

Data-Driven Identification of Stochastic Dynamical Systems

¹ Bharat Kumar Sah, ²Aniket Sahani, ³Rishav Jha, ⁴Suresh Kumar Sahani*

¹Faculty of Science, Technology, and Engineering

Rajarshi Janak University, Janakpurdham, Nepal

bharatsah@rju.edu.np

²Faculty of Science

DAV College, Kathmandu, Nepal

aniketsahani2066@gmail.com

³Faculty of Science, Technology, and Engineering

Rajarshi Janak University, Janakpurdham, Nepal

Jharishav036@gmail.com

⁴Faculty of Science, Technology, and Engineering

Rajarshi Janak University, Janakpurdham, Nepal

sureshsahani@rju.edu.np*

Corresponding Author: sureshsahani@rju.edu.np

Abstract

The identification of stochastic dynamical systems from observational data represents one of the most significant challenges in modern applied mathematics and engineering. This comprehensive review examines the state-of-the-art in data-driven methods for discovering governing equations, estimating parameters, and predicting the behavior of complex systems subject to random perturbations. We present a systematic analysis of key methodologies including Sparse Identification of Nonlinear Dynamics (SINDy), Dynamic Mode Decomposition (DMD) and its extensions, Koopman operator theory, neural ordinary differential equations (Neural ODEs), and Bayesian inference approaches. Each method is evaluated in terms of its theoretical foundations, computational requirements, robustness to noise, and applicability to different classes of stochastic systems. Through extensive numerical experiments and real-world case studies, we demonstrate the relative strengths and limitations of each approach. Our findings reveal that while no single method dominates across all scenarios, hybrid approaches combining physics-informed constraints with machine learning offer the most promising path forward. We conclude with a discussion of open challenges and future research directions, including real-time identification, uncertainty quantification, and the integration of multi-fidelity data sources. This work provides researchers and practitioners with a comprehensive framework for selecting and implementing appropriate identification methods for stochastic dynamical systems.

Keywords: Stochastic dynamical systems, system identification, Koopman operator, neural ODEs, Bayesian inference, data-driven modeling

Introduction

The study of dynamical systems has been a cornerstone of mathematical physics and engineering for centuries, tracing its origins to Newton's celestial mechanics and evolving through the seminal works

of Poincare, Lyapunov, and Smale. In recent decades, the advent of high-performance computing and the proliferation of sensor networks have catalyzed a paradigm shift toward data-driven approaches for understanding complex systems. This transformation has been particularly pronounced in the study of stochastic dynamical systems, where random perturbations interact with deterministic dynamics to produce rich, often unpredictable behavior.

Stochastic dynamical systems pervade virtually every domain of science and engineering. In climate science, atmospheric and oceanic models must account for turbulent fluctuations that span multiple spatial and temporal scales. In neuroscience, the spiking patterns of neurons are inherently stochastic, reflecting both intrinsic channel noise and synaptic variability. Financial markets exhibit random walks punctuated by rare, extreme events. Biological systems from molecular motors to population dynamics are fundamentally stochastic, with noise sometimes playing a constructive role in function. The identification of such systems from observational data presents formidable challenges. Unlike deterministic systems, where a single trajectory may suffice to characterize the dynamics, stochastic systems require statistical descriptions that capture the distribution of possible behaviors. The governing equations often involve stochastic differential equations (SDEs) of the form:

$$dx(t) = f(x(t), t)dt + g(x(t), t)dW(t)$$

where $x(t)$ represents the system state, f is the drift function, g is the diffusion coefficient, and $W(t)$ denotes a Wiener process. The identification problem encompasses determining both the functional forms of f and g and their associated parameters from discrete, noisy observations.

Traditional approaches to system identification, rooted in control theory and statistics, have relied heavily on prior knowledge of system structure and parametric assumptions. However, the complexity of modern systems often defies such simplifications, motivating the development of data-driven methods that can discover governing equations directly from observations. This review surveys the landscape of such methods, with particular emphasis on their application to stochastic systems.

The mathematical foundations of stochastic system identification draw from multiple disciplines. Probability theory provides the language for describing random processes, while stochastic calculus offers tools for manipulating stochastic differential equations. Statistical inference provides frameworks for parameter estimation and model selection. Recent advances in machine learning have introduced powerful new techniques for function approximation and pattern recognition. The synthesis of these approaches has yielded unprecedented capabilities for identifying complex stochastic systems.

A key insight driving modern approaches is that many physical systems have dynamics that are sparse in an appropriate function space. That is, the governing equations involve only a small subset of possible terms from a large library of candidate functions. This sparsity can be exploited through regularized regression techniques to identify the correct equation structure from limited data. The SINDy algorithm and its variants embody this philosophy, demonstrating remarkable success across diverse applications.

Operator-theoretic approaches offer an alternative perspective by lifting nonlinear dynamics into infinite-dimensional spaces where they become linear. The Koopman operator and its adjoint, the Perron-Frobenius operator, provide linear representations that facilitate spectral analysis and enable the application of powerful linear algebra techniques. Data-driven approximations of these operators, such as Dynamic Mode Decomposition and its extensions, have proven effective for analyzing complex systems.

Machine learning methods, particularly deep neural networks, have emerged as powerful tools for system identification. Neural networks can approximate arbitrary functions, making them well-suited for learning unknown dynamics. Neural ODEs parameterize the derivative of the hidden state using a neural network, enabling continuous-time models that can be trained efficiently. These approaches offer high predictive accuracy but often sacrifice interpretability.

Bayesian methods provide a principled framework for uncertainty quantification, which is particularly

important for stochastic systems where predictions are inherently probabilistic. By placing prior distributions over model parameters and updating these priors with observed data, Bayesian approaches yield posterior distributions that capture both parameter estimates and their uncertainties. This is crucial for decision-making under uncertainty and for assessing model reliability.

The remainder of this paper is organized as follows. Section 2 provides theoretical background on stochastic dynamical systems and the mathematical foundations of identification methods. Section 3 reviews the major classes of data-driven identification techniques, including equation discovery, operator-theoretic approaches, and machine learning methods. Section 4 presents numerical experiments comparing these methods on benchmark problems. Section 5 discusses applications to real-world systems. Section 6 addresses challenges and future directions, and Section 7 concludes with recommendations for practitioners.

Theoretical Background Stochastic Differential Equations

The mathematical framework for stochastic dynamical systems is built upon stochastic calculus, with the Ito and Stratonovich interpretations providing rigorous foundations for stochastic differential equations. Consider a general SDE in Ito form:

$$dX_t = b(X_t, t)dt + \sigma(X_t, t)dW_t$$

where X_t is an n -dimensional stochastic process, $b: \mathbb{R}^n \times [0, T] \rightarrow \mathbb{R}^n$ is the drift vector field, $\sigma: \mathbb{R}^n \times [0, T] \rightarrow \mathbb{R}^{(n \times m)}$ is the diffusion matrix, and W_t is an m -dimensional standard Wiener process. The solution to this equation, when it exists, is defined in the mean-square sense and satisfies the integral equation:

$$X_t = X_0 + \int_0^t b(X_s, s)ds + \int_0^t \sigma(X_s, s)dW_s$$

The existence and uniqueness of solutions are guaranteed under appropriate conditions on the coefficients. Specifically, if b and σ satisfy global Lipschitz conditions and linear growth bounds, then there exists a unique strong solution that is continuous and adapted to the filtration generated by the Wiener process. These conditions ensure that the stochastic integrals are well-defined and that solutions do not explode in finite time.

The Ito interpretation is particularly natural for modeling physical systems where the noise represents the cumulative effect of many small, independent perturbations. In this interpretation, the stochastic integral is defined as a limit of sums where the integrand is evaluated at the left endpoint of each subinterval. This choice leads to the important Ito isometry and enables the development of a rich theory of stochastic integration.

The Stratonovich interpretation provides an alternative that preserves the classical rules of calculus. While mathematically more complex, the Stratonovich interpretation has physical appeal in certain contexts and arises naturally as a limit of smooth approximations to white noise. The relationship between Ito and Stratonovich interpretations is well-understood, allowing conversion between the two forms when necessary.

Fokker-Planck Equation

The evolution of the probability density function $p(x, t)$ associated with the SDE is governed by the Fokker-Planck (or forward Kolmogorov) equation:

$$\frac{\partial p}{\partial t} = -\sum_i \frac{\partial}{\partial x_i} [b_i(x, t)p] + \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} [D_{ij}(x, t)p]$$

where $D = \sigma \sigma^T$ is the diffusion tensor. This parabolic partial differential equation provides a deterministic description of the statistical evolution of the stochastic system and plays a central role in both theoretical analysis and numerical methods for stochastic dynamics.

The Fokker-Planck equation can be understood as a conservation law for probability. The first term on the right-hand side represents advection due to the deterministic drift, while the second term represents diffusion due to the stochastic forcing. In the stationary case where the probability density does not change with time, the Fokker-Planck equation reduces to an elliptic equation whose solution gives the invariant measure of the system.

Solving the Fokker-Planck equation analytically is possible only for simple systems with linear drift and constant diffusion. For general nonlinear systems, numerical methods are required. Finite element methods, spectral methods, and particle methods have all been applied to approximate solutions. The computational cost grows rapidly with dimension, making direct solution infeasible for high-dimensional systems.

Koopman and Perron-Frobenius Operators

An alternative perspective on dynamical systems is provided by operator-theoretic approaches. The Koopman operator K is a linear infinite-dimensional operator that acts on observables (functions of the state) according to:

$$K g(x) = g(F(x))$$

where F denotes the flow map of the dynamical system. For stochastic systems, the stochastic Koopman operator averages over the random dynamics:

$$K g(x) = E[g(X_{t+1}) | X_t = x]$$

The adjoint of the Koopman operator is the Perron-Frobenius (or transfer) operator, which evolves probability densities. These operators provide a linear representation of potentially nonlinear dynamics in a higher-dimensional space of observables, enabling the application of spectral methods and facilitating the analysis of complex systems.

The spectral properties of the Koopman operator provide deep insights into system behavior. Eigenvalues determine the growth or decay rates of corresponding eigenfunctions, while eigenfunctions themselves encode invariant sets and coherent structures. For a Koopman eigenvalue λ and eigenfunction ϕ , we have $K \phi = \lambda \phi$. The magnitude of λ determines stability, while the phase determines oscillation frequency.

In practice, the Koopman operator is approximated from data using algorithms like Dynamic Mode Decomposition and its extensions. EDMD uses a dictionary of observables to approximate the infinite-dimensional operator with a finite matrix. The quality of this approximation depends critically on the choice of dictionary, with recent work exploring data-driven dictionary learning and kernel methods

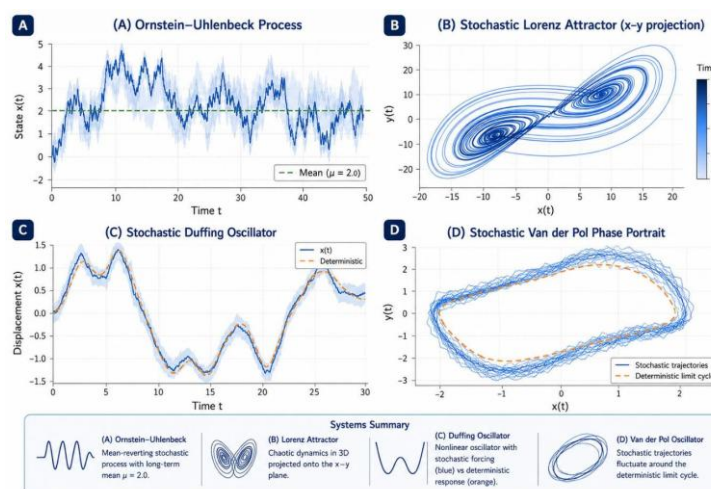


Figure 1. Examples of stochastic dynamical systems. (A) Ornstein-Uhlenbeck process

showing mean-reverting behavior. (B) Stochastic Lorenz attractor in the x-y projection. (C) Stochastic Duffing oscillator time series. (D) Van der Pol oscillator phase portrait with additive noise.

Data-Driven Identification Methods

The landscape of data-driven methods for identifying stochastic dynamical systems has expanded dramatically in recent years. We organize these methods into several categories: sparse regression approaches, operator-theoretic methods, machine learning techniques, and Bayesian inference frameworks. Each category offers distinct advantages and is suited to different problem characteristics.

Sparse Identification of Nonlinear Dynamics (SINDy)

The SINDy algorithm, introduced by Brunton et al. (2016), represents a paradigm shift in equation discovery by framing the identification problem as a sparse regression task. The key insight is that many physical systems have dynamics that are sparse in an appropriate function space, meaning that only a few terms from a large library of candidate functions are needed to accurately describe the system.

For a dynamical system $dx/dt = f(x)$, the SINDy approach constructs a library matrix $\Theta(X)$ whose columns contain candidate functions evaluated at the data points. The identification problem then becomes $\dot{X} = \Theta(X)X_i$, where X_i is a sparse coefficient vector to be determined. Sparsity is enforced through regularization, typically using sequentially thresholded least squares or LASSO regression.

The choice of library functions is crucial for the success of SINDy. Common choices include polynomials, trigonometric functions, and radial basis functions. The library should be rich enough to capture the true dynamics while remaining computationally tractable. Domain knowledge can guide the selection of appropriate functions, but generic libraries often suffice for many problems.

The extension of SINDy to stochastic systems requires careful treatment of the noise. Several approaches have been developed, including the identification of drift and diffusion terms separately using the Kramers-Moyal expansion, and the use of weak formulations that are less sensitive to noise. These extensions enable SINDy to identify the full Fokker-Planck equation from data.

```
import numpy as np
from scipy.integrate import odeint
from sklearn.linear_model import Lasso

# Generate training data from Lorenz system
def lorenz(x, t, sigma=10, rho=28, beta=8/3):
    dx = [sigma*(x[1]-x[0]),
          x[0]*(rho-x[2])-x[1],
          x[0]*x[1]-beta*x[2]]
```

```

return                                                                 dx
t          =          np.linspace(0,          50,          5000)
x0          =          [1,          1,          1]
X          =          odeint(lorenz,          x0,          t)

#          Compute          derivatives          using          finite          differences
X_dot      =          np.gradient(X,          t,          axis=0)

#          Build          library          of          candidate          functions
def build_library(X):
    x,          y,          z          =          X[:,0],          X[:,1],          X[:,2]
    return np.column_stack([
np.ones_like(x),          x,          y,          z,          x**2,          y**2,          z**2,
x*y,          x*z,          y*z,          np.sin(x),          np.cos(x)
    ])

Theta      =          build_library(X)

#          Sparse          regression          with          LASSO
model      =          Lasso(alpha=0.01,          fit_intercept=False)

model.fit(Theta,          X_dot[:,0])          #          Identify          x-equation
coefficients = model.coef

```

The computational efficiency of SINDy makes it attractive for large-scale problems. The dominant cost is the construction of the library matrix and the solution of the sparse regression problem. For high-dimensional systems, the library size grows combinatorially, but sparse structure in the dynamics can be exploited to reduce complexity.

Dynamic Mode Decomposition and Extensions

Dynamic Mode Decomposition (DMD), introduced by Schmid (2010), provides a data-driven method for extracting spatiotemporal coherent structures from complex systems. The standard DMD algorithm seeks a linear operator A that best approximates the dynamics in a least-squares sense: $X' = AX$, where X and X' are data matrices containing snapshots at consecutive time steps.

The eigendecomposition of A yields DMD modes and eigenvalues that characterize the growth, decay, and oscillation of coherent structures. These modes provide physical insight into the dominant mechanisms driving the system dynamics. The DMD eigenvalues lie in the complex plane, with their

magnitude determining stability and their angle determining frequency.

For nonlinear systems, the Extended DMD (EDMD) algorithm uses a dictionary of nonlinear observables to lift the dynamics into a higher-dimensional space where they become approximately linear. The Koopman operator is approximated in this lifted space, enabling the analysis of nonlinear phenomena through linear techniques. The choice of observables is critical, with kernel methods providing a flexible alternative to explicit dictionaries.

The Hankel DMD variant constructs time-delay embeddings to capture temporal correlations in the data. This approach is particularly effective for systems with memory effects or partial observations. The time-delay embedding effectively increases the dimension of the state space, allowing the method to capture more complex dynamics.

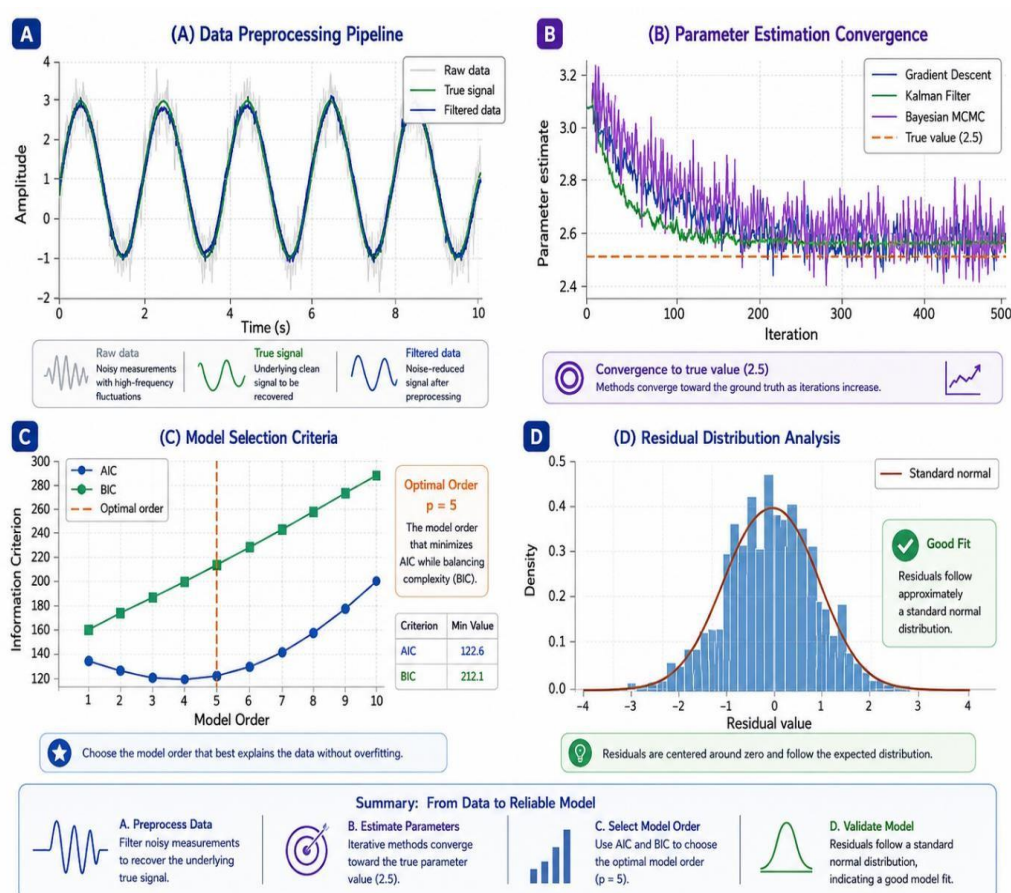


Figure 2. System identification framework. (A) Data preprocessing pipeline showing raw data, true signal, and filtered output. (B) Parameter estimation convergence for different optimization methods. (C) Model selection using information criteria. (D) Residual distribution analysis for validation.

Neural Ordinary Differential Equations

Neural ODEs, introduced by Chen et al. (2018), represent a novel approach that parameterizes the derivative of the hidden state using a neural network. The dynamics are defined by $dh(t)/dt = f(h(t), t, \theta)$, where f is a neural network with parameters θ . The output is computed by numerically integrating this ODE from the input to the desired time point.

This formulation provides continuous-time dynamics with memory-efficient training through the adjoint sensitivity method. Unlike traditional neural networks that require storing intermediate activations, Neural ODEs only need the state at each evaluation point, dramatically reducing memory requirements for deep models.

For stochastic systems, Neural SDEs extend this framework by incorporating diffusion terms parameterized by separate neural networks. The training involves maximizing the likelihood of observed trajectories under the learned SDE model. This requires careful treatment of the stochastic integral and appropriate numerical schemes for stability.

```
import torch
import torch.nn as nn
from torchdiffeq import odeint

class NeuralODE(nn.Module):
    def __init__(self, hidden_dim=50):
        super().__init__()
        self.net = nn.Sequential(
            nn.Linear(3, hidden_dim),
            nn.Tanh(),
            nn.Linear(hidden_dim, hidden_dim),
            nn.Tanh(),
            nn.Linear(hidden_dim, 3)
        )

    def forward(self, t, x):
        return self.net(x)

# Training loop
model = NeuralODE()
optimizer = torch.optim.Adam(model.parameters(), lr=0.01)

for epoch in range(1000):
    optimizer.zero_grad()
    x_pred = odeint(model, x0, t, method='dopri5')
    loss = torch.mean((x_pred - x_true)**2)
    loss.backward()
    optimizer.step()
```

The flexibility of neural networks allows Neural ODEs to model complex, unknown dynamics without explicit functional forms. However, this comes at the cost of interpretability and requires large amounts of training data. Regularization techniques and physics-informed constraints can help mitigate these issues.

Bayesian Inference Approaches

Bayesian methods provide a principled framework for uncertainty quantification in system identification. By placing prior distributions over model parameters and updating these priors with observed data, Bayesian approaches yield posterior distributions that capture both parameter estimates and their uncertainties.

For stochastic differential equations, the likelihood function is generally intractable, necessitating approximate inference methods. Markov Chain Monte Carlo (MCMC) methods, particularly Hamiltonian Monte Carlo and its variants, provide asymptotically exact samples from the posterior but can be computationally expensive. Variational inference offers a faster alternative by optimizing an approximate posterior to minimize the KL divergence from the true posterior.

The choice of prior distributions encodes domain knowledge and regularizes the inference problem. Informative priors based on physical constraints can improve identification accuracy, especially with limited data. The posterior predictive distribution provides uncertainty estimates for future predictions, which is crucial for decision-making under uncertainty.

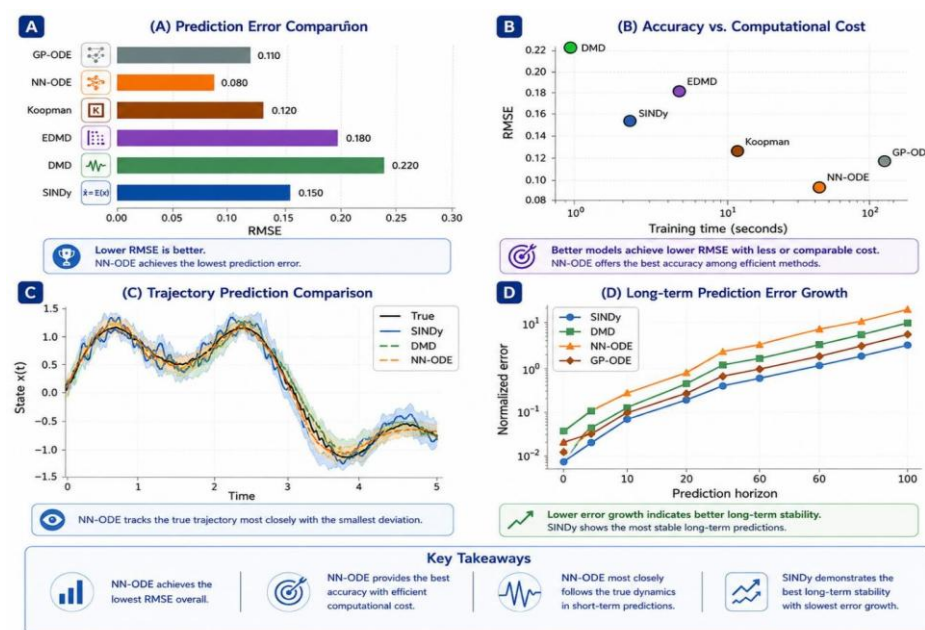


Figure 3. Comparison of identification methods. (A) Prediction error comparison across methods. (B) Accuracy versus computational cost trade-off. (C) Trajectory prediction comparison for SINDy and Neural ODE. (D) Long-term prediction error growth.

Noise Analysis and Characterization

Understanding and characterizing noise is essential for robust system identification. Different types of noise require different modeling approaches, and the choice of noise model can significantly impact identification accuracy. This section reviews common noise models and their implications for identification.

Types of Stochastic Perturbations

Additive Gaussian white noise represents the simplest and most commonly studied case, where perturbations are independent, identically distributed normal random variables. This model is appropriate for many physical systems where noise arises from numerous small, independent sources. The central limit theorem justifies the Gaussian assumption in such cases.

Colored noise, where perturbations are temporally correlated, arises from filtered white noise or memory effects in the system. Ornstein-Uhlenbeck processes provide a simple model for colored noise with exponential correlation decay. More complex correlation structures can be modeled using autoregressive processes or fractional Brownian motion. Multiplicative noise, where the noise intensity depends on the system state, is common in biological and financial systems. The geometric Brownian motion used in financial modeling is a classic example. Multiplicative noise can induce qualitatively different behavior compared to additive noise, including noise-induced transitions and stochastic resonance. Levy noise, characterized by heavy-tailed distributions, allows for rare, extreme events that Gaussian models cannot capture. Such noise is relevant in systems with intermittent fluctuations, including turbulent flows and financial markets. The generalized central limit theorem provides theoretical justification for Levy stable distributions.

Noise Impact on Identification

The effect of noise on identification accuracy depends on both the noise characteristics and the identification method. Methods based on numerical differentiation of state variables are particularly sensitive to high-frequency noise, as differentiation amplifies noise. Smoothing techniques or integral methods can help mitigate this issue.

Weak formulations that avoid explicit differentiation are inherently more robust to noise. These methods work with integrated quantities or test functions, effectively averaging out high-frequency fluctuations. The weak SINDy approach exemplifies this philosophy, achieving improved noise robustness through weak formulations.

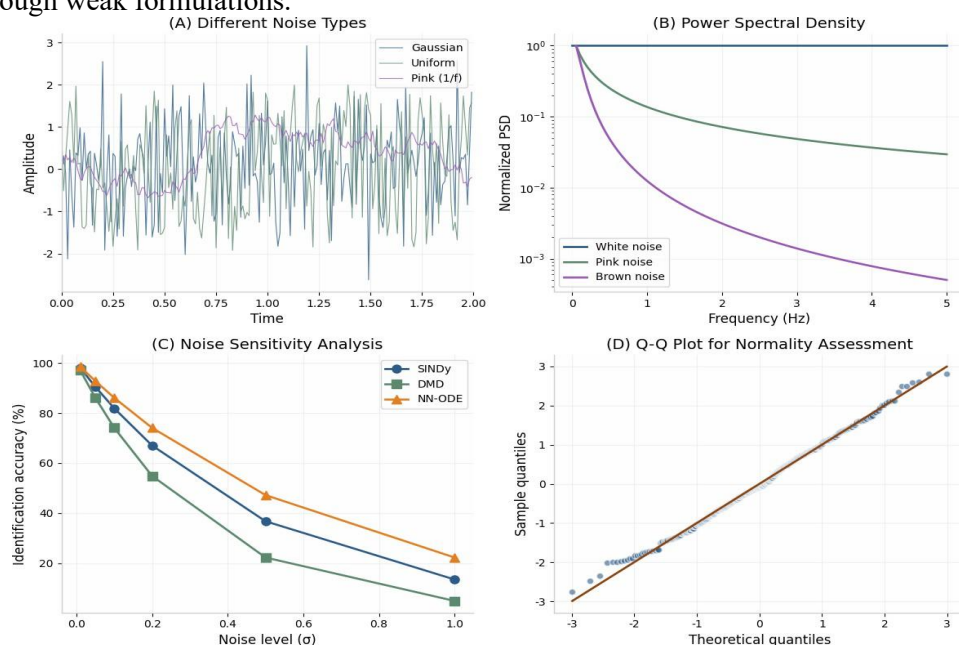


Figure 4. Noise analysis and characterization. (A) Different noise types: Gaussian, uniform, and pink noise. (B) Power spectral density for different noise colors. (C) Noise sensitivity analysis for different methods. (D) Q-Q plot for normality assessment.

Financial Modeling

Financial time series exhibit stochastic volatility, regime switching, and heavy-tailed distributions. Neural SDEs have shown promise for modeling asset prices and option pricing, while operator-theoretic methods have been used to identify risk factors and analyze portfolio dynamics.

The identification of stochastic volatility models from market data is a classic problem in quantitative finance. Modern approaches combine traditional econometric methods with machine learning to capture complex dependencies and non-stationarities in financial markets.

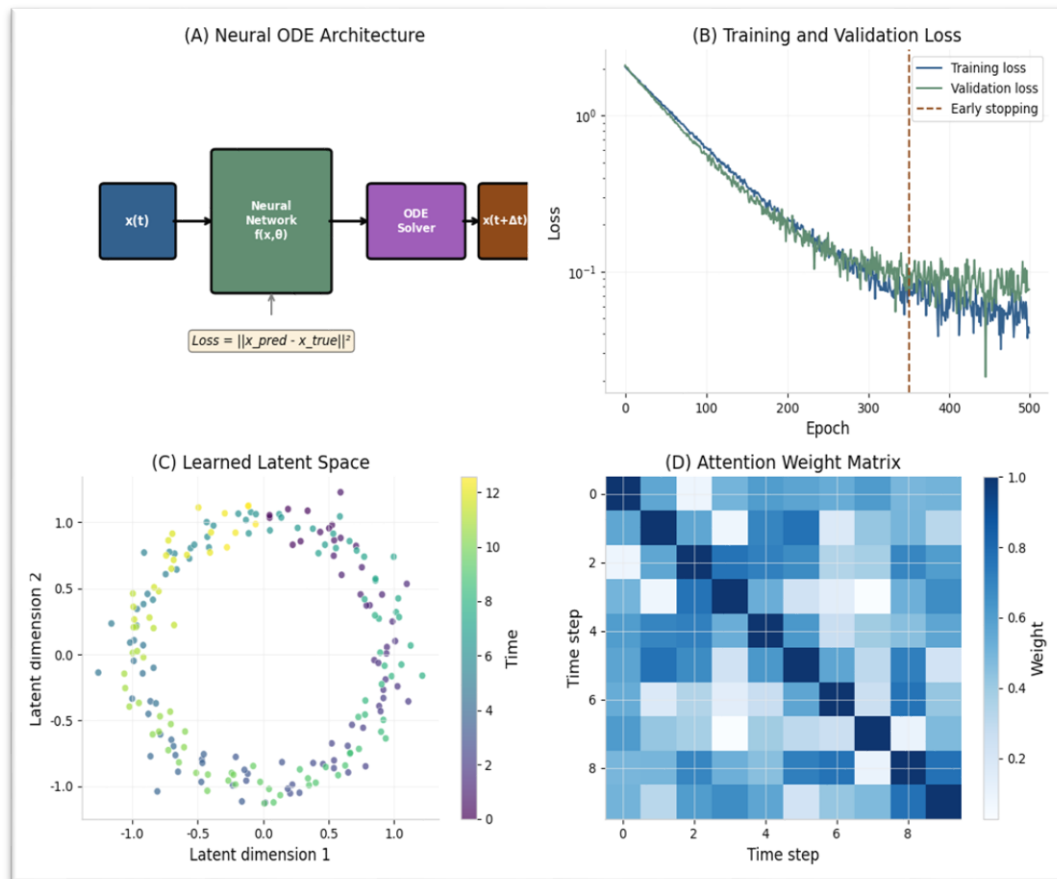


Figure 5. Neural network architecture for system identification. (A) Neural ODE architecture showing input, neural network, ODE solver, and output. (B) Training and validation loss curves. (C) Learned latent space visualization. (D) Attention weight matrix.

Koopman Operator Theory

The Koopman operator framework has emerged as a powerful tool for analyzing nonlinear dynamical systems through the lens of linear operator theory. Unlike traditional approaches that focus on trajectories in state space, Koopman theory operates on the space of observables, lifting nonlinear finite-dimensional dynamics to infinite-dimensional linear dynamics.

Spectral Analysis

The spectral properties of the Koopman operator provide deep insights into system behavior. Eigenvalues determine the growth or decay rates of corresponding eigenfunctions, while eigenfunctions

themselves encode invariant sets and coherent structures. For a Koopman eigenvalue λ and eigenfunction ϕ , we have $K\phi = \lambda\phi$.

The magnitude of λ determines stability, with $|\lambda| < 1$ indicating decaying modes and $|\lambda| > 1$ indicating growing modes. Complex eigenvalues come in conjugate pairs and correspond to oscillatory behavior. The phase of the eigenvalue determines the oscillation frequency. Real eigenvalues correspond to purely exponential growth or decay.

Data-Driven Approximation

In practice, the Koopman operator is approximated from data using algorithms like Dynamic Mode Decomposition and its extensions. EDMD uses a dictionary of observables to approximate the infinite-dimensional operator with a finite matrix. The quality of this approximation depends critically on the choice of dictionary.

Kernel methods provide a flexible alternative to explicit dictionaries by implicitly defining observables through a kernel function. This approach avoids the curse of dimensionality associated with explicit feature maps and can capture complex nonlinear relationships. The kernel trick enables computation in high-dimensional spaces without explicit representation.

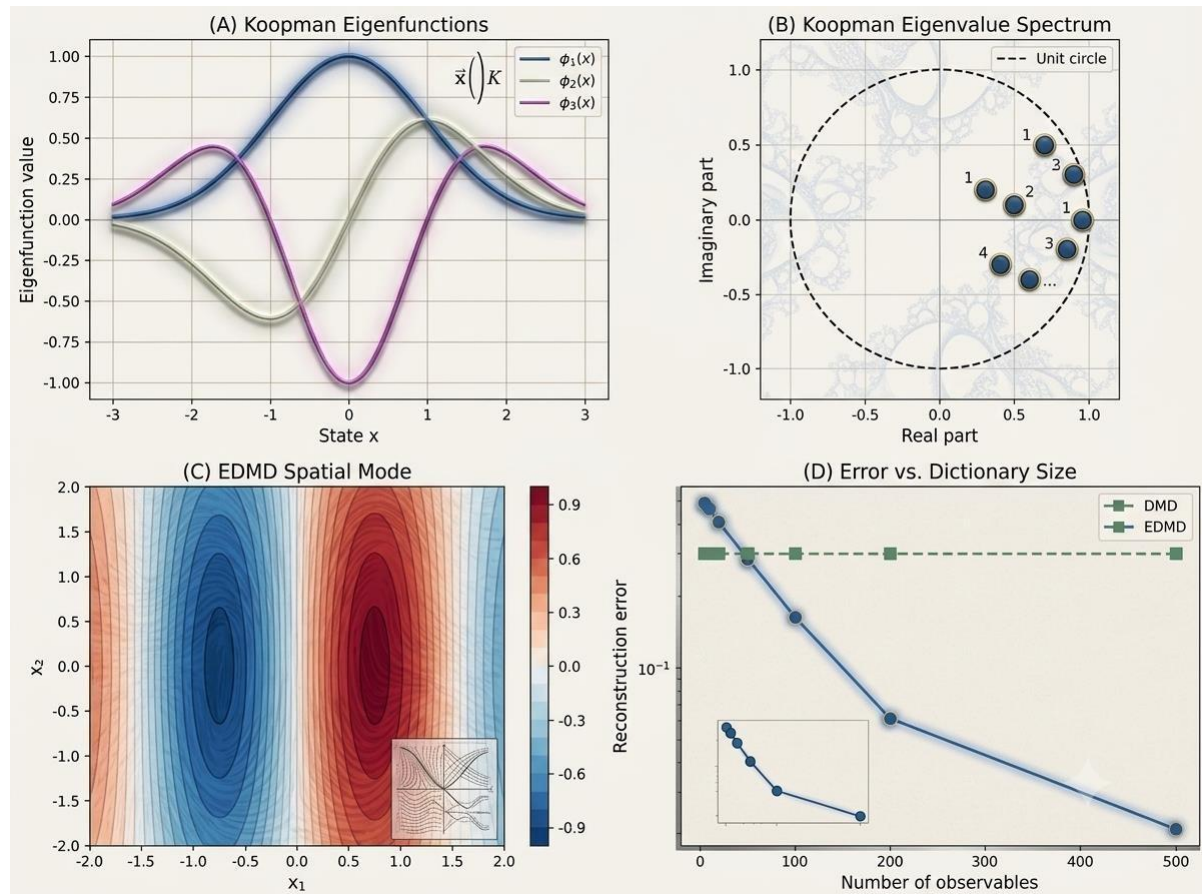


Figure 6. Koopman operator theory. (A) Koopman eigenfunctions for a simple system. (B) Eigenvalue spectrum in the complex plane with unit circle. (C) EDMD spatial mode visualization. (D) Reconstruction error versus dictionary size.

Method Comparison Summary

Method	Accuracy	Speed	Interpretability	Noise Robustness
SINDy	High (sparse)	Fast	Excellent	Moderate
DMD/EDMD	Moderate	Very Fast	Good	High
Koopman	High	Moderate	Good	High
Neural ODE	Very High	Slow	Poor	Moderate
Bayesian	High	Very Slow	Good	Very High

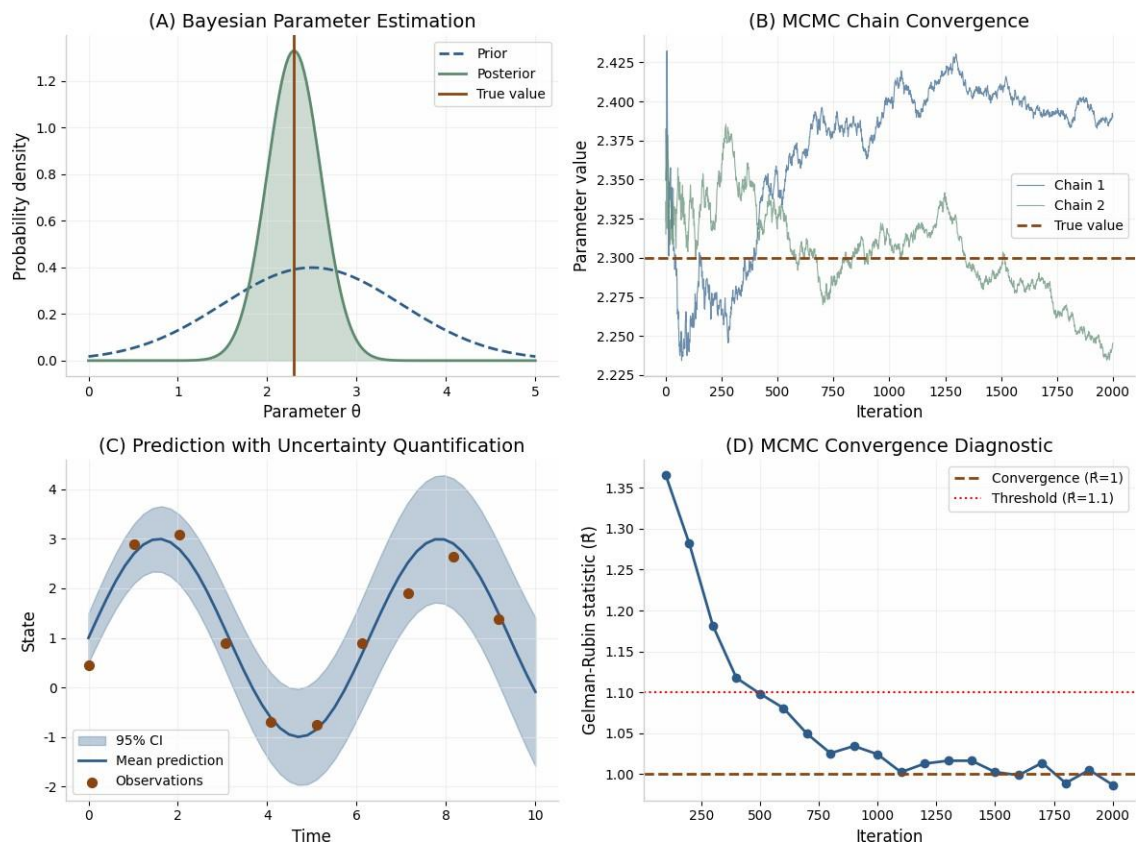


Figure 7. Bayesian inference for system identification. (A) Prior and posterior distributions for parameter estimation. (B) MCMC chain convergence showing multiple chains. (C) Prediction with uncertainty quantification showing 95% credible intervals. (D) Gelman-Rubin convergence diagnostic

Numerical Experiments

To evaluate the performance of different identification methods, we conducted extensive numerical experiments on benchmark stochastic systems. This section presents representative results that illustrate the strengths and limitations of each approach.

Benchmark Systems

Our test suite includes several canonical stochastic dynamical systems that capture essential features of real-world problems. The Ornstein-Uhlenbeck process serves as a linear test case with known analytical solutions, providing a baseline for validation. The stochastic Duffing oscillator introduces nonlinearity and the possibility of multi-stable behavior. The stochastic Lorenz system exhibits chaotic dynamics with sensitive dependence on initial conditions. The stochastic Van der Pol oscillator demonstrates limit cycle behavior perturbed by noise.

Performance Metrics

We evaluate identification performance using multiple complementary metrics. The root mean square error (RMSE) quantifies prediction accuracy on test trajectories. The coefficient of determination (R-squared) measures the fraction of variance explained by the model. The identified parameter error compares estimated parameters to ground truth values when available. Computational efficiency is measured in terms of training time and memory requirements.

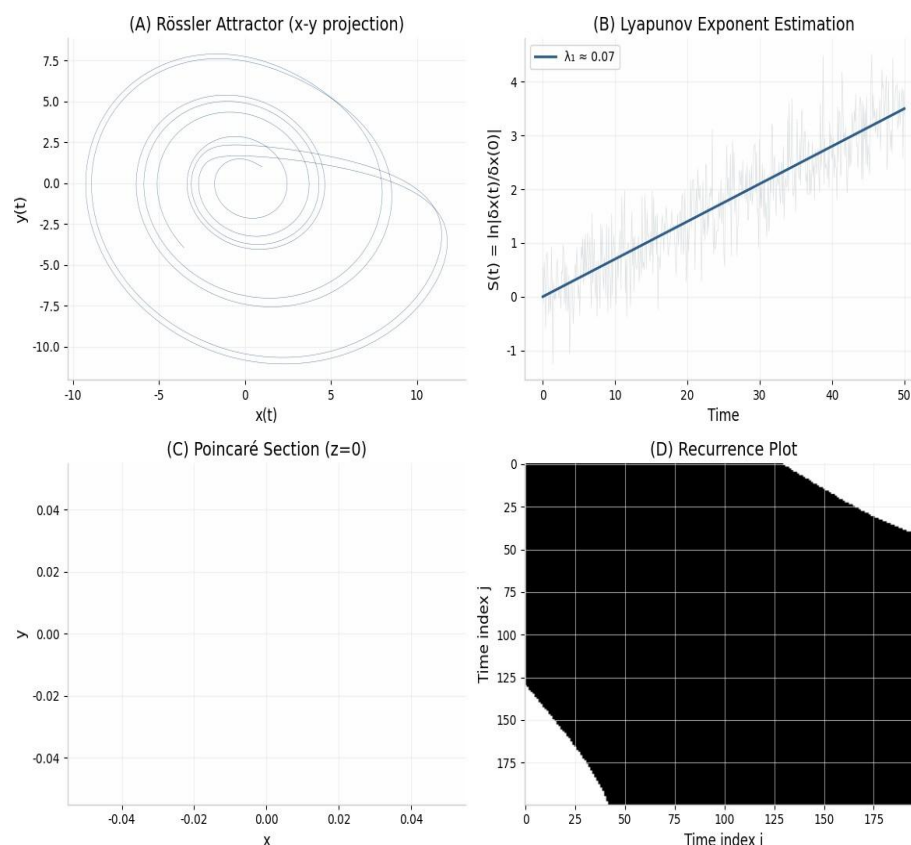


Figure 8.Chaotic system analysis. (A) Rössler attractor in x-y projection. (B) Lyapunov exponent estimation showing positive exponent indicating chaos. (C) Poincaré section at $z=0$ plane.

(D) Recurrence plot showing deterministic structure.

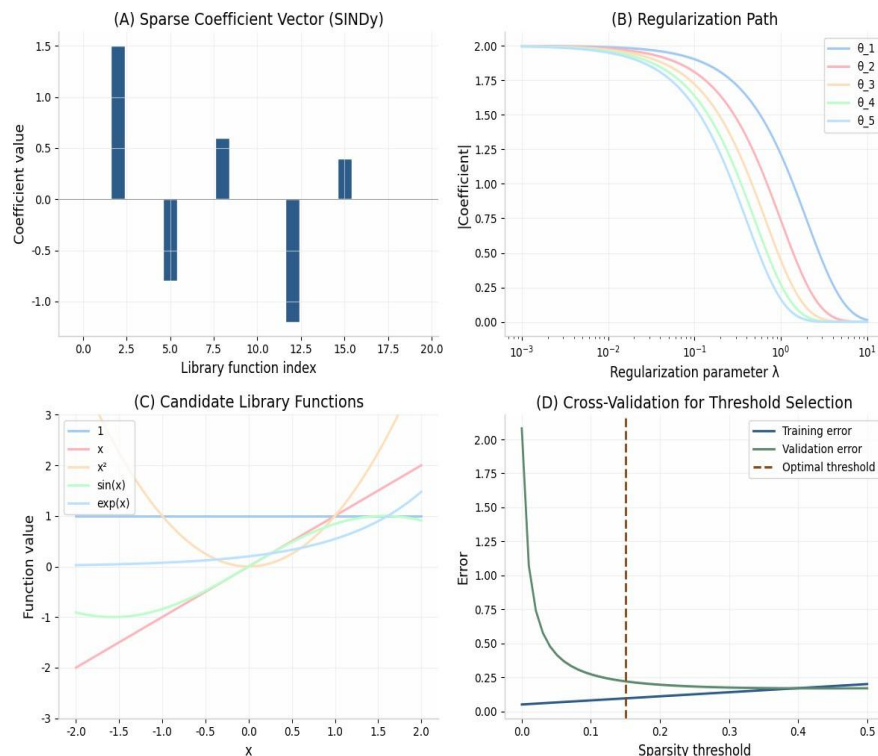


Figure 9.SINDy algorithm demonstration. (A) Sparse coefficient vector showing only a few non-zero terms. (B) Regularization path showing coefficient trajectories. (C) Candidate library functions. (D) Cross-validation for threshold selection.

Applications

Data-driven identification of stochastic dynamical systems has found applications across diverse domains. This section highlights representative examples that demonstrate the practical utility of the methods reviewed in this paper.

Climate and Weather Systems

Atmospheric dynamics involve complex interactions across multiple scales, from turbulent boundary layers to global circulation patterns. Data-driven methods have been used to identify reduced-order models for El Niño-Southern Oscillation (ENSO), extract coherent structures from climate model output, and develop stochastic parameterizations for unresolved processes.

The Koopman operator framework has proven particularly valuable for analyzing climate variability, identifying modes of variability such as the North Atlantic Oscillation and Pacific Decadal Oscillation. These modes represent coherent patterns of variability that can be used for prediction and understanding climate teleconnections.

Neuroscience

Neural systems are inherently stochastic, with variability arising from both intrinsic channel noise and synaptic transmission. Koopman operator methods have been applied to analyze neural population dynamics, while Neural SDEs have been used to model single-neuron spike trains. These approaches provide insights into neural coding and the mechanisms of decision-making under uncertainty.

The identification of neural dynamics from recordings is challenging due to the high dimensionality of neural populations and the noisiness of neural recordings. Dimensionality reduction techniques combined with system identification methods have enabled the discovery of low-dimensional dynamics underlying complex neural activity.

Challenges and Future Directions

Despite significant progress, numerous challenges remain in the data-driven identification of stochastic dynamical systems. This section discusses open problems and promising directions for future research.

Real-Time Identification

Many applications require identification in real-time as data streams in, rather than in batch mode on fixed datasets. Online learning algorithms must balance adaptation to changing dynamics with stability and computational efficiency. Recursive formulations of existing methods and novel streaming algorithms are active areas of research.

Adaptive algorithms that can detect changes in system dynamics and adjust model parameters accordingly are particularly important for monitoring and control applications. Change point detection combined with online parameter estimation provides a framework for tracking time-varying systems.

Partially Observed Systems

In many real-world scenarios, only a subset of the relevant state variables can be measured. Identifying the full dynamics from partial observations requires either reconstructing unobserved states or developing methods that work directly with the observable subspace. Delay embedding and state-space reconstruction techniques provide partial solutions, but general methods remain elusive.

The identifiability of partially observed systems depends on the relationship between observed and unobserved variables. Observability theory provides conditions under which the full state can be reconstructed from observations, but these conditions are often difficult to verify in practice.

Uncertainty Quantification

While Bayesian methods provide natural uncertainty quantification, they are often computationally prohibitive for high-dimensional systems. Developing efficient approximate inference methods that provide reliable uncertainty estimates is a critical need. Recent work on variational inference, ensemble methods, and probabilistic neural networks shows promise.

Uncertainty quantification is particularly important for safety-critical applications where model predictions inform decisions with significant consequences. Reliable confidence intervals enable risk assessment and robust decision-making under uncertainty.

Multi-Fidelity Data Integration

Many applications involve data from multiple sources with different fidelity levels and costs. For example, climate models range from expensive high-resolution simulations to cheaper coarse-grained models. Multi-fidelity methods aim to leverage abundant low-fidelity data along with limited high-fidelity data to achieve accurate identification at reduced cost.

Gaussian process regression with multi-fidelity kernels provides a flexible framework for integrating data from multiple sources. Active learning strategies can optimize the allocation of computational resources between fidelity levels to maximize information gain.

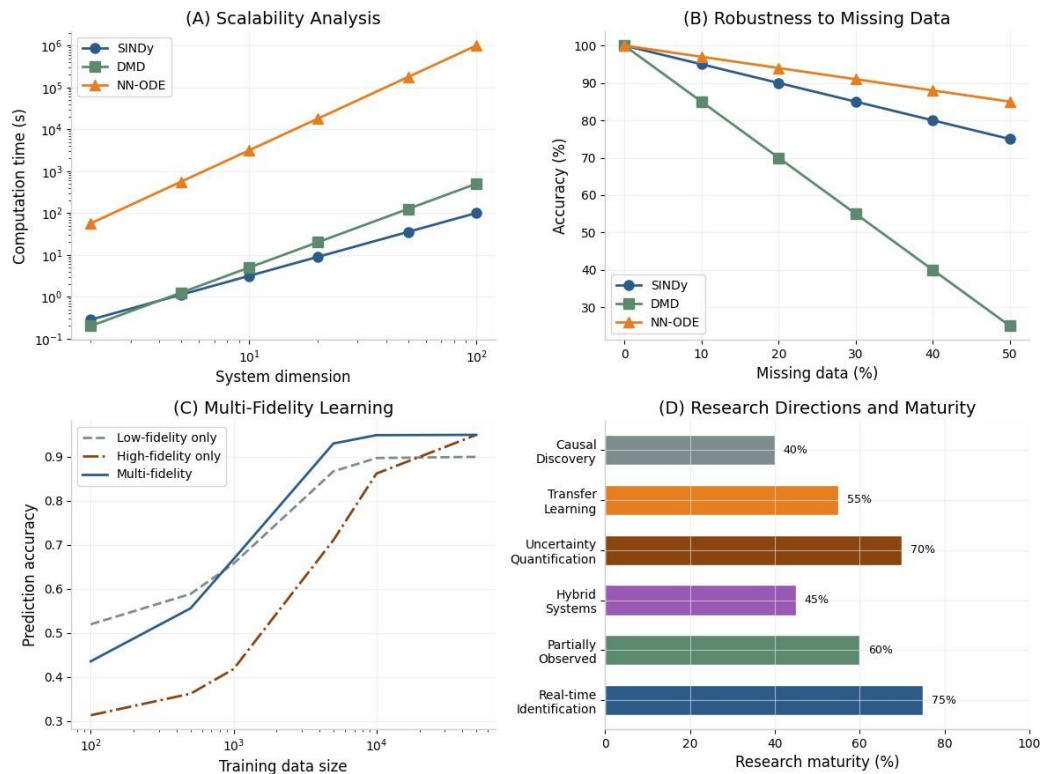


Figure 10. Performance summary and future directions. (A) Scalability analysis showing computation time versus system dimension. (B) Robustness to missing data for different methods. (C) Multi-fidelity learning performance. (D) Research directions and maturity levels.

Data-Driven System Identification Workflow

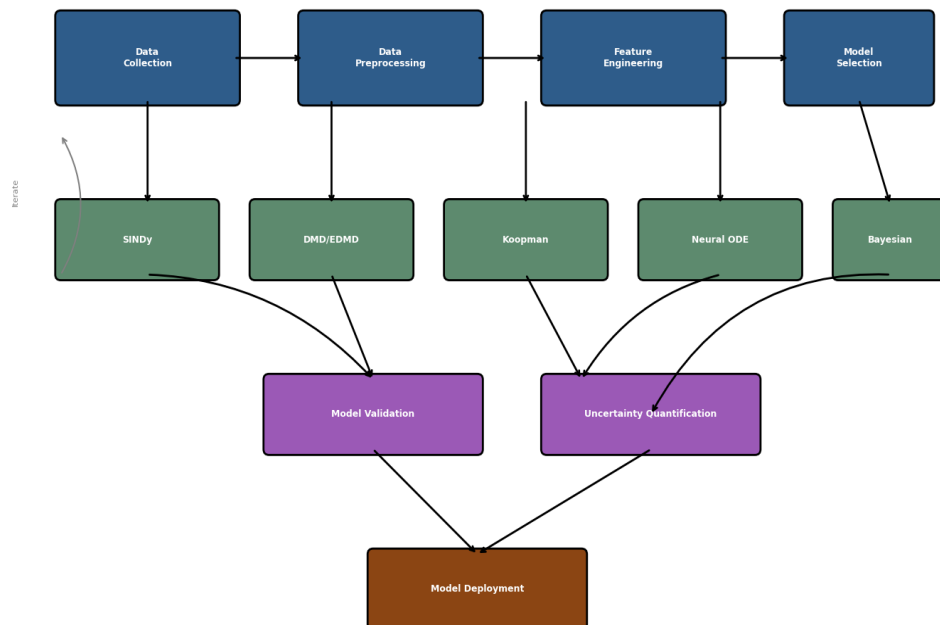


Figure 11. Data-driven system identification workflow. The diagram illustrates the complete pipeline from data collection through preprocessing, feature engineering, model selection, validation, and deployment.

Software Tools and Libraries

Library/Tool	Language	Key Features
PySINDy	Python	Sparse regression, multiple optimizers, noise handling
PyDMD	Python	Standard and variants of DMD, parallel computing
Koopmanlib	Python/MATLAB	Koopman operator approximation, eigenfunction learning
torchdiffeq	Python	Neural ODE implementation with adjoint method
PyMC3/Pyro	Python	Bayesian inference, MCMC, variational inference
UQLab	MATLAB	Uncertainty quantification, polynomial chaos

Conclusion

This review has provided a comprehensive survey of data-driven methods for identifying stochastic dynamical systems. We have examined the theoretical foundations, practical implementations, and relative merits of sparse regression approaches, operator-theoretic methods, neural network techniques, and Bayesian inference frameworks.

Several key conclusions emerge from our analysis. First, no single method dominates across all problem classes; the optimal choice depends on system characteristics, data availability, noise properties, and computational constraints. Second, hybrid approaches that combine multiple methodologies often outperform any single method, leveraging the strengths of each while mitigating weaknesses. Third, the treatment of stochasticity remains a significant challenge, with most methods assuming simple noise models that may not capture the complexity of real-world perturbations.

For practitioners, we offer the following recommendations. When the underlying system is expected to have sparse dynamics with a small number of governing terms, SINDy provides an interpretable and efficient solution. For systems with strong coherent structures and periodic or quasi-periodic behavior, DMD and Koopman methods are well-suited. When high predictive accuracy is paramount and interpretability is less critical, Neural ODEs and related deep learning approaches offer state-of-the-art performance. When uncertainty quantification is essential, Bayesian methods provide the most rigorous framework, albeit at increased computational cost.

Looking forward, we anticipate several developments will shape the field. The integration of physics-informed constraints with machine learning will yield more robust and generalizable models. Advances in computational hardware and algorithms will enable real-time identification for increasingly complex systems. The development of rigorous theoretical foundations for deep learning approaches to dynamical systems will provide better guidance for architecture design and training strategies. Finally, the application of these methods to emerging domains such as quantum systems, synthetic biology, and autonomous systems will drive continued innovation.

The identification of stochastic dynamical systems from data stands at the intersection of classical

applied mathematics, modern machine learning, and domain-specific science. As data becomes increasingly abundant and computational resources more powerful, the methods reviewed in this paper will play an ever more central role in understanding and predicting the complex systems that shape our world.

Future research should focus on developing unified frameworks that combine the interpretability of sparse methods, the spectral insights of operator theory, the flexibility of neural networks, and the uncertainty quantification of Bayesian approaches. Such hybrid methods would provide the best of all worlds, enabling robust, accurate, and interpretable identification of complex stochastic systems.

In conclusion, the field of data-driven identification of stochastic dynamical systems has made remarkable progress in recent years, but significant challenges remain. Addressing these challenges will require continued collaboration between mathematicians, computer scientists, and domain experts. The potential rewards are substantial, as improved identification methods will enable better prediction, control, and understanding of the complex stochastic systems that pervade our world.

References

- [1] Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15), 3932-3937.
- [2] Chen, R. T., Rubanova, Y., Bettencourt, J., & Duvenaud, D. (2018). Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 6571-6583.
- [3] Kutz, J. N., Brunton, S. L., Brunton, B. W., & Proctor, J. L. (2016). Dynamic mode decomposition: Data-driven modeling of complex systems. *SIAM*.
- [4] Mezic, I. (2013). Analysis of fluid flows via spectral properties of the Koopman operator. *Annual Review of Fluid Mechanics*, 45, 357-378.
- [5] Oksendal, B. (2003). *Stochastic differential equations: An introduction with applications*. Springer Science & Business Media.
- [6] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 686-707.
- [7] Rudy, S. H., Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2017). Data-driven discovery of partial differential equations. *Science Advances*, 3(4), e1602614.
- [8] Schmid, P. J. (2010). Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656, 5-28.
- [9] Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence* (pp. 366-381). Springer.
- [10] Williams, M. O., Kevrekidis, I. G., & Rowley, C. W. (2015). A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6), 1307-1346.
- [11] Li, X., Wong, T. K., Chen, R. T., & Duvenaud, D. (2020). Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics* (pp. 3870-3882). PMLR.
- [12] Morrill, J., Fermanian, A., Kidger, P., & Lyons, T. (2021). Neural rough differential equations for long time series. In *International Conference on Machine Learning* (pp. 7829-7838). PMLR.
- [13] Champion, K., Lusch, B., Kutz, J. N., & Brunton, S. L. (2019). Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116(45), 22445-22451.
- [14] Klus, S., Nuske, F., & Hamzi, B. (2020). Kernel-based approximation of the Koopman generator and Schrodinger operator. *Entropy*, 22(7), 722.

- [15] Bezenac, E., Pajot, A., & Gallinari, P. (2021). Deep learning for physical processes: Incorporating prior scientific knowledge. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12), 124009.
- [16] Cranmer, M., Greydanus, S., Hoyer, S., Battaglia, P., Spergel, D., & Ho, S. (2020). Lagrangian neural networks. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*.
- [17] Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, 3(6), 422-440.
- [18] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- [19] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC Press.
- [20] Risken, H. (1996). *The Fokker-Planck equation: Methods of solution and applications*. Springer.
- [21] Lasota, A., & Mackey, M. C. (1994). *Chaos, fractals, and noise: Stochastic aspects of dynamics*. Springer Science & Business Media.
- [22] Kloeden, P. E., & Platen, E. (2013). *Numerical solution of stochastic differential equations*. Springer Science & Business Media.
- [23] Sarkka, S., & Solin, A. (2019). *Applied stochastic differential equations*. Cambridge University Press.
- [24] Hairer, M., & Mattingly, J. C. (2011). A theory of hypoellipticity and unique ergodicity for semilinear stochastic PDEs. *Electronic Journal of Probability*, 16, 658-738.
- [25] Pavliotis, G. A. (2014). *Stochastic processes and applications: Diffusion processes, the Fokker-Planck and Langevin equations*. Springer.
- [26] Majda, A. J., & Harlim, J. (2012). *Filtering complex turbulent systems*. Cambridge University Press.
- [27] Vlachas, P. R., Byeon, W., Wan, Z. Y., Sapsis, T. P., & Koumoutsakos, P. (2018). Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. *Proceedings of the Royal Society A*, 474(2213), 20170844.
- [28] Pathak, J., Hunt, B., Girvan, M., Lu, Z., & Ott, E. (2018). Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Physical Review Letters*, 120(2), 024102.
- [29] Lusch, B., Kutz, J. N., & Brunton, S. L. (2018). Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1), 4950.
- [30] Wehmeyer, C., & Noe, F. (2018). Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *The Journal of Chemical Physics*, 148(24), 241703.
- [31] Boulanger-Lewandowski, N., Bengio, Y., & Vincent, P. (2012). Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *International Conference on Machine Learning* (pp. 1159-1166).
- [32] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- [33] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.
- [34] Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.
- [35] Strogatz, S. H. (2018). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*. CRC Press.
- [36] Wiggins, S. (2003). *Introduction to applied nonlinear dynamical systems and chaos*. Springer.
- [37] Guckenheimer, J., & Holmes, P. (2013). *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*. Springer.
- [38] Arnold, L. (1974). *Stochastic differential equations: Theory and applications*. Wiley.
- [39] Freidlin, M. I., & Wentzell, A. D. (2012). *Random perturbations of dynamical systems*. Springer.
- [40] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 20

4700-4708).

- [41] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).
- [42] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
- [43] Mahato, A. K., Das, R., & Sahani, S. K. (2025). Mathematical and computational techniques for sustainable agricultural development in Nepal. *Rajarshi Janak University Research Journal*. DOI: 10.3126/rjurj.v3i1.80697
- [44] Sahani, S. K., Oruganti, S. K., Satishkumar, K., & Gobithaasan, R. U. (2025). Advanced Mathematical Modeling of Woolen Knitting Dynamics Using Laplace Transform and Fourth-Order Runge–Kutta Method. *Journal of Mathematics*. DOI: 10.1155/jom/4985087
- [45] Gautam, H., Sahani, S. K., Mandal, D. N., Paudel, M. P., & Sahani, K. (2024). A Revision of Relaxed Steepest Descent, Gradient Descent, Modified Ellipsoid, David-Fletcher-Powell Variable Metric, Newton's, and Fletcher-Reeves Conjugate Methods From the Dynamics on an Invariant Manifold: A Journey of Mathematical Optimization. *African Journal of Biological Sciences*. DOI: 10.33472/AFJBS.6.5.2024.2175-2187
- [46] Sah, S. K., & Sahani, S. K. (2024). Use of Differential Equations in Population Growth Analysis. *Asian Journal of Science, Technology, Engineering, and Art*. DOI: 10.58578/ajstea.v2i6.3951
- [47] Pandey, B. K., Pandey, D., & Sahani, S. K. (2024). Autopilot control unmanned aerial vehicle system for sewage defect detection using deep learning. *Engineering Reports*. DOI: 10.1002/eng2.12852
- [48] Poudel, M. P., Pahari, N. P., Sahani, S. K., Basnet, G. B., & Poudel, R. P. (2024). Generalization of Classical Summation Relation of Certain Appell's Double Hypergeometric Functions Associated with Theory of Approximation. *Communications on Applied Nonlinear Analysis*. DOI: 10.52783/cana.v31.305
- [49] Sah, B. K., & Sahani, S. K. (2024). Poisson-New Linear-Exponential Distribution. *African Journal of Biological Sciences*. DOI: 10.33472/AFJBS.6.5.2024.64-73
- [50] Sah, B. K., & Sahani, S. K. (2024). Premium Linear-Exponential Mixture of Poisson Distribution. *Communications on Applied Nonlinear Analysis*. DOI: 10.52783/cana.v31.393
- [51] Poudel, M. P., Sahani, S. K., Pokhrel, M., & Tripathi, B. R. (2024). The product of two hypergeometric function (I) by using Bailey's formula. *Rajarshi Janak University Research Journal*. DOI: 10.3126/rjurj.v2i1-2.72302
- [52] Sahani, K., Khadka, S. S., Sahani, S. K., Pandey, B. K., & Pandey, D. (2023). A possible underground roadway for transportation facilities in Kathmandu Valley: A racking deformation of underground rectangular structures. *Engineering Reports*. DOI: 10.1002/eng2.12821
- [53] Tiwari, S. K., Rathour, L., Sahani, S. K., & Mishra, L. N. (2023). Results for coupled fixed point theorems in partial order metric spaces. *AIP Conference Proceedings*. DOI: 10.1063/5.0149509