支配方程式の発見とドメイン知識を持つ学習システムに向けて

大坪洋介，中島伸一

Discovery of Governing Equations and Learning Systems with Domain Knowledge†

Yosuke OTSUBO and Shinichi NAKAJIMA

機械学習は近年様々な分野で成功を収めてきている．しかし一般に，従来のデータ駆動型の機械学習は次のような課題がある．1）解釈性が乏しい，2）不十分なデータやラベルに対して十分な精度がでない．本稿ではまず，解釈可能な予測モデルを構成するため，時系列データから常微分方程式（ODE）を発見する問題に焦点を当て，スパース推定とカーネルリッジ回帰を用いた新しいアルゴリズムを提案する．ODE はこれまで専門家の深い洞察によってモデル化されてきたが，データ駆動で ODE の関数形を発見することは，解釈性を備えた予測モデルを学習するという観点から，物理，化学，生物分野などの様々な科学分野において価値がある．さらに上記 1）と 2）の課題を解決するため，近年提案されたドメイン知識を活用した機械学習のフレイムワークである Informed Machine Leaning について簡単に紹介し，ものづくり企業の立場から機械学習に活用できる知識を整理する．このような試みは，解釈性が高く，不十分なデータについても対応可能な機械学習システムの開発に役立つと考えられる．

Machine learning has been great successful in many areas in recent years. However, in general, the conventional data driven approaches in machine learning may have limitations for the following senses: 1) Lack of interpretability, 2) Low accuracy in insufficient data and annotations. To develop predictive model with rich interpretabilities, we focus on ordinary differential equation (ODE) discovery problem and propose a novel algorithm using kernel ridge regression with sparsity inducing regularizer. The ODEs have been modeled by domain experts based on theoretical deduction and empirical observations. So, automatic discovery of ODEs through data-driven is of great significance in various scientific fields, such as those of physics, chemistry, and biology in terms of interpretable predictions. Furthermore, to remedy the issues 1) and 2), we shortly introduce Informed Machine Learning, a machine learning pipeline framework with prior knowledge, and provide useful knowledge for further development of the learning system from the viewpoint of manufacturing companies. Such an attempt will help us to develop the interpretable learning systems that can deal with insufficient data.

Key words 常微分方程式，スパース推定，再生核ヒルベルト空間，ドメイン知識，知識が導入された機械学習
ordinary differential equations, sparse inference, reproducing kernel space, domain knowledge, Informed Machine Learning

1 Introduction

Many methodologies in machine learning make some inference by using data efficiently. In general, the conventional data driven approaches may have limitations for the following senses.

I. Lack of interpretability. If the machine learning techniques work well, interpretation and explanation are often required for the resultant algorithms and models. Understanding the natural phenomena in science, particularly, can be more important than making accurate predictions. For example, during anomaly detection in manufacturing processes, it is important to interpret the results and suggest next action for engineers.

II. Low accuracy in insufficient data and annotations. We must deal with the lack of enough data or their labels. For example, an adequate amount of customer data cannot be obtained owing to confidentiality and privacy issues or limitations related to the biological and medical experimental environment. Despite the advanced knowledge for annotations, it is difficult to obtain enough labeled datasets.

To remedy these issues, we mainly focus on 1) Ordinary differential equation (ODE) discovery problem for developing predictive model with rich interpretabilities, and shortly

† This article contains a summary of [14]

provide 2) taxonomy of useful prior knowledge for developing further develop the learning systems in manufacturing companies.

1.1. ODE discovery problem

Various types of nonlinear dynamical systems have been developed for characterizing the natural phenomena in science and engineering. For example, Newtonian dynamics, i.e., Newton's second law describes the dynamics of particles, and enzyme kinetics provides insights into the catalytic mechanisms of enzymes in the biochemical context. Such dynamics are often described as nonlinear ODE in the following form:

$$\dot{\boldsymbol{x}} = \frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{x}$ is the state variable, t is the time, and $\boldsymbol{f}$ is a nonlinear function parameterized by $\boldsymbol{\theta}$. Historically, many important ODEs, e.g., Newton' law, Maxwell equations, enzyme kinetics, were discovered by domain experts based on theoretical deduction and empirical observations.

A question we try to answer in this paper is whether such discovery process can be automated, i.e., we try to find ODEs that the observed time-series data satisfy, automatically by training machine learning models*1. To accomplish this, the following two issues need to be addressed: parameter specification and inference. The former, known as the ODE parameter inference problem, corresponds to the determination of the internal parameter $\boldsymbol{\theta}$, and the latter, known as the ODE discovery problem, corresponds to the identification of the functional form of $\boldsymbol{f}$ in Eq. (1).

We tackle the ODE discovery problem in the first half of this article. In practice, most of the possible applications include the identification of the dynamics of biopathways, which are usually described as ODEs based on their biochemical reactions[1]. Even though various computational models of regulatory and metabolic networks have been proposed by domain experts (e.g.[2]), determining the essential connectivity and structures of these dynamics remains an extremely challenging task. In computer aided engineering (CAE) processes, the dynamics of the flow and temperature on materials need to be mathematically modeled to design and construct mechanical architectures. Thus, inferring the structures and nonlinear dynamics in large systems is a challenging problem.

Additionally, if the predictive models are trained in the form of an ODE function, they can provide rich interpretabilities to domain experts. That is, the terms in the ODEs can be considered relevant in a physical or chemical context (e.g., friction strength or reaction intensity). Therefore, our study may be closely related to the estimation of interpretable predictive models.

1.2. Taxonomy of useful knowledge

To deal with the subjects mentioned in 1) and 2), several works incorporated prior knowledge into machine learning processes[3)~10)]. For example, knowledge gained from a scientific or mechanical perspective can help us to improve the learning accuracy and interpretability. As concrete examples, physics guided neural networks, where a penalty term inspired by scientific knowledge is added to the loss function as regularizer, provide more accurate results than purely data driven approaches[9]. A recent study introduced a systematic taxonomy of integrating knowledge into learning systems, called *Informed Machine Learning*[11]. The study provides definitions of the prior knowledge, its representation, and integration into the machine learning pipeline. In this study, we introduce new useful domain knowledge to further develop the learning system in the context of manufacturing. Such an attempt will help us to develop useful and efficient learning systems for manufacturing companies.

2 ODE discovery in RKHS

In this section, we introduce the ODE discovery problem, our approach, and report on experimental results.

2.1. Problem definition and related works

Consider N time points $t_1 < t_2 \cdots < t_N$ and their corresponding state variables

$$\boldsymbol{X} = \left[\boldsymbol{x}(t_1),\ldots,\boldsymbol{x}(t_N)\right]^T = \left[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\right]^T \tag{2}$$

where $\boldsymbol{x}_s \in \mathbb{R}^d$ represents the states at the s-th time-point. Similarly, the matrix of derivatives can be described as

$$\dot{\boldsymbol{X}} = \left[\dot{\boldsymbol{x}}(t_1),\ldots,\dot{\boldsymbol{x}}(t_N)\right]^T = \left[\dot{\boldsymbol{x}}_1,\ldots,\dot{\boldsymbol{x}}_N\right]^T \tag{3}$$

Our problem includes estimating the functional form of $\boldsymbol{f}$ in Eq. (1) from the data $\boldsymbol{X}$ given by Eq. (2). This can be accomplished using a method in which sparse inference is applied to fit the numerical derivatives of a linear regression model using a large set of possible ODE candidate functions[12)13)]. Such methods are called sparse identification of nonlinear dynamics (SINDy). In SINDy, the library of candidates of the nonlinear functions constructed by their states are set as

*1 Although Eq. (1) involves only the first derivative, it covers any finite degree ODE: for expressing an R-th degree ODE, the state vector $\boldsymbol{x}$ should be augmented by its first to the R-1-th degree derivatives. This procedure adequately set the degree of freedom (i.e., the dimension of the state) of the model.

$$\Theta(\boldsymbol{X})=\left[\mathbf{1}\ \boldsymbol{X}\ \boldsymbol{X}^2\ \boldsymbol{X}^3\cdots\right], \tag{4}$$

where $\boldsymbol{X}^k$ denotes the matrix containing all possible column vectors obtained from the time series of k-th degree polynomials in the state vector $\boldsymbol{x}$. Note that the dimension D of $\Theta(\boldsymbol{X})$ with k-th polynomial terms in d variables can be computed as ${}_{k+d}C_d$; the possible combination is given by $N_p=\sum_{i=1}^{D}{}_DC_i$.

The ODE with the possible candidate bases can be modeled in the parametric form:

$$\dot{\boldsymbol{X}}=\Theta(\boldsymbol{X})\boldsymbol{B}, \tag{5}$$

where $\boldsymbol{B}=\left[\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d\right]$ corresponds to the coefficients of the ODE. Let us focus on the l-th column,

$$\dot{\boldsymbol{X}}_l=\Theta(\boldsymbol{X})\boldsymbol{\beta}_l \tag{6}$$

where $\boldsymbol{X}_l=\left[\boldsymbol{x}_l(\mathrm{t}_1),\ldots,\boldsymbol{x}_l(t_N)\right]$. Then, it is evident that if $\beta_{l,s}=0$, the s-th feature is not effective in the l-th state. Thus, the ODE discovery problem in the formulation is reduced to the inference problem where the coefficients of matrix $\boldsymbol{B}$ contain many zero components. The original SINDy algorithm uses the sequential thresholded least squares (see Algorithm 1 in[14]). Lasso also can be employed as an alternative approach to enforce sparsity:

$$\boldsymbol{\beta}_l=\arg\min_{\boldsymbol{\beta}_l}\left\|\dot{\boldsymbol{X}}_l-\Theta(\boldsymbol{X})\boldsymbol{\beta}_l\right\|_2^2+\lambda\left\|\boldsymbol{\beta}_l\right\|_1,\forall l, \tag{7}$$

where λ denotes the strength of the L1 regularization term that controls the sparsity. Note that in this approach, the derivatives should be computed numerically from the noisy observations, for which stable implementation is non-trivial. However, the numerical computation of the derivatives is not trivial, and several studies have focused on differentiating the variables precisely. Among many methods proposed, SINDy employed the total variation regularized derivatives (TVDiff)[15] method, a well-known robust method for computing the derivatives from noisy data.

Many methods have been proposed to solve the parameter inference problem for ODE, given the function form of $\boldsymbol{f}$. Gradient matching methods, which are effective for inferring the parameters of ODEs[16)~18)], consist of two steps: a smoothing process to fit the data and an optimization process to minimize some metric between the smooth model and the derivatives predicted from ODEs. A recently proposed gradient matching method defined in the reproducing kernel Hilbert space (RKHS)[18] achieved significant improvements in comparison to alternative probabilistic methods[16)17)]. The method minimizes the loss term for the kernel regression term and the gradient matching term simultaneously:

$$\left\{\boldsymbol{A}^*,\boldsymbol{\theta}^*\right\}=\mathrm{argmin}_{\boldsymbol{A},\boldsymbol{\theta}}\,E(\boldsymbol{A},\boldsymbol{\theta}), \tag{8}$$

$$E(\boldsymbol{A},\boldsymbol{\theta})=\sum_{l=1}^{d}\left\|\boldsymbol{g}_l(\boldsymbol{\alpha}_l)-\boldsymbol{X}_l\right\|_2^2 +\rho\sum_{l=1}^{d}\left\|\dot{\boldsymbol{g}}_l(\boldsymbol{\alpha}_l)-\boldsymbol{f}_l(\boldsymbol{g}(\boldsymbol{A});\boldsymbol{\theta})\right\|_2^2, \tag{9}$$

where kernel regression can be expressed as

$$\boldsymbol{g}_l(t;\boldsymbol{\alpha}_l)=\textstyle\sum_{i=1}^{N}\boldsymbol{\alpha}_{l,i}k(t,t_i)=\boldsymbol{\alpha}_l^T\boldsymbol{k}_l(t), \tag{10}$$

and

$$\dot{\boldsymbol{g}}_l(t;\boldsymbol{\alpha}_l)=\boldsymbol{\alpha}_l^T\dot{\boldsymbol{k}}_l(t). \tag{11}$$

Here, $\boldsymbol{\alpha}_l=\left[\alpha_{l,1},\ldots,\alpha_{1,N}\right]^T$ denotes the vector of the kernel regression coefficients of the l-th variable. The vector of l-th kernels of t, $\boldsymbol{k}_l(t)=\left[k_l(t,t_1),\ldots,k_l(t,t_N)\right]^T$ is specified by the hyperparameter $\boldsymbol{\phi}_l$, i.e., $\boldsymbol{k}_l(t)=\boldsymbol{k}(t;\boldsymbol{\phi}_l)$. The first term in Eq. (9) encourages reconstructed by $g_l(t;\boldsymbol{\alpha})$ of the data $\boldsymbol{X}_l$, while the second term penalizes the inconsistency with the ODE model.

2.2. Proposed method

Inspired by the methods provided in the previous subsection, we propose a hybrid algorithm of sparse inference and a gradient matching algorithm in RKHS.

First, we impose L1 and L2 regularization to Eq. (9):

$$E(\boldsymbol{A},\boldsymbol{B})=\sum_{l=1}^{d}E_l(\boldsymbol{A},\boldsymbol{\beta}_l), \tag{12}$$

$$E_l(\boldsymbol{A},\boldsymbol{\beta}_l)=\left\|\dot{\boldsymbol{g}}_l(\boldsymbol{\alpha}_l)-\Theta(\boldsymbol{A})\boldsymbol{\beta}_l\right\|_2^2+\lambda_1\left\|\boldsymbol{\beta}_l\right\|_1+\lambda_2\boldsymbol{\alpha}_l^T\boldsymbol{K}_l\boldsymbol{\alpha}_l+\rho\left\|\boldsymbol{g}_l(\boldsymbol{\alpha}_l)-\boldsymbol{X}_l\right\|_2^2, \tag{13}$$

where the first term corresponds to the gradient matching term, in which the ODE is represented as a library of candidates of possible bases, $\Theta(\boldsymbol{A})$, similar to SINDy. The interpolant functions $\boldsymbol{g}_l(\boldsymbol{\alpha}_l)$ and $\dot{\boldsymbol{g}}_l(\boldsymbol{\alpha}_l)$ are defined by Eqs. (10) and (11), respectively. $\{\lambda_1,\ \lambda_2,\ \rho\}$ are the regularization parameters and the Gram matrix K_l depends on the kernel parameter $\boldsymbol{\phi}_l$. The minimization of Eq. (12) with respect to $\boldsymbol{A}$ encounters a problem of the complicated dependence of the first term on $\boldsymbol{A}$. Introducing an auxiliary variable $\tilde{\boldsymbol{A}}$ detangles the dependency:

$$\tilde{\boldsymbol{E}}(\boldsymbol{A},\tilde{\boldsymbol{A}},\mathbf{B})=\sum_{l=1}^{d}\boldsymbol{E}_l(\boldsymbol{A},\tilde{\boldsymbol{A}},\boldsymbol{\beta}_l), \tag{14}$$

$$\boldsymbol{E}_l(\boldsymbol{A},\tilde{\boldsymbol{A}},\boldsymbol{\beta}_l)=\left\|\dot{\boldsymbol{g}}_l(\boldsymbol{\alpha}_l)-\Theta(\tilde{\boldsymbol{A}})\boldsymbol{\beta}_l\right\|_2^2+\lambda_1\left\|\boldsymbol{\beta}_l\right\|_1+\lambda_2\boldsymbol{\alpha}_l^T\boldsymbol{K}_l\boldsymbol{\alpha}_l+\rho\left\|\boldsymbol{g}_l(\boldsymbol{\alpha}_l)-\boldsymbol{X}_l\right\|_2^2+\lambda_3\left\|\boldsymbol{\alpha}_l-\tilde{\boldsymbol{\alpha}}_l\right\|_2^2. \tag{15}$$

The last term forces $\tilde{\boldsymbol{A}}$ to match $\boldsymbol{A}$ when λ_3 is sufficiently large, thereby leading to $\boldsymbol{E}_l(\boldsymbol{A},\tilde{\boldsymbol{A}},\boldsymbol{B})=E_l(\boldsymbol{A},\boldsymbol{B})$. We optimize each parameter; Eq. (15) can be analytically minimized with respect to $\boldsymbol{A}$ as follows:

$$\boldsymbol{\alpha}_l^{\mathrm{new}}=\left[\boldsymbol{K}_l^T\left(\lambda_2\boldsymbol{I}_N+\rho\boldsymbol{K}_l\right)+\dot{\boldsymbol{K}}_l^T\dot{\boldsymbol{K}}_l+\lambda_3\boldsymbol{I}_N\right]^{-1}\times\left[\rho\boldsymbol{K}_l\boldsymbol{X}_l+\dot{\boldsymbol{K}}_l^T\Theta(\tilde{\boldsymbol{A}})\boldsymbol{\beta}_l+\lambda_3\tilde{\boldsymbol{\alpha}}_l\right]. \tag{16}$$

Then, $\tilde{\boldsymbol{A}}$ is replaced with $\boldsymbol{A}$ for the next iteration, giving $\tilde{\boldsymbol{E}}(\boldsymbol{A}^{\mathrm{new}},\boldsymbol{A}^{\mathrm{new}},\boldsymbol{B})$. Minimization with respect to $\boldsymbol{B}$ can be per-

formed by the standard-lasso algorithm, such as coordinate descent, least angle regression, or alternating direction method of multipliers[19]. It is known that L1 regularizer tends to give a significant bias to the LASSO estimator. To remove the bias, we reapply the least squares method for the non-zero components $\Omega_l' = \{i \mid \beta_l \neq 0\}$ as the final step.

$$\beta_l = \left(\Theta'^T \Theta'\right)^{-1} \Theta' \boldsymbol{K}_l \boldsymbol{\alpha}_l, \tag{17}$$

where $\Theta' = \Theta_{:,\Omega_l'}$ and $\beta_l = \beta_{l,\Omega_l'}$. The hyperparameters $\boldsymbol{h} = \{\lambda, \Phi\}$, where $\lambda = \{\lambda_1, \lambda_2, \lambda_3, \rho\}$ and $\Phi = \{\phi_1, \ldots, \phi_d\}$, which correspond to the regularization and kernel parameters, respectively, are determined in preliminary experiment.

2.3. Numerical experimental settings

Baseline methods

To compare the algorithm performances for the ODE discovery problem, we selected the following methods as baselines.

· TVSINDy[12]: the sequential thresholded least-squares method for selecting variables with the total variation method for numerical differentiation.

· TVLasso: the lasso for selecting variables with the total variation method for numerical differentiation.

· RKHS-Lasso (1): special case of proposed method without the iteration, i.e., the solution obtained after a single epoch.

· RKHS-Lasso: our proposed method.

TVSINDy was the first method proposed for the ODE discovery problem, as demonstrated in Section 2.1; it was implemented using the MATLAB code provided by the authors. TVLasso represents our minor modifications to the TVSINDy method; we used the lasso algorithm in the MATLAB library after the total variation method for numerical differentiation. The third and fourth methods are proposed by us; note that the former corresponds to the easy version of our method. The hyperparameters were tuned manually.

Benchmark ODE models

· *1D-Spring model*, given by

$$\dot{x} = v, \quad \dot{v} = -kx - vv, \tag{18}$$

where $\boldsymbol{x}(t) \equiv \left[x(t), v(t)\right]$ consisting of the position and velocity, k and v are the model parameters expressing spring constant and air resistance constant, respectively.

· *Lotka-Volterra model* [20] is a model for ecological system that is used to describe the interactions between two species corresponding to predators and preys. The accurate ODE can be described as follows:

$$\dot{H} = H(\alpha - \beta P), \quad \dot{P} = -P(\gamma - \delta H) \tag{19}$$

· *Lorentz system*[21] was developed as a simplified mathematical model for atmospheric convection. The true ODE can be described by

$$\dot{x} = \sigma(y - x), \dot{y} = x(\rho - z) - y, \dot{z} = xy - \beta y \tag{20}$$

where $\boldsymbol{x}(t) \equiv \left[x(t), y(t), z(t)\right]$ correspond to the rate of convection, horizontal temperature, and vertical temperature, respectively.

· *Enzyme kinetics*[2] is a well-known mathematical formulation for enzyme-catalyzed reactions that can be described by four-dimensional ODE systems:

$$\begin{aligned}
\left[\dot{\mathrm{S}}\right] &= -\mathrm{k}_1[E][S] + k_{-1}[ES], \\
\left[\dot{\mathrm{E}}\right] &= -\mathrm{k}_1[E][S] + (k_{-1} + k_2)[ES], \\
\left[\dot{ES}\right] &= \mathrm{k}_1[E][S] - (k_{-1} + k_2)[ES], \\
\left[\dot{P}\right] &= k_2[ES]
\end{aligned} \tag{21}$$

where $\boldsymbol{x}(t) \equiv \left[[S], [E], [ES], [P]\right]$ correspond to a substrate, enzyme, complex, and product, respectively.

Settings of kernels and library

In this study, the least-square kernel, $k(t,t') = a\exp\left((t-t)^2 / 2b^2\right)$, was used for the spring, Lotka-Volterra, and Lorentz systems and the sigmoid kernel, $k(t,t') = \sigma^2 \arcsin(a + btt'/Z)$, where $Z = \sqrt{(a+bt^2+1)(a+bt'^2+1)}$, was used for enzyme kinetics. Note that the derivatives of each kernel with respect to t can be analytically computed (see supplement in[17]). The library of the candidates of nonlinear functions were set to be second order polynomials: $\Theta(\mathbf{X}) = \left[\mathbf{1}, \boldsymbol{X}, \boldsymbol{X}^2\right]$. Thus, the numbers of features, D, in each variable was 4 for the spring and Lotka-Volterra models, 10 for the Lorentz system, and 15 for enzyme kinetics.

2.4. Results

Two criteria were used to compare the performances: the MSE of $\boldsymbol{B}$ defined by $\Delta\boldsymbol{B} = 1/Dd \sum_{i=1}^{Dd} \left\|\hat{\boldsymbol{B}} - \boldsymbol{B}_{\text{true}}\right\|_2^2$ and Fscores defined by the harmonic mean of precision and recall, where $\hat{\boldsymbol{B}}$ denotes the value estimated by each method.

In Fig. 1, each component of $\hat{\boldsymbol{B}}$ is compared with the ground truth $\boldsymbol{B}_{\text{true}}$ in two cases with different noise levels for the (a) Lorenz system and (b) Enzyme system, where the regularization parameter λ_s is tuned so as to give the best Fscore by changing it systematically. Note that λ_s corresponds to the threshold of the iterative scheme in TVSINDy and the L1 regularized parameter in TVLasso, RKHS-Lasso(1), and RKHS-Lasso respectively. When the noise is large, more misidentifications occur; while, the parameters obtained with RKHS-Lasso are close to the true value.

Table 1 summarizes the performance in four benchmark ODE models with two different noise levels. It is evident that the proposed method outperforms the baselines in most cases.

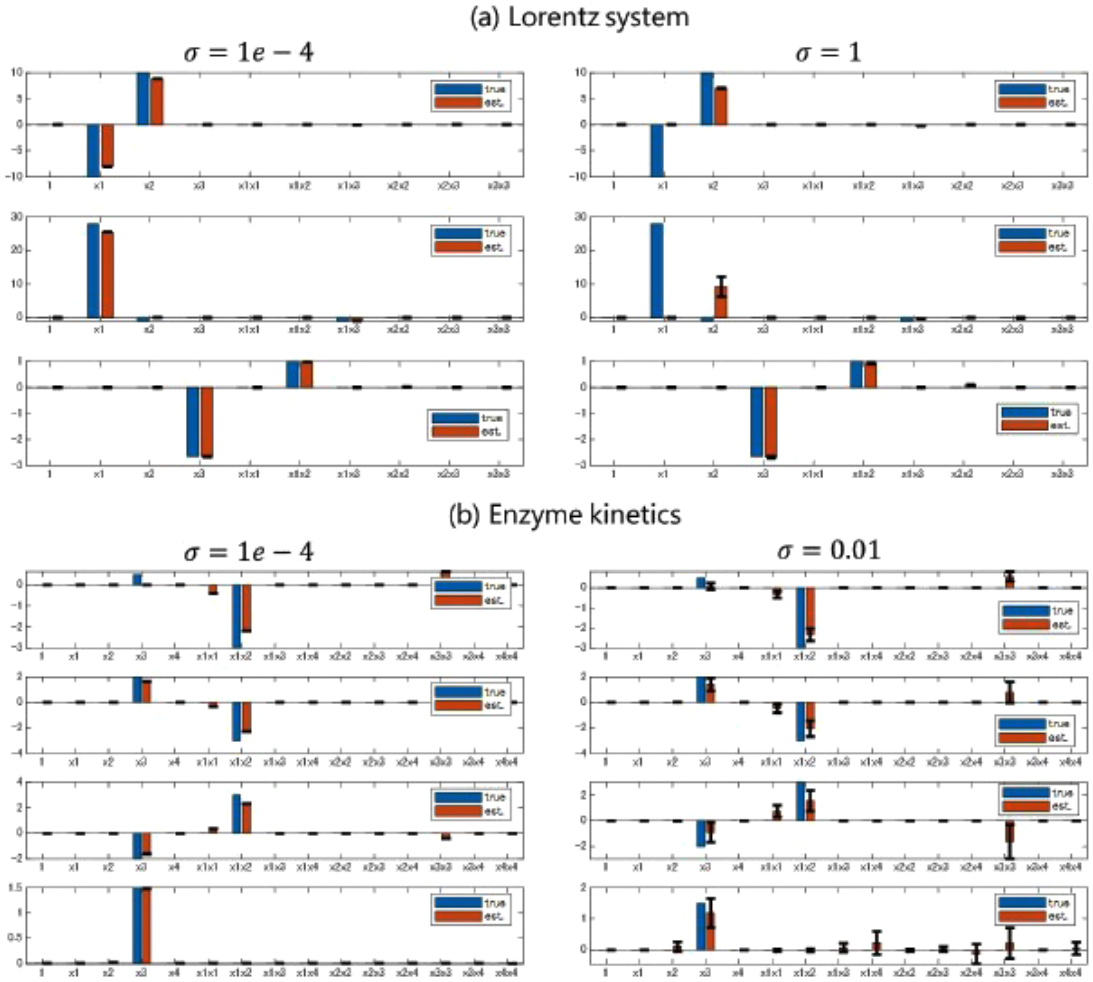

Fig. 1 Each panel shows both the ground truth (blue bars) and the estimated parameters (red bars) in each candidate function.

Table 1 MSE and Fscores in four benchmark ODE models. The means and standard deviations (values in the bracket) over 10 simulation trials are shown with the regularization parameter optimized for the Fscores.

			TVSINDy	TVLasso	RKHS-Lasso (1)	RKHS-Lasso
Spring model	sigma = 1	MSE (min)	0.3128 (0.0943)	0.0485 (0.0522)	0.0334 (0.0322)	**0.0148 (0.0215)**
		Fscore (max)	0.7250 (0.0876)	**1.0000 (< 1e-6)**	**1.0000 (< 1e-6)**	**1.0000 (< 1e-6)**
	sigma = 5	MSE (min)	1.9133 (2.0197)	0.6157 (0.6381)	0.4620 (0.365127)	**0.4096 (0.3473)**
		Fscore (max)	0.6810 (0.0698)	0.7490 (0.1797)	0.8062 (0.1946)	**0.8157 (0.1838)**
Lotka-Volterra model	sigma = 1	MSE (min)	0.1667 (< 1e-6)	5.2e-4 4 (1.3e-4)	2.3e-4 4 (6.3e-5)	**1.3e-5 (1.2e-5)**
		Fscore (max)	0.6667 (< 1e-6)	**0.8000 (< 1e-6)**	**0.8000 (< 1e-6)**	0.7782 (0.0351)
	sigma = 5	MSE (min)	0.1667 (< 1e-6)	0.0014 (0.0012)	3.4e-4 (3.7e-4)	**2.8e-4 (4.0e-4)**
		Fscore (max)	0.6667 (< 1e-6)	0.7442 (0.0534)	0.7648 (0.0486)	**'0.8004 (0.1244)**
Lorentz system	sigma = 1e-4	MSE (min)	29.8885 (1.7e-4)	5.5922 (1.2e-4)	0.5169 (2.9e-5)	**0.0091 (2.4e-5)**
		Fscore (max)	0.6667 (< 1e-6)	0.7143 (< 1e-6)	0.7368 (< 1e-6)	**0.7778 (< 1e-6)**
	sigma = 1	MSE (min)	23.4052 (12.9023)	9.1514 (0.8997)	3.620779 (0.5580)	**1.964616 (0.5135)**
		Fscore (max)	0.6808 (0.0403)	**0.7143 (< 1e-6)**	**0.71433 (< 1e-6)**	**0.7143 (< 1e-6)**
Enzyme kinetics	sigma = 1e-4	MSE (min)	> 1e + 3	0.5599 (1.7e-4)	'0.3145 (0.2169)	**0.04343 (2.0e-4)**
		Fscore (max)	0.2093 (0.0010)	0.2908 (0.0333)	36 (0.0685)	**0.57711 (0.0120)**
	sigma = 0.01	MSE (min)	> 1e + 3	0.6229 (0.004699)	0.2953 (0.2719)	**0.2491 (0.2606)**
		Fscore (max)	0.2285 (0.0238)	0.2000 (< 1e-6)	0.4651 (0.0901)	**0.4831 (0.0849)**

2.5. Short summary

We proposed a new method to solve the ODE discovery problem that combined the RKHS-based method for interpolating the signals from the time series measurements and the sparse inference for selecting relevant bases from the library of possible features.

Our simulation studies showed that the proposed method compared favorably with the baseline methods based on sparse inference with the total variation regularized derivatives.

3 Useful domain knowledge to introduce in machine learning systems

We briefly review the Informed Machine Learning and provide the useful prior knowledge in manufacturing for developing interpretable learning systems that deal with insufficient data.

3.1. Overview of Informed Machine Learning

Informed Machine Learning is a framework where prior knowledge is explicitly integrated into the machine learning pipeline (Fig. 2). Rueden et al. defined "*knowledge*" as validated information about the relations between entities in certain contexts[11]. Such additional information will make the conventional machine learning techniques performed by data driven approaches more powerful in the following aspects:

・Incorporating what we have accumulated in a domain so far into a new system; prediction accuracy may be higher.

・By effectively utilizing the "knowledge and human resources" assets of a domain, unique systems can be developed.

・Interpretability for the learning processes and predicted results can be improved.

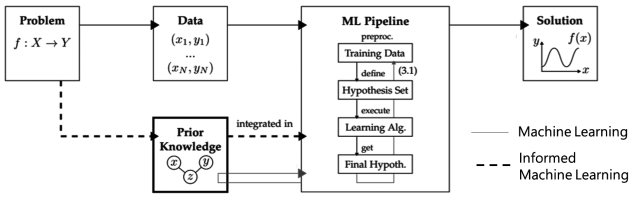

Fig. 2 Concept of Informed Machine Learning[11].

The taxonomy of the knowledge source, knowledge representation, and knowledge integration was elaborated[11), *2]. Here, we reconsider the part of "knowledge sources" from the manufacturing perspective.

3.2. Useful domain knowledge

Knowledge source refers to the origin of prior knowledge, which means various types of knowledge. They can be categorized as follows.

Natural Science: it is typically validated explicitly through scientific experiments (e.g., the universal laws of physics, bio-molecular descriptions of genetic sequences, or material-forming production processes).

Design information*: it denotes the specifications and mechanical information used in product design, including component dimensions, design layout, and structure of the products.

Process flows*: it represents the product manufacturing process and manufacturing settings. It sometimes includes staffing, inspection equipment settings, factory size, and inventory for the production.

Confidence*: it denotes the reliability for annotations or data. Sometimes, data labels may vary; in such cases, the confidence is given at the same time as data.

Domestic reports/ Past cases*: it denotes the past cases and considerations confirmed in-house, which are often given heuristically.

Aggregated human knowledge: it represents the facts from everyday life that are known to almost everyone and can also be called general knowledge.

(Expert's) Intuition: it denotes the knowledge based on the experiences and insights of experts that may not always have scientific evidence.

It should be noted that the knowledge sources marked with * are newly introduced in addition to the original form (Fig. 2 in[11]) from the viewpoint of manufacturing companies. The framework enriches our approaches of machine learning development.

4 Summary

We introduced two topics in this article: 1) ODE discovery problem and 2) useful domain knowledge to introduce in machine learning systems. In the former, we proposed an algorithm for discovering the functional form of ODE from time-series data that combined the gradient matching method and sparse inference; the proposed method outperformed other baseline methods. The latter provided an effec-

*2 In the original paper, the process of integrating prior knowledge into the machine learning pipeline was systematically investigated as following perspectives: 1) "what type of knowledge is integrated?", 2) "how is the knowledge represented or transformed?", 3) "where is the knowledge integrated in the machine learning pipeline?".

tive and usable domain knowledge in manufacturing processes for developing the Informed Machine Learning that is effective framework for integrating the prior knowledge into the machine learning pipeline. Our proposition will be useful for future machine learning techniques and data science development.

References

1) James D. Murray: Mathematical Biology: I. An Introduction, Third ed. (Springer, 2002).
2) G. M. Suel, J. Garcia-Ojalvo, L. M. Liberman and M. B. Elowitz: "An excitable gene regulatory circuit induces transient cellular differentiation", *Nature*, **440** (2006), 545–550.
3) A. Karpatne, G. Atluri, J. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova and V. Kumar: "Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data", *IEEE Transactions on Knowledge and Data Engineering*, **29** (2017), 2318–2331.
4) Geoffrey G. Towell and Jude W. Shavlik: "Knowledge-based artificial neural networks", *Artificial Intelligence*, **70** (1994), 119–165.
5) Y. Otsubo: "AI Research in Manufacturing—Thinking Away from Science", *Bulletin of the Japan Society for Industrial and Applied Mathematics (in Japanese)*, **29** (2019), 26–30.
6) M. L. Minsky: "Logical versus analogical or symbolic versus connectionist or neat versus scruffy", *AI magazine*, **12** (1991).
7) W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl and B. Yu: "Definitions, methods, and applications in interpretable machine learning", *PNAS,* **116** (2019), 22071–22080.
8) T. Yu: "Incorporating Prior Domain Knowledge Into Inductive Machine Learningg: Its implementation in contemporary capital markets", Ph.D. thesis, Faculty Inf. Technol., Univ. Technology, Sydney, Australia (2007).
9) A. Karpatne, W. Watkins, J. Read and V. Kumar: "Physics-guided neural networks (pgnn): An application in lake temperature modeling", *arxiv*., 1711.11157 (2017).
10) Y. Otsubo, N. Otani, M. Chikasue and M. Sugiyama: "Failure factor detection in production process", JSAI2020, 2I4-GS-2-04, (2020) (in Japanese).
11) L. von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, R. Ramamurthy, M. Walczak, J. Garcke, C. Bauckhage and J. Schuecker: "Informed Machine Learning - A Survey and Taxonomy of Integrating Knowledge into Learning Systems", *arxiv*., 1903.12394*v*2 (2016).
12) Steven L. Brunton, Joshua L. Proctor and N. Kutz: "Discovering governing equations from data by sparse identification of nonlinear dynamical systems", *PNAS*, **113** (2016), 3932–3937.
13) R. Tibshirani: "Regression shrinkage and selection via", *Journal of the Royal Statistical Society. Series B*, **58** (1996), 267–288.
14) Y. Otsubo and S. Nakajima: "Discovery of Governing equations in Reproducing Kernel Hilbert space", *IEICE Technical Report (IBISML2019-65)*, **118** (2018), 159–166.
15) R. Chartrand: "Numerical Differentiation of Noisy, Nonsmooth Data", *International Scholarly Research Network, ISRN Applied Mathematic*, **2011** (2011).
16) B. Calderhead, M. Girolami and N. D Lawrence: "Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes", *In Proceedings of Advances in Neural Information Processing Systems (NIPS)* (2013), 217–224.
17) F. Dondelinger, M. Filippone, S. Rogers and D. Husmeier: "ODE parameter inference using adaptive gradient matching with Gaussian processes", *In Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS)* (2013), 216–228.
18) M. Niu, S. Rogers, M. Filippone and D. Husmeier: "Fast inference in nonlinear dynamical systems using gradient matching", *In Proceedings of the 33rd International Conference on Machine Learning (ICML)* (2016), 1699–1707.
19) S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein: "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers", *Foundations and Trends in Machine Learning*, **3** (2011), 1–122.
20) Alfred J. Lotka: "Analytical Note on Certain Rhythmic Relations in Organic Systems", *PNAS*, **6** (1920), 410–415.
21) Edward N. Lorenz: "Deterministic nonperiodic flow", *Journal of the atmospheric sciences*, **20** (1963), 130–141.

大坪洋介
Yosuke OTSUBO
研究開発本部
数理技術研究所
Mathematical Sciences Research Laboratory
Research & Development Division

中島伸一
Shinichi NAKAJIMA
ベルリン工科大学
理化学研究所
Technische Universität Berlin
RIKEN