

Master projects

Aliaksandr Hubin, postdoc, Office 810

Department of Mathematics
University of Oslo

aliaksah@math.uio.no



UiO : Universitetet i Oslo

Oslo
01.10.2021

Bayesian Generalized Nonlinear Model (JAIR 2021)

Sample of observations $i = 1, \dots, n$

- Y_i ... response data;
- $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$... p -dimensional vector of input covariates.

Specification of the model

From input variables a huge (but finite) number of features can be generated: $F_j(\mathbf{x}_i)$, $j = 1, \dots, q$ (consider ordering w.r.t. complexity)

The **model** is then specified as GLM:

$$Y_i|\mu_i, \phi \sim f(y|\mu_i, \phi), i = 1, \dots, n; \quad (1)$$

$$h(\mu_i) = \beta_0 + \sum_{j=1}^q \gamma_j \beta_j F_j(\mathbf{x}_i, \boldsymbol{\alpha}_j). \quad (2)$$

- $f(\cdot|\mu, \phi)$ - density from exponential family with mean μ_i and dispersion parameter ϕ ;
- h - link function;
- $\beta_j \in \mathbb{R}$ - regression coefficients of j -th feature;
- $\gamma_i \in \{0, 1\}$ - indicator variable for j -th feature.

Hierarchy of the features

A feature $F_j(\mathbf{x}, \alpha_j)$, $j \in \{p+1, \dots, q\}$ can be constructed recursively through:

$$F_j(\mathbf{x}, \alpha_j) = v(\alpha_j^T F_{1:j-1}(\mathbf{x}))$$

$v \in \mathcal{G}$ is one of the allowed basic function from set $\mathcal{G}$.

Types and meaning of functions in $\mathcal{G}$

- Neural Networks: $\text{logit}(x)$, $\tanh(x)$, $\text{erf}(x)$, $\text{ReLU}(x)$;
- Polynomials: $F_k(\mathbf{x}) * F_l(\mathbf{x}) = \exp(\log(F_k(\mathbf{x})) + \log(F_l(\mathbf{x})))$;
- CART: $I(x \geq 1)$;
- MARS: $\max\{0, x - t\}$ and $\max\{0, t - x\}$;
- Fractional polynomials: $x^{\frac{1}{a}} = \exp(b \log(x))$, $b = \frac{1}{a}$;
- **Logical AND, OR and NOT:** $L_k \wedge L_l = L_k * L_l$,
 $L_k \vee L_l = L_k + L_l - L_k * L_l$, and $\bar{L}_k = 1 - L_k$.

Examples on Inference

Dataset: $n = 223$ exoplanets. We want to recover 2 basic physical laws.

Input variables include: *TypeFlag*, *RadiusJpt*, *PeriodDays*, *PlanetaryMassJpt*, *Eccentricity*, *HostStarMassSlrMass*, *HostStarRadiusSlrRad*, *HostStarMetallicity*, *HostStarTempK*, *PlanetaryDensJpt* denoted as x_1 - x_{10} .

Planetary mass

$$m_p \approx K_1 R_p^3 \times \rho_p.$$

Planetary mass m_p is proportional to cube of radius R_p times the density of the planet ρ_p

Kepler's third law

The square of the orbital period P of a planet is directly proportional to the cube of the semi-major axis a of its orbit:

$$a \approx K_2 (P^2 M_h)^{\frac{1}{3}}.$$

Examples for Prediction

Three binary classification tasks

- Asteroids data: Mean anomaly, Inclination, Argument of perihelion, Longitude of the ascending node, Rms residual, Semi major axis, Eccentricity, Mean motion, Absolute magnitude;
- Breast cancer data: Radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension;
- Spam emails: 58 characteristics, including 57 continuous and 1 nominal variable, where most of these are concerned with the frequency of particular words or characters. 3 provide different measurements on the sequence length of consecutive capital letters.

Regression task

- Abalone age prediction: sex, length, diameter, height, whole weight, shucked weight, viscera weight and shell weight.

Open questions for master theses

Scalability within the populations:

- Try variational inference within the populations;
- Likelihood-free MCMC with amortized approximate ratio estimators;
- Further develop subsampling in MCMC;

Choice of weights α :

- Needs more methodology and systematic simulation studies to get a better idea what is really needed or only assume with specific models e.g. fractional polynomials to start with;

Choice of priors:

- Priors for consistency of the discoveries and controlling FDR;

Other topics:

- Bayesian neural networks/model selection;
- Fraud detection in Bank transactions.

Literature

1. Hubin, A., Storvik G. (2018). **Mode jumping MCMC for Bayesian variable selection in GLMM.** *Journal of Computational Statistics and Data Analysis*; 2018 November; 127:281-297.
2. Hubin, A., Storvik G., Frommlet F. (2018). **A novel algorithmic approach to Bayesian Logic Regression (with Discussion).** *Bayesian analysis*; 2020.
3. Hubin, A., Storvik G., Frommlet F. (2018). **Flexible Bayesian Nonlinear Model Configuration.** *Accepted to JAIR for publication.*