Auditable Bayesian modeling for quality measurements

Nicolas Bousquet

EDF R&D & Sorbonne Université

- **Quantifying and separating uncertainties in model-based decision-helping processes** (often through Bayesian and machine/deep learning approaches)
  - computing cautious assessments
  - conducting risk and reliability studies

Dealing with measurement and model uncertainties seems to bring me closer to the concerns of meteorologists

A shared interest ?

Selecting good data / good measurements to say something about the (claimed) behavior of a system / model, in a broad (but formalized) sense

- **Quantifying and separating uncertainties in model-based decision-helping processes** (often through Bayesian and machine/deep learning approaches)
  - computing cautious assessments
  - conducting risk and reliability studies

Dealing with measurement and model uncertainties seems to bring me closer to the concerns of meteorologists

A shared interest ?

Selecting good data / good measurements to say something about the (claimed) behavior of a system / model, in a broad (but formalized) sense

- **Quantifying and separating uncertainties in model-based decision-helping processes** (often through Bayesian and machine/deep learning approaches)
  - computing cautious assessments
  - conducting risk and reliability studies

Dealing with measurement and model uncertainties seems to bring me closer to the concerns of meteorologists

A shared interest ?

Selecting good data / good measurements to say something about the (claimed) behavior of a system / model, in a broad (but formalized) sense

Thanks

Works and thoughts shared with several statisticians : Fabrizio Ruggeri, Adrian Raftery, Anne Philippe, Bertrand Iooss, Mélanie Blazère, Sophie Ancelet, Eric Parent, etc.

Selecting a good measurement is a decision that might be formalized as follows

Let $Y = y_i$ be an (indirect) measurement of a quantity $X = x_i$, understood as

$$y_i \quad = \quad g_\Sigma(x_i, \varepsilon_i)$$

where

- g_Σ is an operator modeling a measurement process Σ
- $\varepsilon \sim P(\varepsilon)$ is a random "noise" summarizing the influence of external factors

For a same (hidden) source $X = x_i$, several values of y_i due to ε_i

Example : lung cancer screening by thoracic scanner

Source : [32]

- X = tumor features
- Y = table of pixels
- ε = patient position + setting chosen by the operator

Based on the GUM, from repeated observations $\boldsymbol{Y}(x)$, assess the quality of a measurement Y by estimating (for instance) the conditional variance

$$\begin{aligned} \mathrm{Var}[Y|X=x] &= \int \ell(g_\Sigma(x,\varepsilon))dP(\varepsilon) \quad \text{with } \ell(u(\varepsilon)) = \mathrm{E}_\varepsilon[u^2(\varepsilon)] - \mathrm{E}_\varepsilon^2[u(\varepsilon)] \\ &= \textit{indicator of measurement uncertainty in } X = x \end{aligned}$$

Assuming $X \sim P_X$, a global indicator of quality for Σ could legitimately be

$$Q_\Sigma = \mathrm{E}_X\left[\mathrm{Var}[Y|X]\right]$$

(note that is can be estimated only with a sample $\boldsymbol{Y} = \{\boldsymbol{Y}_{ij}(x_i)\}_{i,j}$ without knowing the real x_i)

Now, having two competing measurement processes Σ_1 and Σ_2, may we compare the Q_{Σ_i} to check if "Σ_1 is better than Σ_2" ?

Based on the GUM, from repeated observations $\boldsymbol{Y}(x)$, assess the quality of a measurement Y by estimating (for instance) the conditional variance

$$\begin{aligned} \mathrm{Var}[Y|X=x] &= \int \ell(g_\Sigma(x,\varepsilon))dP(\varepsilon) \quad \text{with } \ell(u(\varepsilon)) = \mathrm{E}_\varepsilon[u^2(\varepsilon)] - \mathrm{E}_\varepsilon^2[u(\varepsilon)] \\ &= \textit{indicator of measurement uncertainty in } X = x \end{aligned}$$

Assuming $X \sim P_X$, a global indicator of quality for Σ could legitimately be

$$Q_\Sigma = \mathrm{E}_X\left[\mathrm{Var}[Y|X]\right]$$

(note that is can be estimated only with a sample $\boldsymbol{Y} = \{\boldsymbol{Y}_{ij}(x_i)\}_{i,j}$ without knowing the real x_i)

Now, having two competing measurement processes Σ_1 and Σ_2, may we compare the Q_{Σ_i} to check if "Σ_1 is better than Σ_2" ?

What we want from using each Σ_i is to reconstruct X, or rather P_X (in a concern of generality), using $\boldsymbol{Y}_{\Sigma_i}$ (*stochastic inversion*)

Classical approach.

1. Assume $X \sim P_X(.|\theta)$ parameterized by θ (e.g., a multivariate Gaussian)
2. Estimate θ from $\boldsymbol{Y}_{\Sigma_i}$ (e.g., using missing data, EM-type algorithms [13, 5])

$$\theta \quad \Rightarrow \quad \hat{\theta}(\boldsymbol{Y}_{\Sigma_i})$$

Then

$$Q_{\Sigma_i} \quad = \quad Q_{\Sigma}\left(\hat{\theta}(\boldsymbol{Y}_{\Sigma_i})\right)$$

but we cannot be sure to have a total order between the Q_{Σ_i} [44]
$\Leftrightarrow$ we cannot properly compare Σ_1 and Σ_2

What we want from using each Σ_i is to reconstruct X, or rather P_X (in a concern of generality), using $\boldsymbol{Y}_{\Sigma_i}$ (*stochastic inversion*)

Bayesian approach.

1. Note that θ is a summary of the features of $X \sim P_X$, endowed with epistemic uncertainty
2. Model this uncertainty by defining technically θ as a random variable with **prior measure**
$$\theta \sim \pi(\theta)$$
3. Estimate the **posterior** $\pi(\theta|\boldsymbol{Y}_{\Sigma_i})$ (e.g., using Monte Carlo-type algorithms [19, 20])
$$\pi(\theta) \quad \Rightarrow \quad \pi(\theta|\boldsymbol{Y}_{\Sigma_i}) \qquad \textit{(Bayesian updating)}$$

Then

$$Q_{\Sigma_i} \quad = \quad \mathrm{E}_\theta \left[\mathrm{E}_X \left[\mathrm{Var}[Y|X]|\theta \right] | \boldsymbol{Y}_{\Sigma_i} \right]$$

It is a Bayes estimator then we are sure to get a total order between the Q_{Σ_i}
$\Leftrightarrow$ we can compare Σ_1 and Σ_2

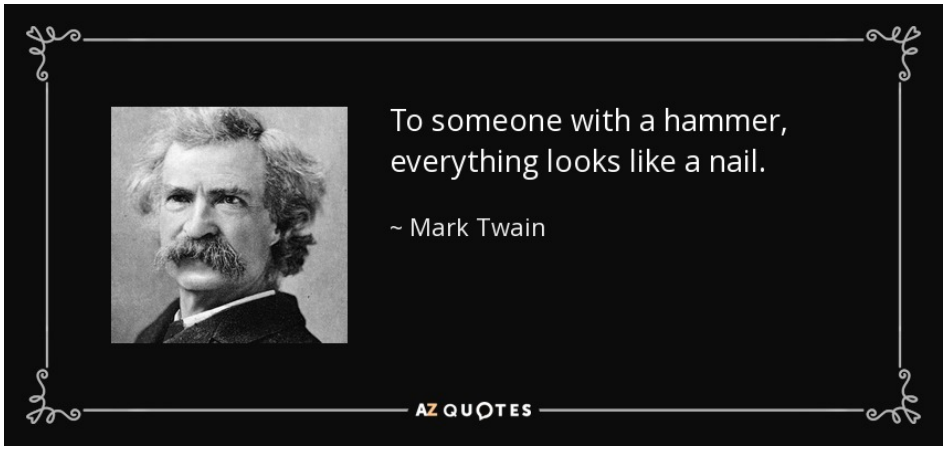
The Bayesian paradigm can be useful to have a global quality and correctly treat uncertainties

And for selecting good measurements ? Does the prior requires significant work ?

Analogy

Testing a claim on an industrial component or system is somewhat similar to testing whether that component or system can withstand certain stresses

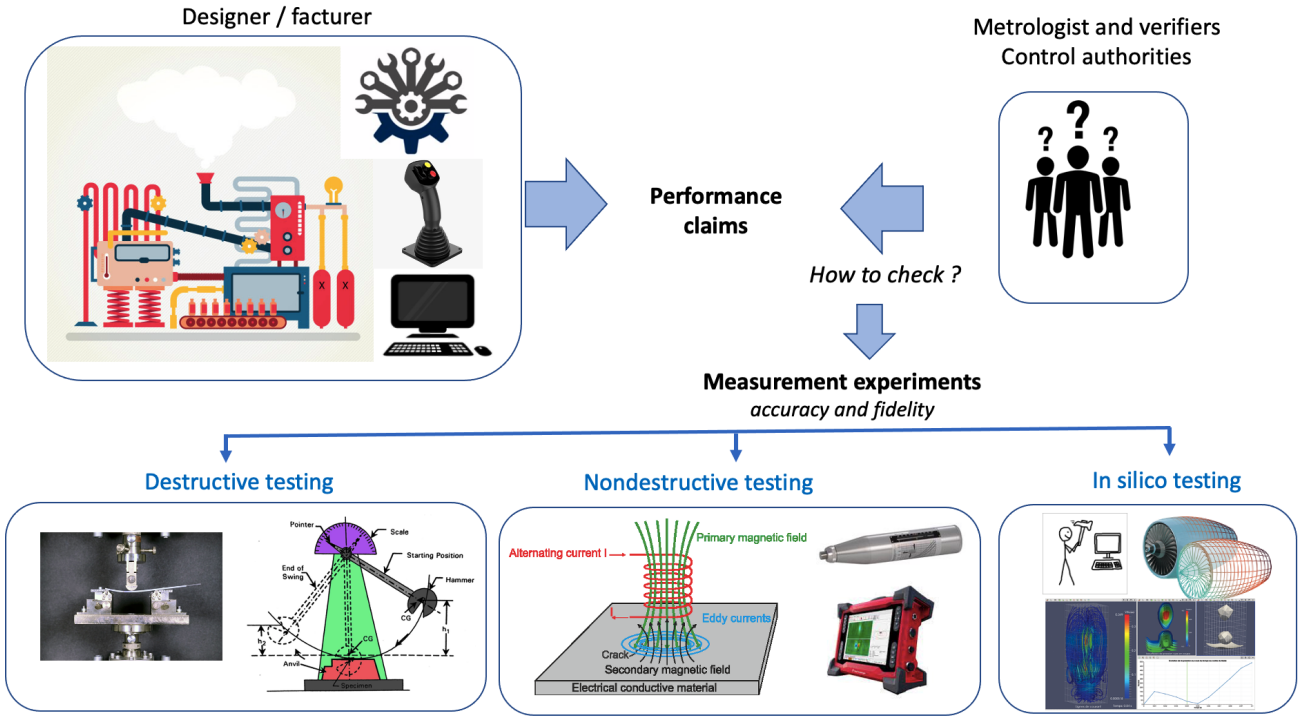
Probably a limited analogy, but can be helpful



(or maybe Abraham Maslow)

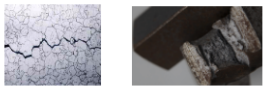
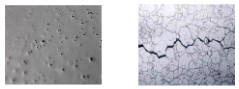
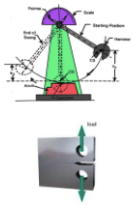
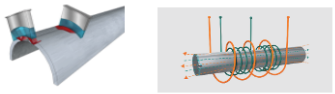
Analogy

Testing a claim on an industrial component or system is somewhat similar to testing whether that component or system can withstand certain stresses



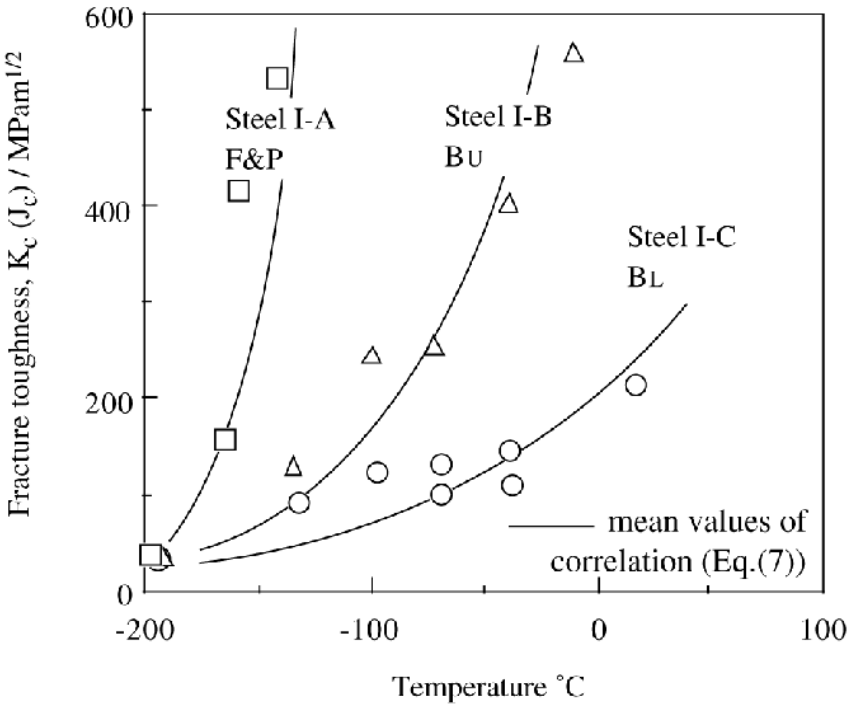
In each case, how designing *good experiments* (ie to have good quality measurements) ?

Designing good (informative) experiments

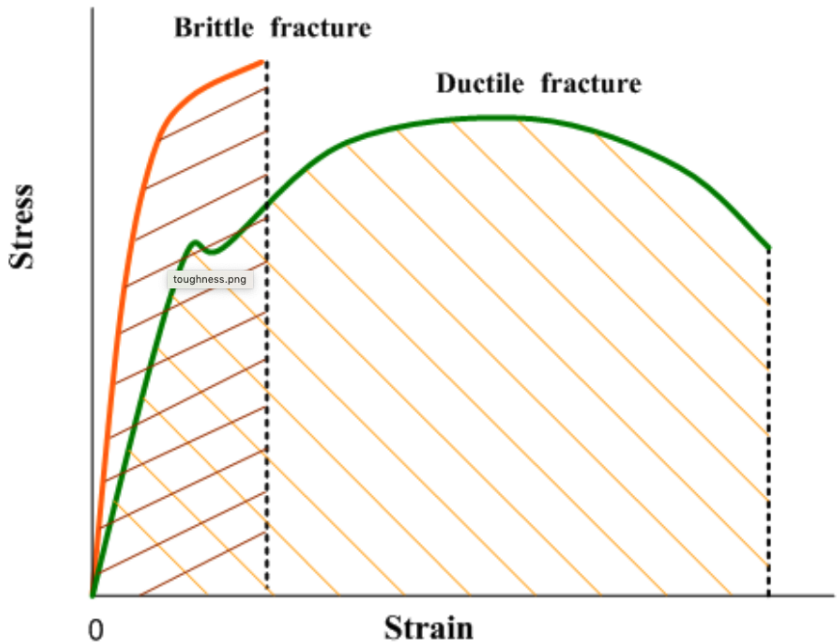
	Destructive experiments	**Nondestructive experiments**	**In silico experiments**
Examples of tested properties	• Fracture toughness • Stress corrosion cracking 	• Stable undercoating defects • Stress corrosion cracking 	• Robustness of an artificial intelligence (AI) tool • Fidelity of a physically-based digital twin
Some experimental techniques	• Charpy-type experiments  • ALT chemical testing, etc.	• Ultrasonic inspections • Eddy current testing 	• Selecting among collected observations • Designing numerical experiments using optimization
Typical cost constraints	• Very limited number of specimen • Selecting stress levels and specimen features • Strong environemental conditions (e.g. T°) • Noise removal requirement • Availability of experts		• Prohibitive training time • Prohibitive inference / simulation time related to model complexity

Fracture toughness of steel (FTOS) characterizes the capacity of the material to resist to cracking through plastic deformation when a load is applied (e.g., a transient cooling such as water injection)

It is part of the most influential material attributes in structural safety studies [46].

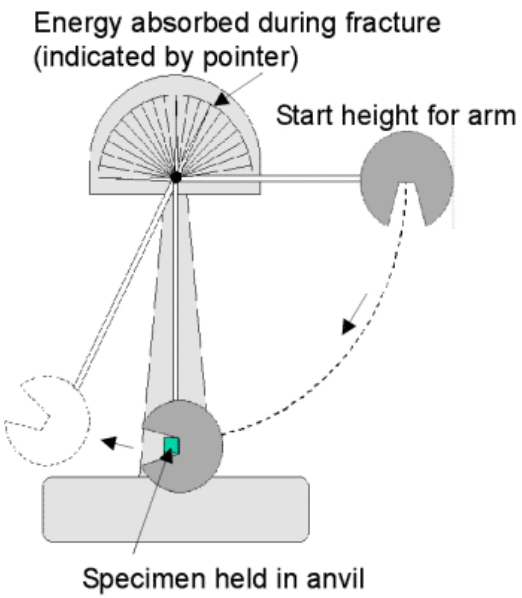


Source : [35]



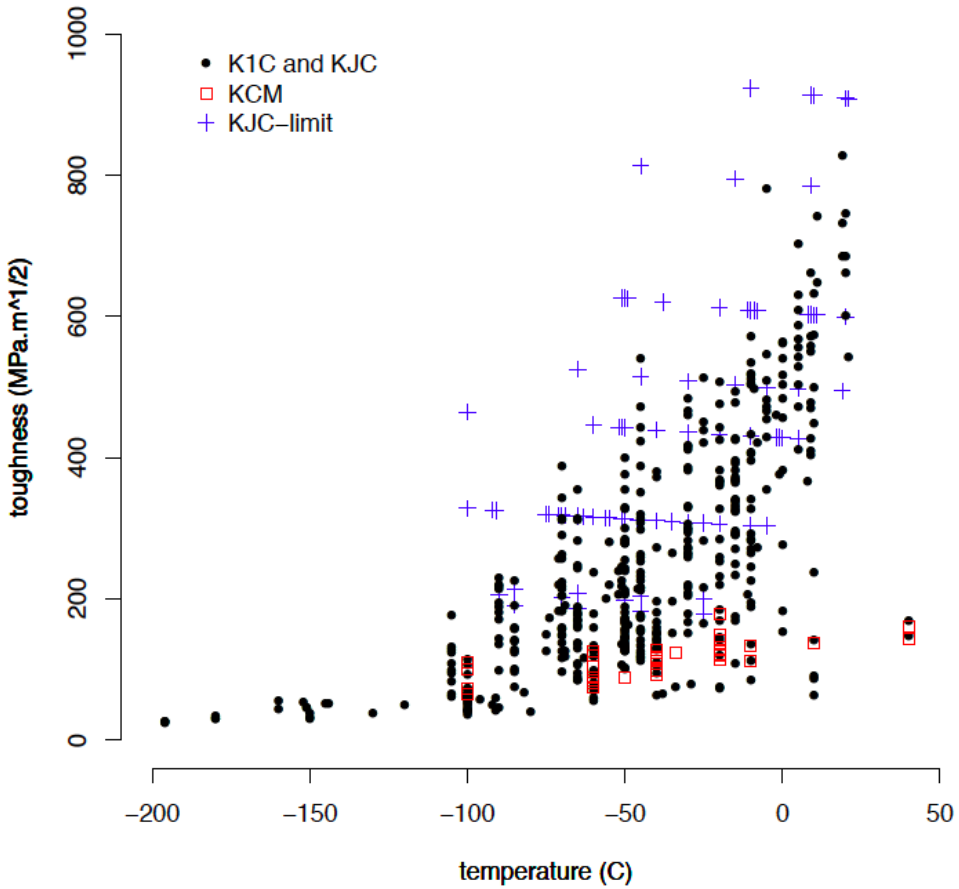
Source : https://www.substech.com

Charpy impact tests [4] $\Rightarrow$ indirect toughness values (megapascal square root meter) with different qualities



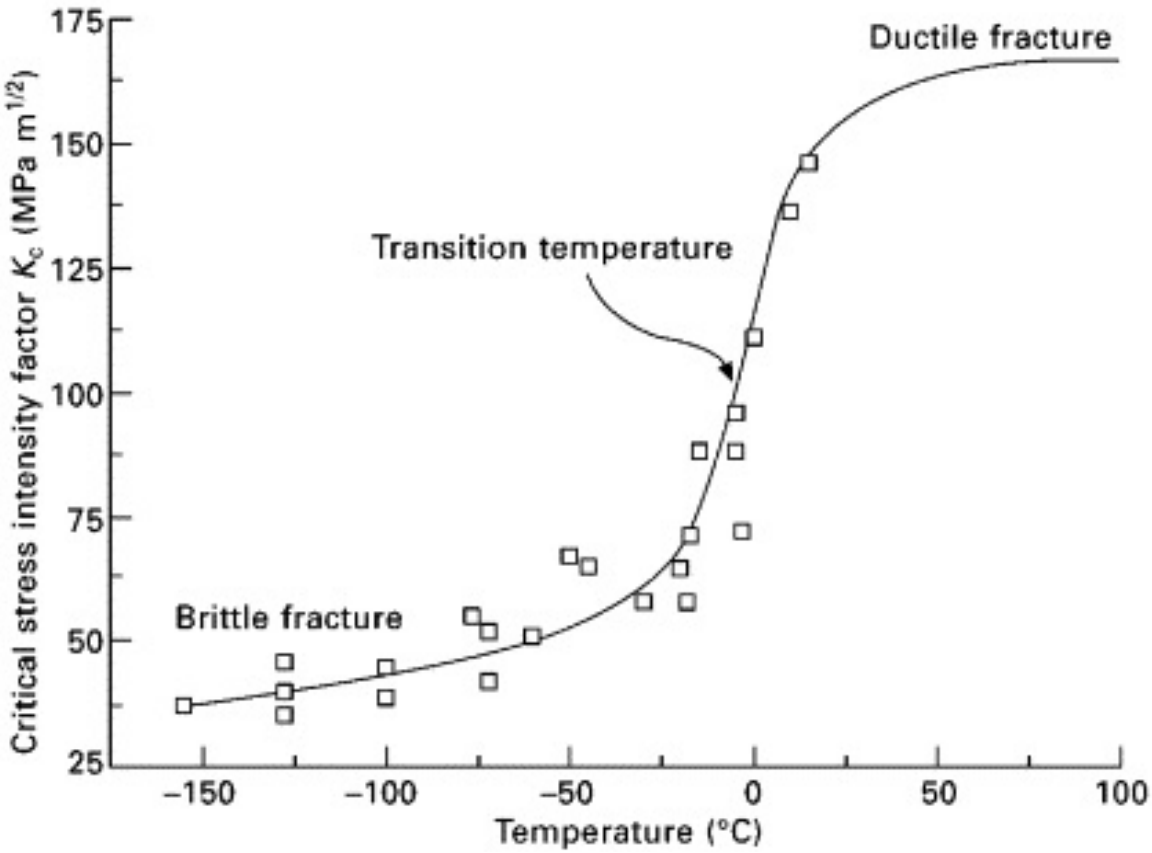
Source : `https://theconstructor.org`

European FTOS database (ferritic steels) from Oak Ridge National Laboratory (ASTM E399-90 [3])



Goal

Checking if the brittle-ductile transition temperature T_0 is as claimed by the supplier



Source : [36]

T_0 is a deterministic function of the distribution P of FTOS

$\Updownarrow$

It is tantamount to select a (very) limited of measurements
that offer the best possible knowledge on P (with lowest uncertainties)
$\Leftrightarrow$ the most informative measurements

If P were known / simulable, quantization techniques like Maximum Mean Discrepancy minimization could be used (ie., using kernel herding, grid search or Sequential Bayesian Quadrature [42])

1. Consider a well recognized theoretical statistical model (e.g., from *weakest link theory [26]*) linking a FTOS measure $y_i^j \in \Omega$ at a given temperature T_j and $T_0 = g(\theta)$

$$P(Y_i^j < y | T_j, \theta) \quad = \quad 1 - \exp\left(- \left\{ \frac{y_i^j - \alpha)}{\mu(T_j)} \right\}^{\beta} \right) \qquad \textit{(simple Master Curve [47])}$$

with $\mu(T_j) = \lambda_1 + \lambda_2 \exp(\lambda_3 T_j)$ and $\theta = (\alpha, \{\lambda_i\}_i, \beta)$

2. Elicit a good prior distribution $\Pi(\theta)$
3. Formalize a design of experiments for fixed n standard Charpy specimen [25 mm]

$$\varepsilon = \left\{ J, \left\{ \begin{array}{cccc} T_1 & T_2 & \dots & T_J \\ \eta_1 & \eta_2 & \dots & \eta_J \end{array} \right\} \right\}$$

with $\eta_j = \frac{n_j}{n} \in [0, 1]$ for all $j = 1, \dots, J$ and $\sum_{j=1}^{J} \eta_j = 1$

For tractability, relax the assumption $n \times \eta_j \in \mathbb{N} \Rightarrow$ find a probability measure $\eta = (\eta_1, \dots, \eta_J)$ with optimal J^* and temperatures T_j^*

1. Consider a well recognized theoretical statistical model (e.g., from *weakest link theory [26]*) linking a FTOS measure $y_i^j \in \Omega$ at a given temperature T_j and $T_0 = g(\theta)$

$$P(Y_i^j < y | T_j, \theta) \quad = \quad 1 - \exp\left(- \left\{ \frac{y_i^j - \alpha)}{\mu(T_j)} \right\}^{\beta} \right) \qquad \textit{(simple Master Curve [47])}$$

with $\mu(T_j) = \lambda_1 + \lambda_2 \exp(\lambda_3 T_j)$ and $\theta = (\alpha, \{\lambda_i\}_i, \beta)$

2. Elicit a good prior distribution $\Pi(\theta)$
3. Formalize a design of experiments for fixed n standard Charpy specimen [25 mm]

$$\varepsilon = \left\{ J, \left\{ \begin{array}{cccc} T_1 & T_2 & \dots & T_J \\ \eta_1 & \eta_2 & \dots & \eta_J \end{array} \right\} \right\}$$

with $\eta_j = \frac{n_j}{n} \in [0, 1]$ for all $j = 1, \dots, J$ and $\sum_{j=1}^{J} \eta_j = 1$

For tractability, relax the assumption $n \times \eta_j \in \boldsymbol{N} \Rightarrow$ find a probability measure $\eta = (\eta_1, \dots, \eta_J)$ with optimal J^* and temperatures T_j^*

Last ingredient : an utility function

$U_1(\varepsilon)$ = expected utility function quantifying the expected gain in knowledge about θ provided by data collected under the experimental design ε

$U_2(\varepsilon)$ = expected utility function quantifying the opposite of the expected experimental cost under ε

Generic (compound) weighted (dimensionless) utility [2] (similar idea in [25])

$$U(\varepsilon) = \omega \times \Delta U_1(\varepsilon) + (1 - \omega) \times \Delta U_2(\varepsilon)$$

where

$$\Delta U_k(\varepsilon) = \frac{U_k(\varepsilon) - U_k(\varepsilon_0)}{|U_k(\varepsilon_0)|} \qquad \text{for} \qquad k = 1, 2$$

- $\Delta U_k(\varepsilon)$ = relative change in expected utility
- ε_0 = fixed baseline experimental design for which the total expected utility $U(\varepsilon_0)$ is set to zero
- for instance (using typical temp values within the brittle-ductile transition zone)

$$\varepsilon_0 = \left\{ 4, \left\{ \begin{array}{cccc} -150 & -100 & -50 & 0 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\} \right\}$$

1. Quantifying the opposite of the number of days of work required for collecting data at a design point :

$$U_2(\varepsilon) = -\sum_{j=1}^{J} n_j \times \left(2 - \mathbf{1}_{\{T^- < T_j < T^+\}}\right)$$

where $T^- = -130^\circ C$, $T^+ = -60^\circ C$

(one day of work to make a test when $T \in [T^-, T^+]$ but two days to homogenize the room temperature in more extreme conditions)

2. Quantifying the expected gain in knowledge provided by data collected under an experimental design ε about θ

[Ex.1] Posterior-prior KL divergence $\Rightarrow$ all dimensions of θ

$$U_1^1(\varepsilon) = \int_\Omega \int_\Theta \log \frac{\pi(\theta|\mathbf{y}, \varepsilon)}{\pi(\theta)} \pi(\theta|\mathbf{y}, \varepsilon) d\theta d\mathbf{y}$$

1. Quantifying the opposite of the number of days of work required for collecting data at a design point :

$$U_2(\varepsilon) = -\sum_{j=1}^{J} n_j \times \left(2 - \mathbf{1}_{\{T^- < T_j < T^+\}}\right)$$

where $T^- = -130^\circ C$, $T^+ = -60^\circ C$

(one day of work to make a test when $T \in [T^-, T^+]$ but two days to homogenize the room temperature in more extreme conditions)

2. Quantifying the expected gain in knowledge provided by data collected under an experimental design ε about θ

[Ex.2] Opposite of the quadratic loss function $\Rightarrow$ selected linear combination of dimensions of θ

$$U_1^2(\varepsilon) = -\int_{\Omega}\int_{\Theta} (\theta - \hat{\theta})^T A(\theta - \hat{\theta}) f(\mathbf{y}, \theta|\varepsilon) d\theta d\mathbf{y}$$

with $\hat{\theta}$ = Bayesian point estimate of θ and A symmetric nonnegative definite matrix

One may write

$$U_1^k(\varepsilon) = \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\theta|Y}\left(u^k(\theta,\mathbf{Y},\varepsilon)\right) \quad \text{with} \quad \left\{ \begin{array}{rcl} u^1(\theta,\mathbf{y},\varepsilon) & = & \log\pi(\theta|\mathbf{y},\varepsilon) \\ u^2(\theta,\mathbf{y},\varepsilon) & = & -(\theta-\hat{\theta})^T A(\theta-\hat{\theta}) \end{array} \right\}$$

and under the asymptotic approximation (Bernstein-von Mises)

$$\theta|\mathbf{y},\varepsilon \simeq \mathcal{N}_d\left(\widehat{\theta}, \Sigma(\widehat{\theta},\varepsilon) = [nI(\widehat{\theta},\varepsilon) + R]^{-1}\right)$$

(with $\widehat{\theta}$ = posterior mode, $I(.,.)$ = Fisher matrix and R = prior precision matrix), it comes

$$\begin{array}{rcl} \mathbb{E}_{\theta|Y}\left(u^1(\theta,\mathbf{Y},\varepsilon)\right) & \simeq & cte + \frac{1}{2}\log\left(\det(\Sigma(\widehat{\theta},\varepsilon)^{-1})\right) \\ \mathbb{E}_{\theta|Y}\left(u^2(\theta,\mathbf{Y},\varepsilon)\right) & \simeq & -\text{tr}(A\Sigma(\widehat{\theta},\varepsilon)) \end{array}$$

- Maximizing the expected $\mathbb{E}_{\theta|Y}\left(u^1(\theta,\mathbf{Y},\varepsilon)\right) \Leftrightarrow D-$optimal design
- Maximizing the expected $\mathbb{E}_{\theta|Y}\left(u^2(\theta,\mathbf{Y},\varepsilon)\right) \Leftrightarrow A-$optimal design

See [31, 24] for convergence results to sequential maximum likelihood-based adaptive designs for treatment allocation)

One may write

$$U_1^k(\varepsilon) = \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\theta|Y}\left(u^k(\theta,\mathbf{Y},\varepsilon)\right) \quad \text{with} \quad \left\{ \begin{array}{rcl} u^1(\theta,\mathbf{y},\varepsilon) & = & \log\pi(\theta|\mathbf{y},\varepsilon) \\ u^2(\theta,\mathbf{y},\varepsilon) & = & -(\theta-\hat{\theta})^T A(\theta-\hat{\theta}) \end{array} \right\}$$

and under the asymptotic approximation (Bernstein-von Mises)

$$\theta|\mathbf{y},\varepsilon \simeq \mathcal{N}_d\left(\widehat{\theta}, \Sigma(\widehat{\theta},\varepsilon) = [nI(\widehat{\theta},\varepsilon) + R]^{-1}\right)$$

(with $\widehat{\theta}$ = posterior mode, $I(.,.)$ = Fisher matrix and R = prior precision matrix), it comes

$$\begin{array}{rcl} \mathbb{E}_{\theta|Y}\left(u^1(\theta,\mathbf{Y},\varepsilon)\right) & \simeq & cte + \frac{1}{2}\log\left(\det(\Sigma(\widehat{\theta},\varepsilon)^{-1})\right) \\ \mathbb{E}_{\theta|Y}\left(u^2(\theta,\mathbf{Y},\varepsilon)\right) & \simeq & -\mathrm{tr}(A\Sigma(\widehat{\theta},\varepsilon)) \end{array}$$

- Maximizing the expected $\mathbb{E}_{\theta|Y}\left(u^1(\theta,\mathbf{Y},\varepsilon)\right) \Leftrightarrow D-$optimal design
- Maximizing the expected $\mathbb{E}_{\theta|Y}\left(u^2(\theta,\mathbf{Y},\varepsilon)\right) \Leftrightarrow A-$optimal design

See [31, 24] for convergence results to sequential maximum likelihood-based adaptive designs for treatment allocation)

$$I(\theta,\epsilon) = \begin{pmatrix} I11 & I12 & I13 & I14 \\ I12 & I22 & I23 & I24 \\ I13 & I23 & I33 & I34 \\ I14 & I24 & I34 & I44 \end{pmatrix}.$$

with

$$I_{11} = (\beta-1)^2\Gamma\left(1-\frac{2}{\beta}\right)\sum_{j=1}^{J}\frac{n_j}{\mu(T_j)^2}$$

$$I_{22} = \beta^2\sum_{j=1}^{J}\frac{n_j}{\mu(T_j)^2}$$

$$I_{33} = \beta^2\sum_{j=1}^{J}\frac{n_j\exp^2(\lambda_3 T_j)}{\mu(T_j)^2}$$

$$I_{44} = \beta^2\lambda_2^2\sum_{j=1}^{J}\frac{n_j T_j^2 exp^2(\lambda_3 T_j)}{\mu(T_j)^2}$$

$$I_{12} = \beta^2\left(1-\frac{1}{\beta}\right)\Gamma\left(1-\frac{1}{\beta}\right)\sum_{j=1}^{J}\frac{n_j}{\mu(T_j)^2}$$

$$I_{13} = \beta^2\left(1-\frac{1}{\beta}\right)\Gamma\left(1-\frac{1}{\beta}\right)\sum_{j=1}^{J}\frac{n_j\exp(\lambda_3 T_j)}{\mu(T_j)^2}$$

$$I_{14} = \beta^2\lambda_2\left(1-\frac{1}{\beta}\right)\Gamma\left(1-\frac{1}{\beta}\right)\sum_{j=1}^{J}\frac{n_j T_j\exp(\lambda_3 T_j)}{\mu(T_j)^2}$$

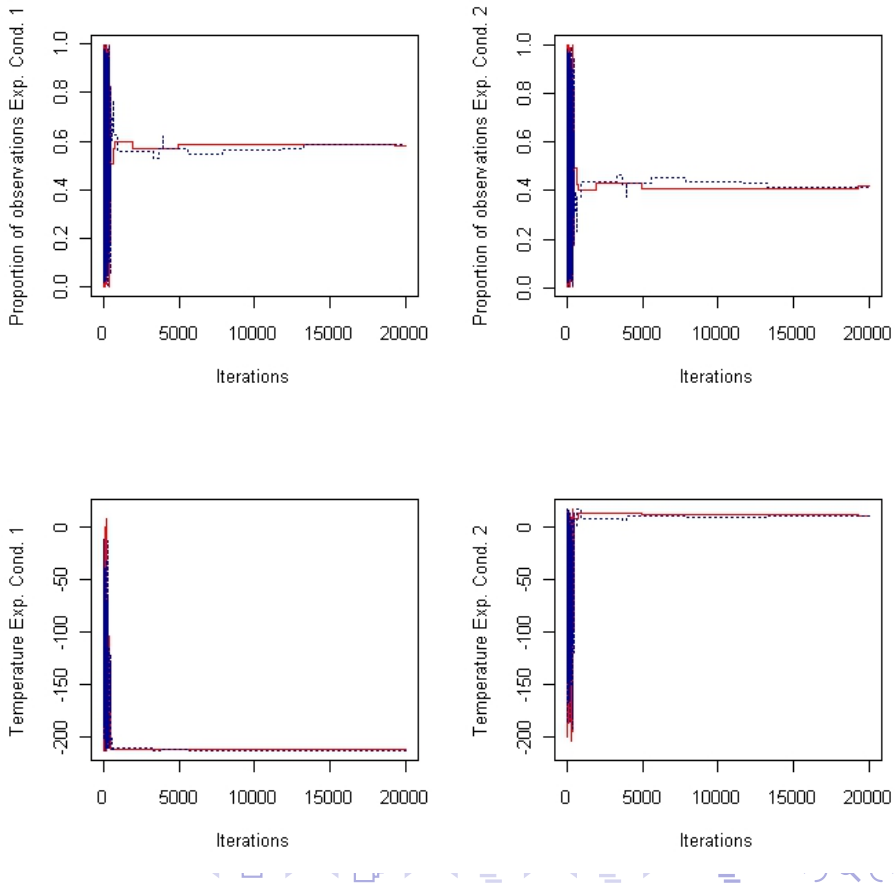
$$I_{23} = \beta^2\sum_{j=1}^{J}\frac{n_j exp(\lambda_3 T_j)}{\mu(T_j)^2}$$

$$I_{24} = \beta^2\lambda_2\sum_{j=1}^{J}\frac{n_j T_j\exp(\lambda_3 T_j)}{\mu(T_j)^2}$$

$$I_{34} = \beta^2\lambda_2\sum_{j=1}^{J}\frac{n_j T_j\exp^2(\lambda_3 T_j)}{\mu(T_j)^2}$$

1. Choose a Gaussian prior computed as an approximation of a posterior from European FTOS data (flat baseline prior)
2. Solve the problem by computational techniques like simulated annealing [2] or more recently the approximate coordinate exchange algorithm [40]

$$\varepsilon_0 = \left\{ 4, \begin{array}{cccc} -150 & -100 & -50 & 0 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array} \right\}$$



Application [2]

ω	u_1	J^*	η^*	T^*	$\widetilde{U}(\epsilon^*)$
1	D	J=3	(0.55,0.27,0.18)	(-213.84,-97.52,17.80)	0.046
	A_1	J=2	(0.31,0.69)	(-213.80,9.21)	0.156
	A_2	J=2	(0.58,0.42)	(-213.70,12.48)	0.102
0.9	A_1	J=2	(0.31,0.69)	(-213,91,7.62)	0.126
	A_2	J=3	(0.54,0.10,0.36)	(-213.96,-60.21,17.71)	0.079
0.5	A_1	J=3	(0.49,0.42,0.09)	(-129.51,-60.10,17.92)	0.164
	A_2	J=2	(0.92,0.08)	(-129.97,-60.37)	0.200

Charpy impact tests are defined by an intensity η that may be insufficient ($\eta < \eta_0$) or too high ($\eta > \eta_1$) $\Rightarrow$ Fracture too shallow or deep

Taking account of this experimental difficulty by considering rather

$$\left\{1 - \exp\left(-\left\{\frac{y_i^j(\eta_i^j) - \alpha)}{\mu(T_j)}\right\}^{\beta}\right)\right\} \mathbb{1}_{[\eta_0, \eta_1]}(\eta_i^j)$$

as the cdf of an acceptable measurement y_i^j (providing information on θ)

Uncertain parameters (η_0, η_1) are nuisance parameters from the point of view of the statistician that wants to estimate θ

A prior $\pi(\eta_0, \eta_1)$ should be chosen within the whole design approach to integrate the limits of the measurement device

- More generally, the addition of noises and measurements limits will decrease the quantity of information yielded by planned experiments
- Asymptotic assumptions and prior (Gaussian) assumptions behind A- and D-optimal design criteria can strongly be not realistic
- In such cases $\Rightarrow$ bad pseudo-Bayesian designs :
  - [27] for generalized linear situations
  - [25] for clinical trials with low convergence of the optimization algorithms (effect of highly concentrated pseudo-posterior)
- Modern computational techniques can tackle the problem of computing repeatedly posteriors to solve the optimization problem of the design ε
  - multi-stages mixing stochastic gradient optimisation and automatic differentiation [41]

We are technically "allowed" to focus now on prior modeling

- Priors ("best guesses") can significantly help to produce useful designs (e.g., [8] for clinical studies)
- All the more when the planned design is small-sized (since costly)
- Remind that priors are used for the randomized planning stage, but they are not required for inference

Producing defensible priors take part in a more general, growing approach of questioning the formalization of prior choices

- A. Gelman and J. Sprenger on the objectivity and reproducibility of Bayesian assessments : [Holes in Bayesian Statistics] [22, 45, 23]
- Contemporary concerns for the auditability of deep learning [18] and artificial intelligence [48]

We are technically "allowed" to focus now on prior modeling

- Priors ("best guesses") can significantly help to produce useful designs (e.g., [8] for clinical studies)
- All the more when the planned design is small-sized (since costly)
- Remind that priors are used for the randomized planning stage, but they are not required for inference

Producing defensible priors take part in a more general, growing approach of questioning the formalization of prior choices

- A. Gelman and J. Sprenger on the objectivity and reproducibility of Bayesian assessments : [Holes in Bayesian Statistics] [22, 45, 23]
- Contemporary concerns for the auditability of deep learning [18] and artificial intelligence [48]

Goal

Produce clear, accountable, repeatable, formal rules for informative Bayesian modeling, consistent with the reality of a prior uncertain information

A (very) long history in objective Bayes [7]

- e.g., Jeffreys prior, reference priors (and variants), maxent priors, MDIP, etc. dedicated to specific tasks or not

$\Rightarrow$ A quick view of the most shared contemporary approaches and challenges

Posterior prior-type assumption [33, 15, 37, 43] ⇒ algebraic structure

- Denote $\tilde{\mathbf{x}}_\mathbf{m} = (\tilde{x}_1, \ldots, \tilde{x}_m) \sim f(x|\theta)$ an *imaginary* iid sample bringing prior information
- Let $\pi^J(\theta)$ an objective (noninformative) prior for $f(x|\theta)$ with support Θ

Then an "ideal" prior is (if integrable)

$$\pi(\theta) \quad = \quad \pi^J(\theta|\tilde{\mathbf{x}}_\mathbf{m})$$

Prior information assumption [6, 21, 29, 34] ⇒ information content

The informative content of $\tilde{\mathbf{x}}_\mathbf{m}$ is mostly provided through a set K of estimates of *marginal prior predictive quantiles (Berger-Kadane assumption)*

$$P_{f_K}(X_i < x_{i,\alpha_{ij}}) \quad = \quad \alpha_{ij} \quad \overset{\text{ideally}}{=} \quad \int P(X_i < x_{i,\alpha_j}|\theta)\pi(\theta)\, d\theta$$

for $i = 1, \ldots, d = \dim X$

[Existence ensured and interpretability through expected pinball loss function [28]]

1 - If $\pi(\theta) = \pi^J(\theta|\tilde{x}_m)$ is tractable

- it provides a natural (coherent) dependence structure on θ

$$\pi(\theta) \uparrow \quad \Rightarrow \quad \det I(\theta) \uparrow$$

- it defends a reasonable way of aggregating M independent priors

$$\pi(\theta) \quad \propto \quad \left(\prod_{i=1}^{M} f(\tilde{x}_{m_i}|\theta) \right) \pi^J(\theta) \qquad (\textit{logarithmic pooling } [16])$$

Example

$f \in$ exponential family, usual choice for $\pi^J \Rightarrow \pi$ is conjugate and writes as

$$\pi(\theta) \quad = \quad \pi(\theta|t(\tilde{x}_m))$$

where $t(.)$ is a sufficient statistics of dimension $< m$

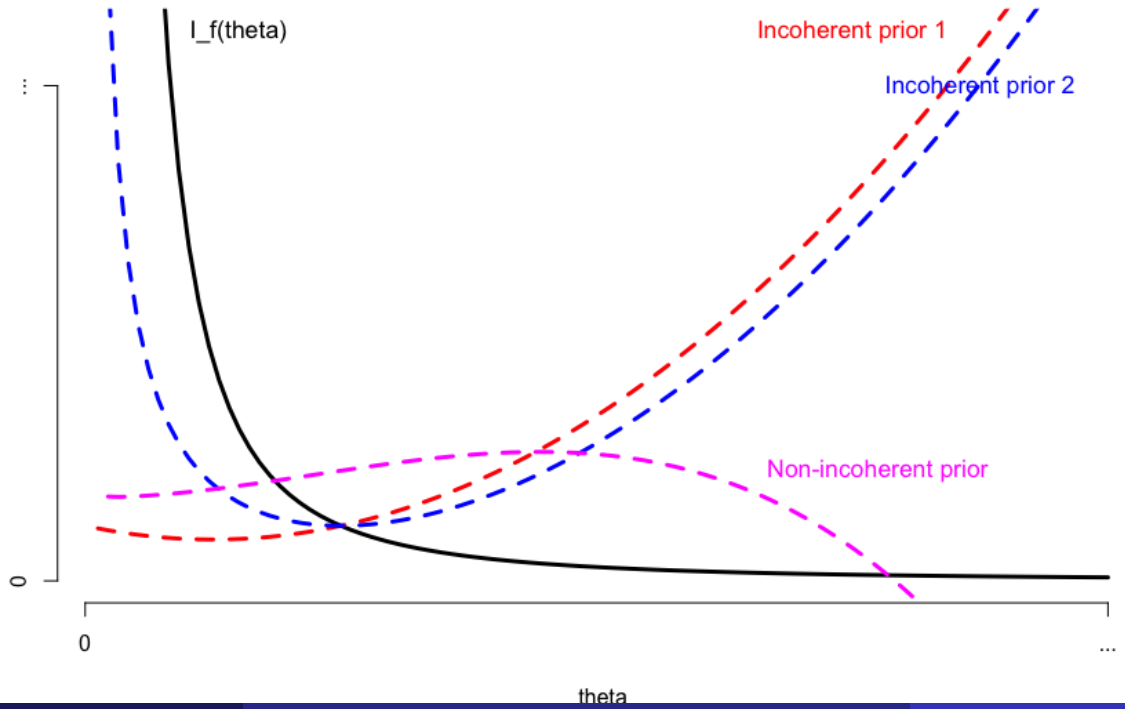
A - Exponential model

Let $f(x|\theta) = \theta \exp(-\theta x)$ and $\Theta = \boldsymbol{R}_*^+$ and $x > 0$

Then $I(\theta) = 1/\theta^2$

We should avoid any prior $\pi(\theta)$ such that $\exists c > 0$ and $c < \infty$ and

$$\frac{\pi(\theta)}{\theta^2} \xrightarrow{\theta \to \infty} c$$



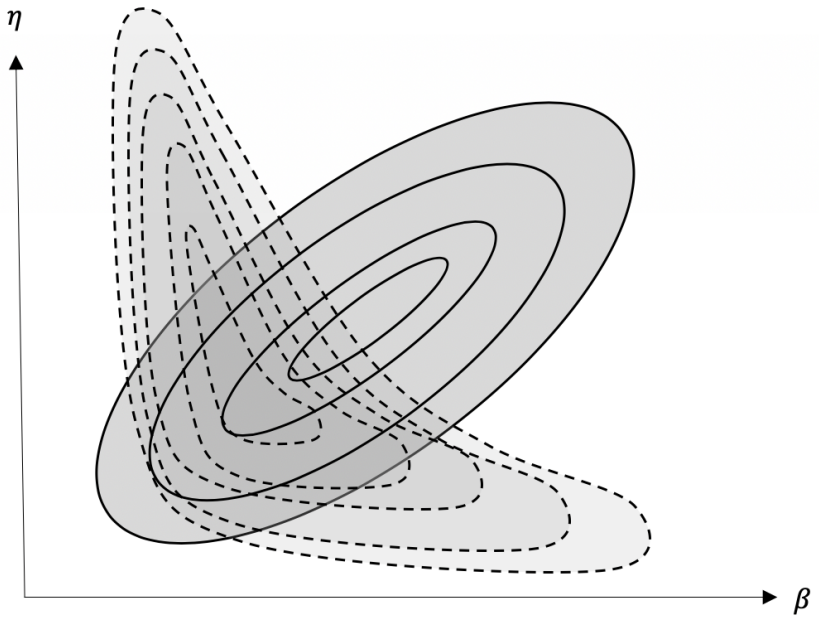
B - Weibull model

$\theta = (\eta, \beta) \in \boldsymbol{R}_*^+ \times \boldsymbol{R}_*^+$ and $x > 0$

$$\mathrm{f}(\mathrm{x}|\theta) = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1} \exp\left(-\left\{\frac{x}{\eta}\right\}^{\beta}\right)$$

$$\det I(\theta) \propto \beta^{-1}\eta^{\beta-1}$$

In survival analysis : high $\beta \Leftrightarrow$ short *Mean Time To Failure*, weak η



Light grey : isodensity curves of a coherent prior
Dark grey : isodensity curves of an incoherent prior

2 - If $\pi(\theta) = \pi^J(\theta|\tilde{\mathbf{x}}_\mathbf{m})$ is untractable, consider a *variational approximation* hyperparameterized by the virtual size m

$$\pi(\theta|m,t) \quad \simeq \quad \pi^J(\theta|\tilde{\mathbf{x}}_\mathbf{m})$$

where t is a summary statistics

3 - Given m, in both previous situations, *calibrate* t using a discrepancy $\mathcal{D}$

$$t^*(m) \quad = \quad \arg\min_t \mathcal{D}\left(f_K, f(.|m,t)\right)$$

where

- $f(x|m,t)$ is the feasible prior predictive measure

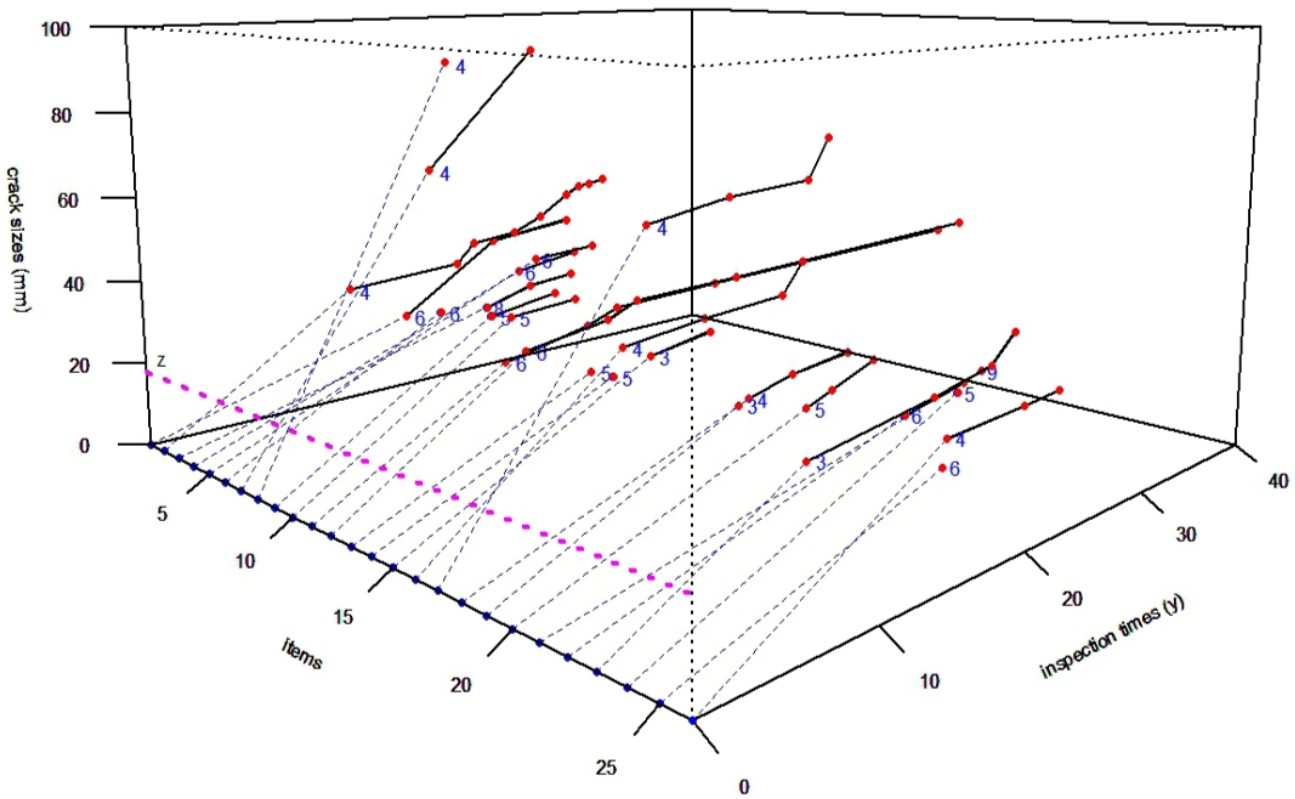
$$f(x|m,t) \quad = \quad \int_\Theta f(x|\theta)\pi(\theta|m,t)\ d\theta$$

- f_K is the "empirical" measure provided by the available marginal quantiles on X

A crack size $Z_{k,t}$ on a component k is monotonically increasing with time t

The increments (assumed independent) $X_{k,i} = Z_{k,t_i} - Z_{k,t_{i-1}}$ are assumed to obey gamma laws

$$f_{\alpha(t-s),\beta}(x) = \frac{1}{\Gamma(\alpha(t-s))} \cdot \frac{x^{\alpha(t-s)-1} e^{-\frac{x}{\beta}}}{\beta^{\alpha(t-s)}} \mathbb{1}_{\{x \geq 0\}}$$



Consider Jeffreys' prior $\pi^J(\alpha,\beta) \propto \frac{1}{\beta}\sqrt{\alpha\Psi_1(\alpha)-1}$

$\pi(\theta)$ beneath is the first-order (Taylor) approximation of the posterior of an imaginary sample of crack increments $\tilde{\mathbf{x}}_\mathbf{m} = (\tilde{x}_1,\ldots,\tilde{x}_m)$ observed at times $\tilde{\mathbf{t}}_\mathbf{m} = (\tilde{t}_1,\ldots,\tilde{t}_m)$:

$$\begin{aligned} \beta|\alpha &\sim \mathcal{IG}\left(\alpha m\tilde{t}_{e,1}, m\tilde{x}_e\right) \\ \alpha &\sim \mathcal{G}\left(m/2, m\tilde{t}_{e,2}\right) \end{aligned}$$

with the meanings

$$\tilde{t}_{e,1} = \frac{1}{m}\sum_{i=1}^{m}\tilde{t}_i \quad \textit{(mean observation time)}$$

$$\tilde{x}_e = \frac{1}{m}\sum_{i=1}^{m}\tilde{x}_i \quad \textit{(mean increase)}$$

$$\tilde{t}_{e,2} = \frac{1}{m}\sum_{i=1}^{m}\tilde{t}_i \log\frac{\sum_{j=1}^{m}\tilde{x}_j/\tilde{x}_i}{\sum_{j=1}^{m}\tilde{t}_j/\tilde{t}_i} \quad \textit{(tuning hyperparameter)}$$

Other similar ideas can come from the rich literature on Edgeworth expansions for posterior densities [30]

The "expert" distribution f_K is supposed to be only known by several marginal quantiles

$$P_{f_k}(X_i < x_{i,\alpha_j}) \quad = \quad \alpha_j$$

for $i = 1, \ldots, d = \dim X$

The feasible prior predictive measure

$$f(x|m,t) \quad = \quad \int_\Theta f(x|\theta)\pi(\theta|m,t)\ d\theta$$

can be continuous

The (usual) KL divergence cannot be used for comparing two distributions that are not absolutely continuous with respect to each other

$\Rightarrow$ use the Wasserstein distance, that allows to measure the "closeness" of two measures defined on arbitrary sets

The "expert" distribution f_K is supposed to be only known by several marginal quantiles

$$P_{f_k}(X_i < x_{i,\alpha_j}) \quad = \quad \alpha_j$$

for $i = 1, \ldots, d = \dim X$

The feasible prior predictive measure

$$f(x|m,t) \quad = \quad \int_\Theta f(x|\theta)\pi(\theta|m,t)\ d\theta$$

can be continuous

The (usual) KL divergence cannot be used for comparing two distributions that are not absolutely continuous with respect to each other

$\Rightarrow$ use the Wasserstein distance, that allows to measure the "closeness" of two measures defined on arbitrary sets

The p-Wasserstein distance between f_K and f on respective supports $\mathcal{X}_K$ and $\mathcal{X}$ is the quantity defined by

$$W_p(f_K, f) = \inf_{f_c \in \Pi_c(f_k, f)} \left\{ \int_{\mathcal{X}_K \times \mathcal{X}} \|x - y\|_p^p df_c(x, y) \right\} \tag{1}$$

where $\|.\|_p$ denotes the ℓ^p norm and $\Pi_c(f_K, f)$ the set of probability couplings, with f_K and f as its marginals, i.e.,

$$\Pi_c(f_K, f) = \left\{ f_c \in \mathcal{P}(\mathcal{X}_K \times \mathcal{X}) \mid \int_{\mathcal{X}} df_c(x, y) = f(dx), \int_{\mathcal{X}_K} df_c(x, y) = f_K(dy) \right\},$$

Theorem ([28] using a result from [1])

If f_k and f share the same dependence structure (copula), then

$$W_p^p(f_K, f) = \sum_{i=1}^{d} W_p^p(f_{K,i}, f_i). \tag{2}$$

Working on the real line (each dimension of X), the choice of the 2−Wasserstein distance (W_2) leads to

$$W_2(f_{K,i}, f_i) = \sqrt{\int_0^1 \left(F^{\rightarrow}_{K,i}(x) - F^{\rightarrow}_i(x)\right)^2 dx}, \quad f_{K,i}, f_i \in \mathcal{P}_2(\mathbb{R})$$

, with $F^{\rightarrow}$ denoting the generalized inverse cdf, which

- metricizes weak convergence on $\mathcal{P}_2(\mathbb{R}) \Leftrightarrow W_2$ is a measure of proximity on a broad set of probability measures
- simplifies solving

$$t^*(m) \quad = \quad \arg\min_t W_2\left(f_K, f(.|m,t)\right)$$

by

- estimating the $F^{\rightarrow}_{K,i}$ using isotonic polynomials between marginal quantiles, with controlled regularity, which requires to solve a convex quadratic program
- using gradient descent

Technical details in our recent preprint [28]

Another simple example [11]

Consider the Fréchet distribution (used for extreme value analysis)

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

with $\theta = (\mu, \nu, \xi) \in \boldsymbol{R} \times \boldsymbol{R}_+^*$ and $x \geq \mu$

Variational approximation of $\pi^J(\theta|\tilde{\mathbf{x}}_m)$

With π^J = BB's reference prior, an approximate posterior prior of imaginary data of size m is

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

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

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

Virtual size m	x_{e_1}	x_{e_2}	Order of prior predictive quartiles (75,100,150)
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%]

Close results obtained with Cooke's criterion (a fully discretized version of KL)

These elements are part of a broad methodological program to build robust and justified formal rules

It is clear that many other approaches can be used to elicit priors (e.g., using generative algorithms)

Nonetheless, having such rules can help to quantify the weight of prior information and shrink subjectivity within specific hyperparameters

For instance :

Q-vague convergence property [9] $\Rightarrow$ constraints on algebraic structure

A sequence of (variational) priors $(\pi_m)_m$ indexed by m should converge to π^J in the $q-$vaguely Radon sense :

$$\exists \ \{a_m\}_m \in \boldsymbol{R}^m \quad s.t. \quad \forall \text{ function } h \text{ with compact support,}$$
$$\lim_{m \to 0^+} \int h d(a_m \pi_m) \quad = \quad \int h d\pi.$$

Others :

1. Prior-data agreement or conflict $\Leftrightarrow$ constraints on calibration [17, 10]
2. Estimating the Effective Sample Size m for complex variational situations [43, 38]
3. Controlling the errors of prior approximations
4. etc.

Key messages

- Producing priors is certainly useful to guide experiments in view of improving the quality of measurements
- Disposing of a clear corpus of rules would be nice, but it's hard
- Published Bayesian workflows seem insufficient today
- The research landscape on the subject seems still very fragmented, despite the many tools and methods available
- Ongoing work on several points of this "program", motivated by increasingly pressing questions about the overall intelligibility of quantification and the management of uncertainties

Thank You !

Major references

[1] Alfonsi, A. and Jourdain, B. (2014). A remark on the optimal transport between two probability measures sharing the same copula. *Statistics & Probability Letters*, 84 :131–134.

[2] Ancelet, S., Bousquet, N., and Parent, E. (2022). Optimal nonlinear bayesian designs for structural reliability estimation under experimental cost constraints. application to a functional weibull model of steel fracture toughness. *submitted*.

[3] ASTM (1997). *E399-90 : Standard Test Method for Plane-Strain Fracture Toughness of Metallic Materials. Annual Book of ASTM Standards*. American Society for Testing and Materials International.

[4] AWS (2008). Best practices : destructive testing for material toughness. *Inspection Trends, American Welding Society*, pages 30–31.

[5] Barbillon, P., Celeux, G., Grimaud, A., Lefebvre, Y., and Rocquigny (de), E. (2011). Non linear methods for inverse statistical problems. 55 :132–142.

[6] Berger, J. (1993). *Statistical Bayesian Theory and Decision Analysis (Second Edition)*. Springer.

[7] Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3) :385 – 402.

[8] Berry, S., Carlin, B., Lee, J., and Muller, P. (2011). *Bayesian Adaptive Methods for Clinical Trials*. FDA Guidelines.

[9] Bioche, C. (2015). *Approximation de lois impropres et applications*. Ph.D. thesis, Université de Clermont-Ferrand.

[10] Bousquet, N. (2008). Diagnostics of prior-data agreement in applied Bayesian analysis. *Journal of Applied Statistics*, 35 :1011–1029.

[11] Bousquet, N. (2021). *Bayesian Extreme Value Theory*. Springer Nature : Heidelberg.

[12] Bousquet, N., Fouladirad, M., Grall, A., and Paroissin, C. (2015). Bayesian gamma processes for optimizing condition-based maintenance under uncertainty. *Applied Stochastic Models in Business and Industry*, 31 :360–379.

[13] Celeux, G., Grimaud, A., Lefebvre, Y., and Rocquigny (de), E. (2010). Identifying intrinsic variability in multivariate systems through linearised inverse methods. 18 :401–415.

Major references (cont.)

[14] Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design : a review. *Statistical Science*, 10(3) :273–304.

[15] Clarke, B. (1996). Implications of reference priors for prior information and for sample size. *Journal of the American Statistical Association*, 91 :173–184.

[16] Clemen, R. and Winkler, R. (2007). Aggregating probability distributions. In *Advances in Decision Analysis*. Cambridge University Press.

[17] Evans, M. and Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, 1 :893–914.

[18] Fortuin, V. (2022). Priors in Bayesian Deep Learning : A Review. *International Statistical Review*.

[19] Fu, S., Celeux, G., Bousquet, N., and Couplet, M. (2015). Bayesian inference for inverse problems occurring in uncertainty analysis. 5 :73–98.

[20] Fu, S., Couplet, M., and Bousquet, N. (2017). An adaptive kriging method for solving nonlinear inverse statistical problems. 28.

[21] Gelfand, A., Mallick, B., and Dey, D. (1995). Modeling expert opinion arising as a partial probabilistic specification. *Journal of the American Statistical Association*, 90(430) :598–604.

[22] Gelman, A., Simpson, D., and Betancourt, M. (2017). The Prior can often only be understood in the context of the Likelihood. *Entropy*, 19(10).

[23] Gelman, A. and Yao, Y. (2020). Holes in Bayesian statistics. *Journal of Physics G : Nuclear and Particle Physics*, 48(1) :014002.

[24] Giovagnoli, A. (2021). The bayesian design of adaptive clinical trials. *Int J Environ Res Public Health*.

[25] Giovagnoli, A. and Verdinelli, I. (2018). Bayesian randomized adaptive designs with a compound utility function. *Proceedings of the CIRM Conference on New Challenges on Designs of Experiments*.

[26] Hasofer, A. (1968). A statistical theory of the brittle fracture of steel. *International Journal of Fracture*, 4 :439–452.

[27] Hassler, E. (2015). Bayesian d-optimal issues and optimal design construction.

Major references (cont.)

[28] Il Idrissi, M., Bousquet, N., Loubes, J.-M., Iooss, B., and Gamboa, F. (2022). Quantile-constrained Wasserstein projections for robust interpretability of numerical and machine learning models. *https://arxiv.org/abs/2209.11539*.

[29] Kadane, J. and Wolfson, J. (1998). Experiences in elicitation. *The Statistician*, 47 :3–19.

[30] Kolassa, J. and Kuffner, T. (2020). On the validity of the formal edgeworth expansion for posterior densities. *Annals of Statistics*, 48(4) :1940–1958.

[31] Komaki, F. and Biswas, A. (2018). Bayesian optimal response-adaptive design for binary responses using stopping rule. *Statistical Methods in Medical Research*, 27(3) :891–904. PMID : 27142983.

[32] Lazor, R., Lovis, A., Nicod, L., and Cornuz, J. (2012). Dépistage du cancer pulmonaire par scanner thoracique. *Rev Med Suisse*, 363 :2206–2211.

[33] Lindley, D. (1983). Reconciliation of probability distributions. *Operations Research*, 31 :866–880.

[34] Mikkola, P., Martin, O., Chandramouli, S., Hartmann, M., Pla, O., Thomas, O., Pesonen, H., Corander, J., Vehtari, A., Kaski, S., Bürkner, P.-C., and Klami, A. (2021). Prior knowledge elicitation : The past, present, and future. arXiv :2112.01380.

[35] Miyata, T. and Tagawa, T. (2002). Mezzo-scopic analysis of fracture toughness in steels. *Materials Research*, 5.

[36] Mouritz, A. (2012). 18 - fracture processes of aerospace materials. In Mouritz, A. P., editor, *Introduction to Aerospace Materials*, pages 428–453. Woodhead Publishing.

[37] Neal, R. (2001). Transferring prior information between models using imaginary data. *Technical Report 0108, Dept. Statistics, Univ. Toronto, 2001*.

[38] Neuenschwander, B., Weber, S., Schmidli, H., and O'Hagan, A. (2019). Predictively consistent Prior effective sample sizes. *to find*.

[39] Overstall, A. and McGree, J. (2021). Bayesian Decision-Theoretic Design of Experiments Under an Alternative Model. *Bayesian Analysis*, pages 1 – 21.

Major references (cont.)

[40] Overstall, A., McGree, J., and Drovandi, C. (2018). An approach for finding fully bayesian optimal designs using normal-based approximations to loss functions. *Statistics and Computing*.

[41] Prangle, D., Harbisher, S., and Gillespie, C. S. (2022). Bayesian Experimental Design Without Posterior Calculations : An Adversarial Approach. *Bayesian Analysis*, pages 1 – 31.

[42] Pronzato, L. (2021). Performance analysis of greedy algorithms for minimising a maximum mean discrepancy. *arXiv :2101.07564*.

[43] Reimherr, M., Meng, X.-L., and Nicolae, D. (2021). Prior sample size extensions for assessing prior impact and prior-likelihood discordance. *to find*.

[44] Robert, C. (2007). *The Bayesian Choice : From Decision-Theoretic Foundations to Computational Implementation (2nd edition)*. Springer.

[45] Sprenger, J. (2018). The objectivity of subjective Bayesianism. *European Journal for Philosophy of Science*, 69 :539–558.

[46] Strnadel, B. and Haušild, P. (2008). Statistical scatter in the fracture toughness and charpy impact energy of pearlitic steel. *Material Science and Engineering*, 486 :208–214.

[47] Wallin, K. (2002). Master curve analysis of the "euro" fracture toughness dataset. *Engineering Fracture Mechanics*, 69 :451–481.

[48] Yang, S., Vong, W., Sojitra, R., Folke, T., and Shafto, P. (2021). Mitigating belief projection in explainable artificial intelligence via bayesian teaching. *Scientific Reports*, 11.