Lecture 12

Gaussian Process Models

10/1~~6~~7/2018

Multivariate Normal

Multivariate Normal Distribution

For an n-dimension multivate normal distribution with covariance $\boldsymbol{\Sigma}$ (positive semidefinite) can be written as

$$\underset{n\times 1}{\mathbf{Y}} \sim N(\underset{n\times 1}{\boldsymbol{\mu}},\ \underset{n\times n}{\boldsymbol{\Sigma}}) \text{ where } \{\boldsymbol{\Sigma}\}_{ij} = \sigma^2_{ij} = \rho_{ij}\,\sigma_i\,\sigma_j$$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, \begin{pmatrix} \rho_{11}\sigma_1\sigma_1 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \cdots & \rho_{nn}\sigma_n\sigma_n \end{pmatrix} \right)$$

For the n dimensional multivate normal given on the last slide, its density is given by

$$(2\pi)^{-n/2}\ \det(\boldsymbol{\Sigma})^{-1/2}\ \exp\left(-\frac{1}{2}\underset{1\times n}{(\mathbf{Y}-\boldsymbol{\mu})'}\underset{n\times n}{\boldsymbol{\Sigma}^{-1}}\underset{n\times 1}{(\mathbf{Y}-\boldsymbol{\mu})}\right)$$

and its log density is given by

$$-\frac{n}{2}\log 2\pi - \frac{1}{2}\log\det(\boldsymbol{\Sigma}) - -\frac{1}{2}\underset{1\times n}{(\mathbf{Y}-\boldsymbol{\mu})'}\underset{n\times n}{\boldsymbol{\Sigma}^{-1}}\underset{n\times 1}{(\mathbf{Y}-\boldsymbol{\mu})}$$

To generate draws from an n-dimensional multivate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$,

Sampling

To generate draws from an n-dimensional multivate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$,

- Find a matrix $\mathbf{A}$ such that $\boldsymbol{\Sigma} = \mathbf{A}\,\mathbf{A}^t$, most often we use $\mathbf{A} = \text{Chol}(\boldsymbol{\Sigma})$ where $\mathbf{A}$ is a lower triangular matrix.

Sampling

To generate draws from an n-dimensional multivate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$,

- Find a matrix $\mathbf{A}$ such that $\boldsymbol{\Sigma} = \mathbf{A}\,\mathbf{A}^t$, most often we use $\mathbf{A} = \text{Chol}(\boldsymbol{\Sigma})$ where $\mathbf{A}$ is a lower triangular matrix.

- Draw n iid unit normals ($\mathcal{N}(0,1)$) as $\mathbf{z}$

Sampling

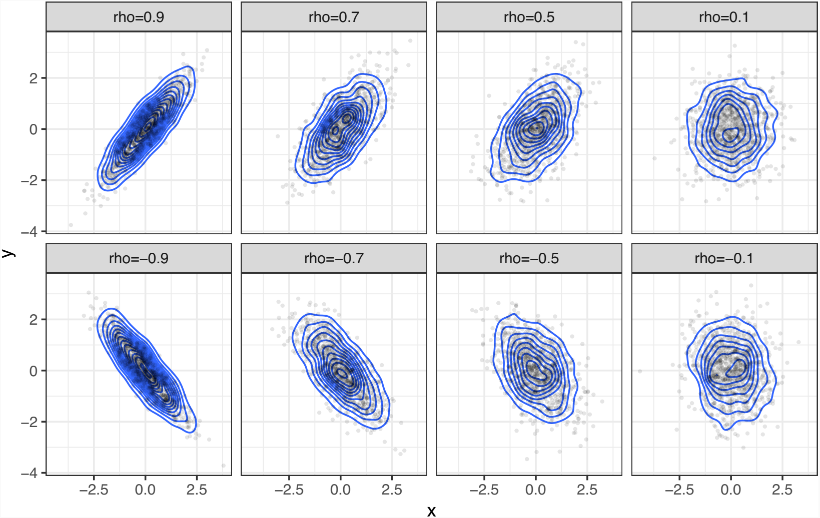
To generate draws from an n-dimensional multivate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$,

- Find a matrix $\mathbf{A}$ such that $\boldsymbol{\Sigma} = \mathbf{A}\,\mathbf{A}^t$, most often we use $\mathbf{A} = \text{Chol}(\boldsymbol{\Sigma})$ where $\mathbf{A}$ is a lower triangular matrix.
- Draw n iid unit normals ($\mathcal{N}(0,1)$) as $\mathbf{z}$
- Obtain multivariate normal draws using

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times 1}{\boldsymbol{\mu}} + \underset{n\times n}{\mathbf{A}}\,\underset{n\times 1}{\mathbf{z}}$$

Bivariate Example

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$



Marginal distributions

Proposition - For an n-dimensional multivate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, any marginal or conditional distribution of the y's will also be (multivariate) normal.

Marginal distributions

Proposition - For an n-dimensional multivate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, any marginal or conditional distribution of the y's will also be (multivariate) normal.

For a univariate marginal distribution,

$$y_i = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii})$$

Marginal distributions

Proposition - For an n-dimensional multivate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, any marginal or conditional distribution of the y's will also be (multivariate) normal.

For a univariate marginal distribution,

$$y_i = \mathcal{N}(\boldsymbol{\mu}_i,\, \boldsymbol{\Sigma}_{ii})$$

For a bivariate marginal distribution,

$$\mathbf{y}_{ij} = \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_i \\ \boldsymbol{\mu}_j \end{pmatrix},\ \begin{pmatrix} \boldsymbol{\Sigma}_{ii} & \boldsymbol{\Sigma}_{ij} \\ \boldsymbol{\Sigma}_{ji} & \boldsymbol{\Sigma}_{jj} \end{pmatrix} \right)$$

Proposition - For an n-dimensional multivate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, any marginal or conditional distribution of the y's will also be (multivariate) normal.

For a univariate marginal distribution,

$$y_i = \mathcal{N}(\boldsymbol{\mu}_i,\, \boldsymbol{\Sigma}_{ii})$$

For a bivariate marginal distribution,

$$\mathbf{y}_{ij} = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}_i \\ \boldsymbol{\mu}_j\end{pmatrix},\, \begin{pmatrix}\boldsymbol{\Sigma}_{ii} & \boldsymbol{\Sigma}_{ij} \\ \boldsymbol{\Sigma}_{ji} & \boldsymbol{\Sigma}_{jj}\end{pmatrix}\right)$$

For a k-dimensional marginal distribution,

$$\mathbf{y}_{i,\cdots,k} = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}_i \\ \vdots \\ \boldsymbol{\mu}_k\end{pmatrix},\, \begin{pmatrix}\boldsymbol{\Sigma}_{ii} & \cdots & \boldsymbol{\Sigma}_{ik} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{ki} & \cdots & \boldsymbol{\Sigma}_{kk}\end{pmatrix}\right)$$

If we partition the n-dimensions into two pieces such that $\mathbf{Y} = (\mathbf{Y}_1,\, \mathbf{Y}_2)^t$ then

$$\underset{n\times 1}{\mathbf{Y}} \sim \mathcal{N}\left(\underset{n\times 1}{\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}},\, \underset{n\times n}{\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}} \right)$$

$$\begin{aligned} \underset{k\times 1}{\mathbf{Y}_1} &\sim \mathcal{N}(\underset{k\times 1}{\boldsymbol{\mu}_1},\, \underset{k\times k}{\boldsymbol{\Sigma}_{11}}) \\ \underset{n-k\times 1}{\mathbf{Y}_2} &\sim \mathcal{N}(\underset{n-k\times 1}{\boldsymbol{\mu}_2},\, \underset{n-k\times n-k}{\boldsymbol{\Sigma}_{22}}) \end{aligned}$$

Conditional Distributions

If we partition the n-dimensions into two pieces such that $\mathbf{Y} = (\mathbf{Y}_1,\, \mathbf{Y}_2)^t$ then

$$\underset{n\times 1}{\mathbf{Y}} \sim \mathcal{N}\left(\underset{n\times 1}{\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}},\ \underset{n\times n}{\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}} \right)$$

$$\underset{k\times 1}{\mathbf{Y}_1} \sim \mathcal{N}(\ \underset{k\times 1}{\boldsymbol{\mu}_1}\ ,\ \underset{k\times k}{\boldsymbol{\Sigma}_{11}}\ )$$

$$\underset{n-k\times 1}{\mathbf{Y}_2} \sim \mathcal{N}(\ \underset{n-k\times 1}{\boldsymbol{\mu}_2}\ ,\ \underset{n-k\times n-k}{\boldsymbol{\Sigma}_{22}}\ )$$

then the conditional distributions are given by

$$\mathbf{Y_1} \,|\, \mathbf{Y}_2 = \mathbf{a} \ \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{1} + \boldsymbol{\Sigma}_\mathbf{12}\, \boldsymbol{\Sigma}_\mathbf{22}^{-1}\, (\mathbf{a} - \boldsymbol{\mu}_\mathbf{2}),\ \boldsymbol{\Sigma}_\mathbf{11} - \boldsymbol{\Sigma}_\mathbf{12}\, \boldsymbol{\Sigma}_\mathbf{22}^{-1}\, \boldsymbol{\Sigma}_\mathbf{21})$$

k×1   k×1   k×n-k   n-k×n-k   k×1   k×1   k×k   k×n-k   n-k×n-k   n-k × k

$$\mathbf{Y_2} \,|\, \mathbf{Y}_1 = \mathbf{b} \ \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{2} + \boldsymbol{\Sigma}_\mathbf{21}\, \boldsymbol{\Sigma}_\mathbf{11}^{-1}\, (\mathbf{b} - \boldsymbol{\mu}_\mathbf{1}),\ \boldsymbol{\Sigma}_\mathbf{22} - \boldsymbol{\Sigma}_\mathbf{21}\, \boldsymbol{\Sigma}_\mathbf{11}^{-1}\, \boldsymbol{\Sigma}_\mathbf{21})$$

Gaussian Processes

From Shumway,

> *A process, $\mathbf{Y} = \{Y(t) \,:\, t \in T\}$, is said to be a Gaussian process if all possible finite dimensional vectors $\mathbf{y} = \cancel{(y_{t_1}, y_{t_2}, \dots, y_{t_n})^t}$, for every collection of time points $t_1, t_2, \dots, t_n$, and every positive integer n, have a multivariate normal distribution.*

$y = (y(t_1), y(t_2), \dots y(t_n))$

From Shumway,

> *A process,* $\mathbf{Y} = \{Y(t) \,:\, t \in T\}$*, is said to be a Gaussian process if all possible finite dimensional vectors* $\mathbf{y} = (y_{t_1}, y_{t_2}, ..., y_{t_n})^t$*, for every collection of time points* $t_1, t_2, \ldots, t_n$*, and every positive integer* n*, have a multivariate normal distribution.*

So far we have only looked at examples of time series where T is discete (and evenly spaces & contiguous), it turns out things get a lot more interesting when we explore the case where T is defined on a *continuous* space (e.g. $\mathbb{R}$ or some subset of $\mathbb{R}$).

Gaussian Process Regression

Imagine we have a Gaussian process defined such that
$\mathbf{Y} = \{Y(t) \,:\, t \in [0,1]\}$,

Parameterizing a Gaussian Process

Imagine we have a Gaussian process defined such that $\mathbf{Y} = \{Y(t) \,:\, t \in [0,1]\}$,

- We now have an uncountably infinite set of possible t's and $Y(t)$s.

Parameterizing a Gaussian Process

Imagine we have a Gaussian process defined such that
$\mathbf{Y} = \{Y(t) \,:\, t \in [0,1]\}$,

- We now have an uncountably infinite set of possible t's and $Y(t)$s.
- We will only have a (small) finite number of observations $Y(t_1), \ldots, Y(t_n)$ with which to say something useful about this infinite dimensional process.

Parameterizing a Gaussian Process

Imagine we have a Gaussian process defined such that $\mathbf{Y} = \{Y(t)\ :\ t \in [0,1]\}$,

- We now have an uncountably infinite set of possible t's and $Y(t)$s.
- We will only have a (small) finite number of observations $Y(t_1), \dots, Y(t_n)$ with which to say something useful about this infinite dimensional process.
- The unconstrained covariance matrix for the observed data can have up to $n(n+1)/2$ unique values*

$\Sigma =$

Parameterizing a Gaussian Process

Imagine we have a Gaussian process defined such that
$\mathbf{Y} = \{Y(t) \,:\, t \in [0,1]\}$,

- We now have an uncountably infinite set of possible t's and $Y(t)$s.
- We will only have a (small) finite number of observations $Y(t_1), \ldots, Y(t_n)$ with which to say something useful about this infinite dimensional process.
- The unconstrained covariance matrix for the observed data can have up to $n(n+1)/2$ unique values*
- Necessary to make some simplifying assumptions:
    - Stationarity → Const $E(Y)$, Finite $Var(Y)$, Cov depends on distance
    - Simple parameterization of Σ

Covariance Functions

More on these next week, but for now some simple / common examples

More on these next week, but for now some simple / common examples

Exponential Covariance:

scale

dist.

$$\Sigma(y_t, y_{t'}) = \sigma^2 \exp\left(-\,|t-t'|\;l\,\right)$$

range

More on these next week, but for now some simple / common examples

Exponential Covariance:

$$\Sigma(y_t, y_{t'}) = \sigma^2 \exp\big(-|t - t'|\; l\,\big)$$

Squared Exponential Covariance:

$$\Sigma(y_t, y_{t'}) = \sigma^2 \exp\big(-(|t - t'|\; l\,)^2\big)$$

More on these next week, but for now some simple / common examples

Exponential Covariance:

$$\Sigma(y_t, y_{t'}) = \sigma^2 \exp\big(-|t - t'|\; l\big)$$

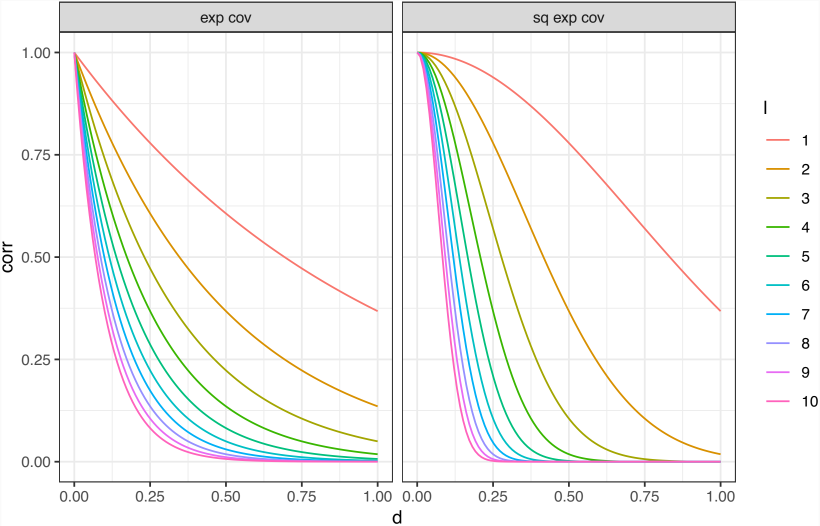
Squared Exponential Covariance:

$$\Sigma(y_t, y_{t'}) = \sigma^2 \exp\big(-(|t - t'|\; l)^2\big)$$

Powered Exponential Covariance ($p \in (0, 2]$):

$$\Sigma(y_t, y_{t'}) = \sigma^2 \exp\big(-(|t - t'|\; l)^p\big)$$

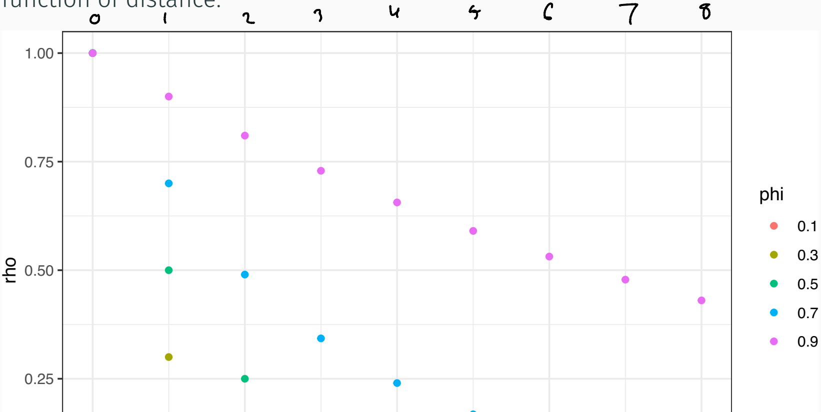
Covariance Function - Correlation Decay



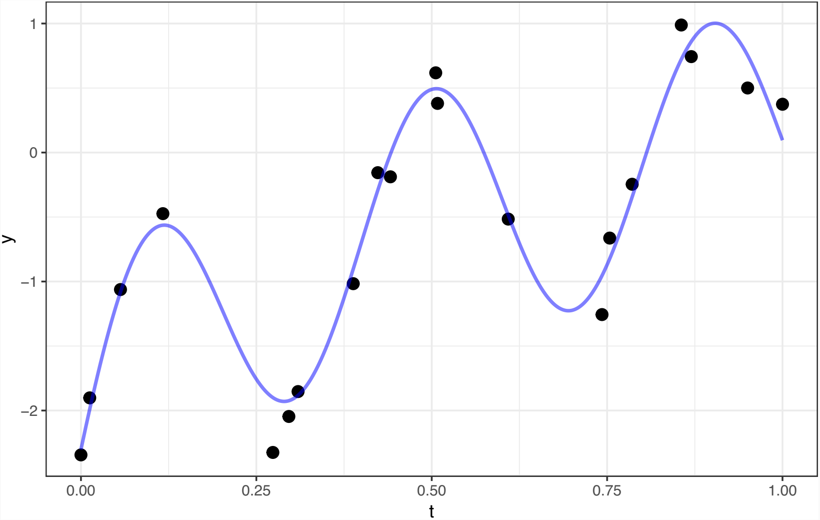
Recall that for a stationary AR(1) process:

$$\gamma(h) = \sigma_w^2 \phi^{|h|} \text{ and } \rho(h) = \phi^{|h|}$$

therefore we can draw a somewhat similar picture about the decay of ρ as a function of distance.



Example



Prediction

Our example has 15 observations which we would like to use as the basis for predicting $Y(t)$ at other values of t (say a sequence of values from 0 to 1).

Our example has 15 observations which we would like to use as the basis for predicting $Y(t)$ at other values of t (say a sequence of values from 0 to 1).

10 15

For now lets use a square exponential covariance with $\sigma^2 = 10$ and $l = 5$

Our example has 15 observations which we would like to use as the basis for predicting $Y(t)$ at other values of t (say a sequence of values from 0 to 1).

For now lets use a square exponential covariance with $\sigma^2 = 10$ and $l = 5$

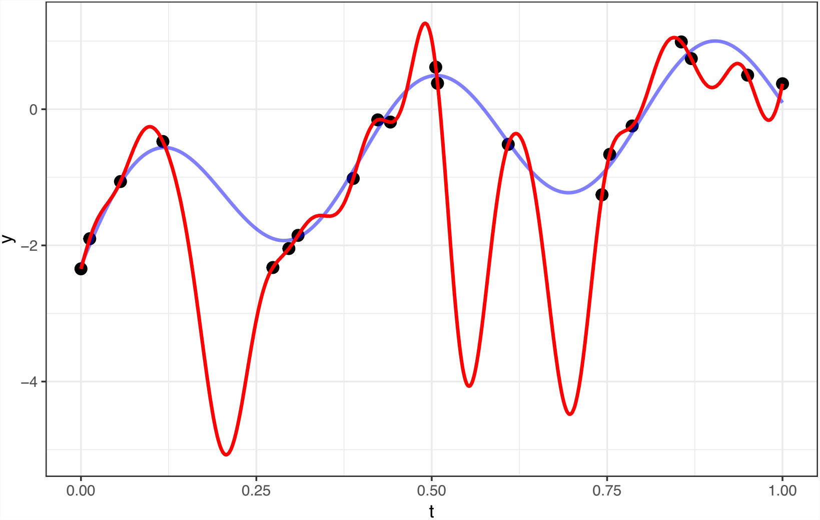
We therefore want to sample from $\mathbf{Y}_{pred} | \mathbf{Y}_{obs}$.

$$\mathbf{Y}_{pred} \mid \mathbf{Y}_{obs} = \mathbf{y} \sim \mathcal{N}(\boldsymbol{\Sigma}_{po}\, \boldsymbol{\Sigma}_{obs}^{-1}\, \mathbf{y},\ \boldsymbol{\Sigma}_{\mathbf{pred}} - \underset{n_p \times n}{\boldsymbol{\Sigma}_{po}}\, \underset{n \times n}{\boldsymbol{\Sigma}^{-1}_{\cancel{pred}\ obs}}\, \underset{n \times n_p}{\boldsymbol{\Sigma}_{op}})$$

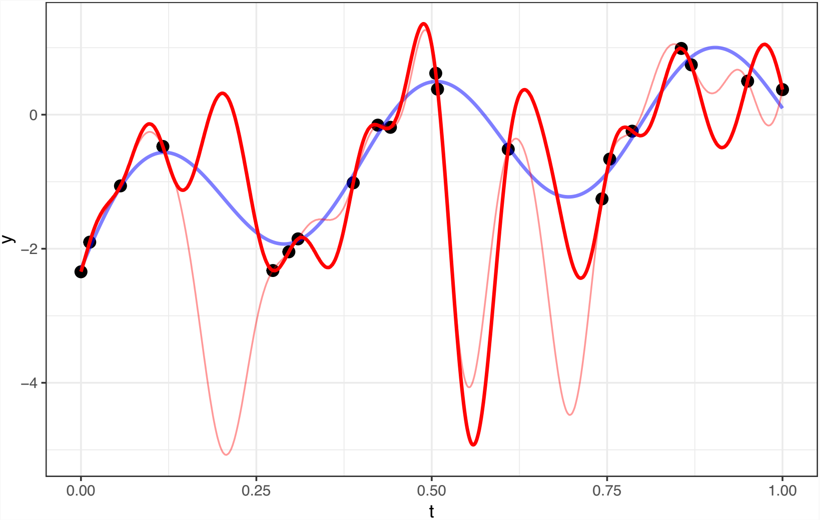
$$Y \sim N(0, \Sigma)$$

$$\{\Sigma\}_{ij} = 10 \exp\left(-\left(|t_i - t_j|/5\right)^2\right)$$

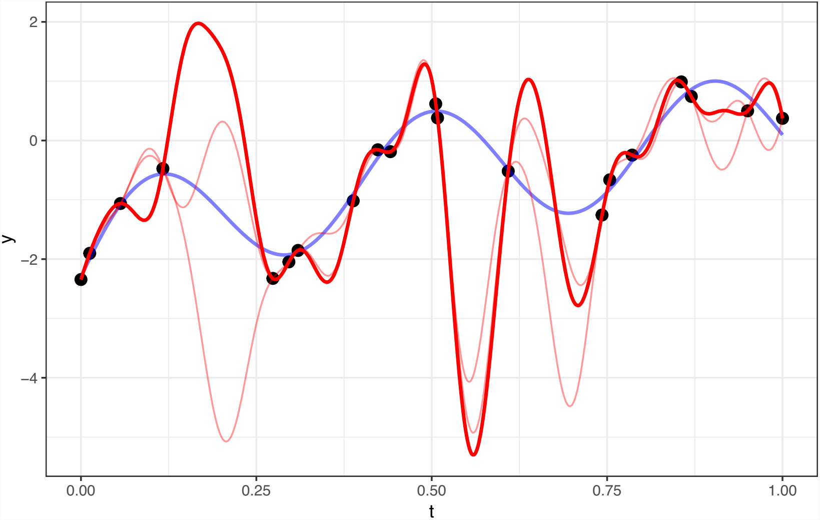
Draw 1



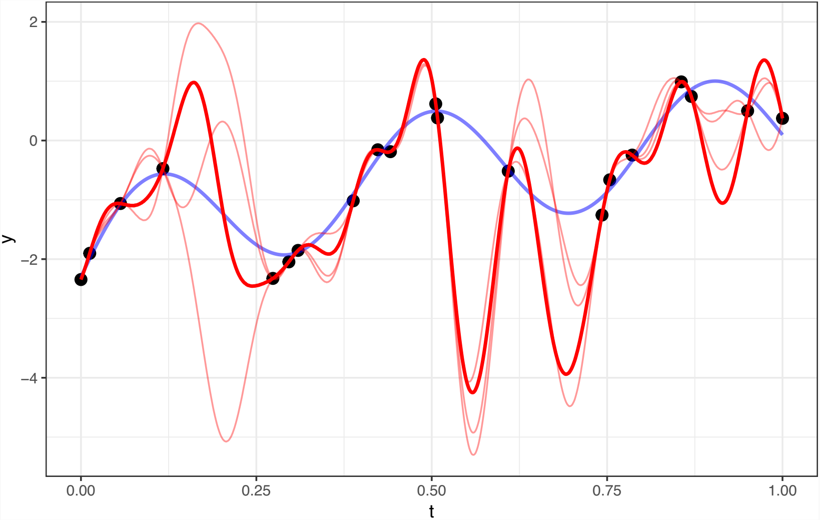
Draw 2



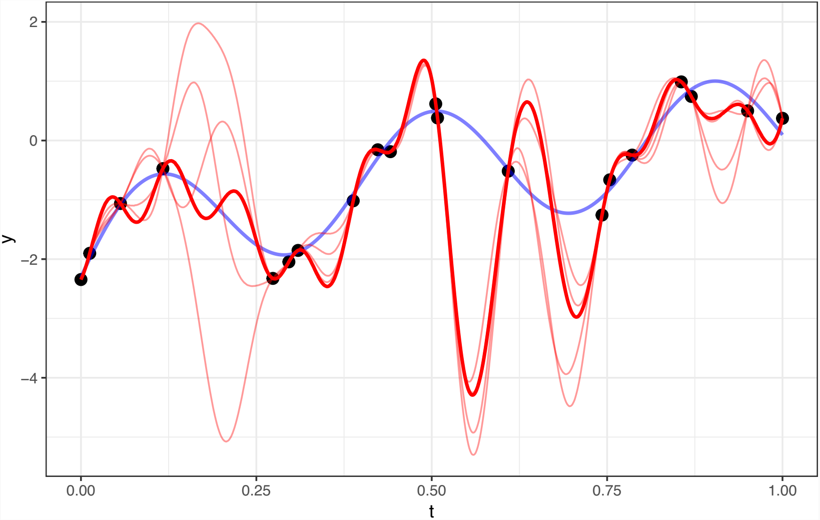
Draw 3



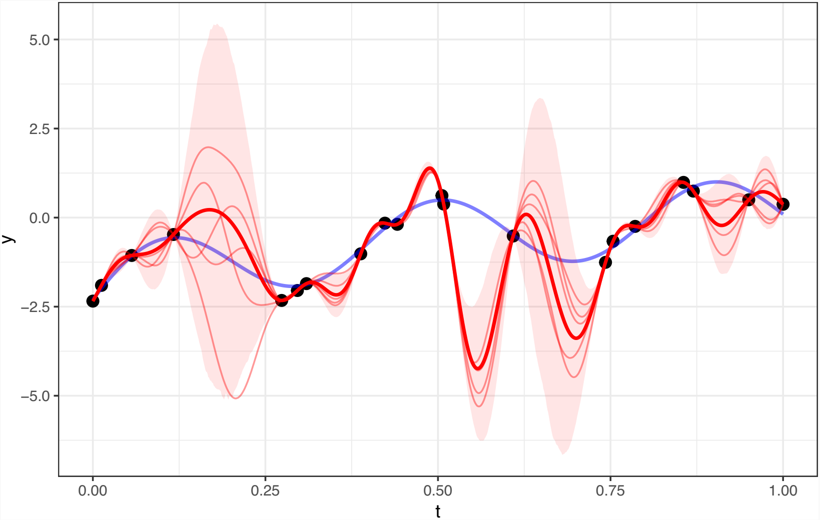
Draw 4



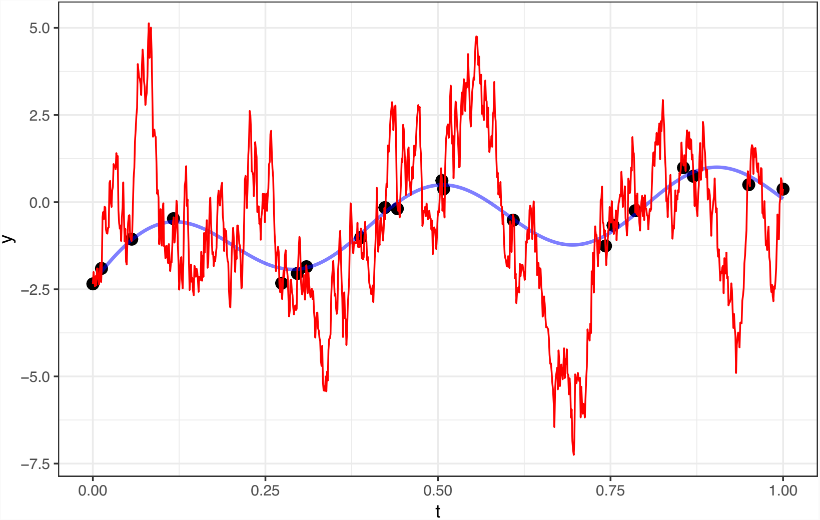
Draw 5



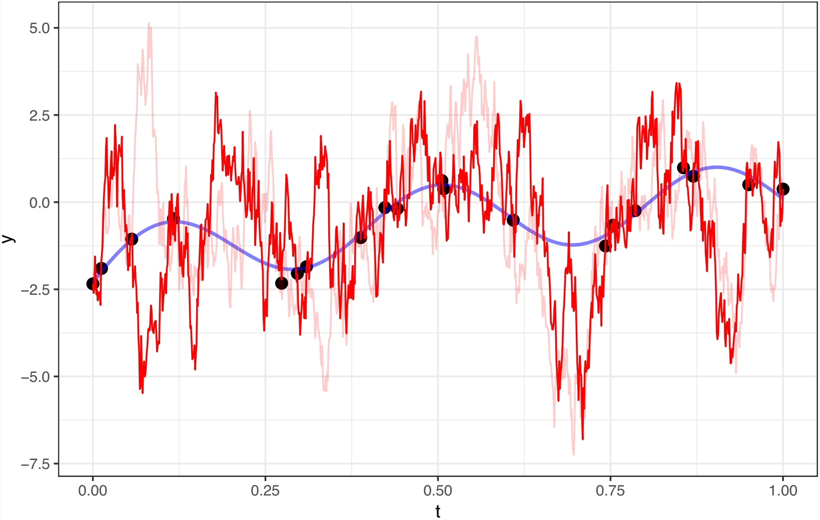
Many draws later



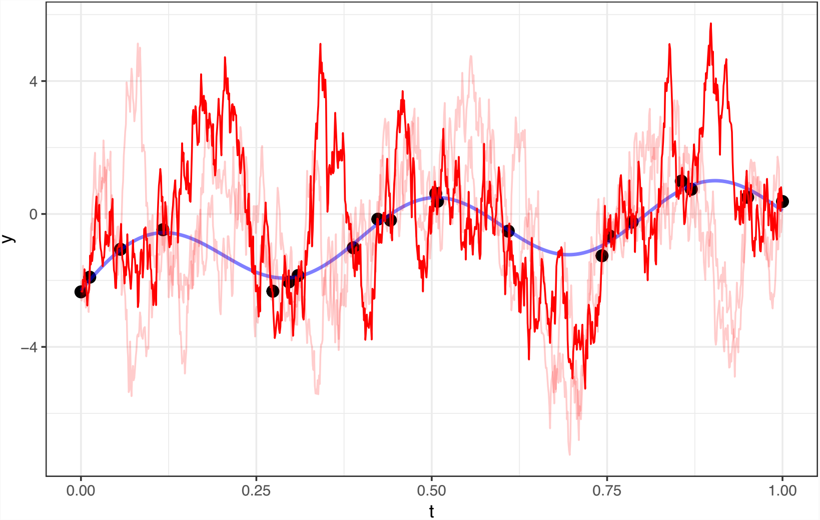
Exponential Covariance



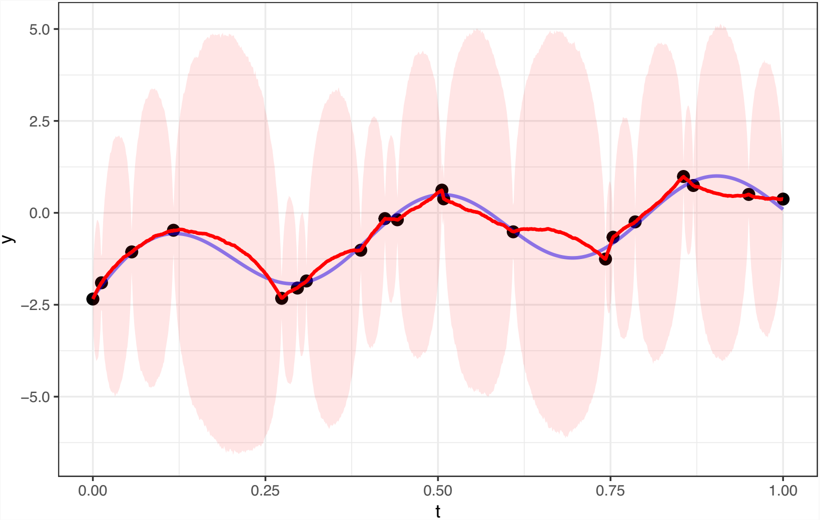
Exponential Covariance - Draw 2



Exponential Covariance - Draw 3

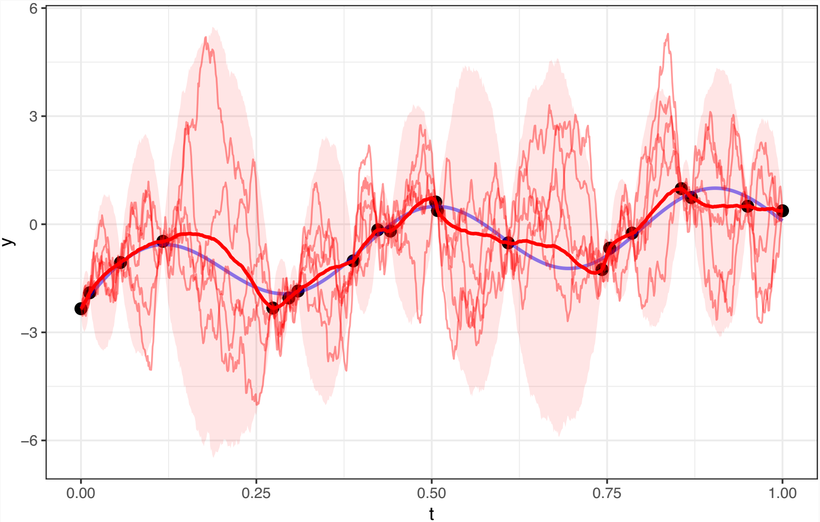


Exponential Covariance - Posterior

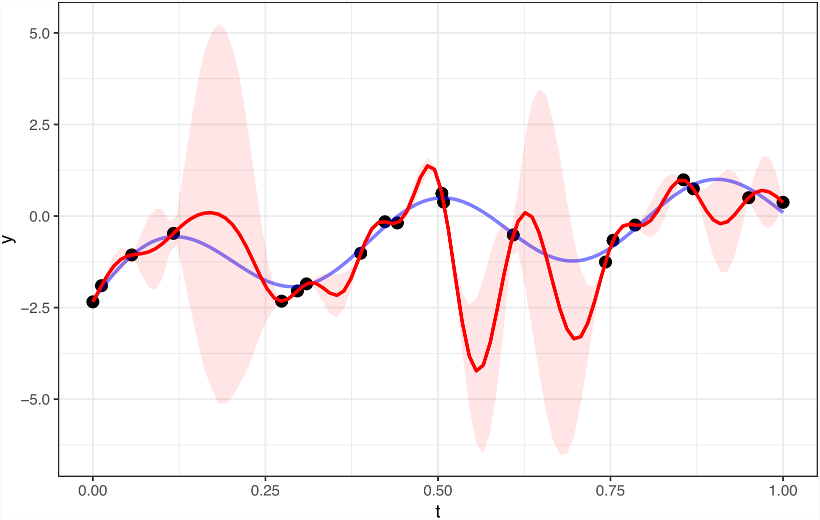


Powered Exponential Covariance ($p = 1.5$)

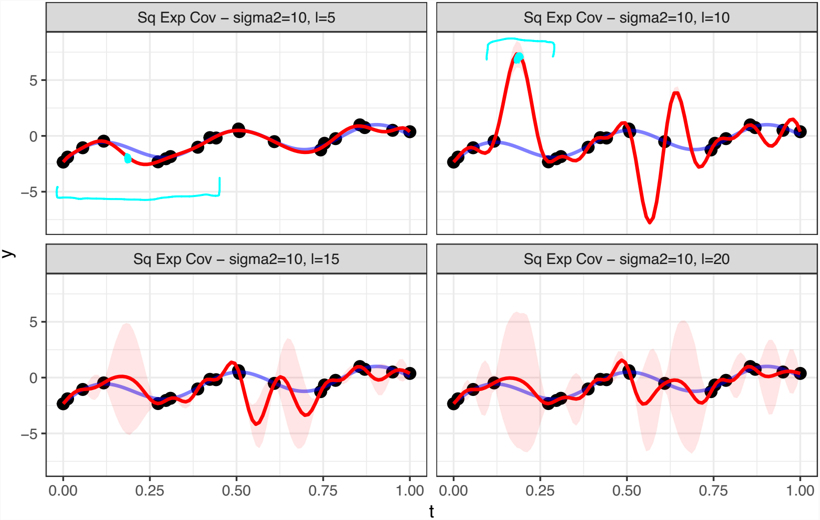
smoothness



Back to the square exponential



Changing the range (l)



For the square exponential covariance

$$Cov(d) = \sigma^2 \exp\left(-(l \cdot d)^2\right)$$
$$Corr(d) = \exp\left(-(l \cdot d)^2\right)$$

we would like to know, for a given value of l, beyond what distance apart must observations be to have a correlation less than 0.05?

$$Corr(d) = \exp(-(l \cdot d)^2) < 0.05$$

$$-(l \cdot d)^2 < \ln(0.05) = -3$$

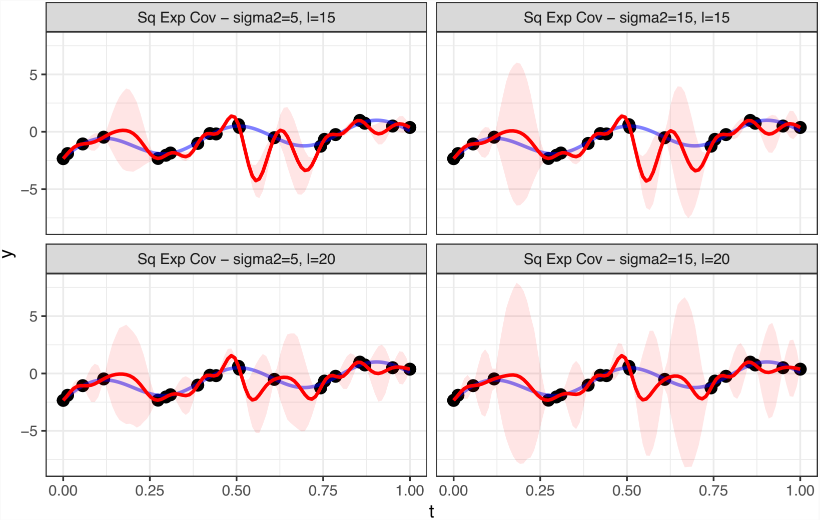
$$d < \frac{\sqrt{3}}{l}$$

Exp Cov

$$\rho(d) = \exp(-(l(d))) < 0.05$$

$$d < \frac{3}{l}$$

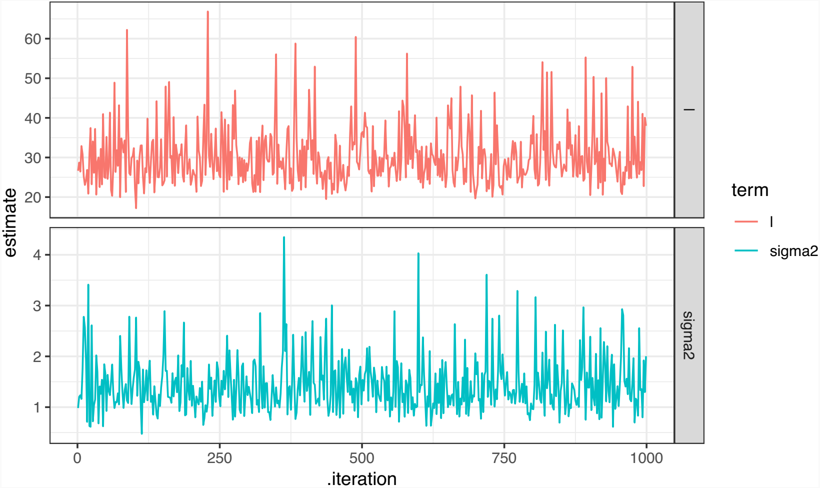
Changing the scale (σ^2)



Fitting

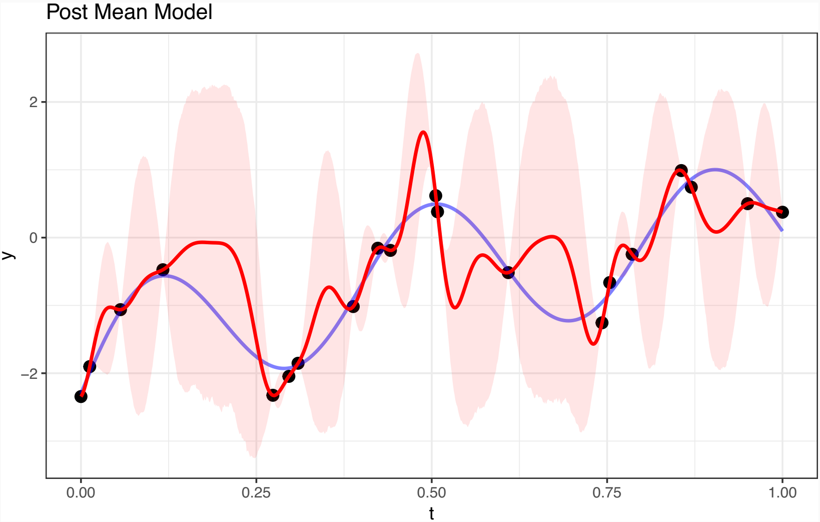
```
gp_sq_exp_model = "model{
  y ~ dmnorm(mu, inverse(Sigma))

  for (i in 1:N) {
    mu[i] <- 0
  }

  for (i in 1:(N-1)) {
    for (j in (i+1):N) {
      Sigma[i,j] <- sigma2 * exp(- pow(l*d[i,j],2))
      Sigma[j,i] <- Sigma[i,j]
    }
  }

  for (k in 1:N) {
    Sigma[k,k] <- sigma2 + 0.00001
  }

  sigma2   ~ dlnorm(0, 1.5)
  l        ~ dt(0, 2.5, 1) T(0,) # Half-cauchy(0,2.5)
}"
```

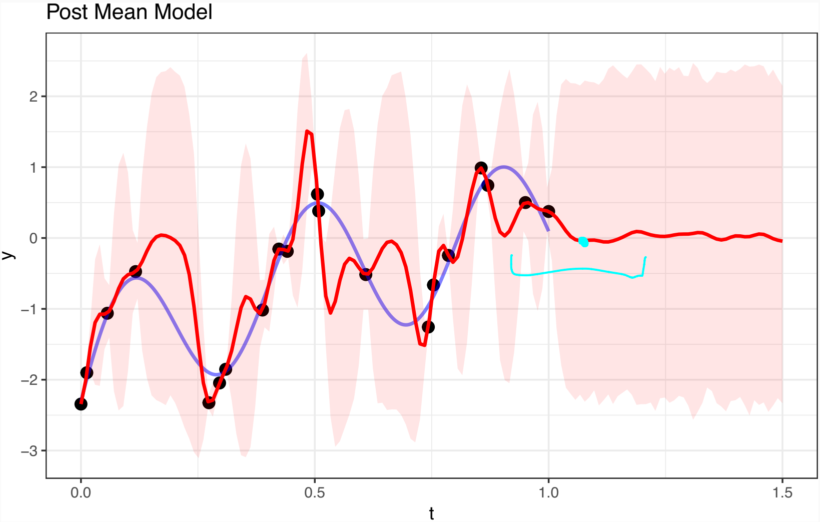
Trace plots



param	post_mean	post_med	post_lower	post_upper
l	30.20	28.70	20.63	51.51
sigma2	1.44	1.33	0.72	2.78

Fitted models




Post Mean Model
y
t
2
1
0
−1
−2
−3
0.0
0.5
1.0
1.5

Improving the model

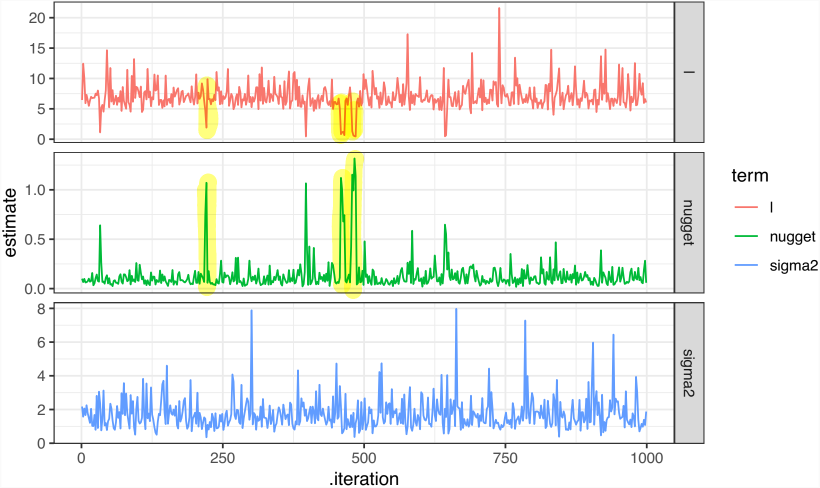
```
gp_sq_exp_model2 = "model{
  y ~ dmnorm(mu, inverse(Sigma))

  for (i in 1:N) {
    mu[i] <- 0
  }

  for (i in 1:(N-1)) {
    for (j in (i+1):N) {
      Sigma[i,j] <- sigma2 * exp(- pow(l*d[i,j],2))
      Sigma[j,i] <- Sigma[i,j]
    }
  }

  for (k in 1:N) {
    Sigma[k,k] <- sigma2 + nugget
  }

  sigma2   ~ dlnorm(0, 1.5)
  l        ~ dt(0, 2.5, 1) T(0,) # Half-cauchy(0,2.5)

  nugget ~ dlnorm(0, 1)
}"
```

$$y(t) = w(t) + \epsilon(t)$$

$$w(t) \sim MVN(0, \Sigma)$$

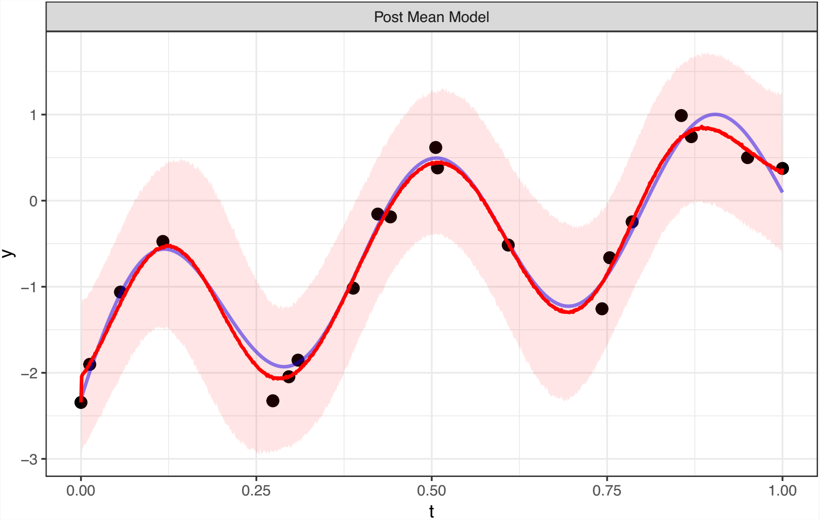
$$\epsilon(t) \overset{iid}{\sim} N(0, \sigma_w^2)$$

Trace plots



param	post_mean	post_med	post_lower	post_upper
l	7.01	6.75	2.17	11.79
nugget	0.13	0.09	0.03	0.57
sigma2	1.73	1.53	0.64	4.04

Fitted models



Forcasting