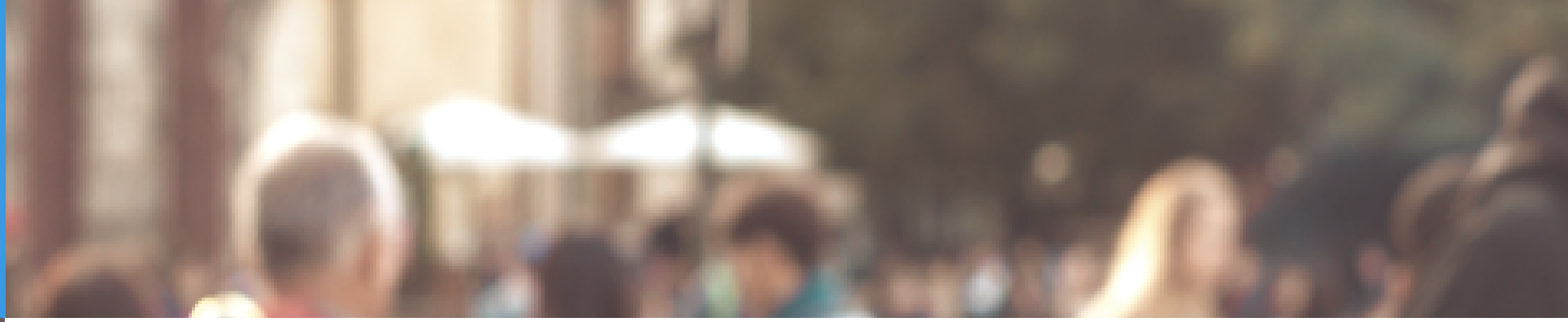


Elements of Structural Modelling

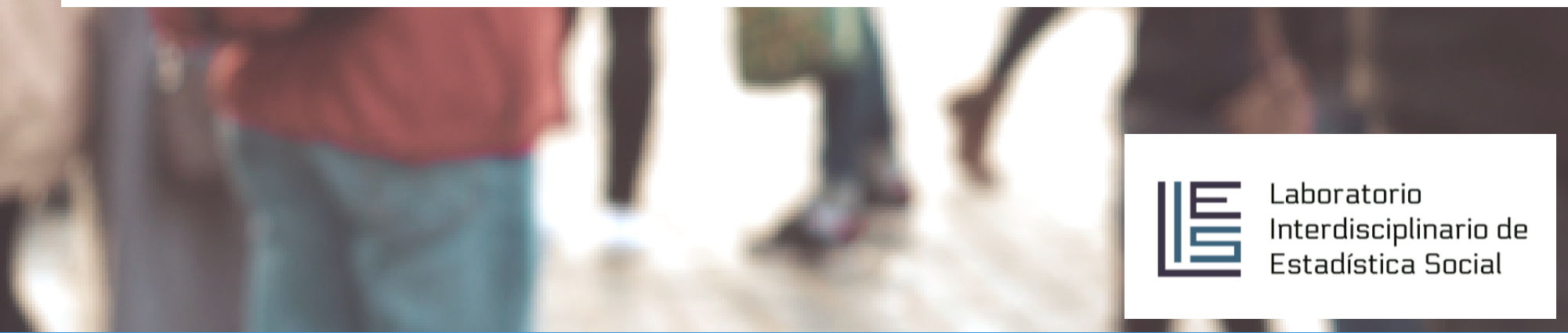
Bayesian Model

Cursillo de Verano 2020, Facultad de Matemáticas, UC

Ernesto San Martín

Laboratorio Interdisciplinario de Estadística Social LIES, UC
The Economics School of Louvain, Université Catholique de Louvain, Belgium

January 2021



Construction of a Bayesian model

- We consider a statistical model:

$$\{p(x \mid a) : a \in A\}$$

- We consider that the density (distribution) $p(x \mid a)$ is a function of $a \in A$: for each x,

$$a \longmapsto p(x \mid a)$$

- The statistical model describes the way on which the data are in fact, or supposed to be, generated by the family of distributions $\{p(x \mid a) : a \in A\}$.

Construction of a Bayesian model

- We consider a statistical model:

$$\{p(x \mid a) : a \in A\}$$

- We consider that the density (distribution) $p(x \mid a)$ is a function of $a \in A$: for each x,

$$a \longmapsto p(x \mid a)$$

- The statistical model describes the way on which the data are in fact, or supposed to be, generated by the family of distributions $\{p(x \mid a) : a \in A\}$.
- Are we agree on that?

Construction of a Bayesian model

- We define a probability distribution on the parameter space A, denoted as $\mu(a)$.
- Therefore, we have $\{p(x \mid a) : a \in A\}$ and μ defined on $(A, \mathcal{A})$, where $\mathcal{A}$ is the σ-algebra of subsets of the parameter space A.
- These two ingredients allow to define/construct a unique probability distribution on the product space "observations $\times$ parameters" defined as

$$p(x, a) = p(x \mid a)\mu(a).$$

- μ is called *prior distribution*.

Construction of a Bayesian model

- Therefore, a Bayesian model is a unique probability distribution defined on the product space "observations $\times$ parameters".

Construction of a Bayesian model

- Therefore, a Bayesian model is a unique probability distribution defined on the product space "observations $\times$ parameters".
- The unique joint distribution $p(x, a)$ can also be decomposed as

$$p(x, a) = p(x)\mu(a \mid x),$$

where $\mu(a \mid x)$ is called *posterior distribution* and $p(x)$ is called *predictive distribution*.

- Thus,

$$\begin{aligned} p(x, a) &= p(x \mid a)\mu(a) \\ &= \mu(a \mid x)p(x). \end{aligned}$$

- Typically, the inference is based on the posterior distribution $\mu(x \mid a)$.

Example 1

- Let us consider the following example developed by Poirier (1998, Sectio 3.2) and discussed in San Martín et al. (2013): let $(Y \mid \psi, \lambda) \sim \mathcal{N}(\psi, \sigma_1^2)$ be the likelihood or sampling distribution, and let

$$\begin{pmatrix} \psi \\ \lambda \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_2^2 + 1 & 1 \\ 1 & 1 \end{pmatrix} \right)$$

  be the prior specification. For simplicity, it is assumed that σ_1^2 and σ_2^2 are known constants.

- It is clear that ψ is identified because the mapping $\psi \longmapsto \mathcal{N}(\psi, \sigma_1^2)$ is injective.

- Similarly, λ is unidentified and therefore it can not be said something about it.

Example 1

- It is nevertheless possible to compute the posterior distribution of λ, which is given by

$$(\lambda \mid Y) \sim \mathcal{N}\left(\frac{Y}{\sigma_1^2 + \sigma_2^2 + 1}, \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2 + 1}\right).$$

  This distribution provides information about λ.

Example 1

- It is nevertheless possible to compute the posterior distribution of λ, which is given by

$$(\lambda \mid Y) \sim \mathcal{N}\left(\frac{Y}{\sigma_1^2 + \sigma_2^2 + 1}, \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2 + 1}\right).$$

  This distribution provides information about λ.

- Has the situation actually been saved?

Example 1

- It is nevertheless possible to compute the posterior distribution of λ, which is given by

$$(\lambda \mid Y) \sim \mathcal{N}\left(\frac{Y}{\sigma_1^2 + \sigma_2^2 + 1}, \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2 + 1}\right).$$

  This distribution provides information about λ.

- Has the situation actually been saved?

- This question is meaningful because the sampling process is fully characterized by ψ in the sense that

$$Y \perp\!\!\!\perp \lambda \mid \psi.$$

  So, what are we updating about the sampling process?

Example 1

- It is nevertheless possible to compute the posterior distribution of λ, which is given by

$$(\lambda \mid Y) \sim \mathcal{N}\left(\frac{Y}{\sigma_1^2 + \sigma_2^2 + 1}, \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2 + 1}\right).$$

  This distribution provides information about λ.

- Has the situation actually been saved?

- This question is meaningful because the sampling process is fully characterized by ψ in the sense that

$$Y \perp\!\!\!\perp \lambda \mid \psi.$$

  So, what are we updating about the sampling process?

- Where is the trap?

Example 1

- The posterior distribution of λ is fully characterized by the two first moment because it is a normal distribution.

Example 1

- The posterior distribution of λ is fully characterized by the two first moment because it is a normal distribution.
- Well, the two first posterior moments are a function of the identified parameter ψ:

$$E[\lambda \mid Y] = \frac{1}{\sigma_2^2 + 1} E[\psi \mid Y], \qquad V(\lambda \mid Y) = \frac{1}{\sigma_2^2 + 1} \left[\frac{1}{\sigma_2^2 + 1} V(\psi \mid Y) + \sigma_2^2 \right].$$

Example 1

- But "we are Bayesian", so ψ and λ are endowed of a probability distribution and then we can marginalize the sampling distribution w.r.t ψ (by the way, this is the solution to the nuisance parameter problem from a Bayesain perspective).

Example 1

- But "we are Bayesian", so ψ and λ are endowed of a probability distribution and then we can marginalize the sampling distribution w.r.t ψ (by the way, this is the solution to the nuisance parameter problem from a Bayesain perspective).
- The induced sampling process is now described by the sampling distribution

$$(Y \mid \lambda) \sim \mathcal{N}(\lambda, \sigma_1^2 + \sigma_2^2),$$

  where the prior distribution is given by $\lambda \sim \mathcal{N}(0, 1)$.
- In this case, λ is identified ... situation saved!

Example 1

- Where is the trap?

Example 1

- Where is the trap?
- Which is the statistical model,

$$(Y \mid \psi, \lambda) \sim \mathcal{N}(\psi, \sigma_1^2)$$

or

$$(Y \mid \lambda) \sim \mathcal{N}(\lambda, \sigma_1^2 + \sigma_2^2)$$

?

Example 1

- Where is the trap?
- Which is the statistical model,

$$(Y \mid \psi, \lambda) \sim \mathcal{N}(\psi, \sigma_1^2)$$

or

$$(Y \mid \lambda) \sim \mathcal{N}(\lambda, \sigma_1^2 + \sigma_2^2)$$

?

- Is it true that we accept the definition of a statistical model as the one that describes the data we are going to analyze?

Remark

- When Bayesian statisticians change arbitrarily the model specification with involved constructions, I feel that modelling is arbitrary rather than rigorous.
- Science is not objective, is rigorous.

Example 2

- Let us consider an example widely discussed in the Bayesian literature: Poirier (1998), Carlin and Loui (2000), Xie and Carlin (2006) and Kass et al. (1998).

Example 2

- Let us consider an example widely discussed in the Bayesian literature: Poirier (1998), Carlin and Loui (2000), Xie and Carlin (2006) and Kass et al. (1998).
- The data generating process is characterized by the sampling distribution

$$(Y_i \mid a, b) \overset{\text{iid}}{\sim} \mathcal{N}(a + b, \sigma_Y^2),$$

where σ_Y^2 is known. The parameters of interest are (a, b), but it is not identified because the mapping $(a, b) \longmapsto \mathcal{N}(a + b, \sigma_Y^2)$ is not injective.

- The prior specification is given by

$$(a \mid \mu_a, \sigma_a^2) \sim \mathcal{N}(\mu_a, \sigma_a^2), \quad (b \mid \mu_b, \sigma_b^2) \sim \mathcal{N}(\mu_b, \sigma_b^2), \qquad a \perp\!\!\!\perp b \mid \mu_a, \mu_b, \sigma_a^2, \sigma_b^2.$$

Example 2

- The Bayesian model is, consequently, characterized by a unique probability distribution given by

$$\left(\begin{array}{c|c} \begin{array}{c} Y \\ a \\ b \end{array} & \omega \end{array} \right) \sim \mathcal{N}_3 \left(\left(\begin{array}{c} \mu_a + \mu_b \\ \mu_a \\ \mu_b \end{array} \right), \left(\begin{array}{ccc} \sigma_a^2 + \sigma_b^2 + \sigma_Y^2 & \sigma_a^2 & \sigma_b^2 \\ \sigma_a^2 & \sigma_a^2 & 0 \\ \sigma_b^2 & 0 & \sigma_b^2 \end{array} \right) \right),$$

where $\omega = (\mu_a, \mu_b, \sigma_a^2, \sigma_b^2)$ are the hyper-parameters.

Example 2

- The Bayesian model is, consequently, characterized by a unique probability distribution given by

$$\left(\begin{array}{c|c} Y & \\ a & \omega \\ b & \end{array} \right) \sim \mathcal{N}_3 \left(\left(\begin{array}{c} \mu_a + \mu_b \\ \mu_a \\ \mu_b \end{array} \right), \left(\begin{array}{ccc} \sigma_a^2 + \sigma_b^2 + \sigma_Y^2 & \sigma_a^2 & \sigma_b^2 \\ \sigma_a^2 & \sigma_a^2 & 0 \\ \sigma_b^2 & 0 & \sigma_b^2 \end{array} \right) \right),$$

  where $\omega = (\mu_a, \mu_b, \sigma_a^2, \sigma_b^2)$ are the hyper-parameters.

- The parameters of interest are updated through their posterior distributions

$$(a \mid Y, \omega) \sim \mathcal{N}\left(\frac{(\sigma_b^2+\sigma_Y^2)\mu_a+\sigma_a^2(Y-\mu_b)}{\sigma_a^2+\sigma_b^2+\sigma_Y^2}, \sigma_a^2 \left[1 - \frac{\sigma_a^2}{\sigma_a^2+\sigma_b^2+\sigma_Y^2} \right] \right),$$

$$(b \mid Y, \omega) \sim \mathcal{N}\left(\frac{(\sigma_a^2+\sigma_Y^2)\mu_b+\sigma_b^2(Y-\mu_a)}{\sigma_a^2+\sigma_b^2+\sigma_Y^2}, \sigma_b^2 \left[1 - \frac{\sigma_b^2}{\sigma_a^2+\sigma_b^2+\sigma_Y^2} \right] \right).$$

Example 2

- However, these posterior distributions are functions of the posterior distribution of the identified parameter since

$$E[\, a \mid Y \,] = \eta_{a,b} + \frac{\sigma_a^2}{\sigma_b^2 + \sigma_a^2} E[\, a + b \mid Y \,],$$

where $\eta_{a,b} \doteq \frac{\sigma_b^2}{\sigma_b^2+\sigma_a^2}\mu_a - \frac{\sigma_a^2}{\sigma_b^2+\sigma_a^2}\mu_b$; and similarly for $E[\, b \mid Y \,]$

Example 3

- Wechsler, Izbicki and Esteves (2013), A Bayesian look at nonidentifiability: A simple example.
- They consider a problem proposed by Ross (2009):

  "A ball is in any one of n boxes and is in the ith box with probability p_i. If the ball is in box i, a search of that box will uncover it with probability α_i. Show that the conditional probability that the ball is in box j, given that a search of box i did not uncover it, is

$$\frac{p_j}{1-\alpha_i p_i} \text{ if } j \neq i, \quad \text{and} \quad \frac{(1-\alpha_i)p_i}{1-\alpha_i p_i} \text{ if } j = i."$$

Example 3

- The problem rephrased:

  "A missing plane has crashed in one of n possible regions. A search for the plane is then performed in region i; p_i is the probability that the plane has gone down in region i and α_i the probability that the plane will be found in a search of the i-th region when it is, in fact, in that region. We consider $0 < p_i < 1$ and $0 < \alpha_i < 1$ to make the problem reasonable and avoid trivialities."

Example 3

- The problem rephrased:

  "A missing plane has crashed in one of n possible regions. A search for the plane is then performed in region i; p_i is the probability that the plane has gone down in region i and α_i the probability that the plane will be found in a search of the i-th region when it is, in fact, in that region. We consider $0 < p_i < 1$ and $0 < \alpha_i < 1$ to make the problem reasonable and avoid trivialities."

- We note that the parameter of interest is the region in which the plane has gone down; let us denote it as θ.

- The parameter space can be represented by a set of integers because each region is labelled by a number. Thus, the parameter space Θ corresponds to the set $\{1, 2, \ldots, n\}$ and the prior specification is given by

$$p_j = P[\theta = j], \quad j = 1, 2, \ldots, n.$$

Example 3

- The observed data is represented by the random variable X: $X = 1$ if the plain is found in region i; 0 otherwise.
- Thus, $X \in \{0, 1\}$ and, according to the conditions of the problem,

$$P[X = 0 \mid \theta = j] = 1 \text{ if } j \neq i, \qquad P[X = 1 \mid \theta = i] = \alpha_i.$$

Example 3

- The observed data is represented by the random variable X: $X = 1$ if the plain is found in region i; 0 otherwise.
- Thus, $X \in \{0, 1\}$ and, according to the conditions of the problem,

$$P[X = 0 \mid \theta = j] = 1 \text{ if } j \neq i, \qquad P[X = 1 \mid \theta = i] = \alpha_i.$$

- An application of Bayes theorem gives us (for $j \neq i$)

$$P[\theta = i \mid X = 1] = 1, \quad P[\theta = i \mid X = 0] = \frac{(1-\alpha_i)p_i}{1-\alpha_i p_i},$$

$$P[\theta = j \mid X = 1] = 0, \quad P[\theta = j \mid X = 0] = \frac{p_j}{1-\alpha_i p_i}.$$

Example 3

- Sample space: $(\{0,1\}, \{\varnothing, \{0\}, \{1\}, \{0,1\}\})$.
- Sampling probabilities: for simplicity, let me suppose that the label i of the original problem corresponds to 1; then the sampling probabilities are the following:

$$P[X = 1 \mid \theta = 1] = \alpha_1, \qquad P[X = 1 \mid \theta = j] = 0 \text{ for all } j \geqslant 2.$$

  It is clear how these probabilities are indexed by the parameter θ.

- Parameter space: $\Theta = \{1, 2, \ldots, n\}$.
- Thus, compactly written, the statistical model corresponding to the problem at hand is given by

$$\mathcal{E} = \{ (\{0,1\}, \{\varnothing, \{0\}, \{1\}, \{0,1\}\}), \ P[\bullet \mid \theta] : \theta \in \Theta \},$$

  where

$$P[X = 0 \mid \theta] = (1 - \alpha_1)\mathbb{1}_{\{\theta=1\}} + \sum_{j=2}^{n} \mathbb{1}_{\{\theta=j\}}.$$

Example 3

- Prior specification: the corresponding measurable space is given by $(\Theta, \mathcal{P}(\Theta))$.
- In order to define a probability on it, it is enough to take n real numbers $p_1, p_2, \ldots, p_n$ such that

$$0 \leqslant p_j \leqslant 1, \quad \sum_{j=1}^{n} p_j = 1.$$

- We interpret p_j as $P[\theta = j]$. The prior distribution corresponds, therefore, to the vector $(p_1, p_2, \ldots, p_n)'$: there are as many prior distributions as numbers p_j's we choose. Each person –as Bayesian statisticians say– chooses as its initial opinion.

Example 3

- These elements allow us to construct a Bayesian model, which is characterized by a *unique* probability defined on the product space "parameters $\times$ observations".
- In our case, the Bayesian model is given by

$$(\Theta \times \{0,1\}, \mathcal{P}(\Theta) \otimes \mathcal{P}(\{0,1\}), \Pi),$$

where Π can be fully described by the following table:

	$\{X=0\}$	$\{X=1\}$
$\{\theta=1\}$	$(1-\alpha_1)p_1$	$\alpha_1 p_1$
$\{\theta=2\}$	p_2	0
$\{\theta=3\}$	p_3	0
$\vdots$	$\vdots$	$\vdots$
$\{\theta=n\}$	p_n	0

Example 3

- Let us consider the sampling distributions. It is easy to verify that the parameter θ is unidentified.
- In fact,

$$P[X = 1 \mid \theta = 2] = P[X = 1 \mid \theta = 3] = 0,$$
$$P[X = 0 \mid \theta = 2] = P[X = 0 \mid \theta = 3] = 1.$$

- That is, we found at least two different parameters such that the respective sampling probabilities are equal.

Example 3

- In spite of this identification problem, it is still possible to make inferences on all the parameters θ, in particular on those that are unidentified.
- Thus, for instance,

$$P[\theta = 2 \mid X = 0] = \frac{p_2}{1 - \alpha_1 p_1}, \qquad P[\theta = 2 \mid X = 1] = 0.$$

Example 3

- In spite of this identification problem, it is still possible to make inferences on all the parameters θ, in particular on those that are unidentified.
- Thus, for instance,

$$P[\theta = 2 \mid X = 0] = \frac{p_2}{1 - \alpha_1 p_1}, \qquad P[\theta = 2 \mid X = 1] = 0.$$

- Those computations show the advantage of the Bayesian approach when compared to the sampling theory framework approach:

  "We reinforce that even a nonidentifiable model may bring information about the parameters of interest. Bayesian inference is easily performed in cases of nonidentifiability [...] *Nonidentifiability may bring undesirable nonuniqueness to the ML estimation method" (Weshler et al., 2013, p.93).*

Example 3

- Let us still focus our attention on the sampling distributions, which can be written in the following terms:

$$P[X = 0 \mid \theta] = (1 - \alpha_1)\mathbb{1}_{\{\theta=1\}} + \mathbb{1}_{\{\theta\neq 1\}},$$

  where $\{\theta \neq 1\} = \{\theta \in \{2, 3, \ldots, n\}\}$.

- Who can describe the corresponding minimal sufficient parameter?

- The sampling process is fully characterized by the measurable partition

$$\{\{\theta = 1\}, \{\theta \neq 1\}\}.$$

- This is the identified parameter.

Example 3

- Let go to the updating process. I can update $\{\theta = 1\}$. This is easy.
- But I can also update $\{\theta = j\}$ for $j \geqslant 2$, in spite that

$$\{\theta = j\} \notin \{\{\theta = 1\}, \{\theta \neq 1\}\}.$$

Example 3

- This statement need to be carefully considered.
- In fact, when we compute the posterior probability of $\{\theta = j\}$ with $j = 1, 2, \ldots, n$, we consider the measurability structure given by $\mathcal{P}(\{1, 2, \ldots, n\})$.
- In this context, we can compute the posterior distribution not only of the identified parameters $\{\theta = 1\}$ and $\{\theta \neq 1\}$, but also of $\{\theta = j\}$ for $j = 2, 3, \ldots, n$:

$$P[\{\theta = 1\} \mid X] = \frac{(1 - \alpha_1)p_1}{1 - \alpha_1 p_1} \mathbb{1}_{\{X=0\}} + \mathbb{1}_{\{X=1\}};$$

$$P[\{\theta \neq 1\} \mid X] = \frac{p_2 + \cdots + p_n}{1 - \alpha_1 p_1} \mathbb{1}_{\{X=0\}} + 0 \cdot \mathbb{1}_{\{X=1\}};$$

$$P[\{\theta = j\} \mid X] = \frac{p_j}{1 - \alpha_1 p_1} \mathbb{1}_{\{X=0\}} + 0 \cdot \mathbb{1}_{\{X=1\}} \text{ for } j \geqslant 2.$$

Example 3

- Now: when we update unidentified parameters, do we learn something different from what we learn from the updating of the identified parameter?
- That is, that what we learn about unidentified parameters in *not* more than what we learn about the identified one: is it correct?
- Note the following: for $j = 2, 3, \ldots, n$,

$$\begin{aligned} P[\theta = j \mid X = 0] &= E[\mathbb{1}_{\{\theta=j\}} \mid \{X = 0\}] \\ &= E\left\{E[\mathbb{1}_{\{\theta=j\}} \mid \{X = 0\}, \{\theta = 1\}, \{\theta \neq 1\}]\{X = 0\}\right\} \\ &= E\left\{E[\mathbb{1}_{\{\theta=j\}} \mid \{\theta = 1\}, \{\theta \neq 1\}] \mid \{X = 0\}\right\}. \end{aligned}$$

- But

$$\begin{aligned} P[\theta = j \mid \{\theta = 1\}, \{\theta \neq 1\}] &= P[\theta = j \mid \theta = 1]\mathbb{1}_{\{\theta=1\}} + P[\theta = j \mid \theta \neq 1]\mathbb{1}_{\{\theta\neq 1\}} \\ &= 0 \cdot \mathbb{1}_{\{\theta=1\}} + \frac{p_j}{1 - p_1}\mathbb{1}_{\{\theta\neq 1\}}. \end{aligned}$$

Example 3

- Therefore,

$$P[\theta = j \mid X = 0] = E\left[\frac{p_j}{1-p_1}\mathbb{1}_{\{\theta \neq 1\}} \mid X = 0\right] = \frac{p_j}{1-p_1}P[\theta \neq 1 \mid X = 0].$$

Example 3

- Therefore,

$$P[\theta = j \mid X = 0] = E\left[\frac{p_j}{1-p_1}\mathbb{1}_{\{\theta \neq 1\}} \mid X = 0\right] = \frac{p_j}{1-p_1} P[\theta \neq 1 \mid X = 0].$$

- This type of statements should be complemented by a remark previously done, namely that the original Bayesian model can be replaced by a *reduced* model without losing relevant information: we can reduce our original Bayesian model to one which only includes the identified parameters.

Example 3

- Therefore,

$$P[\theta = j \mid X = 0] = E\left[\frac{p_j}{1-p_1}\mathbb{1}_{\{\theta \neq 1\}} \mid X = 0\right] = \frac{p_j}{1-p_1}P[\theta \neq 1 \mid X = 0].$$

- This type of statements should be complemented by a remark previously done, namely that the original Bayesian model can be replaced by a *reduced* model without losing relevant information: we can reduce our original Bayesian model to one which only includes the identified parameters.

- By doing so, we avoid redundant information! The original Bayesian model can, therefore, be reduced to the following one:

	$\{X = 0\}$	$\{X = 1\}$
$\{\theta = 1\}$	$(1-\alpha_1)p_1$	$\alpha_1 p_1$
$\{\theta \neq 1\}$	$p_2 + \cdots + p_n$	0

(1)

Example 3

- This reduction can be done *without losing relevant information*.
- Consequently, the likelihood is *exhaustively* described by the identified parameter, and the learning-by-observing process is fully concentrated on the minimal sufficient parameter in the sense that, conditionally on it, *there is nothing more to learn on* $\{\theta = j\}$ *for* $j = 2, \ldots, n$.
- Moreover, the posterior probability of the unidentified parameters are less than the posterior probability of the identified ones: for each $j = 2, \ldots, n$,

$$P[\theta = j \mid X = 0] < P[\theta \neq 1 \mid X = 0], \quad P[\theta = j \mid X = 1] < P[\theta \neq 1 \mid X = 1].$$

- In other words, it is more informative the posterior distribution of an identified parameter than that of the unidentified parameter.

Example 3

- The reduced Bayesian model is now defined on the product space

$$(\{1,\ldots,n\}\times\{0,1\},\ \mathcal{A}\otimes\mathcal{S},\Pi),$$

where

  - $\mathcal{A}=\{\varnothing,\{\theta=1\},\{\theta\neq 1\},\Theta\}$;
  - $\mathcal{S}=\{\varnothing,\{X=0\},\{X=1\},\{0,1\}\}$;
  - and Π is the matrix that Alpha computed.

- In this model, the events

$$\{\theta=j\},\quad j=2,3,\ldots,n$$

are *not* measurable and, consequently, they don't provide information on the observations are not informative: all the relevant information regarding the data is concentrated on the identified parameter.

Example 3

- What we prove is that all the information generated by the sampling process is represented by the partition

$$\{\{\theta = 1\},\ \{\theta \neq 1\}\}.$$

- Let us define the functions Y_0 and Y_1 as follows:

$$P[X = 0 \mid \theta] = (1 - \alpha_1)\mathbb{1}_{\{\theta=1\}} + \mathbb{1}_{\{\theta\neq 1\}} \doteq Y_0$$

$$P[X = 1 \mid \theta] = \alpha_1 \mathbb{1}_{\{\theta=1\}} + 0 \cdot \mathbb{1}_{\{\theta\neq 1\}} \doteq Y_1.$$

Therefore,

$$\{\theta = 1\} = Y_0^{-1}\left[\{1 - \alpha_1\}\right] \cap Y_1^{-1}\left[\{\alpha_1\}\right], \qquad \{\theta \neq 1\} = Y_0^{-1}\left[\{1\}\right] \cap Y_1^{-1}\left[\{0\}\right].$$

Example 3

- It can be seen that the parameters $\{\theta = j\}$ for $j = 2, \ldots, n$ can no be expressed as functions of the sampling distributions.
- The parameter $\{\theta = 1\}$ that the plain is in region 1, provides the same information as the probability of finding it and of not finding it there. The parameter $\{\theta \neq 1\}$ that the plain is not in region 1 provides the same information as the probability of finding and of no finding the plain in a region different from 1.

Mutual Ancillarity, Exogeneity and Cut

- Consider a Bayesian model $p(X, \theta)$, which can be decomposed as follows

$$p(X, \theta) = m(\theta)q(X \mid \theta) = q(X)\, m(\theta \mid X).$$

- Let us decompose both the observations and the parameters into two components: $X = (Y, Z)$ and $\theta = (\psi, \lambda)$, where ψ is a nuisance parameter and λ is a parameter of interest. Therefore, we only have interest in the posterior distribution of λ which can be decomposed as

$$m(\lambda \mid X) \propto m(\lambda)\, q(Z \mid \lambda)\, q(Y \mid Z, \lambda).$$

- We are interesting in the following question: how far can we forget the term $p(z \mid \lambda)$ for the inference on λ? Let us introduce three sufficient conditions in a decreasing order of generality.

Mutual Ancillarity

- If Z and λ are *mutually ancillary*, that is

$$Z \perp\!\!\!\perp \lambda,$$

then $q(Z \mid \lambda) = q(Z)$ and $m(\lambda \mid Z) = m(\lambda)$ and, therefore,

$$m(\lambda \mid X) \propto m(\lambda)\, q(Y \mid Z, \lambda).$$

Mutual Exogeneity

- However, the use of

$$m(\lambda \mid X) \propto m(\lambda)\, q(Y \mid Z, \lambda).$$

implies integrating the conditional sampling distribution of $(Y \mid Z)$ with respect to the conditional posterior distribution of ψ given Z and λ, namely

$$q(Y \mid Z, \lambda) = \int_{\Psi} q(Y, \psi \mid Z, \lambda)\, d\psi = \int_{\Psi} q(Y \mid Z, \theta)\, m(\psi \mid Z, \lambda)\, d\psi.$$

- But if λ is a sufficient parameter of the conditional sampling process $(Y \mid Z)$, i.e.,

$$Y \perp\!\!\!\perp \psi \mid Z, \lambda,$$

then $q(Y \mid Z, \lambda) = q(Y \mid Z, \psi, \lambda)$ and, therefore,

$$\begin{aligned} q(Y \mid Z, \lambda) &= \int_{\Psi} q(Y \mid Z, \theta)\, m(\psi \mid Z, \lambda)\, d\psi \\ &= \int_{\Psi} q(Y \mid Z, \lambda)\, m(\psi \mid Z, \lambda)\, d\psi \\ &= q(Y \mid Z, \lambda) \int_{\Psi} m(\psi \mid Z, \lambda)\, d\psi = q(Y \mid Z, \lambda). \end{aligned}$$

Mutual Exogeneity

- When both

$$Z \perp\!\!\!\perp \lambda, \qquad Y \perp\!\!\!\perp \psi \mid Z, \lambda$$

  are verified, λ and Z are said to be *mutually exogenous*.

- Thus, if only λ is of interest, and if λ and Z are mutually exogenous, the process generating Z becomes irrelevant, and the data-generating process may be directly conditioned on Z.

- This is precisely the meaning of the so-called "exogenous variables" used in econometric models and introduced by Koopmans (1950).

Bayesian cut

- When λ is a sufficient parameter of the sampling distribution conditional on Z, a rather natural way of obtaining the mutual exogeneity of λ and Z is through the structure of a *Bayesian cut* which means that along with

$$Y \perp\!\!\!\perp \psi \mid Z, \lambda,$$

ψ is furthermore a sufficient parameter of the sampling marginal process of Z, i.e.

$$Z \perp\!\!\!\perp \lambda \mid \psi,$$

and is also a priori independent of λ:

$$\psi \perp\!\!\!\perp \lambda.$$

Bayesian cut

- The structure of a Bayesian cut is given by the following decomposition:

$$\begin{aligned} q(Y, Z, \lambda, \psi) &= p(Y \mid Z, \lambda, \psi)\, q(Z \mid \lambda, \psi)\, m(\lambda, \psi) \\ &= [p(Y \mid Z, \lambda)\, m(\lambda)]\, [q(Z \mid \psi)\, m(\psi)] \end{aligned}$$

- In a sampling theory framework, the corresponding concept of cut (Barndorff-Nielsen, 1978; Engle, Hendry and Richard, 1980) replaces the prior independence between ψ and λ by the condition of being variation-free, that is,

$$(\psi, \lambda) \in \Lambda \times \Psi.$$

Bayesian cut

- The Bayesian cut implies different consequences:

1. Conditions $Z \perp\!\!\!\perp \lambda \mid \psi$ and $\psi \perp\!\!\!\perp \lambda$ are jointly equivalent to $(Z, \psi) \perp\!\!\!\perp \lambda$, which implies condition $Z \perp\!\!\!\perp \lambda$. In other words, Bayesian cut implies mutual ancillarity between Z and λ.
2. $(Z, \psi) \perp\!\!\!\perp \lambda$ also implies that $\psi \perp\!\!\!\perp \lambda \mid Z$, that is, the posterior independence of λ and ψ conditionally on Z.
3. Condition $Z \perp\!\!\!\perp \lambda \mid \psi$, along with $\psi \perp\!\!\!\perp \lambda \mid Z$, are jointly equivalent to $\psi \perp\!\!\!\perp (Y, \lambda) \mid Z$. This latter condition implies that $\psi \perp\!\!\!\perp \lambda \mid (Y, Z)$. In other words, the Bayesian cut implies the posterior independence of ψ and λ given (Y, Z).

Bayesian cut

- Let us consider bivariate parameters and observations: $a = (b, c)$ and $X = (T, U)$. Suppose now that each coordinate may assume two values only: $b \in \{b_1, b_2\}$, $c \in \{c_1, c_2\}$, $T \in \{T_1, T_2\}$ and $U \in \{U_1, U_2\}$. The Bayesian model characterized by the joint probability $\pi(a, X)$ is suitable defined by the following 15 numbers (assumed to be different from zero):

$$\begin{aligned} &\mu(c_1) = \mu_0 \\ &\mu(b_1 \mid c_i) = \mu_i, && i = 1, 2 \\ &p(T_1 \mid c_i, b_j) = p_{ij}, && i, j = 1, 2 \\ &p(U_1 \mid T_i, c_j, b_k) = q_{ijk}, && i, j, k = 1, 2. \end{aligned}$$

Bayesian cut

- Mutual ancillarity between c and T is equivalent to $p(T_1 \mid c_1) = p(T_1 \mid c_2)$, that is,

$$\mu_1 p_{11} + (1 - \mu_1) p_{12} = \mu_2 p_{21} + (1 - \mu_2) p_{22}. \tag{2}$$

- For mutual exogeneity the supplementary condition $(a \perp\!\!\!\perp X \mid T, c)$ is equivalent to the four equalities

$$q_{ij1} = q_{ij2} \quad i, j = 1, 2. \tag{3}$$

Bayesian cut

- For a cut, it is necessary to verify condition (3) plus the following three:

$$\mu_1 = \mu_2, \qquad \text{i.e., } b \perp\!\!\!\perp c, \tag{4}$$

$$p_{1j} = p_{2j}, \quad j = 1, 2, \qquad \text{i.e., } a \perp\!\!\!\perp T \mid b. \tag{5}$$

- Even under prior independence (4), it should be clear that mutual exogeneity (conditions (2) and (3)) does not imply a cut (conditions (3), (4) and (5)).