Bayesian Online Changepoint Detection

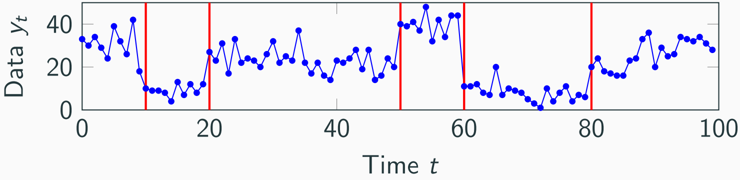
for Multivariate Point Processes

Jay Ng

15 March 2019

BSc Data Science

Motivation



What is it?

- Time-series segmentation; Partition into sub-sequences.
- Boundaries between partitions are *Changepoints (CPs)*.
- Identify abrupt changes in a data sequence.

How is this useful?

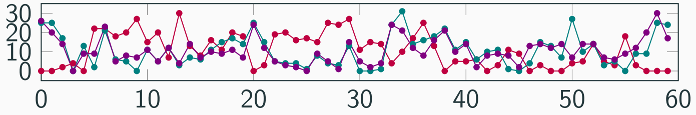
- Climate change detection.
- Financial forecasting.
- Medical condition monitoring.

... and more ...

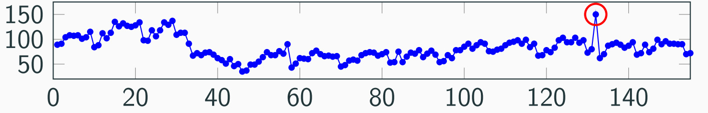
Challenges

Why is it difficult?

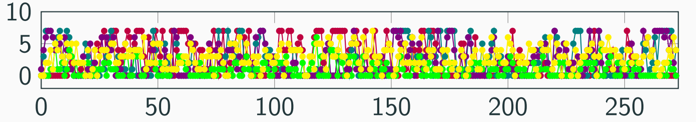
- Modelling dependencies in multiple data streams, in space and time.



- Robustness, i.e. dealing with outliers.



- Scalability to high-dimensional data. Speed is a concern.



Contributions

Literature

1. Changepoint Detection
2. Point Processes, Gaussian Processes

Novelty

1. Implemented a non-parametric model for univariate count data.
    - Built in Python3.6 using libraries GPy and GPFlow.
    - 3 inference methods: Laplace, VI, SVI.
2. Speed up the model using sparse approximations.
3. Extended the model to a multivariate setting.
4. Tested the model using two real-world and one artificial dataset.

Key Concepts

Why Bayesian Inference?

$$\overbrace{p(\theta|X,y)}^{\text{posterior}} = \frac{p(y|X,\theta)p(\theta)}{p(y|X)} \propto \overbrace{p(y|X,\theta)}^{\text{likelihood (update beliefs!)}} \overbrace{p(\theta)}^{\text{prior}}$$

Maximum A Posteriori (MAP) estimate

$$\theta_{\text{MAP}} = \arg\max_{\theta \in \Theta} p(y|X,\theta)p(\theta)$$

Conjugacy

Assume posterior and prior are in the same family of distributions
$\implies$ available closed-form expression for posterior.

Stationarity

Modelling a time series $y := (y_1, \dots y_t)$ as a piecewise stationary process.

strictly stationary if $\mathbb{P}(y_1, \dots y_t) \stackrel{\text{def}}{=} \mathbb{P}(y_{i+1}, \dots, y_{i+t})$, $i \in \mathbb{N}$, $\forall t \in \mathbb{N}$.

weakly stationary if y has constant mean and variance $\forall t \in \mathbb{N}$.

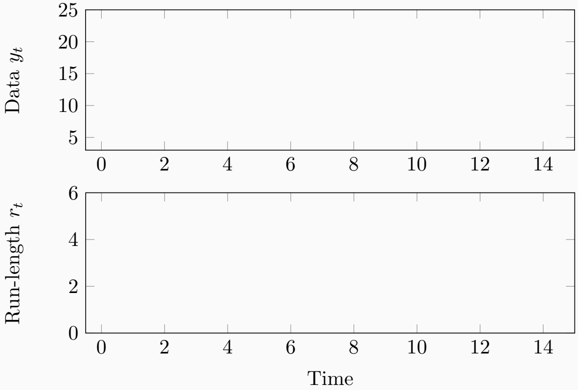
Outline

1. Bayesian On-line Changepoint (CP) Detection
2. Point Processes & Gaussian Processes
3. Applications
4. Multivariate Extensions

Bayesian On-line Changepoint (CP) Detection I/III

Standard BOCPD (Adams and MacKay, 2007) [5]

- Run-length at $t = r_t \iff$ a change-point occured at time $t - r_t$.
- Inference on last change-point $p(r_t|y_{1:t})$.
- $\mathcal{O}(t)$ linear time complexity instead of $\mathcal{O}(\prod_{i=1}^{t} i)$ factorial.



Bayesian On-line Changepoint (CP) Detection I/III

Standard BOCPD (Adams and MacKay, 2007) [5]

- Run-length at $t = r_t \iff$ a change-point occured at time $t - r_t$.
- Inference on last change-point $p(r_t | y_{1:t})$.
- $\mathcal{O}(t)$ linear time complexity instead of $\mathcal{O}(\prod_{i=1}^{t} i)$ factorial.

Posterior predictive:

$$f(y_t | y_{1:(t-1)}, r_{t-1}) = \int_{\Theta} f(y_t|\theta)\pi(\theta|y_{(t-r_{t-1}):(t-1)})d\theta \tag{1}$$

Inference via Recursion:

$$p(y_1, r_1 = 0) = \int_{\Theta} f(y_1|\theta)\pi(\theta)d\theta = f(y_1|y_0)$$

$$p(y_{1:t}, r_t) = \sum_{r_{t-1}} \left\{ \underbrace{f(y_t|y_{1:(t-1)}, r_{t-1})}_{\text{posterior predictive}} \underbrace{H(r_t, r_{t-1})}_{\text{CP Hazard prior}} \underbrace{p(y_{1:(t-1)}, r_{t-1})}_{\text{recursive term}} \right\} \tag{2}$$

Limitations:

1. Assumes a single model. Hard to infer best parameters!

Bayesian On-line Changepoint (CP) Detection II/III

Model Selection

Idea by Knoblauch and Damoulas (2018) [3], unifying AM [5] and FL [2].

Introduce m_t at time t, within a model universe $\mathcal{M}$.

New Recursion:

$$p(y_1, r_1 = 0, m_1) = q(m_1) f_{m_1}(y_1 | y_0)$$

$$p(y_{1:t}, r_t, m_t) = \sum_{m_{t-1}, r_{t-1}} \Big\{ f_{m_t}(y_t | y_{1:(t-1)}, r_{t-1}) \overbrace{q(m_t | y_{1:(t-1)}, r_{t-1}, m_{t-1})}^{\text{new term for model dist.}}$$

$$H(r_t, r_{t-1}) p(y_{1:(t-1)}, r_{t-1}, m_{t-1}) \Big\}$$

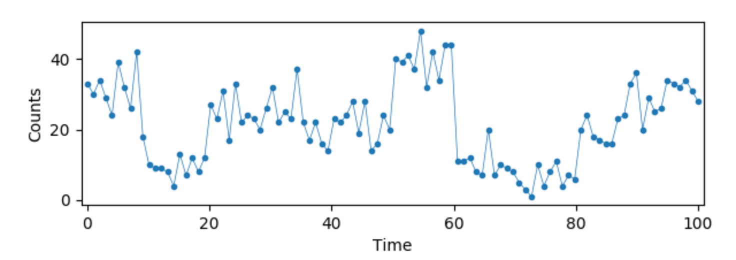
Bayesian On-line Changepoint (CP) Detection III/III

Inference:

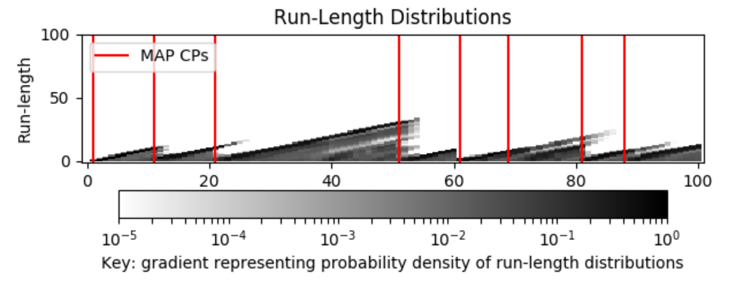
1. Evidence $p(y_{1:t}) = \sum_{r_t, m_t} p(y_{1:t}, r_t, m_t)$
2. Run-length & model posterior: $p(r_t, m_t | y_{1:t}) = p(y_{1:t}, r_t, m_t) / p(y_{1:t})$
3. Predictive: $p(y_{t+1} | y_{1:t}) = \sum_{r_t, m_t} f_{m_t}(y_{t+1} | y_{1:t}, r_t) p(r_t, m_t | y_{1:t})$
4. Run-length marginal posterior: $p(r_t | y_{1:t}) = \sum_{m_t} p(r_t, m_t | y_{1:t})$
5. Model marginal posterior: $p(m_t | y_{1:t}) = \sum_{r_t} p(r_t, m_t | y_{1:t})$
6. Maximum A Posteriori segmentation:

$$MAP_t = \arg\max_{r,m} \{ p(r_t = r, m_t = m | y_{1:t}) MAP_{t-r-1} \}$$

Demo



Counts
0
20
40
Time
0
20
40
60
80
100



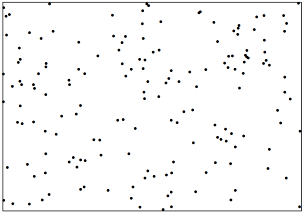
Run-Length Distributions
Run-length
0
50
100
MAP CPs
0
20
40
60
80
100
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
Key: gradient representing probability density of run-length distributions

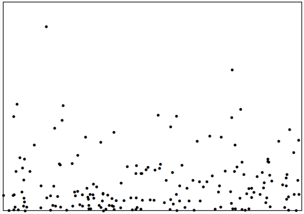
1. Parametric models assume some finite set of parameters θ
2. Non-parametric models can often be defined by assuming an infinite dimensional θ. Usually we think of θ as a function. $\implies$ flexibility.
3. Introduce a specific model class for spatial and spatio-temporal point process data, known as the log Gaussian cox process (LGCP).

Outline

1. Bayesian On-line Changepoint (CP) Detection
2. Point Processes & Gaussian Processes
3. Applications
4. Multivariate Extensions

Point Processes & Gaussian Processes I/III





Poisson Process

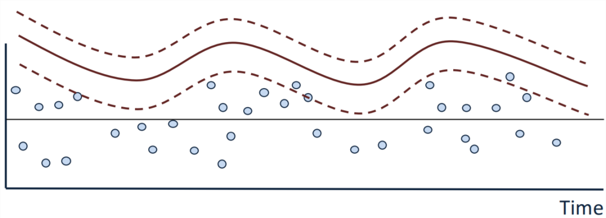
X defined on S with intensity measure μ and intensity function ρ satisfies any bounded region $B \subseteq S$ with $\mu(B) > 0$.

- $N(B)$ is Poisson distributed with mean $\mu(B)$
- Homogeneous if $\rho(u)$ is constant $\forall u \in S$

Cox Process

An in-homogeneous Poisson process driven by a stochastic intensity Λ.

Point Processes & Gaussian Processes II/III



Log Gaussian Cox Process (Moller et al., 1998) [4] [1]

$$f \sim \mathcal{GP}(m(x), k(x, x'))$$
$$\Lambda = \exp(f)$$
$$y|\Lambda \sim \text{Poisson}(\Lambda)$$

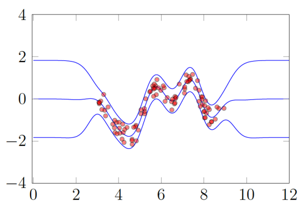
Take exponential for positive intensity. f is a Gaussian Process, a collection of random variables, any finite number of which have a joint Gaussian distribution.

Point Processes & Gaussian Processes III/III

Gaussian Processes (Rasmussen and Williams, 2006) [6]

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

kernel function $k(x, x')$ computes the covariance matrix, which must be positive semi-definite



Why? Predictive distribution is

$$f_* | x_*, x, f \sim \mathcal{N}(k(x_*, x){\color{red}k(x, x)^{-1}}f,\quad k(x_*, x_*) - k(x_*, x){\color{red}k(x, x)^{-1}}k(x, x_*))$$

Many kernels! Radial Basis Function with hyperparameters (ℓ, σ^2)

$$k(x, x') = \sigma^2 \exp\left\{ -\frac{1}{2\ell^2}(x - x')^2 \right\}$$

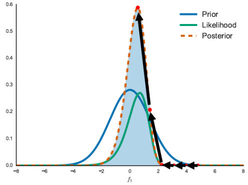
However...

- ${\color{red}\mathcal{O}(N^3)}$ time complexity to invert covariance matrix!
- Intractable for non-Gaussian likelihoods.

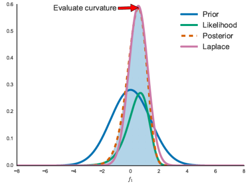
Solution: Approximate the posterior

Laplace approximation

1. 'Try' to find mode via Newton's method.
2. Second order Taylor expansion around the mode.
3. Posterior $p(f|y)$ approximates as a Gaussian distribution $\mathcal{N}(f|f_{mode}, A^{-1})$ $s.t.$ A is a negative Hessian matrix.



(a) Mode finding via Newton optimisation



(b) Evaluate curvature at mode, use mode and curvature to approximate the posterior

✔ Fast, scalable!

✖ Information loss.

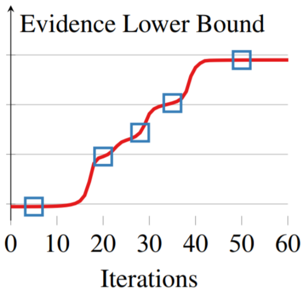
Solution: Approximate the posterior

Variational Inference

1. Goal: minimize $\mathcal{KL}(q(f|\theta_V) \parallel p(f|y))$.
2. Optimize $ELBO(q(f|\theta_V))$ via Coordinate Ascent VI.
3. Mean-field variational $q(\theta_V) = \prod_i q(\theta_{V\,i})$.

✔ Greater accuracy.

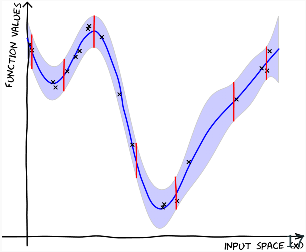
✖ Expensive, non-deterministic.



Sparse Variational Inference

1. Low rank approximation using inducing inputs $|Z| = M$.
2. Find $p(f_*|u, Z, x_*)$ instead of $p(f_*|f, x, x_*)$.
3. $\mathcal{O}(NM^2)$ time complexity $s.t.$ $M \ll N$.

⇄ Trade-off between speed and accuracy.

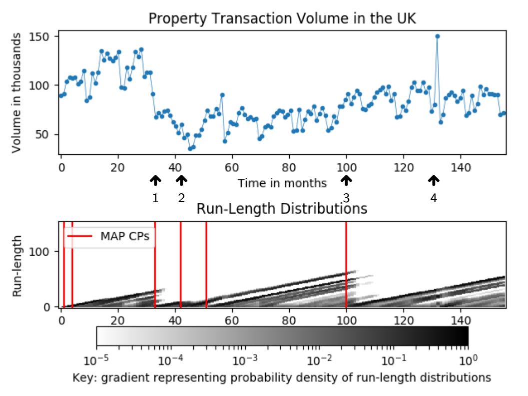


Applications I/III

Table 1: Effects of Sparse Approximations on VI.

Inducing Points	Time
M = X	115s
M = 10	90s
M = 5	88s

≈ 24% speed increase!



Data from April 2005 to February 2018

Key events:

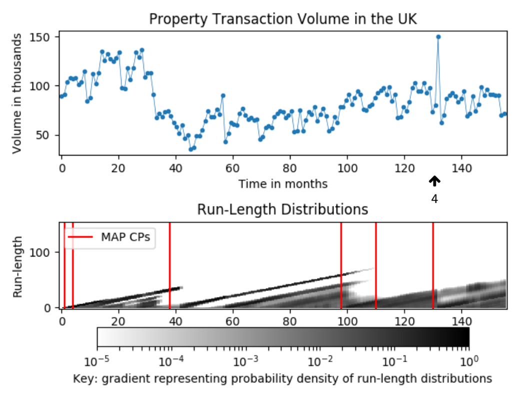
1. Global financial crisis
2. Second largest percentage decrease in house prices
3. Mortgage costs fall due to the BoE.
4. David Cameron announces Brexit referendum.

Applications I/III

Table 1: Effects of Sparse Approximations on VI.

Inducing Points	Time
$M = X$	115s
$M = 10$	90s
$M = 5$	88s

$\approx$ 24% speed increase!

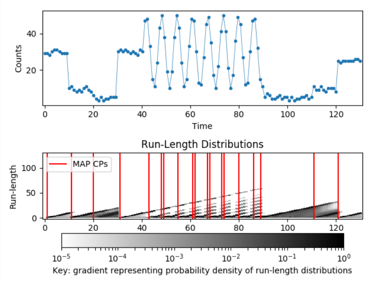


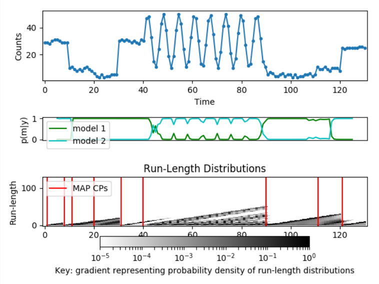
Data from April 2005 to February 2018

Key events:

1. Global financial crisis
2. Second largest percentage decrease in house prices
3. Mortgage costs fall due to the BoE.
4. David Cameron announces Brexit referendum.

Applications II/III





Parametric bottlenecks

✖ Models such as the Poisson Gamma don't work!

✖ No way to capture periodicity.

Model Selection

✔ Model 1: Poisson Gamma.

✔ Model 2: Log Gaussian Cox Process using a periodic kernel.

✔ Track model posterior $p(m|y)$

Multivariate: Intrinsic Coregionalization Model

Consider two outputs $f_1(x)$ and $f_2(x)$ with $x \in \mathbb{R}^p$.
Sample twice from a GP to get $u^1(x)$ and $u^2(x)$.
Group it into a vector-valued function.

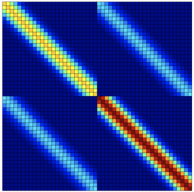
$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = \begin{bmatrix} a_1^1 u^1(x) + a_1^2 u^2(x) \\ a_2^1 u^1(x) + a_2^2 u^2(x) \end{bmatrix} = \begin{bmatrix} a^1 u^1(x) & a^2 u^2(x) \end{bmatrix}$$

Model dependencies by showing covariance $f(x)$ is

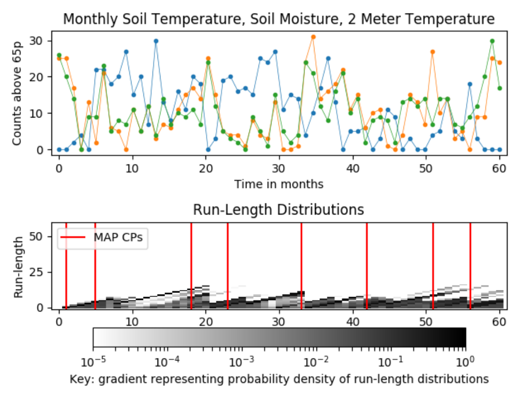
$$\begin{aligned} \operatorname{cov}(f(x), f(x')) &= \dots \\ &= [a^1(a^1)^T + a^2(a^2)^T] \; k(x, x') \\ &= B k(x, x') \end{aligned}$$

Take the Kronecker product of coregionalization matrix and kernel

$$\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, B \otimes K)$$

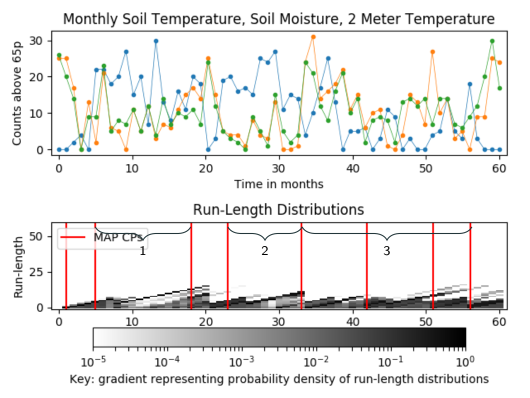


Cervest Bio-climatic Data



Data from April 2007 to August 2018.
Counts above the 65th percentile on a single spatial grid.
Using the Log Gaussian Cox Process model.

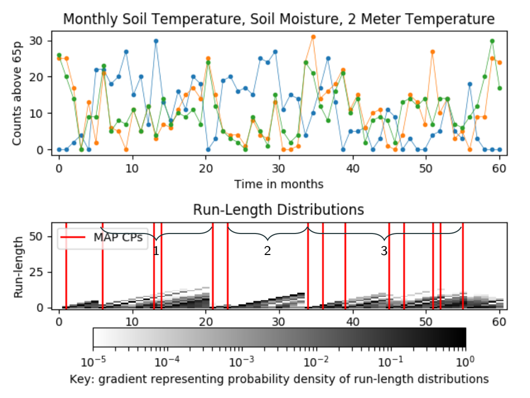
Cervest Bio-climatic Data



Data from April 2007 to August 2018.
Counts above the 65th percentile on a single spatial grid.
Using the Log Gaussian Cox Process model.

Applications III/III

Cervest Bio-climatic Data



Data from April 2007 to August 2018.
Counts above the 65th percentile on a single spatial grid.
Using the Multinomial Dirichlet model.

Summary

What have we achieved	Future Directions
Proposed the LGCP model. 1. Demonstrate flexible use-cases; why/when do you want to use this model? 2. Speed up via sparse approximations. 3. Model multivariate dependencies via ICM.	Still slow! 1. Current VI implementation naively rebuilds itself from scratch. 2. Kernel approximation methods, i.e. $\mathcal{O}(N)$ time, KISS-GP (Wilson & Nickisch, 2015).

References I

David Clifton.
Chi Square Group Meeting Slides, 2016.
http://www.robots.ox.ac.uk/~davidc/pubs/ht2016_kn.pdf.

P. Fearnhead and Z. Liu.
On-line inference for multiple changepoint problems.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(4):589–605.

J. Knoblauch and T. Damoulas.
Spatio-temporal Bayesian On-line Changepoint Detection with Model Selection.
arXiv e-prints, page arXiv:1805.05383, May 2018.

J. Moller, A. R. Syversveen, and R. P. Waagepetersen.
Log gaussian cox processes.
Scandinavian Journal of Statistics, 25(3):451–482.

R. Prescott Adams and D. J. C. MacKay.
Bayesian Online Changepoint Detection.
arXiv e-prints, page arXiv:0710.3742, Oct 2007.

C. Rasmussen and C. Williams.
Gaussian Processes for Machine Learning.
Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, Jan. 2006.

Appendix I: Sampling from a GP using different kernels.

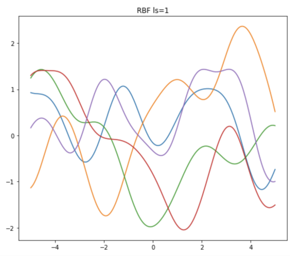
Many kernels:

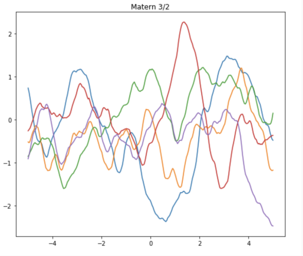
1. Radial Basis Function
2. Matern32
3. Matern52
4. Brownian
5. Bias
6. Linear
7. StdPeriodic
8. Exponential

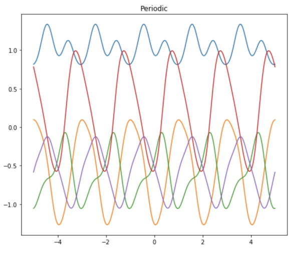
Can take sums or products since PSD.

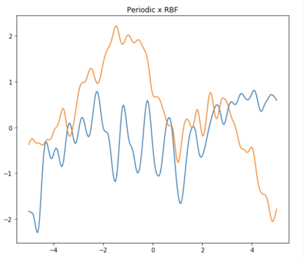
$k_1(x, x') + k_2(x, x')$

$k_1(x, x')k_2(x, x')$









Appendix II: Recursive MAP Segmentation

For simplicity, take

$$MAP_t = \arg\max_{r,m}\{p(r_t = r, m_t = m|y_{1:t})MAP_{t-r}\}$$

Consider for a single model $|\mathcal{M}| = 1$ only, then for $\forall t \in \mathbb{N}$

$$MAP_1 = MAP_0 \times p(r_1 = 1|y_1)$$

$$MAP_2 = \max\left\{\begin{matrix} MAP_0 \times p(r_2 = 2|y_{1:2}) \\ MAP_1 \times p(r_2 = 1|y_{1:2}) \end{matrix}\right\}$$

$$\ldots$$

$$MAP_t = \arg\max_{r}\{MAP_{t-r} \times p(r_t = r|y_{1:t})\}$$

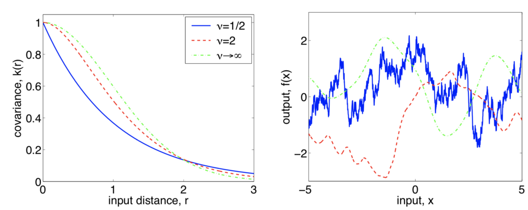
Hence for $|\mathcal{M}| \geq 2$, W.L.O.G.

$$MAP_1 = MAP_0 \times \arg\max_{m\in\mathcal{M}} p(r_1 = 1, m_t = m|y_1)$$

$$\ldots$$

$$MAP_t = \arg\max_{(r^*,m^*\in\mathcal{S})}\{MAP_{t-r} \times p(r_t = r, m_t = m|y_{1:t})\}$$

Appendix III: Matern covariance functions



Functions from Matern forms are floor of $v - 1$ times differentiable. For classes 3/2 (once differentiable) and 5/2 (twice differentiable). Smooth, infinitely differentiable if $v \to \infty$.

$$k_{v=3/2}(x, x') = (1 + \frac{\sqrt{3}|x - x'|}{\ell}) \exp(-\frac{\sqrt{3}|x - x'|}{\ell})$$

$$k_{v=5/2}(x, x') = (1 + \frac{\sqrt{5}|x - x'|}{\ell} + \frac{5r^2}{3\ell^2}) \exp(-\frac{\sqrt{5}|x - x'|}{\ell})$$

$$k_{v\to\infty}(x, x') = \exp(-\frac{|x - x'|^2}{2\ell^2})$$

Appendix IV: More Applications

Major events in 2018:

1. 4th June:
MSFT announces GitHub acquisition for $7.5 billion.

2. 10th July:
Surface Go is revealed to the public.

3. 10th October:
Big tech sell-off in Wall Street extends to the lowest point.

4. 18th December:
Bleakest Christmas in Wall-Street since 1930s.