Gaussian process regression based on dimension reduction for time series outputs

Baptiste Kerleguer [1]

[1] *CEA, DAM, DIF, F-91297, Arpajon, France & baptiste.kerleguer@cea.fr*

Résumé. La régression par processus gaussien est largement utilisée pour émuler la sortie d'un code coûteux. Nous nous intéressons à des codes dont la sortie est une série temporelle. Pour réaliser des métamodèles de ces codes, il est usuel de réduire la dimension. Ensuite, une régression est réalisée dans l'espace latent, par exemple régression par processus Gaussien. Ce travail étudie l'influence de l'indépendance des paramètres dans l'espace latent sur les sorties. Pour cela, à l'aide de données simulées et d'un système physique, nous étudions les processus ainsi créés dans l'espace latent et nous les comparons aux hypothèses de décorrélation des coefficients dans l'espace latent faites lors de la métamodélisation.

Mots-clés. Reduction de la dimension, Processus Gaussien, Métamodèle

Abstract. Gaussian process regression is commonly used to emulate the output of an expensive code. The code outputs are assumed to be a time series. To produce surrogate models of these codes, it is common practice to reduce the dimension. Regression is then performed in the latent space using, for example, Gaussian process regression. This work studies the influence of the independence of the parameters in the latent space. We use simulated data and a physical system. We study the processes created in the latent space and compare them with the assumptions made during surrogate modeling (decorrelation and same variance).

Keywords. Dimension reduction, Gaussian process, surrogate model

1 Introduction

Advances in scientific modeling have led to the development of more complex and computationally expensive codes. It has therefore become necessary to use surrogate models, constructed from the outputs of these codes, in order to study the uncertainties of these codes. One method widely used in the uncertainty quantification community to produce surrogate models is Gaussian process (GP) regression, see Gramacy (2022) and Williams and Rasmussen (2006). This method, also known as Kriging, was introduced for geostatistics before being used for numerical experiments and uncertainty quantification.

As the complexity of scientific modeling has increased, model outputs have become more and more sophisticated. One of the difficulties is the high dimension of the outputs. Among the codes with high-dimensional outputs, we are interested in those whose outputs are time-dependent functions. When they are sampled, they are called time series. We assume that

the sampling is done on a fine, regular grid, so the output is very high-dimensional. It is assumed that time series are available at all times for each call to the code.

When the regression problem has high-dimensional outputs, such as time series, then the natural solution is to reduce the dimension of the outputs to return to the previous case. The problem is therefore divided into two parts: dimension reduction and Gaussian process regression. Usually, these two parts are treated one after the other, with a certain degree of decoupling. The question is: to what extent is it possible to have a regression model that is independent of dimension reduction? In other words, can we change the dimension reduction method without changing the model in latent space?

The problem we are interested in is the following: let us consider a computational code of which outputs are of the form $z(\mathbf{x}, t)$ with $\mathbf{x} \in \mathbb{R}$ and $t \in [0, 1]$. In the following the stochastic process emulating the code in $Z(\mathbf{x}, t)$. We know this code in a limited number of $\mathbf{x}$ but for t discretized on a regular grid. We try to reduce the output $z(\mathbf{x}, t)$ by r coefficients $a_1(\mathbf{x}), \cdots, a_r(\mathbf{x})$. For that we use a dimension reduction noted $\mathcal{F}$. The principle of dimension reduction is illustrated in Figure 1. For the regression, the objective is to predict in the space of $z(\mathbf{x}, t)$, it is thus necessary to be able to return to the output space of the code. This is how we look for a pseudo-inverse of $\mathcal{F}$ denoted $\tilde{\mathcal{F}^{-1}}$. Note that $\tilde{\mathcal{F}^{-1}}$ is not an inverse of $\mathcal{F}$ because there would be no reduction of dimension. The only exception would be the case where r is equal to the number of points on the time grid. Thus, we obtain an approximation $\hat{z}(\mathbf{x}, t)$ of $z(\mathbf{x}, t)$. Our objective is to have r as small as possible and to have $\hat{z}(\mathbf{x}, t)$ as close as possible to $z(\mathbf{x}, t)$. In order to solve the regression problem, GP regression will be used on the coefficients $a_1(\mathbf{x}), \cdots, a_r(\mathbf{x})$. The space of these coefficients is called the latent space.

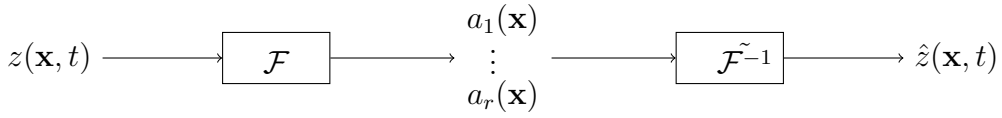

Figure 1: Illustration of the dimension reduction for time series outputs. $\mathcal{F}$ is the dimension reduction function and $\mathcal{F}^{-1}$ its pseudo inverse.

A survey of surrogate model methods for high-dimensional outputs with uncertainty quantification is available in Kontolati et al. (2022). We restrict ourselves to methods that allow an inverse reconstruction in the original space (existence of a pseudo-inverse introduced before).

All that remains, given the dimension reduction, to emulate the coefficients $a_1(\mathbf{x}), \cdots, a_r(\mathbf{x})$ to obtain a surrogate model of the output. We assume that $\mathbf{A} = \{A_1(\mathbf{x}), \cdots, A_r(\mathbf{x})\}$ is a Gaussian process for which the coefficients $a_1(\mathbf{x}), \cdots, a_r(\mathbf{x})$ are a realisation. The common assumption for surrogate modeling the vector $\mathbf{A}$ is that its coefficients are uncorrelated. With an appropriate covariance kernel it is possible to predict $\mathbf{A}$ and therefore to have a surrogate model of $Z(\mathbf{x}, t)$.

In the following section, we present the pendulum attached to a mass-spring system we will use to illustrate this work. Then, in the third section, we study Gaussian processes in the latent space. Using the dimension reduction part of these methods, we will then see how the stochastic process $Z(\mathbf{x}, t)$ behaves compare to a Gaussian process with known mean and

covariance.

2 Physical data

Figure 2 shows the system we propose to study. The data comes from a numerical simulation. For the numerical simulation the spring is assumed to have a restoring force proportional to k and the pendulum is assumed to have a mass m at one point. The inputs are k, M, θ, $\dot{\theta}$ and y_0. The output is the position of the pendulum during the first 10 seconds. The other values are fixed. The system with the variations presented was introduced in Perrin (2020). A study of this system with the concidered ranges of variations was carried out in Kerleguer (2023). Perrin (2020) shows that it is possible to build a efficient surrogate model of this system either by dimension reduction and GP regression or by tensorization of the covariance on a regular grid for GP regression.

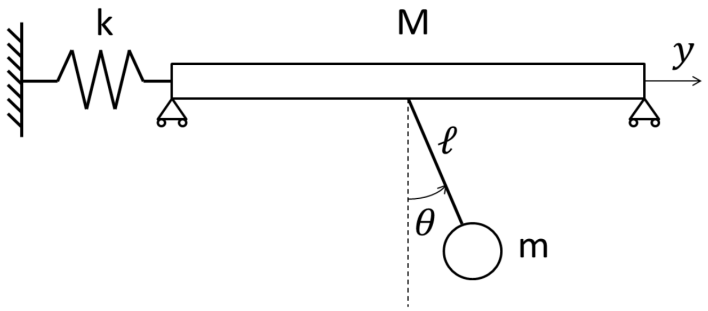

Figure 2: Diagram of a mass-spring system with a simple pendulum attached to it.

This model is used to construct the data studied below. To do this, a sample of 1500 realizations of the system drawn uniformly in the space of the inputs is carried out. The temporal properties of the output signal are given in Figure 3 and 4. Figure 3 shows that the system has a non-zero trend. Furthermore, the convariance, shown in Figure 4, seems to indicate that the covariance is time-dependent.

All the figures of this document showing stochastic processes use the same pattern. Figures 3 and 5 to 9 show stochastic processes with 10 realisations in colour and the mean in dotted grey. The 95% interval centred on the mean is shown in grey.

3 Gaussian process in the latent space

In this section, we present dimension reduction methods used in more detail. We then propose a method for understanding the behavior of Gaussian processes in latent space. Then, two approaches to constructing stochastic processes in latent space are compared. The first approach consists of generating Gaussian processes in real space and then passing them into

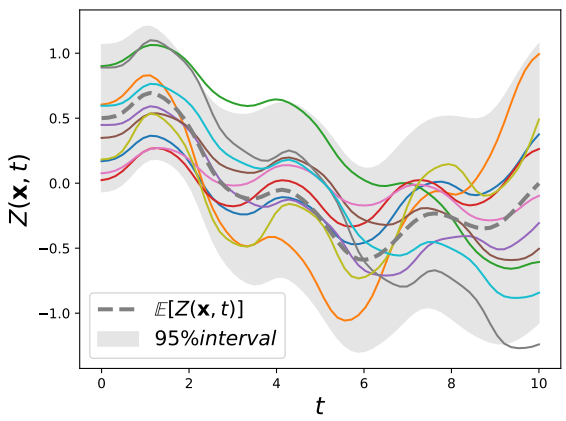

Figure 3: Output of the mass-pendulum system

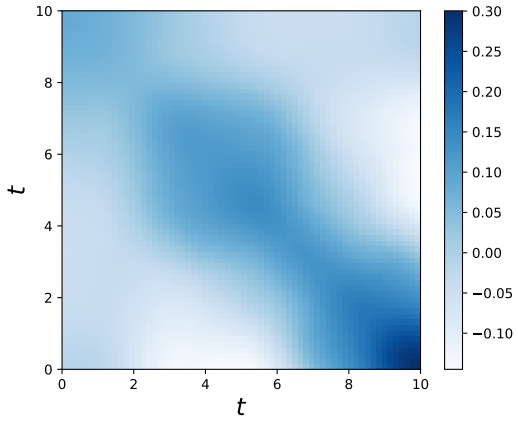

Figure 4: Covariance of the time output of the system

latent space. The process laws are then studied in latent space. The second approach is to assume that the coefficients are independent in the latent space. We can then generate realizations of a Gaussian vector in latent space and study the processes in arrival space. Finally, the GPs are compared with the pendulum data.

3.1 Dimension reduction

The different dimension reduction techniques for surrogate modeling used in this work are:

- Singular value decomposition (SVD), with GP regression presented in Nanty (2017),
- Wavelet transform, with GP regression applications given in Perrin et al. (2021) and Rohmer et al. (2023),
- Autoencoder, with an application in Donnelly et al. (2024).

These three types of dimension reduction are studied because they make it easy to find $\tilde{\mathcal{F}^{-1}}$. In addition, these methods are widely used in the literature, see Kontolati et al. (2022). Moreover, studies on the latent space in the case of wavelets with computation of the corresponding covariance kernel have already been carried out in Kerleguer (2022).

The SVD gives an expression of the Gaussian process Z depending on the basis and coefficients:

$$Z(\mathbf{x}, t) = \sum_{i=1}^{r} A_i(\mathbf{x})\Gamma_i(t) + \varepsilon(\mathbf{x}, t), \tag{1}$$

with N is the number of elements, $\mathbf{\Gamma}_N = \{\Gamma_i\}_{i=1,\dots,N}$ is the basis, the coefficients are A_i and ε is the error of dimension reduction. The base $\mathbf{\Gamma}_N$ is computed using the training data. The

decomposition is then performed each time the surrogate model is called. The equation (1) can be detailed for the case of wavelet decomposition. The equation thus becomes:

$$Z(\mathbf{x}, t) = \sum_{m,n \in \mathcal{I}} A_{i=\{m,n\}}(\mathbf{x}) \psi_{m,n}(t) + \varepsilon(\mathbf{x}, t), \tag{2}$$

with I the space of the wavelet coefficients, and ψ the wavelet function. In the following, the Haar decomposition will be used.

For autoencoder, a $\mathbf{\Gamma}_N$ basis representing the latent space is not available. This complicates the expression of the equations (1) & (2). However, it is always possible to return from latent space to time series space. To compute the autoencoder, the network is small: 2 hidden layers with 20% dropout during training. Optimization time is of the order of a minute on a desktop computer. The latent space of the autoencoder is of dimension 16.

3.2 Gaussian process in dimension reduction

$Z(\mathbf{x}, t)$ is assumed to be a Gaussian process with mean zero and covariance Matérn kernel $5/2$ and $\mathbb{V}[Z(\mathbf{x}, t)] = 1$. This process is then applied to the various dimension reduction methods. This gives figures 5, 7 and 9 for SVD, autoencoder and wavelets respectively. The two linear methods show similar behaviour and a rapid decrease in the importance of the coefficients. This decrease is faster for the SVD method, which uses the data to calculate an adapted base. For the autoencoder, the behaviour is different and the coefficients all seem to behave similarly. As the coefficients are exchangeable in the model, this is also to be expected.

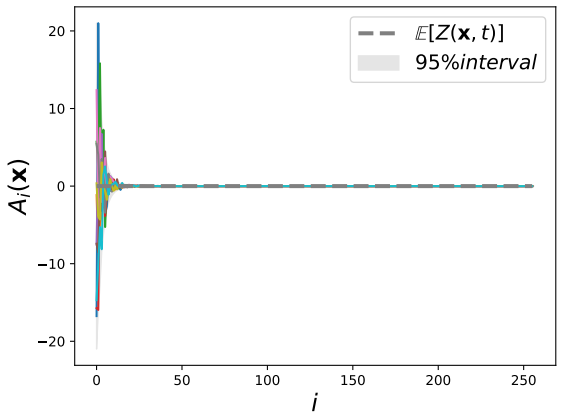

Figure 5: Gaussian Process in the SVD latent space

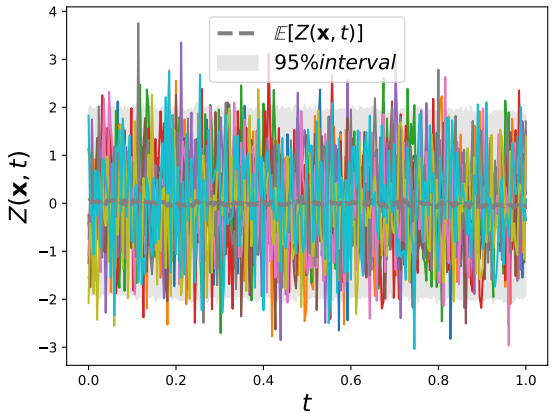

Figure 6: Prediction of a process in the time space with SVD

Figures 6, 8 and 10 are constructed assuming that $A = [A_1(x), \ldots, A_N(x)]$ is a Gaussian vector with mean zero and covariance identity. We can see that for the SVD and wavelet methods, we have high-frequency curves in the time series space. This is because, in the case of a regression, a truncation would be performed, which would eliminate the high-frequency

coefficients. In addition, the decrease in variance will lead to coefficients becoming less and less important as it increases. The mean and variance of the processes are coherent with the input process. In the case of the autoencoder, the samples are much closer to Gaussian processes with Matèrn kernel. However, the variance is lower by an order of two, as can be seen from the 95% interval.

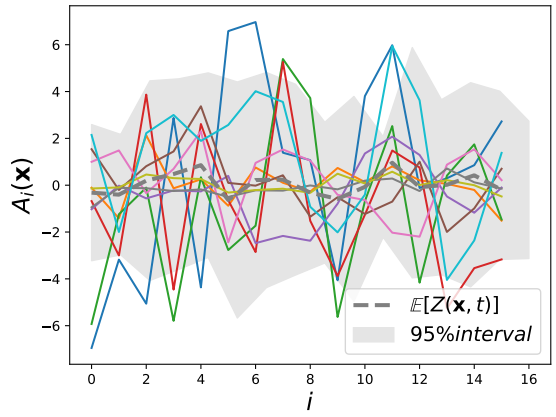

Figure 7: Gaussian Process in the autoencoder latent space

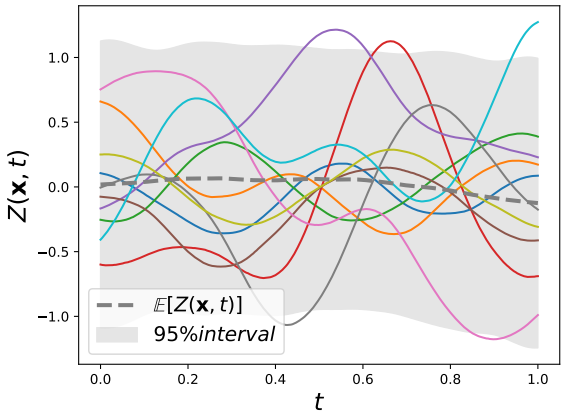

Figure 8: Prediction of a process in the time space with autoencoder

In this example, there is a close relationship between SVD and wavelet decomposition. The latter being a deterministic decomposition, this motivates the approach presented in Perrin (2021).

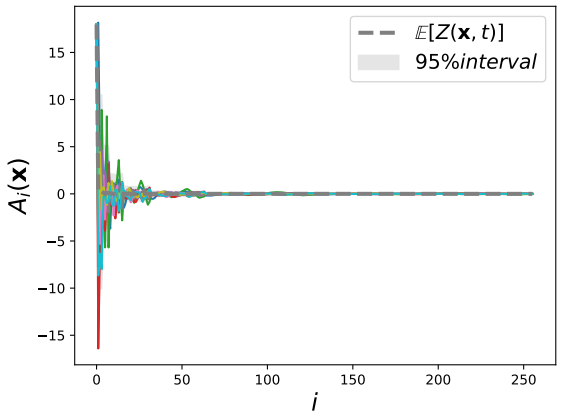

Figure 9: Gaussian Process in the wavelet space

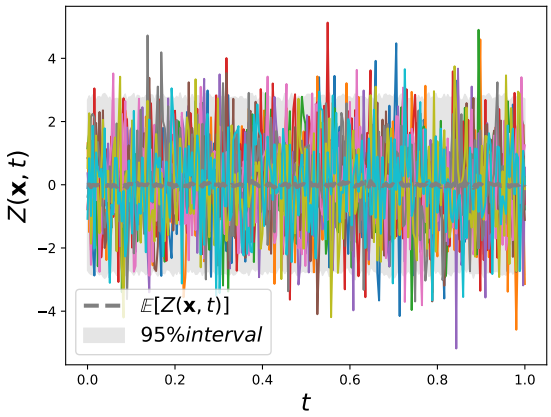

Figure 10: Prediction of a process in the time space with wavelet decomposition

The covariance matrices of the dimension reduction methods also explain the differences between the methods. For SVD, we obtain a diagonal matrix, as expected in theory. The diagonal has decreasing coefficients, which are rapidly negligible numerically. The wavelet

decomposition also shows rapid decrease on the diagonal. Note that this decrease is also visible in Figures 2 and 6. However, while many non-diagonal terms are zero as predicted in Morel et al. (2023), there are still some non-zero terms, see Figure 11, but it seems possible to neglect it. The analytical calculation of these terms is available in Kerleguer (2022). For autoencoding, the covariance matrix is full, see Figure 12, as you would expect from a non-linear method. However, this has no influence on the model's ability to generate Gaussian processes in the time series space, see Figure 5.

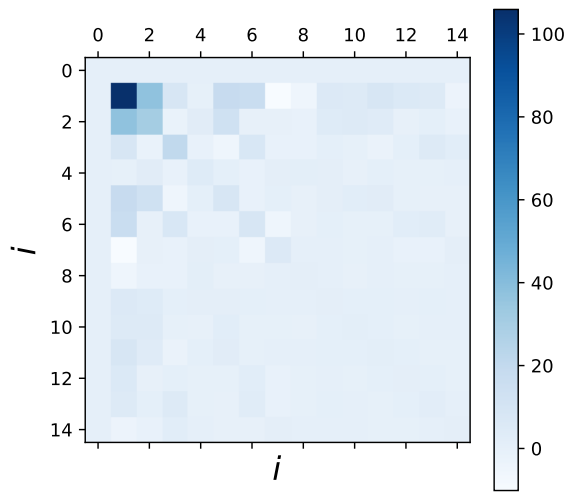

Figure 11: Covariance in the wavelet space

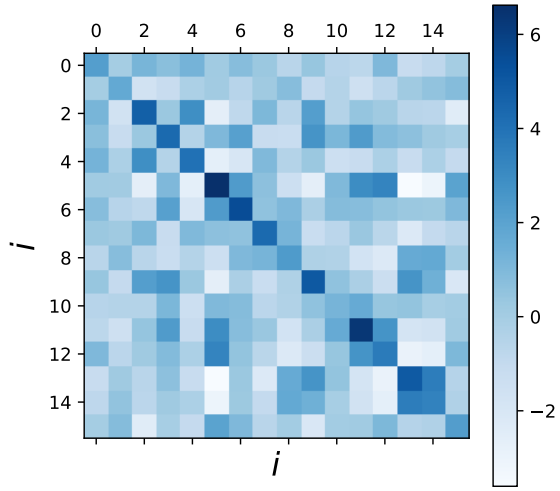

Figure 12: Covariance in the autoencoder space

3.3 Application of methods to pendulum systems

Now that the assumptions made in the dimension-reduced GP surrogate modeling have been fixed, these assumptions are compared with the pendulum data. Figures 13, 14 and 15 show the same methods as Figures 5, 7 and 9 but for pendulum data. It should be noted for Figures 13 and 14 that the latent space is of a smaller dimension than the latent space of a GP because the data is less sampled than the realizations of the GP. The mean is the same in the latent space for pendulum data as for GP, but the variance is much smaller. Consequently, it will be straightforward to model A processes in latent space by GP. The surrogate models built using these 3 dimension reduction methods give an R^2 scores for 200 training time series between 0.99 and 1, with a test base of 1300 time series.

These results seem to validate the independence hypotheses proposed in Kontolati et al. (2022). However, the empirical covariance matrix in wavelet space proposed in Figure 16 shows a strong correlation between coefficients. The same is true for the autoencoders, although this problem has already been identified for the GP. This is likely to be a problem

in cases of small data sets where the quantification of uncertainties is important. The A_i correlation should be taken into account in the creation of a surrogate model quantifying the uncertainties.

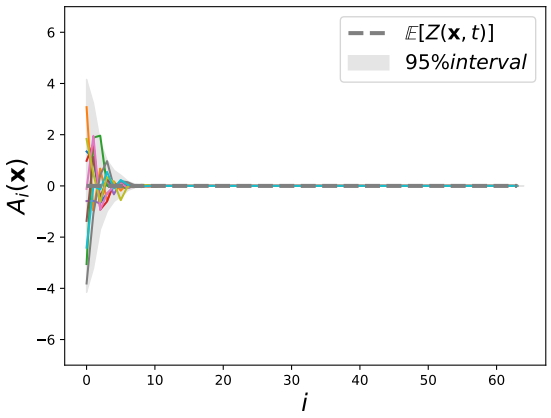

Figure 13: Pendulum data in the SVD

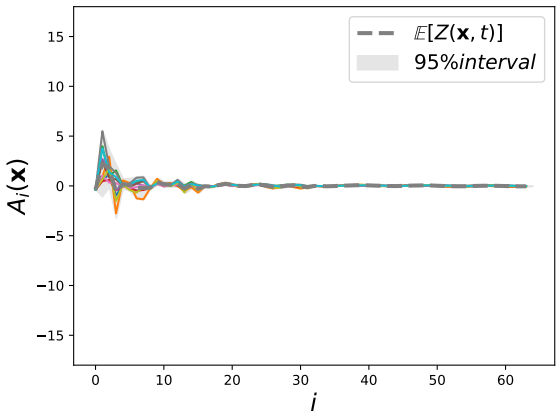

Figure 14: Pendulum data in the wavelet space

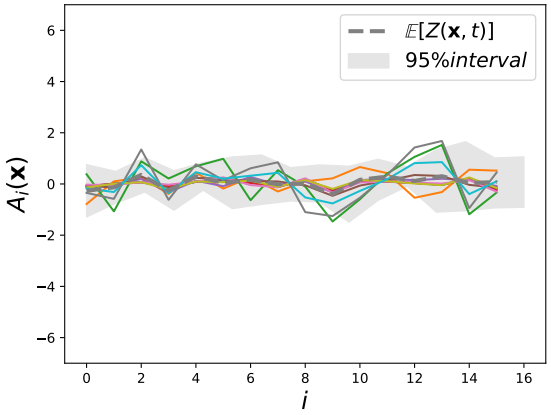

Figure 15: Pendulum data in the autoencoder latent space

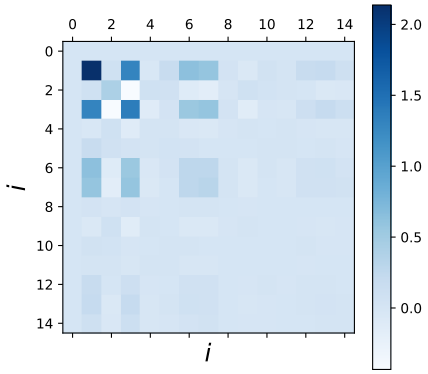

Figure 16: Pendulum data covariance in the wavelet space

4 Conclusion

To perform a regression with a time series output, it is possible to reduce the dimension and then perform a Gaussian process regression in the latent space. This method has been shown numerically to be effective on numerous occasions, see Kontolati et al. (2022). However, since dimension reduction does not specifically imply that the coefficients are uncorrelated, it is open to question whether this assumption should be made. We have investigated exactly how this assumption can be made. The data used to study the latent space are of two

kinds: simulated data and data from a physical system. In both cases, it was shown that the processes constructed from these hypotheses can be used for time series regression. However, these hypotheses are not always verified and the quantification of uncertainties in surrogates models could benefit from correlated GP in latent space. In a future work we would like to construct covariance kernels adapted to the different latent spaces in order to be able to take into account correlation between coefficients.

Bibliographie

Donnelly, J., Daneshkhah, A., & Abolfathi, S. (2024). Forecasting global climate drivers using Gaussian processes and convolutional autoencoders, *Engineering Applications of Artificial Intelligence*, 128, 107536.

Gramacy, R. B. (2020). *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*, CRC press, Boca Raton.

Kerleguer, B. (2022). Multi-fidelity surrogate modeling adapted to functional outputs for uncertainty quantification of complex models *(Doctoral dissertation, Institut Polytechnique de Paris)*.

Kerleguer, B. (2023). Multifidelity Surrogate Modeling for Time-Series Outputs, *SIAM/ASA Journal on Uncertainty Quantification*, 11(2), 514-539.

Kontolati, K., Loukrezis, D., Giovanis, D. G., Vandanapu, L., & Shields, M. D. (2022). A survey of unsupervised learning methods for high-dimensional uncertainty quantification in black-box-type problems, *Journal of Computational Physics*, 464, 111313.

Morel, R., Rochette, G., Leonarduzzi, R., Bouchaud, J. P., & Mallat, S. (2023). Scale dependencies and self-similar models with wavelet scattering spectra, *Available at SSRN* 4516767.

Nanty, S., Helbert, C., Marrel, A., Pérot, N., & Prieur, C. (2017). Uncertainty quantification for functional dependent random variables,*Computational Statistics*, 32, 559-583.

Perrin, T. V. E., Roustant, O., Rohmer, J., Alata, O., Naulin, J. P., Idier, D., Pedreros, R., Moncoulon, D. & Tinard, P. (2021). Functional principal component analysis for global sensitivity analysis of model with spatial output. *Reliability Engineering & System Safety*, 211, 107522.

Perrin, G. (2020). Adaptive calibration of a computer code with time-series output, *Reliability engineering & system safety*, 196, 106728.

Rohmer, J., Sire, C., Lecacheux, S., Idier, D., & Pedreros, R. (2023). Improved metamodels for predicting high-dimensional outputs by accounting for the dependence structure of the latent variables: application to marine flooding, *Stochastic Environmental Research and Risk Assessment*, 1-23.

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, Cambridge, MA: MIT press.