Automatic Construction and Natural-language Description of Additive Nonparametric Models

James Robert Lloyd
University of Cambridge
jrl44@cam.ac.uk

David Duvenaud
University of Cambridge
dkd23@cam.ac.uk

Roger Grosse
M.I.T.
rgrosse@mit.edu

Joshua B. Tenenbaum
M.I.T.
jbt@mit.edu

Zoubin Ghahramani
University of Cambridge
zoubin@eng.cam.ac.uk

Abstract

To complement recently introduced automatic model-construction and search methods, we demonstrate an automatic model-summarization procedure. After building an additive nonparametric regression model, our method constructs a report which visualizes and explains in words the meaning and relevance of each component. These reports enable human model-checking and the understanding of complex modeling assumptions and structure. We demonstrate this procedure on two time-series, showing that the automatically constructed models identify clearly interpretable structures that can be automatically described in simple natural language.

1 Introduction

Simple parametric regression models such as linear regression are usually easily interpretable, and have well-established methods for model-checking. In constrast, non-parametric models may be reasonably viewed with suspicion by the non-expert, since their assumptions may be difficult to check, and their predictive implications difficult to explain. The aim of this work is to develop tools which make complex nonparametric models more accessible and intelligible to non-experts.

Several recently proposed methods search over a large class of structured nonparametric regression models [1, 2, 3, 4]. In [1], it was noted that these models could be decomposed into sums of diverse components, and that the different components corresponded to different features of the data. In that work, components discovered on real datasets were interpreted post-hoc by the authors.

We extend this work by demonstrating that the structure of the models searched over allows for automatic natural-language description of patterns within a data set. We have also expanded the components used in the model-construction process to improve the expressivity and interpretability of the models.

This paper contains extracts of human-readable reports automatically generated by our procedure. The automatically generated text descriptions of the components of the models clearly communicate interpretable features of the data. The supplementary material to this paper is a pair of complete reports generated by our method.

2 Gaussian Process Structure Search

The family of models described by our procedure are Gaussian process (GP) regression models [5]. GPs use a kernel to define the covariance between any two function values, y, y' evaluated at two inputs, x, x' i.e. $\text{Cov}(y, y') = k(x, x')$. The kernel specifies which structures are likely under the GP prior, which in turn determines the generalization properties of the model.

Different kernels can express a wide variety of covariance structures, such as local smoothness or periodicity. New kernels can be constructed by taking the product of a set of base kernels to express richer structures, such as functions which are locally periodic, or heteroscedastic.

The Gaussian process structure search (GPSS) procedure [1] searches over sums and products of a set of simple base kernels to produce an appropriate model for a given data set. To produce the reports exhibited in this paper we used 6 base kernels that represent the following structures: smooth functions (SE), periodic functions (PER), linear functions (LIN), constant functions (C), changepoints (CP), and white noise (WN).

3 Description of kernel compositions

For a given dataset, the GPSS method produces a compound kernel composed of sums and products of the 6 base kernels. In this section, we describe how the properties of kernel functions reduce the task of natural-language summarization into relatively simple subproblems.

3.1 Distributivity of multiplication over addition

By distributivity of multiplication over addition, any kernel expression can be converted into a sum of products of base kernels. For example, when the first version of GPSS [1] was applied to airline passenger data, the syntax of the learned kernel structure was

$$\text{SE} \times \big(\text{LIN} + \text{LIN} \times (\text{PER} + \text{SE})\big) + \text{WN} \tag{3.1}$$

which can be distributed into a sum of products

$$(\text{SE} \times \text{LIN}) + (\text{SE} \times \text{LIN} \times \text{PER}) + (\text{SE} \times \text{LIN}) + \text{WN}. \tag{3.2}$$

If $f_1(x) \sim \text{GP}(0, k_1)$ and $f_2(x) \sim \text{GP}(0, k_2)$ then $f_1(x) + f_2(x) \sim \text{GP}(0, k_1 + k_2)$. Therefore, a sum of kernels can be described as a sum of independent functions. Thus, we can always exactly describe our model as a sum of these components, each of which is a product of base kernels. The only remaining task is to produce a procedure for describing products of base kernels.

3.2 Description of products of base kernels

Since the individual additive components produced by the GPSS method are produced by an open-ended grammar, it is impossible to write a lookup table to handle every type of component that GPSS might produce. In this section, we demonstrate how to describe the properties of an arbitrary product of kernels.

First, we will show that multiplying by each kernel modifies the properties of the resulting prior on functions in a consistent way. This means that, for most cases, multiplying by another base kernel means we can simply append a set of adjectives to the description of the resulting function.

3.2.1 Multiplication by a squared exponential kernel

Multiplying by an SE kernel means that the resulting prior over functions now varies smoothly. The degree of smoothness depends on the lengthscale of the kernel. A product of two SE kernels is exactly equivalent to a single SE with different parameters. This means that multiplication by an SE kernel is an idempotent operation.

For example, the kernel SE × PER gives rise to functions which are *locally* periodic: they only approximately repeat. If the input space is restricted to a grid with spacing equal to the period of the periodic kernel, then $\text{SE} \times \text{PER} = \text{SE}$. We therefore see that the functional form of the periodicity varies like SE, giving rise to the local nature of the periodicity.

3.2.2 Multiplication by a white noise kernel

For stationary kernels, WN behaves like a multiplicative zero. For example, $\text{WN} \times \text{SE} = \text{WN} \times \text{PER} = \text{WN}$. Thus the product of white noise with any stationary kernel is simply white noise.

3.2.3 Multiplication by a constant kernel

The constant kernel C behaves like a multiplicative identity: $\text{C} \times k = k$ for any k.

3.2.4 Multiplication by a periodic kernel

Multiplication by a periodic kernel means that each $f(x)$ will co-vary with points near $f(x + n\tau)$, where n is an integer and τ is the period of the kernel. This property also applies to products of periodic kernels. Suppose that $f_1(x) \sim \text{GP}(0, k_1)$ and $f_2(x) \sim \text{GP}(0, k_2)$. Then

$$\text{Cov}\left[f_1(x)f_2(x), f_1(x')f_2(x')\right] = k_1(x, x')k_2(x, x'). \tag{3.3}$$

Therefore $\text{PER} \times \text{PER}$ defines a prior on functions whose covariance is the same as the product of independent periodic functions. However, note that a product of periodic functions drawn from GP priors will not be distributed according to a GP.

3.2.5 Multiplication by linear and changepoint kernels

The linear (LIN) and changepoint (CP) kernels both have the form $k(x, x') = a(x)a(x')$. We can interpret kernels of this form in the following way: Suppose that $f(x) \sim \text{GP}(0, k)$ and $a : \mathcal{X} \to \mathcal{Y}$ is a known function. Then $a(x)f(x) \sim \text{GP}\left(0, a(x)k(x, x')a(x')\right)$.

For example, the linear kernel, LIN, has the form $\text{LIN}(x, x') = c(x - a)(x' - a)$. Setting $a(x) = \sqrt{c}(x - a)$ we see that multiplying a kernel, k, by the linear kernel is equivalent to multiplying $f(x) \sim \text{GP}(0, k)$ by a linear function. Similarly, multiplying by a changepoint kernel is equivalent to multiplying a GP by a sigmoid function.

This interpretation means that we can separately describe the properties of the model arising from multiplication with kernels of the form $k(x, x') = a(x)a(x')$. For example, multiplying by a linear kernel means that the amplitude of the function grows linearly away from a central point. Multiplication by two linear kernels results in quadratic growth, etc.

Changepoints are expressed in a similar manner: By multiplying a function by a sigmoid the contribution of any given kernel will go to zero in certain regions. This modulation can be trivially expressed in natural-language e.g. "this component applies from 1700 to 1800".

4 Automatic Report Generation

Our report generation procedure starts with a dataset and a composite kernel, which together define a joint posterior probability distribution over a sum of functions. The procedure summarizes properties of this complex distribution to the user through a comprehensive report. These reports are designed to be intelligible to non-experts, illustrating the assumptions made by the model, describing the model's posterior distribution, and most importantly, enabling model-checking.

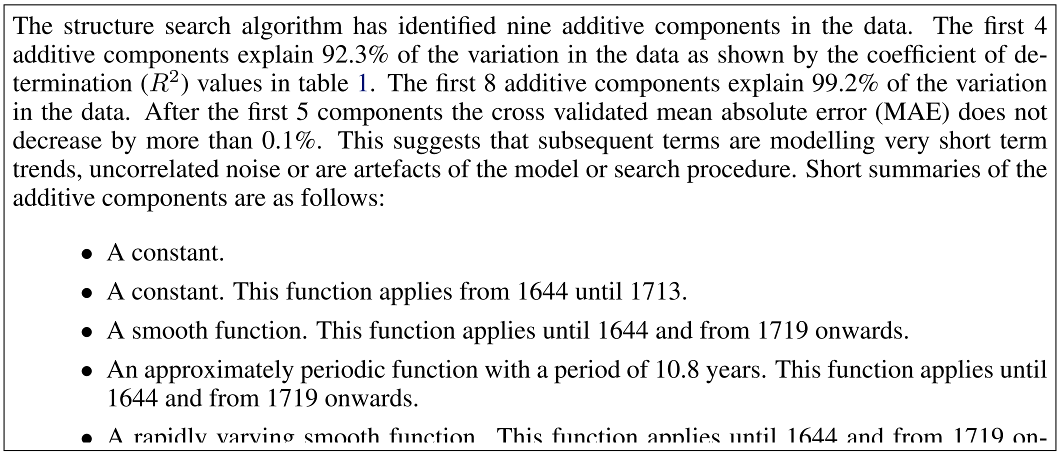
These reports have three sections: an executive summary, a detailed discussion of each component, and a section discussing how the model extrapolates beyond the range of the data.

- **Executive Summary** The first section of each report summarizes the components of the model, and the relative importance of the different components in explaining the data.
- **Decomposition plots** The second section of each report contains a detailed discussion of each component. Every component is plotted, and properties of the covariance structure are described. Each component's posterior is plotted in two ways. First, the posterior mean and variance of each component is plotted on its own. Second, the posterior mean and variance of all components shown so far is plotted against the data. This progression of plots shows qualitatively how each component contributes to an overall explanation of the data.
- **Extrapolation plots** The third section of each report shows extrapolations into the future, as well as posterior samples from each individual component of the model. These samples help to characterize the uncertainty expressed by the model, and the extent to which different components contribute to predicting the future behavior of a time series. The predictive mean and variance of the signals shown in the summary plots are useful, but do not capture the joint correlation structure in the posterior. Showing posterior samples is a simple and universal way to illustrate joint statistical structure.

An additional section that automatically describes model-checking procedures is work in progress.

4.1 Example: Summarizing 400 Years of Solar Activity

To give an example of the capabilities of our procedure, we show excerpts from the report automatically generated on annual solar irradiation data from 1610 to 2011. This time series has two obvious 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 [6]. The GPSS search procedure and automatic summary clearly identify these two features, as discussed below.

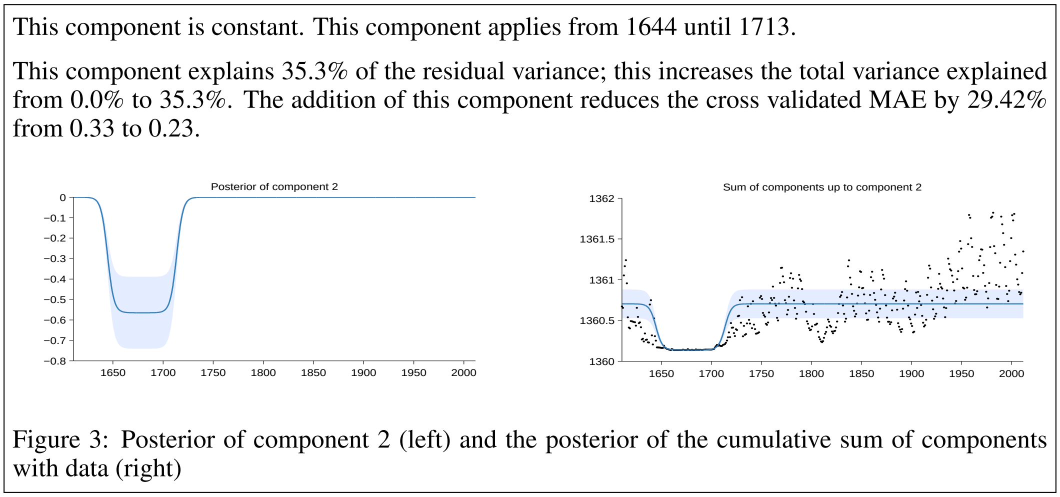


The structure search algorithm has identified nine additive components in the data. The first 4 additive components explain 92.3% of the variation in the data as shown by the coefficient of determination (R^2) values in table 1. The first 8 additive components explain 99.2% of the variation in the data. 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 1644 until 1713.
- A smooth function. This function applies until 1644 and from 1719 onwards.
- An approximately periodic function with a period of 10.8 years. This function applies until 1644 and from 1719 onwards.
- A rapidly varying smooth function. This function applies until 1644 and from 1719 on-

Figure 1: An example of an automatically-generated summary of a dataset. The dataset is decomposed into diverse types of structures, and each structure is explained in simple terms.

Figure 1 shows the automatically-generated summary of the solar dataset. The model uses 9 additive components to explain the data, and reports that the first 4 components explain more than 90% of the variance in the data. Just from the short summaries of the additive components we can see that the model has identified the Maunder minimum (second component) and 11-year solar cycle (fourth component).



This component is constant. This component applies from 1644 until 1713.

This component explains 35.3% of the residual variance; this increases the total variance explained from 0.0% to 35.3%. The addition of this component reduces the cross validated MAE by 29.42% from 0.33 to 0.23.

Figure 3: Posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

Figure 2: Discovering the Maunder minimum. The kernel found by GPSS contained a pair of changepoints bracketing the period of low solar activity.

Figure 2 shows that GPSS has captured the unusual period of decreased solar activity from about 1645 to 1715 and is able to report this in natural language. This feature was captured by the model by multiplying a constant kernel by two changepoint kernels. Figure 3 shows that GPSS has isolated the approximately 11 year solar cycle.

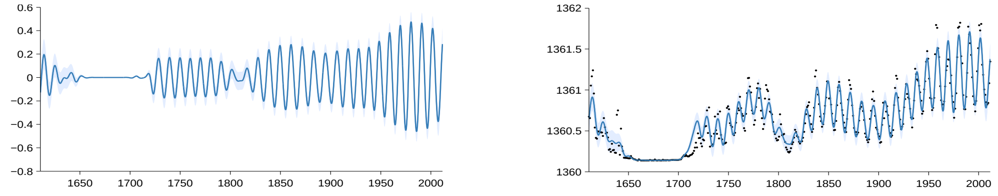


Figure 5: Posterior of component 4 (left) and the posterior of the cumulative sum of components with data (right)

Figure 3: Isolating the periodic component of the dataset. By isolating this aspect of the statistical structure, we can easily observe additional features, such as the shape of the peaks and troughs, or the fact that the amplitude changes over time.

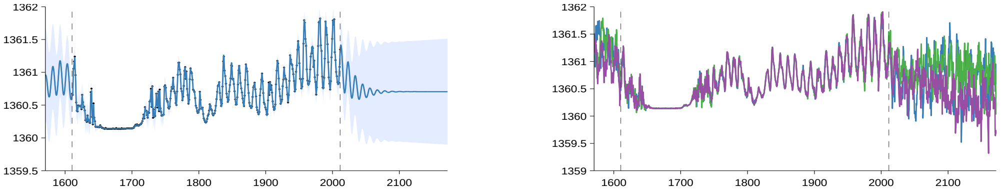


Figure 11: Full model posterior. Mean and pointwise variance (left) and three random samples (right)

Figure 4: Sampling from the posterior. These samples help show not just the predictive mean and variance, but also the predictive covariance structure. Note, for example, that the predictive mean (left) does not exhibit periodicity, but the samples (right) do.

The posterior samples help illustrate detailed structure in the posterior not captured by the pointwise predictive mean and variance. For example, it is not clear from the left-hand plot in figure 4 whether or not the periodicity of the dataset is expected to continue into the future. However, from the samples on the right-hand size, we can see that this is indeed the case.

Source Code Python code to perform all experiments is available on github, available at `github.com/jamesrobertlloyd/gpss-research`.

References

[1] David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning*, June 2013.

[2] Gabriel Kronberger and Michael Kommenda. Evolution of covariance functions for gaussian process regression using genetic programming. *arXiv preprint arXiv:1305.3794*, 2013.

[3] L. Diosan, A. Rogozan, and J.P. Pecuchet. Evolving kernel functions for SVMs by genetic programming. In *Machine Learning and Applications, 2007*, pages 19–24. IEEE, 2007.

[4] W. Bing, Z. Wen-qiong, C. Ling, and L. Jia-hong. A GP-based kernel construction and optimization method for RVM. In *International Conference on Computer and Automation Engineering (ICCAE)*, volume 4, pages 419–423, 2010.

[5] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA, USA, 2006.

[6] J. Lean, J. Beer, and R. Bradley. Reconstruction of solar irradiance since 1610: Implications for climate change. *Geophysical Research Letters*, 22(23):3195–3198, 1995.