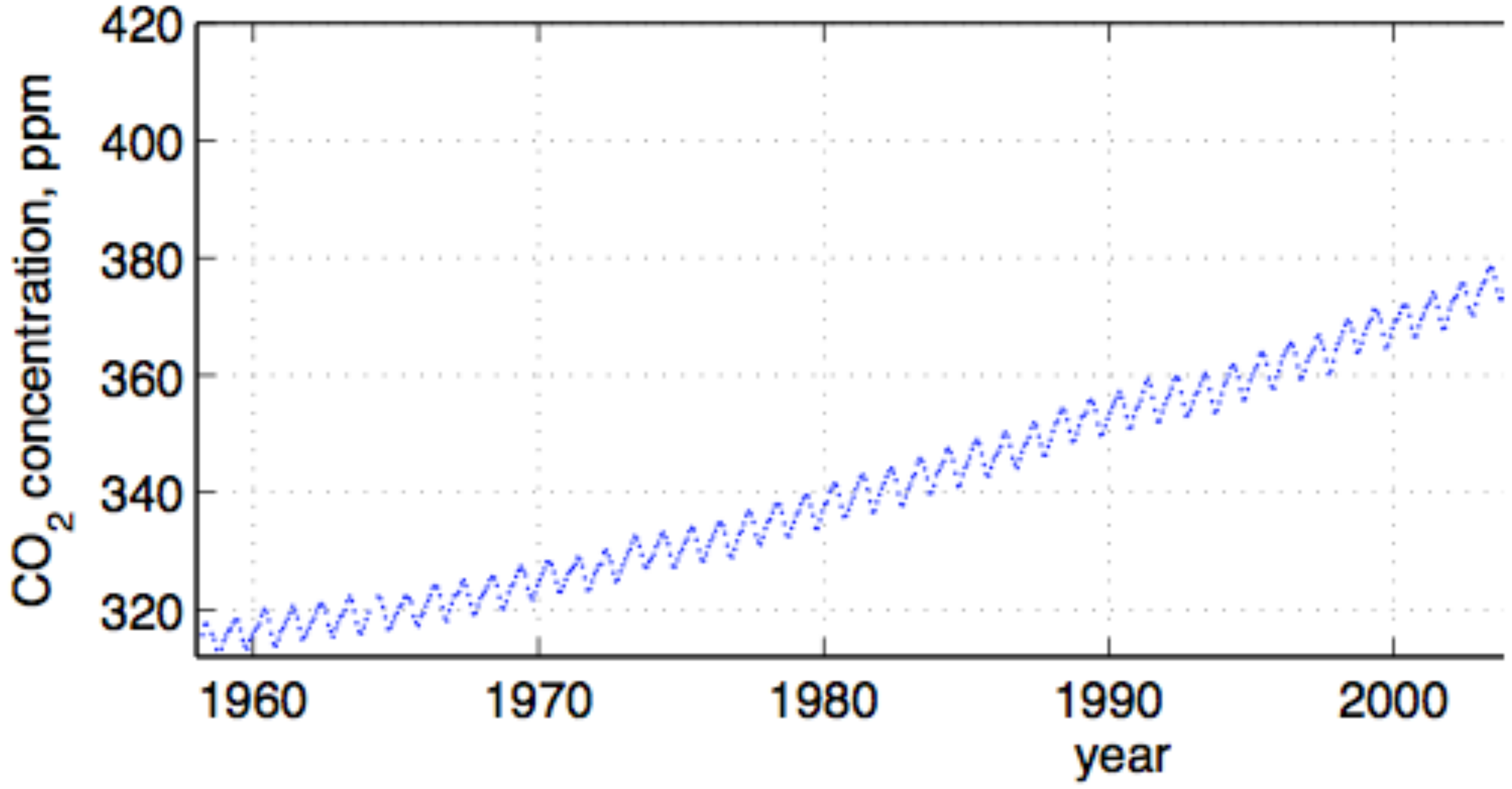
Statistical Machine Learning

Luca Bortolussi

University of Trieste

Manua Loa CO2 atmospheric data



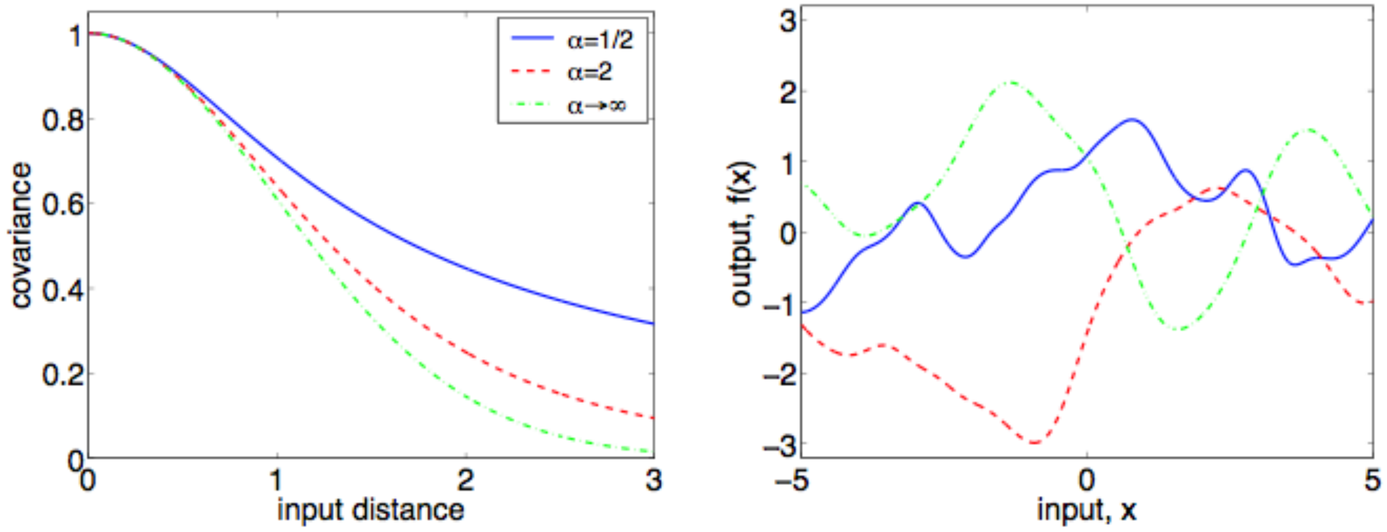
CO2 Data

Measured over 40 years. We observe a seasonal periodic behaviour and an increasing trend.

We want to construct a GP model of it.

The goal is to build a good kernel.

Rational Quadratic Kernel



(a) (b)

Figure 4.3: Panel (a) covariance functions, and (b) random functions drawn from Gaussian processes with rational quadratic covariance functions, eq. (4.20), for different values of α with $\ell = 1$. The sample functions on the right were obtained using a discretization of the x-axis of 2000 equally-spaced points.

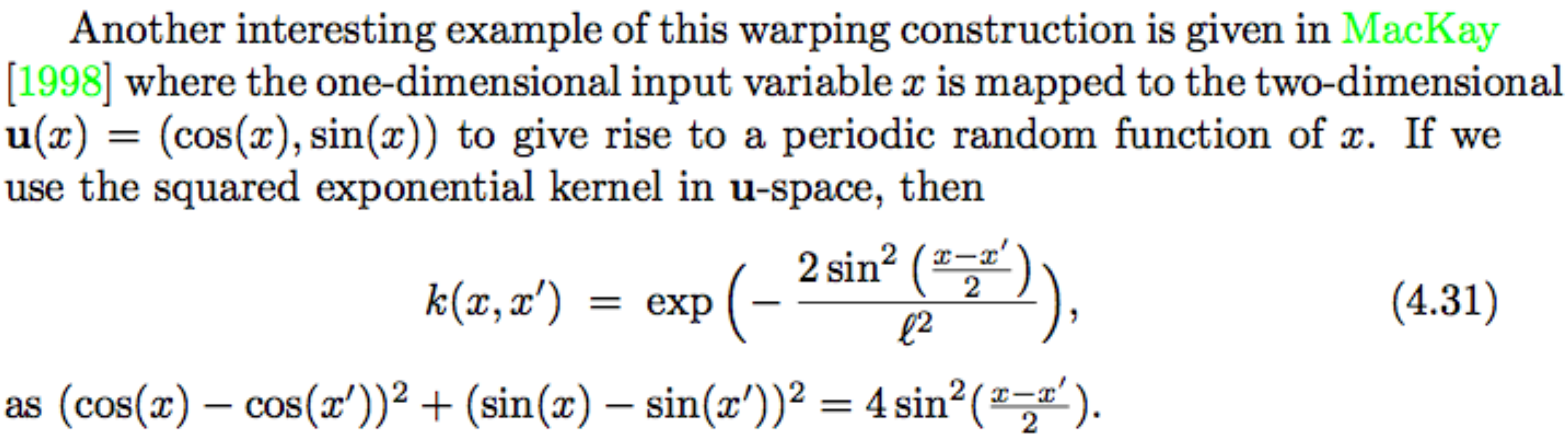
$$k_{\mathrm{RQ}}(r) = \left(1 + \frac{r^2}{2\alpha\ell^2}\right)^{-\alpha}$$

with α, $\ell > 0$ can be seen as a *scale mixture* (an infinite sum) of squared exponential (SE) covariance functions with different characteristic length-scales (sums of covariance functions are also a valid covariance, see section 4.2.4). Parameterizing now in terms of inverse squared length scales, $\tau = \ell^{-2}$, and putting a gamma distribution on $p(\tau|\alpha, \beta) \propto \tau^{\alpha-1}\exp(-\alpha\tau/\beta)$,[5] we can add up the contributions through the following integral

$$\begin{aligned} k_{\mathrm{RQ}}(r) &= \int p(\tau|\alpha,\beta) k_{\mathrm{SE}}(r|\tau)\, d\tau \\ &\propto \int \tau^{\alpha-1} \exp\left(-\frac{\alpha\tau}{\beta}\right) \exp\left(-\frac{\tau r^2}{2}\right) d\tau \propto \left(1 + \frac{r^2}{2\alpha\ell^2}\right)^{-\alpha}, \end{aligned} \tag{4.20}$$

where we have set $\beta^{-1} = \ell^2$. The rational quadratic is also discussed by Matérn

Periodic Kernel

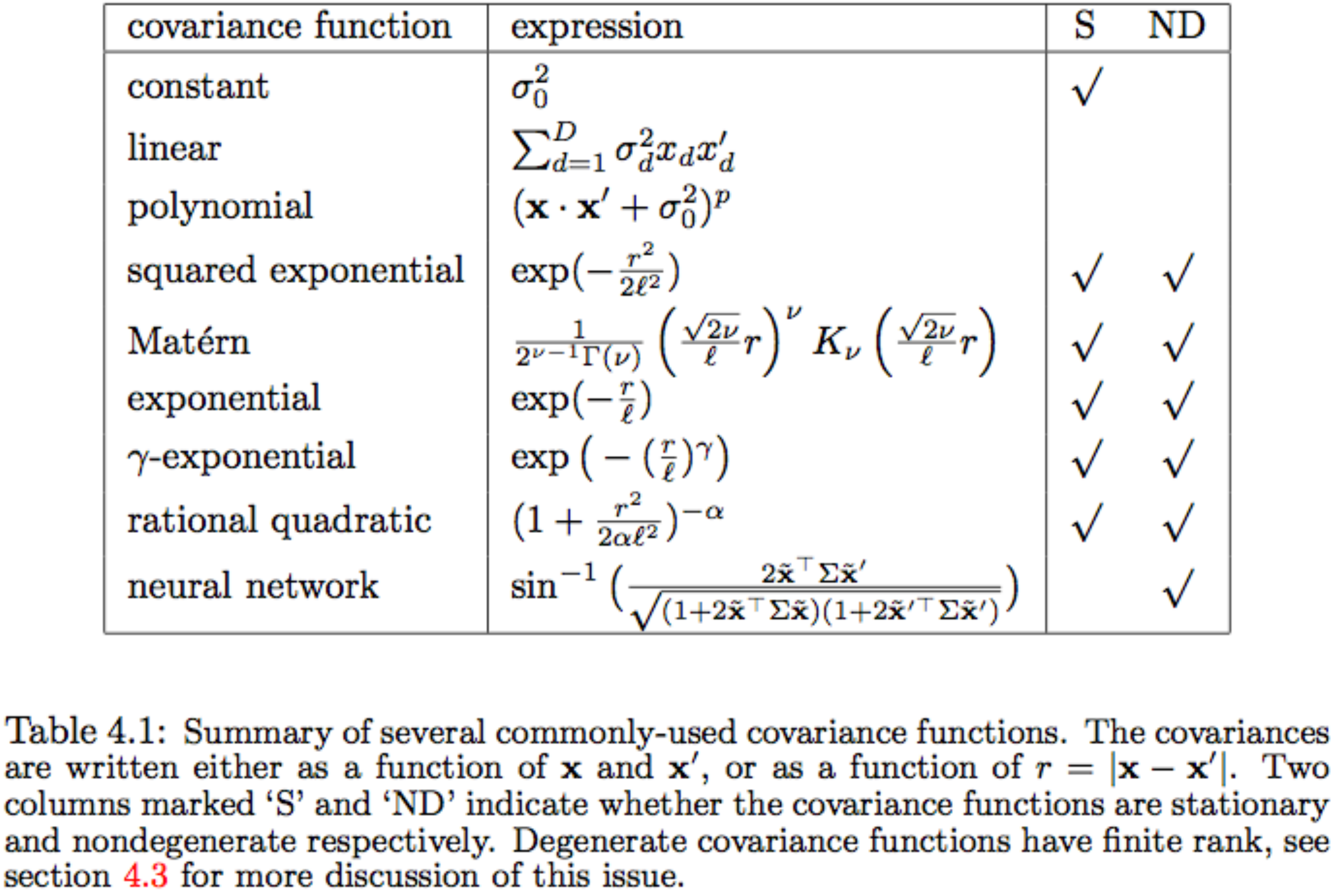


Another interesting example of this warping construction is given in MacKay [1998] where the one-dimensional input variable x is mapped to the two-dimensional $\mathbf{u}(x) = (\cos(x), \sin(x))$ to give rise to a periodic random function of x. If we use the squared exponential kernel in $\mathbf{u}$-space, then

$$k(x, x') = \exp\left(-\frac{2\sin^2\left(\frac{x-x'}{2}\right)}{\ell^2}\right), \qquad (4.31)$$

as $(\cos(x) - \cos(x'))^2 + (\sin(x) - \sin(x'))^2 = 4\sin^2(\frac{x-x'}{2})$.

Summary of Kernels



covariance function	expression	S	ND
constant	σ_0^2	✓	
linear	$\sum_{d=1}^{D} \sigma_d^2 x_d x'_d$		
polynomial	$(\mathbf{x} \cdot \mathbf{x}' + \sigma_0^2)^p$		
squared exponential	$\exp(-\frac{r^2}{2\ell^2})$	✓	✓
Matérn	$\frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} r\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\ell} r\right)$	✓	✓
exponential	$\exp(-\frac{r}{\ell})$	✓	✓
γ-exponential	$\exp\left(-\left(\frac{r}{\ell}\right)^{\gamma}\right)$	✓	✓
rational quadratic	$(1 + \frac{r^2}{2\alpha\ell^2})^{-\alpha}$	✓	✓
neural network	$\sin^{-1}\left(\frac{2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}}'}{\sqrt{(1+2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}})(1+2\tilde{\mathbf{x}}'^\top \Sigma \tilde{\mathbf{x}}')}}\right)$		✓

Table 4.1: Summary of several commonly-used covariance functions. The covariances are written either as a function of $\mathbf{x}$ and $\mathbf{x}'$, or as a function of $r = |\mathbf{x} - \mathbf{x}'|$. Two columns marked 'S' and 'ND' indicate whether the covariance functions are stationary and nondegenerate respectively. Degenerate covariance functions have finite rank, see section 4.3 for more discussion of this issue.

Manua Loa CO2 atmospheric data

Idea: combination of simple kernels

$$k_1(x,x') = \theta_1^2 \exp\left(-\frac{(x-x')^2}{2\theta_2^2}\right). \qquad (5.15)$$

smooth trend

$$k_2(x,x') = \theta_3^2 \exp\left(-\frac{(x-x')^2}{2\theta_4^2} - \frac{2\sin^2(\pi(x-x'))}{\theta_5^2}\right), \qquad (5.16)$$

periodic kernel with decay of periodicity

$$k_3(x,x') = \theta_6^2\left(1+\frac{(x-x')^2}{2\theta_8\theta_7^2}\right)^{-\theta_8},$$

medium term irregularities

$$k_4(x_p,x_q) = \theta_9^2 \exp\left(-\frac{(x_p-x_q)^2}{2\theta_{10}^2}\right) + \theta_{11}^2\delta_{pq}, \qquad (5.18)$$

noise terms

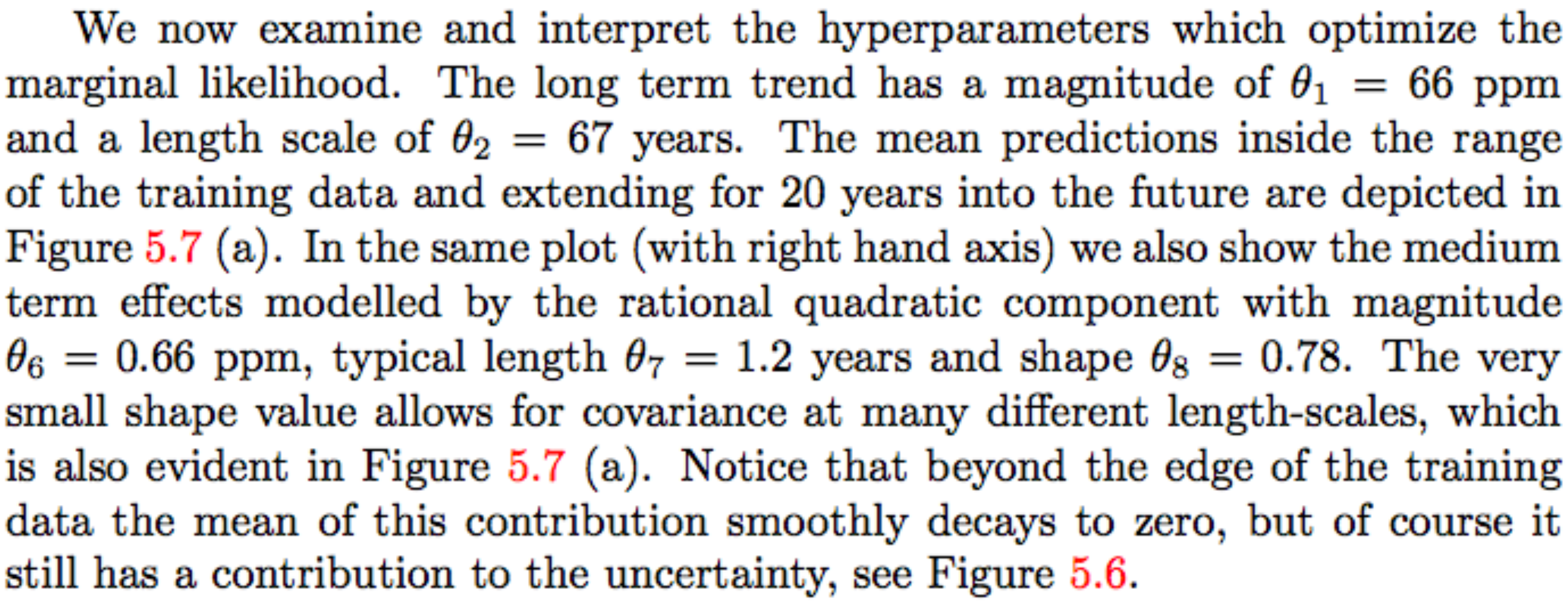
$$k(x,x') = k_1(x,x') + k_2(x,x') + k_3(x,x') + k_4(x,x'), \qquad (5.19)$$

Final kernel

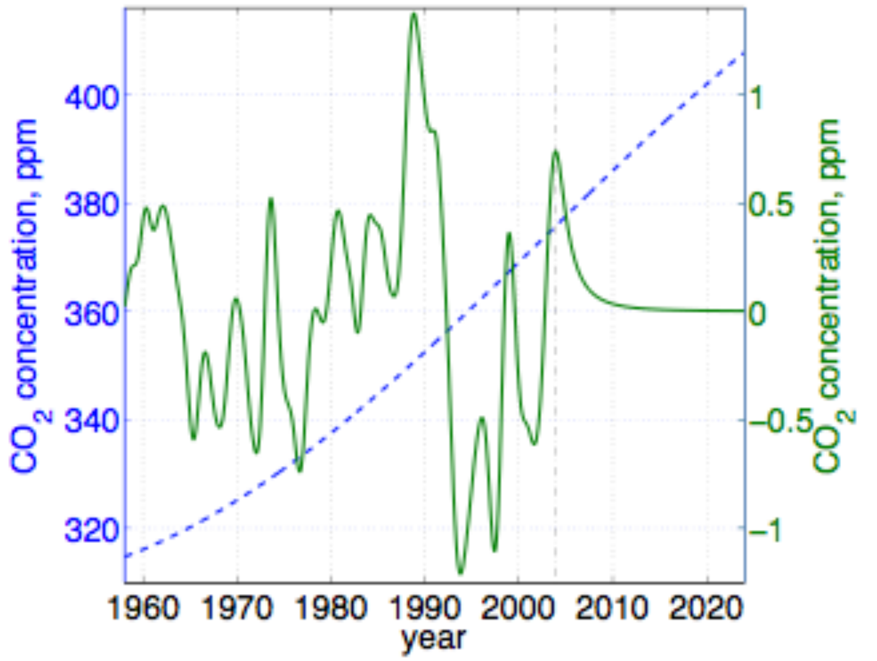
Manua Loa CO2 atmospheric data

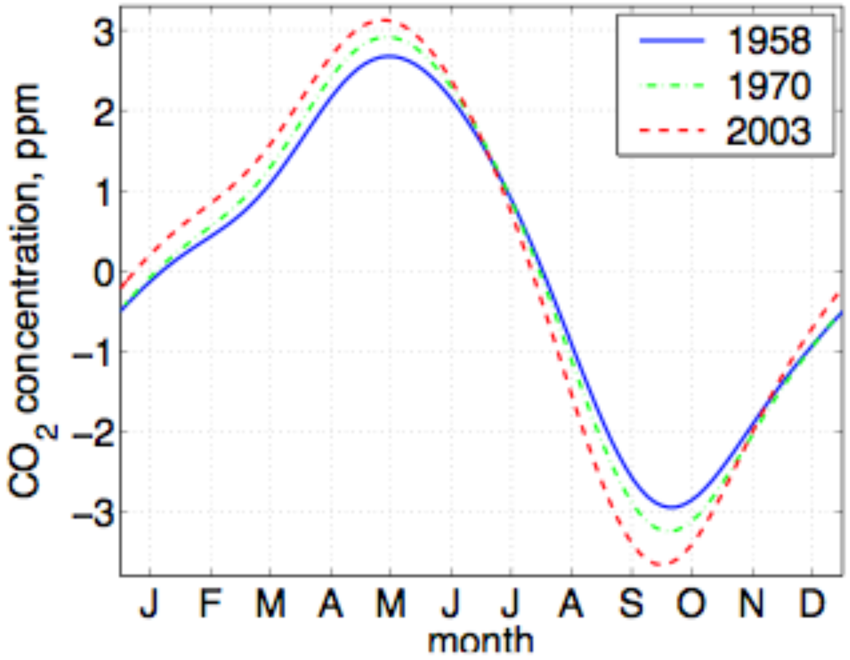
11 hyperparameters, which are optimised.

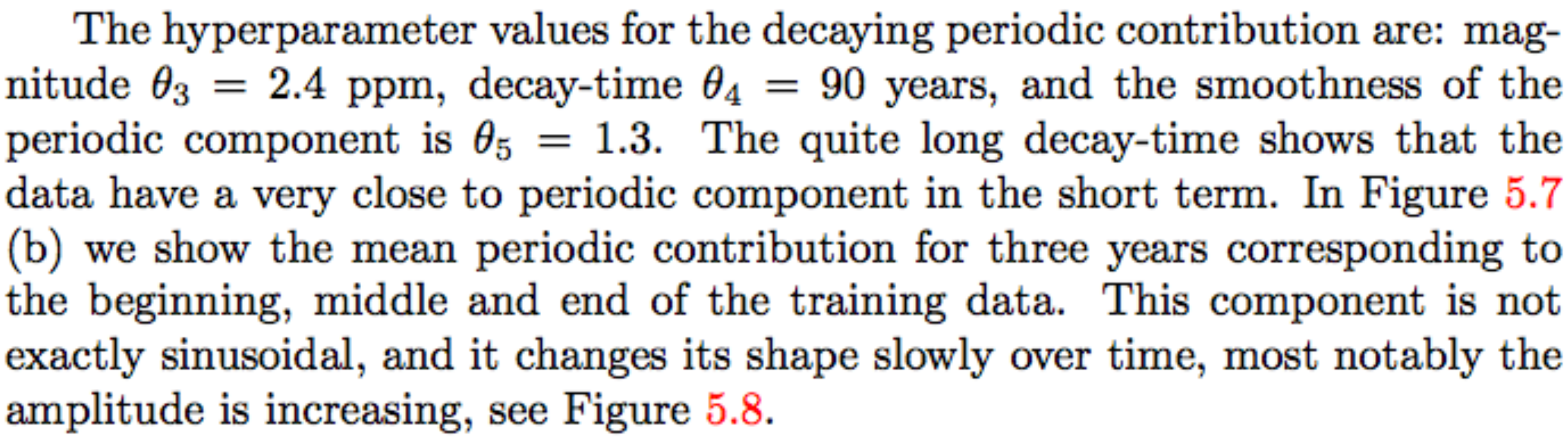
Optimal marginal likelihood: $\log p(\mathbf{y}|X, \boldsymbol{\theta}) = -108.5.$



We now examine and interpret the hyperparameters which optimize the marginal likelihood. The long term trend has a magnitude of $\theta_1 = 66$ ppm and a length scale of $\theta_2 = 67$ years. The mean predictions inside the range of the training data and extending for 20 years into the future are depicted in Figure 5.7 (a). In the same plot (with right hand axis) we also show the medium term effects modelled by the rational quadratic component with magnitude $\theta_6 = 0.66$ ppm, typical length $\theta_7 = 1.2$ years and shape $\theta_8 = 0.78$. The very small shape value allows for covariance at many different length-scales, which is also evident in Figure 5.7 (a). Notice that beyond the edge of the training data the mean of this contribution smoothly decays to zero, but of course it still has a contribution to the uncertainty, see Figure 5.6.

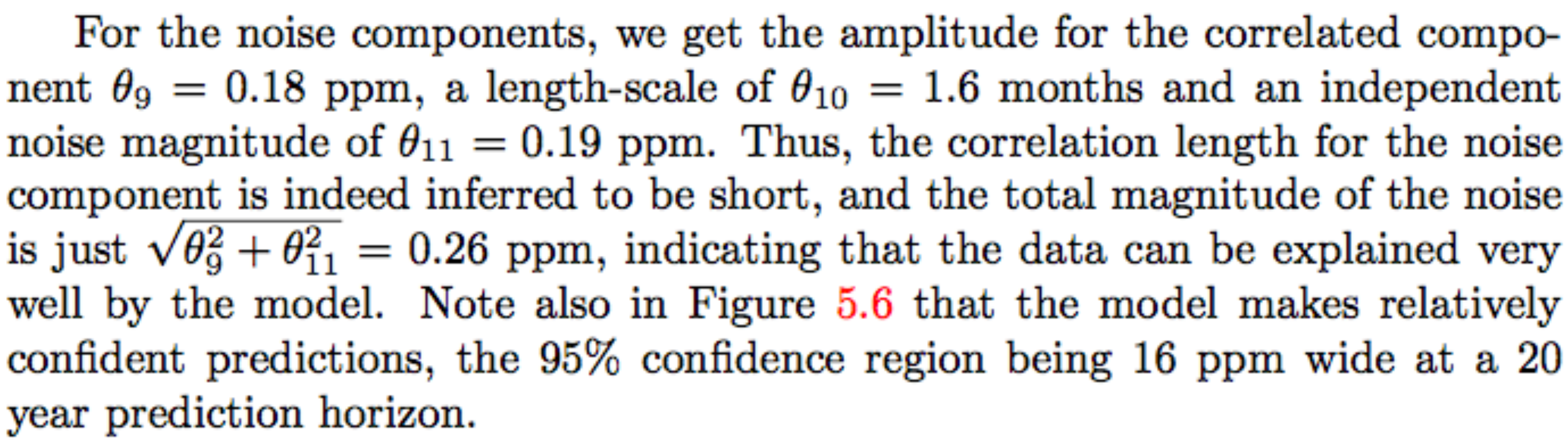






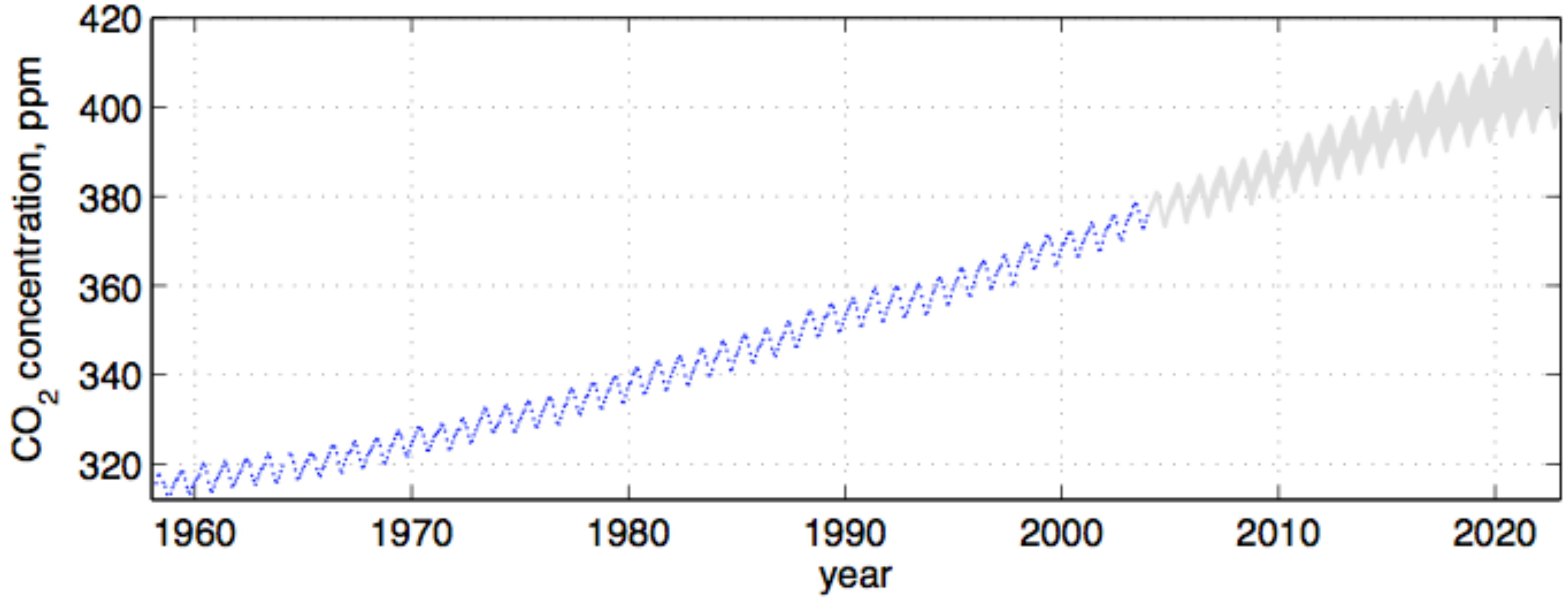
The hyperparameter values for the decaying periodic contribution are: magnitude $\theta_3 = 2.4$ ppm, decay-time $\theta_4 = 90$ years, and the smoothness of the periodic component is $\theta_5 = 1.3$. The quite long decay-time shows that the data have a very close to periodic component in the short term. In Figure 5.7 (b) we show the mean periodic contribution for three years corresponding to the beginning, middle and end of the training data. This component is not exactly sinusoidal, and it changes its shape slowly over time, most notably the amplitude is increasing, see Figure 5.8.

Manua Loa CO2 atmospheric data



For the noise components, we get the amplitude for the correlated component $\theta_9 = 0.18$ ppm, a length-scale of $\theta_{10} = 1.6$ months and an independent noise magnitude of $\theta_{11} = 0.19$ ppm. Thus, the correlation length for the noise component is indeed inferred to be short, and the total magnitude of the noise is just $\sqrt{\theta_9^2 + \theta_{11}^2} = 0.26$ ppm, indicating that the data can be explained very well by the model. Note also in Figure 5.6 that the model makes relatively confident predictions, the 95% confidence region being 16 ppm wide at a 20 year prediction horizon.

Prediction


CO_2 concentration, ppm
year