Multivariate Data Analysis Coursework

Elliot Jones (C21029768) Holly Ford (C1908545) Katherine Zverovich (C2029069)

April 2024

Contents

1 Introduction

This report aims to investigate properties of varying kernel regression models such as Gaussian, Laplace and Matérn-3/2. We will consider how the choice of kernel and its parameters impact the kernel regression model and how maximum-likelihood estimation and leave-one-out cross-validation can be performed to estimate the covariance kernel parameters.

We will also explore and experiment with the volcano data set. We will consider applying kernel regression, investigating the impact and optimisation of the kernel parameter ρ. We also will consider additional kernel parameters, such as a scale parameter α and a noise parameter σ^2 on the volcano data set. In addition to this, we will optimise the volcano data set for α and ρ in conjunction, as well as α, ρ and σ^2 simultaneously, to see the effects of these parameters.

2 First Experiments with Functions from $\mathbb{R}$ to $\mathbb{R}$

In this section we discuss building the confidence region related to a given Gaussian-process regression model, in the $\mathbb{R}$ to $\mathbb{R}$ case.

Consider the Gaussian kernel, we begin by discretising on the interval $[0, 10]$ and building our covariance matrix accordingly to our discretisation. By extracting the columns of the kernel matrix grid, we are able to fix $\tilde{x} \in \mathbb{R}$ and plot the function $x \to K(x, \tilde{x})$, this will allow us to simulate some realisations of the centered Gaussian process.

Selecting n sample locations on our grid and using a simulated realisation we can generate our response vector, and build our optimal prediction $x \to E(Zx|z = z)$. It is possible to then generate realisations of the conditional process and simulate a centered Gaussian process with covariance kernel C.

Using the formula in the lecture notes, we can now construct the 95%-confidence region of our Gaussian kernel regression model, see Figure 1. Plotting 100 realisations, we can use these to compare the validity our 95%-confidence interval. We observe that around 95 of the 100 realisations do indeed fall within our confidence interval region, thus leading us to believe that this 95%-confidence interval is accurate.

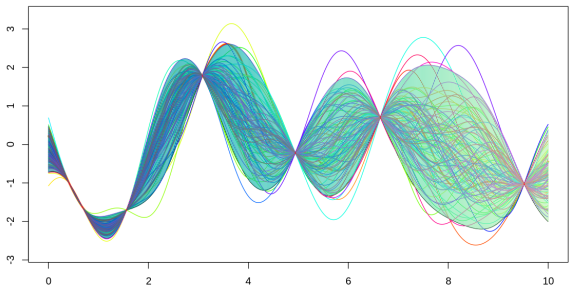

Figure 1: 95%-Confidence interval (blue) of the Gaussian kernel regression model with 100 realisations.

We can observe that the Gaussian kernel produces smooth conditional paths and that the confidence interval converges at our observations and expands as it moves further away from the observations. We would expect to observe this due to the fact that we have used these points to create our realisations. All of our realisations will pass through our observations and as the gap grows bigger between each observation, there will be a larger 95%-confidence interval and the prediction will become less uncertain due to the distance between observations.

3 Impact of Choice of Kernel and Kernel Parameters

In this section we investigate how the choice of kernel affects the kernel regression model. We have chosen to consider the Gaussian, Laplace and Matérn-3/2 kernels.

3.1 Choice of Kernel on the Kernel Regression Model

Using the previous execution of kernel regression, we can replicate this using the Laplace, and Matérn-3/2 kernels instead of the Gaussian kernel. We will again simulate 100 realisations to help verify this.

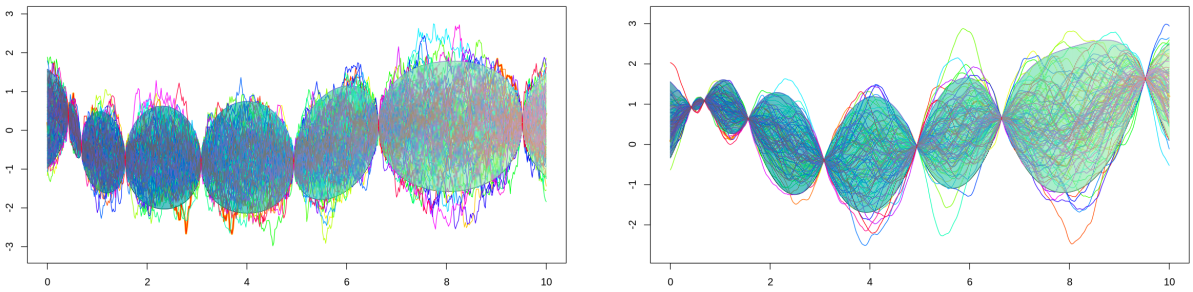

Figure 2: 95%-Confidence regions for Laplace (left) and Matérn-3/2 (right) kernel regression models. [could we use the structure that he has done in chapter 3 for his plots please?]

Comparing Figures 1 and 2 we can see that the Gaussian and Matérn-3/2 confidence regions follow a similar form, where the Gaussian model has slightly smoother realisations in contrast to the Matérn-3/2 kernel, whose curves hold less of a formation and have periods where they change their curvature slightly and tend to expand further than the Gaussian. We can also observe that the 95%-confidence region is slightly wider for the Matérn-3/2 kernel compared to the Gaussian kernel.

The Laplace model is the most dissimilar of the three, the confidence region when approaching the points remains wide whereas for Gaussian and Matérn-3/2 the confidence regions are tighter near the points. For Laplace we can see that the confidence interval is wide and shaped more like an oval, and at the observations we can see that the confidence interval expands very wide and quickly from the observations. This is in contrast to Gaussian and Matérn-3/2. The Matérn-3/2 is very similar to Gaussian but slightly different due to the curvature of the lines changing frequently throught the lines.

From observation it appears that the realisations of the Laplace kernel are more densely packed and more jagged in shape, making it more "chaotic" compared to the smooth conditional paths obtained from the Gaussian kernel. Thus it seems the Laplace kernel realisations resemble more similarities to that of a random walk rather than to the Gaussian or Matérn-3/2 kernels.

We can also observe that the Laplace kernel doesn't form as well to the confidence region as the Gaussian and Matérn-3/2 kernels. This can be seen as the confidence region between two close points compared to two far points is similar in size, whereas, we can see a significant change in the size of the confidence regions in these cases for the Gaussian and Matérn-3/2 kernels.

3.2 Impact of Additional Kernel Parameters

3.2.1 Scale Parameter α

We want to consider what happens when the original kernel K is replaced by αK, where α is an additional scale parameter. We will explore the effect of alpha using the Gaussian Kernel. We will consider the 95%-

confidence interval for the regression model for increasing values of α. Take, for example, the values α= 0.1 and $\alpha = 100$, respectively (Figure 3).

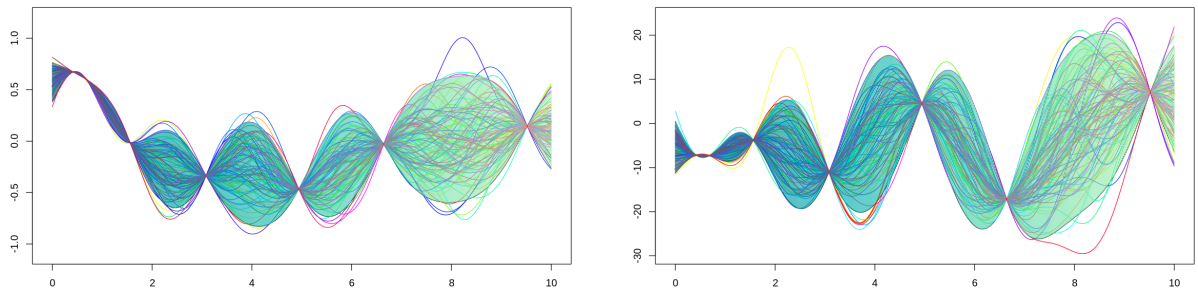

Figure 3: 95%-Confidence intervals for Laplace (left) and Matérn-3/2 (right) kernels.

We observe that α determines the scale of the kernel, so as the value of α rises, the amplitude also increases relative to this. This is presented in Figure 3. We can see that increasing the value of α stretches the confidence region vertically, such that it now spans over a larger y interval. We will later see how α can be estimated in conjunction with ρ on the volcano data set.

3.2.2 Noise Parameter σ^2

We want to consider adding a noise parameter σ^2 on the kernel regression model, so the kernel regression model is replaced by K+σ^2I. We are able to do this by adding $\sigma^2 I$ to the kernel matrix, where we make an identity matrix I which has the same dimension as the kernel matrix. We can build a 95%-confidence interval, which we can plot over a set of conditional paths, see Figure 4.

We will consider two cases, when σ^2=0.02 and σ^2=0.5.

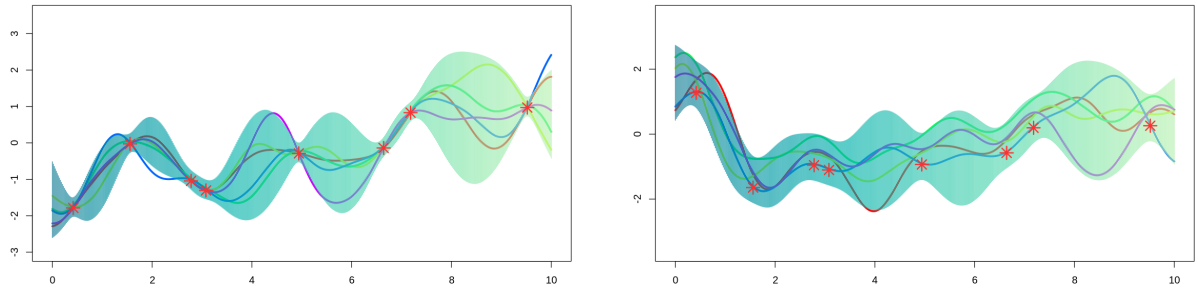

Figure 4: 95%-Confidence intervals the for Gaussian kernel regression model with added noise parameter for σ^2=0.02 (left) and σ^2=0.5 (right).

We can observe that close to the points, the confidence region becomes tighter but never converges at the observations, as it would without noise. By comparing the two graphs at a given point we can see that as more noise is added, the confidence region gets further from reaching the observations.

4 Maximum-likelihood Estimation of the Covariance Kernel Parameter(s)

In this section we explore the maximum-likelihood estimation of the covariance kernel parameter ρ. We aim to maximise the likelihood function

$$(L(p|z = z)) \tag{1}$$

so that we can observe which data is most probable under the statistical model. We use a Gaussian kernel with unknown parameter ρ, where $\rho > 0$. We calculate this in R by creating a function which takes ρ as our input and calculates the likelihood function on the given parameter. Inside this function we have initialise a matrix and find its inverse. We define two variables which calculate the terms-based determinant of the matrix and the exponential function, respectively. This generates the likelihood values with respect to ρ and we can use this to find the maximum-likelihood. Observing this in Figure 5 we can see that the distribution of ρ where maximum-likelihood is reached is approximately 0.8. The curve shows a steep incline to the maximum and a comparatively slow decline after reaching this, tailoring off steadily as ρ increases. This exact graph may vary based on the random realisations observed but the overall shape is what we would expect for the Gaussian kernel.

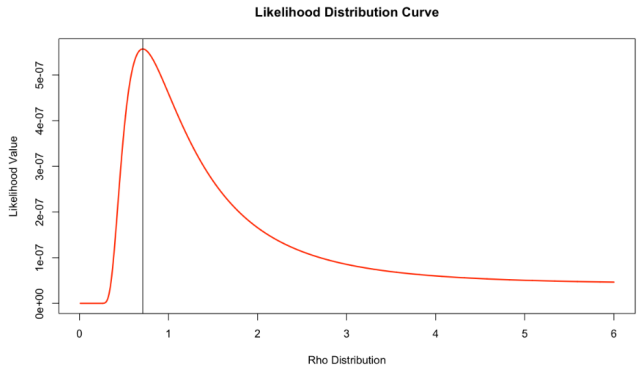

Figure 5: Likelihood curve for the covariance kernel parameter ρ.

Following this we can investigate obtaining an approximation of the maximum-likelihood estimator. The aim here is to take a fixed sample location and record the response vector for this. By slightly modifying our likelihood function to include our response vector we can see how the value of ρ evolves. We know the true value of ρ to be 1 but each time we repeat the experiment we will obtain a new value of ρ for the maximum-likelihood and our aim is to see how these values are distributed. To do this, we consider 3000 realisations and for each experiment we can generate the maximum likelihood and record the value of ρ that we obtain. We can then plot these maximum likelihood values of ρ to obtain a histogram. This can be done using six, eight and ten sample locations, the histograms obtained for these are shown in Figure 6, respectively. In theory as our value of ρ is equal to 1 we expect our estimations to accumulate around 1 given a large enough sample. We observe that as the number of sample locations increases, we can more precisely estimate ρ. When we have ten sample locations we have a higher frequency of results reaching the maximum-likelihood estimation at approximately $\rho = 0.8$. This is as opposed to 0.3 with six locations and 0.5 with eight locations. They each follow a similar curve with more of the maximum-likelihoods shown at lower values of ρ and then tailoring off to the end. It should be noted that due to our maximum likelihood estimation taking place over the interval $[0, 6]$ we would see a peak at 6. This is due to all values of ρ obtained that are larger than six being accumulated at six due to the restriction of our interval. We solved this by choosing not to display six in our results.

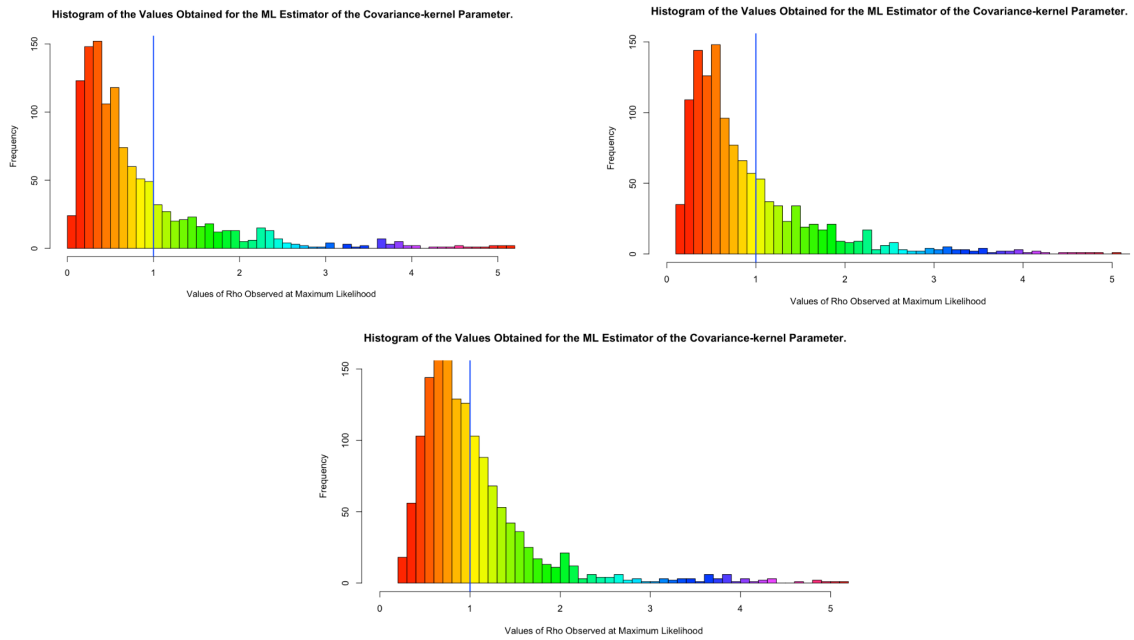

Figure 6: Histogram of values of ρ that achieve maximum-likelihood when number of sample locations are: six (top left), eight (top right), and ten (bottom).

5 Exploration of Volcano Data Set

5.1 Kernel Regression on Volcano Data

In this section, using a 50-point data set of a 50 x 3 matrix with information about a volcano, we will use kernel regression to predict z as a function of x and y and visually represent this. The data used consists of the (x, y, z) coordinates of 50 points in $[0, 1]^2$ x $\mathbb{R}$. The first two columns in our dataset are the sample locations, $\{x_1, \cdots, x_{50}\} \subset [0, 1]^2$ and the last column is the response vector. Using these we can apply kernel regression and known formulas to predict the shape of the volcano. Once the dataset has been separated into appropriate variables, we can begin by defining a set of M grid points $\{g_1, \cdots, g_M\} \subset [0, 1]^2$, as we want to predict the shape of the volcano at all grid points from the data set. We choose M to be a 40 x 40 matrix in our case and use the sequence operation built into R to achieve this. We will model this using our Gaussian kernel, previously denoted as K, while choosing $\rho = 35$, to ensure a sensible prediction of the volcano. Using known theory we can gain the optimal prediction p_l at the grid point g_l, where $l \in \{1, \cdots, M\}$, is

$$p_l = k^T(g_l)K^{-1}z$$

and $k(g_l) = ((K(x_1, g_l), \cdots, K(x_n, g_l))^T$, where K is the n x n kernel matrix defined by kernel k and the sample locations, for the volcano data set $n = 50$. Using the sample locations from our data set we can build the matrix K with i, j entries $K(x_i, x_j)$ for $i, j \in \{1, \cdots, n\}$. We can then compute 1 x n row vectors $k^T(g_l)$ for all $l \in \{1, \cdots, M\}$ (for all the grid points). We can do this by computing all M vectors at once and storing them in an M x n matrix labelled C, where C is the cross-covariance matrix. In order to do this we use the Gaussian kernel denoted K, the set of grid points g_i previously defined, and the sample locations x_j from our dataset, where $i \in \{1, \cdots, M\}$ and $j \in \{1, \cdots, n\}$. Using this we can define a vector $P = (p_1, \cdots, p_M)^T \in \mathbb{R}^M$ to store the prediction p_l of the volcano at grid points g_l, with $l \in \{1, \cdots, M\}$. This is mathematically computed within the code as

$$p = CK^{-1}z$$

Finally, as we have obtained our prediction points from the volcano dataset using kernel regression, we can visualise the predicted shape of the volcano. Choosing $\rho = 35$ as our Gaussian kernel parameter, we can see the shape of the volcano with the observations alongside a heat-map showing the depth of the volcano, see Figure 7.

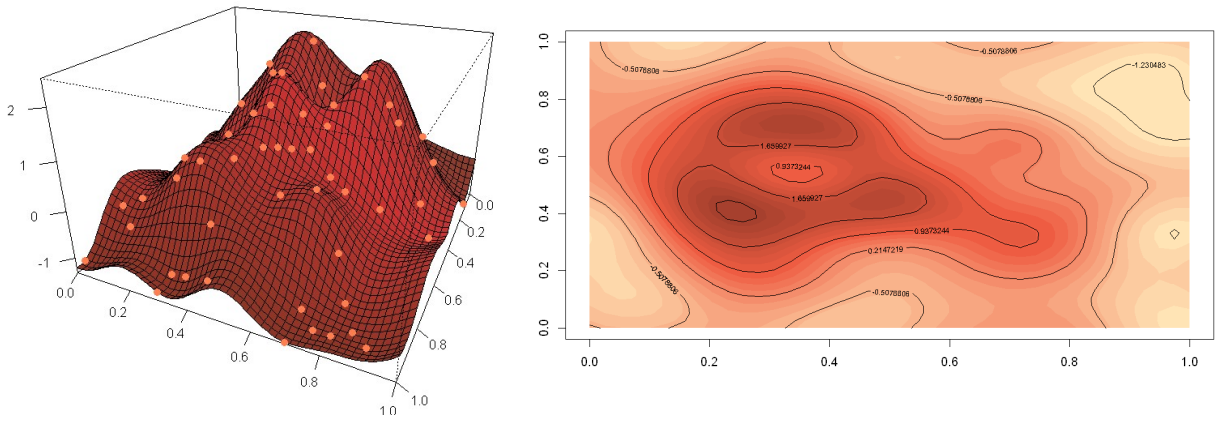

Figure 7: Visualisation of the predicted shape of the volcano with observations (orange spheres) and its respective heat-map for a Gaussian kernel.

5.2 Impact and Optimisation of ρ on Kernel Regression of the Volcano Dataset

To continue our investigations, we can experiment with the effect ρ has on predicting the volcano from the data set. As we can see in Figure 8, when $\rho = 1.2$ our prediction of the volcano becomes like a flat plane, however at two particular points, goes up and down drastically. We can notice this effect when we compare the scale of the Z axis of [-500,500] compared to the other plots of [-2,2]. This suggest that with this ρ we have significant over-fitting of the predicted values. When we increase the value to $\rho = 50$, we can see that the plot displays a lot more "mountainous" like peaks and begins to resemble that of a volcano shaped structure. We can say that this value of ρ is neither over-fitted or under-fitted but nor is it optimal. Finally, we can see that at an extreme value of $\rho = 500$ there are extremely sharp points and troughs. This is due to under-fitting of the prediction. We can see that the graph is going directly to the prediction points and not fitting a sensible prediction in-between these points. We can conclude from this that when ρ is extremely small, the prediction points are over-fitted. However when ρ is extremely large, the prediction points are under-fitted.

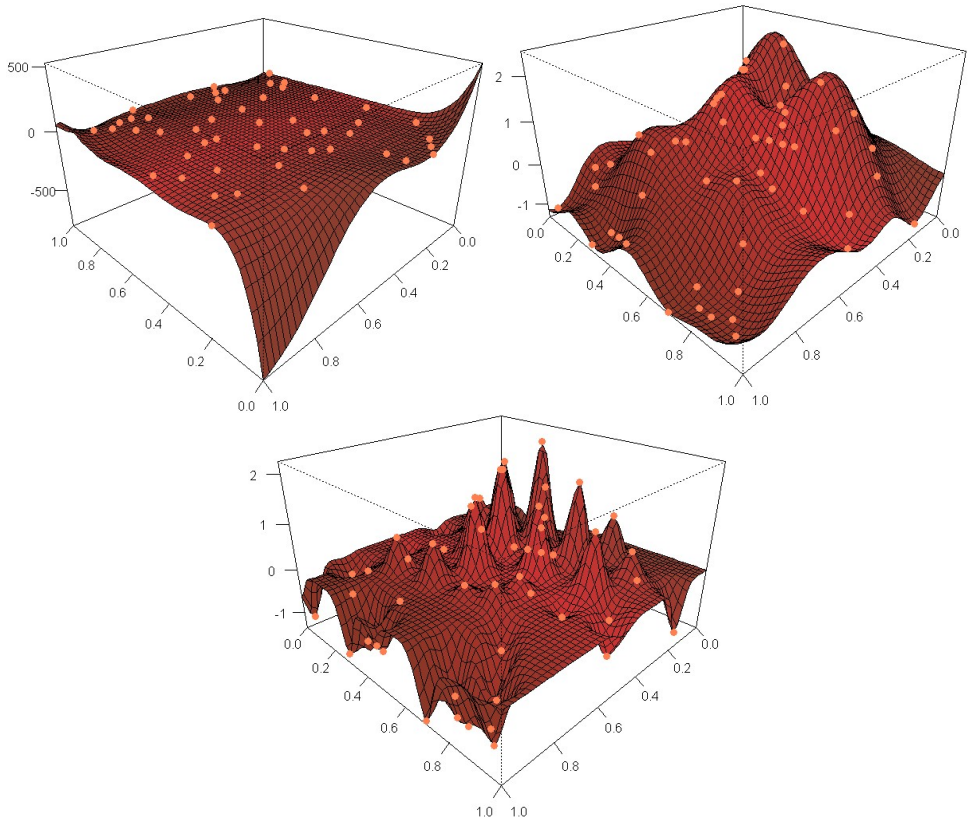

Figure 8: Visualisation of the predicted shape of the volcano with predicted ρ values of, $\rho = 1.2$ (top left), $\rho = 50$ (top right), $\rho = 500$ (bottom), with observations (orange spheres)

Furthermore, we can try to achieve the optimal value for ρ that would output the best prediction of the volcano using maximum-likelihood. Using our sample locations for the volcano and the Gaussian kernel, we can build a function such that we can compute the maximum-likelihood of ρ given a number of observations. Consider, for example, 3000 observations of ρ to be analysed. This results in a maximum-likelihood with $\rho = 34.2187$ (see Figure 9). Plugging this into our Gaussian kernel we may observe the volcano plot and heat-map shown in Figure 10. Using this value of ρ for our volcano prediction, we can identify that the highest point of the volcano is more defined and better connected at the peak than the one produced with $\rho = 35$. However, it is only marginally different. This is because as we can see from the maximum-likelihood estimation plot, there are multiple values which have a high likelihood for ρ and as such 35 is very close to the optimal so it will still produce a close-to optimal prediction.

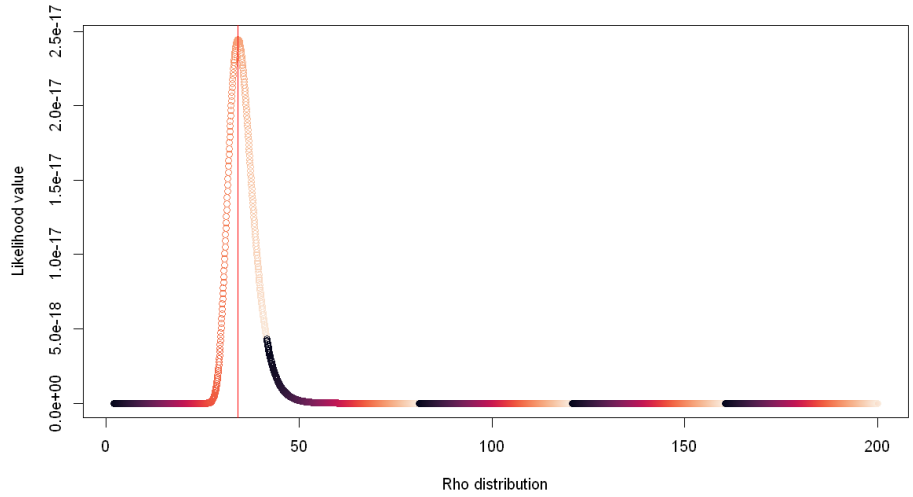

Figure 9: Likelihood curve for predicting the optimal value of ρ of the volcano.

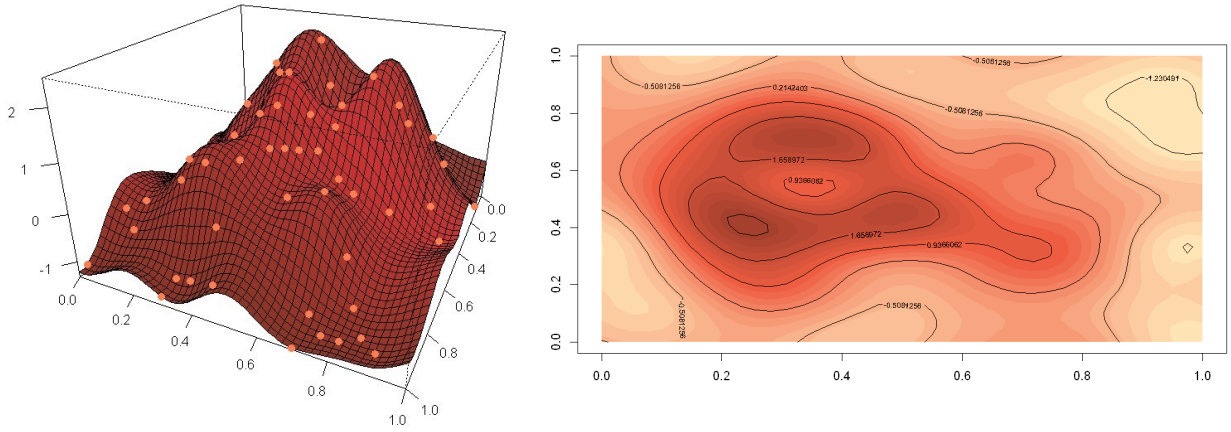

Figure 10: Volcano plot and heat-map for our Gaussian kernel at the optimal value $\rho = 34.2187$.

5.3 Kernel regression with αK and optimising the Volcano dataset for α and ρ

As seen previously in Section 3, we have considered the impact of multiplying an additional parameter α to our kernel. Following on from this we can use maximum-likelihood to estimate the optimal value for both ρ and α simultaneously. Our method involves testing multiple combinations of α and ρ values, while trying to maximise the likelihood of the observed data. Through this process, we can derive the optimal values: $\alpha = 0.8979798$ and $\rho = 31.87879$. Figure 11 graphically demonstrates this optimal point.

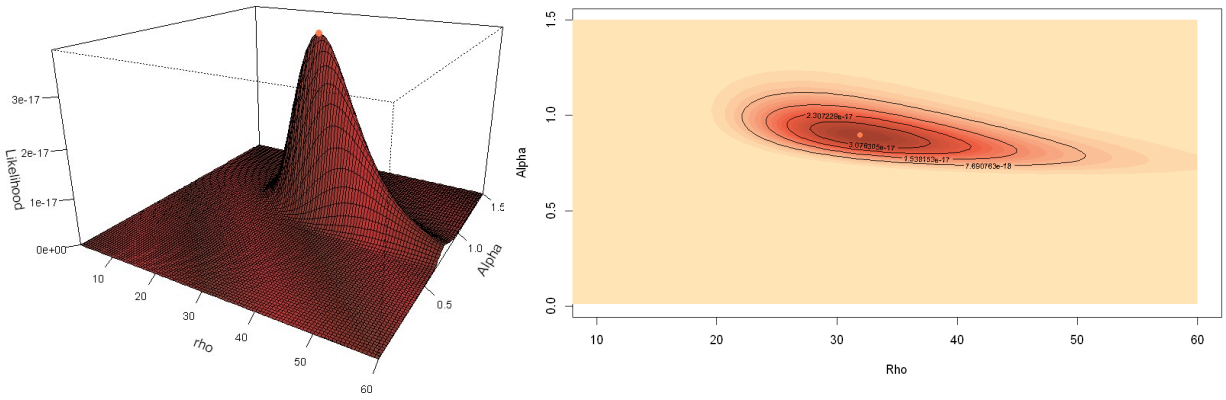

Figure 11: Maximum-likelihood estimation of the optimal value of α and ρ simultaneously (orange).

These plots demonstrate the optimised combination of α and ρ, these are crucial in ensuring the models predictive accuracy. This experiment not only allows us to analyse the effect kernel scaling has but also allows us to optimise our parameters for a given data set. We can visualise the impact these values have by applying the same kernel regression to the volcano dataset using these specified parameters for αK. We can see in Figure 12 that there is a small change in our volcano which uses the optimal values compared to that of the volcano with values $\alpha = 1$ and $\rho = 35$. This is due to the fact that the initial produced volcano had values of ρ chosen such that they were still close to the optimal prediction. This can be seen from the heat map for optimal values for ρ and α in Figure 11. Overall, we may conclude that it is critical that we optimise our parameters of the kernel, as choosing the wrong value of parameters might lead to over-fitting or under-fitting of the kernel regression prediction.

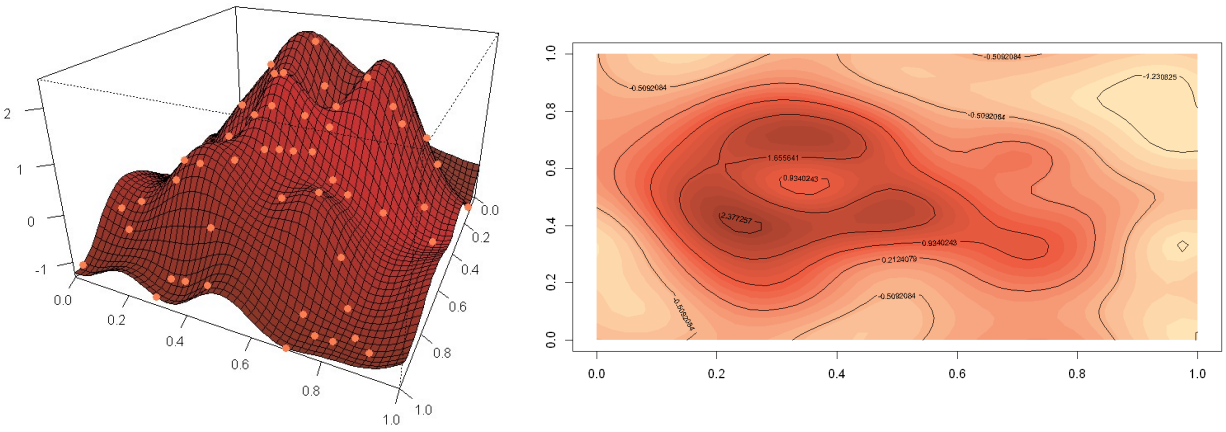

Figure 12: Volcano plot and heat-map for our Gaussian kernel at the optimal values $\alpha = 0.8979798$ and $\rho = 31.87879$.

5.4 Estimating the Volcano Data Set on the Matérn-3/2 and Laplace Kernels

To measure how well the Gaussian kernel models our data, we can look into how other kernels process the volcano data set. We may consider, for example, the Laplace and Matérn-3/2 kernels. To accommodate for the different kernels we will observe the volcano when $\rho = 1$ and $\rho = 10$, respectively. These values allow us

to more accurately represent the volcano as we found when using ρ=35 both the models were under-fitted. (see Figure 13).

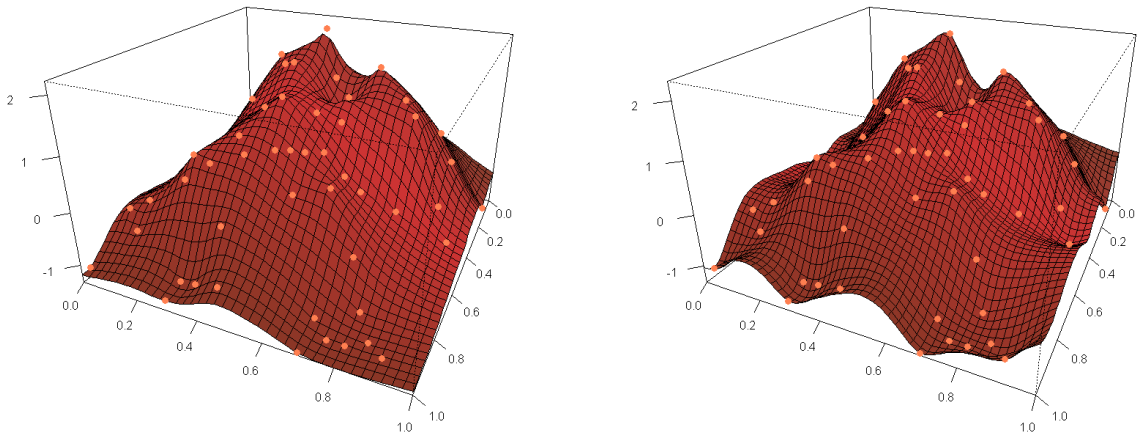

Figure 13: Visualisations of the predicted shape of the volcano for a Laplace kernel with $\rho = 1$ (left), and Matérn-3/2 kernel with $\rho = 10$ (right).

Immediately we can see that both of these volcanoes have less curvature than that created from the Gaussian kernel. The Laplace volcano emphasises points more directly with sharper and straighter lines compared to the volcanoes produced by the other kernels. The Matérn-3/2 prediction produces a volcano almost in between these two, with more curves than Laplace but with steeper peaks whereas Gaussian has a more natural volcano like curve. Overall this may incline us to conclude that the Gaussian prediction is the most accurate estimator out of the three kernels tested for the volcano data set.

5.5 Estimating σ^2, α and ρ simultaneously on the Volcano Data Set

To conclude our investigations into the volcano data set we want to estimate the values of ρ, α, and σ^2 simultaneously. To do so we create our function using maximum likelihood estimation on the variables ρ, σ^2 and α using similar methodology as in the previous sections. However, due to us estimating in an additional dimension on a 2-dimensional matrix this would not suffice. In order to change this we used a 3-Dimensional array to calculate and store our maximum likelihood predictions. We could display these results on a 3-Dimensional heat map however due to this figure having to be printed out and not being able to move as it would on a computer, this image would not be very informative and as such has been omitted. The results of our optimal predictions were as follows, $\alpha = 1.157895, \rho = 32.60526, \sigma^2 = 0.3157895$. Our scale parameter, α, is very close to 1 implying that the data is already close to the optimal scale, it has an almost negligible effect on the prediction. Furthermore, it can be noted that our noise parameter σ^2 is relatively small. However it still implies that adding some noise to our data does indeed help the prediction, and that our value of ρ is similar to the optimal when estimating it alone. Using these optimal values in conjunction we can reproduce the volcano and its heat map. Given the above analysis it is perhaps not surprising to see an almost identical volcano to the one produced in previous sections.

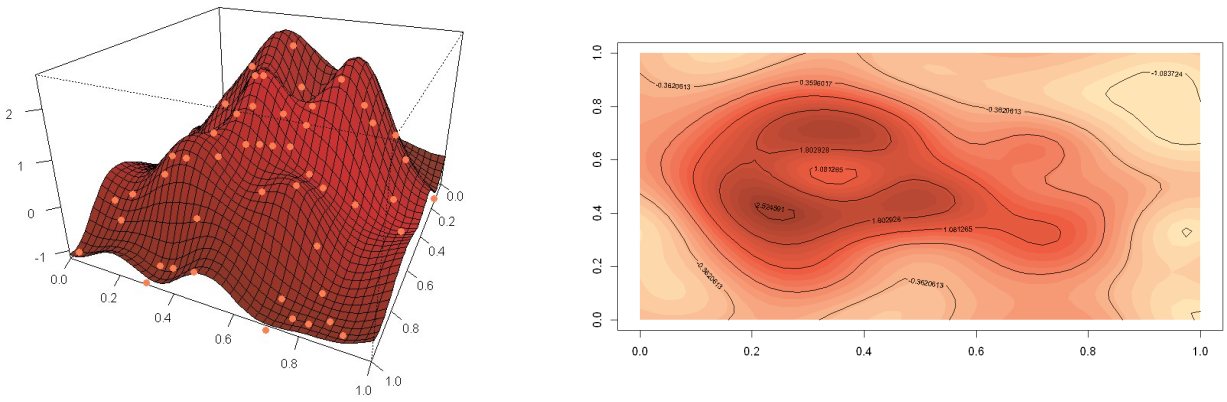

Figure 14: Volcano plot and heat-map for our Gaussian kernel at the optimal values $\alpha = 1.157895$, $\rho = 32.60526$ and $\sigma^2 = 0.3157895$.

6 Experiments on Leave-one-out Cross-validation (LOO-CV)

To further advance our investigations into kernel parameters we can conduct experiments using leave-one-out cross-validation on the Gaussian kernel. This algorithm works by removing any one entry of the data set, in our case the sample location and corresponding vector. Using the remaining entries we can predict this point and then determine how close the predicted value is to the true value. This algorithm can be applied to all data points and then the results can be combined to give an overall validation score. Here we aim to minimise the score as a function of ρ, which in turn identifies the optimal value of ρ. Minimising the score will determine how reliable the kernel is in predicting entries, given a data set and training.

To do this in R we begin by initialising our data which involves checking the length of our response vector and creating a square matrix of zeroes, where our covariance values will go. We can use the loop function to calculate the co-variances between each pair of sample locations, which we then use to store our prediction errors. Finally, the function returns the sum of absolute squares. To visualise this graphically we again use the loop function to calculate the LOO-CV score for each value of ρ. This is illustrated in Figure 15. From this we can identify that the minimum score is 6.752 achieved at $\rho = 38$, thus we may conclude that this is the optimal value of ρ from the LOO-CV score. From Figure 15 we can see that there is a sharp increase in the score on either side of the optimal value. This implies that there is only a small range of ρ for which the model reliably works, as this is where we have struck the best balance between bias and variance. This pattern occurs because at low values of ρ, over-fitting occurs as opposed to under-fitting at large values of ρ. These correspond to not considering enough of the observations or too much of the observations, respectively. Using this optimal value of ρ we can produce the volcano seen in Figure 15. We may observe the overall shape looks like very similar to our optimal volcano seen earlier.

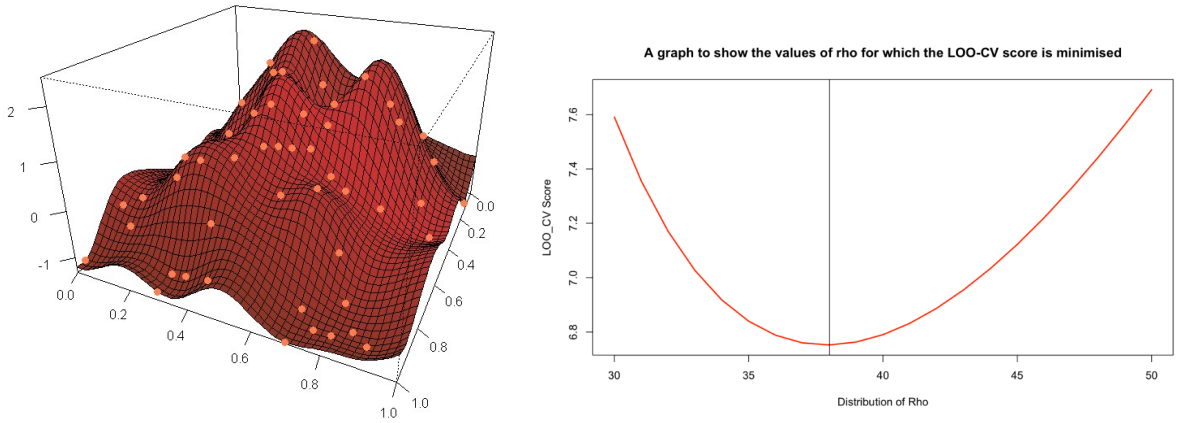

Figure 15: Leave-one-out cross-validation curve which finds the value of ρ that optimises the cross-validation score and the respective volcano prediction for this value.

This method can take a long time for the algorithm to calculate when our number of observations n is large, as it must go through each possible value of n. An alternative method is to approximate the leave-one-out cross-validation by k-fold-cross validation. This involves dividing the sample into equally sized sub-samples, k, however we could not achieve this due to our dataset being so small so it was pointless to perform this.

7 Conclusion

Throughout this report we have explored and analysed various results including, but not limited to, the impact of various kernel regression models; the impact of maximum-likelihood estimations, the impact of adding additional kernel parameters, using leave-one-out cross-validation, and the effects of observational noise. All of this has helped aid us in our exploration and visualisation of the volcano data set and of the theory we have covered within lectures. To further our investigations it would be interesting to explore the effects of kernel regression with finite dimensional kernels, considering specific functions, and introducing observational noises to this to see how the model reacts.