Gaussian Process Models

Lecture 13

Dr. Colin Rundel

Multivariate Normal

Multivariate Normal Distribution

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

$$\underset{n\times 1}{\boldsymbol{y}} \sim \mathrm{N}(\underset{n\times 1}{\boldsymbol{\mu}}, \underset{n\times n}{\boldsymbol{\Sigma}})$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, \begin{pmatrix} \rho_{11}\sigma_1\sigma_1 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \cdots & \rho_{nn}\sigma_n\sigma_n \end{pmatrix}\right)$$

Density

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

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

and its log density is given by

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

Sampling

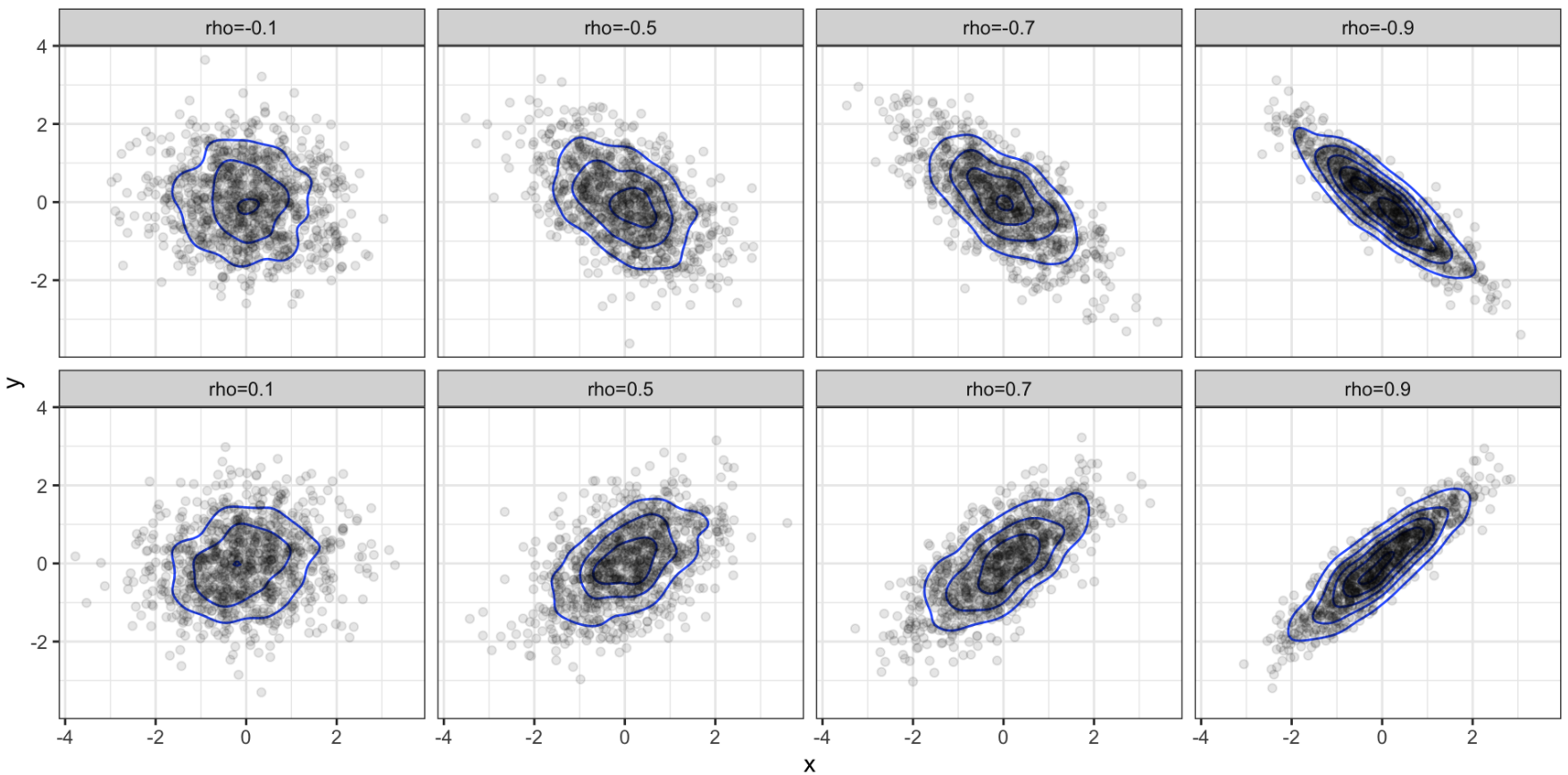
To generate draws from an n-dimensional multivate normal with mean $\underset{n\times 1}{\boldsymbol{\mu}}$ and covariance matrix $\underset{n\times n}{\boldsymbol{\Sigma}}$,

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

$$\underset{n\times 1}{\boldsymbol{y}} = \underset{n\times 1}{\boldsymbol{\mu}} + \underset{n\times n}{\boldsymbol{A}}\ \underset{n\times 1}{\boldsymbol{z}}$$

Bivariate Examples

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



Marginal / conditional 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.

Univariate marginal distribution:

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

Bivariate marginal distribution:

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

k-*dimensional marginal distribution*:

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

$\mathrm{i}, \ldots, \mathrm{k}$ do not need to be ordered in any particular way

Conditional Distributions

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

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

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

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

then the conditional distributions are given by

$$\boldsymbol{y}_1 \mid \boldsymbol{y}_2 = \boldsymbol{a} \sim N(\boldsymbol{\mu_1} + \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}(\boldsymbol{a} - \boldsymbol{\mu_2}), \boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}\boldsymbol{\Sigma_{21}})$$

$$\boldsymbol{y}_2 \mid \boldsymbol{y}_1 = \boldsymbol{b} \sim N(\boldsymbol{\mu_2} + \boldsymbol{\Sigma_{21}}\boldsymbol{\Sigma_{11}^{-1}}(\boldsymbol{b} - \boldsymbol{\mu_1}), \boldsymbol{\Sigma_{22}} - \boldsymbol{\Sigma_{21}}\boldsymbol{\Sigma_{11}^{-1}}\boldsymbol{\Sigma_{12}})$$

Gaussian Processes

From Shumway,

> A process, $\boldsymbol{y} = \{y(t) \ : \ t \in T\}$, is said to be a Gaussian process if all possible finite dimensional vectors $\boldsymbol{y} = (y_{t_1}, y_{t_2}, \ldots, y_{t_n})^t$, 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

Parameterizing a Gaussian Process

Imagine we have a Gaussian process defined such that

$$\boldsymbol{y} = \{\mathrm{y(t)} \ : \ \mathrm{t} \in [0, 1]\},$$

- We now have an uncountably infinite set of possible t's and $\mathrm{y(t)}$s.
- We will only have a (small) finite number of observations $\mathrm{y}(\mathrm{t}_1), \ldots, \mathrm{y}(\mathrm{t}_n)$ with which to say something useful about this infinite dimensional process.
- The unconstrained covariance matrix for the observed data can have up to $\mathrm{n}(\mathrm{n}+1)/2$ unique values*
- Necessary to make some simplifying assumptions:
  - Stationarity
  - Simple(r) parameterization of Σ

Covariance Functions

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

Exponential covariance:

$$\Sigma\left(y(t), y(t')\right) = \sigma^2 \exp\left(-|t - t'|\,1\right)$$

Squared exponential covariance (Gaussian):

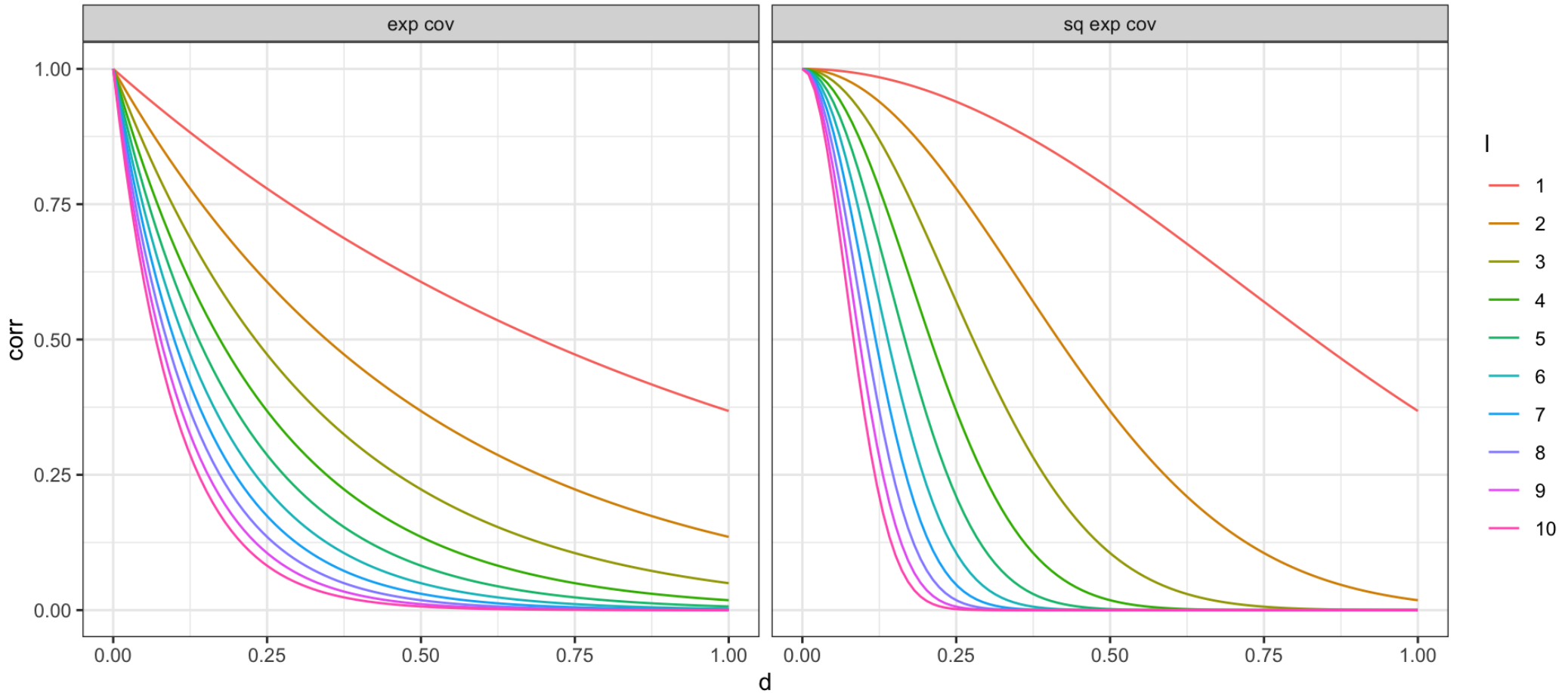
$$\Sigma\left(y(t), y(t')\right) = \sigma^2 \exp\left(-\left(|t - t'|\,1\right)^2\right)$$

Powered exponential covariance $(p \in (0, 2])$:

$$\Sigma\left(y(t), y(t')\right) = \sigma^2 \exp\left(-\left(|t - t'|\,1\right)^p\right)$$

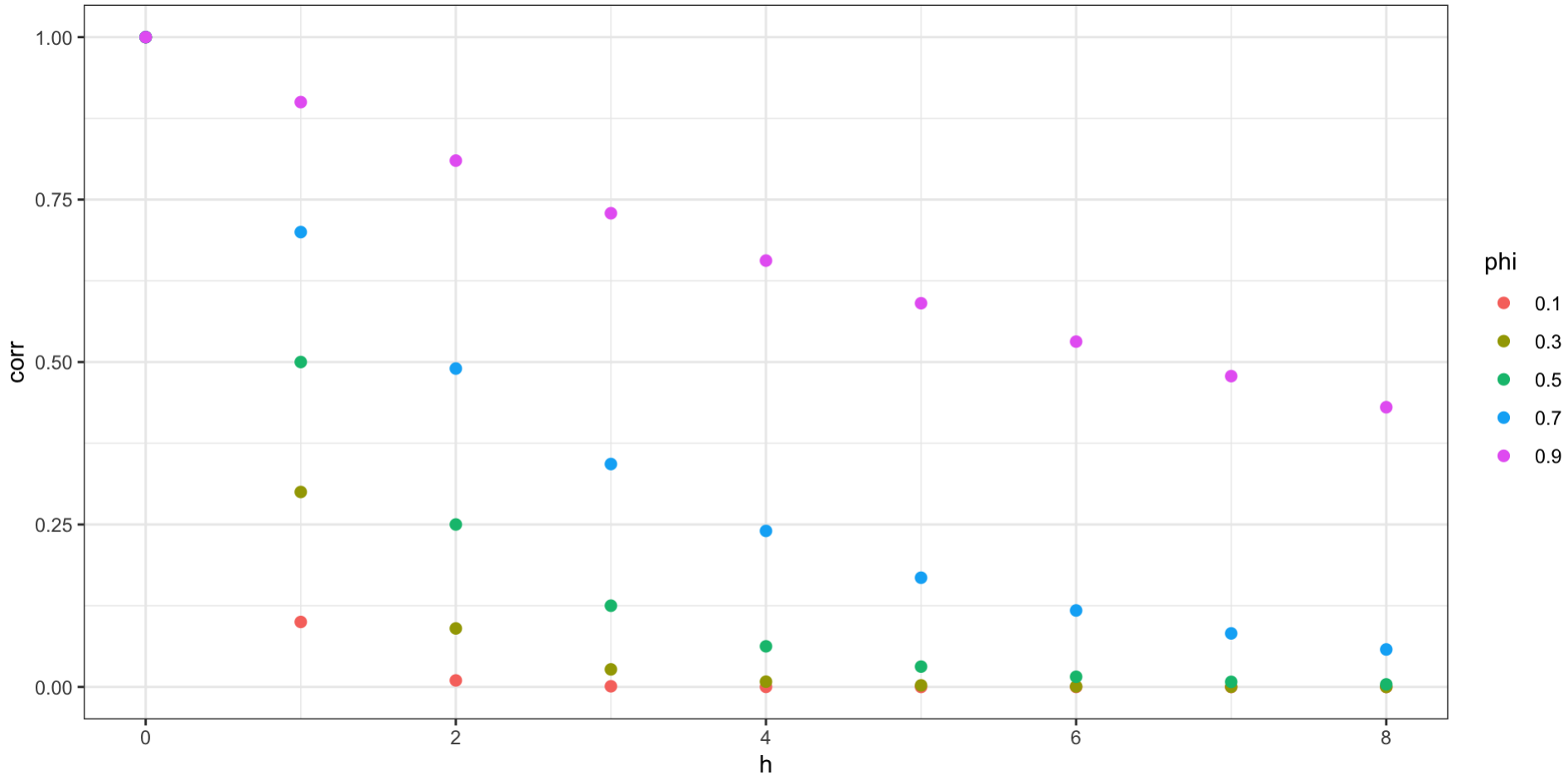
Correlation Decay

Letting $\sigma^2 = 1$ and trying different values of the inverse lengthscale l,

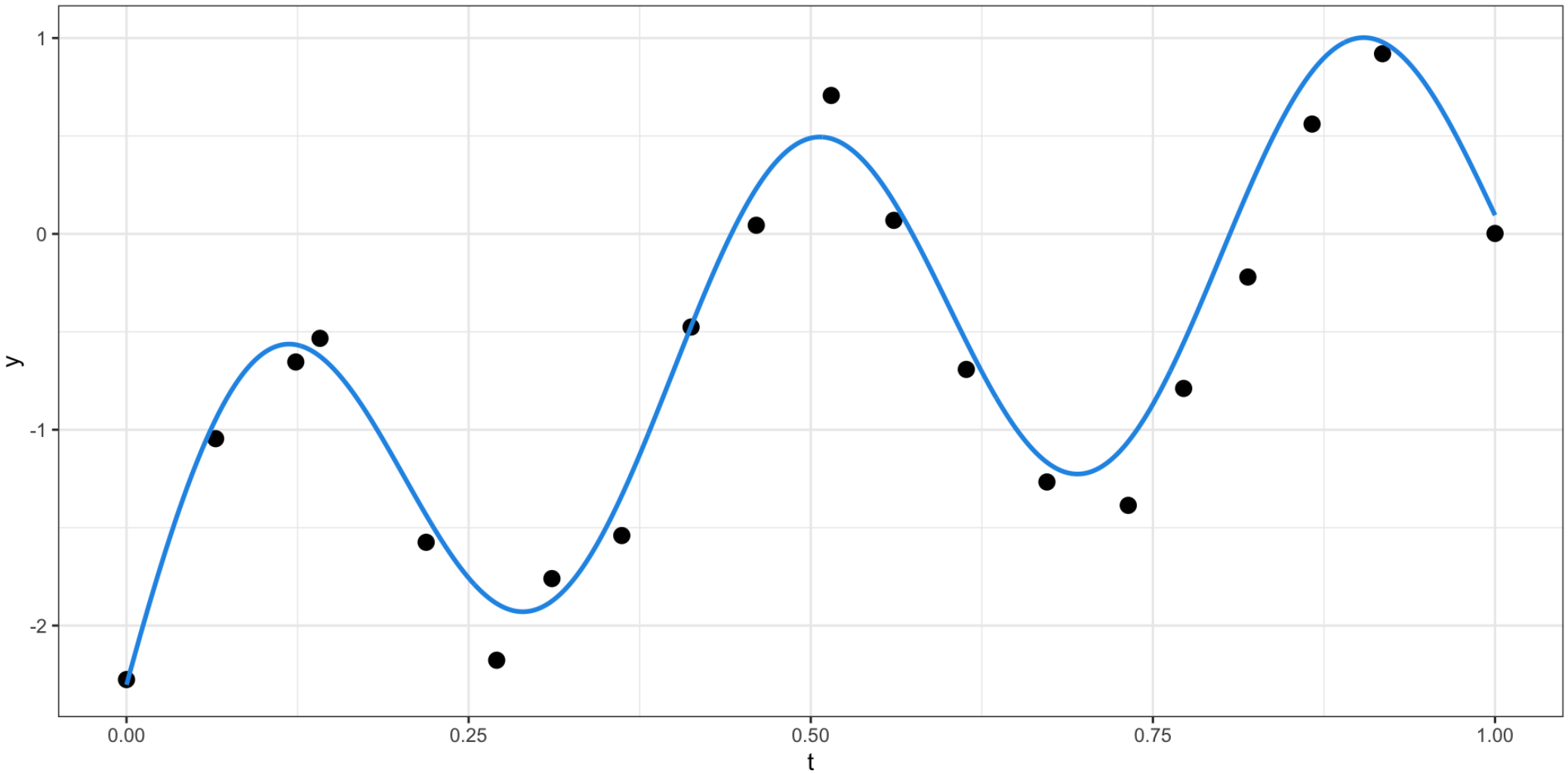


Correlation Decay - AR(1)

Recall that for a stationary AR(1) process: $\gamma(h) = \sigma_w^2 \phi^{|h|}$ and $\rho(h) = \phi^{|h|}$ we can draw a similar picture of the decay of correlation as a function of "distance".


1.00
0.75
0.50
0.25
0.00
corr
0
2
4
6
8
h
phi
0.1
0.3
0.5
0.7
0.9

GP Example



Prediction

Our example has 20 observations $\left(\boldsymbol{y}_{\text{obs}} = (\text{y}(\text{t}_1), \ldots, \text{y}(\text{t}_{20}))'\right)$, which we would like to use as the basis for predicting $\text{y}(\text{t})$ at other values of t $\left(\boldsymbol{y}_{\text{pred}}\right)$ at a regular sequence of values from 0 to 1.

For now lets use a squared exponential covariance with $\sigma^2 = 10$ and $\text{l} = 15$, as such the covariance is given by:

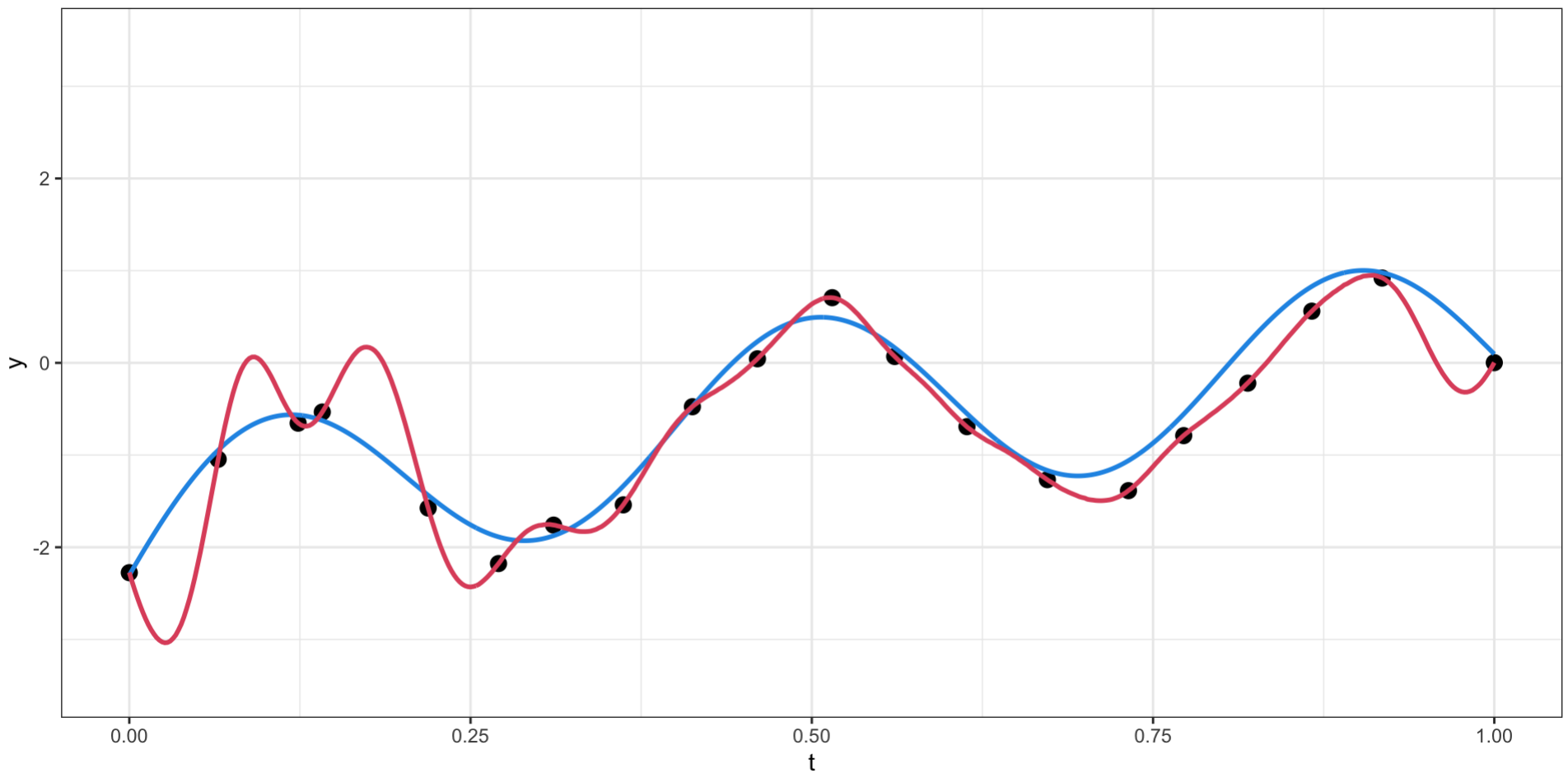
$$\Sigma\left(\text{y}(\text{t}), \text{y}(\text{t}')\right) = \sigma^2 \exp\left(-|\text{t} - \text{t}'|\,\text{l}\right)$$

Our goal is to obtain samples from $\boldsymbol{y}_{\text{pred}}|\boldsymbol{y}_{\text{obs}}$ which has the following distribution,

$$\boldsymbol{y}_{\text{pred}} \mid \boldsymbol{y}_{\text{obs}} = \boldsymbol{y} \sim \text{N}(\boldsymbol{\Sigma}_{\text{po}}\,\boldsymbol{\Sigma}_{\text{obs}}^{-1}\,\boldsymbol{y},\ \boldsymbol{\Sigma}_{\text{pred}} - \boldsymbol{\Sigma}_{\text{po}}\,\boldsymbol{\Sigma}_{\text{pred}}^{-1}\,\boldsymbol{\Sigma}_{\text{op}})$$

Squared exponential covariance

Draw 1 | Draw 2 | Draw 3 | Draw 4 | Draw 5 | Mean | CI



Exponential covariance

Now lets consider an exponential covariance model instead where $\sigma = 10$, $l = \sqrt{15}$,

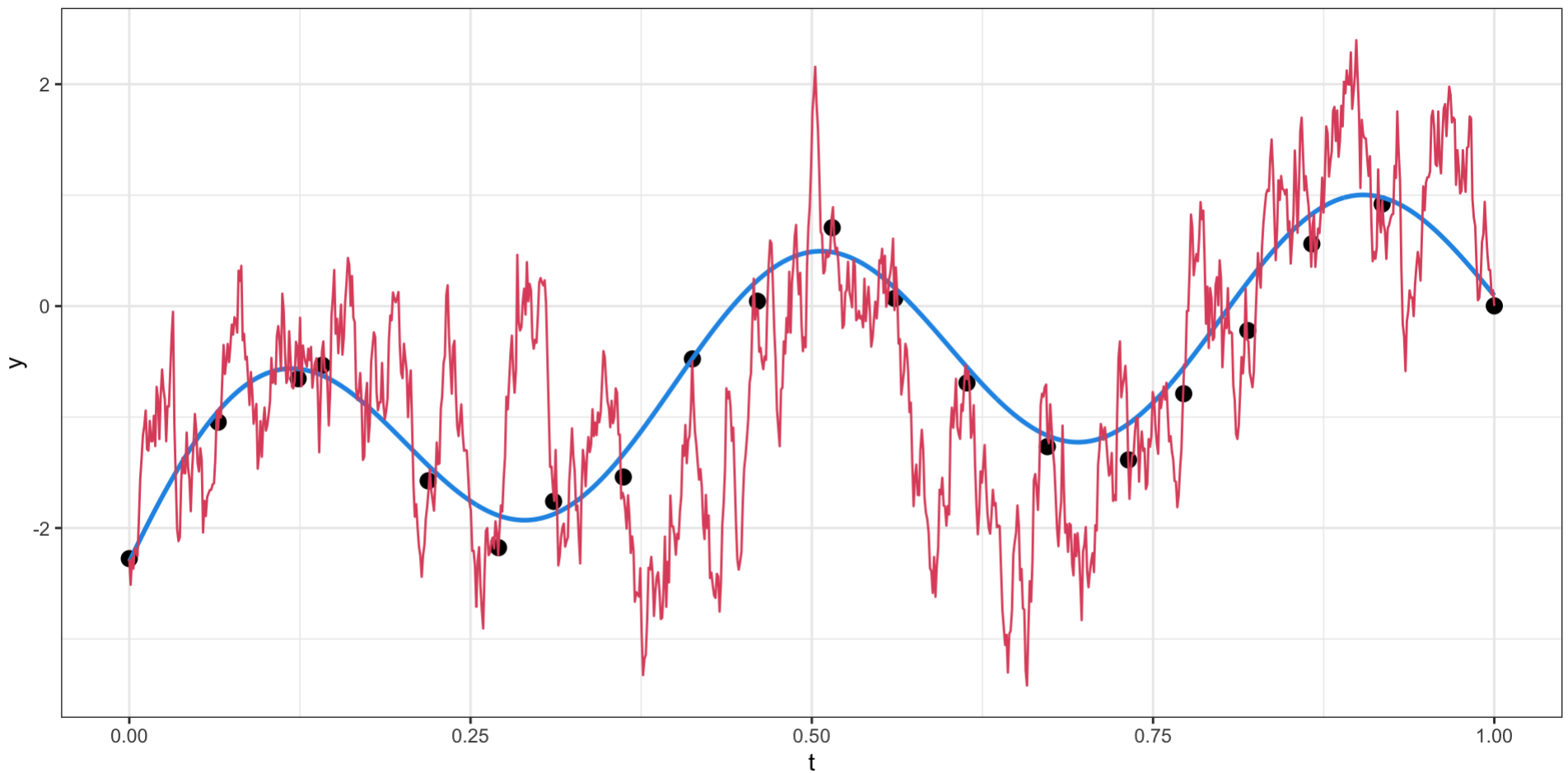
$$\Sigma(y_t, y_{t'}) = \sigma^2 \exp\left(-|t - t'|\, l\right)$$

We are still sampling from $\boldsymbol{y}_{\text{pred}} \mid \boldsymbol{y}_{\text{obs}}$, all that has changed is the values of the covariance matrices.

What "paths" do we get with this covariance? How are they similar and how are they different from the squared exponential covariance?

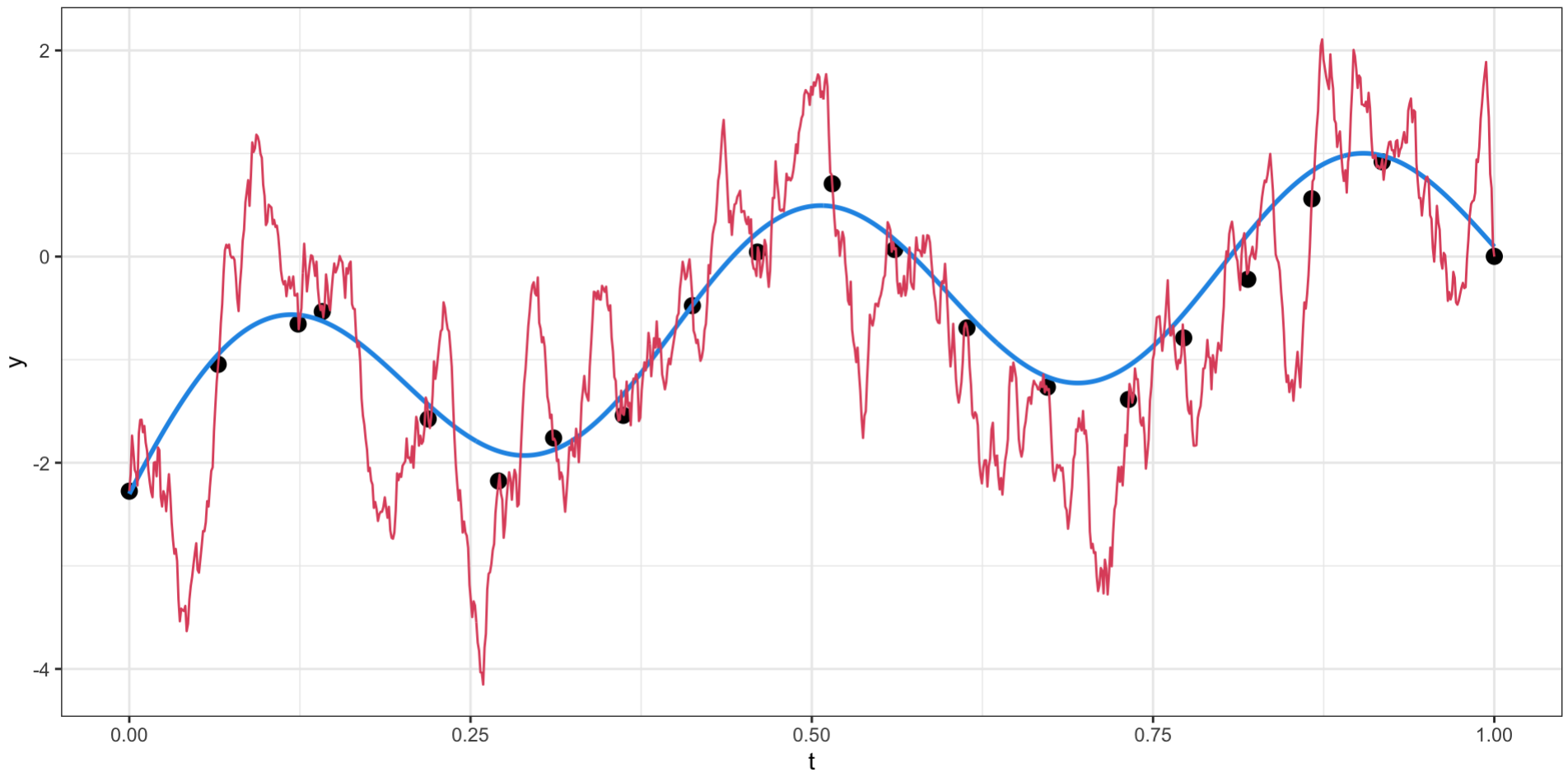
Exponential Covariance

Draw 1 | Draw 2 | Draw 3 | Draw 4 | Draw 5 | Mean | CI

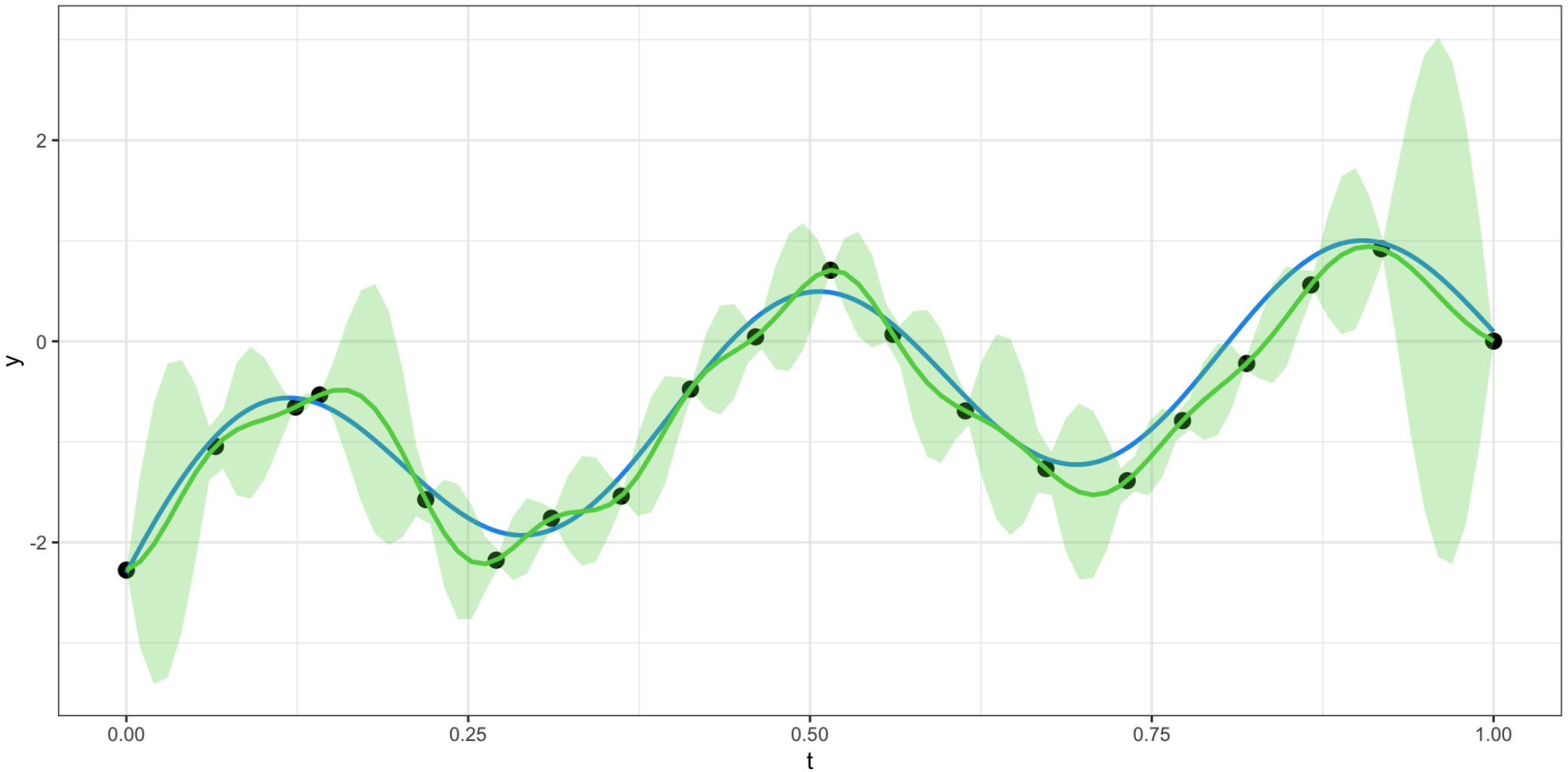


Powered exponential covariance ($p = 1.5$)

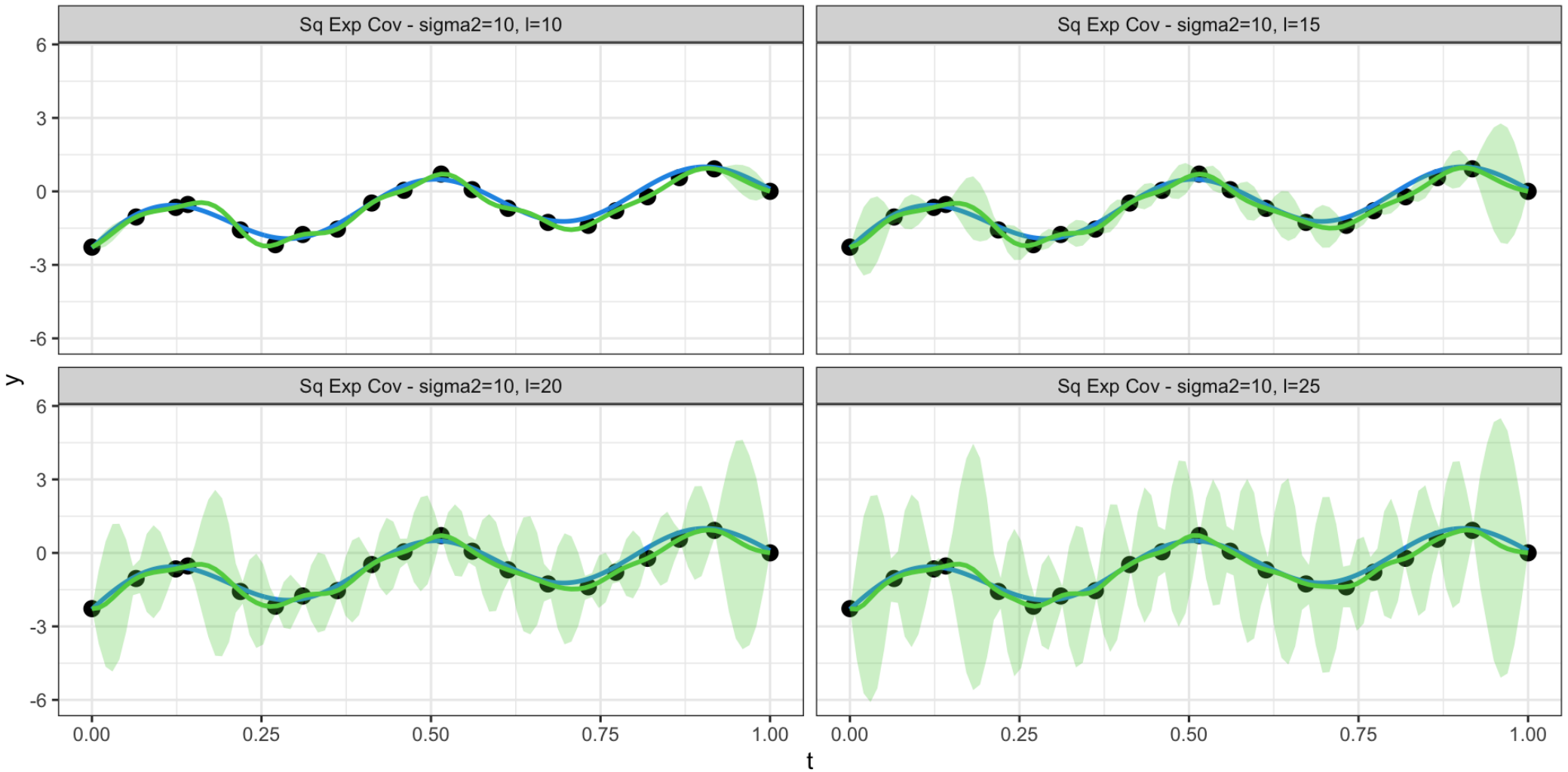
Draw 1 | Draw 2 | Draw 3 | Draw 4 | Draw 5 | Mean | CI



Back to the square exponential



Changing the inverse lengthscale (l)



Effective range

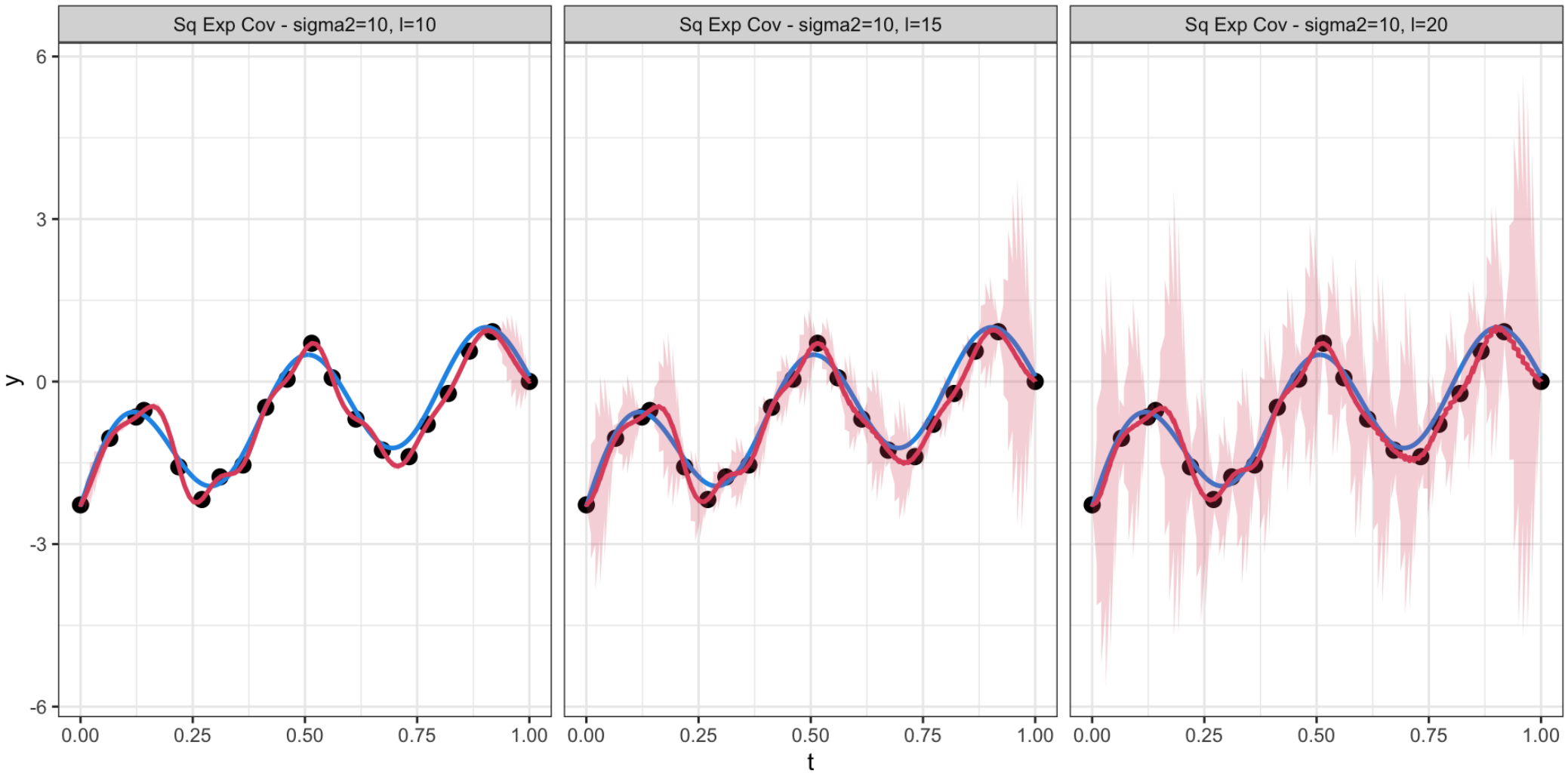
For the square exponential covariance

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

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

$$\begin{aligned}
\exp\left(-(l \cdot d)^2\right) &< 0.05 \\
-(l \cdot d)^2 &< \log 0.05 \\
-(l \cdot d)^2 &< 3 \\
l \cdot d &< \sqrt{3} \\
d &< \sqrt{3}/l
\end{aligned}$$

Changing the scale (σ^2)



Fitting w/ BRMS

```r
library(brms)
gp = brm(y ~ gp(t), data=d, cores=4, refresh=0)
summary(gp)
```

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: y ~ gp(t)
   Data: d (Number of observations: 20)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Gaussian Process Terms:
            Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sdgp(gpt)       1.71      0.70     0.83     3.57 1.01      467      623
lscale(gpt)     0.13      0.03     0.08     0.18 1.02      316      570

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    -1.06      0.86    -2.96     0.49 1.01     1167      631

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.18      0.05     0.11     0.31 1.00     1510     1835

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
```

Some notes

`gp()` serves a similar function to `arma()` within a brms model - it specifies the structure of the model which is then translated into stan.

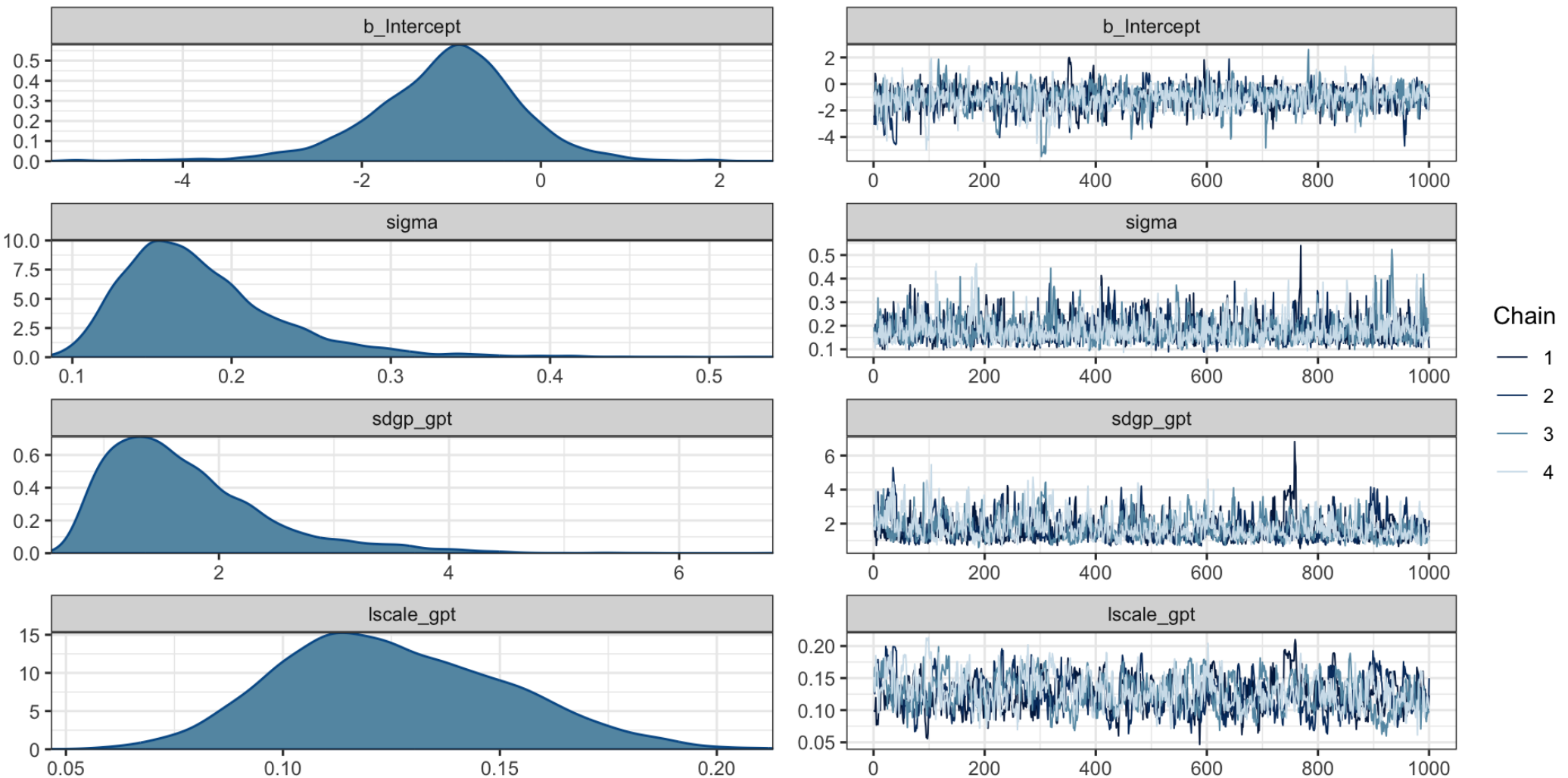
- The covariance function is specified using the `cov` argument to `gp()`, currently `"exp_quad"` is the default and *only* supported covariance.
- The covariance parameterization differs slightly from what was given previous (but is equivalent),

$$k(x_i, x_j) = \{\text{sdgp}\}^2 \exp(-||x_i - x_j||^2/(2\ \{\text{lscale}\}^2))$$

- Model fitting scales very poorly in terms of n (this is a stan issue and not just a brms issue)

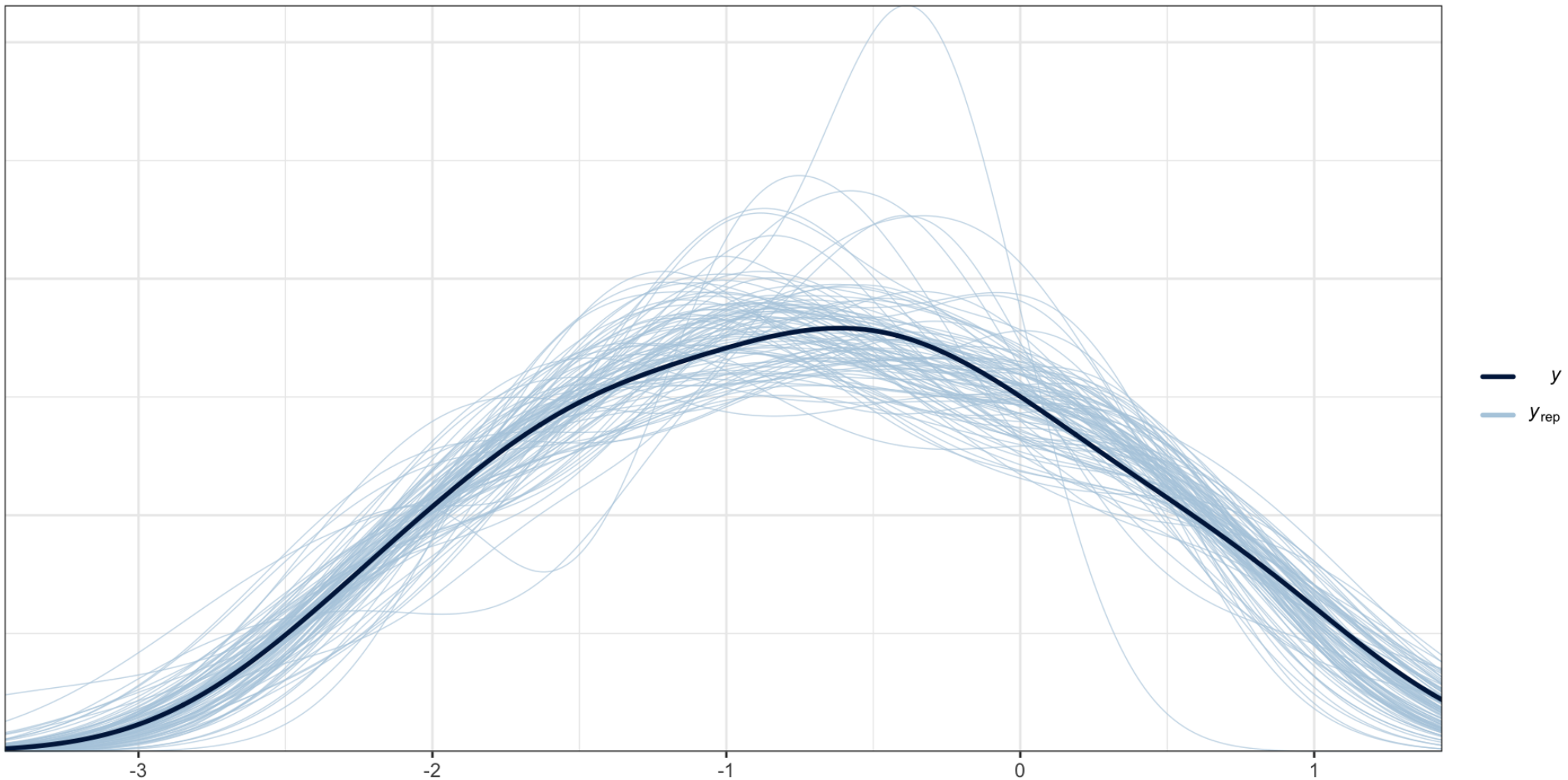
Trace plots

```
1 plot(gp)
```

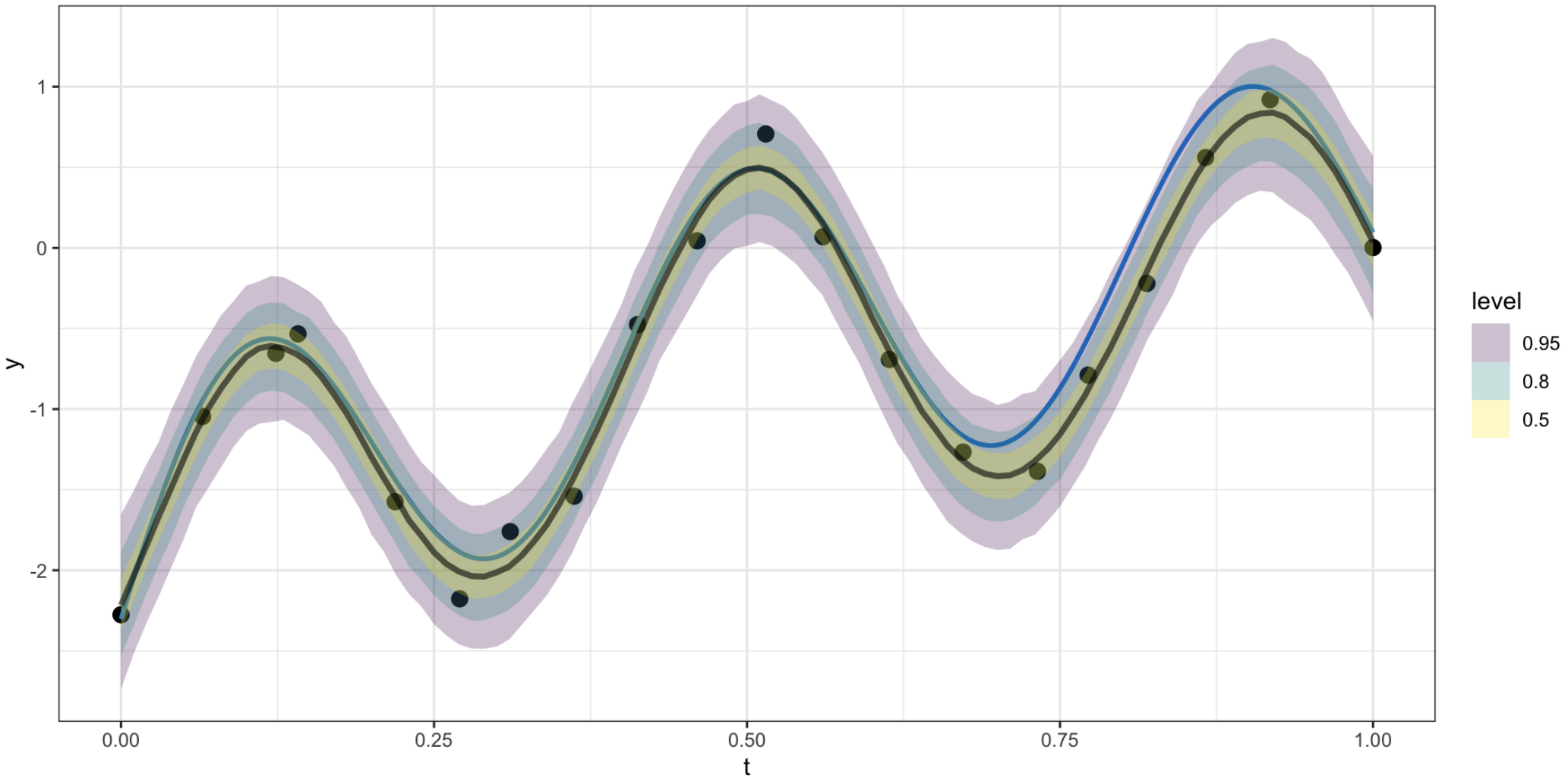


PP Checks

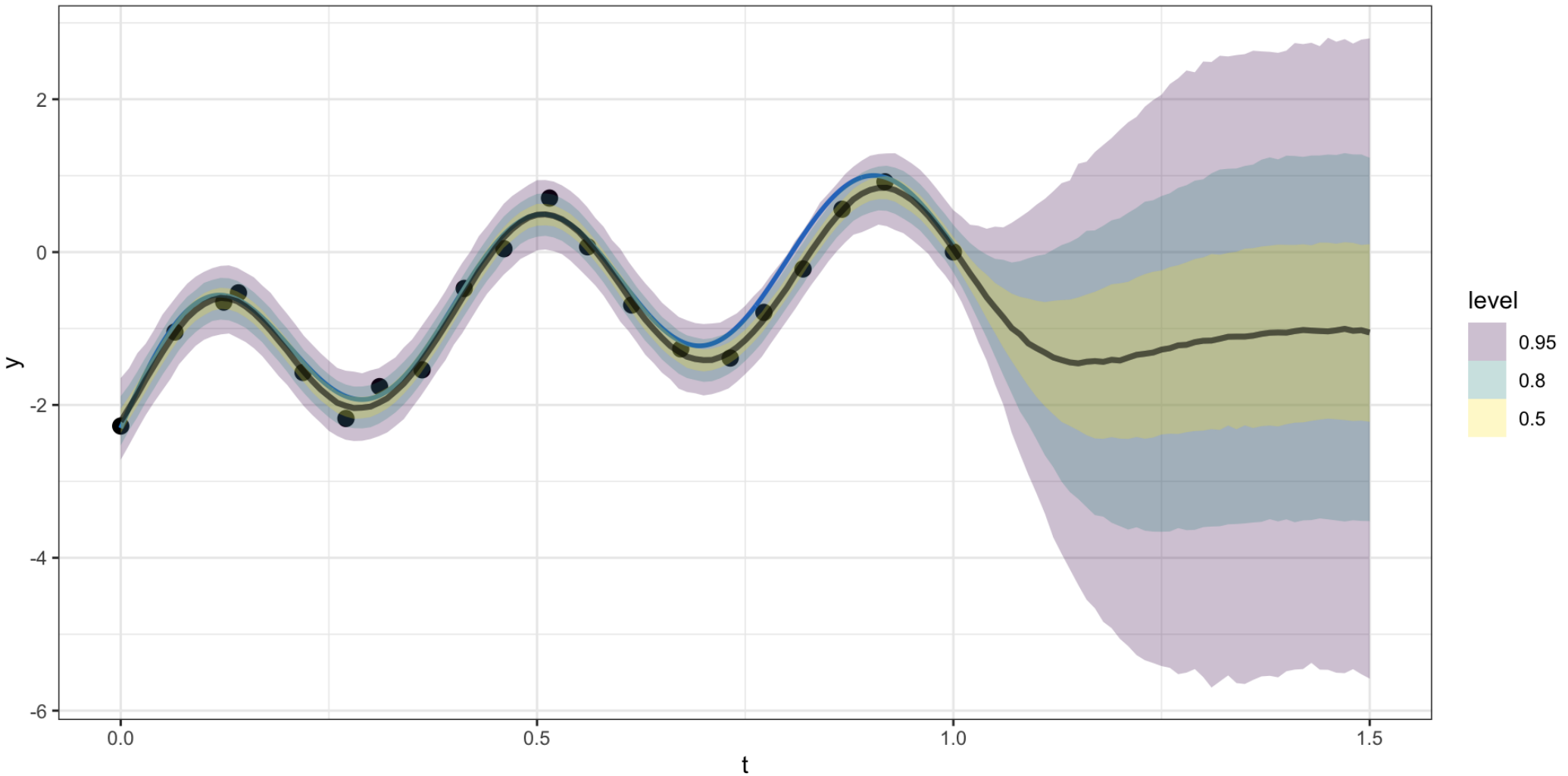
```
1 pp_check(gp, ndraws = 100)
```



Model predictions

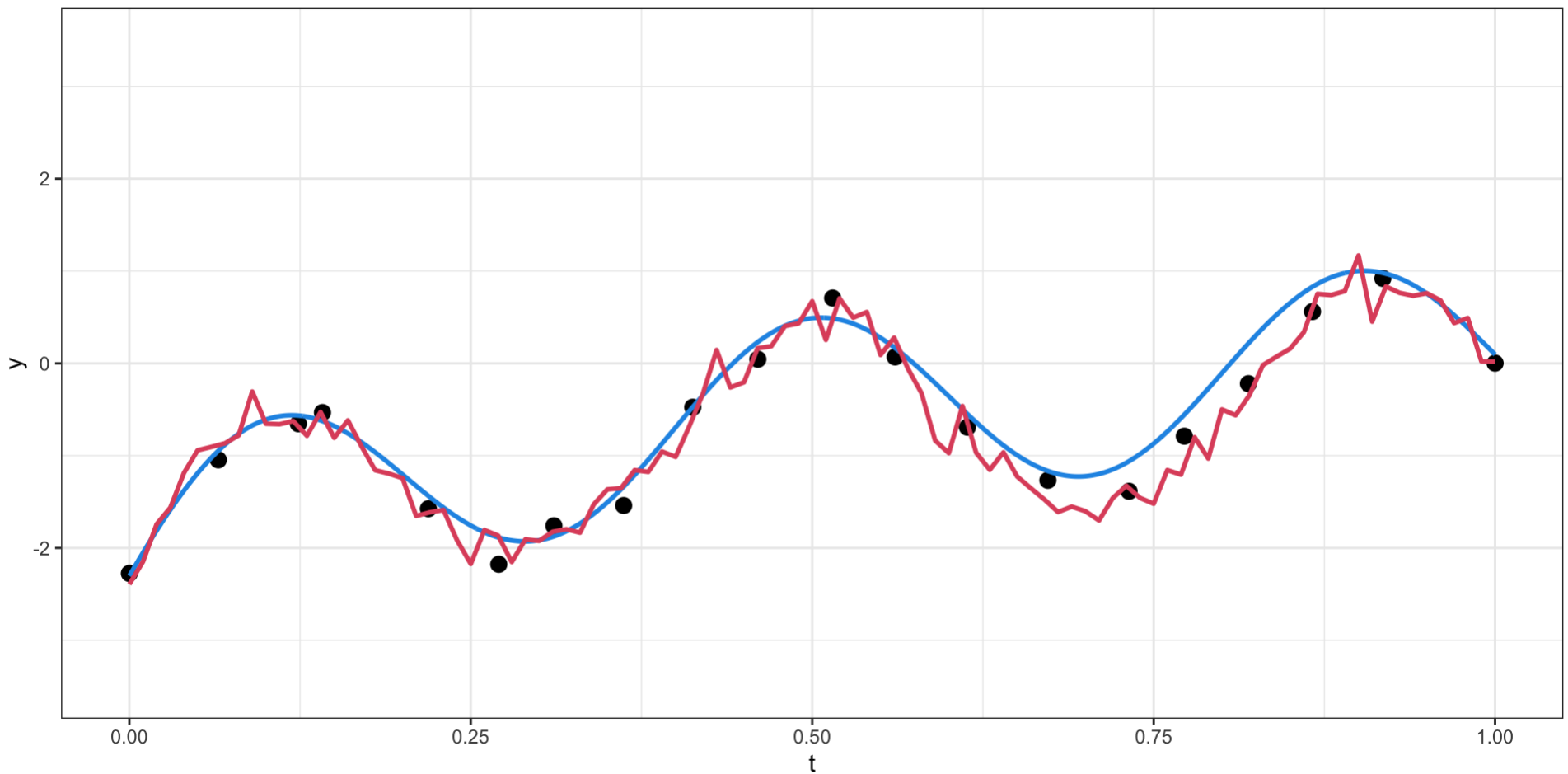


Forecasting



brms paths

Draw 1 | Draw 2 | Draw 3 | Draw 4 | Draw 5 | Mean | CI



Why does this look different?

Unlike our previous conditional samples these predictions no longer pass exactly through each observation and so we get a bit of a better behaved curve and much more reasonable intervals.

The model being fit by brms has one additional parameter that our previous example did not have - a nugget covariance (`sigma` in the brms summary).

So what does the covariance look like?

$$k(x_i, x_j) = \{\text{sdgp}\}^2 \exp(-||x_i - x_j||^2/(2\ \{\text{lscale}\}^2)) + \{\text{sigma}\}^2 \mathbb{1}_{i=j}$$

Squared exponential w/ nugget ($\sigma_w^2 = 0.1$)

Draw 1 | Draw 2 | Draw 3 | Draw 4 | Draw 5 | Mean | CI