Gaussian Processes (Presentation of Roberts et al., 2012)

Alan Aw

Department of Statistics

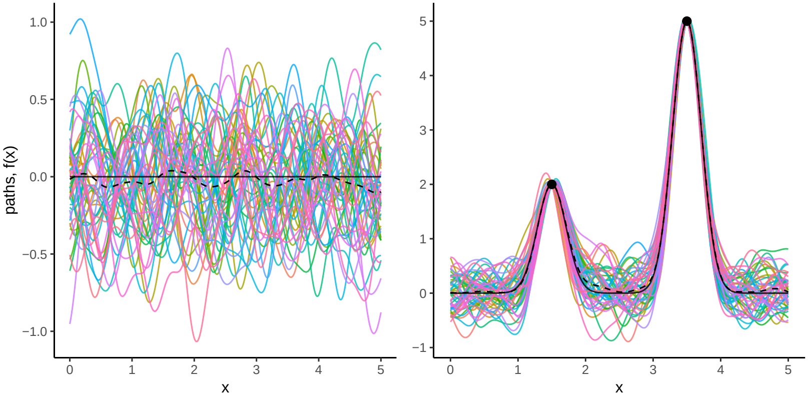
March 11, 2019

I am more familiar with R than Matlab, so...

Feel free to download code to play with Gaussian Processes with your bare hands:

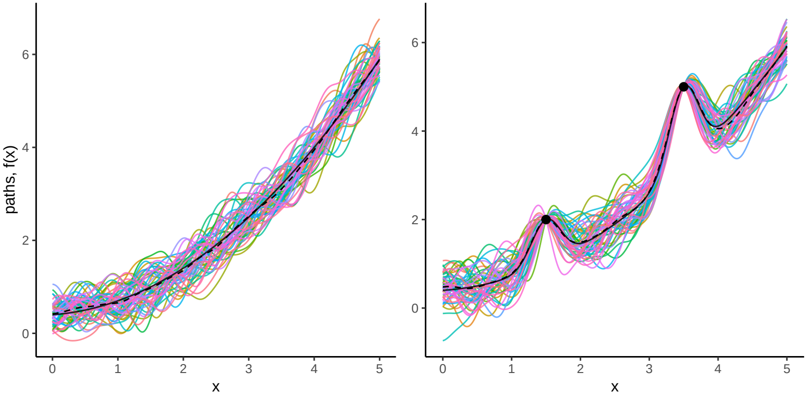
1. Go to `https://github.com/alanaw1/gaussianprocesses/`
2. Download the **R** script

Conditioning on Observed Data



SE kernel, zero mean, w/o and with two points observed

Effect of Changing the Mean



SE kernel, quadratic mean, w/o and with two points observed

Hyperparameters matter

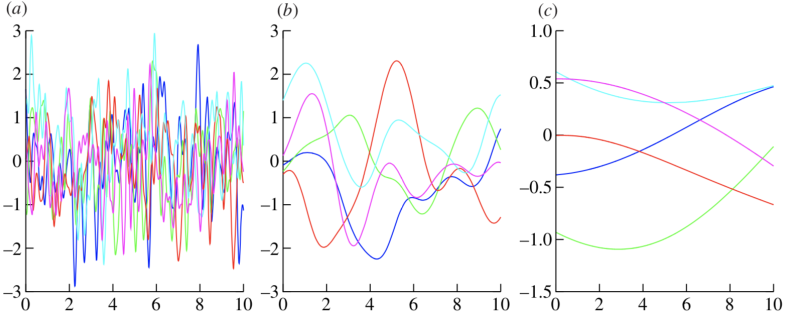
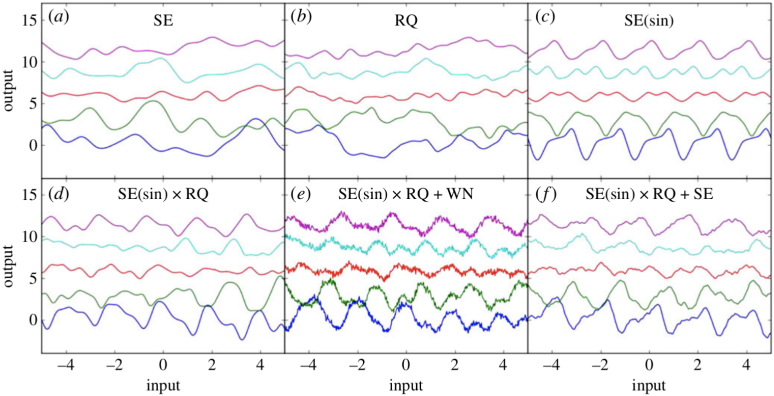


Figure 5. (*a*–*c*) Functions drawn from a GP with a squared exponential covariance function with output scale $h = 1$ and length scales $\lambda = 0.1$ (*a*), 1 (*b*), 10 (*c*). (Online version in colour.)

SE kernel with varying hyperparameters. Source: Roberts et al. (2012)

Some kernel recipes



Combinations of different kernels allow flexible modelling. Source: Roberts et al. (2012)

Modelling light curves of transiting exoplanets

Based on Gibson et al. (2012)

Context: Modeling instrumental systematics with application to archival NICMOS transmission spectroscopy of the hot Jupiter HD 189733

Data: Transit light data, external state variables (temperature of light detector, orbital phase, position of host star, etc.), flux variability of host star

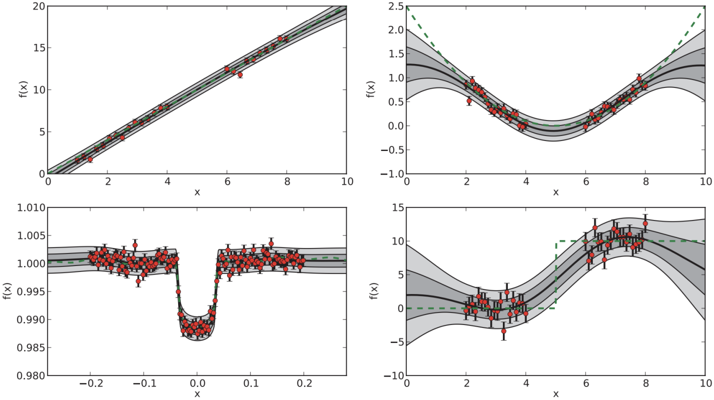
Modelling light curves of transiting exoplanets

Goal: Discover and characterise extra-solar planets through observing light curves, while accounting for instrumental systematics

Method: Fit a multi-input GP

- Use a complex mean function that encodes physical relationship between transit light curves and time (aka "planetary transit function")
- GP with SE covariance kernel used to model instrumental systematics

Effect of different GP mean functions on inference

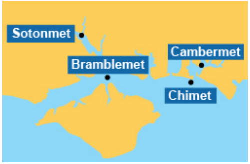


(Clockwise, from top left) linear, quadratic, step function, planetary transit function. Source: Gibson et al. (2012)

Multi-dimensional weather sensor data

Based on Osborne et al. (2007)

Context: Four sensors on South coast of the UK



Source: Weather Reports from Bramble Bank
(`www.bramblemet.co.uk`)

Data: Environmental variables per sensor: wind speed/direction, air temperature, sea temperature, height, etc. (*two* data streams, one is real-time and other is retrospective)

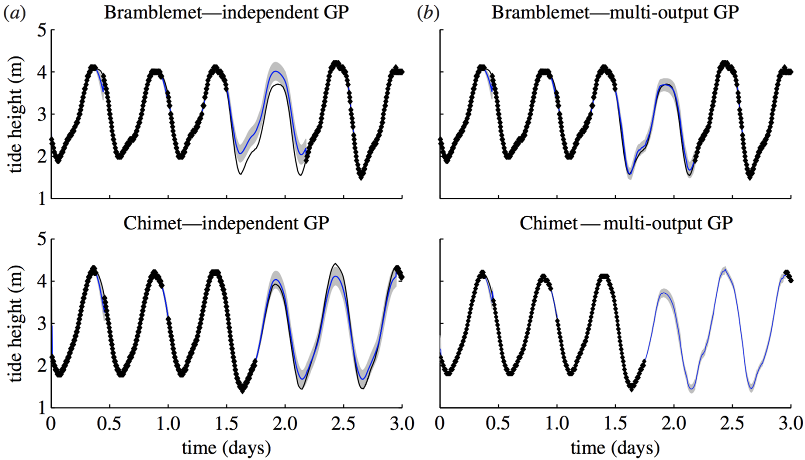
Multi-dimensional weather sensor data

Goal: Build a system that can adaptively sample and process information cost-effectively (including relying on fewer points to predict missing or future values)

Method: Fit a multi-input and multi-output GP with active data selection

- Uses spherical decomposition kernel for $K_L(\ell_m, \ell_n)$ (see Pinheiro & Bates, 1996)
- Active data selection: algorithm is simply induced to make a reading whenever the uncertainty grows beyond a pre-specified threshold (p. 5 of paper)

Four independent GPs vs multi-output GP



Black is actual data, blue is predicted from GP. Can you tell which is which? Source: Roberts et al. (2012)

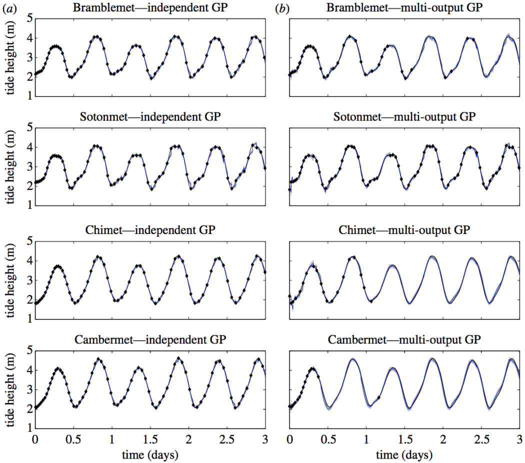
Comparison of GP with other time series methods

Table 1. Predictive performances for 5 day Bramblemet tide height dataset. We note the superior performance of the GP compared with a more standard Kalman filter model. Error metrics shown are root mean square error (r.m.s.e) and normalized mean square error (n.m.s.e.), which is presented on a logarithmic, decibel scale.

algorithm	r.m.s.e (m)	n.m.s.e. (dB)
naive	7.5×10^{-1}	-2.1
Kalman filter	1.7×10^{-1}	-15.2
independent GPs	8.7×10^{-2}	-20.3
multi-output GP	3.8×10^{-2}	-27.6

Source: Roberts et al. (2012)

Active data selection



Sampler always chooses Sotonmet readings, because of their complexity owing to existence of "young flood stand" and "double high tide." Source: Roberts et al. (2012)

Whither, from here?

1. Rasmussen & Williams (2005) contains plentiful information for building models from basic GP blocks.
2. Applications to Biology (potential project ideas)
   - Population genetics: inferring demographic histories. See Palacios, Wakeley & Ramachandran (2015) *Genetics*; and Palacios & Minin (2013) *Biometrics*
   - Influenza dynamics: work by Palacios, Wang and Hernandez using Twitter data to create multi-input GP to better model and predict seasonal influenza rates (motivated by the CDC "Predict the Influenza Season Challenge," `www.cdc.gov/flu/news/predict-flu-challenge.htm`)
   - Gene expression: Profiling transcriptome-wide time series expression. See McDowell et al. (2018) *PLoS Comp. Biol.*
   - Gene-specific branching dynamics: Fit branching Gaussian process to single-cell RNA-seq data to infer branching times. See Boukouvalas, Hensman & Rattray (2018) *Genome Biol.*

R Packages for GPs

- `mlegp` is a popular package
- `INLA` (Integrated Nested Laplace Approximation) can fit GPs with Laplace approximation for integration over hyperparameter space
  - Tutorial (`http://www.maths.bath.ac.uk/ jjf23/brinla/gpreg.html`)
- `GPfit` based on a new optimisation algorithm (see MacDonald, Ranjan & Chipman, 2013)

Concluding Remarks

1. GPs are flexible models for dynamic, probabilistic (aka, stochastic) processes
2. Means and covariances can be specified by user before performing inference
3. Including priors on covariance hyperparameters leads to complicated issues with integration, requiring tools like Laplace approximation, MCMC over posterior, and Bayesian quadrature
4. Has potential applications to biological questions