Louis Sharrock Imperial College London

Gaussian Process Regression

1. Summary

In this project, we investigate a probabilistic model that is often used within the atmospheric and environmental sciences, namely, Gaussian process regression. We apply this model to the Mauna Loa dataset to construct a predictive model for carbon dioxide (CO2) emissions over a 20 year period.

2. Preliminaries

2.1. The Data

We consider the Mauna Loa dataset, which consists of monthly atmospheric carbon dioxide (CO2) concentrations derived from the Scripps Institution of Oceanography's continuous monitoring program at the Mauna Loa Observatory, Hawaii, from 1958 to 1993. This record constitutes the longest continuous record of atmospheric CO2 concentrations anywhere in the world. The data is shown in Figure 1.

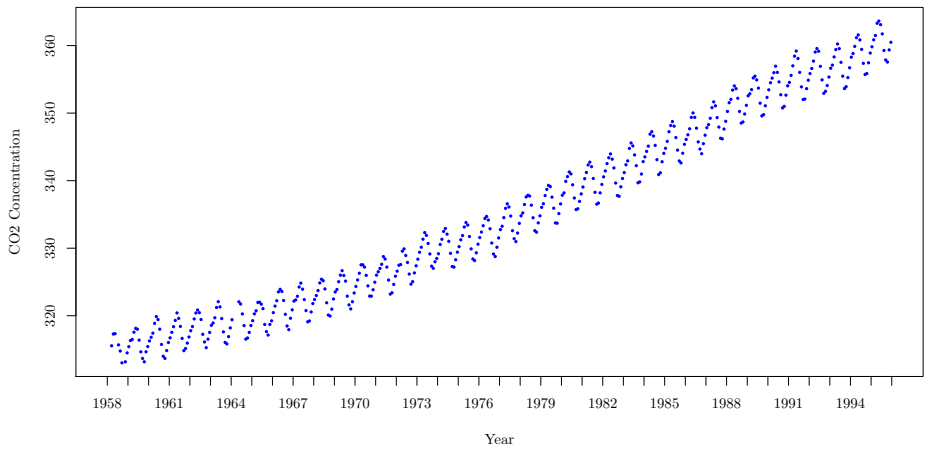

Figure 1: Plot of the Mauna Loa dataset, consisting of monthly averages of atmospheric concentration of CO_2 between the start of 1958 and the end of 1995.

Two features of this dataset are of particular note. Namely, a long term rising trend, and a pronounced seasonal fluctuation. There are, of course, other variations and irregularities which we may also hope to encompass in our model, but it is these two features which we would particularly like our model to capture.

3. Methods

3.1. The Bayesian Approach to Regression

3.1.1 Linear Regression

We begin by reviewing the Bayesian approach to linear regression, which may be used to describe an underlying hidden function of time given noisy data points. Suppose we have a (training set) $\mathcal{D}$ containing n observations, that is,

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n = (X, \mathbf{y}) \tag{1}$$

where $\mathbf{x}_i = (x_i^1, \dots, x_i^p)^T$ denotes a p-dimensional input vector, $X = (\boldsymbol{x}_1^T, \dots, \boldsymbol{x}_n^T)^T$ denotes the $p \times n$ matrix with columns containing all such input vectors, y denotes a scalar output, and $\mathbf{y} = (y_1, \dots, y_n)^T$ denotes the vector containing all such outputs. We are interested in obtaining the conditional distribution of these outputs, given the inputs.

The standard linear regression model can be summarised in the form

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w}, \qquad y = f(\mathbf{x}) + \varepsilon \tag{2}$$

where $\mathbf{x}$ is a vector of inputs, $\mathbf{w}$ is a vector of weights (i.e. parameters), f is the function value, and y is the observed target value. Note that we have assumed that the observed data y differ from the function values by additive noise. We will henceforth assume that this noise is independent and identically distributed (i.i.d.) Gaussian with zero mean and variance σ_n^2. That is, in a slight abuse of notation, $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$. Under these assumptions, the likelihood of the observed data is determined straightforwardly as

$$p(\mathbf{y}|X, \mathbf{w}) = \prod_{i=1}^n p(y_i|\mathbf{x}_i, \mathbf{w}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{(y_i - \mathbf{x}_i^T\mathbf{w})^2}{2\sigma_n^2}\right] = \mathcal{N}(X^T\mathbf{w}, \sigma_n^2 I). \tag{3}$$

It remains only to specify a prior distribution over the parameters. In this case, we will assume an independent Gaussian prior with mean zero and covariance Σ_w: $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_w)$.

Inference in the Bayesian linear regression model is derived from the posterior distribution over the weights, which is given by

$$p(\mathbf{w}|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \mathbf{w})p(\mathbf{w})}{\int p(\mathbf{y}|X, \mathbf{w})p(\mathbf{w})d\mathbf{w}}. \tag{4}$$

Let us ignore, for the moment, the normalising 'constant' in the denominator. That is, for now, we consider only terms depending on the weights. Then completing the square, we have

$$p(\mathbf{w}|X, \mathbf{y}) \propto \exp\left[-\frac{1}{2\sigma_n^2}(\mathbf{y} - X^T\mathbf{w})^T(\mathbf{y} - X^T\mathbf{w})\right] \cdot \exp\left[-\frac{1}{2}\mathbf{w}^T\sigma_w^{-1}\mathbf{w}\right] \tag{5}$$

$$\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_n^2}XX^T + \sigma_w^{-1}\right)(\mathbf{w} - \bar{\mathbf{w}})^T(\mathbf{w} - \bar{\mathbf{w}})\right], \tag{6}$$

where we have defined

$$\bar{\mathbf{w}} = \frac{1}{\sigma_n^2}\underbrace{\left(\frac{1}{\sigma_n^2}XX^T + \sigma_w^{-1}\right)^{-1}}_{\equiv A} X\mathbf{y} \equiv \frac{1}{\sigma_n^2}A^{-1}X\mathbf{y}. \tag{7}$$

This posterior is immediately recognisable as another normal distribution with mean $\bar{\mathbf{w}}$ and covariance matrix A^{-1}:

$$p(\mathbf{w}|X,\mathbf{y}) \sim \mathcal{N}(\bar{\mathbf{w}}, A^{-1}). \tag{8}$$

To make predictions for test data, say $\mathbf{x}^*$, we average over all possible parameter values, weighted according to their posterior probability. In particular, the predictive distribution for $f^* \equiv f(\mathbf{x}^*)$ at $\mathbf{x}^*$ is given by averaging the outputs of all linear models with respect to the normal posterior. That is,

$$p(f^*|\mathbf{x}^*, X, \mathbf{y}) = \int p(f^*|\mathbf{x}^*, \mathbf{w}) p(\mathbf{w}|X,\mathbf{y}) d\mathbf{w} = \mathcal{N}\left(\frac{1}{\sigma_n^2}\mathbf{x}^{*T}A^{-1}X\mathbf{y}, \mathbf{x}^{*T}A^{-1}\tilde{\mathbf{x}}\right) \tag{9}$$

We note that the predictive distribution is also Gaussian, with mean equal to the posterior mean of the weights, $\bar{\mathbf{w}}$, multiplied by the test data input, and variance given by the quadratic form of the test input and posterior covariance matrix.

3.1.2 Generalised Linear Regression

It is, in fact, possible to generalise this formulation of Bayesian linear regression via the introduction of a function $\boldsymbol{\phi}(\mathbf{x})$, which maps the p-dimensional input vector $\mathbf{x}$ into an N-dimensional 'feature space'. In this case, our model is now given by

$$f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^T\mathbf{w}, \qquad y = f(\mathbf{x}) + \varepsilon \tag{10}$$

where, of course, the vector of weights $\mathbf{w}$ now has length N.

Analysis of this more general model is entirely analogous to the standard linear model, now replacing any ocurrence of the design matrix X with $\Phi \equiv \Phi(X)$, the matrix with columns given by the feature vectors $\boldsymbol{\phi}(\mathbf{x})$. Hence the predictive distribution is now specified by

$$f^*|\mathbf{x}^*, X, \mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\phi}(\mathbf{x}_*)^TA^{-1}\boldsymbol{\phi}(\mathbf{x}_*)\mathbf{y}, \boldsymbol{\phi}(\mathbf{x}_*)^TA^{-1}\boldsymbol{\phi}(\mathbf{x}_*)\right) \tag{11}$$

where $A = \sigma_n^{-2}\Phi\Phi^T + \Sigma_w^{-1}$.

Given that A is of dimension $N \times N$, where N is the (potentially very large) dimension of the feature space, it is of interest to avoid computation of A^{-1}, as required by this distribution. To do so, we consider the following scheme. First, let us define $K = \Phi\Sigma_w\Phi$. Then, from the definitions of A and K, we have

$$\sigma_n^{-2}\boldsymbol{\Phi}(K + \sigma_n^2 I) = \sigma_n^{-2}\Phi(\Phi^T\Sigma_w\Phi + \sigma_n^2 I) = A\Sigma_w\Phi. \tag{12}$$

It follows straightforwardly that

$$\sigma_n^{-2}A^{-1}\Phi = \Sigma_w\Phi(K + \sigma_n^2 I)^{-1}. \tag{13}$$

Moreover, we can observe that

$$A^{-1} = \Sigma_w - \Sigma_w\Phi(K + \sigma_n^2 I)^{-1}\Phi^T\Sigma_w \tag{14}$$

Hence, writing $\boldsymbol{\phi}_* \equiv \boldsymbol{\phi}(\mathbf{x}_*)$, we obtain

$$f^*|\mathbf{x}^*, X, \mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\phi}_*^T\Sigma_w\Phi(K + \sigma_n^2 I)^{-1}\mathbf{y}, \boldsymbol{\phi}_*^T\Sigma_w\boldsymbol{\phi}_* - \boldsymbol{\phi}_*^T\Sigma_w\Phi(K + \sigma_n^2 I)^{-1}\Phi^T\Sigma_w\boldsymbol{\phi}_*\right) \tag{15}$$

This involves the inversion of matrices of dimension $n \times n$, which is computationally preferrable whenever n, the number of data points, is less than N, the dimension of the feature space.

To conclude this exposition, it is worth noting that the components of this distribution can be defined solely in terms of inner products in the input space, and thus we may apply the 'kernel trick' in computing predictive distributions. Observe that all occurrences of the feature space in the predictive distribution are of the form $\Phi^T\Sigma_w\Phi$, $\boldsymbol{\phi}_*^T\Sigma_w\Phi$ or $\boldsymbol{\phi}_*^T\Sigma_w\boldsymbol{\phi}_*$. It follows that the entries of these matrices, now dropping our shorthand notation, are always of the form $\boldsymbol{\phi}(\mathbf{x})^T\Sigma_w\boldsymbol{\phi}(\mathbf{x}')$, where $\mathbf{x}$ and $\mathbf{x}'$ are vectors in the training and test sets respectively. We thus define a kernel function

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T\Sigma_w\phi(\mathbf{x}') \tag{16}$$

3.2. The Gaussian Process Approach to Linear Regression

The Bayesian approach to linear regression is in fact precisely equivalent to the Gaussian process approach with a suitable choice of kernel function. To show this, we begin by recalling the definition of a Gaussian process:

Definition. *A Gaussian process is an infinite collection of random variables, any finite number of which have a joint Gaussian distribution.*

We will write

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right) \tag{17}$$

where the mean function $m(\mathbf{x})$ and covariance (or kernel) function $k(\mathbf{x}, \mathbf{x}')$ of the real process $f(\mathbf{x})$ fully specify the Gaussian Process, and are defined via

$$m(\mathbf{x}) = \mathbb{E}\left[f(\mathbf{x})\right] \tag{18}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right] \tag{19}$$

For our Bayesian linear regression model, we recall that $f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^T\mathbf{w}$, with prior distribution on the weights given by $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_w)$. Thus we have mean,

$$\mathbb{E}\left[f(\mathbf{x})\right] = \boldsymbol{\phi}(\mathbf{x})\mathbb{E}\left[\mathbf{w}\right] = 0 \tag{20}$$

and covariance (at the two time points $\mathbf{x}$ and $\mathbf{x}'$)

$$\mathbb{E}\left[f(\mathbf{x})f(\mathbf{x}')\right] = \boldsymbol{\phi}(\mathbf{x})^T\mathbb{E}\left[\mathbf{w}\mathbf{w}^T\right]\boldsymbol{\phi}(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T\Sigma_w\boldsymbol{\phi}(\mathbf{x}') \tag{21}$$

This covariance, or kernel, function corresponds precisely to an inner product of basis functions $\boldsymbol{\phi}(\cdot)$, with respect to Σ_w. Equivalently, writing

$$\boldsymbol{\varphi}(\mathbf{x}) = \Sigma_w^{1/2}\boldsymbol{\phi}(\mathbf{x}), \tag{22}$$

which is well defined since Σ_w is positive definite, this kernel function can in fact be expressed more simply using an inner product of the basis functions $\boldsymbol{\varphi}(\cdot)$:

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\varphi}(\mathbf{x}) \cdot \boldsymbol{\varphi}(\mathbf{x}') \tag{23}$$

The derived mean and covariance functions clearly define a valid Gaussian Process, from which it follows immediately that the Bayesian linear regression approach is equivalent to the Gaussian process approach, with kernel function as specified previously.

4. A Predictive Model for CO2 Emissions.

Having reviewed the necessary theory, we now proceed to use a Gaussian process, trained on the Mauna-Loa dataset, to construct a predictive model of CO2 emissions over the subsequent 20 years. The details of our approach our outlined below.

4.1. The Data

We begin by reviewing our training data, which has (scalar) inputs corresponding to the time, measured as the number of years since 01-01-1960, and targets corresponding to the atmospheric carbon dioxide concentration at each of these timepoints. There are 449 such observations, corresponding to measurements for the thirty year period from 1958 to 1993. Thus we can write our data-set as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{449} = \big(X_{\{1\times 449\}}, \mathbf{y}\big)$.

4.2. The Model

4.2.1 Summary

We will model each datapoint as some function of the corresponding input, plus some noise term. We thus have

$$y_i = f(x_i) + \varepsilon_i, \tag{24}$$

We assume i.i.d. Gaussian noise for each of the inputs, so $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2)$, and a Gaussian process with mean zero and kernel function k over the functions f, so $\mathbf{f}(\mathbf{x}) \sim \mathcal{GP}(m = 0, k)$. We can thus write the 'prior on noisy observations' as

$$\text{Cov}(y_p, y_q) = k(\mathbf{x}_p, \mathbf{x}_q) + \sigma_n^2 \delta_{pq} \iff \text{Cov}(\mathbf{y}) = K(X, X) + \sigma_n^2 I_n \tag{25}$$

where $K(X, X)$ denotes the the $n \times n$ matrix of covariances evaluated at all pairs of training points (x_i, x_j).

4.2.2 Predictions

We wish to make predictions of y_i^* for test inputs x_i^* corresponding to the twenty year period from 1993 to 2013. To do so, we must consider the joint distribution of the observed targets and the function evaluated at these test inputs, f_*, which under the given prior is precisely

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right), \tag{26}$$

where $K(X, X_*)$ etc. are defined analogously to $K(X, X)$. Using standard properties of multivariate normal distributions, we obtain $\mathbf{f}_* | X, \mathbf{y}, X_* \sim \mathcal{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$, where

$$\boldsymbol{\mu}_* = K(X_*, X)\left[K(X, X) + \sigma_n^2 I_n\right]^{-1} \mathbf{y} \tag{27}$$

$$\boldsymbol{\Sigma}_* = K(X_*, X_*) - K(X_*, X)\left[K(X, X) + \sigma_n^2 I\right]^{-1} K(X, X_*) \tag{28}$$

It follows that the predictive distribution of $\mathbf{y}_*$ is given by $\mathbf{y}_* | X, \mathbf{y}, X_* \sim \mathcal{N}\left(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_* + \sigma_n^2 I_n\right)$.

4.2.3 The Kernel

It remains to specify an appropriate kernel function for the analysis at hand. This particular data-set has been widely studied in the literature, and in particular, one kernel has been used extensively in existing analyses. We will base our own analysis on a (modified) version of this kernel. Our choice is also motivated by the primary features of the underlying training data: namely, a long term rising trend, and a pronounced seasonal fluctuation.

We will construct our combined kernel function as a sum of several component functions, which should respectively account for the different modes of variations in our data.

(i) To model the smooth, long term increase in CO_2 concentration as a function of time, we will implement a squared exponential term:

$$k_1(x, x') = \theta_1^2 \exp\left(-\frac{(x - x')^2}{2\theta_2^2}\right) \tag{29}$$

The hyper-parameters θ_1 and θ_2 here correspond to the amplitude and characteristic length scale respectively.

(ii) We model the seasonal variation via

$$k_2(x, x') = \theta_3^2 \exp\left(-\frac{(x - x')^2}{2\theta_4^2} - \frac{2\sin^2(\pi(x - x'))}{\theta_5^2}\right). \tag{30}$$

where here θ_3 corresponds to the magnitude, θ_4 the decay time for the periodic component, and θ_5 the 'smoothness' of the periodic component. Note that the inclusion of the first term in the exponent is to allow for a decay away from exact periodicity.

(iii) Finally, for ease of implementation, we will include noise in our model explicitly as a component kernel function

$$k_3(x_p, x_q) = \theta_6^2 \delta_{pq} \equiv \sigma_n^2 \delta_{pq} \tag{31}$$

Our final covariance function is then given by

$$k(x, x') = k_1(x, x') + k_2(x, x') + k_3(x, x') \tag{32}$$

Having established a suitable covariance function, we are required to determine optimal values of the hyperparameters $\boldsymbol{\theta} = (\theta_1, \ldots, \sigma_n)$ on the basis of the available training data. It is in this setting that we introduce the marginal likelihood, or evidence, $p(\mathbf{y}|X)$. This is defined by

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f}, X) p(\mathbf{f}|X) d\mathbf{f}. \tag{33}$$

Under the Gaussian process model, the prior is Gaussian, $\mathbf{f}|X \sim \mathcal{N}(0, K_f)$, where we use K_f to denote the covariance matrix for the noise-free latent $\mathbf{f}$, or equivalently

$$\log p(\mathbf{f}|X) = -\frac{1}{2}\mathbf{f}^T K_f^{-1} \mathbf{f} - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi, \tag{34}$$

and the likelihood is a factorised Gaussian, $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(0, \sigma_n^2 I)$. It follows from standard results that the log-marginal likelihood is given by

$$\log p(\mathbf{y}|X, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^T K_y^{-1} \mathbf{y} - \frac{1}{2}\log|K_y| - \frac{n}{2}\log 2\pi \tag{35}$$

where $K_y = K_f + \sigma_n^2 I$ is the covariance matrix for the noisy targets. Note the choice to now explicitly include the dependence of this function on the hyperparameters $\boldsymbol{\theta}$. Specification of the optimal hyperparameters then corresponds precisely to maximisation of the log-marginal likelihood, with respect to all such parameters.

For this optimisation task, we implement the 'Nelder-Mead' algorithm via the `optim` routine in R. Though somewhat slower to converge than alternative method such as conjugate gradients, Nelder-Mead requires only the computation of function values and as such is relatively robust. Despite this, we should note that there is no guarantee that the marginal likelihood does not suffer from multiple local optima, and that care must be taken to ensure that we do not end up in a bad local optimum. Indeed, our own empirical results indicate that the optimal hyper-parameters determined by the

chosen algorithm are highly dependent on the initial parameter values.[1] To address this issue, we consider several approaches. These include using:

(i) Knowledge of the 'physical' interpretation of each of our hyperparameters. In particular, on the basis of our training data, and considering the parameters encoded in each of our component kernel functions in turn, it is reasonable to expect:
   - $k_1(x, x')$: θ_1 and θ_2 both to be relatively large. These two parameters correspond respectively to the magnitude and length scale of the long term trend in our data, which are clearly somewhat significant on the basis of the training set (see Figure 2).
   - $k_2(x, x')$: θ_3 to be relatively large, θ_4 to be 'very large'. The former denotes the magnitude of the seasonal variation in the data, which is very notable, while the latter represents the decay time of this variation. The data are very close to periodic in the short term, suggesting this decay time should be very long. Interpretation of θ_5, the 'smoothness' of the periodic component, is somewhat less apparent.
   - $k_3(x, x')$: $\theta_6 \equiv \sigma_n$ to be very small. The data-set does not appear to be at all 'noisy', suggesting this parameter will be optimised by some value close to zero.

(ii) A 'step-wise' approach to optimisation. Explicitly, we seek to maximise the marginal likelihood derived from each component of our kernel function individually. The parameter spaces of these optimisation problems (respectively, $\mathbb{R}^2$ and $\mathbb{R}^3$), are clearly of lower-dimension than the parameter space corresponding to optimisation of the full kernel function ($\mathbb{R}^6$), and as such, we are much less likely to suffer from the issue of multiple optima. Clearly it is still necessary in the final instance to maximise the marginal likelihood with respect to all hyperparameters simultaneously, but the values specified by our previous optimisations now provide highly suitable initial inputs, more likely to result in convergence to the required global optimum.

Irrespective of these arguments, we still opt to run `optim` with several sets of initial values, and choose the output obtaining the maximal value of the marginal likelihood. The results obtained at each stage of our optimisation procedure are presented in Figures 2 - 4.

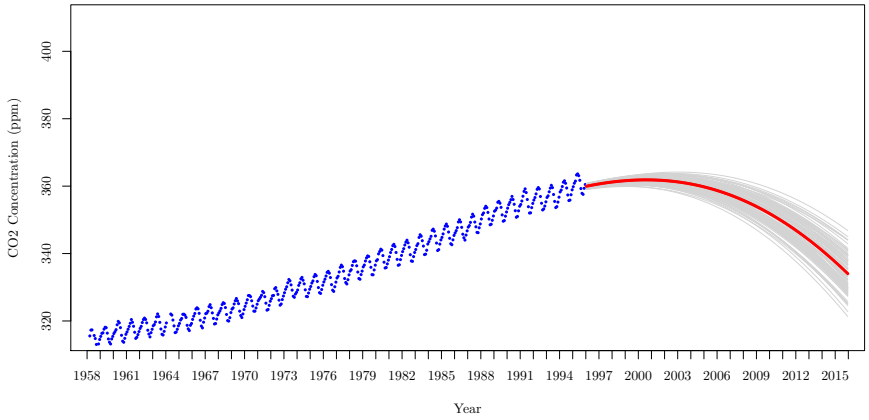

Figure 2: Gaussian process regression model for atmospheric CO_2 concentration, for twenty years after the end of the training data, based on the kernel function $k(x, x') = k_1(x, x') + k_3(x, x')$. The 'optimised' hyperparameters are given by $\theta_1 = 25.92$, $\theta_2 = -32.19$ and $\theta_6 \equiv \sigma_n = 2.09$.

[1] They are also highly dependent on the `parscale` option within the `optim` function, which provides a vector of scaling values for the parameters.

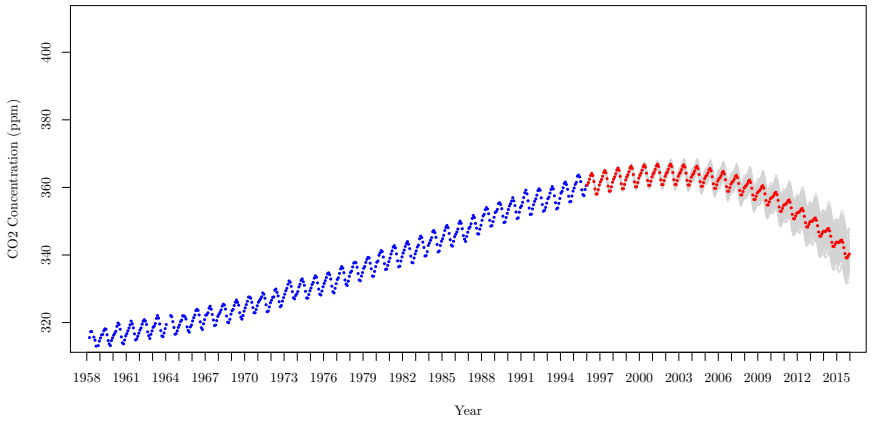

Figure 3: Gaussian process regression model for atmospheric CO_2 concentration, for twenty years after the end of the training data, based on the kernel function $k(x, x') = k_2(x, x') + k_3(x, x')$. The 'optimised' hyperparameters are given by $\theta_3 = 19.59$, $\theta_4 = -27.99$, $\theta_5 = 4.82$ and $\theta_6 \equiv \sigma_n = 0.45$.

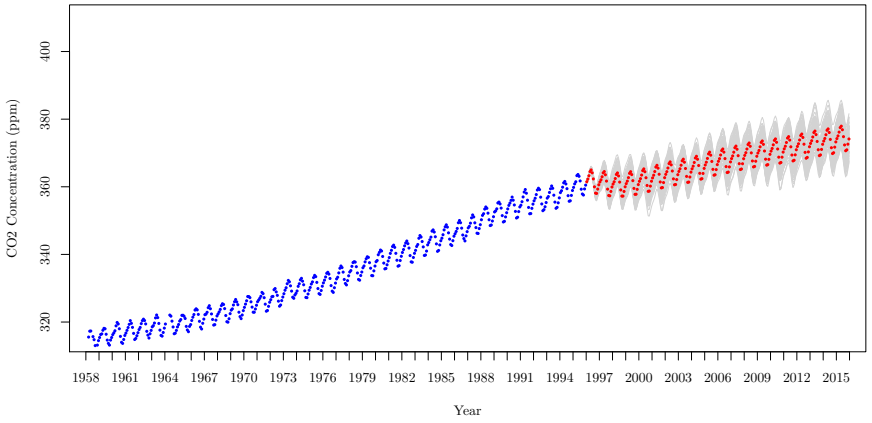

Figure 4: Gaussian process regression model for atmospheric CO_2 concentration, for twenty years after the end of the training data, based on the kernel function $k(x, x') = k_1(x, x') + k_2(x, x') + k_3(x, x')$.The 'optimised' hyperparameters are given by $\theta_1 = 1.60$, $\theta_2 = -1.19$, $\theta_3 = 43.65$, $\theta_4 = -107.33$, $\theta_5 = 3.08$ and $\theta_6 \equiv \sigma_n = 0.32$.

The log marginal likelihood of each of these three models is given by

$$\log p(\mathbf{y}|X, \boldsymbol{\theta}_{1,2,6}) = -982.6924 \tag{36}$$

$$\log p(\mathbf{y}|X, \boldsymbol{\theta}_{3,4,5,6}) = -334.2191 \tag{37}$$

$$\log p(\mathbf{y}|X, \boldsymbol{\theta}_{1,2,3,4,5,6}) = -177.559 \tag{38}$$

As expected, this quantity increases as a function of the complexity and relevance of the kernel function. Furthermore, somewhat reassuringly, it is increasing as a function of our qualitative assessment of the quality of the model fit. This is to say, somewhat more explicitly, that of the three models, the third (Figure 4) has the largest log marginal likelihood, and would also be assessed to have the best fit. Similarly, of the three models, the first (Figure 2) has the smallest log marginal likelihood, and would be assessed as having the worst fit.

It is clear that even our best model (Figure 4), computed on the basis of the full kernel function

$k(x, x') = k_1(x, x') + k_2(x, x') + k_3(x, x')$, while generally of the expected form, leaves room for significant improvement. In this context, it is of interest to investigate the introduction of a mean function $m(\mathbf{x})$ into our Gaussian process. In particular, we will now consider

$$g(\mathbf{x}) = f(\mathbf{x}) + \mathbf{h}(\mathbf{x})^T \boldsymbol{\beta} \quad , \quad f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')) \tag{39}$$

where $f(\mathbf{x})$ is a zero-mean Gaussian process, as above, $\mathbf{h}(\mathbf{x})$ are a set of fixed basis functions, and β are additional parameters to be inferred from the data. This formulation can be seen as representing that the data is close to a global linear model, with the residuals now modelled by a Gaussian process. Not only do we hope this may provide a a better predictive model in the case of the given data, but it has the advantage of providing a somewhat more interpretable model.

With the introduction of this mean function, the predictive distribution of interest remains of the form $\mathbf{f}_*|X, \mathbf{y}, X_* \sim \mathcal{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$. The predictive covariance is unchanged, while the predictive mean is now given by

$$\boldsymbol{\mu}_* = \mathbf{m}(X_*) + K(X_*, X)\left[K(X, X) + \sigma_n^2 I_n\right]^{-1}(\mathbf{y} - \mathbf{m}(X)) \tag{40}$$

The log marginal likelihood, which we now wish to maximise with respect to the full set of parameters $(\boldsymbol{\theta}, \boldsymbol{\beta})$ is similarly now given by

$$\log p(\mathbf{y}|X, \boldsymbol{\theta}) = -\frac{1}{2}(\mathbf{y} - \mathbf{m}(X))^T K_y^{-1}(\mathbf{y} - \mathbf{m}(X)) - \frac{1}{2}\log|K_y| - \frac{n}{2}\log 2\pi \tag{41}$$

It remains to specify a suitable set of fixed basis functions for the mean, which (as with the kernel function) is best determined on the basis of the behaviour exhibited by the training data. As such, in this instance we choose simply to specify $\mathbf{h}(x) = (1, x)$, corresponding to a linear mean function.

The results of the inclusion of a mean function in our Gaussian Process are provided in Figure 5. As anticipated, the behaviour exhibited by the new model is clearly favourable to its predecessor, with log marginal likelihood now given by

$$\log p(\mathbf{y}|X, \boldsymbol{\theta}, \boldsymbol{\beta}) = -175.9161 \tag{42}$$

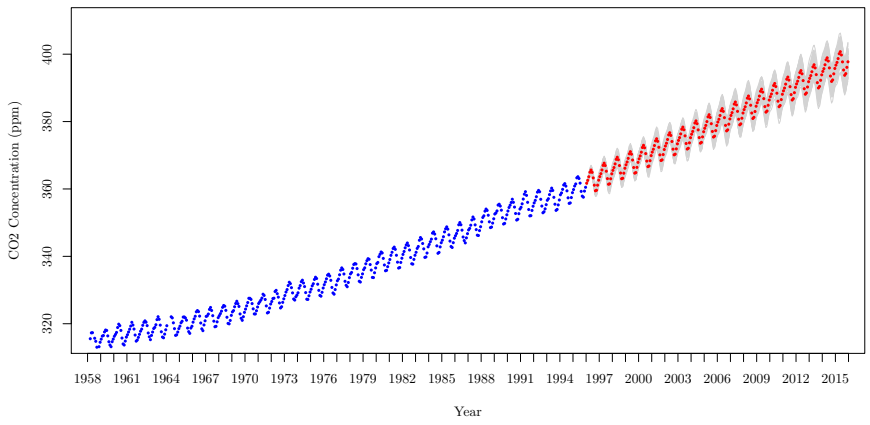

Figure 5: Gaussian process regression model for atmospheric CO_2 concentration, for twenty years after the end of the training data, based on the kernel function $k(x, x') = k_1(x, x') + k_2(x, x') + k_3(x, x')$, and linear mean function $m(x)$.The 'optimised' hyperparameters are given by $\theta_1 = 0.83$, $\theta_2 = -0.74$, $\theta_3 = 74.62$, $\theta_4 = -105.50$, $\theta_5 = 3.43$, $\theta_6 \equiv \sigma_n = 0.31$, $\beta_0 = 313.83$ and $\beta_1 = 0.50$.

4.3. Additional Remarks

In 2013, the atmospheric concentration of CO_2 was measured to have exceeded 400ppm for the first time in human history. This observation compares very closely with the predictions made by our model, for which the peak value in this year is given by 396.8563ppm. The slight negative discrepancy between the peak value predicted by our model, and the true peak value, can perhaps be explained in the context of a recent acceleration in the rate of global warming, and, thus, presumably, an acceleration in the increase in atmospheric concentration of CO_2. This trend, ocurring within the last twenty years, is clearly not encoded in our training data, and as such it is unreasonable to expect that it would be represented in our predictive model. This being said, the difference between the the true peak value in 2013 and our prediction is, on the whole, insignificant. Indeed, it is not unreasonable to suggest that it may be entirely reconciled if the model were to be fit using very slightly different hyper-parameters, which our optimisation procedure may have determined were it to have been re-run a vast number of times with a wider array of slightly varying initial parameters.