Bayesian Generalized Nonlinear Models Offer Basis Free SINDy with model uncertainty

Aliaksandr Hubin[1]

[1] NMBU, HiOF, UiO, Norway

E-mail for correspondence: aliaksandr.hubin@nmbu.no

Abstract: Sparse Identification of Nonlinear Dynamics (SINDy) has become a standard methodology for inferring governing equations of dynamical systems from observed data using statistical modeling. However, classical SINDy approaches rely on predefined libraries of candidate functions to model nonlinearities, which limits flexibility and excludes robust uncertainty quantification. This paper proposes Bayesian Generalized Nonlinear Models (BGNLMs) as a principled alternative for more flexible statistical modeling. BGNLMs employ spike-and-slab priors combined with binary inclusion indicators to automatically discover relevant nonlinearities without predefined basis functions. Moreover, BGNLMs quantify uncertainty in selected bases and final model predictions, enabling robust exploration of the model space. In this paper, the BGNLM framework is applied three dimensional canonical SINDy, including the Lorenz system, linear ODEs, and a hybrid Rössler-Lorenz system.

Keywords: BGNLM; GMJMCMC; SINDy; Flexible Statistical Modeling.

1 Introduction

The sparse recovery of dynamical systems has recently become a useful tool in applied sciences, with SINDy offering a particularly effective method for reconstructing governing equations (Brunton et al, 2016). However, SINDy simply applies l_1 regularized lasso variable selection for predefined libraries of candidate basis functions (e.g., polynomials, trigonometric terms), which may limit its adaptability. Furthermore, SINDy requires tuning of the regularization strength for the l_1 norm and does not quantify uncertainty in the selected models, which can be problematic when data are noisy or sparse. The latter was resolved in a Bayesian version of SINDy (Hirsh et al, 2022), yet this solution still relies on a predefined library limiting the flexibility of statistical modeling.

Bayesian Generalized Nonlinear Models (BGNLMs) (Hubin et al, 2021) can address both of these challenges. BGNLMs on one hand allow for

flexible statistical modeling of nonlinear features, but on the other hand employ a probabilistic framework with spike-and-slab priors, allowing for sparse selection of bases while accounting for uncertainty. Binary inclusion indicators are used to define the model priors, and Bayesian inference through a geneticaly modified mode jumping MCMC explores the vast nonlinear model space, providing marginal distributions over plausible governing equations. These capabilities allow to hypothesize that BGNLM may offer a novel, powerful, and robust tool for solving SINDy problems.

2 Mathematical Framework

Consider a system of ordinary differential equations (ODEs) of the form:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t)),$$

where $\boldsymbol{x}(t) \in \mathbb{R}^m$ is the state vector, and $\boldsymbol{f}(\boldsymbol{x})$ is the unknown governing function. Assume that we observe a sample of noisy realizations of the stochastic process associated with the derivatives $Y_i = \dot{\boldsymbol{x}}(t_i) + \epsilon$ with $Y_i \in \mathbb{R}^m, i \in \{1, ..., n\}$ and $\epsilon \sim N_m(\mathbf{O}, \sigma^2 \mathbf{I_m})$. Further we observe the process itself at these points $\boldsymbol{x_i} = \boldsymbol{x}(t_i), i \in \{1, ..., n\}$.
BGNLM, allows to statistically model each component $\dot{x_j}(t) = \boldsymbol{f}_j(\boldsymbol{x}(t))$ as $Y_{ji} \sim N(f_j(\boldsymbol{x_i}), \sigma^2), j \in \{1, ..., m\}$ with

$$f_j(\boldsymbol{x_i}) = \sum_{k=1}^{q} \gamma_{jk} \beta_{jk} g_k(\boldsymbol{x_i}),$$

where $g_j(\boldsymbol{x})$ are the space of BGNLM's nonlinear features, $\beta_{jk} \in \mathbb{R}$ are coefficients for the k-th basis of the j-th equation, $\gamma_{jk} \in \{0, 1\}$ are binary inclusion indicators for them. The inclusion indicators γ_{jk} are governed by Bernoulli priors: $\gamma_{jk} \sim$ Bernoulli(π_k), where π_k represents the prior probability of including $g_k(\boldsymbol{x})$ in the models, which fully follow Hubin et al (2021). The coefficients β_{jk} and σ^2 are assigned Jeffreys conditional on the configuration of the inclusion of the respective bases (Hubin et al, 2021).
Compared to classical SINDy, BGNLM automatically discovers relevant nonlinearities and estimates the probabilities of inclusions of the components into the basis. This allows to select a robust solution through the median probability model, which selects the features with $p(\gamma_{jk}|\mathbf{D}) > 0.5$.

3 Applications

3.1 Experimental Systems

Table 1 summarizes three nonlinear systems we will be identifying using statistical modeling with BGNLMs: (a) Linear 3D, (b) Lorenz 3D, and (c) Rössler-Lorenz Hybrid 3D.

System	Equations	Parameters	Init. Cond.
Linear 3D	$\dot{x} = ax + bxy$ $\dot{y} = cx + dxy$ $\dot{z} = ez$	$a = -1,\ b = 20,$ $c = -20, d = -1,$ $e = -3$	$x = 2,$ $y = 0,$ $z = 1$
Lorenz 3D	$\dot{x} = \sigma(y - x)$ $\dot{y} = x(\rho - z) - y$ $\dot{z} = xy - \beta z$	$\sigma = 10,$ $\rho = 28,$ $\beta = \frac{8}{3}$	$x = -0.5,$ $y = -2,$ $z = 3$
Hybrid 3D	$\dot{x} = -y - z + a\sin(x)$ $\dot{y} = x + by$ $\dot{z} = c + z(x - d)$	$a = 0.2,$ $b = 0.1,$ $c = 0.4, d = 5.7$	$x = 0.5,$ $y = -1,$ $z = 2$

TABLE 1. Governing equations, parameters, and initial conditions for three nonlinear dynamical systems.

The simulations were performed for the systems using a fine-grained time step of $\Delta t = 0.0001$ and a total simulation time of $T = 50$, resulting in $n = 500,000$ observations per trajectory. The simulations were conducted at multiple distinct noise levels, defined as $0.1 * 2^k, k \in 0, ..., 7$. Furthermore, 10 independent repetitions were run per noise level, resulting in a comprehensive exploration of model performance under diverse conditions. The finite difference method was used to compute derivatives from noisy trajectories. For each replication at each noise level, the dataset was divided into the following subsets for evaluation and training: **Training data:** Randomly sampled $1,000$ observations from $t \in (0.05, 49.5)$. **In-sample predictions:** A subset of 1,000 observations randomly sampled from $t \in (0.05, 49.5)$ but non-overlapping with the training data. **Out-of-sample predictions:** 1000 samples from two disjoint intervals, covering the first and the last parts of the trajectories $t \in [0, 0.05] \cup [49.5, 50]$ allowing us to estimate the generalizability of the found solutions to completely unobserved parts of the domain.

For each noise level and repetition, model performance was evaluated based on Power/FDR analysis of identified terms, as well as R^2 for the training, in-sample predictions, and out-of-sample predictions sets. Median probability model was used as the finally selected model for the identified terms and predictions. Predictions were made using the posterior mode estimates of the parameters of the median probability model under Jeffreys priors. The tuning parameters for the `fbms` function were set as follows (without additional tuning): The population size (`pop.max`) was limited to 15, with original features retained (`keep.org = TRUE`). The number of MJMCMC iterations per population was set to 500. Nonlinear transformations included `sin_deg`, `cos_deg`, `p0`, `p2`, `p3`, `p05`, `pm1`, `pm2`, and `pm05`. The maximal population size was set to 20. Finally, 10 chains of the GMJMCMC were run for every simulation. Implementation with all details is available at `github.com/aliaksah/BGNLM-for-SINDy`.

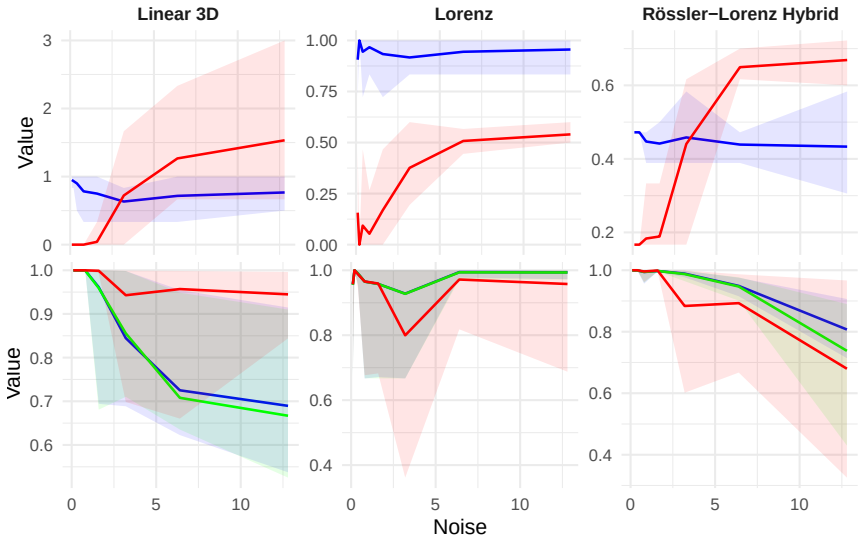

FIGURE 1. **Top:** Power (blue) FDR (red) curves for the identification of the the three systems. **Bottom:** R^2 curves for the predictions of the the three systems for train (blue), test (green) and out-of-sample test (red) data.

4 Results and Discussion

BGNLM showed a promising trade off between statistical Power and FDR for identifying the governing equations for each system, especially for lower noise levels, see Figure 1. Predictions also seem to be reasonable not only for the in-sample domain, but also for the out-of-distributions data, yet again their quality often drop for higher levels of noise. This also comes with quantified uncertainties in the selected terms, allowing to make uncertainty aware decisions. Thus, flexible Bayesian statistical modeling allows exploration of multiple plausible nonlinear models and avoids overfitting to noisy data. Unlike traditional SINDy, which requires predefining a library of functions, BGNLM adapts to the structure of the underlying dynamics. Thanks to automatically discovering relevant bases, quantifying uncertainties, and exploring plausible model spaces, BGNLM may well-suited for solving real-world identification problems.

In the future, it is of interest to obtain theoretical results for the signal to noise ratios allowing for identifibility of the system. Further, de-noising of the derivatives will become crucial for accurate discoveries under small and sparse samples. Finally, applications to real-world dynamic systems will be of interest to check the practical use of BGNLMs for applied problems.

References

Brunton, S.L., Proctor, J.L., and Kutz, J.N. (2016). Discovering governing equations from data: Sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U.S.A.*, **113**(15), 3932–3937.

Hirsh, S.M., Barajas-Solano, D.A., and Kutz, J.N. (2022). Sparsifying priors for Bayesian uncertainty quantification in model discovery. *R. Soc. Open Sci.*, **9**(2), 211823.

Hubin, A., Storvik, G., and Frommlet, F. (2021). Flexible Bayesian nonlinear model configuration. *J. Artif. Intell. Res.*, **72**, 901–942.