How to build an automatic statistician



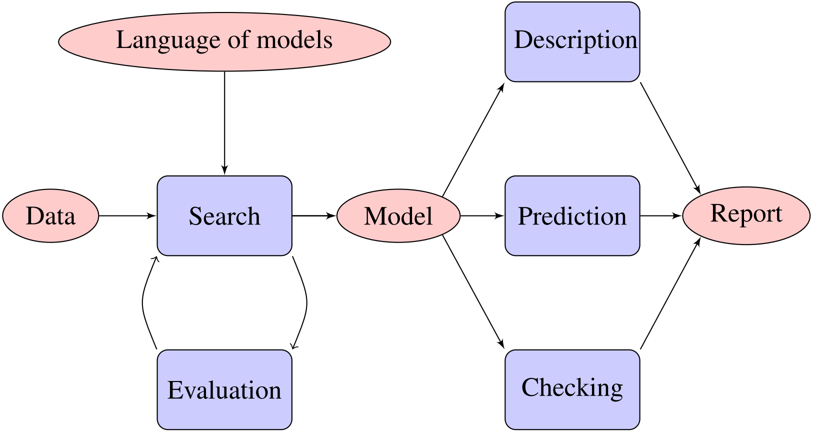
James Robert Lloyd[1], David Duvenaud[1], Roger Grosse[2],


Joshua Tenenbaum[2], Zoubin Ghahramani[1]

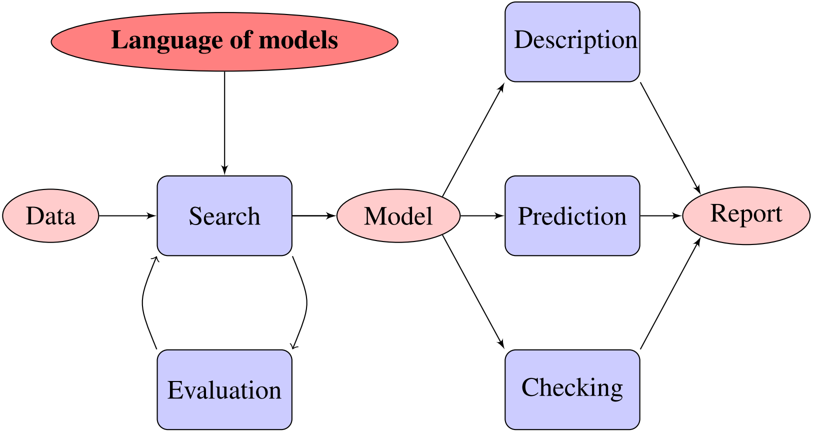
1: Department of Engineering, University of Cambridge, UK
2: Massachusetts Institute of Technology, USA

August 8, 2014

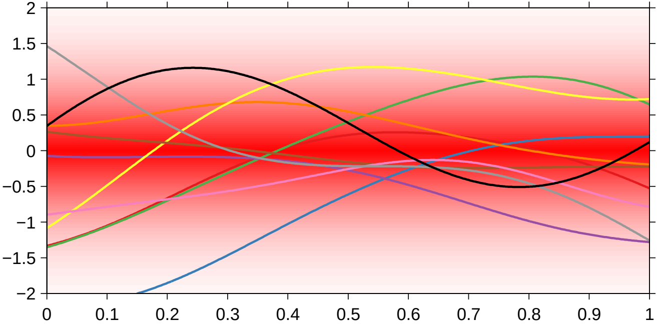
A SYSTEM FOR AUTOMATIC DATA ANALYSIS



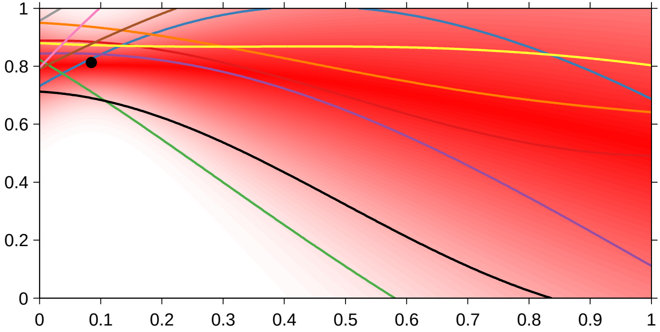
DEFINING A LANGUAGE OF MODELS



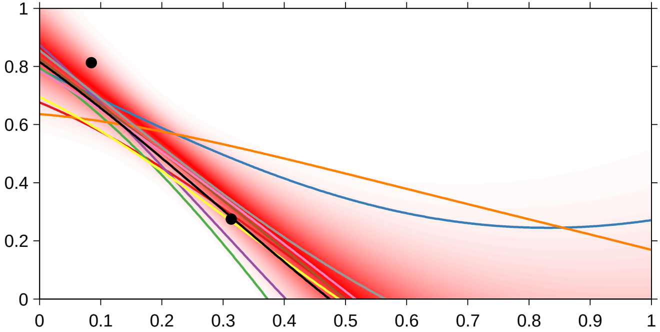
We can use Gaussian processes to place priors on functions and perform a Bayesian regression analysis



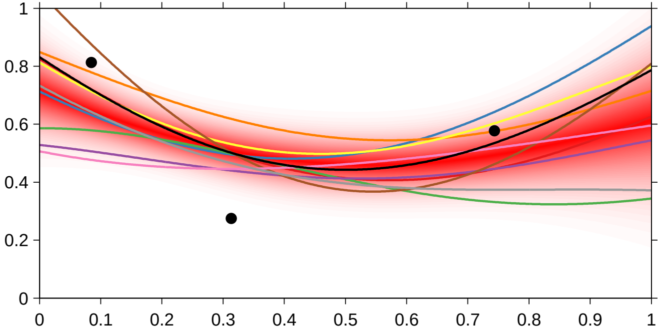
We can use Gaussian processes to place priors on functions and perform a Bayesian regression analysis



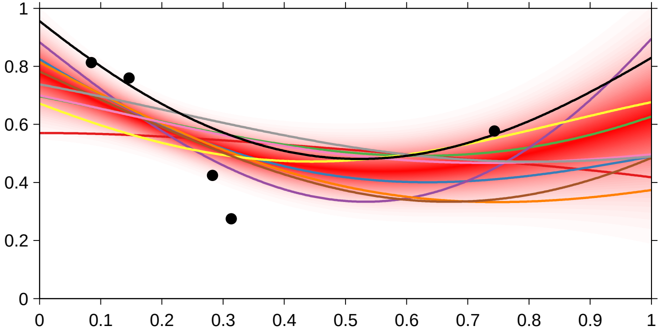
We can use Gaussian processes to place priors on functions and perform a Bayesian regression analysis



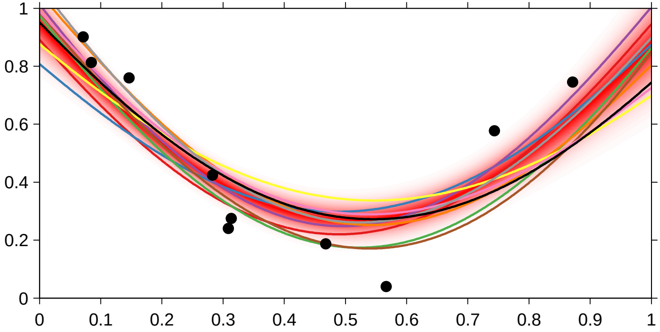
We can use Gaussian processes to place priors on functions and perform a Bayesian regression analysis



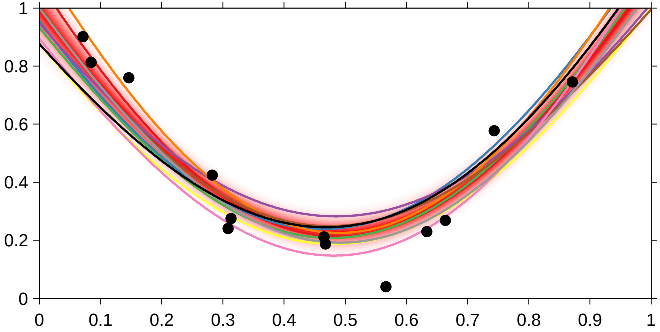
We can use Gaussian processes to place priors on functions and perform a Bayesian regression analysis



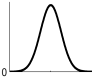
We can use Gaussian processes to place priors on functions and perform a Bayesian regression analysis



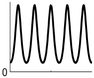
We can use Gaussian processes to place priors on functions and perform a Bayesian regression analysis



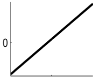
Five base kernels



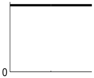
Squared exp. (SE)



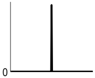
Periodic (PER)



Linear (LIN)

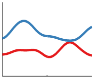


Constant (C)

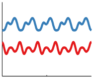


White noise (WN)

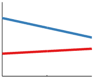
Encoding for the following types of functions



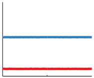
Smooth functions



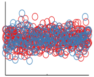
Periodic functions



Linear functions



Constant functions

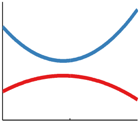


Gaussian noise

THE COMPOSITION RULES OF OUR LANGUAGE

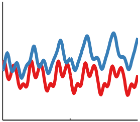
- Two main operations: addition, multiplication

LIN × LIN



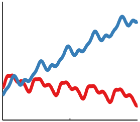
quadratic functions

SE × PER



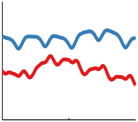
locally periodic

LIN + PER



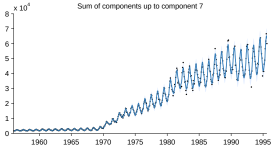
periodic plus linear trend

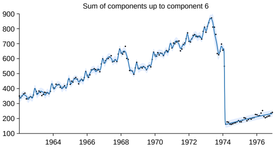
SE + PER



periodic plus smooth trend

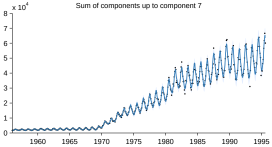
Time series data often exhibit changepoints:

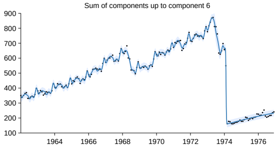




MODELING CHANGEPOINTS

Time series data often exhibit changepoints:





We can model this by assuming $f_1(x) \sim \mathrm{GP}(0, k_1)$ and $f_2(x) \sim \mathrm{GP}(0, k_2)$ and then defining

$$f(x) = (1 - \sigma(x))f_1(x) + \sigma(x)f_2(x)$$

where σ is a sigmoid function between 0 and 1.

MODELING CHANGEPOINTS

We can model this by assuming $f_1(x) \sim \text{GP}(0, k_1)$ and $f_2(x) \sim \text{GP}(0, k_2)$ and then defining

$$f(x) = (1 - \sigma(x)) f_1(x) + \sigma(x) f_2(x)$$

where σ is a sigmoid function between 0 and 1.

Then $f \sim \text{GP}(0, k)$, where

$$k(x, x') = (1 - \sigma(x)) \, k_1(x, x') \, (1 - \sigma(x')) + \sigma(x) \, k_2(x, x') \, \sigma(x')$$

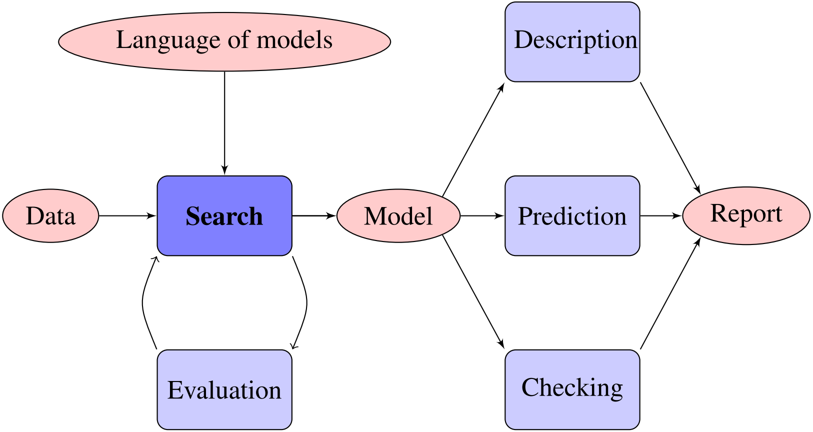
We define the changepoint operator $k = \text{CP}(k_1, k_2)$.

AN EXPRESSIVE LANGUAGE OF MODELS

Regression model	Kernel
GP smoothing	$\text{SE} + \text{WN}$
Linear regression	$\text{C} + \text{LIN} + \text{WN}$
Multiple kernel learning	$\sum \text{SE} + \text{WN}$
Trend, cyclical, irregular	$\sum \text{SE} + \sum \text{PER} + \text{WN}$
Fourier decomposition	$\text{C} + \sum \cos + \text{WN}$
Sparse spectrum GPs	$\sum \cos + \text{WN}$
Spectral mixture	$\sum \text{SE} \times \cos + \text{WN}$
Changepoints	e.g. $\text{CP}(\text{SE}, \text{SE}) + \text{WN}$
Heteroscedasticity	e.g. $\text{SE} + \text{LIN} \times \text{WN}$

Note: cos is a special case of our version of PER

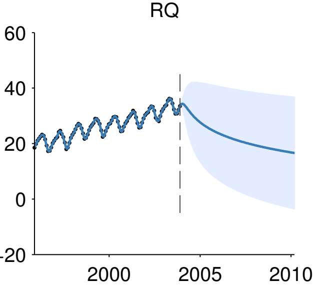
Discovering a good model via search

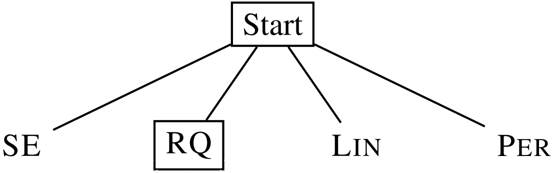


DISCOVERING A GOOD MODEL VIA SEARCH

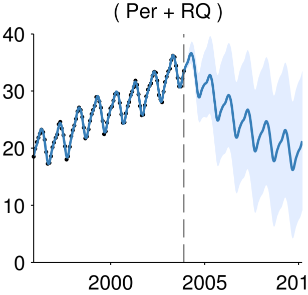
- Language defined as the arbitrary composition of five base kernels (WN, C, LIN, SE, PER) via three operators ($+, \times$, CP).
- The space spanned by this language is open-ended and can have a high branching factor requiring a judicious search
- We propose a greedy search for its simplicity and similarity to human model-building

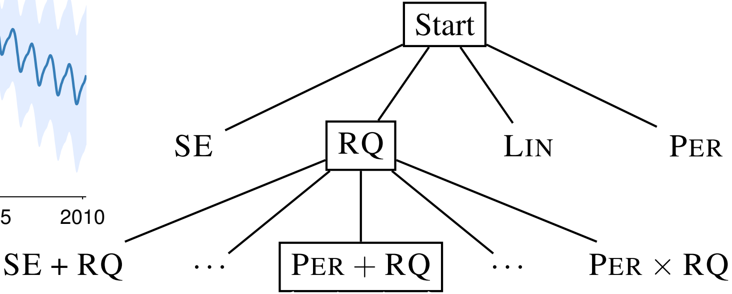
Example: Mauna Loa Keeling Curve



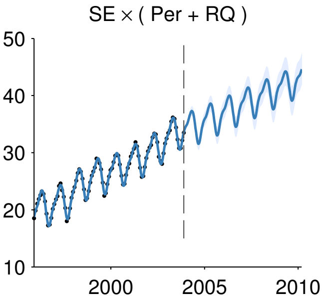


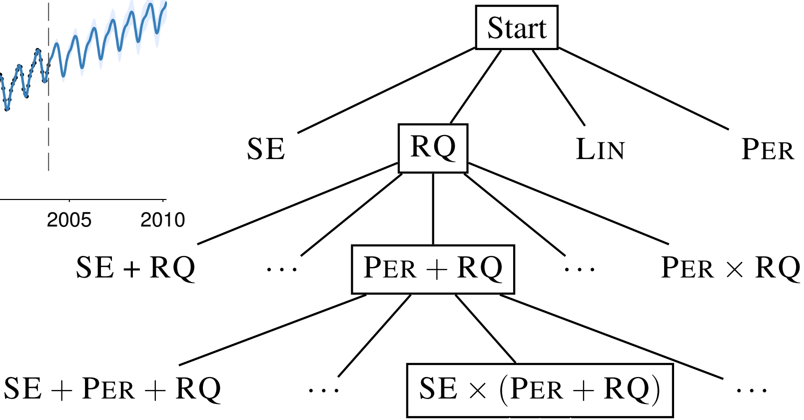
EXAMPLE: MAUNA LOA KEELING CURVE



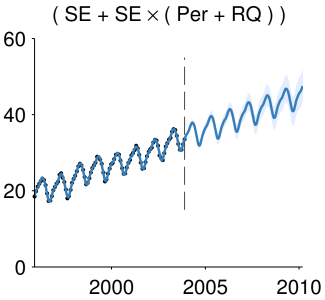


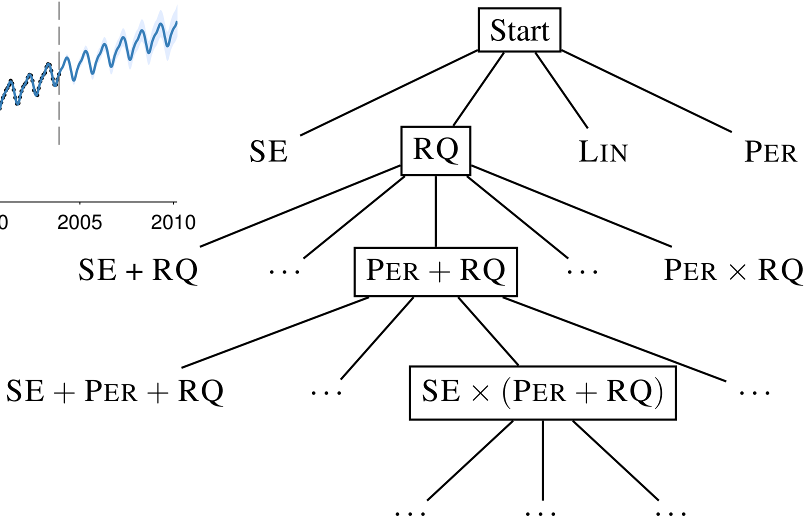
Example: Mauna Loa Keeling Curve



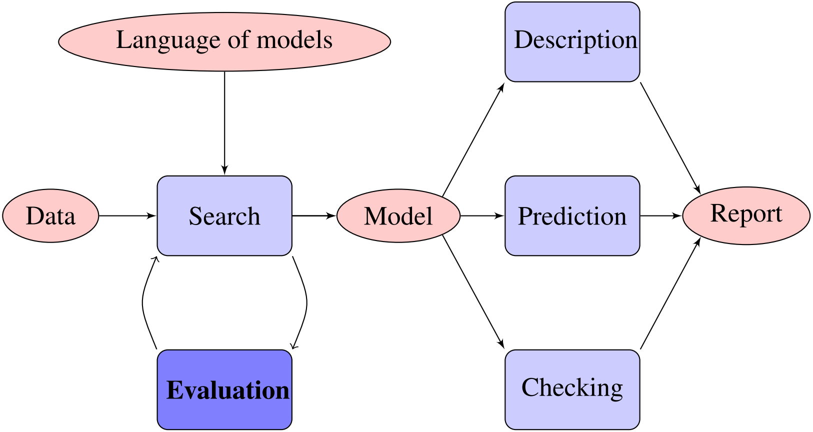


EXAMPLE: MAUNA LOA KEELING CURVE





MODEL EVALUATION



Bayesian model selection

Suppose we have a collection of models $\{M_i\}$ and some data D

Bayesian model selection

Suppose we have a collection of models $\{M_i\}$ and some data D

Bayes rule tells us

$$p(M_i \,|\, D) = \frac{p(D \,|\, M_i) p(M_i)}{p(D)}$$

Bayesian model selection

Suppose we have a collection of models $\{M_i\}$ and some data D

Bayes rule tells us

$$p(M_i \mid D) = \frac{p(D \mid M_i) p(M_i)}{p(D)}$$

If $p(M_i)$ is equal for all i (prior ignorance) then

$$p(M_i \mid D) \propto p(D \mid M_i) = \int p(D \mid \theta_i, M_i) p(\theta_i \mid M_i) \mathrm{d}\theta_i$$

BAYESIAN MODEL SELECTION

Suppose we have a collection of models $\{M_i\}$ and some data D

Bayes rule tells us

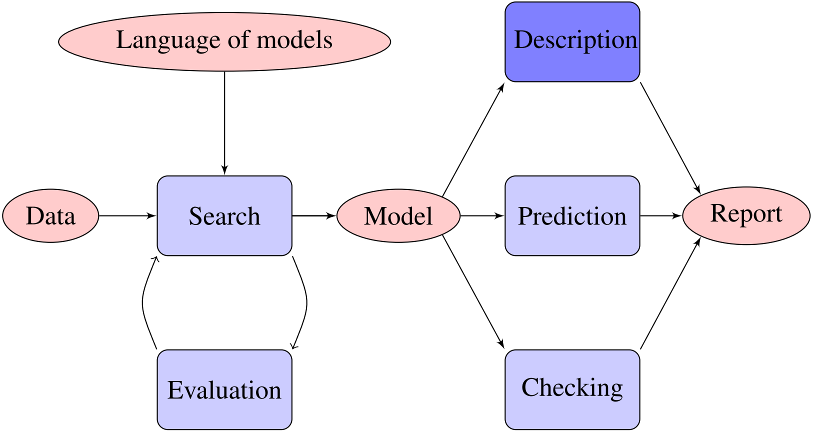
$$p(M_i \mid D) = \frac{p(D \mid M_i)p(M_i)}{p(D)}$$

If $p(M_i)$ is equal for all i (prior ignorance) then

$$p(M_i \mid D) \propto p(D \mid M_i) = \int p(D \mid \theta_i, M_i)p(\theta_i \mid M_i)\mathrm{d}\theta_i$$

i.e. The most likely model has the highest marginal likelihood

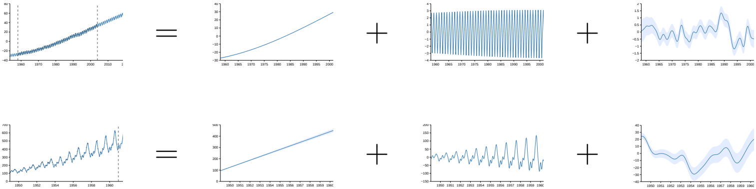
Automatic translation of models



If $f_1 \sim \mathrm{GP}(0, k_1)$ and independently $f_2 \sim \mathrm{GP}(0, k_2)$ then

$$f_1 + f_2 \sim \mathrm{GP}(0, k_1 + k_2)$$

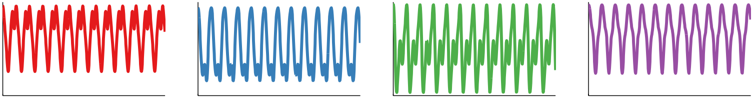
e.g.



We can therefore describe each component separately

$$\underbrace{\text{PER}}_{\text{periodic function}}$$

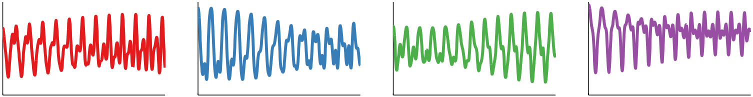
On their own, each kernel is described by a standard noun phrase



PRODUCTS OF KERNELS - SE

$$\underbrace{\text{SE}}_{\text{approximately}} \times \underbrace{\text{PER}}_{\text{periodic function}}$$

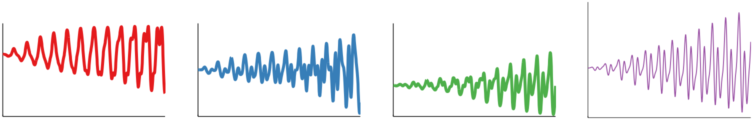
Multiplication by SE removes long range correlations from a model since $\text{SE}(x, x')$ decreases monotonically to 0 as $|x - x'|$ increases.



PRODUCTS OF KERNELS - LIN

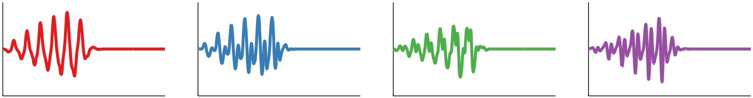
$$\underbrace{\text{SE}}_{\text{approximately}} \times \underbrace{\text{PER}}_{\text{periodic function}} \times \underbrace{\text{LIN}}_{\text{with linearly growing amplitude}}$$

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.



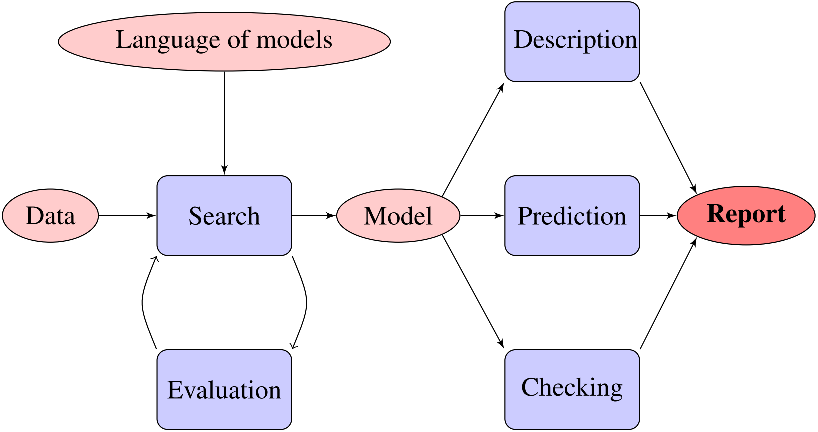
$$\underbrace{\text{SE}}_{\text{approximately}} \times \underbrace{\text{PER}}_{\text{periodic function}} \times \underbrace{\text{LIN}}_{\text{with linearly growing amplitude}} \times \underbrace{\boldsymbol{\sigma}}_{\text{until 1700}}$$

Multiplication by $\boldsymbol{\sigma}$ is equivalent to multiplying the function being modeled by a sigmoid.

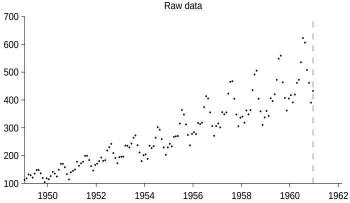


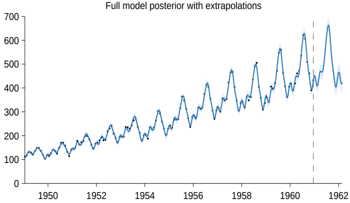
Kernel	Noun phrase	Postmodifier phrase
WN	uncorrelated noise	n/a
C	constant	n/a
SE	smooth function	whose shape changes smoothly
PER	periodic function	modulated by a periodic function
LIN	linear function	with linearly varying amplitude
$\prod_k \text{LIN}^{(k)}$	polynomial	with polynomially varying amplitude
$\prod_k \boldsymbol{\sigma}^{(k)}$	n/a	which applies until / from [changepoint]

AUTOMATICALLY GENERATED REPORTS



EXAMPLE: AIRLINE PASSENGER VOLUME



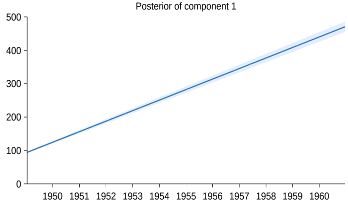


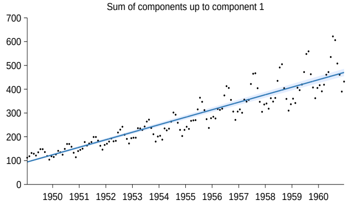
Four additive components have been identified in the data

- 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.

Example: Airline passenger volume

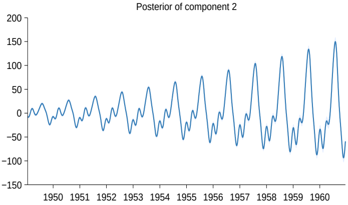
This component is linearly increasing.

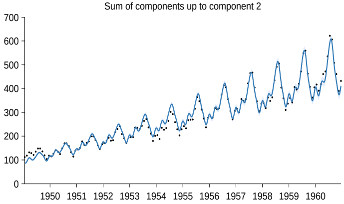




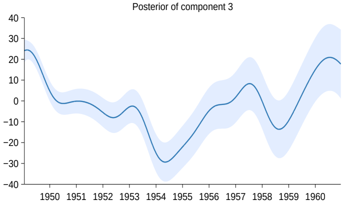
EXAMPLE: AIRLINE PASSENGER VOLUME

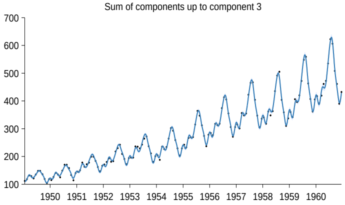
This component is approximately periodic with a period of 1.0 years and varying amplitude. Across periods the shape of this function varies very smoothly. The amplitude of the function increases linearly. The shape of this function within each period has a typical lengthscale of 6.0 weeks.





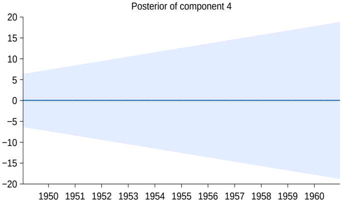
This component is a smooth function with a typical lengthscale of 8.1 months.

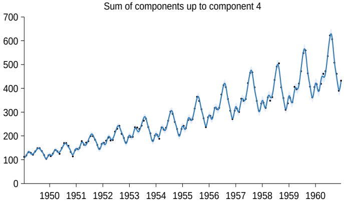




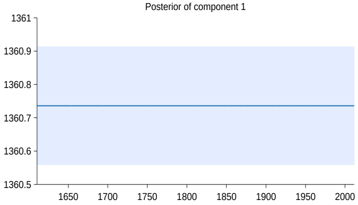
Example: Airline passenger volume

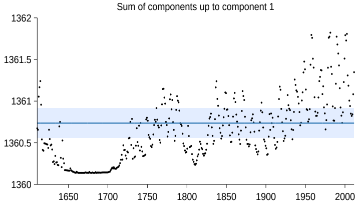
This component models uncorrelated noise. The standard deviation of the noise increases linearly.



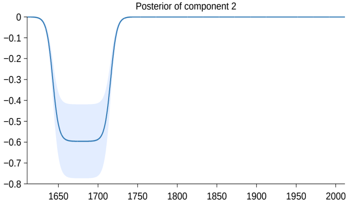


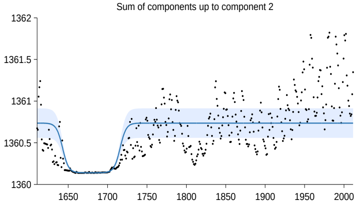
This component is constant.





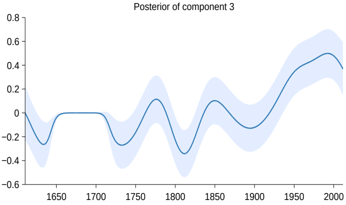
This component is constant. This component applies from 1643 until 1716.

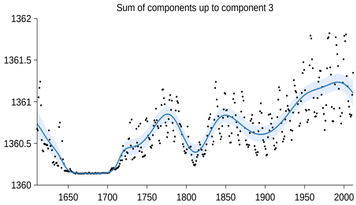




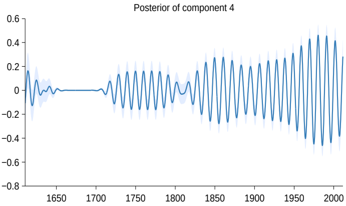
Example: Solar irradiance

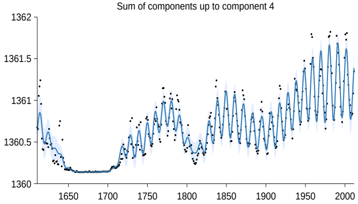
This component is a smooth function with a typical lengthscale of 23.1 years. This component applies until 1643 and from 1716 onwards.





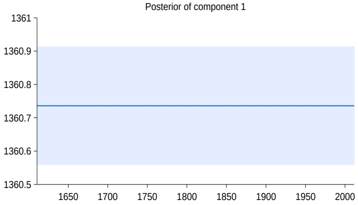
This component is approximately periodic with a period of 10.8 years. Across periods the shape of this function varies smoothly with a typical lengthscale of 36.9 years. The shape of this function within each period is very smooth and resembles a sinusoid. This component applies until 1643 and from 1716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

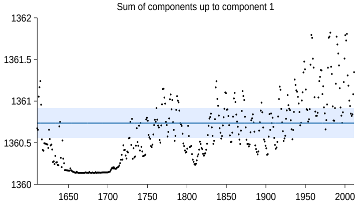




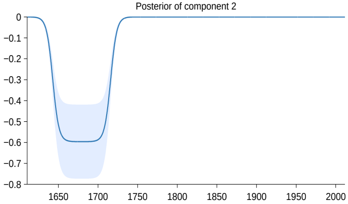
EXAMPLE: SOLAR IRRADIANCE

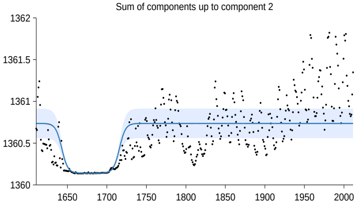
This component is constant.



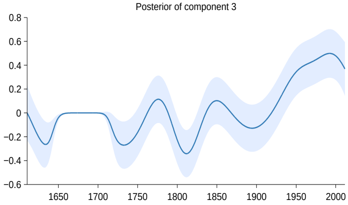


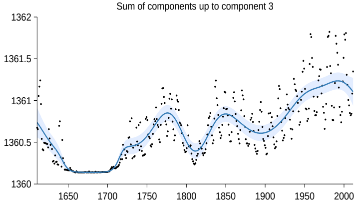
This component is constant. This component applies from 1643 until 1716.





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





EXAMPLE: SOLAR IRRADIANCE

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.