Introduction to probabilistic modelling

Simon Rouchier

LOCIE UMR CNRS 5271
Université Savoie Mont-Blanc
simon.rouchier@univ-smb.fr

21 mai 2026

Contents

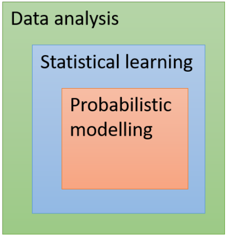
History

Rethinking: A little history. Bayesian statistical inference is much older than the typical tools of introductory statistics, most of which were developed in the early 20th century. Versions of the Bayesian approach were applied to scientific work in the late 1700s and repeatedly in the 19th century. But after World War I, anti-Bayesian statisticians, like Sir Ronald Fisher, succeeded in marginalizing the approach. All Fisher said about Bayesian analysis (then called *inverse probability*) in his influential 1925 handbook was:

> [...] the theory of inverse probability is founded upon an error, and must be wholly rejected.[27]

Bayesian data analysis became increasingly accepted within statistics during the second half of the 20th century, because it proved not to be founded upon an error. All philosophy aside, it worked. Beginning in the 1990s, new computational approaches led to a rapid rise in application of Bayesian methods.[28] Bayesian methods remain computationally expensive, however. And so as data sets have increased in scale—millions of rows is common in genomic analysis, for example—alternatives to or approximations to Bayesian inference remain important, and probably always will.

Glossary



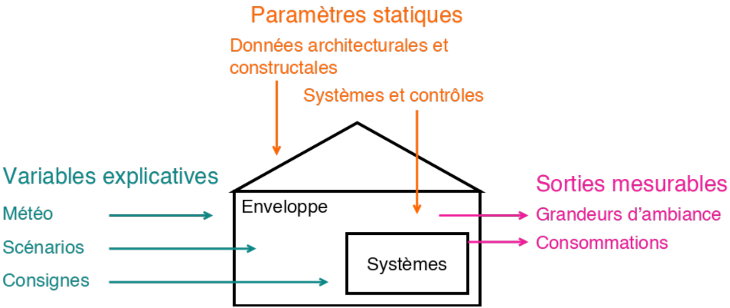
Data analysis: the process of inspecting, cleansing, transforming, and modeling data to extract useful information and support decision-making.

Statistical learning (machine learning): a set of tools for *understanding data*.

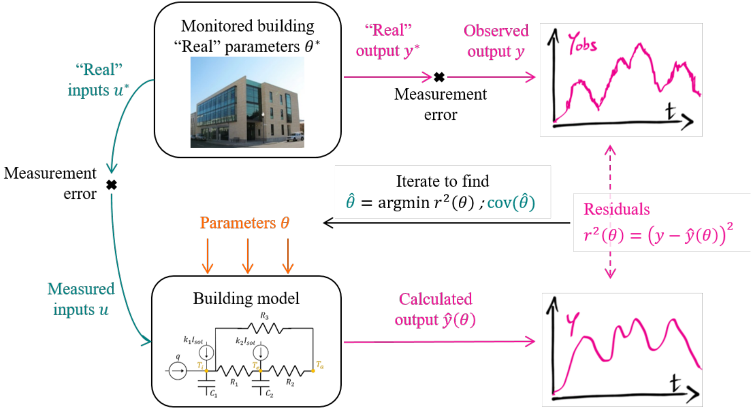
Probabilistic modelling (Bayesian modelling, Bayesian data analysis, Bayesian inference): explicit use of probability for quantifying uncertainty in inferences.

Artificial Intelligence (AI): cool words to use in grant applications.

Glossary


Paramètres statiques
Données architecturales et constructales
Systèmes et contrôles
Variables explicatives
Météo
Scénarios
Consignes
Enveloppe
Systèmes
Sorties mesurables
Grandeurs d'ambiance
Consommations

Why probabilistic modelling

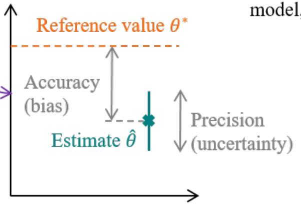


Why probabilistic modelling

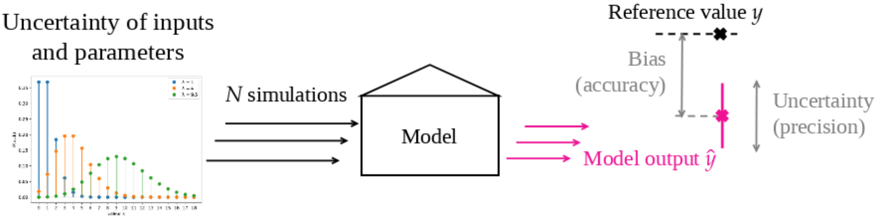
	Systematic errors	Random errors
Experimental	Calibration error Nonrepresentative sampling Sensor intrusiveness	Measurement noise Finite resolution Missing data
Numerical	Excessive discretisation Unaccounted influences Fixed parameter error	Process noise

Systematic errors that are not corrected or explicitly accounted for will affect accuracy

Random errors, or systematic errors accounted for by the model, will affect precision



Why probabilistic modelling



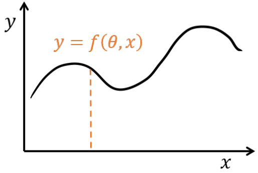
Deterministic vs probabilistic modelling

Deterministic models: the output is entirely determined by the inputs.

Inputs x

Parameters θ

Model $y = f(x, \theta)$



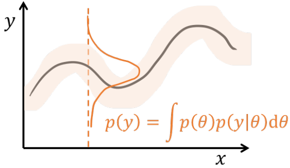
Probabilistic models:

- Unobserved variables θ are stochastic.
- Interdependence between variables is recorded in a probability distribution.

Explanatory variables x

Prior distribution $p(\theta)$

Observational model $p(y|\theta)$



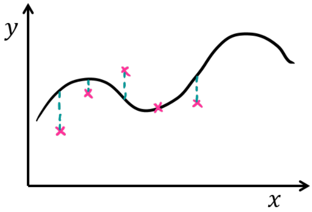
Deterministic vs probabilistic learning

Data (x_i, y_i)

Model $\hat{y} = f(\theta, x)$

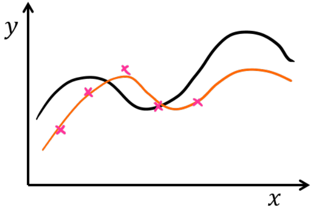
Residuals

$r_i(\theta) = y_i - \hat{y}(\theta, x_i)$



Model fitting: $\hat{\theta} = \operatorname{argmin}\left(\sum r_i^2(\theta)\right)$

Trained model: $\tilde{y} = f(\hat{\theta}, \tilde{x})$



Deterministic vs probabilistic learning

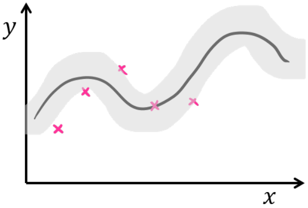
Data (x_i, y_i)

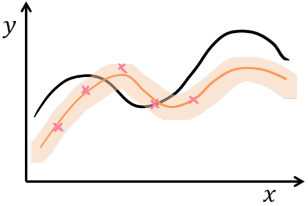
Model $p(\theta, y) = p(\theta)p(y|\theta)$

Bayesian inference

Posterior density: $p(\theta|y) \propto p(\theta)p(y|\theta)$

Posterior predictive distribution: $p(\tilde{y}|y) = \int p(\theta|y)p(\tilde{y}|\theta)\mathrm{d}\theta$





Bayesian data analysis

The **model** is a joint probability distribution for θ and y, written as a product of two densities:

- A prior distribution $p(\theta)$ which encodes eventual assumptions regarding model parameters, independently of the observed data.
- An observational model $p(y|\theta)$, or data distribution, which describes the relationship between the data y and the model parameters θ;

$$p(\theta, y) = p(\theta)p(y|\theta)$$

Prior predictive distribution: distribution of the observable y before the data are considered.

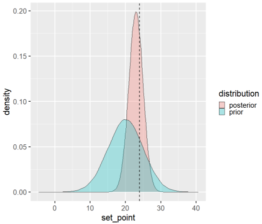
$$p(y) = \int p(y, \theta)\mathrm{d}\theta = \int p(\theta)p(y|\theta)\mathrm{d}\theta$$

Bayesian data analysis

Bayesian statistical conclusions about a parameter θ, or unobserved data $\tilde{y}$, are made in terms of probability statements conditional on the observed value of y.

Bayesian inference: conditioning on the known value of the data y using Bayes' rule yields the **posterior** density.

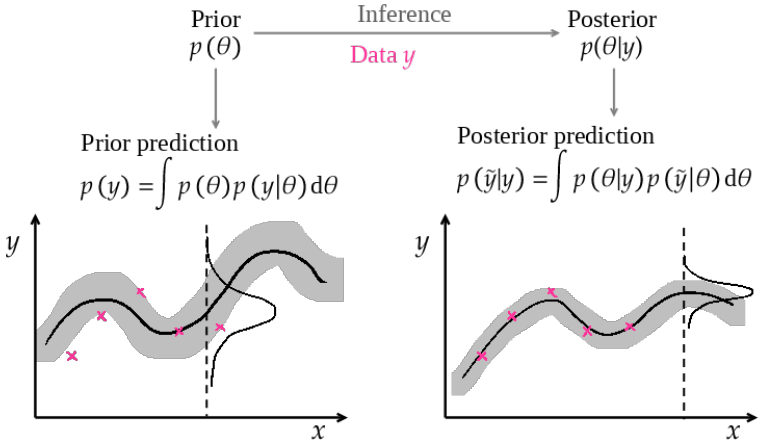
$$p(\theta|y) \propto p(\theta)p(y|\theta)$$



Posterior predictive distribution: prediction of an unknown observable $\tilde{y}$ conditional on the observed y.

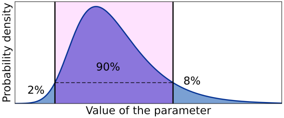
$$p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y)\mathrm{d}\theta = \int p(\tilde{y}|\theta)p(\theta|y)\mathrm{d}\theta$$

Bayesian data analysis

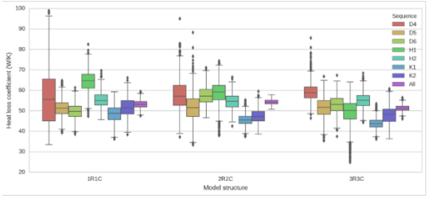


Outcome

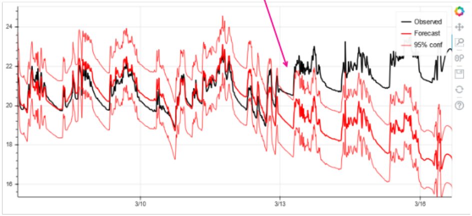
High Density Intervals (HDI): intervalle le plus étroit ayant une certaine probabilité de contenir la grandeur cible.



Physical parameter identification



Fault detection: observations stray away from the predictions of a calibrated model



Contents

The three steps of Bayesian data analysis

1. Model specification
$p(\theta, y) = p(\theta)p(y|\theta)$ → Prior predictive checking
$p(y) = \int p(\theta)p(y|\theta)\, \mathrm{d}\theta$

2. Learning
$p(\theta|y) \propto p(\theta)p(y|\theta)$ → MCMC convergence diagnostics
ESS, $\hat{R}$

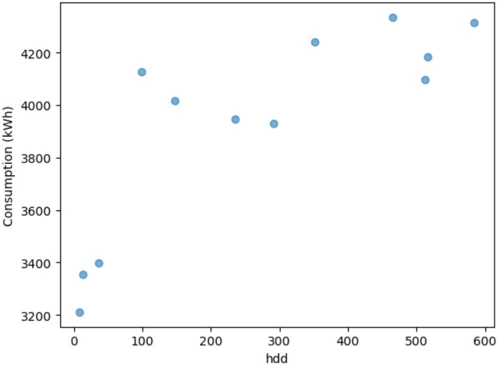
3. Posterior prediction
$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)\, \mathrm{d}\theta$ → Model checking (ACF)
Model evaluation (CRPS)
Model comparison (LOO)

1. Setting up a full probability model - a joint probability distribution for all observable and unobservable quantities in a problem.
2. Conditioning on observed data: calculating and interpreting the appropriate posterior distribution.
3. Evaluating the fit of the model and the implications of the resulting posterior distribution.

Setting up a model

Setting up a full probability model - a joint probability distribution for all observable and unobservable quantities in a problem. The model should be consistent with knowledge about the underlying scientific problem and the data collection process.

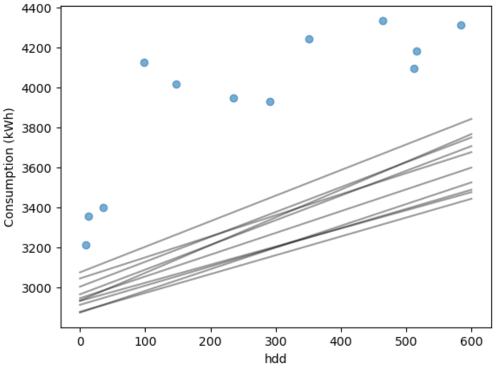
Setting up a model



- Energy consumption e
- heating degree days hdd

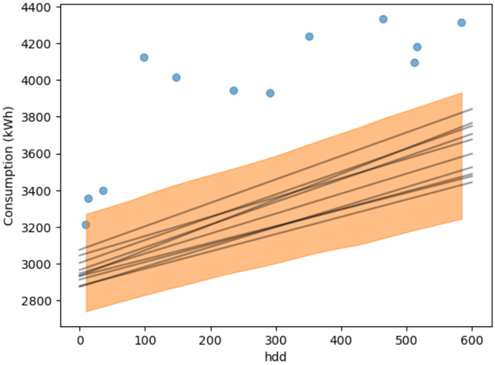
$$e_i = a + b \cdot \mathrm{hdd}_i$$

Setting up a model



$$e_i = a + b \cdot \mathrm{hdd}_i$$
$$a \sim \mathrm{N}(3000, 100)$$
$$b \sim \mathrm{N}(1, 0.2)$$

Setting up a model



$$e_i = a + b \cdot \mathrm{hdd}_i$$
$$p(y_i|a, b, \sigma) = \mathrm{N}(e_i, \sigma)$$
$$p(a) = \mathrm{N}(3000, 100)$$
$$p(b) = \mathrm{N}(1, 0.2)$$
$$p(\sigma) = \mathrm{HalfN}(100)$$

Prior predictive distribution: $p(y) = \int p(y, \theta)\mathrm{d}\theta = \int p(\theta)p(y|\theta)\mathrm{d}\theta$

Conditioning on observed data

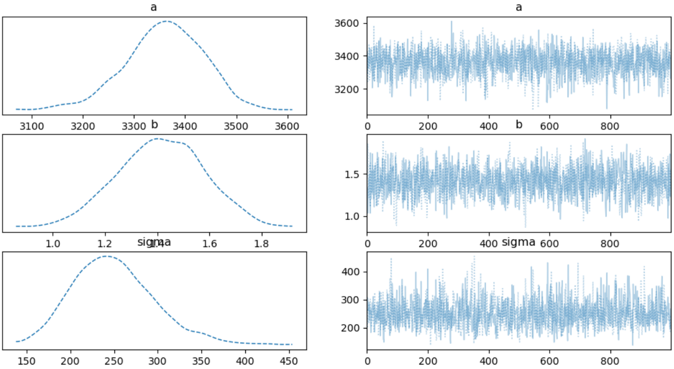
Conditioning on observed data: calculating and interpreting the appropriate posterior distribution - the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data.

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

Markov Chain Monte Carlo algorithms: Metropolis-Hastings, Gibbs, Hamiltonian Monte-Carlo (HMC), No U-Turn Sampling (NUTS), JAX... generate a sequence $\left\{\theta^{(1)}, ..., \theta^{(S)}\right\}$ which approximates the posterior $p(\theta|y)$.

Variational Bayesian methods provide a locally-optimal, exact analytical solution to an approximation of the posterior.

Conditioning on observed data



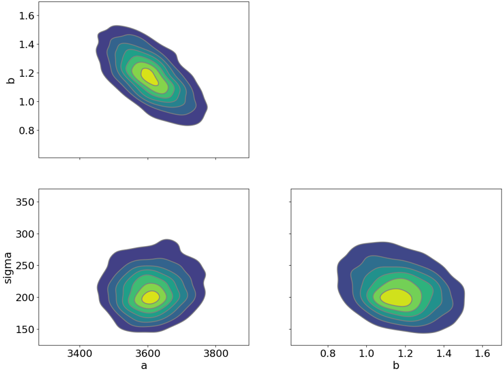
$$p(a) \to p(a|y)$$
$$p(b) \to p(b|y)$$
$$p(\sigma) \to p(\sigma|y)$$

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
a	3359.899	76.656	3209.777	3492.630	2.163	1.530	1297.0	1071.0	1.0
b	1.414	0.167	1.101	1.719	0.004	0.003	1469.0	1269.0	1.0
sigma	250.186	47.700	156.443	334.459	1.395	0.986	1183.0	1077.0	1.0

Conditioning on observed data

$\left\{\theta^{(1)}, ..., \theta^{(S)}\right\}$ approximates a *joint* distribution of all parameters $p(a, b, \sigma | y)$

$$\theta^{(s)} = (a, b, \sigma)^{(s)}$$



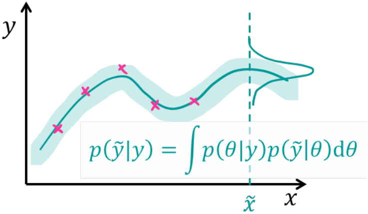
Evaluating the fit of the model

Evaluating the fit of the model and the implications of the resulting posterior distribution: how well does the model fit the data, are the substantive conclusions reasonable, and how sensitive are the results to the modeling assumptions in step 1? In response, one can alter or expand the model and repeat the three steps.

Evaluating the fit of the model

The posterior predictive density for new data $\tilde{y}$ given observed data y is

$$p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y)\mathrm{d}\theta = \int p(\tilde{y}|\theta)p(\theta|y)\mathrm{d}\theta$$



Given draws from the posterior $\theta^{(m)} \sim p(\theta|y)$, draws from the posterior predictive $\tilde{y}^{(m)} \sim p(\tilde{y}|y)$ can be generated by randomly generating from the sampling distribution with the parameter draw plugged in,

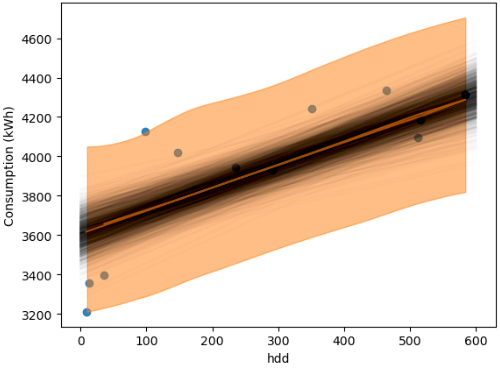
$$\tilde{y}^{(m)} \sim p(y|\theta^{(m)})$$

The posterior distribution or mean of any function f can be approximated as well:

$$\mathrm{E}\left[f(\tilde{y}, \theta)|y\right] \approx \frac{1}{M}\sum_{m=1}^{M} f(\tilde{y}^{(m)}, \theta^{(m)})$$

Evaluating the fit of the model

$$p(\tilde{y}|y) = \int \underbrace{p(\tilde{y}|\theta)}_{\text{sampling uncertainty}} \underbrace{p(\theta|y)}_{\text{estimation uncertainty}} \mathrm{d}\theta$$



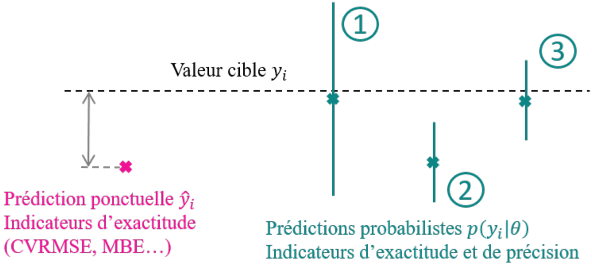
Estimation uncertainty:

$$e = a + b \cdot \mathrm{hdd}$$
$$a \sim p(a|y)$$
$$b \sim p(b|y)$$

Sampling uncertainty:

$$p(\tilde{y}|a, b, \sigma) = \mathrm{N}(a + b \cdot \tilde{\mathrm{hdd}}, \sigma)$$
$$\sigma \sim p(\sigma|y)$$

Scoring rules



Keilman Nico. Évaluer les prévisions probabilistes de population / Evaluating Probabilistic Population Forecasts. In: Economie et Statistique / Economics and Statistics, n°520-521, 2020. pp. 51-68;

Scoring rules

Continuous Ranked Probability Score (CRPS)

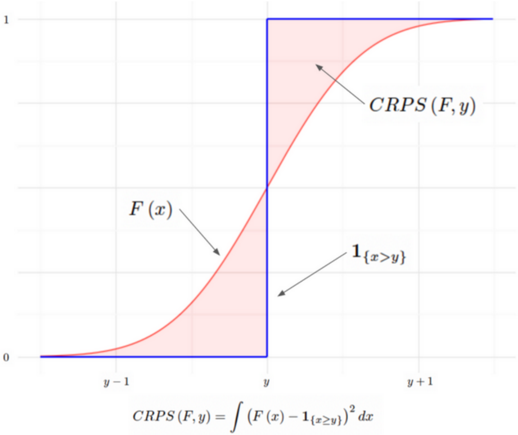
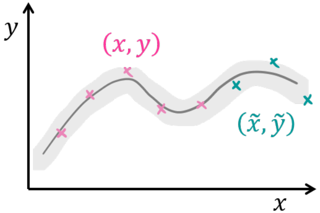


Image : Itamar Faran, CRPS - A Scoring Function for Bayesian Machine Learning Models

Bayesian model comparison and selection

Given training data (x, y) and test data $(\tilde{x}, \tilde{y})$



The log pointwise predictive density is a model comparison and selection metric.

$$\log p(\tilde{y}|\tilde{x}, x, y) = -\log M + \mathrm{logsumexp}_{m=1}^{M} \log p(\tilde{y}|\tilde{x}, \theta^{(m)})$$

Built-in methods to estimate the expected log pointwise predictive density (elpd) for a new dataset:

- Leave-One-Out Cross Validation (LOO-CV)
- Widely-applicable Information Criterion (WAIC)

Contents

Change-point model for M&V

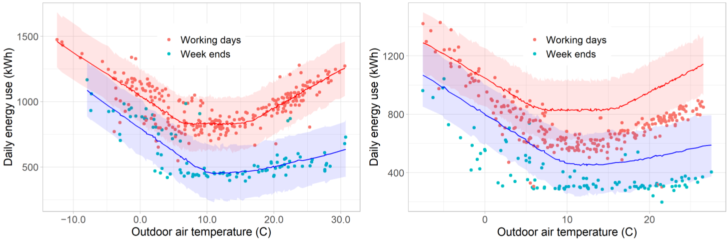
Energy use y as function of ambient temperature T_a

$$p(y|\theta, T_a) = \mathrm{Normal}\left[E_0 + H_1(T_1 - T_a)^+ + H_2(T_a - T_2)^+, \sigma\right]$$

The model is trained on baseline data

```
model {
  // Prior distributions
  E ~ normal(800, 150);
  H ~ normal([40, 40], [15, 15]);
  T ~ normal([8, 18], [5, 5]);
  // Observational model
  for (n in 1:N) {
    y[n] ~ normal(E + H[1]*fmax(T[1]-t[n],0) + H[2]*fmax(t[n]-T[2],0),
    sigma);
  }
```

Simulating draws from the parameter posterior generates posterior distributions for predictions

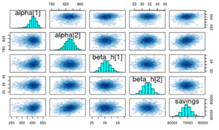


Change-point model for M&V

Simulation draws	Parameters θ $(\theta_1, \ldots, \theta_p)$		Posterior predictions $\tilde{y}_{repo} = (\tilde{y}_1, \ldots, \tilde{y}_n)$		Savings Δe
1	$(\theta_1, \ldots, \theta_p)^{(1)}$	Each draw produces a prediction.	$(\tilde{y}_1, \ldots, \tilde{y}_n)^{(1)}$	Each prediction estimates a value of the savings.	$\Delta e^{(1)}$
.	.		.		.
.	.	$\longrightarrow$	.	$\longrightarrow$	.
.	.		.		.
S	$(\theta_1, \ldots, \theta_p)^{(S)}$	$\tilde{y}^{(s)} \sim p(y \mid \theta = \theta^{(s)})$	$(\tilde{y}_1, \ldots, \tilde{y}_n)^{(S)}$	$\Delta e^{(s)} = \sum_{i=i}^{n} \left(\tilde{y}_{repo,i}{}^{(s)} - y_{repo,i}\right)$	$\Delta e^{(S)}$

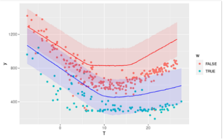
The draws approximate the posterior of each parameter.

$$p(\theta_j \mid y_{base}) \approx \left\{\theta_j^{(1)}, \ldots, \theta_j^{(S)}\right\}$$



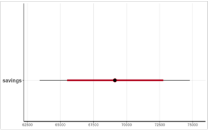
Each data point of the reporting period has prediction intervals.

$$p(\tilde{y}_{repo} \mid y_{base}) \approx \{\tilde{y}^{(1)}, \ldots, \tilde{y}^{(S)}\}$$



Savings are described as a probability density

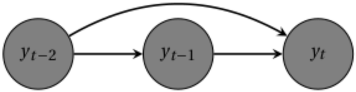
$$p(\Delta e \mid y_{base}) \approx \{\Delta e^{(1)}, \ldots, \Delta e^{(S)}\}$$



Time series: ARMAX models

Autoregressive (AR) model

$$y_t \sim N(\alpha + \beta_1 y_{t-1} + \beta_2 y_{t_2} + ..., \sigma^2)$$

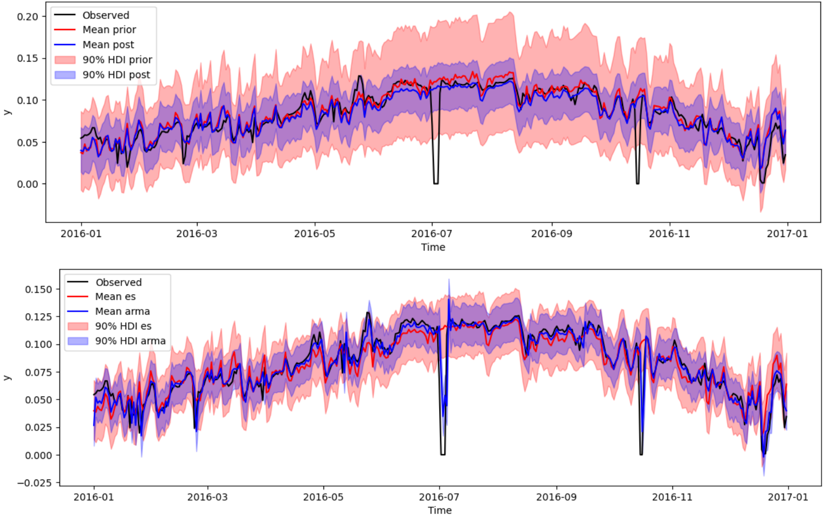


AutoRegressive Moving Average with eXogenous inputs

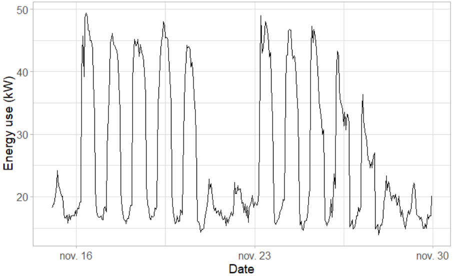
$$E(y_t|\theta, X) = \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{j=0}^{q} \gamma_j \varepsilon_{t-j} + \sum_{k=1}^{K} \left[\sum_{i=0}^{p} \theta_{k,i} x_{t-i,k} \right]$$

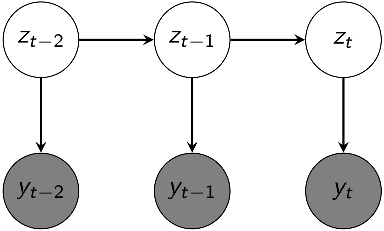
$$\varepsilon_t \sim N(0, \sigma^2)$$

Energy signature + ARMA models



Hidden Markov models





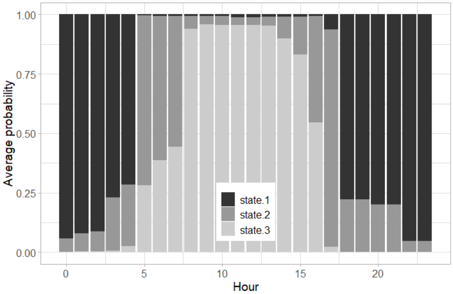
- The occupancy z_t state at each time t is unknown, and described by a hidden Markov chain with a transition probability matrix:

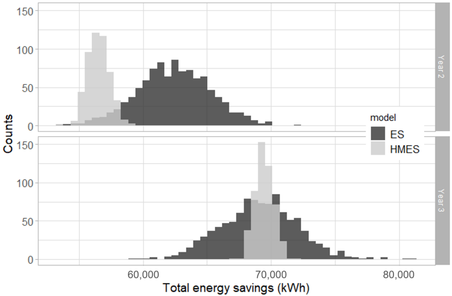
$$a_{ij}(h, d) = p\left(z_{h,d} = j | z_{h-1,d} = i\right)$$

- The energy use y_t follows a different model for each possible occupancy state z_t

$$b_i(y_t) = p(y_t|\theta, T_a, z_t = i)$$

Hidden Markov Models



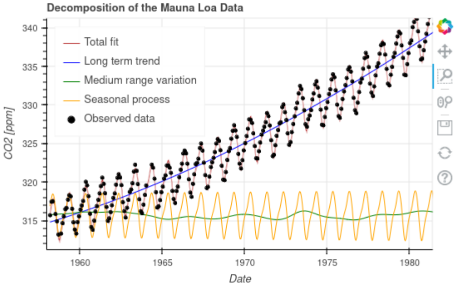


Rouchier S. Bayesian Workflow and Hidden Markov Energy-Signature Model for Measurement and Verification. Energies 2022, 15(10), 3534

Gaussian Process models

Prior distribution over any regression function

$$y \sim \mathrm{multivariate\,normal}\left(m(x), K(x|\theta)\right)$$



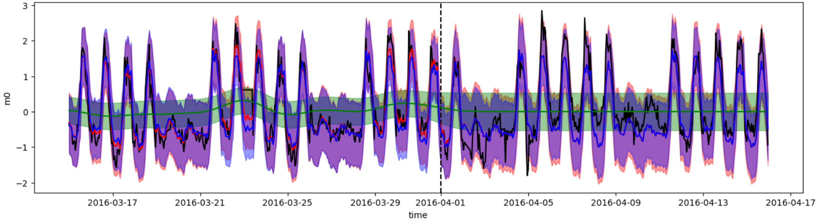
Gaussian processes are additive and can be used as components in a larger model

$$y_t(t) = f_1(t) + f_2(t) + \varepsilon_t$$

$$f_1(t) \sim \mathrm{GP}(0, k_1), \quad k_1(t, t') = \sigma_1^2 \mathrm{exp}\left(-\frac{|t - t'|^2}{2 l_1^2}\right)$$

$$f_2(t) \sim \mathrm{GP}(0, k_2), \quad k_2(t, t') = \sigma_2^2 \mathrm{exp}\left(-\frac{2 \sin^2(\pi(t - t')/7)}{l_2^2}\right)$$

Gaussian Process models

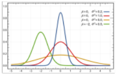


Avant les données : génération d'une base de donnée de simulation représentative

p variables explicatives sur n points de mesure

x_1	...	x_p
x_1^1		x_p^1
.		.
.		.
x_1^n		x_p^n

Q paramètres avec distributions a priori



1) LHS à N échantillons

θ_1	...	θ_Q
θ_1^1		θ_Q^1
.		.
.		.
θ_1^N		θ_Q^N

2) N simulations du modèle numérique

Sorties $\eta \in \mathbb{R}^{(n\times N)}$

3) Réduction de la dimension

- Analyse de sensibilité : sélection de q paramètres parmi Q
- Sélection d'un ensemble représentatif de x : m points parmi $n \times N$

-> Base de données simulée $D^S = [\eta^S\ x^S\ \theta^S] \in \mathbb{R}^{m\times(p+q+1)}$

η	x_1	...	x_p	θ_1	...	θ_q
η^1	x_1^1		x_p^1	θ_1^1		θ_q^1
.	.		.	.		.
.	.		.	.		.
η^m	x_1^m		x_p^m	θ_1^m		θ_q^m

Après les données : combinaison des bases de données mesurée et simulée

Base de données simulée
$D^S = [\eta^S\ x^S\ \theta^S] \in \mathbb{R}^{m\times(p+q+1)}$

η	x_1	...	x_p	θ_1	...	θ_q
η^1	x_1^1		x_p^1	θ_1^1		θ_q^1
.	.		.	.		.
.	.		.	.		.
η^m	x_1^m		x_p^m	θ_1^m		θ_q^m

Base de données mesurées
$D^F = [\eta^S\ x^F\ ] \in \mathbb{R}^{n\times(p+1)}$

y	x_1	...	x_p
y^1	x_1^1		x_p^1
.	.		.
.	.		.
y^n	x_1^n		x_p^n

- Combinaison des bases de données avec la méthode Kennedy et O'Hagan

$$y = \eta(x,\theta) + \delta(x) + \epsilon$$

- MCMC pour estimer les distributions a posteriori $p(\theta|y)$

-> modèle probabiliste calibré utilisable pour des ajustements périodiques ou non

Contents

Material

Books

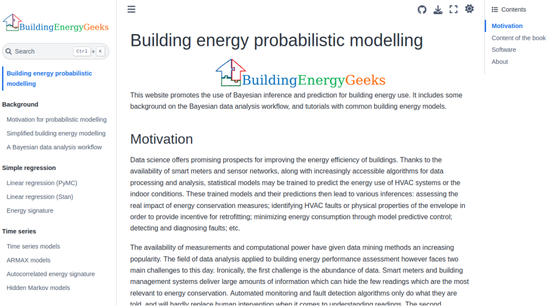
- Mc Elreath, Statistical Rethinking
- Gelman et al. Bayesian Data Analysis
- Betancourt, `https://betanalpha.github.io/`

Probabilistic programming libraries

PyMC	`https://www.pymc.io/`	Python
Stan	`https://mc-stan.org/`	R, Python, Julia...
Pyro	`https://pyro.ai/`	Python

More content

- This tutorial: `https://github.com/srouchier/simurex2026`
- *Building energy statistical modelling* book: `https://BuildingEnergyGeeks.com/`
- Bayesian M&V: `https://srouchier.github.io/bayesmv/`



Introduction to probabilistic modelling

Simon Rouchier

LOCIE UMR CNRS 5271
Université Savoie Mont-Blanc
simon.rouchier@univ-smb.fr

21 mai 2026