M1 CogSUP – M2 Éco/Psycho
Bayesian Models for Cognition and Behavior
Semestre 1

Université Paris Cité — SORBONNE UNIVERSITÉ

Cours 3: Bayesian Inference

Responsables : Jonathan Vacher (jonathan.vacher@u-paris.fr)
Jean Daunizeau (jean.daunizeau@gmail.com)

1 Introduction

1.1 A Brief History and a Bit of Philosophy

The notion of probability emerged during the XVI$^{\text{th}}$ century with Cardano defining odds. During the XVII$^{\text{th}}$ century Fermat, Pascal and Huygens exchange letters about probabilities talking about games (How to fairly split rewards in game of chance ?). Until the XX$^{\text{th}}$ century people like de Moivre, Franklin, Simpson, Laplace, Bernoulli, Gauss, Legendre, Lacroix, Dedekind, Pearson, De Morgan, Boole contributed to developing and formalizing the theory.

The modern theory of probability is based on measure (of sets) theory and was mainly established by Andreï Kolmogorov during the 1920s/1930s. In his theory events are sets and the theory is based on three axioms :

(i) Probability of an event is positive;

(ii) Probability of at least one event is one;

(iii) Probability of disjoint events equals the sum of the probabilities of these events.

Those axioms are sometimes derived from Cox's assumptions mentioned by Jean during the 1$^{\text{st}}$ class (comparability, common sense, consistency).

There are philosophical debates about how to interpret the theory of probabilities. This is how originated the Frequentists *vs* Bayesianists clash. In short, for Frequentists probabilities are observable frequencies while for Bayesianists probabilities represent tangible states of knowledge (uncertainty). Contrary to Frequentists, Bayesianists attribute probabilities to hypotheses.

Regarding cognitive sciences, Thurstone 1927 brought the concept of random variable to psychology. He thought that we have continuous internal representation of physical variables.

1.2 Framework

In this course, we denote $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space and $(E, \mathcal{E})$ a measurable space. A E-valued random variable X is a measurable function $X : \Omega \longrightarrow E$. The law of X is the image measure of the reference measure $\mathbb{P}$ by the random variable X *i.e.* for any measurable set $A \subset E$ the probability that X belongs to A is

$$\mathbb{P}_X(A) \overset{\text{def.}}{=} \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A\}) = \mathbb{P}(X^{-1}(A)) = \int_{X^{-1}(A)} \mathrm{d}\mathbb{P}(\omega) = \int_A \mathrm{d}\mathbb{P}_X(x).$$

The density of a random variable X with respect to a reference measure μ over E is the Radon-Nikodym derivative of $\mathbb{P}_X$, $\mathrm{f}_X = {}^{\mathrm{d}\mathbb{P}_X}/_{\mathrm{d}\mu}$ *i.e.* for any measurable set $A \subset E$

$$\mathbb{P}_X(A) = \int_A \mathrm{d}\mathbb{P}_X(x) = \int_A \mathrm{f}_X(x)\mathrm{d}\mu(x).$$

For any measurable function g over E, the expectation of $g(X)$ is given by

$$\mathbb{E}_X(g(X)) = \int_E g(x)\mathrm{d}\mathbb{P}_X(x) = \int_E g(x)\mathrm{f}_X(x)\mathrm{d}\mu(x).$$

Those theoretical considerations are not necessary to understand the content of the course, yet they provide a rigorous framework. In particular, this level of generality encompasses both discrete and continuous random variables and justify the correspondence between $\int$ and $\sum$. In practice, the most simple case is when $E = \mathbb{R}$ and $\mathrm{d}\mu$ is the Lebesgue measure $\mathrm{d}x$.

Exercise 1. Write f_X, $\mathbb{P}_X$ and $\mathbb{E}_X$ when $E = \{1, \dots, n\}$ with $n \in \mathbb{N}$.

Solution 1.

2 Conditional Probabilities

The notion of conditional probabilities is a convenient tool to make sense of the probabilistic dependence between variables.

Definition 1 (Joint Law)**.** *Let X and Y be two E-valued random variables . The joint law of X and Y is the law of the $E \times E$-valued random variable $Z = (X, Y)$. We denote $\mathbb{P}_Z = \mathbb{P}_{X,Y}$ for the probability measures and $\mathrm{f}_Z = \mathrm{f}_{X,Y}$ for the densities.*

Definition 2 (Marginal Laws)**.** *Let X and Y be two E-valued random variables. The marginal law of X and Y are the following laws defined for any measurable set $A \subset E$ by*

$$\mathbb{P}_X(A) = \int_{A\times E} \mathrm{d}\mathbb{P}_{X,Y}(x,y) \quad \textit{and} \quad \mathbb{P}_Y(A) = \int_{E\times A} \mathrm{d}\mathbb{P}_{X,Y}(x,y).$$

Similarly, the marginal densities of X and Y are

$$\mathrm{f}_X = \int_E \mathrm{f}_{X,Y}(\cdot, y)\mathrm{d}y \quad \textit{and} \quad \mathrm{f}_Y = \int_E \mathrm{f}_{X,Y}(x,\cdot)\mathrm{d}x.$$

Remark 1. In practice Definition 2 is well-justified by coming back to the definition of $\mathbb{P}_{X,Y}$ in Section 1.2. The core is reason is that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

Equipped with the notions of joint law and marginal laws, we can now define conditional probabilities.

Definition 3 (Conditional Probabilities)**.** *Let X and Y be two E-valued random variables. The conditional probability of X knowing Y is defined for any pair of measurable sets $A, B \subset E$ such that $\mathbb{P}_Y(B) > 0$ by*

$$\mathbb{P}_{X|Y}(A|B) = \frac{\mathbb{P}_{X,Y}(A,B)}{\mathbb{P}_Y(B)}$$

Additionally, when for all $y \in E$, $\mathrm{f}_Y(y) > 0$, the conditional density of X knowing Y is defined for all $(x, y) \in E^2$ by

$$\mathrm{f}_{X|Y}(x|y) = \frac{\mathrm{f}_{X,Y}(x,y)}{\mathrm{f}_Y(y)}.$$

Definition 4 (Independence)**.** *Let X and Y bet two E-valued random variables. The random variables X and Y are independent if for any pair of measurable sets $A, B \subset E$,*

$$\mathbb{P}_{X,Y}(A,B) = \mathbb{P}_X(A)\mathbb{P}_Y(B) \quad \textit{or} \quad \mathbb{P}_{X|Y}(A|B) = \mathbb{P}_X(A) \quad \textit{or} \quad \mathbb{P}_{Y|X}(B|A) = \mathbb{P}_Y(B).$$

Similarly for the densities, X and Y are independent if for all $(x, y) \in E^2$,

$$\mathrm{f}_{X,Y}(x,y) = \mathbb{P}_X(x)\mathbb{P}_Y(y) \quad \textit{or} \quad \mathrm{f}_{X|Y}(x|y) = \mathrm{f}_X(x) \quad \textit{or} \quad \mathrm{f}_{Y|X}(y|x) = \mathbb{P}_Y(y).$$

Example 1. For the $2D$-Gaussian case, we can write everything in closed form. Let $(X, Y) \sim \mathcal{N}(\mu, \Sigma)$ with

$$\mu = (\mu_X, \mu_Y) \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_X^2 & \rho \\ \rho & \sigma_Y^2 \end{pmatrix}.$$

The joint density is

$$\mathrm{f}_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right).$$

The marginal densities are

$$\mathrm{f}_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right) \quad \text{and} \quad \mathrm{f}_f Y(y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}\right).$$

The conditional density of X knowing Y is

$$\mathrm{f}_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi\sigma_{X|Y}^2}} \exp\left(-\frac{(x-\mu_{X|Y})^2}{2\sigma_{X|Y}^2}\right)$$

with

$$\mu_{X|Y} = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y) \quad \text{and} \quad \sigma_{X|Y}^2 = \sigma_X^2(1-\rho^2).$$

From the symmetry of Definition 3, we immediately derive Bayes rule/Theorem/formula ... But this is definitely almost a definition !

Proposition 1 (Bayes Rule)**.** *Let X and Y be two E-valued random variables. Then for any pair of measurable sets $A, B \subset E$ such that $\mathbb{P}_Y(A) > 0$,*

$$\mathbb{P}_{Y|X}(B|A) = \frac{\mathbb{P}_{X|Y}(A|B)\mathbb{P}_Y(B)}{\mathbb{P}_X(A)}.$$

Additionally, when for all $x \in E$, $\mathrm{f}_X(x) > 0$, then for all $(x, y) \in E^2$,

$$\mathrm{f}_{Y|X}(y|x) = \frac{\mathrm{f}_{X|Y}(x|y)\mathrm{f}_Y(y)}{\mathrm{f}_X(x)}.$$

Bayes rule is the rule of probabilistic induction. To make sense of this, we can index events by time t and assume causality, then

$$\mathbb{P}_{Y|X}(A_t|A_{t+1}) = \frac{\mathbb{P}_{X|Y}(A_{t+1}|A_t)\mathbb{P}_Y(A_t)}{\mathbb{P}_X(A_{t+1})}.$$

Written this way, we calculate the probability of A_t causing A_{t+1}.

3 Estimation Theory

In this section, we place ourselves in the standard statistical setting. We consider a set of $N \in \mathbb{N}$ observations $\boldsymbol{x} = (x_1, \dots, x_N) \in E^N$ that are independent identically distributed (iid) samples of an E-valued random variable X. The random variable X has conditional density $\mathrm{f}_{X|\Theta}$ where Θ is a Π-valued random variable with Π being the parameter space. The samples $\boldsymbol{x}$ can also be viewed as a realization of the a random variable $\boldsymbol{X} = (X_1, \dots, X_N)$ where for all $i \in \{1, \dots, N\}$, X_i has the same distribution as X. In the present Bayesian setting, the random variable X depends on the random variable Θ. This contrasts with the frequentist approach in which θ is a deterministic parameter of the density $\mathrm{f}_{X,\theta}$ of X (in practice we will only use the conditional density *i.e.* for any $\theta \in \Pi$, $\mathrm{f}_{X,\theta} = \mathrm{f}_{X|\Theta}(\cdot|\theta)$). In addition, we denote by $L : \Pi \times \Pi \to \mathbb{R}$ the loss function. A loss function is also called a cost function, the opposite of the loss function is called gain or utility function. The lost/cost function is minimized while the gain/utility function is maximized. We can now define different notion of risks.

3.1 Frequentist Estimators

In the frequentist setting, an estimator is a deterministic function of the observation.

Definition 5 (Frequentist Risk)**.** *Let $\tilde{\theta} : E^N \longrightarrow \Pi$ be an estimator of $\theta \in \Pi$ the density parameter of X. The frequentist risk is*

$$R_f(\theta, \tilde{\theta}) = \mathbb{E}_{\boldsymbol{X}|\Theta}\left(L(\theta, \tilde{\theta}(\boldsymbol{X})) \mid \theta\right) = \int_{E^N} L(\theta, \hat{\theta}(\boldsymbol{x}))\mathrm{f}_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)\mathrm{d}\boldsymbol{x}.$$

This definition assumes that the true parameter θ is known, the interest is mostly theoretical. In practice, it is impossible to compute the frequentist risk. Though, the notion of frequentist risk provides a way to compare estimators. Indeed, for any pair of estimators $(\tilde{\theta}_1, \tilde{\theta}_2)$, we can say that $\tilde{\theta}_1$ is better than $\tilde{\theta}_2$ if for all $\theta \in \Pi$,

$$R_f(\tilde{\theta}_1, \theta) \leqslant R_f(\tilde{\theta}_2, \theta). \tag{1}$$

Such a relation of comparison is very restrictive because the inequality must hold for all $\theta \in \Pi$, it is only defining a partial order on the estimators. Additional assumptions are required to define the best estimator in this frequentist sense.

A first relaxation of the above relation of comparison is given by the minimax criterion. The minimax criterion aims at choosing an estimator that has a minimal risk in the worst case scenario *i.e.* an estimator $\hat{\theta}$ is minimax if for all $\tilde{\theta} \in \Pi$,

$$\max_{\theta \in \Pi} R_f(\theta, \hat{\theta}) \leqslant \max_{\theta \in \Pi} R_f(\theta, \tilde{\theta}).$$

This can also be written, $\hat{\theta}$ is minimax if

$$\max_{\theta \in \Pi} R_f(\theta, \hat{\theta}) = \min_{\tilde{\theta} \in \Pi} \max_{\theta \in \Pi} R_f(\theta, \tilde{\theta}). \tag{2}$$

The minimax criterion (2) is less restrictive than the criterion (1), though it tends to favor the least risky estimators.

3.2 Bayesian Estimators

In the Bayesian setting, the parameter to be estimated is a random variable with density $\mathbb{P}_\Theta$. The Bayesian risk takes this into account by averaging the risk with respect to θ. In a way, the hypothesis that the parameter θ is known when defining the frequentist risk is relaxed, it is only assumed that the distribution of Θ is known. The distribution of Θ is called the prior distribution.

Definition 6 (Bayesian Risk)**.** *Let $\tilde{\theta} : E^N \longrightarrow \Pi$ be an estimator of $\theta \in \Pi$. The Bayesian risk is*

$$R_b(\tilde{\theta}) = \mathbb{E}_\Theta\left(R_f(\Theta, \tilde{\theta})\right) = \mathbb{E}_\Theta\left(\mathbb{E}_{\boldsymbol{X}|\Theta}\left(L(\Theta, \tilde{\theta}(\boldsymbol{X})) \mid \Theta\right)\right) = \int_{E^N \times \Pi} L(\theta, \tilde{\theta}(\boldsymbol{x}))\mathrm{f}_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)\mathbb{P}_\Theta(\theta)\mathrm{d}\boldsymbol{x}\mathrm{d}\theta.$$

Thanks to this relaxed definition of the risk, we can now define what is a Bayesian estimator.

Definition 7 (Bayesian Estimator)**.** *A estimator $\hat{\theta} : E^N \longrightarrow \Pi$ is a Bayesian estimator if it minimizes the Bayesian risks i.e.*

$$R_b(\hat{\theta}) = \min_{\tilde{\theta} \in \Pi} R_b(\tilde{\theta}).$$

Interestingly, the Bayesian risk can be rewritten thanks to Bayes rule,

$$R_b(\tilde{\theta}) = \int_{E^N \times \Pi} L(\theta, \tilde{\theta}(\boldsymbol{x}))\mathrm{f}_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)\mathbb{P}_\Theta(\theta)\mathrm{d}\boldsymbol{x}\mathrm{d}\theta = \int_{E^N}\int_{\Pi} L(\theta, \tilde{\theta}(\boldsymbol{x}))\mathrm{f}_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\mathrm{d}\theta\, \mathbb{P}_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

This last expression lets appear a quantity named posterior risk which we can calculate and minimize in practice.

Definition 8 (Posterior Risk)**.** *The posterior risk is*

$$R(\tilde{\theta}) = \mathbb{E}_{\Theta|\boldsymbol{X}}\left(L(\Theta, \tilde{\theta}) \mid \boldsymbol{x}\right) = \int_{\Pi} L(\theta, \tilde{\theta})\mathrm{f}_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\mathrm{d}\theta.$$

The posterior risk is key because of the following theorem that states that it is sufficient to minimize the posterior risk to obtain a Bayesian estimator.

Theorem 1 (Minimizers of the posterior risk are Bayesian)**.** *Suppose that $R_b(\tilde{\theta})$ for some $\tilde{\theta} \in \Pi$. For all $\hat{\theta} \in \Pi$ such that*

$$\hat{\theta} \in \operatorname*{argmin}_{\tilde{\theta} \in \Pi} R(\tilde{\theta})$$

then $\hat{\theta}$ is a Bayesian estimator.

3.3 Standard Loss Functions

Loss functions come with different flavors though they must quantify the difference between an estimate $\tilde{\theta}$ and the possible values of θ. A general class of loss functions is given by the $\mathcal{L}_p$-norms, for all $(\tilde{\theta}, \theta) \in \Pi^2$,

$$L_p(\theta, \tilde{\theta}) = \|\theta - \tilde{\theta}\|_p^p.$$

Among those the case $p = 2$ and $p = 1$ are specifically interesting.

Quadratic loss The quadratic loss is the case $p = 2$ *i.e.* $L_2(\theta, \tilde{\theta}) = \|\theta - \tilde{\theta}\|_2^2$. The posterior risk, also called Mean Square Error (MSE) is

$$R(\tilde{\theta}) = \mathbb{E}_{\Theta|\boldsymbol{X}}\left(\|\Theta - \tilde{\theta}\|_2^2 \mid \boldsymbol{x}\right).$$

A potential minimizer can be found by finding the zeroes of the gradient of R *i.e.*

$$\nabla R(\hat{\theta}) = 2\mathbb{E}_{\Theta|\boldsymbol{X}}\left(\Theta - \hat{\theta} \mid \boldsymbol{x}\right) = 0 \Longrightarrow \hat{\theta} = \mathbb{E}_{\Theta|\boldsymbol{X}}\left(\Theta \mid \boldsymbol{x}\right).$$

It is indeed a minimizer because for any $\tilde{\theta} \in \Pi$,

$$\begin{aligned} R(\hat{\theta}) &= \mathbb{E}_{\Theta|\boldsymbol{X}}\left(\|\Theta - \tilde{\theta} + \tilde{\theta} - \hat{\theta}\|_2^2 \mid \boldsymbol{x}\right) = \mathbb{E}_{\Theta|\boldsymbol{X}}\left(\|\Theta - \tilde{\theta}\|_2^2 + \|\tilde{\theta} - \hat{\theta}\|_2^2 + 2\langle\Theta - \tilde{\theta},\, \tilde{\theta} - \hat{\theta}\rangle \mid \boldsymbol{x}\right) \\ &= R(\tilde{\theta}) + \|\tilde{\theta} - \hat{\theta}\|_2^2 + 2\langle\mathbb{E}_{\Theta|\boldsymbol{X}}\left(\Theta \mid \boldsymbol{x}\right) - \tilde{\theta},\, \tilde{\theta} - \hat{\theta}\rangle = R(\tilde{\theta}) + \|\tilde{\theta} - \hat{\theta}\|_2^2 - 2\langle\tilde{\theta} - \hat{\theta},\, \tilde{\theta} - \hat{\theta}\rangle \\ &= R(\tilde{\theta}) - \|\tilde{\theta} - \hat{\theta}\|_2^2 \leqslant R(\tilde{\theta}). \end{aligned}$$

The Bayesian estimator under the quadratic loss function is the posterior expectation of Θ.

$\mathcal{L}_1$ **loss** The $\mathcal{L}_1$ loss is the case $p = 1$ *i.e.* $L_1(\theta, \tilde{\theta}) = \|\theta - \tilde{\theta}\|_1$. The posterior risk, also called Meas Absolute Error (MAE) is

$$R(\tilde{\theta}) = \mathbb{E}_{\Theta|\boldsymbol{X}}\left(\|\Theta - \tilde{\theta}\|_1 \mid \boldsymbol{x}\right).$$

If Π has dimension 1, L_1 is the absolute value, we can write

$$\begin{aligned} R(\tilde{\theta}) &= \mathbb{E}_{\Theta|\boldsymbol{X}}\left(|\Theta - \tilde{\theta}| \mid \boldsymbol{x}\right) = \mathbb{E}_{\Theta|\boldsymbol{X}}\left(|\Theta - \tilde{\theta}|\left(\mathbb{1}_{\Theta>\tilde{\theta}}(\Theta) + \mathbb{1}_{\Theta<\tilde{\theta}}(\Theta)\right) \mid \boldsymbol{x}\right) \\ &= \mathbb{E}_{\Theta|\boldsymbol{X}}\left((\Theta - \tilde{\theta})\mathbb{1}_{\Theta>\tilde{\theta}}(\Theta) - (\Theta - \tilde{\theta})\mathbb{1}_{\Theta<\tilde{\theta}}(\Theta) \mid \boldsymbol{x}\right). \end{aligned}$$

Then, we can find a potential minimizer by finding the zeroes of the derivative of R *i.e.*

$$\begin{aligned} R'(\hat{\theta}) &= \mathbb{E}_{\Theta|\boldsymbol{X}}\left(-\mathbb{1}_{\Theta>\hat{\theta}}(\Theta) + \mathbb{1}_{\Theta<\hat{\theta}}(\Theta) + 2(\Theta - \hat{\theta})\delta_{\hat{\theta}}(\Theta) \mid \boldsymbol{x}\right) \\ &= -\mathbb{E}_{\Theta|\boldsymbol{X}}\left(\mathbb{1}_{\Theta>\hat{\theta}}(\Theta) \mid \boldsymbol{x}\right) + \mathbb{E}_{\Theta|\boldsymbol{X}}\left(\mathbb{1}_{\Theta<\hat{\theta}}(\Theta) \mid \boldsymbol{x}\right) \\ &= -\mathbb{P}_{\Theta|\boldsymbol{X}}\left(\Theta > \hat{\theta}\right) + \mathbb{P}_{\Theta|\boldsymbol{X}}\left(\Theta < \hat{\theta}\right) = 0 \Longrightarrow \hat{\theta} = \mathrm{Med}_{\Theta|\boldsymbol{X}}(\Theta|\boldsymbol{x}). \end{aligned}$$

Since the expectation of the absolute value is convex, $\hat{\theta}$ is the only possible global minimum (this argument also works for the quadratic loss). The Bayesian estimator under the $\mathcal{L}_1$ loss function is the posterior median of Θ.

Generated 0-1 loss The generated 0-1 loss is not a $\mathcal{L}_p$ norm. It is useful to interpret the Maximum A Posteriori (MAP) estimator as a Bayesian estimator. This loss function writes $L_{\mathrm{MAP}}(\theta, \tilde{\theta}) = 1 - \delta_{\tilde{\theta}}(\theta)$. Therefore the posterior risk writes

$$R(\tilde{\theta}) = \mathbb{E}_{\Theta|\boldsymbol{X}}\left(1 - \delta_{\tilde{\theta}}(\theta) \mid \boldsymbol{x}\right) = 1 - \int_{\Pi} \delta_{\tilde{\theta}}(\theta)\mathrm{f}_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\mathrm{d}\theta = 1 - \mathrm{f}_{\Theta|\boldsymbol{X}}(\tilde{\theta}|\boldsymbol{x}).$$

Then, minimizing R corresponds to maximize the posterior distribution. When the posterior distribution is unimodal we have

$$\hat{\theta} = \underset{\tilde{\theta}\in\Pi}{\mathrm{argmax}}\ \mathrm{f}_{\Theta|\boldsymbol{X}}(\tilde{\theta}|\boldsymbol{x}).$$

Hence, in this context the MAP estimator is a Bayesian estimator (not everyone agrees on that...).

To conclude this section about Bayesian estimation it is important to state the Bernstein-Von Mises theorem. The hypotheses are quite technical so it will only be a vague but memorable statement.

Theorem 2 (Bernstein-Von Mises)**.** *Let $\boldsymbol{x} = (x_1, \ldots, x_N)$ an idd sample with distribution $\mathrm{f}_{X|\Theta}(\cdot|\theta_0)$ where $\theta_0 \in \Pi$. Under some regularity and convergence conditions the posterior distribution of $\sqrt{N}(\Theta - \Theta_N)$ knowing $\boldsymbol{X} = \boldsymbol{x}$ converges toward a Gaussian with mean 0 and covariance $I(\theta_0)$ as N goes to infinity.*

The Bernstein-Von Mises theorem means that a Bayesian estimator is asymptotically independent from the choice of the prior distribution.

4 Application to Cognition and Bayesian Intuition

4.1 Cue Combination

We place ourselves in the experimental context of Ernst and Banks (2002) that is a human observer has to compare repeatedly the sizes of two bars. To estimate the size of a bar the human observer can use both haptic and visual information. The experimenter is able to control the level of noise in the visual modality. Moreover, the experimenter can set different sizes for one bar depending on the modality (larger when viewed, smaller when touched). In the following, we present the very simple Bayesian observer model that is used by Ernst and Banks (2002). We suppose that the Bayesian observer makes visual measurements S_v and haptic measurements S_h. Those measurements are random variables assumed to be independent. Their conditional distributions knowing the stimulus true sizes S are Gaussian with the same mean S and different standard deviations σ_v and σ_h. The Bayesian observer is assumed to have a flat degenerated prior about the size of the bars.

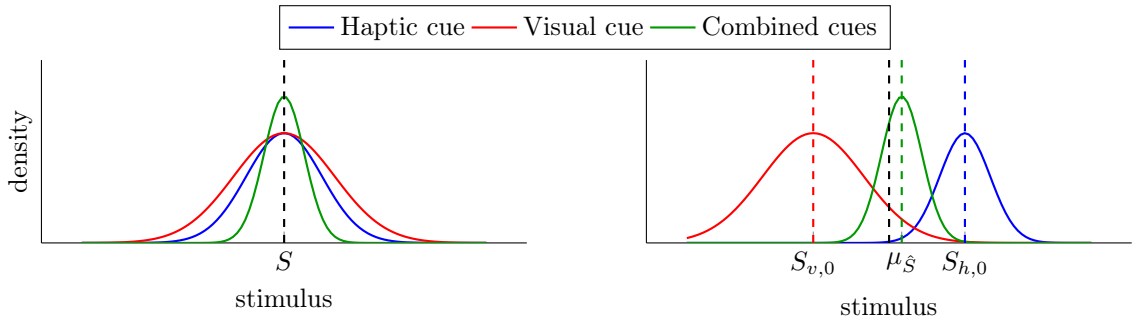

Figure 1. Estimations in the case where both modalities have the same presented sizes (left) and different sizes (right).

Estimation Thanks to Bayes rule, the posterior density about the size of the bar is

$$\mathrm{f}_{S|S_v,S_h}(s|s_v,s_h) = \frac{\mathrm{f}_{S_v,S_h|S}(s_v,s_h|s)\mathrm{f}_S(s)}{\mathrm{f}_{s_v,s_h}(s_v,s_h)} \propto \mathrm{f}_{S_v|S}(s_v|s)\mathrm{f}_{S_h|S}(s_h|s).$$

Let us use the MAP estimator. We need to maximize $\mathrm{f}_{S|S_v,S_h}(s|s_v,s_h)$ though it is convenient to minimize the negative logarithm of $\mathrm{f}_{S|S_v,S_h}(s|s_v,s_h)$. We write

$$\ell(s) = -\ln\left(\mathrm{f}_{S|S_v,S_h}(s|s_v,s_h)\right) = -\ln\left(\mathrm{f}_{S_v|S}(s_v|s)\right) - \ln\left(\mathrm{f}_{S_h|S}(s_h|s)\right) = \frac{(s_v-s)^2}{2\sigma_v^2} + \frac{(s_h-s)^2}{2\sigma_h^2} + C(\sigma_v,\sigma_h).$$

Taking the derivative with respect to s and looking for its zeroes, we obtain

$$\frac{s_v-\hat{s}}{\sigma_v^2} + \frac{s_h-\hat{s}}{\sigma_h^2} = 0 \Longrightarrow \hat{s} = \frac{\tau_v s_v + \tau_h s_h}{\tau_v+\tau_h} \quad \text{where} \quad \tau_i = \frac{1}{\sigma_i^2} \quad \text{for} \quad i \in \{v,h\}.$$

The variable τ is the inverse of the variance, it is known as the precision. This means that a Bayesian observer combines his visual and haptic estimates optimally by weighting each estimate by its sensitivity. The most sensitive the cue, the most reliable it is. The equality above can be written with random variables

$$\hat{S} = \frac{\tau_v S_v + \tau_h S_h}{\tau_v+\tau_h}.$$

From this we can conclude that $\hat{S}|S$ is a Gaussian random variable with mean $\mu_{\hat{S}}$ and variance $\sigma_{\hat{S}}^2$ given by

$$\mu_{\hat{S}}(S) = \mathbb{E}_{\hat{S}|S}(\hat{S}|S) = \frac{\tau_v \mathbb{E}_{S_v|S}(S_v|S) + \tau_h \mathbb{E}_{S_h|S}(S_h|S)}{\tau_v+\tau_h} = S \tag{3}$$

and

$$\sigma_{\hat{S}}^2(S) = \mathbb{V}_{\hat{S}|S}(\hat{S}|S) = \frac{\tau_v \mathbb{V}_{S_v|S}(S_v|S) + \tau_h \mathbb{V}_{S_h|S}(S_h|S)}{\tau_v+\tau_h} = \frac{1}{\tau_v+\tau_h}. \tag{4}$$

In the protocol, the experimenter is able to control the level of noise in the visual modality *i.e.* σ_v^2 or $1/\tau_v$. In addition, they are able to control independently the sizes of the bars presented in both modalities say $S_{v,0}$ and $S_{h,0}$ such that $S_{v,0}+S_{h,0}=2S$. The observer still believe that there is a unique bar with a single size S so that the estimation is still the same. However the mean of the estimate $\hat{S}$ is different

$$\mu_{\hat{S}}(S_{v,0},S_{h,0}) = \frac{\tau_v \mathbb{E}_{S_v|S}(S_v|S_{v,0}) + \tau_h \mathbb{E}_{S_h|S}(S_h|S_{h,0})}{\tau_v+\tau_h} = \frac{\tau_v S_{v,0} + \tau_h S_{h,0}}{\tau_v+\tau_h}. \tag{5}$$

There are two partial conclusions:

1. Cue combination provides a more accurate estimate than the ones obtained under any single modality

$$\frac{1}{\tau_v+\tau_v} \leqslant \min\left(\frac{1}{\tau_v},\frac{1}{\tau_h}\right).$$

2. The estimate is biased toward the most accurate estimate (equation (5)).

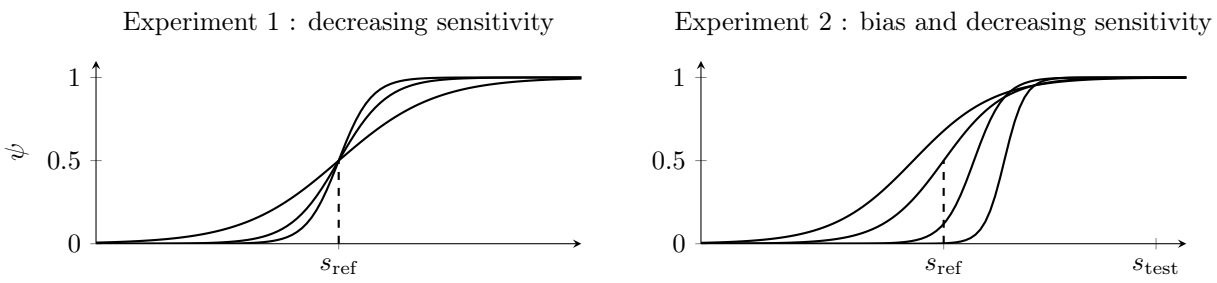

Figure 2. Expected psychometric functions in experiment 1 and 2.

Decision The experiment does not provide a direct measure of the estimate performed by the observer. Instead, the participant is asked to repeatedly compare pairs of stimuli in a two alternative forced choice paradigm. How does the Bayesian behavior is reflected in this experiment ? The Bayesian observer makes two estimates, one for each stimuli $\hat{S}_{\text{ref}}$ and $\hat{S}_{\text{test}}$. Then, they have to decide which is larger *i.e.* whether $\hat{S}_{\text{ref}} > \hat{S}_{\text{test}}$. Remind that the estimates $\hat{S}_{\text{ref}}$ and $\hat{S}_{\text{test}}$ are conditioned on the value of the presented stimuli S_{ref} and S_{test} which are deterministically chosen by the experimenter. Hence, we should observe

$$\psi(s_{\text{ref}}, s_{\text{test}}) = \mathbb{P}_{\hat{S}_{\text{ref}},\hat{S}_{\text{test}}}\left(\hat{S}_{\text{ref}} < \hat{S}_{\text{test}} \; ; \; \mu_{\hat{S}_{\text{ref}}}(s_{\text{ref}}), \mu_{\hat{S}_{\text{test}}}(s_{\text{test}})\right)$$

Under our Gaussian assumptions the variable $Z = \hat{S}_{\text{test}} - \hat{S}_{\text{ref}}$ is also Gaussian with mean $\mu_{\hat{S}_{\text{test}}} - \mu_{\hat{S}_{\text{ref}}}$ and variance $\sigma^2_{\hat{S}_{\text{test}}} + \sigma^2_{\hat{S}_{\text{ref}}}$. Hence

$$\psi(s_{\text{ref}}, s_{\text{test}}) = \Psi\left(\frac{\mu_{\hat{S}_{\text{test}}}(s_{\text{test}}) - \mu_{\hat{S}_{\text{ref}}}(s_{\text{ref}})}{\sqrt{\sigma^2_{\hat{S}_{\text{test}}} + \sigma^2_{\hat{S}_{\text{ref}}}}}\right).$$

Consequently, using equations (3), (4), (5) we obtain for the first experimental condition in which both cues have the same mean

$$\psi(s_{\text{ref}}, s_{\text{test}}) = \Psi\left(\sqrt{\frac{\tau_v + \tau_h}{2}}(s_{\text{test}} - s_{\text{ref}})\right).$$

In the second experimental condition, the reference condition has conflicting cues *i.e.* $s_{\text{ref}} = (s_{\text{ref},v,0}, s_{\text{ref},h,0})$ thus,

$$\psi(s_{\text{ref}}, s_{\text{test}}) = \Psi\left(\sqrt{\frac{\tau_v + \tau_h}{2}}\left(s_{\text{test}} - \frac{\tau_v s_{\text{ref},v,0} + \tau_h s_{\text{ref},h,0}}{\tau_v + \tau_h}\right)\right).$$

5 Prior Attraction

For this second example, we place ourselves in the experimental context of Stocker and Simoncelli (2006) that is a human observer has to compare the speed of moving textures. The experimenter is controlling the level of contrast of the texture which tends to increase the perceptual noise. Again, we present in the following the Bayesian observer model that is used by Stocker and Simoncelli (2006). We suppose that the Bayesian observer makes visual measurements M of the speed of the speed V. The distribution of the measurements knowing the stimulus speed V is Gaussian with mean V and standard deviation σ. In natural viewing conditions nothing is really moving, most motion in the visual field is in fact due to the observer self-motion. Objects are therefore expected to move slowly, hence we assumed that the Bayesian observer has a prior for slow speed, say a Laplace distribution with parameter $a > 0$ (*i.e.* density $v \mapsto e^{-a|v|}$).

Estimation Thanks to Bayes rule, the posterior density about the speed of the moving texture is

$$\mathrm{f}_{V|M}(v|m) = \frac{\mathrm{f}_{M|V}(m|v)\mathrm{f}_V(v)}{\mathrm{f}_M(m)} \propto \mathrm{f}_{M|V}(m|v)\mathrm{f}_V(v).$$

Let us use the MAP estimator. Again we will minimize the negative log-posterior which is similar but easier. We write

$$\ell(v) = -\ln\left(\mathrm{f}_{V|M}(v|m)\right) = -\ln\left(\mathrm{f}_{M|V}(m|v)\right) - \ln\left(\mathrm{f}_V(v)\right) = \frac{(m-v)^2}{2\sigma^2} + a|v| + C(\sigma).$$

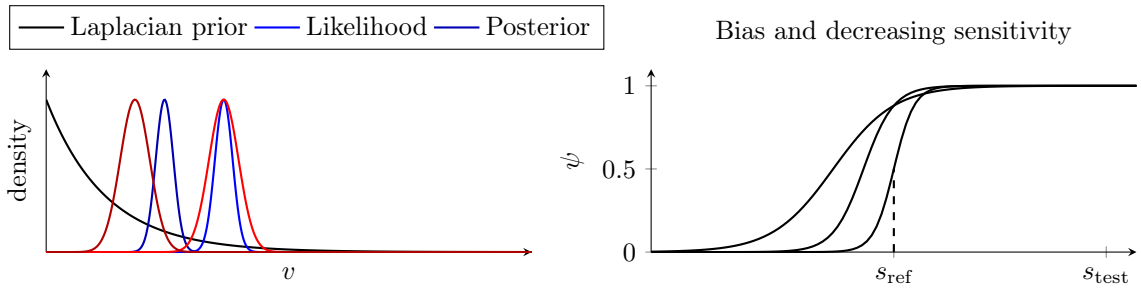

Figure 3. Left : Bayesian estimation with a prior for low values, the more the likelihood uncertainty the more the prior attracts the likelihood. Right : Consequences on the psychometric functions.

We assume in addition that $v > 0$ and that $v \gg \sigma$. We can now compute the derivative to find its zeroes, we obtain

$$-\frac{m-\hat{v}}{\sigma^2} + a = 0 \Longrightarrow \hat{v} = m - a\sigma^2.$$

This equation means that the estimated speed is biased toward slow speeds (negative sign with $a > 0$ *i.e.* underestimation of the speed). We can rewrite this equation with random variables $\hat{V} = M - a\sigma^2$. Therefore $\hat{V}|V$ is a Gaussian random variable with mean $\mu_{\hat{V}}$ and variance $\sigma^2_{\hat{V}}$ given by

$$\mu_{\hat{V}}(V) = V - a\sigma^2 \quad \text{and} \quad \sigma^2_{\hat{V}}(V) = \sigma^2.$$

Knowing the distribution of $\hat{V}|V$ we can now move to the decision paragraph.

Decision Similarly to Ernst and Banks 2002, Stocker and Simoncelli 2006 asked the participants to repeatedly compare pairs of stimuli in a two-alternative forced choice paradigm. We can therefore evaluate how the Bayesian behavior is reflected in this experiment. There is still a reference stimulus v_{ref} and a test stimulus v_{test} that are both estimated by the observer. Hence for any pair of test and reference stimuli, under our hypotheses (Gaussian/Laplacian) we have

$$\psi(v_{\text{ref}}, v_{\text{test}}) = \Psi\left(\frac{\mu_{\hat{V}_{\text{test}}}(v_{\text{test}}) - \mu_{\hat{V}_{\text{ref}}}(v_{\text{ref}})}{\sqrt{\sigma^2_{\hat{V}_{\text{test}}} + \sigma^2_{\hat{V}_{\text{ref}}}}}\right) = \Psi\left(\frac{v_{\text{test}} - v_{\text{ref}} + a(\sigma^2_{\text{ref}} - \sigma^2_{\text{test}})}{\sqrt{\sigma^2_{\text{test}} + \sigma^2_{\text{ref}}}}\right).$$

Therefore, using different contrasts for the reference and test stimuli generates a bias and a change in sensitivity that is visible in the psychometric function.

References

Ernst, M. O. and M. S. Banks (2002). "Humans integrate visual and haptic information in a statistically optimal fashion". In: *Nature* 415.6870, pp. 429–433.

Stocker, A. A. and E. P. Simoncelli (2006). "Noise characteristics and prior expectations in human visual speed perception". In: *Nature neuroscience* 9.4, pp. 578–585.

Thurstone, L. L. (1927). "A law of comparative judgment." In: *Psychological review* 34.4, p. 273.