Sparse Regression and Adaptive Feature Generation for the Discovery of Dynamical Systems

Chinmay S. Kulkarni, Abhinav Gupta, and Pierre F. J. Lermusiaux(✉)

Massachusetts Institute of Technology, Cambridge, MA, USA
{chinmayk,guptaa,pierrel}@mit.edu

Abstract. We study the performance of sparse regression methods and propose new techniques to distill the governing equations of nonlinear dynamical systems from data. We start from the recently proposed generic methodology of learning interpretable equation forms from data, followed by performance of least absolute shrinkage and selection operator (LASSO) for this purpose. We first develop an algorithm that uses the dual of LASSO optimization for higher accuracy and stability. We then derive a second algorithm that learns the candidate function library in a dynamic data driven applications systems (DDDAS) manner to distill the governing equations of the dynamical system. This is achieved via sequentially thresholded ridge regression (STRidge) over a orthogonal polynomial space. The performance of the methods is illustrated using the Lorenz 63 system and a marine ecosystem model.

Keywords: Machine learning · DDDAS · Sparse regression · Nonlinear dynamical systems · Dual LASSO · System identification

1 Introduction and Overview

Data today are no longer used mostly to verify models derived from first principles but also to dynamically adapt and learn such models [9]. This is particularly important for non-autonomous nonlinear dynamical systems that describe a multitude of problems from science and engineering. Recent groundbreaking methods leverage the fact that most dynamical equations governing physical systems contain a few terms, making them sparse in high-dimensional nonlinear function space [2,12]. By constructing an appropriate feature library based on the data coordinates, one can apply sparse regression to discover the governing equations of the dynamical system. Few studies however try to improve upon the sparse regression algorithm at the core of the approach. This is exactly the first focus area of the present work. We examine the sparse regression method most commonly employed in this field: Least Absolute Shrinkage and Selection Operator (LASSO) [13]. Although LASSO works well with assured fast convergence rates for uncorrelated features, it converges more slowly for highly correlated features,

© Springer Nature Switzerland AG 2020
F. Darema et al. (Eds.): DDDAS 2020, LNCS 12312, pp. 208–216, 2020.
https://doi.org/10.1007/978-3-030-61725-7_25

and tends to choose a feature at random from each of the correlated groups [7]. To alleviate these difficulties, we propose to solve the dual of LASSO to learn the governing equations. Even in the case of correlated features, the dual LASSO has a unique solution, which allows us to correctly choose the features. The second part of this work deals with the case when the exact function blocks that describe the dynamical system are not present in the feature library. We develop a way to handle such cases by using an appropriate family of orthogonal functional basis to span the feature library combined with an approach to adaptively increase and decrease the dimension of the feature space. This allows us to add new components to the feature space that are orthogonal to the existing features while discarding those that do not have any projection of the dynamical system along them. We employ this algorithm iteratively, while adding or removing appropriate features to dynamically adapt our feature space for the best approximation of the equations from data. These Dynamic Data Driven Applications Systems (DDDAS) [3] approaches are demonstrated on the Lorenz 63 system [11] and a marine ecosystem model [6,8] with a non-polynomial nonlinearity. We show that our dynamic data driven algorithms robustly and accurately learn the presence of active features and of the nonlinearities without requiring any explicit feature information.

1.1 General Methodology

Let us assume that we have n state space parameters $(x_1, \ldots, x_n)$, with measurements for x_i and $\dot{x_i} = dx/dt$ at times $t = 1, \ldots, T$ (denoted by a superscript). If only state observations are available, the rate parameters can be computed using finite difference. This is followed by constructing a nonlinear library of features using the state space parameters. The span of these features now describes the feature space. Typically we would construct this feature space through a class of functions that are dense in the space that our dynamical system lives in. In this work, we assume a polynomial feature library, however the methodology is agnostic towards the choice of functional basis and would apply to any other feature library as well. After constructing the feature library (say X), we formulate the regression problem as $\dot{X} = XW + \varepsilon$, where $\dot{X}_{(t,j)} = \dot{x_j}^t$ and W are the unknown weights, with ε being the noise. Often in dynamical models, not all the features in the library that we consider are required to explain the dynamical model. Thus, as in [2], we utilize sparse regression to select the relevant features. However, unlike the aforementioned work, we dynamically build an suitable feature library which allows us to infer the nonlinear terms in the governing equations effectively, without knowing the type of functional space they live in. We also use the dual of LASSO optimization for higher accuracy and stability. These features, with their corresponding coefficients describe the functional form of the governing equations.

2 Regression Over Fixed Feature Space

In this section, we assume that the feature library is fixed, and that we wish to find either the exact sparse equation form from this library or the closest approximation to the governing equation only from the terms in the library. The highest polynomial degree in the feature space (X) is p. Then, the feature space contains terms of the form $(x_1^t)^{p_1} \dots (x_i^t)^{p_i} \dots (x_n^t)^{p_n}$, such that $p_1 + \dots + p_n \leq p$. The number of terms in the feature library is $m = \binom{n+p}{n} = \frac{(n+p)!}{n!p!}$ (*i.e.* $X \in \mathbb{R}^{\mathrm{T} \times \mathrm{m}}$). Empirically the number of distinct terms in the governing equations is $\mathcal{O}(n)$. Thus even for small enough p, the terms in the feature library are much more in number than those to be chosen, which justifies sparse regression to select the features. Let us denote the coefficient matrix obtained from the sparse regression by W. The optimization problem with some penalty ($\mathcal{P}$) is:

$$\min_{W} \mathcal{L}(W) = \left[\left(\dot{X} - XW \right)^2 + \mathcal{P}(W) \right] . \tag{1}$$

To further select the features appropriately, we use our knowledge of the underlying physics of the dynamical system. We select features by looking at their net characteristic magnitude instead of just the regression coefficients. We refer to this as ‘scale based thresholding’.

As is well-known, the LASSO penalty is $\mathcal{P}(W) = \lambda ||W||_1$ (hyperparameter λ), which serves as a convex counterpart to the non-convex L_0 norm. The pitfalls of LASSO (even after removing the irrelevant features using the SAFE bounds [15]) are that it requires significant hyperparameter tuning and it is extremely sensitive to λ for correlated features (observed empirically). These motivate us to instead formulate a new approach to solve the sparse regression problem.

To overcome the difficulties in the application of LASSO (along with the SAFE rules), we formulate and solve its dual problem. For the LASSO solution to be unique, the feature matrix must satisfy the irrepresentability condition (IC) and beta-min condition [15]. The feature library violates the IC for highly correlated columns, leading to an unstable feature selection. However, even for highly correlated features, the corresponding dual LASSO solution is always unique [13]. The dual problem is given by Eq. (2), which is strictly convex in θ (implying a unique solution).

$$\max_{\theta} \mathcal{D}(\theta) = ||\dot{X}||_2^2 - ||\theta - \dot{X}||_2^2 \text{ such that } ||X^T \theta||_\infty \leq \lambda . \tag{2}$$

Let $\hat{W}$ be a solution of Eq. (1) with LASSO penalty and $\hat{\theta}$ be the unique solution to the corresponding dual problem Eq. (2). Then a stationarity condition implies:

$$\hat{\theta} = \dot{X} - X\hat{W} . \tag{3}$$

Even though LASSO does not have a unique $\hat{W}$, the fitted value $X\hat{W}$ is unique, as the optimization problem Eq. (1) is strongly convex in XW for $\mathcal{P}(W) = \lambda ||W||_1$. We make use of this by first computing a solution to the primal LASSO problem and then computing the unique dual solution by using the primal fitted value and Eq. (3). Once we have the unique dual solution $\hat{\theta}$, we complete the feature selection by using the dual active set, which is same as the primal active set with high probability under the IC [7]. The KKT conditions imply:

$$\hat{\theta}^T X_i = \text{sign}(\hat{W}_i) \text{ if } \hat{W}_i \neq 0 \text{ and } \hat{\theta}^T X_i \in (-1, 1) \text{ if } \hat{W}_i = 0\,. \tag{4}$$

Equation (4) gives us a direct way to compute the active dual set once we have $\hat{\theta}$. We discard the features for which $\hat{\theta}^T X_i \in (-1, 1)$ and retain the others. This does not give us a good fit of the solution, so to compute the coefficients accurately, we perform ridge regression ($\mathcal{P}(W) = \lambda_2 ||W||_2^2$ over the active features. We refer to this new algorithm as 'dual LASSO feature selection'.

3 Regression Over a Dynamic Data Driven Feature Space

In this section, we consider cases where the feature library is not known and learned using DDDAS. If we have no prior belief over the form of the equations, we may not be able to construct an efficient feature library. In such situations, learning this library from data might be the most advantageous choice. The naïve approach of adding any new functions to the feature library until convergence can be very expensive and ill conditioned. A more principled and efficient approach is to make the use of orthogonal functions of some parametric family to construct this library, ensuring that the problem is always well conditioned. The drawback in this case is that the regressor may not be sparse over this feature library.

Starting with an empty library, we recursively add a feature to it and compute the corresponding loss function of the resulting fit by using STRidge (as will be described). If the loss function decreases by more than a certain fraction, we keep this feature. Otherwise, we discard it and look at the next orthogonal feature. Once every few addition timesteps, we perform a removal step to discard the feature(s) that do not result in a significant increase in the loss function. This ensures that we do not keep lower order functions that may not be required to describe the equations as higher order functions are added. Our algorithm is inspired by previous greedy feature development algorithms such as FoBa [14]. However, these algorithms require pre-determined full possible feature space, whereas we construct new features on the fly. Once the equations are obtained in terms of these orthogonal polynomials, we distill their sparse forms by using symbolic equation simplification [1].

To compute regressors over the orthogonal feature space, we use sequentially thresholded ridge regression (STRidge), developed by [12]. The idea is simple: we iteratively compute the ridge regression solution with decreasing penalty proportional to the condition number of X, and discard the components using scale based thresholding (Sect. 2). We iterate with ridge regression until there is no change in the feature space. As the feature matrix is orthonormal by construction, the analytical solution is $W = (1 + \lambda)^{-1} X^T \dot{X}$. The overall pseudocode for learning the governing equations through adaptively growing the feature library is given by Algorithm 1, and the corresponding results are presented in Sect. 4.

Algorithm 1. Learning Governing Equations through Adaptive Feature Library

```
Require: state parameters: $\boldsymbol{x} = x_i^t, \dot{\boldsymbol{x}} = \dot{x}_i^t$; orthogonal family $F_j(\bullet)$; feature addition
  / removal thresholds: $r_a\ (\leq 1), r_r\ (\geq 1), \lambda_0$; removal step frequency $k_r$
  Initialize: $X = \emptyset, W = \mathbf{0}, t = 0, \mathcal{L} = \infty, k = 0$
  while True do
      $X_t = [X, F_k(\boldsymbol{x})]$; solve the STRidge problem: $W_t = \text{STRidge}(\dot{X}, X_t, \lambda_0)$
      Compute the loss $\mathcal{L}_t = \left(\dot{X} - X_t W_t\right)^2$
      if $\mathcal{L}_t \leq r_a \mathcal{L}$ then $X = X_t$ ; $W = W_t$
      if $\mod(k, k_r) == 0$ then
          for $i = 1, \ldots, X.\text{shape}[1]$ do
              $X_t = [X[:, 1 : i-1], X[:, i+1 : \text{end}]]$; solve: $W_t = \text{STRidge}(\dot{X}, X_t, \lambda_0)$
              Compute the loss $\mathcal{L}_t = \left(\dot{X} - X_t W_t\right)^2$
              if $\mathcal{L}_t \leq r_r \mathcal{L}$ then $X = X_t$ ; $W = W_t$
      $k = k + 1$.
      break if no change in feature space over multiple iterations.
  Perform symbolic simplification of $\dot{X} = XW$ to obtain the final form of the equations
```

4 Results

4.1 Lorenz 63 System

For the first applications, our testbed will be the Lorenz 63 system ($n = 3$) given by Eq. (5), and fixed polynomial feature libraries with $p = 3$, 10 and 20 ($m = 20, 286$ and 1771). The idea behind considering larger orders (p) is that it highlights the poor performance of LASSO for highly correlated features.

$$\dot{x} = 10(yz - x) \ ; \quad \dot{y} = x(28 - z) \ ; \quad \dot{z} = xy - 2.667z \tag{5}$$

Figure 1a plots the number of non-zero features in the equations for different p values. LASSO has a much higher number of non-zero terms, and this number increases significantly with p (and m), indicating instability of the solution. Dual LASSO feature selection performs very well, and the number of present features does not change for the most part with p. Figure 1b plots the absolute weights for the components for the $p = 3$ case for the $\dot{y}$ equation. Dual LASSO feature selection retrieves the correct features (with accurate weights), while LASSO detects the correct features but also detects high order features that have low weights and are highly correlated to each other. This serves as a great validation of the superiority of dual LASSO feature selection over conventional LASSO feature selection for model discovery.

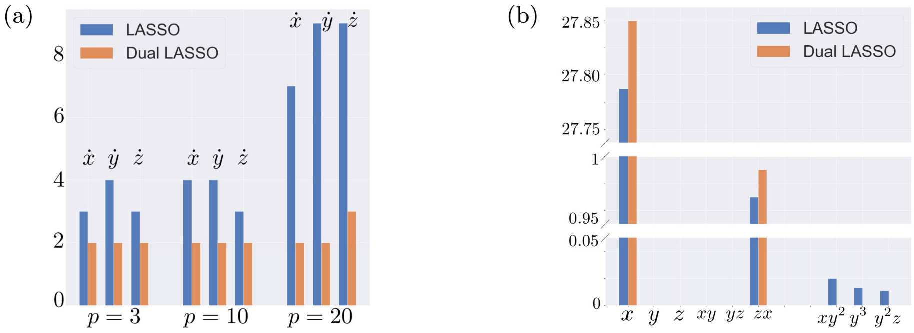

Fig. 1. (a) Number of nonzero terms for $\dot{x}, \dot{y}, \dot{z}$, and (b) absolute weights in the ODE for $\dot{y}$ ($p = 3$) for the Lorenz 63 system.

4.2 Marine Ecosystem Model

To demonstrate the capabilities of adaptive feature library growth algorithm (Sect. 3), we evaluate the learning scheme in a more complicated and realistic scenario. Hence, we now try to learn marine ecosystem models, which contain non-polynomial non-linearities. Realistic ecosystem models are very complex, but in broad-terms they can be seen as flow of food energy from nutrients, to phytoplanktons, to zooplanktons, to fishes, and finally recycling back to nutrients. Due to the lack of governing laws, and empirical nature of the development of these models, there are many different options in-terms of complexity and model parameterizations available, which could be highly nonlinear. But given the regional and seasonal differences at different locations in the world's oceans, one can quickly run out of all the options suggested by different biologists, and there is a need for DDDAS that adapt and learn new models from data [4,5,9,10]. Such models could be further adapted to run 'online', *i.e.* the inferred models can be updated as more data comes in, thereby improving and assimilating the observations on the fly. For the present test case, we consider a 3-component Nutrients-Phytoplankton-Detritus (NPD) model [6], given by,

$$\dot{N} = -\frac{r_{max}NP}{(k_N) + N} + l_N^P P + l_N^D D; \dot{P} = \frac{r_{max}NP}{(k_N) + N} - l_N^P P - l_D^P P; \dot{D} = l_D^P P - l_N^D D \qquad (6)$$

where N, P and D are normalized biological concentrations. The involved parameters are the nutrient uptake rate for phytoplanktons, r_{max}, losses by respiration, l_N^P, and mortality l_D^P. Mineralization is simulated by the rate l_N^D. The choice of the parameter values determine the dynamical stability of this system, and it can vary between stable point, spiral to stable point, and stable limit cycle.

The parameter values chosen for the testcase are: $r_{max} = 1\ day^{-1}$, $K_N = 0.3\,\text{mmol m}^{-3}$, $l_N^P = 0.50\ day^{-1}$, $l_D^P = 0.05\ day^{-1}$, $l_N^D = 0.06\ day^{-1}$, and $T = 1$ mmol m^{-3}, which makes the system spiral towards a stable point. Noise free data of the states and derivatives computed using a forward Euler scheme are extracted at a time-step of $\Delta t = 0.01\ \mathit{day}$ for the time period of $t = 0$ to $t = 50\ \mathit{days}$, and used for learning the system from scratch.

We start with an empty feature library and $W = \mathbf{0}$ and iteratively grow the feature space using Algorithm 1 with Legendre polynomials (denoted by $\mathbb{L}_p^{(\bullet)}$), $r_a = 0.85$, $r_r = 1.10$, $\lambda_0 = 1$, and removal step working after every addition step ($k_r = 1$). We adaptively grow the feature space until a maximum polynomial degree of 6 is reached. Finally, we use symbolic simplification followed by scale based thresholding to obtain the governing equations in an interpretable form.

This equation discovery problem presents a challenging paradigm as the exact evolution equations for the N and P states contain a non-polynomial nonlinearity. We expect that our algorithm captures an approximation of this term in the space spanned by (multivariable) polynomials. Equation (7) describes the final active terms of the governing equations obtained after the adaptive growth of the feature space along with their corresponding coefficients.

$$\begin{aligned}
\frac{dN}{dt} &= 27.92\mathbb{L}_1^{(P)} + 0.053\mathbb{L}_1^{(D)} - 199.18\mathbb{L}_1^{(N)}\mathbb{L}_1^{(P)} + 77.13\mathbb{L}_2^{(N)}\mathbb{L}_1^{(P)} \\
&\quad - 194.94\mathbb{L}_3^{(N)}\mathbb{L}_1^{(P)} + 27.90\mathbb{L}_4^{(N)}\mathbb{L}_1^{(P)} + 1.12\mathbb{L}_4^{(P)}\mathbb{L}_2^{(D)} - 51.50\mathbb{L}_5^{(N)}\mathbb{L}_1^{(P)} \\
\frac{dP}{dt} &= -28.65\mathbb{L}_1^{(P)} + 199.18\mathbb{L}_1^{(N)}\mathbb{L}_1^{(P)} - 77.13\mathbb{L}_2^{(N)}\mathbb{L}_1^{(P)} + 196.71\mathbb{L}_3^{(N)}\mathbb{L}_1^{(P)} \\
&\quad - 0.94\mathbb{L}_3^{(N)}\mathbb{L}_3^{(D)} - 27.22\mathbb{L}_4^{(N)}\mathbb{L}_1^{(P)} + 52.12\mathbb{L}_5^{(N)}\mathbb{L}_1^{(P)} \\
\frac{dD}{dt} &= 0.0502\mathbb{L}_1^{(P)} - 0.061\mathbb{L}_1^{(D)} - 0.0003\mathbb{L}_3^{(N)}\mathbb{L}_2^{(D)}
\end{aligned} \tag{7}$$

Amongst the $\binom{9}{3} = 84$ terms, only a few are determined to be active for each of the evolution equations. Once Eq. (7) is simplified using symbolic simplification and scale based thresholding (cutoff 0.1%), we obtain the functional form of the governing equations:

$$\begin{aligned}
\frac{dN}{dt} &= 0.51P - 3.40NP + 11.55N^2P - 36.30N^3P + 124.69N^4P - 382.72N^5P \\
\frac{dP}{dt} &= -0.56P + 3.30NP - 10.78N^2P + 37.76N^3P - 127.16N^4P + 378.60N^5P \\
\frac{dD}{dt} &= 0.0505P - 0.062D - 0.0002N^2D
\end{aligned} \tag{8}$$

We write Eq. (8) in a more concise form as given by Eq. (9). The terms within the parentheses in the first two expressions is the truncated Taylor series for $0.3/(0.3 + N)$ (expanded around $N = 0$, with $\tilde{N} = N/0.3$.) that our algorithm learns. This is the best representation of the non-polynomial nonlinearity in the available subspace. Thus, without any prior information, our adaptive algorithm infers the presence and the best approximation of the present nonlinearity. Unfortunately, our algorithm does not recognize the presence of the $l_N^D D$ term in the equation for dN/dt, but it does capture it in the dD/dt equation. It also incorrectly adds a term ND^2 to the evolution equation of D with a very small coefficient. However, all other active terms are correctly chosen and their corresponding learned coefficients are very close to the actual values from Eq. (6).

This example effectively shows the superiority of our algorithm in identifying the nonlinearities present in the governing equations without any prior information.

$$\begin{aligned}
&\frac{dN}{dt} = 0.51P - P\tilde{N}\left(1.02 - 1.04\tilde{N} + 0.98\tilde{N}^2 - 1.01\tilde{N}^3 + 0.93\tilde{N}^4\right)\\
&\frac{dP}{dt} = -0.56P + P\tilde{N}\left(0.99 - 0.97\tilde{N} + 1.02\tilde{N}^2 - 1.03\tilde{N}^3 + 0.92\tilde{N}^4\right)\\
&\frac{dD}{dt} = 0.0505P - 0.062D + 0.00067ND^2\\
&\dot{N} \approx 0.5P - \frac{PN}{0.3+N}; \dot{P} \approx -0.50P - 0.06P + \frac{PN}{0.3+N}; \dot{D} \approx 0.0505P - 0.062D
\end{aligned} \tag{9}$$

5 Conclusions and Future Work

We investigated the LASSO and developed the dual LASSO feature selection algorithm and dynamic data driven feature learning approaches to solve the problem of discovering governing equations only from state parameter data. After defining the problem and the solution methodology, we addressed the limitations of LASSO in feature selection through a new algorithm, referred to as 'dual LASSO feature selection', that relies on the uniqueness of the dual solution for the active set selection. This was followed by proposing a new methodology to learn the governing equations from scratch by dynamically building the feature library using appropriate orthogonal functional basis. We showcased results of the learning schemes on the classic Lorenz 63 system and also a marine ecosystem model with a non-polynomial nonlinearity. We found that our adaptive subspace algorithm effectively learns a Taylor series approximation of such a nonlinearity, even when no prior information about the presence and the nature of this nonlinearity is provided. Future directions involve extending the ideas of feature library building to the construction of the functions to be added through a mix of a larger family of orthogonal functions. It would be interesting to study the applications of these algorithms in the presence of model and observation noise, and to higher dimensional systems often encountered in science and engineering. Further, using the learned system to guide future observations would also close the loop for the DDDAS paradigm.

Acknowledgments. We are grateful to the Office of Naval Research for support under grants N00014-19-1-2693 (IN-BDA) and N00014-20-1-2023 (MURI ML-SCOPE) to the Massachusetts Institute of Technology.

References

1. Bailey, D.H., Borwein, J.M., Kaiser, A.D.: Automated simplification of large symbolic expressions. J. Symbol. Comput. **60**, 120–136 (2014)
2. Brunton, S.L., Proctor, J.L., Kutz, J.N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems. PNAS **13**, 3932–3937 (2016)

3. Darema, F.: Dynamic data driven applications systems: a new paradigm for application simulations and measurements. In: Bubak, M., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2004. LNCS, vol. 3038, pp. 662–669. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24688-6_86
4. Davis, C.S., Steele, J.H.: Biological/physical modeling of upper ocean processes. Technical report, Woods Hole Oceanographic Institution (1994)
5. Evangelinos, C., Chang, R., Lermusiaux, P.F.J., Patrikalakis, N.M.: Rapid real-time interdisciplinary ocean forecasting using adaptive sampling and adaptive modeling and legacy codes: component encapsulation using XML. In: Sloot, P.M.A., Abramson, D., Bogdanov, A.V., Gorbachev, Y.E., Dongarra, J.J., Zomaya, A.Y. (eds.) ICCS 2003. LNCS, vol. 2660, pp. 375–384. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-44864-0_39
6. Fennel, W., Neumann, T.: Introduction to the modelling of marine ecosystems. Elsevier (2014)
7. Gauraha, N.: Dual lasso selector. arXiv preprint arXiv:1703.06602 (2017)
8. Gupta, A., Haley, P.J., Subramani, D.N., Lermusiaux, P.F.J.: Fish modeling and Bayesian learning for the Lakshadweep Islands. In: OCEANS 2019 MTS/IEEE SEATTLE. pp. 1–10 (2019)
9. Lermusiaux, P.F.J.: Adaptive modeling, adaptive data assimilation and adaptive sampling. Physica D **230**(1), 172–196 (2007). https://doi.org/10.1016/j.physd.2007.02.014
10. Lermusiaux, P.F.J., et al.: Adaptive coupled physical and biogeochemical ocean predictions: a conceptual basis. In: Bubak, M., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2004. LNCS, vol. 3038, pp. 685–692. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24688-6_89
11. Lorenz, E.N.: Deterministic nonperiodic flow. J. Atmos. Sci. **20**(2), 130–141 (1963)
12. Rudy, S.H., Brunton, S.L., Proctor, J.L., Kutz, J.N.: Data-driven discovery of partial differential equations. Sci. Adv. **3**(4), e1602614 (2017)
13. Tibshirani, R.J., et al.: The lasso problem and uniqueness. Electr. J. Stat. **7**, 1456–1490 (2013)
14. Zhang, T.: Adaptive forward-backward greedy algorithm for sparse learning with linear models. In: Advances in Neural Information Processing Systems (2009)
15. Zhao, P., Yu, B.: On model selection consistency of lasso. J. Mach. Learn. Res. **7**, 2541–2563 (2006)