Chapter 7
Multi-Task Equation Discovery

S.C. Bee, K. Worden, N. Dervilis, and L.A. Bull

Abstract Equation discovery provides an approach to system identification by uncovering the governing dynamics of a system from observed data. By applying a Bayesian relevance vector machine (RVM) within a multi-task (MT) framework for simultaneous parameter identification across multiple datasets, this work seeks to improve the challenge of generalising models. In this work, responses of the same structure under different excitations are modelled as related tasks with shared parameters and task-specific noise. A simulated single degree-of-freedom oscillator with linear and cubic stiffness presents the case study, with datasets generated under three excitation regimes. This work shows that while single-task RVMs reproduce responses, they often miss governing terms when nonlinear dynamics were weakly excited. In contrast, the MT-RVM leverage information across tasks, improving parameter recovery under weak and moderate excitations while retaining accuracy at high excitation. These results show how multi-task Bayesian inference can reduce over-fitting and improve generalisation in equation discovery, with particular relevance to structural health monitoring under variable loading.

Keywords Sparse regression · Equation discovery · Multi-task learning · System identification

Introduction

Equation discovery has been applied to the field of system identification to identify equations of motion that capture the behaviour of measured systems, along with the parameters associated with each of their terms [1]. In practice, however, any selected model is only an approximation of reality, and uncertainty is inevitably present in the predictive equations [2]. To address this, Bayesian approaches have been widely explored for equation discovery, which quantify parameter uncertainty [2–6].

A potential challenge in equation discovery is the risk of over-fitting, where the discovered model fits the training data closely but fails to represent the underlying physical system. The antithesis of over-fitting is generalisation, which can be encouraged via several strategies. One way to aid generalisation is via *multi-task learning* (MTL), which learns multiple tasks simultaneously and exploits shared information to capture underlying commonalities. Caruana outlines several domains of MTL [7], some of which have since been translated into the context of structural health monitoring (SHM) [8].

In this paper, MTL is applied by treating datasets from the same structure under different excitation levels as separate, but related, tasks. By sharing parameters across tasks, the model exploits common features while still accounting for differences introduced by varying excitations. This design ensures that the model does not over-fit to one excitation condition (one dataset) and instead learns a representation of the system that is valid across multiple responses.

This approach is valuable in SHM, where different loading conditions often highlight different aspects of the same underlying physics. For example, low-level excitations may reveal the system's linear response, while higher excitations expose nonlinear effects. By training across multiple excitation scenarios simultaneously, the model is able to combine these complementary insights into a single, coherent description of the structure.

S.C. Bee · K. Worden · N. Dervilis
Dynamics Research Group, School of Mechanical, Aerospace and Civil Engineering,
University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK
e-mail: scbee1@sheffield.ac.uk; k.worden@sheffield.ac.uk; n.dervilis@sheffield.ac.uk

L.A. Bull
School of Mathematics and Statistics, University of Glasgow, Glasgow, G12 8QQ, UK
e-mail: Lawrence.Bull@glasgow.ac.uk

Kristin B. Zimmerman, PhD (eds.), *Data Science in Engineering, 2026, Vol. 9*, Conference Proceedings of the Society for Experimental Mechanics Series, https://doi.org/10.1007/978-8-743-81448-1_7

System Identification with Bayesian Inference

System identification characterises the dynamics of a system from observations of the evolution of states through time. The current work follows the approach proposed by [1] and treats equation discovery as sparse linear regression, known as SINDy.

For a multi-degree-of-freedom system, the system states are typically displacement and velocity,

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} y \\ \dot{y} \end{bmatrix} \tag{1}$$

where y is displacement, $\dot{y}$ is velocity and $\dot{}$ denotes the first order differentiation with respect to time.

Within the framework of equation discovery, the above formulation is combined with a predefined dictionary, or design matrix, of M candidate terms,

$$\mathbf{D} = [d_1(\boldsymbol{X}), d_2(\boldsymbol{X}), ..., d_M(\boldsymbol{X})] \tag{2}$$

where $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_N}]$ and $\mathbf{D} \in R^{N\times M}$. This corresponds to N data-points and M candidate components (features). The design matrix is intentionally over-defined, incorporating a wide range of potential components that could contribute to the equation of motion of the system.

By utilising shrinkage [9], the algorithm identifies only the base elements that are most relevant, effectively revealing the governing dynamics of the system.

RVM Formulation of SINDy

The relevance vector machine (RVM), devised by Tipping [9], is a Bayesian algorithm which has a similar functional form to the deterministic support vector machine [10]. RVMs have been used within various applications to analyse systems, and the approach is well-suited to equation discovery problems. The formulation used in this research is demonstrated graphically in Figure 1.

In the context of identifying forced (non-autonomous) duynamical systems, a function $\mathbf{F}$ is present, which represents the input force acting on the structure. Arguably, this could be embedded within the design matrix, defined in equation (2); however, in this work it is treated as a separate observed variable,

$$\boldsymbol{y} = \mathbf{D}\mathbf{w} + \mathbf{F} + \boldsymbol{\epsilon} \tag{3}$$

where $\mathbf{w} \in R^M$ is the vector representing the weights associated with each candidate basis function and $\boldsymbol{\epsilon}$ is the noise that distorts the signal.

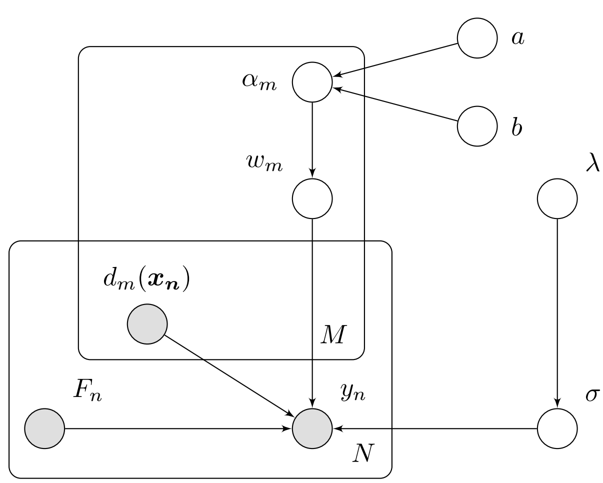

Fig. 1 A graphical representation of the RVM. Grey nodes correspond to observed variables and white nodes represent latent variables. The M plate represents the number of candidate features and the N plate represents the number of data points

Each weight is assumed to follow a zero-mean Gaussian prior, $w_m \sim \mathcal{N}(0, \alpha)$ where the variance is governed by an inverse gamma distribution, $\alpha_m^2 \sim \Gamma^{-1}(a, b)$.

This hierarchical prior structure introduces hyper-parameters a and b, which control the distribution of weight variances and encourage sparsity.

The signal noise is assumed to be Gaussian, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, with an exponential prior placed on the variance of the noise, $\sigma^2 \sim \text{Exp}(\lambda)$, where λ is a hyper-parameter. This prior has a preference for data with generally low levels of noise.

RVM Extension to Multiple Tasks

When analysing L tasks, the RVM can be adapted as shown in Figure 2.

Figure 2 shows that the weight vector, $\mathbf{w}$, is shared between *all* tasks, while the noise variance is task-specific. A shared weight distribution reflects the assumption that measurements across tasks have common underlying physical properties. In contrast, the noise is modelled separately for each task as it typically arises from environmental conditions or measurement settings that vary between tasks.

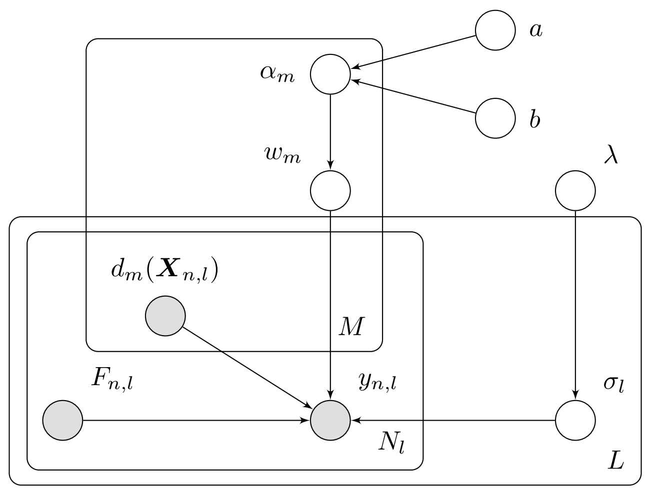

Fig. 2 A graphical representation of a multi-task RVM. Grey nodes correspond to observed variables and white nodes represent latent variables. The L plate represents the number of tasks, the M plate represents the number of candidate features and the N plate represents the number of data points

Case Study

A simulated dataset was developed, based on a single degree-of-freedom (SDOF) system with nonlinear stiffness, shown by Figure 3.

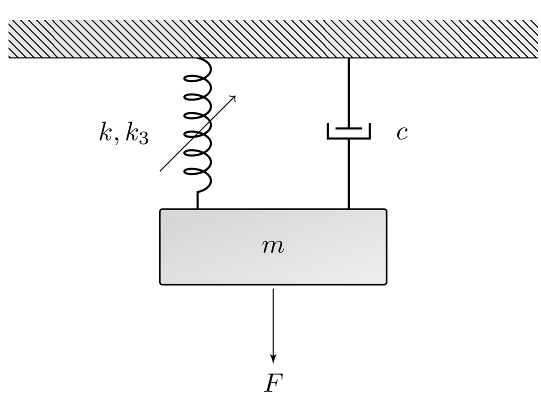

Fig. 3 SDOF oscillator with stiffness (linear, k and cubic, k_3), mass m, damping coefficient c and force acting on the system, F

The equation of motion for this system is given by,

$$m\ddot{y} + c\dot{y} + k_1 y + k_3 y^3 = F \tag{4}$$

where $m = 1kg$ is mass, $c = 0.2Ns/m$ is the damping coefficient, $k_1 = 1N/m$ is the linear stiffness, $k_3 = 1N/m^3$ is the cubic stiffness, F is the system forcing, and, $\ddot{}$ denotes the second order differentiation with respect to time. The values for the damping coefficient and stiffnesses were selected to demonstrate the algorithm and not to be representative of a real system.

The system was integrated forward in time via a fourth-order Runge-Kutta method, with a sampling frequency of 100kHz. To reduce dataset size, the signal was downsampled by selecting every thousandth point, resulting in an effective sampling frequency of 100Hz over an interval of 10s. A Butterworth filter [11] was applied with a low cut-off of 3Hz.

To generate a suitable dataset, it was necessary to construct data under varying conditions. Since the objective is to use these datasets to gain insight into the structure, the underlying physics, represented by the right-hand side of equation (4), must remain unchanged. Consequently, only the input forcing can be modified to produce different datasets. The forcing is modelled as random excitation in the form of noise, drawn from a uniform distribution whose amplitude can be scaled by a factor, f, such that $F \sim f\mathcal{U}(-0.5, 0.5)$. To create three distinct datasets, data were generated over a duration of 10s with $f = \{10^1, 10^2, 10^3\}$. Three datasets allow for a single task (ST) scenario with low, medium and high excitation, one at each of the values of f, and one multi-task (MT) scenario with all three datasets utilised in a joint inference for the identification of the shared underlying system.

For each of the three datasets, the dictionary of candidate basis functions required for the RVM was defined as,

$$D(\mathbf{X}) = \begin{bmatrix} | & | & | & | & | & | & | \\ \mathbf{y} & \dot{\mathbf{y}} & \mathbf{y}^2 & \dot{\mathbf{y}}^2 & \mathbf{y}^3 & \dot{\mathbf{y}}^3 & \mathbf{1} \\ | & | & | & | & | & | & | \end{bmatrix} \tag{5}$$

Results and Discussion

It is worth noting that the target vector, $\boldsymbol{y}$ (equation (3)), for this problem is acceleration, $\ddot{y}$. Hence, Equation (4) will take the form,

$$m\ddot{y} = F - c\dot{y} - k_1 y - k_3 y^3 \tag{6}$$

The system has unit mass, and therefore, the target value of $\boldsymbol{w}$ will be: $[-k_1, \quad -c, \quad 0, \quad 0, \quad -k_3, \quad 0, \quad 0]$.

Figure 4 shows the results for $f = 10^1$. The lower plot shows that the RVM output tracks the target acceleration over the 10-second training dataset reasonably well. However, the estimated values of the elements in $\boldsymbol{w}$, shown in the upper plot, deviate from the true values. The mean of the displacement cubic term, y^3, is 0.79, much closer to unity than the true value of minus unity and the mean of the velocity cubic term, $\dot{y}^3$, is -0.77, closer to -1 than the true value of 0, this outcome is challenging to interpret in physical terms.

Figure 5 shows the results for $f = 10^2$ and Figure 6 shows the results for $f = 10^3$. As with the lower plot in Figure 4, the corresponding plots in both Figure 5 and Figure 6 accurately tracks the target acceleration over the 10-second training dataset.

However, for $f = 10^2$ the estimated weight matrix again diverges from the *True* values. In this case, y^3, is 0.01, much closer to zero than the true value of -1, this discrepancy can be attributed to the insufficient excitation of the system's nonlinearities. The model misidentified several terms because of insufficient excitation, yielding values that deviate from the *True* dynamics and lack clear physical interpretation.

The weight matrix for $f = 10^3$ is in good agreement with the *True* values of $\boldsymbol{W}$, the value for y^3 is slightly lower than the *True* value but is much closer than when $f = 10^2$ and $f = 10^3$.

The results for the combination of all the data as three distinct tasks is shown in Figure 7. The prediction of values over the 10s interval for the three tasks remains good and there is improvement in the prediction of W for $f = 10^1$ and $f = 10^2$. The results of $f = 10^3$ and the multi-task learner are similar indicating that there has been positive knowledge transfer from $f = 10^3$ to the other two tasks, without negative transfer from the lower excitation tasks impacting the strongest performing task.

The parameter used to evaluate the predictive performance of each model is the normalised mean-square error metric,

$$NMSE = \frac{100}{N\sigma_z^2} \Sigma_{n=1}^{N} (z_n - \hat{z}_n)^2 \tag{7}$$

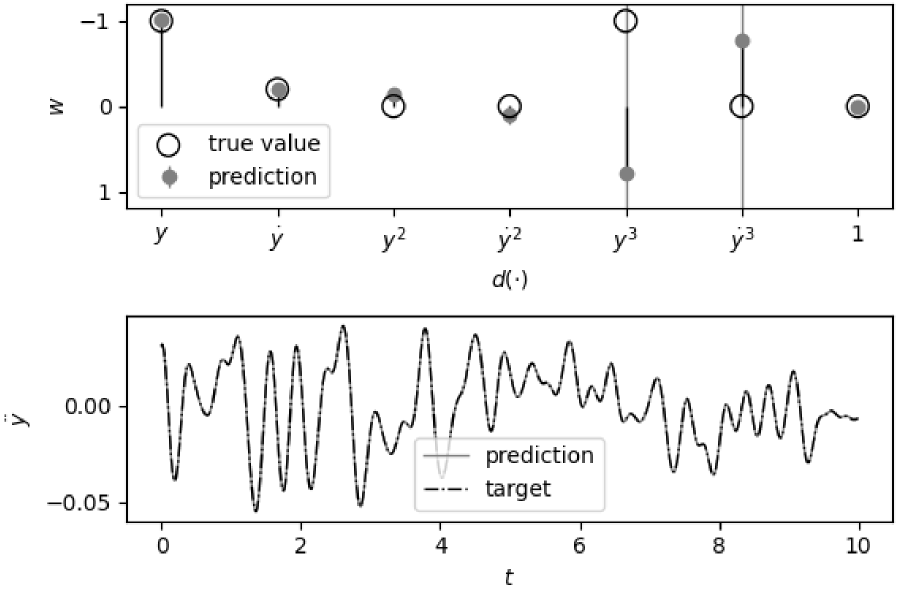

Fig. 4 Results of the ST scenario with medium excitation, $f = 10^2$ with a standard RVM.(*Upper*) Weight matrix, $\boldsymbol{W}$, results; (*Lower*) the predicted response

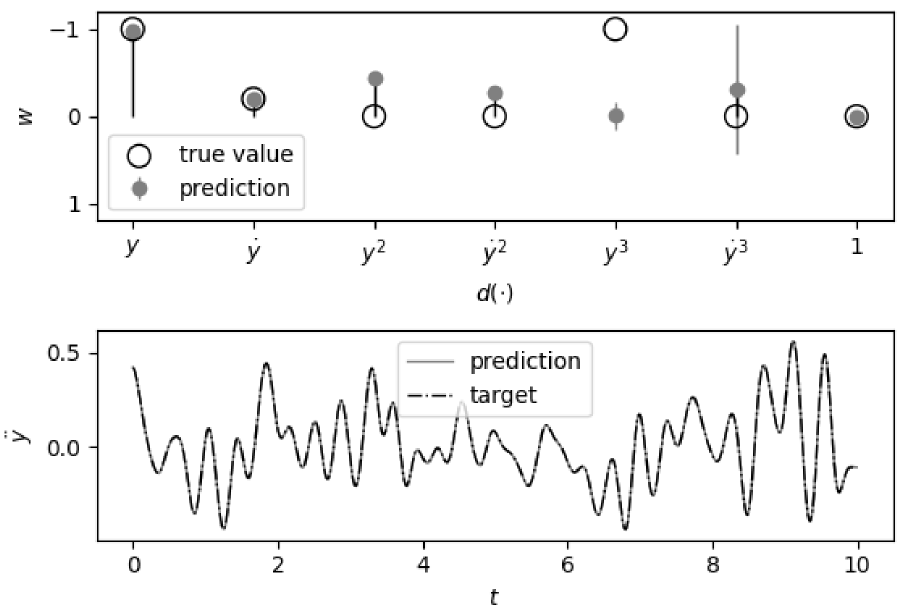

Fig. 5 Results of the ST scenario with medium excitation, $f = 10^1$ with a standard RVM as per Figure 1).(*Upper*) Weight matrix, $\boldsymbol{W}$, results; (*Lower*) the predicted response

σ_z^2 is the variance of the target, $\mathbf{z}$, $\hat{\cdot}$ represents the predicted value, and, in this study $\mathbf{z} = \ddot{\boldsymbol{y}}$.

Figure 8 shows the NMSE results from the ST and MT scenarios for the three excitation levels. One hundred datasets were generated at each excitation level and the NMSE for each dataset was calculated. The graph shows the average and distribution of the NMSE for the different excitation levels and the different models.

For the low and high excitation cases, the NMSE is similar between the ST and MT models. The benefit of the MT model is demonstrated in the medium excitation case as the average NMSE is lower, but also, the variance is lower. With medium excitation, the data does not contain information on the nonlinear components of the structure, as these nonlinear components have not been excited. Therefore, the ST model with medium excitation has been unable to identify this nonlinearity and

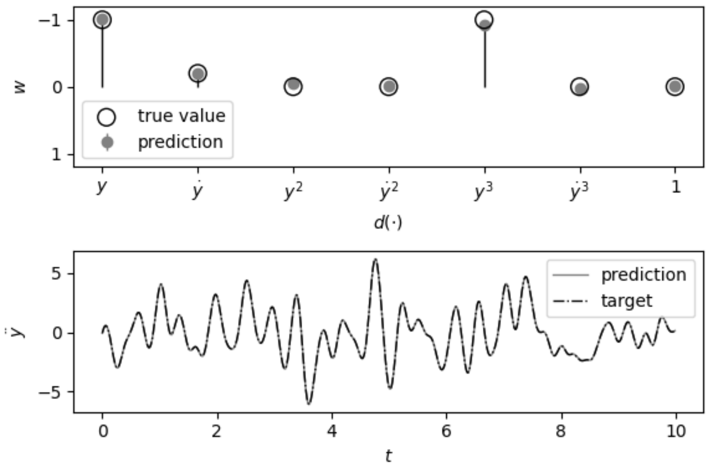

Fig. 6 Results of the ST scenario with high excitation, $f = 10^3$ with a standard RVM.(*Upper*) Weight matrix, $\boldsymbol{W}$, results; (*Lower*) the predicted response

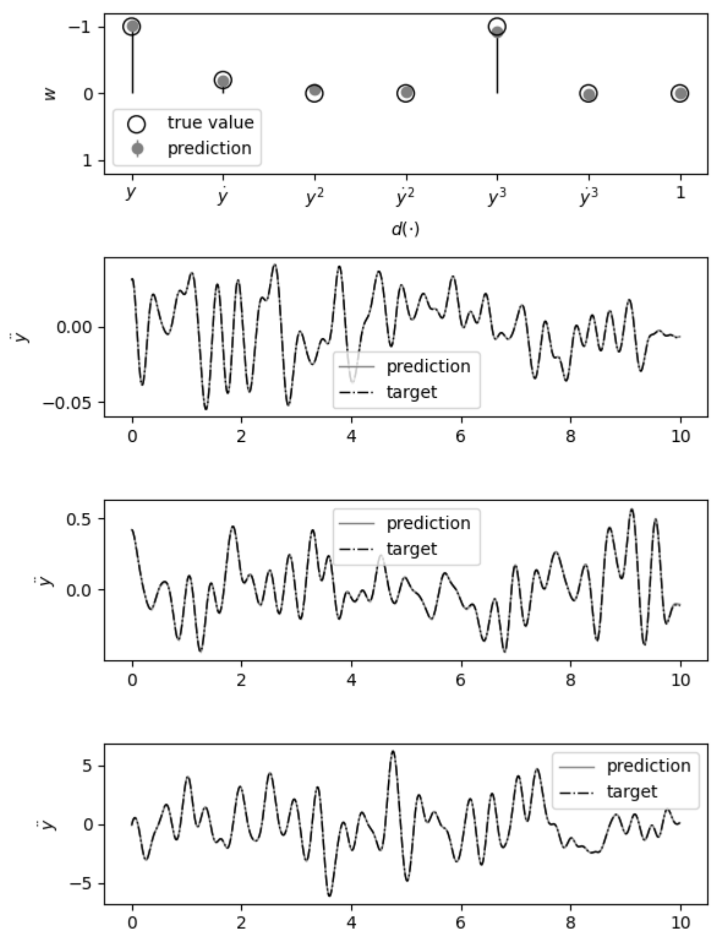

Fig. 7 Results of the MT scenario with $f = [10^1 \quad 10^2 \quad 10^3]$ with a multi-task RVM (as per Figure 2); (*Upper*) Weight matrix, $\boldsymbol{W}$, results shared across all tasks; (*Second*) the predicted response for $f = 10^1$; (*Third*) the predicted response for $f = 10^2$; and, (*Fourth*) the predicted response for $f = 10^3$

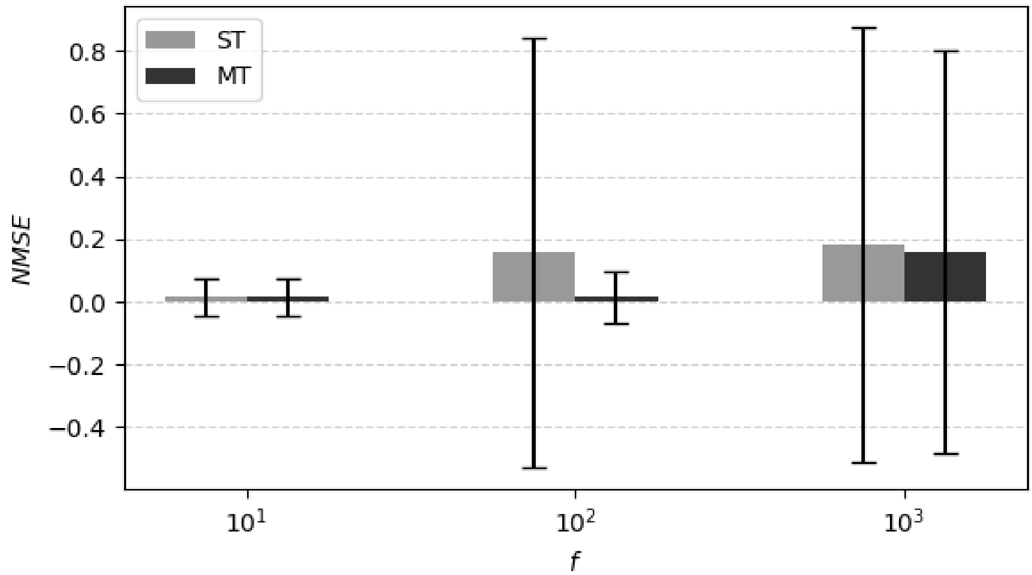

Fig. 8 The distribution of NMSE for the MT and ST scenarios showing 3σ

instead, has fit a model based on structurally incomplete data (missing nonlinear information). Hence, the ST model's weight matrix does not match well with actual parameters. In contrast, the MT model *does* have a dataset which includes both linear and nonlinear excitation, the parameters are much closer to the *True* values.

Conclusion

This provides a strong example of how multi-task learning can enhance equation discovery by promoting generalisation across tasks. When working with a single task, it is often difficult to assess whether the identified model truly generalises. For instance, the solutions obtained for low levels of forcing, $f = 10^1$ and $f = 10^2$, differ considerably between the two single-task models. By contrast, the multi-task RVM yields a solution that performs well across all three datasets. This consistency suggests that the model has captured a representation that better reflects the underlying physics of the system. When considering the NMSE of the models, the MT model outperforms single-task models with a lower standard deviation in the medium and high excitation cases.

Acknowledgments The authors wish to gratefully acknowledge support for this work through grants from the Engineering and Physical Sciences Research Council (EPSRC), UK, and Natural Environment Research Council, UK via grant number, EP/S023763/1. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

1. S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.
2. J. L. Beck, "Bayesian system identification based on probability logic," *Structural Control and Health Monitoring*, vol. 17, no. 7, pp. 825–847, 2010.
3. K. Worden and J. Hensman, "Parameter estimation and model selection for a class of hysteretic systems using Bayesian inference," *Mechanical Systems and Signal Processing*, vol. 32, pp. 153–169, 2012. Uncertainties in Structural Dynamics.
4. R. Fuentes, N. Dervilis, K. Worden, and E. Cross, "Efficient parameter identification and model selection in nonlinear dynamical systems via sparse Bayesian learning," *Journal of Physics: Conference Series*, vol. 1264, p. 012050, jul 2019.
5. R. Fuentes, R. Nayek, P. Gardner, N. Dervilis, T. Rogers, K. Worden, and E. Cross, "Equation discovery for nonlinear dynamical systems: A Bayesian viewpoint," *Mechanical Systems and Signal Processing*, vol. 154, p. 107528, 2021.
6. Y.-C. Zhu, P. Gardner, D. J. Wagg, R. J. Barthorpe, E. J. Cross, and R. Fuentes, "Robust equation discovery considering model discrepancy: A sparse Bayesian and Gaussian process approach," *Mechanical Systems and Signal Processing*, vol. 168, p. 108717, 2022.

7. R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, pp. 41–75, 07 1997. Copyright - Kluwer Academic Publishers 1997; Last updated - 2023-11-29.
8. S. Bee, L. Bull, N. Dervilis, and K. Worden, "When is an SHM problem a multi-task-learning problem?," 2023.
9. M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, p. 211–244, Sept. 2001.
10. D. Kong, Y. Chen, N. Li, C. Duan, L. Lu, and D. Chen, "Relevance vector machine for tool wear prediction," *Mechanical Systems and Signal Processing*, vol. 127, pp. 573–594, 2019.
11. S. Butterworth, "On the theory of filter amplifiers," *Experimental Wireless & the Wireless Engineer*, vol. 7, pp. 536–541, Oct. 1930.