

# Chapter 1

## Automatically Describing Structured Covariance Functions

This paper presents the beginnings of an automatic statistician, focusing on regression problems. Our system explores an open-ended space of possible statistical models to discover a good explanation of the data, and then produces a detailed report with figures and natural-language text.

Our approach treats unknown functions nonparametrically using Gaussian processes, which has two important consequences. First, Gaussian processes model functions in terms of high-level properties (e.g. smoothness, trends, periodicity, changepoints). Taken together with the compositional structure of our language of models, this allows us to automatically describe functions through a decomposition into additive parts. Second, the use of flexible nonparametric models and a rich language for composing them in an open-ended manner also results in state-of-the-art extrapolation performance evaluated over 13 real time series data sets from various domains.

### 1.0.1 Attribution

Joint work with James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum and Zoubin Ghahramani.

## 1.1 Automatic description of regression models

**Overview** In this section, we describe how ABCD generates natural-language descriptions of the models found by the search procedure. There are two main features of our

language of GP models that allow description to be performed automatically.

First, the sometimes complicated kernel expressions found can be simplified into a sum of products. A sum of kernels corresponds to a sum of functions so each product can be described separately. Second, each kernel in a product modifies the resulting model in a consistent way. Therefore, we can choose one kernel to be described as a noun, with all others described using adjectives.

**Sum of products normal form** We convert each kernel expression into a standard, simplified form. We do this by first distributing all products of sums into a sum of products. Next, we apply several simplifications to the kernel expression: The product of two SE kernels is another SE with different parameters. Multiplying WN by any stationary kernel (C, WN, SE, or Per) gives another WN kernel. Multiplying any kernel by C only changes the parameters of the original kernel.

After applying these rules, the kernel can as be written as a sum of terms of the form:

$$K \prod_m \text{Lin}^{(m)} \prod_n \sigma^{(n)}, \quad (1.1)$$

where  $K$  is one of WN, C, SE,  $\prod_k \text{Per}^{(k)}$  or  $\text{SE} \prod_k \text{Per}^{(k)}$  and  $\prod_i k^{(i)}$  denotes a product of kernels, each with different parameters.

**Sums of kernels are sums of functions** Formally, if  $f_1(x) \sim \text{GP}(0, k_1)$  and independently  $f_2(x) \sim \text{GP}(0, k_2)$  then  $f_1(x) + f_2(x) \sim \text{GP}(0, k_1 + k_2)$ . This lets us describe each product of kernels separately.

**Each kernel in a product modifies a model in a consistent way** This allows us to describe the contribution of each kernel in a product as an adjective, or more generally as a modifier of a noun. We now describe how each kernel modifies a model and how this can be described in natural language:

- **Multiplication by SE** removes long range correlations from a model since  $\text{SE}(x, x')$  decreases monotonically to 0 as  $|x - x'|$  increases. This can be described as making an existing model's correlation structure 'local' or 'approximate'.
- **Multiplication by Lin** is equivalent to multiplying the function being modeled by a linear function. If  $f(x) \sim \text{GP}(0, k)$ , then  $xf(x) \sim \text{GP}(0, k \times \text{Lin})$ . This

causes the standard deviation of the model to vary linearly without affecting the correlation and can be described as e.g. ‘with linearly increasing standard deviation’.

- **Multiplication by  $\sigma$**  is equivalent to multiplying the function being modeled by a sigmoid which means that the function goes to zero before or after some point. This can be described as e.g. ‘from [time]’ or ‘until [time]’.
- **Multiplication by Per** modifies the correlation structure in the same way as multiplying the function by an independent periodic function. Formally, if  $f_1(x) \sim \text{GP}(0, k_1)$  and  $f_2(x) \sim \text{GP}(0, k_2)$  then

$$\text{Cov}[f_1(x)f_2(x), f_1(x')f_2(x')] = k_1(x, x')k_2(x, x').$$

This can be loosely described as e.g. ‘modulated by a periodic function with a period of [period] [units]’.

**Constructing a complete description of a product of kernels** We choose one kernel to act as a noun which is then described by the functions it encodes for when unmodified e.g. ‘smooth function’ for SE. Modifiers corresponding to the other kernels in the product are then appended to this description, forming a noun phrase of the form:

Determiner + Premodifiers + Noun + Postmodifiers

As an example, a kernel of the form  $\text{SE} \times \text{Per} \times \text{Lin} \times \sigma$  could be described as an

$$\underbrace{\text{SE}}_{\text{approximately}} \times \underbrace{\text{Per}}_{\text{periodic function}} \times \underbrace{\text{Lin}}_{\text{with linearly growing amplitude}} \times \underbrace{\sigma}_{\text{until 1700.}}$$

where Per has been selected as the head noun.

In principle, any assignment of kernels in a product to these different phrasal roles is possible, but in practice we found certain assignments to produce more interpretable phrases than others. The head noun is chosen according to the following ordering:

$$\text{Per} > \text{WN, SE, C} > \prod_m \text{Lin}^{(m)} > \prod_n \sigma^{(n)}$$

i.e. Per is always chosen as the head noun when present.

**Ordering additive components** The reports generated by ABCD attempt to present the most interesting or important features of a data set first. As a heuristic, we order components by always adding next the component which most reduces the 10-fold cross-validated mean absolute error.

### 1.1.1 Worked example

Suppose we start with a kernel of the form

$$\text{SE} \times (\text{WN} \times \text{Lin} + \text{CP}(\text{C}, \text{Per})).$$

This is converted to a sum of products:

$$\text{SE} \times \text{WN} \times \text{Lin} + \text{SE} \times \text{C} \times \boldsymbol{\sigma} + \text{SE} \times \text{Per} \times \bar{\boldsymbol{\sigma}}.$$

which is simplified to

$$\text{WN} \times \text{Lin} + \text{SE} \times \boldsymbol{\sigma} + \text{SE} \times \text{Per} \times \bar{\boldsymbol{\sigma}}.$$

To describe the first component, the head noun description for WN, ‘uncorrelated noise’, is concatenated with a modifier for Lin, ‘with linearly increasing standard deviation’. The second component is described as ‘A smooth function with a lengthscale of [lengthscale] [units]’, corresponding to the SE, ‘which applies until [changepoint]’, which corresponds to the  $\boldsymbol{\sigma}$ . Finally, the third component is described as ‘An approximately periodic function with a period of [period] [units] which applies from [changepoint]’.

Automating the process of statistical modeling would have a tremendous impact on fields that currently rely on expert statisticians, machine learning researchers, and data scientists. While fitting simple models (such as linear regression) is largely automated by standard software packages, there has been little work on the automatic construction of flexible but interpretable models. What are the ingredients required for an artificial intelligence system to be able to perform statistical modeling automatically? In this paper we conjecture that the following ingredients may be useful for building an AI system for statistics, and we develop a working system which incorporates them:

- **An open-ended language of models** expressive enough to capture many of the modeling assumptions and model composition techniques applied by human statisticians to capture real-world phenomena

This component is approximately periodic with a period of 10.8 years. Across periods the shape of this function varies smoothly with a typical lengthscale of 36.9 years. The shape of this function within each period is very smooth and resembles a sinusoid. This component applies until 1643 and from 1716 onwards.

This component explains 71.5% of the residual variance; this increases the total variance explained from 72.8% to 92.3%. The addition of this component reduces the cross validated MAE by 16.82% from 0.18 to 0.15.

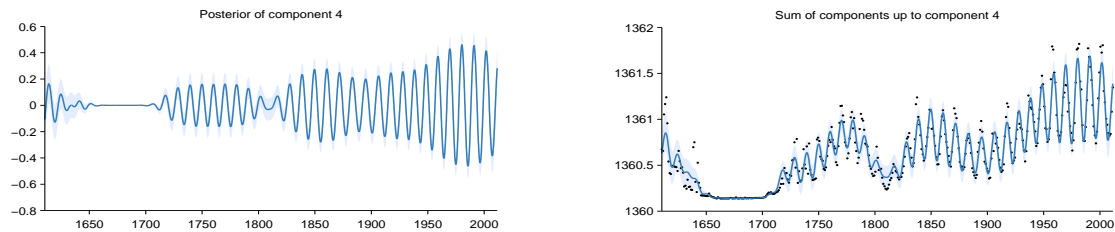


Figure 8: Pointwise posterior of component 4 (left) and the posterior of the cumulative sum of components with data (right)

Fig. 1.1 Extract from an automatically-generated report describing the model components discovered by automatic model search. This part of the report isolates and describes the approximately 11-year sunspot cycle, also noting its disappearance during the 16th century, a time known as the Maunder minimum (?).

- **A search procedure** to efficiently explore the space of models spanned by the language
- **A principled method for evaluating models** in terms of their complexity and their degree of fit to the data
- **A procedure for automatically generating reports** which explain and visualize different factors underlying the data, make the chosen modeling assumptions explicit, and quantify how each component improves the predictive power of the model

The compositional structure of the language allows us to develop a method for automatically translating components of the model into natural-language descriptions of patterns in the data.

We show examples of automatically generated reports which highlight interpretable features discovered in a variety of data sets (e.g. figure 1.3). The supplementary material to this paper includes 13 complete reports automatically generated by ABCD.

Good statistical modeling requires not only interpretability but predictive accuracy.

**Airline passenger data** Figure ?? shows the decomposition produced by applying our method to monthly totals of international airline passengers (?). We observe similar components to the previous dataset: a long term trend, annual periodicity and medium-term deviations. In addition, the composite kernel captures the near-linearity of the long-term trend, and the linearly growing amplitude of the annual oscillations.

We demonstrate the ability of our procedure to discover and describe a variety of patterns on two time series. Full automatically-generated reports for 13 data sets are provided as supplementary material.

### 1.1.2 Summarizing 400 Years of Solar Activity

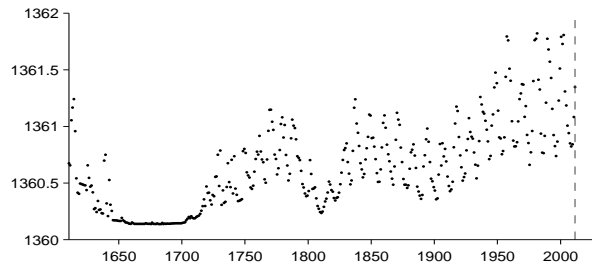


Fig. 1.2 Solar irradiance data.

We show excerpts from the report automatically generated on annual solar irradiation data from 1610 to 2011 (figure 1.2). This time series has two pertinent features: a roughly 11-year cycle of solar activity, and a period lasting from 1645 to 1715 with much smaller variance than the rest of the dataset. This flat region corresponds to the Maunder minimum, a period in which sunspots were extremely rare (?). ABCD clearly identifies these two features, as discussed below.

Figure 1.4 shows the natural-language summaries of the top four components chosen by ABCD. From these short summaries, we can see that our system has identified the Maunder minimum (second component) and 11-year solar cycle (fourth component). These components are visualized in figures 1.5 and 1.3, respectively. The third component corresponds to long-term trends, as visualized in figure 1.6.

### 1.1.3 Finding heteroscedasticity in air traffic data

Next, we present the analysis generated by our procedure on international airline passenger data (figure 1.7). The model constructed by ABCD has four components:  $\text{Lin} + \text{SE} \times \text{Per} \times \text{Lin} + \text{SE} + \text{WN} \times \text{Lin}$ , with descriptions given in figure 1.8.

This component is approximately periodic with a period of 10.8 years. Across periods the shape of this function varies smoothly with a typical lengthscale of 36.9 years. The shape of this function within each period is very smooth and resembles a sinusoid. This component applies until 1643 and from 1716 onwards.

This component explains 71.5% of the residual variance; this increases the total variance explained from 72.8% to 92.3%. The addition of this component reduces the cross validated MAE by 16.82% from 0.18 to 0.15.

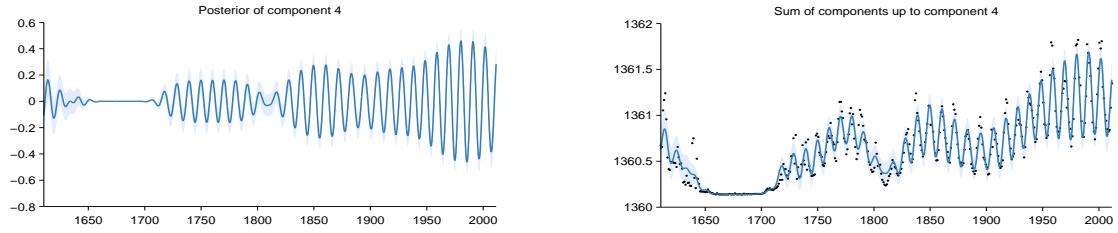


Figure 8: Pointwise posterior of component 4 (left) and the posterior of the cumulative sum of components with data (right)

Fig. 1.3 Extract from an automatically-generated report describing the model components discovered by automatic model search. This part of the report isolates and describes the approximately 11-year sunspot cycle, also noting its disappearance during the 16th century, a time known as the Maunder minimum (?).

The second component (figure 1.9) is accurately described as approximately (SE) periodic (Per) with linearly growing amplitude (Lin). By multiplying a white noise kernel by a linear kernel, the model is able to express heteroscedasticity (figure 1.10).

## 1.2 Related work

### 1.2.1 Natural-language output

To the best of our knowledge, our procedure is the first example of automatic description of nonparametric statistical models. However, systems with natural language output have been built in the areas of video interpretation (?) and automated theorem proving (?).

### 1.2.2 Interpretability versus accuracy

BIC trades off model fit and complexity by penalizing the number of parameters in a kernel expression. This can result in ABCD favoring kernel expressions with nested

The structure search algorithm has identified eight additive components in the data. The first 4 additive components explain 92.3% of the variation in the data as shown by the coefficient of determination ( $R^2$ ) values in table 1. The first 6 additive components explain 99.7% of the variation in the data. After the first 5 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 constant.
- A constant. This function applies from 1643 until 1716.
- A smooth function. This function applies until 1643 and from 1716 onwards.
- An approximately periodic function with a period of 10.8 years. This function applies until 1643 and from 1716 onwards.

Fig. 1.4 Automatically generated descriptions of the components discovered by ABCD on the solar irradiance data set. The dataset has been decomposed into diverse structures with simple descriptions.

products of sums, producing descriptions involving many additive components. While these models have good predictive performance the large number of components can make them less interpretable. We experimented with distributing all products over addition during the search, causing models with many additive components to be more heavily penalized by BIC. We call this procedure ABCD-interpretability, in contrast to the unrestricted version of the search, ABCD-accuracy.

## 1.3 Conclusion

Towards the goal of automating statistical modeling we have presented a system which constructs an appropriate model from an open-ended language and automatically generates detailed reports that describe patterns in the data captured by the model. We have demonstrated that our procedure can discover and describe a variety of patterns on several time series. Our procedure's extrapolation and interpolation performance on time-series are state-of-the-art compared to existing model construction techniques. We believe this procedure has the potential to make powerful statistical model-building techniques accessible to non-experts.



This component is constant. This component applies from 1643 until 1716.

This component explains 37.4% of the residual variance; this increases the total variance explained from 0.0% to 37.4%. The addition of this component reduces the cross validated MAE by 31.97% from 0.33 to 0.23.

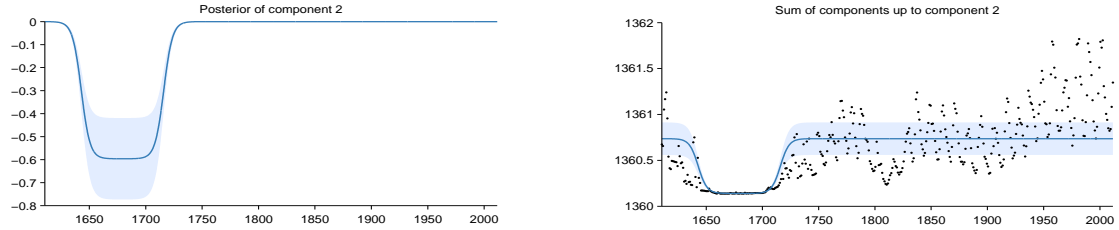


Figure 4: Pointwise posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

Fig. 1.5 One of the learned components corresponds to the Maunder minimum.

This component is a smooth function with a typical lengthscale of 23.1 years. This component applies until 1643 and from 1716 onwards.

This component explains 56.6% of the residual variance; this increases the total variance explained from 37.4% to 72.8%. The addition of this component reduces the cross validated MAE by 21.08% from 0.23 to 0.18.

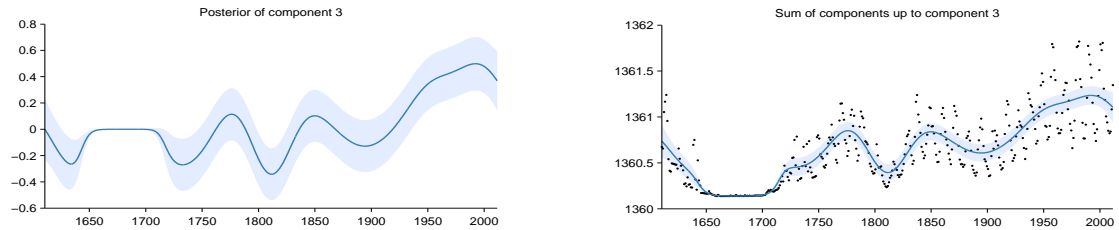


Figure 6: Pointwise posterior of component 3 (left) and the posterior of the cumulative sum of components with data (right)

Fig. 1.6 Characterizing the medium-term smoothness of solar activity levels. By allowing other components to explain the periodicity, noise, and the Maunder minimum, ABCD can isolate the part of the signal best explained by a slowly-varying trend.

## 1.4 Discussion

Towards the goal of automating the choice of kernel family, we introduced a space of composite kernels defined compositionally as sums and products of a small number of base kernels. The set of models included in this space includes many standard regression

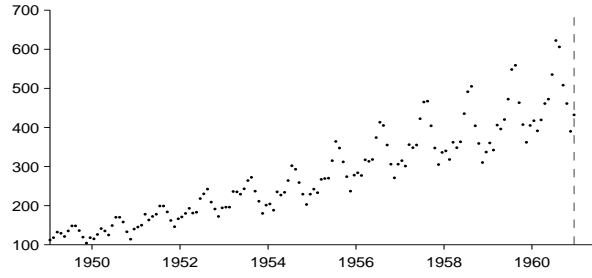


Fig. 1.7 International airline passenger monthly volume (e.g. ?).

models. We proposed a search procedure for this space of kernels which parallels the process of scientific discovery.

We found that the learned structures are often capable of accurate extrapolation in complex time-series datasets, and are competitive with widely used kernel classes and kernel combination methods on a variety of prediction tasks. The learned kernels often yield decompositions of a signal into diverse and interpretable components, enabling model-checking by humans. We believe that a data-driven approach to choosing kernel structures automatically can help make nonparametric regression and classification methods accessible to non-experts.

1

---

<sup>1</sup>All GP parameter optimisation was performed by automated calls to the GPML toolbox available at <http://www.gaussianprocess.org/gpml/code/>.

The structure search algorithm has identified four additive components in the data. The first 2 additive components explain 98.5% of the variation in the data as shown by the coefficient of determination ( $R^2$ ) values in table 1. The first 3 additive components explain 99.8% 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.
- An approximately periodic function with a period of 1.0 years and with linearly increasing amplitude.
- A smooth function.
- Uncorrelated noise with linearly increasing standard deviation.

#	$R^2$ (%)	$\Delta R^2$ (%)	Residual $R^2$ (%)	Cross validated MAE	Reduction in MAE (%)
-	-	-	-	280.30	-
1	85.4	85.4	85.4	34.03	87.9
2	98.5	13.2	89.9	12.44	63.4
3	99.8	1.3	85.1	9.10	26.8
4	100.0	0.2	100.0	9.10	0.0

Fig. 1.8 Short descriptions and summary statistics for the four components of the airline model.

## 2.2 Component 2 : An approximately periodic function with a period of 1.0 years and with linearly increasing amplitude

This component is approximately periodic with a period of 1.0 years and varying amplitude. Across periods the shape of this function varies very smoothly. The amplitude of the function increases linearly. The shape of this function within each period has a typical lengthscale of 6.0 weeks.

This component explains 89.9% of the residual variance; this increases the total variance explained from 85.4% to 98.5%. The addition of this component reduces the cross validated MAE by 63.45% from 34.03 to 12.44.

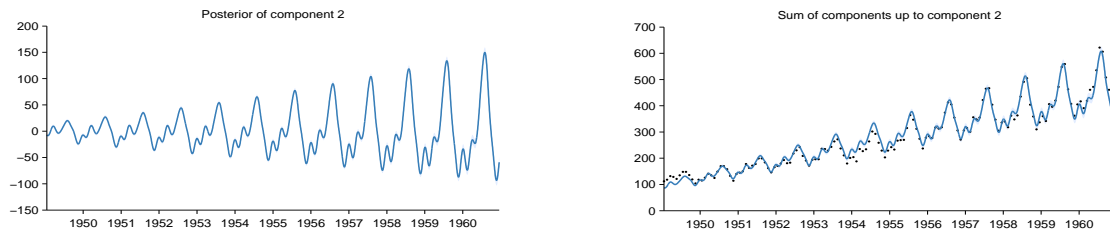


Figure 4: Pointwise posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

Fig. 1.9 Capturing non-stationary periodicity in the airline data

#### 2.4 Component 4 : Uncorrelated noise with linearly increasing standard deviation

This component models uncorrelated noise. The standard deviation of the noise increases linearly.

This component explains 100.0% of the residual variance; this increases the total variance explained from 99.8% to 100.0%. The addition of this component reduces the cross validated MAE by 0.00% from 9.10 to 9.10. 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.

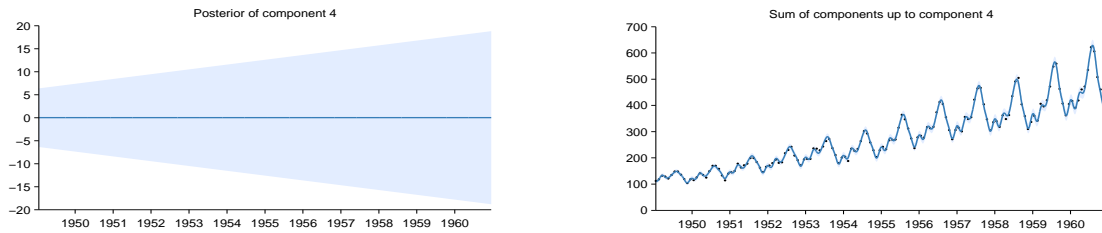


Figure 8: Pointwise posterior of component 4 (left) and the posterior of the cumulative sum of components with data (right)

Fig. 1.10 Modeling heteroscedasticity in the airline dataset.