Automatic construction and natural-language summarization 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 a recently introduced automatic model-construction and search method, we demonstrate an automatic model-summarization procedure. After automatically 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 highly 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.

Advances in nonparametric and compositional modeling have allowed fully automatic model search and construction of complex, structured, nonparametric regression models [1]. These automatically-constructed models have been used for extrapolation without further human intevention, but for model-checking or dataset exploration, extrapolations alone are insufficient. Fortunately, the generated models are sufficiently structured to capture human-interpretable features in a given dataset, and the compositional nature of this structure allows for simple translation into natural language.

This paper exhibits extracts of the models and human-readable reports which are 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 (some sections are work in progress).

2 Gaussian Process Structure Search

Gaussian process (GP) models [2] 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, periodicity, and changepoints. New kernels can be constructuted by taking the product of a set of base kernels to express richer structures, such as functions which are periodic only locally, or functions whose marginal prior variance varies with the input (i.e. heteroskedasticity).

The Gaussian process structure search (GPSS) procedure searches over sums and products of a set of simple base kernels, in order to find complex structures in time-series or multidimensional functions [1]. To produce the reports exhibited in this paper we used 6 base kernels that represent the following structures : smooth functions (squared-exponential kernel - SE), periodic functions (periodic kernel - PER), linear functions (linear kernel - LIN), constant functions, changepoints, and white noise.

The GPSS method produces a composite kernel which can be distributed into a sum of products of the base kernels. A sum of kernels directly corresponds to a sum of functions which allows for the models produced by GPSS to be decomposed naturally into a sum of parts.

3 Automatic Report Generation

Our report generation procedure takes as its starting point 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.

Automatic natural language descriptions of functions The compositional nature of kernel structures produced by GPSS allows for simple translation of individual components. Each type of kernel in a product of kernels modifies the form of the posterior in a consistent way, preventing the need for an infinite lookup table of descriptions. Table 1 gives some examples of high level summaries of functions. Since the posterior distribution also depends on the data, our summarizer includes heuristics to describe the shape of the components being modeled, such as "smooth" or "monotonically increasing".

Kernel	Description
PER	An exactly periodic function
PER $\times$ SE	An approximately periodic function
PER $\times$ SE $\times$ LIN	An approximately periodic function with linearly varying amplitude
LIN	A linear function
LIN $\times$ LIN	A quadratic function
PER $\times$ LIN $\times$ LIN	An exactly periodic function with quadratically varying amplitude

Table 1: A demonstration of how it is possible to write an open-ended summarizer. Different kernel structures modify the overall statistical structure of each component in independent ways.

3.1 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 [3]. The GPSS search procedure and automatic summary clearly identify these two features, as discussed below.

Executive Summary The first section of each report describes which structures were discovered in the model, and the relative importance of the different components in explaining the data.

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).

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.

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.

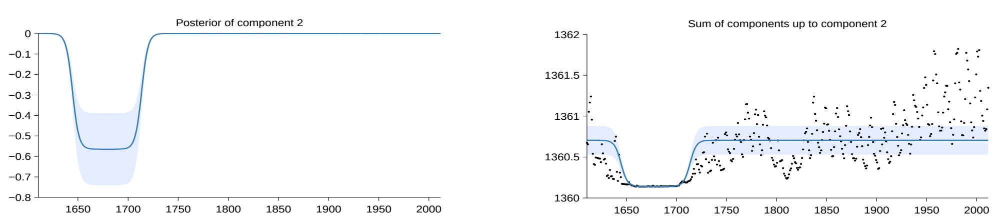

Figure 3: Posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

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.

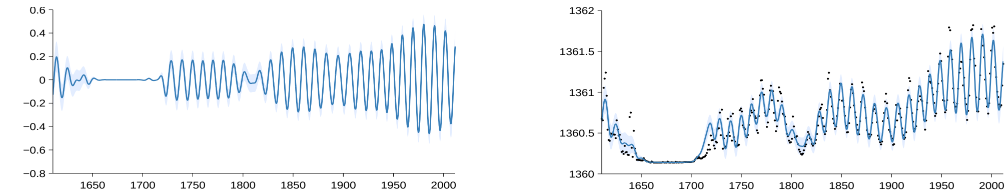

Figure 5: Posterior of component 4 (left) and the posterior of the cumulative sum of components with data (right)

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.

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. For example, it is not clear from the left-hand plot in figure 4 whether or not the periodicity of the

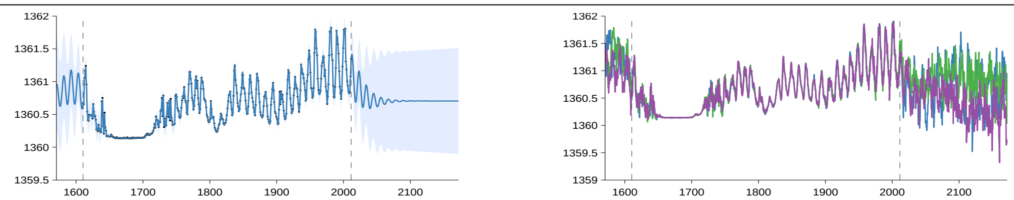

Figure 11: Full model posterior. Mean and pointwise variance (left) and three random samples (right)

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.

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.[1]

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

[3] 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.

[1] Available at `github.com/jamesrobertlloyd/gpss-research`