Proceedings Paper:
Cross, E.J. and Rogers, T.J. (2021) Physics-derived covariance functions for machine learning in structural dynamics ∗. In: Pillonetto, G., (ed.) IFAC-PapersOnLine. 19th IFAC Symposium on System Identification SYSID 2021, 13-16 Jul 2021, Padova, Italy. Elsevier, pp. 168-173. ISSN: 2405-8963.

https://doi.org/10.1016/j.ifacol.2021.08.353

Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 54-7 (2021) 168–173

Physics-derived covariance functions for machine learning in structural dynamics ⋆

Elizabeth J. Cross, Timothy J. Rogers

Dynamics Research Group, University of Sheffield, Sheffield, UK
(e-mail: {e.j.cross, tim.rogers}@sheffield.ac.uk).

Abstract: This paper attempts to bridge the gap between standard engineering practice and machine learning when modelling stochastic processes. For a number of physical processes of interest, derivation of the (auto)covariance is achievable. This paper suggests their use as priors in a standard Gaussian process regression as a means of enhancing predictive capability in situations where they are reflective of the process of interest. A covariance function of a linear oscillator under random load is derived and used in a regression context to predict the displacements of a vibratory system. A simulation case study is used to demonstrate the enhancement over a standard Gaussian process regression model.

Keywords: Bayesian methods, mechanical and aerospace estimation, grey-box modelling, physics-informed machine learning, time-series modelling, stochastic systems

1. INTRODUCTION

The use of stochastic processes for modelling phenomena of interest is pervasive across engineering and scientific disciplines. These processes, which capture the evolution of probability distributions through time or across a domain of interest, provide a means of describing systems that are uncertain or which have a stochastic/random element. Although the mathematics for describing and making inferences over stochastic processes is universal, the means in which they are employed can significantly vary, particularly when comparing a machine learning approach with a classical mechanical one.

In this paper we will first draw comparison between machine learning and physics-based approaches to stochastic processes before attempting to unify them by positing the physics-based approach as a principled means of establishing an informative prior for a Gaussian process regression.

Physics-informed machine learning is a growing area of interest. In terms of Gaussian process regression, physical insight can be expressed either through the mean or covariance function or both. Here we consider zero mean processes and, therefore, focus on the covariance function. The design of covariance functions to exhibit appropriate/useful behaviour has been considered by a number of researchers. Pillonetto et al. (2014) provide a review in the context of system identification and particularly highlights the use of a stable spline kernel for linear parameter estimation (Pillonetto and De Nicolao (2010)). An alternative means of building in insight can be achieved through the multiple output framework, where relationships between multivariate targets are encoded in cross-covariance terms between standard machine learning kernels (Solin et al. (2018); Jidling et al. (2018); Cross et al. (2019)). This paper suggests an alternative approach where the covariance functions are directly derived based on (partial) knowledge of the process of interest via equations of motion. The example shown here assumes a random excitation. A more general, yet somewhat involved, approach avoiding this assumption can be found in Alvarez et al. (2009), which employs a multiple output GP.

⋆ The authors would like to acknowledge the support of the EPSRC, particularly through grant reference number EP/S001565/1

2. STOCHASTIC PROCESSES IN PHYSICS AND MACHINE LEARNING

The fundamental elements for describing a stochastic process are the mean and autocorrelation, which are functions over time or the domain of interest. Considering a process $y(t)$, its mean $\mu_y(t)$ and autocorrelation $\phi_y(t_1, t_2)$ functions are

$$\begin{aligned} \mu_y(t) =& \mathbb{E}[y(t)] \\ \phi_y(t_1, t_2) =& \mathbb{E}[y(t_1)y(t_2)] \\ =& \int\int_{-\infty}^{\infty} y(t_1)y(t_2)g(y(t_1), y(t_2))dy(t_1)dy(t_2) \end{aligned} \tag{1}$$

where $\mathbb{E}$ is the expectation operator. The autocorrelation requires integration of the product of $y(t_1)y(t_2)$ and their joint probability density, g, at times t_1 and t_2. Higher order moment functions are similarly defined.

Following on from this, the (auto)covariance of a process, $k(t_1, t_2)$, is

$$k(t_1, t_2) = \mathbb{E}[(y(t_1) - \mu(t_1))(y(t_2) - \mu(t_2))] \tag{2}$$

Clearly, the autocorrelation and (auto)covariance are one in the same for a process with a zero-mean.

A *Gaussian* process is one where at each instance or iteration, the value of the variable of interest follows a normal/Gaussian distribution, with the joint distribution of a finite collection of these also normal. It is completely defined by its mean and the covariance function, i.e. one need only consider the joint density between two points (second order density).

10.1016/j.ifacol.2021.08.353

2.1 Physics-based perspective

The description of physical systems as stochastic processes is well established; the first use of the term 'stochastic process' arose in the 1930s (see Khintchine (1934); Doob (1934)), but the response of a physical system to random excitation had been under study since at least the turn of the 20th century, for example, in 1905 Einstein derived the probability distributions of the displacement through time of particles suspended in fluid (Einstein (1905)). For the interested reader, two review papers on Brownian motion by Uhlenbeck and co-authors provide an excellent discussion of the work around this time (Uhlenbeck and Ornstein (1930); Wang and Uhlenbeck (1945)).

From the mechanistic or physics-based view point, one may take the approach of assuming the form of the process $Y(t)$, and then derive the moment functions. As a simple example, the harmonic process $Y(t) = A\cos(\omega t + \Phi)$, with A and Φ random variables, will be Gaussian if A follows a Rayleigh distribution ($A \sim R(\sigma)$) and Φ a uniform distribution over $(-\pi, \pi)$. In this case, one can derive the mean and autocorrelation functions from (1), which are $\mu_Y(t) = 0$ and $\phi_Y(t_1, t_2) = \sigma^2 \cos(\tau)$ with $\tau = t_i - t_j$. Where a spectral representation is more appropriate, as may often be the case, the autocorrelation may be derived from the power spectral density of the process, as the two are Fourier duals. In a later section we will show the autocovariance of a linear oscillator under random excitation before employing it as a prior in a Gaussian process regression.

2.2 Machine-learning perspective

Stochastic processes are also a popular modelling choice for machine learning tasks, probably the most common of which is Gaussian process regression (Rasmussen and Williams (2006)). Here, one adopts a Gaussian process prior which is conditioned on a set of training data, the conditioned posterior is then used in a regression setting (see Appendix A for mathematical detail).

The use of Gaussian process regression is now fairly common in any engineering research disciplines where measured data are available from a structure or system. In structural dynamics they are commonly used for health monitoring tasks (Farrar and Worden (2012)), such as predicting features of interest to enable inference over a damage state (Bull et al. (2020); Kullaa (2011)), or to infer unmeasured loads (Holmes et al. (2016); Rogers et al. (2020)), for example.

In this data-driven approach the prior mean and (auto)-covariance functions are selected as modelling choices. The mean function is often set to zero and the covariance function selected from either squared-exponential (SE) or Matérn kernel classes.

The posterior GP mean is a weighted sum of observations in the training set (see Appendix A), with the weights determined by the covariance function. Selecting an SE or Matérn covariance function allows the regression model to be data-driven in nature; Figure 1a illustrates how the influence of a training point on a prediction decays as the distance in the input space increases when using an SE covariance function (hyperparameters arbitrarily selected). This shows how the covariance between points with similar inputs will be high, as is entirely appropriate for a data-based learner.

In the absence of training data in an area of the input space, the mean value of the GP will return to the prior mean (usually zero).

2.3 Physics-derived covariance functions in a machine learning setting

A benefit of the machine learning approach described above is that one requires little to no insight of the process of interest. In addition, a GP with an SE prior is a universal approximator (Micchelli et al. (2006)). The implicit assumption, however, when taking this approach is that we have sufficient data to characterise all behaviours of interest and that we are able to encompass these in our training dataset. In areas where there is insufficient coverage of the input domain in the training dataset, the predictive distribution will return to its (potentially uninformative) prior.

Although monitoring data of engineering systems and structures are increasing in availability, in many situations it is unlikely that we would be able to collect a fully representative dataset of all behaviours of interest. Monitoring data may be sparse due to cost limitations, or where structures operate in a complex environment, we may not have observed extremes, for example. In a structural health monitoring setting, one may wish to make predictions about an ageing structure for a prognosis task, here one most certainly would not have access to monitoring data that would allow an entirely data-driven approach.

In such cases, using a more informative prior that is representative of our (partial) knowledge of the process as engineers seems a pragmatic and sensible approach (of course, entirely befitting of a Bayesian view point). Happily, those covariance functions derived under a physics-based view point as discussed above may be readily employed in a Gaussian process regression as the prior covariance. We argue here that, when our prior knowledge may be encapsulated in a covariance function, it is appropriate and useful to do just this.

This short paper will explore a particular example of where a physics-derived covariance function may be used to improve inference over a structural system. In the next section, the covariance of a linear oscillator is derived and employed in a regression setting. We will consider its use for making predictions about linear systems, and also briefly, nonlinear systems. The paper concludes with a more general discussion of the approach suggested.

3. DERIVATION OF THE COVARIANCE FUNCTION OF A LINEAR OSCILLATOR

Consider a linear SDOF, single-degree of freedom, oscillator (with mass, damping and stiffness parameters, m, c, k respectively) driven by a forcing process $F(t)$:

$$m\ddot{y}(t) + c\dot{y}(t) + ky(t) = F(t) \tag{3}$$

The impulse response function $h(t)$, assuming $t > t_0$, is

$$h(t) = \frac{e^{-\zeta\omega_n t}}{m\omega_d}\sin(\omega_d t) \tag{4}$$

where standard notation has been used; $\omega_n = \sqrt{k/m}$, the natural frequency, $\zeta = c/2\sqrt{km}$, the damping ratio, $\omega_d = \omega_n\sqrt{1-\zeta^2}$, the damped natural frequency. The response of the system, $Y(t)$, is the convolution of the impulse response function, $h(t)$, and the excitation:

$$Y(t) = \int_{-\infty}^{\infty} F(t-r)h(r)dr \tag{5}$$

If we consider this to be a stochastic process, then the first and second order moments are

$$\mu_{Y(t)} = \mathbb{E}[Y(t)] = \mathbb{E}\left[\int_{-\infty}^{\infty} F(t-r)h(r)dr\right] \tag{6}$$

$$\begin{aligned}\phi_{Y(t_1)Y(t_2)} &= \mathbb{E}[Y(t_1)Y(t_2)] \\ &= \mathbb{E}\left[\int\int_{-\infty}^{+\infty} F(t_1-r_1)h(r_1)F(t_2-r_2)h(r_2)dr_1dr_2\right]\end{aligned} \tag{7}$$

The simplest formulation of $Y(t)$ as a stochastic process is to consider the system as deterministic and the forcing as a random process. Then Eq (7) simplifies to:

$$\begin{aligned}\mu_{Y(t)} &= \int_{-\infty}^{\infty} \mu_{F(t-r)}h(r)dr \\ \phi_{Y(t_1)Y(t_2)} &= \int\int_{-\infty}^{+\infty} \phi_{F(t_1-r_1)F(t_2-r_2)}h(r_1)h(r_2)dr_1dr_2\end{aligned} \tag{8}$$

Under a Gaussian white noise assumption, $\mu_F(t) = 0$ and $\phi_{F(t_1)F(t_2)} = \sigma^2\delta(t_1-t_2) = \sigma^2\delta(\tau)$, so $\mu_{Y(t)} = 0$ and (8) becomes:

$$\phi_{Y(\tau)} = \frac{\sigma^2}{m^2\omega_d^2}\int_{-\infty}^{\infty} e^{-\zeta\omega_n(2r_1-\tau)}\sin(\omega_d(r_1-\tau))\sin(\omega_d r_1)dr_1 \tag{9}$$

For an alternative derivation we can make use of the Fourier duality between power spectral density and autocorrelation.The power spectral density, S_{YY} is

$$S_{YY} = \frac{S_{FF}}{|-\omega^2 + i(2\zeta\omega_n)\omega + \omega_n^2|^2} \tag{10}$$

The autocorrelation is the Fourier transform of S_{YY}

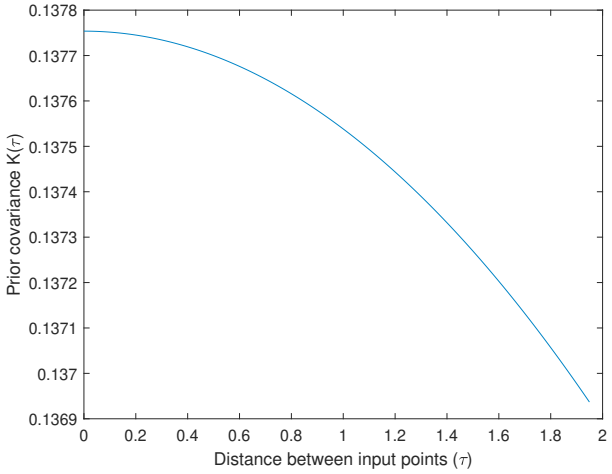
$$\phi_{Y(\tau)} = \int_{-\infty}^{\infty} e^{i\omega\tau}\frac{\sigma^2}{|-\omega^2 + i(2\zeta\omega_n)\omega + \omega_n^2|^2}d\omega, \tag{11}$$

again assuming Gaussian white noise, $S_{FF} = \sigma^2$. Through either approach the integration (which can be long winded unless one resorts to contour integrals and the residue theorem) leads to the covariance function:

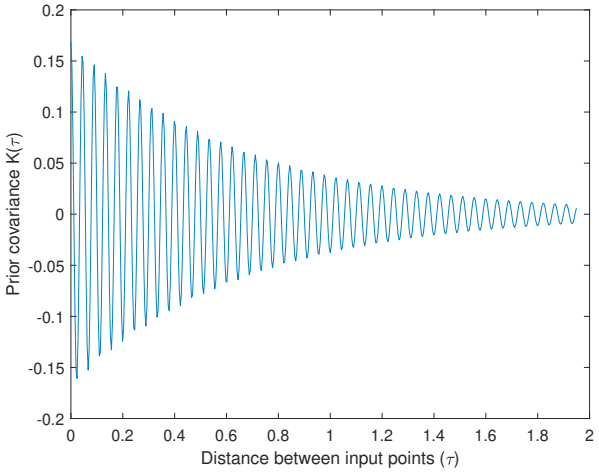
$$\phi_{Y(\tau)} = \frac{\sigma^2}{4m^2\zeta{\omega_n}^3}e^{-\zeta\omega_n|\tau|}(\cos(\omega_d\tau) + \frac{\zeta\omega_n}{\omega_d}\sin(\omega_d|\tau|)) \tag{12}$$

See also Papoulis (1965); Caughey (1971). For comparison with Figure 1a, Figure 1b shows the influence of an input point on a prediction for this covariance function.

If one is able to access modal coordinates and has well separated modes, the extension of this covariance to multi-degrees of freedom may be gained simply through a sum of the same covariance term over the multiple frequencies $w_n^{(i)}$, with $\zeta^{(i)}$ corresponding to modal damping ratios.



(a) SE



(b) SDOF

Fig. 1. Measure of influence of an input point on a prediction for the squared-exponential(SE) and SDOF covariance functions

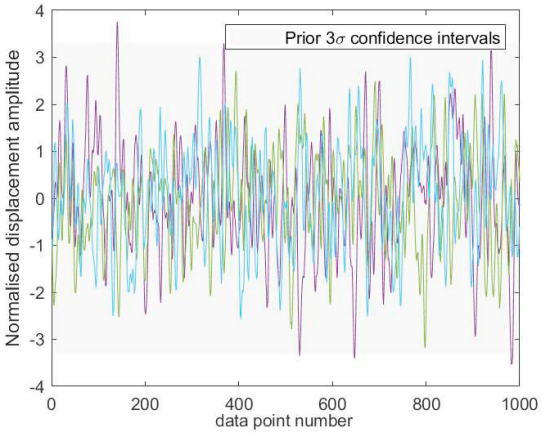
Here, for demonstration we will work with the SDOF representation.

We are now in a position to be able to use this function as our prior covariance a standard GP regression and will do so in Section 4.

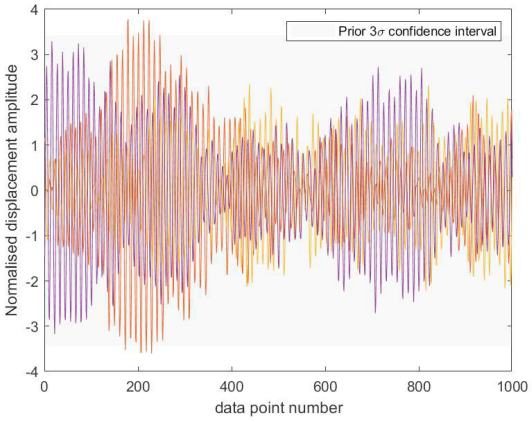
3.1 Related work - kernel design in machine learning

A number of researchers in machine learning have considered the design and construction of different kernel functions and their properties to enhance regression performance. A particularity interesting subset of the research papers on this topic refer to 'expressive kernels'. These kernels have a richer spectral content than the more standard ones, such as the squared-exponential discussed above.

In Wilson and Adams (2013), the authors use the Fourier duality between PSD and autocorrelation (referred to there as Bochner's theorem) to construct the Spectral Mixture (SM) kernel from a mixture of Gaussians in the frequency domain. They show that this expressive covariance is able considerably outperform the standard



(a) Prior draws with a squared-exponential covariance



(b) Prior draws with the SDOF covariance

Fig. 2. Prior draws from different covariance functions

kernels for extrapolation. This is extended in Parra and Tobar (2017) to the multiple output case, where phase-lags between variables are also accounted for (see also Boyle and Frean (2005)). The covariance structure in the SM is similar to that shown in (12) (notably the weighted sinusoid is absent in the SM), indicating that it is likely that (12) will be useful in a generic regression task. This idea is not pursued here, where the interest is much more in what may be gained from employing physical insight of the system in question to derive priors that are useful for the specific regression task in hand.

4. EXPLORATION OF USING PHYSICS-DERIVED COVARIANCE FUNCTIONS IN GPR

In this section we will explore the characteristics of the SDOF covariance function in a regression setting.

Figure 2 shows the draws from Gaussian process priors with a squared-exponential and the SDOF covariance function respectively (both zero mean). As is to be expected, the draws from the SDOF covariance resemble responses of a linear oscillator under white noise, with different amplitudes and phases accounted for. One can see that there is much more structure in the prior draws of the SDOF covariance process than from the squared-exponential.

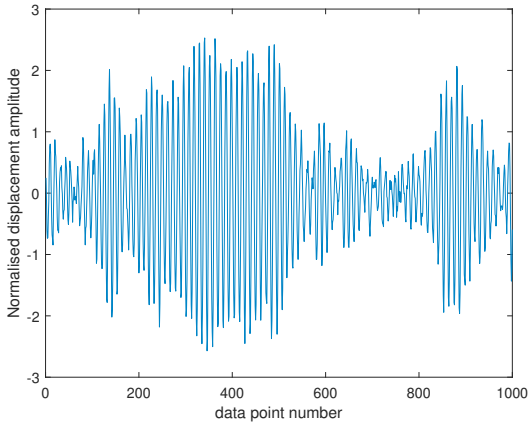


Fig. 3. Simulated SDOF system under random load

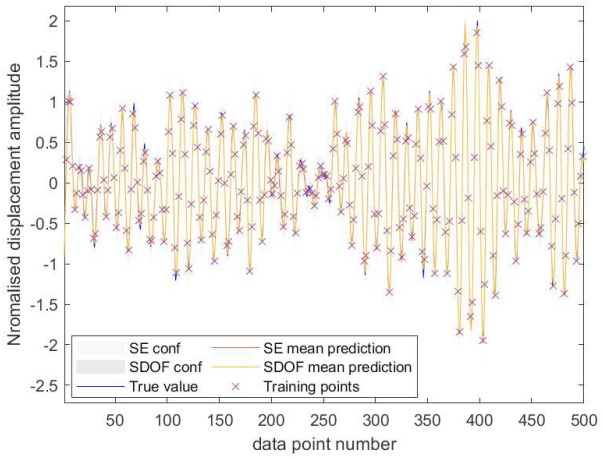
To explore further a simulation is employed, here a linear oscillator with $\omega_n = 141.4$ and $\zeta = 0.01$ is excited under white noise with $\sigma^2 = 1e-6$. The simulation time history is shown in Figure 3.

Gaussian process regression is attempted using subsets of the simulation for training/conditioning. Figure 4 shows the posterior predictions of two GPs conditioned on every second simulation point until data point 500, one has an SE covariance, the other the derived SDOF covariance. One can see that both GPs have an excellent fit in the training regime.

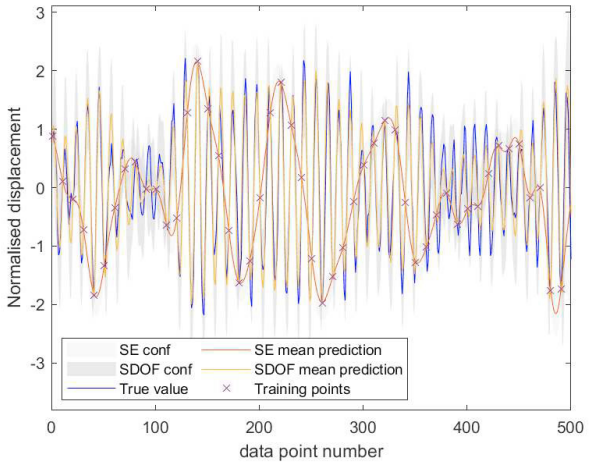
The potential benefits of employing the physical covariance structure become clearer when training data are less abundant. Figure 4b shows each GP conditioned now on every tenth data point. The prediction accuracy of the SDOF GP far outweighs that of the SE in this case and the benefit of our knowledge of the physical system is brought to bear.

The system (hyper)parameters may be fixed or learned based on the available knowledge. The optimal hyper-parameters are sought here via maximising the marginal likelihood of the predictions. In this paper a particle swarm optimisation is used following Rogers (2019). In the case above, hyperparameter optimisation is able to reproduce the same prediction results as when they are fixed to match the parameters of the simulation. Although not the focus of this paper, the authors note that accurate parameter estimation is possible using a small number of conditioning points, but does require adaption of the optimisation approach. This will be the topic for a separate paper.

The example above represents the situation where one's partial knowledge is of the structure of a system and not of its parameters. Where our structural knowledge is partial, our intention with this approach is to pursue a combination of covariance functions for the prior, building in additional flexibility to account for unknown behaviour. One might envisage the combination of a covariance function representing an underlying linear system with a data-driven covariance function to account for an unknown non-linearity, for example. These combination will be demonstrated in future work, however, here we briefly explore

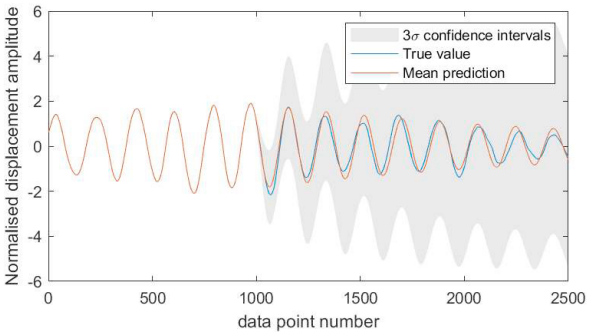


(a) Conditioned on every data point up until data point 500

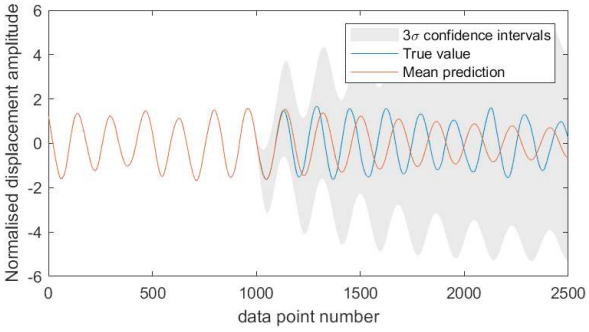


(b) Conditioned on every 10th data point up until data point 500

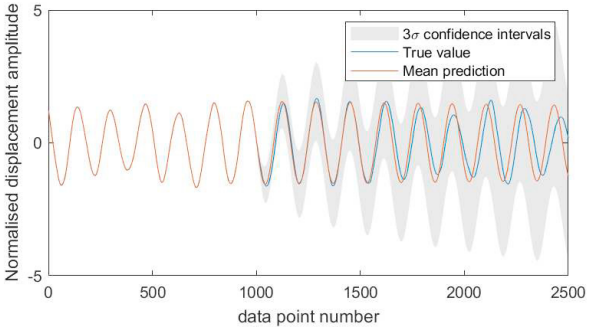
Fig. 4. Comparison between SE and SDOF kernels when conditioned on simulated vibration data



(a) Fit to linear simulation



(b) Fit to nonlinear simulation with hyperparameters fixed to those of the linear system



(c) Fit to nonlinear simulation with learned hyperparameters

Fig. 5. A comparison when using the SDOF covariance for linear and nonlinear systems. GP training in all cases is on data points 1:2:1000.

the use of the current covariance function for a nonlinear system.

The simulation above is repeated with and without a nonlinear (cubic) spring added between ground and the mass (the nonlinear spring has the same stiffness coefficient as the linear one). The same random forcing was applied to the linear and nonlinear system. Figure 5 compares the fit to the linear simulation with that of the nonlinear simulation (every second point up to data point 1000 was used for training). In terms of modelling the nonlinear system, Figure 5b shows the case when the covariance hyperparameters are fixed to those of the known linear system. Figure 5c shows the fit when the hyperparameters are learned. One can see that in all cases, the fit in interpolation is perfect, demonstrating the flexibility of the covariance function. It is the extrapolative ability that is affected by the misspecification of the hyperparameters, where one can see that the GP defined by the linear system parameters loses predictive capability particularly in terms of the phase.

5. DISCUSSION

The previous section highlights the usefulness of a physics-informed prior for a regression task. The derived covariance function brings a structure which encodes the expected behaviour and is able to account for missing information in the training set. An additional benefit from the approach is that the hyperparameters are interpretable physically, meaning that the learning of them may now be guided by our insight. For example, the frequency component may be set based on prior engineering analysis, equally a hyper-prior distribution could be set reflecting our belief in the possible values our system may take.

The usefulness of this approach will naturally depend on the derived covariance adopted and the particular application. The SDOF covariance function adopted here is expressive; it is able to represent a wide range of behaviours, including some nonlinear behaviours. In the

example shown, the covariance function is able to perform well for a simulated Duffing oscillator. This indicates that the higher order moments in the process induced by the nonlinearity are not contributing significantly to the response, rendering the Gaussian process assumption useful in this case.

Future work will consider encapsulating partial knowledge of the process of interest through combinations of derived and data-driven covariance functions. Here, the benefits of working in a GP framework are evident; as covariance functions remain valid under linear operation, one can easily manipulate and combine candidate covariance terms to tailor the model to a particular task at hand.

REFERENCES

Alvarez, M., Luengo, D., and Lawrence, N.D. (2009). Latent force models. In *Artificial Intelligence and Statistics*, 9–16.

Boyle, P. and Frean, M. (2005). Dependent Gaussian processes. In *Advances in neural information processing systems*, 217–224.

Bull, L.A., Gardner, P., Rogers, T.J., Cross, E.J., Dervilis, N., and Worden, K. (2020). Probabilistic inference for Structural Health Monitoring: New modes of learning from data. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 7(1), 03120003.

Caughey, T. (1971). Nonlinear theory of random vibrations. In *Advances in applied mechanics*, volume 11, 209–253. Elsevier.

Cross, E., Gibbons, T., and T.J, R. (2019). Grey-box modelling for structural health monitoring; physical constraints on machine learning algorithms. In *Proceedings of the International Workshop on Strucutral Health Monitoring*.

Doob, J.L. (1934). Stochastic processes and statistics. *Proceedings of the National Academy of Sciences of the United States of America*, 20(6), 376.

Einstein, A. (1905). On the motion of small particles suspended in liquids at rest required by the molecular-kinetic theory of heat. *Annalen der physik*, 17(549-560), 208.

Farrar, C.R. and Worden, K. (2012). *Structural Health Monitoring: a machine learning perspective*. John Wiley & Sons.

Holmes, G., Sartor, P., Reed, S., Southern, P., Worden, K., and Cross, E. (2016). Prediction of landing gear loads using machine learning techniques. *Structural Health Monitoring*, 15(5), 568–582.

Jidling, C., Hendriks, J., Wahlström, N., Gregg, A., Schön, T.B., Wensrich, C., and Wills, A. (2018). Probabilistic modelling and reconstruction of strain. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 436, 141–155.

Khintchine, A. (1934). Korrelationstheorie der stationären stochastischen prozesse. *Mathematische Annalen*, 109(1), 604–615.

Kullaa, J. (2011). Distinguishing between sensor fault, structural damage, and environmental or operational effects in Structural Health Monitoring. *Mechanical Systems and Signal Processing*, 25(8), 2976–2989.

Micchelli, C.A., Xu, Y., and Zhang, H. (2006). Universal kernels. *Journal of Machine Learning Research*, 7(Dec), 2651–2667.

Papoulis, A. (1965). *Probability, random variables, and stochastic processes*. McGraw-Hill Education.

Parra, G. and Tobar, F. (2017). Spectral mixture kernels for multi-output Gaussian processes. In *Advances in Neural Information Processing Systems*, 6681–6690.

Pillonetto, G. and De Nicolao, G. (2010). A new kernel-based approach for linear system identification. *Automatica*, 46(1), 81–93.

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., and Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50(3), 657–682.

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*, volume 38. The MIT Press, Cambridge, MA, USA.

Rogers, T.J. (2019). *Towards Bayesian System Identification: With Application to SHM of Offshore Structures*. Ph.D. thesis, University of Sheffield.

Rogers, T., Worden, K., and Cross, E. (2020). On the application of Gaussian process latent force models for joint input-state-parameter estimation: With a view to Bayesian operational identification. *Mechanical Systems and Signal Processing*, 140, 106580.

Solin, A., Kok, M., Wahlström, N., Schön, T.B., and Särkkä, S. (2018). Modeling and interpolation of the ambient magnetic field by Gaussian processes. *IEEE Transactions on robotics*, 34(4), 1112–1127.

Uhlenbeck, G.E. and Ornstein, L.S. (1930). On the theory of the Brownian motion. *Physical review*, 36(5), 823.

Wang, M.C. and Uhlenbeck, G.E. (1945). On the theory of the Brownian motion (II). *Reviews of modern physics*, 17(2-3), 323.

Wilson, A. and Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, 1067–1075.

Appendix A. GAUSSIAN PROCESS REGRESSION

Here we follow the notation used in Rasmussen and Williams (2006); $k(\mathbf{x}_p, \mathbf{x}_q)$ defines a covariance matrix K_{pq}, with elements evaluated at the points $\mathbf{x}_p$ and $\mathbf{x}_q$, where $\mathbf{x_i}$ may be multivariate.

Assuming a zero-mean function, the joint Gaussian distribution between measurements/observations $\mathbf{y}$ with inputs X and unknown/testing targets $\mathbf{y}^*$ with inputs X^* is

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{y}^*\end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix}K(X,X)+\sigma_n^2 I & K(X,X^*)\\ K(X^*,X) & K(X^*,X^*)\end{bmatrix}\right) \tag{A.1}$$

The distribution of the testing targets $\mathbf{y}^*$ conditioned on the training data (which is what we use for prediction) is also Gaussian:

$$\mathbf{y}^*|X_*, X, \mathbf{y} \sim \mathcal{N}(K(X^*,X)(K(X,X)+\sigma_n^2 I)^{-1}\mathbf{y}, \\ K(X^*,X^*) - K(X^*,X)(K(X,X)+\sigma_n^2 I)^{-1}K(X,X^*)) \tag{A.2}$$

See Rasmussen and Williams (2006) for the derivation. The mean and covariance here are that of the posterior Gaussian process.