An introduction to Bayesian System Identification

Some fundamental ideas and recent methods

Andreas Lindholm
May 28, 2019

Department of Information Technology
• **Classical** (parametric) **approach**: Find a value $\hat{\theta}$ that well fits data $y_{1:T}$ (*Loss functions, prediction error methods, maximum likelihood, etc.*)

$$\hat{\theta} = \arg \min_{\theta} L(\theta; y_{1:T})$$

• **Bayesian** (parametric) **approach**: Handle $\theta$ as a random variable using probability theory

$$p(\theta | y_{1:T}) = \frac{p(y_{1:T} | \theta)p(\theta)}{p(y_{1:T})}$$

• $p(\theta)$ **prior**
• $p(\theta | y_{1:T})$ **posterior**
• $p(y_{1:T} | \theta)$ **likelihood**
• $p(y_{1:T})$ **marginal likelihood**
1. Start with \( p(\theta) \), your **prior belief** about \( \theta \)

2. **Update your belief** about \( \theta \) after observing the data \( y_{1:T} \) (probability terminology: *conditioning* on the data) by computing the posterior \( p(\theta | y_{1:T}) \)

3. **Extract** whatever **information** you want about \( \theta \) from the posterior \( p(\theta | y_{1:T}) \), for example its mean, maximum mode, variance, ...
Why care about the Bayesian approach?

+ Gives a new perspective
+ Results in a different method with a different solution
+ Has a different and more intuitive statistical interpretation (in my personal opinion, feel free to disagree!)
  - Is often (but not always) more computationally intense

(compared to the classical approach)
Example: data, model, prior

Data $y_{1:T}$

Model $p(y_{1:T} | \theta)$
AR(2) with Gaussian noise

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + v_t,$$
$$v_t \sim \mathcal{N}(0, \sigma^2)$$

$\{a_1, a_2, \sigma^2\} = \theta$ (unknown)

Prior $p(\theta)$
Normal-inverse-gamma distribution

$$\sigma^2 \sim \mathcal{IG}(2, 2), \ a_i | \sigma^2 \sim \mathcal{N}(0, 0.1\sigma^2)$$
Example: posterior

Prior $p(\theta)$ (dashed gray)
Normal-inverse-gamma distribution

$$\sigma^2 \sim IG(a, b), \begin{bmatrix} a_1 \ a_2 \end{bmatrix} \mid \sigma^2 \sim \mathcal{N}(\mu, V\sigma^2)$$

Posterior $p(\theta \mid y_{1:T})$ (solid red)
Normal-inverse-gamma distribution

$$\sigma^2 \mid y_{1:T} \sim IG \left( a + \frac{T}{2}, b + \frac{1}{2} \left( \sum_{t=1}^{T} y_t^2 + \mu V^{-1} \mu - \mu V V^{-1} \mu \right) \right)$$

$$\begin{bmatrix} a_1 \ a_2 \end{bmatrix} \mid \sigma^2, y_{1:T} \sim \mathcal{N} \left( \mu_{VT}, \mu_{VT} V_{VT}^{-1} \right),$$

$$\mu_{VT} = \sqrt{\frac{V_{VT}^{-1} + \sum_{t=1}^{T} \begin{bmatrix} y_{t-1} \\ y_{t-2} \end{bmatrix} \begin{bmatrix} y_{t-1} & y_{t-2} \end{bmatrix}}{V_{VT}}} - 1$$
Example (second version): data, model, prior

Model $p(y_{1:T} \mid \theta)$

AR(2) with Laplacian noise (instead of Gaussian)

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + v_t,$$

$$v_t \sim \mathcal{L}(0, \beta)$$

$\{a_1, a_2, \beta\} = \theta$ (unknown)

Prior $p(\theta)$

Normal-inverse-gamma distribution

$$\beta \sim IG(2, 2), \ a_i \mid \beta \sim \mathcal{N}(0, 0.1\beta)$$
There is **no analytical expression for the posterior** $p(\theta \mid y_{1:T})$.

(The first version of the example had a so-called *conjugate prior*, meaning the posterior has the same form as the prior.)

In general, we cannot work out posteriors by pen and paper.
Many probability distributions are defined with a finite set of parameters (Gaussian: mean and variance, Bernoulli: success-rate, Laplace: mean and scale, ...).

Posterior distributions $p(\theta | y_{1:T})$ are not necessarily one of these ‘pre-defined’ distributions.

How to work with a general posterior distribution $p(\theta | y_{1:T})$?

- **Represent the distribution using samples** (Monte Carlo)
- **Approximate** the distribution with some parametric distribution (variational inference)
Representing an arbitrary distribution $\pi(\theta)$ with samples
Representing an arbitrary distribution $\pi(\theta)$ with samples
Representing an arbitrary distribution $\pi(\theta)$ with samples
Representing an arbitrary distribution $\pi(\theta)$ with samples

Histograms = approximations of probability distributions

Need a way to generate those samples!
One option: Markov chain Monte Carlo (MCMC)
Markov chain Monte Carlo (MCMC)

Idea: Construct a Markov chain (a random walk) in $\theta$ such that its stationary distribution is $p(\theta \mid y_{1:T})$, meaning the Markov chain will eventually generate samples from $p(\theta \mid y_{1:T})$.

(Widely used also for other purposes than system identification)
Example (second version): posterior

**Prior** $p(\theta)$ (dashed gray)
Normal-inverse-gamma distribution

$$\beta \sim \mathcal{IG}(a, b), \begin{bmatrix} a_1 & a_2 \end{bmatrix} | \beta \sim \mathcal{N}(\mu, \nu \sigma^2)$$

**Posterior** $p(\theta | y_{1:T})$(histograms)
Sampled using MCMC
(Metropolis-Hastings sampler)

1. Initialize $\theta[0]$
   - for $k = 0, 1, \ldots, K$
2. Propose $\theta' \sim q(\theta'; \theta[k])$
3. Sample $d \sim \mathcal{U}[0, 1]$
   - if $\frac{p(\theta' | y_{1:T}) q(\theta[k]; \theta')} {p(\theta[k] | y_{1:T}) q(\theta'; \theta[k])} > d$
4. $\theta[k + 1] = \theta'$
   - else
5. $\theta[k + 1] = \theta[k]$
   - end for
A state-space model:

\[ x_t = f_\theta(x_{t-1}) + w_t, \quad w_t \sim p(w_t), \quad \leftrightarrow \quad p(x_t | x_{t-1}, \theta), \]

\[ y_t = g_\theta(x_t) + e_t, \quad e_t \sim p(e_t), \quad \leftrightarrow \quad p(y_t | x_t, \theta). \]

But we actually need

\[
p(y_{1:T}, x_{1:T} | \theta) = \prod_{t=1}^{T} p(y_t | x_t, \theta) p(x_t | x_{t-1}, \theta)
\]

\[
p(y_{1:T} | \theta) = \int p(y_{1:T}, x_{1:T} | \theta) dx_{1:T}
\]
Bayesian system identification for state-space models

\[
p(y_{1:T} | \theta) = \int p(y_{1:T}, x_{1:T} | \theta) dx_{1:T}
\]

The integral can be solved with the help of a

- **Kalman filter**, if \(f_\theta, g_\theta\) are linear and \(p(w_t), p(e_t)\) Gaussian
- **particle filter**; only approximate, but applicable also to non-linear non-Gaussian models

Combining particle filter and MCMC (to generate the posterior \(\theta\)-samples) gives **particle-Markov chain Monte Carlo (PMCMC)**, a (family of) methods for Bayesian system identification of general state-space models.
State-space model example

\[ F_s = -ks^p, \]
\[ F_d = -f_c \text{sign}(s) - c_0 \dot{s}, \]
\[ x_{t+1}^1 = x_t^1 + T_s x_t^2; \]
\[ x_{t+1}^2 = x_t^2 + T_s (-f_c \text{sign}(x_t^2) - c_0 x_t^2 - k(x_t^1)^p) + v_t; \]
\[ y_t = x_t^1 + e_t; \]

measured displacement vs. time (s)
State-space model example

Posterior samples obtained using PMCMC:
Idea

- **Classical approach**: Find a value \( \hat{\theta} \) that well fits data \( y_{1:T} \)

\[
\hat{\theta} = \arg\min_{\theta} L(\theta; y_{1:T})
\]

*Workhorses: optimization methods*

- **Bayesian approach**: Handle \( \theta \) as a random variable using probability theory

\[
p(\theta | y_{1:T}) = \frac{p(y_{1:T} | \theta)p(\theta)}{p(y_{1:T})}
\]

*Workhorses: methods for solving integrals*
