An automatic report for the dataset : lgchem

The Automatic Statistician

March 13, 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 1.

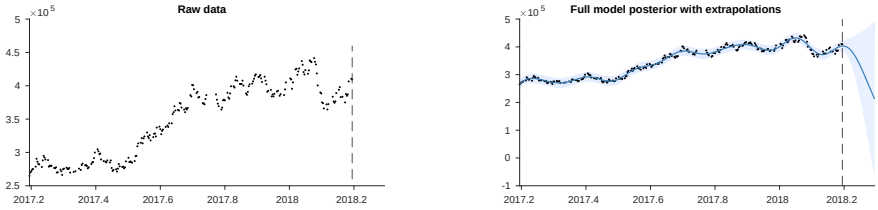

Figure 1: Raw data (left) and model posterior with extrapolation (right)

The structure search algorithm has identified two additive components in the data. The first additive component explains 98.1% of the variation in the data as shown by the coefficient of determination (R^2) values in table 1. The 2 additive components explain 100.0% of the variation in the data. After the first component 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 smooth function.
- Uncorrelated noise.

Model checking statistics are summarised in table 2 in section 4. These statistics have revealed highly statistically significant discrepancies between the data and model in component 1.

#	R^2 (%)	ΔR^2 (%)	Residual R^2 (%)	Cross validated MAE	Reduction in MAE (%)
-	-	-	-	347954.82	-
1	98.1	98.1	98.1	23227.70	93.3
2	100.0	1.9	100.0	23227.70	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.

The rest of the document is structured as follows. In section 2 the forms of the additive components are described and their posterior distributions are displayed. In section 3 the modelling assumptions of each component are discussed with reference to how this affects the extrapolations made by the model. Section 4 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 smooth function

This component is a smooth function with a typical lengthscale of 4.4 weeks.

This component explains 98.1% of the total variance. The addition of this component reduces the cross validated MAE by 93.3% from 347954.8 to 23227.7.

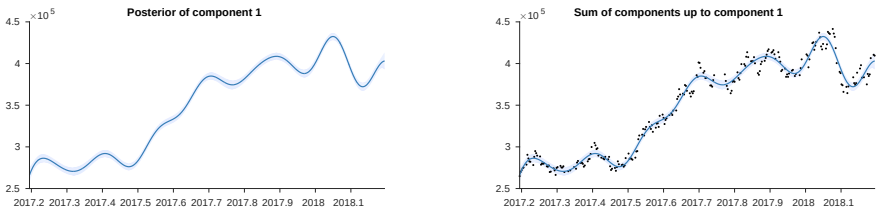

Figure 2: Pointwise posterior of component 1 (left) and the posterior of the cumulative sum of components with data (right)

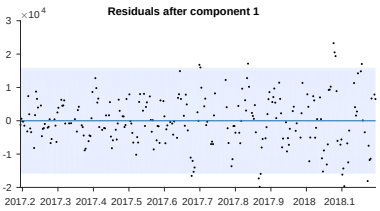

Figure 3: Pointwise posterior of residuals after adding component 1

2.2 Component 2 : Uncorrelated noise

This component models uncorrelated noise.

This component explains 100.0% of the residual variance; this increases the total variance explained from 98.1% to 100.0%. The addition of this component reduces the cross validated MAE by 0.00% from 23227.70 to 23227.70. 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.

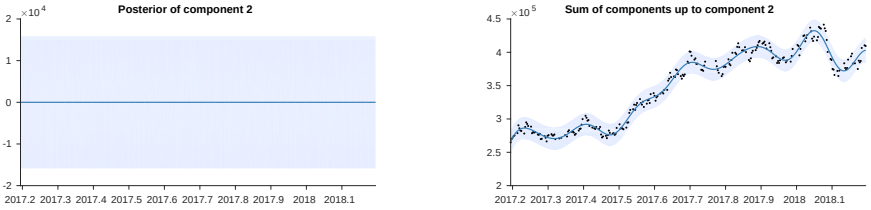

Figure 4: Pointwise posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

3 Extrapolation

Summaries of the posterior distribution of the full model are shown in figure 5. 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.

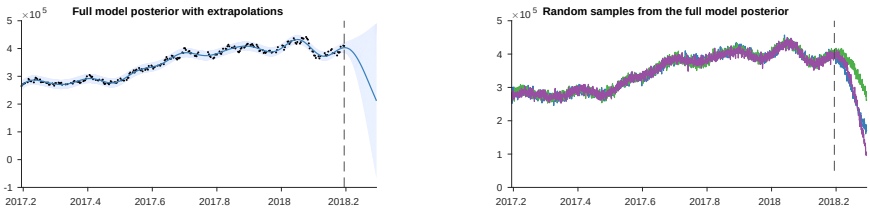

Figure 5: Full model posterior with extrapolation. Mean and pointwise variance (left) and three random samples (right)

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 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 4.4 weeks. [This is a placeholder for a description of how quickly the posterior will start to resemble the prior].

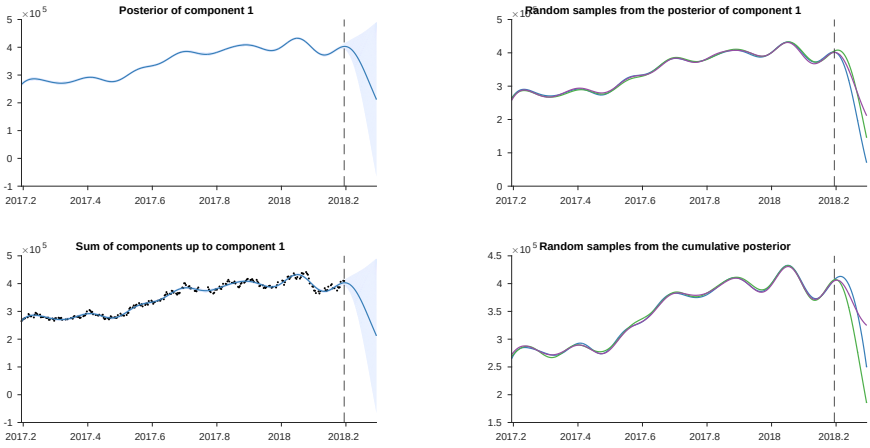

Figure 6: Posterior of component 1 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.2 Component 2 : Uncorrelated noise

This component assumes the uncorrelated noise will continue indefinitely.

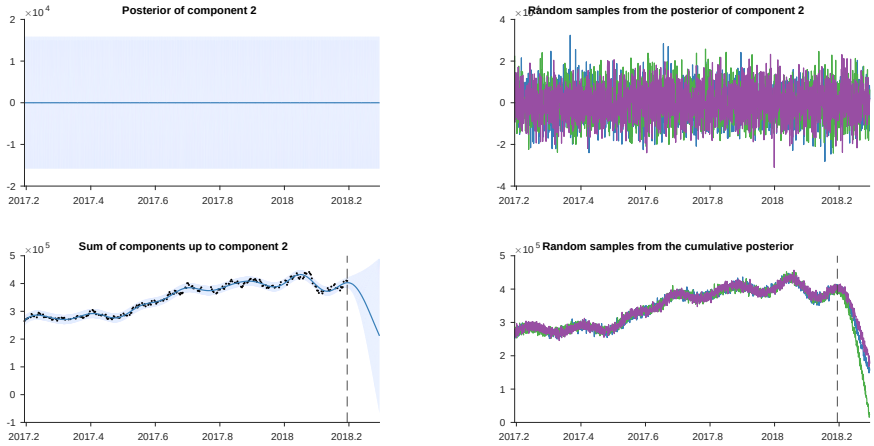

Figure 7: Posterior of component 2 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

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 2 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.

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 Highly statistically significant discrepancies

4.1.1 Component 1 : A smooth function

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

	ACF		Periodogram		QQ	
#	min	min loc	max	max loc	max	min
1	0.377	0.827	0.993	0.062	**0.000**	0.690
2	0.507	0.502	0.520	0.512	0.445	0.666

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 qq plot has an unexpectedly large positive deviation from equality ($x = y$). This discrepancy has an estimated p-value of **0.000**.
- The maximum value of the periodogram is unexpectedly high. This discrepancy has an estimated p-value of 0.014.

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. The large maximum value of the periodogram can indicate periodicity that is not being captured by the model.

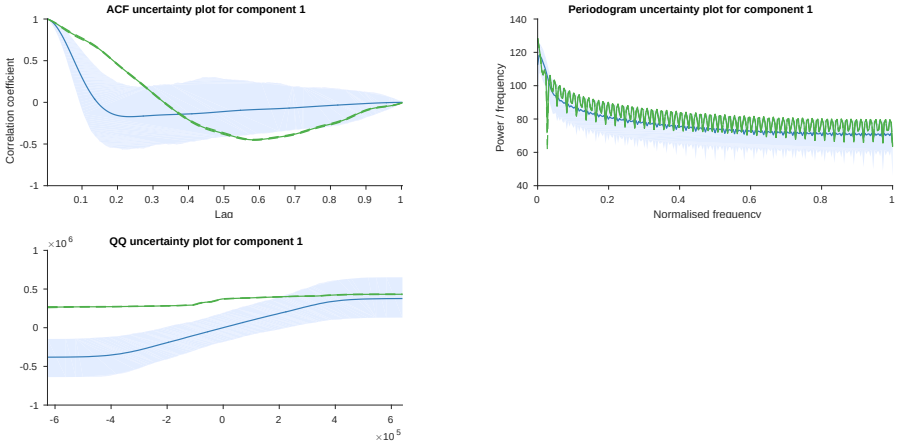

Figure 8: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 1. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2 Model checking plots for components without statistically significant discrepancies

4.2.1 Component 2 : Uncorrelated noise

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

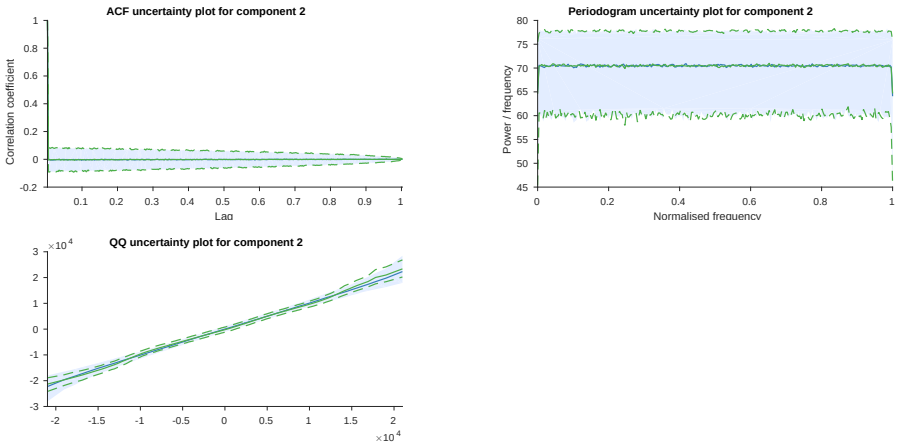

Figure 9: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 2. The green line and green dashed lines are the corresponding quantities under the posterior.

5 MMD - experimental section

#	mmd
1	**0.000**
2	**0.000**

Table 3: MMD p-values

5.0.2 Component 1 : A smooth function

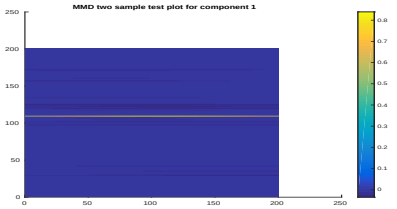

Figure 10: MMD plot

5.0.3 Component 2 : Uncorrelated noise

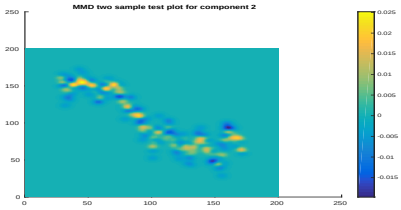

Figure 11: MMD plot