Bayesian equation discovery of non-linear stochastic dynamical systems through sparse linear regression using spike and slab prior

Tapas Tripura[1] **and Souvik Chakraborty**[1,2,*]

[1]Department of Applied Mechanics, Indian Institute of Technology Delhi, 110016, India
[2]School of Artificial Intelligence (ScAI), Indian Institute of Technology Delhi, 110016, India
[*]souvik@am.iitd.ac.in

ABSTRACT

We propose a novel data-driven framework for discovering the governing physics of a non-linear dynamical system from the noisy measurement of system states in this work. A complete understanding about the governing physics of a dynamical system significantly enhances our ability to model and predict the systems behaviour in an unseen environment. The governing physics of many natural processes are generally modelled by putting together the first principles with scientific laws. These mathematical models are many a case approximate in nature. Often in complex environments the minimal knowledge of the physical process are not obtainable. In such cases, the modern advancement in the machine learning techniques has enabled the researchers and scientists to identify the governing physics of a complicated system from noisy data without invoking scientific assumptions. However recently developed physics discovery techniques requires the information about both the input force and output state measurement data. The driving input force in almost any natural process is not measurable directly due to limitation in the current state of art of the force estimation algorithms. To add, the forces in a naturally occurring process continuously evolves over time, making them intractable to measure. In such cases these techniques are bound to fail. Also measurement of all the systems states such as acceleration, displacement and velocities of a dynamical system can impose additional cost. Thus a novel framework should make use of as minimum states as possible and discover the governing physics without explicit measurement of the input force data. The absence of external force information essentially transfers the physics discovery problem from deterministic to stochastic domain. Hence we leverage the stochastic calculus theory and propose a novel framework to discover the governing physics of a non-linear stochastic dynamical system, in terms of the stochastic differential equations (SDE). In the proposed framework, a sparse Bayesian linear regression is formulated for identification of an SDE in terms of its drift and diffusion components. The target vector in the sparse linear regression is formulated as linear and quadratic variations of the noisy displacement time history. The spike and slab distributions are used as prior over the weight vector for promotion of sparsity in the solution. The statistical properties of the posterior distribution are utilized to obtain the final governing physics and associated confidence bounds. The fidelity of the proposed work is tested on variety of examples involving both fully and partially observable state measurement data. The results indicate the successful identification of actual governing physics.

Introduction

Physical systems are governed by the physical law, often represented either in form of ordinary differential equations (ODE) or partial differential equations (PDE). Once the governing ODE/PDE is defined, there exist a plethora of methods including finite element and finite volume methods for solving the same. The literature on solution of ODE/PDE is quite matured and given sufficient computational resources, it is possible to solve the governing ODE/PDE with sufficient accuracy. Modern techniques such a physics-informed deep learning and neural network based operator learning can also be used. However, due to various assumptions and approximations, the governing equations often fails to represent the exact physics of the system and results in modelling and prediction errors. With advancement in the sensor technology and the internet of things (IoT), we have access to data in abundance whereas the underlying governing physics often remains elusive specifically is domains such as climate science[1], biology[2], physics[3], chemistry[4] and finance[5] to name a few. One possible alternative is obviously to train a purely data-driven machine learning based model to obtain an input-output mapping; however, such models often do not generalize to unseen environment and out-of-distribution inputs. To address this issue, methods for data-driven equation discovery has recently been proposed. A seminal work in this area was carried out by[6] wherein symbolic regression was utilized for discovering underlying structure of the data. Another seminal work in this area includes work by[7]. However, most of the work in this area is limited by the fact that information on both input and state vector are needed. However, in practice, the input is often not measurable and only few states are accurately observable; this greatly limits the applicability of the available physics discovery techniques. To address this issue, we propose a novel approach that leverages sparse learning, Bayesian

statistics, and stochastic calculus and enables discovery of governing physics from only the state measurements.

The development in the equation discovery techniques has made significant progress from early 1970s when the equation discovery was mostly dependent on the expertise in the interested area, to the modern machine learning techniques that can extract the physical law of the underlying process with only little human supervision. In the early 1970s the models were selected by maintaining a balance between the model complexity and its fitness. The fitness of the models were measured using Akaike (AIC) and Bayesian information criterion (BIC)[8,9]. Later advances in the machine learning tools in combination with new sophisticated data measurement instruments gave rise to the data-driven techniques for discovery of governing physics. Towards this, first attempt was made by combining symbolic regression with genetic programming to select the best combination of the functions from candidate functional forms that accurately represent the data[6,10]. There are also equation-free methods that bypasses the need of formulating constitutive equations to track the time evolution of the dynamical systems in closed form. These methods provide a practical recipe for multiscale simulations through local microscopic simulations in time and space[11].

Following the work of symbolic regression[6,10], Sparse Identification of Nonlinear Dynamics (SINDy) was proposed later for discovery of the governing physics of a non-linear dynamical system[7]. In terms of accuracy and intepretability of the discovered physics, the SINDy framework overcame the shortcomings of the previous discovery techniques. The idea behind SINDy was to use sparse linear regression in the purview of least-square fitting for selecting the most dominating candidate functions that best represent the data. Due to sparsity promoting approach, SINDy proved to be computationally efficient and scalable with the increase in the input dimension. In the later years, SINDy has shown tremendous applications and area-specific developments of sparse linear regression in various fields. Few example are, sparse identification of biological networks in biology[12] and sparse identification of chemical reaction in chemistry[13,14], sparse modelling for state estimation in fluid mechanics[15,16], system identification of structures with hysteresis and sparse learning of aerodynamics of bridges in structural systems[17,18], sparse model selection using an integral formulation[19], sparse model selection of dynamical system using information criteria[20], sparse identification for predictive control[21], identification of structured dynamical systems with limited data[22], model identification using recovery of differential equations from short impulse response time-series data[23], identifying stochastic dynamic equation[24], and discovery of partial differential equations[25,26], among others.

Apart from SINDy, identification of non-linear dynamical systems using deep neural networks were also proposed by the researchers[27–29]. The deep learning approaches for non-linear dynamical systems exists as a black-box which takes the data as a input and provides the prediction as output. Later, a different approach within the Bayesian framework for discovery of governing physics from noisy and/or limited data was proposed[30,31]. In this approach the least-square based sparse linear regression was replaced by more accurate sparse Bayesian linear regression technique. Due to the ability of the Bayesian inference to perform simultaneous model selection and parameter estimation, this approach eliminated the possibility of overfitting of the governing model. The sparsity in the solution was promoted by assigning the appropriate sparse promoting priors over the weight vector of the sparse regression problem. This approach provided a natural elimination of the basis functions that do not present the data accurately. As a output, this framework provides the posterior distributions of the weight vector; this means one could accurately able to identify the uncertainty in the parameter estimation and construct a confidence interval for future predictions.

Despite the progress in equation discovery from data in recent times, almost all the approaches require measurements for both state and input variables; the only exceptions is perhaps the Stochastic SINDy[24]. Unfortunately, we often only have access to the state variables only; this significantly limits the applicability of the available approaches. We hereby propose a equation discovery framework rooted in stochastic calculus, sparse learning, and Bayesian statistics. that require only the state estimates. The basic premise here is to treat the unknown input as a stochastic process, model it as a Gaussian white noise, and identify the stochastic differential equation (SDE); this essentially leads to identifying the drift component and diffusion component of a stochastic differential equation. We utilize stochastic calculus to decouple the drift and diffusion part; this allows parallel identification of the two components. Sparse learning and Bayesian statistics, on the other hand, ensures that the governing equation identified is sparse and interpretable; this is achieved by using the spike and slab prior[32,33]. The proposed approach has several key features that can be encapsulated into the following points:

- First and foremost, unlike other approaches, the proposed approach require only the state measurements; no input measurements are needed.
- The proposed approach being Bayesian in nature estimates the posterior distribution. This allows computation of the epistemic uncertainty (aka predictive uncertainty) due to limited data. This feature is particularly important in design or decision making.
- Unlike non-Bayesian approaches, the proposed approach is autonomous in the sense that it doesn't require any human calibration and cross-validation.

- The proposed framework falls under the broad umbrella of interpretable machine learning and explainable AI. This allows generalization of the model to unseen environment and out-of-distribution data.

Results

Discovery of governing physics without input measurement

Natural dynamical processes are generally expressed in terms of higher order differential equations. However, in most of the data-driven machine learning techniques the actual system is not observed through its original space but identified in a projected space. In the current work, the actual dynamics of the underlying process is expressed in terms of the first order Itô . Let the projection can be realized by a map $\mathcal{T} : \mathbb{R}^d \rightarrow \mathbb{R}^m$ such that $m < d$. If there exists a mapping $\mathcal{T}$, then any higher order system can be reduced to a set of SDEs of the form:

$$\dot{\boldsymbol{X}} = f(\boldsymbol{X}_t, t) + g(\boldsymbol{X}_t, t)\xi(t) \tag{1}$$

where, $f(\boldsymbol{X}_t, t)$ is the deterministic dynamics of the underlying process, $g(\boldsymbol{X}_t, t)$ is the volatility associated with the dynamics arising due to the stochastic external input, and $\xi(t)$ is the white noise. Let $\{B(t); t \geq 0\}$ be the Brownian motion with the properties $\langle B(t) \rangle = 0$, and $\langle B(s), B(t) \rangle = \min(s, t)$. Then the white noise can be defined as, $\xi(t) = \dot{B}(t)$. With the above definitions, an m-dimensional SDE is expressed as:

$$d\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{X}_t, t)\, dt + \boldsymbol{g}(\boldsymbol{X}_t, t)\, d\boldsymbol{B}(t); \quad \boldsymbol{X}(t = t_0) = \boldsymbol{X}_0; \quad t \in [0, T] \tag{2}$$

where, $\boldsymbol{X}_t \in \mathbb{R}^m$ denotes the states of the process, $\boldsymbol{f}(\boldsymbol{X}_t, t) : \mathbb{R}^m \mapsto \mathbb{R}^m$ is the drift vector, $\boldsymbol{g}(\boldsymbol{X}_t, t) : \mathbb{R}^m \mapsto \mathbb{R}^{m \times n}$ is the diffusion matrix and $\boldsymbol{B}_j(t) \in \mathbb{R}^n$ is the n-dimensional Brownian motion. The Brownian motion $B(t)$ is generally not differential with respect to the process $X(t)$. However, their continuity is assumed to exists in mean square sense. Formally, if the interval $s \in [0, t]$ is partitioned into n-parts as, $\mathcal{P}_n(0, t) := [0 = s_0 < s_1 < \ldots < s_n = t]$, then for a random process w the quadratic variation $Q_n(w, t) = \sum_i^n |w(s_i) - w(s_{i-1})|^2$ converges to $Q_n(w, t) \rightarrow Q(w, t)$. From the property of Brownian motions it can be found that $Q(w, t)$=t. This states that $B(t)$ has zero finite variation but non-vanishing quadratic variation. On the other hand, the deterministic functions have finite variation but zero quadratic variations. Leveraging this facts the drift and diffusion components of an SDE given in Eq. (2) can be expressed in terms of the sample time history as follows (more details are present in the *Methods* section and Appendix A):

$$\begin{aligned} f_i(\boldsymbol{X}_t, t) &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} E\left[X_i(t + \Delta t) - \xi_i\right]\Big|_{x_k(t) = \xi_k} \quad \forall\, k = 1, 2, \ldots N \\ \Gamma_{ij}(\boldsymbol{X}_t, t) &= \frac{1}{2} \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} E\left[|x_i(t + \Delta t) - \xi_i| \left|X_j(t + \Delta t) - \xi_j\right|\right]\Big|_{x_k(t) = \xi_k} \quad \forall\, k = 1, 2, \ldots N \end{aligned} \tag{3}$$

where, $f_i(\boldsymbol{X}_t, t)$ is the i^{th} drift component and $\boldsymbol{\Gamma}_{ij}$ is the $(ij)^{th}$ component of the diffusion covaraince matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{n \times n} := g(t, \boldsymbol{X}_t) g(t, \boldsymbol{X}_t)^T$. Let us assume that the analytical form of the drift and diffusion components can be obtained from measurement sample paths in terms of some basis functions. Let $\mathbf{L}^f \in \mathbb{R}^{N \times K} := [\upsilon_1^f(\boldsymbol{X}_t), \ldots, \upsilon_K^f(\boldsymbol{X}_t)]$ and $\mathbf{L}^g \in \mathbb{R}^{N \times K} := [\upsilon_1^g(\boldsymbol{X}_t), \ldots, \upsilon_K^g(\boldsymbol{X}_t)]$ are the library of candidate functions $\{\upsilon_k(\boldsymbol{X}_t), k = 1, \ldots, K\}$. Here N is the length of sample path and the candidate functions can contain several functional forms such as polynomial, trigonometric, etc. Then, the drift and diffusion components can be expressed as linear combination of the library functions as,

$$\begin{aligned} f_i(\boldsymbol{X}_t, t) &= \theta_{i,1}^f \upsilon_1^f(\boldsymbol{X}_t) + \ldots + \theta_{i,k}^f \upsilon_k^f(\boldsymbol{X}_t) + \ldots + \theta_{i,K}^f \upsilon_K^f(\boldsymbol{X}_t) \\ \Gamma_{ij}(\boldsymbol{X}_t, t) &= \theta_{ij,1}^g \upsilon_1^g(\boldsymbol{X}_t) + \ldots + \theta_{ij,k}^g \upsilon_k^g(\boldsymbol{X}_t) + \ldots + \theta_{ij,K}^g \upsilon_K^g(\boldsymbol{X}_t) \end{aligned} \tag{4}$$

where, $\theta_{i,k}^f$ and $\theta_{ij,k}^g$ are the weights associated with the corresponding basis functions of drift and diffusion covariance components, respectively. In general, the libraries $\boldsymbol{L}^f$ and $\boldsymbol{L}^g$ can be different but one can use the same library for discovery of both drift and diffusion terms. The governing physics in the absence of input measurement is discovered in the proposed framework in terms of the SDEs. This requires identification of both the drift and diffusion components. Thus, one can construct two but separate sparse linear regression frameworks for the drift and diffusion terms and solve them independently. This yields,

$$\begin{aligned} \boldsymbol{Y}_i &= \mathbf{L}^f \boldsymbol{\theta}_i^f + \boldsymbol{\varepsilon}_i \\ \boldsymbol{Y}_{ij} &= \mathbf{L}^g \boldsymbol{\theta}_{ij}^g + \boldsymbol{\eta}_{ij} \end{aligned} \tag{5}$$

where, $\boldsymbol{Y}_i = f_i(\boldsymbol{X}_t, t)$ and $\boldsymbol{Y}_{ij} = \Gamma_{ij}(\boldsymbol{X}_t, t)$ are the target vectors of the sparse regression problem associated with i^{th}-drift component and $(ij)^{th}$-diffusion covariance term, respectively, and, $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\eta}_{ij}$ are the corresponding residual error vectors. The above equations are solved using sparse Bayesian linear regression. To understand, let us represent Eq. (5) as follows,

$$\boldsymbol{Y} = \mathbf{L}\boldsymbol{\theta} + \epsilon \tag{6}$$

where, $\boldsymbol{Y} \in \mathbb{R}^N$ denotes the N-dimensional target vector, $\mathbf{L}$ denotes the library of candidate functions, $\boldsymbol{\theta}$ is the weight vector, and $\epsilon \in \mathbb{R}^N$ is the residual error vector representing the model mismatch error. On application of the Bayes formula yields,

$$P(\theta|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|\boldsymbol{\theta})}{P(\boldsymbol{Y})} \tag{7}$$

Modeling the mismatch error ϵ as i.i.d Gaussian random variable with zero mean and variance σ^2, the likelihood function is written as,

$$\boldsymbol{Y}|\theta, \sigma^2 \sim \mathcal{N}\left(\mathbf{L}\boldsymbol{\theta}, \sigma^2 \mathbf{I}_{N\times N}\right) \tag{8}$$

where, $\mathbf{I}_{N\times N}$ denotes the $N \times N$ identity matrix. The aim of the sparse regression is to find a governing model which contains only few terms in the final model. This is obtained by imposing sparsity promoting priors on the weight vector. In this work, the spike and slab (SS) distributions are used for promotion of sparsity in the solution. This suggest to take into account the effect of parameters of SS-prior on the probability of the weights, this is taken into account by constructing a bigger Bayesian hierarchical model. The structure of the Bayesian hierarchical model for the SS-prior is shown in Fig. 1. For the classification

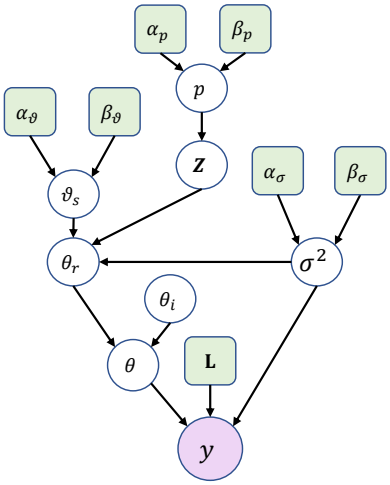

Figure 1. Hierarchical Bayesian network of the discontinuous spike and slab model for sparse linear regression. The variables in the green square boxes indicate the deterministic parameters and the variables in the white circles represents random variables. The term $\mathbf{L}$ represents the library of candidate functions, and the parameters α_ϑ, β_ϑ, α_p, β_p, α_σ, and β_σ indicates the hyperparameters of the priors of the hierarchical DSS model. Here p, ϑ_s and σ^2 scalar valued, and, the the variables $\boldsymbol{Z}$ and $\boldsymbol{\theta}$ are vector valued variables.

of the weights into spike and slab consider a latent variable Z such that it takes a value 1 if the weight fall into slab, otherwise, takes a value 0. Denoting $\boldsymbol{\theta}_r$ as the weight vector containing only those variables from $\boldsymbol{\theta}$ for which $Z_k = 1$, the DSS-prior is defined as[31,32],

$$p(\boldsymbol{\theta}|\boldsymbol{Z}) = p_{slab}(\theta_r) \prod_{k, Z_k=0} p_{spike}(\theta_k) \tag{9}$$

where, the spike and slab distributions are defined as, $p_{spike}(\theta_k) = \delta_0$ and $p_{slab}(\theta_r) = \mathcal{N}\left(\mathbf{0}, \sigma^2 \vartheta_s \mathbf{R}_{0,r}\right)$ with $\mathbf{R}_{0,r} = \mathbf{I}_{r\times r}$. The hyperparameters α_ϑ, β_ϑ, α_p, β_p, α_σ, and β_σ in the Fig. 1 are provided as a deterministic constants in the hierarchical model.

The random variables p_0, ϑ_s, σ^2 and Z_k are simulated as follows,

$$p(\vartheta_s) = IG(\alpha_\vartheta, \beta_\vartheta) \tag{10}$$
$$p(Z_k|p_0) = Bern(p_0); k = 1 \ldots K \tag{11}$$
$$p(p_0) = Beta\left(\alpha_p, \beta_p\right) \tag{12}$$
$$p\left(\sigma^2\right) = IG(\alpha_\sigma, \beta_\sigma) \tag{13}$$

The joint distribution of the random variables $\boldsymbol{\theta}$, $\boldsymbol{Z}$, ϑ_s, σ^2, $p_0|\boldsymbol{Y}$ are obtained from the Fig. 1 as,

$$\begin{aligned} p\left(\boldsymbol{\theta}, \boldsymbol{Z}, \vartheta_s, \sigma^2, p_0|\boldsymbol{Y}\right) &= \frac{p\left(\boldsymbol{Y}|\boldsymbol{\theta}, \sigma^2\right) p\left(\boldsymbol{\theta}|\boldsymbol{Z}, \vartheta_s, \sigma^2\right) p(\boldsymbol{Z}|p_0)\, p(\vartheta_s)\, p\left(\sigma^2\right) p(p_0)}{p(\boldsymbol{Y})} \\ &\propto p\left(\boldsymbol{Y}|\boldsymbol{\theta}, \sigma^2\right) p\left(\boldsymbol{\theta}|\boldsymbol{Z}, \vartheta_s, \sigma^2\right) p(\boldsymbol{Z}|p_0)\, p(\vartheta_s)\, p\left(\sigma^2\right) p(p_0) \end{aligned} \tag{14}$$

where, $p\left(\boldsymbol{\theta}, \boldsymbol{Z}, \vartheta_s, \sigma^2, p_0|\boldsymbol{Y}\right)$ denotes the joint distribution of the random variables, $p\left(\boldsymbol{Y}|\boldsymbol{\theta}, \sigma^2\right)$ denotes the likelihood function, $p\left(\boldsymbol{\theta}|\boldsymbol{Z}, \vartheta_s, \sigma^2\right)$ is the prior distribution for the weight vector $\boldsymbol{\theta}$, $p(\boldsymbol{Z}|p_0)$ is the prior distribution for the latent vector $\boldsymbol{Z}$, $p(\vartheta_s)$ is the prior distribution for the slab variance ϑ_s, $p\left(\sigma^2\right)$ is the prior distribution for the noise variance, $p(p_0)$ is the prior distribution for the success probability p_0 and $p(\boldsymbol{Y})$ is the marginal likelihood. The Gibbs sampling technique is used to draw the random samples from the above joint distribution[44]. For the Gibbs sampling the conditional distributions of the random variables are derived in Appendix B. For final selection of the model basis function marginal posterior probabilities are used, the details of which are briefly provided in *Methods* section.

The Gibbs sampler is initialised with the following initial values of the hyperparameters: $p_0^{(0)}$=0.1, $\vartheta^{(0)}$=10, and $\sigma^{2(0)}$ is set equal to the residual variance from ordinary least-squares regression. The initial vector of binary latent variables $\boldsymbol{Z}^{(0)}$ is computed by setting $\boldsymbol{Z}^{(0)} = [Z_1, \ldots, Z_K]$ to zero and then activating the components of $\boldsymbol{Z}$ that reduce the mean-squared error between the training data and the obtained model from ordinary least-squares. For this purpose a forward followed by a backward search algorithm is devised, where the backward search iterates through the activated components of initial latent vector in similar fashion to forward search. Given all the other parameters, the initial value of $\theta^{(0)}$ is obtained from the Eq. (14). For the commencement of the algorithm the deterministic prior parameters are set to the following values: a_p=0.1 and b_p=1 for the Beta prior on p_0, a_v=0.5 and b_v=0.5 for inverse-Gamma prior on slab variance, and, a_σ=10^4 and b_σ=10^4 for inverse-Gamma prior on measurement noise. A Markov chains with 3000 samples is used for Gibbs sampling. The first 1000 samples are discarded as burn-in, and the remaining 2000 samples are used for posterior computation.

For all the demonstrations in this work, the data are simulated using the Euler-Maruyama scheme at a frequency of 1000Hz using the parameters listed in the Table 1. The noise in the measurements is modeled as N-dimensional sequence of zero-mean Gaussian white noise with a standard deviation equal to 5% of the standard deviation of the simulated quantities. In this work, the dictionary $\mathbf{L} \in \mathbb{R}^{N \times K}$ is constructed from 5 types of mathematical functions, each function representing a mapping of the m-dimensional state vector $\boldsymbol{X} = \{X_1, X_2, \ldots X_m\}$:

$$\mathbf{L}(\boldsymbol{X}) = \begin{bmatrix} \mathbf{1} & P^1(\boldsymbol{X}) & P^2(\boldsymbol{X}) & \ldots & P^{\mathcal{P}}(\boldsymbol{X}) & \text{sgn}(\boldsymbol{X}) & |\boldsymbol{X}| & \boldsymbol{X} \otimes |\boldsymbol{X}| \end{bmatrix} \tag{15}$$

Here, $\mathbf{1} \in \mathbb{R}^N$ denotes the N-dimensional vector of 1, $P^{\mathcal{P}}(\boldsymbol{X}) \in \mathbb{R}^{N \times m}$ denotes the set of terms present in the multinomial expansion $(X_1 + X_2 + \ldots + X_m)^{\mathcal{P}}$, $sgn(\boldsymbol{X}) \in \mathbb{R}^{N \times m}$ represents the signum functions of the form $sgn(X_i)\ \forall i = 1 \ldots m$, $|\boldsymbol{X}| \in \mathbb{R}^{N \times m}$ denotes the absolute mapping of the states: $|X_i|\ \forall i = 1 \ldots m$. The tensor product term $\boldsymbol{X} \otimes |\boldsymbol{X}| \in \mathbb{R}^{N \times 2m}$ represents the set of functions: $X_i \otimes \left|X_j\right|\ \forall i, j = 1 \ldots m$. For the study, the length of the polynomial $\mathcal{P}$ is chosen as 6. The cardinality of the library is found as $K = (1 + |(X_1 + X_2 + \ldots + X_m)^n| + 4m)$, where, the number of terms in the multinomial expansion can be found as, $|(X_1 + X_2 + \ldots + X_m)^n| = {}^{n+m-1}C_{m-1}$. The complete architecture of the propose framework is illustrated in the Fig. 2.

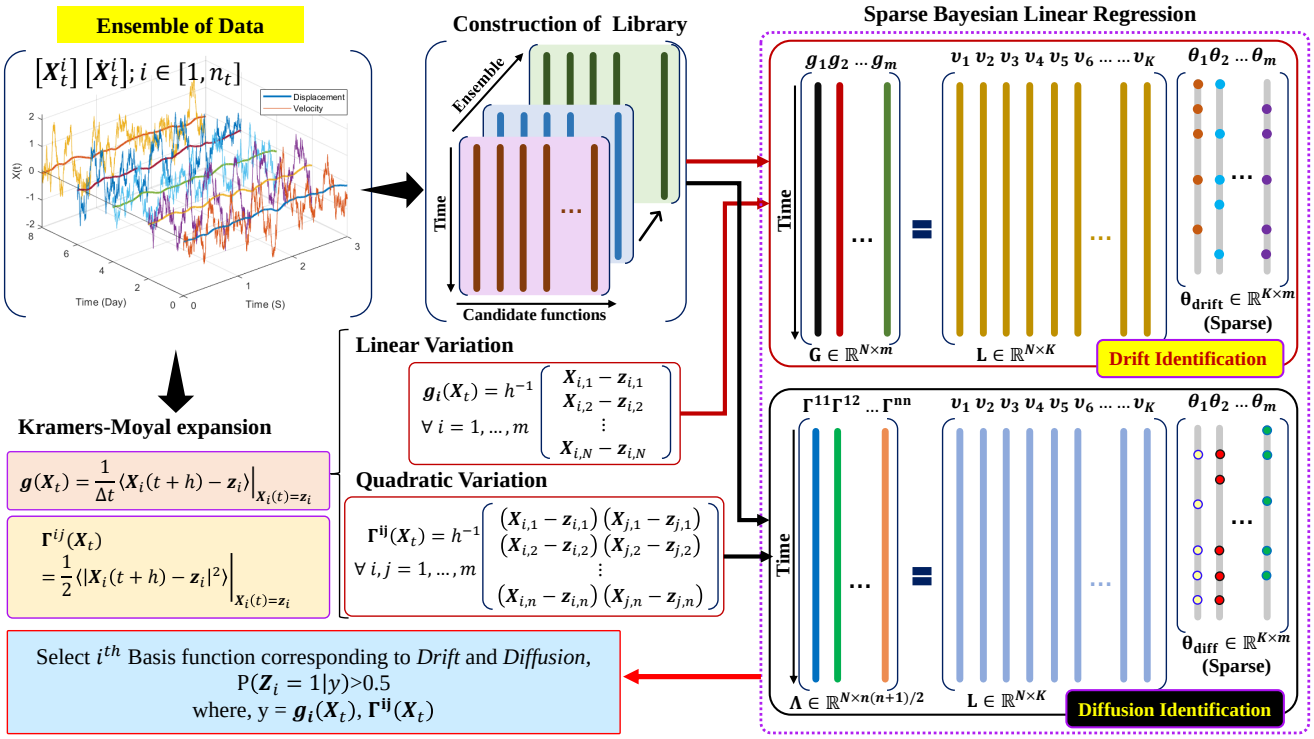

Figure 2. Schematic illustration of the proposed Bayesian-SDE framework for discovery of stochastic dynamical systems. The proposed framework integrates two key ideas for identification of the stochastic systems in terms of the stochastic differential equations. The first key step, is the construction of the target vector for the sparse linear regression. For this purpose the proposed framework utilises the Kramers-Moyal expansion to estimate the drift $g(\boldsymbol{X}_t, t)$ and diffusion terms $f(\boldsymbol{X}_t, t)$ in terms of the linear and quadratic variations of the measured sample paths. The linear variation: $\lim_{\Delta t \to 0} \langle X(t+\Delta t) - z \rangle; X(t) = z$ indicates sum of the increments of the sample paths and the quadratic variation: $\lim_{\Delta t \to 0} \left\langle |X(t+\Delta t) - z|^2 \right\rangle; X(t) = z$ indicates sum of the square of the increments of the sample paths in limiting sense. The second step, includes formulation of the sparse Bayesian linear regression framework for the above target vector. Given an ensemble of time history of data the framework obtains a sequence of libraries by evaluating the candidate basis functions $\{\upsilon_k(\mathbf{X}_t);\ k = 1, 2, \ldots K\}$ over the ensembles. Ensemble mean is taken over all the libraries and the averaged library $\mathbf{L} \in \mathbb{R}^{N \times K}$ is utilized for performing the sparse linear regression. The candidate function in the library are parameterized by a weight vector $\boldsymbol{\theta}$. To identify each of the drift terms in an m-dimensional diffusion process this regression are performed m-times. The diffusion terms are not directly recoverable but the discovery of diffusion terms are possible in terms of its covariances. This requires $m(m+1)/2$ number of regression for discovery of all the terms in a covariance matrix. The elements of the weight vector $\boldsymbol{\theta}$ is assigned a latent variable $Z_k; k = 1 \ldots K$ in order to classify the weights into spikes and slabs. Post the Bayesian regression the marginal posterior inclusion probabilities (PIP) $p(Z_k = 1|\boldsymbol{Y})$ is estimated. In the final model only the basis functions whose corresponding marginal PIP values are higher than a threshold value is included.

Discovery of governing physics of example problems

In this section, the efficacy and robustness of the proposed Bayesian physics discovery framework is observed on a variety of representative stochastic non-linear dynamical systems. The examples taken includes: (a) Black–Scholes SDE, (b) Duffing Van-der pol oscillator, and, (c) Two degree-of-freedom (DOF) base isolated shear building. For the equation discovery, it is assumed that the input forces $\dot{B}(t)$ are not measurable and only the noisy measurements of the displacements are available for physics discovery. The velocity component is obtained from the displacement vector through numerical differentiation such as fourth order central finite difference formula. Additionally a case study using Bouc-Wen oscillator is undertaken, where it is assumed that the hysteresis measurement is also not available. Therefore the proposed framework treats it as a partially observed system and tries to estimate the effect of the unobserved state without explicitly considering it in to the library of candidate functions. The results of the case studies undertaken in this work are presented in Fig. 3, 4 and Table 2. These results demonstrate sufficient accuracy in the discovered physics and robustness of the proposed framework to learn governing law from limited and noisy data.

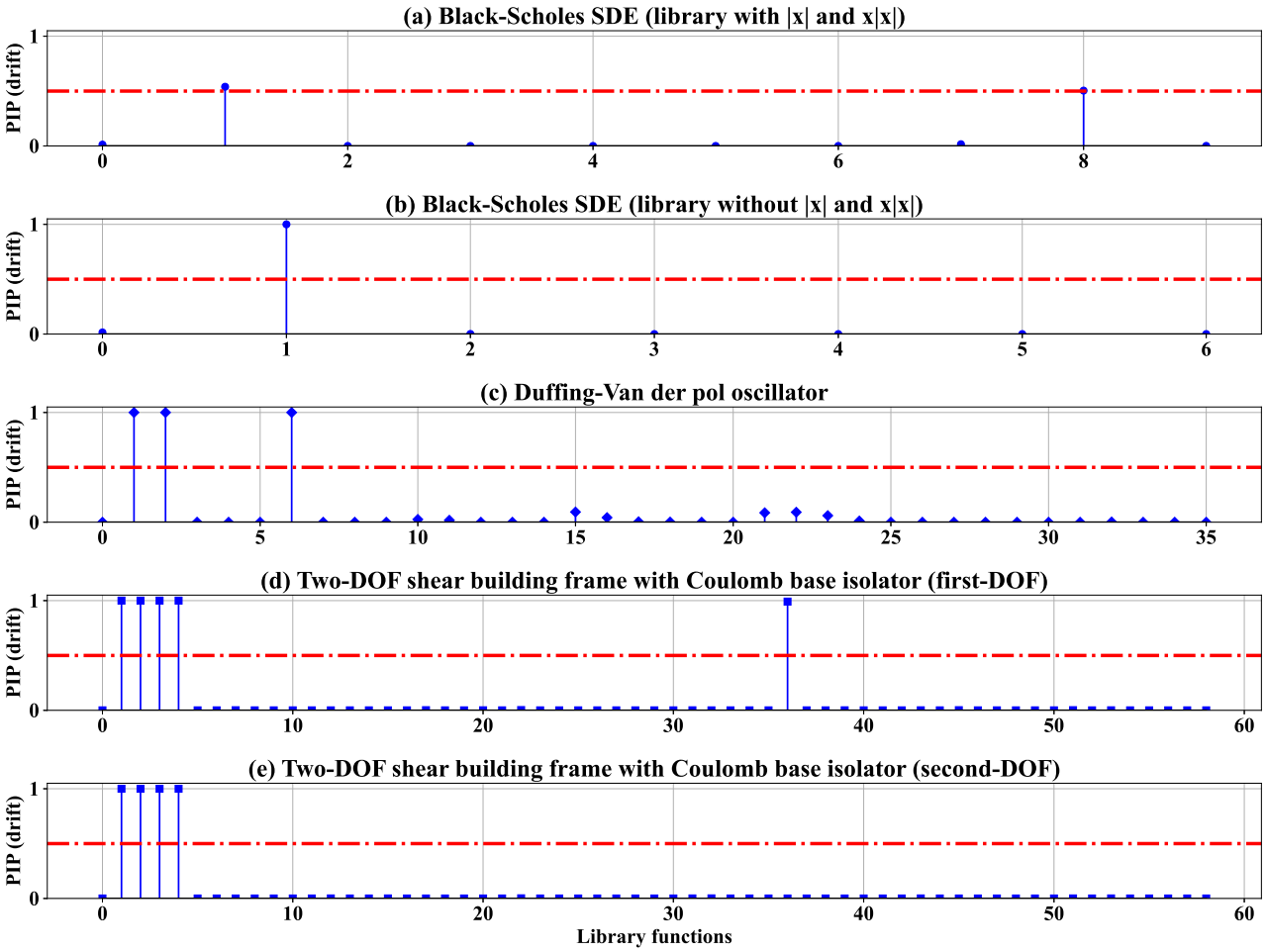

Figure 3. Discovery of the drift components of the governing physics from data corrupted with 5% white Gaussian noise based on marginal posterior inclusion probability (PIP), P($Z_k = 1|\boldsymbol{Y}$). (a) Black-Scholes SDE: the library $\mathbf{L} \in \mathbb{R}^{N\times 9}$ consists 9 candidate basis functions out of which the basis functions x and $|x|$ are identified based on the criteria, PIP>0.5. (b) Black-Scholes SDE: the library $\mathbf{L} \in \mathbb{R}^{N\times 6}$ is constructed as a collection of 6 basis functions. The correct drift component $X(t)$ is accurately identified with the almost sure probability $\mathrm{P}(Z_1 = 1) = 1$. (c) Duffing-Van der pol oscillator: the library $\mathbf{L} \in \mathbb{R}^{N\times 36}$ has 36 candidate basis functions. The drift component is correctly identified as, $\upsilon(1) = X_1$, $\upsilon(2) = X_2$ and $\upsilon(6) = X_1^3$. (d) Two-DOF shear building (drift component of first DOF): the library $\mathbf{L} \in \mathbb{R}^{N\times 58}$ consists a total of 58 candidate basis functions, out of which 5 basis are selected for discovery of first drift component as, $\upsilon(1) = Y_1$, $\upsilon(2) = Y_2$, $\upsilon(3) = Y_3$, $\upsilon(4) = Y_4$ and $\upsilon(36) = sgn(Y_2)$. (e) Two-DOF shear building (drift component of first DOF): from the library $\mathbf{L} \in \mathbb{R}^{N\times 58}$ the basis functions are discovered as, $\upsilon(1) = Y_1$, $\upsilon(2) = Y_2$, $\upsilon(3) = Y_3$, $\upsilon(4) = Y_4$. The basis functions for both the drift components are selected with almost full probability $\mathrm{P}(Z_k = 1|\boldsymbol{Y})$.

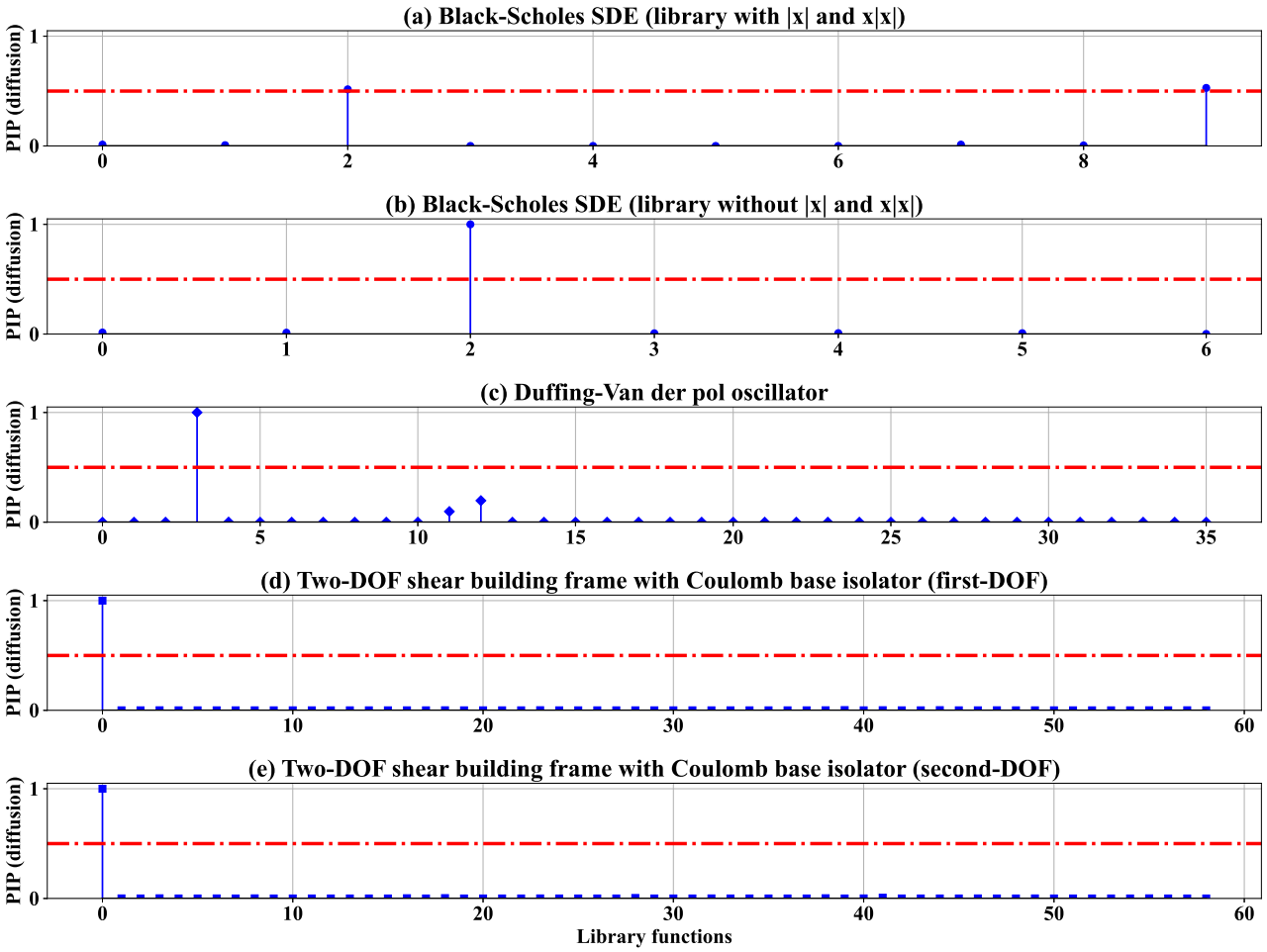

Figure 4. Discovery of the diffusion components of the governing physics from data corrupted with 5% white Gaussian noise based on marginal posterior inclusion probability (PIP), P($Z_k = 1|\boldsymbol{Y}$). (a) Black-Scholes SDE: the weight vector $\theta \in \mathbb{R}^9$ for the library $\mathbf{L} \in \mathbb{R}^{N\times 9}$ is shown. The covariance of the diffusion component is identified as a combination of the terms $X^2(t)$ and $X|X|$. (b) Black-Scholes SDE: the weight vector $\theta \in \mathbb{R}^6$ has 6 elements out of which the the correct covariance term $X^2(t)$ is discovered with almost sure probability P($Z_2 = 1$) = 1. (c) Duffing-Van der pol oscillator: there are 36 elements in the weight vector $\theta \in \mathbb{R}^{36}$. The diffusion term is discovered correctly as $X_1^2(t)$ with almost sure probability P($Z_2 = 1$) = 1. (d) Two-DOF shear building (drift component of first DOF): the weight vector $\theta \in \mathbb{R}^{58}$ has 58 elements corresponding to the 58 basis functions. The first diffusion component is accurately identified as $X_1^2(t)$ with almost sure probability P($Z_2 = 1$) = 1. (e) Two-DOF shear building (drift component of first DOF): the second diffusion component $X_2^2(t)$ is accurately identified with almost sure probability P($Z_2 = 1$) = 1.

Simulated systems	**Drift parameters**		**Diffusion parameters**
	Linear	**Non-linear**	
Black-Scholes	$\lambda = 2$	-	$\mu = 1$
Duffing-Van der pol	$m = 1, k = 2\times 10^3, c = 2$	$\alpha = 10^5$	$\sigma = 10$
2-DOF non-linear	$m_1 = m_2 = 1, k_1 = 4\times 10^3, k_2 = 2\times 10^3$ $c_1 = c_2 = 2$	$\mu = 1, g = 9.81$	$\sigma_1 = 10, \sigma_2 = 10$
Bouc-Wen	$m = 1, k = 1\times 10^4, c = 20, \lambda = 0.5$	$A_1 = A_2 = 0.5, A_3 = 1, \bar{n} = 3$	$\sigma = 2$

Table 1. Simulation parameters of the systems.

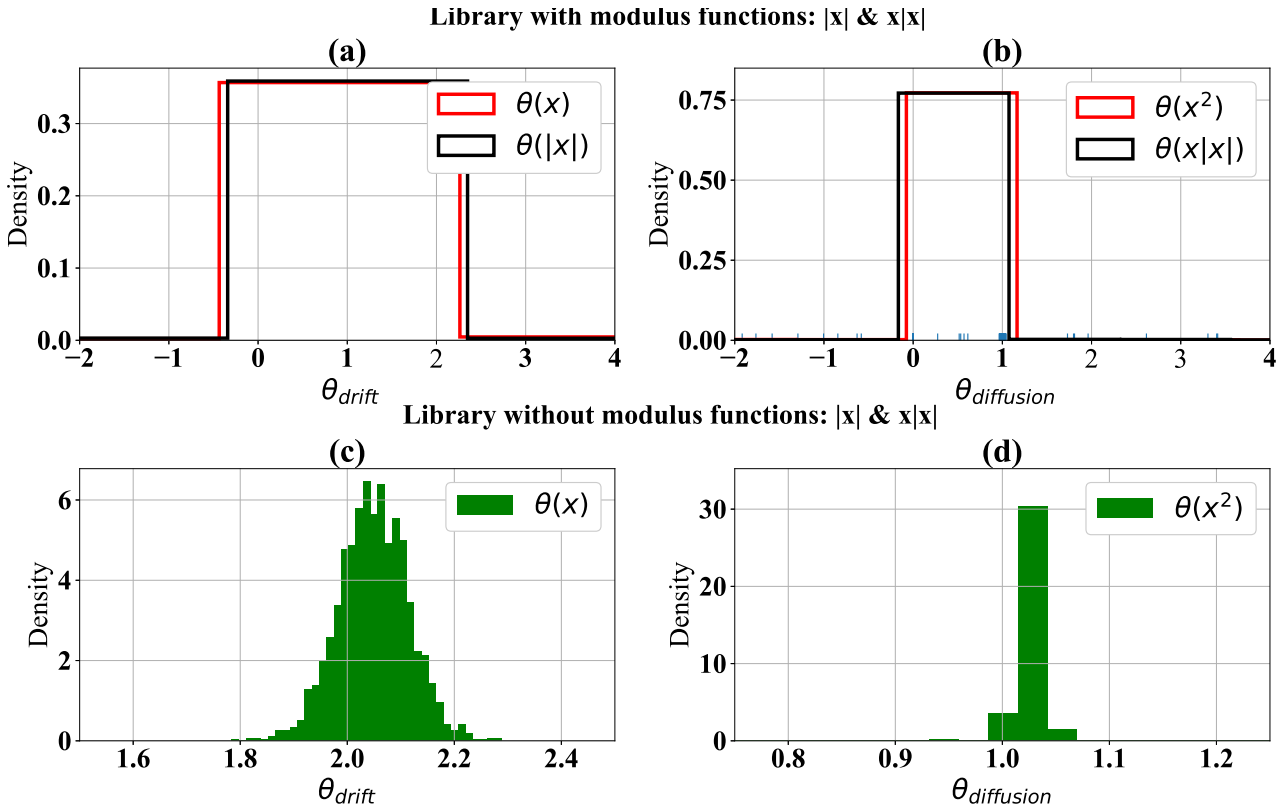

Figure 5. Black-Scholes SDE: posterior distributions of the weights of discovered basis functions. (a) The posterior distributions of θ_X and $\theta_{|X|}$. Posterior of both emulates the uniform distribution, whose means lie around the value 1. As a combination the basis θ_X and $\theta_{|X|}$ identifies the drift. (b) The posterior distributions of θ_{X^2} and $\theta_{X|X|}$. The posteriors emulate uniform distribution. The mean gives a value around 0.5. Thus, as a combination they discovered the diffusion. (c) The posterior of weight θ_X. The mean is obtained as 2.04, which exactly identifies the drift with almost sure probability $\mathrm{P}(Z_1 = 1) = 1$. (d) The posterior of weight θ_{X^2} whose mean provides the value of the parameters μ as 1.02.

Correct SDEs:	
Black-Schole:	$dX(t) = 2X(t)dt + X(t)dB(t)$
Duff.-Van der pol:	$dX_2(t) = -\left(1000X_1(t) + 2X_2(t) + 100000X_1^3(t)\right)dt + 10X(t)dB(t)$
2-DOF non-linear:	$\begin{cases} dY_2(t) = (-6000Y_1(t) - 4Y_2(t) - 9.81\mathrm{sgn}(Y_2(t)) + 2000Y_3(t) + 2Y_4(t))\,dt + 10dB_1(t) \\ dY_4(t) = (2000Y_1(t) + 2Y_2(t) - 2000Y_3(t) - 2Y_4(t))\,dt + 10dB_2(t) \end{cases}$
Identified SDEs:	
Black-Schole:	$\begin{cases} dX(t) = 2.05X(t)dt + 1.02X(t)dB(t) \\ dX(t) = (0.47X(t) + 0.53\,\lvert X(t)\rvert)\,dt + \left(1.09X^2(t) + 0.95X(t)\,\lvert X(t)\rvert\right)dB(t) \end{cases}$
Duff.-Van der pol:	$dX_2(t) = -\left(1000.02X_1(t) + 1.99X_2(t) + 99885.10X_1^3(t)\right)dt + 10.31X(t)dB(t)$
2-DOF non-linear:	$\begin{cases} dY_2(t) = (-6000.30Y_1(t) - 3.97Y_2(t) - 9.89\mathrm{sgn}(Y_2(t)) + 1999.73Y_3(t) + 2.09Y_4(t))\,dt \\ \qquad\qquad +10.03dB_1(t) \\ dY_4(t) = (2000.65Y_1(t) + 1.99Y_2(t) - 1999.06Y_3(t) - 1.99Y_4(t))\,dt + 10.11dB_2(t) \end{cases}$

Table 2. Summary of the results of Bayesian equation discovery of stochastic dynamical systems. The parameters of the identified systems denote the expected value of the weights after discarding the burn-in samples. Here, the first 1000 samples are taken as burn-in samples and therefore discarded for obtaining the final stationary distribution. In this work, the burn-in samples are taken as first 1000 MCMC samples.

Black–Scholes SDE

A Black–Scholes SDE (formulated as geometric Brownian motion) is a continuous-time stochastic process where the logarithm of the randomly varying quantity follows a Brownian motion. The Black-Scholes SDE is frequently used in stock market for modelling of the evolution of stock price of an underlying asset[34]. The Black–Scholes SDE has the drift $f(t, X_t) = \lambda X$ and diffusion $g(t, X) = \mu X$ with $\lambda > 0$ and $\mu > 0$ being the real constants. Towards this the Black–Scholes SDE is defined as a geometric Brownian motion as follows:

$$\frac{dX(t)}{X(t)} = \lambda dt + \mu dB(t) \quad X(t = t_0) = X_0; \quad t \in [0, T] \tag{16}$$

where $B = \{B(t); t \geq 0\}$ is the Brownian motion. The solution to the above SDE is attempted using Euler Maruyama (EM) with a uniform time step size Δt. The one-step approximation of the above SDE is then expressed by EM scheme as:

$$X(n+1) = X(n) + \lambda X(n)\Delta t + \mu X(n)\Delta B(n) \tag{17}$$

where, $\Delta t = t_{n+1} - t_n$ is the time increment and $\Delta B(n) = B(t_{n+1}) - B(t_n)$ is the Brownian increment. Identifying the diffusion term as, $g(\mathbf{X}_t, t) = \mu X(n)$, the variance of the diffusion is obtained: $\Gamma = \mu^2 X^2(n)$. The results of this case study are plotted in the Figures 3(a) and 4(a). Figure 3(a) depicts that there are two identified basis function X and $|X|$. However, from the Eq. (16) it can understood that only one function X should have been identified as the driving function of the Black-Scholes SDE. To understand this discrepancy one needs to consider that the evolution of the random variable $X(t)$ in the Black-Scholes SDE follows Geometric Brownian motion (GBM) and GBMs always assume positive values (e.g. real stock price). Since the characteristics of both the functions X and $|X|$ in this case are same, thus the algorithm identifies both the functions with almost equal probability. The similar phenomenon is observed in case of the identification of the diffusion term too. Figures 4(b) depicts almost equal contribution from the functions X^2 and $X|X|$, however, only the term X^2 should have been actually identified. Upon considering the corresponding parameter values of the basis functions X and $|X|$ for the drift component, and, X^2 and $X|X|$ for the diffusion component, from the Fig. 5(a) and (b), the Black-Scholes SDE in Eq. (16) can be easily simulated for any unseen environmental conditions.

To verify that this is not a limitation of the proposed scheme, this numerical demonstration is repeated in the Fig. 3(b) and 4(b) without considering the functions $|X|$ and $X|X|$ in the library. The parameters of the basis functions for the drift and diffusion components are portrayed in Fig. 5(c) and 5(d), respectively. The results clearly shows that the proposed scheme is able to identify the basis functions along with their associated parameters λ and μ without any significant error. As a consequence of the above case study, one can ask a question about what basis functions are to be considered in the library. One of the possibilities could be to visualize the the time series data and then take scientific and engineering judgements. For example, if the time history data show no zero crossing then one can opt for not-considering the functions which shares similar properties (for e.g. $|X|$ and X, and $X|X|$ and X^2). Besides one can also choose to consider due to the fact that the similar terms will have equal probability of occurrence and as a combination will demonstrate the observed phenomenon. However, for future prediction the model will be applicable to process depicting zero crossing only.

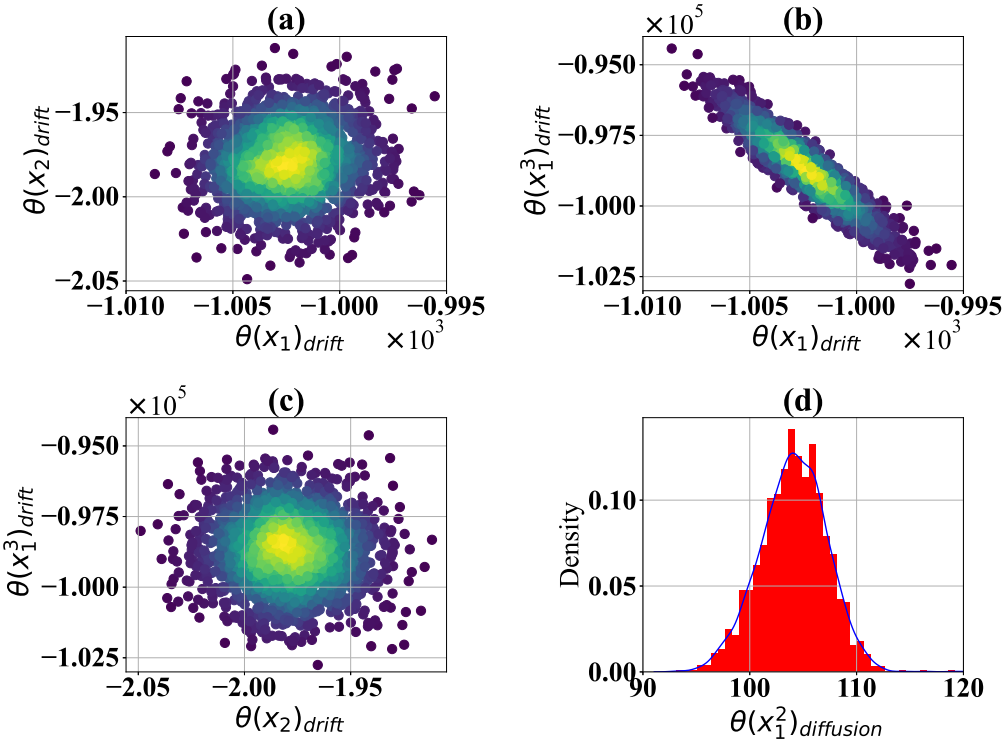

Figure 6. Duffing-Van der pol oscillator: Posterior distributions of the weights of the selected basis functions. (a) Joint posterior distribution of $\theta(X_1)$ and $\theta(X_2)$. The yellow region indicates the mean of the joint distribution. This weights $\theta(X_1)$ and $\theta(X_2)$ represents the parameters k and c. The expected values of the identified weights are, $\mathrm{E}(\theta(X_1)) = 1000.2$ and $\mathrm{E}(\theta(X_2)) = 1.99$. (b) Joint posterior distribution of $\theta(X_1)$ and $\theta(X_1^3)$. The expected values of the weights are $\mathrm{E}(\theta(X_1)) = 1000.2$ and $\mathrm{E}(\theta(X_1^3)) = 99885.1$. (c) Joint posterior distribution of $\theta(X_2)$ and $\theta(X_1^3)$. he expected values of the weights are $\mathrm{E}(\theta(X_2)) = 1.99$ and $\mathrm{E}(\theta(X_1^3)) = 99885.1$. (d) The histogram of the diffusion parameter. The mean value of the histogram is 106.47 which represents the term $g^2(\boldsymbol{X}_t, t)$. Upon conversion the parameters σ/m can be obtained as 10.29.

Duffing-Van der pol oscillator

The second order non-linear hardening Duffing-Van der pol oscillator is considered, in this section. The Duffing-Van der pol oscillator draws its physical relevance from the models in flow-induced structural vibration problems. The Duffing-Van der pol oscillator has a cubic dissipating force and it is driven by multiplicative white noise. Towards this, the equation of motion of the undertaken system is expressed as:

$$m\ddot{X}(t) + c\dot{X}(t) + kX(t) + \alpha X^3(t) = \sigma X(t)\dot{B}(t) \quad \boldsymbol{X}(t = t_0) = \boldsymbol{X}_0; \quad t \in [0, T] \tag{18}$$

where, m, c and k are the mass, damping and stiffness parameters of the oscillator. Here, α is a real-valued parameter associated with the cubic non-linearity, $\sigma \geq 0$ is the strength of the multiplicative white noise, and $B = \{B(t); t \geq 0\}$ is the Brownian motion. The time derivative of the Brownian $\dot{B}(t)$ here represents the white noise. With a statespace transformation $X = X_1$, and $\dot{X} = X_2$, where X_1 and X_2 represents the displacement and velocity, the corresponding two first order Itô-stochastic SDEs for the dynamical system are derived as:

$$\begin{bmatrix} dX_1(t) \\ dX_2(t) \end{bmatrix} = \begin{bmatrix} X_2(t) \\ \frac{-1}{m}\left(kX_1(t) + cX_2(t) + \alpha X_1^3(t)\right) \end{bmatrix} dt + \begin{bmatrix} 0 \\ \frac{\sigma}{m}X(t) \end{bmatrix} dB(t) \tag{19}$$

In the above equation it be noticed that the first equation provides only information of the velocity component, thus not considered for discovery of equation. For estimating the drift and diffusion terms only the evolution of second variable $X_2(t)$ is used to construct the target linear and quadratic variation vectors from the sample paths using the Kramers-Moyal formulae. Here, the variance of the diffusion term is identified as, $\Gamma = (\sigma^2 X^2(t))/m^2$. The results for the basis function selection are displayed in Fig. 3(c) and 4(c), respectively, while their associated parameters are plotted in Fig. 6. From the Fig. 3(c) and 4(c),

it is evident that the proposed framework is able to accurately discover the basis functions for the drift and diffusion terms as, $\left\{X_1(t), X_2(t), X_1^3(t)\right\}$ and $\{X_1(t), X_2(t)\}$, respectively. Afterwards, using the reverse statespace transformation the governing physics in terms of the second order differential equation can be easily obtained.

The pairwise joint posterior distributions of weights associated with the discovered basis functions are presented in Fig. 6. In the subplots (a), (b) and (c), the weights of discovered drift functions are shown. It is evident in this figures that the actual parameters associated with the drift component of DVP oscillator as listed in Table 1 is correctly identified with very small relative error. The subplot (d) depicts the posterior distribution of the identified diffusion basis function. The mean value of the posterior distribution represents the term $g^2(\boldsymbol{X}_t, t)$, which in this case can be noticed as $g(\boldsymbol{X}_t, t) = X(t)$. The mean value here is obtained as approximately 106.47. By performing the square root operation over the mean, one can approximately get the diffusion value $\sigma = 10.31$. This results, verifies that the proposed scheme can be successfully applied to discover the governing physics of non-linear oscillators when subjected to parametric excitation effectively.

Two-DOF base isolated Shear Building

The responses of civil engineering structures often exhibit the characteristics of stochastic non-linear dynamical systems due to the random external excitation. For the purpose of dynamical analysis the building structures are dominantly represented through a spring-mass-dashpot model. Here, a simple 2-DOF base isolated structure idealised as spring-mass-dashpot is considered. The system is a linear base isolated structure where the structure is connected to the foundation through a Coulomb friction-type base isolator. Under this considerations, the governing equation of motion of the system is written as,

$$\begin{aligned} &m_1\ddot{X}_1(t) + c_1\dot{X}_1(t) + \mu m_1 g \mathrm{sgn}\left(\dot{X}_1(t)\right) + k_1 X_1(t) + c_2\left(\dot{X}_1(t) - \dot{X}_2(t)\right) + k_2\left(X_1(t) - X_2(t)\right) = \sigma_1 \dot{B}_1(t) \\ &m_2\ddot{X}_2(t) + c_2\left(\dot{X}_2(t) - \dot{X}_1(t)\right) + k_2\left(X_2(t) - X_1(t)\right) = \sigma_2 \dot{B}_2(t) \\ &\boldsymbol{X}(t = t_0) = \boldsymbol{X}_0; \quad t \in [0, T] \end{aligned} \tag{20}$$

where, $\mu m_1 g \mathrm{sgn}(\dot{X}_1(t))$ is the Coulomb friction force arising due to the sliding of bearings in Coulomb oscillator, and, $\{m_i, i = 1,2\}$, $\{c_i, i = 1,2\}$ and $\{k_i, i = 1,2\}$ are the mass, damping and stiffness of i^{th}-DOF. Here, $\sigma_1 \geq 0$ and $\sigma_2 \geq 0$ are the strength of the white noise $\dot{B}_1(t)$ and $\dot{B}_2(t)$, where $B_1(t)$ and $B_2(t)$ are the independent Brownian motion. A statespace transformation is considered for the shear building as, $X_1 = Y_1$, $\dot{X}_1 = Y_2$, $X_2 = Y_3$, and $\dot{X}_2 = Y_4$, where Y_1 and Y_3 denotes the displacements, and, Y_2 and Y_4 denotes the velocities of the two DOF. By representing the complete state vector as $\boldsymbol{Y} = [Y_1, Y_2, Y_3, Y_4]$ the first order Itô-stochastic SDEs are obtained as follows:

$$\begin{bmatrix} dY_1(t) \\ dY_2(t) \\ dY_3(t) \\ dY_4(t) \end{bmatrix} = \begin{bmatrix} Y_2(t) \\ -\frac{k_1}{m_1}Y_1(t) - \frac{c_1}{m_1}Y_2(t) - \mu g \mathrm{sgn}(Y_2(t)) - \frac{k_2}{m_1}(Y_1(t) - Y_3(t)) - \frac{c_2}{m_1}(Y_2(t) - Y_4(t)) \\ Y_4(t) \\ -\frac{k_2}{m_2}(Y_3(t) - Y_1(t)) - \frac{c_2}{m_2}(Y_4(t) - Y_2(t)) \end{bmatrix} dt + \begin{bmatrix} 0 & 0 \\ \frac{\sigma_1}{m_1} & 0 \\ 0 & 0 \\ 0 & \frac{\sigma_2}{m_2} \end{bmatrix} \begin{bmatrix} dB_1(t) \\ dB_2(t) \end{bmatrix} \tag{21}$$

For the interest of the proposed scheme, the covariance of the diffusion matrix is obtained in Eq. (22). In this case study, the equation discovery involves identification of two drift and two covariance terms (one for each of the DOFs). Following to the case study of Duffing-Van der pol oscillator, the target linear variation vector for discovery of i^{th}; $\forall i = 1, \ldots, 2$ drift term needs to be constructed from the velocity component of the corresponding DOF. For discovery of the $(ij)^{th}$; $\forall i, j = 1, \ldots, 2$ component of the covariance matrix, the quadratic variation matrix is constructed from the pointwise product of i^{th} and j^{th} response vectors.

$$\Gamma = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{\sigma_1^2}{m_1^2} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\sigma_2^2}{m_2^2} \end{bmatrix} \tag{22}$$

The results for the identification of corresponding basis functions for both the drift terms are presented in Fig. 3(d) and (e). From the Fig. 3(d), it can be observed that the identified basis functions are $\{Y_1(t), Y_2(t), Y_3(t), Y_4(t), sgn(Y_2(t))\}$, which correctly matches with the terms in the equation of motion of first DOF. Similarly, from Fig. 3(e) it can be verified that the proposed scheme has correctly identified the basis functions of the second drift component as $\{Y_1(t), Y_2(t), Y_3(t), Y_4(t)\}$. Figure 7 and 8, shows the pairwise joint posterior distributions of the weights associated with identified first and second drift basis functions,

respectively. From these figures it is evident that the proposed scheme is able to identify the parameters of the basis functions as listed in Table 1 with sufficient accuracy. In order to find the explicit values of system stiffness and damping parameters, the parameters of second DOF can be identified first from the Figures 7(h), 7(e) and 7(f). Thereafter, the parameters of first DOF can be extracted easily by following the relations given in Eq. (21).

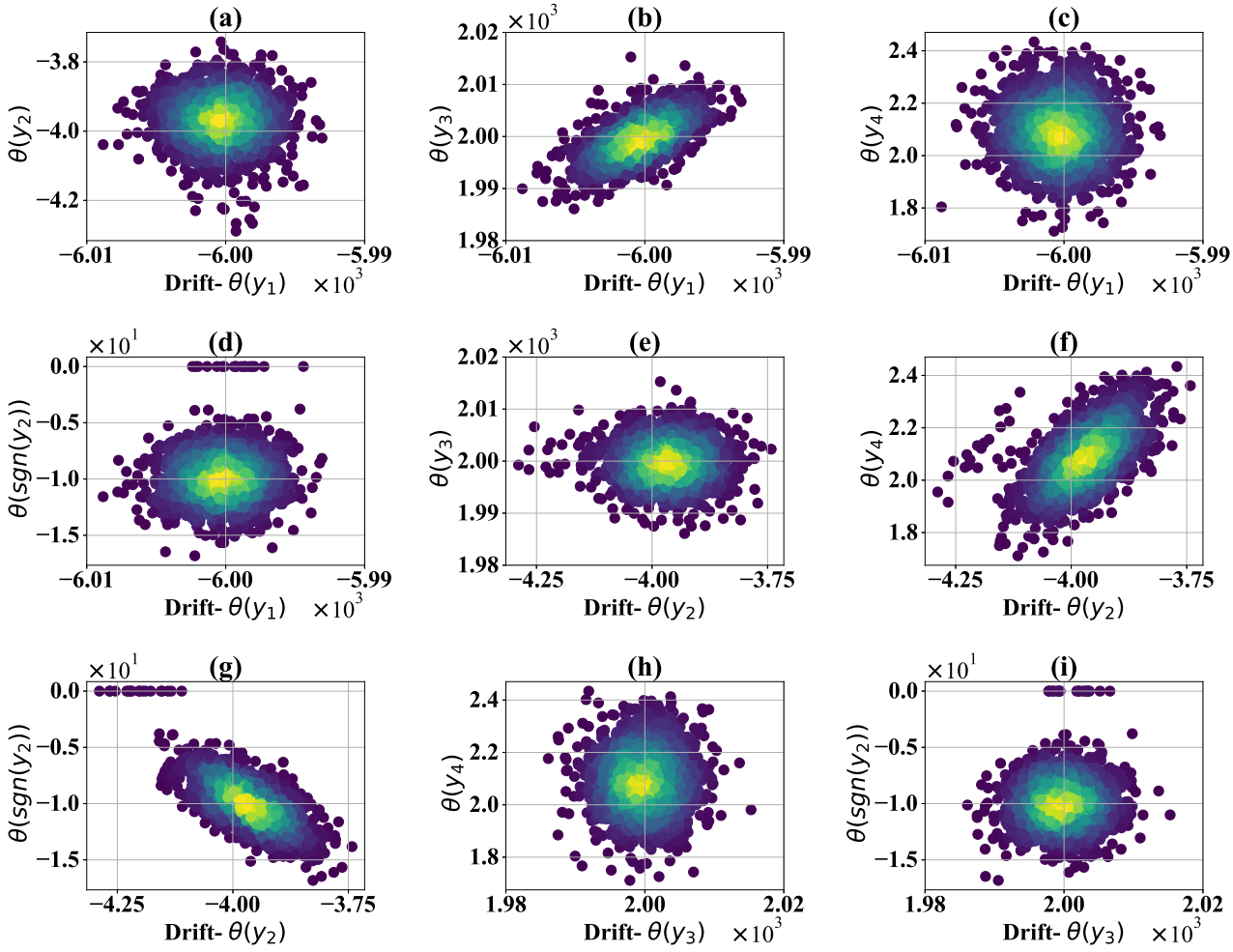

Figure 7. Non-linear two-DOF oscillator (first drift component): Pairwise joint posterior distributions of the parameters θ_k of the selected basis functions. (a) Joint posterior of the weights $\theta(Y_1)$ and $\theta(Y_2)$. The yellow color denotes the mean value of the joint distribution. The expected values of the parameters are identified as E($\theta(Y_1)$)= -6000.30 and E($\theta(Y_2)$)= -3.97. (b) Joint posterior of the weights θ_{Y_1} and θ_{Y_3}. The expected values of the parameters are identified as E($\theta(Y_1)$)= -6000.30 and E($\theta(Y_3)$)= 1999.73. (c) Joint posterior of the weights θ_{Y_1} and θ_{υ_1}. The expected values of the parameters are identified as E($\theta(Y_1)$)= -6000.30 and E($\theta(Y_4)$)= 2.09. (d) Joint posterior of the weights θ_{Y_1} and $\theta_{sgn(Y_2)}$. The expected values of the parameters are identified as E($\theta(Y_1)$)= -6000.30 and E($\theta(sgn(Y_2))$)= -9.89. (e) Joint posterior of the weights θ_{Y_2} and θ_{Y_3}. The expected values of the parameters are identified as E($\theta(Y_2)$)= -3.97 and E($\theta(Y_3)$)= 1999.73. (f) Joint posterior of the weights θ_{Y_2} and θ_{Y_4}. The expected values of the parameters are identified as E($\theta(Y_2)$)= -3.97 and E($\theta(Y_4)$)= 2.09. (g) Joint posterior of the weights θ_{Y_2} and $\theta_{sgn(Y_2)}$. The expected values of the parameters are identified as E($\theta(Y_2)$)= 1999.73 and E($\theta(sgn(Y_2))$)= -9.89. (h) Joint posterior of the weights θ_{Y_3} and θ_{Y_4}. The expected values of the parameters are identified as E($\theta(Y_3)$)= 1999.73 and E($\theta(Y_4)$)= 2.09. (i) Joint posterior of the weights θ_{Y_3} and $\theta_{sgn(Y_2)}$. The expected values of the parameters are identified as E($\theta(Y_3)$)= 1999.73 and E($\theta(sgn(Y_2))$)= -9.89. The weights $\{\theta_{Y_1},\theta_{Y_2},\theta_{Y_3},\theta_{Y_4},\theta_{sgn(Y_2)}\}$ corresponds to the basis functions $\{\upsilon_1,\upsilon_2,\upsilon_3,\upsilon_4,\upsilon_36\}$. They represents the relative value of the parameters $\{k_1,k_2,k_3,k_4,\mu mg\}$. The negative values arises due to the coupling of the systems states.

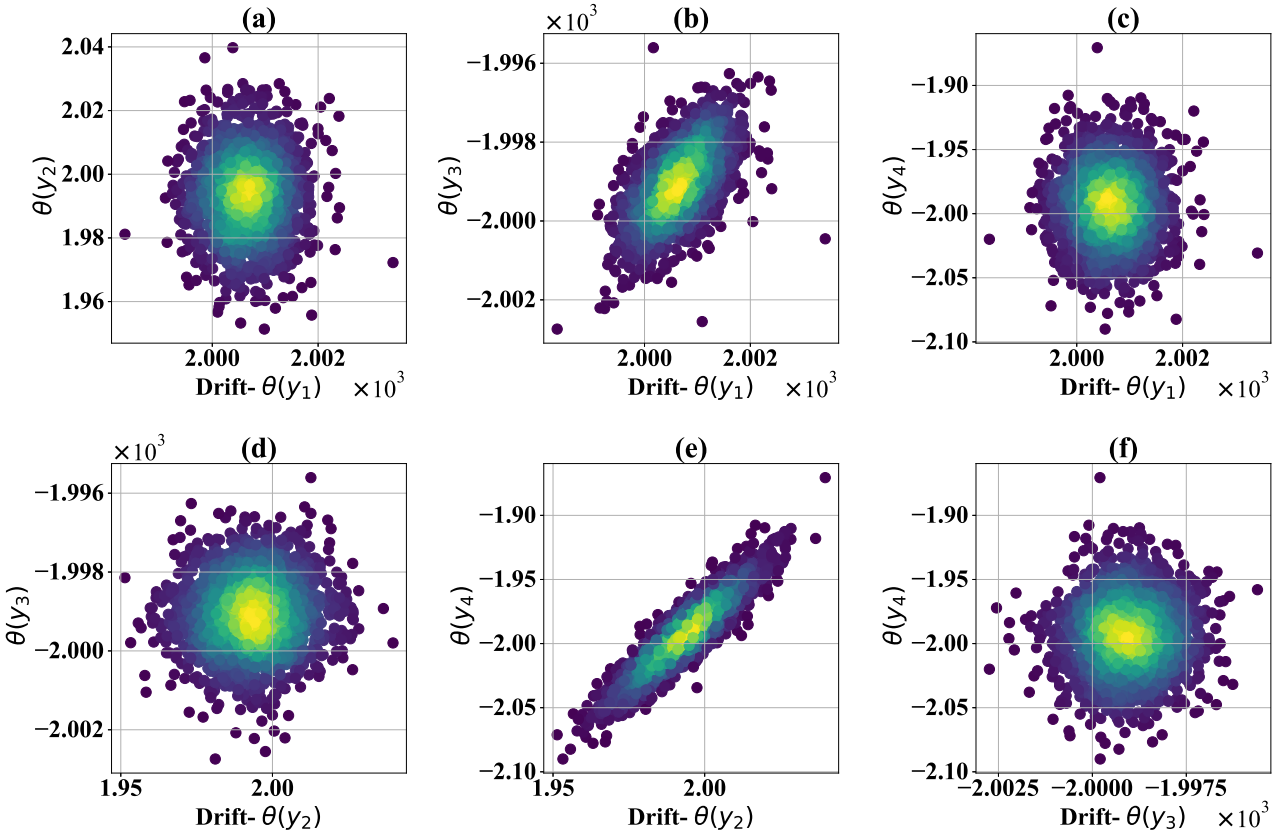

Figure 8. Non-linear two-DOF oscillator (second drift component): Pairwise joint posterior distributions of the parameters θ_k of the selected basis functions. (a) Joint posterior of the weights $\theta(Y_1)$ and $\theta(Y_2)$ corresponding to the basis functions υ_1 and υ_2. The yellow color denotes the mean value of the joint distribution. The expected values of the parameters are identified as E($\theta(Y_1)$)= 2000.65 and E($\theta(Y_2)$)= 1.99. (b) Joint posterior of the weights θ_{Y_1} and θ_{Y_3} corresponding to the basis functions υ_1 and υ_3. The expected values of the parameters are identified as E($\theta(Y_1)$)= 2000.65 and E($\theta(Y_3)$)= -1999.06. (c) Joint posterior of the weights θ_{Y_1} and θ_{Y_4} corresponding to the basis functions υ_1 and υ_4. The expected values of the parameters are identified as E($\theta(Y_1)$)= 2000.65 and E($\theta(Y_4)$)= -1.99. (d) Joint posterior of the weights θ_{Y_2} and θ_{Y_3} corresponding to the basis functions υ_2 and υ_3. The expected values of the parameters are identified as E($\theta(Y_2)$)= 1.99 and E($\theta(Y_3)$)= -1999.06. (e) Joint posterior of the weights θ_{Y_2} and θ_{Y_4} corresponding to the basis functions υ_2 and υ_4. The expected values of the parameters are identified as E($\theta(Y_2)$)= 1.99 and E($\theta(Y_4)$)= -1.99. (f) Joint posterior of the weights θ_{Y_3} and θ_{Y_4} corresponding to the basis functions υ_3 and υ_4. The expected values of the parameters are identified as E($\theta(Y_3)$)= -1999.06 and E($\theta(Y_4)$)= -1.99. The weights $\{\theta_{Y_1},\theta_{Y_2},\theta_{Y_3},\theta_{Y_4}\}$ represents the relative value of the parameters $\{k_1,k_2,k_3,k_4\}$. The negative values arises due to the coupling of the systems states.

The identification results for the basis functions along with the posterior distribution of their parameters for both the diffusion terms are depicted in Fig. 9. Figures 9(a) and 9(b) presents the results of the first diffusion term. The identification results of the second diffusion term are shown in Figures 9(c) and 9(d). In both the cases, it is evident that the proposed scheme has correctly identified the basis function as 1. The posterior distributions the covariance terms, $\mathbf{\Gamma}^{11}$ and $\mathbf{\Gamma}^{22}$ is plotted in Fig. 9(b) and 9(d) whose mean values represents σ_1^2/m_1^2 and σ_2^2/m_2^2. The summery of the results are presented in Table 2.

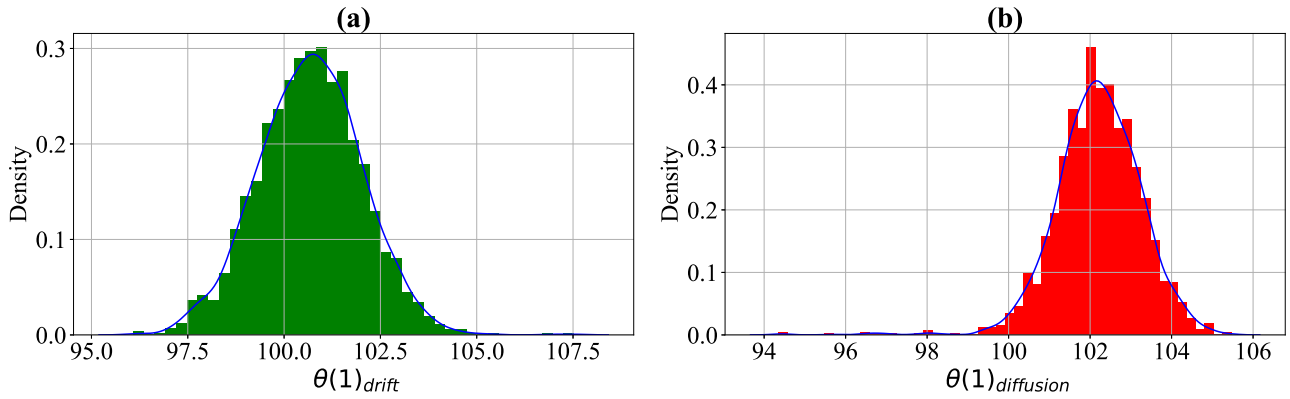

Figure 9. Two-DOF shear building: posterior distributions of the weights associated with the discovered basis functions of diffusion components. (a) First DOF: posterior distribution of the basis function θ_1. The mean of the distribution denotes the expected value of σ_1^2/m_1^2, which in this case is observed from the figure as approximately 101. (b) Second-DOF: posterior distribution of the basis function θ_1. The mean is observed as 102, which is equivalent to the term σ_2^2/m_2^2. By performing the square root operation over the mean values one can easily obtain the desired diffusion components.

Case study: Stochastic non-linear system with partially observed variables

In this section, a case study has been undertaken when a measurement of the states of a system is not possible. Let us consider a Bouc-Wen osillator where the system is described using three states namely, displacement and velocity of the main mass, and the hysteresis displacement associated with the isolator[35,36]. In this context, it can understood that the measurement of the hysteresis displacement is practically intractable and only the displacement and velocity states of the main system is available. Further, it can be appreciated that measuring the velocity state of a mechanical system using sensors are practically challenging task which leaves the only available option is displacement. Thus, a noble methodology should be able to identify the physics of the system using only the directly measurable states only, since the derivative quantities may not represent the motion of the original system appropriately. Towards this, a first step is taken using the proposed framework in which a base isolated sdof spring-mass-dashpot system is undertaken and only the displacement state is observed. Consider, the governing equation of motion of a sdof Bouc-Wen oscillator as[35],

$$m\ddot{X}(t) + c\dot{X}(t) + k\lambda X(t) + k(1-\lambda)Z(t) = \sigma\dot{B}(t) \qquad \boldsymbol{X}(t=t_0) = \boldsymbol{X}_0; \quad t \in [0,T] \tag{23}$$

where, m, c, and k are the mass, damping and stiffness of the system, $X(t)$ is the system state, λ is a factor that defines the participation of the elastic force F_e and hysteresis force F_h, $F_e(t) = k\lambda X(t)$, $F_h(t) = k(1-\lambda)Z(t)$, and $Z(t)$ is the hysteresis displacement. The evolution of the hysteresis parameter is defined using the non-linear differential equation:

$$Z(t) = -A_1 Z(t)\left|\dot{X}(t)\right||Z(t)|^{\bar{n}-1} - A_2\dot{X}(t)|Z(t)|^{\bar{n}} + A_3\dot{X}(t) \qquad \boldsymbol{Z}(t=t_0) = \boldsymbol{Z}_0; \quad t \in [0,T] \tag{24}$$

The positive exponential parameter n defines the smoothness of the transition from elastic to the post-elastic branch is smooth, and the parameters A_1, A_2 and A_3 control the size and shape of the hysteresis loop. For the simulation a statespace transformation is effected as $X = X_1$, $\dot{X} = X_2$, $Z = X_3$ to obtain the first order Itô-stochastic differential equations as,

$$\begin{bmatrix} dX_1(t) \\ dX_2(t) \\ dX_3(t) \end{bmatrix} = \begin{bmatrix} X_2(t) \\ \frac{-1}{m}\left(k\lambda X_1(t) + cX_2(t) + k(1-\lambda)X_3(t)\right) \\ -A_1X_3(t)|X_2(t)||X_3(t)|^{\bar{n}-1} - A_2X_2(t)|X_3(t)|^{\bar{n}} + A_3X_2(t) \end{bmatrix} dt + \begin{bmatrix} 0 \\ \frac{\sigma}{m} \\ 0 \end{bmatrix} dB(t) \tag{25}$$

One can easily simulate the system using the following Euler discretization,

$$\begin{bmatrix} X_1(n+1) \\ X_2(n+1) \\ X_3(n+1) \end{bmatrix} = \begin{bmatrix} X_1(n) \\ X_2(n) \\ X_3(n) \end{bmatrix} + \begin{bmatrix} X_2(t) \\ \frac{-1}{m}\left(k\lambda X_1(t) + cX_2(t) + k(1-\lambda)X_3(t)\right) \\ -A_1X_3(t)|X_2(t)||X_3(t)|^{\bar{n}-1} - A_2X_2(t)|X_3(t)|^{\bar{n}} + A_3X_2(t) \end{bmatrix} \Delta t + \begin{bmatrix} 0 \\ \frac{\sigma}{m} \\ 0 \end{bmatrix} \Delta B(t) \tag{26}$$

where, the drift and diffusion terms can be identified as,

$$f(t,\boldsymbol{X}_t)=\begin{bmatrix} X_2(t) \\ \frac{-1}{m}\left(k\lambda X_1(t)+cX_2(t)+k(1-\lambda)X_3(t)\right) \\ -A_1X_3(t)|X_2(t)||X_3(t)|^{\bar{n}-1}-A_2X_2(t)|X_3(t)|^{\bar{n}}+A_3X_2(t) \end{bmatrix};\quad g(t,\boldsymbol{X}_t)=\begin{bmatrix} 0 \\ \frac{\sigma}{m} \\ 0 \end{bmatrix} \tag{27}$$

Then, the covariance of the diffusion matrix is obtained as,

$$\Gamma=g(t,\mathbf{X}_t)g(t,\mathbf{X}_t)^T=\begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{\sigma^2}{m^2} & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{28}$$

It is straightforward to note that the key equation is, $dX_2(t)=\frac{-1}{m}\left(k\lambda X_1(t)+cX_2(t)+k(1-\lambda)X_3(t)\right)dt+\frac{\sigma}{m}dB(t)$, because the first equation in Eq. (25) does not convey anything about system motion and the variable $Z(t)$ is not observable. Thus, for the identification of the system only the displacement value $X_2(t)$ is considered as input to the algorithm. The system is simulated for $T=1$s at a frequency of 1000Hz using the parameter values, $m=1$, $c=20$, $k=100000$, $\lambda=0.5$, $A_1=0.5$, $A_2=0.5$, $A_3=1$, $\bar{n}=3$ and $\sigma_1=2$. Once, the displacement time history of the state $X_2(t)$ is obtained, the linear and quadratic variation of the displacement history is obtained using the formula in Eq. (34)[37]. The identified basis functions for the drift and diffusion terms are presented in Fig. 10. In sub-figure (a) it is evident that there are more number of basis functions than the input equation. It is straightforward to note that this extra basis functions arises to take care of the non-observable hysteresis parameter $Z(t)$. One can then identify the basis functions $\Theta_f(\mathbf{X})$ and $\Theta_g(\mathbf{X})$ for the drift and diffusion, respectively, that best represent the data as:

$$\begin{aligned} \Theta_f(\mathbf{X})&=\begin{bmatrix} 1 & X_1 & X_2 & X_1^2 & X_2^2 & X_1^3 & X_1^4 & sgn(X_1) & |X_1| & |X_2| & X_1|X_1| \end{bmatrix} \\ \Theta_g(\mathbf{X})&=\begin{bmatrix} 1 & X_1 & |X_1| \end{bmatrix} \end{aligned} \tag{29}$$

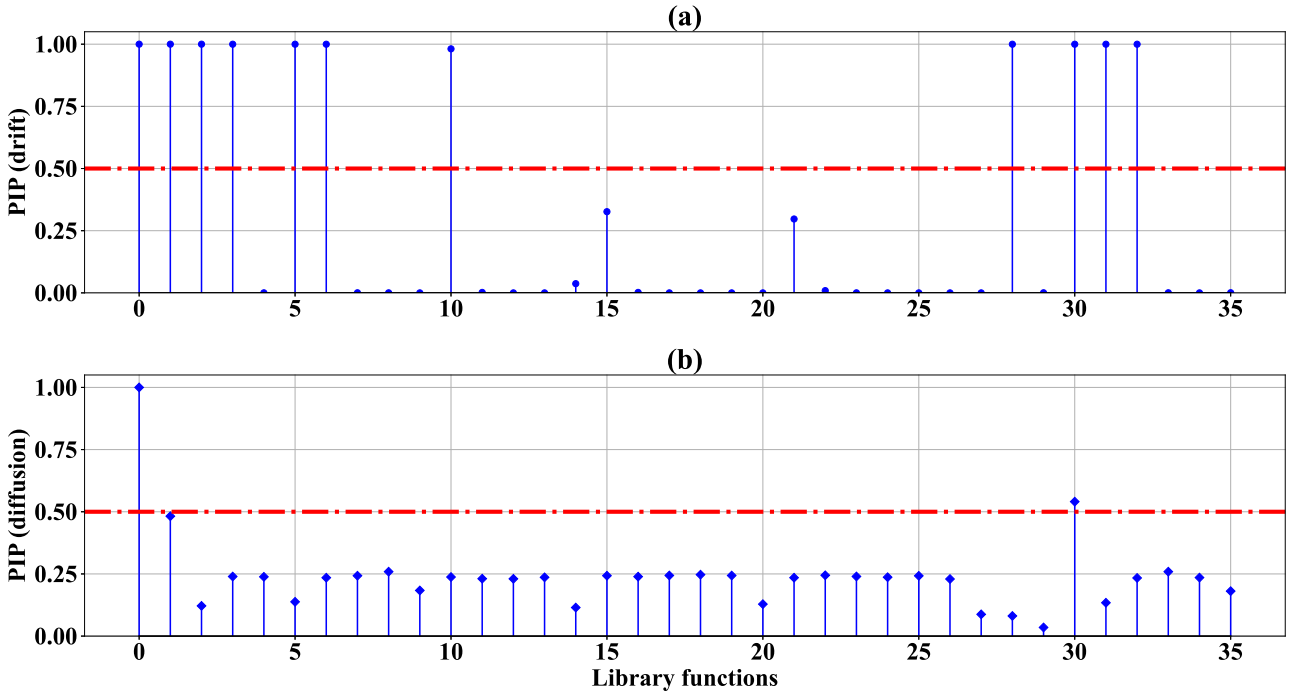

Figure 10. Bouc-Wen oscillator: Basis function selection for the Bouc-Wen base isolator system based on marginal posterior inclusion probability (PIP), $\mathbf{P}(Z_k=1|\boldsymbol{Y});k=1,\ldots,36$. (a) The identified basis functions for the drift component. The black horizontal axes represent the collection of 36 basis functions, and the red line represents marginal PIP > 0.5. The identified basis functions are $\{\upsilon(0),\upsilon(1),\upsilon(2),\upsilon(3),\upsilon(5),\upsilon(6),\upsilon(10),\upsilon(28),\upsilon(30),\upsilon(31),\upsilon(32)\}$. These basses corresponds to the functions $\{1,X_1,X_2,X_1^2,X_2^2,X_1^3,X_1^4,sgn(X_1),|X_1|,|X_2|,X_1|X_1|\}$. (b) The identified basis functions for the diffusion component. The identified function are $\{1,X_1,|X_1|\}$.

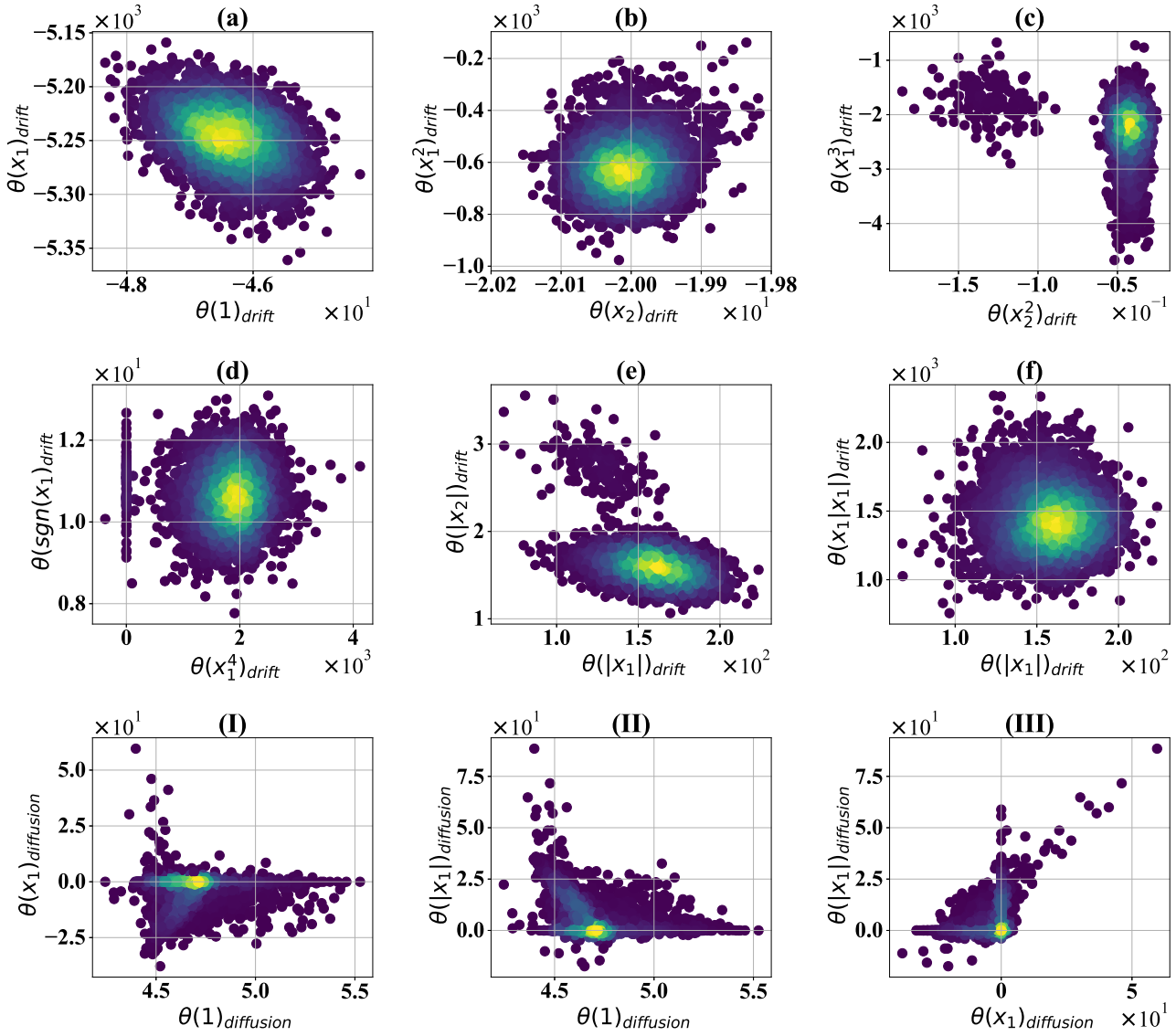

Figure 11. Bouc-Wen oscillator: Pairwise joint posterior distributions of the parameters θ_k of the selected basis functions. The subplots (a)→(f) presents the joint posterior distributions of drift parameters whereas the same is presented in the subplots (I)→(III). (a) Joint posterior distribution of $\theta(1)$ and $\theta(X_1)$. The parameters are identified as E($\theta(1)$)= -46.4 and E($\theta(X_1)$)= -5246. (b) Joint posterior distribution of $\theta(X_2)$ and $\theta(X_1^2)$. The expected value of the parameters are identified as E($\theta(X_2)$)= -20 and E($\theta(X_1^2)$)= -614. (c) Joint posterior distribution of $\theta(X_2^2)$ and $\theta(X_1^3)$. The parameters are identified as E($\theta(X_2^2)$)= -0.05 and E($\theta(X_1^3)$)= -2328. (d) Joint posterior distribution of $\theta(X_1^4)$ and $\theta(sgn(X_1))$. The expected value of the parameters are identified as E($\theta(X_1^4)$)= 1829 and E($\theta(sgn(X_1))$)= 10.64. (e) Joint posterior distribution of $\theta(|X_1|)$ and $\theta(|X_2|)$. The parameters are identified as E($\theta(|X_1|)$)= 159 and E($\theta(sgn(X_1))$)= 10.64. (f) Joint posterior distribution of $\theta(|X_1|)$ and $\theta(X_1|X_1|)$. The parameters are identified as E($\theta(|X_1|)$)= 159 and E($\theta(X_1|X_1|)$)= 1439. (I) The joint posterior of the weights θ_1 and θ_2 corresponding to the basis functions 1 and X_1 is shown. The yellow color denotes the mean values of the joint distribution. (II) The joint posterior of the weights θ_1 and θ_{30} corresponding to the basis functions 1 and $|X_1|$ is shown. (III) The joint posterior of the weights θ_2 and θ_30 corresponding to the basis functions X_1 and $|X_1|$ is shown.

The joint posterior distributions of the retained basis functions in Fig. 10, (a) and (b) are plotted in the Fig. 11. Based on the basis selection in Fig. 10 and from the joint posteriors of the parameter θ_i plotted in Fig. 11. the final drift and diffusion

fields for the Bouc-Wen system can identified as follows:

$$\begin{aligned} f(t,X) = \;& -46.4 - 5246X_1 - 20X_2 - 614X_1^2 - 0.05X_2^2 - 2328X_1^3 + 1829X_1^4 + 10.64\,sgn(X_1) + 159\,|X_1| \\ & + 1.625\,|X_2| + 1439X_1\,|X_1| \\ g(t,X) = \;& 4.65 - 4.872X_1 + 6.786\,|X_1| \end{aligned} \tag{30}$$

Correct SDEs:	$\begin{cases} dX_2(t) = -\left(5000X_1(t) + 20X_2(t) + 5000X_3(t)\right)dt + 2dB(t) \\ dX_3(t) = \left(-0.5X_3(t)\,\lvert X_2(t)\rvert\,\lvert X_3(t)\rvert^2 - 0.5X_2(t)\lvert X_3(t)\rvert^3 + X_2(t)\right)dt \end{cases}$
Identified SDEs:	$\begin{cases} dX_2(t) = \left(-46.4 - 5246X_1 - 20X_2 - 614X_1^2 - 2328X_1^3 + 1829X_1^4 + 10.64\,sgn(X_1) + 159\,\lvert X_1\rvert \right. \\ \qquad\qquad \left. + 1.63\,\lvert X_2\rvert + 1439X_1\,\lvert X_1\rvert\right)dt + \sqrt{4.65 - 3.79X_1 + 5.76\lvert X_1\rvert}\,dB(t) \end{cases}$

Table 3. Identification results of the Bayesian physics discovery for Bouc-Wen oscillator

For further treatment the term $0.05X_2^2$ is neglected due to its non-significant parameter value. With the above assumptions, finally the identified equation of motion of the Bouc-Wen system can be derived as in Eq. (31). In order to verify the fidelity of the identified system the learned the Bouc-Wen equation in Eq. (31) is simulated using the Euler Maruyama (EM) scheme with a time step of $\Delta t = 0.001$. However, in ahead of time the external excitation will not remain same. In the present framework this is modelled as Brownian force with different intensity. A Brownian noise intensity of $\sigma = 6$ is used to simulate the new unseen environment. The results of the comparison are plotted in Fig. 12.

$$\begin{aligned} \ddot{X} + 20\dot{X} + 5246X + 614X^2 + 2328X^3 - 1829X^4 - 10.64\,sgn(X) - 159\,|X| \\ - 1.625\,\left|\dot{X}\right| - 1439X\,|X| + 46.4 = 4.65 - 4.872X + 6.786\,|X| \end{aligned} \tag{31}$$

Figure 12 provides evidence that the proposed system correctly learns the SDEs corresponding to Bouc-Wen system with only 1 second of data (1000 samples). The learned system not only identifies the input responses of the system but also able to emulates the predicted response of the original system by learning the system from only 1 second of data. This presents the ability of the proposed scheme to not only learn a partially observed system but additionally, to be able predict the future in an unseen environment using the new basis functions. The summery of the cases studies undertaken in this work is presented in the Table 2.

Discussion

We propose a novel data-driven framework for discovery of governing physics of multi-dimensional non-linear stochastic dynamical systems from limited and noisy data. In many natural process the estimate of external input is intractable due to the limitations in the present measurement technologies. Further, measurement of all the systems are often not possible due to high operational cost. The proposed novel framework tries to emulate such situations by assuming that the input information is not required and only the measurements of noisy displacement data is available for physics discovery. The velocity state is obtained by performing the numerical differentiation on the displacement vector. The library of candidate functions are then constructed from the functional evaluated over obtained displacement and velocity states. To the knowledge of authors the discovery of governing physics of without the explicit knowledge of input excitation is not present in the literature, and, therefore a key novelty of the proposed novel framework. The absence of information about external force renders the physics discovery problem to a stochastic equation discovery problem. Towards this the proposed framework discovers the governing physics of the underlying stochastic process in terms of the stochastic differential equations (SDEs). The deterministic drift component of an SDE captures the dynamics of the underlying dynamical whereas the external stochastic force is identified in terms of the diffusion component.

The proposed framework employs the sparse Bayesian linear regression in conjunction with the Gibbs sampler to simultaneously obtain the basis functions of the model and their associated parameters. For this purpose the Kramers-Moyal expansion is utilised to express the drift and diffusion dynamics of an SDE in terms of the measured samples paths. The drift and diffusion vectors obtained from the sample paths are then used as a target vector in the sparse regression. Use of the Kramers-Moyal expansion within the purview of sparse Bayesian linear regression for discovery of an explainable governing physics is second novelty of the proposed framework. The unified framework possesses the ability of the Bayesian inference to update the probability of observing the correct basis function based on the new information. The ability to update the information is missing in recently published least-square based physics discovery schemes. This further introduces the natural elimination of the basis functions that do not present the observed data. In a noisy and unseen environment, one would be more interested in

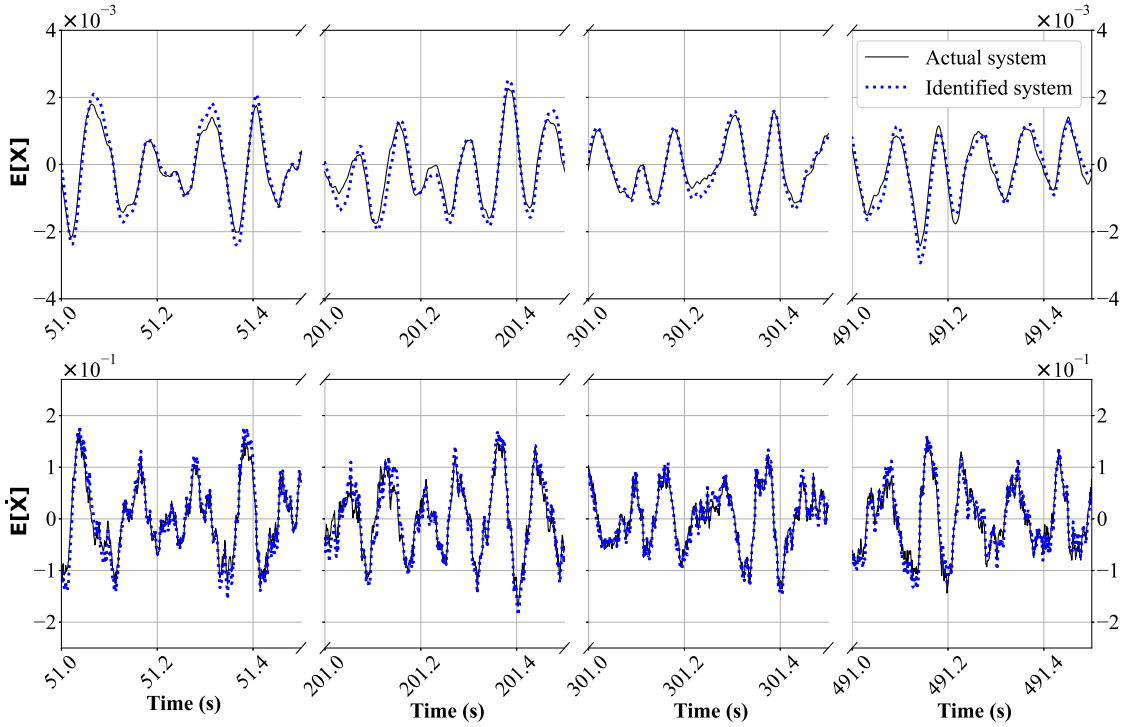

Figure 12. Bouc-Wen oscillator: Future prediction in an unseen scenario using the discovered equation in the absence of the hysteresis data. The governing equation of the Bouc-Wen system learned using 1 second of data sampled at 1000Hz. The prediction is performed for 500 seconds ahead of time using the discovered Bouc-Wen model. In the figure, the ensemble of the displacement and velocity prediction data over 500 seconds is presented. The simulation is performed using a different Brownian noise intensity of $\sigma = 6$, which is significantly higher than the training data. The dotted red line shows the prediction using the discovered equation and the solid blue line denotes the accurate simulated data. The predicted data almost accurately emulates the original system even though the measurements for all the states are not available.

dealing with the probability distribution of a random variable rather than a deterministic value. The proposed framework thus takes a leap in such situations and provides a probability distribution for each of the weights associated with corresponding basis functions. The obtained standard deviations of the distributions can then be used for defining the lower and uppers bounds on the estimates of a unknown parameters. Another aspect of the proposed framework is its advantages over the Neural network based grey models, where the explicit expression of the discovered governing physics is not known. In cases of partially observed processes, the proposed framework seems to provide an explainable equation for the governing physics. Although the basis functions in the obtain equation is significantly different from the one determined from first principal laws, the obtained model is able to predict the distant future without significant error. To the knowledge of authors, the discovery of physics in a partially observed process from the purview of probability theory is one of the key contributions of the proposed novel framework.

Despite of the advantages of the proposed framework over presently available physics discovery techniques there are certain issues that requires special attention for accurate identification of the governing physics. The first issue is the judicious selection of the candidate basis function. The presence of basis functions with high correlation may sometimes yield a physical law different from the actual physics. This is evident from the identification example on the Black-Schole SDE. The second issue is the associated computational cost. Although the computational cost of the proposed scheme is significantly less than the time required to train its neural network alternatives, further improvements can be made. One such alternative is to use the marginal standard deviation criteria, instead of taking the mean of the latent vector. Another improvement in terms of computational efficiency can be made by devising an appropriate filter to process out the signal noise synchronously with the sparse Bayesian linear regression. The third issue is related to the quality of the data. The low-fidelity data are less expensive to obtain but the representation of the physical model in the data poor due to the presence of noise. On the other hand, the

the the high-fidelity data are highly accurate but expensive to obtain. The proposed framework in its present state is not well equipped to make use of both the low and high-fidelity data. Thus, to further enhance the fidelity of the proposed framework for field applications a more robust algorithm needs to be formulated. The use of low-fidelity data ensure low operational cost while the high-fidelity data will ensure the accuracy of the discovered physics. Lastly, the proposed framework leverages the Kramers-Moyal expansion to express the drift and diffusion components of an SDE in terms of the first and second order moments of the sample paths. The moments are approximated in a limiting sense by making use of the absolute and quadratic identities of Brownian motion. In this context, future improvements in the proposed framework can be made by exploiting the higher order moments to retain the higher order terms from Stochastic-Taylor expansions of the drift and diffusion terms. Overall, the proposed framework presented a novel methodology for discovering the governing physics of a stochastic process whenever there is not possible to measure the input force data. The learned equations have shown accurate identification of the physical process and demonstrated good prediction ability in distant future.

Methods

The structure of the proposed Bayesian physics discovery framework is presented in the Fig. 2. The proposed framework treats the state measurements from the sensors as stochastic process and applies Kramers-Moyal formula to obtain the target vector for sparse linear regression. In the sparse regression, the library is obtained by taking ensemble mean over all the libraries constructed from the collected data set. Then the proposed framework constructs two sparse regression problem using these libraries and target vectors for identification of the governing physics, which are independent and can be solved simultaneously. The discovery of governing physics here involves identification of the system dynamics and the volatility associated arising due to environmental uncertainty. The system dynamics and the volatility are represented in this framework as drift and diffusion components of an SDE. The methodology of the proposed framework is further illustrated though mathematical equations as follows. Let $\mathbf{L} \in \mathbb{R}^{N\times K}$ be the library of candidate functions, where, K is the total number of independent basis functions in the library and N is the sample length. Then the key idea of the sparse Bayesian linear regression is to select k basis functions from the total number of K candidate library functions such that $k \ll K$. If $\boldsymbol{Y} \in \mathbb{R}^N$ is the target vector then the linear combinations of the identified basis functions should represent the target vector $\boldsymbol{Y}$ in best possible manner. The coefficients of the linear coefficient is denoted by a weight vector $\boldsymbol{\theta}$. For the Bayesian model discovery, the condition $k \ll K$ is maintained by assigning certain type of prior distributions on the weight vector $\boldsymbol{\theta}$, such that they favour the sparsity in the solution. In this context, the SS-prior have shown high shrinkage property due to its sharp spike at zero and a diffused density spanned over a large range of possible parameter value. The Dirac-delta spike concentrates most of the probability mass at zero thus allowing most of the samples to take value zero. On the other hand the diffused tail distributes a small amount of probability mass over a large range of possible values allowing only very few samples with very high probability to escape the shrinkage. The alternate flavours of the SS-prior constructed from various combination of candidate spike and slab functions are well documented in the literature[32,33,38]. In this work, the authors have considered the discontinuous spike and slab prior (DSS-prior) where the spike at zero is modelled as Dirac-delta function and the tail distribution as independent Student's-t distribution. Next, a sparse regression framework is formulated for stochastic differential equations.

Sparse learning of Stochastic differential equations

The natural dynamical systems are often not observed directly but through some projected space. The projected space contains all the states of a system which are directly observable. For instance, the second order dynamical systems are often expressed in terms of its displacement and velocity components. In cases of non-linear stochastic dynamical systems the stochastic differential equations provide an splendid approach for expressing the behavior of the underlying dynamical system. Stochastic differential equations (SDEs) arises naturally in non-linear dynamical systems subjected to stochastic excitation such as earthquake, wind force, wave force etc.[39,40]. An SDE is defined in term of its deterministic drift and the additive stochastic diffusion components where both the components are allowed to depend on time and systems states. Let $(\Omega, \mathcal{F}, P)$ be the probability space and $\{\mathcal{F}_t, 0 \le t \le T)\}$ be the natural filtration constructed from sub σ-algebras of $\mathcal{F}$. Consider the m-dimensional n-factor SDE driven by n-dimensional Brownian motion $\{\boldsymbol{B}_j(t), j = 1, \dots n\}$:

$$d\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{X}_t, t)\,dt + \sum_{j=1}^{n} \boldsymbol{g}_j(\boldsymbol{X}_t, t)\,d\mathbf{B}_j(t); \quad \boldsymbol{X}(t = t_0) = \boldsymbol{X}_0; \quad t \in [0, T] \tag{32}$$

where, $\boldsymbol{X}_t \in \mathbb{R}^m$ denotes the $\mathcal{F}_t$-measurable state vector, $\boldsymbol{f}(\boldsymbol{X}_t, t) : \mathbb{R}^m \mapsto \mathbb{R}^m$ is the drift vector, $\boldsymbol{g}(\boldsymbol{X}_t, t) : \mathbb{R}^m \mapsto \mathbb{R}^{m\times n}$ is the diffusion matrix and $\boldsymbol{B}_j(t) \in \mathbb{R}^n$ is the Brownian motion. In a compact matrix notation the SDE can be expressed as:

$$d\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{X}_t, t)\,dt + \boldsymbol{g}(\boldsymbol{X}_t, t)\,d\boldsymbol{B}(t); \quad \boldsymbol{X}(t = t_0) = \boldsymbol{X}_0; \quad t \in [0, T] \tag{33}$$

The solution to Eq. (33) can be attempted using various stochastic integration schemes[35,41–43]. However, in the interest of the present work one can try to find the probability distribution function (PDF) of the random variable $X(t)$. In general the

stochastic Brownian motion $B(t)$ is not a well defined mathematical function since it is not differential everywhere with respect to the process $X(t)$. This kind of non-differentiable functions have zero finite variation but non-vanishing quadratic variation. The Kramers-Moyal formula suggests that the drift and diffusion components of an SDE can be expressed in terms of the linear and quadratic variations of the sample time history as (the simplified detailed derivation is presented in A):

$$\begin{aligned} f_i(\boldsymbol{X}_t,t) &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}E\left[X_i(t+\Delta t)-\xi_i\right]\Big|_{x_k(t)=\xi_k} \quad \forall\, k=1,2,\dots N \\ \Gamma_{ij}(\boldsymbol{X}_t,t) &= \frac{1}{2}\lim_{\Delta t\to 0}\frac{1}{\Delta t}E\left[|X_i(t+\Delta t)-\xi_i|\left|X_j(t+\Delta t)-\Xi_j\right|\right]\Big|_{x_k(t)=\xi_k} \quad \forall\, k=1,2,\dots N \end{aligned} \tag{34}$$

where, $f_i(\boldsymbol{X}_t,t)$ is the i^{th} drift component and $\boldsymbol{\Gamma}_{ij}$ is the $(ij)^{th}$ component of the diffusion covaraince matrix $\boldsymbol{\Gamma}\in\mathbb{R}^{n\times n} := g(t,\boldsymbol{X}_t)g(t,\boldsymbol{X}_t)^T$. The discovery of an SDE requires identification of the drift and diffusion components in terms of the candidate basis functions from the library. Since the discovery problem demands the identified physics to have a interpretable form, sparse regression using Bayesian framework can be utilized in this context. Thus for the discovery of the drift and diffusion terms two independent linear regression problems are constructed next.

For the discovery of the i^{th} drift terms, the target vector is defined as, $\boldsymbol{Y}_i := \lim_{\Delta t\to 0}\frac{1}{\Delta t}E\left[X_i(t+\Delta t)-\xi_i\right]\Big|_{x_k(t)=\xi_k}$. Further, in a m-dimensional diffusion process each of the drift term can be expressed in terms of the basis functions as, $f_i(\boldsymbol{X}_t,t) = \sum_{k=1}^{K}\upsilon_k(\boldsymbol{X}_t)\theta_{i,k};\ i=1\dots m$ where, $\{\upsilon_k(\boldsymbol{X}_t);\ k=1,2,\dots K\}$ denotes the candidate basis functions and $\{\theta_{i,k};k=1\dots K\}$ denotes the weight vector corresponding to the i^{th} drift component. A large library is constructed from the candidate basis functions as, $\mathbf{L}=\{\upsilon_1(\boldsymbol{X}_t),\dots,\upsilon_K(\boldsymbol{X}_t)\}$, where, $\upsilon_k(\boldsymbol{X}_t)$ represents the various linear and non-linear mathematical functions evaluated on the system states. In a general setting, the identification of the drift terms culminates in the following regression problem.

$$\begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \dots & \boldsymbol{Y}_m \end{bmatrix} = \underbrace{\begin{bmatrix} \upsilon_1(X_{1,1}\dots X_{m,1}) & \upsilon_2(X_{1,1}\dots X_{m,1}) & \cdots & \upsilon_K(X_{1,1}\dots X_{m,1}) \\ \upsilon_1(X_{1,2}\dots X_{m,2}) & \upsilon_2(X_{1,2}\dots X_{m,2}) & \cdots & \upsilon_K(X_{1,2}\dots X_{m,2}) \\ \vdots & \vdots & \ddots & \vdots \\ \upsilon_1(X_{1,N}\dots X_{m,N}) & \upsilon_2(X_{1,N}\dots X_{m,N}) & \cdots & \upsilon_K(X_{1,N}\dots X_{m,N}) \end{bmatrix}}_{\mathbf{L}\in^{N\times K}} \underbrace{\begin{bmatrix} \theta_{1,1} & \theta_{2,1} & \cdots & \theta_{m,1} \\ \theta_{1,2} & \theta_{2,2} & \cdots & \theta_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1,k} & \theta_{2,k} & \cdots & \theta_{m,k} \end{bmatrix}}_{\theta\in^{K\times m}} \tag{35}$$

As a subset of the above regression one can solve the following equation and identify the drift components independently:

$$\boldsymbol{Y}_i = \mathbf{L}\boldsymbol{\theta}_i^f + \boldsymbol{\varepsilon}_i \tag{36}$$

where, the i^{th} target vector $\boldsymbol{Y}_i$ is constructed from the linear variations of the i^{th} sample path as: $\boldsymbol{Y}_i = [(X_{i,1}-\xi_{i,1})\dots(X_{i,N}-\xi_{i,N})]^T$ and the drift weight vector is $\theta_i^g = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \dots & \theta_{1,K} \end{bmatrix}^T$. The straightforward application of the Gibbs sampling mentioned in the Eqs. (47) → (52) can be performed to discover the drift components of a system of SDEs.

The diffusion components does not have finite variations but are bounded by their quadratic variations. Thus the diffusion components unlike the drifts of an SDE are recoverable only through its covariation terms. In the formulation of the sparse regression problem for diffusion components the target vector is defined as, $\boldsymbol{Y}_{ij} := \frac{1}{2}\lim_{\Delta t\to 0}\frac{1}{\Delta t}E\left[|X_i(t+\Delta t)-\xi_i|\left|X_j(t+\Delta t)-\xi_j\right|\right]\Big|_{x_k(t)=\xi_k}$. Also, the $(ij)^{th}$ component of the covariance matrix $\boldsymbol{\Gamma}$ can be expressed in terms of the basis functions as: $\boldsymbol{\Gamma}_{ij} = \sum_{k=1}^{K}\upsilon_k(\mathbf{X}_t)\theta_{ij,k}$; $i,j=1,2,\dots m$. Noting that the covariance matrix of the $m\times n$-diffusion matrix has a dimension $m\times m$ the following regression problem can be constructed,

$$\boldsymbol{Y}_{ij} = \mathbf{L}\boldsymbol{\theta}_{ij}^g + \boldsymbol{\eta}_{ij} \tag{37}$$

where, $\boldsymbol{Y}_{ij} = \boldsymbol{\Gamma}_{ij} = \left[(X_{i,1}-\xi_{i,1})\left(X_{j,1}-\xi_{j,1}\right)\dots(X_{i,N}-\xi_{i,N})\left(X_{j,N}-\xi_{j,N}\right)\right]^T$ is the quadratic covariation of the i^{th} and j^{th} sample path, and $\theta_{ij}^f = \begin{bmatrix} \theta_1^{ij} & \theta_2^{ij} & \dots & \theta_K^{ij} \end{bmatrix}^T$ is the weight vector for the diffusion terms. With the above regression statement the Gibbs sampling can be performed to obtain the sparse solution. Due to the symmetry of the covariance matrix the above sparse regression is required to be performed for $m(m+1)/2$ times to completely identify the diffusion space.

Discovery of SDE by sparse Bayesian regression

The equation discovery using Bayesian linear regression essentially involves classification of the weights of the basis functions in the library into either spike or slab components of the SS-priors. This is done by introducing a latent indicator variable $\boldsymbol{Z}=[Z_1,\dots,Z_K]$ for each of the component $\theta_k;k=1,\dots,K$ of the weight vector $\boldsymbol{\theta}$. The latent indicator variable Z_k is assigned

a value 1 if the weight corresponds to the slab component, otherwise, it takes a value 0 when the weight belongs to spike component. For the understanding on the sparse Bayesian linear regression, consider the following one-dimensional regression,

$$\boldsymbol{Y} = \mathbf{L}\boldsymbol{\theta} + \epsilon \tag{38}$$

where, $\boldsymbol{Y} \in \mathbb{R}^N$ denotes the N-dimensional target vector, $\mathbf{L}$ denotes the library of candidate functions, $\boldsymbol{\theta}$ is the weight vector, and $\epsilon \in \mathbb{R}^N$ is the residual error vector representing the model mismatch error. The model error ϵ is modelled as i.i.d Gaussian random variable with zero mean and variance σ^2. As the name suggests only a few function from the candidate library will actively participate in the final representation of the target vector $\boldsymbol{Y}$. This will yield a sparse solution such that only the components of weight vector that consistently takes a value sufficiently different from zero will be considered. For estimating the weight vector $\boldsymbol{\theta}$, the Bayes formula can be applied as,

$$P(\theta|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|\boldsymbol{\theta})}{P(\boldsymbol{Y})} \tag{39}$$

However, the effect of the parameters of the DSS-prior on the probability of the weights also needs to accounted. For this one can construct a Bayesian hierarchical model as presented in Fig. 1. Then, the likelihood function is written as,

$$\boldsymbol{Y}|\theta,\sigma^2 \sim \mathcal{N}\left(\mathbf{L}\boldsymbol{\theta}, \sigma^2\mathbf{I}_{N\times N}\right) \tag{40}$$

where, $\mathbf{I}_{N\times N}$ denotes the $N\times N$ identity matrix. For the estimate of the posterior distribution of the parameters of the hierarchical the components of the weight vector that belongs to spike does not contribute. For this purpose let us consider a vector $\boldsymbol{\theta}_r \in \mathbb{R}^r : \{r \ll K\}$, composed from the elements of the weight vector $\boldsymbol{\theta}$ for which $Z_k = 1$. Thus, if $\boldsymbol{\theta}_r$ denotes the weight vector corresponding to slab component of DSS-prior then the SS-prior is defined as[31,32],

$$p(\boldsymbol{\theta}|\boldsymbol{Z}) = p_{slab}(\theta_r) \prod_{k,Z_k=0} p_{spike}(\theta_k) \tag{41}$$

where, p_{spike} and p_{slab} denotes the spike and slab distributions. These are defined as, $p_{spike}(\theta_k) = \delta_0$ and $p_{slab}(\theta_r) = \mathcal{N}\left(\mathbf{0}, \sigma^2\vartheta_s\mathbf{R}_{0,r}\right)$ with $\mathbf{R}_{0,r} = \mathbf{I}_{r\times r}$. It is to be noted that the spike and slab distributions are assumed as independent. Given the noise variance σ^2, the the slab variance ϑ_s is assigned the Inverse-gamma prior with the hyperparameters α_ϑ and β_ϑ. The latent variables Z_k takes a value from the set $\{0,1\}$, thus each of the element in the latent vector is assigned the Bernoulli prior with common hyperparameter p_0. However, the hyperparameter p_0 is allowed to adapt according to the Beta prior with the hyperparameters α_p, and β_p. Similarly the variance of the measurement noise is simulated from the Inverse-gamma distribution with the hyperparameters α_σ, and β_σ. The hyperparameters $\alpha_\vartheta, \beta_\vartheta, \alpha_p, \beta_p, \alpha_\sigma$, and β_σ are provided as a deterministic constants in the hierarchical model. The summery of the DSS-prior model is specified in the Fig. 1 whose corresponding equations are given in Eqs. (42), (43), (44) and (45).

$$p(\vartheta_s) = IG(\alpha_\vartheta, \beta_\vartheta) \tag{42}$$
$$p(Z_k|p_0) = Bern(p_0); k = 1\ldots K \tag{43}$$
$$p(p_0) = Beta\left(\alpha_p, \beta_p\right) \tag{44}$$
$$p\left(\sigma^2\right) = IG(\alpha_\sigma, \beta_\sigma) \tag{45}$$

From the DAG structure of the DSS-model in Fig. 1, the joint distribution of the random variables $\boldsymbol{\theta}$, $\boldsymbol{Z}$, ϑ_s, σ^2, $p_0|\boldsymbol{Y}$ can be expanded using the Bayes formula as,

$$\begin{aligned} p\left(\boldsymbol{\theta}, \boldsymbol{Z}, \vartheta_s, \sigma^2, p_0|\boldsymbol{Y}\right) &= \frac{p\left(\boldsymbol{Y}|\boldsymbol{\theta},\sigma^2\right) p\left(\boldsymbol{\theta}|\boldsymbol{Z},\vartheta_s,\sigma^2\right) p(\boldsymbol{Z}|p_0)\, p(\vartheta_s)\, p\left(\sigma^2\right) p(p_0)}{p(\boldsymbol{Y})} \\ &\propto p\left(\boldsymbol{Y}|\boldsymbol{\theta},\sigma^2\right) p\left(\boldsymbol{\theta}|\boldsymbol{Z},\vartheta_s,\sigma^2\right) p(\boldsymbol{Z}|p_0)\, p(\vartheta_s)\, p\left(\sigma^2\right) p(p_0) \end{aligned} \tag{46}$$

where, $p\left(\boldsymbol{\theta}, \boldsymbol{Z}, \vartheta_s, \sigma^2, p_0|\boldsymbol{Y}\right)$ denotes the joint distribution of the random variables, $p\left(\boldsymbol{Y}|\boldsymbol{\theta},\sigma^2\right)$ denotes the likelihood function, $p\left(\boldsymbol{\theta}|\boldsymbol{Z},\vartheta_s,\sigma^2\right)$ is the prior distribution for the weight vector $\boldsymbol{\theta}$, $p(\boldsymbol{Z}|p_0)$ is the prior distribution for the latent vector $\boldsymbol{Z}$, $p(\vartheta_s)$ is the prior distribution for the slab variance ϑ_s, $p\left(\sigma^2\right)$ is the prior distribution for the noise variance, $p(p_0)$ is the prior distribution for the success probability p_0 and $p(\boldsymbol{Y})$ is the marginal likelihood. Direct sampling from the joint distribution function is intractable in this case due to the spike and slab distribution. Thus, Gibbs sampling technique is used to draw the random

samples from the joint distribution[44]. For the Gibbs sampling the conditional distributions of the random variables are derived in Appendix B. The weights corresponding to the spike distribution does not contribute to the selection of the correct basis functions. Thus only the weights corresponding to $Z_k \neq 0$ i.e. the $\boldsymbol{\theta}_r$ vector is samples using the Gibbs sampling. Referring to the Eqs. (42), (43), (44) and (45), the sequence of the random variables $\boldsymbol{\theta}^{(0)}, \sigma^{2(0)}, \vartheta_s^{(0)}, p_0^{(0)}, \boldsymbol{Z}^{(0)}, \ldots, \boldsymbol{\theta}^{(1)}, \sigma^{2(1)}, \vartheta_s^{(1)}, p_0^{(1)}, \boldsymbol{Z}^{(1)}, \ldots$, using the Gibbs sampling technique is obtained by following steps,

1. The weight vector $\boldsymbol{\theta}_r^{(i)}$ is sampled from the Gaussian distribution with mean $\boldsymbol{\mu}_\theta$ and variance $\boldsymbol{\Sigma}_\theta$ as,

$$\boldsymbol{\theta}_r^{(i)} | \boldsymbol{Y}, \vartheta_s^{(i)}, \sigma^{2(i)} \sim N\left(\boldsymbol{\mu}_\theta^{(i)}, \boldsymbol{\Sigma}_\theta^{(i)}\right) \tag{47}$$

   where, the mean and covariance is defined as, $\boldsymbol{\mu}_\theta^{(i)} = \boldsymbol{\Sigma}_\theta^{(i)} \mathbf{L}_r^{(i)T} \boldsymbol{Y}$ and $\boldsymbol{\Sigma}_\theta^{(i)} = \sigma^{2(i)} \left(\mathbf{L}_r^{(i)T} \mathbf{L}_r^{(i)} + \vartheta_s^{(i)-1} \mathbf{R}_{0,r}^{(i)-1}\right)^{-1}$, respectively.

2. The latent variable $Z_k^{(i+1)}$ is assigned the values from the set, $\{0, 1\}$ by using the Bernoulli distribution as,

$$Z_k^{(i+1)} | \boldsymbol{Y}, \vartheta_s^{(i)}, p_0^{(i)} \sim Bern(u_k) \tag{48}$$

   where, $u_k = \dfrac{p_0}{p_0 + \lambda(1 - p_0)}$ and $\lambda = \dfrac{p\left(\boldsymbol{Y} | Z_k^{(i)} = 0, \boldsymbol{Z}_{-k}^{(i)}, \vartheta_s^{(i)}\right)}{p\left(\boldsymbol{Y} | Z_k^{(i)} = 1, \boldsymbol{Z}_{-k}^{(i)}, \vartheta_s^{(i)}\right)}$. Here, $\boldsymbol{Z}_{-k}^{(i)} \in \mathbb{R}^{K-1}$ denotes the latent variable vector $\boldsymbol{Z}$ consisting of all the elements except the k^{th} component. The probability that the k^{th} latent variable $Z_k^{(i)}$ takes a value 0 or 1 is estimated as follows,

$$\begin{aligned} p\left(\boldsymbol{Y} | \boldsymbol{Z}^{(i)}, \vartheta_s^{(i)}\right) &= \frac{\Gamma\left(\alpha_\sigma + \dfrac{N}{2}\right) \beta_\sigma^{\alpha_\sigma}}{\Gamma(\alpha_\sigma)(2\pi)^{\frac{N}{2}} \left(\beta_\sigma + \dfrac{1}{2}\boldsymbol{Y}^T \boldsymbol{Y}\right)^{\left(\alpha_\sigma + \frac{N}{2}\right)}}; \quad \text{when all } \left\{Z_k^{(i)} : k = 1, \ldots, K\right\} = 0 \\ &= \frac{\Gamma\left(\alpha_\sigma + \dfrac{N}{2}\right) \beta_\sigma^{\alpha_\sigma} \left(\left|\mathbf{R}_{0,r}^{(i)-1}\right| \left|\boldsymbol{\Sigma}_\theta^{(i)}\right|\right)^{\frac{1}{2}}}{\Gamma(\alpha_\sigma)(2\pi)^{\frac{N}{2}} \vartheta_s^{\frac{h_z}{2}} \left(\beta_\sigma + \dfrac{1}{2}\boldsymbol{Y}^T \left(\mathbf{I}_{N\times N} - \mathbf{L}_r^{(i)} \boldsymbol{\Sigma}_\theta^{(i)} \mathbf{L}_r^{(i)T}\right) \boldsymbol{Y}\right)^{\left(\alpha_\sigma + \frac{N}{2}\right)}}; \quad \text{otherwise} \end{aligned} \tag{49}$$

3. The noise variance $\sigma^{2(i+1)}$ is simulated from the Inverse-gamma distribution as,

$$\sigma^{2(i+1)} | \boldsymbol{Y}, \boldsymbol{Z}^{(i+1)}, \vartheta_s^{(i)} \sim IG\left(\alpha_\sigma + \frac{N}{2}, \beta_\sigma + \frac{1}{2}\left(\boldsymbol{Y}^T \boldsymbol{Y} - \boldsymbol{\mu}_\theta^{(i)T} \boldsymbol{\Sigma}_\theta^{(i)-1} \boldsymbol{\mu}_\theta^{(i)}\right)\right) \tag{50}$$

4. The slab variance $\vartheta_s^{(i+1)}$ is sampled from the Inverse-gamma distribution as,

$$\vartheta_s^{(i+1)} | \boldsymbol{\theta}^{(i)}, \boldsymbol{Z}^{(i+1)}, \sigma^{2(i+1)} \sim IG\left(\alpha_\vartheta + \frac{h_z}{2}, \beta_\vartheta + \frac{1}{2\sigma^2} \boldsymbol{\theta}_r^{(i)T} \mathbf{R}_{0,r}^{(i)-1} \boldsymbol{\theta}_r^{(i)}\right) \tag{51}$$

5. The success rate $p_0^{(i+1)}$ is sample from the Beta distribution as,

$$p_0^{(i+1)} | \boldsymbol{Z}^{(i+1)} \sim Beta\left(\alpha_p + h_z, \beta_p + K - h_z\right) \tag{52}$$

   where, $h_z = \sum_{k=1}^{K} Z_k{}^{(i+1)}$.

6. The weight vector $\boldsymbol{\theta}_r^{(i+1)}$ is updated using the step 1.

This MCMC is performed for a total of N_M simulations out of which initial 1000 samples are discarded as the burn-in samples. Let N_s denote the number of MCMC required to achieve the stationary distribution after the burn-in samples are discarded. Then the marginal posterior inclusion probability (PIP):= $p(Z_k = 1 | \boldsymbol{Y})$ for each of the K basis functions can be estimated by taking mean over the Gibbs samples for each of the k^{th} latent vector Z_k[31] as,

$$p(Z_k = 1 | \boldsymbol{Y}) \approx \frac{1}{N_s} \sum_{j=1}^{N_s} Z_k^j; k = 1, \ldots, K \tag{53}$$

The basis functions whose corresponding PIP values are more than 0.5 i.e. $p(Z_k = 1|\boldsymbol{Y}) > 0.5$ are included in the final model of discovered equation. The PIP value greater than 0.5 indicates that the selected basis functions are observed more than half of the times in the MCMC simulations. Higher PIP value suggests that in case of unseen scenarios the corresponding basis functions are highly likely to occur in the data representation target vector. The mean and covariance of the weight vector gives the expected value and standard deviation of the system parameters. The standard deviation provides the confidence interval of the parameters which can be used to design the lower and upper bounds of a random variable in an unseen environment. A pseudo code for the proposed framework, is provided in the Algorithm 1.

Algorithm 1 Pseudo code of the proposed Bayesian framework for discovery of governing physics of stochastic non-linear systems

Input: Sample paths: $\mathbf{X}(t) \in \mathbb{R}^{N\times m}$, hyperparameters: $\alpha_p, \beta_p, \alpha_\sigma, \beta_\sigma, \alpha_\vartheta, \beta_\vartheta, p_0^{(0)}, \vartheta_s^{(0)}$

1: Estimate the linear and quadratic variation vectors. ▹ Eq. (34)
2: For drift obtain the target vectors $\boldsymbol{Y}_i$ and the library $\mathbf{L}$ using the candidate basis functions. ▹ Eq. (36)
3: For diffusion obtain the target vectors $\boldsymbol{Y}_{ij}$ and the library $\mathbf{L}$. ▹ Eq. (37)
4: Estimate the initial variance of noise from residual variance: $\sigma^{2,(0)} = \text{Var}(\mathbf{L}\boldsymbol{\theta} - \boldsymbol{Y})$
5: Estimate the initial latent vector $\boldsymbol{Z}^{(0)}=\left[Z_1^{(0)}, Z_2^{(0)}, \ldots, Z_K^{(0)}\right]$ such that $\text{MSE}(\mathbf{L}\boldsymbol{\theta} - \boldsymbol{Y})$ is minimum.
6: Find $\theta^{(0)}$, $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$ ▹ Eq. (47)
7: **for** $i = 1, \ldots, N_M$ **do**
8: Update the latent variable vector $\boldsymbol{Z}^{(i)}$. ▹ Eq. (48)
9: Update the noise variance $\sigma^{2(i)}$. ▹ Eq. (50)
10: Update the slab variance $\vartheta_s^{(i)}$. ▹ Eq. (51)
11: Update the success rate $p_0^{(i)}$. ▹ Eq. (52)
12: Update the weight vector $\boldsymbol{\theta}^{(i)}$. ▹ Eq. (47)
13: Repeat steps 8→12
14: Discard the burn-in MCMC samples
15: Estimate the marginal PIP values $p(Z_k = 1|\boldsymbol{Y})$ ▹ Eq. (53)
16: Include the basis functions with higher PIP values
17: Estimate the expected value $\text{E}[\theta_k]$ and standard deviation $\sigma[\theta_k]$ of the system parameters

Output: $\{\boldsymbol{\theta}_k; k = 1 \ldots K\}$, $\{\vartheta_k(\boldsymbol{X}_j); j = 1 \ldots \bar{m}\}$, where, K and $\bar{m}$ are the number of library functions and process states, respectively.

Data availability

On acceptance all the used datasets in this study will be made public on GitHub by the corresponding author.

Code availability

On acceptance all the source codes to reproduce the results in this study will be made available to public on GitHub by the corresponding author.

References

1. Zanna, L. & Bolton, T. Data-driven equation discovery of ocean mesoscale closures. *Geophys. Res. Lett.* **47**, e2020GL088376 (2020).
2. Zheng, Y. *et al.* The maximum likelihood climate change for global warming under the influence of greenhouse effect and lévy noise. *Chaos: An Interdiscip. J. Nonlinear Sci.* **30**, 013132 (2020).
3. Jia, C., Zhang, M. Q. & Qian, H. Emergent lévy behavior in single-cell stochastic gene expression. *Phys. Rev. E* **96**, 040402 (2017).
4. Noé, F. & Clementi, C. Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Curr. opinion structural biology* **43**, 141–147 (2017).
5. Zhang, X. F. Information uncertainty and stock returns. *The J. Finance* **61**, 105–137 (2006).
6. Bongard, J. & Lipson, H. Automated reverse engineering of nonlinear dynamical systems. *Proc. Natl. Acad. Sci.* **104**, 9943–9948 (2007).

7. Brunton, S. L., Proctor, J. L. & Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. national academy sciences* **113**, 3932–3937 (2016).
8. Akaike, H. A new look at the statistical model identification. *IEEE transactions on automatic control* **19**, 716–723 (1974).
9. Schwarz, G. Estimating the dimension of a model. *The annals statistics* 461–464 (1978).
10. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *science* **324**, 81–85 (2009).
11. Kevrekidis, I. G. & Samaey, G. Equation-free multiscale computation: Algorithms and applications. *Annu. review physical chemistry* **60**, 321–344 (2009).
12. Mangan, N. M., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Transactions on Mol. Biol. Multi-Scale Commun.* **2**, 52–63 (2016).
13. Hoffmann, M., Fröhner, C. & Noé, F. Reactive sindy: Discovering governing reactions from concentration data. *The J. chemical physics* **150**, 025101 (2019).
14. Bhadriraju, B., Narasingam, A. & Kwon, J. S.-I. Machine learning-based adaptive model identification of systems: Application to a chemical process. *Chem. Eng. Res. Des.* **152**, 372–383 (2019).
15. Loiseau, J.-C. & Brunton, S. L. Constrained sparse galerkin regression. *J. Fluid Mech.* **838**, 42–67 (2018).
16. Loiseau, J.-C., Noack, B. R. & Brunton, S. L. Sparse reduced-order modelling: sensor-based dynamics to full-state estimation. *J. Fluid Mech.* **844**, 459–490 (2018).
17. Lai, Z. & Nagarajaiah, S. Sparse structural system identification method for nonlinear dynamic systems with hysteresis/inelastic behavior. *Mech. Syst. Signal Process.* **117**, 813–842 (2019).
18. Li, S. *et al.* Discovering time-varying aerodynamics of a prototype bridge by sparse identification of nonlinear dynamical systems. *Phys. Rev. E* **100**, 022220 (2019).
19. Schaeffer, H. & McCalla, S. G. Sparse model selection via integral terms. *Phys. Rev. E* **96**, 023302 (2017).
20. Mangan, N. M., Kutz, J. N., Brunton, S. L. & Proctor, J. L. Model selection for dynamical systems via sparse regression and information criteria. *Proc. Royal Soc. A: Math. Phys. Eng. Sci.* **473**, 20170009 (2017).
21. Kaiser, E., Kutz, J. N. & Brunton, S. L. Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proc. Royal Soc. A* **474**, 20180335 (2018).
22. Schaeffer, H., Tran, G., Ward, R. & Zhang, L. Extracting structured dynamical systems using sparse optimization with very few samples. *Multiscale Model. & Simul.* **18**, 1435–1461 (2020).
23. Stender, M., Oberst, S. & Hoffmann, N. Recovery of differential equations from impulse response time series data for model identification and feature extraction. *Vibration* **2**, 25–46 (2019).
24. Boninsegna, L., Nüske, F. & Clementi, C. Sparse learning of stochastic dynamical equations. *The J. chemical physics* **148**, 241723 (2018).
25. Rudy, S. H., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Data-driven discovery of partial differential equations. *Sci. Adv.* **3**, e1602614 (2017).
26. Zhang, S. & Lin, G. Robust data-driven discovery of governing physical laws with error bars. *Proc. Royal Soc. A: Math. Phys. Eng. Sci.* **474**, 20180305 (2018).
27. Rudy, S. H., Kutz, J. N. & Brunton, S. L. Deep learning of dynamics and signal-noise decomposition with time-stepping constraints. *J. Comput. Phys.* **396**, 483–506 (2019).
28. Raissi, M. Deep hidden physics models: Deep learning of nonlinear partial differential equations. *The J. Mach. Learn. Res.* **19**, 932–955 (2018).
29. Raissi, M., Perdikaris, P. & Karniadakis, G. E. Multistep neural networks for data-driven discovery of nonlinear dynamical systems. *arXiv preprint arXiv:1801.01236* (2018).
30. Fuentes, R. *et al.* Equation discovery for nonlinear dynamical systems: A bayesian viewpoint. *Mech. Syst. Signal Process.* **154**, 107528 (2021).
31. Nayek, R., Fuentes, R., Worden, K. & Cross, E. J. On spike-and-slab priors for bayesian equation discovery of nonlinear dynamical systems via sparse linear regression. *Mech. Syst. Signal Process.* **161**, 107986 (2021).
32. Mitchell, T. J. & Beauchamp, J. J. Bayesian variable selection in linear regression. *J. american statistical association* **83**, 1023–1032 (1988).

33. George, E. I. & McCulloch, R. E. Approaches for bayesian variable selection. *Stat. sinica* 339–373 (1997).

34. Oksendal, B. *Stochastic differential equations: an introduction with applications* (Springer Science & Business Media, 2013).

35. Tripura, T., Gogoi, A. & Hazra, B. An ito-taylor weak 3.0 method for stochastic dynamics of nonlinear systems. *Appl. Math. Model.* (2020).

36. Tripura, T., Bhowmik, B., Pakrashi, V. & Hazra, B. Real-time damage detection of degrading systems. *Struct. Heal. Monit.* **19**, 810–837 (2020).

37. Risken, H. Fokker-planck equation. In *The Fokker-Planck Equation*, 63–95 (Springer, 1996).

38. O'Hara, R. B. & Sillanpää, M. J. A review of bayesian variable selection methods: what, how and which. *Bayesian analysis* **4**, 85–117 (2009).

39. Calin, O. *An informal introduction to stochastic calculus with applications* (World Scientific, 2015).

40. Klebaner, F. C. *Introduction to stochastic calculus with applications* (World Scientific Publishing Company, 2005).

41. Kloeden, P. E. & Platen, E. Higher-order implicit strong numerical schemes for stochastic differential equations. *J. statistical physics* **66**, 283–314 (1992).

42. Tripura, T., Imran, M., Hazra, B. & Chakraborty, S. A change of measure enhanced near exact euler maruyama scheme for the solution to nonlinear stochastic dynamical systems. *arXiv preprint arXiv:2108.10655* (2021).

43. Tripura, T., Hazra, B. & Chakraborty, S. Generalized weakly corrected milstein solutions to stochastic differential equations. *arXiv preprint arXiv:2108.10681* (2021).

44. Casella, G. & George, E. I. Explaining the gibbs sampler. *The Am. Stat.* **46**, 167–174 (1992).

Acknowledgements

SC acknowledges the financial support received from IIT Delhi in form of seed grant.

Author contributions statement

T. Tripura: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft. S. Chakraborty: Conceptualization, Methodology, Software, review and editing, Supervision, Funding acquisition.

Competing interests

The authors declare no competing interests. The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.

Additional information

Correspondence and requests for materials should be addressed to Souvik Chakraborty.

A Kramers-Moyal expansion for estimation of the drift and diffusion terms of an SDE from sample paths

Let us consider, $p(X,t)$=P$(X,t|X_0,t_0)$ be the transition probability density of the solution of the SDE in Eq. (33). Then the Kramers-Moyal expansion is written as[37]:

$$\frac{\partial P(X,t)}{\partial t} = \sum_{n=1}^{\infty} \left(-\frac{\partial}{\partial X}\right)^n D^{(n)}(X,t)p(X,t) \tag{54}$$

where the coefficients in the expansion are given as:

$$D^{(n)}(X) = \left.\frac{1}{n!}M_n(t)\right|_{t=0} = \frac{1}{n!}\lim_{\Delta t\to 0}\frac{1}{\Delta t}\left\langle |X(t+\Delta t)-z|^n\right\rangle\Big|_{X(t)=z} \tag{55}$$

In order to know how many terms in the Eq. (54) will be active for the SDE in Eq. (33) an Fokker-Planck equation for the pdf of the solution of Eq. (33) needs to be constructed. Towards this, let us consider a well behaved generic function $F(X,t)$ which is at least twice differential. For this function the Itô's lemma is written as:

$$dF(X,t) = \frac{\partial F(X,t)}{\partial X} dX(t) + \frac{1}{2}\frac{\partial^2 F(X,t)}{\partial X^2} g^2(X,t)dt \tag{56}$$

Substituting the Eq. (33) in the above equation gives the following:

$$dF(X,t) = \left(\frac{\partial F(X,t)}{\partial X} f(X,t) + \frac{1}{2}\frac{\partial^2 F(X,t)}{\partial X^2} g^2(X,t)\right)dt + \frac{\partial F(X,t)}{\partial X} g(X,t)dB(t) \tag{57}$$

Upon taking the derivative with respect to t on both sides yields:

$$\frac{dE[F(X,t)]}{dt} = E\left(\frac{\partial F(X,t)}{\partial X} f(X,t) + \frac{1}{2}\frac{\partial^2 F(X,t)}{\partial X^2} g^2(X,t)\right) \tag{58}$$

Expanding the above equation using expectation operator and noting that the last term in the expression is a stochastic integral, one can invoke the result $\mathrm{E}\left[\int_0^t \frac{\partial F(X,s)}{\partial X} g(X,s)dB(s)\right] = 0$, where E(.) is the expectation operator to obtain the following,

$$\int_{\infty}^{\infty} p(X,t)\frac{dF(X)}{dt}dX = \int_{\infty}^{\infty} p(X,t)\left(\frac{\partial F(X,t)}{\partial X} f(X,t) + \frac{1}{2}\frac{\partial^2 F(X,t)}{\partial X^2} g^2(X,t)\right)dX \tag{59}$$

By expanding the terms in the above equation using integration by parts and assuming that $\lim_{X\to 0} p(X,t) \to 0$, the following weak form is obtained:

$$\int_{\infty}^{\infty} F(X)\left(\frac{\partial p(X,t)}{\partial t} + \frac{\partial (p(X,t)f(X,t))}{\partial X} - \frac{1}{2}\frac{\partial^2 (p(X,t)g^2(X,t))}{\partial X^2}\right)dX = 0 \tag{60}$$

As the function $F(X,t)$ is arbitrarily selected, the FPK equation is obtained:

$$\frac{\partial p(X,t)}{\partial t} = -\frac{\partial (p(X,t)f(X,t))}{\partial X} + \frac{1}{2}\frac{\partial^2 (p(X,t)g^2(X,t))}{\partial X^2};\ p(0,X) = p_0(X) \tag{61}$$

In a general setting the above equation for higher dimensional diffusion process is expressed as:

$$\frac{\partial p(X,t)}{\partial t} = L_t p(X,t) = -\sum_i^m \frac{\partial (p(X,t)f_i(X,t))}{\partial X_i} + \frac{1}{2}\sum_i^m \sum_j^m \sum_k^n \frac{\partial^2 (p(X,t)g_{i,k}(X,t)g_{j,k}(X,t))}{\partial X_i \partial X_j} \tag{62}$$

where, the random variable X satisfies the SDE in Eq. (33). At this point it is clear that there will be two terms active in the Eq. (33). Thus for n=2, the Eq. (54) gives,

$$\frac{\partial P(X,t)}{\partial t} = -\frac{\partial}{\partial X}\left[D^{(1)}(X,t)p(X,t)\right] + \frac{\partial^2}{\partial X^2}\left[D^{(2)}(X,t)p(X,t)\right] \tag{63}$$

On comparison of Eq. (54) and 63, it is straightforward to note that to estimate the drift and diffusion terms it suffices to estimate the first and second order moments of the variations of the random variable $X(t)$ and then calculate the coefficients $D^{(1)}$ and $D^{(2)}$, formally,

$$f(X,t) = D^{(1)},\ g^2(X,t) = D^{(2)} \tag{64}$$

To derive these coefficients in the Kramers-Moyal expansion, it is imperative to understand the one-step Itô-Taylor expansion of the random variable $X(t)$. Referring to the Eq. (57), the integral form the Itô-lemma can be expressed as follows:

$$F(X_{t+h}, t+h) = F(X_t,t) + \int_t^{t+h}\left\{f(X_s,s)F'(X_s,s) + \frac{1}{2}g^2(X_s,s)F''(X_s,s)\right\}ds + \int_t^{t+h} g(X_s,s)F'(X_s,s)dB(s) \tag{65}$$

where, $F'(X_s,s)$ and $F''(X_s,s)$ denotes the first and second order partial derivatives with respect to system states. For the simplicity two stochastic operators are defined in the following form:

$$\begin{aligned}\Im^0(.) &= \frac{\partial(.)}{\partial t} + \sum_i^m f_i(X_t,t)\frac{\partial(.)}{\partial X_i} + \frac{1}{2}\sum_i^m\sum_j^m\sum_k^n g_{i,k}(X_t,t)g_{j,k}(X_t,t)\frac{\partial^2(.)}{\partial X_i \partial X_j}\\ \Im^1(.) &= \sum_i^m\sum_k^n g_{i,k}(X_t,t)\frac{\partial(.)}{\partial X_i}\end{aligned} \tag{66}$$

Under which the Eq. (65) can be rephrased as:

$$F(X_{t+h},t+h) = F(X_t,t) + \int_t^{t+h}\Im^0\left(F(X_s,s)\right)ds + \int_t^{t+h}\Im^1\left(F(X_s,s)\right)dB(s) \tag{67}$$

Then substituting $F(X_t,t) = X(t)$, one verifies that $\Im^0 X(t) = f(X_t,t)$, and $\Im^1 X(t) = g(X_t,t)$. This yields the first iteration:

$$\begin{aligned}X(t+h) &= X(t) + \int_t^{t+h}\Im^0\left(X(s)\right)ds + \int_t^{t+h}\Im^1\left(X(s)\right)dB(s)\\ &= X(t) + \int_t^{t+h} f(X_s,s)ds + \int_t^{t+h} g(X_s,s)dB(s)\end{aligned} \tag{68}$$

In order to perform the second iteration it is required to find the stochastic expansion of the terms $F(X_t,t) = f(X_s,s)$, and $F(X_t,t) = g(X_s,s)$. Thus, expanding $f(X_t,t)$, and $g(X_t,t)$ and using the operators in Eq. (66) yields,

$$\begin{aligned}f(X_s,s) &= f(X,t) + \int_t^{s_1}\Im^0\Big(f(X_{s_2},s_2)\Big)ds_2 + \int_t^{s_1}\Im^1\Big(f(X_{s_2},s_2)\Big)dB(s_2)\\ g(X_s,s) &= g(X,t) + \int_t^{s_1}\Im^0\Big(g(X_{s_2},s_2)\Big)ds_2 + \int_t^{s_1}\Im^1\Big(g(X_{s_2},s_2)\Big)dB(s_2)\end{aligned} \tag{69}$$

On substituting the above result in Eq. (68) and further iterating gives,

$$X(t+h) - X(t) = f(X,t)\int_t^{t+h} ds_1 + g(X,t)\int_t^{t+h} dB(s_1) + \Im^1\left(g(X,t)\right)\int_t^{t+h}\int_t^{s_1} dB(s_2)dB(s_1) + R \tag{70}$$

where the remainder term R is,

$$\begin{aligned}R = \;& \int_t^{t+h}\int_t^{s_1}\Im^0\Big(f(X_{s_2},s_2)\Big)ds_2ds_1 + \left\{\int_t^{t+h}\int_t^{s_1}\Im^1\Big(f(X_{s_2},s_2)\Big)dB(s_2)ds_1 + \int_t^{t+h}\int_t^{s_1}\Im^0\Big(g(X_{s_2},s_2)\Big)ds_2dB(s_1)\right\} +\\ & \int_{t_0}^{t}\int_{t_0}^{s_1}\int_t^{s_2}\Im^1\Im^0\Big(g(X_{s_3},s_3)\Big)ds_3dB(s_2)dB(s_1) + \int_{t_0}^{t}\int_{t_0}^{s_1}\int_t^{s_2}\Im^1\Im^1\Big(g(X_{s_3},s_3)\Big)dB(s_3)dB(s_2)dB(s_1)\end{aligned} \tag{71}$$

It can be noticed that the above equation is infinitely expandable using the Eq. (66). For further treatment, the first moment is taken on both side. Noting the results $\left\langle\int_t^{t+h} dB(s_1)\right\rangle = 0$, and $\left\langle\int_{t_0}^{t}\int_{t_0}^{s_1} dB(s_2)dB(s_1)\right\rangle = \left\langle\int_{t_0}^{t} B(s_1)dB(s_1) - B(t)\int_{t_0}^{t} dB(s_1)\right\rangle = 0$,

$$\begin{aligned}\langle X(t+h) - X(t)\rangle &= f(X,t)h + g(X,t)\left\langle\int_t^{t+h} dB(s_1)\right\rangle + \Im^1\left(g(X,t)\right)\left\langle\int_t^{t+h}\int_t^{s_1} dB(s_2)dB(s_1)\right\rangle + \langle R\rangle\\ &= f(X,t)h + \langle R\rangle\end{aligned} \tag{72}$$

If the number of iteration is k then the Brownian integrals $\int_t^{t+h}\int_t^{s_1}\ldots\int_t^{s_k} dB(s_{k+1})\ldots dB(s_2)dB(s_1)$ of multiplicity k have a contribution proportional to $h^{k/2}$, the time integrals $\int_t^{t+h}\int_t^{s_1}\ldots\int_t^{s_k} ds_{k+1}\ldots ds_2ds_1$ have a contribution proportional to h^k and the combination between them shares a contribution between $h^{k/2}$ and h^k. Thus it is easy to infer that as $h\to 0$, the higher order terms in the remainder R will vanish. After invoking this facts in the above expression the first coefficient in Kramers-Moyal expansion is obtained as,

$$D^{(1)} = f(X,t) = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\left\langle X^{(1)}(t+\Delta t) - z\right\rangle\Big|_{X(t)=z} \tag{73}$$

To derive the second coefficient it is only required to find the quadratic variation of the increment process of $X(t)$. Upon taking second moment on both sides of Eq. (70) the following is obtained,

$$\begin{aligned}\left\langle|X(t+h) - X(t)|^2\right\rangle &= f(X,t)h + g(X,t)\Delta B + \langle R\rangle\\ &= f^2(X,t)h^2 + 2f(X,t)g(X,t)h\Delta B + g^2(X,t)(\Delta B)^2 + \langle R\rangle\\ &= g^2(X,t)h + \langle R\rangle\end{aligned} \tag{74}$$

In the above prove the Itô identifies are utilized which states that under the mean square convergence theory the quadratic variation of the time and Brownian increments are given as $ds_2 ds_1 = 0$, $ds_1 dW(s_1) = 0$, $dW(s_2) ds_1 = 0$, $dW(s_1) dW(s_2) = 0$. The deterministic time integrals will vanish automatically since they have finite variation and zero quadratic covariation and the other higher order Brownian integrals will vanish as $h \to 0$. Thus the second coefficient in the Kramers-Moyal expansion is obtained as,

$$D^{(2)} = g^2(X,t) = \frac{1}{2} \lim_{\Delta t \to 0} \frac{1}{\Delta t} \left\langle \left| X^{(1)}(t+\Delta t) - z \right|^2 \right\rangle \Big|_{X(t)=z} \tag{75}$$

From the ergodic assumption of the evolution of process $X(t)$ the time average $\langle . \rangle$ is often replaced by expectation operator E(.). For m-variables $\boldsymbol{X} = \{X_1, X_2, \ldots X_m\}$, the diffusion process has the form,

$$dX_i(t) = f_i(\boldsymbol{X}_t, t) + \sum_j^n g_{ij}(\boldsymbol{X}_t, t) dB_j(t);\ i = 1, 2, \ldots m \tag{76}$$

The properties of the Brownian motion are: $\langle B_i(t) \rangle = 0$ and $\left\langle B_i(t) B_j(s) \right\rangle = \min(t, s)$. If the covariation matrix of the diffusion components is given by $\Gamma(\boldsymbol{X}_t, t) = g(\boldsymbol{X}_t, t) g(\boldsymbol{X}_t, t)^T$, then, the drift $f_i(\boldsymbol{X}, t)$ of i^{th}-diffusion process and the ij^{th}-element of the covariation matrix $\Gamma(\boldsymbol{X}, t)$ can be estimated as,

$$\begin{aligned} f_i(\boldsymbol{X}, t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} E\left[X_i(t+\Delta t) - z_i \right] \Big|_{X_k(t)=z_k} \quad \forall\ k = 1, 2, \ldots m \\ \Gamma_{ij}(\boldsymbol{X}, t) &= \frac{1}{2} \lim_{\Delta t \to 0} \frac{1}{\Delta t} E\left[|X_i(t+\Delta t) - z_i| \left| X_j(t+\Delta t) - z_j \right| \right] \Big|_{X_k(t)=z_k} \quad \forall\ k = 1, 2, \ldots m \end{aligned} \tag{77}$$

B Conditional probability distributions of the discontinuous Spike and Slab priors (DSS)

The DSS prior model is described in section . Drawing sample from the joint distribution is intractable but possible through Monte Carlo Markov Chain methods. In the present work, the Gibbs sampling technique is utilized to draw the sample for the random variables $\boldsymbol{Y}$, $\boldsymbol{\theta}$, $\boldsymbol{Z}$, ϑ_s, σ^2 and p_0. For drawing samples using Gibbs sampling the random variables need to be conditioned on other variables. However, due to the DAG structure in Fig. 1, the dependencies of the conditional distributions on other random variables can be relaxed to some extent as follows:

$$\begin{aligned} p\left(p_0 \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{Z}, \vartheta_s, \sigma^2\right) &= p\left(p_0 \mid \boldsymbol{Z}\right) \\ p\left(\vartheta_s \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{Z}, p_0, \sigma^2\right) &= p\left(\vartheta_s \mid \boldsymbol{\theta}, \boldsymbol{Z}, \sigma^2\right) \\ p\left(\sigma^2 \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{Z}, p_0, \vartheta_s\right) &= p\left(\sigma^2 \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{Z}, \vartheta_s\right) \\ p\left(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{Z}, p_0, \vartheta_s, \sigma^2\right) &= p\left(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{Z}, \vartheta_s, \sigma^2\right) \\ p\left(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta}, p_0, \vartheta_s, \sigma^2\right) &= p\left(\boldsymbol{Z} \mid \boldsymbol{\theta}, p_0, \vartheta_s, \sigma^2\right) \end{aligned} \tag{78}$$

The elements of the weight vector $\boldsymbol{\theta}$ for which the latent variable $Z_k \neq 1$, becomes an absorbing state in the Markov chain. However to achieve a stationary distribution one needs to construct a irreducible Markov chain. Thus the conditionals over weight vector is eliminated by marginalizing the conditional distributions with respect to the vector $\boldsymbol{\theta}$.

$$\begin{aligned} \int p\left(\sigma^2 \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{Z}, \vartheta_s\right) d\boldsymbol{\theta} &= \int \frac{p\left(\sigma^2, \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{Z}, \vartheta_s\right)}{p\left(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{Z}, \vartheta_s\right)} d\boldsymbol{\theta} = p\left(\sigma^2 \mid \boldsymbol{Y}, \boldsymbol{Z}, \vartheta_s\right) \\ \int p\left(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta}, p_0, \vartheta_s, \sigma^2\right) d\boldsymbol{\theta} &= \int \frac{p\left(\boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{\theta}, p_0, \vartheta_s, \sigma^2\right)}{p\left(\boldsymbol{Y}, \boldsymbol{\theta}, p_0, \vartheta_s, \sigma^2\right)} d\boldsymbol{\theta} = \frac{p\left(\boldsymbol{Z}, \boldsymbol{Y}, p_0, \vartheta_s, \sigma^2\right)}{p\left(\boldsymbol{Y}, p_0, \vartheta_s, \sigma^2\right)} = p\left(\boldsymbol{Z} | \boldsymbol{Y}, p_0, \vartheta_s, \sigma^2\right) \end{aligned} \tag{79}$$

The conditional distribution $p(p_0 \mid \boldsymbol{Z})$:

$$\begin{aligned} p(p_0 \mid \boldsymbol{Z}) &= p(\boldsymbol{Z} \mid p_0)\, p(p_0) \\ &\propto \left(\prod_{k=1}^K p_0^{Z_k} (1-p_0)^{1-Z_k} \right) \left(p_0^{\alpha_p - 1} (1-p_0)^{\beta_p - 1} \right) \\ &\propto p_0^{\alpha_p + h_Z} (1-p_0)^{\beta_p + K - h_Z} \end{aligned} \tag{80}$$

where, $h_Z = \sum_{k=1}^K Z_k$. Given the latent vector $\boldsymbol{Z}$, p_0 is sampled as, $p_0 \mid \boldsymbol{Z} \sim Beta\left(\alpha_p + h_z, \beta_p + K - h_z\right)$.

The conditional distribution $p\left(\vartheta_s \mid \boldsymbol{\theta}, \mathbf{Z}, \sigma^2\right)$:

$$\begin{aligned} p\left(\vartheta_s \mid \boldsymbol{\theta}, \mathbf{Z}, \sigma^2\right) &\propto p\left(\boldsymbol{\theta} \mid \mathbf{Z}, \vartheta_s, \sigma^2\right) p\left(\vartheta_s\right) p(\mathbf{Z}) p\left(\sigma^2\right) \\ &\propto p\left(\boldsymbol{\theta} \mid \mathbf{Z}, \vartheta_s, \sigma^2\right) p\left(\vartheta_s\right) \\ &\propto \mathcal{N}\left(\boldsymbol{\theta}_r \mid \mathbf{0}, \sigma^2 \vartheta_s \mathbf{R}_{0,r}\right) \mathcal{IG}\left(\alpha_\vartheta, \beta_\vartheta\right) \\ &\propto \frac{1}{\left(\vartheta_s \sigma^2\right)^{r/2} \left|\mathbf{R}_{0,r}\right|^{1/2}} \exp\left(-\frac{\boldsymbol{\theta}_r^T \mathbf{R}_{0,r}^{-1} \boldsymbol{\theta}_r}{2\sigma^2 \vartheta_s}\right) \vartheta_s^{-\alpha_\vartheta - 1} \exp\left(-\frac{\beta_\vartheta}{\vartheta_s}\right) \\ &\propto \vartheta_s^{-\left(\alpha_\vartheta + \frac{r}{2}\right) - 1} \exp\left(-\frac{\beta_\vartheta + \left(\boldsymbol{\theta}_r^T \mathbf{R}_{0,r}^{-1} \boldsymbol{\theta}_r / 2\sigma^2\right)}{\vartheta_s}\right) \end{aligned} \tag{81}$$

where, r is the number of elements of the weight vector $\boldsymbol{\theta}$ for which the latent variable $Z_k = 1$. In the above it is to be noted that $p(\mathbf{Z})$ and $p\left(\sigma^2\right)$ are constants when ϑ_s is sampled. This can be observed from the DAG structure in Fig. 1. Thus given $\boldsymbol{\theta}$, $\mathbf{Z}$ and σ^2 and ϑ_s is sampled as, $\vartheta_s \mid \boldsymbol{\theta}, \mathbf{Z}, \sigma^2 \sim \mathcal{IG}\left(\alpha_\vartheta + \frac{r}{2}, \beta_\vartheta + \frac{\boldsymbol{\theta}_r^T \mathbf{R}_{0,r}^{-1} \boldsymbol{\theta}_r}{2\sigma^2}\right)$

The conditional distribution $p\left(\sigma^2 \mid \boldsymbol{Y}, \boldsymbol{\theta}, \mathbf{Z}, \vartheta_s\right)$:

$$\begin{aligned} p\left(\sigma^2 \mid \boldsymbol{Y}, \mathbf{Z}, \vartheta_s\right) &\propto \int p\left(\boldsymbol{Y}, \boldsymbol{\theta}, \mathbf{Z}, \vartheta_s, \sigma^2\right) d\boldsymbol{\theta} \\ &\propto \left(\int p\left(\boldsymbol{Y} \mid \boldsymbol{\theta}, \sigma^2\right) p\left(\boldsymbol{\theta} \mid \mathbf{Z}, \vartheta_s, \sigma^2\right) d\boldsymbol{\theta}\right) p(\mathbf{Z}) p\left(\vartheta_s\right) p\left(\sigma^2\right) \\ &\propto \left(\int p\left(\boldsymbol{Y} \mid \boldsymbol{\theta}, \sigma^2\right) p\left(\boldsymbol{\theta} \mid \mathbf{Z}, \vartheta_s, \sigma^2\right) d\boldsymbol{\theta}\right) p\left(\sigma^2\right) \end{aligned} \tag{82}$$

From the DAG structure it can be understood that $p(\mathbf{Z})$ and $p\left(\vartheta_s\right)$ will act as a constant when σ^2 is sampled. Since the θ_k-values corresponding to the spike distribution does not contribute to the basis function selection, the remaining weights belonging to the slab denoted by $\boldsymbol{\theta}_r$ is used. Then one can expand the intregrand $p\left(\boldsymbol{Y} \mid \boldsymbol{\theta}, \sigma^2\right) p\left(\boldsymbol{\theta} \mid \mathbf{Z}, \vartheta_s, \sigma^2\right)$, as,

$$\begin{aligned} &p\left(\boldsymbol{Y} \mid \boldsymbol{\theta}, \sigma^2\right) p\left(\boldsymbol{\theta} \mid \mathbf{Z}, \vartheta_s, \sigma^2\right) \\ &= \mathcal{N}\left(\boldsymbol{Y} | \mathbf{L}_r \boldsymbol{\theta}_r, \sigma^2 \mathbf{I}_{N \times N}\right) N\left(\boldsymbol{\theta}_r | 0, \sigma^2 \vartheta_s \mathbf{R}_{0,r}\right) \\ &= \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(-\frac{\left(\boldsymbol{Y} - \mathbf{D}_r \boldsymbol{\theta}_r\right)^T \left(\boldsymbol{Y} - \mathbf{D}_r \boldsymbol{\theta}_r\right)}{2\sigma^2}\right) \frac{\left(\left|\mathbf{R}_{0,r}^{-1}\right|\right)^{1/2}}{\left(2\pi\vartheta_s\sigma^2\right)^{r/2}} \exp\left(-\frac{\boldsymbol{\theta}_r^T \mathbf{R}_{0,r}^{-1} \boldsymbol{\theta}_r}{2\sigma^2 \vartheta_s}\right) \\ &= \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \frac{\left(\left|\mathbf{R}_{0,r}^{-1}\right|\right)^{1/2}}{\left(2\pi\vartheta_s\sigma^2\right)^{r/2}} \exp\left(-\frac{\left(\boldsymbol{\theta}_r - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\theta}_r - \boldsymbol{\mu}\right)}{2\sigma^2}\right) \exp\left(-\frac{\left(\boldsymbol{Y}^T \boldsymbol{Y} - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}\right)}{2\sigma^2}\right) \end{aligned} \tag{83}$$

where, $\boldsymbol{\Sigma}^{-1} = \left(\mathbf{L}_r^T \mathbf{L}_r + \vartheta_s^{-1} \mathbf{R}_{0,r}^{-1}\right)$ and $\boldsymbol{\mu} = \boldsymbol{\Sigma} \mathbf{L}_r^T \boldsymbol{Y}$. Upon integration of the the above expression about $\boldsymbol{\theta}_r$, one obtains:

$$\begin{aligned} p\left(\sigma^2 \mid \boldsymbol{Y}, \mathbf{Z}, \vartheta_s\right) &\propto \frac{1}{\left(\sigma^2\right)^{N/2}} \exp\left(-\frac{\left(\boldsymbol{Y}^T \boldsymbol{Y} - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}\right)}{2\sigma^2}\right) p\left(\sigma^2\right) \\ &\propto \frac{1}{\left(\sigma^2\right)^{N/2}} \exp\left(-\frac{\left(\boldsymbol{Y}^T \boldsymbol{Y} - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}\right)}{2\sigma^2}\right) \left(\sigma^2\right)^{-\alpha_\sigma - 1} \exp\left(-\frac{\beta_\sigma}{\sigma^2}\right) \\ &\propto \left(\sigma^2\right)^{-(\alpha_\sigma + 0.5N) - 1} \exp\left(-\frac{\beta_\sigma + \frac{1}{2}\left(\boldsymbol{Y}^T \boldsymbol{Y} - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}\right)}{\sigma^2}\right) \end{aligned} \tag{84}$$

The random variable σ^2 is sampled as, $\sigma^2 \mid \boldsymbol{Y}, \mathbf{Z}, \vartheta_s \sim \mathcal{IG}\left(\alpha_\sigma + \frac{N}{2}, \beta_\sigma + \frac{1}{2}\left(\boldsymbol{Y}^T \boldsymbol{Y} - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}\right)\right)$.

The conditional distribution $p\left(\boldsymbol{\theta} \mid \boldsymbol{Y}, \mathbf{Z}, \vartheta_s, \sigma^2\right)$: The elements of the weight vector $\boldsymbol{\theta}$ corresponding to the spike distribution i.e. the weights for which the latent variable $z_k = 0; k = 1 \ldots K$, are assigned the value 0. The conditional distribution for the

weights belonging to the slab distribution denoted by $\boldsymbol{\theta}_r$ is derived as follows:

$$\begin{aligned} p\left(\boldsymbol{\theta}_r \mid \boldsymbol{Y}, \vartheta_s, \sigma^2\right) &\propto \mathcal{N}\left(\boldsymbol{Y} \mid \mathbf{L}_r\boldsymbol{\theta}_r, \sigma^2\mathbf{I}_{N\times N}\right)\mathcal{N}\left(\boldsymbol{\theta}_r \mid \mathbf{0}, \sigma^2\vartheta_s\mathbf{R}_{0,r}\right) \\ &\propto \exp\left(-\frac{(\boldsymbol{Y}-\mathbf{L}_r\boldsymbol{\theta}_r)^T(\boldsymbol{Y}-\mathbf{L}_r\boldsymbol{\theta}_r)}{2\sigma^2}\right)\exp\left(-\frac{\boldsymbol{\theta}_r^T\mathbf{R}_{0,r}^{-1}\boldsymbol{\theta}_r}{2\sigma^2 v_s}\right) \\ &\propto \exp\left(-\frac{\boldsymbol{Y}^T\boldsymbol{Y}+\boldsymbol{\theta}_r^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}_r - 2\boldsymbol{\theta}_r^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{(\boldsymbol{\theta}_r-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_r-\boldsymbol{\mu})}{2\sigma^2}\right) \end{aligned} \tag{85}$$

where, $\boldsymbol{\Sigma}^{-1} = \left(\mathbf{L}_r^T\mathbf{L}_r + \vartheta_s^{-1}\mathbf{R}_{0,r}^{-1}\right)$ and $\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{L}_r^T\boldsymbol{Y}$. Then, the random values of $\boldsymbol{\theta}_r$ can be sampled as $\boldsymbol{\theta}_r \mid \boldsymbol{Y}, \vartheta_s, \sigma^2 \sim \mathcal{N}\left(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma}\right)$.

The conditional distribution $p\left(\boldsymbol{Z} \mid \boldsymbol{\theta}, p_0, \vartheta_s, \sigma^2\right)$: The latent variable Z_k corresponding to the k^{th} element of the weight vector are assigned the value 0 or 1 independently. The conditional probability distributions of Z_k are estimated by comparing the probabilities that the k^{th} latent variable $Z_k = 1$ will assume a value 1 or 0, given the values of ϑ_s, p_0 and remaining values of the latent vector $\boldsymbol{Z}_{-k}$. Here, the term $\boldsymbol{Z}_{-k}$ denotes the latent vector $\boldsymbol{Z}$ whose k^{th} element is removed. Let u_k denotes the probability with which the k^{th} latent variable Z_k takes a value 1. The probability u_k is then found as,

$$\begin{aligned} u_k &= \frac{p(Z_k = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}_{-k}, \vartheta_s, p_0)}{p(Z_k = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}_{-k}, \vartheta_s, p_0) + p(Z_k = 0 \mid \boldsymbol{Y}, \boldsymbol{Z}_{-k}, \vartheta_s, p_0)} \\ &= \frac{p(\boldsymbol{Y} \mid Z_k = 1, \boldsymbol{Z}_{-k}, \vartheta_s)\, p(Z_k = 1 \mid p_0)}{p(\boldsymbol{Y} \mid Z_k = 1, \boldsymbol{Z}_{-k}, \vartheta_s)\, p(Z_k = 1 \mid p_0) + p(\boldsymbol{Y} \mid Z_k = 0, \boldsymbol{Z}_{-k}, \vartheta_s)\, p(Z_k = 0 \mid p_0)} \\ &= \frac{p(\boldsymbol{Y} \mid Z_k = 1, \boldsymbol{Z}_{-k}, \vartheta_s)\, p_0}{p(\boldsymbol{Y} \mid Z_k = 1, \boldsymbol{Z}_{-k}, \vartheta_s)\, p_0 + p(\boldsymbol{Y} \mid Z_k = 0, \boldsymbol{Z}_{-k}, \vartheta_s)(1-p_0)} \\ &= \frac{p_0}{p_0 + \lambda_k(1-p_0)} \end{aligned} \tag{86}$$

where, $\lambda_k = \dfrac{p(\boldsymbol{Y} \mid Z_k = 0, \boldsymbol{Z}_{-k}, \vartheta_s)}{p(\boldsymbol{Y} \mid Z_k = 1, \boldsymbol{Z}_{-k}, \vartheta_s)}$. The marginal likelihood function $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \vartheta_s)$ was derived by integrating out the random variables θ and σ^2 from the original likelihood function, which follows,

$$p(\boldsymbol{Y} \mid \boldsymbol{Z}, \vartheta_s) = \int p\left(\boldsymbol{Y} \mid \boldsymbol{Z}, \vartheta_s, \sigma^2\right) p\left(\sigma^2\right) d\sigma^2 \tag{87}$$

where, $p\left(\boldsymbol{Y} \mid \boldsymbol{Z}, \vartheta_s, \sigma^2\right)$ is obtained by marginalizing the likelihood function with respect to the variable $\boldsymbol{\theta}$. Considering only the weights whose corresponding latent variables $Z_k = 0$, denoted by $\boldsymbol{\theta_r}$, the probability $p\left(\boldsymbol{Y} \mid \boldsymbol{Z}, \vartheta_s, \sigma^2\right)$ is obtained as,

$$\begin{aligned} p\left(\boldsymbol{Y} \mid \boldsymbol{Z}, \vartheta_s, \sigma^2\right) &= \int p\left(\boldsymbol{Y}, \boldsymbol{\theta} \mid \boldsymbol{Z}, \vartheta_s, \sigma^2\right) d\theta \\ &= \int p\left(\boldsymbol{Y} \mid \boldsymbol{\theta}_r, \sigma^2\right) p\left(\boldsymbol{\theta}_r \mid \vartheta_s, \sigma^2\right) d\boldsymbol{\theta}_r \\ &= \int \mathcal{N}\left(\boldsymbol{Y} \mid \mathbf{L}_r\boldsymbol{\theta}_r, \sigma^2\mathbf{I}_{N\times N}\right)\mathcal{N}\left(\boldsymbol{\theta}_r \mid 0, \sigma^2\vartheta_s\mathbf{R}_{0,r}\right) d\boldsymbol{\theta}_r \\ &= \int \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{(\boldsymbol{Y}-\mathbf{L}_r\boldsymbol{\theta}_r)^T(\boldsymbol{Y}-\mathbf{L}_r\boldsymbol{\theta}_r)}{2\sigma^2}\right) \frac{\left(|\mathbf{R}_{0,r}^{-1}|\right)^{1/2}}{(2\pi\vartheta_s\sigma^2)^{r/2}} \exp\left(-\frac{\boldsymbol{\theta}_r^T\mathbf{R}_{0,r}^{-1}\boldsymbol{\theta}_r}{2\sigma^2\vartheta_s}\right) d\boldsymbol{\theta}_r \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \frac{\left(|\mathbf{R}_{0,r}^{-1}|\right)^{1/2}\left(|\boldsymbol{\Sigma}^{-1}|\right)^{1/2}}{(\vartheta_s)^{r/2}} \exp\left(-\frac{\left(\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu}\right)}{2\Sigma^2}\right) \end{aligned} \tag{88}$$

where, $\boldsymbol{\Sigma} = \left(\mathbf{L}_r^T\mathbf{L}_r + \vartheta_s^{-1}\mathbf{R}_{0,r}^{-1}\right)^{-1}$ and $\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{L}_r^T\boldsymbol{Y}$. The operator $|.|$ denotes the determinant. With the above result then, $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \vartheta_s)$

can be derived as,

$$
\begin{aligned}
p(\boldsymbol{Y} \mid \boldsymbol{Z}, \vartheta_s) &= \frac{\left(|\mathbf{R}_{0,r}^{-1}|\right)^{1/2}\left(|\boldsymbol{\Sigma}^{-1}|\right)^{1/2}}{(2\pi)^{N/2}(\vartheta_s)^{r/2}} \int \frac{1}{(\sigma^2)^{N/2}} \exp\left(-\frac{\left(\boldsymbol{Y}^T\boldsymbol{Y}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu}\right)}{2\sigma^2}\right) \mathcal{IG}(\alpha_\sigma,\beta_\sigma)\, d\sigma^2 \\
&= \frac{\left(|\mathbf{R}_{0,r}^{-1}|\right)^{1/2}\left(|\boldsymbol{\Sigma}^{-1}|\right)^{1/2}}{(2\pi)^{N/2}(\vartheta_s)^{r/2}} \frac{(\beta_\sigma)^{\alpha_\sigma}}{\Gamma(\alpha_\sigma)} \int \frac{1}{(\sigma^2)^{\alpha_\sigma+N/2+1}} \exp\left(-\frac{\beta_\sigma+\frac{1}{2}(\boldsymbol{Y}^T\boldsymbol{Y}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu})}{\sigma^2}\right) d\sigma^2 \\
&= \frac{\left(|\mathbf{R}_{0,r}^{-1}|\right)^{1/2}\left(|\boldsymbol{\Sigma}^{-1}|\right)^{1/2}}{(2\pi)^{N/2}(\vartheta_s)^{r/2}} \frac{(b_\sigma)^{\alpha_\sigma}}{\Gamma(\alpha_\sigma)} \frac{\Gamma\left(\alpha_\sigma+\frac{N}{2}\right)}{\left(\beta_\sigma+\frac{1}{2}(\boldsymbol{Y}^T\boldsymbol{Y}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu})\right)^{\left(\alpha_\sigma+\frac{N}{2}\right)}}
\end{aligned}
\tag{89}
$$

where, the operator $\Gamma(.)$ denotes the Gamma function. To summarise, the random variables Z_k can be computed from the Bernoulli distribution with the parameter u_k as: $Z_k; k = 1, \ldots, K \quad \mid \boldsymbol{Y}, \vartheta_s, p_0 \sim Bern(u_k)$. Where, the marginalised likelihood function is obtained as,

$$
p(\boldsymbol{Y} \mid \boldsymbol{Z}, \vartheta_s) = \begin{cases}
\dfrac{\Gamma\left(\alpha_\sigma+\dfrac{N}{2}\right)}{(2\pi)^{N/2}\dfrac{(\beta_\sigma)^{\alpha_\sigma}}{\Gamma(\alpha_\sigma)}} \dfrac{1}{\left(\beta_\sigma+\dfrac{1}{2}(\boldsymbol{Y}^T\boldsymbol{y})\right)^{\left(\alpha_\sigma+\dfrac{N}{2}\right)}} & \text{, when all } \{Z_k; k = 1, \ldots, K\} = 0 \\
\dfrac{\Gamma\left(a_\sigma+\dfrac{N}{2}\right)}{(2\pi)^{N/2}(\vartheta_s)^{r/2}} \dfrac{(\beta_\sigma)^{\alpha_\sigma}}{\Gamma(\alpha_\sigma)} \dfrac{\left(|\mathbf{R}_{0,r}^{-1}|\right)^{1/2}\left(|\boldsymbol{\Sigma}^{-1}|\right)^{1/2}}{\left(\beta_\sigma+\dfrac{1}{2}(\boldsymbol{Y}^T\boldsymbol{Y}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu})\right)^{\left(\alpha_\sigma+\dfrac{N}{2}\right)}} & \text{, otherwise}
\end{cases}
\tag{90}
$$