Automatic construction and description of nonparametric models

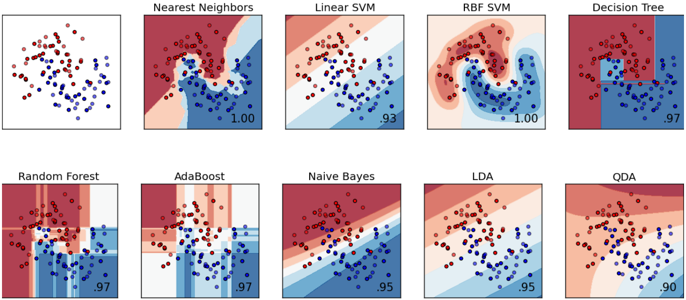
  

James Robert Lloyd, David Duvenaud, Roger Grosse,
Joshua Tenenbaum, Zoubin Ghahramani

December 9, 2013

TYPICAL STATISTICAL MODELLING

- models typically built by hand, or chosen from a fixed set



- Building by hand requires considerable expertise
- 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?

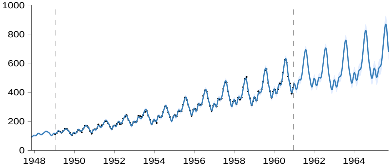
Can we do better?

- Andrew Gelman asks: How can an AI do statistics?

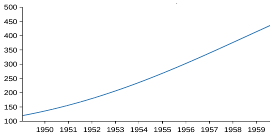
- An artificial statistician would need:
  - a language that could describe arbitrarily complicated models
  - a method of searching over those models
  - a procedure to check model fit

- This talk:We construct 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
  - Working on automatic model-checking…

Example: An automatic analysis

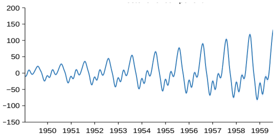


=



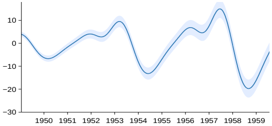
A very smooth, monotonically increasing function

+



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

+



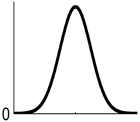
An exactly periodic function with a period of 4.3 years but with linearly increasing amplitude

A language of regression models

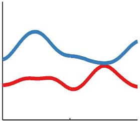
- We define a language of Gaussian process (GP) regression models by defining a language over kernel functions
- We start with a small set of base kernels and create a language with a generative grammar
  - $K \rightarrow K + K$
  - $K \rightarrow K \times K$
  - $K \rightarrow CP(K, K)$
  - $K \rightarrow \{\mathrm{SE}, \textsc{Lin}, \textsc{Per}\}$
- The language is open-ended, but its structure makes natural-language description simple

Kernels determine structure of GPs

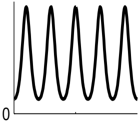
- Kernel determines almost all the properties of a GP prior
- Many different kinds, with very different properties:



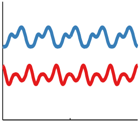
Squared-exp (SE)



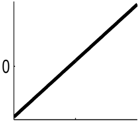
local variation



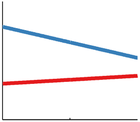
Periodic (PER)



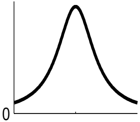
repeating structure



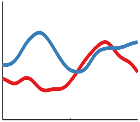
Linear (LIN)



linear functions



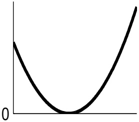
Rational-quadratic(RQ)



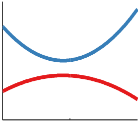
multi-scale variation

Kernels can be composed

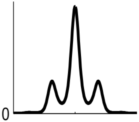
- Two main operations: addition, multiplication



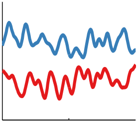
Lin × Lin



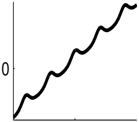
quadratic functions



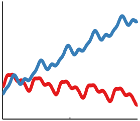
SE × Per



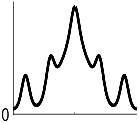
locally periodic



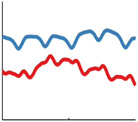
Lin + Per



periodic with trend



SE + Per

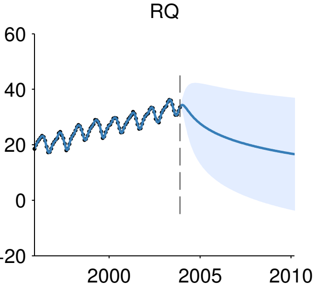


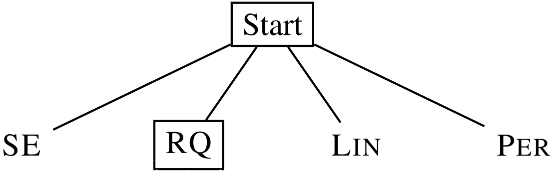
periodic with noise

Special cases in our language

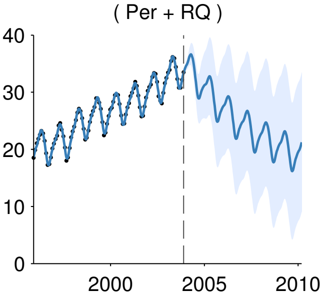
Regression motif	Example kernel
Linear regression	$\textsc{Lin}$
Quadratric regression	$\textsc{Lin} \times \textsc{Lin}$
Fourier analysis	$\sum \cos$
Spectral kernels	$\sum \text{SE} \times \cos$
Changepoints	e.g. $\text{CP}(\textsc{Per}, \text{SE})$
Heteroscedasticity	e.g. $\text{SE} + \textsc{Lin} \times \text{SE}$

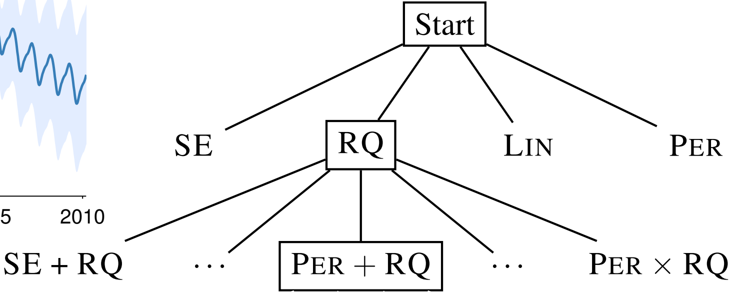
Compositional Structure Search



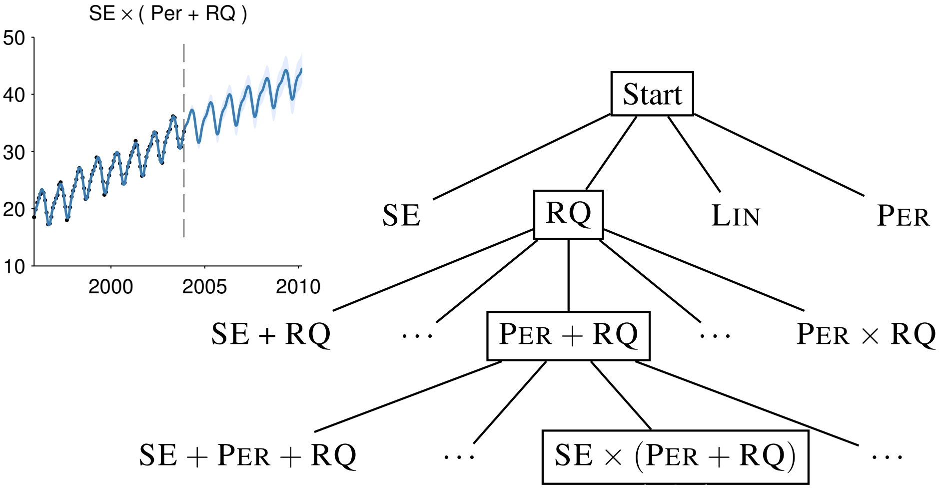


Compositional Structure Search

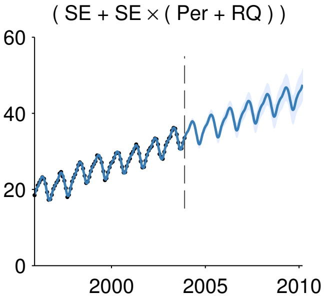


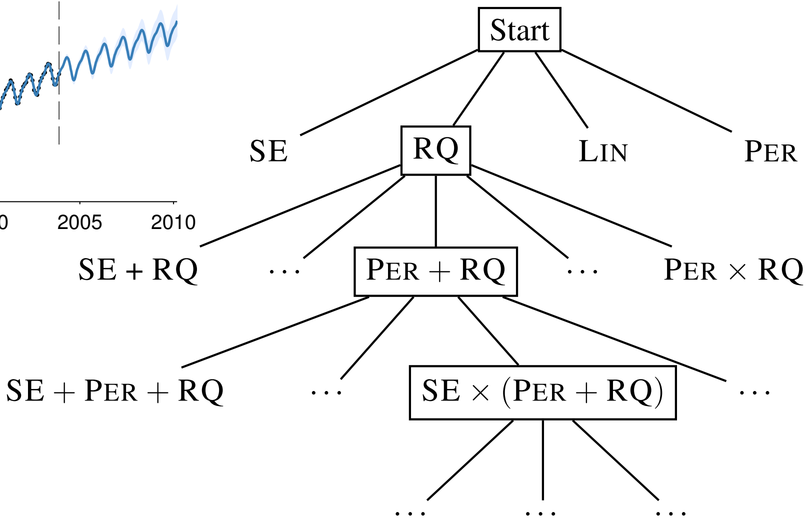


Compositional Structure Search



Compositional Structure Search





DISTRIBUTIVITY HELPS INTERPRETABILITY

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

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

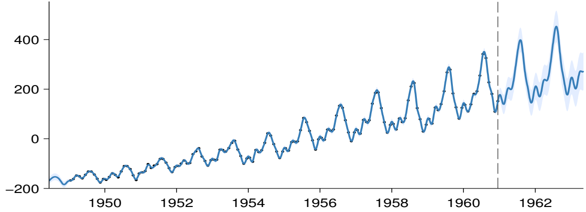
Sums of kernels are equivalent to sums of functions.

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

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

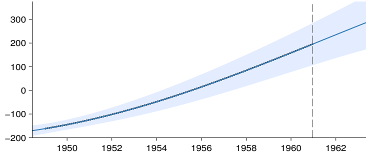
Example Decomposition: Airline

$\mathrm{SE} \times (\mathrm{Lin} + \mathrm{Lin} \times (\mathrm{Per} + \mathrm{RQ}))$

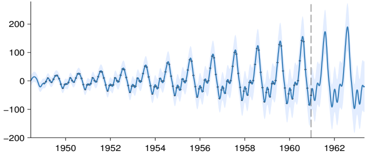


=

$\mathrm{SE} \times \mathrm{Lin}$

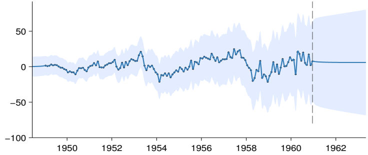


$\mathrm{SE} \times \mathrm{Lin} \times \mathrm{Per}$

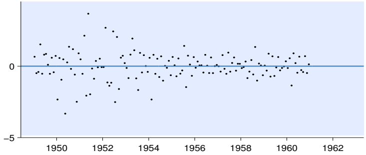


+

$\mathrm{SE} \times \mathrm{Lin} \times \mathrm{RQ}$



Residuals



Describing Kernels

Products of same type of kernel collapse.

Product of Kernels	Becomes
SE $\times$ SE $\times$ SE ...	SE
Lin $\times$ Lin $\times$ Lin ...	A polynomial
Per $\times$ Per $\times$ Per	Same covariance as product of periodic functions

DESCRIBING KERNELS

Each kernel acts as a modifier in a standard way: an “adjective”.

Kernel	Becomes
$K \times \mathrm{SE}$	’locally’ or ’approximately’
$K \times \textsc{Lin}$	’with linearly growing amplitude’
$K \times \textsc{Per}$	’periodic’
$\mathrm{CP}(K1, K2)$	’…changing at x to …’

- Special cases for when they’re on their own
- Extra adjectives depending on hyperparameters

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

THIS ANALYSIS WAS AUTOMATICALLY GENERATED

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

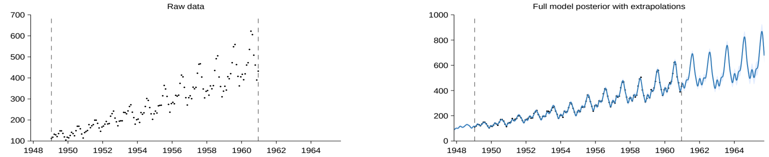


Figure 1: Raw data (left) and model posterior with extrapolation (right)

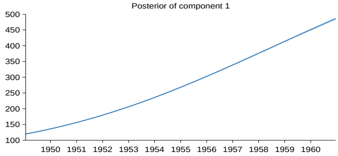
The structure search algorithm has identified four additive components in the data:

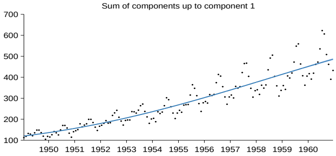
- A very smooth monotonically increasing function.
- An approximately periodic function with a period of 1.0 years and with approximately linearly increasing amplitude.
- An exactly periodic function with a period of 4.3 years but with linearly increasing amplitude.
- Uncorrelated noise with linearly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

THIS ANALYSIS WAS AUTOMATICALLY GENERATED

2.1 Component 1

This component is a very smooth and monotonically increasing function.

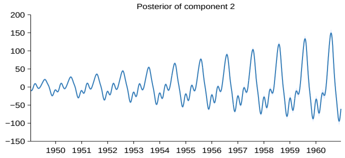


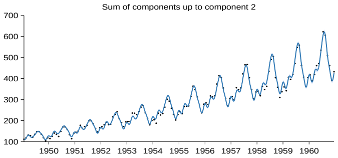


THIS ANALYSIS WAS AUTOMATICALLY GENERATED

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.

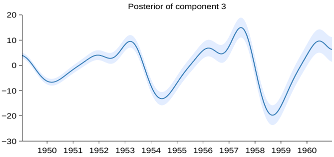


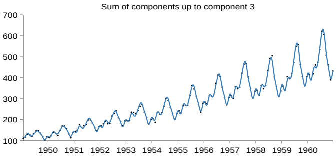


THIS ANALYSIS WAS AUTOMATICALLY GENERATED

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 through kernel composition
- Searched over this language greedily
- Kernels modify prior in predictable ways, allowing automatic natural-language description of models
- Open questions:
  - Interpretability versus flexibility
  - Automatic Model-checking

SUMMARY

- Constructed a language of regression models through kernel composition
- Searched over this language greedily
- Kernels modify prior in predictable ways, allowing automatic natural-language description of models
- Open questions:
  - Interpretability versus flexibility
  - Automatic Model-checking

Thanks!