

# Gaussian Process Models

Lecture 14

Dr. Colin Rundel

# Multivariate Normal

# Multivariate Normal Distribution

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

$$\begin{matrix} \mathbf{y} \\ n \times 1 \end{matrix} \sim N\left(\begin{matrix} \boldsymbol{\mu} \\ n \times 1 \end{matrix}, \begin{matrix} \boldsymbol{\Sigma} \\ n \times n \end{matrix}\right)$$

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

# Density

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

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

and its log density is given by

$$-\frac{n}{2} \log 2\pi - \frac{1}{2} \log \det(\Sigma) - \frac{1}{2} \begin{matrix} \mathbf{y} - \boldsymbol{\mu} \\ 1 \times n \end{matrix}' \begin{matrix} \Sigma^{-1} \\ n \times n \end{matrix} \begin{matrix} \mathbf{y} - \boldsymbol{\mu} \\ n \times 1 \end{matrix}$$

# Sampling

To generate draws from an  $n$ -dimensional multivariate normal with mean  $\mu_{n \times 1}$

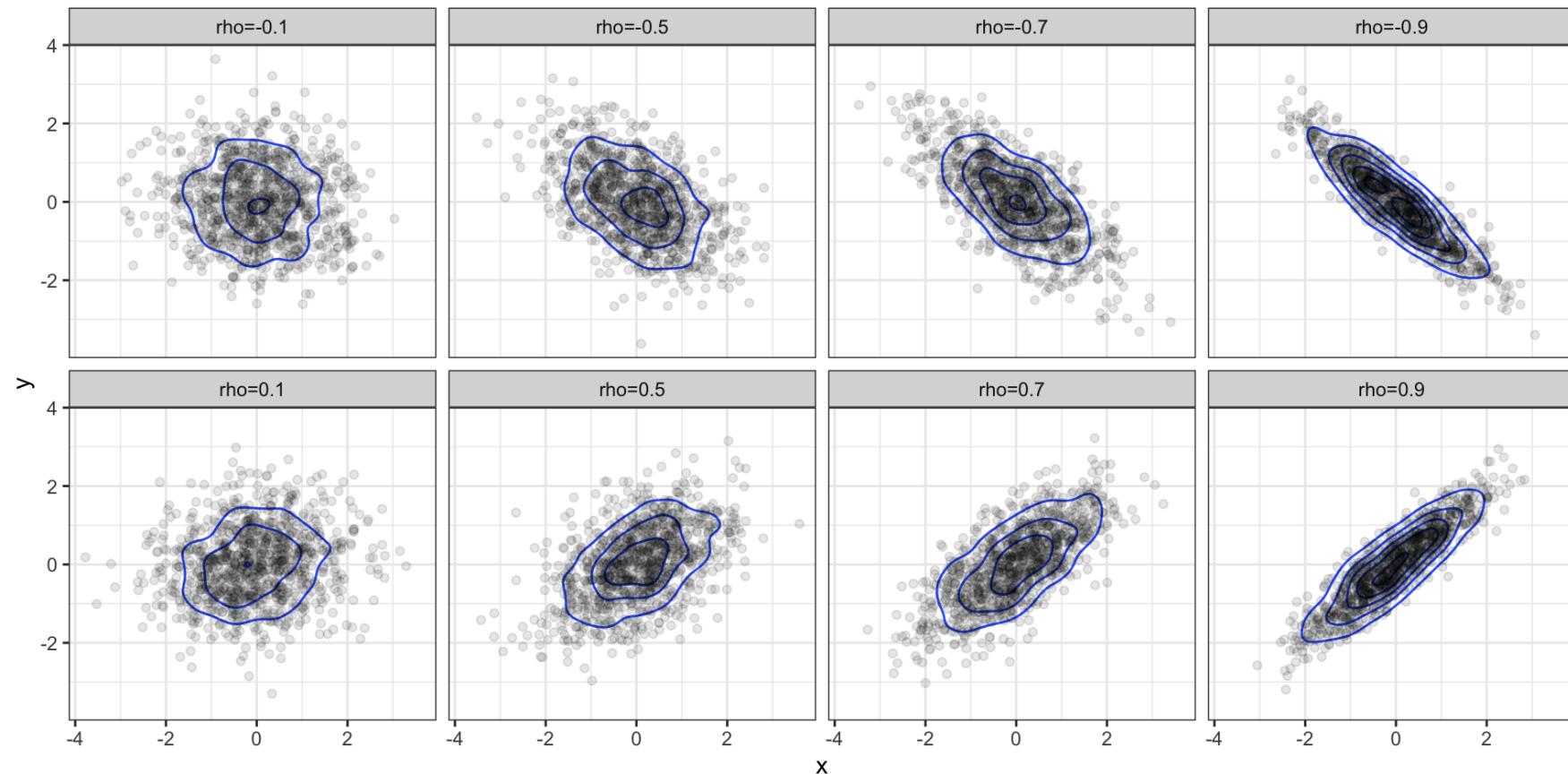
and covariance matrix  $\Sigma_{n \times n}$ ,

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

$$y_{n \times 1} = \mu_{n \times 1} + A_{n \times n} z_{n \times 1}$$

# Bivariate Examples

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



# Marginal / conditional distributions

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

*Univariate marginal distribution:*

$$y_i = N(\mu_i, \Sigma_{ii})$$

*Bivariate marginal distribution:*

$$y_{ij} = N\left(\begin{pmatrix} \mu_i \\ \mu_j \end{pmatrix}, \begin{pmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{pmatrix}\right)$$

*k-dimensional marginal distribution:*

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

# Conditional Distributions

If we partition the  $n$ -dimensions into two pieces such that  $\mathbf{y} = (y_1, y_2)^t$  then

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

$$\underset{k \times 1}{y_1} \sim N(\underset{k \times 1}{\boldsymbol{\mu}_1}, \underset{k \times k}{\boldsymbol{\Sigma}_{11}})$$

$$\underset{n-k \times 1}{y_2} \sim N(\underset{n-k \times 1}{\boldsymbol{\mu}_2}, \underset{n-k \times n-k}{\boldsymbol{\Sigma}_{22}})$$

then the conditional distributions are given by

$$y_1 \mid y_2 = \mathbf{a} \sim N(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{a} - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21})$$

$$y_2 \mid y_1 = \mathbf{b} \sim N(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{b} - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12})$$

# 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  $\mathbf{y} = (y_{t_1}, y_{t_2}, \dots, y_{t_n})^t$ , for every collection of time points  $t_1, t_2, \dots, t_n$ , and every positive integer  $n$ , have a multivariate normal distribution.

So far we have only looked at examples of time series where  $T$  is discrete (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

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

- 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\*
- Necessary to make some simplifying assumptions:
  - Stationarity
  - Simple(r) parameterization of  $\Sigma$

# Covariance Functions

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

*Exponential covariance:*

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

*Squared exponential covariance (Gaussian):*

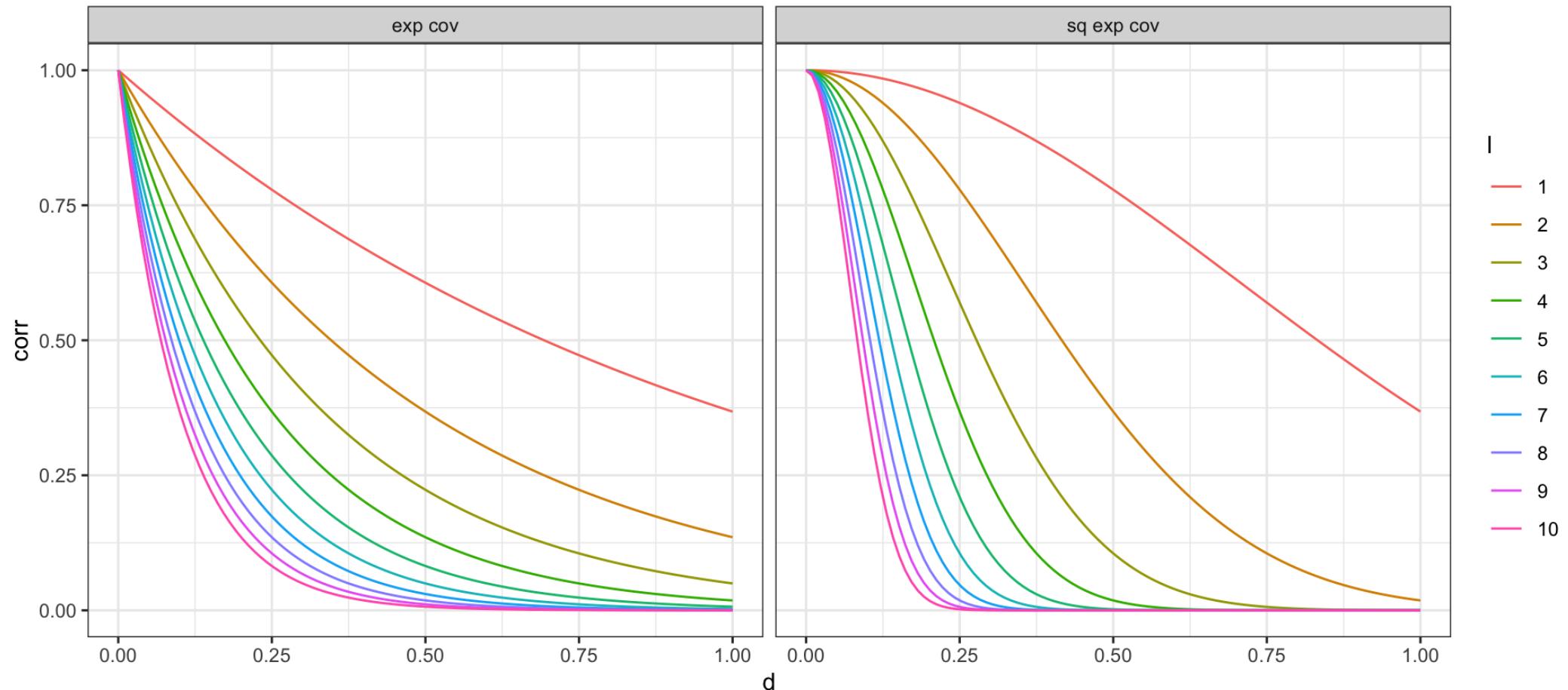
$$\Sigma(y_t, y_{t'}) = \sigma^2 \exp(-(|t - t'| l)^2)$$

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

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

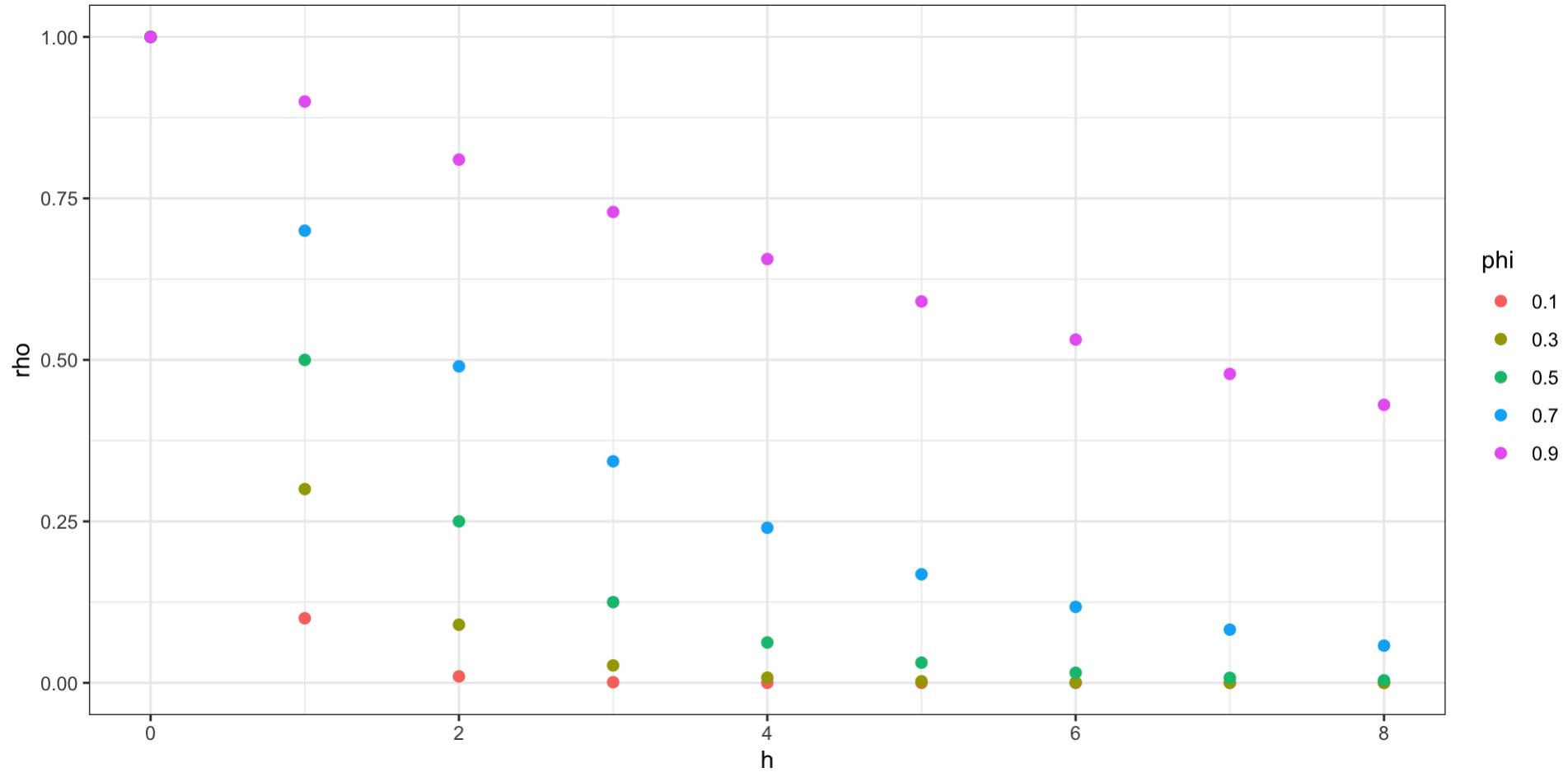
# Covariance Function - Correlation Decay

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

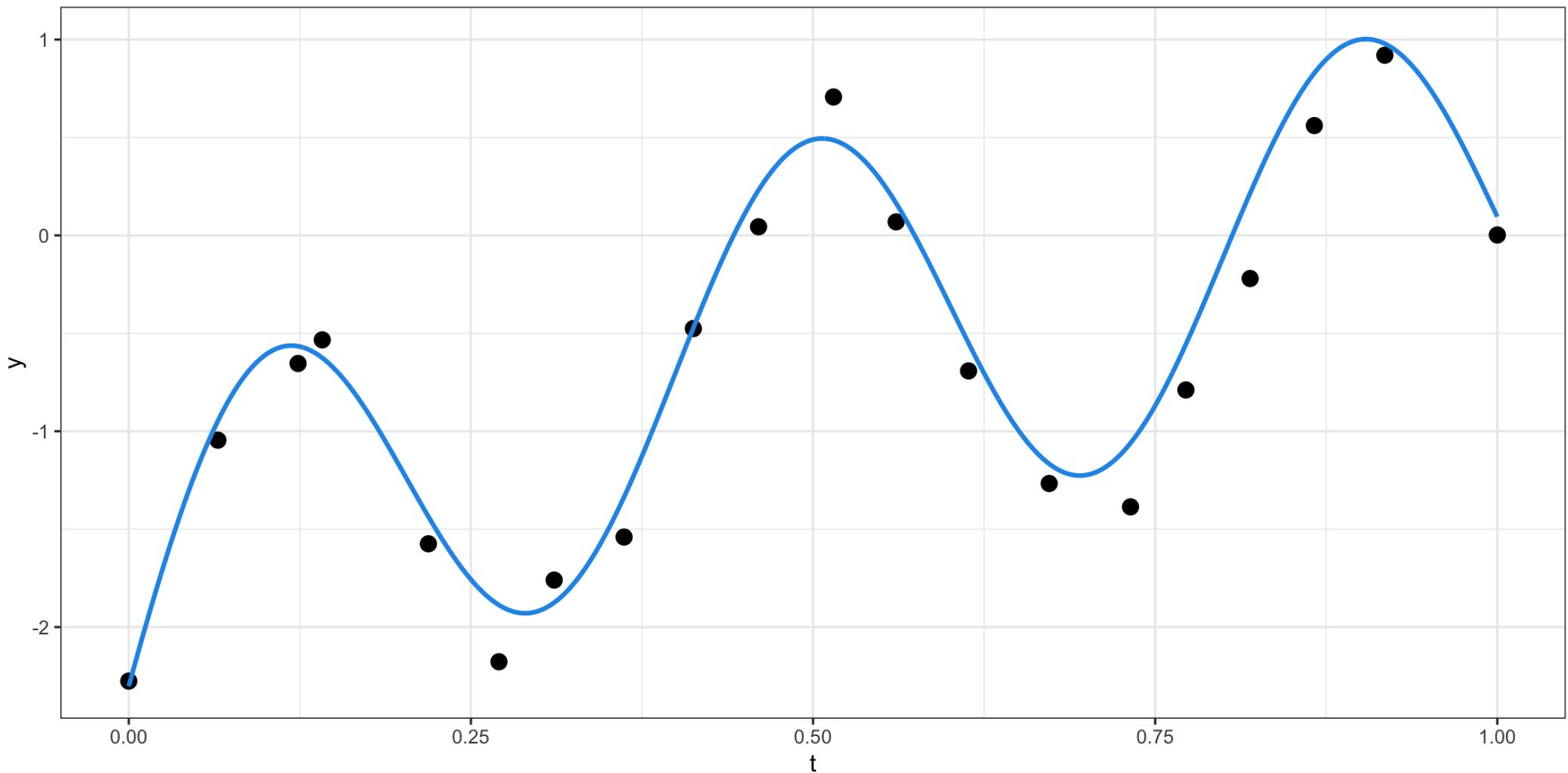


# Correlation Decay - AR(1)

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



# Example



# Prediction

Our example has 20 observations ( $\mathbf{y}_{\text{obs}} = (y(t_1), \dots, y(t_{20}))'$ ), which we would like to use as the basis for predicting  $y(t)$  at other values of  $t$  (say a regular sequence of values from 0 to 1), ( $\mathbf{y}_{\text{pred}}$ ).

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

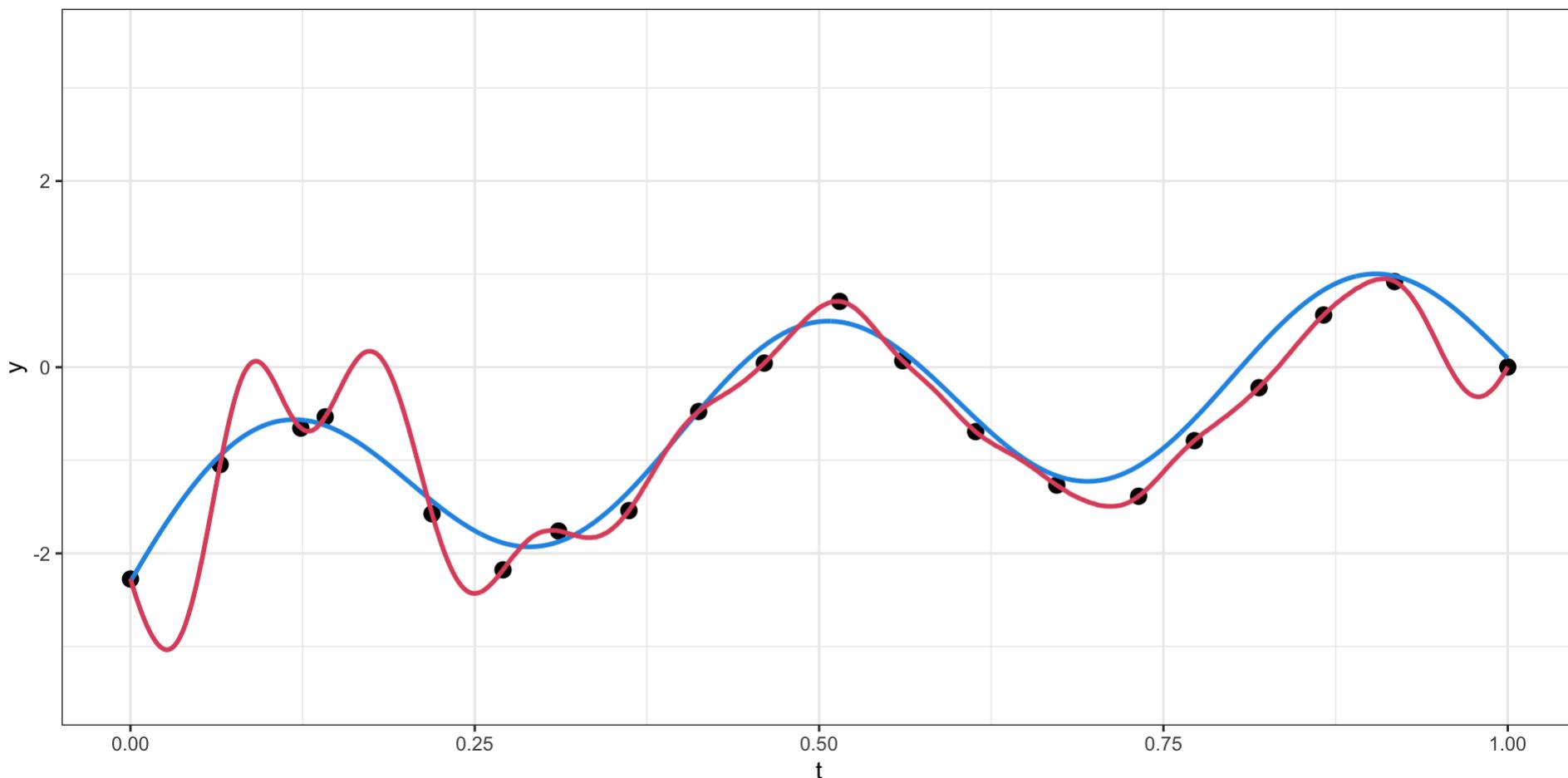
$$\Sigma(y_t, y_{t'}) = \sigma^2 \exp(-|t - t'| l)$$

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

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

# 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  $\sigma = 10$ ,  $l = \sqrt{15}$ ,

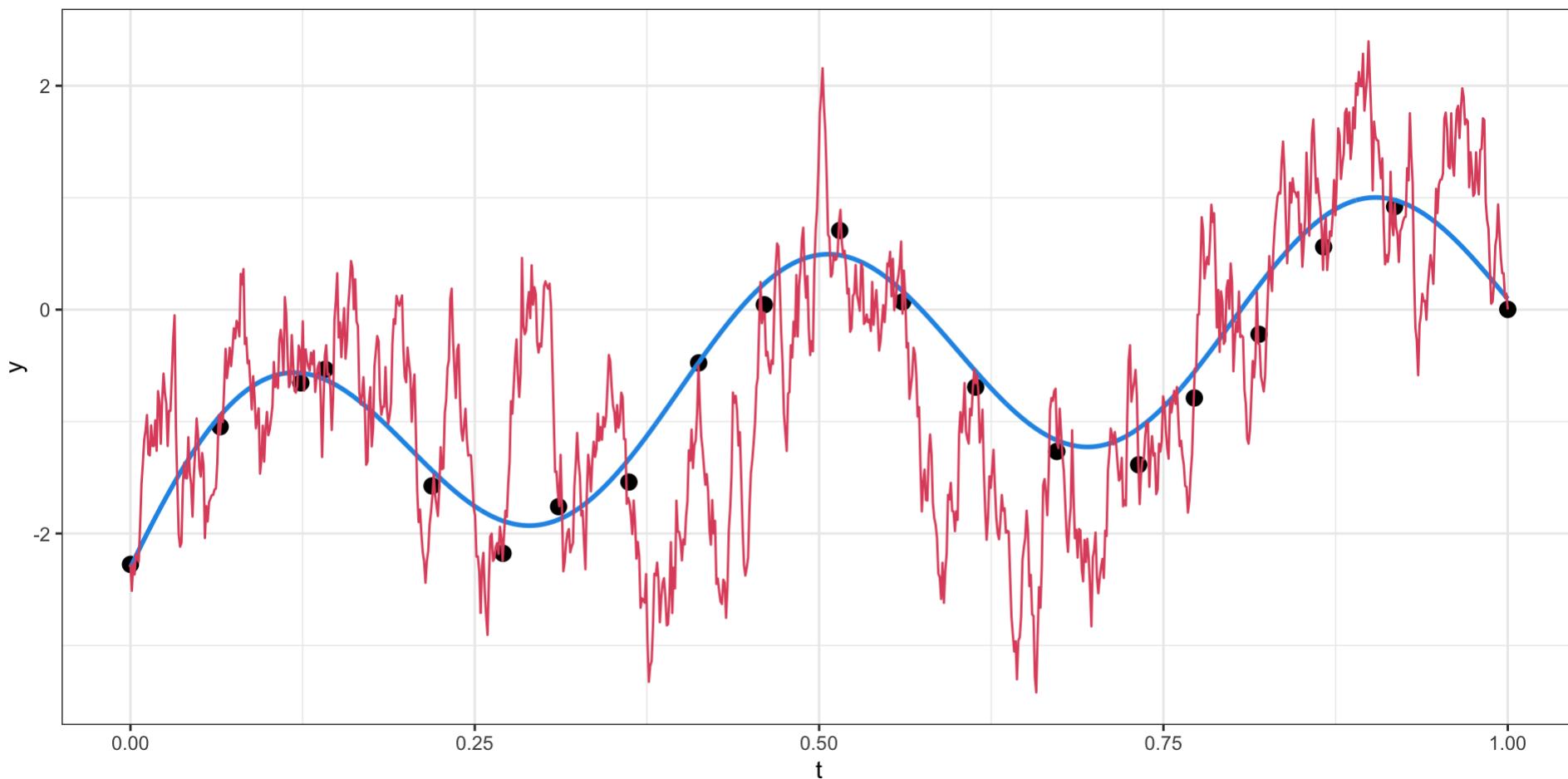
$$\Sigma(y_t, y_{t'}) = \sigma^2 \exp(-|t - t'| l)$$

We are still sampling from  $y_{\text{pred}} | y_{\text{obs}}$ , 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

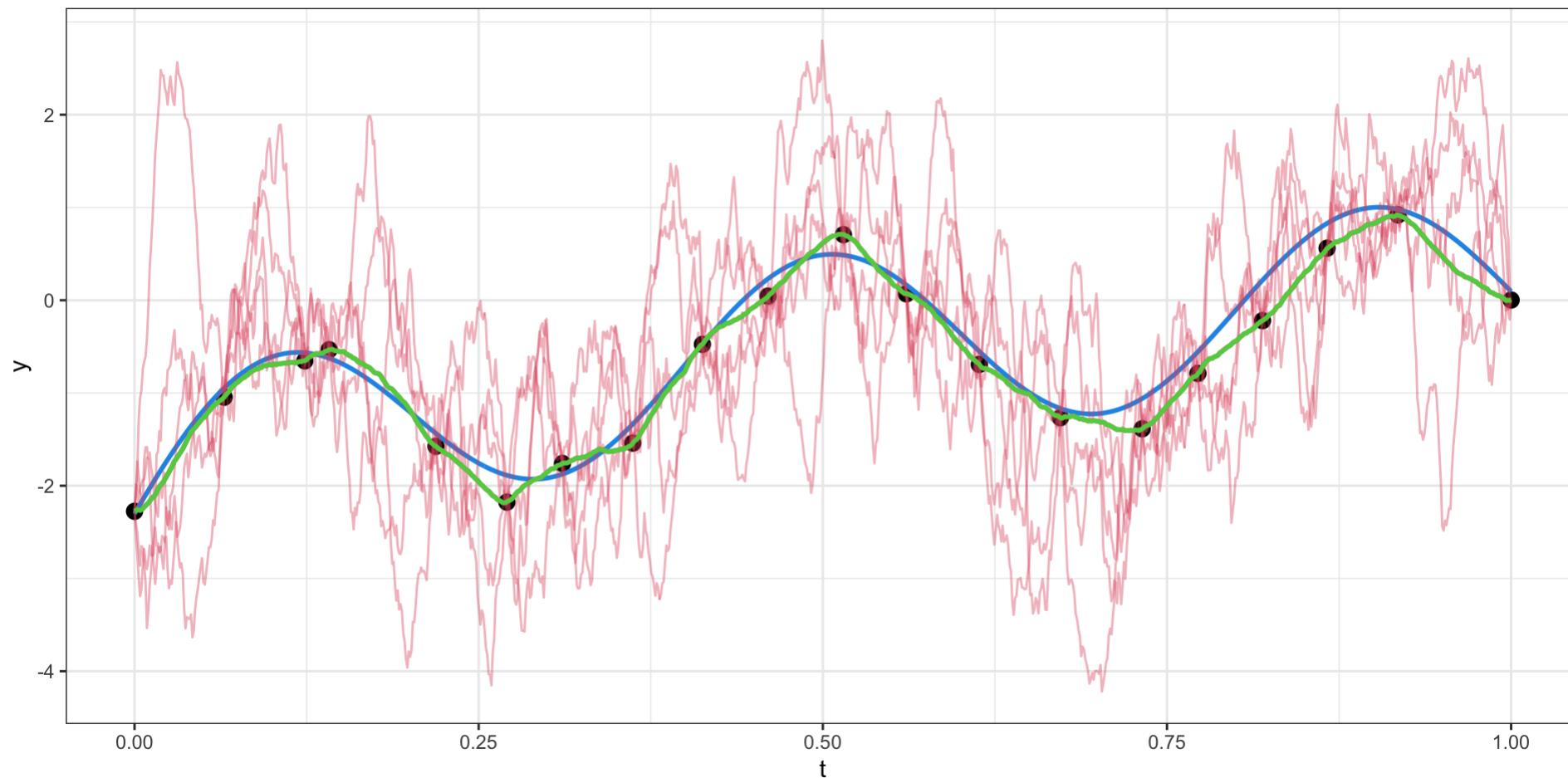
Draw 1   Draw 2   Draw 3   Draw 4   Draw 5   Mean   CI



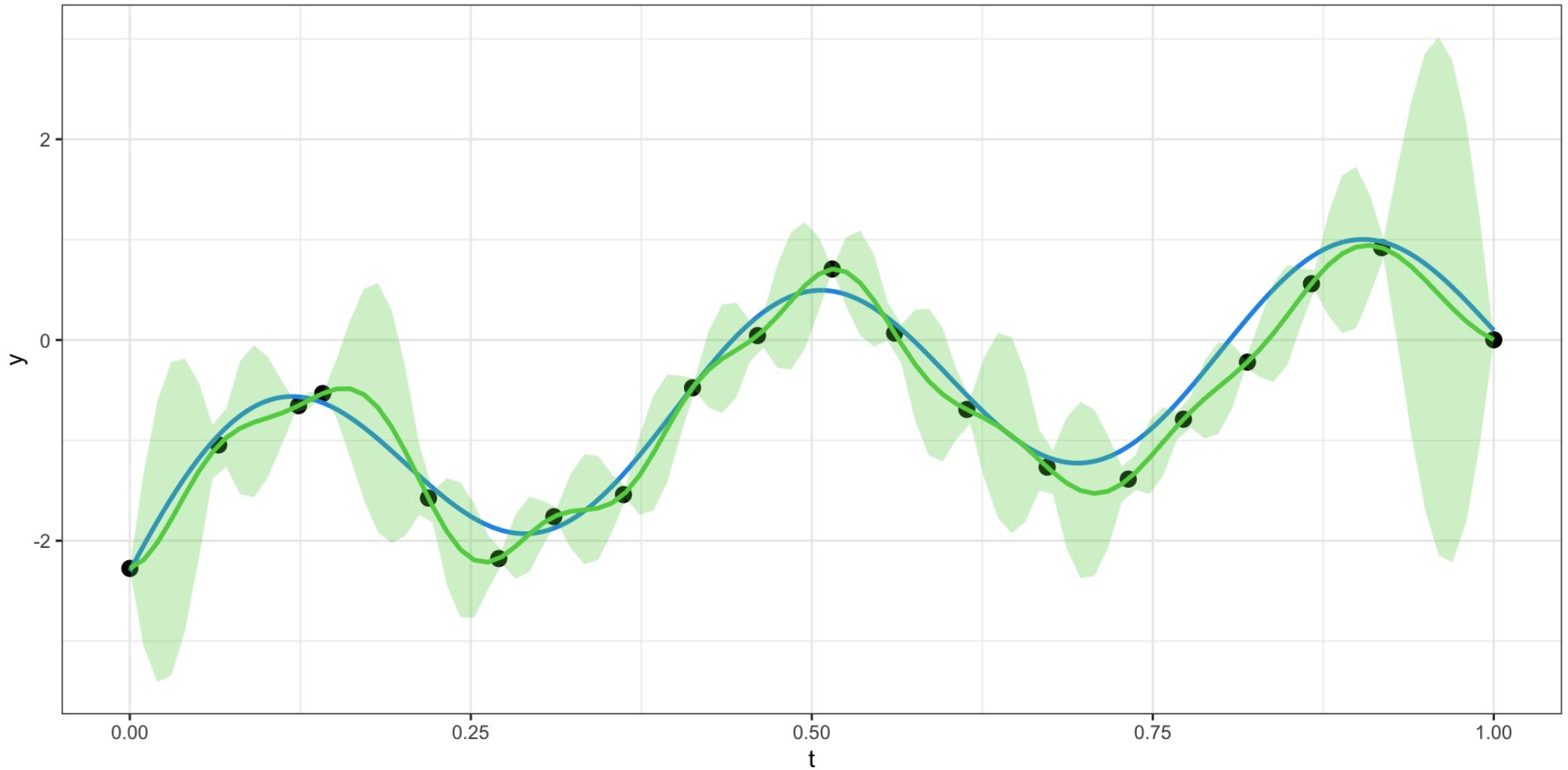
# Powered exponential covariance ( $p = 1.5$ )

Mean

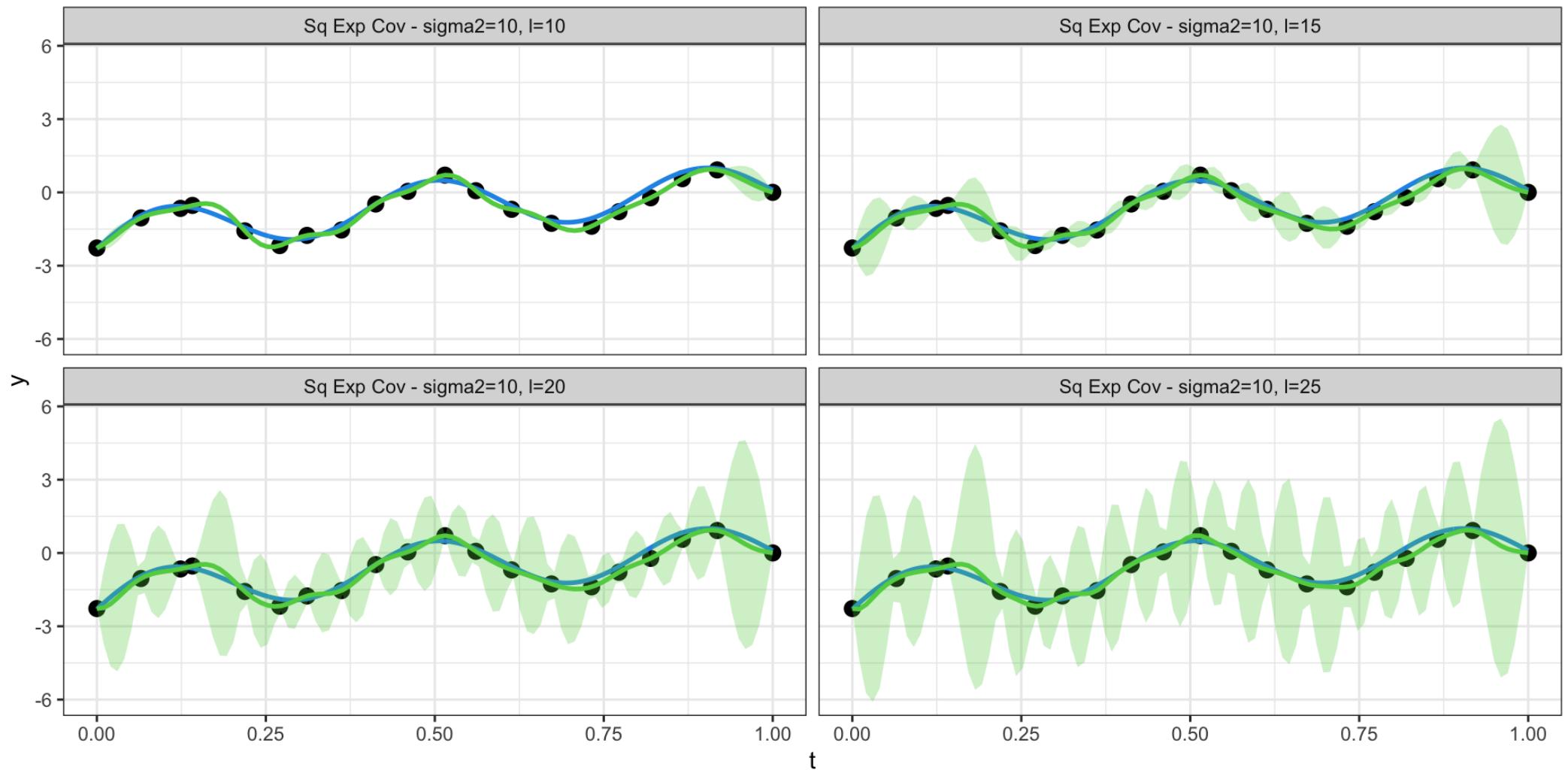
CI



# Back to the square exponential



# Changing the lengthscale (1)



# Effective range

For the square exponential covariance

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

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

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?

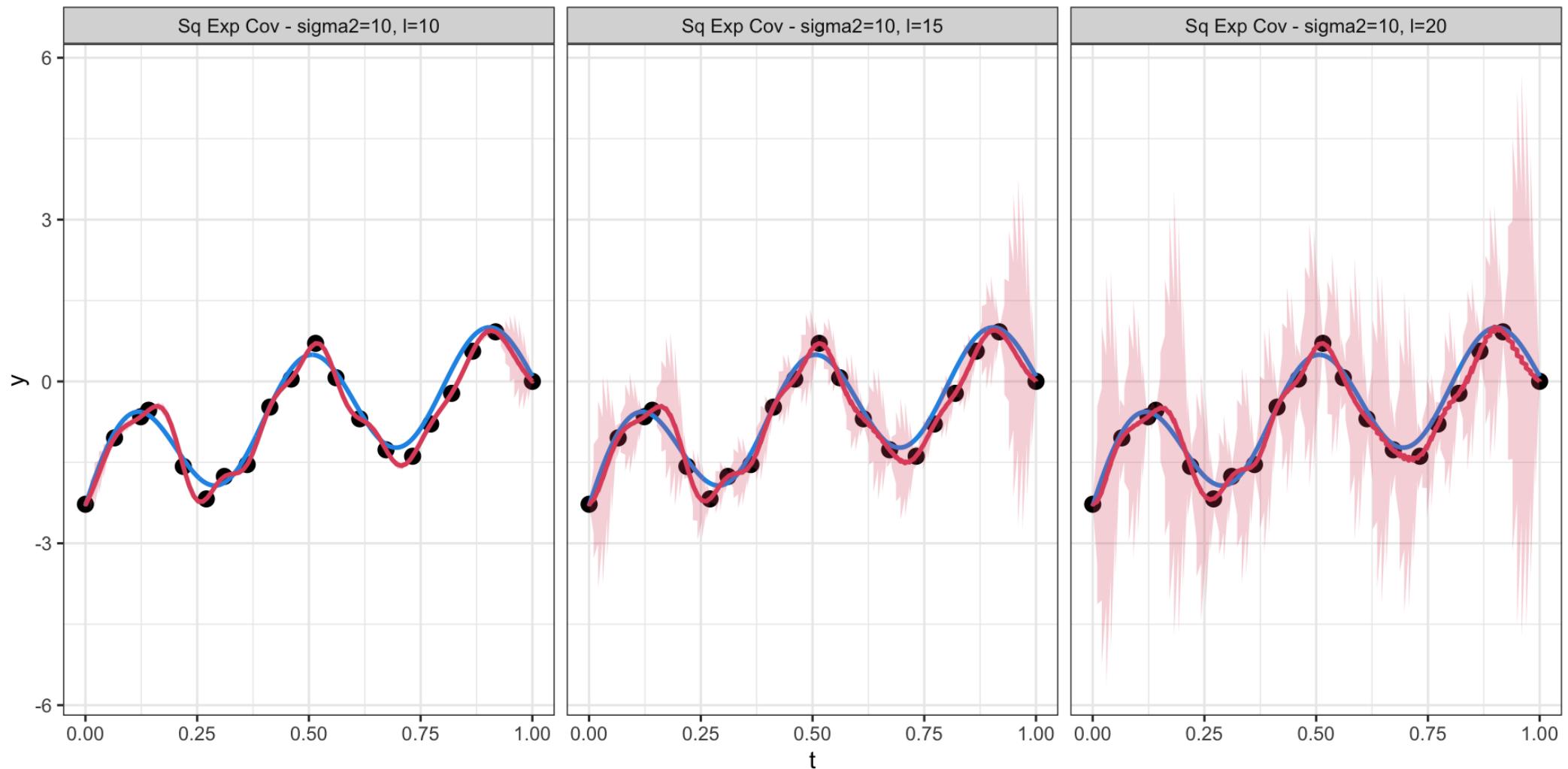
$$\exp(-(l \cdot d)^2) < 0.05$$

$$-(l \cdot d)^2 < \log 0.05$$

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

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

# Changing the scale ( $\sigma^2$ )



# Fitting w/ BRMS

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

```
Running /opt/homebrew/Cellar/r/4.3.1/lib/R/bin/R CMD SHLIB foo.c
using C compiler: 'Apple clang version 15.0.0 (clang-1500.0.40.1)'
using SDK: 'MacOSX14.0.sdk'
clang -I"/opt/homebrew/Cellar/r/4.3.1/lib/R/include" -DNDEBUG  -
I"/Users/rundel/Library/R/arm64/4.3/library/Rcpp/include/"  -
I"/Users/rundel/Library/R/arm64/4.3/library/RcppEigen/include/"  -
I"/Users/rundel/Library/R/arm64/4.3/library/RcppEigen/include/unsupported"  -
I"/Users/rundel/Library/R/arm64/4.3/library/BH/include"  -
I"/Users/rundel/Library/R/arm64/4.3/library/StanHeaders/include/src/"  -
I"/Users/rundel/Library/R/arm64/4.3/library/StanHeaders/include/"  -
I"/Users/rundel/Library/R/arm64/4.3/library/RcppParallel/include/"  -
I"/Users/rundel/Library/R/arm64/4.3/library/rstan/include" -DEIGEN_NO_DEBUG -DBOOST_DISABLE_ASSERTS
-DBOOST_PENDING_INTEGER_LOG2_HPP -DSTAN_THREADS -DBOOST_NO_AUTO_PTR -include
'/Users/rundel/Library/R/arm64/4.3/library/StanHeaders/include/stan/math/prim/fun/Eigen.hpp'  -
D_REENTRANT -DRCPP_PARALLEL_USE_TBB=1 -I/opt/homebrew/opt/gettext/include -
I/opt/homebrew/opt/readline/include -I/opt/homebrew/opt/xz/include -I/opt/homebrew/include  -fPIC
-g -O2 -c foo.c -o foo.o
<built-in>:1:10: fatal error:
'/Users/rundel/Library/R/arm64/4.3/library/StanHeaders/include/stan/math/prim/fun/Eigen.hpp' file
not found
#include
```

```
1 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)	1.60	0.63	0.80	3.28	1.02	282	374
lscale(gpt)	0.12	0.03	0.08	0.17	1.01	249	310

#### Population-Level Effects:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-1.01	0.79	-2.82	0.45	1.01	1118	1075

#### Family Specific Parameters:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	0.19	0.05	0.11	0.32	1.01	941	1403

Draws were sampled using sampling(NUTS). For each parameter, Bulk\_ESS

and Tail\_ESS are effective sample sizes, and Rhat is the potential scale

# 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.

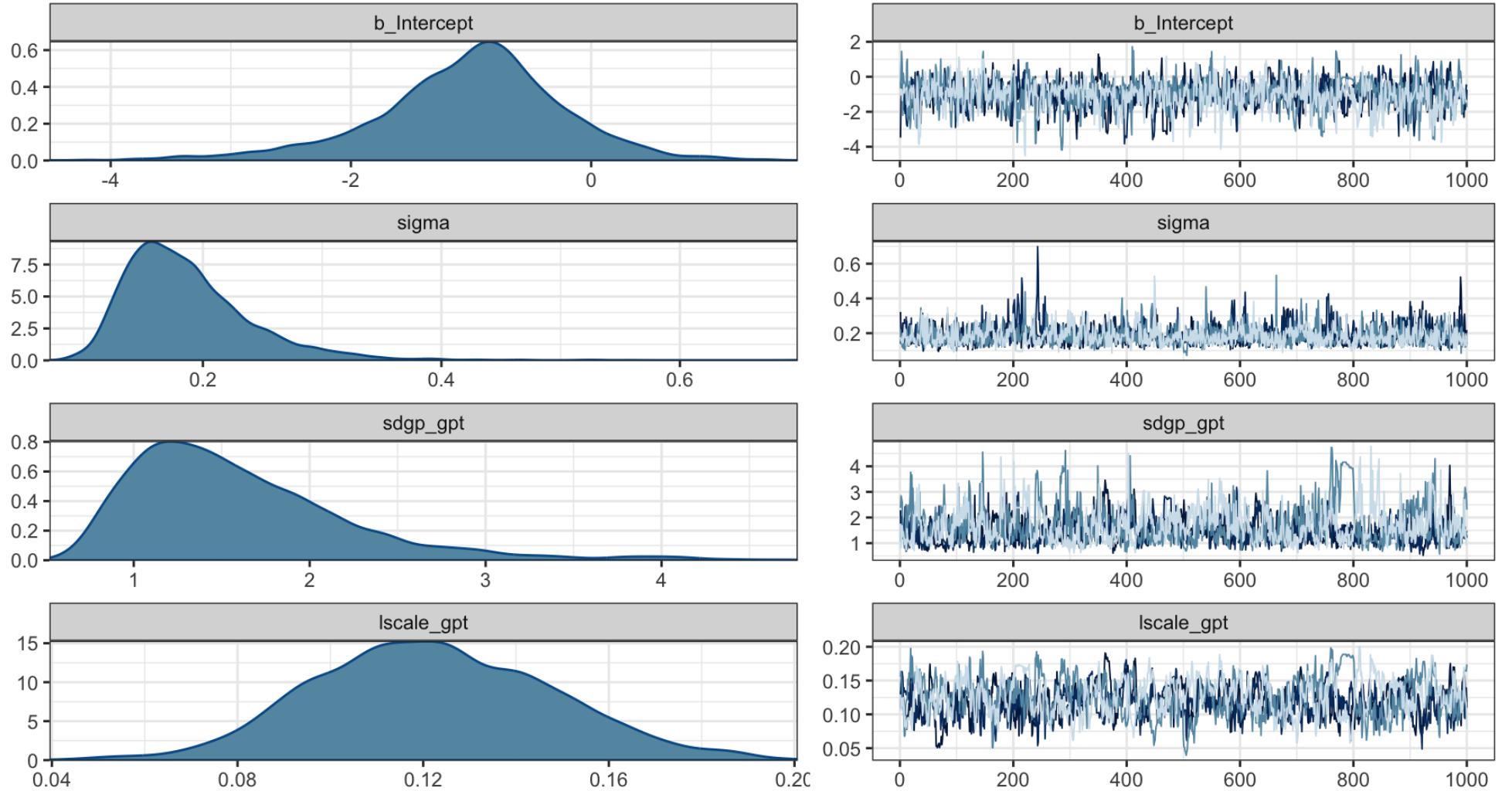
- The covariance function is specified using the `cov` argument, currently only "`exp_quad`" is the default and *only* supported covariance.
- The covariance parameterization used differs slightly from what was given previous (but is equivalent),

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

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

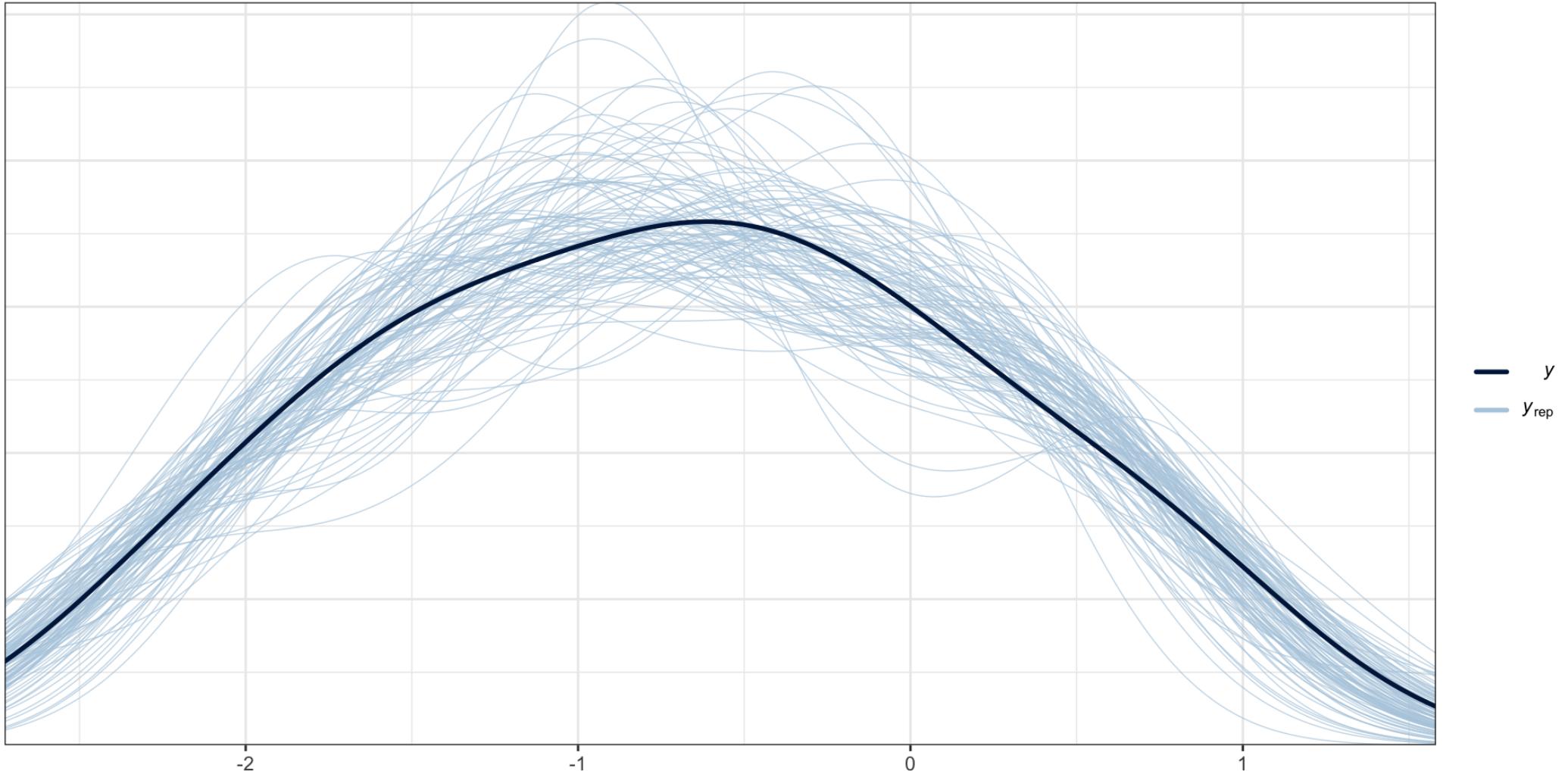
# Trace plots

```
1 plot(gp)
```

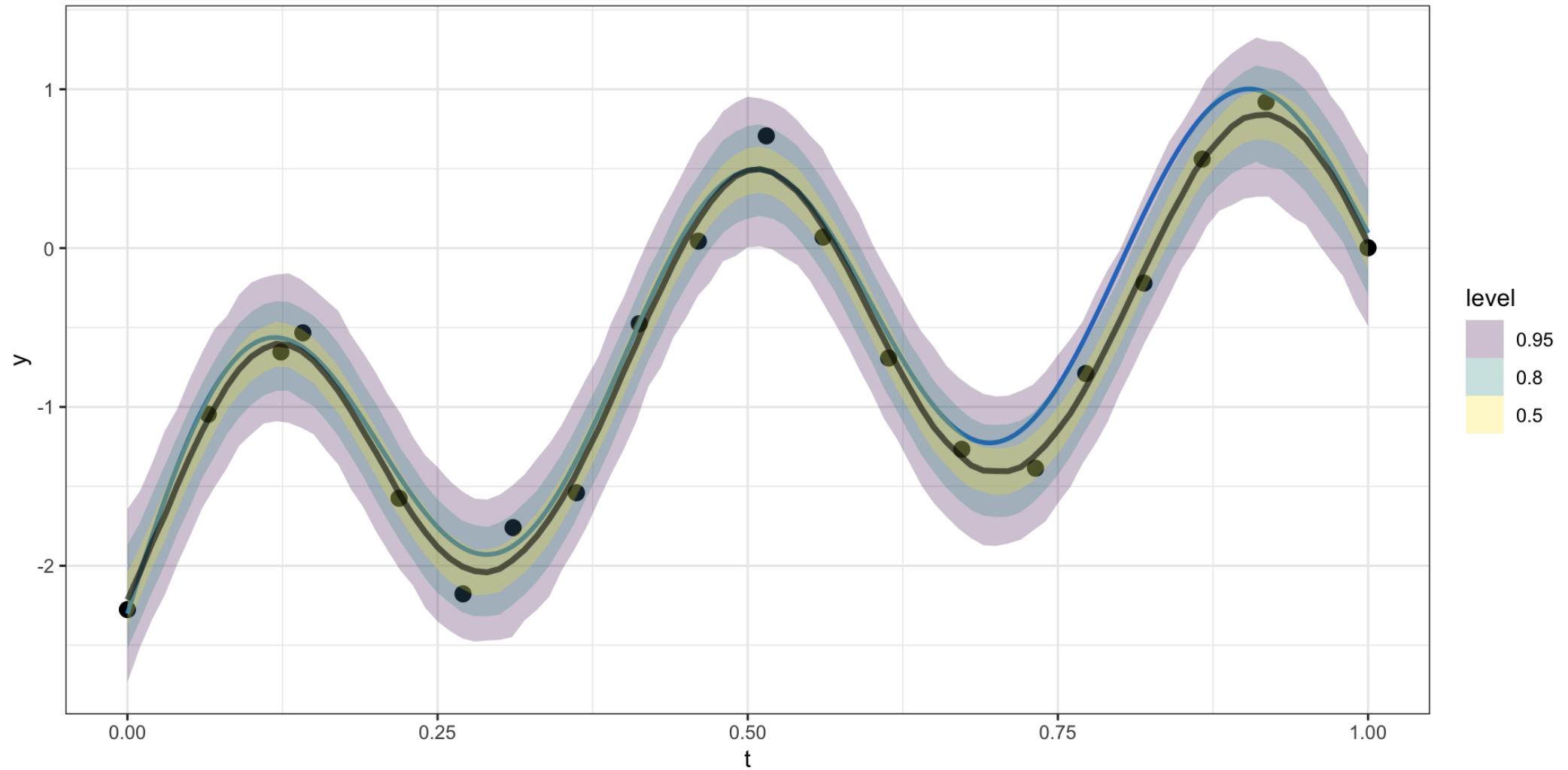


# PP Checks

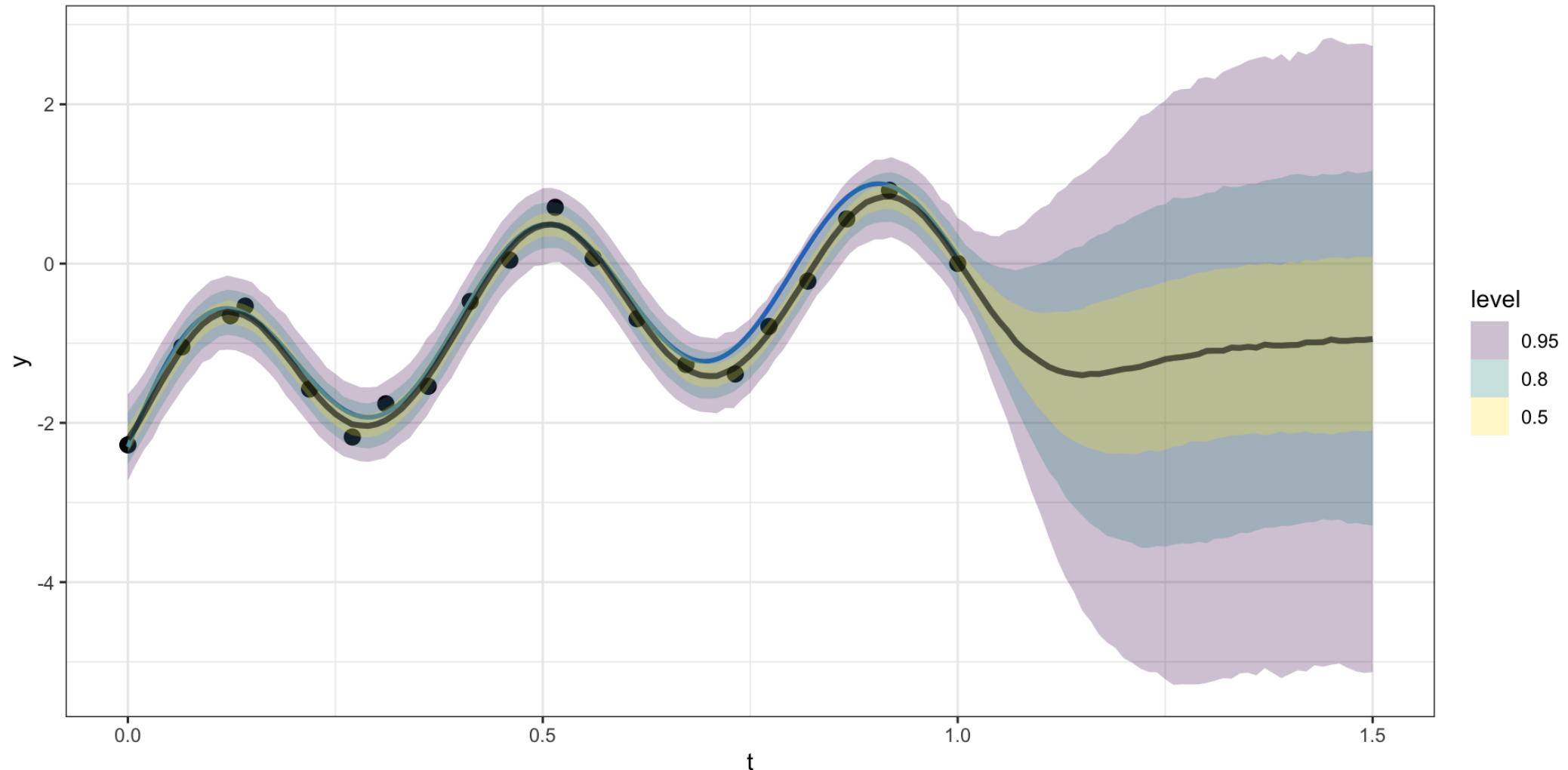
```
1 pp_check(gp, ndraws = 100)
```



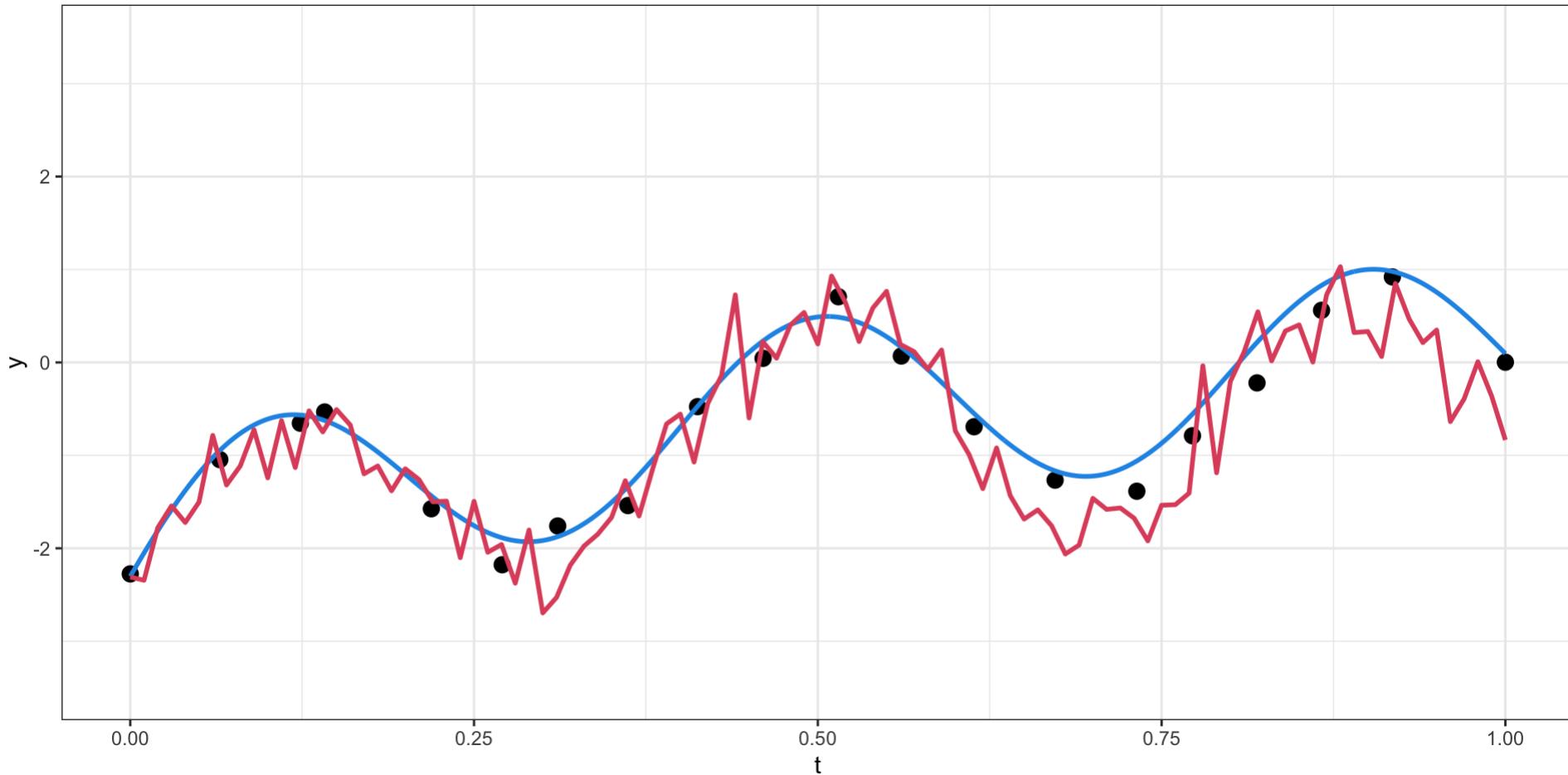
# Model predictions



# Forecasting



# brms paths



# 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?

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

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

Draw 1

Draw 2

Draw 3

Draw 4

Draw 5

Mean

CI

