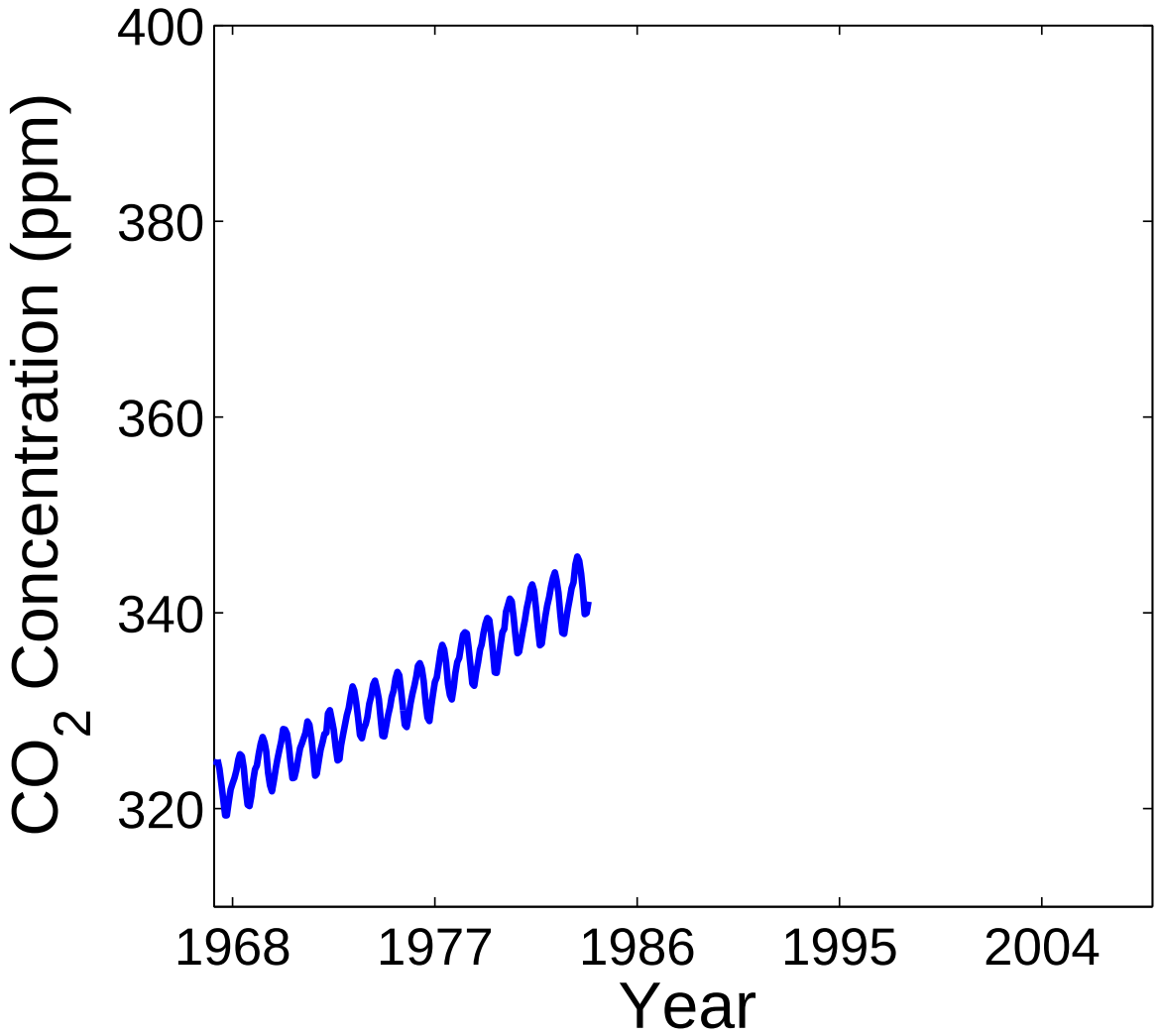
Kernels for Automatic Pattern Discovery and Extrapolation

Andrew Gordon Wilson

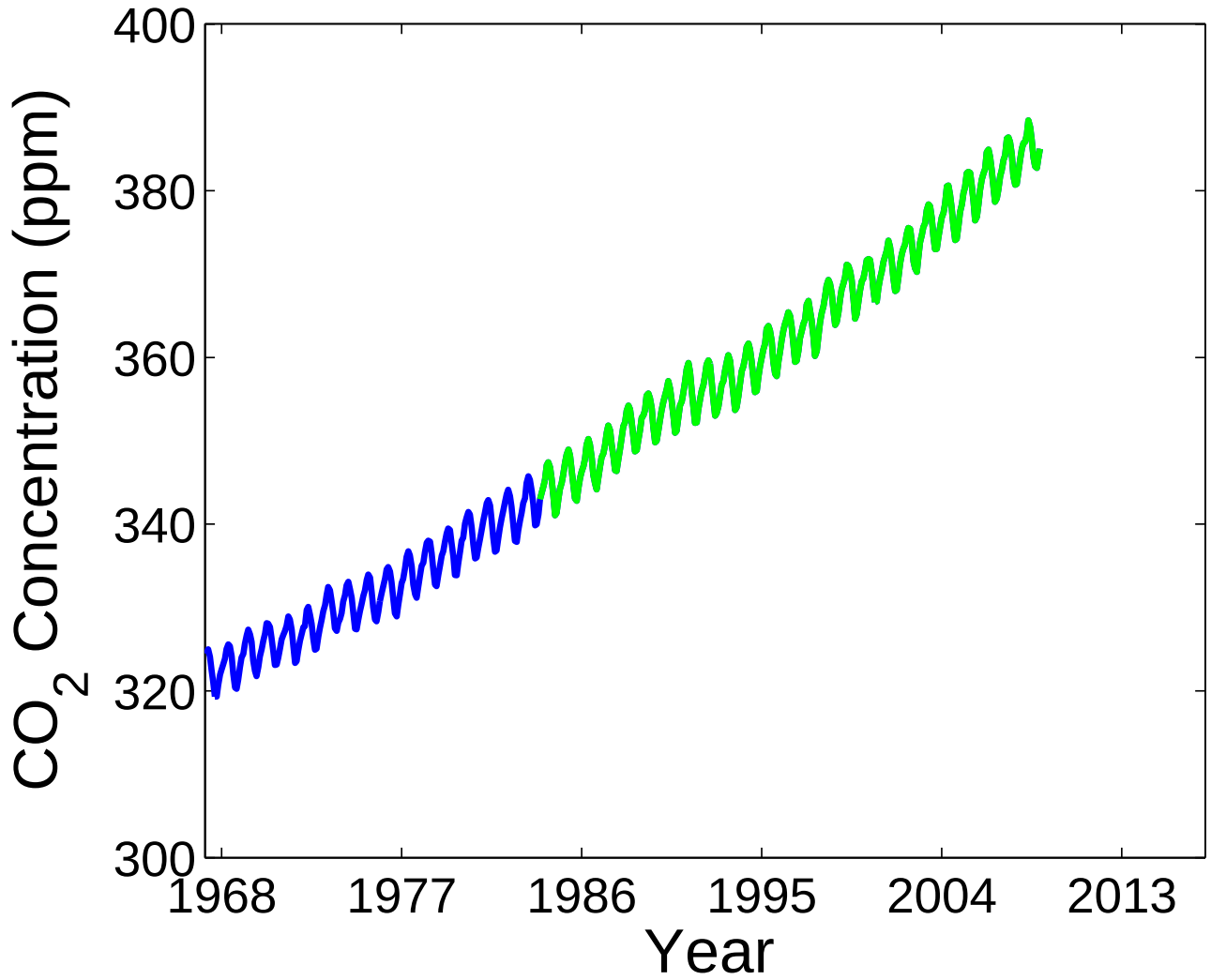
agw38@cam.ac.uk
mlg.eng.cam.ac.uk/andrew
University of Cambridge

Joint work with Ryan Adams (Harvard)

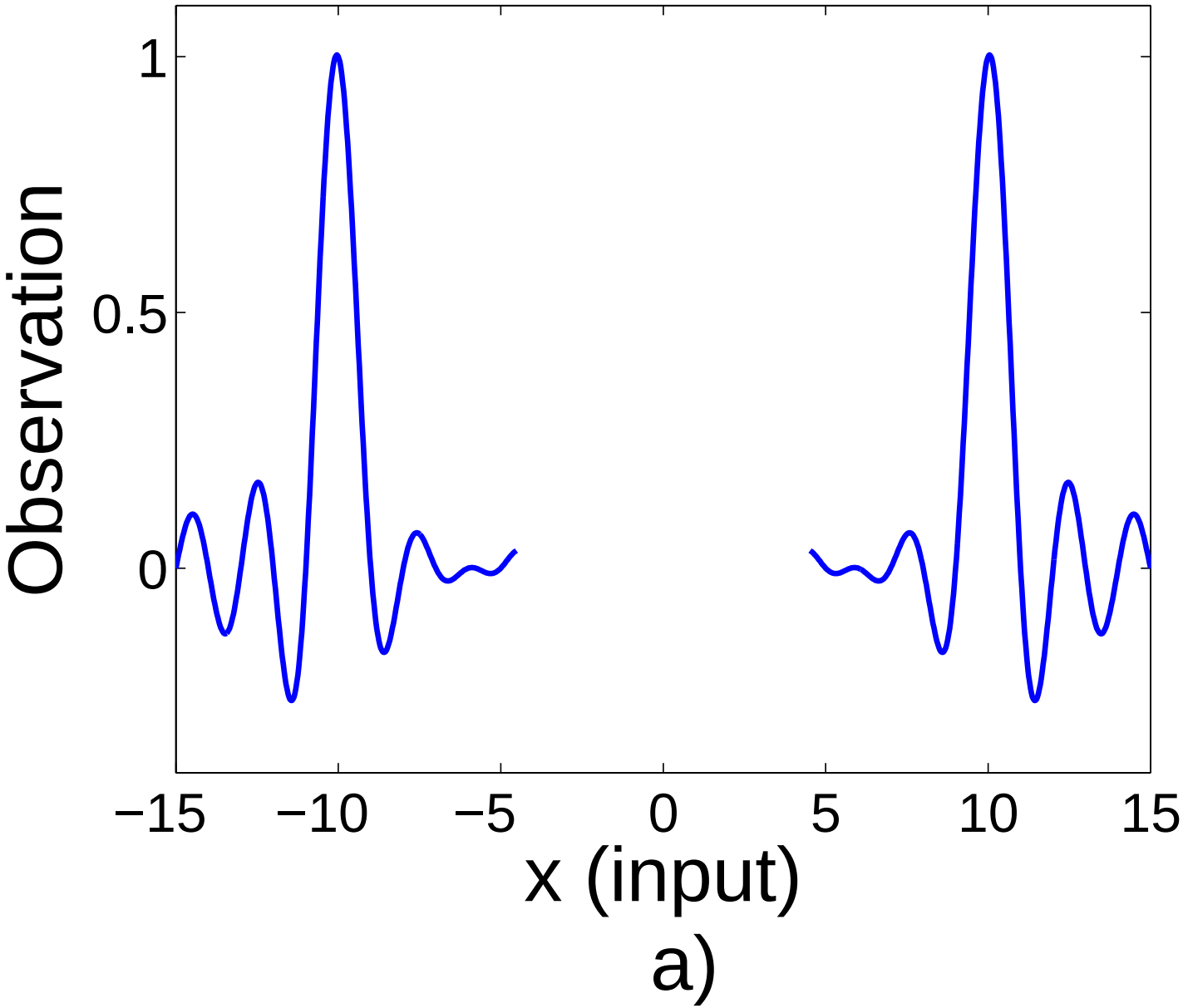
Pattern Recognition



Pattern Recognition



Pattern Recognition



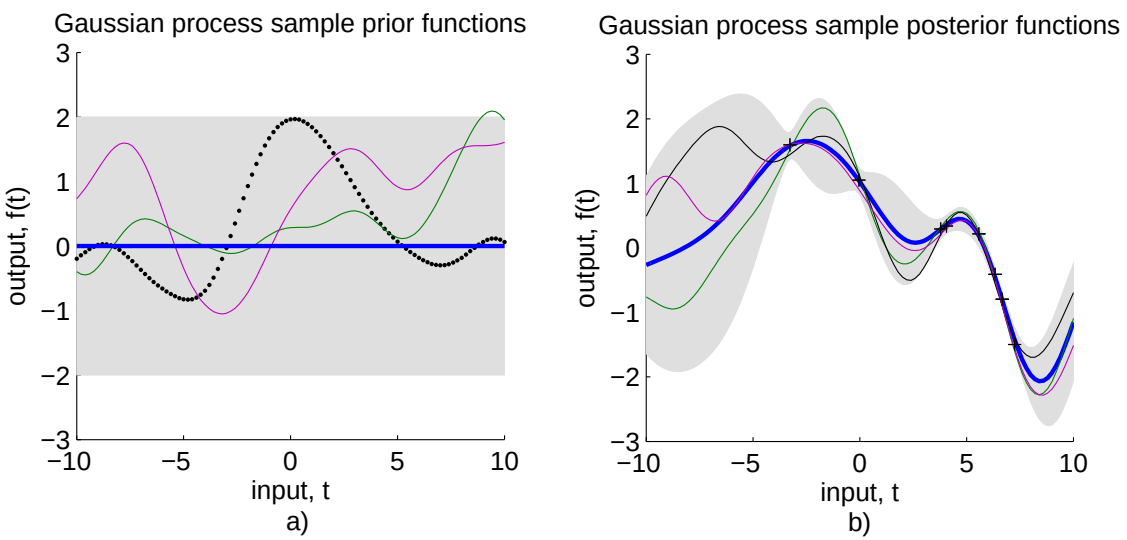
Definition

A Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Nonparametric Regression Model

- Prior: $f(x) \sim \mathcal{GP}(m(x), k(x,x'))$, meaning $(f(x_1), \ldots, f(x_N)) \sim \mathcal{N}(\boldsymbol{\mu}, K)$, with $\boldsymbol{\mu}_i = m(x_i)$ and $K_{ij} = \mathrm{cov}(f(x_i), f(x_j)) = k(x_i, x_j)$.

$$\overbrace{p(f(x)|\mathcal{D})}^{\text{GP posterior}} \propto \overbrace{p(\mathcal{D}|f(x))}^{\text{Likelihood}} \overbrace{p(f(x))}^{\text{GP prior}}$$



Gaussian Process Covariance Kernels

Let $\tau = x - x'$:

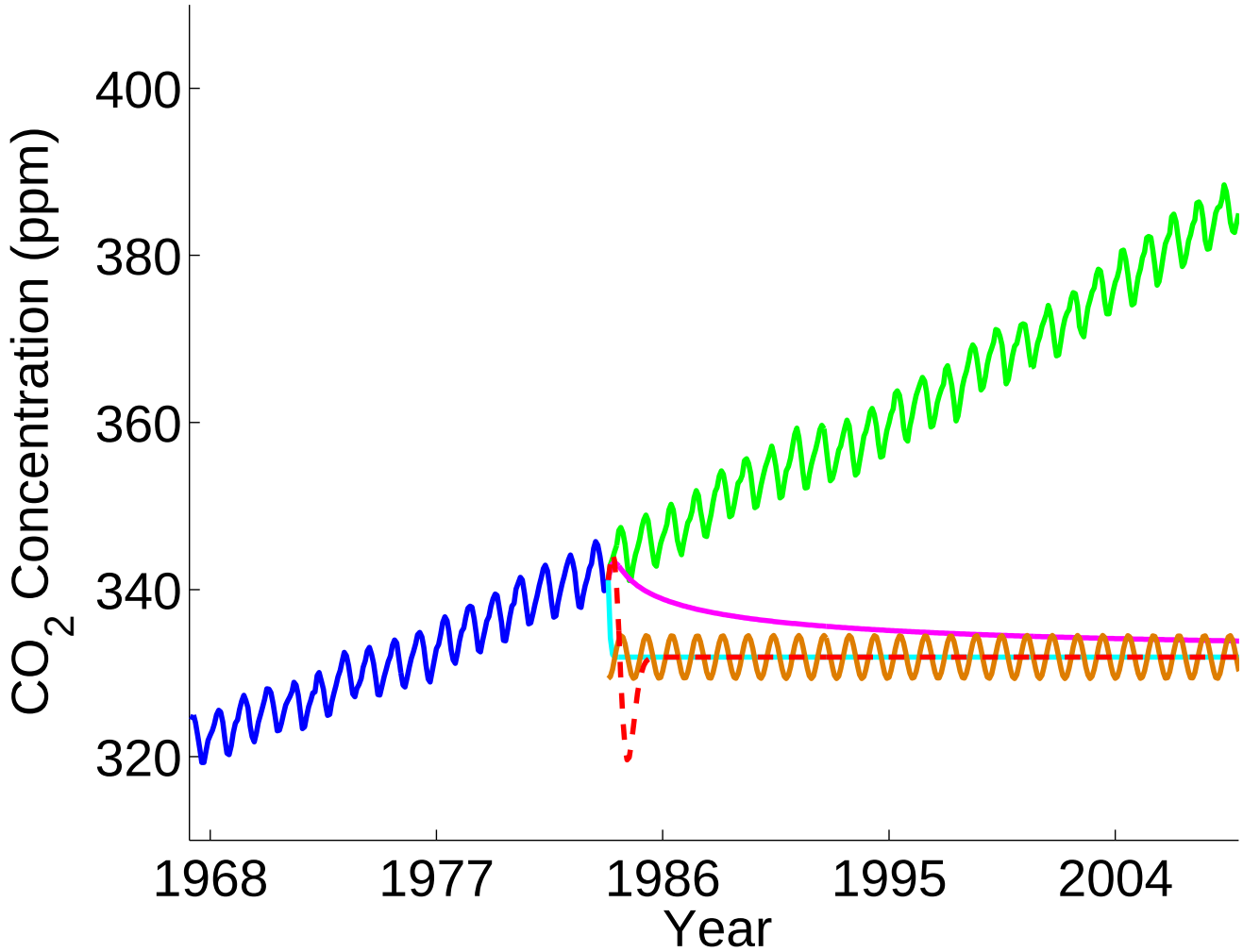
$$k_{\text{SE}}(\tau) = \exp(-0.5\tau^2/\ell^2) \tag{1}$$

$$k_{\text{MA}}(\tau) = a(1 + \frac{\sqrt{3}\tau}{\ell}) \exp(-\frac{\sqrt{3}\tau}{\ell}) \tag{2}$$

$$k_{\text{RQ}}(\tau) = (1 + \frac{\tau^2}{2\,\alpha\,\ell^2})^{-\alpha} \tag{3}$$

$$k_{\text{PE}}(\tau) = \exp(-2\sin^2(\pi\,\tau\,\omega)/\ell^2) \tag{4}$$

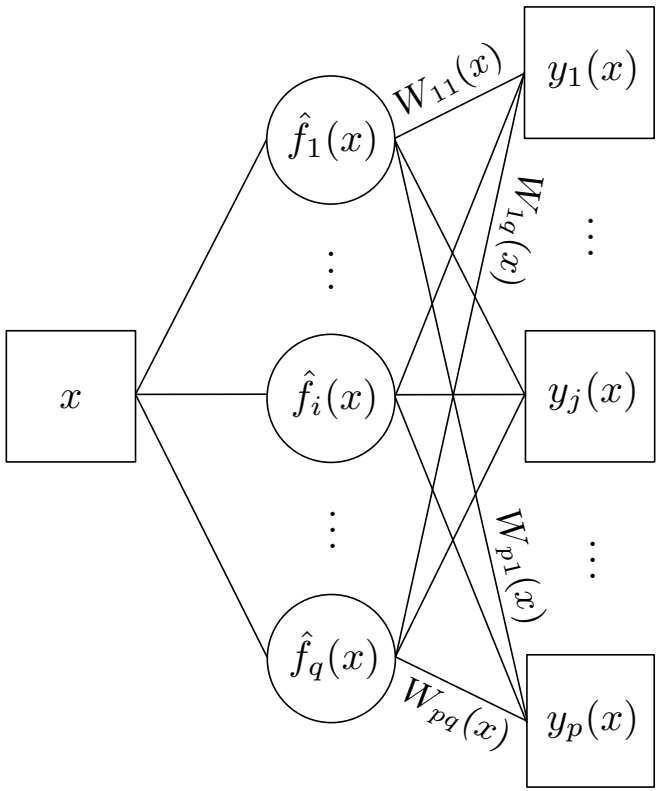
CO_2 Extrapolation with Standard Kernels



"How can Gaussian processes possibly replace neural networks? Did we throw the baby out with the bathwater?"

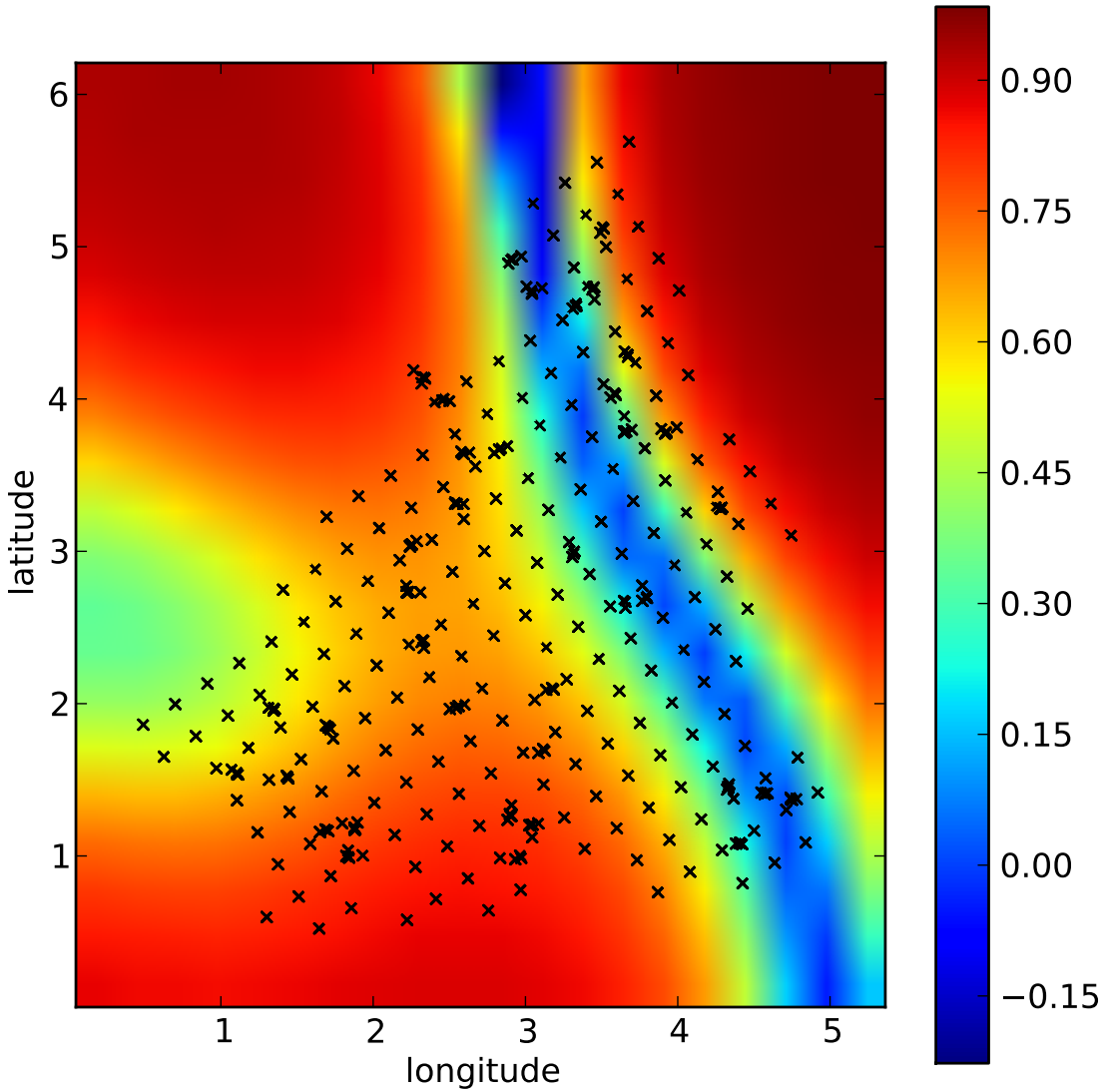
David MacKay, 1998.

More Expressive Covariance Functions



Gaussian Process Regression Networks. Wilson et. al, ICML 2012.

Gaussian Process Regression Network



Expressive Covariance Functions

- GPs in Bayesian neural network like architectures. (Salakhutdinov and Hinton, 2008; Wilson et. al, 2012; Damianou and Lawrence, 2012).
  Task specific, difficult inference, no closed form kernels.
- Compositions of kernels. (Archambeau and Bach, 2011; Durrande et. al, 2011; Rasmussen and Williams, 2006).
  In the general case, difficult to interpret, difficult inference, struggle with over-fitting.

Can learn almost nothing about the covariance function of a stochastic process from a single realization, if we assume that the covariance function could be *any* positive definite function. Most commonly one assumes a restriction to *stationary* kernels, meaning that covariances are invariant to translations in the input space.

Bochner's Theorem

Theorem

(Bochner) A complex-valued function k on $\mathbb{R}^P$ is the covariance function of a weakly stationary mean square continuous complex-valued random process on $\mathbb{R}^P$ if and only if it can be represented as

$$k(\tau) = \int_{\mathbb{R}^P} e^{2\pi i s^T \tau} \psi(\mathrm{d}s) \,, \tag{5}$$

where ψ is a positive finite measure.

If ψ has a density $S(s)$, then S is called the *spectral density* or *power spectrum* of k, and k and S are Fourier duals:

$$k(\tau) = \int S(s) e^{2\pi i s^{\mathrm{T}} \tau} ds \,, \tag{6}$$

$$S(s) = \int k(\tau) e^{-2\pi i s^{\mathrm{T}} \tau} d\tau \,. \tag{7}$$

Idea

k and S are Fourier duals:

$$k(\tau) = \int S(s) e^{2\pi i s^{\mathrm{T}} \tau} ds \,, \tag{8}$$

$$S(s) = \int k(\tau) e^{-2\pi i s^{\mathrm{T}} \tau} d\tau \,. \tag{9}$$

- If we can approximate $S(s)$ to arbitrary accuracy, then we can approximate any stationary kernel to arbitrary accuracy.
- We can model $S(s)$ to arbitrary accuracy, since scale-location mixtures of Gaussians can approximate any distribution to arbitrary accuracy.
- A scale-location mixture of Gaussians can flexibly model many distributions, and thus many covariance kernels, even with a small number of components.

Kernels for Pattern Discovery

Let $\tau = x - x' \in \mathbb{R}^P$. From Bochner's Theorem,

$$k(\tau) = \int_{\mathbb{R}^P} S(s) e^{2\pi i s^{\mathrm{T}} \tau} ds \tag{10}$$

For simplicity, assume $\tau \in \mathbb{R}^1$ and let

$$S(s) = [\mathcal{N}(s; \mu, \sigma^2) + \mathcal{N}(-s; \mu, \sigma^2)]/2\,. \tag{11}$$

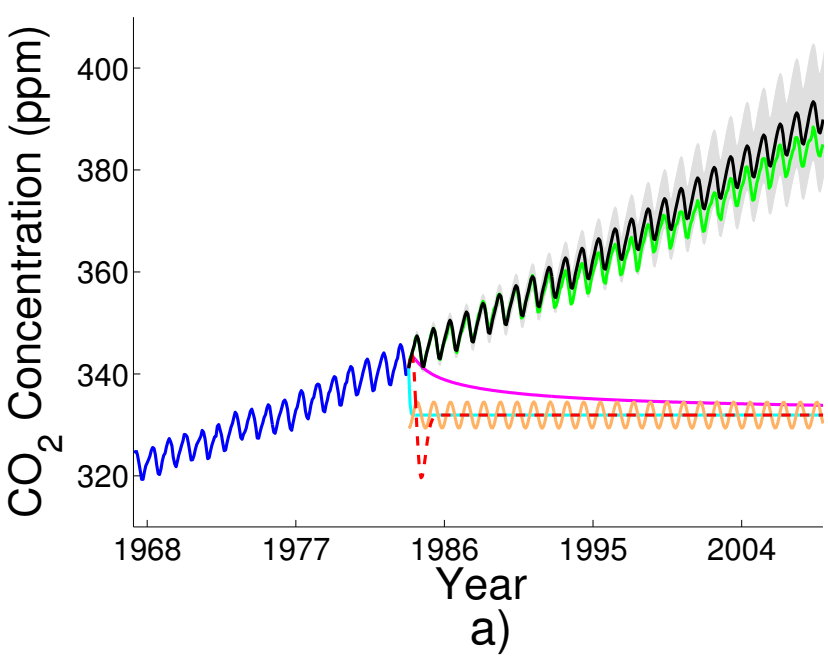
Then

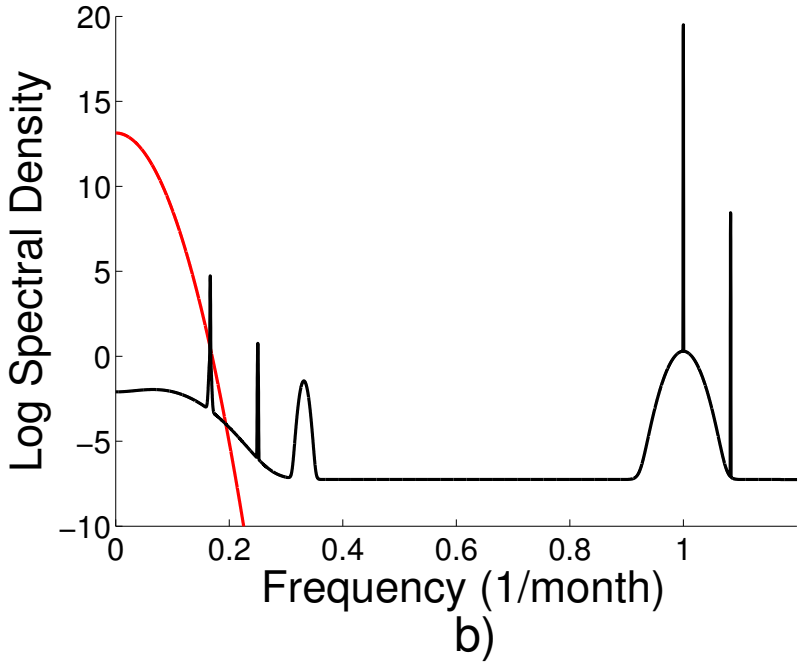
$$k(\tau) = \exp\{-2\pi^2 \tau^2 \sigma^2\} \cos(2\pi\tau\mu)\,. \tag{12}$$

More generally, if $S(s)$ is a symmetrized mixture of diagonal covariance Gaussians on $\mathbb{R}^p$, with covariance matrix $\mathbf{M}_q = \mathrm{diag}(v_q^{(1)}, \ldots, v_q^{(P)})$, then

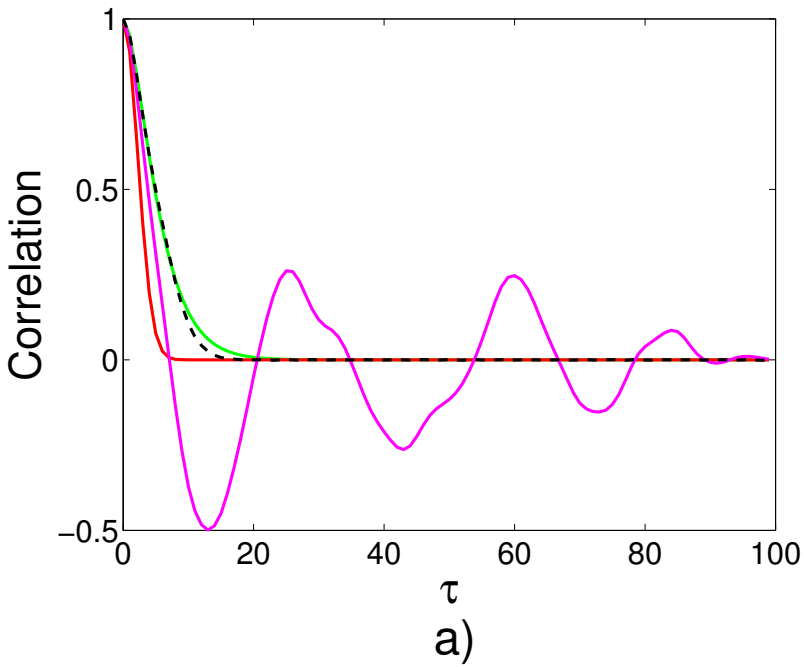
$$k(\tau) = \sum_{q=1}^{Q} w_q \prod_{p=1}^{P} \exp\{-2\pi^2 \tau_p^2 v_q^{(p)}\} \cos(2\pi\tau_p \mu_q^{(p)}). \tag{13}$$

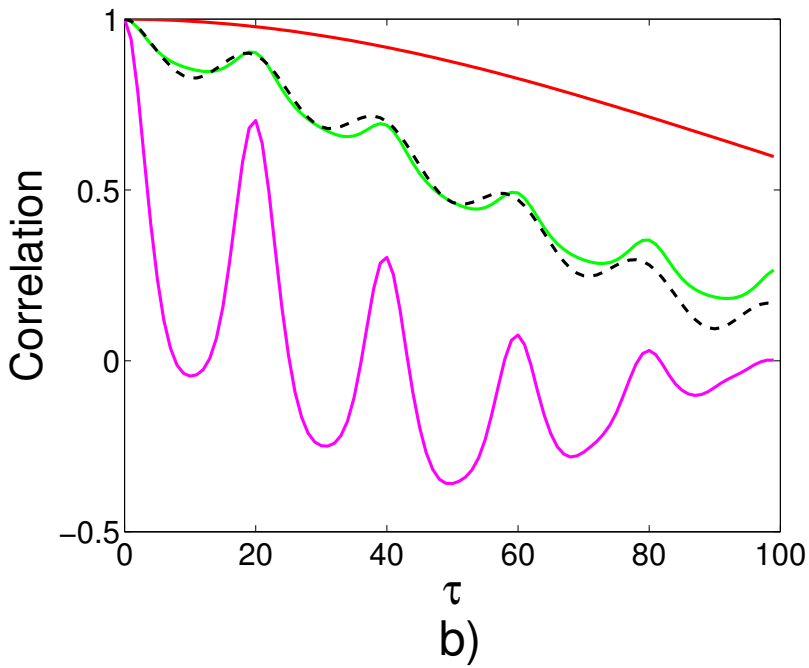
Results, CO_2


CO2 Concentration (ppm)
400
380
360
340
320
1968
1977
1986
1995
2004
Year
a)

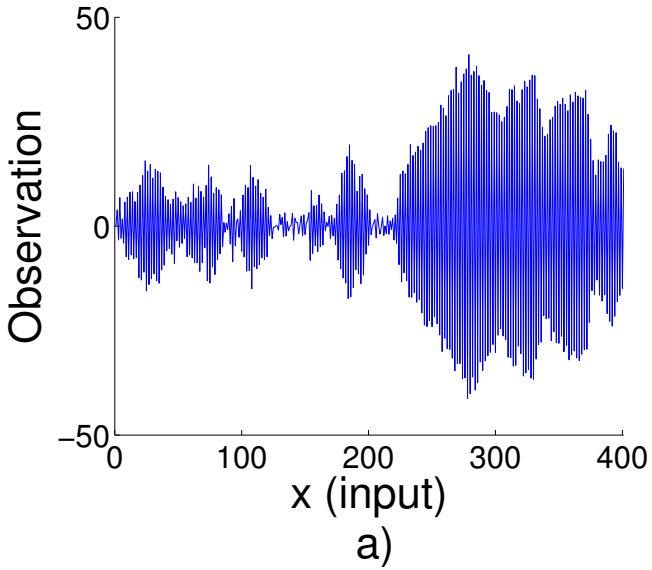

Log Spectral Density
20
15
10
5
0
−5
−10
0
0.2
0.4
0.6
0.8
1
Frequency (1/month)
b)

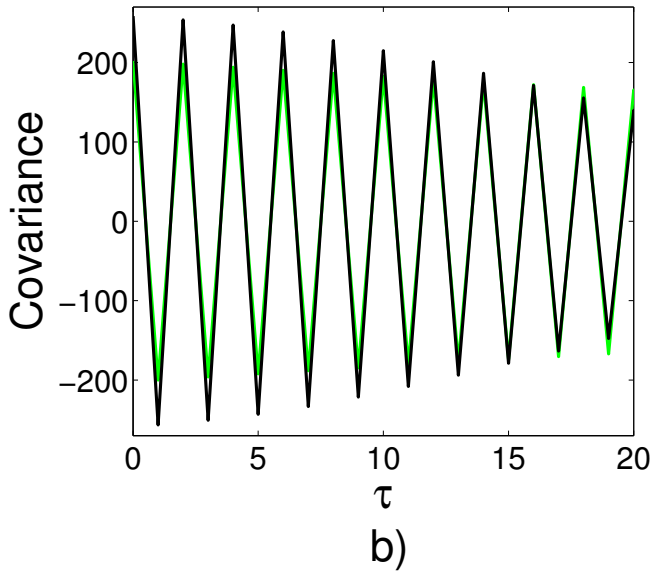
Results, Reconstructing Standard Covariances

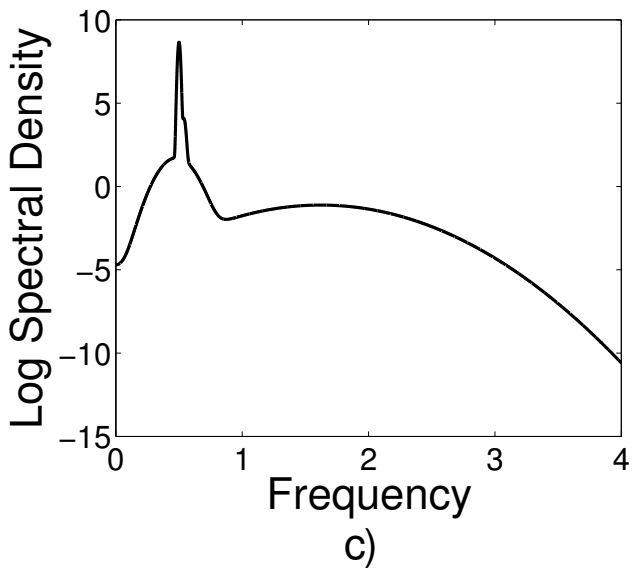




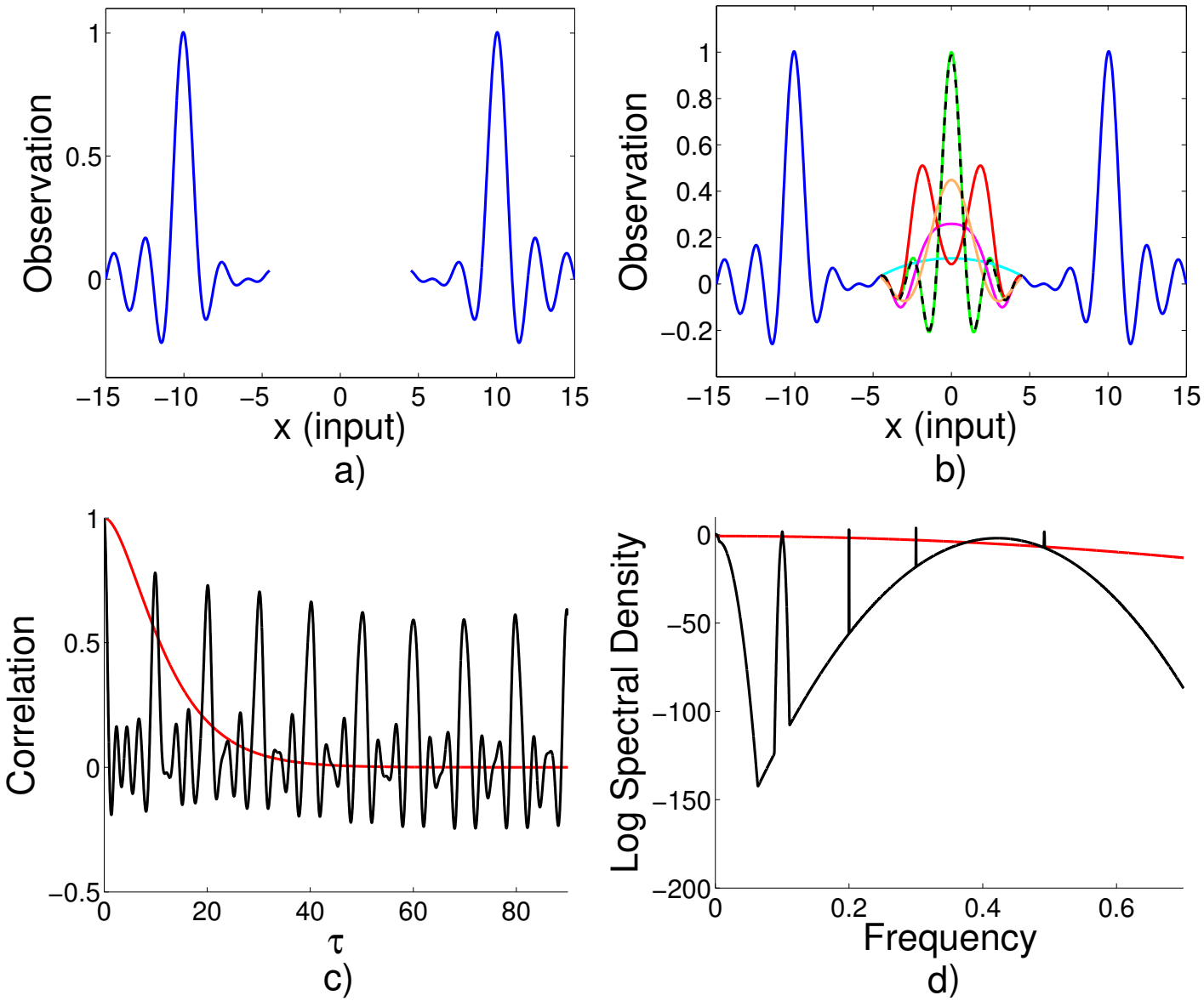
Results, Negative Covariances



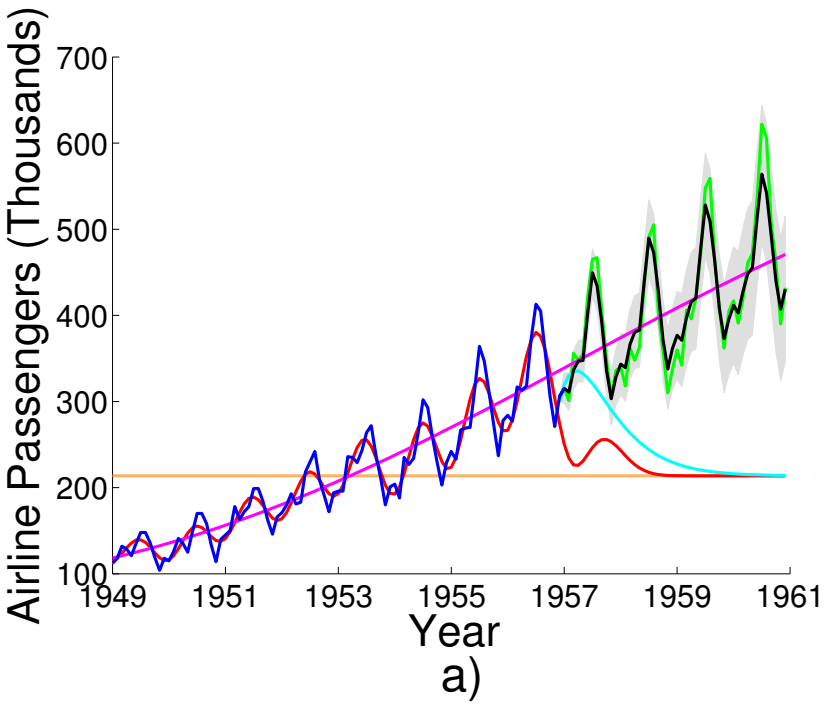


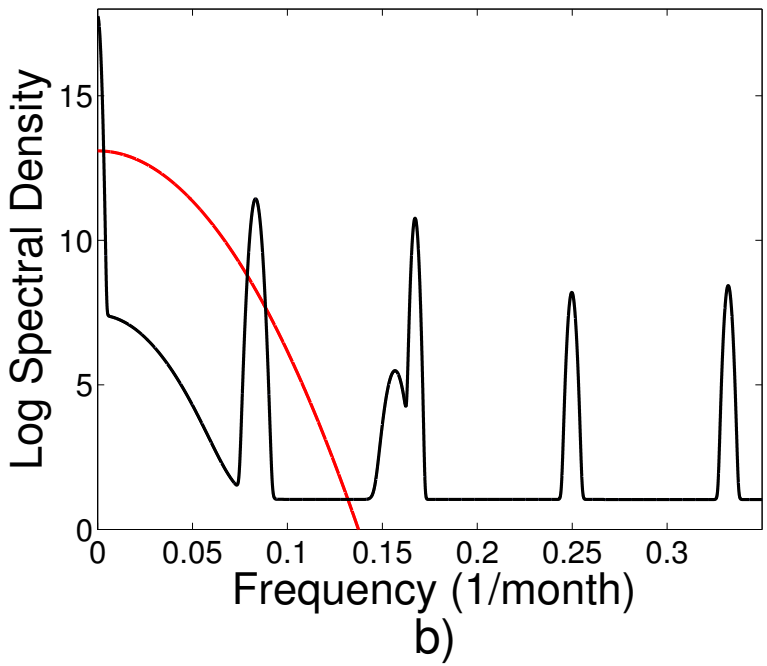


Results, Sinc Pattern



Results, Airline Passengers





Gaussian Process Kernels for Automatic Pattern Discovery

- Gaussian processes are rich distributions over functions, which provide a Bayesian nonparametric approach to smoothing and interpolation.
- We introduce new, *simple*, closed form kernels, which can be used with Gaussian processes to enable automatic pattern discovery and extrapolation. Code available at: `http://mlg.eng.cam.ac.uk/andrew`
- These kernels form a basis for all stationary covariance functions.
- These kernels can be used with non-Bayesian methods.
- In the future, it would be interesting to reverse engineer covariance functions that are induced by expressive architectures, like deep neural networks, to develop powerful interpretable models with simple inference procedures and closed form kernels.

Results

Table: We compare the test performance of the proposed spectral mixture (SM) kernel with squared exponential (SE), Matérn (MA), rational quadratic (RQ), and periodic (PE) kernels. The SM kernel consistently has the lowest mean squared error (MSE) and highest log likelihood ($\mathcal{L}$).

	SM	SE	MA	RQ	PE
CO_2					
MSE	**9.5**	1200	1200	980	1200
$\mathcal{L}$	**170**	−320	−240	−100	−1800
NEG COV					
MSE	**62**	210	210	210	210
$\mathcal{L}$	**−25**	−70	−70	−70	−70
SINC					
MSE	**0.000045**	0.16	0.10	0.11	0.05
$\mathcal{L}$	**3900**	2000	1600	2000	600
AIRLINE					
MSE	**460**	43000	37000	4200	46000
$\mathcal{L}$	**−190**	−260	−240	−280	−370