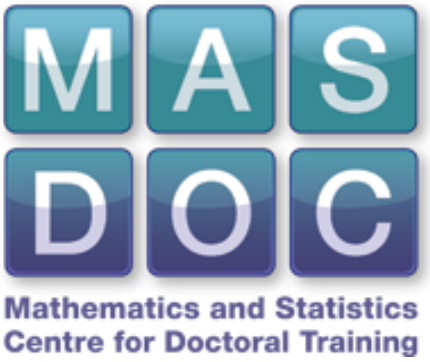


University of Warwick

Model Based Emissions Control

RSG Report

Authors:
Matthew EGGINTON
Jamie LUKINS
Jack SKIPPER

ID number:
0901849
0923890
0903339

Supervisors:
Dr. Tim SULLIVAN
Dr. Florian THEIL

Abstract

An important problem is the removal of noise from input measurements and state estimates. Reverse engineering, or approximation, of black boxes within systems is also a highly relevant topic. Methods are given, including least squares methods and a Bayesian inverse approach, to perform these approximations. Real time methods are also considered, for example the Extended Kalman filter. A particular industrial application, that of emissions control in a car engine, is used to highlight these and test their suitability.

Contents

1 Introduction

1.1 Motivation

Of critical importance in engineering, or any industrial application, is the consideration of data analysis, in particular that of smoothing noisy data and of analysing relationships between variables.

The former is a manifestation of the fact that any measurement taken of a physical setting will have some noise associated to it, essentially due to the fact that in taking a measurement, one disrupts the system one is trying to measure. This can occur throughout applied sciences as one tries to understand the actual state of a system while any data of that system will have associated noise of varying degree and type.

The latter comes from a black box type scenario (see figure 1), an unknown function, where one has input and output data from this system with errors attached to both, coming from inaccuracies in measurements and errors in models, and the goal is to recreate the black box.

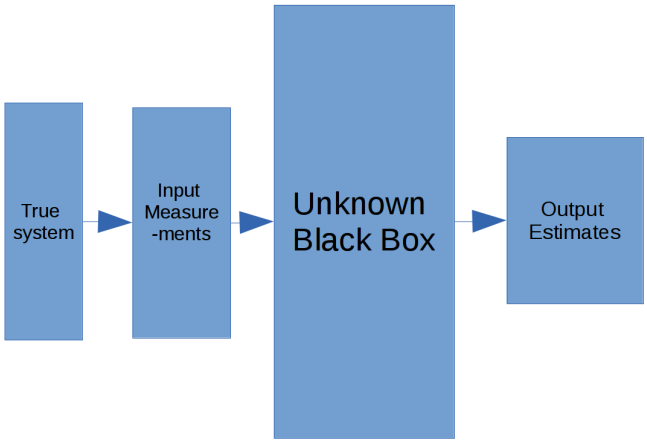


Figure 1: Black Box

This scenario could be due to companies using machinery or technology from external sources who do not divulge the exact workings of the part, or could be from the use of external computer programs to perform specific tasks, where the code is hidden.

1.2 Problem

We focus on the black box scenario. One has a time series $x = (x^i)$ of input measurements where $x^i \in \mathbb{R}^d$ for some d and $x^i \sim \mathcal{N}(\widetilde{x}^i, Q^i)$ for some positive definite symmetric matrices Q^i and $\widetilde{x}^i \in \mathbb{R}^d$, where in general Q^i is independent of Q^j for $i \neq j$.

We assume that the true system is given by $\widetilde{y} = (\widetilde{y}^i)$ and one has some function f that takes the inputs to the true system values.

One also has output measurements y^i where $y^i \sim \mathcal{N}(\widetilde{y}^i, P^i)$ where P^i are positive definite symmetric matrices.

Furthermore, no prior knowledge of the physical system is supposed.

The aim is to remove the noise from the measurements to predict the true system. This motivates two further problems:

1. Recreate, or approximate, the underlying function that maps, at time i, some previous number of estimated physical system measurements $x^{i-l}, ..., x^i$, and gives the true system at the future time y^{i+1}.

2. Specify a method that can run in real time, and with minimal computational power, that gives an estimate $\widetilde{y}^i$ of the physical system y^i given the input data up until that time.

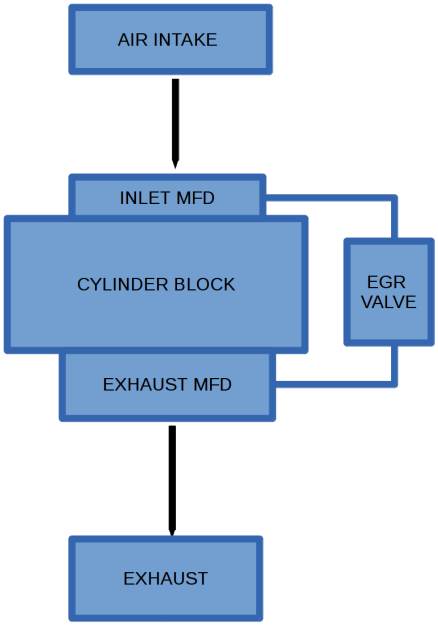
This is a combination of two types of problem. The first is an inverse problem where there is input and output data with noise from a black-box (an object where little is known) and finding or approximating the black-box is the task.

There is a certain sense of linearity with these problems, as it makes more sense to approach the first before the second, as the approximation of the black box can be used in a real time environment where it is coupled with input data to recreate the black box in real time or faster.

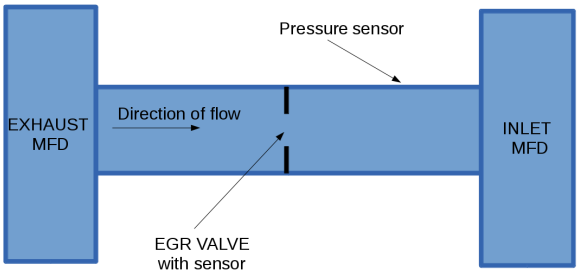
1.3 Exhaust Gas Recirculation in Modern Cars

A particular example of this is to be found within the realms of emissions optimisation within a car power-train.

Figure 2a shows a schematic of the modern car, showing the intake of air into the engine, and the flow through the cylinders, and then with the exhaust gas recirculation (E.G.R.) system to the right, which is highlighted in figure 2b. This system varies the amount of gas that is recirculated around the engine, in a controlled manner, with the control by the electronic control unit (E.C.U.), so as to reduce the emissions of the car. Good estimates on the amount of recirculation are useful so as to optimise this reduction.



(a) Schematic of a Car Engine



(b) Schematic of the EGR system

Figure 2: Diagrams of a Car engine

As such, the mass flow (M) and valve area (A) are desirable data values to know to a high accuracy, as these are used in predicting the future amount of air to be recirculated. The pressures upstream (P_U) and downstream (P_D) are also needed to predict these values.

In this context, we find a client who wishes to understand the mechanism used in the E.C.U. so as to estimate these measurements.

The aim would be, given the input time series of the data $x = ((p_d^i, p_u^i, m^i, a^i))_{i=1}^N$ together with the corresponding outputs $y = ((P_D^i, P_U^i, M^i, A^i))_{i=1}^N$, to construct the function $f : \prod_{k=1}^4 I_k \to \prod_{k=1}^4 I_k$ for suitable closed and bounded intervals I_k of $\mathbb{R}$, that describes the black box procedure, see figure 3. That is, to find the function that attempts to take the input value x^i and map it to the true state of the system $\widetilde{y}^i$. This is a problem in reverse engineering.

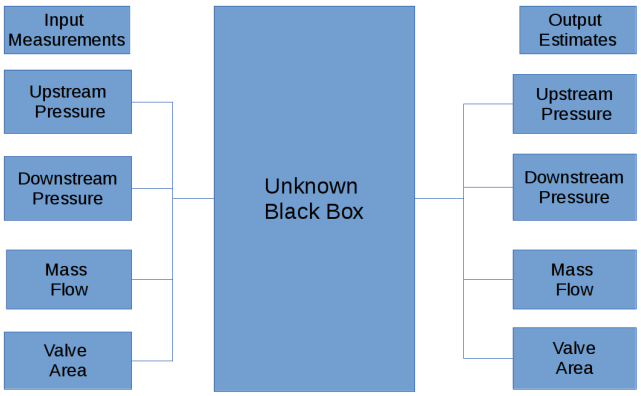


Figure 3: Black Box

We briefly highlight the techniques that can be used for this type of system, and discuss them in more detail in later sections.

1.4 Reverse Engineering of Physical System

1.4.1 Linear Interpolation

A linear interpolation of the data is used to approximate a function, with limitations due to the noise in the system. It is shown that noisy measurements severely restrict the use and results from this method.

1.4.2 Minimisation of Functionals

A least squares approximation of the form

$$I[f] = \sum_{i=1}^{N} \left| f(x^i) - y^i \right|^2 \tag{1.1}$$

is used to approximate functions $f : \mathbb{R}^4 \to \mathbb{R}^4$ in a suitable space of functions $\mathscr{W}$. It is shown that the space must be chosen carefully to ensure the basis functions have certain desired properties for the example in hand.

1.5 Real Time Computation

Considering that the operations run need to be performed in real time, the only feasible method is one which involves linear algebra, and if smoothing is required, it will involve an algorithm like that in the Kalman filter [6]. Methods are suggested for which such an algorithm can be employed.

1.6 Numerical Results

Certain success is met with choosing a Fourier basis of cosines where approximations are taken using a space of dimension $4^4 = 256$ on the first 10000 data points to simulate the remaining data. Note the good conditioning of this method. The results are summarised in table 1, where the values show the relative error improvement in each method in the stated statistic, namely (7.1) and (7.2). Note the fact that we have not used the downstream pressure here, since one cannot give a percentage increase since the input data is exactly the output data. In other words, one calculates

$$\frac{\|\text{input stream} - \text{approximation}\|}{\|\text{input stream} - \text{output stream}\|}$$

Method for $\|\lfloor\cdot\rfloor\|_1$	P_U error change	M error change	A error change
Input Data	1	1	1
Fourier	2.4928	0.3995	0.3787
EKF	1.1511	0.9106	0.6930
Method for $\|\lfloor\cdot\rfloor\|_2$	P_U error change	M error change	A error change
Input Data	1	1	1
Fourier	2.606	0.4658	0.3676
EKF	1.102	0.9654	0.6747
Method for $\|\cdot\|_\infty$	P_U error change	M error change	A error change
Input Data	1	1	1
Fourier	0.9925	1.0483	1.0310
EKF	0.9941	1.0945	0.7083

Table 1: Summary of the improvement of relative error with methods shown.

In the table the red numbers are those that are particularly good, and the blue are values that are a good improvement on the error.

In other words, both our methods improve the approximation by the input data for the mass flow and valve area in $||\lfloor\cdot\rfloor||_1$ and $||\lfloor\cdot\rfloor||_2$ and only slightly improve the approximation in $\|\cdot\|_\infty$. If one would add in constraints upon the valve area and mass flow not being negative, then one would expect better improvements.

2 Data Interpolation

Interpolation is the process where one has a collection (x^i, y^i) of data which is taken from some function $f : \mathbb{R}^d \to \mathbb{R}^d$ and one creates an approximation of the entire function by assuming the function to be linear, or a certain function, between these points. For simplicity, only linear interpolation is considered.

To evaluate the function at a point $x^* \in \mathbb{R}^d$, the process is as follows:

1. Find amongst $\{x^i\}_{i=1}^N$, the $d+1$ points $\{x^{i_1}, x^{i_2}, \ldots, x^{i_{d+1}}\}$ that are closest to x^* such that x^* is contained in the convex hull of $\{x^{i_1}, x^{i_2}, \ldots, x^{i_{d+1}}\}$
2. Calculate the hyperplane spanning those points
3. Calculate the value at the evaluation point $x^\star$ on this hyperplane.

The advantage of this method is the simplicity and speed of the computation of the interpolation. However, due to the fact that the method assumes that the values of the data given are the truth, if the noise in the measurement of the data is significant with regards to its data, then the interpolation will be poor.

If the function to be interpolated is twice differentiable, the error associated to the interpolation is of the order of the second derivative in magnitude, and so for a function that fluctuates wildly, this will be very large. This result is a consequence of Rolle's theorem.

In the situation where the order of magnitude of the variance of the noise is much smaller than the magnitude of the values of the data, linear interpolation is suitable since the noise will not affect the interpolation values so much.

3 Least Squares Minimisation

The method of least squares minimisation is a well studied and well used one. It is a technique that originates with Gauss [4] and Legendre [8]. It consists of the following method.

Given N pairs of data $(x^i, y^i) \in \mathbb{R}^d \times \mathbb{R}^d$ and a chosen function space $\mathscr{W}$, one finds the function f in $\mathscr{W}$ with $f : \mathbb{R}^d \to \mathbb{R}^d$ which minimises:

$$S = \sum_{i=1}^{N} |f_k(x^i) - y_k^i|^2 \tag{3.1}$$

for $k = 1, \ldots, d$. This is of particular interest and of use when one has error with the values y^i, as then one does not try to perfectly fit the information given, as in section 2.

By considering the error in observing y^i, we find ourselves within a probabilistic setting. Suppose the function f can be described by choosing some parameters β which are viewed as random, then since y^i is considered to be a Gaussian random sample drawn from a distribution centred at the true state of the system $\widetilde{y}^i$, we have the following relationship:

$$y_k^i - f_k\left(x^i|\beta\right) \sim \mathcal{N}(0, P^i).$$

where the P^i are the covariances in the outputs, as defined in section 1.2. Thus

$$y_k^i - f_k(x^i|\beta) \propto e^{-\left|f_k(x^i|\beta) - y_k^i\right|^2}.$$

The probabilistic approach to decide which estimate of the parameters β should provide the most accurate function f, is to find the estimate that maximises the likelihood function. Finding the maximum-likelihood estimator then corresponds to maximising

$$\mathcal{L}\left(x^1,\ldots,x^N\right)=\prod_{i=1}^{N} f_k(x^i|\beta) \propto e^{-\sum_{i=1}^{N}\left|f_k(x^i|\beta)-y_k^i\right|^2}$$

and the maximum of this coincides with the minimum least squares distance of $f_k(x^i|\beta)$ with y_k^i for all i which coincides exactly with the approach we are taking.

The following is similar to the book [11] or [12] and is stated in [7], though this goes above and beyond the level of detail we need. In some senses this technique is the most basic, although a more sophisticated constrained minimisation is considered later, this is not implemented, and these two are by no means the only methods that can be considered here. The references are highlighted to demonstrate this.

The minimisation problem in (3.1) of finding a function $f\colon \mathbb{R}^d \to \mathbb{R}^d$ is actually d minimisation problems of finding $f_k\colon \mathbb{R}^d \to \mathbb{R}$ where $\{f_k\}_{k=1}^d$ are the d components of f. To implement on a computer, one must parameterise a projection of the function space onto each of its components by finitely many parameters. In other words, a basis $\{\phi_j : \mathbb{R}^d \to \mathbb{R}^d\}_{j=1}^K$ of $\mathscr{W}$ is chosen and then projected onto each of the coordinate axes of $\mathbb{R}^d$. Then, for each f in $\mathscr{W}$ and $k=1,\ldots,d$, one can write the functions f_k by truncating its expansion:

$$f_k(x)=f_k(x,\beta_k)=\sum_{j=1}^{K}\beta_{j,k}\phi_{j,k}(x),$$

where ϕ_{jk} is clearly the projection of ϕ_j onto the k^{th} coordinate. For the sake of brevity, we will suppress the index k and treat f as a function from $\mathbb{R}^d$ to $\mathbb{R}$. A minimiser, if one exists, of the problem in (3.1) must have differential zero in β, namely

$$\nabla_\beta S(\beta)=0$$

and this is equivalent to

$$\frac{\partial S(x,\beta)}{\partial\beta_j}=2\sum_{i=1}^{N}(f(x^i,\beta)-y^i)\frac{\partial}{\partial\beta_j}(f(x^i,\beta))=0 \qquad \text{for } j=0,\ldots,n \tag{3.2}$$

and if we let

$$X_{ij}=\frac{\partial}{\partial\beta_j}(f(x^i,\beta)) \tag{3.3}$$

then the solution to the above problem is

$$\beta=(X^TX)^{-1}X^Ty$$

using the following theorem, which can be found in [10] or in [11, Theorem 3.1]:

Theorem 3.1 (Gauss-Markov) *Suppose that X^TX is invertible. Then among all unbiased linear estimators K such that $\beta=Ky$ the one that minimises the least squares error is*

$$K=(X^TX)^{-1}X^T$$

Proof Let $\widetilde{\beta} = K_D y$ be another linear estimator of β and let K_D be given by

$$K_D = (X^T X)^{-1} X^T + D$$

where D is a $k \times n$ non-zero matrix. As we only consider unbiased estimators, minimum mean squared error implies minimum variance. The goal is therefore to show that such an estimator has a variance no smaller than that of the least squares estimator.

Supposing that one writes $y = X\beta + \varepsilon$ for some ε, one has that the expectation of $\widetilde{\beta}$ is:

$$\begin{aligned}\mathbb{E}(K_D y) &= \mathbb{E}\Big(((X^T X)^{-1} X^T + D)(X\beta + \varepsilon)\Big) \\ &= ((X^T X)^{-1} X^T + D)X\beta + ((X^T X)^{-1} X^T + D)\mathbb{E}(\varepsilon)\end{aligned}$$

and $\mathbb{E}(\varepsilon) = 0$ since the estimator is unbiased, and then one gets that

$$\mathbb{E}(\widetilde{\beta}) = (X^T X)^{-1} X^T X\beta + DX\beta = (I_k + DX)\beta$$

and thus we have that $\tilde{\beta}$ is unbiased if and only if $DX = 0$. Now, supposing that the variance of y is σ^2, one calculates the variance of $\tilde{\beta}$. We have

$$\mathrm{Var}(\widetilde{\beta}) = \mathrm{Var}(K_D y) = K_D \mathrm{Var}(y) K_D^T = \sigma^2 K_D K_D^T$$

And then explicitly writing the formula for K_D one obtains

$$\begin{aligned}\sigma^2 K_D K_D^T &= \sigma^2((X^T X)^{-1} X^T + D)(X(X^T X)^{-1} + D^T) \\ &= \sigma^2((X^T X)^{-1} X^T X (X^T X)^{-1} + (X^T X)^{-1} X^T D^T + DX(X^T X)^{-1} + DD^T) \\ &= \sigma^2 (X^T X)^{-1} + \sigma^2 (X^T X)^{-1} (DX)^T + \sigma^2 DX(X^T X)^{-1} + \sigma^2 DD^T \\ &= \sigma^2 (X^T X)^{-1} + \sigma^2 DD^T.\end{aligned}$$

since we have from above that $DX = 0$.

Since DD^T is a positive semi-definite matrix, we have that the variance of $\tilde{\beta}$ is larger than the variance of Ky. As D was arbitrary, K must be the minimum. *Q.E.D.*

Observe that this method explicitly gives a unique solution to the above minimisation problem if $K \leq N$ and so the method is suitable to be used.

We have thus proved the following, which is the form of this method that is used in later sections.

Theorem 3.2 *Let I_k be a closed and bounded interval in $\mathbb{R}$ for $k = 1, \ldots, d$. Suppose that $x^i, y^i \in I_1 \times \cdots \times I_d$ for $i = 1, ..., N$. Then, for each k the functional*

$$f \mapsto \sum_{i=1}^{N} |f(x^i) - y_k^i|^2$$

has a minimum in the class of functions

$$\mathscr{W} = \left\{ f(x_1, x_2, \ldots, x_d) = \prod_{k=1}^{d} \cos(c_k x_k) \,\middle|\, c_k \text{ in } \mathbb{R} \text{ for each } k \right\}$$

In addition, the functional has a minimum in the space of polynomials of degree at most L which is unique if $L \leq N$ but not necessarily if $L > N$.

3.1 Constrained Minimisation

Furthermore, if it is physically relevant to consider some constraint in the situation, then a constrained least squares minimisation will be more suitable. While no results are given numerically later on this, it is still a useful consideration to make.

We now prove existence and uniqueness of the minimisation.

$$\text{Minimise } \sum_{i=1}^{N} |f(x^i) - y_k^i|^2 \text{ in } \mathscr{W} \text{ subject to } \|\nabla f\|_{L^2}^2 \leq \gamma \tag{3.4}$$

for some $\gamma > 0$.

This can be rewritten using a basis for the coefficients in the manner shown above, and the constraint can also be rewritten in this form, since we have

$$\|\nabla f\|_{L^2}^2 = \left\| \nabla \left(\sum_{j=1}^{K} \beta_j \phi_j \right) \right\|_{L^2}^2 \leq \sum_{j=1}^{K} |\beta|^2 \|\nabla \phi_j\|_{L^2}^2 .$$

Thus the constrained minimisation can be written as

$$\text{Minimise } \|C\beta - d\|_2^2 \text{ in } \mathscr{W} \text{ subject to } A\beta \leq \gamma$$

and this is equivalent to

$$\text{Minimise } \|C\beta - d\|_2^2 + \|A\beta\|_2^2 \text{ in } \mathscr{W}. \tag{3.5}$$

This equivalence can be seen by considering the minimisation problem

$$\min_{\gamma} \left(\min_{\|C\beta\| \leq \gamma} \|A\beta - d\|_2^2 + \gamma^2 \right).$$

This is the situation of Tikhonov regularisation:

Theorem 3.3 (Tikhonov Regularisation) *Let $A : \mathcal{H} \to \mathcal{K}$ be a linear operator between Hilbert Spaces such that $R(A)$ is a closed subspace of $\mathcal{K}$. Let $Q : \mathcal{H} \to \mathcal{H}$ be self adjoint and positive definite, and $b \in \mathcal{K}$ and $x_0 \in \mathcal{H}$ be given as well. Then*

$$\hat{x} \in \text{argmin}_{x \in \mathcal{H}} \left(\|Ax - b\|^2 + \|x - x_0\|_Q^2 \right) \iff (A^\star A + Q)\hat{x} = A^\star b + Q x_0$$

Proof We modify the problem by considering $B : \mathcal{H} \to \mathcal{K} \otimes \mathcal{K}$ defined by

$$Bx = Ax \oplus Lx$$

where $Lx = x - x_0$. and by considering $d = b \oplus 0$. Then one gets that

$$\|Bx - d\|_{\mathcal{K}\otimes\mathcal{K}}^2 = \|Ax - b\|^2 + \|x - x_0\|_Q^2$$

Then considering the least squares minimisation in $\mathcal{K} \otimes \mathcal{K}$ one finds that

$$(A^\star A + L^\star L)x = A^\star b$$

Inserting the definition of L yields

$$(A^\star A + Q)\hat{x} = A^\star b + Q x_0$$

as required. *Q.E.D.*

A minor digression here motivates this in the Bayesian framework, which is considered in more detail in section 5. One can think of the constraint as some form of prior distribution characterising the beliefs in the system, and one then considers the posterior distribution calculated using squared loss on the information characterised by the relation $Ax = b$.

3.2 Choice of Basis for Minimisation Space

An important part of the least squares minimisation is the choice of the space $\mathscr{W}$ over which one finds the minimum. In part this is decided by the data at hand, and the following points must be considered:

1. How the data is spread about the domain upon which one considers the functions to act upon,
2. The number of data points given.

The first is important due to the fact that if the data is clustered in a small region of the domain, then a choice of basis such that each function has local support is not ideal, since the coefficients of the basis functions cannot be produced accurately away from the clusterings of the data.

On the other hand, if the data is well spread out throughout the domain, then a choice of basis with local support is ideal, since then the coefficient matrix is sparse. This reduces the computation required to find the least squares minimiser.

Furthermore, to ensure that the problem is well conditioned and that the approximation does not over fit the data, the dimension of the basis needs to be taken to be much smaller than the number of data points given.

3.2.1 Principal Component Analysis

Over-fitting the data is a very dangerous situation to be found in. It will force the function to fit too well to the noise associated to the specific data set it is found from and it won't generalise to other sets very well. Therefore, the performance of the approximating function depends upon the number of basis elements used in its expansion, it is desirable to have methods to determine a sensible dimensions of the basis. One such method can be found in the study of principal component analysis (P.C.A.) [5].

P.C.A. is a statistical method that studies the modes of variation in the data. It achieves this by performing a transformation on the data of variables that are potentially highly correlated into a new coordinate system whose variables are now uncorrelated. The advantage of P.C.A. is that information about which directions contain as much of the data's variability as possible is easily available.

To employ this technique, it is necessary to estimate the covariance matrix C for the data. It is then possible to find the singular value decomposition of C. A study of the singular values of C, in particular the largest singular values, then gives a good indication of the number of basis elements that are needed to capture the behaviour of the function without the risk of over-fitting the data.

Two different spaces are now considered, to show the difference between a locally supported basis and a globally supported basis.

3.2.2 Schauder Type Finite Dimensional Subspace

We define $\mathscr{W}_R$ as the space spanned by the following functions. We suppose that our functions are defined on the space $\prod_{k=1}^{d} I_k$ for I_k some closed and bounded intervals of $\mathbb{R}$.

For each interval I_k choose a discretisation of it into l_k points, so that $l_1 \dots l_d << N$ for good conditioning.

For each point $z = (z_1, \ldots, z_d)$ in this discretisation of $\prod_k I_k$ we define the function ψ_z to be the function that is 1 at the point z and linearly decays to zero at the neighbouring points and is given by the formula:

$$\psi_z(w_1, \cdots, w_d) = \begin{cases} 0 & \text{if } \max\limits_{k=1,\ldots,d} \frac{|w_k - z_k|}{l_k} \geq 1 \\ 1 - \max\limits_{k=1,\ldots,d} \frac{|w_k - z_k|}{l_k} & \text{else} \end{cases} \tag{3.6}$$

Observe that this is similar to Haar wavelets and Schauder functions and also that each basis function is locally supported in the region

$$\left[z_1 - \frac{l_1}{|K_1|}, z_1 + \frac{l_1}{|K_1|}\right] \times \cdots \times \left[z_d - \frac{l_d}{|K_d|}, z_d + \frac{l_d}{|K_d|}\right]$$

and also at any given point in $\prod K_i$ there are at most $2d$ non-zero basis functions.

Using this basis, our value to minimise then becomes

$$\sum_{i=1}^{R} \left| f(x^i) - y_k^i \right|^2 = \left\| C\underline{\alpha} - \underline{y}_k \right\|_2^2$$

where $C = \left(\psi_j(x^i)\right)_{i,j}$ for i over the number of data points, and j over the number of basis functions, and $\underline{y}_k, \underline{\alpha}$ are the vectors of the y_k^i and α_j respectively. The matrix C here has $4d$ non-zero elements in each row, since the function above has support on an L^∞ ball, and so if the dimension of the function space is much larger than this, one can see that we have a sparse matrix.

3.2.3 Fourier Space

We again suppose that the functions are defined on the space $\prod_{k=1}^d I_k$ for I_k some closed and bounded intervals of $\mathbb{R}$. Indices $i_1, i_2, \ldots i_d$ are chosen so that $i_k = 1, \ldots, l_k$, and such that $l_1 \ldots l_d << N$. Then one defines

$$\gamma_{i_1,\ldots,i_d}(x_1, x_2, \ldots, x_d) = \cos\left(\frac{x_1 - \min I_1}{2\pi|I_1|} i_1\right) \times \cdots \times \cos\left(\frac{x_d - \min I_d}{2\pi|I_d|} i_d\right), \tag{3.7}$$

This basis is an example of a globally supported basis, as, apart from finitely many points, the basis is non-zero.

Furthermore the method above gives the same form of the minimisation problem, except that the matrix C is no longer sparse, and one would expect each element of the matrix to have a non-zero entry.

3.3 Other Minimisation Spaces

Two spaces for which minimisation has been implemented have been introduced, but these are by no means the only spaces in which minimisation can occur. One can consider spline functions as the basis to minimise, or radial functions instead. Furthermore, the specific situation considered may well provide an insight into the choice of the space to consider.

The example of a radial basis is particularly useful, although not implemented later, which is described in [1]. One defines this as follows.

One has a set Ξ of "centres" which are points at which one wishes to approximate at. In our case $\Xi = (x^i)_{i=1}^N$. Then one has the values $f(x^i)$ at these values, where one

assumes that f is smooth. Then a radial basis function is the composition of a continuous function $\phi\colon \mathbb{R}_+ \to \mathbb{R}$ with the Euclidean norm. One then writes a function s in this space as

$$s(x) = \sum_{\xi \in \Xi} \lambda_\xi \phi\left(\|x - \xi\|_2\right)$$

for some λ_ξ. By construction, the approximation is equal to the function at these points. Common uses of ϕ are Gaussian functions $\phi(r) = e^{-c^2 r^2}$ for some parameter c, or thin plate splines $\phi(r) = r^2 \log r$.

These are particularly good at approximating multivariate functions, especially when there is an absence of grid data. One can also give explicit error bounds on the quality of the approximation.

The method also allows efficient computation for large data sets, in particular, it is easy to calculate the coefficients needed in the expansion.

4 Kalman Filtering

We employ applications of the Kalman filter in our attempt to approximate data of an unknown function that describes some system. Kalman filtering is an algorithm that tries to reconcile outputs from a mathematical model of a physical system and observations of the same system. It combines these two in such a way as to smooth out noise coming from inaccurate observations in an effort to more accurately estimate the true state of the system at the present time. The original paper by Kalman is [6].

Consider the process $x \in \mathbb{R}^d$ given by the stochastic difference equation

$$x^i = Ax^{i-1} + Bu^{i-1} + w^{i-1} \tag{4.1}$$

with u the control variables, and consider also measurements

$$z^i = Hx^i + v^i \tag{4.2}$$

where w and v represent noise terms, and are assumed to be independent multivariate normals with distributions

$$w \sim \mathrm{MVN}(0, Q) \qquad\qquad v \sim \mathrm{MVN}(0, R). \tag{4.3}$$

It should be recognised that this is somewhat of a special case of the usual Kalman filter, because one may well expect the noise to vary over time, and so then the above would be indexed by i.

The algorithm is as follows

1. At time n, given the previous a posteriori estimates $x^{n-1}, ..., x^{n-l}$ of the system, a prediction $\hat{x}^{n|n-1}$ is made based upon the prior belief or physical dynamics of the system at time n.

2. The system is observed at time n and this observation z^n is used to correct the a priori estimate $\hat{x}^{n|n-1}$ and produce an updated estimate $\hat{x}^{n|n}$.

3. Repeat for time $n + 1$.

4.1 The Extended Kalman Filter

The above algorithm handles the case that the update as linear. There is absolutely no reason to expect this to be the case and the algorithm can be altered to allow for non-linearities. In the non-linear case we must add a step in the procedure so that we can make the linear Kalman filter applicable. This new procedure is called the extended Kalman filter. The extended Kalman filter is discussed in detail in [3]. Here, instead of having a relation as in equation (4.1) we have a relation

$$x^{i+1} = f(x^i, u^i) + w^i \tag{4.4}$$

where f is non-linear, and w is again the noise term, and we have measurements

$$z^i = h(x^i) + v^i. \tag{4.5}$$

Now we assume, if $f, h \in C^1$, with derivatives F and H respectively, that the estimate is given by

$$\hat{x}^{i+1} \sim \mathcal{N}(Df(m^i), Df(m^i)^T C^i Df(m^i))$$

for some m^i and C^i to be determined and we have

$$\hat{x}^{i+1|i} = \hat{x}^{i+1}|y^i \sim \mathcal{N}(\hat{x}^{i+1|i}, K(y^i - h(\hat{x}^{i+1|i})))$$

where K is the so-called Kalman gain matrix. Observe that this linearises the problem, and we thus use the linear Kalman filter as above on this model. This however may be poor if our function f is highly non-linear.

The main advantage of using the Kalman filter in practice is that it finds solutions to an estimation problem sequentially and thus reduces the computational cost and time so that it can be performed as and when an observation is made.

4.2 Other Filters

There exist several useful generalisations of the Kalman filter. Another very important variant of the Kalman filter is the Kalman filter with fading memory. This has an immense number of practical applications. It takes into account the fact that a system can change dramatically so that the current state should depend more strongly on the recent observations and in effect forgetting the error contributed by older observations. For example, if you are interested in the current state of some system within the engine of a car and that car has just pulled away after having come to a completing stop, then you would probably not be so worried about the state of the engine before the stop. Including this information at full weight might throw the estimation off. This version of the Kalman filter is said to have "fading memory" and differs from the original by multiplying older measurements by progressively higher powers of a constant $\lambda \in (0, 1)$

5 Bayesian Inverse Approach

An alternative clear choice of method is the approach of Bayesian inverse problems which is discussed in detail in [13]. This is now briefly outlined.

The inverse problem approach for finding the update function for the Kalman filter would take the following form. The object is to find an unknown function f in a Banach space $\mathscr{W}$, possibly the space of continuous functions. A Gaussian prior distribution $\mathcal{N}(m, C)$ where $C\colon \mathscr{W} \to \mathscr{W}$ is the covariance operator is a natural choice to model an

unknown function due to their nice computational qualities as discussed in [2]. Bayes' formula can then be used to calculate the posterior from which we aim to sample in order to produce our guess for what the function should be.

Crucial to sampling from a distribution on a function space is the Karhunen-Loève expansion of a random process in terms of the eigenfunctions of the covariance operator. Suppose $K \subseteq \mathbb{R}^n$ is a compact domain in space and let $(\Omega, \mathcal{F}, \mu)$ be a probability space.

Theorem 5.1 (Karhunen-Loève) *[9] Let $U\colon K \times \Omega \to \mathbb{R}$ be a square integral mean-zero stochastic process with continuous covariance function $C_U(x, y) = \mathbb{E}_\mu[U(x)U(y)]$ satisfying the following: C_U is continuous, symmetric and positive definite. Then U can be decomposed as*

$$U = \sum_{n\in\mathbb{N}} Z_n \phi_n$$

where $\{\phi_n\}_{n\in\mathbb{N}}$ are the orthonormal eigenfunctions of the covariance operator and

$$Z_n = \int_K U(x)\phi_n(x)\mathrm{d}x.$$

The convergence of U is uniform in x.

This theorem can be utilised to sample from a distribution on a function space as follows. Take, for example, the Laplacian operator on $[0, 1]$ and consider the covariance operator given by $(-\Delta)^{-1}$. Then a sample path, u from the normal distribution $\mathcal{N}(0, (-\Delta)^{-1})$ on $\mathcal{H} = L^2([0, 1])$ can be expanded as

$$u(t) = \frac{\sqrt{2}}{\pi} \sum_{k=1}^{\infty} \frac{\xi_k}{k} sin(k\pi t)$$

where $\xi_1, \xi_2, ...$ are independent and identically distributed according to the standard Gaussian distribution $\mathcal{N}(0, 1)$ on the real line.

With this tool in the arsenal, a standard approach to tackling inverse problems and to sample from the posterior is to implement an Markov Chain Monte Carlo (M.C.M.C.) algorithm, examples of which are outlined in [2]. The general idea behind M.C.M.C. is to build a Markov chain whose equilibrium distribution is the distribution of the posterior that is desired. The most standard example is a Random Walk or Metropolis-Hastings M.C.M.C. in which a sample u^k is drawn from the prior and used to propose a candidate draw v^k for the posterior by taking a random step (hence the name) which is accepted with a certain probability. If the sample is accepted, the procedure is repeated after setting $u^{k+1} = v^k$. Otherwise, the procedure is repeated with $u^{k+1} = u^k$.

This approach looked promising, and would have been implemented fully if time permitted.

6 Numerical Data

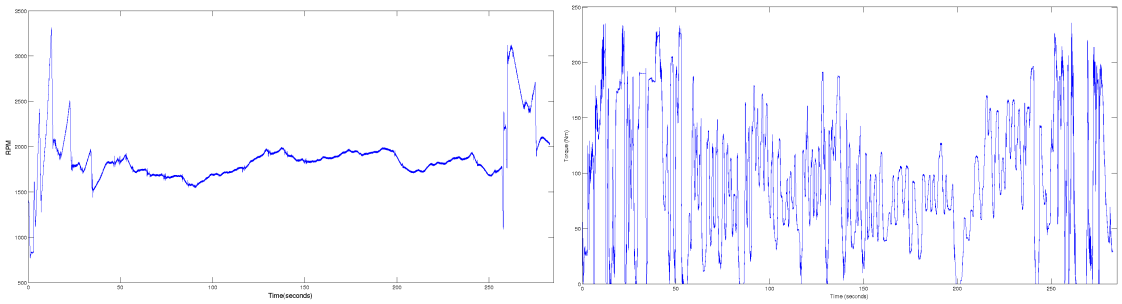
If the reader is not so interested in the specifics of the data, then this section is self contained and can be left out, barring figures 7 and 8 which are the important diagrams to see.

The data used was supplied by an interested client, and describes the E.G.R. system in a car, as shown in figure 2. The data consists of upstream pressure P_U, downstream pressure P_D, mass flow through the system M and the area of a valve A. Henceforth

these notations are sacred and used only to describe these values. The system that is to be approximated is described in figure 3.

A brief analysis of the data is given before any attempts in section 7 to use it. The data consists of a number of time series, with data indexed by the natural numbers from 1 to 28260, where we have a time step of 0.01s in between each natural number.

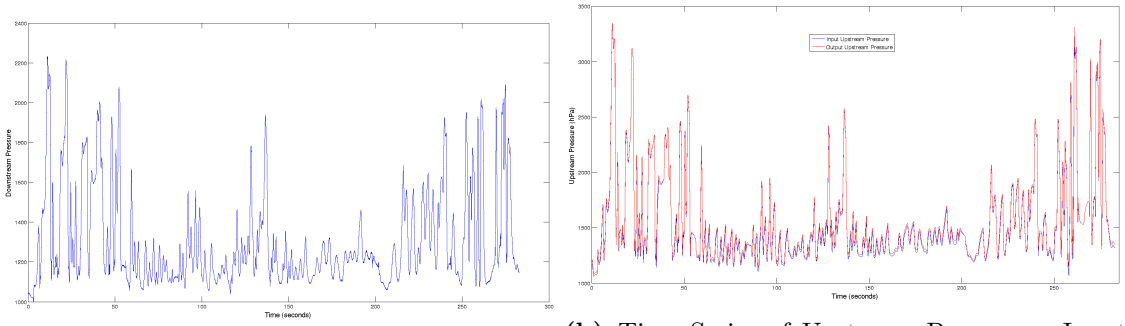
The figures 4, 5, 6a and 6b show time series of the data.



(a) Time Series of Engine RPM **(b)** Time Series of Engine Torque

Figure 4: Working Conditions of Engine

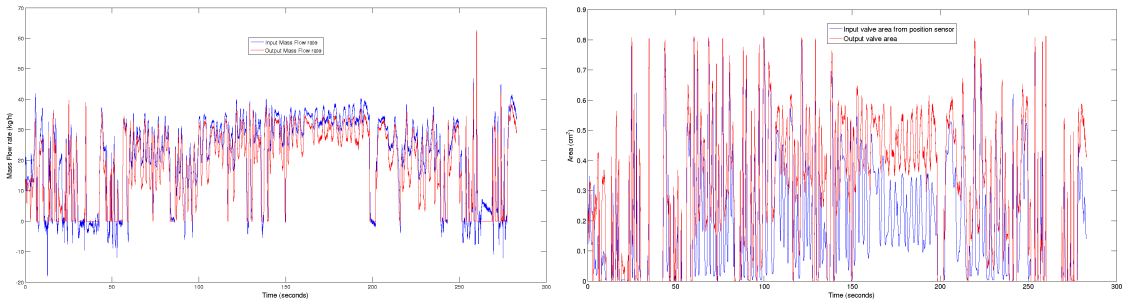
The working conditions of the engine describe a typical journey. The speed of the engine (figure 4a) shows the engine starting up, with a period of constant motion, followed by a sharp acceleration at the end. The force applied by the engine (figure 4b) shows constantly changing torque applied as if the car were performing lots of little changes in relative speed with the ground.



(a) Time Series of Downstream Pressures **(b)** Time Series of Upstream Pressures. Input blue output red

Figure 5: Time Series of Pressure data

As can be seen, the input and output pressures from the system are essentially the identity, but the input and output areas and mass flow values are really noisy and vary wildly.



(a) Time Series of Mass Flows **(b)** Time Series of Valve Areas.

Figure 6: Mass flow and area input and output values

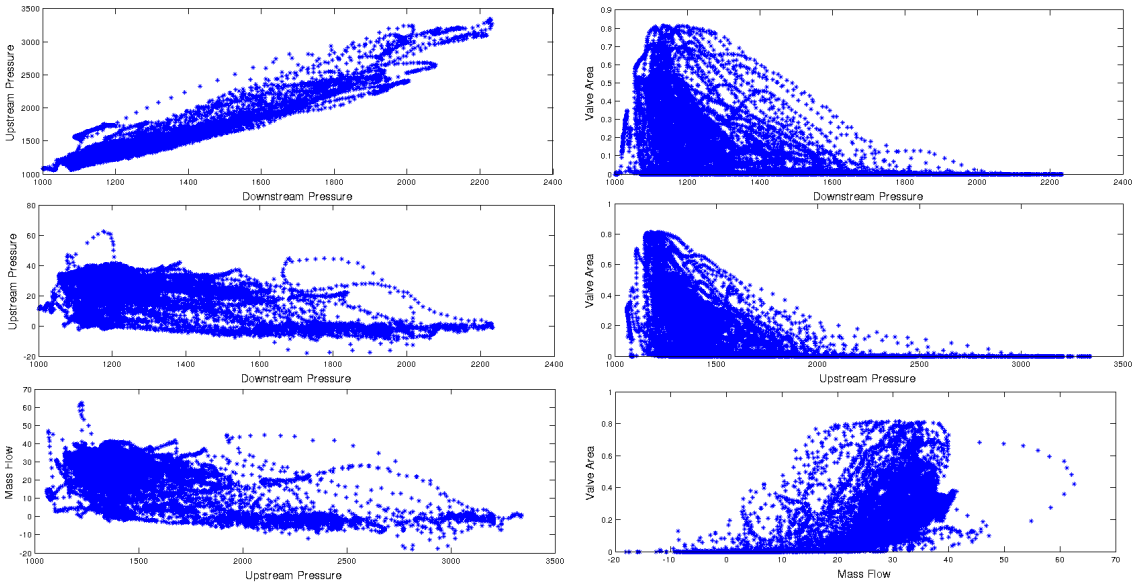


Figure 7: Coordinate Projections of Input Data

Figure 7 shows the relationships between the different data streams, and as one can see, there is very little relationship between all of them, apart from between the two pressure values, where there is roughly a linear relationship. One notes further that the data seems to only cover subsections of the domain so there seems specific areas where engine operations seem not to occur.

If our methods suggested below are to be any good, the minimum requirement is that they produce a smaller error between the approximation and the estimate. As such, logarithmic plots of the normalised difference between the input and output data are given in figure 8. As one can see, there is zero difference between the downstream pressure input and output values, and so one cannot expect to predict a better method here. The error in the upstream pressure is 10^{-2} which is already good. The error in the mass flow is of the order of 10^{-1} and the error in the valve area is $10^{-1/2}$ and one would expect to be able to improve this.

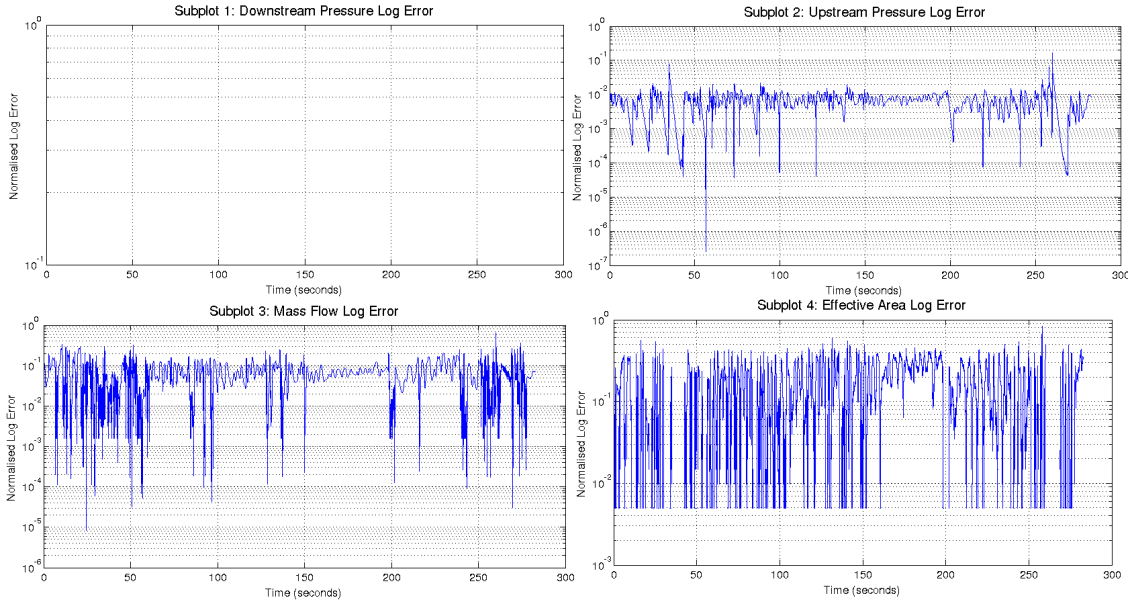


Figure 8: Log Normalised difference of Input and Output data streams

7 Numerical Results

The above methods are now applied to the data as explained in section 6.

The following statistics are used to describe time series throughout:

Definition 7.1 *We define, for each coordinate direction $k = 1, 2, ..., d$ of $\mathbb{R}^d$, the following values:*

$$||\lfloor x_k \rfloor||_p = \frac{1}{N^{1/p}} \left(\sum_{i=1}^{N} |x_k^i|^p \right)^{1/p} \tag{7.1}$$

$$||\lfloor x_k \rfloor||_\infty = \sup_{i=1,...,N} \{|x_k^i|\} \tag{7.2}$$

Observe that the scaling of these values is chosen so that if x were a collection of points in L^p then this value would converge to the L^p norm as $N \to \infty$.

Definition 7.2 *We define, for each coordinate direction, the* **total variation** *to be the value*

$$\|x_k\|_{TV} = \frac{1}{N-1} \sum_{i=1}^{N} \left| x_k^{i+1} - x_k^i \right|$$

7.1 Linear Interpolation

We analyse the difference between two different time series by the values as given in (7.1) and (7.2) so as the first value gives an estimation of the average difference between two elements of $\mathbb{R}^N$ and the second gives a difference of the maximum difference of two elements of $\mathbb{R}^N$. In the former, larger values of p, $(p > 1)$ give more weight to larger differences between two values, and smaller values of p $(p < 1)$ give less weight to larger differences than smaller ones. This interpolation gives, as output, the following graphs:

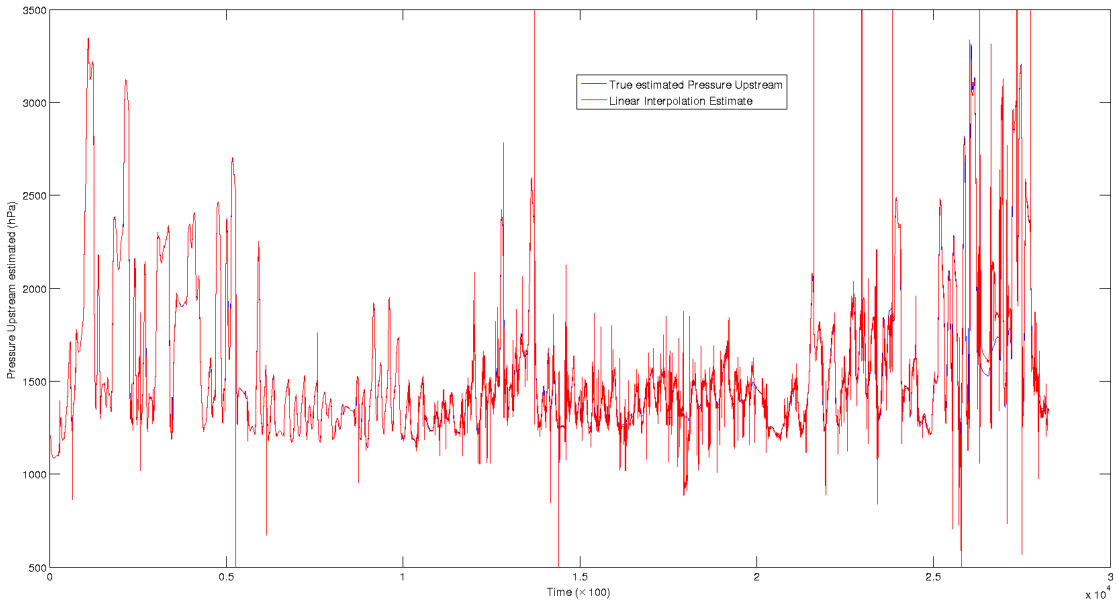


Figure 9: Upstream Pressure Linear interpolation on first 10000 data values

Figure 9 shows an interpolation where the authors have taken one in every 4 points in the first 10000 data points and used this as an approximation for the rest of the data. This produces a much more noisy collection of the data, but one can still see the general trend of the data, except for the final 5000 data points.

One also observes in figure 9 that there are points that the interpolation predicts excessively large values, uncharacteristic within what one can see to be the case. This is due to the fact that we have noise in the system, so the interpolation is affected by this.

7.2 Least Squares Minimisation Method

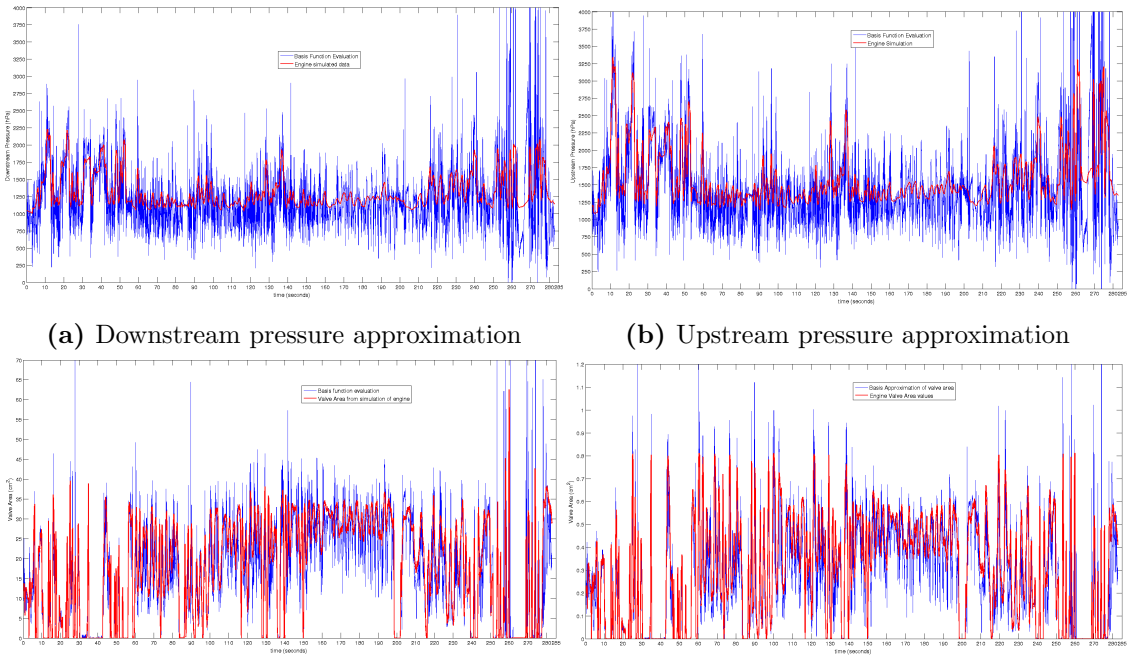
7.2.1 Schauder Type finite dimensional subspace

As has been said above, a local basis would be a very poor choice of basis since the data is clustered in the operating space, see figure 7. However, we feel it important to give a definitive example to show just how poor this method is in such a case, and so first approximate in this manner.

We ran the method as described in section 3.2.2 with 10 basis points in each direction, as this was the limit of the computation power we had. Applying this method, one sees figure 10 as the outputs.

It is clear to see that the approximated values are much worse than the true values. The authors believe that this is due to the low number of basis elements that are being used, and if many more (circa 100 in each direction) are used then a much better approximation will be found. However, one is limited to the number of basis functions to use be the amount of data given.

One has the error values as shown in table 2 and considering P_D and P_U take values between 1000 and 2000 hPa, A has values between 0 and 1 and M has values between 0 and 80 these are huge errors, and so this low dimensional approximation is not good.



(a) Downstream pressure approximation

(b) Upstream pressure approximation

(c) Mass Flow approximation

(d) Valve Area approximation

Figure 10: Approximation with Schauder type functions

The basis functions have bounded support and thus as data is only over a subsection of the domain they cannot predict engine conditions in this area so if input values occur out of the data set we expect failure of the system.

7.2.2 Fourier space

We instead take $\mathscr{W}_R$ to be the space spanned by the functions

$$\gamma_{i,j,k,l}(x_1, x_2, x_3, x_4) = \cos\left(\frac{x_1 - 1000}{2500\pi} i\right) \cos\left(\frac{x_2 - 1050}{4600\pi} j\right) \cos\left(\frac{x_3 + 20}{170\pi} k\right) \cos\left(\frac{x_4}{1.7\pi} l\right)$$

defined on all of the space $[1000, 2250] \times [1050, 3350] \times [-20, 65] \times [0, 0.85]$ and we vary i, j, k, l to be the frequencies in each direction. Taking the frequencies to be the values 1,2,3,4 we produce figure 11 as the approximation to the data. This approximation used a space with dimension $4^4 = 256$ to approximate the first 10000 data points, and then approximated the final 18000 data points.

Table 3 shows the differences of values in various statistics between the approximation and the data given by the client. Recall the norms used in equations (7.1),(7.2) and as you can see this is a much better approximation than the Schauder type basis, see table 2.

We also varied the size of the dimension of the space for this prediction for various differing values of R, with $R = 4^4, 6^4, 8^4, 10^4$ and we found that the error is more or less constant for all these values, and so we chose the smallest dimension for our testing.

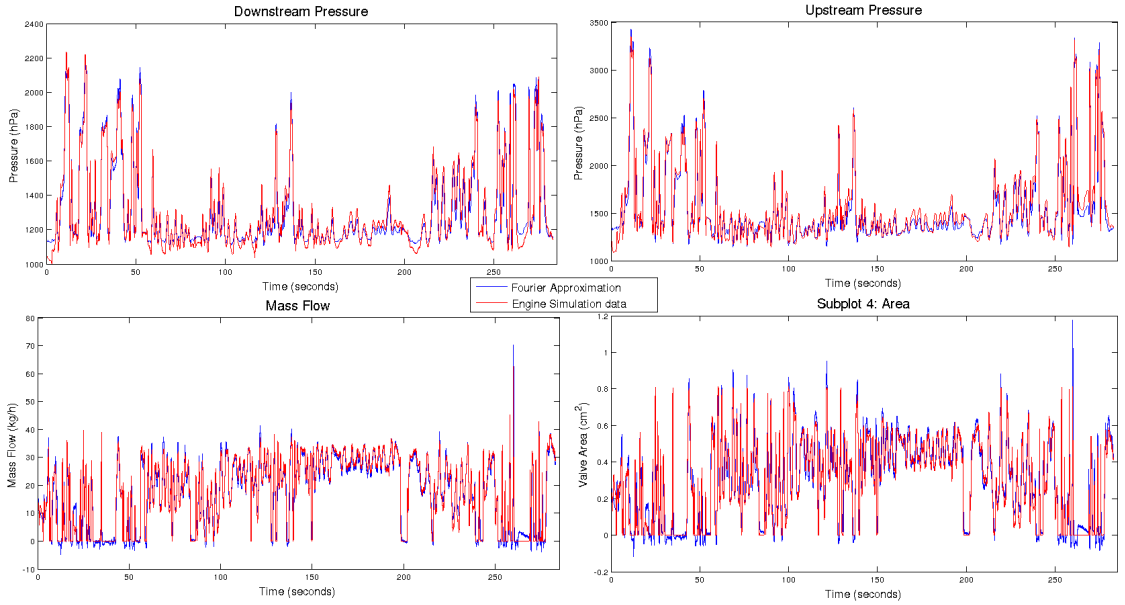


Figure 11: Approximation of the data with Fourier space

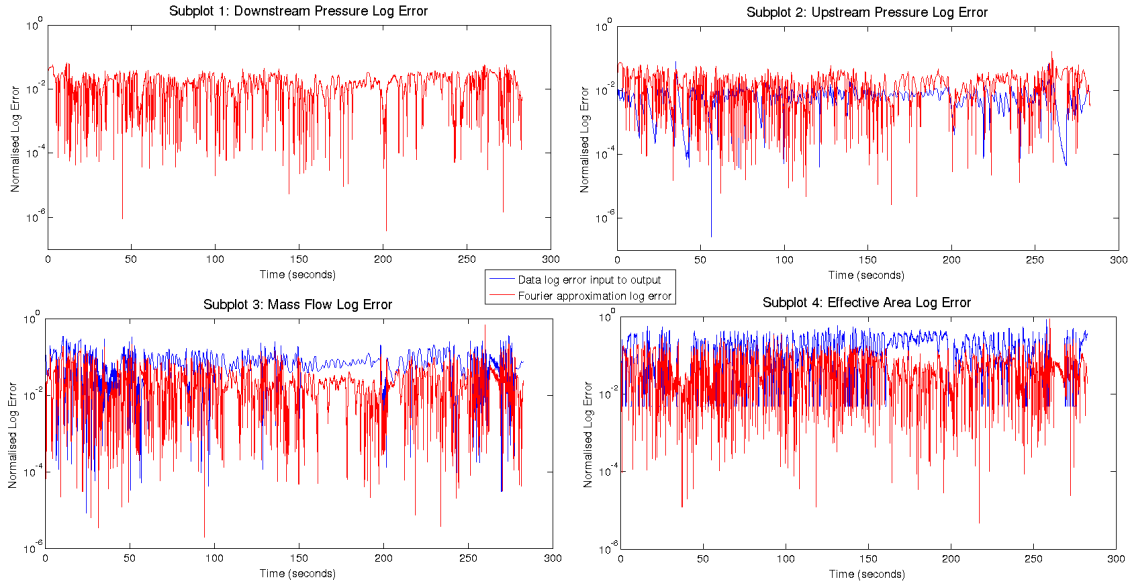


Figure 12: Log Normalised error in data and Fourier approximation

Since figure 11 does not really show how good our approximation was, a log base 10 normalised error plot was produced and is shown in figure 12. This shows that the normalised log error values are around 10^{-2} or smaller for the pressures and between 10^{-1} and 10^{-2} for the valve area and mass flow. As one can see, the approximation improves significantly on the input data for the mass flow and valve area, but doesn't improve on the input for the upstream pressure.

Low dimensional slices of the response surface of this approximation are shown in appendix B. It is useful to envisage this as one can easily see how the differing variables are related. One can also gain an idea of how much spread there is in an output value dependent on various inputs. For example, figure 17 shows the variation of mass flow

with regards to input valve area and downstream pressure for constant input mass flow of 40 kg/h. Thus one can see that the output varies from between 30 kg/h to 45 kg/h due to different values of the downstream pressure and valve area.

7.3 Kalman Filtering

The authors implemented a four dimensional extended Kalman filter using the Fourier approximation and figure 13 shows the output. As one can see, our approximation follows the two pressures well, and the mass flow adequately. It keeps the shape of the valve area but seems to have a downward shift. This is due to the fact that the input data is always below the estimated value, and so which our Fourier approximation is faithful to the true estimate, the data isn't, and so this skews the combination of the two.

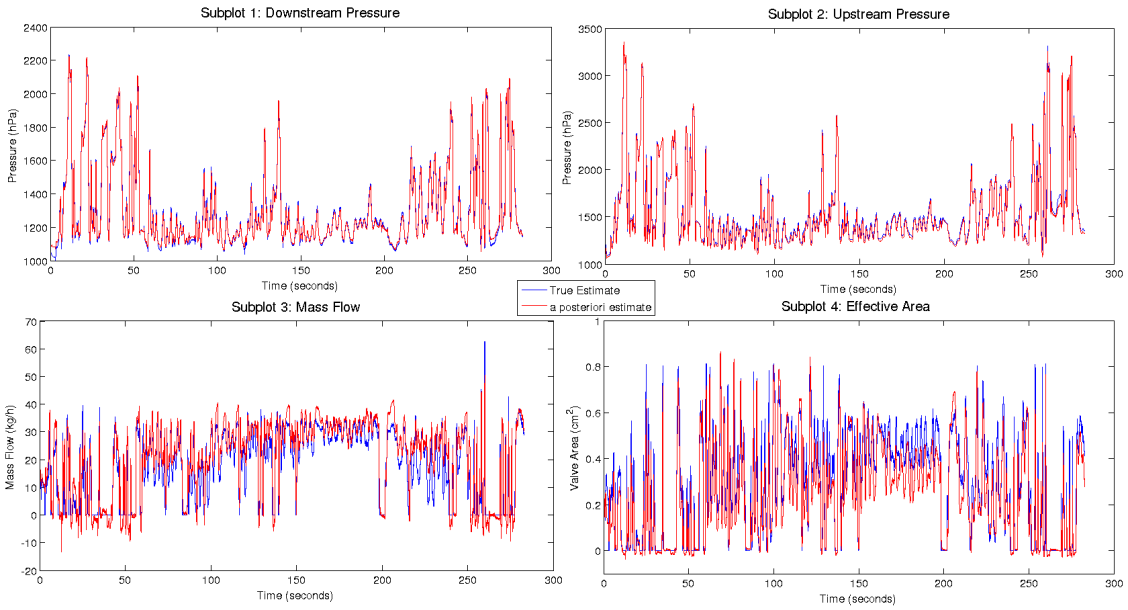


Figure 13: Approximation of the data with Extended Kalman Filter

Figure 14 shows the log of normalised error between the data from the engine and the Kalman filter approximation. As one can see, the log error of 10^{-2} for both the pressure values is no different from the accuracy of the input data, and a value of 10^{-1} is again no different for the mass flow. However, the valve area approximation is a little better than the input data here.

7.4 Comparison of Approximations

Figure 15 shows the superposition of the two log errors of the Fourier and Kalman approximations. As one can see, for the upstream and downstream pressures, the Kalman filter approach produced much better estimates, almost an order of magnitude better. However, the Fourier approximations produce an order of magnitude better than the Kalman filter with the mass flow and valve area.

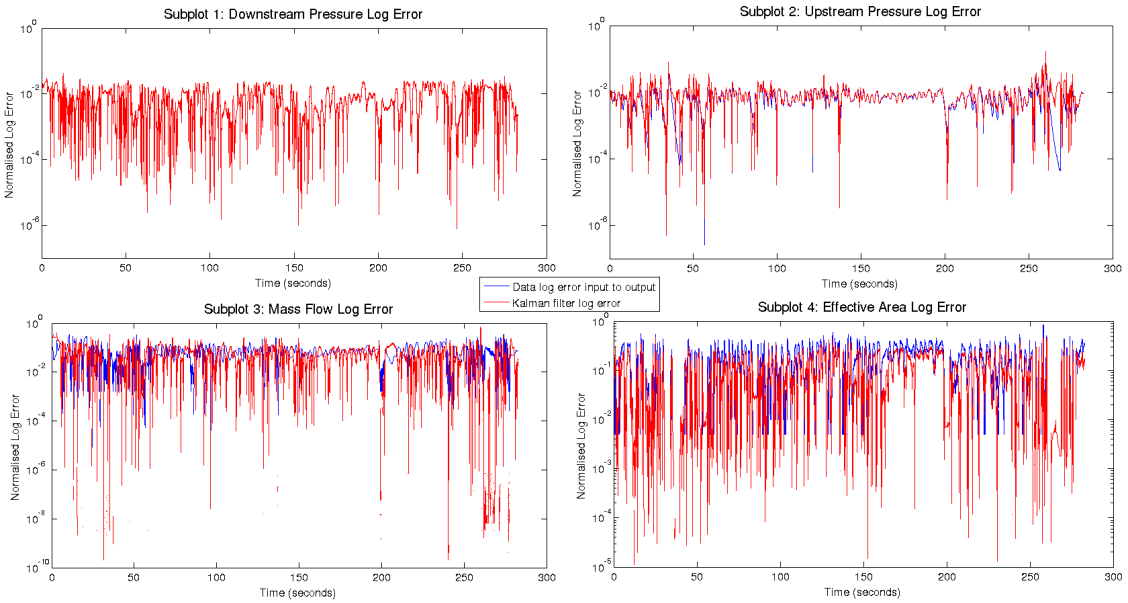


Figure 14: Log Normalised error in data and Kalman filter approximation

The latter may well be due to particularly poor measurement data for these data streams, or from the Kalman filter not being set correctly.

Numeric values comparing these methods are shown in tables in appendix A.

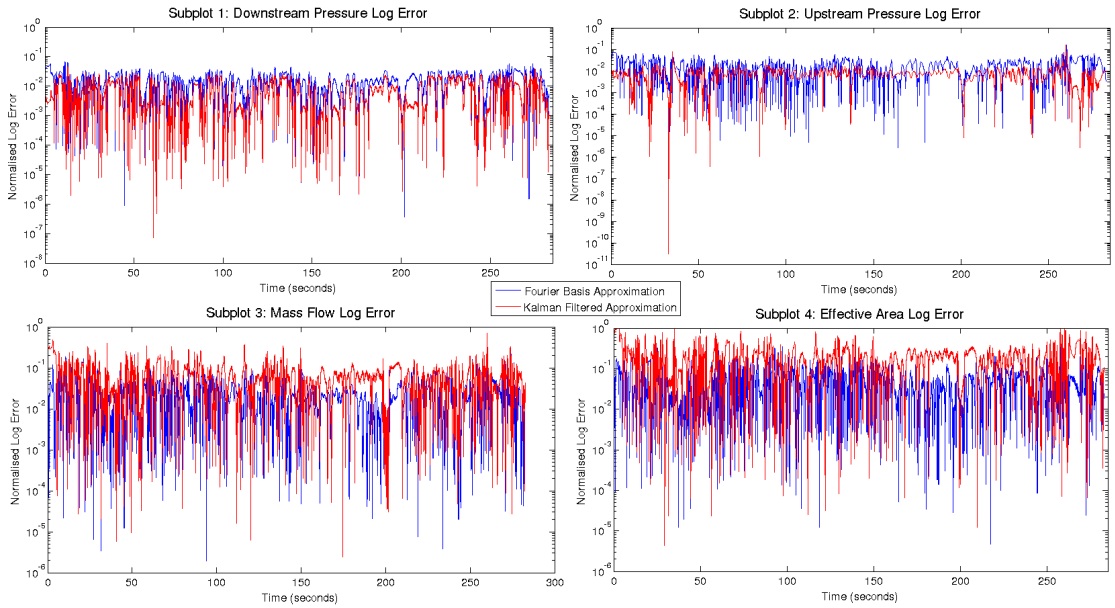


Figure 15: Shows the difference in log error of the Fourier and Kalman filter approximations

7.5 Conclusion

As one can see, methods have been given which produce good accuracy of estimation of the engine sample data. However, the important values of mass flow and valve area

still leave some improvement in the estimation, particularly with regards to the Kalman filtering method. It is clear from the log plots that an approach that involved a Fourier approximation method would be preferable to approximate the mass flow and effective valve area, as the approximations here were significantly better than the Kalman filter approximations, and were much better than the input data. It would be interesting to see whether the other methods suggested, namely the Bayesian approach, or the use of different spaces to minimise in, give a better approximation of the data.

A Approximation Data Tables

We include for reference some analysis of the various approximations that were made. Table 2 describes the Schauder approximation, table 3 the Fourier approximation, and table 4 the extended Kalman filtered approximation with the Fourier approximation as the prior.

	$\|\lfloor\cdot\rfloor\|_1$	$\|\lfloor\cdot\rfloor\|_2$	$\|\lfloor\cdot\rfloor\|_\infty$
P_D (hPa)	307.2364	466.7144	9.0285×10^3
P_U (hPa)	259.6355	40.2119	8.7715×10^3
M (kg/h)	3.7484	7.7528	321.3820
A (cm^2)	0.0666	0.1218	4.1875

Table 2: Differences in various norms between Schauder approximation and engineers data

	$\|\lfloor\cdot\rfloor\|_1$	$\|\lfloor\cdot\rfloor\|_2$	$\|\lfloor\cdot\rfloor\|_\infty$
P_D (hPa)	42.1523	55.7961	248.0025
P_U (hPa)	30.8054	38.0617	176.9373
M (kg/h)	1.6553	2.4824	44.2158
A (cm^2)	0.0432	0.0583	0.6852

Table 3: Differences in various norms between Fourier approximation and engineers data

	$\|\lfloor\cdot\rfloor\|_1$	$\|\lfloor\cdot\rfloor\|_2$	$\|\lfloor\cdot\rfloor\|_\infty$
P_D (hPa)	17.7113	22.7266	94.6994
P_U (hPa)	27.2795	33.5953	95.4926
M (kg/h)	3.6139	5.0557	43.5403
A(cm^2)	0.0786	0.1065	0.3840

Table 4: Differences in various norms between Kalman filter approximation and engineers data

B Slices of the Response Surface

We include as well for reference some images of the response surfaces. These are lower dimensional slices of the

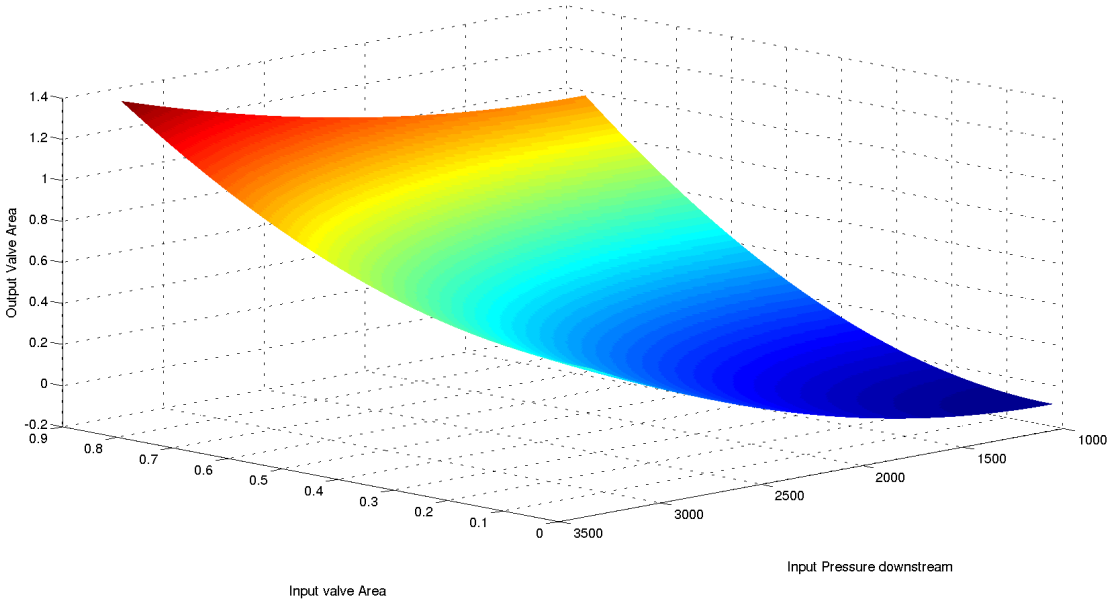


Figure 16: Response surface for $m = 40kg/h$ and $p_U = 1500hPa$ with output valve area

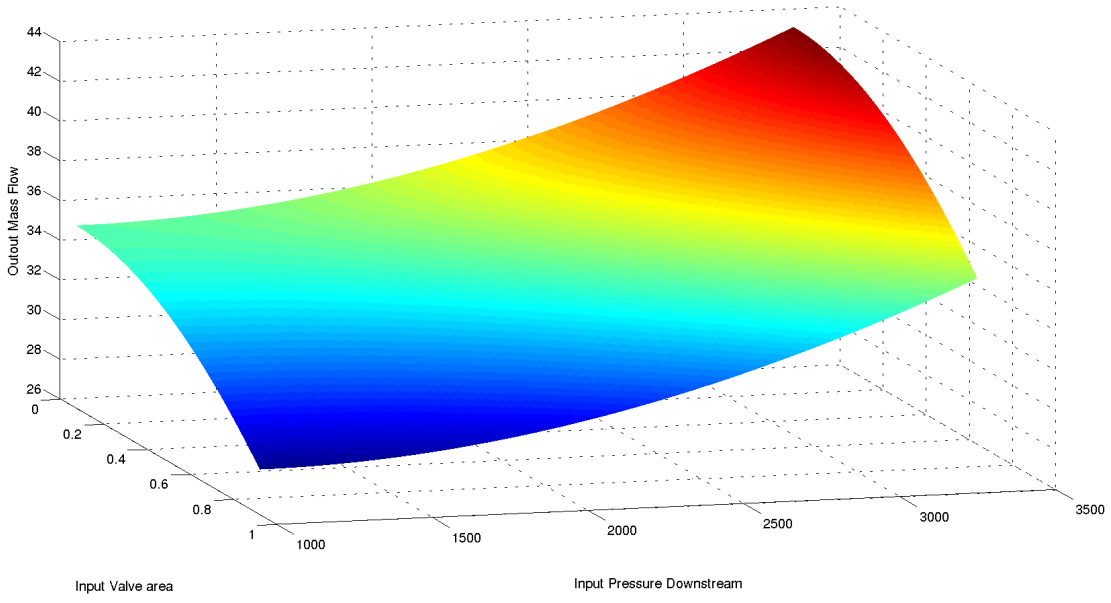


Figure 17: Response surface for $m = 40kg/h$ and $p_U = 1500hPa$ with output mass flow M

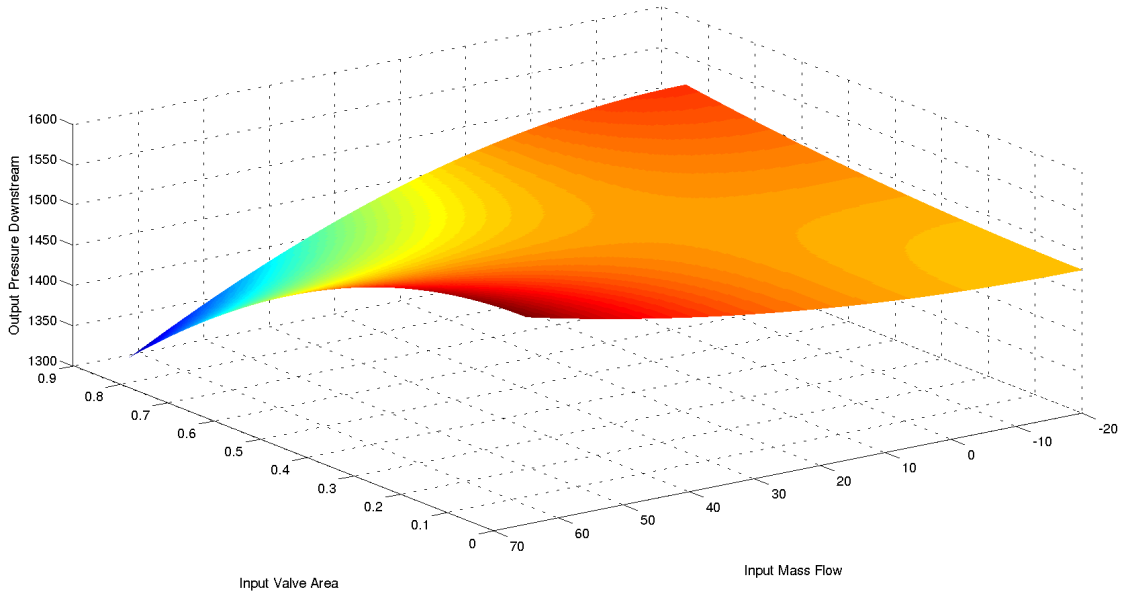


Figure 18: Response surface for $p_U = 1500$ and $p_D = 2000$ with output downstream pressure P_D

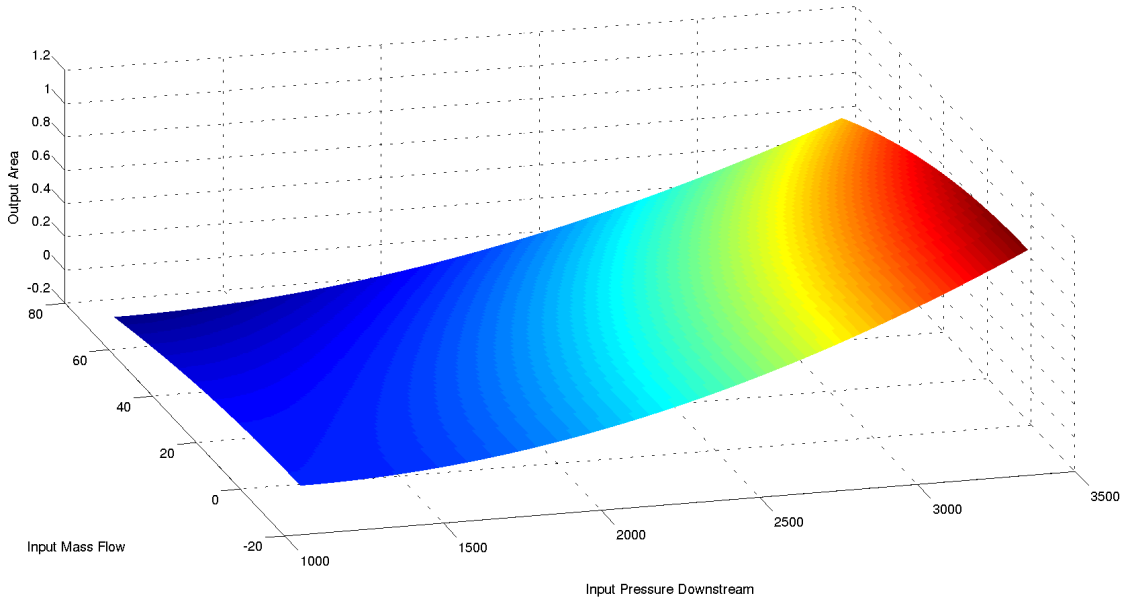


Figure 19: Response surface for $a = 0.4$ and $p_U = 1500$ with output valve area A

References

[1] Martin D Buhmann. Radial basis functions. *Acta Numerica 2000*, 9:1–38, 2000.

[2] S.L. Cotter, G.O. Roberts, A.M. Stuart, and D. White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28(3):424–446, 2013.

[3] G.A Einicke. *Smoothing, Filtering and Prediction - Estimating the Past, Present and Future*. InTech, 2012.

[4] Carl-Friedrich Gauss. *Theoria combinationis observationum erroribus minimis obnoxiae.-Gottingae, Henricus Dieterich 1823.* Henricus Dieterich, 1823.

[5] Ian Jolliffe. *Principal component analysis.* Wiley Online Library, 2005.

[6] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[7] T. Kariya and H. Kurata. *Generalized least squares.* John Wiley & Sons, 2004.

[8] Adrien Marie Legendre. *Nouvelles méthodes pour la détermination des orbites des cometes.* F. Didot, 1805.

[9] Michel Loève. *Probability Theory. Foundations. Random Sequences.* New York: D. Van Nostrand Company, 1955.

[10] R. L. Plackett. Some theorems in least squares. *Biometrika*, 37(1-2):149–157, 1950.

[11] C.R. Rao and H. Toutenburg. *Linear models.* Springer, 1995.

[12] T. Strutz. *Data fitting and uncertainty: a practical introduction to weighted least squares and beyond.* Vieweg and Teubner, 2010.

[13] A. M. Stuart. The Bayesian approach to inverse problems. *arXiv preprint arXiv:1302.6989*, 2013.