Sparse Bayesian ARX models with flexible noise distributions

The 18th IFAC Symposium on System Identification, Stockholm, Sweden.

Johan Dahlin, University of Newcastle, Australia.

THE UNIVERSITY OF NEWCASTLE AUSTRALIA

This work in collaboration with

Dr. Adrian Wills (University of Newcastle, Australia)
Prof. Brett Ninness (University of Newcastle, Australia)

Autoregressive exogenous (ARX) model

$$y_t = \sum_{k=1}^{n_a} a_k y_{t-k} + \sum_{k=1}^{n_b} b_k u_{t-k} + e_t, \qquad e_t \sim g_\theta(\cdot).$$

Two practical problems:

- Finding the model orders n_a and n_b.
- Deciding on the noise distribution $g_\theta(\cdot)$.

Bayesian autoregressive exogenous (BARX) model

$$y_t = \sum_{k=1}^{n_a} a_k y_{t-k} + \sum_{k=1}^{n_b} b_k u_{t-k} + e_t, \qquad e_t \sim g_\theta(\cdot).$$

Modelling choices

- Prior for $\{a_k\}$ and $\{b_k\}$ to encode sparsity.
- Prior for $g_\theta(\cdot)$ to encode sparsity and flexibility.

Inference method

- Compute posterior of $\{a_k, b_k, g\}$ given $\{y_t, u_t\}_{t=1}^{T}$.

What are we going to do?

- Automatically select model orders and noise distributions.
- Employ an efficient sample scheme for inference.

Why are we doing this?

- Data-driven and automated method for Bayesian ARX modelling.
- Handle general noise distributions and blind identification.
- Provide predictors with uncertainty quantification for MPC.

How will we do this?

- Employ model averaging and sparseness priors.
- Model the noise with a Gaussian mixture.
- Sample from the posterior using Hamiltonian Monte Carlo.

Model averaging and sparseness priors.

Bayesian model averaging

The hyper-parameters η are marginalised out of the posterior,

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)}{p(\mathcal{D})} \int p(\theta|\eta)p(\eta)\,\mathsf{d}\eta \propto \sum_{k=1}^{K} p(\mathcal{D}|\theta_k),$$

with $\theta_k \sim p(\theta_k|\eta_k)$ and $\eta_k \sim p(\eta)$ obtained by Monte Carlo.

This takes into account all the values of η supported by $\mathcal{D}$.

Sparseness priors

$$y_t = \sum_{k=1}^{n_a} a_k y_{t-k} + \sum_{k=1}^{n_b} b_k u_{t-k} + e_t, \qquad e_t \sim g_\theta(\cdot).$$

Two common choices are

$$a_k, b_k \sim \mathcal{L}(0, \lambda), \qquad a_k, b_k \sim \mathcal{N}(0, \lambda^2),$$

where $\lambda > 0$ denotes the strength of the regularisation.

Need to select λ. Put a prior on it!

Sparsity with the horseshoe prior

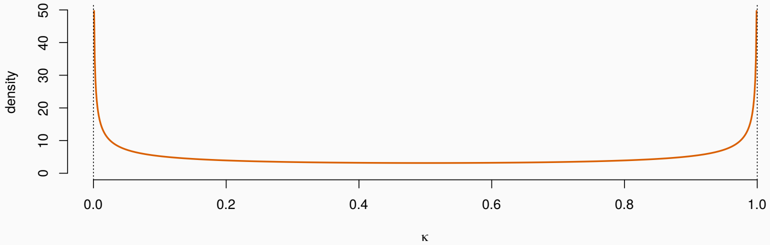
For a linear regression problem with Gaussian noise and

$$a_k \sim \mathcal{N}(0, \sigma_a^2), \quad \sigma_a \sim \mathcal{C}_+(0, 1),$$
$$b_k \sim \mathcal{N}(0, \sigma_b^2), \quad \sigma_b \sim \mathcal{C}_+(0, 1),$$

the prior has the effect that

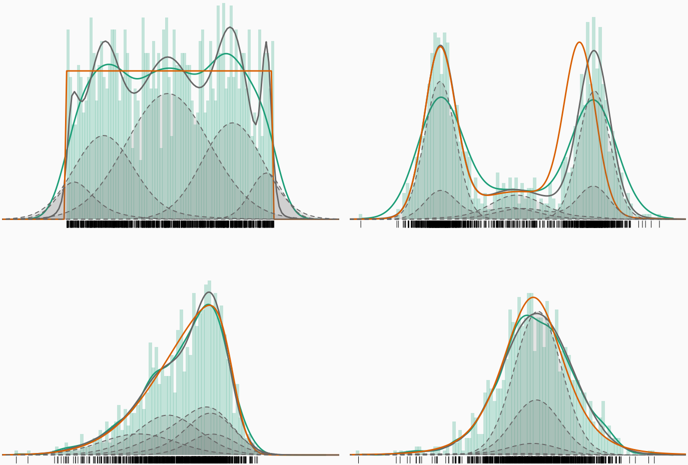
$$\bar{\beta}_j = (1 - \kappa_j)\widehat{\beta}_j,$$

where $\widehat{\beta}_j$ is the LS estimate.



Modelling noise using Gaussian mixture models.

Gaussian mixture models

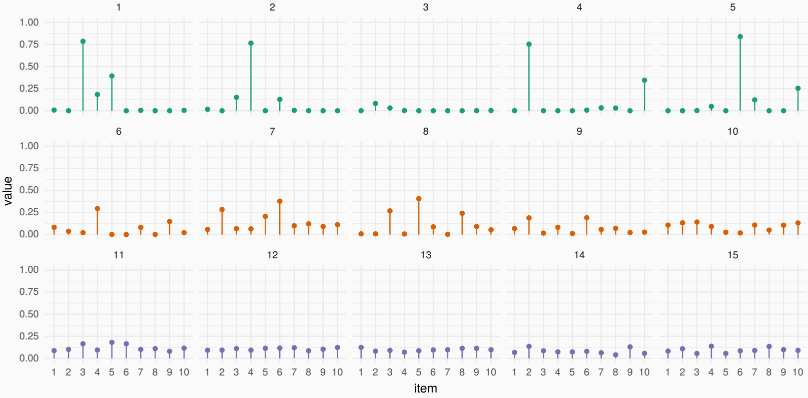


$$e_t \sim \sum_{k=1}^{n_e} w_k \mathcal{N}(e_t; \mu_k, \sigma_k^2), \quad \sum_{k=1}^{n_e} w_k = 1,$$

Dirichlet distribution

A distribution over the unit simplex,

$$x_1, x_2, \ldots, x_n \sim \mathcal{D}(\alpha_1, \alpha_2, \ldots, \alpha_n), \quad \text{such that } \sum_{i=1}^{n} x_i = 1.$$



Bayesian Gaussian mixture models

$$e_t \sim \sum_{k=1}^{n_e} w_k \mathcal{N}(e_t; \mu_k, \sigma_k^2), \quad \sum_{k=1}^{n_e} w_k = 1,$$

with sparsity promoting priors

$$\mu_{1:n_e} \sim \mathcal{N}(0, \sigma_\mu^2), \quad \sigma_\mu \sim \mathcal{C}_+(0, 1),$$
$$\sigma_{1:n_e} \sim \mathcal{C}_+(0, 5),$$
$$w_{1:n_e} \sim \mathcal{D}(e_0, \ldots, e_0), \quad e_0 \sim \mathcal{G}(\alpha_w, n_e \alpha_w).$$

Inference using Hamiltonian Monte Carlo.

Bayesian ARX model

$$y_t = \sum_{k=1}^{n_a} a_k y_{t-k} + \sum_{k=1}^{n_b} b_k u_{t-k} + e_t,$$

$$e_t \sim \sum_{k=1}^{n_e} w_k \mathcal{N}(e_t; \mu_k, \sigma_k^2),$$

$$\mu_{1:n_e} \sim \mathcal{N}(0, \sigma_\mu^2),$$

$$\sigma_\mu \sim \mathcal{C}_+(0, 1),$$

$$\sigma_{1:n_e} \sim \mathcal{C}_+(0, 5),$$

$$w_{1:n_e} \sim \mathcal{D}(e_0, \ldots, e_0),$$

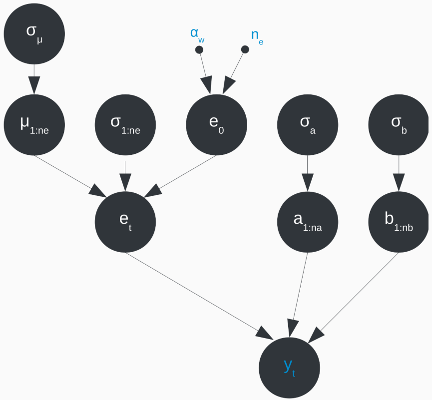
$$e_0 \sim \mathcal{G}(\alpha_w, n_e \alpha_w),$$

$$a_k \sim \mathcal{N}(0, \sigma_a^2),$$

$$\sigma_a \sim \mathcal{C}_+(0, 1),$$

$$b_k \sim \mathcal{N}(0, \sigma_b^2),$$

$$\sigma_b \sim \mathcal{C}_+(0, 1),$$



The inference problem in the BARX model

$$\pi(\theta, \eta) \triangleq p(\theta, \eta | \mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta|\eta)p(\eta)}{p(\mathcal{D})}$$

$$\mathcal{D} = \{x_t, y_t\}_{t=1}^{T},$$

$$\theta = \{a_{1:n_a}, b_{1:n_b}, w_{1:n_e}, \mu_{1:n_e}, \sigma_{1:n_e}\}$$

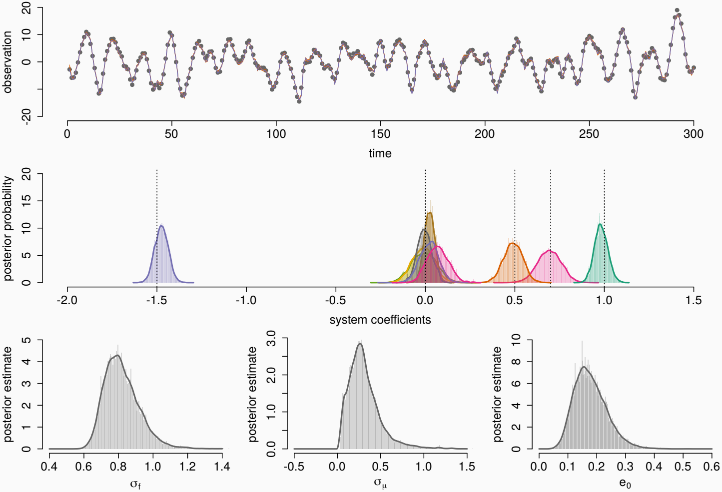
$$\eta = \{\sigma_u, e_0, \sigma_a, \sigma_b\}$$

Need to estimate $\pi(\theta, \eta)$ which is high-dimensional.

Use efficient Hamiltonian Monte Carlo!

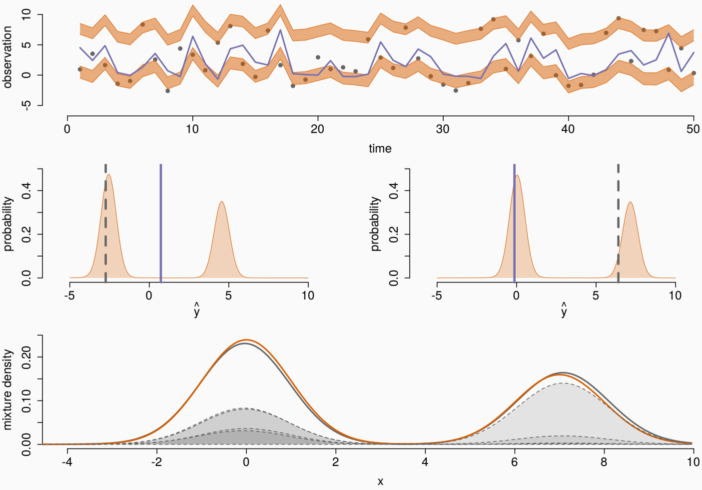
Numerical illustrations.

ARX model with Gaussian noise



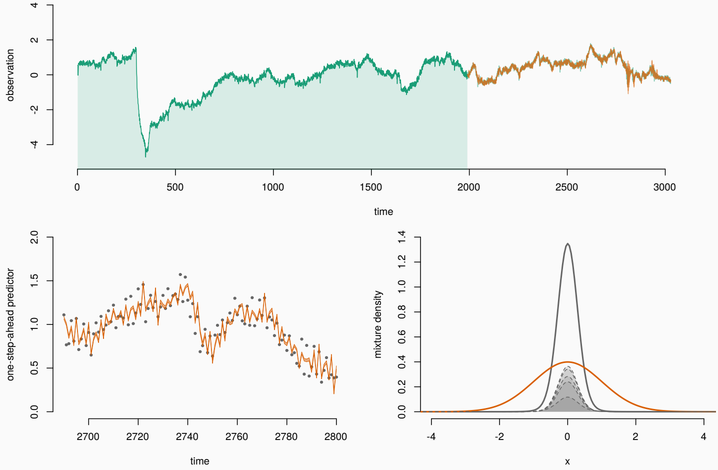
$$y_t = -1.5 y_{t-1} + 0.7 y_{t-2} + 1.0 u_{t-1} + 0.5 u_{t-2} + z_t, \quad z_t \sim \mathcal{N}(0, 1).$$

ARX model with mixture noise



$$y_t = -0.25 y_{t-1} + 0.2 y_{t-2} + 1 u_{t-1} + 0.5 u_{t-2} + z_t, \, z_t \sim 0.4 \cdot \mathcal{N}(7,1) + 0.6 \cdot \mathcal{N}(0,1).$$

Real-world EEG data



Model fit: 92.3% using BARX versus 85.6% using ARX with Student's t-noise.

What did we do?

- Automatically selected model orders using Horseshoe priors.
- Modelled the noise using a sparse Gaussian mixture.
- Employed efficient Hamiltonian Monte Carlo for inference.

Why did we do this?

- Data-driven and automated method for Bayesian ARX modelling.
- Handle general noise distributions and blind identification.
- Provide predictors with uncertainty quantification for MPC.

What are you going to do now?

- Remember hierarchical models and model averaging.
- Remember that simple Gaussian mixtures can be very useful.
- Read the paper and look at the code on GitHub.

Thank you for listening

Comments, suggestions and/or questions?

Johan Dahlin

johan.dahlin@newcastle.edu.au
research.johandahlin.com