Lecture 13

Gaussian Process Models - Part 2

Colin Rundel

03/01/2017

EDA and GPs

Variogram

From the spatial modeling literature the typical approach is to examine an *empirical variogram*, first we'll look at the *theoretical variogram* before looking at the connection to the covariance.

Variogram

From the spatial modeling literature the typical approach is to examine an *empirical variogram*, first we'll look at the *theoretical variogram* before looking at the connection to the covariance.

Variogram:

$$
\begin{aligned}
2\gamma(t_i, t_j) &= Var(Y(t_i) - Y(t_j)) \\
&= E([(Y(t_i) - \mu(t_i)) - (Y(t_j) - \mu(t_j))]^2)
\end{aligned}
$$

where $\gamma(t_i, t_j)$ is called the semivariogram.

Variogram

From the spatial modeling literature the typical approach is to examine an *empirical variogram*, first we'll look at the *theoretical variogram* before looking at the connection to the covariance.

Variogram:

$$
\begin{aligned}
2\gamma(t_i, t_j) &= Var(Y(t_i) - Y(t_j)) \\
&= E([(Y(t_i) - \mu(t_i)) - (Y(t_j) - \mu(t_j))]^2)
\end{aligned}
$$

where $\gamma(t_i, t_j)$ is called the semivariogram.

If the process has constant mean (e.g. $\mu(t_i) = \mu(t_j)$ for all i and j) then we can simplify to

$$2\gamma(t_i, t_j) = E([Y(t_i) - Y(t_j)]^2)$$

Some Properties of the theoretical Variogram / Semivariogram

- both are non-negative

$$\gamma(t_i, t_j) \geq 0$$

- both are 0 at distance 0

$$\gamma(t_i, t_i) = 0$$

- both are symmetric

$$\gamma(t_i, t_j) = \gamma(t_j, t_i)$$

- there is no dependence if

$$2\gamma(t_i, t_j) = Var(Y(t_i)) + Var(Y(t_j)) \quad \text{for all } i \neq j$$

- if the process *is not* stationary

$$2\gamma(t_i, t_j) = Var\Big(Y(t_i)\Big) + Var\Big(Y(t_j)\Big) - 2\,Cov\Big(Y(t_i), Y(t_j)\Big)$$

- if the process *is* stationary

$$2\gamma(t_i, t_j) = 2Var\Big(Y(t_i)\Big) - 2\,Cov\Big(Y(t_i), Y(t_j)\Big)$$

Empirical Semivariogram

We will assume that our process of interest is stationary, in which case we will parameterize the semivariagram in terms of $h = \left|t_i - t_j\right|$.

Empirical Semivariogram:

$$\hat{\gamma}(h) = \frac{1}{2\,N(h)} \sum_{|t_i - t_j| \in (h-\epsilon, h+\epsilon)} \left(Y(t_i) - Y(t_j)\right)^2$$

Empirical Semivariogram

We will assume that our process of interest is stationary, in which case we will parameterize the semivariagram in terms of $h = |t_i - t_j|$.

Empirical Semivariogram:

$$\hat{\gamma}(h) = \frac{1}{2\,N(h)} \sum_{|t_i - t_j| \in (h-\epsilon, h+\epsilon)} \big(Y(t_i) - Y(t_j)\big)^2$$

Practically, for any data set with n observations there are $\binom{n}{2} + n$ possible data pairs to examine. Each individually is not very informative, so we aggregate into bins and calculate the empirical semivariogram for each bin.

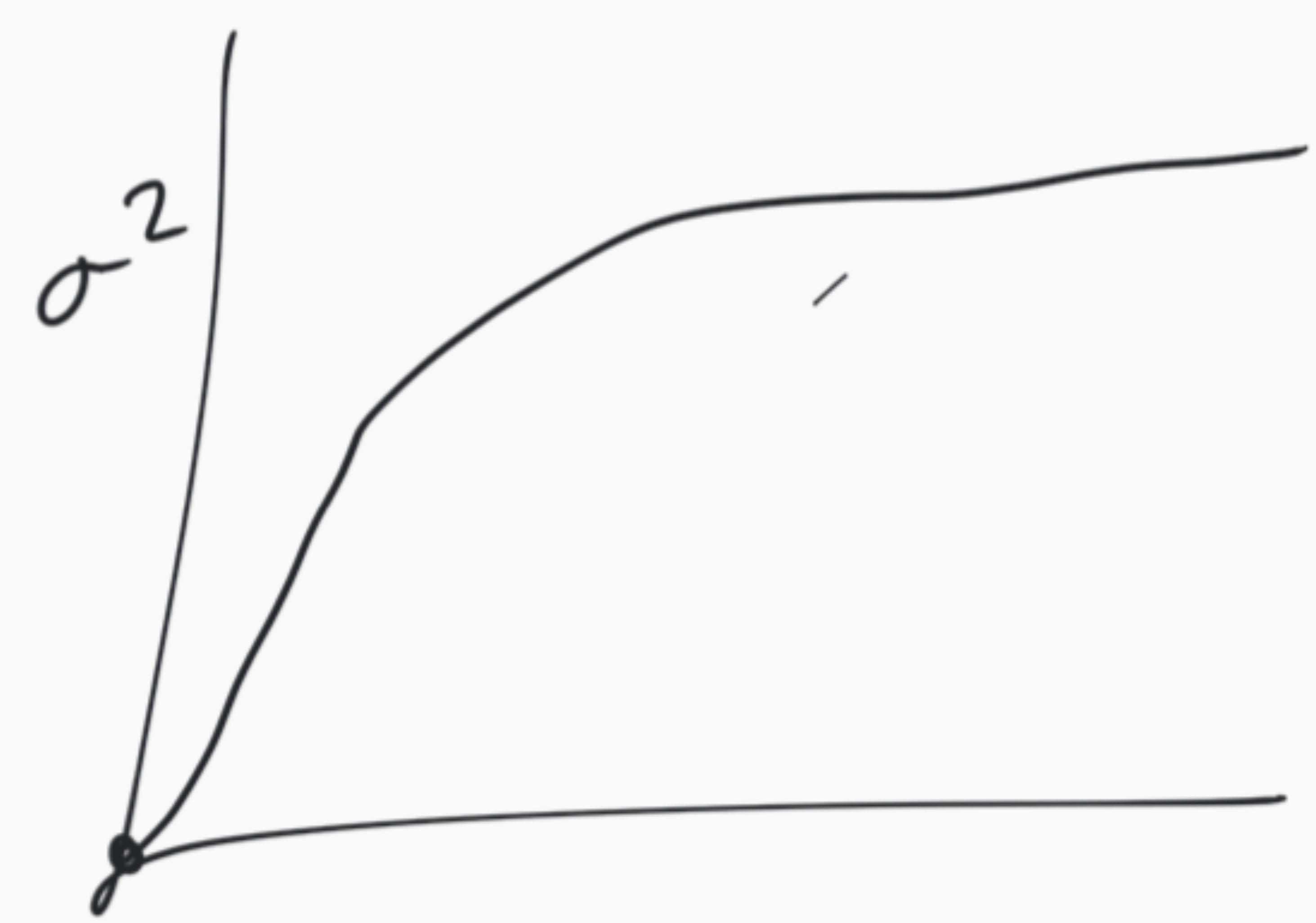
Connection to Covariance

In stationary case

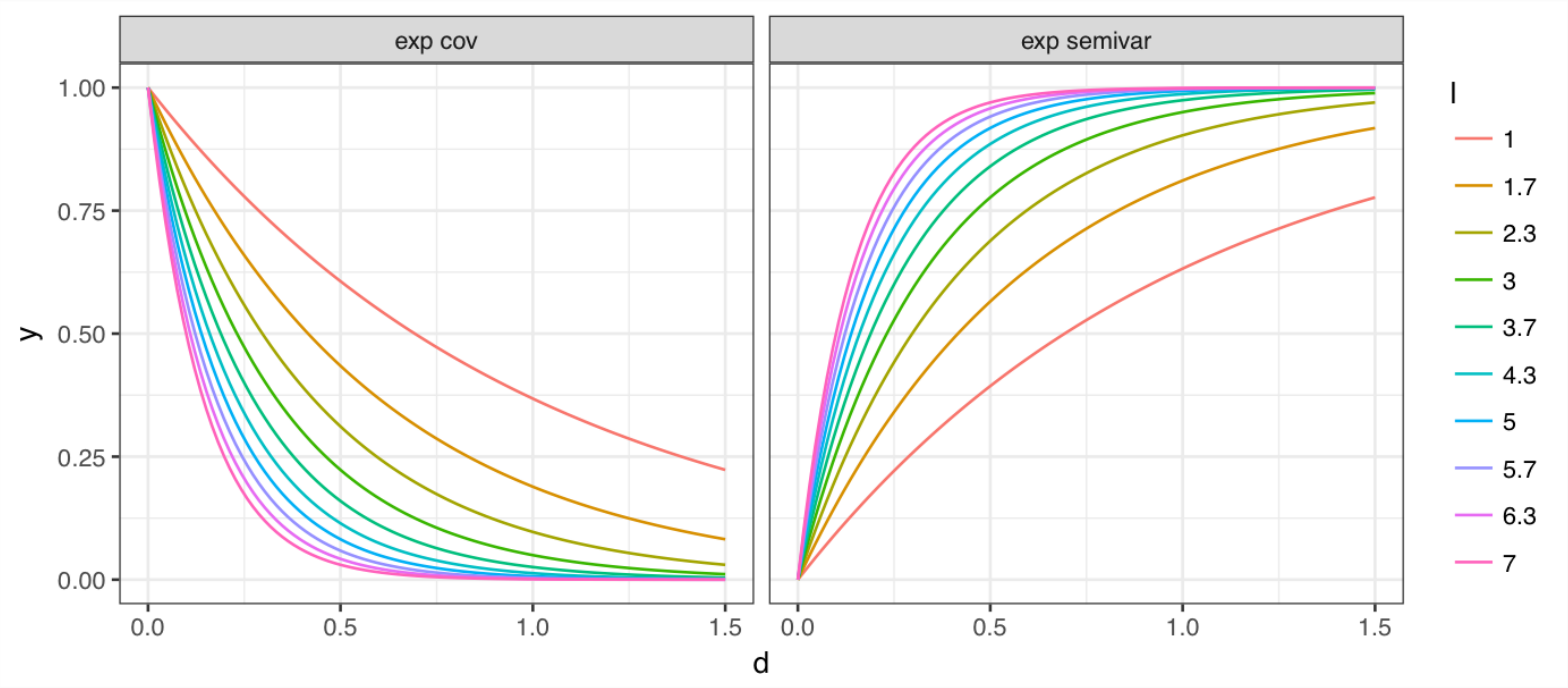
$$\gamma(h) = Var(y_t) - Cov(h)$$
$$= Cov(0) - Cov(h)$$
$$= \sigma^2 - \sigma^2 e^{-h\ell}$$

Exp Cov

$$Cov(h) = \sigma^2 e^{-h\ell}$$

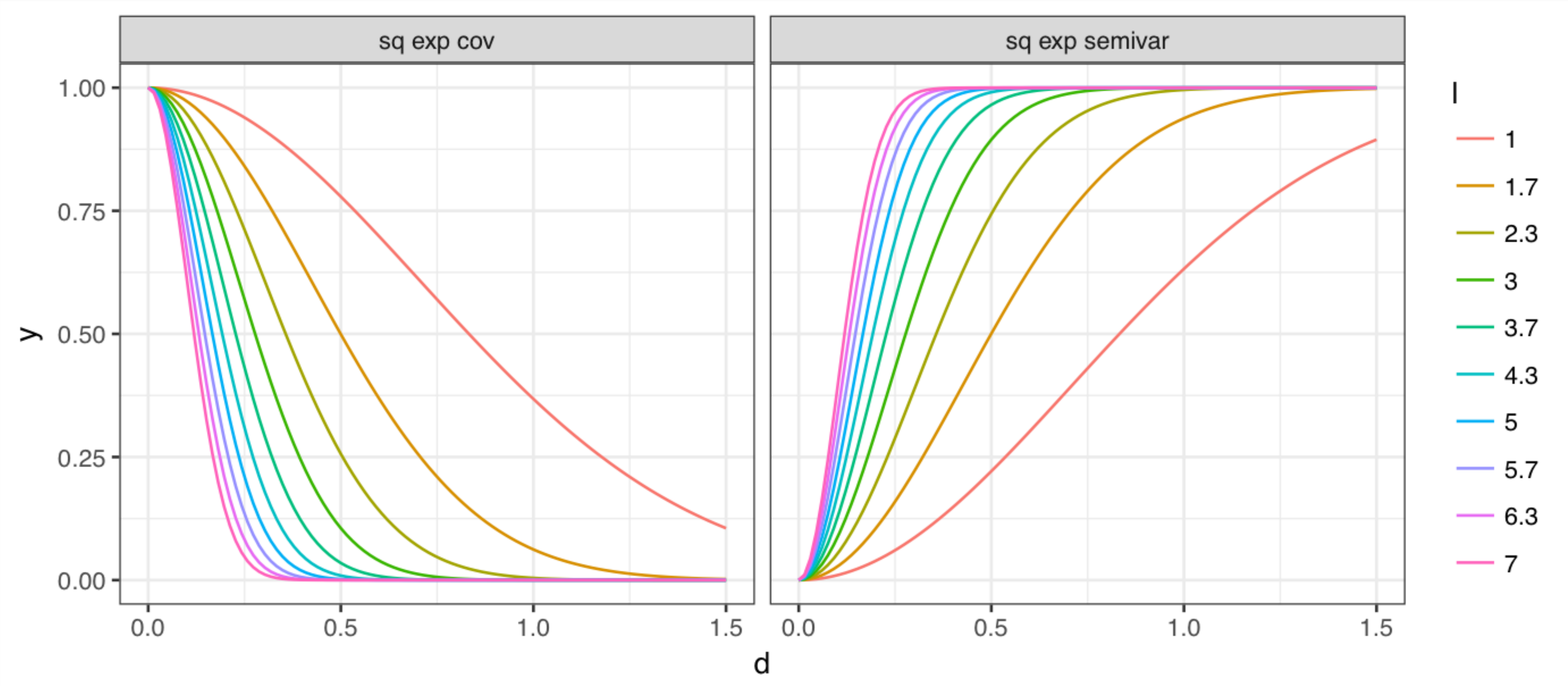


Covariance vs Semivariogram - Exponential

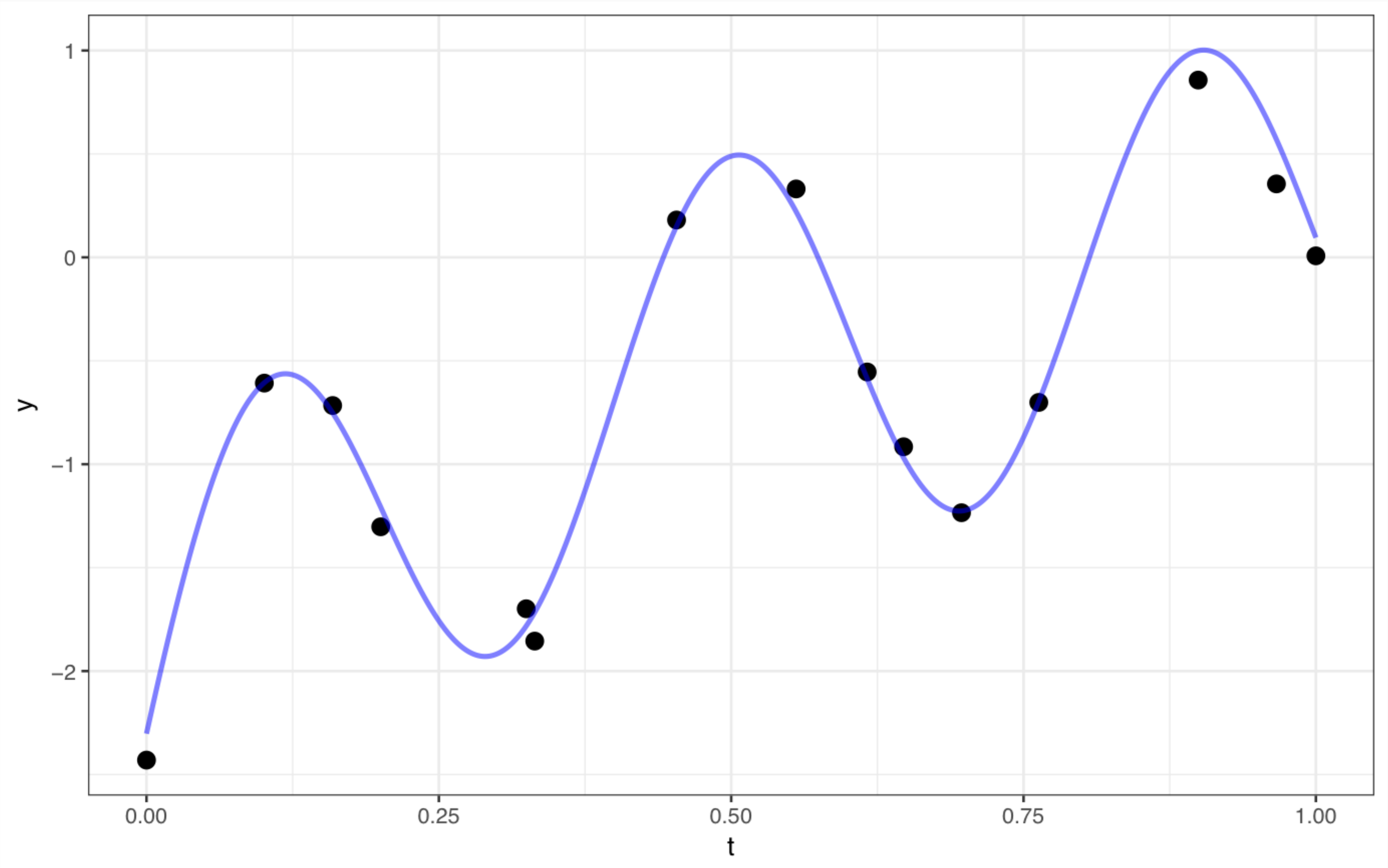


$\sigma^2 = 1$

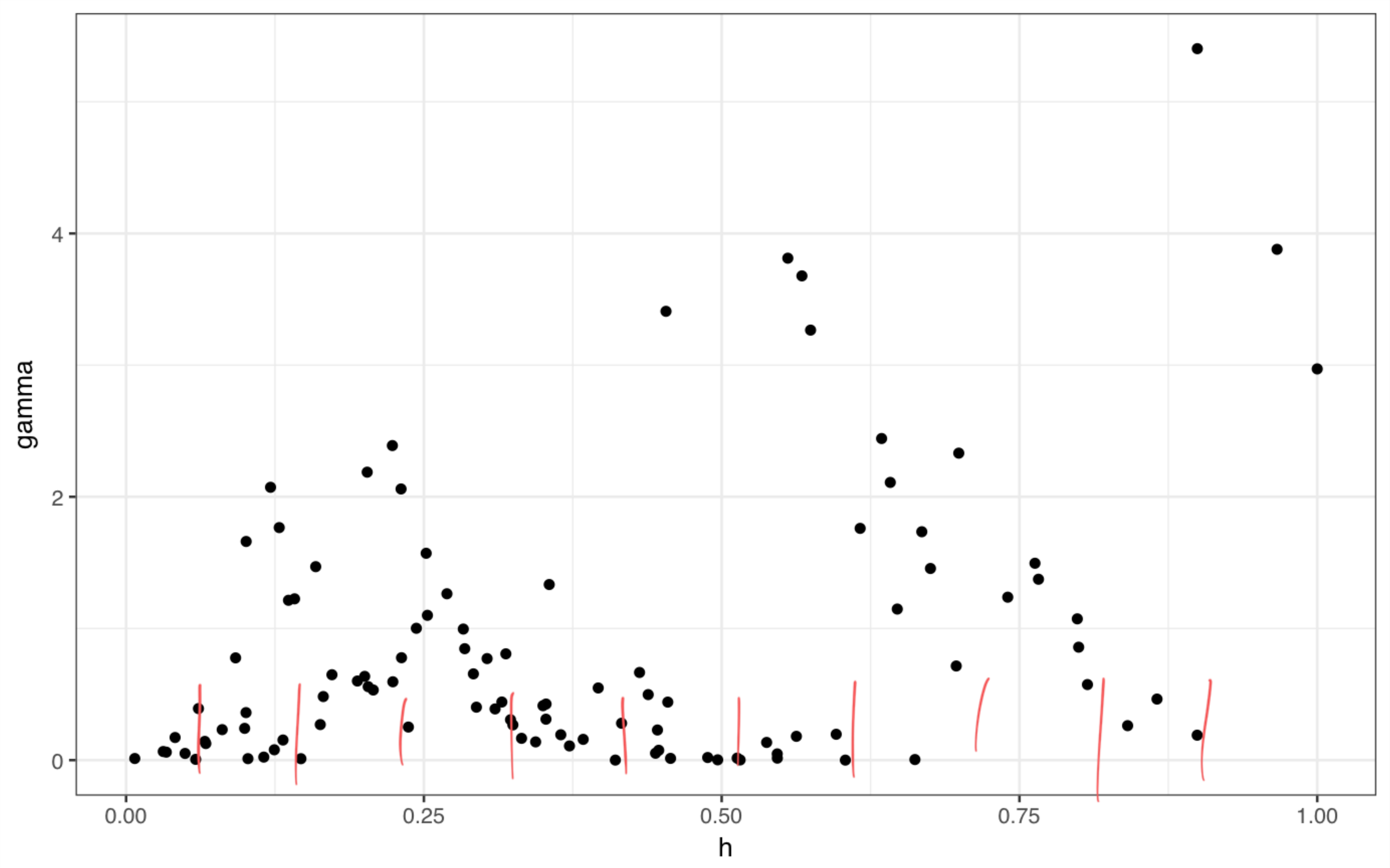
Covariance vs Semivariogram - Square Exponential



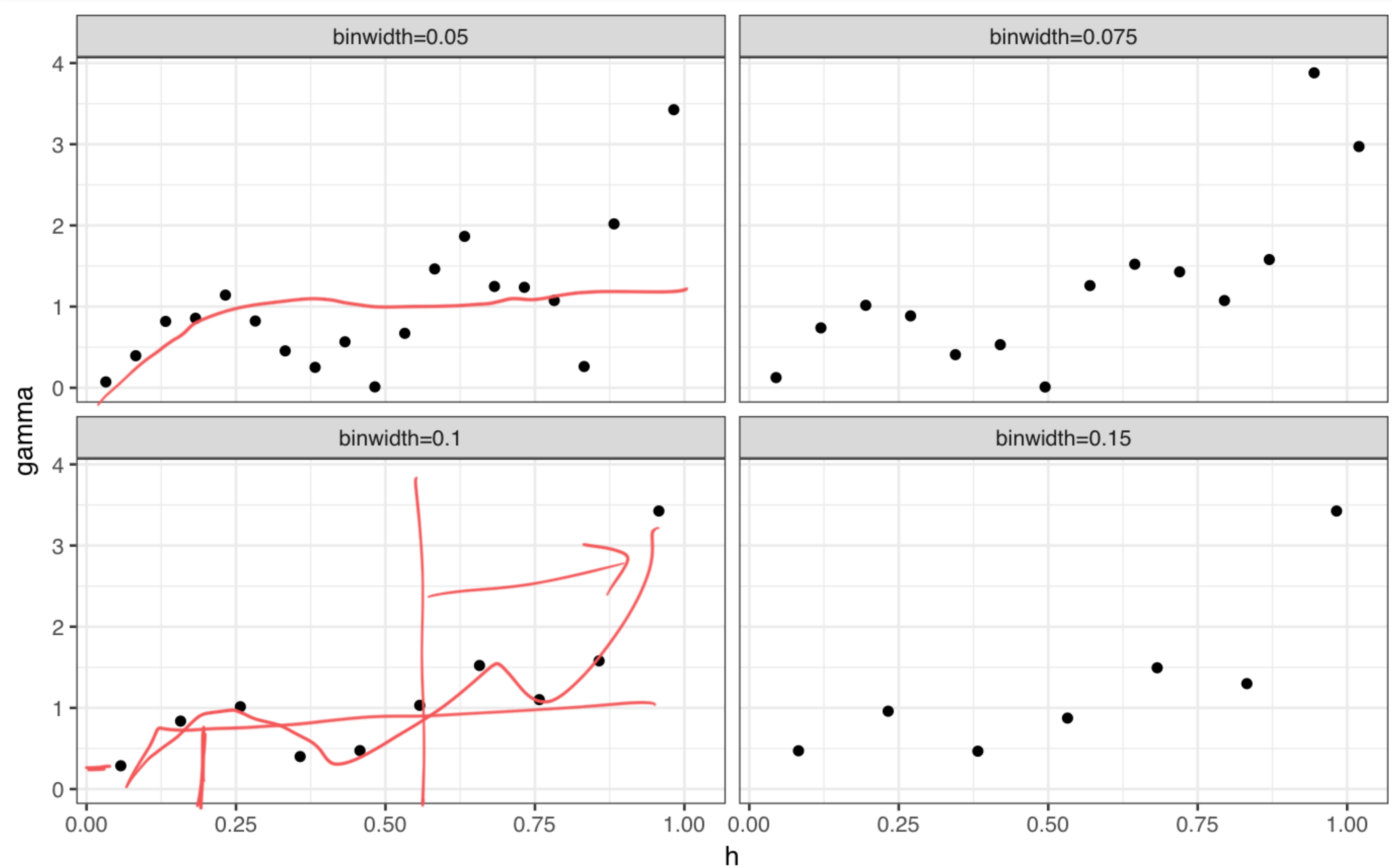
From last time



Empirical semivariogram - no bins / cloud

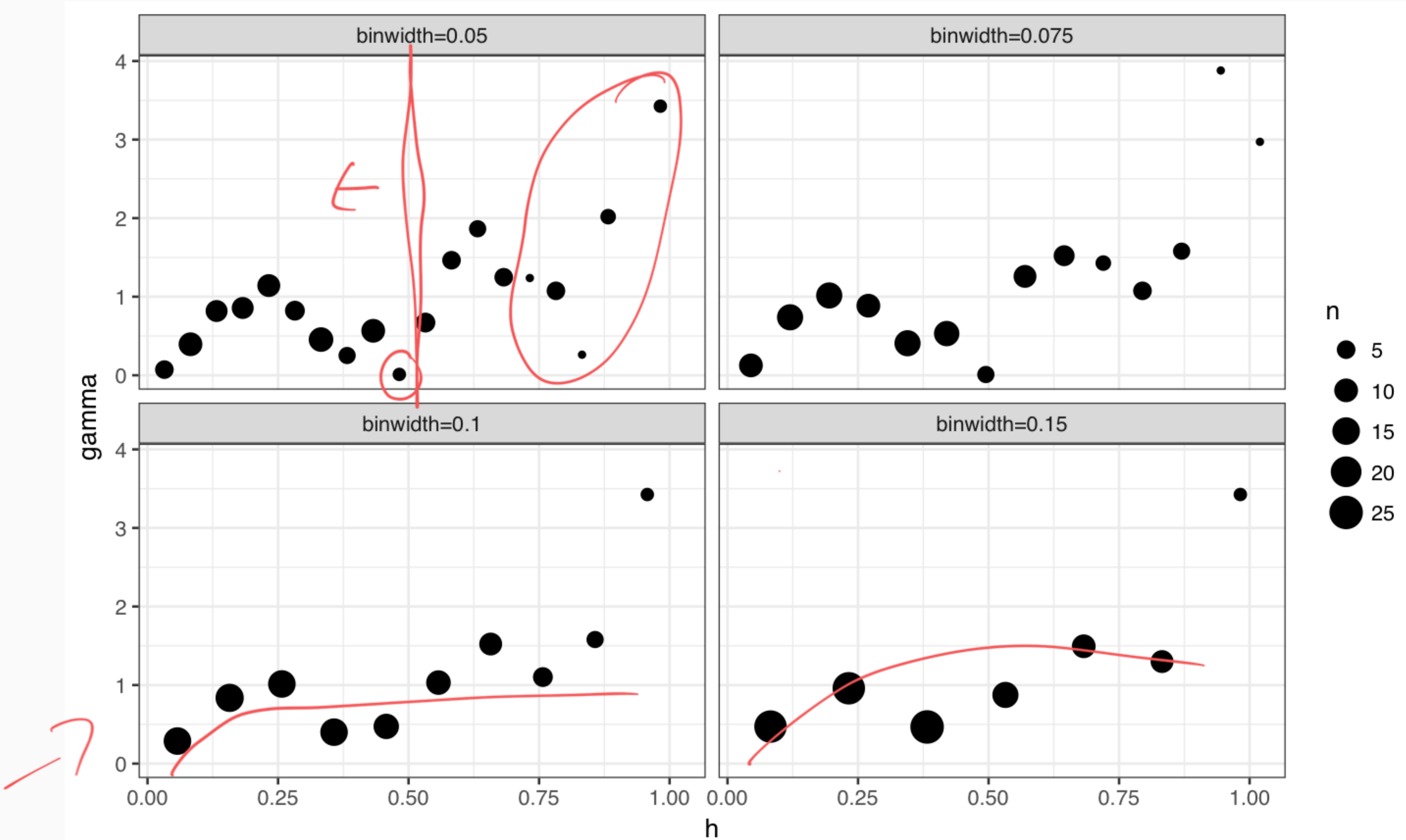


Empirical semivariogram (binned)



$\sigma_v^2 \approx 0.25$ $\quad$ $\sigma^2 = 0.5$ $\quad$ $\ell = \frac{\sqrt{3}}{0.2}$

Empirical semivariogram (binned + n)



Theoretical vs empirical semivariogram

After fitting the model last time we came up with a posterior median of $\sigma^2 = 1.89$ and $l = 5.86$ for a square exponential covariance.

Theoretical vs empirical semivariogram

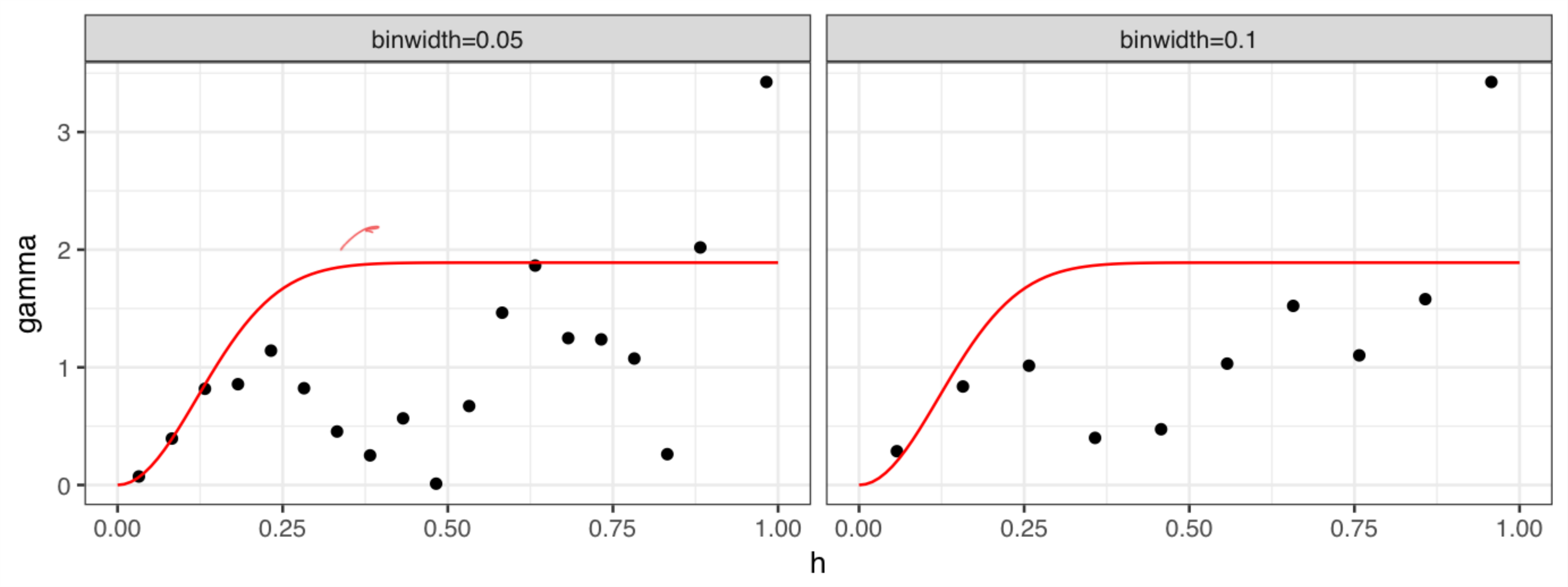
After fitting the model last time we came up with a posterior median of $\sigma^2 = 1.89$ and $l = 5.86$ for a square exponential covariance.

$$
\begin{aligned}
Cov(h) &= \sigma^2 \exp\Big(-(h\, l)^2 \Big) \\
\gamma(h) &= \sigma^2 - \sigma^2 \exp\Big(-(h\, l)^2 \Big) \\
&= 1.89 - 1.89 \exp\Big(-(5.86\, h)^2 \Big)
\end{aligned}
$$

Theoretical vs empirical semivariogram

After fitting the model last time we came up with a posterior median of $\sigma^2 = 1.89$ and $l = 5.86$ for a square exponential covariance.

$$
\begin{aligned}
Cov(h) &= \sigma^2 \exp\Big(-(h\,l)^2\Big) \\
\gamma(h) &= \sigma^2 - \sigma^2 \exp\Big(-(h\,l)^2\Big) \\
&= 1.89 - 1.89 \exp\Big(-(5.86\,h)^2\Big)
\end{aligned}
$$

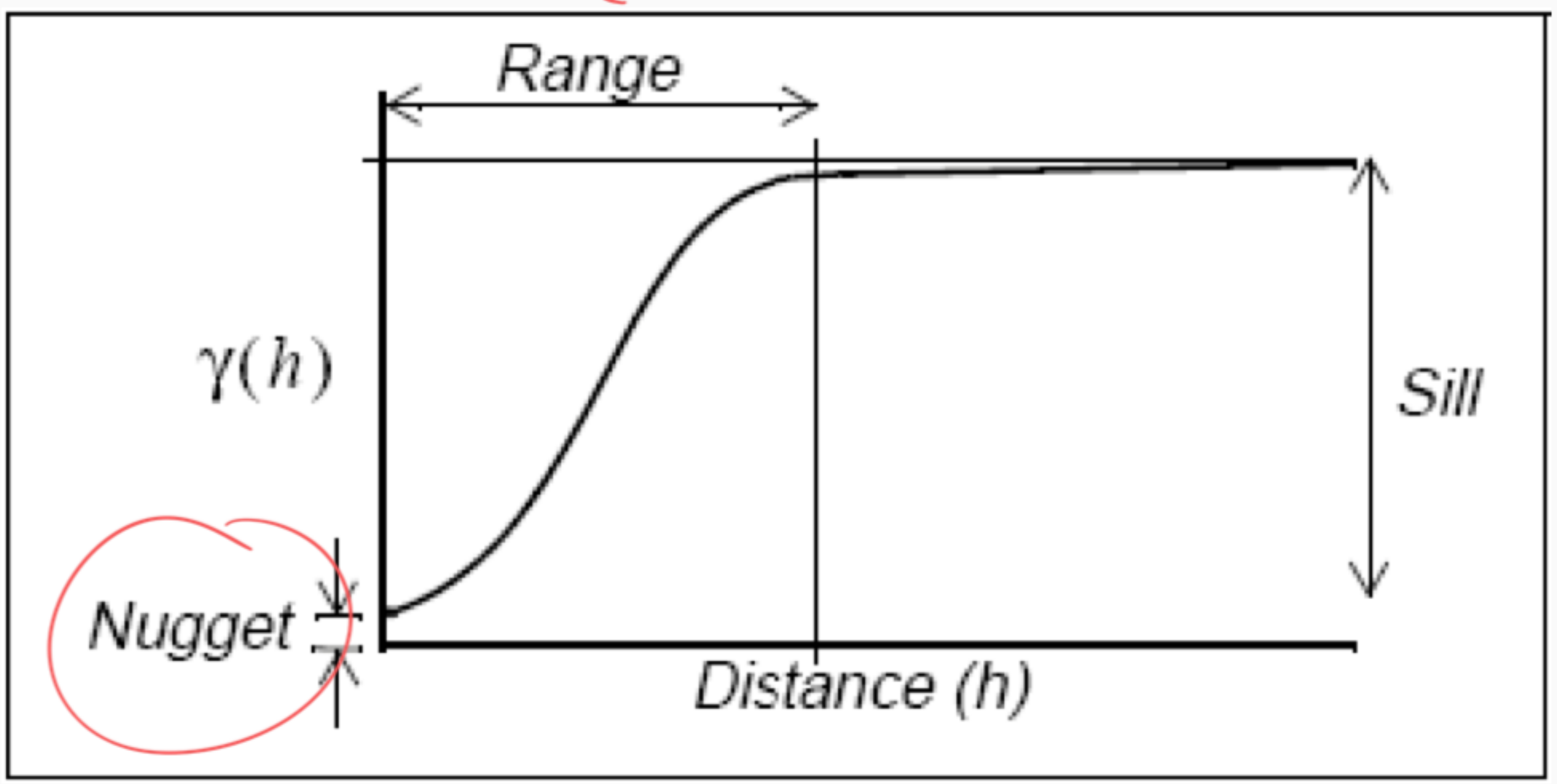


Variogram features

Nugget → σ_ν^2

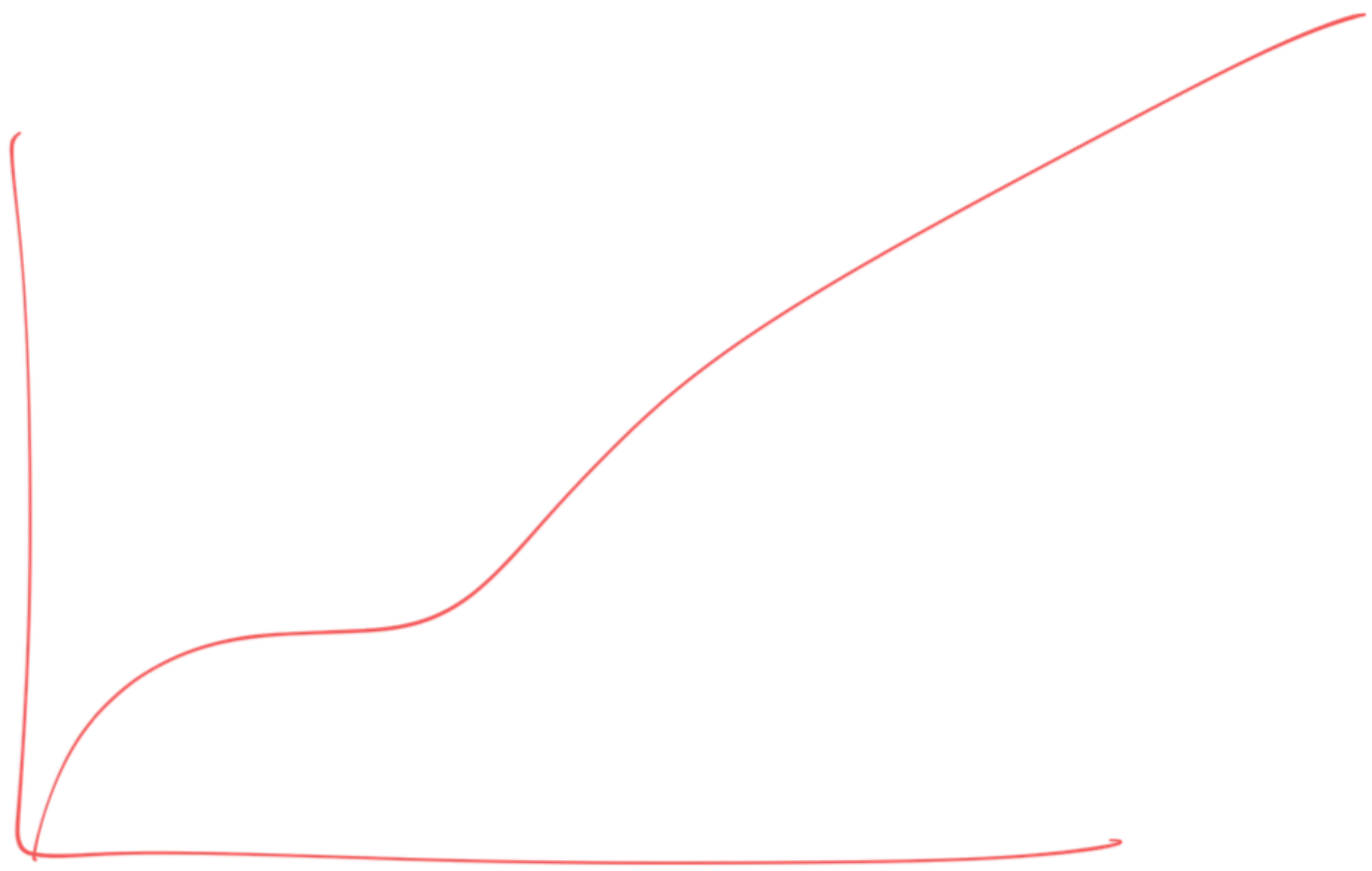
Sill → $\sigma_\nu^2 + \sigma^2$

Range $d = \frac{\sqrt{3}}{\ell}$ (sq exp) $\quad d = \frac{3}{\ell}$ (exp)



$$y(t) = \mu(t) + w(t) + \nu$$

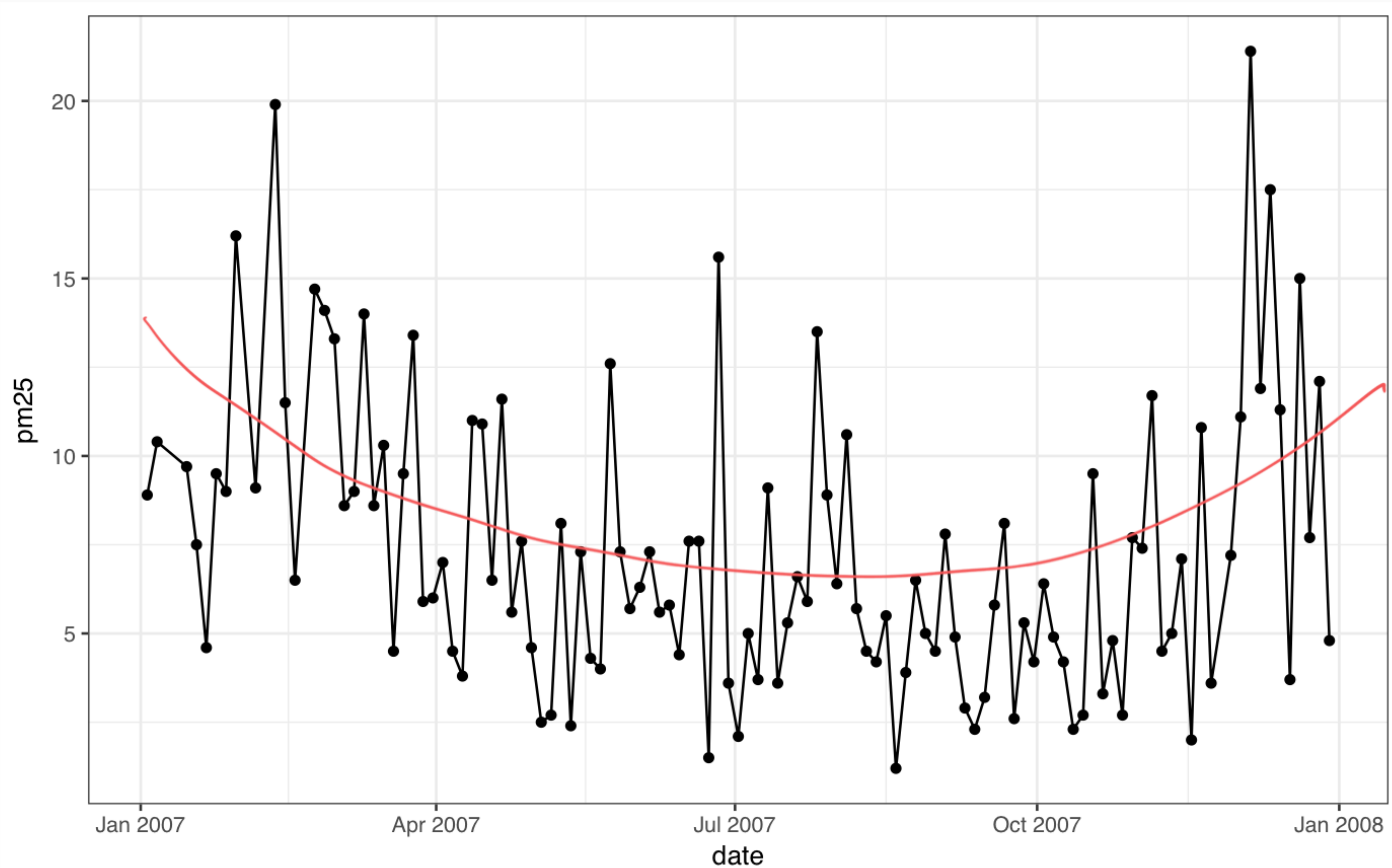
$w \sim GP(0, \Sigma)$ $\quad \nu \sim N(0, \sigma_\nu^2)$



PM2.5 Example

FRN Data

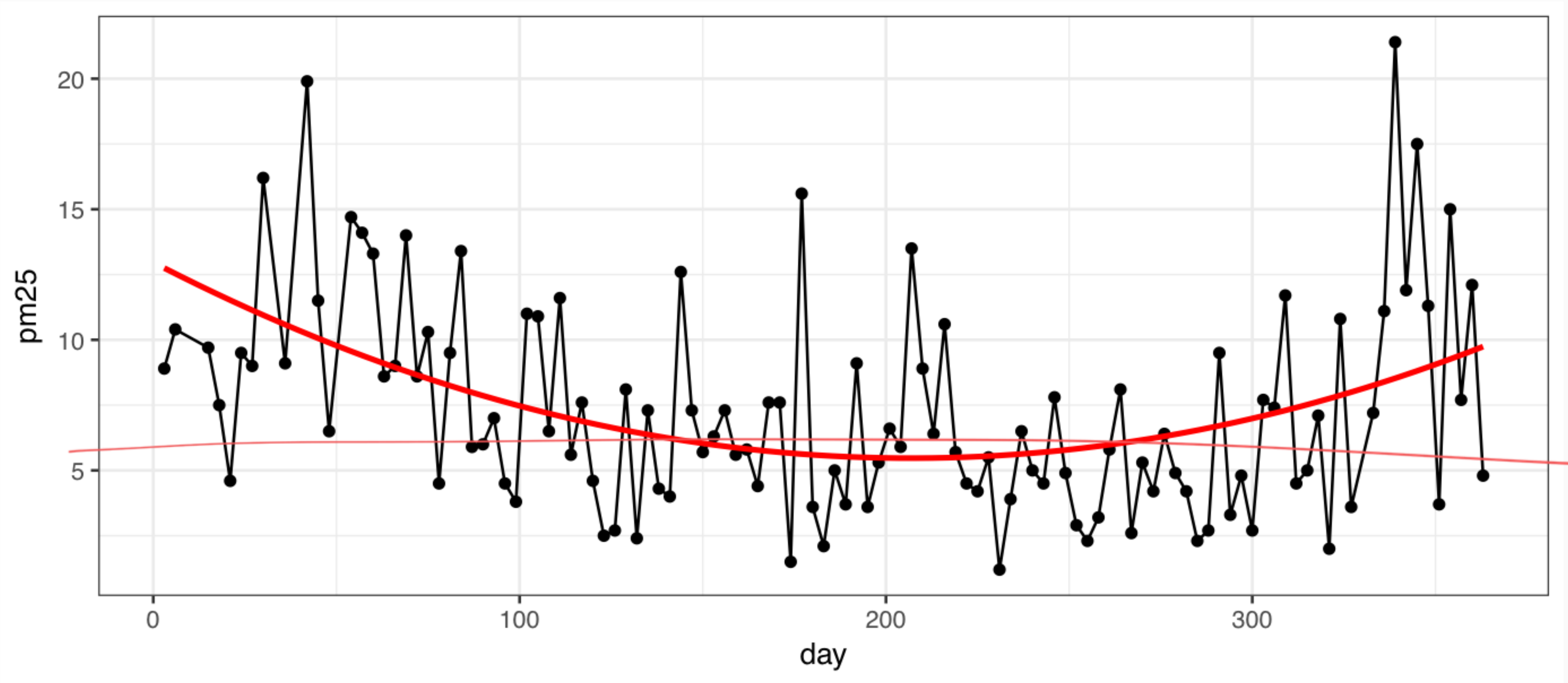
Measured PM2.5 data from an EPA monitoring station in Columbia, NJ.



FRN Data

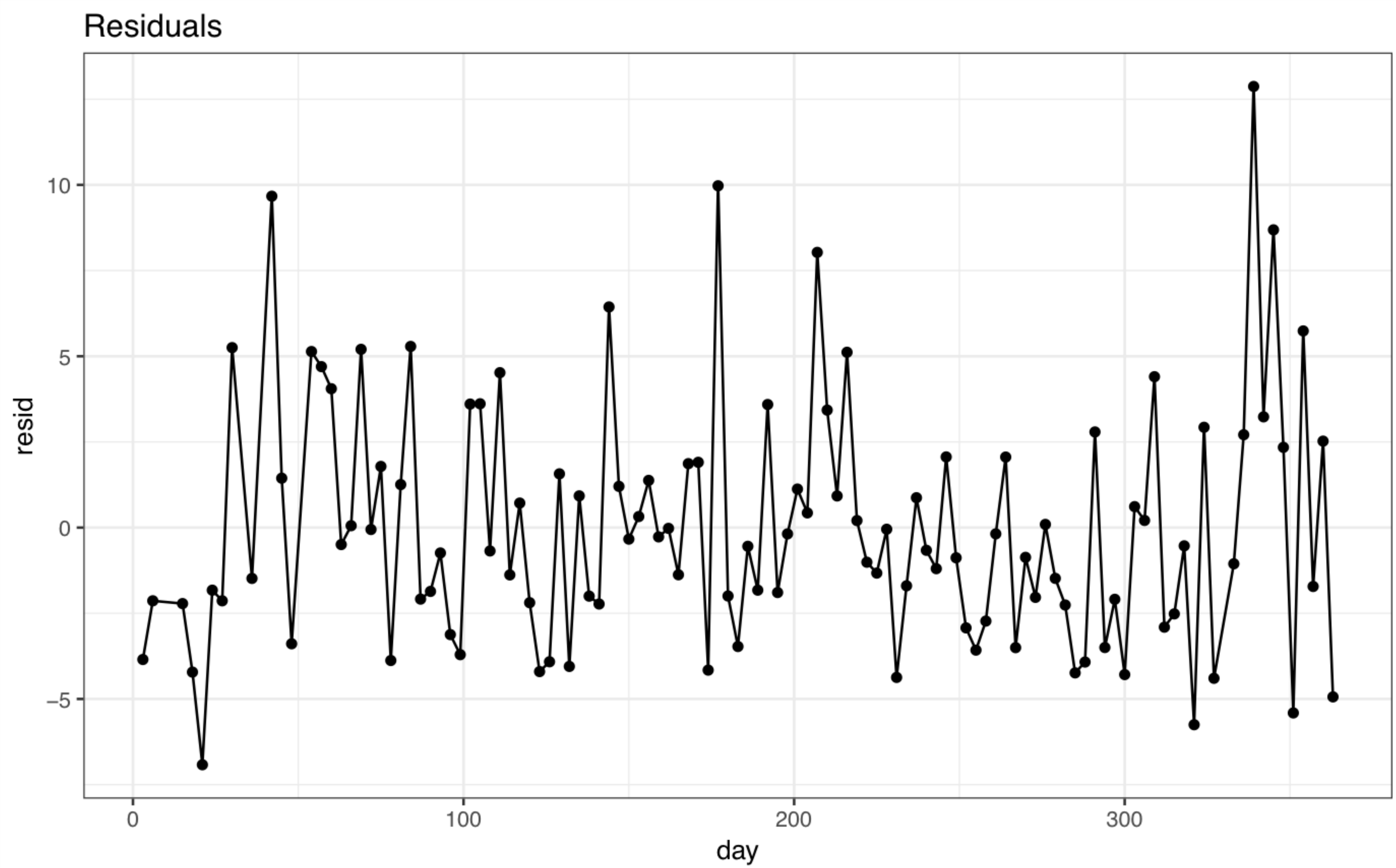
site	latitude	longitude	pm25	date	day
230031011	46.682	-68.016	8.9	2007-01-03	3
230031011	46.682	-68.016	10.4	2007-01-06	6
230031011	46.682	-68.016	9.7	2007-01-15	15
230031011	46.682	-68.016	7.5	2007-01-18	18
230031011	46.682	-68.016	4.6	2007-01-21	21
230031011	46.682	-68.016	9.5	2007-01-24	24
230031011	46.682	-68.016	9.0	2007-01-27	27
230031011	46.682	-68.016	16.2	2007-01-30	30
230031011	46.682	-68.016	9.1	2007-02-05	36
230031011	46.682	-68.016	19.9	2007-02-11	42
230031011	46.682	-68.016	11.5	2007-02-14	45
230031011	46.682	-68.016	6.5	2007-02-17	48
230031011	46.682	-68.016	14.7	2007-02-23	54
230031011	46.682	-68.016	14.1	2007-02-26	57
230031011	46.682	-68.016	13.3	2007-03-01	60
230031011	46.682	-68.016	8.6	2007-03-04	63
230031011	46.682	-68.016	9.0	2007-03-07	66
230031011	46.682	-68.016	14.0	2007-03-10	69
230031011	46.682	-68.016	8.6	2007-03-13	72
230031011	46.682	-68.016	10.3	2007-03-16	75

Mean Model

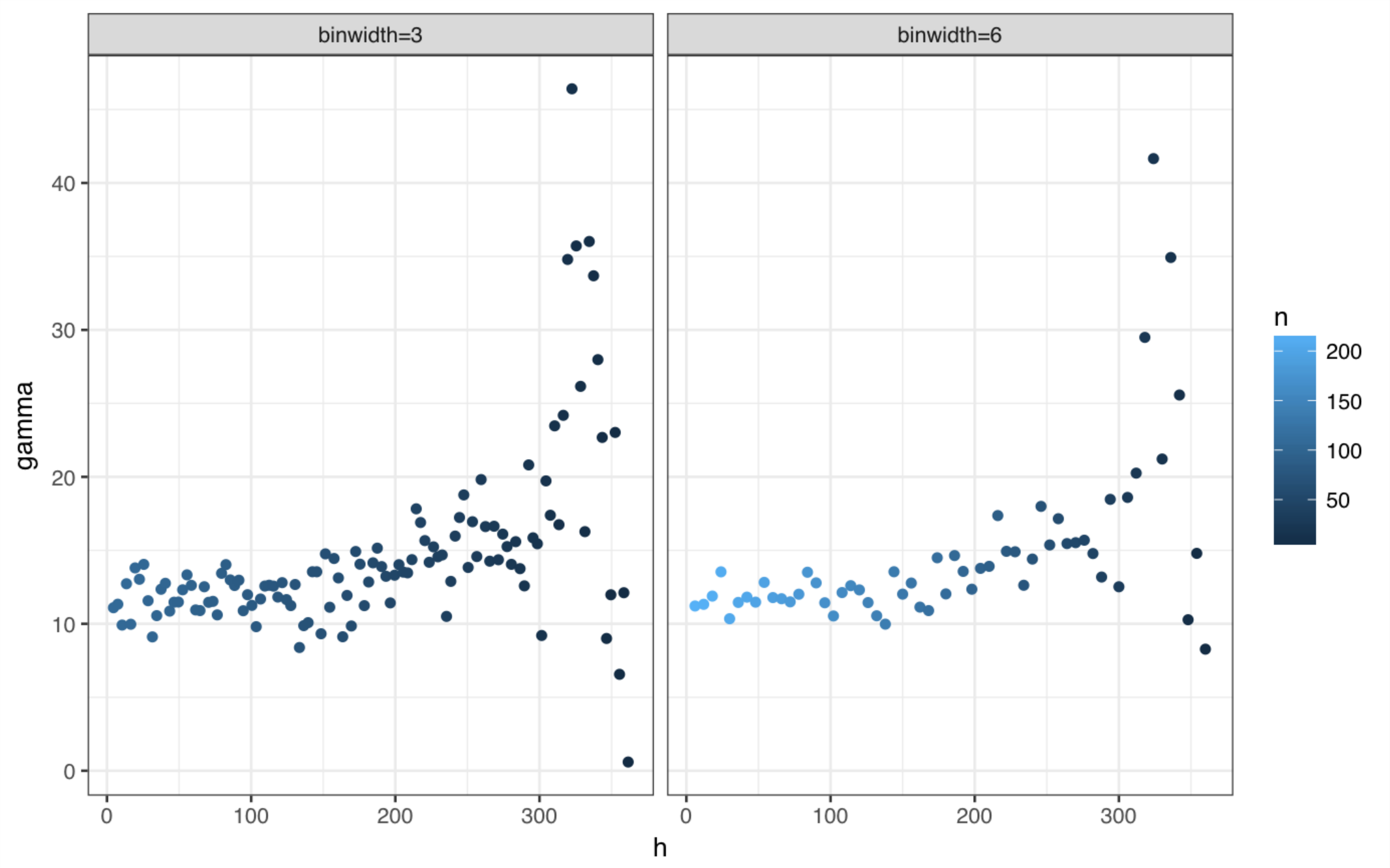



```
##
## Call:
## lm(formula = pm25 ~ day + I(day^2), data = pm25)
##
## Coefficients:
## (Intercept)          day     I(day^2)
##  12.9644351   -0.0724639    0.0001751
##
## Call:
## lm(formula = pm25 ~ day + I(day^2), data = pm25)
```

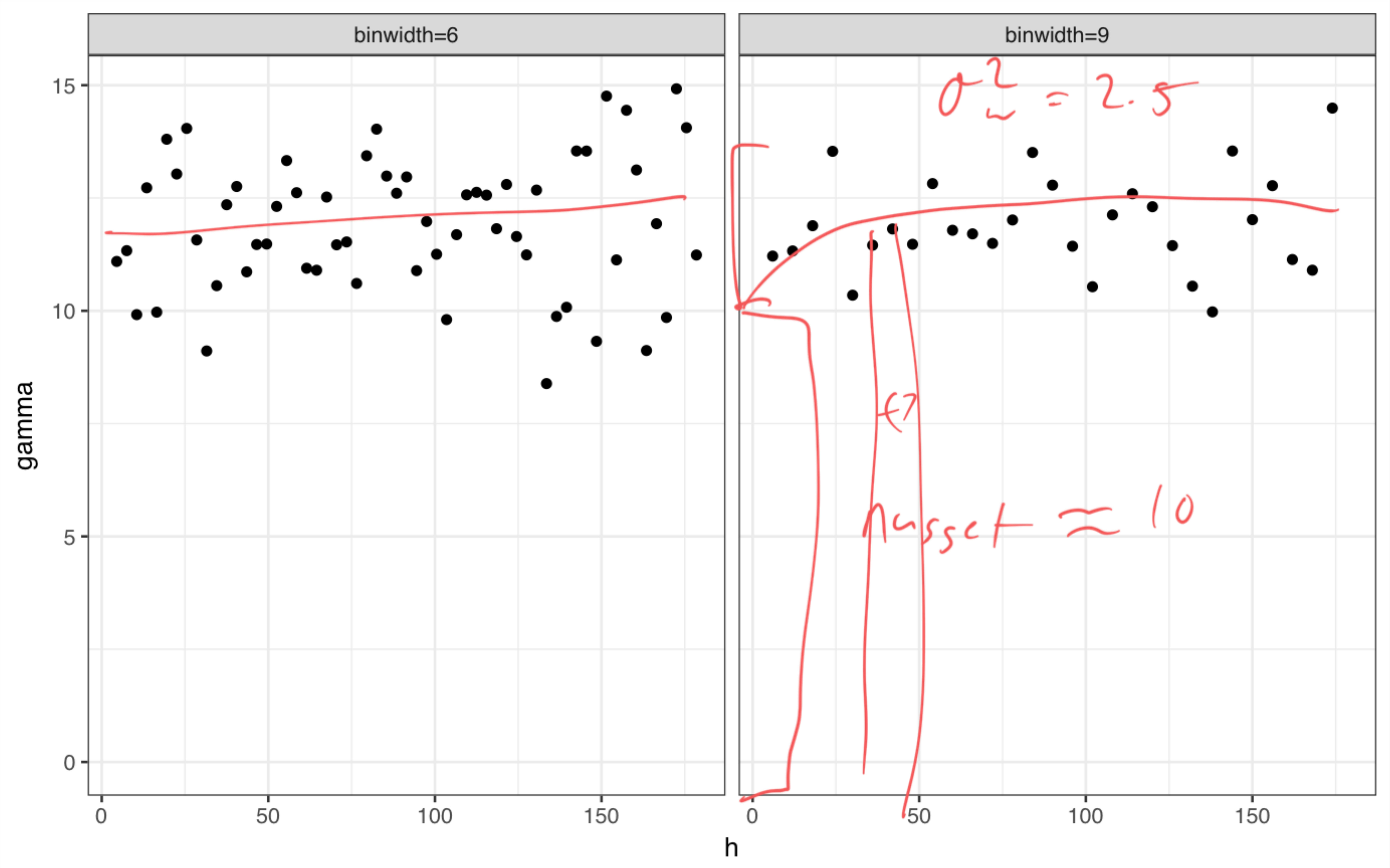
Detrended Residuals



Empirical Variogram



Empirical Variogram



Model

What does the model we are trying to fit actually look like?

Model

What does the model we are trying to fit actually look like?

$$y(d) = \mu(d) + w(d) + w$$

where

$$\mu(d) = \beta 0 + \beta_1 d + \beta_2 d^2$$
$$w(d) \sim \mathcal{GP}(0, \Sigma)$$
$$w \sim \mathcal{N}(0, \sigma_w^2)$$

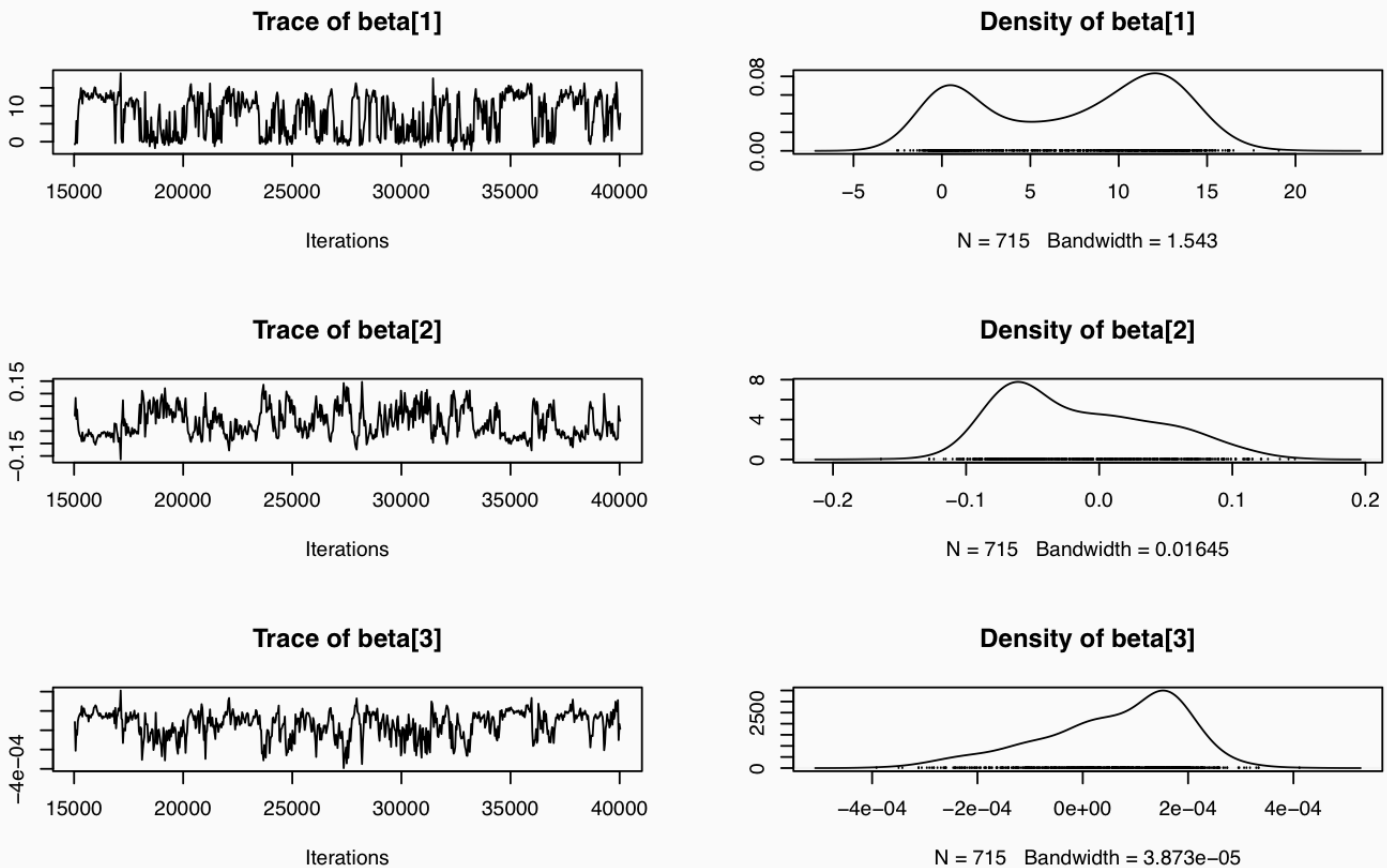
$$y(d) \sim GP\left(\mu(d), \Sigma + \sigma_w^2 I\right)$$

JAGS Model

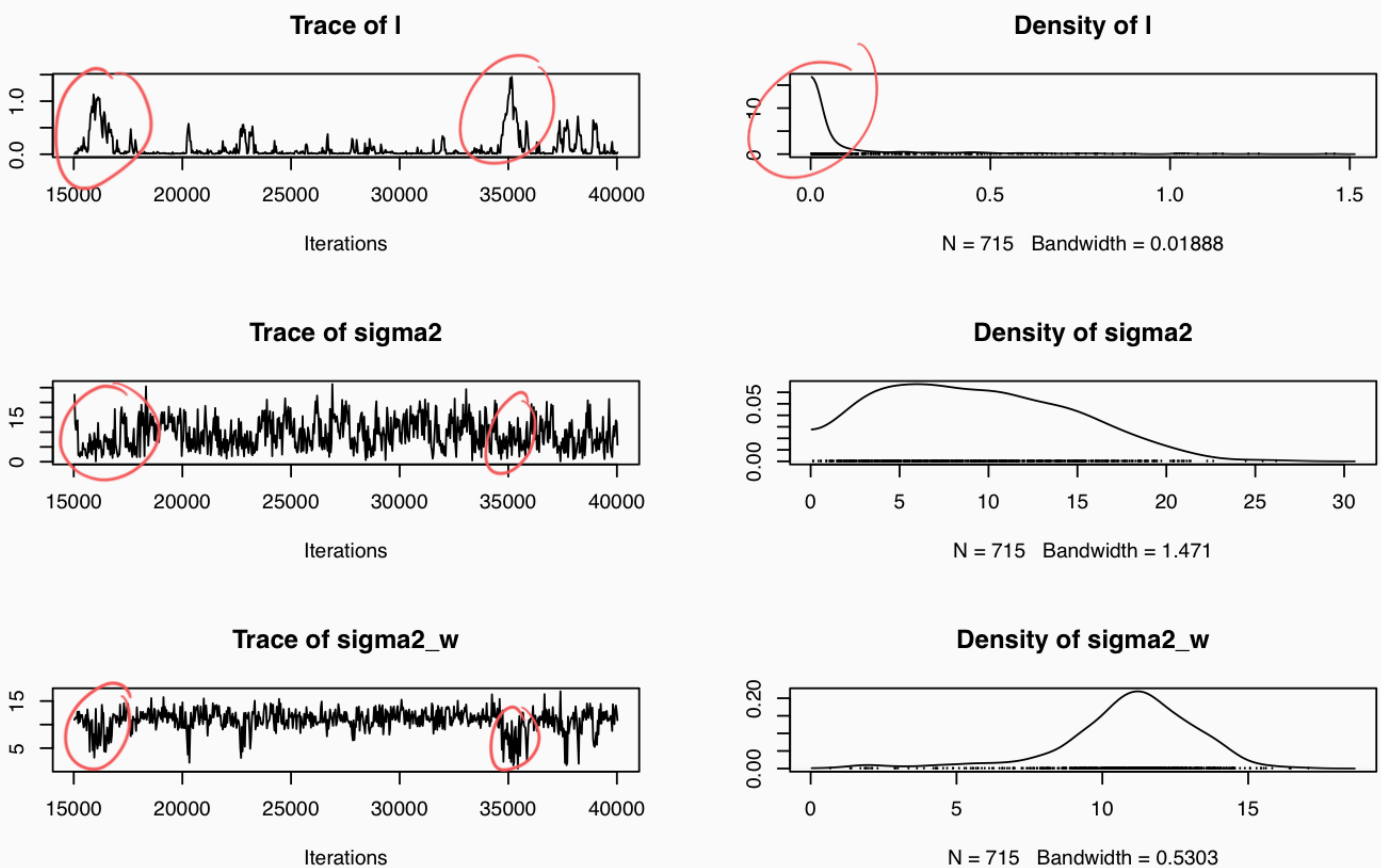
```
## model{
##   y ~ dmnorm(mu, inverse(Sigma))
##
##   for (i in 1:N) {
##     mu[i] <- beta[1]+ beta[2] * x[i] + beta[3] * x[i]^2
##   }
##
##   for (i in 1:(N-1)) {
##     for (j in (i+1):N) {
##       Sigma[i,j] <- sigma2 * exp(- pow(l*d[i,j],2))
##       Sigma[j,i] <- Sigma[i,j]
##     }
##   }
##
##   for (k in 1:N) {
##     Sigma[k,k] <- sigma2 + sigma2_w
##   }
##
##   for (i in 1:3) {
##     beta[i] ~ dt(0, 2.5, 1)
##   }
##   sigma2_w ~ dnorm(10, 1/25) T(0,)
##   sigma2   ~ dnorm(10, 1/25) T(0,)
##   l        ~ dt(0, 2.5, 1) T(0,)
## }
```

Sq Exp

Posterior - Betas



Posterior - Covariance Parameters



Posterior

```
## # A tibble: 6 × 5
##      param     post_mean      post_med    post_lower   post_upper
## *    <chr>         <dbl>         <dbl>         <dbl>        <dbl>
## 1  beta[1]  7.283488e+00  8.667009e+00 -0.7461648059 1.503065e+01
## 2  beta[2] -1.627421e-02 -2.817415e-02 -0.0988863015 1.026401e-01
## 3  beta[3]  5.858818e-05  8.569993e-05 -0.0002481874 2.567976e-04
## 4        l  1.277712e-01  2.433287e-02  0.0060909947 8.443888e-01
## 5   sigma2  9.379213e+00  9.016621e+00  1.5643832453 1.979094e+01
## 6 sigma2_w  1.088809e+01  1.116626e+01  4.2665826402 1.448447e+01
```

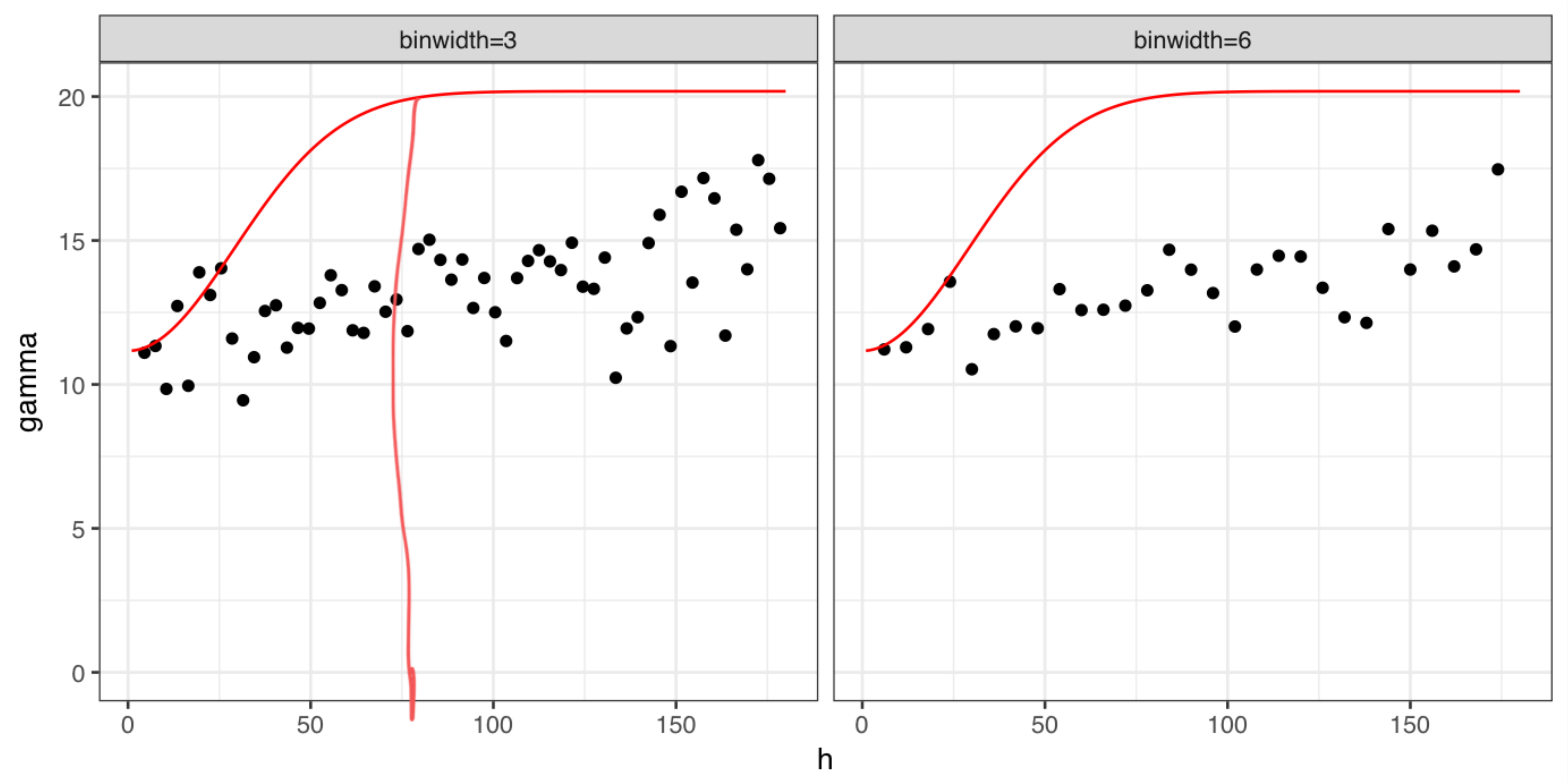
Fitted Variogram

$$Cov(h) = \sigma^2 e^{(-\ell^2 h^2)} + \sigma_v^2 1_{h=0}$$

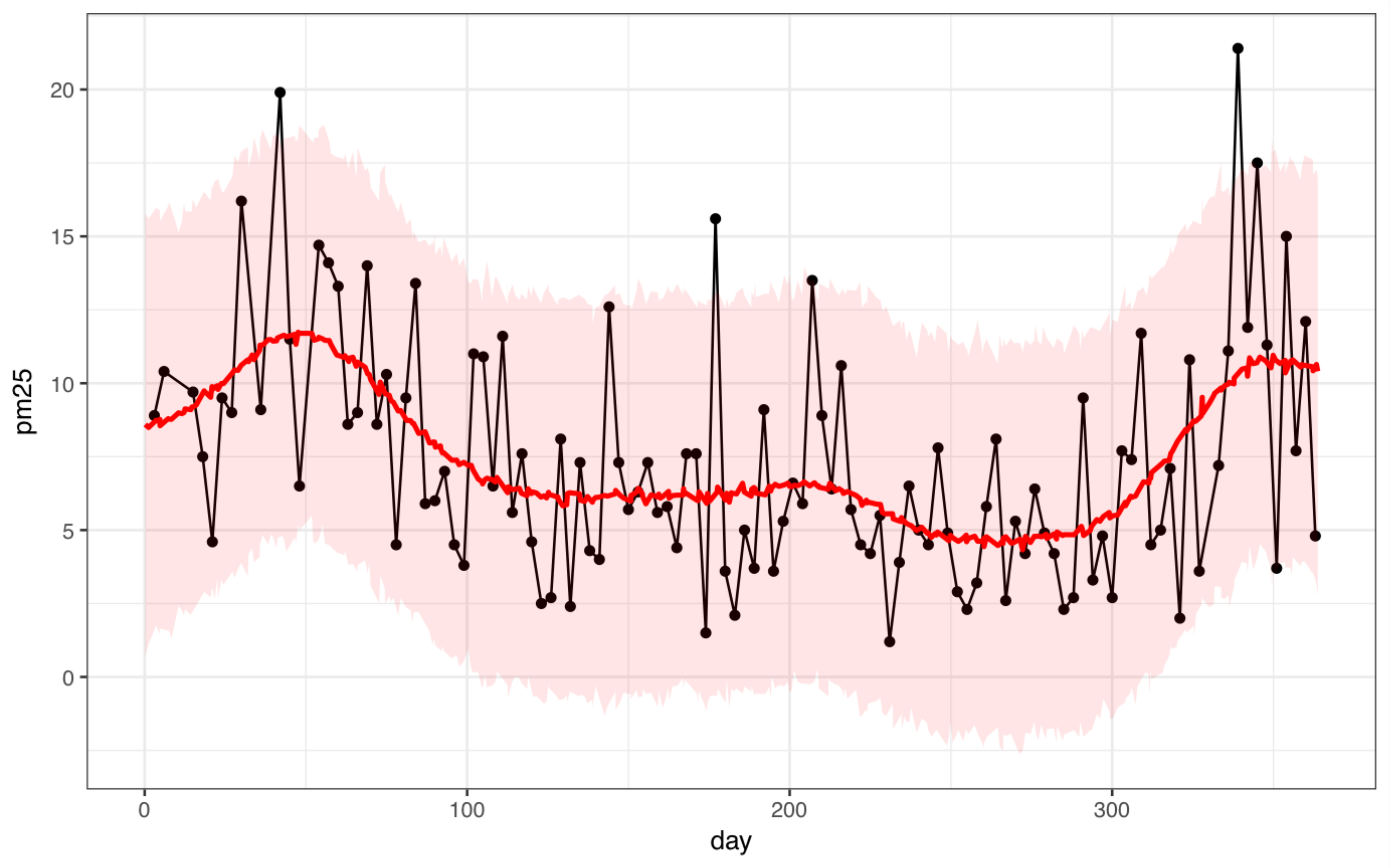
$$\gamma(h) = Cov(0) - Cov(h)$$

$$= \sigma^2 + \sigma_v^2 - \sigma^2 e^{-\ell^2 h^2} - \sigma_v 1_{h=0}$$

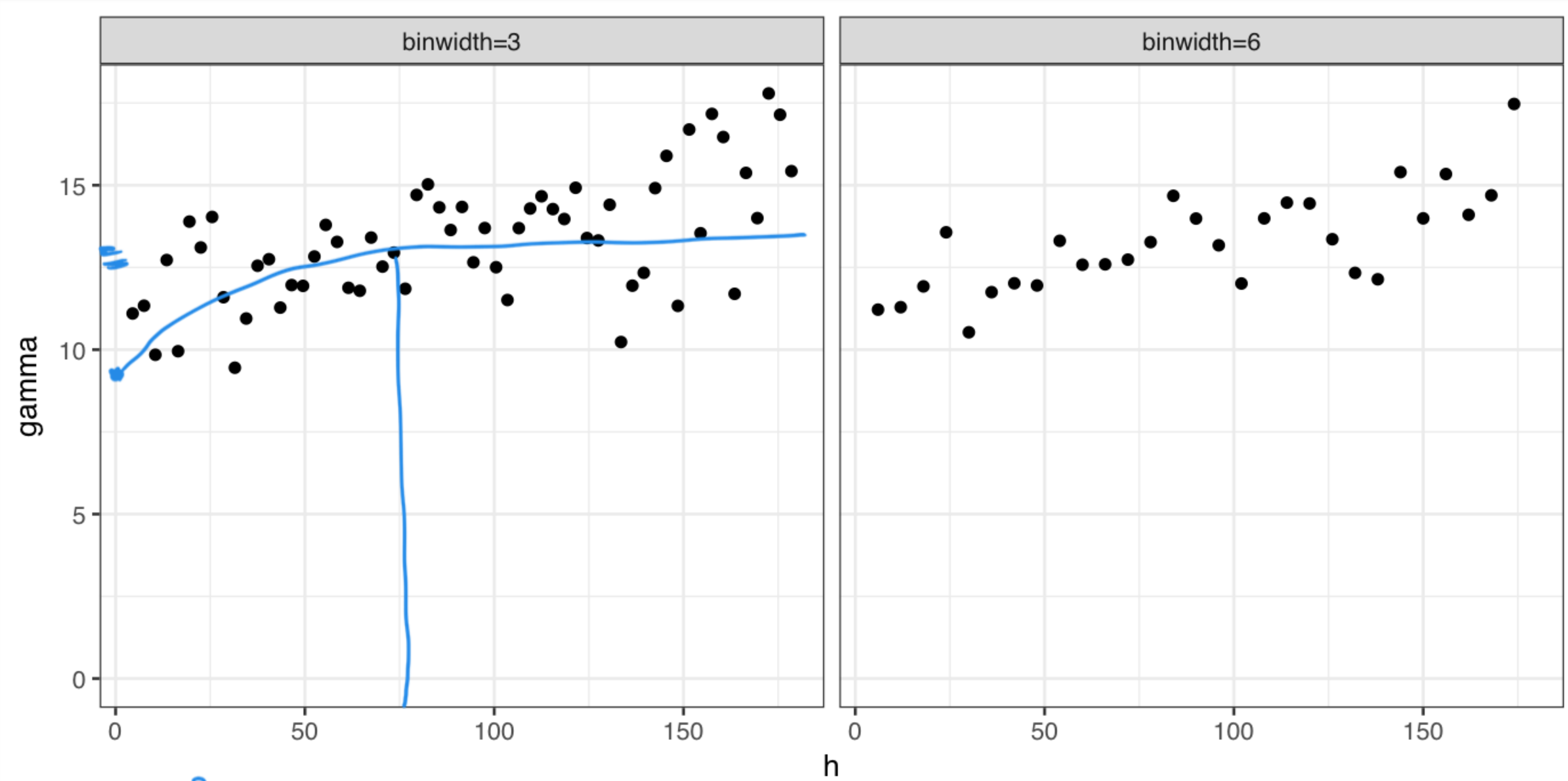
Empirical + Fitted Variogram



Fitted Model + Predictions

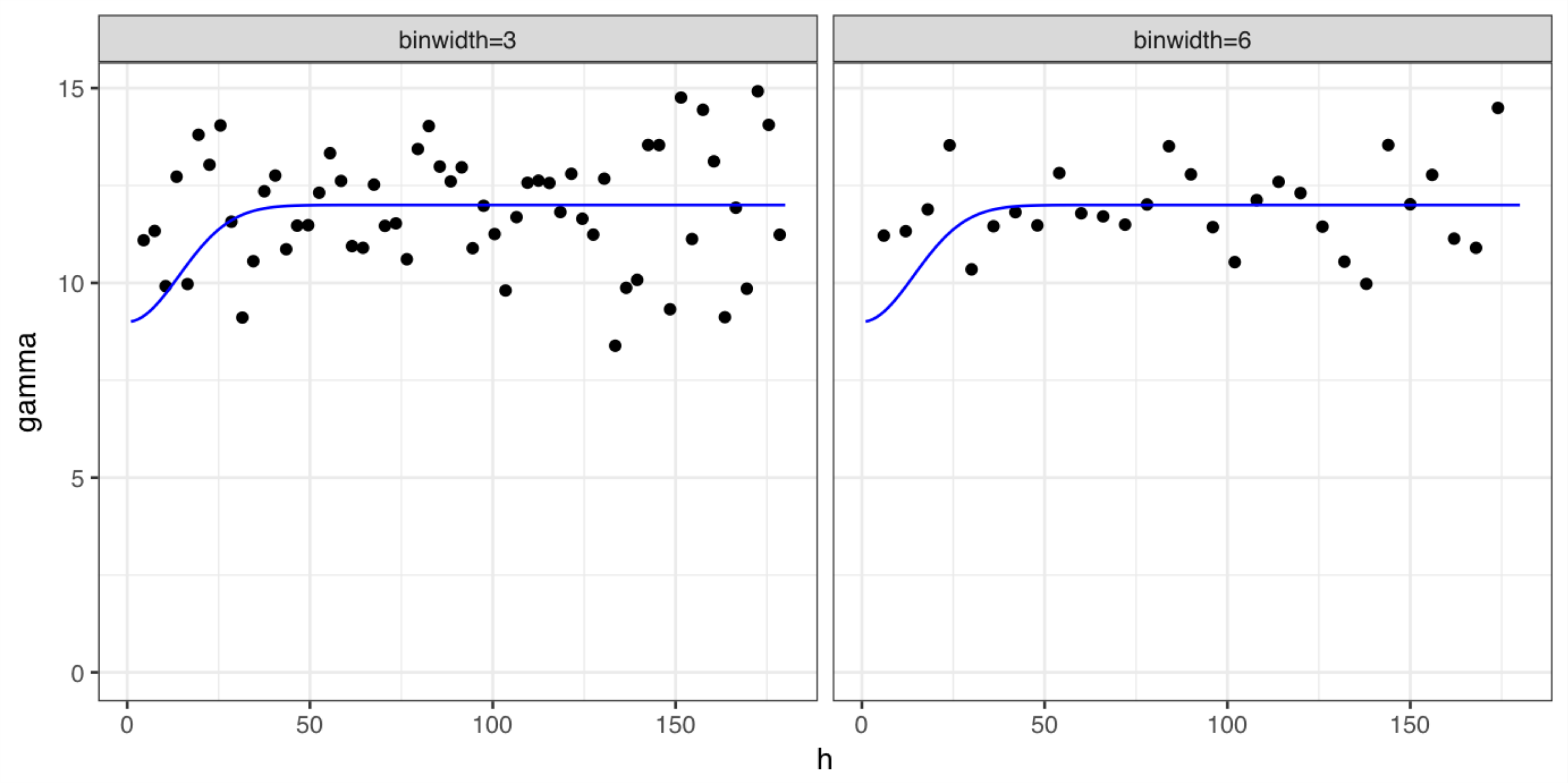


Empirical Variogram (again)



$\sigma_\nu^2 = 9$ $\quad$ sill $= 12$ $\quad$ $\Rightarrow \sigma^2 = 3$ $\quad$ $l = 0.05$

Empirical Variogram Model



Empirical Variogram Model + Predictions