Minimum Reduced Order Modelling

Robert A. Milton, Solomon F. Brown

Department of Chemical and Biological Engineering, University of Sheffield, Sheffield, S1 3JD, United Kingdom

Abstract

Keywords: Gaussian Process, Global Sensitivity Analysis, Sobol' Index, Surrogate Model

1. Introduction

The simulation of a large range of engineering systems requires the application of complex computational models. The use of these models often requires considerable computational effort, and can be prohibitive when attempting to use the underlying model as part of a system optimization. This is commonly mitigated by emulating the complex model response $y(\mathbf{x})$ to its M-dimensional input $\mathbf{x}$ with a surrogate.

A popular class of surrogate is Gaussian Processes (GPs) [1, 2]which are flexible, efficient, non-parametric and analytically tractable (references). Other surrogate methods include Polynomial Chaos Expansions [3, 4, 5], low-rank tensor approximations [6, 7], and support vector regression [8]. These, however, tend to lack the combination of characteristics just mentioned, which make GPs ideal for our purposes.

The development of surrogate models usually requires an exponentially growing number of output results $y(\mathbf{x})$ throughout the input space as M increases: known as the curse of dimensionality. There is therefore a significant driver for methods with which to reduce the dimensionality of the input space, and so a more efficient means of generating the emulator. One way of selecting the directions of most influence is through the application of

Email addresses: `r.a.milton@sheffield.ac.uk` (Robert A. Milton), `s.f.brown@sheffield.ac.uk` (Solomon F. Brown)

Global Sensitivity Analysis, however where these directions are not aligned with the input basis this resulting dimensionality reduction is sub-optimal (references).

As such a number of approaches to obtaining the optimal dimensionality reduction have been developed, which can broadly be catagorized through their use of different sensitivity measures. For example, [9] proposed a means of calculating this optimal reduced dimensional space, the Active Subspace, through a derivative sensitivity measure. This was found to work very effectively for a broad range of problems but requires the information of the derivative which is typically not available for very large or complex systems. Liu and Guillas [10] suggested using gradient-based kernel reduction of a GP surrogate. Minimum Average Variance Estimation (MAVE), such as that proposed by Xia et al [11], on the other hand effectively uses a variance-based sensitivity (statistical independence)

A key variance-based sensitivity measure are the Sobol' indices [12], which is one of the most widely used approaches for GSA. With the increased popularity of using GPs as surrogate models, the main disadvantage to the Sobol' method was solved [13, 14, 15]. The use of GPs provided an alternative method to estimating multidimensional integrals using Monte Carlo schemes, which required 10,000 datapoints to reach 10% precision [16]. GPs enable semi-analytic evaluation of Sobol' indices, introduced by Jin et al. [14]. Before Oakley and O'Hagan [13] used the global stochastic model of a GP, providing the calculations to produce random variables as a new sensitivity measures. Oakley and O'Hagan's [13] model allows the sensitivity indices accuracy to be analysed due to the distribution of the variables. Marrel et al. [15] extended this comparing them and building on the work from Oakley and O'Hagan [13] leading to a novel algorithm which builds confidence intervals for the Sobol' indices. Marrel et al. [15] tested both methods on toy functions providing results that show very accurate sensitivity indices and satisfactory confidence intervals from the second method. However, when the approach was illustrated on real data to provide a sensitivity analysis on radionuclide groundwater transport, it was found that the confidence intervals were inaccurate for very low indices due to overestimation of the lowest Sobol' indices.

The purpose of this work is to present a Global Sensitivity Analysis based model order reduction approach, which uses

- A GP surrogate.

- Semi-analytic Sobol' indices for the surrogate.
- Optimal dimension reduction (essentially locating the active subspace) by Sobol' index.

This paper is organised as follows: Section 2 presents a review of the concepts and measures that are used in this work. Section 3 describes the approach developed, including the details of the calculation of the Sobol' indices and of the optimization of the basis of the optimal low dimensional subspace. Section 4 presents the application of the method to a variety of test problems to assess its performance, while Section 5 summarises our findings and directions for future work.

2. Review of Gaussian Processes and Global Sensitivity Analysis

2.1. Gaussian Process Surrogate

In order to avoid the difficulty and expense in obtaining and analyzing response data from a computationally heavy model we adopt a Gaussian Process (GP) surrogate or emulator. The response $y(\mathbf{x})$ to arbitrarily fixed input is modelled as the sum $f(\mathbf{x})+e(\mathbf{x})$ of two Gaussian random variables encapsulating coherent signal and incoherent noise. The latter is characterized by a zero-mean distribution that is independent of the input:

$$e(\mathbf{x}) \sim \mathsf{N}\left[0, \sigma_{\mathbf{e}}^2\right]$$

The signal $f(\mathbf{x})$ is characterized by its covariance kernel $\sigma_{\mathbf{f}}^2 k(\mathbf{x}_n, \mathbf{x})$ which measures the similarity between inputs $\mathbf{x}_n$ and $\mathbf{x}$, and propagates any similarity to $y(\mathbf{x}_n)$ and $y(\mathbf{x})$. In the majority of applications, the kernel is naturally stationary, a function of $(\mathbf{x}-\mathbf{x}_n)$ alone. We shall further assume that the kernel is twice differentiable at its maximum $(\mathbf{x}=\mathbf{x}_n)$. Hence, the Hessian at the maximum must be symmetric negative semi-definite and therefore diagonalizes to

$$\partial_{\mathbf{x}\mathbf{x}} \log k(\mathbf{x}, \mathbf{x}) =: -\Theta^{\intercal}\Lambda^{-2}\Theta$$

When $|\mathbf{x}-\mathbf{x}_n|$ is large the kernel value is miniscule in any any relevant direction. The kernel details are therefore largely irrelevant to the response *any* time $|\mathbf{x}-\mathbf{x}_n|$ is large, advocating (if not justifying) the Taylor approximation

$$k(\mathbf{x}_n, \mathbf{x}) = \exp\left(-\frac{(\mathbf{x}-\mathbf{x}_n)^{\intercal}\Theta^{\intercal}\Lambda^{-2}\Theta(\mathbf{x}-\mathbf{x}_n)}{2}\left(1+O(|\mathbf{x}-\mathbf{x}_n|)\right)\right)$$

The differentiability we have imposed forces the power spectrum of the signal f to decay rapidly. Modes of response oscillating rapidly with $\mathbf{x}$ are interpreted as noise by the GP, as the kernel smoothes y into f. Such regularization is often, but not always, desirable, to avoid wildly unreliable interpolation of an overfit regression.

2.2. Kernel Optimization

In order to deal with the curse of dimensionality we propose to find orthogonal rotation matrix Θ and diagonal length-scale matrix Λ which best fit observed responses $y(\mathbf{X}^\intercal)$. The largest lengthscales in Λ mark the least relevant directions that can be ignored. However, the best fit must optimize $M(M+1)/2+2$ hyperparameters simultaneously to determine $\Theta, \Lambda, \sigma_{\mathbf{f}}^2$ and $\sigma_{\mathbf{e}}^2$. As a direct optimization of such a large problem may result in obtaining only local optima. Exploratory grid search is astronomically expensive $O(\exp(M(M+1)/2)$, likewise any random sampling which is not hopelessly sparse. Perhaps for these reasons, Θ has always been fixed as identity in the literature. The lengthscales comprising Λ are also usually identical, furnishing a radial basis function (RBF) kernel [1]. The few studies where Λ is not identical speak of an automatic relevance determination (ARD) kernel, with model order reduction in mind [17, 18].

2.3. Global Sensitivity Analysis

This paper proposes to achieve kernel optimization indirectly, via global sensitivity analysis (GSA). The surrogate expectation

$$\mathsf{E}_{\Omega}[y(\mathbf{x})] = \mathsf{E}_{\Omega}[f(\mathbf{x})] =: \bar{f}(\mathbf{x})$$

has a variation (over $\mathbf{x} \in \mathbb{R}^M$) which can be apportioned by Sobol' index

$$S_{\mathbf{m}}((\Theta)_{\mathbf{m}\times\mathbf{M}}) := \mathsf{Var}_{\mathbf{x}}\left[\mathsf{E}_{\mathbf{x}}\left[\bar{f}(\mathbf{x})|(\Theta\mathbf{x})_{\mathbf{m}}\right]\right] / \mathsf{Var}_{\mathbf{x}}\left[\bar{f}(\mathbf{x})\right] \leq 1$$

to subspaces $(\Theta\mathbf{x})_{\mathbf{m}}$ of dimension $m \leq M$. These may be calculated analytically for the exponential quadratic kernel used here. To cure to the curse of dimensionality is to find $(\Theta)_{\mathbf{m}\times\mathbf{M}}$ such that $S_{\mathbf{m}} \approx 1$ for $m \ll M$. The rotation sub-matrix $(\Theta)_{\mathbf{m}\times\mathbf{M}}$ has a manageable number of elements if m is small. This paper takes the most economical approach, maximizing $S_{\mathbf{m}}$ for $m = 1, \ldots, M$ in turn, to find the most relevant direction, then the second most relevant, and so on.

3. Methodology

Let $\mathbf{X}$ be the $(N \times M)$ design matrix of observed inputs eliciting the N response $y(\mathbf{X}^\intercal)$. The observations are standardized such that

$$(\mathbf{0})_{\mathbf{M}} = \mathsf{E}[\mathbf{x}_n] := \sum_{n=1}^{N} (\mathbf{X})^\intercal_{n\times\mathbf{M}} \quad ; \quad 1 = \mathsf{Var}[\mathbf{x}_n] = N^{-1} \sum_{n=1}^{N} (\mathbf{X})_{n\times\mathbf{M}} (\mathbf{X})^\intercal_{n\times\mathbf{M}}$$
$$0 = \mathsf{E}[y(\mathbf{x}_n)] := \sum_{n=1}^{N} (y(\mathbf{X}^\intercal))_n \quad ; \quad 1 = \mathsf{Var}[y(\mathbf{x}_n)] = N^{-1} y(\mathbf{X}^\intercal)^\intercal y(\mathbf{X}^\intercal)$$

where boldface subscripts refer to the multi-indices

$$\emptyset =: \mathbf{0} \subseteq \mathbf{m} := (1, \ldots, m) \subseteq \mathbf{M} \tag{1}$$

which always precede superscript operations (such as transposition or inversion). For brevity, we shall admit vector Gaussian probability densities $p((\mathbf{z})_{\mathbf{m}}; (\mathbf{Z})_{\mathbf{m}\times\mathbf{J}}, \Sigma_{\mathbf{z}})$ such that

$$\begin{aligned} &(p((\mathbf{z})_{\mathbf{m}}; (\mathbf{Z})_{\mathbf{m}\times\mathbf{J}}, \Sigma_{\mathbf{z}}))_j \\ &\qquad := (2\pi)^{-M/2} \left|\Sigma_{\mathbf{z}}\right|^{-1/2} \exp\left(-\frac{(\mathbf{z} - (\mathbf{Z})_{\mathbf{m}\times j})^\intercal \Sigma_{\mathbf{z}}^{-1} (\mathbf{z} - (\mathbf{Z})_{\mathbf{m}\times j})}{2} \right) \end{aligned} \tag{2}$$

naturally collapsing to the (scalar) normal multivariate density when $J = 1$.

3.1. Gaussian Process Surrogate

Non-parametric GP regression fits signal f and noise e Gaussian processes to

$$y(\mathbf{X}^\intercal) = f(\mathbf{X}^\intercal) + e(\mathbf{X}^\intercal) \tag{3}$$

This work exclusively employs objective Bayesian priors

$$\begin{aligned} f(\mathbf{X}^\intercal) &\sim \mathsf{N}\left[(\mathbf{0})_{\mathbf{N}}, \sigma_{\mathbf{f}}^2 k(\mathbf{X}^\intercal, \mathbf{X}^\intercal)\right] \\ e(\mathbf{X}^\intercal) &\sim \mathsf{N}\left[(\mathbf{0})_{\mathbf{N}}, \sigma_{\mathbf{e}}^2 (\mathbf{I})_{\mathbf{N}\times\mathbf{N}}\right] \end{aligned}$$

built on an ARD kernel

$$k(\mathbf{x}_n, \mathbf{x}) := (2\pi)^{M/2} \left|\Lambda\right| p\left(\mathbf{x}; \mathbf{x}_n, \Lambda^2\right) \tag{4}$$

with diagonal positive definite lengthscale matrix Λ. Bayesian conditioning ultimately furnishes the predictive process

$$y(\mathbf{x}) \sim \mathsf{N}\left[\bar{f}(\mathbf{x}), \Sigma_{\mathbf{f}}(\mathbf{x}) + \sigma_{\mathbf{e}}^2\right]$$

with signal mean and variance

$$\begin{aligned} \bar{f}(\mathbf{x}) &:= \sigma_{\mathbf{f}}^2 k(\mathbf{x}, \mathbf{X}^\intercal)\mathbf{K}^{-1} y(\mathbf{X}^\intercal) \\ \Sigma_{\mathbf{f}}(\mathbf{x}) &:= \sigma_{\mathbf{f}}^2 k(\mathbf{x}, \mathbf{x}) - \sigma_{\mathbf{f}}^2 k(\mathbf{x}, \mathbf{X}^\intercal)\mathbf{K}^{-1}\sigma_{\mathbf{f}}^2 k(\mathbf{X}^\intercal, \mathbf{x}) \end{aligned} \tag{5}$$

where

$$\mathbf{K} := \sigma_{\mathbf{f}}^2 k(\mathbf{X}^\intercal, \mathbf{X}^\intercal) + \sigma_{\mathbf{e}}^2 (\mathbf{I})_{\mathbf{N}\times\mathbf{N}} \tag{6}$$

The $M+2$ hyperparameters constituting $\Lambda, \sigma_{\mathbf{f}}$ and $\sigma_{\mathbf{e}}$ are simultaneously optimized for maximium marginal likelihood $\mathsf{p}[y|\mathbf{X}^\intercal]$, using the GPy software library (refeence).

3.2. Global Sensitivity Analysis

Imagine a sample datum $\mathbf{u}$ is drawn from a standardized normal test distribution

$$\mathbf{u} \sim \mathsf{N}[(\mathbf{0})_{\mathbf{M}}, (\mathbf{I})_{\mathbf{M}\times\mathbf{M}}] \tag{7}$$

The datum basis is rotated to

$$\mathbf{x} =: \Theta^\intercal \mathbf{u} \tag{8}$$

eliciting the conditional surrogate responses

$$f_{\mathbf{m}}((\mathbf{u})_{\mathbf{m}}) := \mathsf{E}\left[\bar{f}(\Theta^\intercal \mathbf{u}) | (\mathbf{u})_{\mathbf{m}}\right] \tag{9}$$

Knowledge of $\mathbf{u}$ herein ranges from totally conditional $f_{\mathbf{M}}(\mathbf{u}) = \bar{f}(\mathbf{x})$ to unconditional ignorance $f_{\mathbf{0}} = \mathsf{E}\left[\bar{f}(\mathbf{x})\right]$. Equations (4) to (7) enable analytic integration yielding

$$f_{\mathbf{m}}((\mathbf{u})_{\mathbf{m}}) = \tilde{\mathbf{f}}^\intercal \, \frac{p\big((\mathbf{u})_{\mathbf{m}}; (\mathbf{T})_{\mathbf{N}\times\mathbf{m}}^\intercal, (\Sigma)_{\mathbf{m}\times\mathbf{m}}\big)}{p((\mathbf{u})_{\mathbf{m}}; (\mathbf{0})_{\mathbf{m}}, (\mathbf{I})_{\mathbf{m}\times\mathbf{m}})} \tag{10}$$

where $\tilde{\mathbf{f}}$ is the Hadamard (element-wise) product $\circ$ of two vectors

$$\tilde{\mathbf{f}} := (2\pi)^{M/2} \, |\Lambda| \, p\big(\mathbf{0}; \mathbf{X}^\intercal, \Lambda^2 + \mathbf{I}\big) \circ \big(\sigma_{\mathbf{f}}^2 \mathbf{K}^{-1} y(\mathbf{X}^\intercal)\big) \tag{11}$$

and

$$\mathbf{T} := \mathbf{X}\left(\Lambda^2 + \mathbf{I}\right)^{-1}\Theta^{\intercal} \tag{12}$$

$$\Sigma := \Theta\left(\Lambda^{-2} + \mathbf{I}\right)^{-1}\Theta^{\intercal} \tag{13}$$

According to these formulae, the unconditional surrogate response is

$$f_{\mathbf{0}} = \mathsf{E}\left[\bar{f}(\mathbf{x})\right] = \tilde{\mathbf{f}}^{\intercal}(\mathbf{1})_{\mathbf{N}} \tag{14}$$

which does not depend on Θ of course. Standardization of $y(\mathbf{X}^{\intercal})$ instills an expectation of precisely zero here if $\mathbf{x}_n \sim \mathsf{N}[(\mathbf{0})_{\mathbf{M}}, (\mathbf{I})_{\mathbf{M}\times\mathbf{M}}]$ (which is often not exactly true).

Conditional variances may now be calculated as

$$D_{\mathbf{m}}((\Theta)_{\mathbf{m}\times\mathbf{M}}) := \mathsf{Var}[f_{\mathbf{m}}((\mathbf{u})_{\mathbf{m}})] = \frac{\tilde{\mathbf{f}}^{\intercal}\,\mathbf{W}_{\mathbf{m}}\,\tilde{\mathbf{f}}}{\left|2(\Sigma)_{\mathbf{m}\times\mathbf{m}} - (\Sigma)^2_{\mathbf{m}\times\mathbf{m}}\right|^{1/2}} - f_{\mathbf{0}}^2 \tag{15}$$

where

$$\begin{aligned}(\mathbf{W}_{\mathbf{m}})_{n\times o} := {} & \exp\left(\frac{-(\mathbf{T})_{n\times\mathbf{m}}(\Sigma)^{-1}_{\mathbf{m}\times\mathbf{m}}(\mathbf{T})^{\intercal}_{n\times\mathbf{m}} - (\mathbf{T})_{o\times\mathbf{m}}(\Sigma)^{-1}_{\mathbf{m}\times\mathbf{m}}(\mathbf{T})^{\intercal}_{o\times\mathbf{m}}}{2}\right) \\ & \times \exp\left(\frac{+\left((\mathbf{T})_{n\times\mathbf{m}} + (\mathbf{T})_{o\times\mathbf{m}}\right)(\Psi)^{-1}_{\mathbf{m}\times\mathbf{m}}(\Sigma)^{-1}_{\mathbf{m}\times\mathbf{m}}\left((\mathbf{T})^{\intercal}_{n\times\mathbf{m}} + (\mathbf{T})^{\intercal}_{o\times\mathbf{m}}\right)}{2}\right)\end{aligned} \tag{16}$$

and

$$\Psi := \Theta\left(\Lambda^{-2} + \mathbf{I}\right)^{-1}\left(2\Lambda^{-2} + \mathbf{I}\right)\Theta^{\intercal} \tag{17}$$

The proportion of response variance ascribable to the first m basis directions of $\mathbf{u}$ is given by the Sobol' index

$$S_{\mathbf{m}}((\Theta)_{\mathbf{m}\times\mathbf{M}}) := D_{\mathbf{m}}((\Theta)_{\mathbf{m}\times\mathbf{M}})/D_{\mathbf{M}}(\Theta) \leq S_{\mathbf{M}}(\Theta) = 1 \tag{18}$$

Analytic expressions for $\partial_\Theta D_{\mathbf{m}}((\Theta)_{\mathbf{m}\times\mathbf{M}})$ have been obtained from Eq. (15) using standard formulae for differentating matrix inverses and determinants. As $D_{\mathbf{m}}$ projects M-dimensional $(\mathbf{x})_{\mathbf{M}}$ onto m-dimensional $(\mathbf{u})_{\mathbf{M}}$, the result is affected by just a few components of rotation:

$$(\partial_\Theta D_{\mathbf{m}}((\Theta)_{\mathbf{m}\times\mathbf{M}}))_{i\times j} \neq 0 \quad \Longrightarrow \quad i \leq m < M \tag{19}$$

In particular $D_{\mathbf{M}}(\Theta) = D_{\mathbf{M}}$ and $S_{\mathbf{M}}(\Theta) = 1$ are independent of Θ, as there is no projection, only rotation, in transforming $(\mathbf{x})_{\mathbf{M}}$ into $(\mathbf{u})_{\mathbf{M}}$.

3.3. Basis Optimization

At this point in the analysis, everything has been fixed save the rotation

$$\mathbf{u} := \Theta\mathbf{x} \tag{20}$$

relating sampling distribution $\mathbf{u} \sim \mathsf{N}[(\mathbf{0})_{\mathbf{M}}, (\mathbf{I})_{\mathbf{M}\times\mathbf{M}}]$ to the input of the surrogate response $\bar{f}(\mathbf{x})$. This rotation will now be determined by maximizing the relevance – as measured by Sobol' index – of each $\mathbf{u}$-direction in turn. This means optimizing Θ in Eq. (20) row by row from top to bottom.

Row orthonormality leaves just $(M-m-1)$ elements free in row m, which we encode as

$$(\Theta)_{\mathbf{m}\times\mathbf{M}} =: (\Xi)_{\mathbf{m}\times\mathbf{M}}\tilde{\Theta} \tag{21}$$

where Ξ is orthogonal, and identical on the $(m-1)$ rows already optimized

$$\begin{aligned}
(\Xi)_{\mathbf{m}\setminus\{m\}\times\mathbf{M}} &= \mathbf{I}_{\mathbf{m}\setminus\{m\}\times\mathbf{M}} \\
(\Xi)_{m\times\mathbf{m}\setminus\{m\}} &= (\mathbf{0})_{1\times\mathbf{m}\setminus\{m\}} \\
(\Xi)_{m\times m} &= \left(1 - \sum_{k=m+1}^{M} (\Xi)^2_{m\times k}\right)^{1/2}
\end{aligned} \tag{22}$$

The last line induces a derivative adjustment

$$\frac{\partial}{\partial\,(\Xi)_{m\times k}} = \frac{\partial}{\partial\,(\Xi)_{m\times k}} - \frac{(\Xi)_{m\times k}}{(\Xi)_{m\times m}}\frac{\partial}{\partial\,(\Xi)_{m\times m}} \tag{23}$$

which should be exploited by the optimizer as a powerful repellant to orthonormality violations. This work uses a BFGS optimizer, fed an analytic Jacobian.

Given these constraints, row m is optimally determined by

$$(\Xi)_{m\times\mathbf{M}\setminus\mathbf{m}} = \underset{(\Xi)_{m\times\mathbf{M}\setminus\mathbf{m}}}{\operatorname{argmax}}\ S_{\mathbf{m}}((\Theta)_{\mathbf{m}\times\mathbf{M}}) = \underset{(\Xi)_{m\times\mathbf{M}\setminus\mathbf{m}}}{\operatorname{argmax}}\ D_{\mathbf{m}}((\Theta)_{\mathbf{m}\times\mathbf{M}}) \tag{24}$$

The optimal row m is then incorporated in $\tilde{\Theta}$ and Ξ according to

$$\begin{gathered}
\tilde{\Theta} = \mathbf{Q}^{\intercal} \text{ where } \Theta^{\intercal} = \mathbf{QR} \text{ is the QR factorization of the update} \\
(\Xi)_{\mathbf{m}\times\mathbf{M}} = \mathbf{I}_{\mathbf{m}\times\mathbf{M}}
\end{gathered} \tag{25}$$

ready to optimize row $m+1$. Optimization followed by incorporation is performed for $m = 1, \ldots, M-1$ in turn to entirely optimize Θ. The later rows could be left unoptimized, though they are successively cheaper to obtain.

Algorithm 1 Summary of the basis optimization algorithm.

1: **repeat**
2: Fit GP surrogate to $y(\mathbf{X}^\intercal)$, determining $\bar{f}(\mathbf{x})$ according to Section 3.1
3: Set $\tilde{\Theta} \leftarrow \Theta \leftarrow \Theta_\Pi \leftarrow \mathbf{I}$
4: **for** $m = 1$ **to** M **do**
5: According to Section 3.2, optimize

$$(\Xi)_{m\times\mathbf{M}\backslash\mathbf{m}} \leftarrow \underset{(\Xi)_{m\times\mathbf{M}\backslash\mathbf{m}}}{\operatorname{argmax}} D_{\mathbf{m}}((\Theta)_{\mathbf{m}\times\mathbf{M}})$$

where $(\Theta)_{\mathbf{m}\times\mathbf{M}} =: (\Xi)_{\mathbf{m}\times\mathbf{M}}\tilde{\Theta}$
6: Update $\tilde{\Theta} \leftarrow \mathbf{Q}^\intercal$ where $\Theta^\intercal = \mathbf{QR}$
7: **end for**
8: Update the input basis to $\mathbf{X}^\intercal \leftarrow \Theta\mathbf{X}^\intercal$
9: Update the overall rotation to $\Theta_\Pi \leftarrow \Theta\Theta_\Pi$
10: **until** $\Theta \approx \mathbf{I}$

3.4. Summary

The main loop of the algorithm is described i:

During testing the optimization in Step 5 is found to be prone to converge to local optima, especially in the first iteration or two of the outermost loop. An approach was therefore developed that the early iterations explore the behaviour of $(\Xi)_{m\times\mathbf{M}\backslash\mathbf{m}}$ by grid or randomized search, before attempting exploitation by gradient descent).

As the input basis is updated at each step, $\mathbf{X} = \mathbf{U}$ ultimately. The key output of the algorithm is the overall rotation Θ_Π of the original basis for $\mathbf{x}$ to the optimal basis for $\mathbf{u}$.

4. Results

In this section, the method described in Section 3.4 is applied to a series of test functions to evaluate its performance. Each function takes $N \in 100, 200, 400, 800, 1600$ data from a latin hypercube of $M = 5$ input dimensions. All inputs and outputs are standardized to mean 0, standard deviation 1 before folding. All results are calculated as the mean over two folds (each with N training data and N test data, so predictions are rigorously cross-validated.

In each case an $N \times M$ design matrix $\mathbf{X}$ is sampled from a standard normal distribution (latin hypercube). The input to the test function $f\colon [x_-, x_+]^M \to \mathbb{R}$ is generally constructed as

$$\hat{\mathbf{X}}^\intercal = (x_+ - x_-)c(\Phi \mathbf{X}^\intercal) + x_-(\mathbf{1})_{\mathbf{M}\times\mathbf{N}} \tag{26}$$

where $c\colon \mathbb{R}^M \to \mathbb{R}^M$ is the cumulative density function for M independent standard normal random variables, and Φ is a test rotation matrix. The corresponding optimal input rotation from Section 3.4 is

$$\Theta_\Pi = \begin{cases} \Theta_{\mathbf{1}} & \text{if } \Phi \text{ is identity matrix } \mathbf{1} \\ \Theta_{\mathbf{R}} & \text{if } \Phi \text{ is a random rotation matrix } \Phi_{\mathbf{R}} \end{cases} \tag{27}$$

which should recover the random rotation as

$$\Theta_{\mathbf{R}} \cong \Theta_{\mathbf{1}} \Phi_{\mathbf{R}} \tag{28}$$

However, this is congruence, not equality: different rotations might locate (excactly or nearly exactly) the same active subspace.

For each function, the intial GP fit is assessed by test statistics from independent data (from the other fold), together with errors in the calculated Sobol' indices. The latter are important as they at the heart of subsequent calculations. The input basis is then optimized, calculating $\Theta_{\mathbf{1}}$. A reduced dimensionality $\underline{M}$ for the optimized basis is determined as

$$\min \left\{ \underline{M} \leq M \;\middle|\; S_{\underline{\mathbf{M}}} \geq 0.90 \right\} \tag{29}$$

A GP is fit to this reduced input, and its test statistics compared with the initial GP.

The whole procedure is then repeated (with entirely fresh data) to which a random input rotation $\Phi_{\mathbf{R}}$ is applied. The input basis is optimized, calculating $\Theta_{\mathbf{R}}$, whereas the reduced dimensionality $\underline{M}$ is not re-assessed, but retained from the unrotated analysis.

Finally the rotated and unrotated active subspaces are compared for congruence, using the ordered sigular values $\Sigma_m(\mathbf{u}^\dagger)$ of

$$\mathbf{u}^\dagger = (\Theta_{\mathbf{1}} \Phi_{\mathbf{R}} \Theta_{\mathbf{R}}^\intercal)_{\underline{\mathbf{M}}\times\underline{\mathbf{M}}} \tag{30}$$

This matrix transforms the active subspace according to $\Theta_{\mathbf{R}}$ into the active subspace according to $\Theta_{\mathbf{1}}\Phi_{\mathbf{R}}$ without straying outside the union of two. The basis vector length(s) lost to the inactive subspaces in doing this is $(\mathbf{1} - \Sigma_m(\mathbf{u}^\dagger))$.

4.1. Sine Function

$$f(\hat{\mathbf{x}}) := \sin(\hat{\mathbf{x}}_1) \tag{31}$$

$$[x_-, x_+] := [-\pi, +\pi]$$
$$S_{\mathbf{1}} = 1$$

Fitting an initial GP recovers the exact Sobol' indices to within an accuracy of 0.005 (precision actually decreasing with the number of data). Optimizing the input basis improves this to 10^{-8}, which can only be due to repeating GP regression as 4 iterations were run to optimize Θ, even though convergence is immediate. The predictive improvement is shown in Table 1.

Noise	N	RMSE (%)		$\sigma_{f(\mathbf{x})}$ (%)		Outliers (%)	
0.000	100	0.05	0.06	0.54	0.54	0.00	0.00
0.025	100	1.06	8.36	4.24	30.53	6.00	18.00
0.000	200	0.01	0.01	0.38	0.38	0.00	0.00
0.025	200	0.96	8.26	3.80	34.70	4.75	12.75
0.000	400	2.28	25.08	0.76	99.59	0.00	0.12
0.025	400	0.91	10.38	3.62	42.23	4.75	6.25
0.000	800	0.63	24.98	0.45	99.86	0.00	0.00
0.025	800	1.40	18.19	3.75	72.83	4.75	4.19
0.000	1600	0.00	0.00	0.32	0.32	0.00	0.00
0.025	1600	0.89	20.52	3.57	83.14	4.41	1.75

Table 1: Predictive performance of initial GPs (left sub-columns) after reducing dimensionality (right sub-columns), for the sine function. Three measures are shown: the Root Mean Square Error, the GPs' predictive standard deviation $\sigma_{f(\mathbf{x})}$, and the percentage of observations outside $\pm 2\sigma_{f(\mathbf{x})}$.

Applying a random rotation $\Phi_{\mathbf{R}}$, the exact Sobol' indices are recovered to within 10^{-8} after 3 iterations. The predictive performance is shown in Table 2 , the 1D active subspace measures in Table 3.

Noise	N	RMSE (%)		$\sigma_{f(\mathbf{x})}$ (%)		Outliers (%)	
0.000	100	5.01	0.05	16.68	0.42	9.50	0.00
0.025	100	8.92	20.58	32.48	84.69	6.50	2.50
0.000	200	6.77	0.07	26.45	0.45	5.50	0.00
0.025	200	7.63	15.51	31.56	58.88	3.75	9.75
0.000	400	5.17	0.02	18.71	0.36	5.25	0.00
0.025	400	2.70	20.41	10.30	78.87	4.62	3.50
0.000	800	1.93	0.00	5.51	0.33	4.31	0.00
0.025	800	1.78	15.95	5.45	61.67	4.88	5.81
0.000	1600	2.45	0.00	6.97	0.32	4.59	0.00
0.025	1600	1.40	24.99	5.67	99.97	4.56	0.00

Table 2: Predictive performance of initial GPs (left sub-columns) after reducing dimensionality (right sub-columns), for the sine function with randomly rotated inputs. Three measures are shown: the Root Mean Square Error, the GPs' predictive standard deviation $\sigma_{f(\mathbf{x})}$, and the percentage of observations outside $\pm 2\sigma_{f(\mathbf{x})}$.

Noise	N	$\Sigma_1(\mathbf{u}^\dagger)$
0.0000	100	0.9993
0.0250	100	0.9995
0.0000	200	0.9998
0.0250	200	0.9996
0.0000	400	0.3534
0.0250	400	1.0000
0.0000	800	0.2693
0.0250	800	0.9998
0.0000	1600	0.9999
0.0250	1600	1.0000

Table 3: The active subspace measures $\Sigma_m(\mathbf{u}^\dagger)$ for the sine function, comparing optimization of unrotated and randomly rotated inputs.

4.2. Decoupled Ishigami Function

$$f(\mathbf{x}) := \left(1 + b\mathbf{x}_3^4\right)\sin(\mathbf{x}_1) + a\sin^2(\mathbf{x}_2) \tag{32}$$

$$a = 2.0 \quad ; \quad b = 0$$
$$S_{\mathbf{1}} = 0.5 \quad ; \quad S_{\mathbf{2}} = 1$$

Fitting an initial GP recovers the exact Sobol' indices to within an accuracy of 0.03, which is barely changed on optimizing the input basis. The predictive performance is shown in Table 4.

Noise	N	RMSE (%)		$\sigma_{f(\mathbf{x})}$ (%)		Outliers (%)	
0.000	100	3.55	3.54	7.46	7.46	0.50	0.50
0.025	100	2.82	7.12	8.78	20.27	2.50	19.00
0.000	200	1.65	1.64	3.24	3.26	0.25	0.25
0.025	200	3.50	8.55	7.09	21.56	4.00	19.75
0.000	400	0.93	0.94	1.68	1.68	0.00	0.00
0.025	400	1.53	8.93	4.27	25.04	4.25	17.00
0.000	800	1.16	1.16	0.87	0.87	0.06	0.06
0.025	800	1.46	16.98	3.32	65.07	5.44	6.94
0.000	1600	5.63	12.50	3.92	50.22	0.19	0.09
0.025	1600	1.07	7.95	2.99	24.32	4.59	13.41

Table 4: Predictive performance of initial GPs (left sub-columns) after reducing dimensionality (right sub-columns), for the decoupled Ishigami function. Three measures are shown: the Root Mean Square Error, the GPs' predictive standard deviation $\sigma_{f(\mathbf{x})}$, and the percentage of observations outside $\pm 2\sigma_{f(\mathbf{x})}$.

Applying a random rotation $\Phi_{\mathbf{R}}$, ????. The predictive performance is shown in Table 5 , the 1D active subspace measures are in Table 6.

Noise	N	RMSE (%)		$\sigma_{f(\mathbf{x})}$ (%)		Outliers (%)	
0.000	100	21.10	23.64	73.01	91.81	8.00	3.50
0.025	100	19.04	24.47	76.11	96.51	1.50	2.50
0.000	200	17.85	23.61	73.29	94.06	1.00	3.50
0.025	200	14.22	24.67	51.92	95.72	7.25	4.25
0.000	400	15.55	24.39	62.98	96.83	3.62	2.25
0.025	400	12.25	24.80	48.85	99.05	2.62	1.25
0.000	800	11.46	24.37	42.67	96.18	3.94	2.12
0.025	800	11.87	24.44	46.79	97.10	3.75	2.50
0.000	1600	8.51	24.94	29.81	99.63	4.28	0.25
0.025	1600	7.38	24.51	24.82	97.18	4.09	2.09

Table 5: Predictive performance of initial GPs (left sub-columns) after reducing dimensionality (right sub-columns), for the decoupled Ishigami function with randomly rotated inputs. Three measures are shown: the Root Mean Square Error, the GPs' predictive standard deviation $\sigma_{f(\mathbf{x})}$, and the percentage of observations outside $\pm 2\sigma_{f(\mathbf{x})}$.

Noise	N	$\Sigma_1(\mathbf{u}^\dagger)$	$\Sigma_2(\mathbf{u}^\dagger)$
0.0000	100	0.8081	0.9964
0.0250	100	0.3714	0.8405
0.0000	200	0.5456	0.9963
0.0250	200	0.8322	0.9995
0.0000	400	0.5971	0.9997
0.0250	400	0.7286	0.9982
0.0000	800	0.6712	0.9994
0.0250	800	0.6624	0.9917
0.0000	1600	0.5741	0.9012
0.0250	1600	0.5650	0.9999

Table 6: The active subspace measures $\Sigma_m(\mathbf{u}^\dagger)$ for the decoupled Ishigami function, comparing optimization of unrotated and randomly rotated inputs.

4.3. Ishigami Function

$$f(\mathbf{x}) := \left(1 + b\mathbf{x}_3^4\right)\sin(\mathbf{x}_1) + a\sin^2(\mathbf{x}_2) \tag{33}$$

$$a = 7.0 \quad ; \quad b = 0.1$$
$$S_{\mathbf{1}} = 0.3139 \quad ; \quad S_{\mathbf{2}} = 0.7563 \quad ; \quad S_{\mathbf{3}} = 1$$

Fitting an initial GP recovers the exact Sobol' indices to within an accuracy of 0.04 with $N = 100$ training data, gradually improving to 0.002 with $N = 1600$, which does not improve on optimizing the input basis. The predictive performance is shown in Table 7.

Noise	N	RMSE (%)		$\sigma_{f(\mathbf{x})}$ (%)		Outliers (%)	
0.000	100	11.17	17.46	32.25	50.84	18.00	22.50
0.025	100	10.80	11.06	32.58	34.74	10.50	12.00
0.000	200	7.88	7.52	18.18	17.22	9.75	6.25
0.025	200	6.88	12.33	19.10	36.24	4.00	8.25
0.000	400	3.95	3.99	8.43	8.92	5.50	5.12
0.025	400	4.00	6.79	7.67	19.15	7.50	6.50
0.000	800	2.48	13.51	3.90	50.87	1.06	3.12
0.025	800	2.25	18.43	3.70	69.04	9.75	7.44
0.000	1600	1.60	1.65	1.77	2.22	0.88	0.72
0.025	1600	4.42	20.80	13.36	80.82	5.62	7.09

Table 7: Predictive performance of initial GPs (left sub-columns) after reducing dimensionality (right sub-columns), for the Ishigami function. Three measures are shown: the Root Mean Square Error, the GPs' predictive standard deviation $\sigma_{f(\mathbf{x})}$, and the percentage of observations outside $\pm 2\sigma_{f(\mathbf{x})}$.

Applying a random rotation $\Phi_{\mathbf{R}}$, the exact Sobol' indices are recovered to within 10^{-8} after 3 iterations. The predictive improvement is shown in Table 8 , the 1D active subspace measures are in Table 9.

Noise	N	RMSE (%)		$\sigma_{f(\mathbf{x})}$ (%)		Outliers (%)	
0.000	100	24.40	25.77	83.58	85.12	8.00	15.00
0.025	100	24.41	25.23	80.42	89.14	10.00	8.00
0.000	200	20.03	23.56	63.86	94.80	8.00	7.00
0.025	200	18.49	24.21	59.21	92.08	7.50	6.75
0.000	400	11.99	21.30	40.48	78.07	4.62	8.88
0.025	400	17.77	25.23	65.35	95.37	7.88	6.75
0.000	800	13.33	21.25	45.48	87.14	5.38	6.88
0.025	800	9.10	19.14	30.41	65.32	6.50	10.75
0.000	1600	4.92	18.47	11.63	62.25	5.34	10.53
0.025	1600	8.20	17.80	24.87	63.43	6.47	9.50

Table 8: Predictive performance of initial GPs (left sub-columns) after reducing dimensionality (right sub-columns), for the Ishigami function with randomly rotated inputs. Three measures are shown: the Root Mean Square Error, the GPs' predictive standard deviation $\sigma_{f(\mathbf{x})}$, and the percentage of observations outside $\pm 2\sigma_{f(\mathbf{x})}$.

Noise	N	$\Sigma_1(\mathbf{u}^\dagger)$	$\Sigma_2(\mathbf{u}^\dagger)$	$\Sigma_3(\mathbf{u}^\dagger)$
0.0000	100	0.7008	0.9881	1.0000
0.0250	100	0.4051	0.9722	1.0000
0.0000	200	0.1234	0.8905	1.0000
0.0250	200	0.7299	0.9534	1.0000
0.0000	400	0.4590	0.9630	1.0000
0.0250	400	0.5723	0.9156	1.0000
0.0000	800	0.2627	0.7233	1.0000
0.0250	800	0.4772	0.8070	1.0000
0.0000	1600	0.9264	0.9989	1.0000
0.0250	1600	0.8921	0.9895	1.0000

Table 9: The active subspace measures $\Sigma_m(\mathbf{u}^\dagger)$ for the Ishigami function, comparing optimization of unrotated and randomly rotated inputs.

4.4. Sobol' G Function

$$f(\mathbf{x}) := \prod_{i=1}^{D} \frac{|4\mathbf{x}_i - 2| + \mathbf{a}_i}{1 + \mathbf{a}_i} \tag{34}$$

$$\mathbf{a}_i = (i-1)/2$$

$$S_{\mathbf{1}} = 0.4107 \quad ; \quad S_{\mathbf{2}} = 0.6541 \quad ; \quad S_{\mathbf{3}} = 0.8113 \quad ; \quad S_{\mathbf{4}} = 0.9203 \quad ; \quad S_{\mathbf{5}} = 1$$

Fitting an initial GP recovers the exact Sobol' indices to within an accuracy of 0.03, which does not improve on optimizing the input basis. The predictive improvement is shown in Table 10.

Noise	N	RMSE (%)		$\sigma_{f(\mathbf{x})}$ (%)		Outliers (%)	
0.000	100	25.25	23.21	42.87	45.23	26.00	26.50
0.025	100	18.06	20.68	35.80	38.84	14.50	25.00
0.000	200	17.66	19.21	38.06	42.82	16.50	18.25
0.025	200	14.84	18.04	37.56	47.37	14.00	13.75
0.000	400	13.86	18.28	32.94	45.69	12.50	13.50
0.025	400	13.54	17.44	32.35	44.40	10.75	10.50
0.000	800	11.60	17.94	27.53	44.68	8.56	13.50
0.025	800	11.47	17.09	27.66	43.42	9.88	12.56
0.000	1600	9.93	16.71	21.87	41.78	7.72	13.19
0.025	1600	9.95	16.65	22.66	41.07	8.41	14.50

Table 10: Predictive performance of initial GPs (left sub-columns) after reducing dimensionality (right sub-columns), for the sobol' G function. Three measures are shown: the Root Mean Square Error, the GPs' predictive standard deviation $\sigma_{f(\mathbf{x})}$, and the percentage of observations outside $\pm 2\sigma_{f(\mathbf{x})}$.

Applying a random rotation $\Phi_{\mathbf{R}}$, the exact Sobol' indices are recovered to within 10^{-8} after 3 iterations. The predictive improvement is shown in Table 11 , the 1D active subspace measures are in Table 12.

Noise	N	RMSE (%)		$\sigma_{f(\mathbf{x})}$ (%)		Outliers (%)	
0.000	100	25.87	25.85	58.02	69.09	7.50	11.00
0.025	100	23.52	21.99	42.74	48.76	20.00	16.50
0.000	200	17.99	22.31	43.16	62.20	12.50	13.50
0.025	200	19.22	23.37	47.04	66.77	11.75	10.00
0.000	400	16.77	20.24	36.79	49.04	13.50	14.75
0.025	400	15.75	21.31	37.64	61.27	11.00	9.62
0.000	800	12.46	20.20	30.03	55.50	8.31	11.94
0.025	800	11.92	19.52	28.49	53.85	9.06	11.88
0.000	1600	10.05	18.90	25.44	55.13	8.84	10.97
0.025	1600	10.43	20.37	24.10	60.01	9.19	9.81

Table 11: Predictive performance of initial GPs (left sub-columns) after reducing dimensionality (right sub-columns), for the sobol' G function with randomly rotated inputs. Three measures are shown: the Root Mean Square Error, the GPs' predictive standard deviation $\sigma_{f(\mathbf{x})}$, and the percentage of observations outside $\pm 2\sigma_{f(\mathbf{x})}$.

Noise	N	$\Sigma_1(\mathbf{u}^\dagger)$	$\Sigma_2(\mathbf{u}^\dagger)$	$\Sigma_3(\mathbf{u}^\dagger)$	$\Sigma_4(\mathbf{u}^\dagger)$
0.0000	100	0.7579	1.0000	1.0000	1.0000
0.0250	100	0.3726	1.0000	1.0000	1.0000
0.0000	200	0.5138	1.0000	1.0000	1.0000
0.0250	200	0.5053	1.0000	1.0000	1.0000
0.0000	400	0.5529	1.0000	1.0000	1.0000
0.0250	400	0.4915	1.0000	1.0000	1.0000
0.0000	800	0.3902	1.0000	1.0000	1.0000
0.0250	800	0.0213	1.0000	1.0000	1.0000
0.0000	1600	0.2511	1.0000	1.0000	1.0000
0.0250	1600	0.5513	1.0000	1.0000	1.0000

Table 12: The active subspace measures $\Sigma_m(\mathbf{u}^\dagger)$ for the Sobol' G function, comparing optimization of unrotated and randomly rotated inputs.

5. Conclusion

References

[1] J. Sacks, W. J. Welch, T. J. Mitchell, H. P. Wynn, Design and analysis of computer experiments, Statistical Science 4 (4) (1989) 409–423.
URL `http://www.jstor.org/stable/2245858`

[2] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning series), The MIT Press, 2005.

[3] R. G. Ghanem, P. D. Spanos, Spectral techniques for stochastic finite elements, Archives of Computational Methods in Engineering 4 (1) (1997) 63–100. doi:10.1007/BF02818931.
URL `https://doi.org/10.1007/BF02818931`

[4] D. Xiu, G. E. Karniadakis, The wiener–askey polynomial chaos for stochastic differential equations, SIAM Journal on Scientific Computing 24 (2) (2002) 619–644. doi:10.1137/s1064827501387826.

[5] D. Xiu, Numerical Methods for Stochastic Computations: A Spectral Method Approach, Princeton University Press, 2010.

[6] M. Chevreuil, R. Lebrun, A. Nouy, P. Rai, A least-squares method for sparse low rank approximation of multivariate functions, SIAM/ASA Journal on Uncertainty Quantification 3 (1) (2015) 897–921. arXiv:http://arxiv.org/abs/1305.0030v2, doi:10.1137/13091899X.

[7] K. Konakli, B. Sudret, Global sensitivity analysis using low-rank tensor approximations, Reliability Engineering & System Safety 156 (2016) 64–83. doi:10.1016/j.ress.2016.07.012.

[8] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297. doi:10.1007/bf00994018.

[9] P. G. Constantine, E. Dow, Q. Wang, Active subspace methods in theory and practice: Applications to kriging surfaces, SIAM Journal on Scientific Computing 36 (4) (2014) A1500–A1524. arXiv:1304.2070, doi:10.1137/130916138.

[10] X. Liu, S. Guillas, Dimension Reduction for Gaussian Process Emulation: An Application to the Influence of Bathymetry on Tsunami Heights, SIAM-ASA Journal on Uncertainty Quantification 5 (1) (2017) 787–812. doi:10.1137/16M1090648.

[11] Y. Xia, H. Tong, W. K. Li, L.-X. Zhu, An adaptive estimation of dimension reduction space, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64 (3) (2002) 363–410. doi:10.1111/1467-9868.03411.

[12] I. M. Sobol, Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates, Mathematics and Computers in Simulation 55 (2001) 271–280.

[13] J. E. Oakley, A. O'Hagan, Probabilistic sensitivity analysis of complex models: a bayesian approach, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66 (3) (2004) 751–769. doi:10.1111/j.1467-9868.2004.05304.x.

[14] R. Jin, W. Chen, A. Sudjianto, Analytical metamodel-based global sensitivity analysis and uncertainty propagation for robust design, SAE Transactions Journal of Materials & Manufacturing (2004). doi:10.4271/2004-01-0429.

[15] A. Marrel, B. Iooss, B. Laurent, O. Roustant, Calculations of sobol indices for the gaussian process metamodel, Reliability Engineering & System Safety 94 (3) (2009) 742–751. doi:10.1016/j.ress.2008.07.008.

[16] B. Lamoureux, N. Mechbal, J. R. Massé, A combined sensitivity analysis and kriging surrogate modeling for early validation of health indicators, Reliability Engineering and System Safety 130 (2014) 12–26. doi:10.1016/j.ress.2014.03.007.
URL http://dx.doi.org/10.1016/j.ress.2014.03.007

[17] D. P. Wipf, S. Nagarajan, A new view of automatic relevance determination, in: Proceedings of the 20th International Conference on Neural Information Processing Systems, NIPS'07, Curran Associates Inc., 2007, pp. 1625–1632.
URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.143.8009&rep=rep

[18] R. M. Neal, Bayesian Learning for Neural Networks, Springer New York, 1996.