An automatic report for the dataset : data

The Automatic Statistician

January 31, 2018

Abstract

This report was produced by the Automatic Bayesian Covariance Discovery (ABCD) algorithm.

1 Executive summary

The raw data and full model posterior with extrapolations are shown in figure **??**.

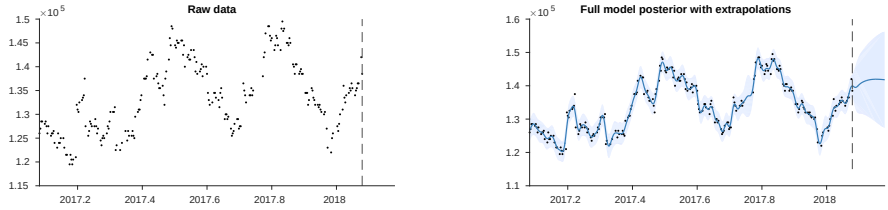

Figure 1: Raw data (left) and model posterior with extrapolation (right)

The structure search algorithm has identified four additive components in the data. The first 3 additive components explain 97.7% of the variation in the data as shown by the coefficient of determination (R^2) values in table **??**. The 4 additive components explain 100.0% of the variation in the data. After the first 3 components the cross validated mean absolute error (MAE) does not decrease by more than 0.1%. This suggests that subsequent terms are modelling very short term trends, uncorrelated noise or are artefacts of the model or search procedure. Short summaries of the additive components are as follows:

- A linearly increasing function.
- A smooth function.
- A smooth function.
- Uncorrelated noise.

#	R^2 (%)	ΔR^2 (%)	Residual R^2 (%)	Cross validated MAE	Reduction in MAE (%)
-	-	-	-	133788.07	-
1	18.5	18.5	18.5	5621.81	95.8
2	85.7	67.2	82.4	3584.75	36.2
3	97.7	12.0	84.1	3498.10	2.4
4	100.0	2.3	100.0	3498.10	0.0

Table 1: Summary statistics for cumulative additive fits to the data. The residual coefficient of determination (R^2) values are computed using the residuals from the previous fit as the target values; this measures how much of the residual variance is explained by each new component. The mean absolute error (MAE) is calculated using 10 fold cross validation with a contiguous block design; this measures the ability of the model to interpolate and extrapolate over moderate distances. The model is fit using the full data and the MAE values are calculated using this model; this double use of data means that the MAE values cannot be used reliably as an estimate of out-of-sample predictive performance.

Model checking statistics are summarised in table **??** in section **??**. These statistics have revealed statistically significant discrepancies between the data and model in component 4.

The rest of the document is structured as follows. In section **??** the forms of the additive components are described and their posterior distributions are displayed. In section **??** the modelling assumptions of each component are discussed with reference to how this affects the extrapolations made by the model. Section **??** discusses model checking statistics, with plots showing the form of any detected discrepancies between the model and observed data.

2 Detailed discussion of additive components

2.1 Component 1 : A linearly increasing function

This component is linearly increasing.

This component explains 18.5% of the total variance. The addition of this component reduces the cross validated MAE by 95.8% from 133788.1 to 5621.8.

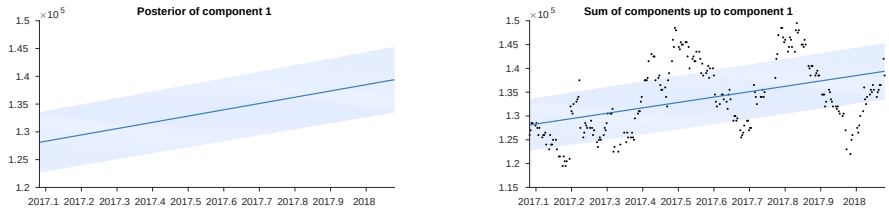

Figure 2: Pointwise posterior of component 1 (left) and the posterior of the cumulative sum of components with data (right)

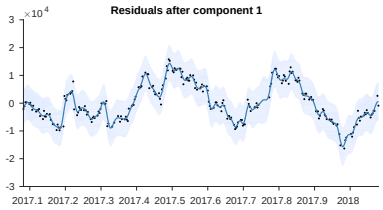

Figure 3: Pointwise posterior of residuals after adding component 1

2.2 Component 2 : A smooth function

This component is a smooth function with a typical lengthscale of 4.0 weeks.

This component explains 82.4% of the residual variance; this increases the total variance explained from 18.5% to 85.7%. The addition of this component reduces the cross validated MAE by 36.24% from 5621.81 to 3584.75.

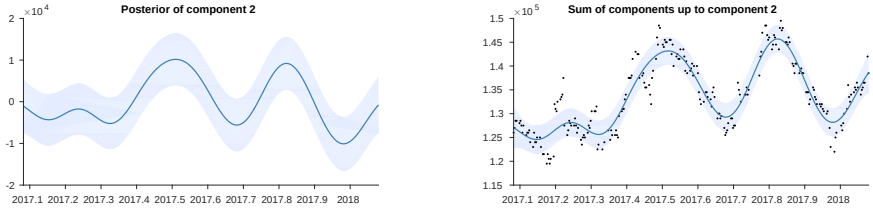

Figure 4: Pointwise posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

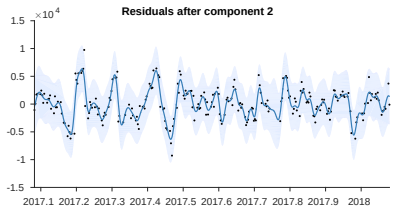

Figure 5: Pointwise posterior of residuals after adding component 2

2.3 Component 3 : A smooth function

This component is a smooth function with a typical lengthscale of 3.3 days.

This component explains 84.1% of the residual variance; this increases the total variance explained from 85.7% to 97.7%. The addition of this component reduces the cross validated MAE by 2.42% from 3584.75 to 3498.10.

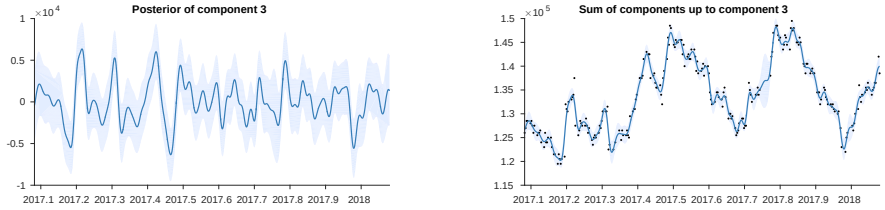

Figure 6: Pointwise posterior of component 3 (left) and the posterior of the cumulative sum of components with data (right)

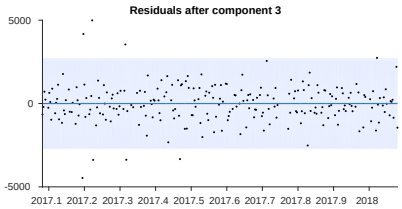

Figure 7: Pointwise posterior of residuals after adding component 3

2.4 Component 4 : Uncorrelated noise

This component models uncorrelated noise.

This component explains 100.0% of the residual variance; this increases the total variance explained from 97.7% to 100.0%. The addition of this component reduces the cross validated MAE by 0.00% from 3498.10 to 3498.10. This component explains

residual variance but does not improve MAE which suggests that this component describes very short term patterns, uncorrelated noise or is an artefact of the model or search procedure.

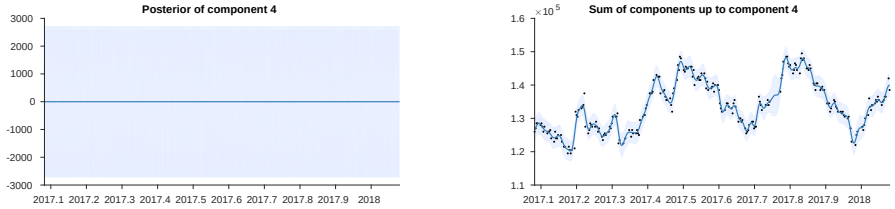

Figure 8: Pointwise posterior of component 4 (left) and the posterior of the cumulative sum of components with data (right)

3 Extrapolation

Summaries of the posterior distribution of the full model are shown in figure **??**. The plot on the left displays the mean of the posterior together with pointwise variance. The plot on the right displays three random samples from the posterior.

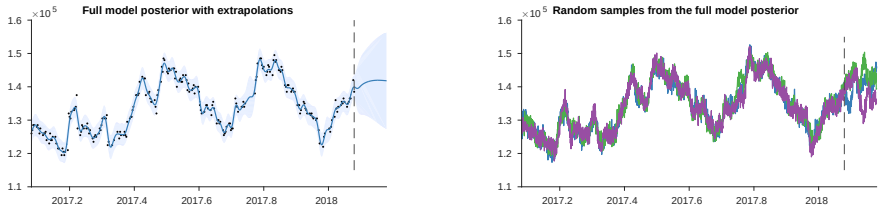

Figure 9: Full model posterior with extrapolation. Mean and pointwise variance (left) and three random samples (right)

Below are descriptions of the modelling assumptions associated with each additive component and how they affect the predictive posterior. Plots of the pointwise posterior and samples from the posterior are also presented, showing extrapolations from each component and the cuulative sum of components.

3.1 Component 1 : A linearly increasing function

This component is assumed to continue to increase linearly.

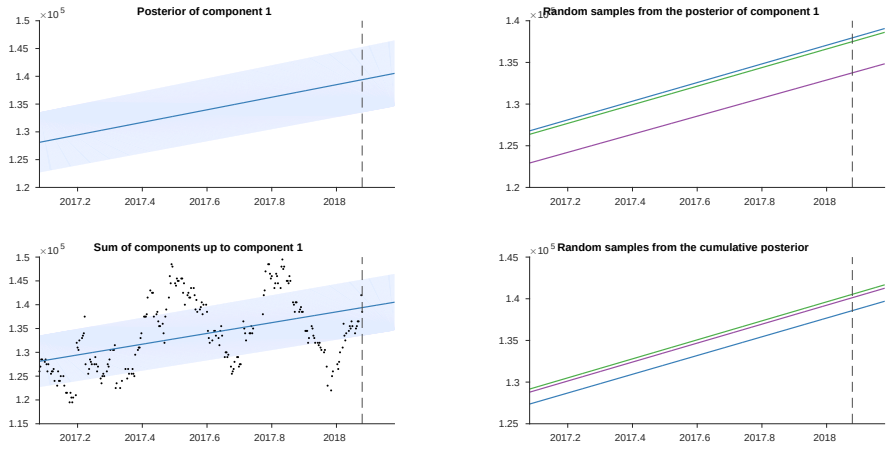

Figure 10: Posterior of component 1 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.2 Component 2 : A smooth function

This component is assumed to continue smoothly but is also assumed to be stationary so its distribution will return to the prior. The prior distribution places mass on smooth functions with a marginal mean of zero and a typical lengthscale of 4.0 weeks. [This is a placeholder for a description of how quickly the posterior will start to resemble the prior].

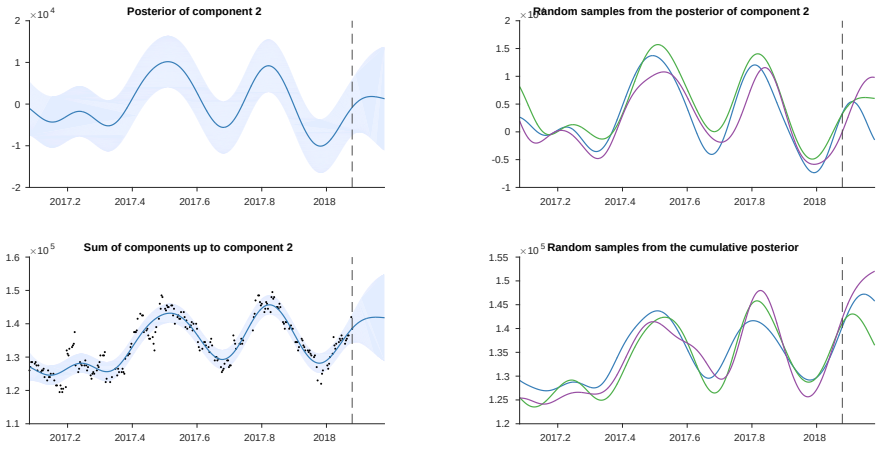

Figure 11: Posterior of component 2 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.3 Component 3 : A smooth function

This component is assumed to continue smoothly but is also assumed to be stationary so its distribution will return to the prior. The prior distribution places mass on smooth functions with a marginal mean of zero and a typical lengthscale of 3.3 days. [This is a placeholder for a description of how quickly the posterior will start to resemble the prior].

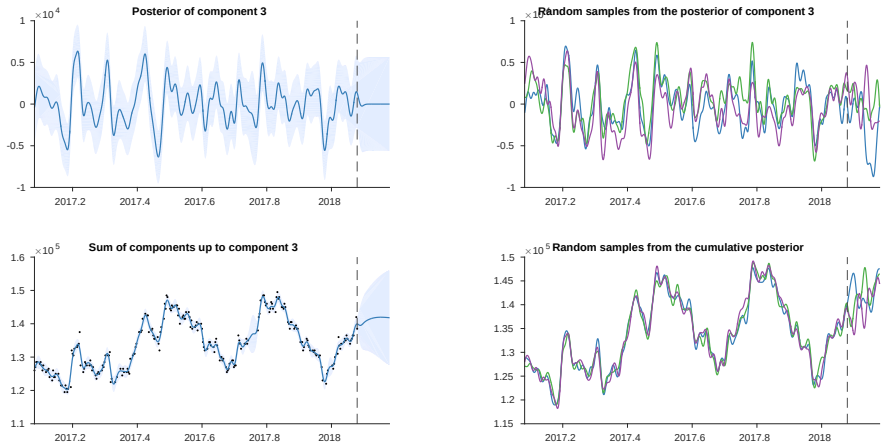

Figure 12: Posterior of component 3 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.4 Component 4 : Uncorrelated noise

This component assumes the uncorrelated noise will continue indefinitely.

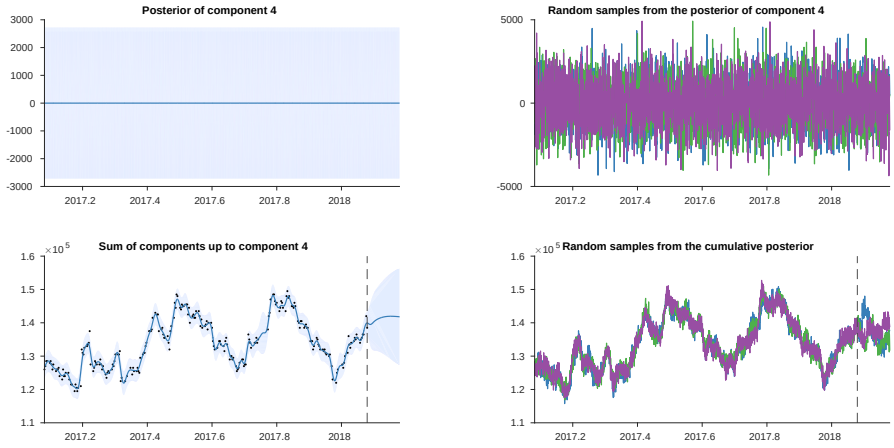

Figure 13: Posterior of component 4 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

4 Model checking

Several posterior predictive checks have been performed to assess how well the model describes the observed data. These tests take the form of comparing statistics evaluated on samples from the prior and posterior distributions for each additive component. The statistics are derived from autocorrelation function (ACF) estimates, periodograms and quantile-quantile (qq) plots.

Table **??** displays cumulative probability and p-value estimates for these quantities. Cumulative probabilities near 0/1 indicate that the test statistic was lower/higher under the posterior compared to the prior unexpectedly often i.e. they contain the same information as a p-value for a two-tailed test and they also express if the test statistic was higher or lower than expected. p-values near 0 indicate that the test statistic was larger in magnitude under the posterior compared to the prior unexpectedly often.

	ACF		Periodogram		QQ	
#	min	min loc	max	max loc	max	min
1	0.456	0.483	0.709	0.500	0.150	0.850
2	0.303	0.196	0.393	0.777	0.463	0.518
3	0.321	0.456	0.691	0.591	0.501	0.567
4	0.500	0.478	0.493	0.493	0.048	0.154

Table 2: Model checking statistics for each component. Cumulative probabilities for minimum of autocorrelation function (ACF) and its location. Cumulative probabilities for maximum of periodogram and its location. p-values for maximum and minimum deviations of QQ-plot from straight line.

The nature of any observed discrepancies is now described and plotted and hypotheses are given for the patterns in the data that may not be captured by the model.

4.1 Moderately statistically significant discrepancies

4.1.1 Component 4 : Uncorrelated noise

The following discrepancies between the prior and posterior distributions for this component have been detected.

- The qq plot has an unexpectedly large positive deviation from equality ($x = y$). This discrepancy has an estimated p-value of 0.048.

The positive deviation in the qq-plot can indicate heavy positive tails if it occurs at the right of the plot or light negative tails if it occurs as the left.

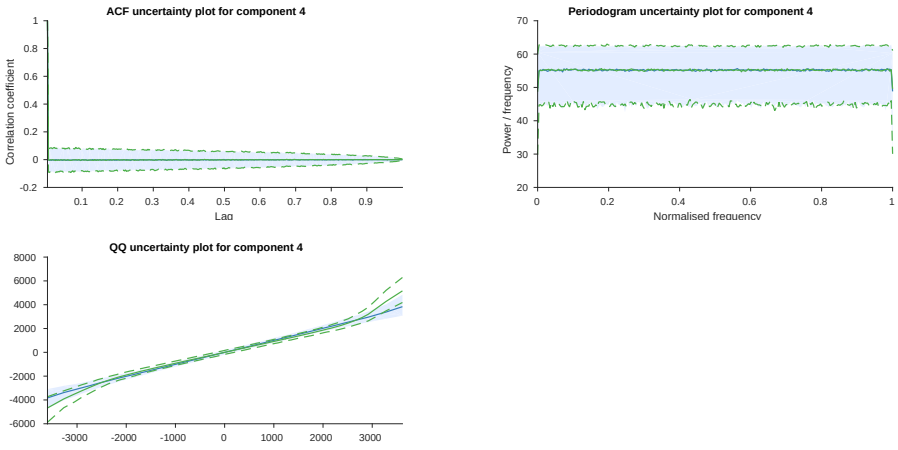

Figure 14: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 4. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2 Model checking plots for components without statistically significant discrepancies

4.2.1 Component 1 : A linearly increasing function

No discrepancies between the prior and posterior of this component have been detected

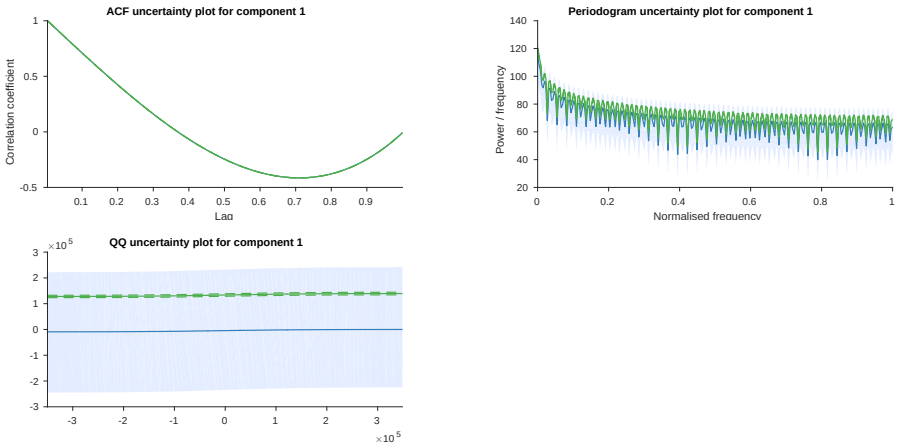

Figure 15: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 1. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2.2 Component 2 : A smooth function

No discrepancies between the prior and posterior of this component have been detected

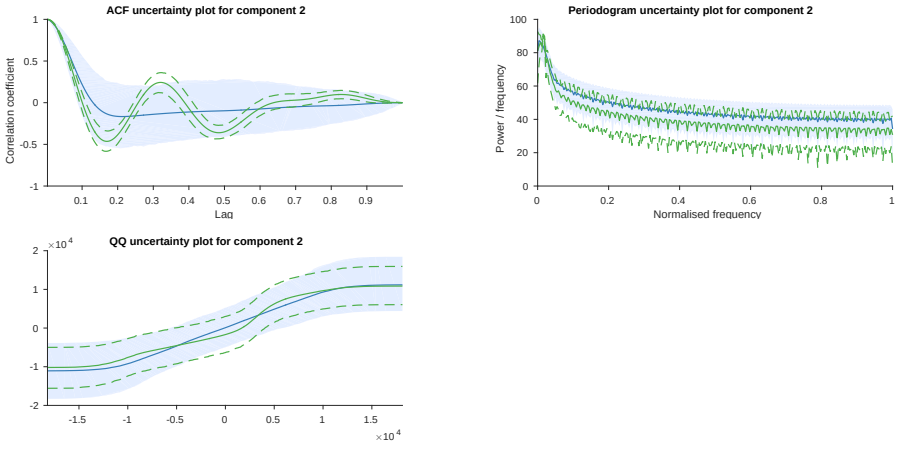

Figure 16: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 2. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2.3 Component 3 : A smooth function

No discrepancies between the prior and posterior of this component have been detected

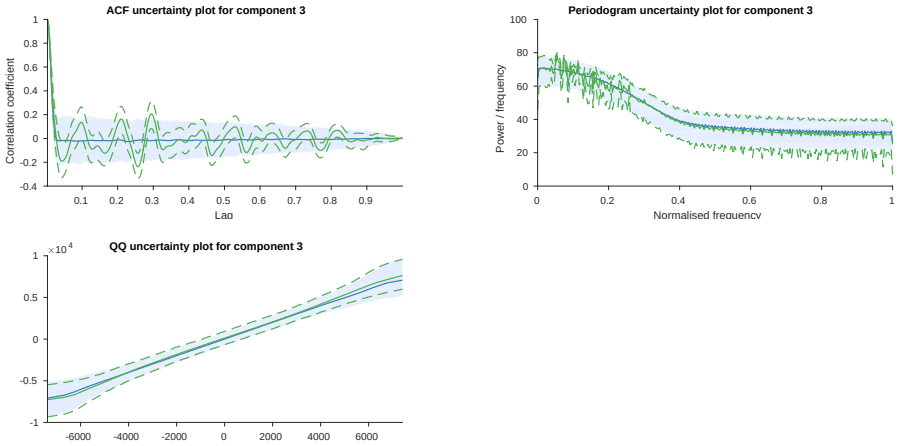

Figure 17: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 3. The green line and green dashed lines are the corresponding quantities under the posterior.

5 MMD - experimental section

#	mmd
1	**0.000**
2	**0.000**
3	**0.000**
4	0.059

Table 3: MMD p-values

5.0.4 Component 1 : A linearly increasing function

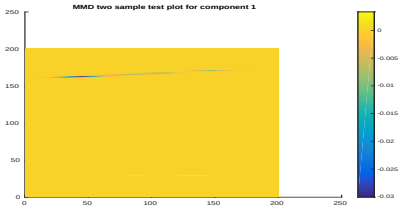

Figure 18: MMD plot

5.0.5 Component 2 : A smooth function

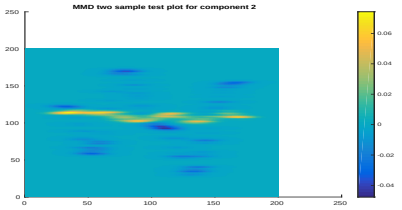

Figure 19: MMD plot

5.0.6 Component 3 : A smooth function

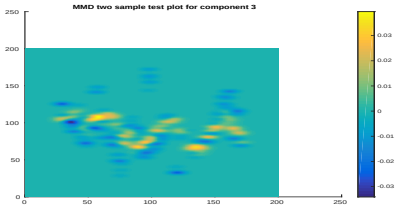

Figure 20: MMD plot

5.0.7 Component 4 : Uncorrelated noise

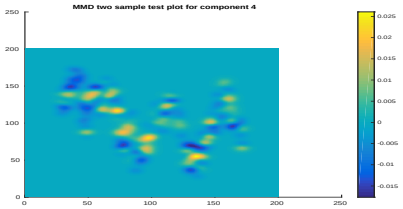

Figure 21: MMD plot