between two measurements conducted on the same experimental unit can be quantified by the intraclass correlation coefficient $\rho = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$. We adopt

$$\mu \sim \mathcal{N}(\mu_o, \sigma_o^2) \tag{2.2}$$

and the inverse gamma (IG) distributions for σ_a^2 and σ_e^2

$$\sigma_a^2 \sim \text{IG}(\beta, 1),\ \sigma_e^2 \sim \text{IG}(\alpha, 1).$$

The last two prior choices assure that

$$\rho \sim \text{Beta}(\alpha, \beta). \tag{2.3}$$

* The conditionally conjugate distributions. We denote the conditional densities by $[X|Y]$, the Gibbs sampling approach is based on the following conditional densities

$$[\boldsymbol{\theta}|\boldsymbol{y}, \mu, \sigma_a^2, \sigma_e^2] = \mathcal{N}_n\Big(\frac{m\sigma_a^2}{m\sigma_a^2 + \sigma_e^2}\overline{\boldsymbol{y}} + \frac{\sigma_e^2}{m\sigma_a^2 + \sigma_e^2}\mu\mathbf{1}_n, \frac{\sigma_a^2\sigma_e^2}{m\sigma_a^2 + \sigma_e^2}\boldsymbol{I}_n\Big)$$

$$[\mu|\boldsymbol{y}, \boldsymbol{\theta}, \sigma_a^2, \sigma_e^2] = [\mu|\sigma_a^2, \boldsymbol{\theta}] = \mathcal{N}\Big(\frac{\sigma_a^2\mu_o + n\sigma_o^2\overline{\theta}}{n\sigma_o^2 + \sigma_a^2}, \frac{\sigma_o^2\sigma_a^2}{n\sigma_o^2 + \sigma_a^2}\Big)$$

$$[\sigma_e^2|\boldsymbol{y}, \mu, \boldsymbol{\theta}, \sigma_a^2] = [\sigma_e^2|\boldsymbol{y}, \boldsymbol{\theta}] = \text{IG}\Big(\alpha + \frac{nm}{2}, 1 + \frac{\sum_{j=1}^m \sum_{i=1}^n (y_{ij} - \theta_i)^2}{2}\Big)$$

$$[\sigma_a^2|\boldsymbol{y}, \mu, \boldsymbol{\theta}, \sigma_e^2] = [\sigma_a^2|\mu, \boldsymbol{\theta}] = \text{IG}\Big(\beta + \frac{n}{2}, 1 + \frac{\sum_{i=1}^n (\theta_i - \mu)^2}{2}\Big)$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$, $\overline{\theta} = n^{-1}\sum_{i=1}^n \theta_i$, $\overline{\boldsymbol{y}} = (\overline{y}_1, \cdots, \overline{y}_n)$, $\overline{y}_i = m^{-1}\sum_{j=1}^m y_{ij}$, $\mathbf{1}_n$ is an $n \times 1$ vector with all elements equal to 1 and $\boldsymbol{I}_n$ is the identity matrix of order n.

Applying the Gibbs sampling procedure in this setting, we obtain the empirical densities of σ_e^2 and σ_a^2 and, consequently, the empirical density of ρ.

VALUES: The function returns

rhobetamodel.bug: a file saved in the working directory.

2. "posteriorprob": computes the empirical probability of the event $\{\rho \leq c\}$.

USAGE: posteriorprob(sample,cust)

ARGUMENTS: The arguments of this function are:

sample: a sample of realizations of ρ;

cust: the desired value of cust c (it could be a vector).

VALUES: The function returns

cust: the vector of custs;

Pcust: empirical estimation of the probability of $\{\rho \leq c\}$.

3. "rhohierarchicalmodel": it is a routine which calls the R function "bugs" from the "R2WinBUGS" package.

USAGE: rhohierarchicalmodel(subs, cust,file,n,m,alpha,beta,muo,tauo)

ARGUMENTS: The arguments of this function are:

subs: integer ($\leq$ n). Number of experimental units that we want to use. If subs="null" it is used all the dataset (file argument). Otherwise "subs" random experimental units are selected without replacement, from the dataset;

cust: vector of c values;

file: dataset is a string of length equal to n $\times$ m. It displays all the experimental units in one line (see the applications);

n: number of experimental units in the dataset;

m: replications;

alpha: (α) parameter of the beta distribution in ρ. The default is alpha=1;

beta: (β) parameter of the beta distribution in ρ. The default is beta=1;

muo: (μ_o) location parameter of the normal in μ where $\mu \sim \mathcal{N}(\mu_o, 1/\tau_o)$. The default is muo=0;

tauo: ($\tau_o = \frac{1}{\sigma_o^2}$) precision parameter of the normal in μ where $\mu \sim \mathcal{N}(\mu_o, 1/\tau_o)$. The default is tauo=1e-06.

VALUES: The function returns

- data: it is the observed data (or subsample) selected. If subs="null" it is equal to the dataset;
- newsam: it is a string with the selected experimental units;
- simulation: inference from Bugs model;
- postrhosample: the posterior sample of ρ;
- PProbrho: posterior probability of $\rho \leq c$.

DETAILS: The code calls the R-function "bugs", from the"R2WinBUGS" package. The main specifications are:

i. number of Markov chains (n.chains parameter) equal to 3;
ii. number of total iterations per chain (n.chain parameter) equal to 500000;
iii. thinning rate (n.thin parameter) equal to 50.

More details about "bugs" function could be obtained from
http://cran.r-project.org/web/packages/R2WinBUGS/index.html

3 Example

A practical example involves the estimation of the average amount of oil contained in lemon juice. Each of 60 samples was obtained from a batch of lemon juice and divided into three portions (haphazardly labeled A, B and C) each of which was analyzed with respect to the amount of oil (kg/ton). The data for the 60 triplicates are displayed in Table 1.

3.1 The problem

Singer et al. (2007) show that for fixed n, a reduction of $100r\%$, $0 < r < 1$, in the length of a confidence interval for μ occurs when ρ is lesser than $c = [m(1-r)^2 - 1]/(m-1)$. Under this setup, the decision to carry out the measurements in m-plicate depends on whether the hypothesis

$$H_0 : \rho \leq c \tag{3.4}$$

is rejected or not.

We need to compute the posterior probability of $\rho \leq c$ for same specified value of c, according to the desired reduction r. If the probability computed is high enough, we decide in favour of taking m-plicates. In our application $m = 3$.

Table 1: Amount of oil in lemon juice (kg of oil / ton of juice)

Sample	A	B	C	Sample	A	B	C	Sample	A	B	C
1	5.29	5.10	5.13	21	5.66	5.64	5.46	41	4.90	4.75	4.84
2	5.34	5.34	5.27	22	5.62	5.49	5.73	42	4.88	4.57	4.54
3	5.20	5.07	5.08	23	5.36	5.33	5.46	43	4.80	4.82	4.94
4	5.43	5.38	5.36	24	4.91	5.01	4.86	44	5.29	5.29	5.10
5	5.18	5.03	5.02	25	5.28	5.35	5.14	45	4.53	4.66	4.63
6	5.33	5.07	5.07	26	5.02	4.80	4.64	46	4.39	4.49	4.39
7	5.16	5.40	5.23	27	5.57	5.54	5.29	47	4.50	4.51	4.52
8	4.91	5.10	4.84	28	5.09	5.22	4.95	48	4.82	4.80	4.66
9	5.07	5.01	4.87	29	5.58	5.45	5.32	49	5.06	4.96	4.94
10	4.85	4.76	4.54	30	5.04	4.90	4.94	50	5.20	4.97	5.11
11	5.31	5.42	5.52	31	5.79	5.65	5.58	51	5.63	5.75	5.63
12	5.12	5.40	5.27	32	5.46	5.38	5.36	52	5.38	5.51	5.14
13	5.29	5.47	5.13	33	5.21	5.20	5.07	53	5.37	5.06	5.13
14	5.04	5.09	4.98	34	4.84	4.98	4.91	54	5.06	5.20	5.07
15	5.11	5.11	5.11	35	5.27	5.11	5.25	55	5.15	5.32	4.99
16	4.96	5.07	4.94	36	5.06	5.08	4.89	56	4.74	4.74	4.64
17	5.36	5.06	5.10	37	5.10	5.24	5.05	57	4.48	4.4	4.37
18	5.36	5.40	5.33	38	5.32	5.51	5.22	58	4.26	4.12	4.37
19	5.39	5.13	5.34	39	4.80	4.70	4.58	59	4.46	4.37	4.62
20	5.49	5.60	5.28	40	5.18	4.83	4.80	60	5.20	4.93	5.07

3.2 Application

First, run the code available in http: www.ime.unicamp.br/∼veronica/intracorr/ in your R console. Save your dataset in your working directory. For our example, save the file “lemondata” available in http: www.ime.unicamp.br/∼veronica/intracorr/ in your working directory.

Digit this command
rhohierarchicalmodel(subs="null",c(0.72,0.74,0.77,0.80,0.83,0.85,0.88,0.90),
"lemondata",60,3)
End
About the specifications:

subs ="null" (we use all the data)

cust =c(0.72,0.74,0.77,0.80,0.83,0.85,0.88,0.90)

n =60 (we have 60 experimental units)

m =3 (triplicates)

muo ="default value"

tauo ="default value"

alpha ="default value"

beta ="default value".

DETAILS: We use the data in Table 1 ("lemondata") to compute the posterior probability for the event $\{\rho \leq c\}$ and different choices of c using the Bayesian method described by the model (2.1), with a noninformative prior on μ, given by equation (2.2), $\mu_o = 0$, $\sigma_o^2 = 1e + 06$ and an uniform (i.e. noninformative) prior distribution for ρ, given by equation (2.3) with $\alpha = \beta = 1$.

RESULTS: The results are presented in Table 2; the prior and posterior sample distributions for ρ are displayed graphically in Figure 1. There, a Gibbs sample of size 15000 was used.

Table 2: Posterior probability of $\rho \leq c$ using an uniform prior distribution

c	$\pi(\rho \leq c \vert data)$	% reduction in length of confidence interval
0.72	0.010	10%
0.74	0.027	9%
0.77	0.111	8%
0.80	0.328	7%
0.83	0.660	6%
0.85	0.860	5%
0.88	0.985	4%
0.91	0.999	3%

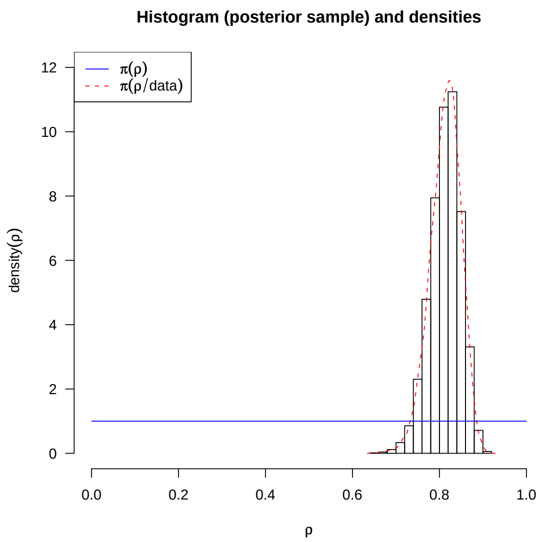

Figure 1: Uniform prior distribution $\pi(\rho)$ and posterior sample distribution $\pi(\rho|data)$

3.3 Application

CODE 1:
```
#### Digit this command
rhohierarchicalmodel(subs="null",c(0.72,0.74,0.77,0.80,0.83,0.85,0.88,0.90),
"lemondata",60,3,2,2)
#### End
```

About the specifications:

subs =“null”

cust =c(0.72,0.74,0.77,0.80,0.83,0.85,0.88,0.90)

n =60

m =3

muo =“default value”

tauo =“default value”

alpha =2

beta =2.

CODE 2:

```
#### Digit this command
rhohierarchicalmodel(subs=“null”,c(0.72,0.74,0.77,0.80,0.83,0.85,0.88,0.90),
“lemondata”,60,3,2,10)
#### End
```

About the specifications:

subs =“null”

cust =c(0.72,0.74,0.77,0.80,0.83,0.85,0.88,0.90)

n =60

m =3

muo =“default value”

tauo =“default value”

alpha =2

beta =10.

CODE 3:

```
#### Digit this command
rhohierarchicalmodel(,subs=“null”,c(0.72,0.74,0.77,0.80,0.83,0.85,0.88,0.90),
“lemondata”,60,3,10,2)
#### End
```

About the specifications:

subs =“null”

cust =c(0.72,0.74,0.77,0.80,0.83,0.85,0.88,0.90)

n =60

m =3

$$
\begin{aligned}
\text{muo} &= \text{“default value”}\\
\text{tauo} &= \text{“default value”}\\
\text{alpha} &= 10\\
\text{beta} &= 2.
\end{aligned}
$$

RESULTS: Graphical representations of the corresponding prior distribution and posterior empirical distribution based on Gibbs samples of size 15000 are displayed in Figure 2 (left-up), Figure 2 (right-up) and Figure 2 (down), respectively.

DETAILS: We considered Beta(2,2) prior distribution for ρ in the CODE 1, Figure 2 (left-up); Beta (2,10) prior distribution for ρ in the CODE 2, Figure 2 (right-up) and Beta (10,2) prior distribution for ρ in the CODE 3, Figure 2 (down).

The posterior probability of $\rho \leq c$ for different choices of c using the Bayesian method described in the model (2.1), equations (2.2) and (2.3) with different beta prior distributions for ρ are presented in Table 3.

Table 3: Posterior probability of $\rho \leq c$ using beta prior distributions

c	$\pi(\rho \leq c \mid data)$			% reduction in length of confidence interval
	Beta(2,2)	Beta(2,10)	Beta(10,2)	
0.72	0.011	0.125	0.001	10%
0.74	0.030	0.244	0.007	9%
0.77	0.129	0.518	0.044	8%
0.80	0.364	0.797	0.180	7%
0.83	0.697	0.958	0.488	6%
0.85	0.877	0.992	0.731	5%
0.88	0.986	1.000	0.961	4%
0.91	0.999	1.000	0.996	3%

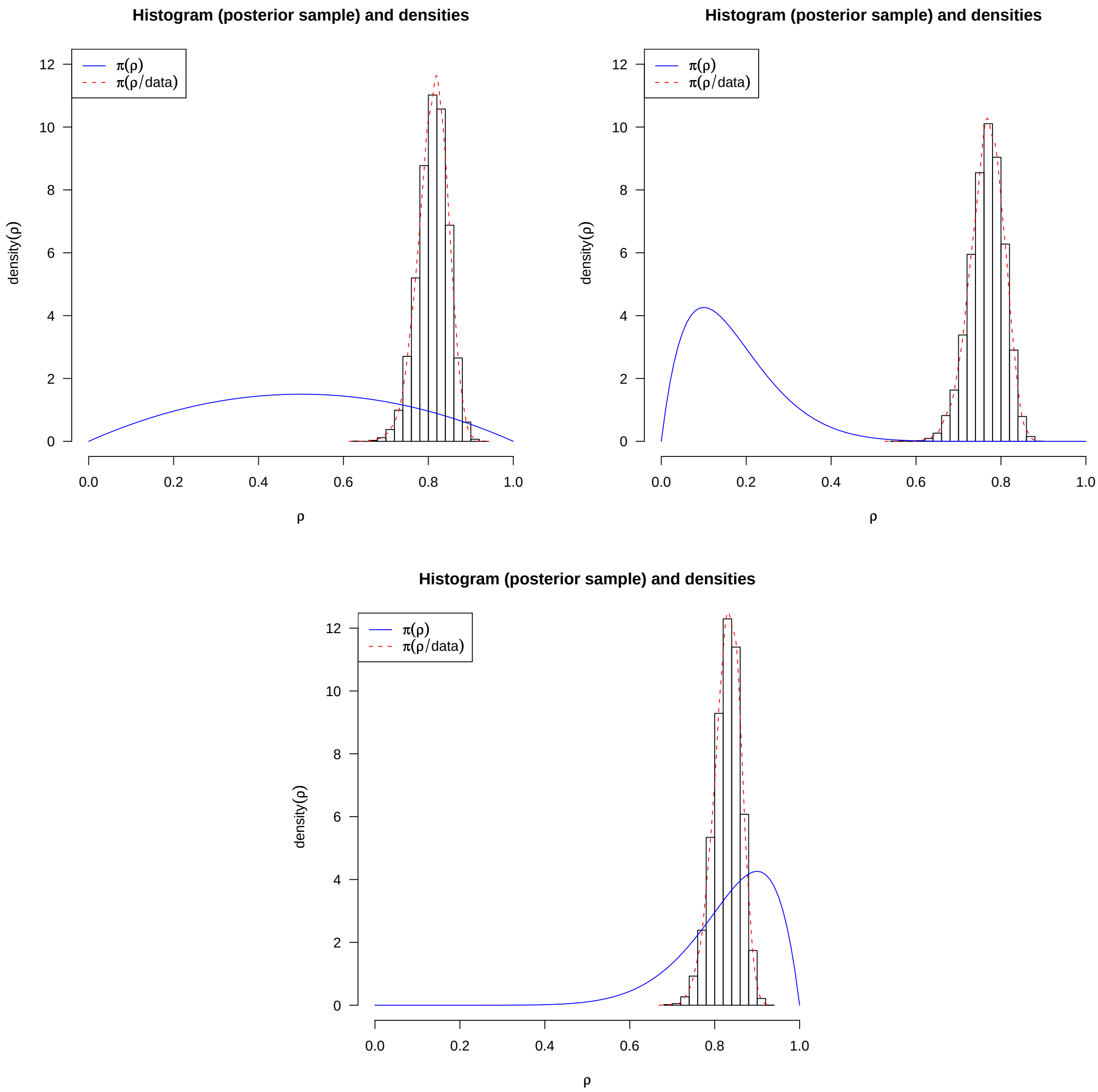

Figure 2: (On the left-up) Beta(2,2) prior distribution $\pi(\rho)$ and posterior sample distribution $\pi(\rho|data)$. (On the right-up) Beta(2,10) prior distribution $\pi(\rho)$ and posterior sample distribution $\pi(\rho|data)$. (Down) Beta(10,2) prior distribution $\pi(\rho)$ and posterior sample distribution $\pi(\rho|data)$

3.4 Application

Using a subsample of size 5 from Table 1 and the "default" options in the function "rhohierarchicalmodel" (see the Application 3.2),

Run this command
M<-rhohierarchicalmodel(subs=5,c(0.72,0.74,0.77,0.8,0.83,0.85,0.88,0.9),
“lemondata”,60,3)
End

The subsample selected was

Table 4: M$data

	[, 1]	[, 2]	[, 3]
[1,]	5.016683	4.796618	4.642930
[2,]	4.902320	4.749127	4.843544
[3,]	5.036630	5.087199	4.978621
[4,]	5.358516	5.401905	5.326429
[5,]	5.369298	5.056737	5.134903

We obtain the Figure 3 and the results could be found in Table 5.

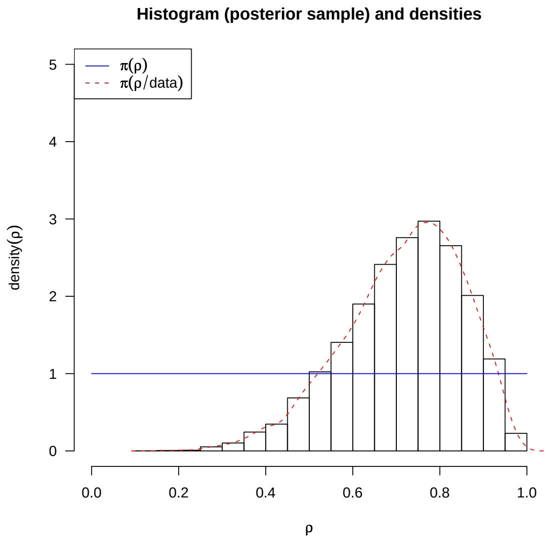

Figure 3: Uniform prior distribution $\pi(\rho)$ and posterior sample distribution $\pi(\rho|data)$

We use the same subsample, given by Table 4 with different prior distributions (see Application 3.3). First we save the subsample in the working directory,
####
sample<-c(t(M$data));

Table 5: Posterior probability of $\rho \leq c$ using an uniform prior distribution

c	$\pi(\rho \leq c \mid data)$	% reduction in length of confidence interval
0.26	0.001	29%
0.51	0.081	18%
0.72	0.462	10%
0.74	0.518	9%
0.77	0.606	8%
0.80	0.696	7%
0.83	0.780	6%
0.85	0.829	5%
0.88	0.894	4%
0.91	0.944	3%

```
# we obtain
# sample
# [1] 5.016683 4.796618 4.642930 4.902320 4.749127 ...
write(sample,file="sample")
####
```

We can now run the next codes putting prior Beta(2,2) in ρ,

CODE 4:
```
####
rhohierarchicalmodel(subs="null",c(0.72,0.74,0.77,0.8,0.83,0.85,0.88,0.9),
"sample",5,3,2,2);
####
```

Prior Beta(2,10) in ρ,

CODE 5:
```
####
rhohierarchicalmodel(subs="null",c(0.72,0.74,0.77,0.8,0.83,0.85,0.88,0.9),
"sample",5,3,2,10)
####
```

Prior Beta(10,2) in ρ

CODE 6: ####

```
rhohierarchicalmodel(subs="null",c(0.72,0.74,0.77,0.8,0.83,0.85,0.88,0.9),
"sample",5,3,10,2)
####
```

The results are presented in the Table 6 and Figure 4.

Table 6: Posterior probability of $\rho \leq c$ using beta prior distributions

c	$\pi(\rho \leq c \mid data)$			% reduction in length of confidence interval
	Beta(2,2)	Beta(2,10)	Beta(10,2)	
0.26	0.002	0.131	<1e-04	29%
0.51	0.104	0.845	0.002	18%
0.72	0.575	0.998	0.113	10%
0.74	0.633	1.000	0.154	9%
0.77	0.719	1.000	0.238	8%
0.80	0.800	1.000	0.346	7%
0.83	0.870	1.000	0.480	6%
0.85	0.908	1.000	0.578	5%
0.88	0.950	1.000	0.731	4%
0.91	0.981	1.000	0.868	3%

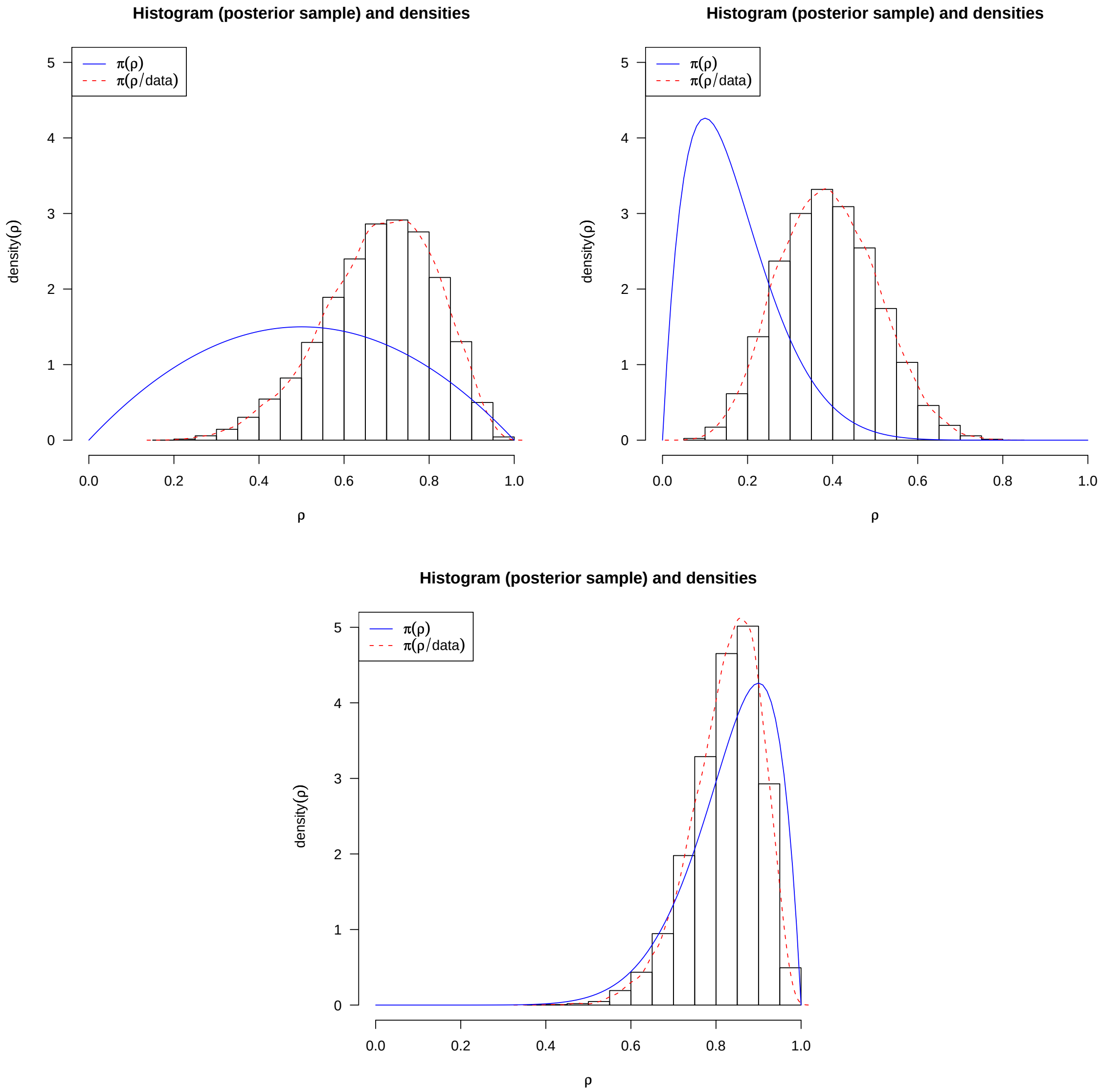

Figure 4: (On the left-up) Beta(2,2) prior distribution $\pi(\rho)$ and posterior sample distribution $\pi(\rho|data)$. (On the right-up) Beta(2,10) prior distribution $\pi(\rho)$ and posterior sample distribution $\pi(\rho|data)$. (Down) Beta(10,2) prior distribution $\pi(\rho)$ and posterior sample distribution $\pi(\rho|data)$.

4 Prior elicitation

We can use as a reference for the specifications of α and β, the relation given by the Table 7. A prior like Beta(2,10) (or Beta(α,β) with $\alpha << \beta$) on ρ means that from the prior knowledge the measurements on the same experimental unit could be considered almost independent (see for example the prior in the Figure 2, right-up). In opposition, the prior given by Beta(10,2) (or Beta(α, β) with $\alpha >> \beta$) on ρ means that from a prior conception we believe that the measurements on the same experimental unit could be considered as being strongly dependent (see Figure 2, down). In addition, several degrees of dependence could be performed taking a prior like Beta(α, β) on ρ and depending on the relation between the hyperparameters α and β, for example a non informative prior is adopted when $\alpha = \beta = 1$ (see Figure 1). As a consequence our approach allows several prior conditions that can represent the expert's uncertainty about ρ.

Table 7: Prior distribution on ρ, Beta(α, β)

α, β	Prior of σ_a^2	Prior of σ_e^2	about the prior on ρ
$\alpha = \beta = 1$	IG(1,1)	IG(1,1)	U(0,1): equiprobability in [0,1]
$\alpha = \beta > 1$	IG(α,1)	IG(α,1)	symmetric around 0.5 and concave
$\alpha = \beta < 1$	IG(α,1)	IG(α,1)	symmetric around 0.5 and convex
$1 < \alpha < \beta$	IG(β,1)	IG(α,1)	asymmetric, mode<0.5
$\alpha > \beta > 1$	IG(β,1)	IG(α,1)	asymmetric, mode>0.5

References

Burch, B. D., and Harris, I. R. (1999). Bayesian Estimators of the Intraclass Correlation Coefficient in the One-Way Random Effects Model. **Communications in Statistics: Theory and Methods, 28,** 1247-1272.

Gelfand, A. and Smith, A. (1990). Sampling-Based Approaches to Calculating Marginal Densities. **Journal of the American Statistical Association, 85,** 398-409.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). **Bayesian Data Analysis** (Texts in Statistical Science). London: Chapman & Hall/CRC.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. **IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,** 721-741.

Kass, R.W. and Raftery, A.E. (1995). Bayes factors. **Journal of the American Statistical Association, 90,** 773-795.

R Development Core Team (2009). **R: A language and environment for statis-**

tical computing (R Foundation for Statistical Computing). Vienna: Austria. URL http://www.R-project.org.

Robert, C. (1996). **The Bayesian Choice. A Decision-Theoretic Motivation**. New York: Springer-Verlag.

Searle, S.R. (1971). **Linear Models.** New York: John Wiley & Sons.

Singer, J.M., Pedroso de Lima, A.C., Tanaka, N.I. and González-López, V.A. (2007). To triplicate or not to triplicate? **Chemometrics and Intelligent Laboratory Systems, 86,** 82-85.

Sturtz, S., Ligges, U. and Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. **Journal of Statistical Software, 12,** 1-16.