# Bayesian Nonparametric Vector Autoregressive models via a logit stick-breaking prior

Mario Beraha[a], Alessandra Guglielmi[a] and Fernando Quintana[b] {mario.beraha@polimi.it}

[a]Politecnico di Milano, Milano & [b] Pontificia Universidad Católica de Chile

## Abstract

Vector Autoregressive (VAR) models may provide a flexible and powerful representation of longitudinal data; moreover extensions such as ARx models account for exogenous covariates. Bayesian nonparametrics has been successfully applied to VAR models in recent years but limited to purely autoregressive cases. We propose a semiparametric VAR model with exogenous covariates for nonstationary multidimensional longitudinal data, where a dependent stick-breaking prior is assumed for the autoregressive component through logit stick-breaking. We develop an efficient Gibbs sampling algorithm for MCMC posterior simulation, leveraging recent results on logit stick-breaking priors.

## Motivation

Open Angle Glaucoma (**OAG**) is an optic neuropathy characterized by progressive retinal ganglion cell death, structural changes to the retina and optic nerve head, and irreversible visual field loss. We consider a vector of three indicators of OAG progression:

1. Mean Deviation (**MD**): measure of the total amount of visual field loss.
2. Horizontal Integrated Rim Width area (**HIRWA**): topography variable of optic nerve head (ONH), assessed via Optical Coherence Tomography (OCT).
3. Retinal Nerve Fiber Layer thickness (**RNFL**): another topography parameter; larger values of RNFL denote patients in better conditions.

Clinical studies showed that the disease behaves differently in patients with different **gender**, **ethnicity**, **age**, **cardiovascular diseases** and **diabetes**.
We consider several measurements of overall and localized blood flows: intraocular pressure (**IOP**), mean arterial pressure (**MAP**), resistive index of the temporal posterior ciliary artery (**TPCAri**), heart rate (**HR**). We add to these the number of treatments the patient is undergoing at the time of the current visit (between 0 and 4).

## The model

### ARx Likelihood

$$\boldsymbol{Y}_{it} \mid \Phi_i, B, \Gamma, \epsilon_{it} = \Phi_i \boldsymbol{Y}_{it-1} + B\boldsymbol{x}_{it} + \Gamma \boldsymbol{z}_i + \epsilon_{it} \quad t = 1:T_i, \;\; n = 1:N$$

$$\epsilon_{it} \mid \Sigma \overset{\text{iid}}{\sim} \mathcal{N}_K(\boldsymbol{0}, \Sigma) \quad t = 1:T_i, \;\; n = 1:N$$

### Dependent logit stick breaking prior

$$\boldsymbol{\Phi}_i | z_i \overset{\text{ind}}{\sim} \sum_{h=1}^{H} w_h(\boldsymbol{z}_i)\delta_{\Phi_{0h}} \quad i = 1, \dots, N$$

$$w_h(\boldsymbol{z}_i) = \nu_h(\boldsymbol{z}_i) \prod_{l=1}^{h-1} (1 - \nu_l(\boldsymbol{z}_i)) \quad h = 2:H, \quad w_1(\boldsymbol{z}_i) = \nu_1(\boldsymbol{z}_i)$$

$$\text{logit}(\nu_h(\boldsymbol{z}_i)) = \boldsymbol{z}_i^T \alpha_h \quad h = 1:H-1, \quad \nu_H(\boldsymbol{z}_i) = 1$$

$$\alpha_h \overset{\text{iid}}{\sim} \mathcal{N}_q(\mu_\alpha, \Sigma_\alpha)$$

The prior specification is completed by:

$$\Phi_{0h} | \Phi_{00}, V_0 \overset{\text{iid}}{\sim} \mathcal{N}_{K^2}(\Phi_{00}, V_0)$$
$$\Phi_{00}, V_0 | \Phi_{000}, \lambda, V_{00}, \tau_0 \sim \mathcal{NIW}(\Phi_{000}, \lambda, V_{00}, \tau_0)$$
$$B \sim \mathcal{N}_{Kp}(\boldsymbol{0}, \Sigma_B) \quad \boldsymbol{\Gamma} \sim \mathcal{N}_{Kq}(\boldsymbol{0}, \Sigma_\Gamma) \quad \Sigma \sim \mathcal{IW}(\Sigma_0, \nu)$$

## Dataset

We consider longitudinal measurements from visits of $N = 108$ glaucoma patients. The data were collected within the Indianapolis Glaucoma Progression Study (IGPS) directed by Prof. Alon Harris at Eugene and Marilyn Glick Eye Institute, Indianapolis (USA). Patients were evaluated every six months, $\min T_i = 1, \;\; \max T_i = 12$.

## References

- Billio, Casarin and Rossini "Bayesian nonparametric sparse VAR models." *Journal of Econometrics* (2019)
- Rigon and Durante "Logit stick-breaking priors for Bayesian density regression." *arXiv preprint arXiv:1701.02969* (2017)

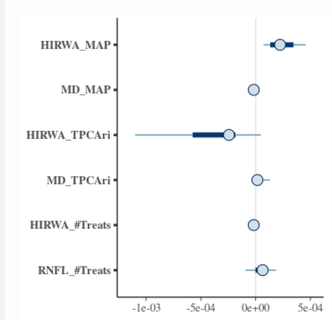
## Gibbs sampler

All the full conditionals are available analitically. In particular, we view the cluster allocation procedure as a sequential Bernoulli decision process and exploit recent result in Bayesian logistic regression. Sampling for the weights parameters $\{\alpha_h\}$ is performed by introducing a latent variable $\omega_{ih}$ and exploiting the Polya-Gamma scheme.
Letting $\omega_{ih} \mid \alpha_h \sim \text{PG}(1, \boldsymbol{z}_i^T \alpha_h)$ yields $\alpha_h \mid rest \sim \mathcal{N}(\widetilde{\mu}, \widetilde{\Sigma})$.

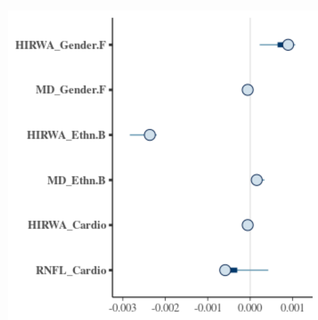
### Algorithm

- Update $B$ using the change of variables:
$$Y_{it} - \Phi_i Y_{it-1} - \boldsymbol{z}_i \Gamma = \boldsymbol{x}_{it}^T B + \boldsymbol{E}$$
- Update $\Gamma$ using the change of variables:
$$Y_{it} - \Phi_i Y_{it-1} - \boldsymbol{x}_{it}^T B = \boldsymbol{z}_i^T \Gamma + \boldsymbol{E}$$
- Update the atoms $\Phi_{0h}$ and the clusters as in standard blocked Gibbs sampling
- Update $\Sigma$ from its full conditional.
- Update the weights parameter $\{\alpha_h\}$
  - sample $\omega_{ih}$ from $\omega_{ih} \mid rest \sim \text{PG}(1, \boldsymbol{z}_i^T \alpha_h)$
  - sample $\alpha_h$ from $\alpha_h \mid rest \sim \mathcal{N}(\widetilde{\mu}, \widetilde{\Sigma})$
- Update of hyperparameters in the prior for $\Phi_{0h}$.

## Posterior Inference




Posterior C.I.s for $B$

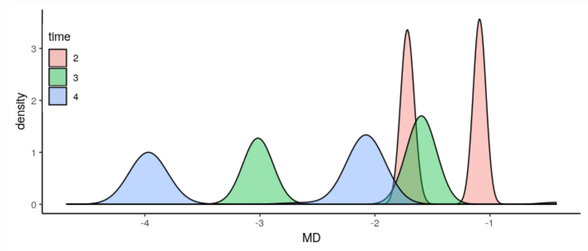



Posterior C.I.s for $\Gamma$

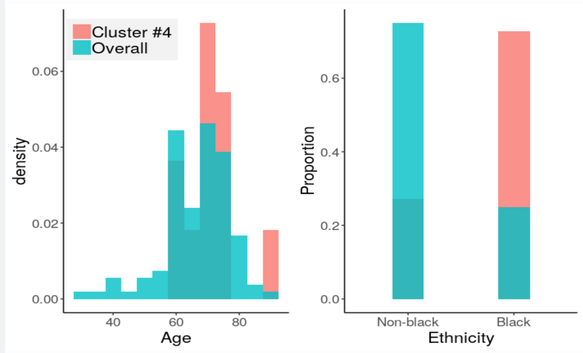
- **strong** effect of gender, ethnicity, TPCAri and MAP on HIRWA
- **milder** effect of ethnicity, and TPCAri on MD
- **strong** effect of cardiovascular diseases on RNFL

**Predictive density** for a new patient: female,age: 57.

- The disease is **progressing**: MD is smaller over time!




## A posteriori interpretation of the clusters




- **Cluster** estimate, found by minimizing Binder loss, shows 6 different clusters.
- The estimated clusters can be "explained" in term of **Age**, **Gender** and **Ethnicity**. That is, the empirical distributions of these covariates is different among clusters.

## Open Problems

- How to monitor and predict disease **progression**?

## Model Comparison

| Model | LPML | MSE |
|---|---|---|
| LSB-VARx | -8.41 | 390.00 |
| DP-VARx | -9.07 | 397.31 |
| Parametric-VARx | -9.64 | $\approx 10^{40}$ |