Automatic construction and description of nonparametric models

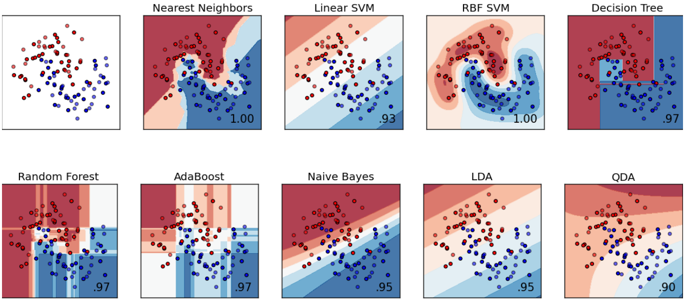
  

James Robert Lloyd, David Duvenaud, Roger Grosse,
Joshua B. Tenenbaum, Zoubin Ghahramani

December 3, 2013

TYPICAL STATISTICAL MODELLING

- Statistical / machine learning models typically built by hand, or chosen from a fixed set
  - Example: Scikit-learn



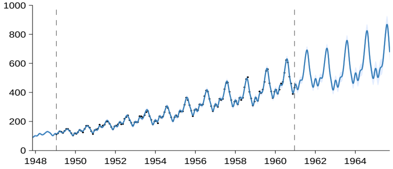
THIS APPROACH IS LIMITED

- Building by hand requires considerable expertise
  - Can become an entire research project
- Just being nonparametric isn't good enough
  - Nonparametric does not mean assumption free!
- Can silently fail
  - If none of the models tried fit the data well, how can you tell?

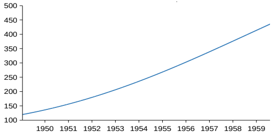
Can we do better?

- Andrew Gelman asks: How can an AI do statistics?
- An artificial statistician would need:
  - a language for describing arbitrarily complicated models
  - a method of searching over those models
  - a procedure to check model fit
- We constructed such a language over regression models, a procedure to search over it, and a method to describe in natural language the properties of the resulting models
  - Actively researching automatic model-checking…

Example: An automatic analysis

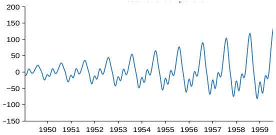


=



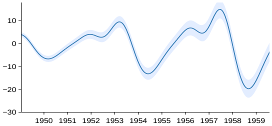
A very smooth, monotonically increasing function

+



An approximately periodic function with a period of 1.0 years and with approximately linearly increasing amplitude

+



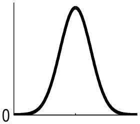
An exactly periodic function with a period of 4.3 years but with linearly increasing amplitude

A LANGUAGE OF REGRESSION MODELS

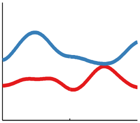
- We define a language of Gaussian process (GP) regression models by defining a language over kernel functions
- We start with a small set of base kernels and create a language with a generative grammar
  - Expansion operators include addition, multiplication and change-points
- The language is open-ended, but its structure makes natural-language description simple

KERNELS DETERMINE STRUCTURE OF GPS

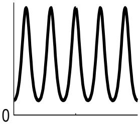
- Kernel determines almost all the properties of a GP prior
- Many different kinds, with very different properties:



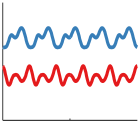
Squared-exp (SE)



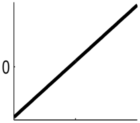
local variation



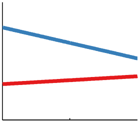
Periodic (PER)



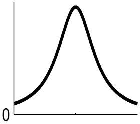
repeating structure



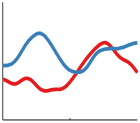
Linear (LIN)



linear functions



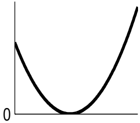
Rational-quadratic(RQ)



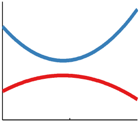
multi-scale variation

KERNELS CAN BE COMPOSED

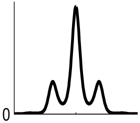
- Two main operations: addition, multiplication



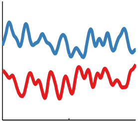
$\textsc{Lin} \times \textsc{Lin}$



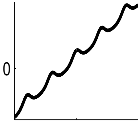
quadratic functions



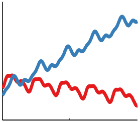
$\textsc{SE} \times \textsc{Per}$



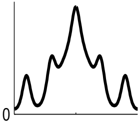
locally periodic



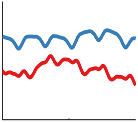
$\textsc{Lin} + \textsc{Per}$



periodic with trend



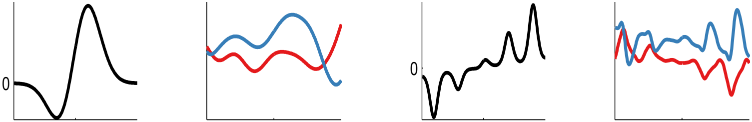
$\textsc{SE} + \textsc{Per}$



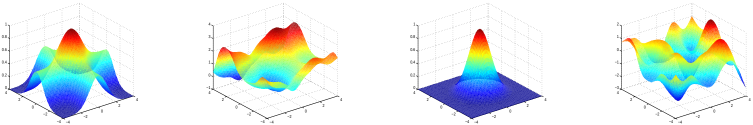
periodic with deviations

Kernels can be composed

- Can be composed across multiple dimensions



Lin × SE | increasing variation | Lin × Per | growing amplitude

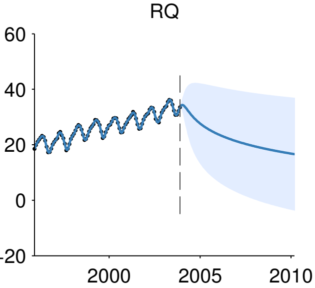


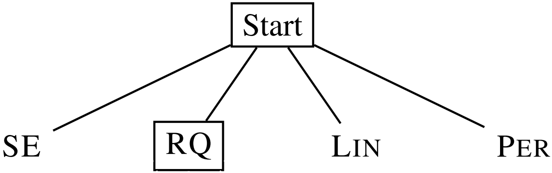
$\mathrm{SE}_1 + \mathrm{SE}_2$ | $f_1(x_1) + f_2(x_2)$ | $\mathrm{SE}_1 \times \mathrm{SE}_2$ | $f(x_1, x_2)$

SPECIAL CASES IN OUR LANGUAGE

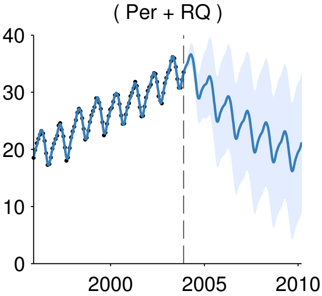
Regression motif	Example kernel
Linear regression	$\text{C} + \text{LIN}$
Fourier analysis	$\text{C} + \sum \cos$
Sparse spectrum GPs	$\sum \cos$
Spectral kernels	$\sum \text{SE} \times \cos$
Changepoints	e.g. $\text{CP}(\text{SE}, \text{SE})$
Kernel smoothing	SE
Heteroscedasticity	e.g. $\text{SE} + \text{LIN} \times \text{WN}$
Trend cyclical irregular	$\sum \text{SE} + \sum \text{PER}$
Additive nonparametric modelling	$\sum \text{SE}$

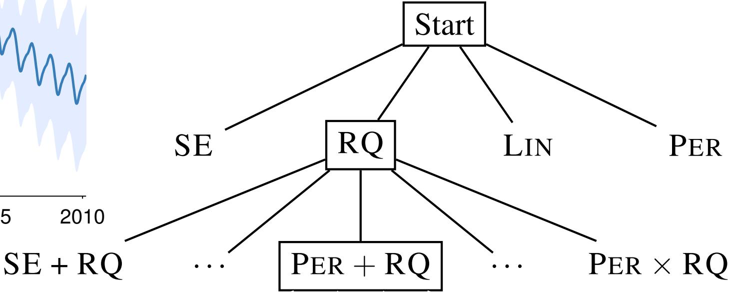
COMPOSITIONAL STRUCTURE SEARCH



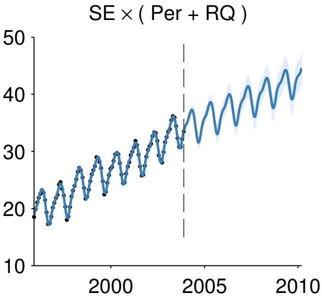


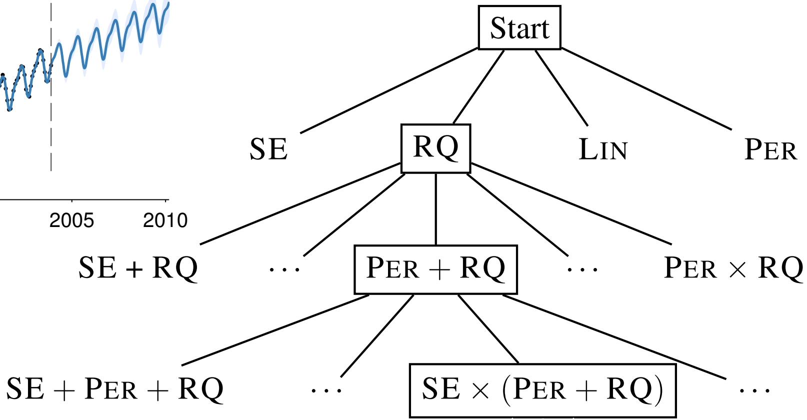
Compositional Structure Search



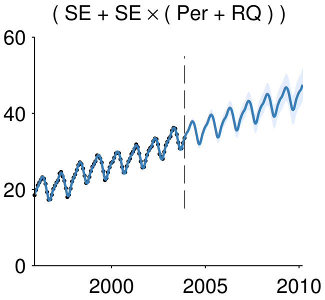


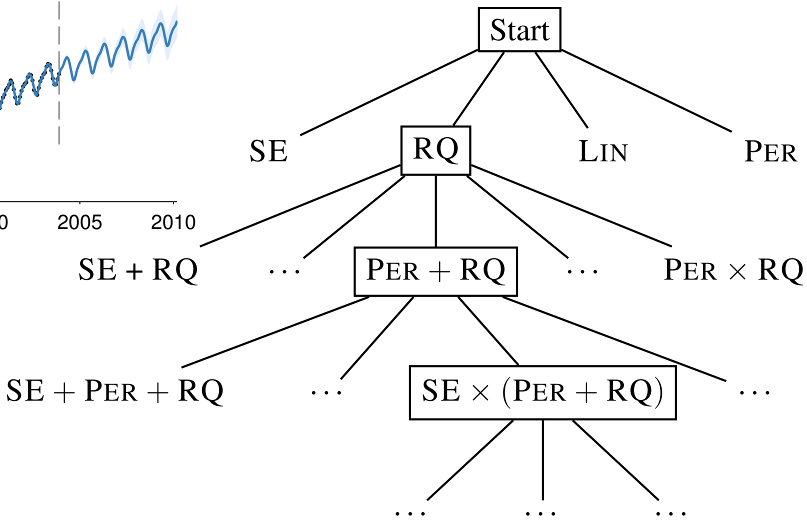
Compositional Structure Search





Compositional Structure Search





DISTRIBUTIVITY HELPS INTERPRETABILITY

We can write all kernels as sums of products of base kernels:

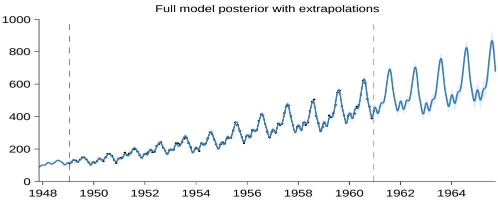
$$\text{SE} \times (\text{RQ} + \text{LIN}) = (\text{SE} \times \text{RQ}) + (\text{SE} \times \text{LIN}).$$

Sums of kernels are equivalent to sums of functions.

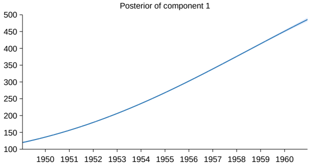
If f_1, f_2 are independent, and $f_1 \sim \mathcal{GP}(\mu_1, k_1), f_2 \sim \mathcal{GP}(\mu_2, k_2)$ then

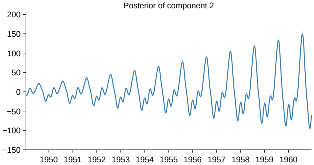
$$(f_1 + f_2) \sim \mathcal{GP}(\mu_1 + \mu_2, k_1 + k_2)$$

EXAMPLE DECOMPOSITION: AIRLINE

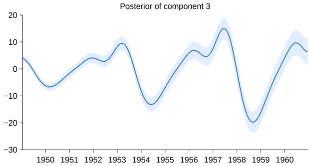


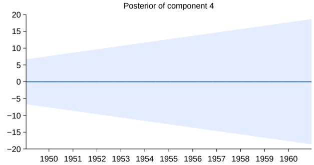
=





+





Example Kernel Descriptions

Product of Kernels	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

THIS ANALYSIS WAS AUTOMATICALLY GENERATED

The raw data and full model posterior with extrapolations are shown in figure 1.

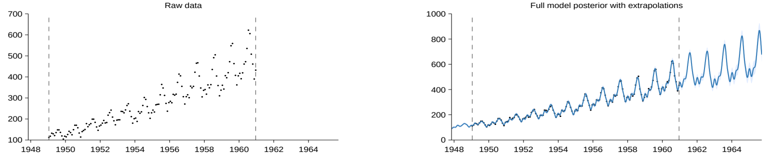


Figure 1: Raw data (left) and model posterior with extrapolation (right)

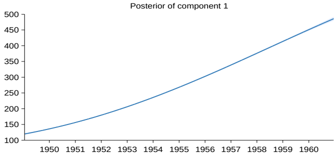
The structure search algorithm has identified four additive components in the data:

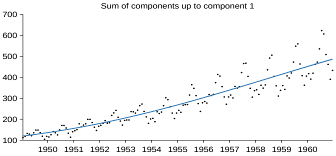
- A very smooth monotonically increasing function.
- An approximately periodic function with a period of 1.0 years and with approximately linearly increasing amplitude.
- An exactly periodic function with a period of 4.3 years but with linearly increasing amplitude.
- Uncorrelated noise with linearly increasing standard deviation.

This analysis was automatically generated

2.1 Component 1

This component is a very smooth and monotonically increasing function.

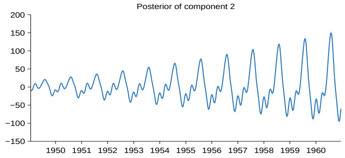


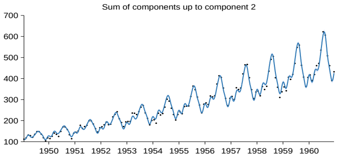


THIS ANALYSIS WAS AUTOMATICALLY GENERATED

2.2 Component 2

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 approximately linearly. The shape of this function within each period has a typical lengthscale of 6.0 weeks.

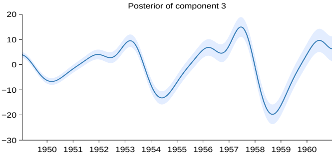


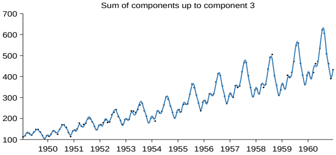


THIS ANALYSIS WAS AUTOMATICALLY GENERATED

2.3 Component 3

This component is exactly periodic with a period of 4.3 years but with varying amplitude. The amplitude of the function increases linearly. The shape of this function within each period has a typical lengthscale of 7.4 months.





SUMMARY

- Constructed a language of regression models via kernel composition
- Searched over this language greedily
- Kernels modify prior in predictable ways, allowing automatic natural-language description of models

SUMMARY

- Constructed a language of regression models via kernel composition
- Searched over this language greedily
- Kernels modify prior in predictable ways, allowing automatic natural-language description of models

Thanks!