Expert elicitation and stochastic prior modeling of uncertain inputs

A rationale and some recipes

nicolas.bousquet@upmc.fr

ETICS 2017 thematics : Uncertainty in Scientific Computing

A key sub-theme : modeling sources of uncertainty with various theoretical tools and applied methodologies

- Probability Theory
- Imprecise Probabilities Theory

Another key sub-theme in scientific computing is the use of expert judgment for assessing uncertain information when there is a lack of experimental data (or other objective source of information)

This lesson is driven by the questions :

Why and how stochastic modeling can be a relevant tool for using expert judgment, and more generally dealing with epistemic uncertainty ?

ETICS 2017 thematics : Uncertainty in Scientific Computing

A key sub-theme : modeling sources of uncertainty with various theoretical tools and applied methodologies

- Probability Theory
- Imprecise Probabilities Theory

Another key sub-theme in scientific computing is the use of expert judgment for assessing uncertain information when there is a lack of experimental data (or other objective source of information)

This lesson is driven by the questions :

Why and how stochastic modeling can be a relevant tool for using expert judgment, and more generally dealing with epistemic uncertainty ?

ETICS 2017 thematics : Uncertainty in Scientific Computing

A key sub-theme : modeling sources of uncertainty with various theoretical tools and applied methodologies

- Probability Theory
- Imprecise Probabilities Theory

Another key sub-theme in scientific computing is the use of expert judgment for assessing uncertain information when there is a lack of experimental data (or other objective source of information)

This lesson is driven by the questions :

Why and how stochastic modeling can be a relevant tool for using expert judgment, and more generally dealing with epistemic uncertainty ?

ETICS 2017 thematics : Uncertainty in Scientific Computing

A key sub-theme : modeling sources of uncertainty with various theoretical tools and applied methodologies

- Probability Theory
- Imprecise Probabilities Theory

Another key sub-theme in scientific computing is the use of expert judgment for assessing uncertain information when there is a lack of experimental data (or other objective source of information)

This lesson is driven by the questions :

Why and how stochastic modeling can be a relevant tool for using expert judgment, and more generally dealing with epistemic uncertainty ?

Expert judgment

Beyond scientific computing and uncertainty propagation, expert judgment has a foremost role in decision-making

-
-
-

Beyond scientific computing and uncertainty propagation, expert judgment has a foremost role in decision-making

- guiding designs of experiments, ordering scientific results [4, 37, 20]
-
-

Beyond scientific computing and uncertainty propagation, expert judgment has a foremost role in decision-making

- guiding designs of experiments, ordering scientific results [4, 37, 20]
- enriching economic [19] and actuarial studies [34] on the impact of financial risks
-

Beyond scientific computing and uncertainty propagation, expert judgment has a foremost role in decision-making

- guiding designs of experiments, ordering scientific results [4, 37, 20]
- enriching economic [19] and actuarial studies [34] on the impact of financial risks
- being determining in legal arbitration, public policies [22] or environmental governance [21, 8]

Beyond scientific computing and uncertainty propagation, expert judgment has a foremost role in decision-making

- guiding designs of experiments, ordering scientific results [4, 37, 20]
- enriching economic [19] and actuarial studies [34] on the impact of financial risks
- being determining in legal arbitration, public policies [22] or environmental governance [21, 8]

Its influence on technological, economic, societal or personal choices when elaborating strategies of gain-winning is explored by many epistemological and psychological authors [12, 20, 11]

What is an expert ?

An infinite number of conceptions

Among them, two main kinds of experts for [37] :

1. (performative expertise)
2.

An infinite number of conceptions

Among them, two main kinds of experts for [37] :

1. *those who expertise is a function of what they do (performative expertise)*
2.

An infinite number of conceptions

Among them, two main kinds of experts for [37] :

1. *those who expertise is a function of what they do (performative expertise)*
2. *those who expertise is a function of what they know (epistemic expertise)*

An infinite number of conceptions

Among them, two main kinds of experts for [37] :

1. *those who expertise is a function of what they do (performative expertise)*
2. *those who expertise is a function of what they know (epistemic expertise)*

An infinite number of conceptions

Among them, two main kinds of experts for [37] :

1. *those who expertise is a function of what they do (performative expertise)*
2. *those who expertise is a function of what they know (epistemic expertise)*

An usual view, with the ability of explaining and transmitting. Furthermore, acccording to Luntley [20] :

I argue that what differentiates the epistemic standpoint of experts is not what or how they know [...], but their capacity for learning

What is an expert ?

Today's question is in fact "what is *formally* an expert ?"

We should rather talk about **"expert systems delivering new knowledge"**

Typically :

- implicit cognitive systems
  - humans
  - some artificial intelligences
- explicit causal systems
  - phenomenological models and their numerical implementation (simulation models)

Capacity for proving expertness $\Leftrightarrow$ capacity of predicting adequately

Capacity for learning $\Leftrightarrow$ capacity of inferring (processing) coherently when new data arrive

Eliciting = assessing her/his/its relevant epistemic information on the behavior of a magnitude of interest $X \in \chi$

elicio, eliciere : to extract from, to drawout (*ex aliquo verbum elicere*)

Immediate difficulties

- bias
- impact of subjectivity in the delivery process
- lack of correct or sharp information
- ...

resulting in epistemic uncertainty

Our work : formalizing the most adapted measure of uncertainty, highlighting clearly the subjective and objective parts of the modeling

Eliciting = assessing her/his/its relevant epistemic information on the behavior of a magnitude of interest $X \in \chi$

elicio, eliciere : to extract from, to drawout (*ex aliquo verbum elicere*)

Immediate difficulties

- bias
- impact of subjectivity in the delivery process
- lack of correct or sharp information
- ...

resulting in epistemic uncertainty

Our work : formalizing the most adapted measure of uncertainty, highlighting clearly the subjective and objective parts of the modeling

Prior information = *information whose the value of truth is justified by considerations independent on experiment on focus* [25]

- other trial results (e.g., on mock-ups)
- technical running specification
- physical bounds
- literature corpus
- and of course, human experts

Often blueincomplete, always blueuncertain, because

- of the non-existence of a system allowing a priori if the expertness is complete or not
- of the non-existence of a system precise enough to specify that $X = x_0$ exactly (except in rare cases)

What means “uncertainty" and especially “epistemic uncertainty" ?

Why probabilities for dealing with uncertainty ?

If we are ok with probabilities, how choosing the probability distributions ?

What means “uncertainty" and especially “epistemic uncertainty" ?
Hard philosophical question ! Providing answering attempts here

Why probabilities for dealing with uncertainty ?

If we are ok with probabilities, how choosing the probability distributions ?

What means “uncertainty" and especially “epistemic uncertainty" ?
Hard philosophical question ! Providing answering attempts here

Why probabilities for dealing with uncertainty ?
Many practical advantages, but how proving they are theoretically relevant ?

If we are ok with probabilities, how choosing the probability distributions ?

What means “uncertainty" and especially “epistemic uncertainty" ?
Hard philosophical question ! Providing answering attempts here

Why probabilities for dealing with uncertainty ?
Many practical advantages, but how proving they are theoretically relevant ?
Raises the question of **auditability** of mathematical procedures = growingly increasing concern

If we are ok with probabilities, how choosing the probability distributions ?

What means “uncertainty" and especially “epistemic uncertainty" ?
Hard philosophical question ! Providing answering attempts here

Why probabilities for dealing with uncertainty ?
Many practical advantages, but how proving they are theoretically relevant ?
Raises the question of **auditability** of mathematical procedures = growingly increasing concern

If we are ok with probabilities, how choosing the probability distributions ?
Use the help of important Bayesian prior modeling techniques

Outline

1. Some arguments in favor of probabilities to deal with expert (and more generally) epistemic uncertainty
2. Some methodological aspects of stochastic modeling for prior elicitation

A rationale for the choice of probabilities to deal with epistemic uncertainty of expert systems

Treating prior information from implicit cognitive systems

If we were omniscient, a causal model could be

$$X = g(Z)$$

where :

- Z is a hidden property of the experiment
- g is a model of information production

The value of Z could be explained by another transformation $\tilde{g}$ of another hidden property $\tilde{\theta}$, etc.

However, there is still a model error between the true values of X and $g(Z)$, since nor g neither Z are known (completely or not)

Hypothesis 1 (epistemological) by Lakatos [18]

- Information on the world is hidden and partially revealed by a consensual theory (*in the sense of Popper [26] : by mutual decision of protagonists*) defining objectivity [13]
- Knowledge is “filtered" from information
- Filtering is performed through the intervention of symbols, or signs, in order to transmit it or even implement it

Hypothesis 2 (arising from neurosciences) [29, 28, 27, 15, 7, 3]

- Face to situations where uncertain information is mobilized, human reasoning produces probabilistic inferences
- Difficulties appear when trying to explicit this inferred knowledge by an interpretative language $\Rightarrow$ providing usable expertness

Hypothesis 1 (epistemological) by Lakatos [18]

- Information on the world is hidden and partially revealed by a consensual theory (*in the sense of Popper [26] : by mutual decision of protagonists*) defining objectivity [13]
- Knowledge is “filtered" from information
- Filtering is performed through the intervention of symbols, or signs, in order to transmit it or even implement it

Hypothesis 2 (arising from neurosciences) [29, 28, 27, 15, 7, 3]

- Face to situations where uncertain information is mobilized, human reasoning produces probabilistic inferences
- Difficulties appear when trying to explicit this inferred knowledge by an interpretative language $\Rightarrow$ providing usable expertness

We don't know what is the "deconvolution" transforming uncertain knowledge backwards into uncertain information, following Lakatos' hypothesis

But we can have ideas about the impact of the addition of uncertain but useful knowledge in the problem of determining X

It should traduce by the increasing of information on X = inference (updating)

$\Rightarrow$ this inference should stands on a reasoning principle

$\Rightarrow$ this principle should stand on a logic = set of **formal rules**

We don't know what is the "deconvolution" transforming uncertain knowledge backwards into uncertain information, following Lakatos' hypothesis

But we can have ideas about the impact of the addition of uncertain but useful knowledge in the problem of determining X

Desirable properties [35]

- Sorting *atomic* assertions of type $X = x_0$ at each addition of information (*exclusive logic*)
  - an initial situation (premise) is less informative than a conclusion (updating)
- Allowing *uncertain* information
  - not only true or false situations can be sorted (*non-boolean logic*)

Definition [35]

Denote S_X a set of atomic propositions of type $X = x_i$. The set B_X of all possible *compound propositions* generated by

$$\neg X = x_i, \quad X = x_i \wedge X = x_j,$$
$$X = x_i \vee X = x_j, \quad X = x_i \Rightarrow X = x_j$$
$$\text{and} \quad X = x_i \Leftrightarrow X = x_j$$

is called a state of information, with $\mathrm{Dom}(B_X) =$ logical closure of S_X

The state of information B_X summarizes the existing information on a set of propositions about X

The same logic should guide how B_X evolves : it is growing following a given metric when information on X is increasing

Definition [35]

Denote S_X a set of atomic propositions of type $X = x_i$. The set B_X of all possible *compound propositions* generated by

$$\begin{aligned} \neg X = x_i, \quad & X = x_i \wedge X = x_j, \\ X = x_i \vee X = x_j, \quad & X = x_i \Rightarrow X = x_j \\ \text{and} \quad & X = x_i \Leftrightarrow X = x_j \end{aligned}$$

is called a state of information, with $\mathrm{Dom}(B_X) =$ logical closure of S_X

The state of information B_X summarizes the existing information on a set of propositions about X

The same logic should guide how B_X evolves : it is growing following a given metric when information on X is increasing

Definition [35]

Denote S_X a set of atomic propositions of type $X = x_i$. The set B_X of all possible *compound propositions* generated by

$$\begin{aligned} \neg X = x_i, \quad & X = x_i \wedge X = x_j, \\ X = x_i \vee X = x_j, \quad & X = x_i \Rightarrow X = x_j \\ \text{and} \quad & X = x_i \Leftrightarrow X = x_j \end{aligned}$$

is called a state of information, with $\mathrm{Dom}(B_X) =$ logical closure of S_X

The state of information B_X summarizes the existing information on a set of propositions about X

The same logic should guide how B_X evolves : it is growing following a given metric when information on X is increasing

Definition

Consider any proposition A on X. Given B_X, the plausibility $[A|B_X]$ is a single real number, upperly bounded by a real (finite or infinite) T

- **Consistency** : B_X is consistent if there is no proposition A for which both $[A|B_X] = T$ and $\neg[A|B_X] = T$
- **Propositional calculus** :

  (i) If $A = A'$ then $[A|B_X] \Leftrightarrow [A'|B_X]$
  (ii) $[A|B_X, C_X, D_X] = [A|(B_X \wedge C_X), D_X]$
  (iii) If B_X consistent and $\neg[A|B_X] < T$, then $A \cup B_X$ is consistent

- **Coherence** : there exists a non-increasing function S_0 such that, for all x and consistent B_X

$$\neg[A|B_X] = S_0([A|B_X])$$

- **Density** : the set $[S_0(T), T]$ admits a non-void, dense and consistent subset

Definition

Consider any proposition A on X. Given B_X, the plausibility $[A|B_X]$ is a single real number, upperly bounded by a real (finite or infinite) T

- **Consistency** : B_X is consistent if there is no proposition A for which both $[A|B_X] = T$ and $\neg[A|B_X] = T$
- **Propositional calculus** :

  (i) If $A = A'$ then $[A|B_X] \Leftrightarrow [A'|B_X]$
  (ii) $[A|B_X, C_X, D_X] = [A|(B_X \wedge C_X), D_X]$
  (iii) If B_X consistent and $\neg[A|B_X] < T$, then $A \cup B_X$ is consistent

- **Coherence** : there exists a non-increasing function S_0 such that, for all x and consistent B_X

$$\neg[A|B_X] = S_0([A|B_X])$$

- **Density** : the set $[S_0(T), T]$ admits a non-void, dense and consistent subset

Definition

Consider any proposition A on X. Given B_X, the plausibility $[A|B_X]$ is a single real number, upperly bounded by a real (finite or infinite) T

- **Consistency** : B_X is consistent if there is no proposition A for which both $[A|B_X] = T$ and $\neg[A|B_X] = T$
- **Propositional calculus** : applicable to any problem domain for which we can formulate useful propositions

  (i) If $A = A'$ then $[A|B_X] \Leftrightarrow [A'|B_X]$
  (ii) $[A|B_X, C_X, D_X] = [A|(B_X \wedge C_X), D_X]$
  (iii) If B_X consistent and $\neg[A|B_X] < T$, then $A \cup B_X$ is consistent

- **Coherence** : there exists a non-increasing function S_0 such that, for all x and consistent B_X

$$\neg[A|B_X] = S_0([A|B_X])$$

- **Density** : the set $[S_0(T), T]$ admits a non-void, dense and consistent subset

Definition

Consider any proposition A on X. Given B_X, the plausibility $[A|B_X]$ is a single real number, upperly bounded by a real (finite or infinite) T

- **Consistency** : B_X is consistent if there is no proposition A for which both $[A|B_X] = T$ and $\neg[A|B_X] = T$
- **Propositional calculus** :

  (i) If $A = A'$ then $[A|B_X] \Leftrightarrow [A'|B_X]$
  (ii) $[A|B_X, C_X, D_X] = [A|(B_X \wedge C_X), D_X]$
  (iii) If B_X consistent and $\neg[A|B_X] < T$, then $A \cup B_X$ is consistent

- **Coherence** : there exists a non-increasing function S_0 such that, for all x and consistent B_X

$$\neg[A|B_X] = S_0([A|B_X])$$

- **Density** : the set $[S_0(T), T]$ admits a non-void, dense and consistent subset

Definition

Consider any proposition A on X. Given B_X, the plausibility $[A|B_X]$ is a single real number, upperly bounded by a real (finite or infinite) T

- **Consistency** : B_X is consistent if there is no proposition A for which both $[A|B_X] = T$ and $\neg[A|B_X] = T$
- **Propositional calculus** :

  (i) If $A = A'$ then $[A|B_X] \Leftrightarrow [A'|B_X]$
  (ii) $[A|B_X, C_X, D_X] = [A|(B_X \wedge C_X), D_X]$
  (iii) If B_X consistent and $\neg[A|B_X] < T$, then $A \cup B_X$ is consistent

- **Coherence** : there exists a non-increasing function S_0 such that, for all x and consistent B_X

$$\neg[A|B_X] = S_0([A|B_X])$$

- **Density** : the set $[S_0(T), T]$ admits a non-void, dense and consistent subset

Axiom

Consider any proposition A on X. Given B_X, the plausibility $[A|B_X]$ is a single real number, upperly bounded by a real (finite or infinite) T

This axiom of non-ambiguity is particularly important

This is an assumption of *universal comparability*

Consequence : an additional information (not a knowledge) can only increase or decrease the plausibility of a proposition

Axiom

Consider any proposition A on X. Given B_X, the plausibility $[A|B_X]$ is a single real number, upperly bounded by a real (finite or infinite) T

This axiom of non-ambiguity is particularly important

This is an assumption of *universal comparability*

As seen later, the differences between probabilistic logic and extra-probabilistic logics arises from the agreement or disagreement with this assumption

Jaynes [17] argues for its validity on pragmatic grounds

Axiom

Consider any proposition A on X. Given B_X, the plausibility $[A|B_X]$ is a single real number, upperly bounded by a real (finite or infinite) T

It is supported when we talk about quantities X with physical meanings and taking a unique value at each instant (possibly given a finite measurement precision)

It may be not supported if we talk about :

- magnitudes considered at the quantum scale (e.g., in neutronics)
- imaginary magnitudes (e.g., latent variables)

Remember that we are dealing with objective information on X, not interpreted knowledge !

Statement

Density : the set $[S_0(T), T]$ admits a non-void, dense and consistent subset

Can be false when the set of all propositions is finite (e.g., discrete and bounded) [16]

Could be partially removed by arguments provided by Snow [32], plaiding for infinite gradations of plausibility within even a single, finite domain

The source of an objective rational measure of belief is external to the cognitive apparatus of the believer. Its value is determined by the vagaries of the real world or by some idealized model of the world. There is no way to tell in advance just which values must arise, and each value may be graduated with arbitrary precision. Any such value can simply be adopted by the believer without recourse to unboundedly precise discrimination between affective states related to credibility... [32]

Working (as usually) with uncountable input spaces for X is not an issue :-)

1. Reproductibility rule : two equivalent assertions about X have the same plausibility
2. Non-contradiction rule : if it exists several ways of coming to the same conclusion about X, all have the same plausibility
3. Consistency rule : the logic cannot reach a conclusion contradicted by the common deductive rules (*e.g.*, *transitivity*)
4. Integrity rule : the logic cannot disregard a part of information to reach to a conclusion about X to come to a conclusion
5. Monotony rule : the plausibility of the non-exclusive union of two assertions is at least equal to the upper plausibility of each
6. Product rule : the plausibility of the intersection of two assertions is at most equal to the lower plausibility of each

Originally proven (erroneously) by Cox [5], corrected by Jaynes [17], extended more rigorously by Paris [24], Van Horn [35] Dupré and Tipler [10] (among others) then finalized by Terenin and Draper [33]

Theorem

Under the previous assumptions, there exists a continuous, increasing function $\mathbb{P}$ such that, for every proposition A, C and consistent B_X,

(i) $\mathbb{P}([A|B_X]) = 0$ iif A is known to be false given the information in X

(ii) $\mathbb{P}([A|B_X]) = 1$ iif A is known to be true given the information in X

(iii) $0 \leq \mathbb{P}([A|B_X]) \leq 1$

(iv) $\mathbb{P}([A \wedge C|B_X]) = \mathbb{P}([A|B_X])\mathbb{P}([C|A, B_X])$

(v) $\mathbb{P}(\neg[A|B_X]) = 1 - \mathbb{P}([A|B_X])$

Any system of plausible reasoning, under the previous assumptions, is isomorphic to probability theory

Goertzel [14] proved that if the consistency rule is weakened, then plausibilities behave approximately like probabilities

The probability theory is relevant to account for uncertainties on a subject explored by a cognitive system (human or machine) which could be not completely consistent

Numerous authors in artificial intelligence [36], epistemology [1] or cognitive sciences [6] recognize the practical relevance of this axiomatic for extracting or updating information, using Bayes rule

Axiom

Consider any proposition A on X. Given B_X, the plausibility $[A|B_X]$ is a single real number, upperly bounded by a real (finite or infinite) T

Its mot common “relaxation" is the assumption that two dimensions are required to represent correctly the plausibility of a proposition

At the origin of belief theory [30, 31] and possibility theory [9]

Experiments show that such a relaxation is clearly supported when the plausibility is understood as the summary of a belief, or a *gamble* [35]

Nonetheless, this “relaxation" remains arbitrary, and usually stands on an interpretation of *the nature of knowledge* (expressed through a language), and not of the *nature of information* (expressed by physical reality or an idealized model of the reality) [32]

Treating uncertain prior information from causal models

Practical models used by engineers (e.g., implemented computer codes Σ''') can produce prior simulations of a phenomenon Σ

$$\underset{\Sigma}{\text{Real phenomenon}} \rightarrow \underset{\Sigma'}{\text{Theoretical model}} \rightarrow \underset{\Sigma''}{\text{Algorithmic model}} \rightarrow \underset{\Sigma'''}{\text{Implemented model}}$$

We want to define what is the conceptual nature of model uncertainty affecting Σ'''

We could ask the question otherwise : what is the conceptual nature of reduction of model uncertainty ?

We need also to define Σ'''

What is Σ''' in usual cases ?

Program. Sequence of operations and instructions

Algorithm. Finite and non-ambiguous sequence of operations and instructions allowing for solving a problem that can be solved exhaustively

Self-delimiting program. A program that ends. Its ending is a command of the program itself

Happens at step Σ''

- Refining the algorithmic description Σ'' by adding new parameters and/or structural equations, necessarily based on improvement of Σ'
- Refining the execution of Σ''' (e.g., improving a tolerance)

Reducing model uncertainty implies to reduce model error

Maybe the nature of model error could say something about the nature of model uncertainty ?

We consider an illustrative example

Consider a real phenomenon Σ with output Y described by

$$\begin{array}{rcl} \chi \times \chi_Z & \rightarrow & \Upsilon \\ \Sigma : X, Z & \mapsto & Y \end{array}$$

where X are known and treated variables, and Z are unknown or untreated variables

Consider a self-delimiting, calculable model of Σ

$$\begin{array}{rcl} \chi_d & \rightarrow \Upsilon_d \\ \Sigma''' : X'' & \mapsto Y'' \end{array}$$

where

- $\chi_d \subsetneq \chi$ is the subset of χ that can be reached by a calculus
- $\Sigma(\chi_d, \chi_Z) = \Upsilon_d$ (Galerkin problem solving)

Assume the following hypotheses

$$\begin{array}{lcl} \text{(H1)} & : & \text{Card}(\chi_d) < \infty, \\ \text{(H2)} & : & \chi_Z \text{ is countable and } \text{Card}(\chi_Z) < \infty. \end{array}$$

Assume that Υ_d is a metric space

It is possible to define a model error $\delta(x, z)$, through a measure $\mathcal{D}$ such that, for all couple $(x, z) \in \chi_d \times \chi_Z$,

$$\delta(x, z) \quad = \quad \mathcal{D}\left\{\Sigma''(x), \Sigma(x, z)\right\} \ \geq \ 0$$

with $\delta(x, z) = 0$ iif $\Sigma'''(x) = \Sigma(x, z)$

Proposition [B. and Denis 2017]

The model error $\delta(x, z)$ cannot be calculated $\forall (x, z) \in \chi_d \times \chi_Z$

Proof : based on tools of computational complexity theory

A more general result can be proved using Turing's machines

The previous proposition (and its extensions) indicate that no algorithm is able to compute all the values of the model error $\delta(x, z)$

- We cannot prove that the error never exists
- Being cautious, we assume its existence

What would be the nature of the best reachable (computable) approximation $\tilde{\delta}(x, z)$ of $\delta(x, z)$?

- $\tilde{\delta}(x, z)$ should be computed by a self-delimiting program
- however there is no recursive function allowing to predict the next value of $\tilde{\delta}(x', z')$ at (x', z')

It comes that any finite sequence of $\tilde{\delta}(x_i, z_i)$ is exhaustively described only by itself

The adapted formalism to describe this property is the following

Kolmogorov's algorithmic complexity

Kolmogorov's complexity $H(s)$ of a program producing a sequence s is the length of the smallest program required to generate s.

A consequence of the impossibility of compressing the information in the sequence of $\tilde{\delta}(x_i, z_i)$ is the following : $\exists c \in \boldsymbol{R}$ such that

$$H(\tilde{\delta}(x_1, z_1), \ldots, \tilde{\delta}(x_n, z_n)) \quad \geq \quad n - c. \tag{1}$$

Result (1) implies that the sequence $\tilde{\delta}(x_i, z_i)$ is in the sense of Chaitin-Levin

Proposition (B. and Denis 2017)

The best computable approximation of model error is random.

Randomness contamines the nature of all concepts incorporating model error

It is arguable to use probabilities for modeling epistemic model uncertainty

Stochastic prior modeling : examples and recipes

We got a rationale for choosing probabilities as relevant tools for yielding uncertain information

The aim of this second part of the lesson is for exploring various methodological approaches to stochastic prior modeling

No reference corpus available ! (not an easy journey)

We will sometimes consider alternatively two situations : quantification of uncertainties and propagation of uncertainties (and take several examples)

L'élicitation est la représentation, par des moyens mathématiques, de l'ontologie des connaissances utiles pour résoudre un problème.

Marc Sancandi, CEA-CESTA, 2011.

In the following, we denote Π and π, respectively, the distribution and density function of a random variable $\theta \in \Theta$

θ is the quantity for which prior uncertain information is provided, directly or indirectly

The notation X will be used too, but possibly not the same X than in the first part of the lesson

Under the prism of scientific computing, we start with the following simple example

Prior information on θ is defined by two deterministic bounds :

$$\Pi(\theta \in [\theta_{\min}, \theta_{\max}]) \quad = \quad 1$$

then

$$\Theta = [\theta_{\min}, \theta_{\max}]$$

and we want to propagate the uncertainty on θ onto X through the deterministic relation

$$X \quad = \quad h(\theta)$$

Which Π to choose ?

We need a rationale for eliciting $\Pi(\theta)$

Laplace's principle of insufficient reason (1773)

In absence of information, all elementary events of a finite Θ are equiprobables, and the same weight must be given to each possible value

Following this principle, $\Pi(\theta)$ should be chosen uniform in $[x_{\mathsf{min}}, x_{\mathsf{max}}]$

This choice is often done in practice in problems of simulation under uncertainty

It is (very) wrong, and you should (a priori) burn in hell for this !

1 - Partitioning paradox.
It is inconsistent to apply the rule to all coarsening and refinings of the parameter space simultaneously

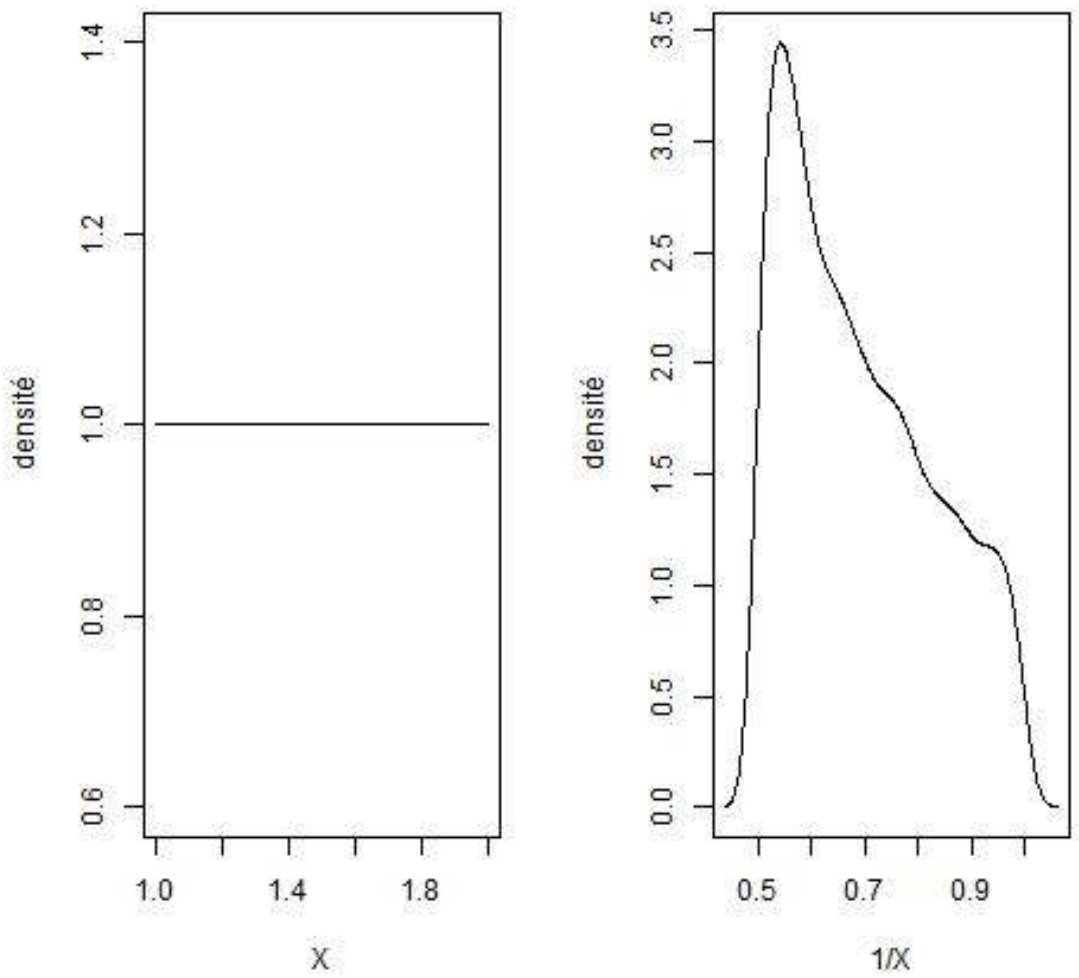
Shafer's example (1976)

Let $\Theta = \{\theta_1, \theta_2\}$, where θ_1 denotes the event that there is life in orbit about the star Sirius and θ_2 denotes the event that there is not. Laplace's rule gives $\mathbb{P}(\theta_1) = \mathbb{P}(\theta_1) = 1/2$.

But now let $R = \{\omega_1, \omega_2, \omega_3\}$ where ω_1 denotes the event that there is life around Sirius, ω_2 denotes the event that there are planets but no life, and ω_3 denotes the event that there are no planets. Then Laplace's rule gives $\mathbb{P}(\omega_1) = \mathbb{P}(\omega_2) = \mathbb{P}(\omega_3) = 1/3$.

The paradox is that the probability of life is $\mathbb{P}(\theta_1) = 1/2$ if we adopt the first formulation, but is $\mathbb{P}(\omega_1) = 1/3$ if we adopt the second formulation

2 - Non-invariance of information.



How avoiding this ?

Consider the enveloppe model

$$X = h(\theta) + \epsilon$$

where ϵ is a known random noise and X^* is an observation

It is clearly enveloppe from the point of view of propagating uncertainties : obviously, even if ϵ is very small, the variance of X increases

How avoiding this ?

Consider the enveloppe model

$$X = h(\theta) + \epsilon$$

where ϵ is a known random noise and X^* is an observation

It is clearly enveloppe from the point of view of propagating uncertainties : obviously, even if ϵ is very small, the variance of X increases

How avoiding this ?

Consider the enveloppe model

$$X \quad = \quad h(\theta) + \epsilon$$

where ϵ is a known random noise and X^* is an observation

It is clearly enveloppe from the point of view of propagating uncertainties : obviously, even if ϵ is very small, the variance of X increases

More generally, consider

$$X \quad = \quad h(\theta, \epsilon)$$

which can rewritten under the classical form

$$X \quad \sim \quad f(x|\theta)$$

f being determined by h and ϵ

X was random thanks to the action of θ, but now $X|\theta$ is still random (because of ϵ)

Example : $X = -\theta^{-1}\log(1-\epsilon) \sim \mathcal{E}(\theta)$ if $\epsilon \sim \mathcal{U}[0,1]$

How choosing $\Pi(\theta)$ to conserve the invariance of information ?

More generally, consider

$$X \quad = \quad h(\theta, \epsilon)$$

which can rewritten under the classical form

$$X \quad \sim \quad f(x|\theta)$$

f being determined by h and ϵ

X was random thanks to the action of θ, but now $X|\theta$ is still random (because of ϵ)

Example : $X = -\theta^{-1} \log(1 - \epsilon) \sim \mathcal{E}(\theta)$ if $\epsilon \sim \mathcal{U}[0, 1]$

How choosing $\Pi(\theta)$ to conserve the invariance of information ?

More generally, consider

$$X \quad = \quad h(\theta, \epsilon)$$

which can rewritten under the classical form

$$X \quad \sim \quad f(x|\theta)$$

f being determined by h and ϵ

X was random thanks to the action of θ, but now $X|\theta$ is still random (because of ϵ)

Example : $X = -\theta^{-1} \log(1 - \epsilon) \sim \mathcal{E}(\theta)$ if $\epsilon \sim \mathcal{U}[0, 1]$

How choosing $\Pi(\theta)$ to conserve the invariance of information ?

Assume to have no prior information (apart possibly the bounds on θ)

Principle of parameterization invariance Transforming θ into $\eta = g(\theta)$ through a bijection g, the prior information still do not exist and nothing should be modified

One has

$$\pi^*(\eta) \quad = \quad \left|Jac(g^{-1}(\eta))\right| \pi(g^{-1}(\eta)) \quad = \quad \left|\det \frac{\partial \eta}{\partial \theta}\right| \pi(g^{-1}(\eta))$$

which (usually) does not stay constant if $\pi(\theta) = 1$

Example : $\eta = -\log(1-\theta) \sim \mathcal{E}(1)$ if $\theta \sim \mathcal{U}[0,1]$

Example 1 : location parameter. If one may write $f(x|\theta) = f(x-\theta)$

- the family f is *invariant by translation* : if $x \sim f$, then $y = x - x_0 \sim f \ \forall x_0$
- it is required that $\pi(\theta)$ be invariant by translation too :

$$\pi(\theta) \quad = \quad \pi(\theta - \theta_0) \qquad \forall \theta_0$$

This rule leads to uniform distribution over Θ

Example 2 : scale parameter. If one may write $f(x|\theta) = \frac{1}{\theta} f(x/\theta)$ with $\theta > 0$

- the family f is *invariant by scale change* : $y = x/\theta_0 \sim f \ \forall \theta_0 > 0$
- it is required that the prior distribution satisfies $\pi(A) = \pi(A/c)$ for any measurable set $A \in]0, +\infty[$ and $c > 0$

$$\pi(\theta) \quad = \quad \frac{1}{c}\pi\left(\frac{\theta}{c}\right)$$

which implies

$$\pi(\theta) \quad \propto \quad 1/\theta$$

We see that in this case, the invariance measure is no longer constant

These approaches imply to choose an invariance structure, in some arbitrary way

To avoid this choice, Jeffreys (1946) interested in the Fisher information matrix $I(\theta)$:

- let $\theta \in \Theta \subset \boldsymbol{R}^d$; the element $(i,j) \in \{1,\ldots,k\}^2$ of I_θ is

$$I_{ij}(\theta) \quad = \quad -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(x|\theta) \right]$$

(under regularity conditions ensuring the existence)

Jeffrey's prior $\pi(\theta) \propto \sqrt{\det I(\theta)}$

For any bijective variable change $\eta = g(\theta)$, one has

$$\pi(\eta) \quad \propto \quad \sqrt{\det I(\eta)}$$

Hence this prior satisfies an *intrinsic* invariance principle, for any prior parameterization choice

- $I(\theta)$ is widely accepted as an indicator of the quantity of information carried by the sampling model (or its average observation) on θ (Fisher, 1956)
- $I(\theta)$ measures the capacity of the sampling model to discriminate between θ and $\theta + / - d\theta$ via the mean slope of $\log f(x|\theta)$
- Favoring the values of θ for which $I(\theta)$ is high is equivalent to minimize the influence of the prior distribution

Unfortunately, Jeffreys' prior is often not a real probability measure (improper prior) : no possibility of simulating (for instance)

$$\int_{\Theta} \pi(\theta) \quad = \quad \infty$$

It can be proper (true probability measure) only if Θ is bounded or discrete

Assume to have data $\mathbf{x_n} = (x_1, \ldots, x_n)$

Using Bayes'rule, the prior measure on θ can be updated by conditioning and become a posterior measure

$$\pi(\theta|\mathbf{x_n}) \quad = \quad \frac{f(\mathbf{x_n}|\theta)\pi(\theta)}{\int f(\mathbf{x_n}|\theta)\pi(\theta)d\theta}$$

This is possible when $\pi(\theta)$ is improper, but for small dimensions of Θ only

Example : exponential lifetime distribution

Consider a $n-$sample from an exponential distribution

$$X \quad \sim \quad \mathcal{E}(\theta)$$

with density

$$f(x|\theta) \quad = \quad \theta \exp(-\theta x)$$

Then the Jeffreys' prior is

$$\pi(\theta) \quad \propto \quad 1/\theta$$

Consequently, the posterior distribution is proportional to $(\propto)$

$$\theta^{n-1} \exp\left(-\theta \sum_{i=1}^{n} x_i\right)$$

and we recognize the general term of the gamma $\mathcal{G}(n, n\bar{x}_n)$ distribution

Usually not so straightforward $\Rightarrow$ typical computational tools : Monte Carlo Markov Chains (MCMC)

Choose now a gamma distribution $\mathcal{G}(a, b)$ as a prior on θ, instead of Jeffreys's choice

$$\pi(\theta) \quad = \quad \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp(-b\theta) \mathbb{1}_{\{\theta \geq 0\}}$$

Then the posterior distribution, given the likelihood

$$f(\mathbf{x_n}|\theta) \quad = \quad \theta^n \exp\left(-\theta \sum_{i=1}^{n} x_i\right),$$

is also a gamma distribution :

$$\theta|\mathbf{x_n} \quad \sim \quad \mathcal{G}\left(a + n, b + \bar{x}_n\right)$$

Jeffreys' prior can be seen as a limiting case of a proper prior, here by choosing $a \to 0$ and $b \to 0$, according to a given topology [2]

To define a particular stochastic prior model $\pi(\theta)$ yielding information, we need a benchmark prior model $\pi^J(\theta)$ (Jeffreys) such that :

- it defines something like "the most objective prior form"
- it yields something like "the minimum amount of prior information"
- its posterior distribution $\pi^J(\theta|\mathbf{x_n})$ is nearly confounded with the distribution of an usual frequentist estimator of θ

There are many other benchmark (say, *noninformative*) priors

Principe

- The Kullback-Leibler divergence ("distance") between posterior and prior

$$KL(\pi, \mathbf{x_n}) \quad = \quad \int_\Theta \pi(\theta|\mathbf{x_n}) \log \frac{\pi(\theta|\mathbf{x_n})}{\pi(\theta)} \; d\mathbf{x_n}$$

  measures the information brought by observed data $\mathbf{x_n}$ on the modelling, independently of the parameterization choice θ

- The idea is to maximize $KL(\mathbf{x_n})$ in π for data $\mathbf{x_n}$ that can be *typically* observed : they are generated by the predictive prior distribution

$$f(\mathbf{x_n}) \quad = \quad \int_\Theta f(\mathbf{x_n}|\theta)\pi(\theta) \; d\theta$$

  and to avoid choosing a size n, let make it tend to ∞

$$\text{soit} \quad \pi^* \quad = \quad \arg\max_\pi \lim_{n\to\infty} \mathbb{E}_{f(\mathbf{x_n})} \left[KL(\pi, \mathbf{x_n})\right]$$

In dimension 1, it is the Jeffreys' prior

Can solve posterior inconsistency problems in higher dimensions

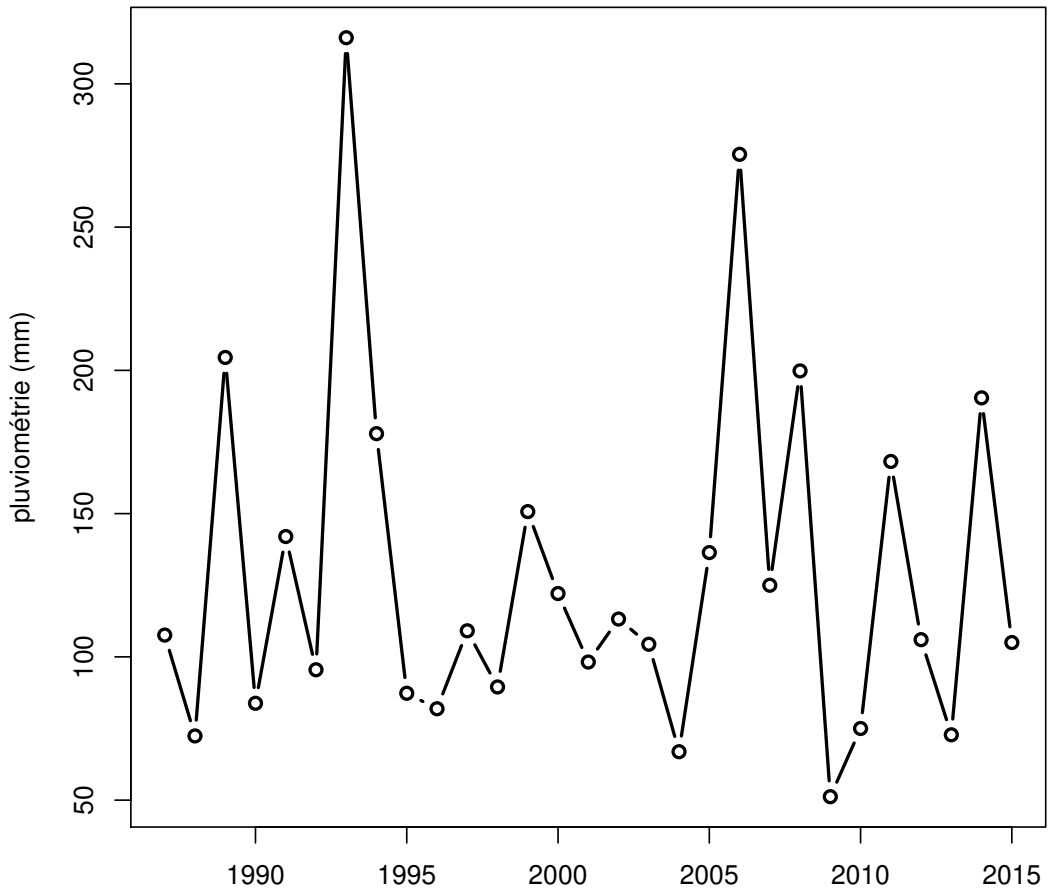
Methodology of prior elicitation much less automated than for Jeffreys

- Let $\theta_n(\alpha)$ be the $\alpha-$order posterior quantile
- $\forall \alpha \in [0,1]$, one may elicit a prior measure such that

$$\underbrace{P_\theta\left(\theta \leq \theta_n(\alpha)\right)}_{\text{frequentist probability}} = \underbrace{P\left(\theta \leq \theta_n(\alpha)|\mathbf{X_n}\right)}_{\text{Bayesian probability}} + \mathcal{O}(n^{-i/2}).$$

Frequentist coverage matching properties allow to discriminate between several benchmark priors

Rainfall annual maxima X at Penta-di-Casinca (Corsica)

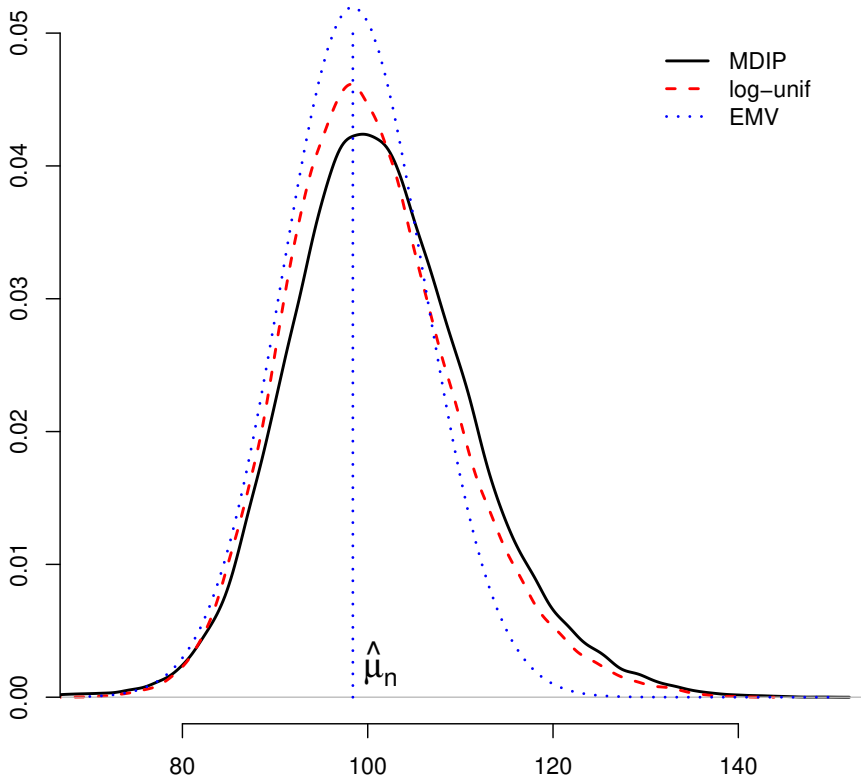


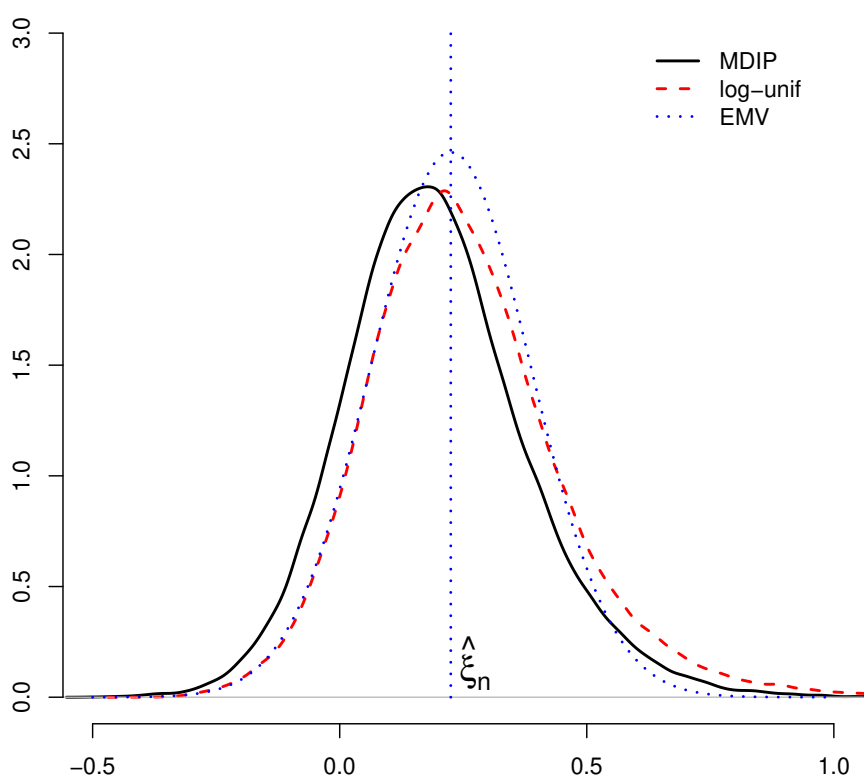
The statistical extreme value theory suggests to select $f(x|\theta)$ in the Generalized Extreme Value (GEV) family, with pdf

$$F(x) \quad = \quad \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}$$

and $[x]_+ = \max(x, 0)$

Action of two benchmark priors with extreme value models and comparison with MLE





Whatever the available information, establishing baseline (noninformative) priors for a Bayesian situation is a prerequisite (having a true distribution is not mandatory)

Objective Bayesian modeling, dedicated to find benchmark (noninformative) priors for statistical models, can provide good ideas

$\Rightarrow$ Exploring Bayesian elicitation

Eliciting a prior form from pure statistical properties

Start from an **application** : the von Bertalanffy curve

$$L(t|\theta) \quad = \quad L_\infty(1 - exp(-g(t,\delta))$$

is frequently used as an **age-length key**, modelling the increasing of length of an organism (e.g., fish) during its life

Denote $\theta = (L_\infty, \delta)$ the vector of unknown parameters

Capture-recapture data : assume to have couples of observation $\{l^*(t_i), l^*(t_{i+\Delta_i})\}$ such that

$$\begin{aligned} l^*(t_i) &= L(t_i|\theta)\exp(\epsilon_1), \\ l^*(t_{i+\Delta_i}) &= L(t_i + \Delta_i|\theta)\exp(\epsilon_2) \end{aligned}$$

where (ϵ_1, ϵ_2) are observational noises

Classical estimations of the asymptotic length L_∞ can be very sensitive to the size of data

How placing a prior on L_∞ ?

L_∞ is given the sense of the maximal length that a fish can reach, on average on all possible observations

Define $L_\infty^* = L_\infty \exp(\epsilon)$ the *observed* maximal length

Denote $\bar{L}$ the medium length of a fish

Theorem (Pickands) When $\bar{L}$ increases, the distribution of $L_\infty^* | \bar{L} = l$ is a generalized Pareto :

$$P\left(L_\infty^* < x | L_\infty^* > \bar{L}, \sigma, \mu\right) = 1 - \left(1 + \mu\left(\frac{x - \bar{L}}{\sigma}\right)\right)^{-1/\mu}$$

We obtain a justification for :

1. choosing a prior form for L_∞ (given ϵ)
2. conditioning to $\bar{L} \Leftrightarrow$ establishing a hierarchical Bayesian approach

Another example : probability of survival

Let X_t be a number of individuals in a population

Denote $\theta = \theta_{t,t+1}$ the probability of surviving between t and $t+1$

The likelihood can be defined by

$$X_{t+1}|X_t, \theta_{t,t-1} \quad \sim \quad \mathcal{B}(X_t, \theta_{t,t+1}) \quad \textit{(Bernoulli dist.)}$$

One may write

$$\theta_{t,t+1} \quad = \quad \prod_{i=0}^{M+1} \theta_{t+i/M,t+(i+1)/M}$$

therefore, from the CLT, when $1 \ll M$,

$$\log(\theta_{t,t+1}) \quad \sim \quad \mathcal{N}(\mu_t, \sigma_t^2)$$

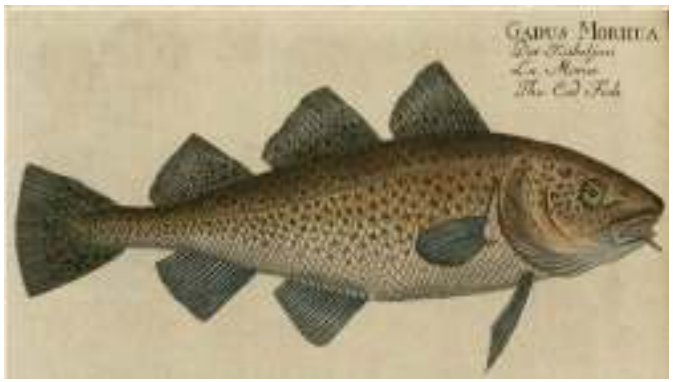
with $\mu_t < -\sigma_t^2/2$ such that $\mathbb{E}[\theta_{t,t+1}] \in [0,1]$

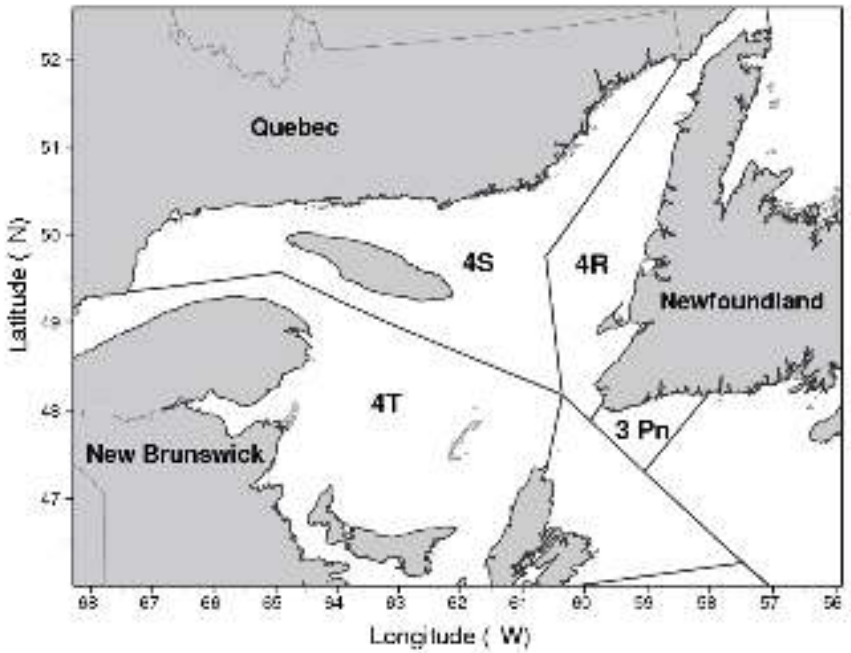
Principle :

- Assume to have an *average observation* $\boldsymbol{Y}^*$ on $Y = g(\theta, c)$ where g is some function and c a set of fixed parameters
- Choose a likelihood linking $\boldsymbol{Y}^*$ and θ
- Choose a noninformative prior $\pi^J(\theta)$ in function of this likelihood
- Select π as the posterior $\pi^J(\theta|\boldsymbol{Y}^*)$

Modelling the cod abundance (*Gadus morhua*) in the Northern Gulf of St Lawrence (Canada)

NAFO division 3Pn4RS



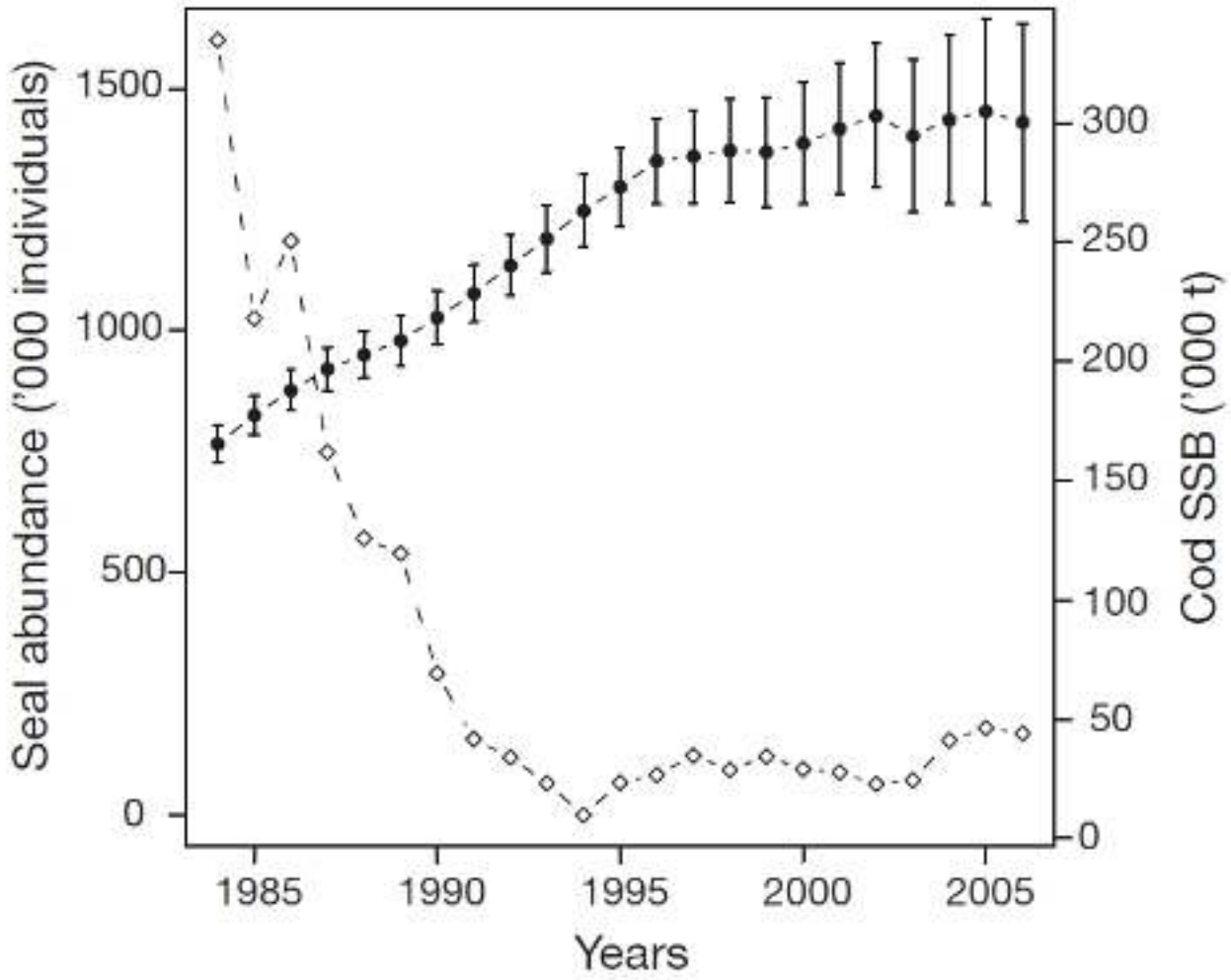


Cod can live 15 years

Main sources of mortality :

- predation by harp seal (*Phoca groenlandica*)
- fishing (especially during the 90's)
- natural (residual) mortality (water layer temperature, etc.)

Observations : seal population increasing and cod decline



(Chassot et al. 2009)

Hidden state-space dynamics of cod abundance $N_{a,t}$

Cod predation	$P_{a,t}$	$=$	$p^c_{a,t} \cdot N_{a,t}$
Residual mortality 1	$N'_{a,t}$	$=$	$p^m_{a,t} \left(N_{a,t} - P_{a,t},\right)$
Commercial fishing	$C_{a,t}$	$=$	$\left(1 - p^f_{a,t}\right) N'_{a,t}$
Middle-year abundance	$N''_{a,t}$	$=$	$N'_{a,t} - C_{a,t}/2$
Residual mortality 2	$N_{a+1,t+1}$	$=$	$p^m_{a,t} \left(N''_{a,t} - C_{a,t}/2\right)$
Total egg production	TEP_t	$=$	$\sum_{a=1}^{A} N_{a,t}\, \xi_{a,\,t}\, \phi_{a,\,t}\, f_{a,t}$
Recruitment at age 0	R_{t+1}	$=$	$p^r_{t+1} \cdot TEP_t$
Recruitment at age 1	$N_{1,t+2}$	$=$	$\left(p^m_{0,t+1}\right)^2 R_{t+1}$

Sex ratio	ξ
Proportion of maturing females	ϕ
Fecundity (NoE cod^{-1})	f

Observations

$I_{a,t} = q\, \varsigma_{a,s}\, N''_{a,t}$	Survey indice
with $\varsigma_{a,s} = \dfrac{1}{1+\exp\left(-\gamma_s\left(a-\delta_s\right)\right)}$	Selectivity
and q	Capturability
$C_t = \sum_{a=1}^{A} C_{a,t}$	Total catch
$p_{a,t,c} = C_{a,t} / \sum_{a=1}^{A} C_{a,t}$	Catch-at-age obs. probability
$p_{a,t,s} = I_{a,t} / \sum_{a=1}^{A} I_{a,t}$	Survey-at-age obs. probability

$$J^*_t = \sum_{a=1}^{A} J^*_{a,t} \overset{iid}{\sim} \mathcal{N}\left(\sum_{a=1}^{A} \log I_{a,t}(\theta), \psi^2\right) \qquad \psi^2 = A\sigma^2 + A^2\tau^2$$

where $J^*_{a,t} = \log(I^*_{a,t}) = \log(I_{a,t}) + \epsilon_{a,t} + \eta_t$

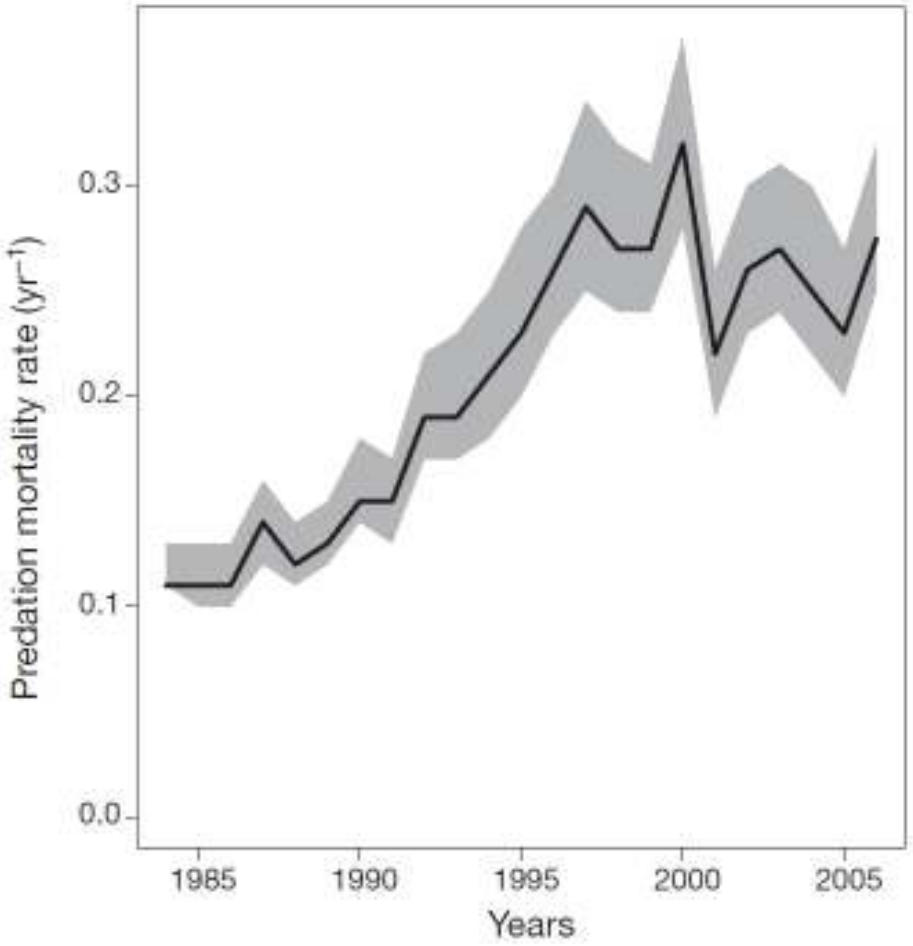
$$\log C^*_t \overset{iid}{\sim} \mathcal{N}\left(\log C_t(\theta) - \frac{\sigma_c^2}{2}, \sigma_c^2\right) \qquad \sigma_c^2$$

List of unknown parameters

ζ_a	Baseline attack rate for age a (nb. attacks seal^{-1})
π	Normalization coefficient of attack rates
m	Shape parameter of the Holling response type
α	Intercept of the natural mortality curve (yr^{-1})
β	Slope of the natural mortality curve
F	Fishing mortality rate of cod (yr^{-1})
$R_{\max}$	Maximum nb. of cod recruits (NoI)
r	TEP needed to produce recruitment $= R_{\max}/2$ (NoE)
$\varsigma_{a,\,c}$	Commercial selectivity-at-age
γ_c^1	Shape parameter of the commercial selectivity (1984-1993)
δ_c^1	Age of half-vulnerability (1984-1993)
γ_c^2	Shape parameter of the commercial selectivity (1994-2006)
δ_c^2	Age of half-vulnerability (1994-2006)
$\varsigma_{a,\,s}$	Survey selectivity-at-age
q	Survey catchability
γ_s	Shape parameter of the survey selectivity
δ_s	Age of half-vulnerability

Frequentist results (Chassot et al. 2009)

Too small “confidence" intervals (bootstrap)



Being Bayesian appears more respectful of all uncertainties around the modelling

$$\varsigma_a = \frac{1}{1 + \exp(-\gamma\{a - \delta\})}.$$

- $\delta =$ age at which 50% of the cod population is sensitive to the fishing gear
- $\gamma =$ shape parameter

Meta-analysis of selectivity estimates obtained from survey / commercial catches-at-age of Atlantic cod with similar gears

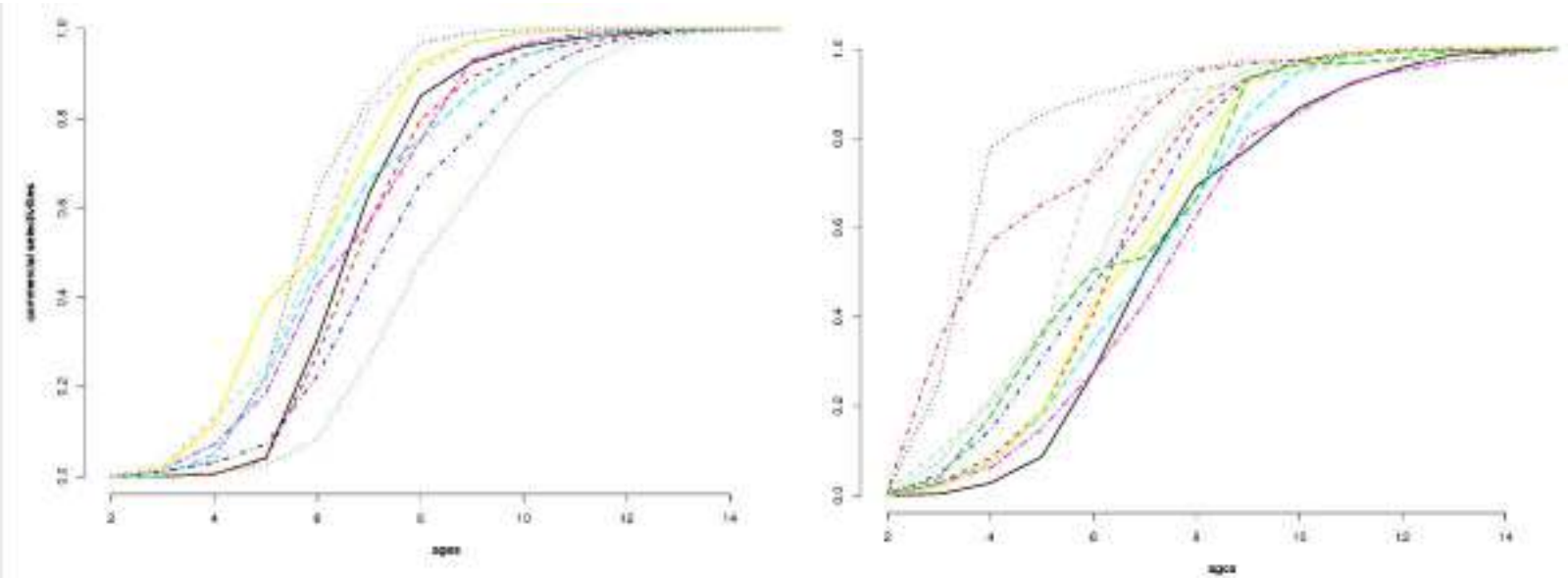
Following an idea of Harley and Myers (2001)

$M = 153$ datasets

Let $c_1, \ldots, c_{A^+}$ be a sample of catches-at-age

Kaplan-Meier estimator = cumulative age frequency

$$\varsigma_a^* = \sum_{i=1}^{A} c_i \mathbb{1}_{\{i \leq a\}} / \sum_{j=1}^{A} c_j.$$



Our aim is

- to define some kind of likelihood $\ell(\varsigma_{1,i}^*, \ldots, \varsigma_{A,i}^*, i = 1, \ldots, M | \gamma, \delta)$
- to define baseline (noninformative) priors for (γ, δ) with respect to ℓ
- to select the final priors on (γ, δ) as posteriors

Consider the reparametrization

$$s_a \quad = \quad -\log(\varsigma_a^{-1} - 1) \; = \; \gamma(a - \delta) \tag{2}$$

and denote s_a^* the corresponding vector of nonparametric estimates

Estimate then test the model hypothesis

$$s_a^* \quad = \quad s_a + \mathcal{N}(0, \sigma_a^2)$$

Classical tests (Shapiro-Wilks, etc.) do not deny this hypothesis (high $p-$values $\in$[0.35,0.86])

Denote $s_I^* = (s_1^*(i_1), \ldots, s_A^*(i_A))$, with $i_j \neq i_k$, the i_j being chosen in $I \subset \{1, \ldots, M\}$, and

$$\bar{s}^* \quad = \quad \frac{1}{A} \sum_{j=1}^{A} s_j^*(i_j) \; = \; \alpha\gamma - \delta + \mathcal{N}(0, \sigma^2)$$

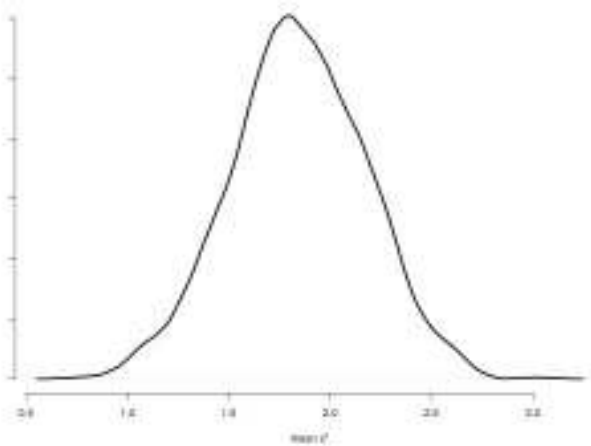
with $\sigma^2 = \sum_{a=1}^{A} \sigma_a^2 / A$ and $\alpha = (A+1)/2$.

There are $(M!)/(M - A^{+}!)$ possible values $\bar{s}^{*}$

They are not independent (since they share data coming from same empirical selectictivies)

Selecting farther ages (a_1, a_2), $s^{*}_{a_1}(i_{a_1})$ and $s^{*}_{a_2}(j_{a_2})$ remain however little correlated

A mixing distribution of random variable $\bar{s}^{*}$ can be empirically simulated



The formal structure appeared relevant

Eliciting an informative prior on (γ, δ)

1. Let $\pi^J(\gamma, \delta)$ be a noninformative prior for the likelihood model
   - *reference prior rule* from Berger & Bernardo (1992)
   - biological experts agreed that the widest interval for δ is $[a_l, a_r] = [1, 6] \subset [1, A]$

$$\pi^J(\gamma, \delta) \quad \propto \quad \mathbb{1}_{\gamma \geq 0} \mathbb{1}_{\{a_l \leq \delta \leq a_r\}}$$

2. Consider the "one-average-data" likelihood emanating from

$$\bar{s} \quad = \quad \alpha\gamma - \delta + \mathcal{N}(0, \nu^2 + \sigma^2)$$

3. Elicit $\pi(\gamma, \delta) = \pi^J(\gamma, \delta | \bar{s}^*)$, ie.

$$\pi(\gamma, \delta) \quad \propto \quad \exp\left\{ -\frac{1}{2(\sigma^2 + \nu^2)} (\alpha\gamma - \delta - \bar{s})^2 \right\} \mathbb{1}_{\{\gamma \geq 0\}} \mathbb{1}_{\{a_l \leq \delta \leq a_r\}}$$

	survey	commercial		survey	commercial
σ^2	1.221	1.510	$\bar{s}$	1.891	1.493
ν^2	0.1146	0.1051	α	6.5	6.5

Once the posterior of generic parameter vector $\boldsymbol{\theta}$ is obtained, we want to make projections studies

Often $\boldsymbol{\theta} = (\boldsymbol{\theta}_I, \boldsymbol{\theta}_N)$ where

- $\boldsymbol{\theta}_I = \{$parameters of interest$\}$ (ex : selectivity parameters, recruitement...)
- $\boldsymbol{\theta}_N = \{$nuisance parameters$\}$ (observational variances, capturability)
  - purely relative to the obtention of data

$$J_t^* = \sum_{a=1}^{A} J_{a,\,t}^* \overset{iid}{\sim} \mathcal{N}\left(\sum_{a=1}^{A} \log I_{a,\,t}(\theta), \psi^2\right)$$

$$\log C_t^* \overset{iid}{\sim} \mathcal{N}\left(\log C_t(\theta) - \frac{\sigma_c^2}{2}, \sigma_c^2\right)$$

No prior information is usually available on $\boldsymbol{\theta}_N = (q, \sigma_c^2, \psi^2)$

The choice of a baseline (noninformative) prior $\pi(\boldsymbol{\theta}_N)$ must be independent on any informative prior choice on $\boldsymbol{\theta}_I$

Kullback-Leibler divergence between posterior and prior

$$\mathrm{KL}(\pi|\mathbf{x}) \quad = \quad \int_{\Theta_N} \pi(\boldsymbol{\theta}_N|\mathbf{x}) \log \frac{\pi(\boldsymbol{\theta}_N|\mathbf{x})}{\pi(\boldsymbol{\theta}_N)} \, d\boldsymbol{\theta}_N$$

with $\pi(\boldsymbol{\theta}_N) = \int \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}_I$

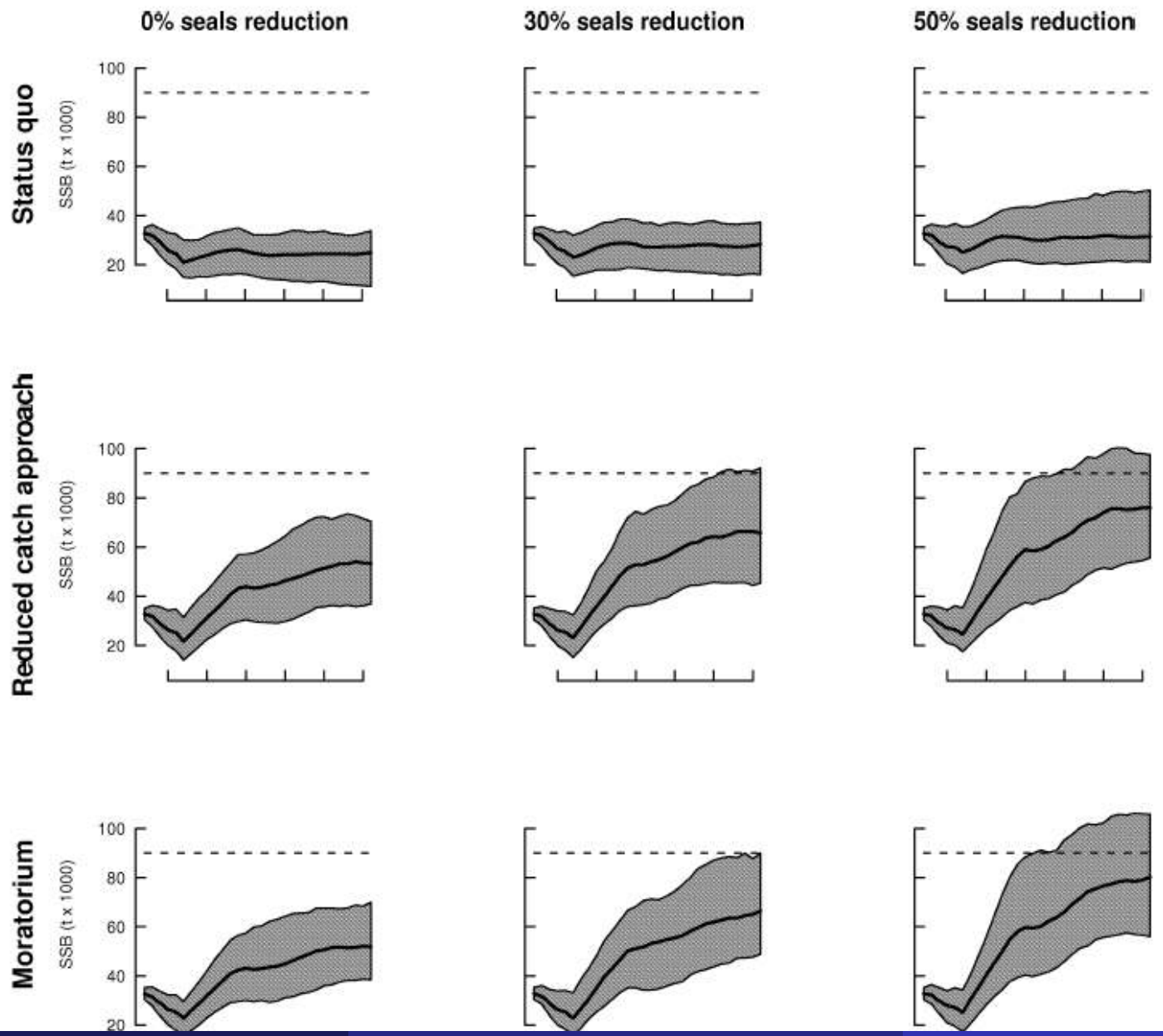
Elicit

$$\pi^* \quad = \quad \arg\max_{\pi} \left\{ \lim_{\mathrm{card}(\mathbf{x}) \to \infty} \mathbb{E}_m \left[\mathrm{KL}(\pi|\mathbf{X}) \right] \right\}$$

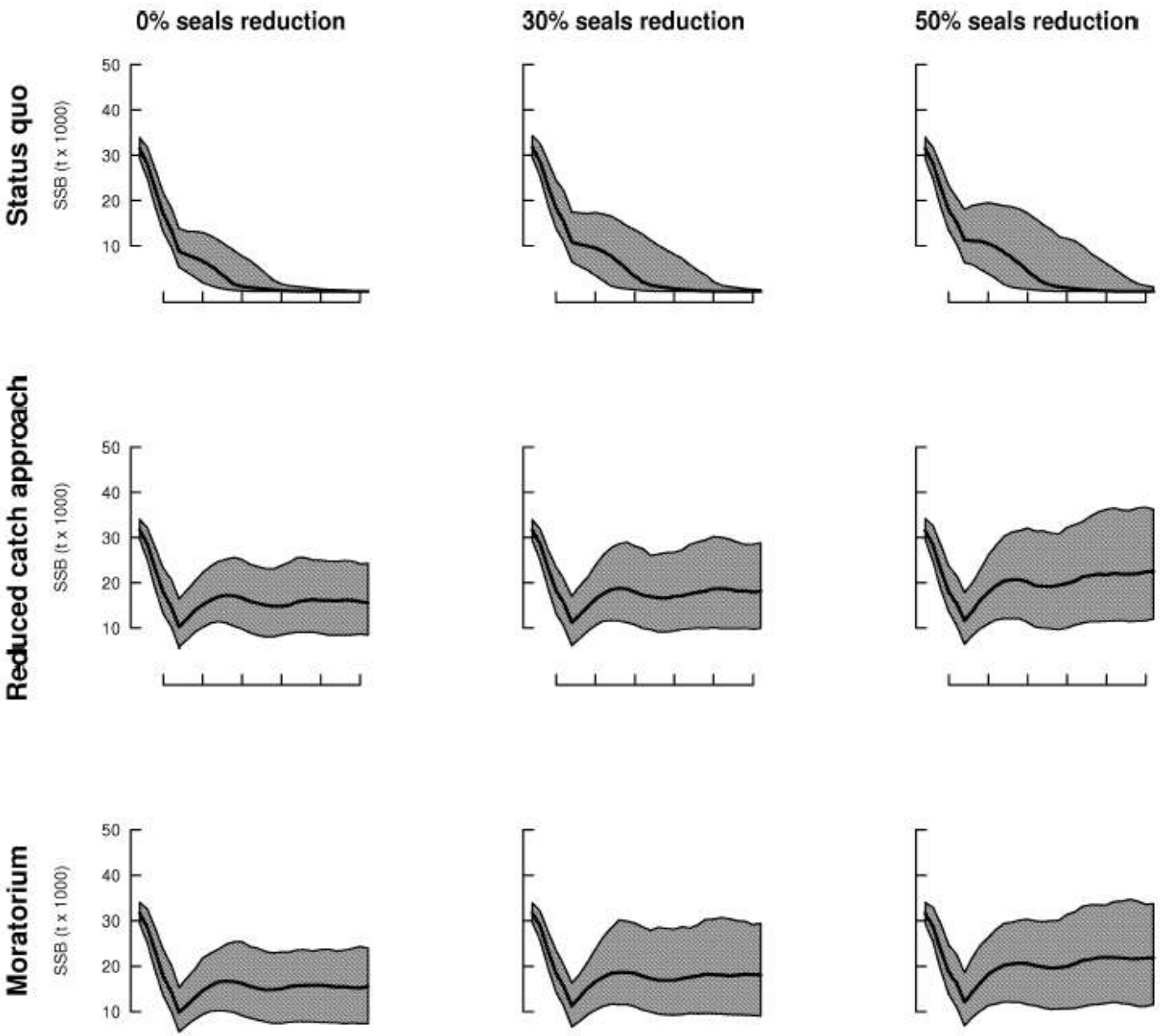
One finds (after boring calculations)

$$\pi^*(\psi^2, \sigma_c^2, q) \quad \propto \quad \psi^{-3} \sigma_c^{-3} q^{-1} \; \mathbb{1}_{\{(\psi,\phi,q) \in R^3_{+,*}\}}.$$

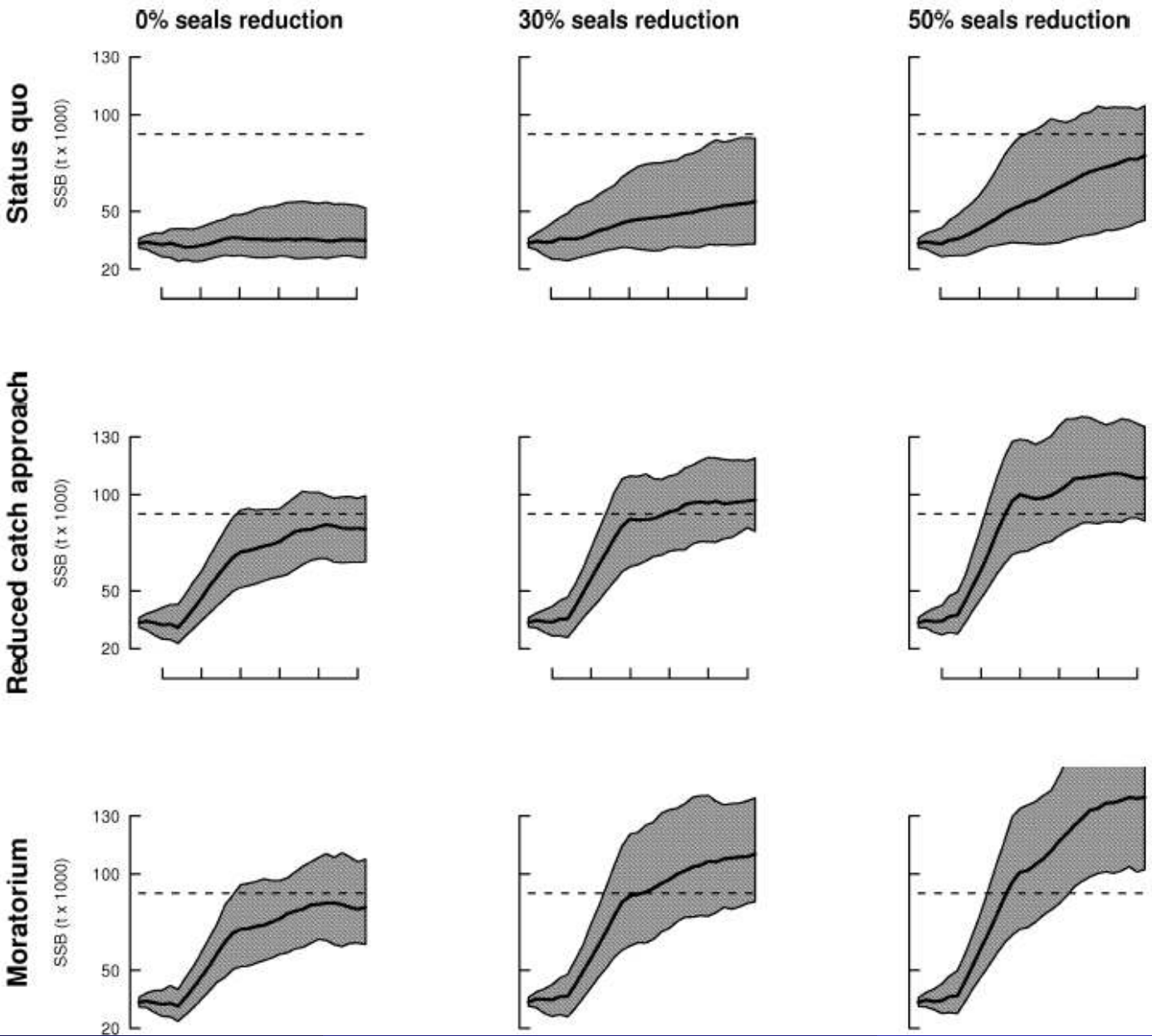
Some expected projections (B., Chassot et al. 2011)



Some expected projections (2)



Some expected projections (3)



Let X represent the lifetime of a device Σ, X is given an exponential distribution $\mathcal{E}(\lambda)$

$\lambda =$ failure rate, which is an object an industrial expert can be familiar with

Dialog with the expert :

1. Consider a management (replacement) decision established on the given value $\bar{\lambda}$ instead of the true λ
2. For a similar cost $|\bar{\lambda} - \lambda|$, there are two possible policies :
   - let C_1 be the mean chance of being too optimistic (assuming $\bar{\lambda} \leq \lambda$)
   - let C_2 be the mean chance of being too pessimistic (assuming $\bar{\lambda} > \lambda$)
3. Can you give an estimate $\hat{\delta}$ of the ratio $\delta = C_2/C_1$?

The rationality axiom says that if the expert is risk-unconcerned, then

$$\bar{\lambda} \quad = \quad \arg\min_x \underbrace{\int_0^\infty |x - \lambda| \left(C_1 \mathbb{1}_{\{x \leq \lambda\}} + C_2 \mathbb{1}_{\{x > \lambda\}} \right) \pi(\lambda)\, d\lambda}_{\text{cost function integrated over all prior possibilities for the true } \lambda}$$

It follows than

$$\int_0^{\bar{\lambda}} d\Pi(\lambda) \quad = \quad \Pi(\lambda < \bar{\lambda}) \; = \; \frac{C_1}{C_1 + C_2}$$

The interpretation of the expert's answer is that $1/(1+\hat{\delta})$ is an estimate of the prior $\alpha-$order quantile with $\alpha = C_1/(C_1 + C_2)$

Remark. The posterior work is similar in spirit : a decision must be adressed by minimizing a cost function integrated over all posterior credible values of λ

Anchoring variable

Unless the expert is very trained, he/she is not a statistician and does not know the existence of a given parametrization θ

Anchoring variables must be selected : often X itself, since **observable**

The previous dialog is more realistic in the context when he/she is questioned about the lifetime X and not λ

In this case, $\bar{x}$ is perceived as the $\alpha-$order prior predictible percentile

$$\int_0^{\bar{x}} m(x)\, dx = P_m(X < \bar{x}) = \alpha$$

$$\text{where} \quad m(x) = \int_0^{\infty} f(x|\lambda)\pi(\lambda)\, d\lambda$$

This interpretation is probably the most accepted in the Bayesian community (O'Hagan 2006), therefore statisticians are motivated to ask questions like

- Given the times x_0 and $x_1 > x_0$, how much chances do the device Σ has to survive after x_0 rather than to survive after x_1 ?

Prospective questions about the auditability of a prior model

expert strenght quantifying the ratio “information yielded by an expert" / "data information"

- needs for an understandable definition

conservative bias do the models $(f_i(.|\boldsymbol{\theta}_i), \pi_i(\boldsymbol{\theta}_i))$ are biased w.r.t. “cautious", “reasonables", “conservative" specifications from the expert ?

coherence W.r.t. consensual qualitative knowledge on Σ, is π coherent ? (ex : exponential aging from a component)

unicity For the model $f_i(.|\boldsymbol{\theta}_i)$, is π_i be defined in a unique way ?

equitability Do the complete Bayesian models $(f_i(.|\boldsymbol{\theta}_i), \pi_i(\boldsymbol{\theta}_i))$ be equitable ?

- a model should not be arbitrarily favorized a priori [Consonni and Veronese 2008]
- the prior of a nested sampling model should be itself nested in the prior of a more complex model

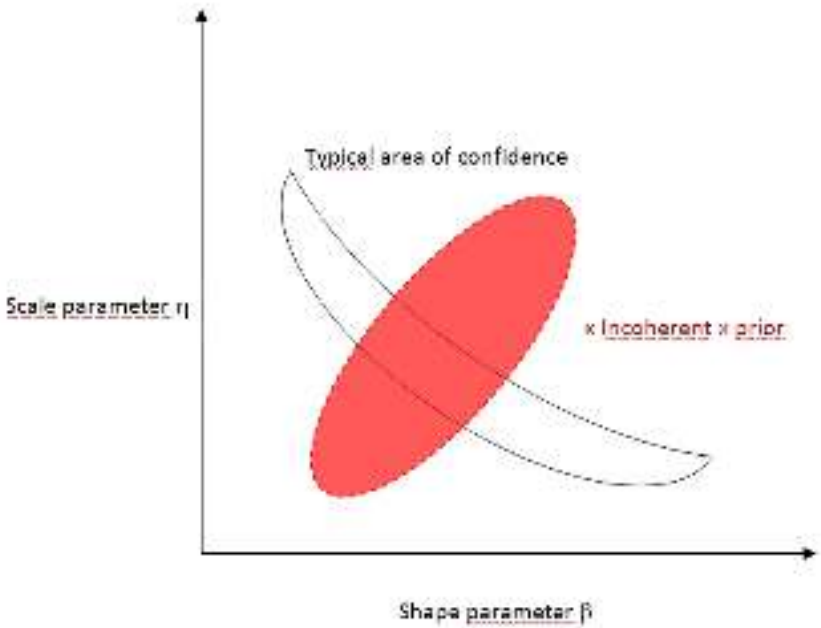
An example of coherence : the Weibull banana shape

Weibull distribution in lifetime data analysis

$$f(t|\eta,\beta) \quad = \quad \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} \exp\left\{-\left(\frac{t}{\eta}\right)^{\beta}\right\} \mathbb{1}_{\{t\geq 0\}}$$

A prior $\pi(\beta,\eta)$ with strongly positive correlation threatens to be incoherent with the meaning of the model :

- high $\beta \Leftrightarrow$ strong ageing $\Rightarrow$ short lifetime $\Leftrightarrow$ small η



Complementary wishes

practicity is π easy to handle ? [Rios Insua and Ruggeri 2000]

- explicit if possible (sensitivity study are simplified)
- easy to sample from (comparisons a posteriori-a priori)

opinion pooling how defining a unique π from several priors $\pi_{(1)}, \ldots, \pi_{(n_e)}$?

no longer available experts information has been summarized in the past. How to deal without questioning again ?

Formal procedure for building a prior $\pi(\theta)$ under a given type of constraints reflecting **quantitative knowledge**

Principle : we are looking for a $\pi(\theta)$ in the widest class of probability measures respecting those constraints

The **entropy** is generally defined as measure of disorder (or uncertainty) associated to a probability distribution

It is a fundamental concept of the information theory

To the origin of this concept, a problem of sorting a discretized information using combinatorics :

- Assume there exists a partition of k geographical areas
- Assume that each area contains $N_i = N \times p_i$ sites, with $i = 1, \ldots, k$, and $\sum_i p_i = 1$
- Assume that on each site you want to find a given information
- Assume that, to find information, you can simply ask *binary questions* (yes/not)
- You want to minimize the number of questions

then it is enough to ask

1. $Q'_i = \log N_i = \log p_i + \log N$ questions to sort the $i-$th area
2. on average on areas, $Q' = \sum_{i=1}^{k} p_i Q'_i$ is the total minimal number of questions needed to find the information

Knowing in probability in which area is the information reduces the average number of questions to ask of the quantity

$$\Delta Q \quad = \quad Q - Q' \quad = \quad -\sum_{i=1}^{k} p_i \log p_i$$

which is positive and maximum when $p_i = 1/k$

The less informative the probability distribution $\Pi = (p_1, \ldots, p_k)$, the higher this quantity

Definition The entropy of a finite random variable with distribution $\Pi = (\pi(\theta_1), \ldots, \pi(\theta_k))$ is

$$\mathcal{H} \quad = \quad -\sum_{i=1}^{k} \pi(\theta_i) \log \pi(\theta_i) \qquad \textit{(Shannon's entropy)}$$

Generalizing to continuous cases :

- the continuous case can be interpreted as a "limit"' discrete case with smallest and smallest intervals
- the entropy must be invariant to any variable change $\theta \mapsto \nu(\theta)$

Definition The entropy of a random variable with probability density $\pi(\theta)$ is

$$\mathcal{H}(\pi) \quad = \quad -\int_{\Theta} \pi(\theta) \log \frac{\pi(\theta)}{\pi_0(\theta)} \, d\theta \qquad \textit{(Kullback's entropy)}$$

where $\pi_0(\theta)$ is a positive benchmark measure on Θ, representing complete ignorance of the value of θ on Θ

Very usually $\pi_0(\theta)$ is chosen as the uniform density over est Θ

Remark : the entropy is not always longer positive, but it remains maximum in $\pi(\theta) = \pi_0(\theta)$

Aim : choose $\pi(\theta)$ as vague as possible

$$\pi^*(\theta) \quad = \quad \arg\max_{\pi \in \mathcal{P}} - \int_\Theta \pi(\theta) \log \frac{\pi(\theta)}{\pi_0(\theta)} \, d\theta \tag{3}$$

in the set $\mathcal{P}$ of positive measures, under M linear-type constraints similar to

$$\int_\Theta g_i(\theta)\pi(\theta) \, d\theta \quad = \quad c_i, \ , \quad i = 1, \ldots, M$$

The first constraint is always a **normalizing** constraint :

$$\int_\Theta \pi(\theta) \quad = \quad 1$$

Solution : if all previous integrals exist, the solution of problem (3) is of the form

$$\pi^*(\theta) \quad \propto \quad \pi_0(\theta) \exp\left(\sum_{i=1}^{M} \lambda_i g_i(\theta)\right)$$

This form characterizes the laws from the exponential family

The parameters $(\lambda_1, \ldots, \lambda_M)$ are Lagrange multipliers and must be calibrated by solving the equations

$$\int_\Theta g_i(\theta)\pi^*(\theta) \quad = \quad c_i, \ , \quad i = 1, \ldots, M$$

When only the normalizing constraint is assumed, then

$$\pi^*(\theta) \quad = \quad \pi_0(\theta)$$

The maximum entropy principle can also be applied to X conditionaly to θ, and leads to the following parametric family

Definition Let $(C, h) : \Theta \times \Omega \mapsto \boldsymbol{R}_+^2$, and $(R, T) : \Theta \times \Omega \mapsto \boldsymbol{R}^k \times \boldsymbol{R}^k$. The family of distributions with density

$$f(x|\theta) \quad = \quad C(\theta)h(x)\exp\{R(\theta)\cdot T(x)\}$$

is called *exponential family* of finite dimension k. When $\Theta \subset \boldsymbol{R}^k$ and $\Omega \subset \boldsymbol{R}^k$, one can use the simpler writing (up to a reparameterisation)

$$f(x|\theta) \quad = \quad h(x)\exp\{\theta\cdot x - \psi(\theta)\}$$

with

$$\begin{aligned} \mathbb{E}_\theta[X] &= \nabla\psi(\theta) \quad \textit{(gradient)} \\ \mathrm{cov}(X_i, X_j] &= \frac{\partial^2\psi}{\partial\theta_i\partial\theta_j}(\theta) \end{aligned}$$

One speak rather about *natural exponential family*

Dirichlet distribution. (extension of the Beta distribution)

$$f(x|\theta) \quad = \quad \frac{\Gamma(\sum_{i=1}^k \theta_i)}{\prod_{i=1}^k \Gamma(\theta_i)} \prod_{i=1}^k x_i^{\theta_i - 1} \ \mathbb{1}_{\{S_k(x)\}}$$

defined on the simplex $S_k(x) = \left\{ x = (x_1, \ldots, x_k);\ \sum_{i=1}^k x_i = 1,\ x_i > 0 \right\}$

Gaussian vector. Si $\mathbf{x_n} = (x_1, \ldots, x_n) \sim \mathcal{N}_p(\mu, \sigma^2 I_p)$, the the joint distribution satisfies

$$f(\mathbf{x_n}|\theta) \quad = \quad C(\theta) h(\mathbf{x_n}) \exp\left(n\bar{x} \cdot (\mu/\sigma^2) + \sum_{i=1}^n \|x_i - \bar{x}\|^2 (-1/2\sigma^2) \right)$$

with $\theta = (\mu, \sigma)$, and the statistics $(\bar{x}, \sum_{i=1}^n \|x_i - \bar{x}\|^2)$ is exhaustive for all $n \geq 2$

Let $X|\theta$ be a maximum entropy distribution, with density of the form :

$$f(x|\theta) \quad = \quad \exp\left(\sum_{j=1}^{L} T_j(x) d_j(\theta)\right)$$

If, moreover, the prior $\pi(\theta)$ is similarly elicited by maximum entropy :

$$\pi(\theta) \quad \propto \quad \nu(\theta) \exp\left(\sum_{i=1}^{M} \lambda_i g_i(\theta)\right)$$

Then, given $\mathbf{x_n} = (x_1, \ldots, x_n)$, the posterior distribution has the same structural form than $\pi(\theta)$:

$$\pi(\theta|\mathbf{x_n}) \quad \propto \quad \nu(\theta) \exp\left(\sum_{i=1}^{M} \lambda_i g_i(\theta) + \sum_{j=1}^{L} \left[\sum_{k=1}^{n} T_j(x_k)\right] d_j(\theta)\right)$$

In this case the prior is said to **conjugate**

Let

$$f(x|\theta) \quad = \quad h(x)\exp(\theta \cdot x - \psi(\theta))$$

then the prior measure automatically generated by

$$\pi(\theta|a,b) \quad = \quad K(a,b)\exp(\theta \cdot a - b\psi(\theta))$$

is naturally conjugate and the corresponding posterior measure, given a data x, is

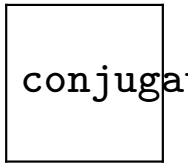
$$\pi(\theta|a+x, b+1)$$

$K(a,b)$ is the normalizing constant

$$K(a,b) \quad = \quad \left[\int_\Theta \exp(\theta \cdot a - b\psi(\theta))\right]^{-1}$$

that is finite if $b > 0$ et $a/b \in \mathring{N}$

Some conjugate prior/posterior distributions for some usual exponential families



courtesy of VS-RSF

Rationale of form invariance :

- the knowledge $x \sim f(x|\theta)$ updating $\pi(\theta)$ into $\pi(\theta|x)$ is limited by nature
- hence it should not lead to modify *all* the structural form of $\pi(\theta)$, but simply of its hyperparameters :

$$\pi(\theta) = \pi(\theta|\delta) \quad \Rightarrow \quad \pi(\theta|x) = \pi(\theta|\delta + s(x))$$

- this modification should remain of finite dimension, and a deeper change of $\pi(\theta)$ is not acceptable

Another justification is the representation using virtual data (see next slide)

In practice, the interest of conjugation is the working convenience

Let the conjugate prior

$$\pi(\theta|x_0, m) \quad \propto \quad \exp\{\theta \cdot x_0 - m\psi(\theta)\} \tag{4}$$

then the prior predictive mean (expectancy) is

$$\mathbb{E}[X] \quad = \quad \mathbb{E}[\mathbb{E}[X|\theta]] \; = \; \mathbb{E}\left[\nabla\psi(\theta)\right] \; = \; \frac{x_0}{m}$$

and the posterior predictive mean, given a i.i.d. sample $\mathbf{x_n} = (x_1, \ldots, x_n)$, is

$$\mathbb{E}[X|\mathbf{x_n}] \quad = \quad \frac{x_0 + n\bar{x}}{m + n} \tag{5}$$

Hence m has the sense of a virtual sample size, offering an indication of the "strength" of information carried through the prior

Theorem (Diaconis & Ylvisaker, 1979) If the dominating measure is continuous with respect to the Lebesgue measure, then (5) $\Rightarrow$ (4)

Remember that the exponential model

$$X \sim f(x|\theta) = \theta \exp(-\theta x) \mathbb{1}_{\{x \geq 0\}}$$

can be useful to model the lifetime of a device only submitted to accidental failures

Modelling very used in reliability engineering

Placing a gamma prior

$$\theta \sim \mathcal{G}(a, b)$$

with density

$$\pi(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp(-b\theta) \mathbb{1}_{\{\theta \geq 0\}}$$

Posterior distribution, given $\mathbf{x_n} = (x_1, \ldots, x_n)$:

$$\theta|\mathbf{x_n} \sim \mathcal{G}(a + n, b + n \cdot \bar{x}_n)$$

Some models accept conjugate priors which does not belong to the exponential family

Example 1 : Pareto distribution with $\alpha > 0$ known, and $\theta > 0$

$$f(x|\theta) = \alpha \frac{\theta^\alpha}{x^{\alpha+1}} \mathbb{1}_{]\theta,\infty[}(x)$$

admits a Pareto conjugate prior over $1/\theta$

Example 2 : uniform distributions

$$f(x|\theta) = \frac{\mathbb{1}_{[-\theta,\theta]}(x)}{2\theta}$$

$$f(x|\theta) = \frac{\mathbb{1}_{[0,\theta]}(x)}{\theta}$$

Another view (idealistic)

Imagine an expert is a statistician and can provide an iid sample $\tilde{\mathbf{x}}_\mathbf{m} \sim f$ of size m

A nice (and logical) prior is $\pi(\theta) = \pi^J(\theta|\tilde{\mathbf{x}}_\mathbf{m})$ where π_i^J is noninformative

It answers to most of our requirements (unicity, assessing correlations within θ, aggregation of opinions without paradoxes...)

We "just" have to care about the subjectivity in data $\tilde{\mathbf{x}}_\mathbf{m}$ (location, size, etc.)

Assuming $\pi(\theta)$ is not conflicting with real data, m is convenient to modulate our trust in the expert opinion

Virtual sample idea = not new

Construction principle of conjugate models, with π entirely explicit

- Gamma prior $\mathcal{G}(m, \sum_{i=1}^{m} \tilde{t}_i)$ for exponential models
- Dirichlet priors for multinomial models

Close idea to Zellner's $g-$ prior (Zellner 1986) for Gaussian regression models

Calibrating with information-theoretic distances
Theoretical works by Clarke (1996), Liu & Clarke (2004), Lin et al. (2007), Morita et al. (2007)

Neal (2001) : imaginary data to equilibrate priors

For a given $f(t|\theta)$

1. select $\pi^J(\theta)$
2. assume there exists a "hidden" (virtual) sample $\tilde{\mathbf{x}}_\mathbf{m}$ of size m
3. give a unique form choosing $\pi(\theta) \equiv \pi^J(\theta|\tilde{\mathbf{x}}_\mathbf{m})$, ie.

$$\pi(\theta) \quad = \quad \pi(\theta|\boldsymbol{\Delta}_m)$$

   with $\boldsymbol{\Delta}_m$ a set of virtual statistics

4. estimate $\boldsymbol{\Delta}_m$ by $\widehat{\boldsymbol{\Delta}}_m = \arg\min_{\boldsymbol{\delta}_m} \mathcal{D}\left(\boldsymbol{\Lambda}_e, \boldsymbol{\Lambda}(\boldsymbol{\delta}_m)\right)$
   - $\boldsymbol{\Lambda}_e$ are prior predictive features given by expert questioning
   - $\boldsymbol{\Lambda}(\boldsymbol{\delta}_m)$ are features of the effective prior predictive distribution with pdf

$$m(x|\boldsymbol{\delta}_m) \quad = \quad \int_\Theta f(x|\theta)\pi(\theta|\boldsymbol{\delta}_m)\ d\theta$$

   - $\mathcal{D}$ is some kind of distance

How choosing $\mathcal{D}$?

1. Ex : Cooke's method of discrete Kullback-Leibler loss (1991).
   - denote $\boldsymbol{\Lambda}_e = \{\lambda_{1,e}, \ldots, \lambda_{q,e})$
   - assume each $\lambda_{i,e}$ is a couple $(x_{i,e}, \alpha_{i,e})$ such that $P_m(X < x_{i,e}) = \alpha_{i,e}$

$$\mathcal{D}\left(\boldsymbol{\Lambda}_e, \boldsymbol{\Lambda}(\boldsymbol{\delta}_m)\right) \quad = \quad \sum_{i=0}^{q} (\alpha_{i+1,e} - \alpha_{i,e}) \log \frac{\alpha_{i+1,e} - \alpha_{i,e}}{\alpha_{i+1}(\boldsymbol{\delta}_m) - \alpha_i(\boldsymbol{\delta}_m)}$$

with $\alpha_{0,e} = \alpha_0 = 0$, $\alpha_{q+1,e} = \alpha_{q+1} = 1$ and

$$\alpha_i(\boldsymbol{\delta}_m) = \int_{-\infty}^{x_{i,e}} m(x|\boldsymbol{\delta}_m)\ dt$$

2. One may weight the Kullback loss such that the most important constraints $\lambda_{i,e}$ are nearly fully respected (most trustworthy, pessimistic (conservative), normative...

One cannot hope all expert specifications are simultaneous coherent with the Bayesian model

Weibull example : lifetime of components from the secondary water circuit of a power plant (B. 2009)

The Weibull model is **not conjugate** : $P(X < x|\theta) = 1 - \exp\left(-\mu x^{\beta}\right)$

Jeffreys' prior : $\pi^J(\mu, \beta) \propto (\mu\beta)^{-1}$

Most trustworthy specification

experts	cred.intervals (5%,95%)	median value
$\mathcal{E}_1$ (exploiter)	[200,300]	250
$\mathcal{E}_2$ (manufacturer)	[100,500]	250

Using the most trustworthy specification $x_e = 250$, the virtual data posterior prior modelling is

$$\begin{aligned} \mu|\beta &\sim \mathcal{G}\left(m, \left((1-\alpha)^{-1/m} - 1\right)^{-1} \left(x_\alpha^{(e)}\right)^{\beta}\right) \\ \beta &\sim \mathcal{G}\left(m, m/\beta_e\right) \end{aligned}$$

where $\alpha = 0.5$ and $\beta_e = \mathbb{E}[\beta]$ and m can be calibrated using the other percentiles (or some qualitative knowledge on aging)

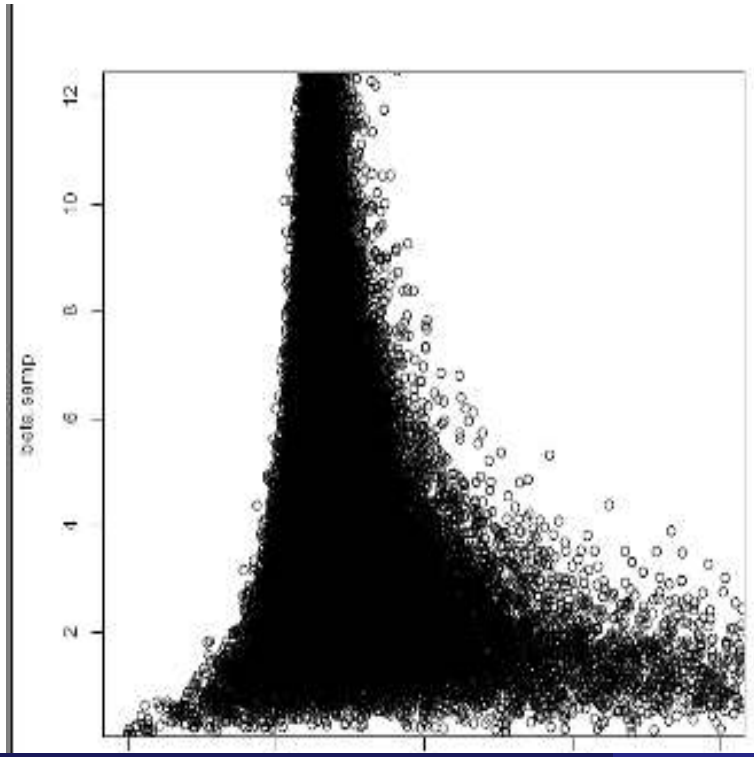
Full calibration : minimizing Cooke's criterion in (m, β_e)

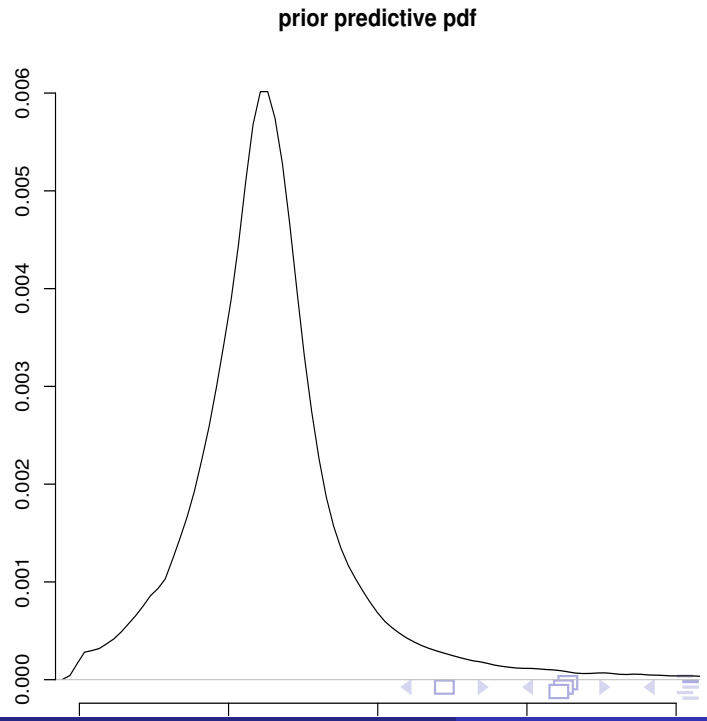
Existence ensured, unicity not formally proven but obtained in practice

Useful especially if the expert cannot be questioned again in practice, or when prior information comes from past summaries

Expert $\mathcal{E}_2$:

m	β_e	coverage error
2.5	4.43	5.10^{-5}





Ideas for calibrating m (to be cautious)

Idea 1 : it depends on the order of prior predictive percentiles

Idea 2 : bisection or hisgram methods (O'Hagan 2006)

Idea 3 : "true" percentiles orders can be corrected

- Example of correction table (Lannoy and Procaccia 2002)

translation of expert opinion	trueness	a_i^*
5%	25%	4
20%	33%	3
25%	40%	2
75%	60%	2
80%	66%	3
95%	75%	4

$$\text{Fréchet } \mathcal{F}(\theta): \quad P(X < x|\theta) \quad = \quad \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\}$$
$$\text{with } \sigma > 0,\ \xi > 0,\ \mu \in \boldsymbol{R} \text{ and } x \geq \mu$$

Reparametrize the Fréchet distribution $\mathcal{F}(\theta)$:

$$P(X < x|\theta) \quad = \quad \exp\left\{-\nu\,(x-\mu)^{-1/\xi}\right\}$$

and denote now $\theta = (\mu, \nu, \xi)$ with $\nu = \sigma^{1/\xi} > 0$.

A nice prior form is given in next proposition

Proposition

Assume the Fréchet prior distribution $\pi(\nu, \mu, \xi)$ defined by

$$\begin{aligned} \nu|\mu,\xi &\sim \mathcal{G}\left(m, s_1(\mu,\xi)\right), \\ \xi|\mu &\sim \mathcal{IG}\left(m, s_2(\mu)\right), \\ \pi(\mu) &\propto \frac{\mathbb{1}_{\{\mu \leq x_{e_1}\}}}{(x_{e_2}-\mu)^m s_2^m(\mu)} \end{aligned} \tag{6}$$

where $\mu < x_{e_1} < x_{e_2}$ and

$$\begin{aligned} s_1(\mu,\xi) &= m(x_{e_1}-\mu)^{-1/\xi}, \\ s_2(\mu) &= m \log\left(\frac{x_{e_2}-\mu}{x_{e_1}-\mu}\right). \end{aligned}$$

Then $\pi(\nu, \mu, \xi)$ is proper for any $m > 0$, is conjugated for ν given (μ, ξ), and when $m \in \mathbb{N}^*$, $\pi(\nu, \mu, \xi) = \pi^R(\nu, \mu, \xi|\tilde{\mathbf{x}}_\mathbf{m})$ where π^R is the Fréchet reference prior and $\tilde{\mathbf{x}}_\mathbf{m}$ is a virtual Fréchet sample of size m with statistics $\{x_{e_1}, x_{e_2}\}$

Percentile order	Pluviometry P (mm)
25%	75
50%	100
75%	150

Table – Prior predictive information on daily maxima per year, extrapolated by an expert from daily maxima measured at a nearby station.

Virtual size m	x_{e_1}	x_{e_2}	μ_{inf} Order of prior predictive quartiles (75,100,150)
1	100.41	130.20	[25%, 50%, 75%]
2	95.30	138.39	[24%, 49%, 74%]
3	91.22	136.93	[23%, 51%, 74%]
4	89.18	135.10	[24%, 50%, 74%]
5	87.72	133.95	[24%, 51%, 75%]
6	87.65	133.88	[24%, 50%, 75%]
7	87.14	133.26	[25%, 50%, 74%]
10	86.63	132.65	[25%, 51%, 75%]
15	85.11	132.24	[26%, 50%, 75%]

Equitability between models

Assume f_{i+1} is nested in f_i

The virtual data elicitation automatically leads to a nested prior $\pi_{i+1} \subset \pi_i$, with virtual sizes $m^{(i)}$ and $m^{(i+1)}$

In absence of data, we should not arbitrarily favor a Bayesian model in the elicitation process

Many proposals [Ibrahim & Laud 1994 ; Dawid & Lauritzen 2000 ; Roverato & Consonni 2004]

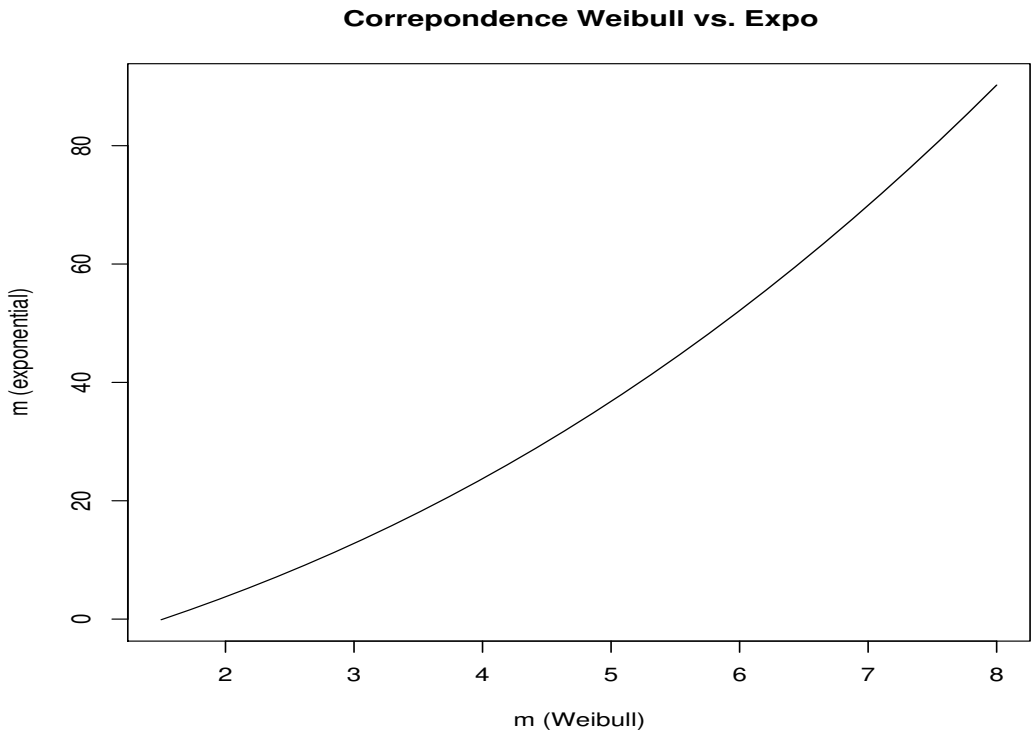
Marin 2006. Assume having elicited π_i. Then π_{i+1} is equitable w.r.t. π_i if

$$\pi_{i+1} \quad = \quad \arg\min \mathrm{KL}\left(m_i(.)|m_{i+1}(.)\right) \quad = \quad \arg\min \int m_i(t) \log \frac{m_i(t)}{m_{i+1}(t)} dt$$

with $m_i(t) = \int f_i(t|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i$

- m is a thought experiment from the Bayesian analyst
- m is compared to $n \Leftrightarrow m$ is linked to the choice of a particular model (or model dimension)
- If $\dim(f_i) > \dim(f_{i+1})$, m data yield more information on θ_{i+1} than θ_i
- For a same marginal information, m_{i+1} should be greater than m_i

Rule : given m_i, minimizing in m_{i+1} the Kullback divergence between the encompassing predictive model and the nested predictive model



In numerous practical cases, one may dispose of several possible priors $\pi_1(\theta), \ldots, \pi_M(\theta)$ assumed to be **independent**

Example : reunions of pharmacologist experts before putting a medicine on the market

A first idea : **weighted linear merging** (arithmetical average)

$$\pi(\theta) \quad = \quad \sum_{i=1}^{M} \omega_i \pi_i(\theta)$$

with $\sum_{i=1}^{M} \omega_i = 1$

Issues :

- the result can be multi-modal
- *not externally Bayesian* :

$$\pi(\theta|\mathbf{x_n}) \quad \neq \quad \sum_{i=1}^{M} \omega_i \pi_i(\theta|\mathbf{x_n})$$

for one or several data $\mathbf{x_n}$

A second idea : **weighted logarithmic merging** (geometrical average)

$$\pi(\theta) \quad = \quad \frac{\prod_{i=1}^{M} \pi_i^{\omega}(\theta)}{\int_{\Theta} \prod_{i=1}^{M} \pi_i^{\omega}(\theta) \; d\theta}$$

with $\sum_{i=1}^{M} \omega_i = 1$

It is externally Bayesian

Issues : it is not *coherent by marginalizing*

- Let A et B be two event such that $A \cap B = \emptyset$ et $C = A \cup B \Rightarrow P(C) = P(A) + P(B)$
- Consider two experts providing their opinions on the occurence of events A and B
- For each expert, one may directly calculate $P(C)$ ou calculate separetely $P(A)$ then $P(B)$
- Only the linear merging allows the equality of both calculus

In reality, the logarithmic merging appears a better choice since it can be explained by an information-theoretic argument

The Kullback-Leibler divergence

$$KL(\pi, \pi_i) = \int_\Theta \pi(\theta) \log \frac{\pi(\theta)}{\pi_i(\theta)}$$

expresses an information loss when the best prior choice π is replaced by π_i

The minimizer of the weighted loss

$$\pi^*(\theta) = \arg\min_\pi \sum_{i=1}^{M} \omega_i KL(\pi, \pi_i)$$

is the logarithmic merging prior

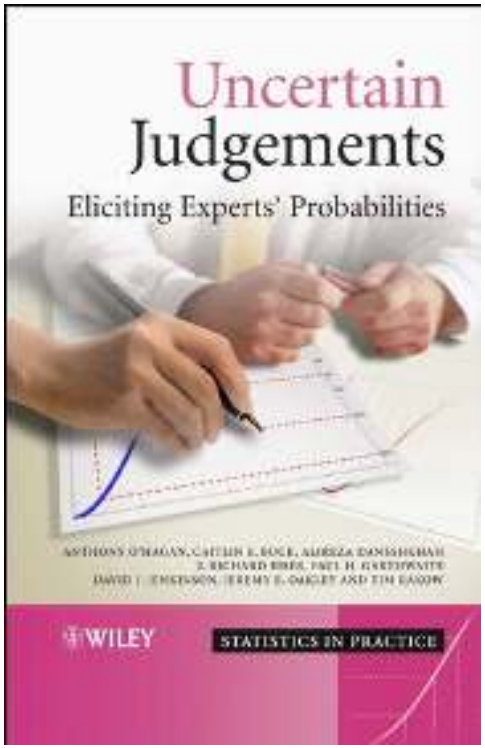
The calibration of weights ω_i is an open problem, although several answers exist

Take-home messages

- Benchmark priors are fundamental : basis for learning from virtual or past data, for criteria...
- Ask the question of “which parameters is my knowledge independent on ?"
- Do not forget that usually, a true expert does not know what a statistical parameter is
- Ask the question “which are the hyperparameters I must give a sense to for defending my prior
- Virtual data posterior priors are nice to quantify the “strength" of subjective information

For other details of elicitation problems...

Most recent overview of elicitation problems in [23]



A. Barberousse.

La valeur de la connaissance approchée. l'épistémologie de l'approximation d'émile borel.

Revue d'Histoire des Mathématiques, 14 :53–75, 2008.

C. Bioche.

Approximation de lois impropres et applications.

Thèse de doctorat de l'Université Blaise Pascal, Clermont-Ferrand, 2015.

S.C.Y. Chan, Y. Niv, and K.A. Norman.

A probability distribution over latent causes, in the orbifrontal cortex.

The Journal of Neuroscience, 36 :7817–7828, 2016.

R.M. Cooke.

Experts in Uncertainty : Opinion and Subjective Probability in Science.

Oxford University Press, 1991.

R.T. Cox.

Probability, frequency and Reasonable Expectation.

American Journal of Physics, 14 :1–10, 1946.

Bibliography (cont.)

S. Dehaene.

The statistician brain : the Bayesian revolution in cognitive science.

Lecture at College de France : Chair in Experimental Cognitive Psychology, 2012.

S. Dehaene.

Consciousness and the Brain : Deciphering How the Brain Codes our Thoughts.

Viking Press, 2014.

M. Drescher, A.H. Perera, C.J. Johson, L.J. Buse, C.A. Drew, and M.A. Burgman.

Toward rigorous use of expert knowledge in ecological research.

Ecosphere, 4 :1–26, 2013.

D. Dubois and H. Prade.

Possibility Theory.

Springer, 2012.

M.J Dupré and F.J. Tipler.

New axioms for rigorous Bayesian probability.

Bayesian Analysis, 4 :599–606, 2009.

Bibliography (cont.)

A. Eagle.

Philosophy of Probability : Contemporary Readings.

Routledge, 2011.

B. Fischhoff and D. MacGregor.

Subjective confidence in forecasts.

Journal of Forecasting, 1 :155–172, 1982.

A. Gelman and C. Hennig.

Beyond subjective and objective in statistics.

Journal of the Royal Statistical Society, Ser. A, 180 :1–31, 2017.

B. Goertzel.

Probability Theory Ensues from Assumptions of Approximate Consistency : A Simple Eerivation and its Implications for AGI.

Proceedings of AGI-13, Springer, 2013.

J.I. Gold and H.R. Heekeren.

Neural Mechanisms for Perceptual Decision Making.

In : Neuroeconomics (chapter 19), P.W. Glimcher and R. Fehr (eds), Second Edition, 2013.

J.Y. Halpern.

A counterexample to theorems of cox and fine.

Journal of Artificial Intelligence Research, 10 :67–85, 1999.

E.T. Jaynes.

Probability Theory : The Logic of Science.

2003.

I. Lakatos.

Falsification and the methodology of scientific research programmes. In : Criticism and the Growth of Knowledge, I. Lakatos and A. Musgrave (eds.).

Cambridge University Press, 1970.

J. Leal, S. Wordsworth, R. Legood, and E. Blair.

Eliciting expert opinion for economic models : an applied example.

Value Health, 10 :195–203, 2007.

M. Luntley.

Understanding Expertise.

Journal of Applied Philosophy, 26 :356–370, 2009.

Bibliography (cont.)

C.A. Miller and P.N. (eds) Edwards.

Changing the Atmosphere. Expert Knowledge and Environmental Governance.

MIT Press, 2001.

M.G. Morgan.

Use (and abuse) of expert elicitation in support of decision making for public policy.

Proceedings of the US National Academy of Sciences, 111 :7176–7184, 2014.

A. O'Hagan, C.E. Buck, A. Daneshkhah, J.R. Eiser, P. Garthwaite, D.J. Jenkinson, J.E. Oakley, and T. Rakow.

Uncertain Judgments : Eliciting Expert's Probabilities.

Wiley, 2006.

J.B. Paris.

The Uncertain Reasoner's Companion : a Mathematical Perspective.

Cambridge University Press, 1994.

M. Pegny.

Les deux formes de la thèse de church-turing et l'épistémologie du calcul.

Philosophia Scientiae, 16 :36–67, 2012.

Bibliography (cont.)

K. Popper.

Objective Knowledge.

Oxford : Clarendon Pr, 1972.

A. Pouget, J.M. Beck, W.J. Ma, and P.E. Latham.

Probabilistic brains : knowns and unknowns.

Nature Neuroscience, 16 :1170–1178, 2016.

E. Salinas.

Prior and prejudice.

Nature Neuroscience, 14 :943–945, 2011.

L. Sanders.

The probabilistic mind.

Science News, 180 :18, 2011.

G. Shafer.

A mathematical theory of evidence.

Princeton University Press, 1976.

P. Smets.

The transferable belief model and other interpretations of Dempster-Schafer's model, pages 375–383.

Elsevier Science Publishers, 1991.

P. Snow.

On the correctness and reasonableness of cox's theorem for finite domains.

Computational Intelligence, 17 :178–192, 1998.

A. Terenin and D. Draper.

Rigorizing and extending the cox-jaynes derivation of probability : Implications for statistical practice.

Submitted. arXiv :1507.06597, 2015.

E.R.W. Tredger, J.T.H. Lo, S. Haria, H.H.K. Lay, N. Bonello, B. Hlavka, and C. Scullion.

Bias, guess and expert judgement in actuarial work.

British Actuarial Journal, 21 :545–578, 2016.

K.S. Van Horn.

Constructing a logic of plausible inference : a guide to cox's theorem.

International Journal of Approximate Reasoning, 34 :3–24, 2003.

Bibliography (cont.)

P. Walley.

Measures of uncertainty in expert systems.

Artificial Intelligence, 83 :1–58, 1996.

B. Weinstein.

What is an expert ?

Theoretical Medicine, 14 :57–73, 1993.