Less greedy equation learning: Balancing interpretability and expressivity through Bayesian model selection

Anonymous authors
Paper under double-blind review

Abstract

In the field of equation learning, exhaustively considering all possible combinations derived from a basis function dictionary is infeasible. Sparse regression and greedy algorithms have emerged as popular approaches to tackle this challenge. However, the presence of strong collinearities poses difficulties for sparse regression techniques, and greedy steps may inadvertently exclude important components of the true equation, leading to reduced identification accuracy. In this article, we present a novel algorithm that strikes a balance between comprehensiveness and efficiency in equation learning. Inspired by stepwise regression, our approach combines the coefficient of determination, R^2, and the Bayesian model evidence, $p(y|\mathcal{M})$, in a novel way. Through three extensive numerical experiments involving random polynomials and dynamical systems, we compare our method against two standard approaches, four state-of-the-art methods, and bidirectional stepwise regression incorporating $p(y|\mathcal{M})$. The results demonstrate that our less greedy algorithm surpasses all other methods in terms of identification accuracy. Furthermore, we discover a heuristic approach to mitigate the overfitting penalty associated with R^2 and propose an equation learning procedure solely based on R^2, which achieves high rates of exact equation recovery.

1 Introduction

Uncovering the underlying laws governing a system is essential for understanding its behavior, and mathematical equations serve as a concise representation of these laws. Equipped with such equations, predictions can be made, and valuable insights can be derived analytically. The pursuit of inferring these governing equations directly from observations has a long history, dating back to Johannes Kepler's deduction of planetary motion laws in 1609. In modern times, significant progress has been made in automated equation inference using machine learning techniques, a discipline commonly referred to as "equation learning" or "symbolic regression."

In contrast to deep learning, equation learning focuses on maximizing model expressivity while minimizing complexity to ensure interpretability. This delicate balance between interpretability and expressivity constitutes the central challenge of equation learning. Other challenges include stable feature selection, solvability of learnt equations, computational feasibility, and handling limited data. Greedy algorithms and regularization techniques applied to regression models built from basis functions are commonly employed to address these challenges.

Greedy algorithms, like stepwise regression, explore the model space iteratively by evaluating scores for individual models. Although these methods are computationally efficient, they tend to significantly reduce the size of the considered model space, often leading to the exclusion of the true model from consideration. On the other hand, regularization transforms the search into an optimization problem, where all candidate models serve as points in the objective function landscape. However, finding the global minimum that corresponds to the true model is not guaranteed, and local minima may only accidentally lead to the true model. Collinearities among basis functions further complicate optimization algorithms due to the jagged nature of the optimization space.

In this work, we tackle the aforementioned challenges of equation learning by enhancing the stepwise regression approach in a less greedy way and exploring multiple score functions. Our method begins with an elimination

process that employs the computationally inexpensive coefficient of determination, R^2, in order to assess almost all individual candidate models belonging to a complexity class. Selecting the best models from that process enables subsequent model selection based on the Bayesian model evidence. The model evidence reflects the probability of a model being the true model for the given data, making it an ideal criterion for penalizing overfitting and addressing the central challenge of equation learning: achieving the optimal balance between interpretability and expressivity. By employing specifically tuned conjugate priors within an empirical Bayes framework, we circumvent the computationally demanding estimation of the evidence and instead derive it analytically.

To evaluate our approach, we employ artificially generated data from known models. This choice allows us to assess whether our strategy can uncover the ground truth model, serving as a testament to its ability to strike the delicate balance between expressivity and interpretability in realistic scenarios involving complex, analytically challenging models. We compare our method against two standard techniques, least absolute shrinkage and selection operator (LASSO) and least-angle regression (LARS), as well as four state-of-the-art methods available in the PySINDy package: sparse relaxed regression (SR3), forward regression orthogonal least squares (FROLS), sequentially thresholded least squares (STLSQ), and best subset selection via mixed-integer optimized sparse regression (MIOSR). Additionally, we propose evidence-based bi-directional stepwise regression (SR) for equation learning using our tuned model evidence. We evaluate the identification accuracy of approximately 80 scenarios for each system and assess the forecasting accuracy based on 100 initial values for each scenario and system, amounting to a total of around 8,000 tests. Our findings demonstrate that our proposed methods outperform other approaches in terms of identifying the correct model and can achieve competitive forecasting accuracy.

Our paper is organized as follows. We begin with a short overview of existing method in Section 2, and introduce our approach and the methods we compare to in Section 3. In Section 4 we present our numerical results, which we discuss in Section 5. We conclude in Section 6 and provide more details of our approach in the appendix.

2 Related work

The literature on equation learning can be roughly divided into three approaches: evolutionary algorithms, tree and neural network representations, and regularized regression. A well-known example of evolutionary algorithms is EUREQA, a commercial software used for symbolic regression Dubčáková (2011); Stoutemyer (2013). A recent open source alternative to EUREQA also using evolutionary algorithms is the python package pySR Cranmer (2023). Tree representations construct equations by combining basic operations and are then employed to minimize regularized objective functions Vaddireddy et al. (2020), using posterior sampling with sparsity-promoting priors Jin et al. (2019), or through mixed-integer linear programming Neumann et al. (2020). Similarly, neural network architectures have been used to represent equations, where network nodes are replaced by expression building blocks Martius & Lampert (2016); Sahoo et al. (2018); Werner et al. (2021). These neural networks are trained with a regularized minimizer applied to objective functions Rackauckas et al. (2020). Another approach, which allows for the incorporation of physics-informed properties such as symmetries, has also been developed Udrescu & Tegmark (2020).

These methods find applications in complementing deep neural networks to enhance generalization Arabshahi et al. (2018) and reducing the data requirement for training Yang et al. (2021). They are also utilized in the analysis of spatio-temporal biological data Nardini et al. (2020) and uncovering complex ecosystem dynamics Chen et al. (2019).

The above approaches are highly non-linear and typically involve complex optimization algorithms. A simpler approach is based on linear regression models, where features are replaced by basis functions derived from observed data. Regularized regression ensures sparse weight estimates on the basis functions Hastie et al. (2009), resulting in concise mathematical expressions. One prominent and widely used method for sparse regression is the LASSO with ℓ_1-regularization Tibshirani (1996). The advantage of ℓ_1-regularization is the convexity of the objective functions, which allows for efficient optimization. However, as we will discuss later, while sparse regression performs well with independent features in reconstruction tasks, the correlations

introduced by basis function expansion can lead to detrimental instability in equation learning. Recent advancements of LASSO, such as SR3, can be found in Zheng et al. (2019); Tibshirani & Friedman (2020).

Greedy algorithms like SR, FROLS, and STLSQ have proven to be more successful than LASSO in equation learning. The latter two algorithms are implemented in the Python package PySINDy, which is designed for the sparse identification of nonlinear dynamics (SINDy) Brunton et al. (2016). SINDy, initially proposed in Brunton et al. (2016) using STLSQ, has seen numerous extensions, including applications to partial differential equations Rudy et al. (2017), improved noise robustness through automated differentiation Kaheman et al. (2020), supplemented with a-posteriori (MAP) estimates Niven et al. (2020), re-weighted ℓ_1-regularization Cortiella et al. (2021), relaxed regularization Champion et al. (2020), and most recently, best subset selection using MIOSR Bertsimas & Gurnee (2023).

In Bayesian linear regression, sparsity-promoting priors are used instead of ℓ_0-regularization for equation learning Nayek et al. (2021). Thresholded sparse regression also employs sparsity-promoting priors Zhang & Lin (2018; 2021). By restricting basis functions to quadratic order, the linear structure of the models allows for deterministic results even in the case of the more challenging ℓ_0-regularization Schaeffer et al. (2018).

3 Less greedy equation learning

The goal of equation learning is to find a function $f(\boldsymbol{x})$ that accurately represents the relationship between the input variables $\boldsymbol{x}$ and the output variable y. In the context of regression models, we consider a dataset consisting of N observations, where the inputs are organized in a design matrix $\boldsymbol{X}$ and the corresponding outputs are modeled by a random variable Y. We can express the relationship between $\boldsymbol{x}$ and y as follows:

$$Y = f(\boldsymbol{x}) + \sigma Z, \tag{1}$$

where $\boldsymbol{x}$ is a row of $\boldsymbol{X}$ (a feature vector), Z is a standard normal random variable, and σ^2 is the variance of the noise term. Our goal is to represent the unknown function $f(\boldsymbol{x})$ using a basis function expansion: We assume that $f(\boldsymbol{x})$ can be represented as a linear combination of p basis functions $k_n(\boldsymbol{x})$,

$$f(\boldsymbol{x}) = \sum_{n=1}^{p} w_n k_n(\boldsymbol{x}), \tag{2}$$

where w_n denotes the weights associated with each basis function. We can construct a basis function matrix $\boldsymbol{K}$, with elements $\boldsymbol{K}_{jn} = k_n(\boldsymbol{x}_j)$. The ordinary least squares (OLS) estimates for the weights w_n are given by Montgomery et al. (2012):

$$\hat{\boldsymbol{w}} = (\boldsymbol{K}^\mathsf{T}\boldsymbol{K})^{-1}\boldsymbol{K}^\mathsf{T}\boldsymbol{y}, \tag{3}$$

where $\boldsymbol{y}$ represents a vector consisting of N samples of the output variable Y. With the estimated weights $\hat{\boldsymbol{w}}$, we can make predictions $\hat{\boldsymbol{y}} = \boldsymbol{K}\hat{\boldsymbol{w}}$.

Common choices for the basis functions $k_n(\boldsymbol{x})$ involve products and powers of the individual features x_j. For example, we can have $k_1(\boldsymbol{x}) = x_1^2$, $k_2(\boldsymbol{x}) = x_1x_2$, $k_3(\boldsymbol{x}) = x_1x_3$, and so on. By imposing restrictions on the maximum power of individual factors and the total number of factors in each $k_n(\boldsymbol{x})$, we can control the total number p of basis functions. Equation learning involves selecting a small subset of $k_n(\boldsymbol{x})$ for which the corresponding weights w_n are estimated, while setting the remaining weights to zero.

Once a choice of basis functions $k_n(\boldsymbol{x})$ is made, the actual learning of the model using equation equation 3 is straightforward. However, the challenging part lies in the selection of the appropriate basis functions. On one hand, we require the model to have enough flexibility (expressivity) to minimize bias, ensuring that it captures the underlying relationship between $\boldsymbol{x}$ and y. On the other hand, we want to avoid overfitting, which can lead to high variance in predictions. This trade-off between bias and variance guides the determination of the number of non-zero weights, denoted as m, or equivalently, the model size. In addition to finding the appropriate model size, we also aim to identify the "correct" set of basis functions $k_n(\boldsymbol{x})$, which corresponds to recovering the true $f(\boldsymbol{x})$ from data generated by equation equation 1.

3.1 Regularized regression

A common way to avoid overfitting is to use regularization,

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w} \in \mathbb{R}^p}{\arg\min} \Big[\big\| \boldsymbol{K}\boldsymbol{w} - \boldsymbol{y} \big\|_2^2 + \lambda \|\boldsymbol{w}\|_q \Big], \tag{4}$$

where $\|\boldsymbol{w}\|_q = [\sum_n |w_n|^q]^{1/q}$, and λ is the Lagrange parameter that sets the strength of the ℓ_q penalty. The standard, sparsity promoting choice is $q = 1$ which is known as the LASSO Tibshirani (1996). The choice $q = 0$, for which $\|\boldsymbol{w}\|_0 = \sum_n \delta_{w_n,0}$ is the number of non-zero weight estimates, is often called *best subset selection* and requires specialized optimization algorithms Zhu et al. (2020); Hastie et al. (2020), like MIOSR Bertsimas & Gurnee (2023).

A relaxed regularization like SR3 can be introduced by letting the penalty act on an auxiliary variable $\boldsymbol{u}$, where distance of $\boldsymbol{u}$ to the actual weights is controlled by an additional regularization term, e.g. $\|\boldsymbol{w} - \boldsymbol{u}\|_q$ Zheng et al. (2019).

3.2 Model selection

Taking a different route, sparse solutions of equation 3 may be realized by model selection. A computationally cheap criteria for model selection is the coefficient of determination,

$$R^2 = \frac{\boldsymbol{y}^{\mathrm{T}}\, \boldsymbol{K}\, (\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K})^{-1}\, \boldsymbol{K}^{\mathrm{T}}\, \boldsymbol{y}}{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}}, \tag{5}$$

here in a simplified form for standardized y Montgomery et al. (2012).

However, R^2 is known for its lack of overfitting penalty as it would just increase with decreasing sparsity. Alternatives like adjusted versions of R^2 exist to incorporate an overfitting penalty, but here we make use of the Bayesian model evidence

$$p(\mathcal{M}) \propto p(y|\mathcal{M}) = \int \mathrm{d}\sigma \int \mathrm{d}\boldsymbol{w}\; p_{\mathrm{li}}(y|\boldsymbol{w}, \sigma, \mathcal{M})\, p_{\mathrm{pr}}(\boldsymbol{w}, \sigma), \tag{6}$$

where a selection of basis functions $k_n(\boldsymbol{x})$ defines a model $\mathcal{M}$ and a likelihood distribution $p_{\mathrm{li}}(y|\boldsymbol{w}, \sigma, \mathcal{M})$ via equation 1 and equation 2. For a conjugate prior $p_{\mathrm{pr}}(\boldsymbol{w}, \sigma)$, the marginalization above can be done analytically and $p(y|\mathcal{M})$ is known exactly, as detailed in appendix B.

The use of $p(y|\mathcal{M})$ is often motivated by its excellent overfitting penalizing properties, which can be ascribed to the fact that $p(y|\mathcal{M})$ is proportional to the probability $p(\mathcal{M})$ of the model $\mathcal{M}$ being the true model for the data $(\boldsymbol{y}, \boldsymbol{X})$ Murphy (2012). The downside of using $p(y|\mathcal{M})$ is the imperative to specify $p_{\mathrm{pr}}(\boldsymbol{w}, \sigma)$ even in cases of scarce prior knowledge, which can have a significant impact on $p(y|\mathcal{M})$. Here we use the empirical Bayes method to fix $p_{\mathrm{pr}}(\boldsymbol{w}, \sigma)$ and exploit the normality property of linear regression which implies that estimates $\hat{\boldsymbol{w}}$ are normally distributed with the mean given by the true values of $\boldsymbol{w}$ and the variance given by

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}})}{N - p} \tag{7}$$

The details of this procedure is included in appendix A.

3.3 Stepwise regression (SR)

A standard greedy algorithm to select basis functions $k_n(\boldsymbol{x})$ is stepwise regression in which a criterion like adjusted R^2, F-statistics or other information criteria (e.g. AIC, BIC) are used Montgomery et al. (2012). Here we promote the evidence $p(y|\mathcal{M})$ as criterion, which is rarely used for SR due to its intricacies in terms of prior selection and computational cost Hastie et al. (2009). Our stepwise procedure is bi-directional, that is, we start with an empty model and select the $k_n(\boldsymbol{x})$ that maximizes $p(y|\mathcal{M})$, add a second $k_n(\boldsymbol{x})$ maximizing $p(y|\mathcal{M})$, and so on (forward selection). Once $p(y|\mathcal{M})$ cannot be increased further by adding more $k_n(\boldsymbol{x})$

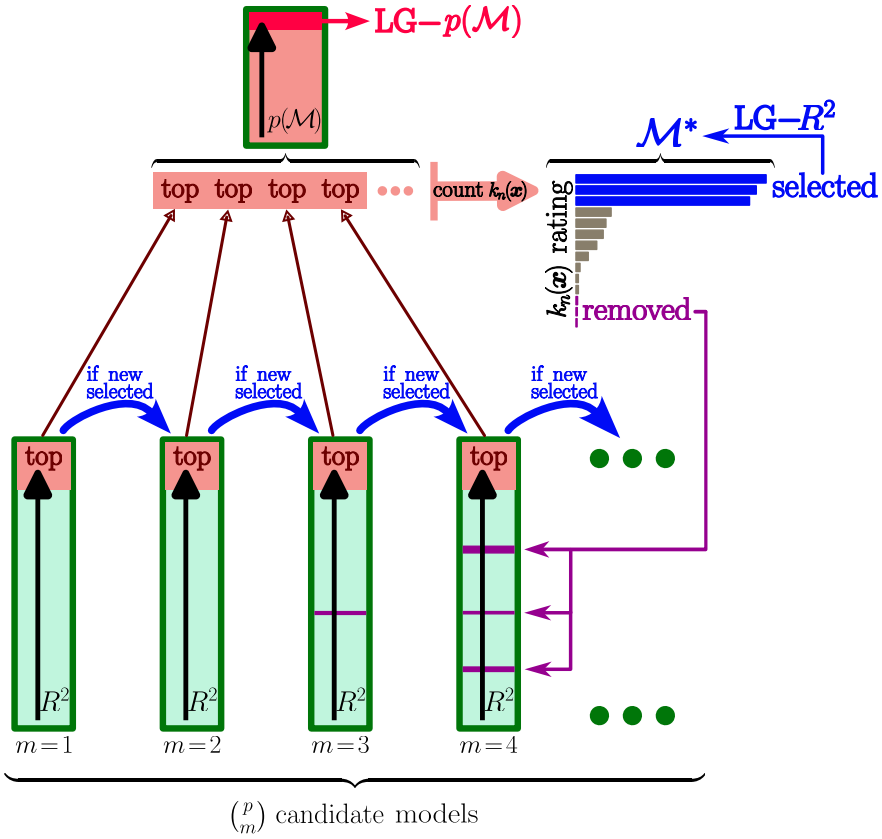

Figure 1: Schematic representation of algorithm LG-R^2 and LG-$p\mathcal{M}$. Starting point is the set of candidate models built from p basis functions $k_n(\boldsymbol{x})$. For fixed model size m, R^2 is computed for all $\binom{p}{m}$ candidate models. Incrementing m, top models in terms of R^2 are collected, from which basis functions $k_n(\boldsymbol{x})$ are rated based on counts of $k_n(\boldsymbol{x})$ in these top models. Typically, incrementing m, new $k_n(\boldsymbol{s})$ with high rates are found and selected for the inferred model $\mathcal{M}^*$ by algorithm LG-R^2 while m is smaller than the true model size. The iteration in m therefore terminates if no new $k_n(\boldsymbol{s})$ are selected for $\mathcal{M}^*$. Basis functions $k_n(\boldsymbol{x}$ that are hardly selected are not considered for building candidate models in the next m-step. The top models selected by R^2 are scored again by $p(\mathcal{M})$ of which the best model is the output of algorithm LG-$p(\mathcal{M})$.

to the model, we start removing $k_n(\boldsymbol{x})$ in the same fashion until again $p(y|\mathcal{M})$ is maximized (backward selection). We continue forward and backward selection until $p(y|\mathcal{M})$ cannot be increased in either selection direction. From including the backward direction and due to the overfitting penalty of $p(y|\mathcal{M})$ we expect our procedure to be more parsimonious in terms of model size.

Other stepwise regression procedures are FROLS and STLSQ. In FROLS, only forward selection is applied, but taking equation 4 with $q=2$ (Ridge regression) as criterion to be minimized Billings (2013). Similarly, but in a backward selection procedure, STLSQ starts with the full model, and alternates between Ridge regression and removing terms $k_n(\boldsymbol{x})$ with weights w_n below a pre-defined threshold Brunton et al. (2016). A similar stepwise algorithm is least-angle regression (LARS) where the correlation between $k_n(\boldsymbol{x})$ and residuals is used to build the model in a forward procedure.

3.4 Less greedy (LG)

A non-greedy algorithm would consider all 2^p-1 combinations of $k_n(\boldsymbol{x})$ which obviously is computationally infeasible. However, since we are interested in parsimonious models, we may choose a small model size m and explore all $\binom{p}{m}$ possible combinations of $k_n(\boldsymbol{x})$ within that budget. For a fixed model size, overfitting

penalty is not important, and we may use R^2 which is particularly cheap to compute for standardized data, c.f. equation 5. In this way, we can single out the best models in terms of R^2 for different budgets m.

To find the best model size m, we employ the overfitting penalty property of the model evidence $p(y|\mathcal{M})$. Starting with $m=1$, we compute $p(y|\mathcal{M})$ for the best models selected by R^2, and continue to do so for increasing m until a stopping criterion is reached.

As a stopping criterion, we may use the first decrease of $p(y|\mathcal{M})$, but we propose a heuristic criterion purely based on R^2 which proved superior in our study. The R^2 criterion we propose builds on the observation that the true $k_n(\boldsymbol{x})$ have a tendency to be consistently selected in the top R^2 models. We therefore keep increasing m until no new $k_n(\boldsymbol{x})$ is selected consistently, and then build the inferred model from those consistently selected terms. Since $\binom{p}{m}$ still becomes very large for larger model sizes, we add an additional pruning step, in which all $k_n(\boldsymbol{x})$ that have consistently not been selected are removed from the basis expansion.

The stopping criterion and the pruning step classifies our procedure as a greedy algorithm. However, due to the exhaustive search for each considered m, we consider a drastically larger model space than other greedy algorithms and are as such significantly less greedy.

While this is a heuristically developed method, it is motivated by the reasonable assumption that the maximization of R^2 is dominated by the true $k_n(\boldsymbol{x})$. Therefore, as soon as we look at the R^2 values of models one term larger than the true model, the procedure is forced to randomly select one extra term in addition to the consistently selected true terms.

For a direct comparison of this rather heuristic selection method with an established method, we also select the model that maximizes $p(y|\mathcal{M})$ out of the top models selected by R^2 across all model sizes. To the first method solely based on R^2 we refer to as LG-R^2, to the second method using a final selection via $p(y|\mathcal{M})$ we refer to as LG-$p(\mathcal{M})$. We illustrate our procedure in figure 1. More details, explicit algorithms, and an illustration of the claims made can be found in appendix C.

It is worth noting that our procedure can also be used in cases where $p(y|\mathcal{M})$ needs to be estimated by computationally costly evidence estimators, as we effectively reduce the pool of candidate models to just a few.

3.5 Model class

Apart from the equations that can directly be written in the form of equation 2, a prominent application of equation learning is the learning of dynamical systems,

$$\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t)), \tag{8}$$

where $\dot{\boldsymbol{x}}(t)$ denotes the time derivative of $\boldsymbol{x}(t)$. To map this problem to equation 2, the response variable can be computed from finite differences $y_i = \frac{x_{i+1}-x_i}{\Delta t}$ for a fixed time step Δt.

A restriction for the regression models to stay linear in its parameters is that parameters of basis functions may only enter as weights w. Basis functions like e^{ax}, $\ln(a+x)$, $\cos ax$, x^a, $\frac{1}{(a+x)^m}$, ... with internal parameter a are not suitable.

This restriction might seem quite limiting. On the other hand, the function $f(\boldsymbol{x}(t), \boldsymbol{w})$ defining a dynamical system typically is linear in its parameters $\boldsymbol{w}$. The reason for that is that these functions often reproduce when differentiated, which can be used to eliminate these functions, retrieving the standard linear form in equation 2. Some special functions like Bessel, Hankel, Struve and Meijer functions are even defined as solutions of differential equations linear in their coefficients. In general, by considering the differentiated response variable, $y \mapsto \frac{\mathrm{d}y}{\mathrm{d}x} \simeq \frac{y(x_{i+1})-y(x_i)}{x_{i+1}-x_i}$, if necessary to higher order, we can learn a surprisingly broad class of equations relating y and $\boldsymbol{x}$, even relations that do not exist in closed form.

4 Numerical experiments

We compare all introduced methods (LASSO, LARS, LG-R^2, LG-$p(\mathcal{M})$, SR, SR3, FROLS, STLSQ, MIOSR) in three numerical experiment. For LASSO and LARS, we use the PYTHON package `scikit-learn` Pedregosa

Table 1: Equation learning techniques used in this study building on regression models as in equations 1, 2.

Acronym	Full name	Description	Reference
LASSO	Least absolute shrinkage and selection operator	Regularized regression as in equation 4 for $q=1$.	Tibshirani (1996).
LARS	Least angle regression	Forward stepwise regression using correlations between basis functions k_n and residuals.	Efron et al. (2004).
LG–R^2	Less-greedy R^2 elimination	For an increasing model size m, the R^2 score is calculated for all candidate models. Based on the number of times basis functions k_n contribute to the models with largest R^2, kernel functions k_n are rated and the lowest rated k_n are excluded. Once no new highly rated R^2 are found, the iteration in m terminates and the model with largest R^2 is returned.	This work, appendix C.
LG–$p(\mathcal{M})$	Less-greedy Bayesian model selection	Same as LG–R^2, but a selection of models is returned based on maximal Bayesian model evidence $p(\mathcal{M})$.	This work, appendix C.
SR	Bi-directional stepwise regression	Iterated forward and backward selection using the model evidence Bayesian model evidence $p(\mathcal{M})$ as score.	Hastie et al. (2009), this work for $p(\mathcal{M})$, appendix B.
SR3	Sparse relaxed regression	The regularization is put on an auxiliary variable u and an extra distance term like $\|w-u\|_q$ is added to equation 4.	Zheng et al. (2019).
FROLS	Forward regression orthogonal least-squares	Forward selection stepwise regression with score given by equation 4 for $q=2$.	Billings (2013)
STLSQ	Sequentially thresholded least squares	Backward selection stepwise regression using results of regularized regression for $q=2$ in equation 4.	Brunton et al. (2016)
MIOSR	Best subset selection via mixed-integer optimized sparse regression	Formulation of regularized regression equation 4 for $q=0$ as a mixed-integer linear program and specialized algorithms.	Bertsimas & Gurnee (2023)

et al. (2011), for LG-R^2, LG-$p(\mathcal{M})$ and SR we use our own implementation, and for SR3, FROLS, STLSQ, MIOSR we use the PYTHON package `PySINDy` de Silva et al. (2020); Kaptanoglu et al. (2022). As mixed-integer optimizer for MIOSR we used GUROBI with an academic license Gurobi Optimization, LLC (2023). For all methods from `scikit-learn` and `PySINDy` we used 5-fold cross validation and 3 refinement steps to determine optimal hyperparameters for each application separately. Our own methods, LG-R^2 and LG-$p(\mathcal{M})$, are not very sensitive to hyperparameters and we worked out the universally best values which, in contrast to the methods we compare to, were used for all applications. The SR method we implemented comes without hyperparameters. Table 1 gives an overview over all considered methods. More details can be found in appendix C.

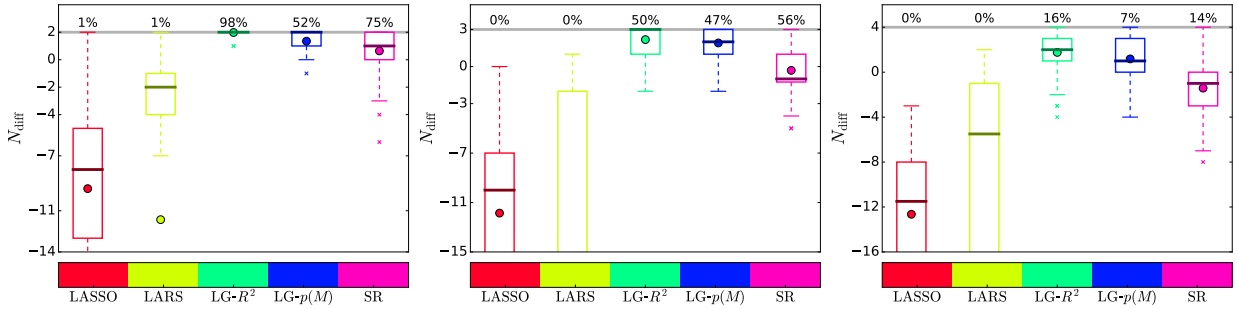

Figure 2: Identification accuracy of learning 100 random polynomials measured in terms of the difference N_{diff} between the number of true terms found and wrong terms found. We generated polynomials with 2, 3 and 4 terms (left to right), which also constitutes the highest possible value for N_{diff} and is indicated by a gray horizontal line. The percentages above this line indicate how often all exact terms and no wrong terms have been recovered.

In all experiments, the data is artificially generated, with the advantage that we know the true model. In the first experiment, we assess the identification accuracy of 100 random polynomials. Since `PySINDy` is tailored to learning dynamical systems and cannot directly be applied to learning polynomials, we omitted those methods in this experiment. In the second and third experiment the methods are applied to two (chaotic) dynamical systems, where, for the sake of readability, we omitted LARS due to its similar performance compared to LASSO.

In all experiments we have three features $\boldsymbol{x} = (x_1, x_2, x_3)$, and we used a basis function expansion of polynomials where the power of individual factors was limited to a maximum of 4, and the combined power of terms $k_n(\boldsymbol{x})$ to a maximum of 6. For instance, $k(\boldsymbol{x}) = x_1^3\, x_2\, x_3^2$ would be a valid basis function, while $k(\boldsymbol{x}) = x_1^5\, x_2$ would be excluded for exceeding the individual power limit of 4, as would $k(\boldsymbol{x}) = x_1^4\, x_2\, x_3^2$ for exceeding the combined power limit of 6. The resulting feature dimension of K is $p = 72$.

The results of each experiment are illustrated statistically by boxplots, where the box indicates the interquartile range, the whiskers extend by a factor 1.5 beyond the box, and outcomes exceeding the whiskers are shown as individual symbols. The median is shown as a darker horizontal line, the mean as a circle.

4.1 Random polynomials

We randomly generated 100 polynomials with 2, 3, and 4 non-zero weights respectively. We restricted the terms of the polynomials to have a maximum collective power $M_2 = 4$, where individual features are restricted to maximum power $M_1 = 2$. The non-zero weights are randomly selected with equal probabilities and their values are uniformly sampled from the set $[-4, -1] \cup [1, 4]$. We generated artificial data $\boldsymbol{X}$ for each polynomial by sampling from normal distributions with means randomly selected from the interval $[-20, 20]$ and standard deviations such that 5% of their probability mass overlap respectively. For the polynomials sized 2, 3, and 4 we generated $N = 20$ and $N = 65$ and $N = 95$ datapoints, respectively. Plugging X into equation 1 with $f(\boldsymbol{x}_i)$ given by the random polynomial, and corrupting the output with normal noise with standard deviation $\sigma = 0.01$, we generate data for the response variable y_i.

The statistics of the identification accuracy for the 100 polynomials are shown in figure 2.

4.2 Lorenz system

As a first dynamical system to test our method, we use the chaotic Lorenz system defined as

$$\begin{aligned} \dot{x}(t) &= \epsilon\big(y(t) - x(t)\big), \\ \dot{y}(t) &= x\big(\rho - z(t)\big) - y, \\ \dot{z}(t) &= x(t)y(t) - \beta z(t). \end{aligned} \tag{9}$$

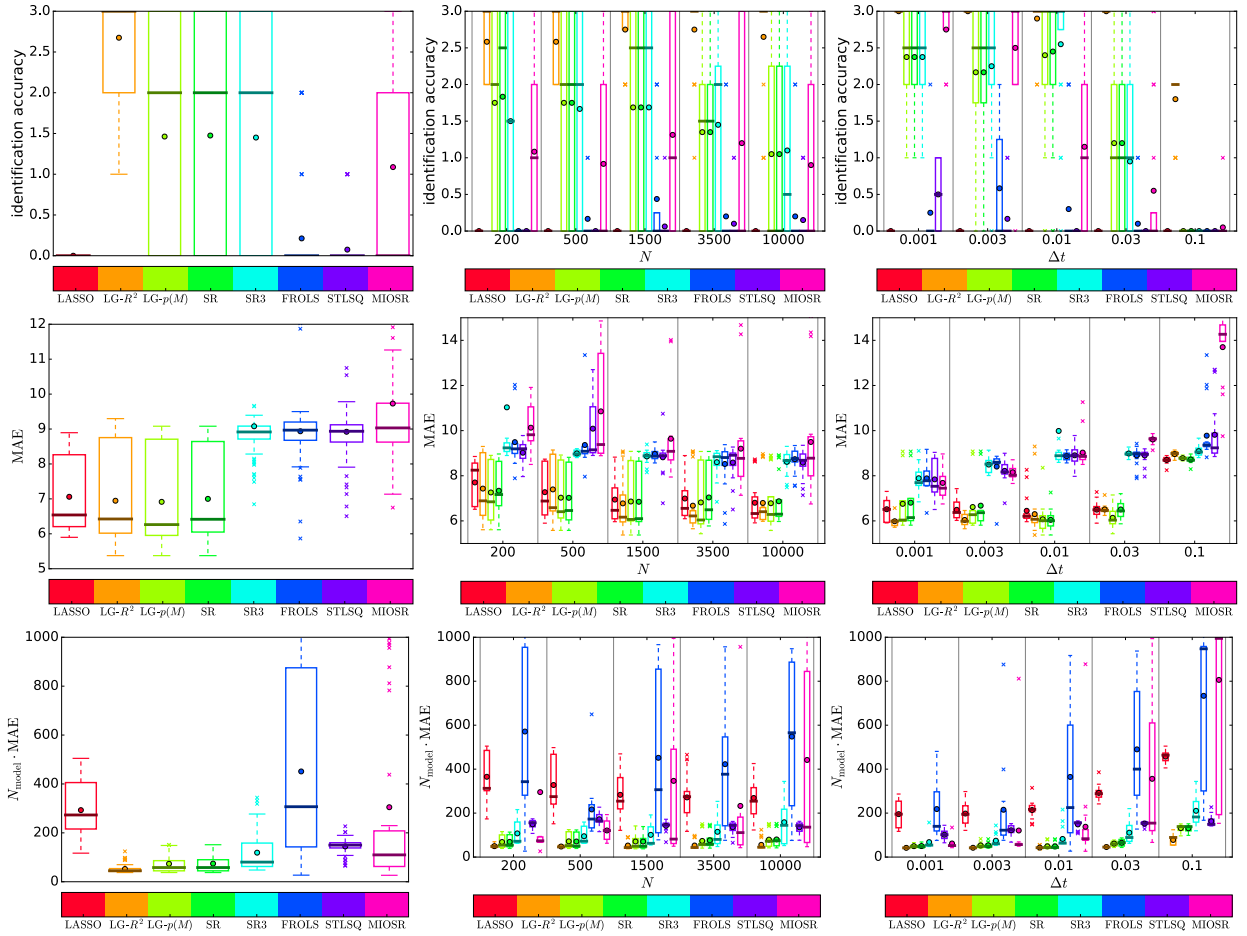

Figure 3: Statistical results for models learnt from data created by solving the Lorenz system equation 9. In each row of the nine plots a different metric is display: the number of equations identified correctly, the MAE to solutions of the true model, the MAE multiplied by the size of the learnt model. The first column uses the statistics across all scenarios, the second column splits it up in terms of number of datapoints N, and the third column in terms of timestep width Δt.

The parameters are fixed to its standard values $\epsilon = 10$, $\rho = 28$ and $\beta = 8/3$. As initial condition, we use $(x_0 = -8, y_0 = 8, z_0 = 27)$. We obtain between $N = 200$ and $N = 10000$ data points by solving equation 9 numerically for timestep widths between $\Delta t = 0.001$ and $\Delta t = 0.1$. We corrupt the solutions with normal noise levels between $\sigma = 0.001$ and $\sigma = 0.1$. Forcing the simulation time to be larger than $T = 2$, we thus have 80 different scenarios with varying N, Δt, σ and T. We apply the equation learning methods to all scenarios to obtain some statistics on identification accuracy and mean-absolute error (MAE).

We measure the identification accuracy as the number of correctly identified equations comprising the model equation 9. The MAE is obtained by solving the learnt dynamical system equations numerically for randomly selected initial values, and comparing the result with the solution we get solving the true equations 9 for the same initial condition.

To further increase the sample size for the statistical evaluation of the methods, and to exclude the possibility to have a particular initial condition that suits a certain method by chance, we sample 100 initial values from the Lorenz attractor and solve each learnt model from all scenarios and equation learning models for the same 100 initial values.

In figure 3 we show the statistical results of our experiment. To emphasize the goal to parsimoniously learn models, we also plot the statistics of the MAE multiplied by the size N_{model} of the learnt model. In addition to the overall statistics, we also show the dependency on N and Δt (splitting the data up in terms of σ or T did not reveal any insights).

4.3 Rabinovich-Fabrikant equations

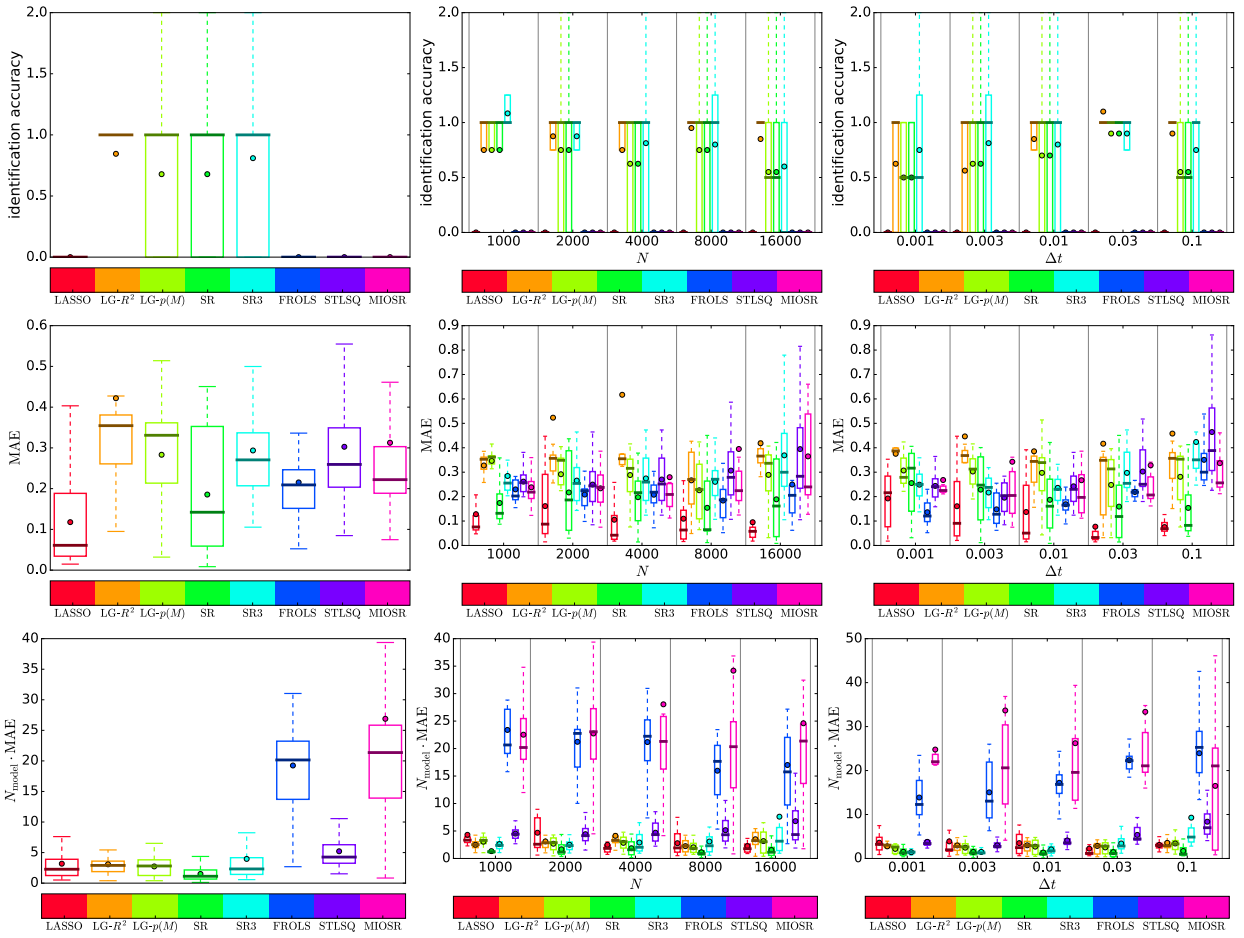

Figure 4: Statistical results from models learnt from data created by solving the Rabinovich-Fabrikant system equation 10. The plots have the same structure as in figure 3: each row shows a different metric, whereas the columns show the overall statistics or split up in terms of datapoints N or Δt.

As a second dynamical system, we use the Rabinovich-Fabrikant equations

$$\begin{aligned}
\dot{x}(t) &= y(t)\big(z(t) - 1 + x(t)^2\big) + \gamma x(t), \\
\dot{y}(t) &= x(t)\big(3z(t) + 1 - x(t)^2\big) + \gamma y(t), \\
\dot{z}(t) &= -2z(t)\big(\alpha + x(t)y(t)\big).
\end{aligned} \tag{10}$$

We choose the parameter values $\alpha = 0.14$ and $\gamma = 0.1$, and the initial conditions $(x_0 = -1.5, y_0 = 0, z_0 = 1)$. We consider 84 scenarios comprising values $N = 1000$ to $N = 16000$, $\Delta t = 0.001$ to $\Delta t = 0.1$, and $\sigma = 0.0001$ and $\sigma = 0.01$, where we enforce $T \geq 5$. We solve equation 10 numerically and corrupt the solutions with noise as for the Lorenz system.

The results are shown in figure 4 in the same fashion as for the Lorenz system.

The PYTHON code to reproduce these results have been included in the supplementary material.

5 Discussion

In this section, we discuss the results of the numerical experiments shown in the previous section.

We begin with the learning of the random polynomials and stress the difficulty of this task, which consists in various aspects: i) Randomly generated polynomials are not hand-picked examples where some fine-tuning is possible. ii) The artificially generated data was not tested to represent the polynomial uniquely. iii) Even if an inferred polynomial deviating from the the true polynomial would describe the data well, it does not contribute positively to the identification accuracy.

With these aspects in mind, it is quite remarkable that overall many of the polynomials could be recovered, as shown in figure 2. Particularly striking is the success rate of LG-R^2 solely based on R^2, whereas the LASSO and LARS clearly overfit in terms of model size. However, the overall accuracy clearly declines with increasing number of non-zero terms. This can be explained with a higher chance of terms being selected spuriously leading to too early or too late stopping of exploring the required number of terms. Also the risk of erroneously removal of true terms from $\boldsymbol{K}$ increases. We are confident that more refined stopping and removal criteria can overcome these inaccuracies. In combining several criteria, we see an opportunity to improve the accuracy even further.

The Lorenz system has between 2 and 3 terms per equation, and as such was learnt quite successfully, as the first row of figure 3 shows. The highest identification accuracy is again achieved by LG-R^2, the LASSO was not able to identify any equation, and of the SINDy methods the relaxed regularized regression SR3 and the best subset-selection using mixed integer optimization MIOSR performed best. In terms of MAE, the LASSO is among the best, but requires significantly larger models is shown by $N_{\text{model}}\,$MAE. Most of the methods are relatively robust against N and Δt. Interestingly, $N_{\text{model}}\,$MAE worsens with more datapoints in the case of MIOSR but improves for the LASSO, signifying that ℓ_0 regression "sees" more in the data than there is if given enough data, while ℓ_1 requires more data to produce smaller models (c.f. equation 4). Regarding robustness against larger timesteps, it turns out that MIOSR is particularly sensitive with deteriorating performance for smaller Δt, as well as FROLS and the LASSO to a lesser extent.

The Rabinovich-Fabrikant system turned out to be particularly difficult to learn. The identification accuracy for the LG methods, the SR using $p(\mathcal{M})$ and the relaxed regularized SR3 was on average just below 1 equation, while the other methods failed to identify any equation in all scenarios. No method was able to learn the complete system of equations correctly. Interestingly, the MAE was still reasonable, in particular the LASSO performed well. Splitting up in N and Δt do not provide any additional insight in particular and are just shown for completeness.

A particular problem in learning dynamical systems is that there is no guarantee that the learnt models are solvable. In cases where the numerical solutions failed, we excluded the results from the statistics and kept a record of how often this happened for the various equation learning methods. For the Lorenz system, LASSO, SR3 and STLSQ failed at about 1%, MIOSR at about 20%, and the other methods always produced solvable systems. This is a little different for the Rabinovich-Fabrikant system, where all methods produced unsolvable models between about 1% and 15%, with STLSQ the most problematic followed by SR3 and LG-R^2.

6 Conclusions

We extensively tested our methods LG-R^2, LG-$p(\mathcal{M})$ and the bi-directional SR employing $p(\mathcal{M})$ against the standard methods LASSO and LARS, as well as state-of-the-art methods SR3, FROLS, STLSQ and MIOSR. To our knowledge, we are the first to explore equation learning based on an exhaustive candidate model evaluation outperforming existing state-of-the-art methods in terms of identification accuracy, and at least on equal terms regarding forecast quality.

A disadvantage of our less greedy approach is the higher computational cost and the absence of clear time complexity measures. However, a direct advantage is that the model evaluation is trivially parallelizable. Also the little amount of data needed for high success rates is striking – tests on two sets of random polynomials where even done with less datapoints than rows in K. Testing candidate models individually also allows for great flexibility when it comes to constraints or conditions on models such as solvability, as well as eliminating

the risk of getting stuck in local minima of an objective function. It also allows combining several criteria for model and feature selection, in particular complementing existing methods in an independent way with the potential of synergy effects.

The observation that learnt chaotic model comprising all true terms plus a few extra terms can decrease the MAE has an interesting implication worth exploring in a future work: It seems possible to learn correction terms from data that lead to a better forecast horizon than the true model itself.

Finally, we contributed towards the utilization of the Bayesian model evidence $p(\mathcal{M})$ in equation learning. Here, we benefit from using a conjugate prior for which $p(\mathcal{M})$ can be computed analytically, and showed in our numerical experiments that our choice of empirical prior is well suited for the tasks considered here. However, in general, one would like to have the freedom to select any prior which may entail particularly computationally costly evidence estimators. Performing the less greedy search with R^2 (and possibly other criteria) can boil down the number of candidate models to a feasible number, an approach planned to be explored in a future work.

Considering these possibilities and the promising identification accuracy achieved, we hope to open a new avenue of equation learning.

Acknowledgements

(after acceptance)

References

Forough Arabshahi, Sameer Singh, and Anima Anandkumar. Combining symbolic expressions and black-box function evaluations in neural programs. In *ICLR*, 2018.

Dimitris Bertsimas and Wes Gurnee. Learning sparse nonlinear dynamics via mixed-integer optimization. *Nonlinear Dynamics*, 1 2023. ISSN 0924-090X. doi: 10.1007/s11071-022-08178-9. URL `https://link.springer.com/10.1007/s11071-022-08178-9`.

Stephen Billings. Nonlinear system identification: Narmax methods in the time, frequency, and spatio-temporal domains. *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains*, 08 2013. doi: 10.1002/9781118535561.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113 (15):3932–3937, apr 2016. ISSN 0027-8424. doi: 10.1073/pnas.1517384113. URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1517384113`.

Kathleen Champion, Peng Zheng, Aleksandr Y. Aravkin, Steven L. Brunton, and J. Nathan Kutz. A Unified Sparse Optimization Framework to Learn Parsimonious Physics-Informed Models From Data. *IEEE Access*, 8:169259–169271, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.3023625. URL `https://ieeexplore.ieee.org/document/9194760/`.

Yize Chen, Marco Tulio Angulo, and Yang Yu Liu. Revealing Complex Ecological Dynamics via Symbolic Regression. *BioEssays*, 41(12):1–9, 2019. ISSN 15211878. doi: 10.1002/bies.201900069.

Alexandre Cortiella, Kwang Chun Park, and Alireza Doostan. Sparse identification of nonlinear dynamical systems via reweighted l1-regularized least squares. *Computer Methods in Applied Mechanics and Engineering*, 376:1–33, 2021. ISSN 00457825. doi: 10.1016/j.cma.2020.113620.

Miles Cranmer. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl, 2023. URL `http://arxiv.org/abs/2305.01582`.

Brian de Silva, Kathleen Champion, Markus Quade, Jean-Christophe Loiseau, J. Kutz, and Steven Brunton. Pysindy: A python package for the sparse identification of nonlinear dynamical systems from data. *Journal*

of Open Source Software, 5(49):2104, 2020. doi: 10.21105/joss.02104. URL https://doi.org/10.21105/joss.02104.

Renáta Dubčáková. Eureqa: software review. *Genetic Programming and Evolvable Machines*, 12(2):173–178, jun 2011. ISSN 1389-2576. doi: 10.1007/s10710-010-9124-z. URL http://link.springer.com/10.1007/s10710-010-9124-z.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004. doi: 10.1214/009053604000000067. URL https://doi.org/10.1214/009053604000000067.

Vincent Fortuin. Priors in Bayesian Deep Learning: A Review. (1):1–28, 2021. URL http://arxiv.org/abs/2105.06868.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023. URL https://www.gurobi.com.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2nd edition, 2009. ISBN 978-0-387-84857-0. doi: 10.1007/978-0-387-84858-7. URL http://link.springer.com/10.1007/978-0-387-84858-7.

Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Statistical Science*, 35, 11 2020. ISSN 0883-4237. doi: 10.1214/19-STS733. URL https://projecteuclid.org/journals/statistical-science/volume-35/issue-4/Best-Subset-Forward-Stepwise-or-Lasso-Analysis-and-Recommendations-Based/10.1214/19-STS733.full.

Ying Jin, Weilin Fu, Jian Kang, Jiadong Guo, and Jian Guo. Bayesian Symbolic Regression. 2019. URL http://arxiv.org/abs/1910.08892.

Kadierdan Kaheman, Steven L. Brunton, and J. Nathan Kutz. Automatic Differentiation to Simultaneously Identify Nonlinear Dynamics and Extract Noise Probability Distributions from Data. pp. 1–30, 2020. URL http://arxiv.org/abs/2009.08810.

Alan A. Kaptanoglu, Brian M. de Silva, Urban Fasel, Kadierdan Kaheman, Andy J. Goldschmidt, Jared Callaham, Charles B. Delahunt, Zachary G. Nicolaou, Kathleen Champion, Jean-Christophe Loiseau, J. Nathan Kutz, and Steven L. Brunton. Pysindy: A comprehensive python package for robust sparse system identification. *Journal of Open Source Software*, 7(69):3994, 2022. doi: 10.21105/joss.03994. URL https://doi.org/10.21105/joss.03994.

Kevin H. Knuth, Michael Habeck, Nabin K. Malakar, Asim M. Mubeen, and Ben Placek. Bayesian evidence and model selection. *Digital Signal Processing: A Review Journal*, 47:50–67, 2015. ISSN 10512004. doi: 10.1016/j.dsp.2015.06.012. URL http://dx.doi.org/10.1016/j.dsp.2015.06.012.

Georg Martius and Christoph H. Lampert. Extrapolation and learning equations. 2016. URL http://arxiv.org/abs/1610.02995.

D. Montgomery, E. Peck, G. Vining, and an O'Reilly Media Company Safari. *Introduction to Linear Regression Analysis, 5th Edition*. John Wiley & Sons, 2012. ISBN 978-0-470-54281-1. URL https://books.google.co.za/books?id=hD46zQEACAAJ.

K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning series. MIT Press, 2012. ISBN 9780262018029. URL https://books.google.co.za/books?id=NZP6AQAAQBAJ.

John T. Nardini, John H. Lagergren, Andrea Hawkins-Daarud, Lee Curtin, Bethan Morris, Erica M. Rutter, Kristin R. Swanson, and Kevin B. Flores. Learning Equations from Biological Data with Limited Time Samples. *Bulletin of Mathematical Biology*, 82(9), 2020. ISSN 15229602. doi: 10.1007/s11538-020-00794-z.

R. Nayek, R. Fuentes, K. Worden, and E. J. Cross. On spike-and-slab priors for Bayesian equation discovery of nonlinear dynamical systems via sparse linear regression. *Mechanical Systems and Signal Processing*, 161:1–22, 2021. ISSN 10961216. doi: 10.1016/j.ymssp.2021.107986.

Pascal Neumann, Liwei Cao, Danilo Russo, Vassilios S. Vassiliadis, and Alexei A. Lapkin. A new formulation for symbolic regression to identify physico-chemical laws from experimental data. *Chemical Engineering Journal*, 387:123412, may 2020. ISSN 13858947. doi: 10.1016/j.cej.2019.123412. URL https://doi.org/10.1016/j.cej.2019.123412https://linkinghub.elsevier.com/retrieve/pii/S1385894719328256.

Robert K. Niven, Ali Mohammad-Djafari, Laurent Cordier, Markus Abel, and Markus Quade. Dynamical System Identification by Bayesian Inference. In *Proceedings of the 22nd Australasian Fluid Mechanics Conference AFMC2020*, dec 2020. doi: 10.14264/692fcb8. URL https://espace.library.uq.edu.au/view/UQ:692fcb8.

A. O'Hagan and M.G. Kendall. *Advanced Theory of Statistics: Bayesian inference. Volume 2B*. Number v. 2, pt. 2 in KENDALL, MAURICE GEORGE//KENDALL'S ADVANCED THEORY OF STATISTICS 6TH ED. Edward Arnold, 1994. ISBN 9780340529225. URL https://books.google.co.za/books?id=DlrEMgEACAAJ.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal Differential Equations for Scientific Machine Learning. 2020. ISSN 2331-8422. URL http://arxiv.org/abs/2001.04385.

Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):1–7, apr 2017. ISSN 2375-2548. doi: 10.1126/sciadv.1602614. URL https://www.science.org/doi/10.1126/sciadv.1602614.

Subham S. Sahoo, Christoph H. Lantpert, and Georg Martius. Learning equations for extrapolation and control. *35th International Conference on Machine Learning, ICML 2018*, 10:7053–7061, 2018.

Hayden Schaeffer, Giang Tran, and Rachel Ward. Extracting Sparse High-Dimensional Dynamics from Limited Data. *SIAM Journal on Applied Mathematics*, 78(6):3279–3295, jan 2018. ISSN 0036-1399. doi: 10.1137/18M116798X. URL https://epubs.siam.org/doi/10.1137/18M116798X.

David R. Stoutemyer. Can the Eureqa Symbolic Regression Program, Computer Algebra, and Numerical Analysis Help Each Other? *Notices of the American Mathematical Society*, 60(06):713, jan 2013. ISSN 0002-9920. doi: 10.1090/noti1000. URL http://www.ams.org/jourcgi/jour-getitem?pii=noti1000.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

Robert Tibshirani and Jerome Friedman. A Pliable Lasso. *Journal of Computational and Graphical Statistics*, 29(1):215–225, 2020. ISSN 15372715. doi: 10.1080/10618600.2019.1648271.

Silviu Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16), 2020. ISSN 23752548. doi: 10.1126/sciadv.aay2631.

Harsha Vaddireddy, Adil Rasheed, Anne E. Staples, and Omer San. Feature engineering and symbolic regression methods for detecting hidden physics from sparse sensor observation data. *Physics of Fluids*, 32 (1):015113, jan 2020. ISSN 1070-6631. doi: 10.1063/1.5136351. URL http://aip.scitation.org/doi/10.1063/1.5136351.

W Von Der Linden, R Preuss, and V Dose. The prior-predictive value: A paradigm of nasty multi-dimensional integrals. In *Maximum Entropy and Bayesian Methods Garching, Germany 1998*, pp. 319–326. Springer, 1999. URL https://link.springer.com/chapter/10.1007/978-94-011-4710-1_31.

Matthias Werner, Andrej Junginger, Philipp Hennig, and Georg Martius. Informed Equation Learning. pp. 1–24, may 2021. URL http://arxiv.org/abs/2105.06331.

Liu Yang, Xuhui Meng, and George Em Karniadakis. B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. *Journal of Computational Physics*, 425:109913, jan 2021. ISSN 00219991. doi: 10.1016/j.jcp.2020.109913. URL https://linkinghub.elsevier.com/retrieve/pii/S0021999120306872.

Sheng Zhang and Guang Lin. Robust data-driven discovery of governing physical laws with error bars. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2217):20180305, sep 2018. ISSN 1364-5021. doi: 10.1098/rspa.2018.0305. URL https://royalsocietypublishing.org/doi/10.1098/rspa.2018.0305.

Sheng Zhang and Guang Lin. SubTSBR to tackle high noise and outliers for data-driven discovery of differential equations. *Journal of Computational Physics*, 428:1–32, 2021. ISSN 10902716. doi: 10.1016/j.jcp.2020.109962.

Peng Zheng, Travis Askham, Steven L. Brunton, J. Nathan Kutz, and Aleksandr Y. Aravkin. A Unified Framework for Sparse Relaxed Regularized Regression: SR3. *IEEE Access*, 7:1404–1423, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2886528. URL https://ieeexplore.ieee.org/document/8573778/.

Junxian Zhu, Canhong Wen, Jin Zhu, Heping Zhang, and Xueqin Wang. A polynomial algorithm for best-subset selection problem. *Proceedings of the National Academy of Sciences*, 117:33117–33123, 12 2020. ISSN 0027-8424. doi: 10.1073/pnas.2014241117. URL https://pnas.org/doi/full/10.1073/pnas.2014241117.

A Linear regression

In this section, we show how linear regression can be applied to equation learning, which sets the basis of this work. Details to regression can be found in Montgomery et al. (2012).

Starting point are N observations (y_i, x_{ij}), where the index i denotes data points, and the index j denotes features. We assume that the response (dependent) variable y is given as a function of explanatory (independent) variable $\boldsymbol{x}$,

$$y = f(\boldsymbol{x}) + \sigma z, \tag{11}$$

where $f(\boldsymbol{x})$ defines the model, z is a standard normal random variable, and σ^2 is the variance of the noise term. The explanatory observations are often organized in terms of a design matrix $\boldsymbol{X}$, with features in columns and datapoints in rows, $X_{ij} = x_{ij}$.

We assume that $f(\boldsymbol{x})$ can be given in terms of p basis functions $k_n(\boldsymbol{x})$,

$$f(\boldsymbol{x}) = \sum_{n=1}^{p} w_n k_n(\boldsymbol{x}), \tag{12}$$

with weights w_n, known as basis function expansion. Similar to the design matrix, we can define a basis function design matrix $\boldsymbol{K}$ with elements $\boldsymbol{K}_{in} = k_n(\boldsymbol{x}_i)$.

The ordinary least squares (OLS) estimates for w_n are known to be

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w} \in \mathbb{R}^p}{\arg\min} \ \left\|\boldsymbol{K}\boldsymbol{w} - \boldsymbol{y}\right\|_2^2 = \left(\boldsymbol{K}^{\mathrm{T}}\,\boldsymbol{K}\right)^{-1}\boldsymbol{K}^{\mathrm{T}}\,\boldsymbol{y} \tag{13}$$

with the q-norm

$$\|\boldsymbol{w}\|_q = \Big[\sum_n |w_n|^q\Big]^{1/q}. \tag{14}$$

Predictions based on the OLS estimates are then given by $\hat{\boldsymbol{y}} = \boldsymbol{K}\hat{\boldsymbol{w}}$.

$$\hat{\boldsymbol{y}} = \boldsymbol{K}\hat{\boldsymbol{w}}. \tag{15}$$

From the normality of linear regression, it is known that the estimates $\hat{\boldsymbol{w}}$ follow a normal distribution with the mean given by the true values for $\boldsymbol{w}$ and the variance given by $\hat{\sigma}^2 = \frac{(\boldsymbol{y}-\hat{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y}-\hat{\boldsymbol{y}})}{N-p}$. We will use these properties later to empirically define the prior in the Bayesian description.

In view of equation 13, once a choice of $k_n(\boldsymbol{x})$ is made, the actual learning of the model is straight forward. The difficult part is the choice of $k_n(\boldsymbol{x})$: on the one hand we require sufficient expressivity of the model to minimize bias, on the other hand we want to avoid overfitting to minimize variance of predictions. This bias-variance trade-off essentially dictates the number of $k_n(\boldsymbol{x})$, i.e. the effective dimension of feature space or complexity. Apart from the appropriate model size, we also seek the "correct" $k_n(\boldsymbol{x})$, in the sense that the true $f(\boldsymbol{x})$ is recovered from data generated by equation 11.

A common approach to avoid overfitting is regularization,

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w} \in \mathbb{R}^M}{\arg\min} \Big[\left\|\boldsymbol{K}\boldsymbol{w} - \boldsymbol{y}\right\|_2^2 + \lambda\|\boldsymbol{w}\|_q\Big], \tag{16}$$

where $\|\boldsymbol{w}\|_q = [\sum_n |w_n|^q]^{1/q}$, and λ is the Lagrange parameter that sets the strength of the ℓ_q penalty. Common choices for q are $q = 2$ (Ridge regression), $q = 1$ (the standard, sparsity promoting choice known as the LASSO), or combinations such as elastic net. A special case of regularization is $q = 0$, for which $\|\boldsymbol{w}\|_0 = \sum_n \delta_{w_n,0}$ is the number of non-zero weight estimates – the regression procedure with this penalty is often called *best subset selection* and requires specialized optimization algorithms.

A standard measure for goodness of fit is the coefficient of determination, R^2, which relates the variance explained by the prediction $\hat{\boldsymbol{y}}$ to the variance of the response variable y,

$$R^2 = 1 - \frac{(\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}})}{(\boldsymbol{y} - \bar{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y} - \bar{\boldsymbol{y}})} \tag{17}$$

Assuming standardized y, R^2 can be simplified to

$$\begin{aligned} R^2 &= 1 - \frac{y^{\mathrm{T}}y - y^{\mathrm{T}}\boldsymbol{K}\hat{w} - \hat{w}^{\mathrm{T}}\boldsymbol{K}^{\mathrm{T}}y + \hat{w}^{\mathrm{T}}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K}\hat{w}}{(y - \bar{y})^{\mathrm{T}}(y - \bar{y})} && (18)\\ &= 1 - \frac{y^{\mathrm{T}}y - y^{\mathrm{T}}\boldsymbol{K}\hat{w}}{(y - \bar{y})^{\mathrm{T}}(y - \bar{y})} && (19)\\ &= 1 - \frac{y^{\mathrm{T}}y - \hat{w}^{\mathrm{T}}\boldsymbol{K}^{\mathrm{T}}y}{(y - \bar{y})^{\mathrm{T}}(y - \bar{y})} && (20)\\ &= 1 - \frac{y^{\mathrm{T}}y - \hat{w}^{\mathrm{T}}\boldsymbol{K}^{\mathrm{T}}y}{y^{\mathrm{T}}y} && (21)\\ &= \frac{y^{\mathrm{T}}\,\boldsymbol{K}\,(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K})^{-1}\,\boldsymbol{K}^{\mathrm{T}}\,y}{y^{\mathrm{T}}y}, && (22) \end{aligned}$$

where we plugged in equation 15 in the first line, in the 2nd line we used that $\boldsymbol{K}^{\mathrm{T}}y = \boldsymbol{K}^{\mathrm{T}}\boldsymbol{K}\hat{w}$, in the 3rd line that $y^{\mathrm{T}}\boldsymbol{K}\hat{w} = \sum_{jk} y_j \boldsymbol{K}_{jk}\hat{w}_k = \sum_{jk} \hat{w}_k \boldsymbol{K}_{jk} y_j = \hat{w}^{\mathrm{T}}\boldsymbol{K}^{\mathrm{T}}y$, in the 4th line we assumed that the mean $\bar{y} = 0$ due to centering to zero, and in the last line we used $\hat{w} = (\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K})^{-1}\boldsymbol{K}^{\mathrm{T}}y$.$R^2$ is essentially is a quadratic form for y where the coefficient matrix is $\boldsymbol{K}$ multiplied with its pseudo-inverse. For sparse weight estimates $\hat{\boldsymbol{w}}$, R^2 is extremely efficient to compute.

A value of R^2 close to 1 signifies good predictions $\hat{\boldsymbol{y}}$. However, it is well known that R^2 can always be brought closer to 1 by increasing the number of features in the model, thus essentially lacking any overfitting penalty if used naively.

For the purpose of adequate model selection, it is helpful to formulate regression in a Bayesian setting. The main step to this end is defining the likelihood distribution for y. In the simple case of equation 11, y is normally distributed,

$$p_{\mathrm{li}}(y|\boldsymbol{w}, \sigma) = \mathcal{N}(\boldsymbol{\mu}, \sigma), \tag{23}$$

where the mean vector is given by $\mu_i = f(\boldsymbol{x}_i)$. Maximization of the log-likelihood reproduces the OLS result equation 13. Encoding existing information on the weights w_n as a prior distribution $p_{\mathrm{pr}}(\boldsymbol{w}, \sigma)$, Bayes' formula implies for the posterior distribution

$$p_{\mathrm{po}}(\boldsymbol{w}, \sigma|y) = \frac{p_{\mathrm{li}}(y|\boldsymbol{w}, \sigma)\, p_{\mathrm{pr}}(\boldsymbol{w}, \sigma)}{p(y)}, \tag{24}$$

where the normalization factor $p(y)$ is known as the *evidence* or marginal likelihood.

The distributions in equation 24 also depend on the choice of model $\mathcal{M}$ given by the representation of $f(\boldsymbol{x}_i)$ in terms of $\boldsymbol{K}$. Adding $\mathcal{M}$ as a condition, and rewriting Bayes' formula on the level of models,

$$p_{\mathrm{po}}(\mathcal{M}|y) = \frac{p(y|\mathcal{M})\, p_{\mathrm{pr}}(\mathcal{M})}{p(y)}, \tag{25}$$

we see that the evidence $p(y)$ is in fact the model-likelihood $p(y|\mathcal{M})$ and as such proportional to the model-posterior (assuming constant model-prior $p_{\mathrm{pr}}(\mathcal{M})$ for simplicity). Therefore, if we could maximize $p(y|\mathcal{M})$ over $\mathcal{M}$, we would in fact identify the model $\hat{\mathcal{M}}$ that most likely explains the observations y.

From equation 24 it follows that the evidence is given by

$$p(y|\mathcal{M}) = \int \mathrm{d}\sigma \int \mathrm{d}\boldsymbol{w}\; p_{\mathrm{li}}(y|\boldsymbol{w}, \sigma, \mathcal{M})\, p_{\mathrm{pr}}(\boldsymbol{w}, \sigma), \tag{26}$$

which in general is not solvable analytically, and also poses a particularly tough numerical challenge Von Der Linden et al. (1999); Knuth et al. (2015). Fortunately, by choosing $p_{\mathrm{pr}}(\boldsymbol{w}, \sigma)$ conjugate to $p_{\mathrm{li}}(y|\boldsymbol{w}, \sigma, \mathcal{M})$, the integral becomes solvable analytically. The conjugate prior, however, is not necessarily the sensible choice from the inference point of view. In fact, making a good choice for the prior is a much debated problem Fortuin (2021). Here, we demonstrate that for the purpose of linear equation learning, the conjugate prior is a suitable choice, if hyper-parameters are distilled from data. The question whether other choices for the prior would perform significantly better is left for future research.

The conjugate prior for the likelihood equation 23 is the gamma-normal distribution O'Hagan & Kendall (1994)

$$p(\boldsymbol{w}, \tau|\boldsymbol{\mu}, \boldsymbol{M}, k, \vartheta) = \frac{\sqrt{\det \boldsymbol{M}}}{(2\pi)^{p/2}\Gamma(k)\vartheta^k}\, \tau^{p/2+k-1}\, e^{-\frac{\tau}{2}(\boldsymbol{w}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{w}-\boldsymbol{\mu})-\tau/\vartheta} \tag{27}$$

with mean vector $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{M}$ for the weights $\boldsymbol{w}$, and shape k and scale ϑ for the precision $\tau = 1/\sigma^2$. Plugging equation 27 and equation 23 into equation 26 and performing the integration, we obtain for the log-evidence per data-point the closed expression

$$\begin{aligned}\frac{1}{N}\ln p(y|\mathcal{M}) = \frac{1}{2N}\ln\frac{\det \boldsymbol{M}}{\det \boldsymbol{A}} - \frac{1}{2}\ln 2\pi - \Big(\frac{1}{2}+\frac{k}{N}\Big)\ln\Big(\frac{\xi}{2}+\frac{1}{\vartheta}\Big) \\ - \frac{k}{N}\ln\vartheta + \frac{1}{N}\ln\Gamma\Big(\frac{N}{2}+k\Big) - \frac{1}{N}\ln\Gamma(k)\end{aligned} \tag{28}$$

with $\boldsymbol{A} = \boldsymbol{K}^{\mathrm{T}}\boldsymbol{K} + \boldsymbol{M}$, $\boldsymbol{b} = \boldsymbol{K}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{\mu}$, and $\xi = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{b}$.

$$\boldsymbol{A} = \boldsymbol{K}^{\mathrm{T}}\boldsymbol{K} + \boldsymbol{M}, \tag{29}$$

$$\boldsymbol{b} = \boldsymbol{K}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{\mu}, \tag{30}$$

$$\xi = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{b} \tag{31}$$

We detail the calculations in appendix B.

Owing to the normality of linear regression, and from standardizing the data, it is reasonable to assume the following parameters for the prior: For the mean vector, we choose $\boldsymbol{\mu} = \hat{\boldsymbol{w}}$, and the precision matrix is $\boldsymbol{M}$ is taken to be diagonal with elements $\mathrm{diag}(\boldsymbol{M}) = \frac{1-p/N}{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}-\hat{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{y}}$, resulting in normal distributions broadened by a factor N to make the prior more uninformative. The gamma distribution entering equation 27 has the mode $(k-1)\vartheta$, which we set to 1 due to standardized y. The scale is set to $\vartheta = 1/2$ which appears to be broad enough for an uninformative prior.

For completeness, we mention a few more selection criteria used for comparison in this work. The adjusted R^2 Montgomery et al. (2012)

$$R^2_{\mathrm{adj}} = 1 - \frac{N-1}{N-p-1}\,(1-R^2) \tag{32}$$

equips the usual R^2 with an overfitting penalty. The Akaike information criterion (AIC) measures the loss of information by using the inferred model instead of the (unknown) true model Murphy (2012), and similarly but derived from the model evidence equation 26 in the big data limit, follows the Bayesian (Schwarz) information criterion,

$$\mathrm{AIC} = 2p_{\mathrm{li}}(y|\hat{\boldsymbol{w}}, \hat{\sigma}) - 2p\,, \quad \mathrm{BIC} = p_{\mathrm{li}}(y|\hat{\boldsymbol{w}}, \hat{\sigma}) - 2p\ln N \tag{33}$$

Similarly, but derived from the model evidence equation 34 in the big data limit, is the Bayesian (Schwarz) information criterion,

$$\mathrm{BIC} = p_{\mathrm{li}}(y|\hat{\boldsymbol{w}}, \hat{\sigma}) - 2p\ln N. \tag{34}$$

Apart from the equations that can directly be written in the form of equation 12, a prominent application of linear equation learning is the sparse identification of dynamical systems,

$$\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t)), \tag{35}$$

where $\dot{x}(t)$ denotes the time derivative of $x(t)$. To map this problem to equation 12, the response variable can be computed from finite differences $y_i = \frac{x_{i+1}-x_i}{\Delta t}$ for a fixed time step Δt.

A restriction for the regression models to stay linear in its parameters is that the parameters of basis functions only enter as weights w. Basis functions like

$$e^{ax},\ \ln(a+x),\ \cos ax,\ x^a,\ \frac{1}{(a+x)^m},\ \ldots \tag{36}$$

with internal parameter a are not suitable.

This restriction might seem quite limiting. On the other hand, the function $f(\boldsymbol{x}(t), \boldsymbol{w})$ defining a dynamical system typically is linear in its parameters $\boldsymbol{w}$. The reason for that is that functions shown in equation 36 often reproduce when differentiated, which can be used to eliminate these functions, retrieving the standard form shown in equation 12 and equation 11. being linear in parameters. Some special functions like Bessel, Hankel, Struve and Meijer functions are even defined as solutions of differential equations.

In general, by considering the differentiated response variable,

$$y \ \mapsto\ \frac{\mathrm{d}y}{\mathrm{d}x} \simeq \frac{y(x_{i+1}) - y(x_i)}{x_{i+1} - x_i}, \tag{37}$$

if necessary to higher order, we can learn a surprisingly broad class of equations relating y and $\boldsymbol{x}$, even relations that do not exist in closed form. Here, we restrict ourselves to dynamical systems and leave the full exploration of learning in this broad model class for future work.

While many equations with non-linear parameters can be rewritten in linear form by differentiation as explained above, some functions like x^a, $\frac{1}{(a+x)^m}$ can only be reduced to fractions or require many differentiations which can give rise to numerical issues. Therefore, if we were able to include fractions in the basis function expansion, we could expand the model class even further. The problem is that basis functions like $\frac{1}{1+x^n}$ are divergent for certain values of x, and are also quite limiting in their form.

It is, however, possible to modify the regression model to also incorporate fractions. In its simplest form, we may consider

$$y = \frac{\sum_{n=1}^{p} w_n\, k_n(\boldsymbol{x}) + \sigma z}{\sum_{m=1}^{p} v_m\, k_m(\boldsymbol{x})} \tag{38}$$

with different (sparse) weights v_m but same set of basis functions for the denominator. For the next step, we assume that $k_n(\boldsymbol{x})$ is part of the numerator, that is $w_1 \neq 0$, and we can rewrite

$$k_1(\boldsymbol{x}) = \sum_{m=1}^{p} \frac{v_m}{w_1}\, y k_m(\boldsymbol{x}) - \sum_{n=2}^{p} \frac{w_n}{w_1}\, k_n(\boldsymbol{x}). \tag{39}$$

In this form, k_1 takes the role of the response variable, and we have a second set of basis functions given by yk_m. For a given model of this form, the weights also follow deterministically from equation 13, only w_1 needs to be determined from a 1-dimensional numerical root-finding algorithm.

A similar idea has been proposed in Kaheman et al. (2020), where a ℓ_0 regularized objective function needs to be minimized for each possible basis function taking the role of k_1 above. Our less greedy strategy naturally includes this procedure as a straight forward possibility, which is planned to be investigated in future work.

B Exact Bayesian model evidence

The model is given by

$$y_i = \sum_n w_n K_{in} + z_i/\sqrt{\tau} \tag{40}$$

where $K_{in} = K_n(x_i)$ is the basis function design matrix, w_n are the weights, $\tau = 1/\sigma^2$ is the precision, and $z \sim \mathcal{N}(0, 1)$. For the whole vector $\boldsymbol{y}$ of N responses, we can use the multivariate normal for the likelihood,

$$p(\boldsymbol{y} \,|\, K, \boldsymbol{w}, \tau) = \frac{\tau^{N/2}}{(2\pi)^{N/2}} \exp\Big(-\frac{\tau}{2} (\boldsymbol{y} - K\boldsymbol{w})^{\mathrm{T}} (\boldsymbol{y} - K\boldsymbol{w}) \Big), \tag{41}$$

where the precision matrix is diagonal with identical τ on the diagonal. Since the weight parameters w_n enter quadratically, we can rewrite this expression in normal form for $\boldsymbol{w}$,

$$\begin{aligned} p(\boldsymbol{y}|K, \boldsymbol{w}, \tau) = \frac{\tau^{N/2}}{(2\pi)^{N/2}} \exp\Big(-\frac{\tau}{2} S \Big) \\ \times \exp\Big(-\frac{\tau}{2} (\boldsymbol{w} - \hat{\boldsymbol{w}})^{\mathrm{T}} K^{\mathrm{T}} K (\boldsymbol{w} - \hat{\boldsymbol{w}}) \Big) \end{aligned} \tag{42}$$

with the residual sum of squares

$$\begin{aligned} S &= (\boldsymbol{y} - K\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - K\hat{\boldsymbol{w}}) && (43) \\ &= \boldsymbol{y}^{\mathrm{T}} \boldsymbol{y} - \hat{\boldsymbol{w}}^{\mathrm{T}} K^{\mathrm{T}} \boldsymbol{y} && (44) \\ &= \boldsymbol{y}^{\mathrm{T}} \boldsymbol{y} - y^{\mathrm{T}} K (K^{\mathrm{T}} K)^{-1} K^{\mathrm{T}} \boldsymbol{y} && (45) \end{aligned}$$

The mixed terms cancel after plugging in $K^{\mathrm{T}} \boldsymbol{y} = K^{\mathrm{T}} K \hat{\boldsymbol{w}}$ from the known OLS solution $\hat{\boldsymbol{w}} = (K^{\mathrm{T}} K)^{-1} K^{\mathrm{T}} \boldsymbol{y}$.

The above is of the form of a gamma distribution for τ multiplied with a normal distribution for $\boldsymbol{w}$ conditioned on τ. If we use a prior of the same form, we keep the form for the posterior, and thus have found the conjugate prior.

As a prior for the weights $\boldsymbol{w}$, we choose

$$p(\boldsymbol{w} \,|\, \boldsymbol{\mu}, M) = \frac{\tau^{p/2} \sqrt{\det M}}{(2\pi)^{p/2}} \exp\Big(-\frac{\tau}{2} (\boldsymbol{w} - \boldsymbol{\mu})^{\mathrm{T}} M (\boldsymbol{w} - \boldsymbol{\mu}) \Big), \tag{46}$$

where τM is the precision matrix with τ split off, and $\boldsymbol{\mu}$ is the mean vector of the multivariate normal prior. Splitting off τ technical means that specifying M is relative to the unknown τ, but τ does not need to be known for that, as we integrate over all possible τ values.

For the posterior, we are interested in the quadratic form involving $\boldsymbol{w}$,

$$\begin{aligned} &(\boldsymbol{w} - \hat{\boldsymbol{w}})^{\mathrm{T}} K^{\mathrm{T}} K (\boldsymbol{w} - \hat{\boldsymbol{w}}) + (\boldsymbol{w} - \boldsymbol{\mu})^{\mathrm{T}} M (\boldsymbol{w} - \boldsymbol{\mu}) && (47) \\ &\quad = \boldsymbol{w}^{\mathrm{T}} A \, \boldsymbol{w} - 2 \, \boldsymbol{w}^{\mathrm{T}} \boldsymbol{b} + c && (48) \end{aligned}$$

with

$$\begin{aligned} A &= K^{\mathrm{T}} K + M && (49) \\ \boldsymbol{b} &= K^{\mathrm{T}} K \hat{\boldsymbol{w}} + M\boldsymbol{\mu} \\ &= K^{\mathrm{T}} \boldsymbol{y} + M\boldsymbol{\mu} && (50) \\ c &= \hat{\boldsymbol{w}}^{\mathrm{T}} K^{\mathrm{T}} K \hat{\boldsymbol{w}} + \boldsymbol{\mu}^{\mathrm{T}} M \boldsymbol{\mu} \\ &= \hat{\boldsymbol{w}}^{\mathrm{T}} K^{\mathrm{T}} \boldsymbol{y} + \boldsymbol{\mu}^{\mathrm{T}} M \boldsymbol{\mu} && (51) \\ &= y^{\mathrm{T}} K (K^{\mathrm{T}} K)^{-1} K^{\mathrm{T}} \boldsymbol{y} + \boldsymbol{\mu}^{\mathrm{T}} M \boldsymbol{\mu}. && (52) \end{aligned}$$

Put into this form, we can use the Gaussian integral $\int \mathrm{d}^n x\, e^{-\frac{1}{2}x^{\mathrm{T}}Ax+b^{\mathrm{T}}x} = \sqrt{\frac{(2\pi)^n}{\det A}}\, e^{\frac{1}{2}b^{\mathrm{T}}A^{-1}b}$ to marginalize,

$$
\begin{aligned}
p(\boldsymbol{y}\,|\,\tau) &= p(\boldsymbol{y}\,|\,K,\tau,\boldsymbol{\mu},M) && (53)\\
&= \int \mathrm{d}\boldsymbol{w}\; p(\boldsymbol{y}\,|\,K,\boldsymbol{w},\tau)\, p(\boldsymbol{w}\,|\,\boldsymbol{\mu},M) && (54)\\
&= \frac{\sqrt{\tau^{N+p}\det M}}{\sqrt{(2\pi)^{N+p}}}\, e^{-\frac{\tau}{2}(S+c)} \int \mathrm{d}\boldsymbol{w}\, e^{-\frac{\tau}{2}(\boldsymbol{w}^{\mathrm{T}}A\boldsymbol{w}-2\boldsymbol{b}^{\mathrm{T}}\boldsymbol{w})} && (55)\\
&= \frac{\sqrt{\tau^{N+p}\det M}}{\sqrt{(2\pi)^{N+p}}}\, e^{-\frac{\tau}{2}(S+c)}\, \frac{\sqrt{(2\pi)^p}}{\sqrt{\tau^p \det A}}\, e^{\frac{\tau}{2}\boldsymbol{b}^{\mathrm{T}}A^{-1}\boldsymbol{b}} && (56)\\
&= \sqrt{\frac{\tau^N \det M}{(2\pi)^N \det A}}\, e^{-\frac{\tau}{2}(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}+\boldsymbol{\mu}^{\mathrm{T}}M\boldsymbol{\mu}-\boldsymbol{b}^{\mathrm{T}}A^{-1}\boldsymbol{b})}. && (57)
\end{aligned}
$$

For the τ-integration, we choose the gamma distribution

$$
p(\tau|k,\vartheta) = \frac{1}{\Gamma(k)\vartheta^k}\, \tau^{k-1}\, \exp\big(-\tau/\vartheta\big) \tag{58}
$$

as the (conjugate) prior for τ, and define

$$
\xi = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{\mu}^{\mathrm{T}}M\boldsymbol{\mu} - \boldsymbol{b}^{\mathrm{T}}A^{-1}\boldsymbol{b}. \tag{59}
$$

The remaining τ-integral follows then from $\int_0^\infty \mathrm{d}\tau\, \tau^{c_0} e^{-c_1\tau} = c_1^{-c_0-1}\,\Gamma(c_0+1)$ as

$$
\begin{aligned}
p(y) &= p(y\,|\,K,\boldsymbol{\mu},M,k,\vartheta) && (60)\\
&= \int_0^\infty \mathrm{d}\tau\, p(\tau|k,\vartheta)\, p(\boldsymbol{y}\,|\,\tau) && (61)\\
&= \frac{1}{\Gamma(k)\vartheta^k(2\pi)^{\frac{N}{2}}} \sqrt{\frac{\det M}{\det A}} \int_0^\infty \mathrm{d}\tau\, \tau^{\frac{N}{2}+k-1}\, e^{-\tau(\frac{z}{2}+\frac{1}{\vartheta})} && (62)\\
&= \frac{\Gamma(\frac{N}{2}+k)}{\Gamma(k)\vartheta^k(2\pi)^{\frac{N}{2}}} \sqrt{\frac{\det M}{\det A}} \Big(\frac{z}{2}+\frac{1}{\vartheta}\Big)^{-\frac{N}{2}-k} && (63)
\end{aligned}
$$

and for the log-evidence per data-point we obtain

$$
\begin{aligned}
\frac{1}{N}\ln p(\boldsymbol{y}) = {} & \frac{1}{2N}\ln\frac{\det M}{\det A} - \frac{1}{2}\ln 2\pi \\
& - \Big(\frac{1}{2}+\frac{k}{N}\Big)\ln\Big(\frac{\xi}{2}+\frac{1}{\vartheta}\Big) - \frac{k}{N}\ln\vartheta \\
& + \frac{1}{N}\ln\Gamma\Big(\frac{N}{2}+k\Big) - \frac{1}{N}\ln\Gamma(k). && (64)
\end{aligned}
$$

C Details and illustration of less greedy stepwise regression

To our knowledge, all equation learning approaches apart from stepwise regression include an optimization step in various representation spaces of equations. Here, we propose a strategy that does without any numerical optimization algorithms, and instead considers candidate models individually in an almost comprehensive manner. Since already small dictionaries of basis functions can lead to tremendous numbers of candidate models, a combination of cheap selection criteria and successive reduction of model space with a suitable stopping criterion is required. We demonstrate how the simple criterion R^2 can be used for such a semi-comprehensive search.

In a first step, a dictionary of basis functions is generated using Algorithm 1. These basis functions consist of all possible products of available features x_j. In these products, the factors are raised to all possible

Algorithm 1 Model ranking R^2

1: Function DesignMatrix($\boldsymbol{X}$)
2: **Input:** data $\boldsymbol{X}$ with N datapoints, l features
3: **Parameters:** maximum degree M_1 for individual features, maximum degree M_2 for term
4: build all powers $x_{ij}^{m_j}$, $m_j = (1, ..., M_1)$ ▷ pre-computed for speed, limited to power M_1
5: initialize counter $p = 0$
6: **for** all unique l-tuples $(m_1, ..., m_l)$ **do**
7: **if** $\sum_j m_j \leq M_2$ **then** ▷ ensure that collective power is limited to M_2
8: $p := p + 1$
9: $\boldsymbol{K}_{:,p} := \prod_j \boldsymbol{X}_{:,j}^{m_j}$ ▷ Design matrix, candidate models in columns
10: **end if**
11: **end for**
12: **Return:** Design matrix $\boldsymbol{K}$, shape $N \times p$

combinations of powers (line 9), where we restrict individual powers to M_1 (line 4) and the combined power of a term to M_2 (line 7). For example, for $M_1 = 3$ and $M_2 = 5$, the term $x_1^4 x_3$ would not be allowed since $4 > M_1$, and the term $x_1^2 x_2 x_3^3$ would not be allowed because $2+1+3 > M_2$.

The less greedy (LG) strategy we propose is based on this basis function expansion and described by Algorithm 3. It begins by considering all regression models with r non-zero weights w_j (line 9), c.f. equation equation 12. To this end, the auxiliary Algorithm 2 produces a list of models with top R^2 values by looping through all candidate models of size r. These models are returned as an index matrix $\boldsymbol{M}$, indicating selected terms with a 1 and deselected terms with a 0, where each column stands for a candidate model (lines 5,11). The models are sorted in descending order with respect to R^2 (line 13).

Back to Algorithm 3, we successively increase r starting from $r = 1$ (lines 7,22). We found that for a fixed r value, R^2 performs particularly well in identifying the best model out of the millions of models (see for instance the left plot in figure 5). To infer a value for r with just R^2, we create a feature rating matrix $\boldsymbol{F}$ defined as the weighted counts of terms being selected across s top models, where the weight is given by R^2. Based on $\boldsymbol{F}$, we check for terms selected by R^2 for two successive model sizes r and $r-1$ (lines 16-20). If for both r and $r-1$ the same terms are selected consistently, we choose these two terms as part of the inferred model and conclude the search.

As for larger values of r the number of candidate models can easily reach hundreds of millions, we implement another strategy to divide out the list of candidate models. If terms have not been selected for two successive model sizes r and $r-1$, we remove these terms from the design matrix $\boldsymbol{K}$ (lines 14-15). In this way, we continuously reduce the model equation space as we go along.

The rationale for this selection and elimination strategy being solely based on R^2 is the following: If the true model has r terms, and we are testing all models with $r-2$ terms, then the models with largest R^2 will consistently be composed of the $r-2$ terms that contribute the most to explaining the variance of y. The other 2 true terms will be selected sporadically but at least once, terms that have not been selected at all can hence be removed from the candidate models. Testing in the next stage all models with $r-1$ terms, one more term will be consistently selected. The same holds for testing models with r terms, but when testing models with $r+1$ terms, no new term can contribute consistently to explaining more of the variance of y. The R^2 measure will increase for models for $r+1$ terms, but compared to models with r terms, no extra term will consistently be selected. Therefore, once no new term is selected consistently when incrementing the number r of terms, we may conclude that all contributing terms have been found. An illustration of this strategy can be found in figure 5, where a typical case with three two terms in the true equation is shown.

Since the cheap computation of R^2 allows to go through millions of models in a matter of minutes on a standard computer, together with the described less greedy strategy, we are able to consider or exclude all candidate models that can be built from the basis function dictionary.

Algorithm 2 Model ranking R^2

Function TopRsq($\boldsymbol{y}, \boldsymbol{K}$)
Input: response data $\boldsymbol{y}$, design matrix $\boldsymbol{K}$
Parameters: number r of terms and t of top models
initialize criterion $\boldsymbol{c}$ ▷ flexible length, to store R^2 for candidate models
initialize model indices $\boldsymbol{M}$ ▷ p rows indicating terms part of models, flexible number of columns
initialize model number $i = 0$
for all $\boldsymbol{n} = (n_1, ..., n_p)$ with $n_j \in \{0, 1\}$, $\sum_j n_j = r$ **do** ▷ possible selections of terms, for fixed r
    $i := i+1$, append $\boldsymbol{c}_i$ and $\boldsymbol{M}_{:,i} = 0$
    reduce $\boldsymbol{K}_{\text{red}} := \boldsymbol{K}_{:,n}$ ▷ extract design matrix for selected terms
    determine R^2 for $\boldsymbol{K}_{\text{red}}$ and $\boldsymbol{y}$ from equation 22
    store $c_i := R^2$ and $\boldsymbol{M}_{:,i} := \boldsymbol{n}$ ▷ each column of M indicates a model with R^2 value c_i
end for
sort columns of $\boldsymbol{M}$ and $\boldsymbol{c}$ according to $\boldsymbol{c}$ (descending) ▷ first columns of M now indicate top models in terms of R^2
Return: top models $M_{:,:t}$ (shape $p \times t$), criterion $c_{:t}$ (length t) ▷ only return the t best models

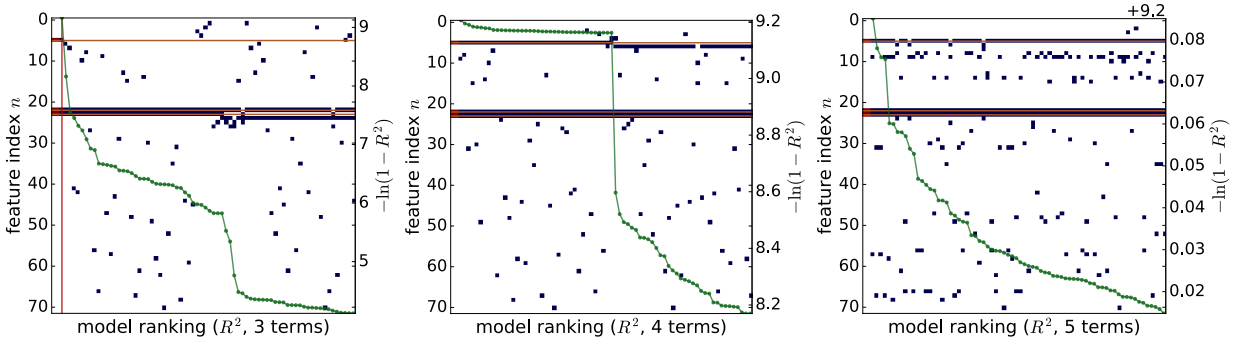

Figure 5: An example of how candidate models with 3, 4, and 5 terms with largest R^2 in each category tend to consistently choose the true terms of the model. The indices of all terms in the dictionary (the basis functions $k_n(\boldsymbol{x})$) are shown on the left vertical axis. Each candidate model along the horizontal axis is represented by squares indicating which terms make up the respective model. The ground truth model in this example has 3 terms, indicated by two lighter squares on the left and the horizontal lines. The case where the candidate model is the true model is indicated by a vertical line in the middle plot. The R^2 value in a logarithmic scale is shown as a line with closed circles, the values are given by the right vertical axis. It can be seen that the true model is chosen by R^2 among all models with 3 terms.

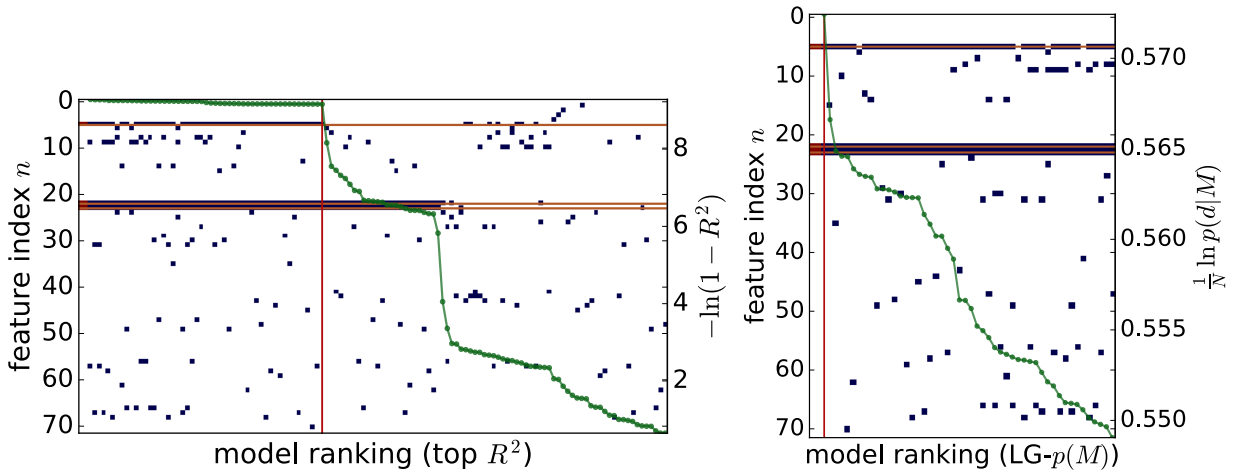

Figure 6: In the plot on the left, the t best models in terms of R^2 from each of the r-sized candidate models in figure 5 have been combined and sorting according to their R^2 value. Clearly, the models with more terms are favored over models with fewer terms, and the true model is not selected. The plot on the rights shows the best models now in terms of the Bayesian model evidence $p(y|\mathcal{M})$, where now the true model is selected illustrating the adequate overfitting penalty held by $p(y|\mathcal{M})$.

Algorithm 3 Semi-comprehensive search with R^2

Function SC-SEARCH($\boldsymbol{y}, \boldsymbol{X}$)
Input: data $\boldsymbol{y}, \boldsymbol{X}$
Parameters: maximum number $r_{\max}$ of terms, number s of top models and threshold $c_{\min}$ for feature selection
$\boldsymbol{K}, p :=$ DESIGNMATRIX($\boldsymbol{X}$)
initialize feature rating $\boldsymbol{F}$ of shape $p \times r_{\max}$ ▷ rate importance of terms for different model sizes r
initialize $search := True$
initialize number of terms $r := 1$
while $search$ and $r \leq r_{\max}$ **do**
  $\boldsymbol{M}$, $\boldsymbol{c}$:= TopRsq($\boldsymbol{y}, \boldsymbol{K}, r$) ▷ obtain list of models and with their R^2 values for fixed model size r
  append $\boldsymbol{M}$ to $\boldsymbol{M}_{\text{all}}$ ▷ append models to index matrix keeping models for all r
  $\boldsymbol{F}_{:,r} := \sum_j^s c_j \, \boldsymbol{M}_{:,j}$ ▷ counts how often terms are selected in s top models, weighted by R^2 criterion
  normalize $\boldsymbol{F}_{:,r} := \boldsymbol{F}_{:,r} \,/\, \max(\boldsymbol{F}_{:,r})$ ▷ to have ratings between 0 and 1
  if $r \geq 2$ **then**
    index $\boldsymbol{i}_0 := (\boldsymbol{F}_{:,r} + \boldsymbol{F}_{:,r-1} = 0)$ ▷ i_0=True if terms not selected for two successive model sizes
    remove $\boldsymbol{K}[:, \boldsymbol{i}_0]$ from $\boldsymbol{K}$ ▷ reduce model equation space by those terms
    index $\boldsymbol{i}_1 := (\boldsymbol{F}_{:,r} \geq c_{\min})$ ▷ indexes terms with significant rating across best s models of size r
    index $\boldsymbol{i}_2 := (\boldsymbol{F}_{:,r-1} \geq c_{\min})$ ▷ same for previously considered model size $r-1$
    if $\boldsymbol{i}_1 = \boldsymbol{i}_2$ **then**
      $search := False$ ▷ if term is selected twice in a row like this, conclude search
    end if
  end if
  $r := r + 1$
end while
Compute criteria $\{\, p(y|\mathcal{M}), \text{AIC}, \text{BIC}, R^2_{\text{adj}} \,\}$ for $\boldsymbol{M}_{\text{all}}$
Return: $\boldsymbol{M}_{\text{all}}$ along with criteria

In a final step, the list of top models from the R^2 evaluation can be combined and each tested with other selection criteria like $p(y|\mathcal{M})$, AIC, BIC or R^2_{adj}. The best choice turned out to be the Bayesian model evidence $p(y|\mathcal{M})$ (c.f. B). In figure 6 the complexity selection power of $p(y|\mathcal{M})$ is demonstrated.

The hyperparameters of our procedure are the number s of models used to count consistent selection of basis functions $k_n(\boldsymbol{x})$ together with the threshold $c_{\min}$ the (normalized) count of a feature must exceed to be selected, the number t of the best models in terms of R^2 that are combined in a new list of top models for re-evaluation with $p(y|\mathcal{M})$ (or other criteria), and the maximum number if iterations, $r_{\max}$. We found that the universally best values are $s = p/2$, where p is the feature dimension, $c_{\min} = 0.75$, $t = 25$, and $r_{\max} = 8$