Introduction to probabilistic modelling

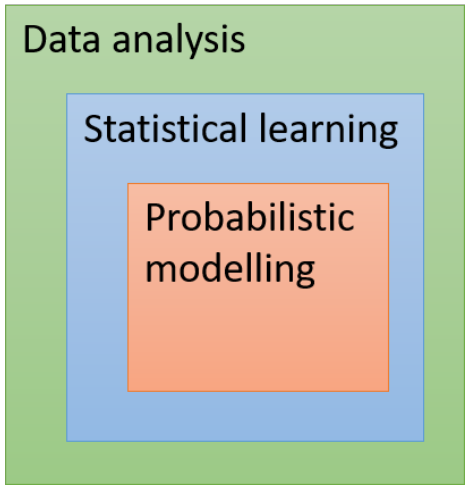
Simon Rouchier

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

16 mai 2024

Contents

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.

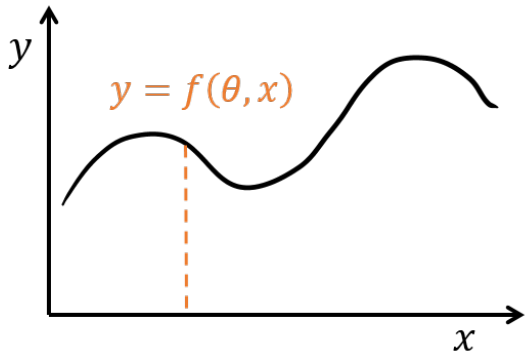
Deterministic vs probabilistic modelling

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

Inputs x

Parameters θ

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



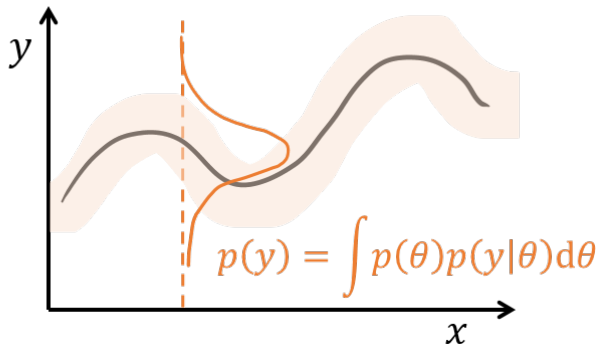
Probabilistic models:

- 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)

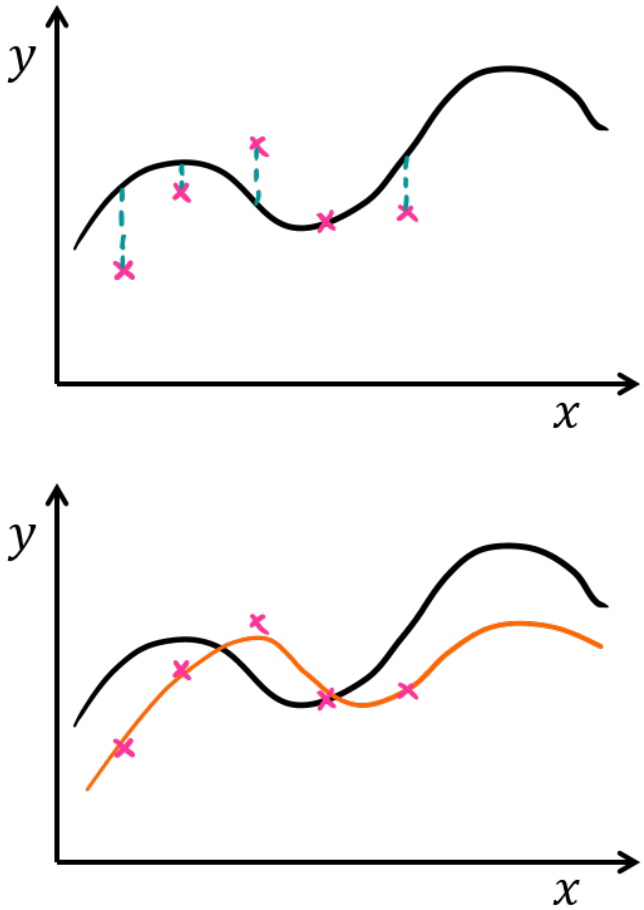
Model $\hat{y} = f(\theta, x)$

Residuals

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

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

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



Deterministic vs probabilistic learning

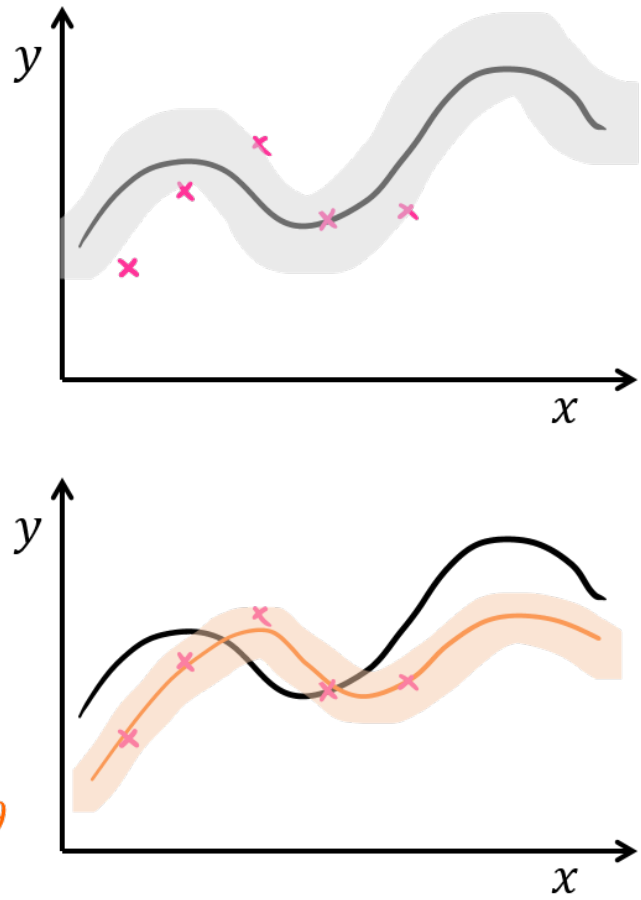
Data (x_i, y_i)

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.

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

The three steps of Bayesian data analysis

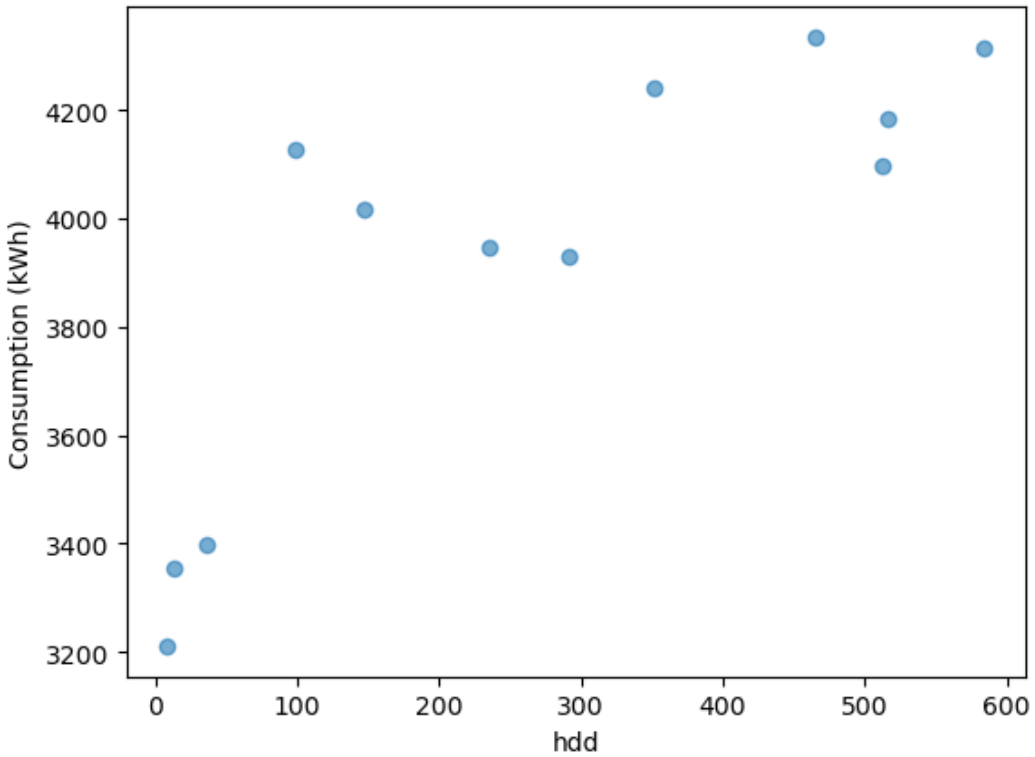
Gelman et al. Bayesian Data Analysis

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.

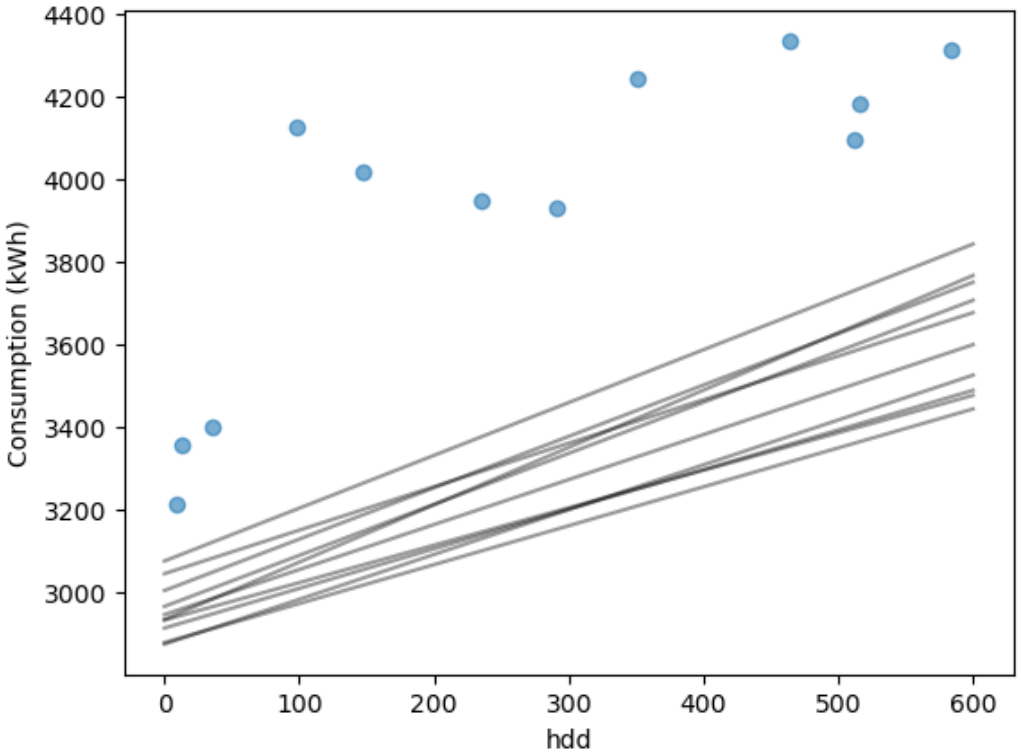
Setting up a model



- Energy consumption e
- heating degree days hdd

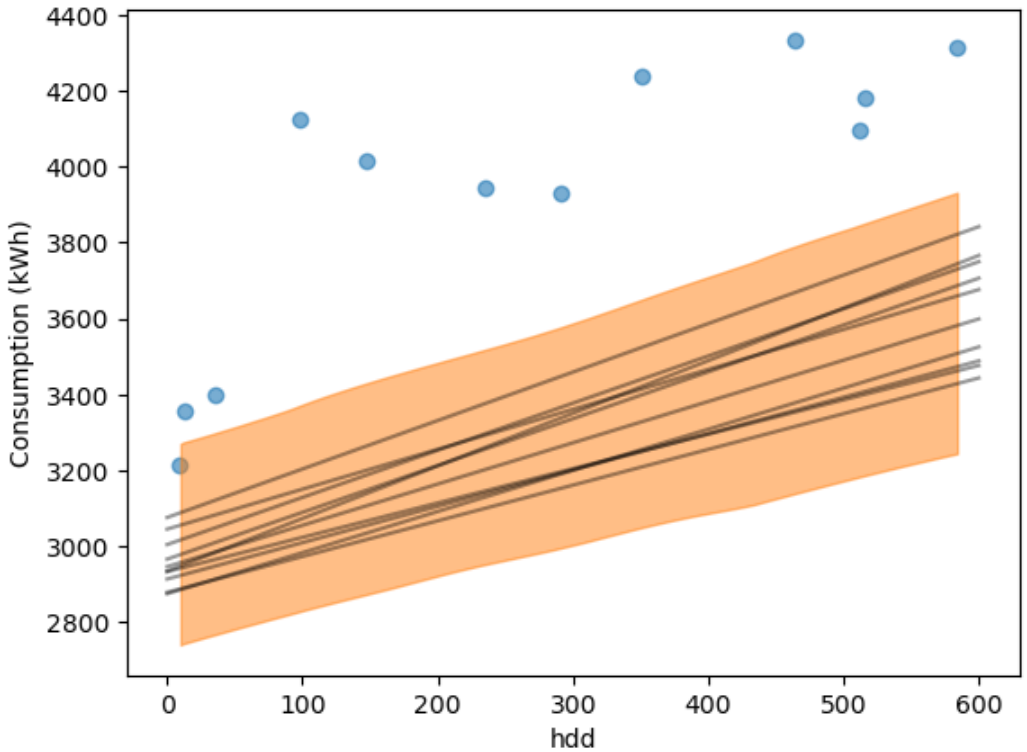
$$e_i = a + b \cdot \mathrm{hdd}_i$$

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



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

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.

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

Conditioning on observed data

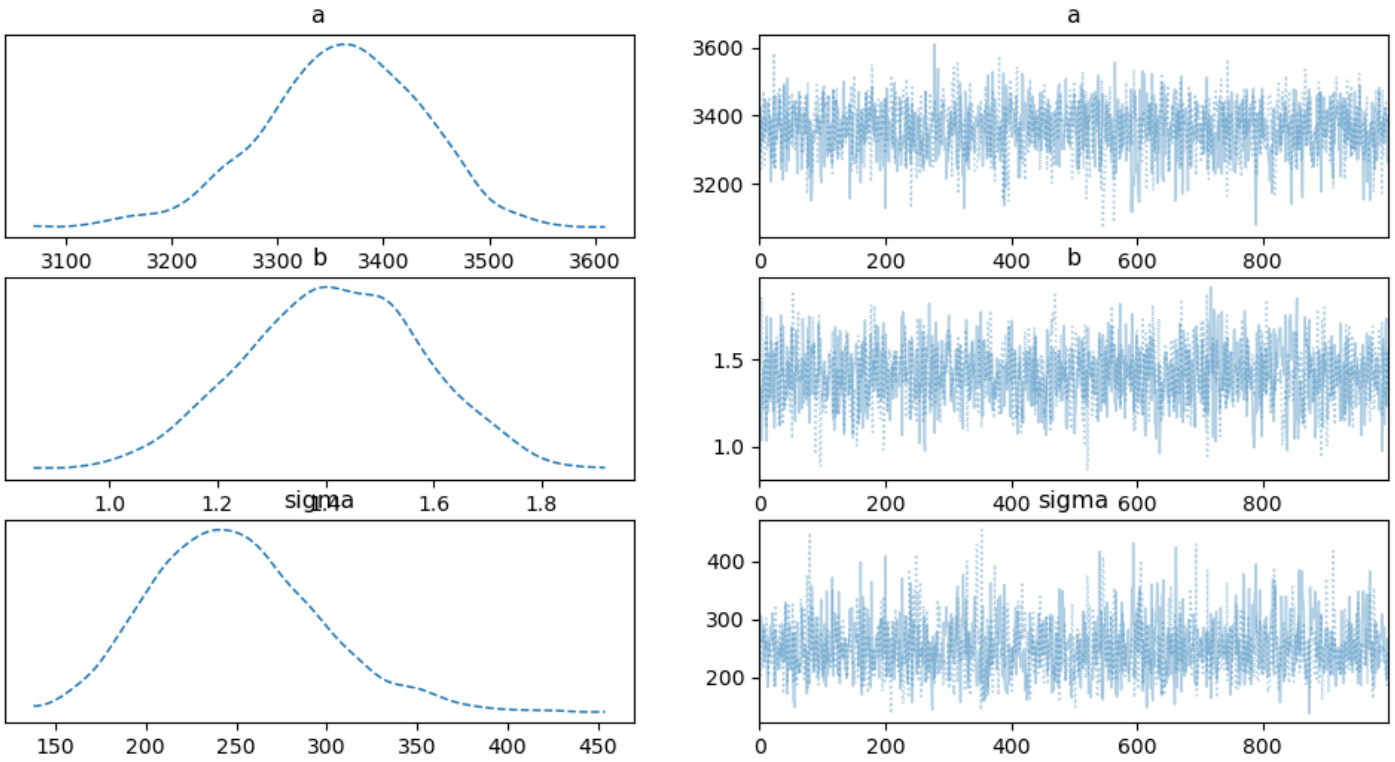
Markov Chain Monte Carlo algorithms:

- Metropolis-Hastings,
- Gibbs,
- Hamiltonian Monte-Carlo (HMC),
- No U-Turn Sampling (NUTS)...

generate a sequence $\left\{\theta^{(1)}, ..., \theta^{(S)}\right\}$ which approximates the posterior distribution $p(\theta|y)$.

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

Conditioning on observed data



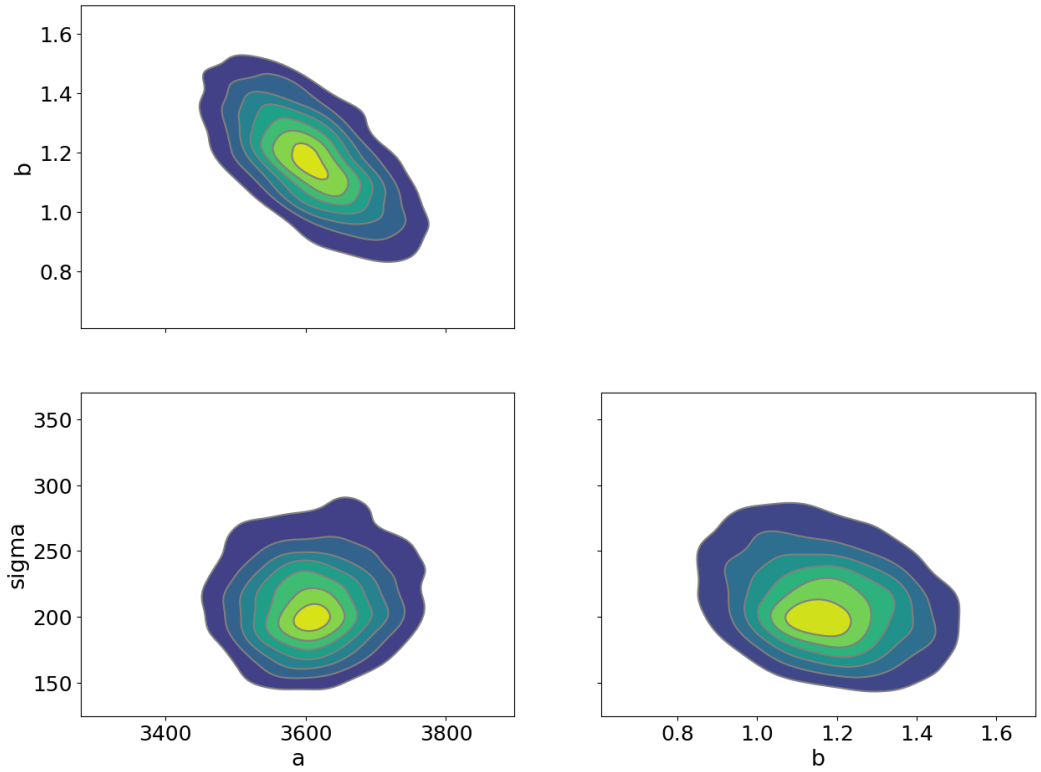
$$p(a) \rightarrow p(a|y)$$
$$p(b) \rightarrow p(b|y)$$
$$p(\sigma) \rightarrow p(\sigma|y)$$

	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

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

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



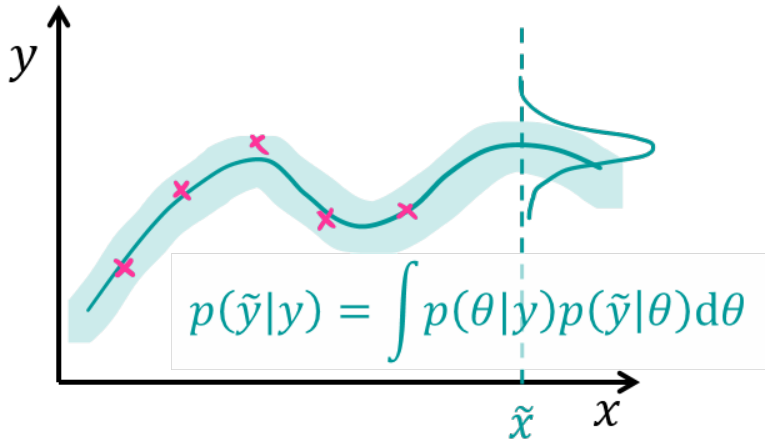
Evaluating the fit of the model

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,

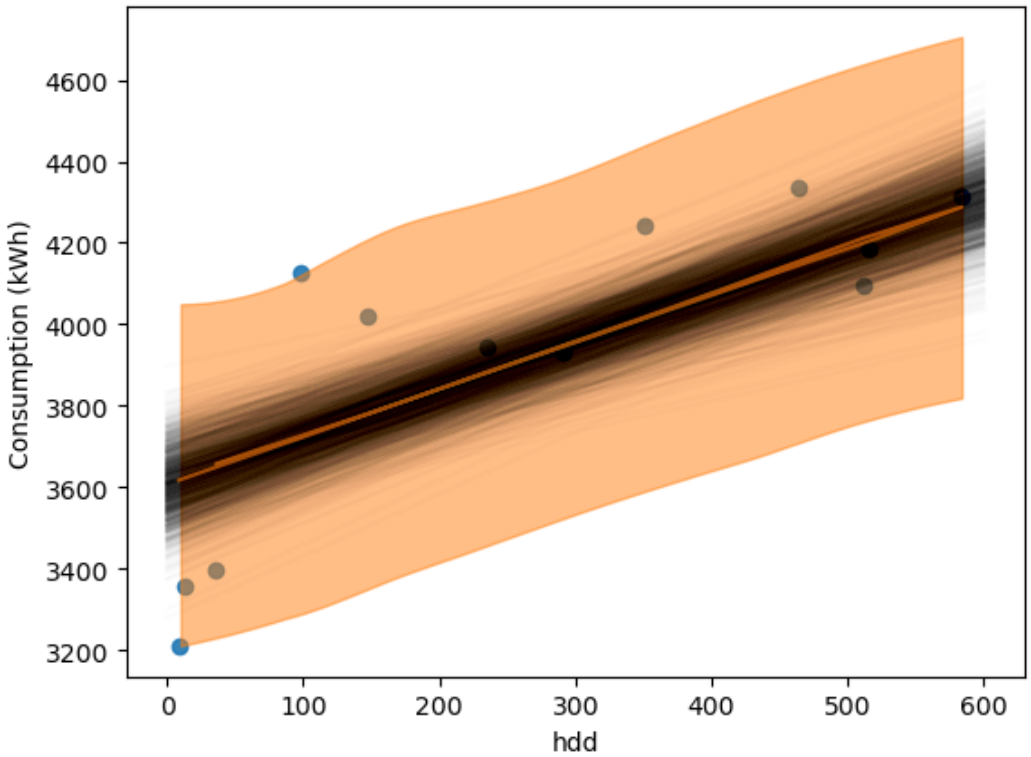
$$\tilde{y}^{(m)} \sim p(y|\theta^{(m)})$$

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

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



Estimation uncertainty:

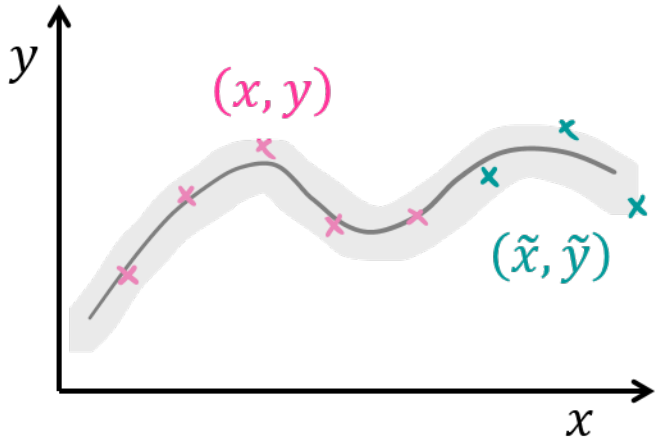
$$e = a + b \cdot \mathrm{hdd}$$
$$a \sim p(a|y)$$
$$b \sim p(b|y)$$

Sampling uncertainty:

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

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.

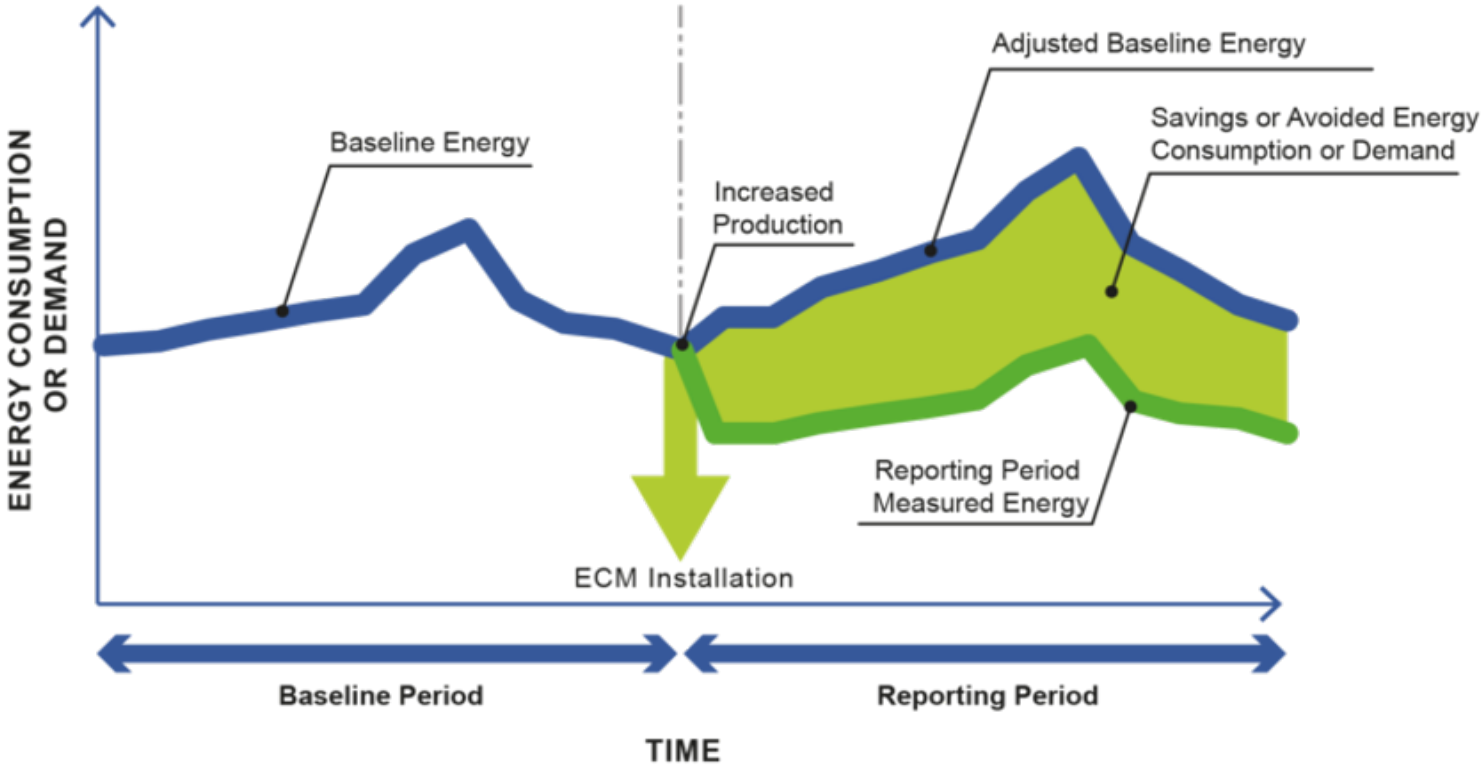
$$\log p(\tilde{y}|\tilde{x}, x, y) = -\log M + \text{logsumexp}_{m=1}^{M} \log p(\tilde{y}|\tilde{x}, \theta^{(m)})$$

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

Measurement and verification



M&V with a change-point model

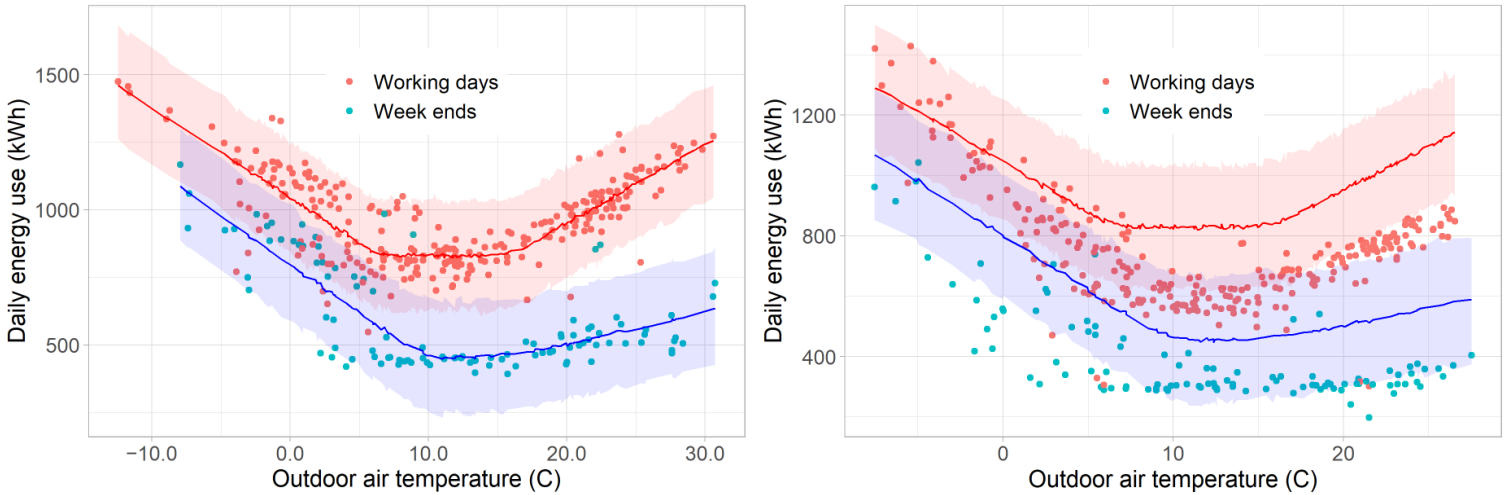
Energy use y as function of ambient temperature T_a

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

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

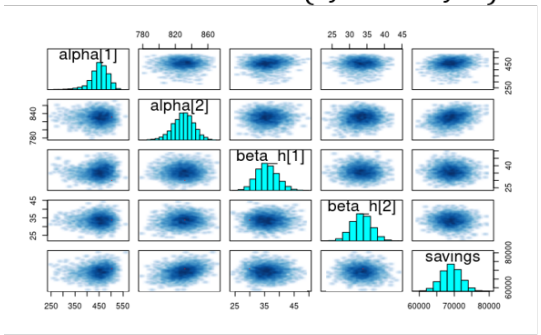


M&V with a change-point model

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

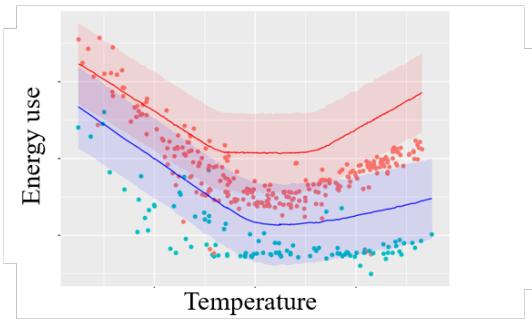
The draws approximate the posterior of each parameter.

$p(\theta_j | y_{base}) \approx \{\theta_j^{(1)}, \dots, \theta_j^{(S)}\}$



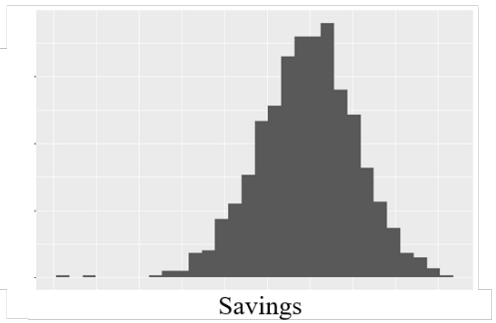
Each data point of the reporting period has prediction intervals.

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



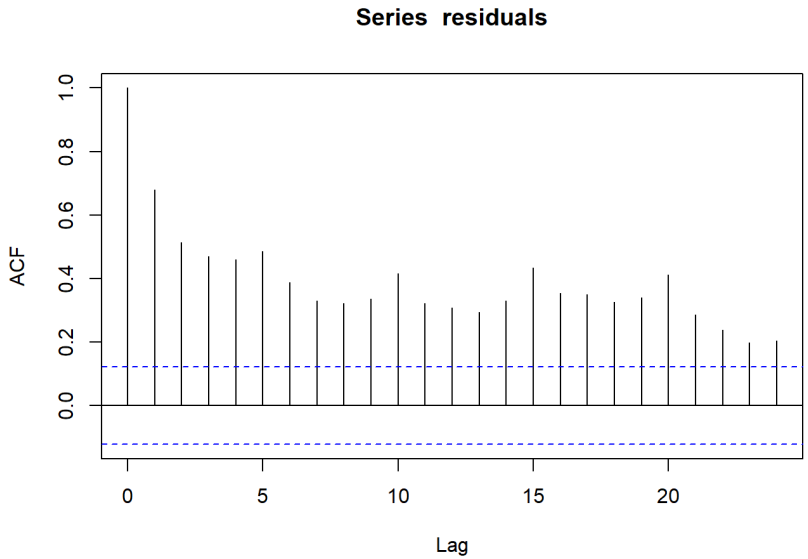
Savings are described as a probability density

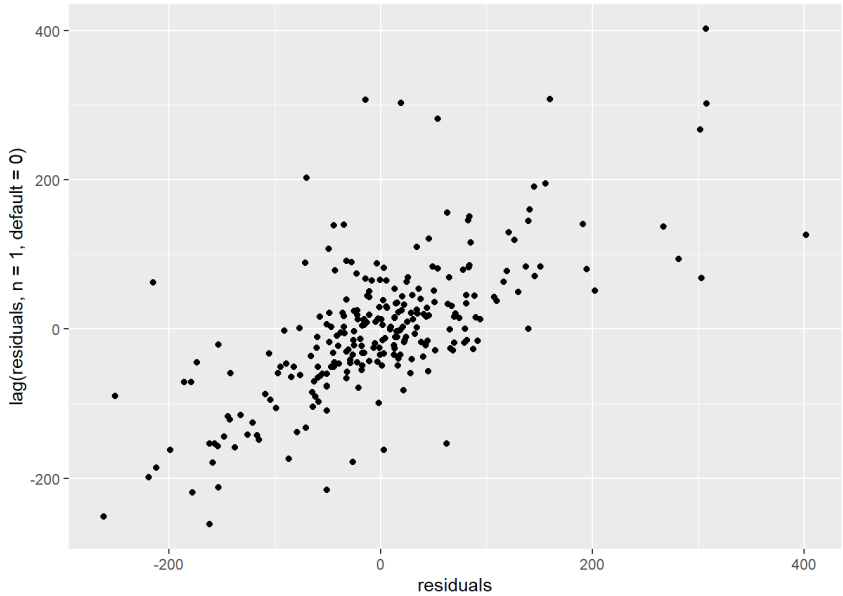
$p(\Delta e | y_{base}) \approx \{\Delta e^{(1)}, \dots, \Delta e^{(S)}\}$



Autocorrelation

Problem: correlated prediction residuals denote a biased model.





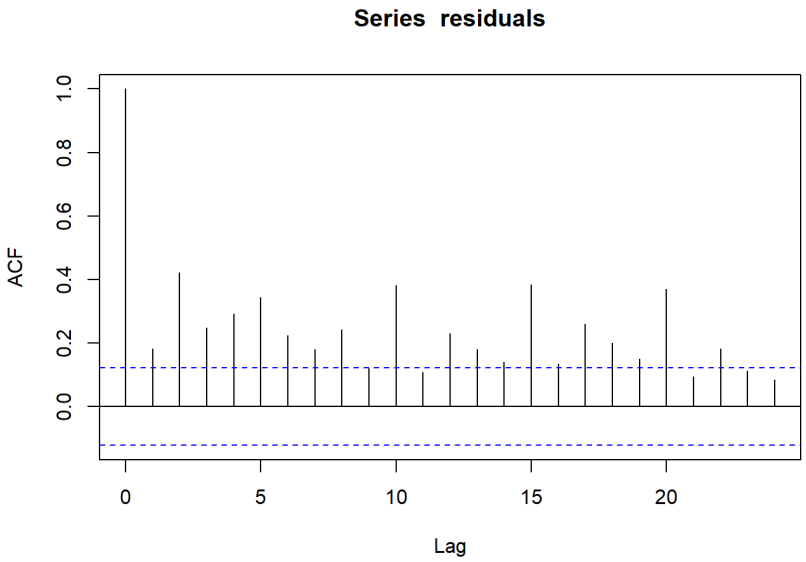
Autocorrelation

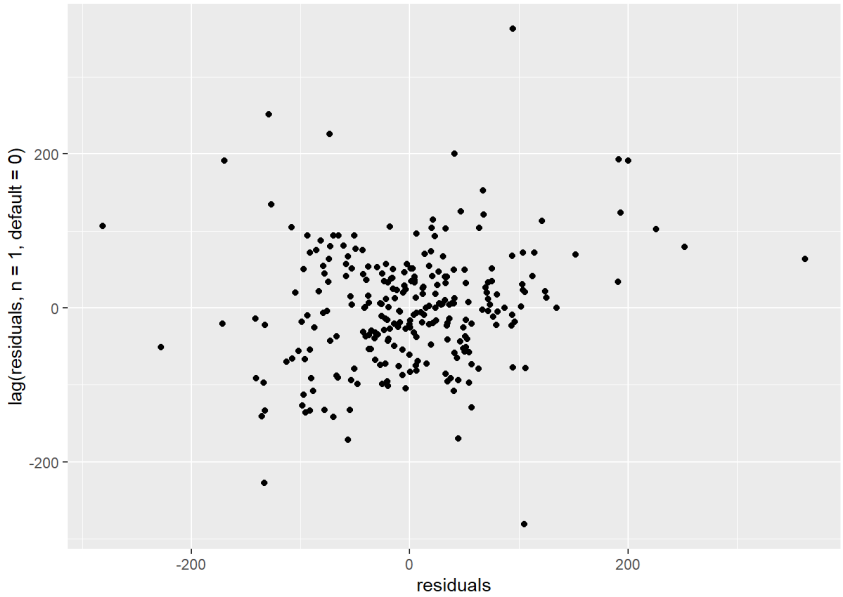
Enhancing the original model f with a Moving Average (MA) model

$$y_n = f(x_n) + \theta_1 \epsilon_{n-1} + \epsilon_n$$
$$\epsilon_n \sim \text{Normal}(0, \sigma)$$

```
model {
  ...
  for (n in 2:N_pre) {
    y_pre[n] ~ normal(f_pre[n] + theta*epsilon[n-1], sigma);
  }
}
```

Autocorrelation



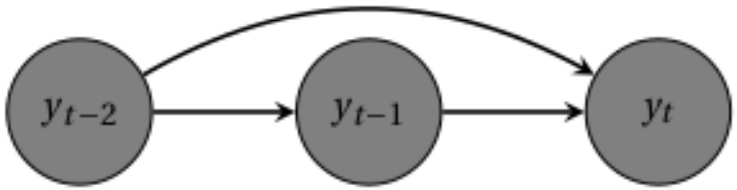


- Decreased model bias
- Increased prediction uncertainty and savings uncertainty

Time series: ARMAX models

Autoregressive (AR) model

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

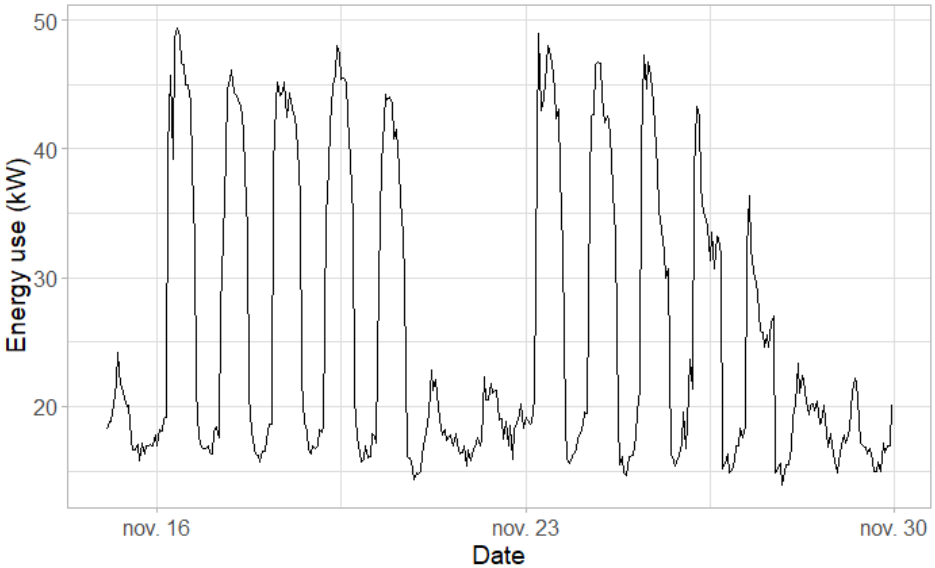


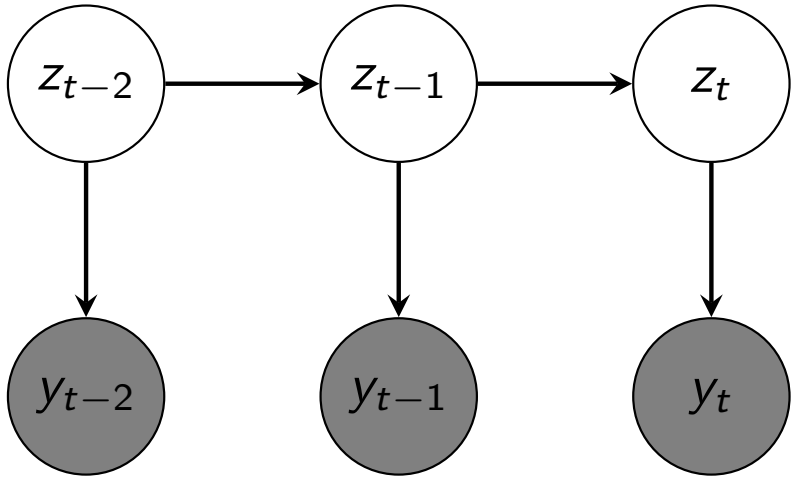
AutoRegressive Moving Average with eXogenous inputs

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

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

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

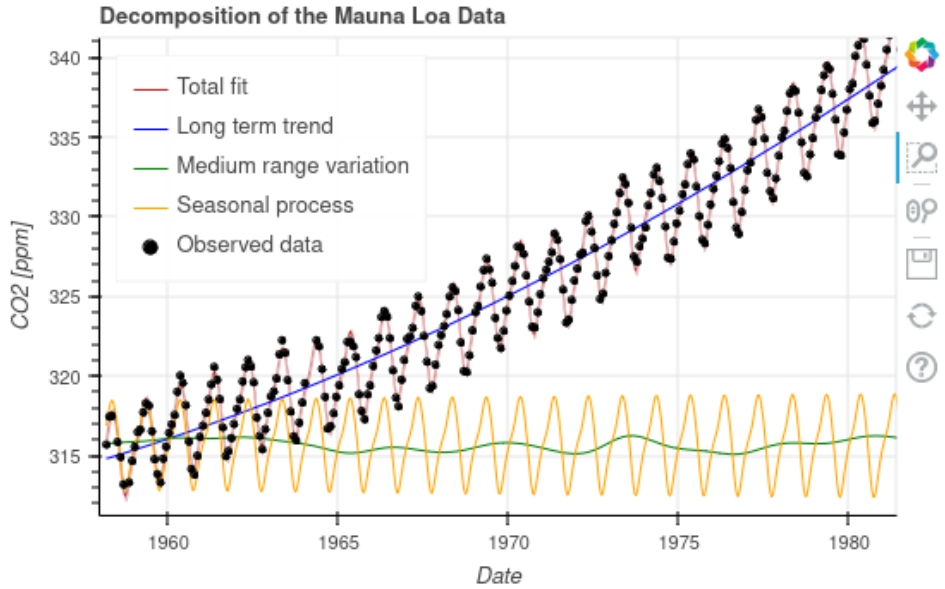
$$b_i(y_t) = p(y_t | \theta, T_a, z_t = i)$$

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

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



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

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

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

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

Other models

- Regression models
- Time-series models: ARMAX, RC, HMM...
- Mixture models
- Hierarchical models
- Gaussian processes
- Missing data, measurement error

Material

Books

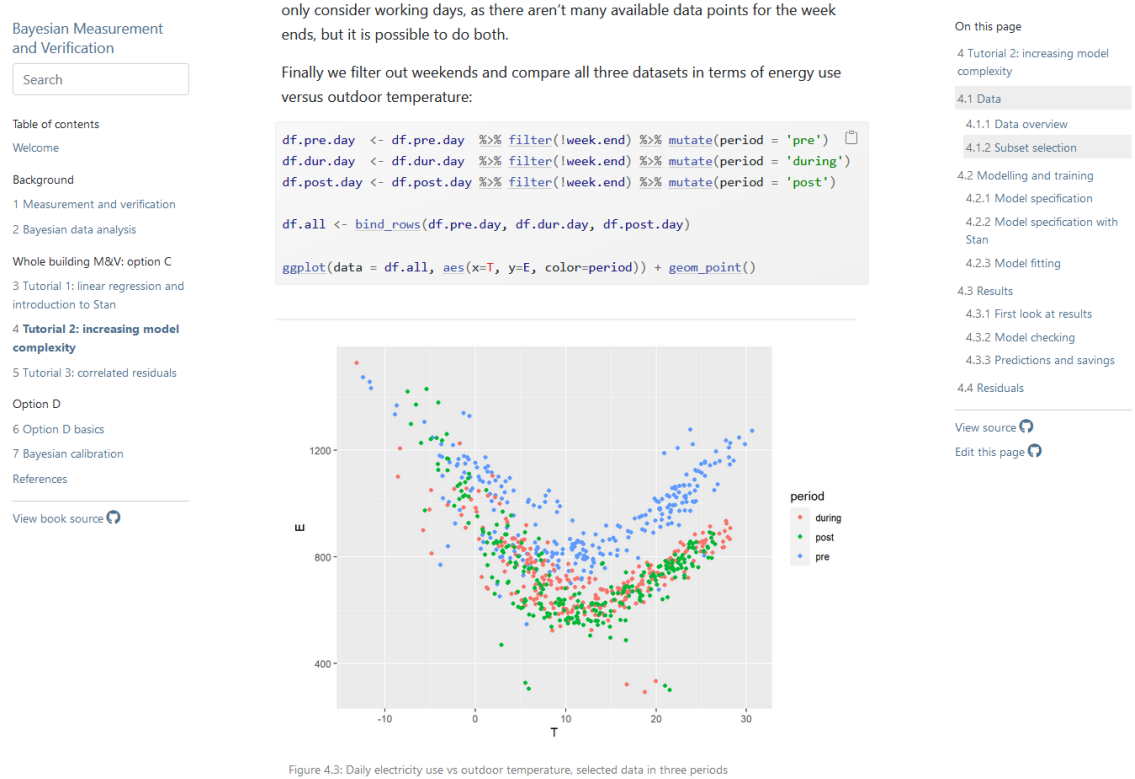
- Gelman et al. Bayesian Data Analysis
- Mc Elreath, Statistical Rethinking
- Betancourt, `https://betanalpha.github.io/`

Probabilistic programming libraries

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

More content

- This tutorial: `https://github.com/srouchier/simurex2024`
- *Building energy statistical modelling* book: `https://BuildingEnergyGeeks.org/`
- 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

16 mai 2024