Spatial Statistics

Session 2

Madlene Nussbaum
m.nussbaum@uu.nl
Department of Physical Geography, Utrecht University

Andreas Papritz
(ehem. ETHZ)

15 Oct 2024

1 Geostatistics: theory, variogram properties, maximum likelihood, prediction

Overview

- Stochastic process
- Realization of stochastic process
- Variogram functions
- Sample variogram and fitting of variogram function
- Maximum likelihood estimation of model parameters
- Model selection/inference
- Predictions for ordinary and universal/external drift kriging

2 Stochastic process

2.1 Terminology and model notation (session 1)

Model for data: $Y_i = S(\mathbf{x}_i) + Z_i$

where

Y_i : i^{th} datum

$S(\mathbf{x}_i)$: “signal” (= true quantity) at location $\mathbf{x}_i$

Z_i : iid. random measurement error

Decomposition of signal into trend $\mu(\mathbf{x}_i)$ and stochastic fluctuation:

$$S(\mathbf{x}_i) = \mu(\mathbf{x}_i) + E(\mathbf{x}_i)$$

where commonly a linear model is used for $\mu(\mathbf{x}_i)$

$$\mu(\mathbf{x}_i) = \sum_k d_k(\mathbf{x}_i)\beta_k = \mathbf{d}(\mathbf{x}_i)^{\mathrm{T}}\beta$$

with $d_k(\mathbf{x}_i)$ denoting (spatial) covariates and $\{E(\mathbf{x}_i)\}$ a zero mean stochastic process (random field).

2.2 Realization of stochastic process

- Spatial phenomena obey laws of physics ⇒ are deterministic, have physical causes
- Numerous processes and interactions thereof produce current complex outcome
- Variation appears random ⇒ in geostatistical terms viewed as a random process
- e.g. rainfall pattern, soil properties, air pollution as a realization of a random process
- Each location x is associated with a suite of values with a known distribution
- Actual value observed at x is viewed as one value of this distribution, allocated at random
- Random function has no mathematical description, but "structure" in the sense of correlation in space (or time) and describes relation of random values at different locations to each other

> **Spatial stochastic process (random process)**
>
> $\{S(\mathbf{x})\}$: Collection (= set) of random variables $S(\mathbf{x}) : \mathbf{x} \in D \subset \mathbb{R}^d$, at location x in area D, with a well defined joint distribution

Stationary and isotropic stochastic processes

Stationarity: Assumption that allows to treat data with same degree of variation over region of interest.

Strictly stationary process: Joint distributions of arbitrary collections of random variables $\{S(\mathbf{x}_1), \ldots, S(\mathbf{x}_n)\}$ are invariant to translations by vector $\mathbf{h} \in \mathbb{R}^d$

$\{S(\mathbf{x}_1), \ldots, S(\mathbf{x}_n)\}$ and $\{S(\mathbf{x}_1 + \mathbf{h}), \ldots, S(\mathbf{x}_n + \mathbf{h})\}$ have the same joint distribution:

$$F(s_1, \ldots, s_n; \mathbf{x}_1, \ldots, \mathbf{x}_n) = F(s_1, \ldots, s_n; \mathbf{x}_1 + \mathbf{h}, \ldots, \mathbf{x}_n + \mathbf{h})$$

Isotropic: Weakly stationary process that is invariant to rotations (opposite: anisotropic).

Gaussian stochastic process: All joint and conditional distributions are normal.

Second-order stationary stochastic processes

Weakly or second-order stationary process:

Distributions of arbitrary pairs of random variables $(S(\mathbf{x}), S(\mathbf{x}+\mathbf{h}))$ satisfy:

1. $\mathbb{E}[S(\mathbf{x})] = \text{constant}$ (independent of $\mathbf{x}$)
2. $\text{Cov}(S(\mathbf{x}+\mathbf{h}), S(\mathbf{x})) = \gamma(\mathbf{h})$ (independent of $\mathbf{x}$)
3. $\text{Var}(S(\mathbf{x})) = \text{constant}$ (independent of $\mathbf{x}$)

⇒ Covariance depends on h and only on h, the separation between samples in both distance and direction

⇒ Strict stationarity implies weak stationarity.

⇒ Stationarity is required for estimation/prediction with a single realization of the stochastic process.

2.3 Covariance function and variogram

Definition of variogram $V(\mathbf{h})$ and covariance function $\gamma(\mathbf{h})$:

$$V(\mathbf{h}) = \frac{1}{2}\mathrm{Var}\left(S(\mathbf{x}+\mathbf{h}) - S(\mathbf{x})\right)$$

$$\gamma(\mathbf{h}) = \mathrm{Cov}\left(S(\mathbf{x}+\mathbf{h}), S(\mathbf{x})\right)$$

Relation between variogram and covariance function:

$$V(\mathbf{h}) = \gamma(0) - \gamma(\mathbf{h}), \quad \text{with} \quad \gamma(0) = \mathrm{Var}(S(\mathbf{x}))$$

⇒ Variogram is preferred.

Variogram is based on differences only (no distribution mean as in covariance). Allows to relax stationarity assumptions even further, instead of $\mathbb{E}[S(\mathbf{x})] = \mathrm{constant}$, we can assume (intrinsic stationarity):

$$\mathbb{E}[S(\mathbf{x}+\mathbf{h}) - S(\mathbf{x})] = 0$$

Variogram and correlogram

Relation between correlogram and covariance function for weakly stationary process:

$$\rho(\mathbf{h}) = \frac{\gamma(\mathbf{h})}{\gamma(0)}$$

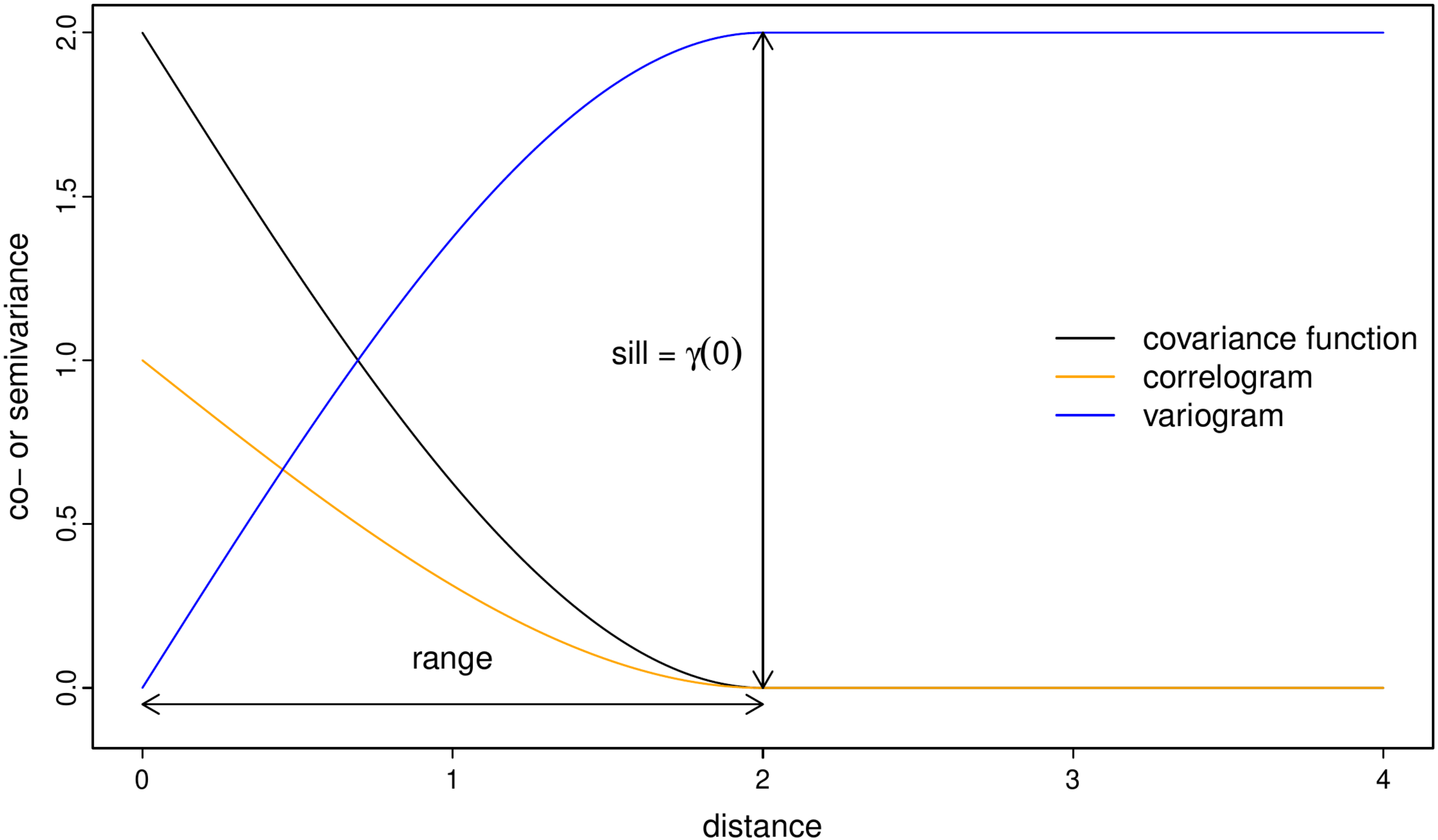
Relation between variogram and correlogram:

$$V(\mathbf{h}) = \gamma(0)(1 - \rho(\mathbf{h}))$$

Symmetry:

$$V(\mathbf{h}) = V(-\mathbf{h}), \quad \gamma(\mathbf{h}) = \gamma(-\mathbf{h}), \quad \rho(\mathbf{h}) = \rho(-\mathbf{h})$$

Covariance and Variogram Plot



Summary stochastic process

- Stochastic process: generalisation of multidimensional random variable
- Stationarity assumption required for estimation from single realisation of stochastic process
- In practice assumption of weak stationarity:
  1. constant mean
  2. constant variance
  3. covariance and semivariance depends only on lag distance but not on location
- Often additional assumption of isotropic auto-correlation

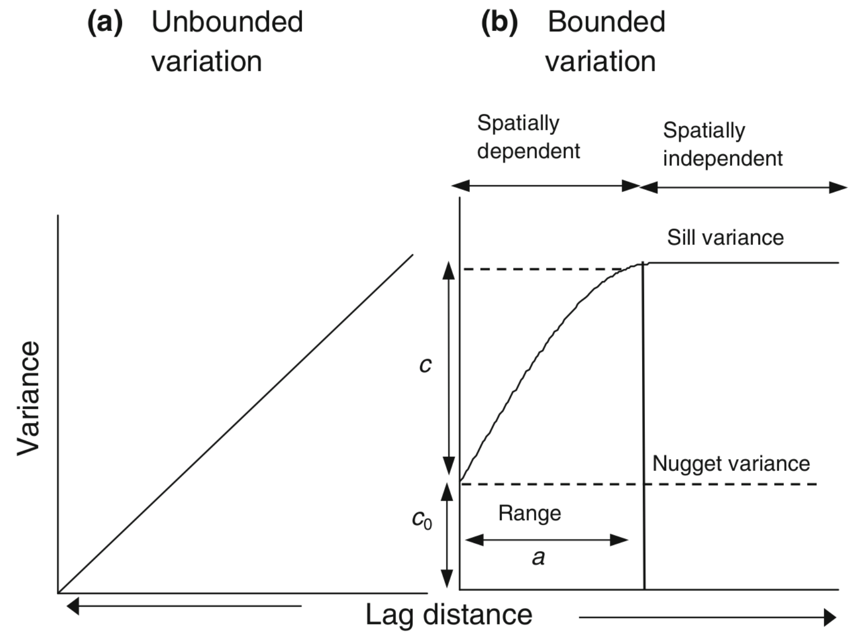
3 Variogram functions and their properties

Principle features of the variogram

- Increase in variance with increasing lag
- Function must guarantee non-negative variances
- **Sill** variance $\mathbf{c_0} + \mathbf{c}$, i.e. an upper bound
- **Range** of spatial correlation $\mathbf{a}$, where auto-correlation becomes 0
- **Nugget** variance $\mathbf{c_0}$, i.e. a positive intercept
- with $\mathbf{c}$ often called **partial sill**, i.e. part of variance with spatial structure up to $\mathbf{a}$

Special cases

- *Anisotropy*, i.e. directional variation depending on angle
- *Unbounded* variogram (not second-order stationary)
- *Pure Nugget* variogram



(Oliver and Webster, 2015, Fig. 3.10)

Pure nugget variogram

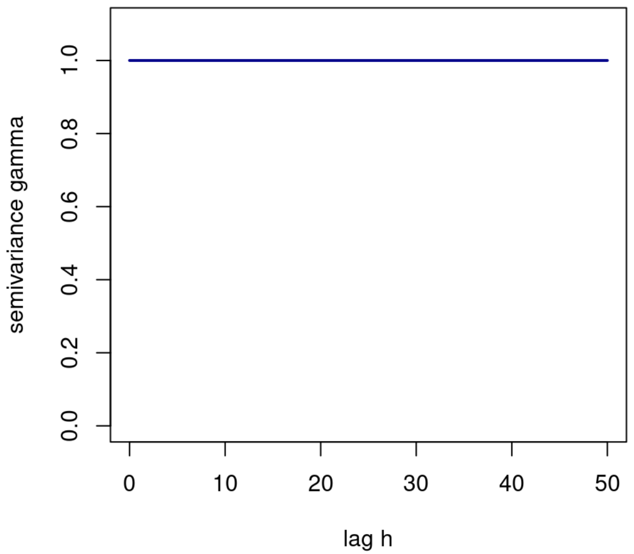
Absence of auto-correlation (nugget effect covariance)

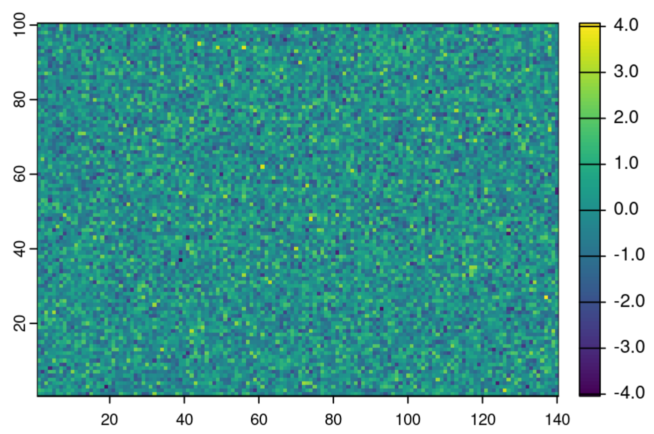
$$V(h) = \begin{cases} 0 & \text{if } h = 0 \\ c_0 & \text{otherwise} \end{cases}$$

Mechanism: measurement error and/or small-scale spatial variation

Pure nugget variogram – realization

```r
library(sp); library(terra); library(gstat)
# unconditional Gaussian simulation on a grid
xy <- expand.grid(1:140, 1:100)
names(xy) <- c("x","y")
gridded(xy) = ~x+y
v.m <- vgm(nugget=1, model = 'Sph', psill = 0, range = 0.0001)
plot(variogramLine(v.m, maxdist = 50), type = "l", ylim = c(0,1.1))
g.sim <- gstat(formula = z~1, dummy = TRUE, beta = 0,   model = v.m, nmax = 100)
r.sim <- predict(g.sim, xy, nsim = 1)
plot(rast(r.sim))
```

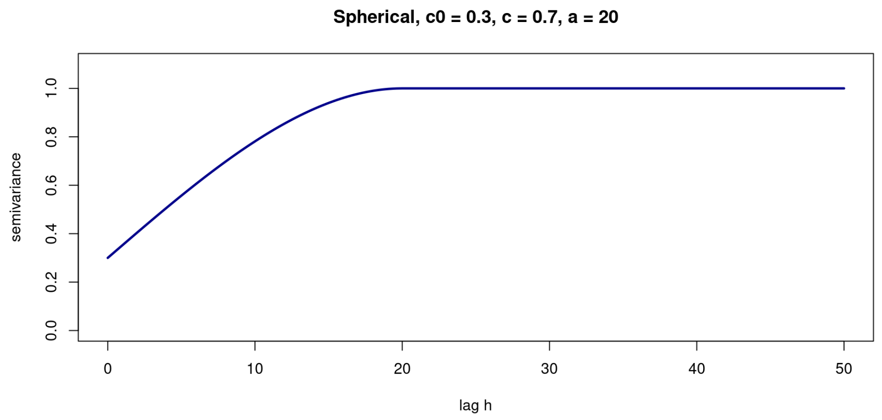




Spherical variogram function

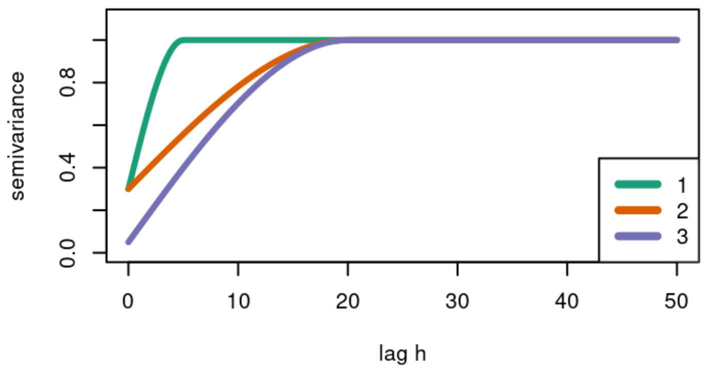
$$V(h) = \begin{cases} c_0 + c\left\{\frac{3h}{2a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right\} & \text{for } 0 < h \leq a \\ c_0 + c & \text{for } h > a \\ c_0 & \text{for } h = 0 \end{cases}$$

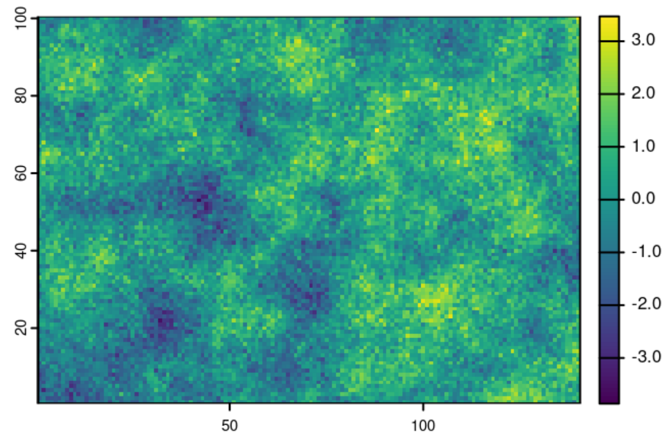
where c_0 is the nugget variance, c the variance of spatially correlated component and a is the range of spatial dependence.

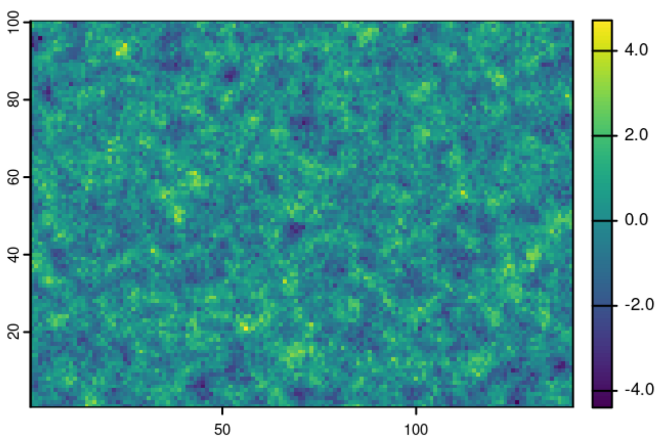


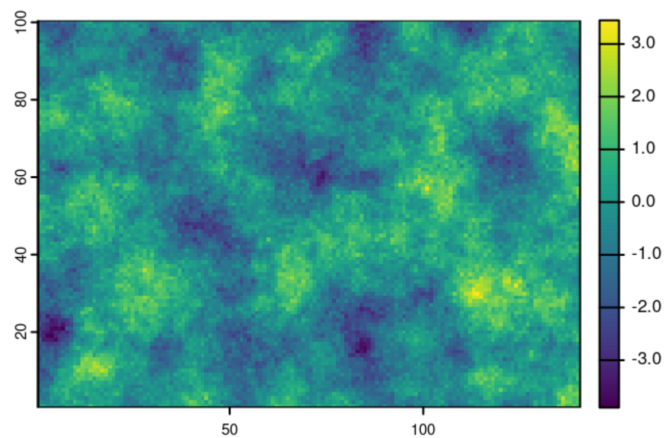
Spherical variogram function – realizations

```r
A <- vgm(nugget=0.3, model = 'Sph', psill = 0.7, range = 20)
B <- vgm(nugget=0.05, model = 'Sph', psill = 0.95, range = 20)
C <- vgm(nugget=0.3, model = 'Sph', psill = 0.7, range = 5)
```







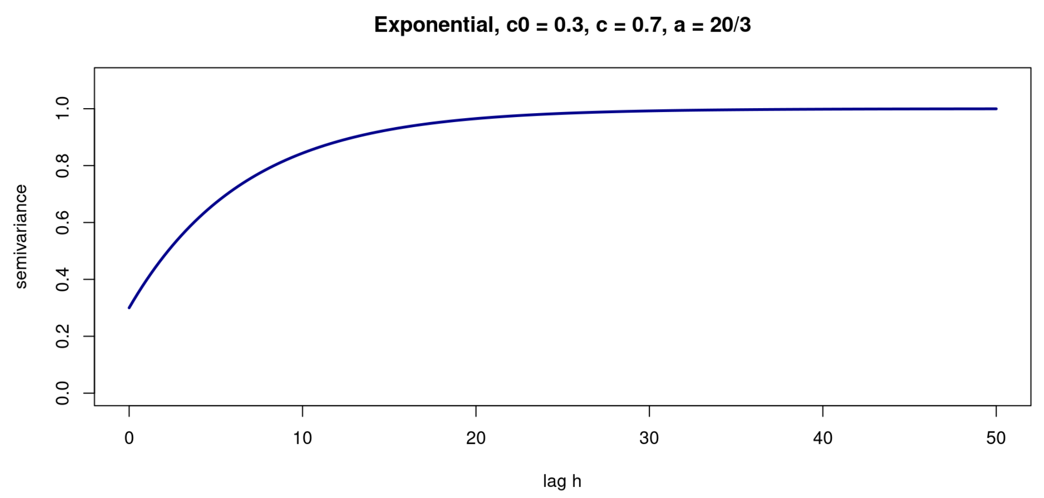


Exponential variogram function

$$V(h) = \begin{cases} c_0 + c\left\{1 - exp\left(\frac{h}{a}\right)\right\} & \text{for } 0 < h \\ c_0 & \text{for } h = 0 \end{cases}$$

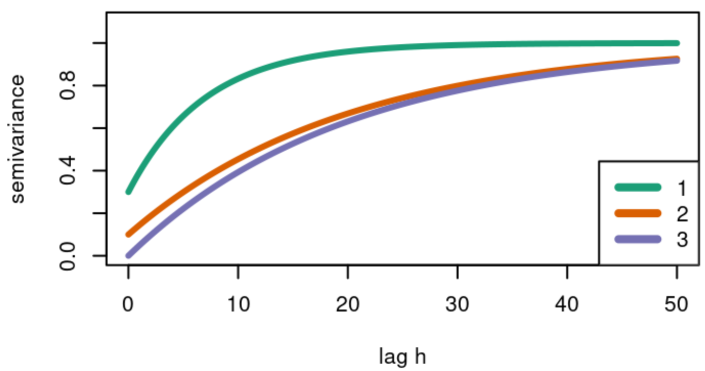
with a here being a distance parameter.

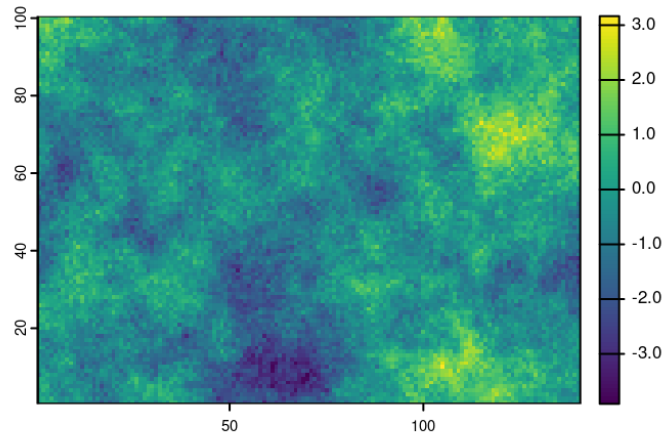
The function approaches the sill asymtotically and does not have a finite range. For practical purposes usually an effective range a' is used which is approximately $3a$.

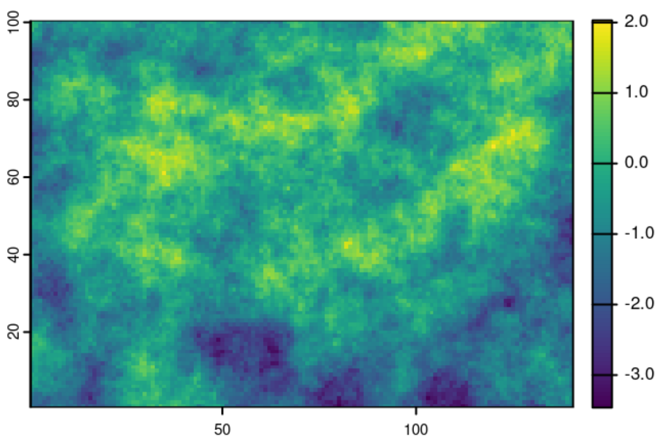


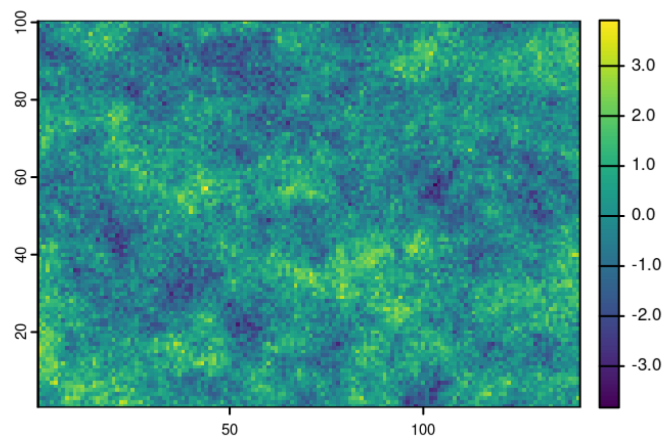
Exponential variogram function – realizations

```r
A <- vgm(nugget=0.3, model = 'Exp', psill = 0.7, range = 7)
B <- vgm(nugget=0.0001, model = 'Exp', psill = 0.999, range = 20)
C <- vgm(nugget=0.1, model = 'Exp', psill = 0.9, range = 20)
```









Matérn (Gaussian) variogram function

$$V(h) = c_0 + c\left\{1 - \frac{1}{2^{v-1}\Gamma(v)}\left(-\frac{h}{a}\right)^{v} K_v\left(\frac{h}{a}\right)\right\}$$

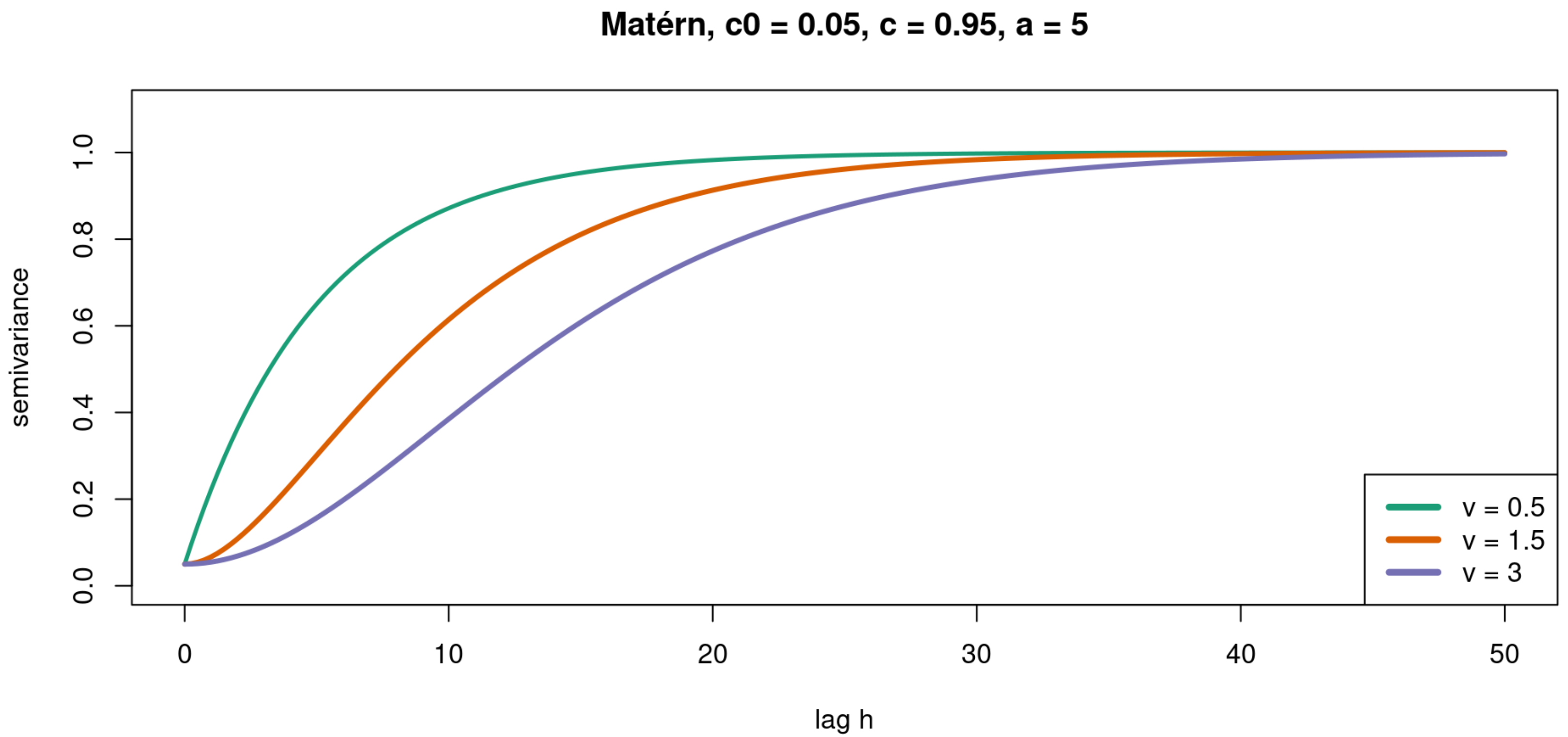
with c_0, c and a being nugget, sill and range parameters as before. here being a distance parameter.

The equation embodies the gamma function Γ with parameter v and the Bessel function of the second kind, K_v for parameter v.

Parameter v describes the smoothness of variation and can vary from 0 (very rough) to infinity (very smooth).

With $v = 0.5$ it becomes the exponential variogram.

Matérn function – Variogram smoothness

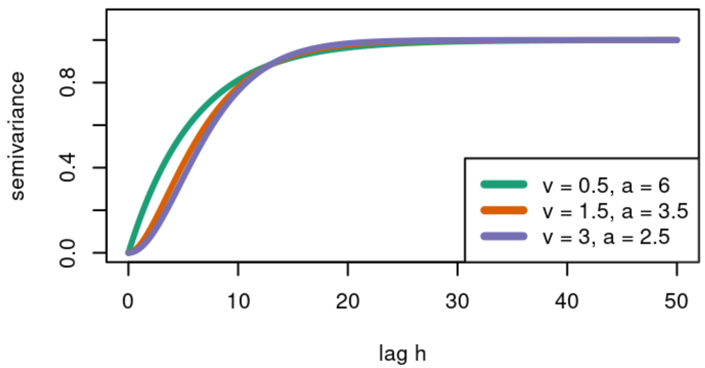


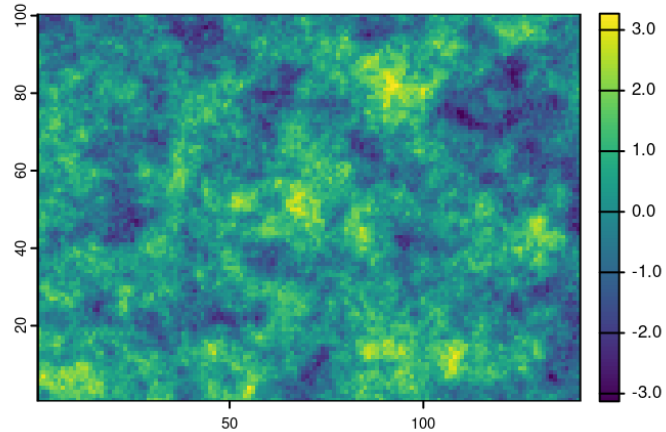
Shape of variogram close to origin controls smoothness of realizations of stochastic processes:

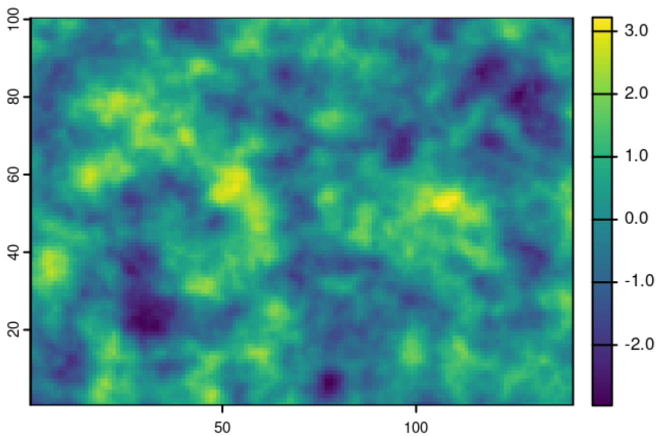
1. Variogram with nugget: realizations non-continuous
2. Variogram grows linearly at origin: realizations continuous, but not everywhere differentiable
3. Variogram grows at at least quadratically at origin: realizations everywhere at least once differentiable

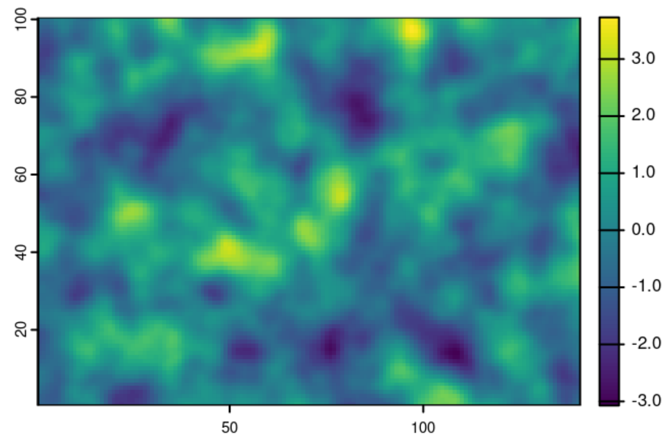
Matérn variogram function – realizations

```
A <- vgm(nugget=0, model = 'Mat', psill = 1, range = 6, kappa = 0.5)
B <- vgm(nugget=0, model = 'Mat', psill = 1, range = 3.5, kappa = 1.5)
C <- vgm(nugget=0, model = 'Mat', psill = 1, range = 2.5, kappa = 3)
```









4 Model for Gaussian spatial data

4.1 Model for Gaussian spatial data

Model for data: $Y_i = S(\mathbf{x}_i) + Z_i = \mu(\mathbf{x}_i) + E(\mathbf{x}_i) + Z_i$

- with Y_i: i^{th} datum; $S(\mathbf{x}_i)$: “signal” (true quantity) at location $\mathbf{x}_i$; $\mu(\mathbf{x}_i)$: trend
- $\{E(\mathbf{x}_i)\}$: a zero-mean Gaussian process, parametrized by covariance function $\gamma(\mathbf{h};\theta)$ or variogram $V(\mathbf{h};\theta)$
- Z_i: iid Gaussian measurement error with variance τ^2

Trend $\mu(\mathbf{x}_i)$ modeled by linear regression model with spatial covariates $d_k(\mathbf{x}_i)$

$$\mu(\mathbf{x}_i) = \sum_k d_k(\mathbf{x}_i)\beta_k = \mathbf{d}(\mathbf{x}_i)^T\beta$$

Unknown elements of the model:

1. Structure and parameters β of the trend model
2. Covariance (or variogram) parameters θ
3. Nugget variance τ^2

4.2 Trend modelling

Ordinary least squares (OLS) trend estimation

Gaussian model in vector notation: $\mathbf{Y} = \mathbf{X}\beta + \mathbf{E} + \mathbf{Z}$

Estimation of trend parameters β by ordinary least squares:

$$\hat{\beta}_{\mathrm{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

For spatially uncorrelated data ($\mathbf{E} = \mathbf{0}$; $\mathrm{Cov}(\mathbf{Y}, \mathbf{Y}^\top) = \tau^2 \mathbf{I}$):

$$\hat{\beta}_{\mathrm{OLS}} \sim \mathcal{N}\left(\beta, \tau^2 (\mathbf{X}^\top \mathbf{X})^{-1}\right)$$

For spatially auto-correlated data
$\mathrm{Cov}(\mathbf{Y}, \mathbf{Y}^\top) = \mathrm{Cov}(\mathbf{Z}, \mathbf{Z}^\top) + \mathrm{Cov}(\mathbf{E}, \mathbf{E}^\top) = \mathbf{\Gamma}_\theta = \tau^2 \mathbf{I} + \mathbf{\Sigma}_\theta$

$$\hat{\beta}_{\mathrm{OLS}} \sim \mathcal{N}\left(\beta, \tau^2 (\mathbf{X}^\top \mathbf{X})^{-1} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Sigma}_\theta \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}\right)$$

Conclusion: Ignoring auto-correlation: $\hat{\beta}_{\mathrm{OLS}}$ unbiased, but the standard errors are too small. Tests based on OLS fit are biased!

Generalized least squares (GLS) trend estimation

Generalized least squares estimates:

$$\hat{\beta}_{\mathrm{GLS}} = (\mathbf{X}^\top \mathbf{\Gamma}_\theta^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Gamma}_\theta^{-1} \mathbf{Y}$$

GLS = OLS with "orthogonalized" response and design matrix.

Sampling distribution:

$$\hat{\beta}_{\mathrm{GLS}} \sim \mathcal{N}\left(\beta, (\mathbf{X}^\top \mathbf{\Gamma}_\theta^{-1} \mathbf{X})^{-1}\right)$$

For spatially uncorrelated data ($\mathbf{\Gamma}_\theta = \tau^2 \mathbf{I}$):

$$\hat{\beta}_{\mathrm{GLS}} = \hat{\beta}_{\mathrm{OLS}}$$

$\hat{\beta}_{\mathrm{GLS}}$ has the smallest standard errors among all linear estimators (Gauss-Markov theorem), hence it's the BLUE (Best Linear Unbiased Estimator).

$\hat{\beta}_{\mathrm{GLS}}$ is the maximum likelihood estimate for Gaussian $\mathbf{Y}$.

Generalised least squares (GLS) is the method of choice for estimating coefficients of the trend model.

4.3 Computing the sample variogram

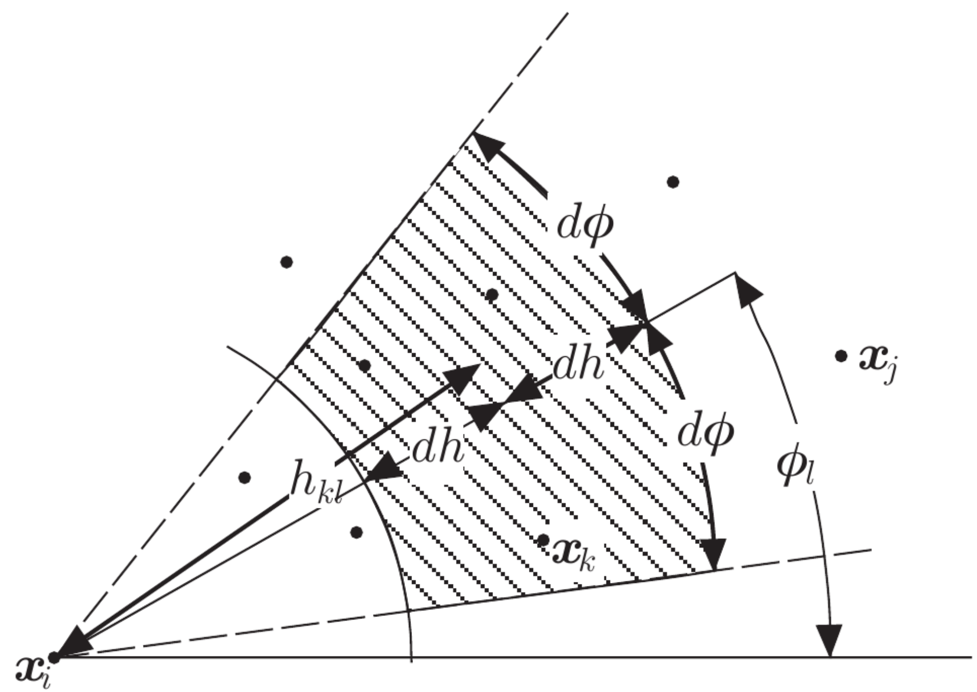
- Sample variogram (also experimental variogram) consists of semivariances at a finite set of discrete lags h.
- Underlying function is continuous for all h.
- Variogram modelling:
  1. Compute sample variogram
  2. Fit smooth function that describes principal features of semivariance sequence

4.4 Computing sample variogram of residuals

Extract residuals $\mathbf{R} = \mathbf{Y} - \mathbf{X}\hat{\beta}$ of the fitted linear model (or use data $\mathbf{Y}$ if the model has constant $\mu(\mathbf{x})$).

Choose bin width dh (and width of angular classes $d\phi$) to define the $(k, l)^{\text{th}}$ lag class, $\mathbf{h}_{kl}$, characterized by:

- Distance: $(h_k - dh, h_k + dh]$
- Angular class: $\phi_l - d\phi, \phi_l + d\phi$



Computing sample variogram – formally

Form all N_{kl} pairs (i, j) with $\mathbf{x}_i - \mathbf{x}_j \approx \mathbf{h}_{kl}$.

Compute for each lag class $\mathbf{h}_{kl}$ the semivariance:

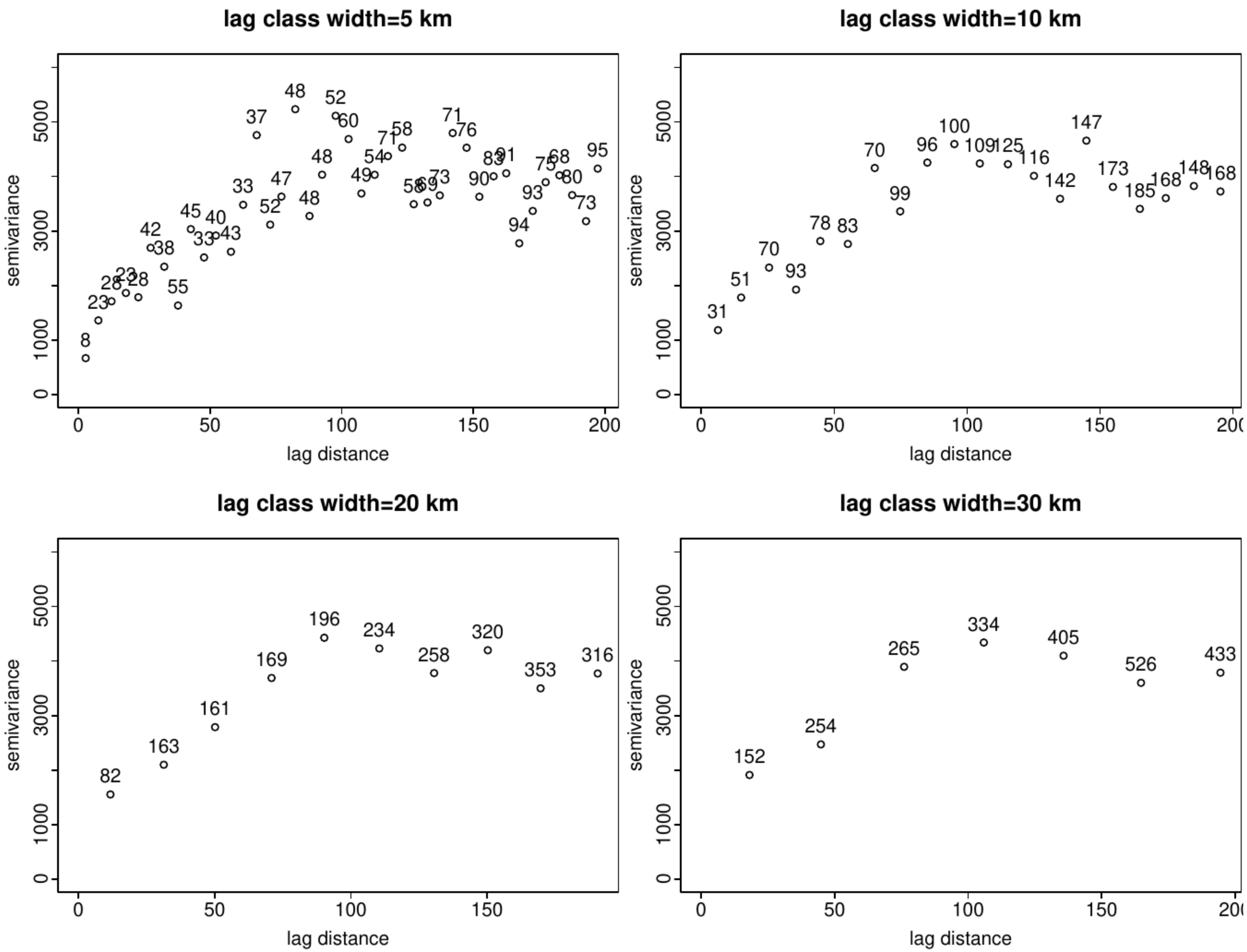
$$\hat{V}(\mathbf{h}_{kl}) = \frac{1}{2N_{kl}} \sum_{(i,j)\in\mathbf{h}_{kl}} [R(\mathbf{x}_i) - R(\mathbf{x}_j)]^2$$

Sample variogram: Plot of $\hat{V}(\mathbf{h}_{kl})$ vs. $\mathbf{h}_{kl}$.

Rules of thumb:

1. Ensure enough data pairs per lag class. Choose dh (and $d\phi$) such that $N_{kl} > 30 - 50$.
2. Only compute the variogram for half the size of the study area.
   Largest $\mathbf{h}_{kl} \leq 0.5 \max(\mathbf{x}_i - \mathbf{x}_j)$.

Examples of different lags (Wolfcamp data)



4.5 Fitting variogram model to sample variogram

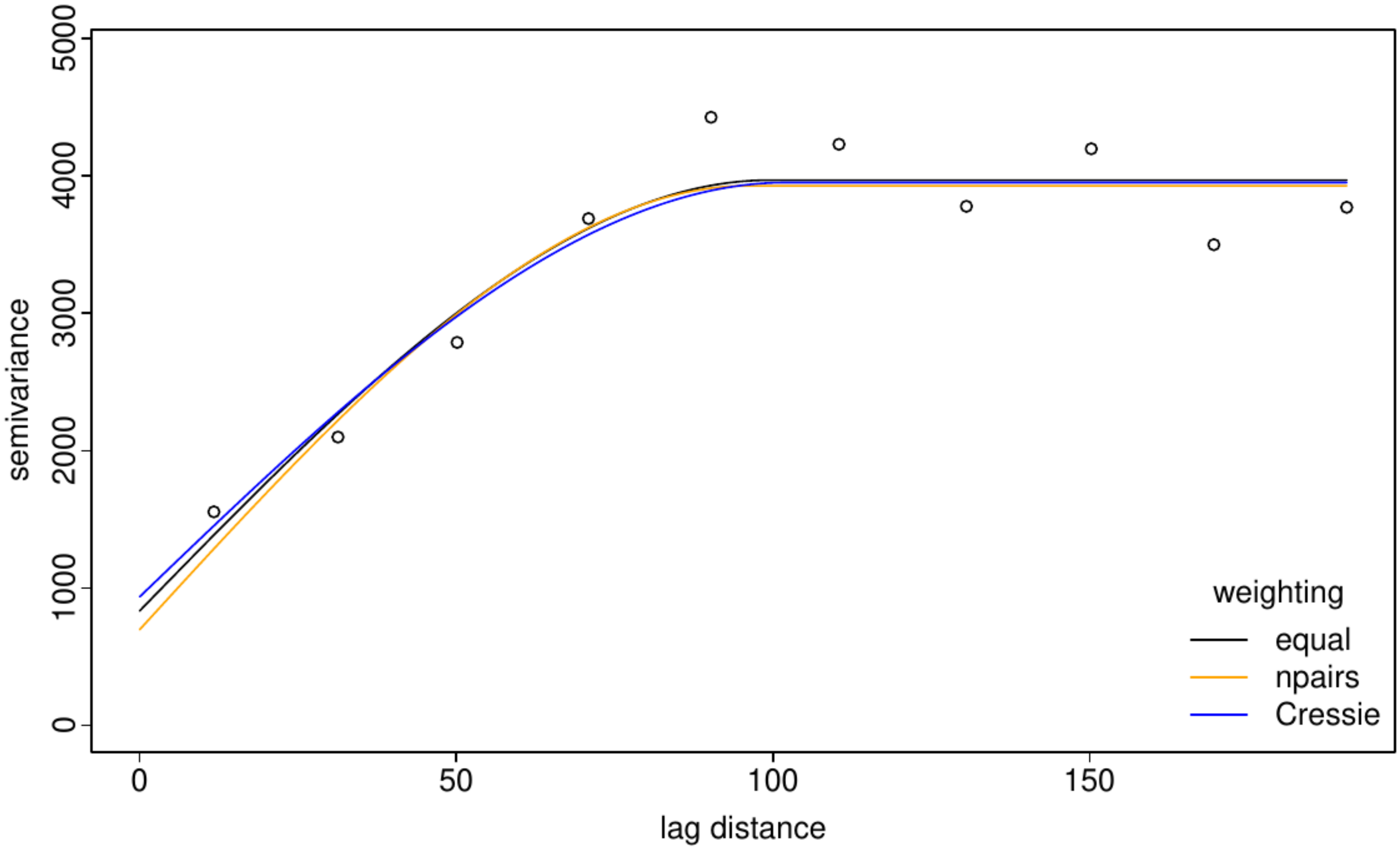
- **Semivariance** is required for arbitrary lag distances when computing predictions.
- **Smoothing** the sample variogram by fitting a parametric variogram function $V(\mathbf{h}, \theta)$.
- Choose a **variogram function** that approximates the shape of the sample variogram well, particularly close to the origin.
- **Fit parameters** θ by (weighted) least squares:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{kl} w(\mathbf{h}_{kl}) \left(V(\mathbf{h}_{kl}) - \hat{V}(\mathbf{h}_{kl}, \theta) \right)^2$$

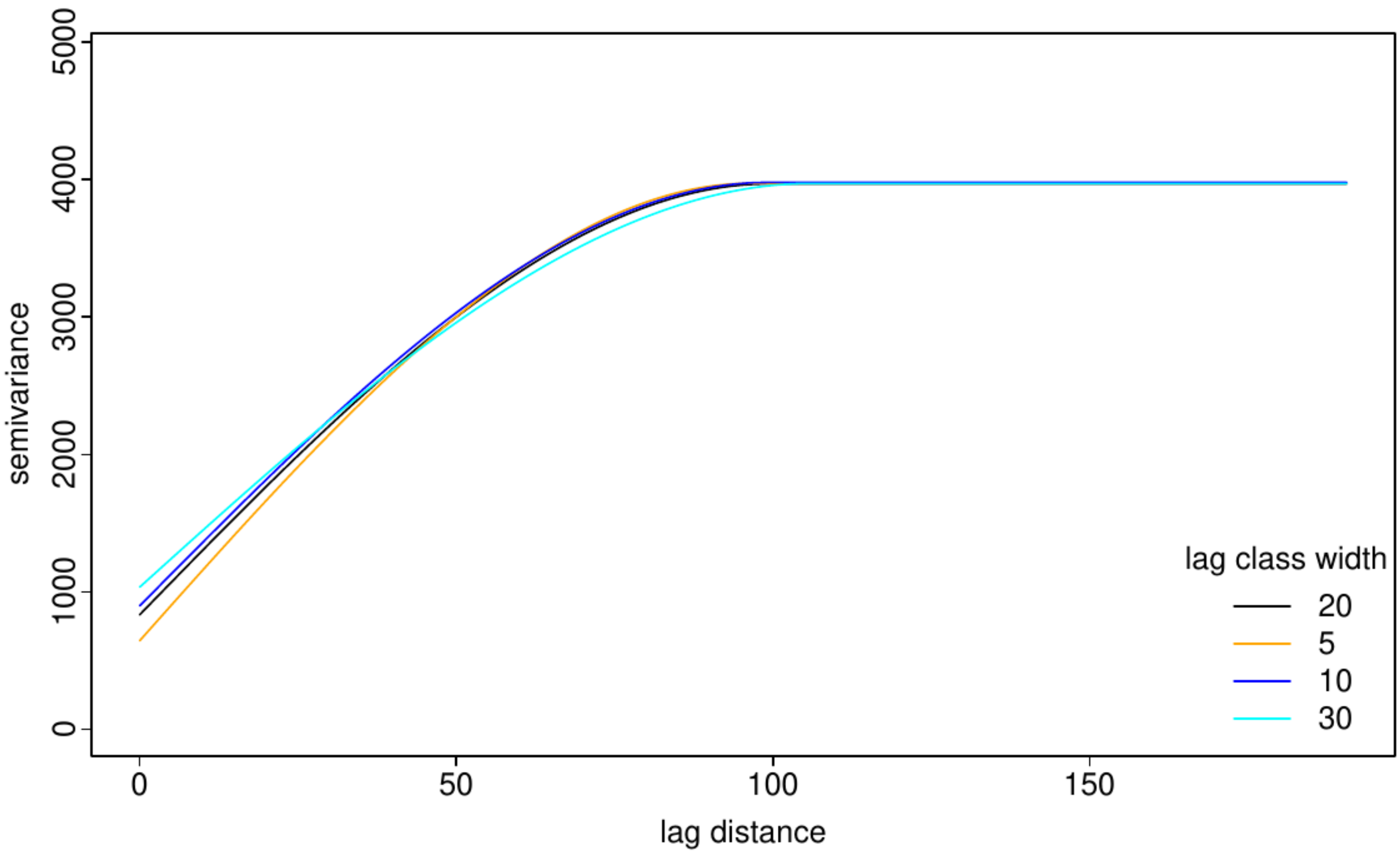
Options for weighing:

1. **Equal weights**: $w(\mathbf{h}_{kl}) = 1$
2. **By number of pairs**: $w(\mathbf{h}_{kl}) = N_{kl}$
3. **Cressie's weights**: $w(\mathbf{h}_{kl}) = \frac{N_{kl}}{V(\mathbf{h}_{kl}, \theta)^2}$

Fits with different weights (Wolfcamp data)



Fits with different lags (Wolfcamp data)



4.6 Problems with ad-hoc model estimation

- **Subjective choice** of lag class width and weighting method for model fitting.
- **Estimates of semivariance** for different lag classes are mutually correlated; the choice of variogram function based on the sample variogram is problematic.
- The sample variogram is **susceptible to outliers**, hence robust estimators are preferred.
- Fitting a model function to the sample variogram requires **further subjective choices**.
- An ad-hoc approach provides **biased estimates** of the variogram of the underlying stochastic process if the trend is modeled.
- Therefore, the estimate of the variogram based on the sample variogram of OLS residuals is biased.
- Thus, estimate trend and variogram parameters **simultaneously** using maximum likelihood.

5 Maximum likelihood estimation

5.1 Maximum likelihood (ML) estimation of parameters of Gaussian model for spatial data

- Principle of maximum likelihood estimation: find parameters that maximize joint probability for observed data
- Properties of maximum likelihood estimates: asymptotically unbiased and fully efficient; asymptotically normally distributed
- profile likelihood useful for exploring shape of likelihood surface and for computing confidence intervals based on likelihood ratio test

Maximum likelihood estimation

Consider a **Gaussian stochastic process** $\{Y(\mathbf{x})\}$ with a linear trend function.

Any arbitrary set of random variables $\mathbf{Y} = (Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n))$ has a multivariate Gaussian distribution with expectation:

$$\mathbb{E}[\mathbf{Y}] = \mathbf{X}\beta$$

and covariance matrix:

$$\mathrm{Cov}(\mathbf{Y}, \mathbf{Y}^{\mathrm{T}}) = \mathbf{\Gamma}_\theta$$

The joint probability density for $\mathbf{Y}$ is given by:

$$f(\mathbf{y}; \beta, \theta) = (2\pi)^{-\frac{n}{2}} |\mathbf{\Gamma}_\theta|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\{\mathbf{y} - \mathbf{X}\beta\}^{\mathrm{T}} \mathbf{\Gamma}_\theta^{-1} \{\mathbf{y} - \mathbf{X}\beta\}\right)$$

Maximum likelihood estimation (cont.)

Unknown model parameters:

1. Regression coefficients β
2. Covariance (or variogram) parameters θ

The log-likelihood function (up to a constant) is given by:

$$L(\beta, \theta; \mathbf{y}) = -\frac{1}{2}\log(|\mathbf{\Gamma}_\theta|) - \frac{1}{2}\{\mathbf{y} - \mathbf{X}\beta\}^{\mathrm{T}}\mathbf{\Gamma}_\theta^{-1}\{\mathbf{y} - \mathbf{X}\beta\}$$

For known variogram parameters θ maximum likelihood estimate for β equals **GLS** estimator:

$$\hat{\beta}_{\mathrm{GLS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{\Gamma}_\theta^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{\Gamma}_\theta^{-1}\mathbf{Y}$$

- Plugging $\hat{\beta}_{\mathrm{GLS}}$ into $L(\beta, \theta; \mathbf{y})$ gives the profile likelihood function for θ.
- No closed form expression for maximum likelihood estimate
- Maximize L numerically by a non-linear optimization method to find $\hat{\theta}$
- Numerical optimization requires initial values of θ
- Make sure numerical optimization converged!

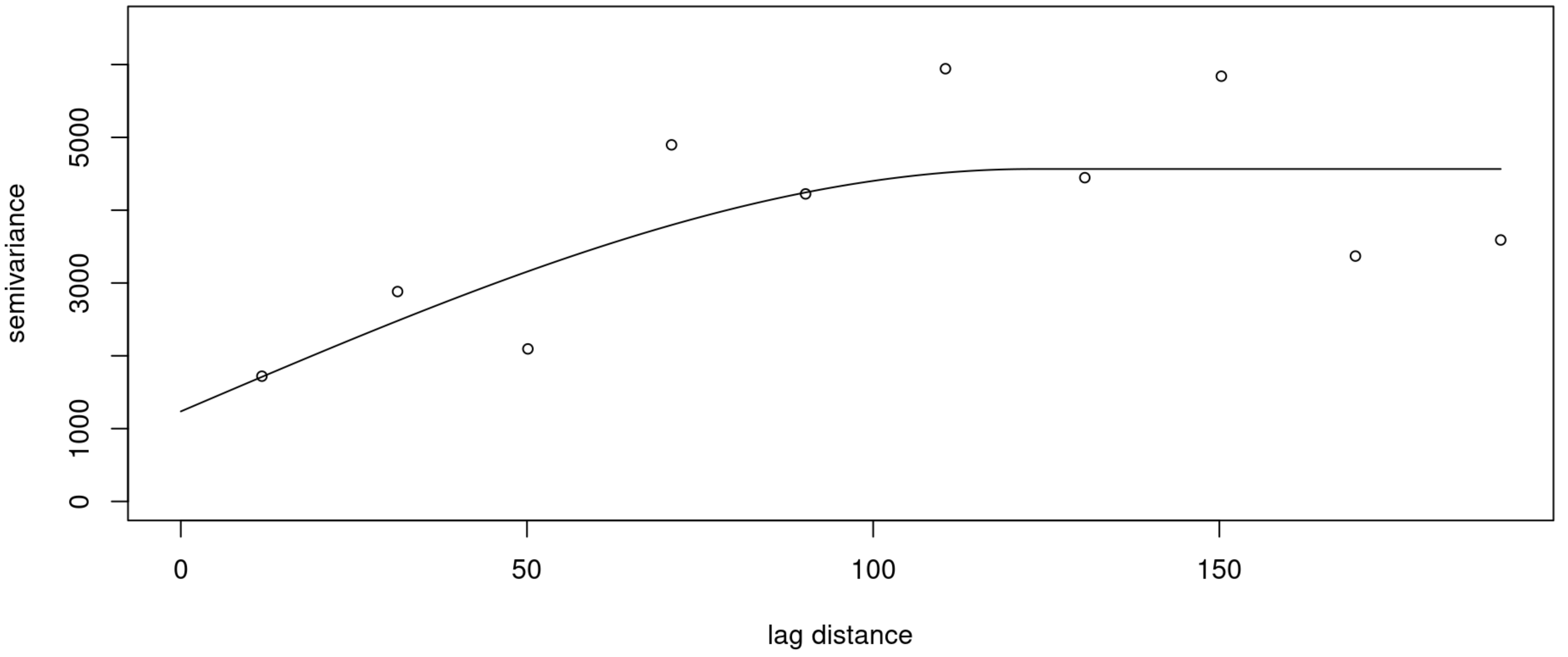
Example: MaxLik estimates Wolfcamp data

```r
library(gstat); library(georob)
data("wolfcamp"); d.w <- wolfcamp
coordinates(d.w) <- c("x", "y")
r.georob.ml <- georob(pressure~x+y, d.w,
  locations=~x+y, variogram.model="RMspheric",
  param=c(variance=3000, nugget=1000, scale=100),
  tuning.psi=1000, control=control.georob(ml.method="ML"))
summary(r.georob.ml)
```

```
Call:georob(formula = pressure ~ x + y, data = d.w, locations = ~x +
    y, variogram.model = "RMspheric", param = c(variance = 3000,
    nugget = 1000, scale = 100), tuning.psi = 1000, control = control.georob(ml.method = "ML"))

Tuning constant:  1000

Convergence in 12 function and 7 Jacobian/gradient evaluations

Estimating equations (gradient)
```

Wolfcamp: MaxLik fitted variogram

```
plot(r.georob.ml, lag.dist.def=20, max.lag=200)
```



5.2 Restricted maximum likelihood estimation

Equivalent number of independent observations of a sample of spatial data often much smaller than nominal sample size ⇒ bias of ML estimates of variance parameters important

The bias of MLEs of variogram parameters θ can be reduced by **restricted maximum likelihood estimation** (REML).

Principle of **REML**:

1. Form linear combinations $\mathbf{Z} = \mathbf{A}\mathbf{Y}$ of the data $\mathbf{Y}$ that have zero expectation (no longer depend on β):

$$\mathbb{E}[\mathbf{Z}] = \mathbf{A}\mathbf{X}\beta = \mathbf{0}$$

The matrix $\mathbf{A}$ must satisfy: $\mathbf{A}\mathbf{X} = \mathbf{0}$

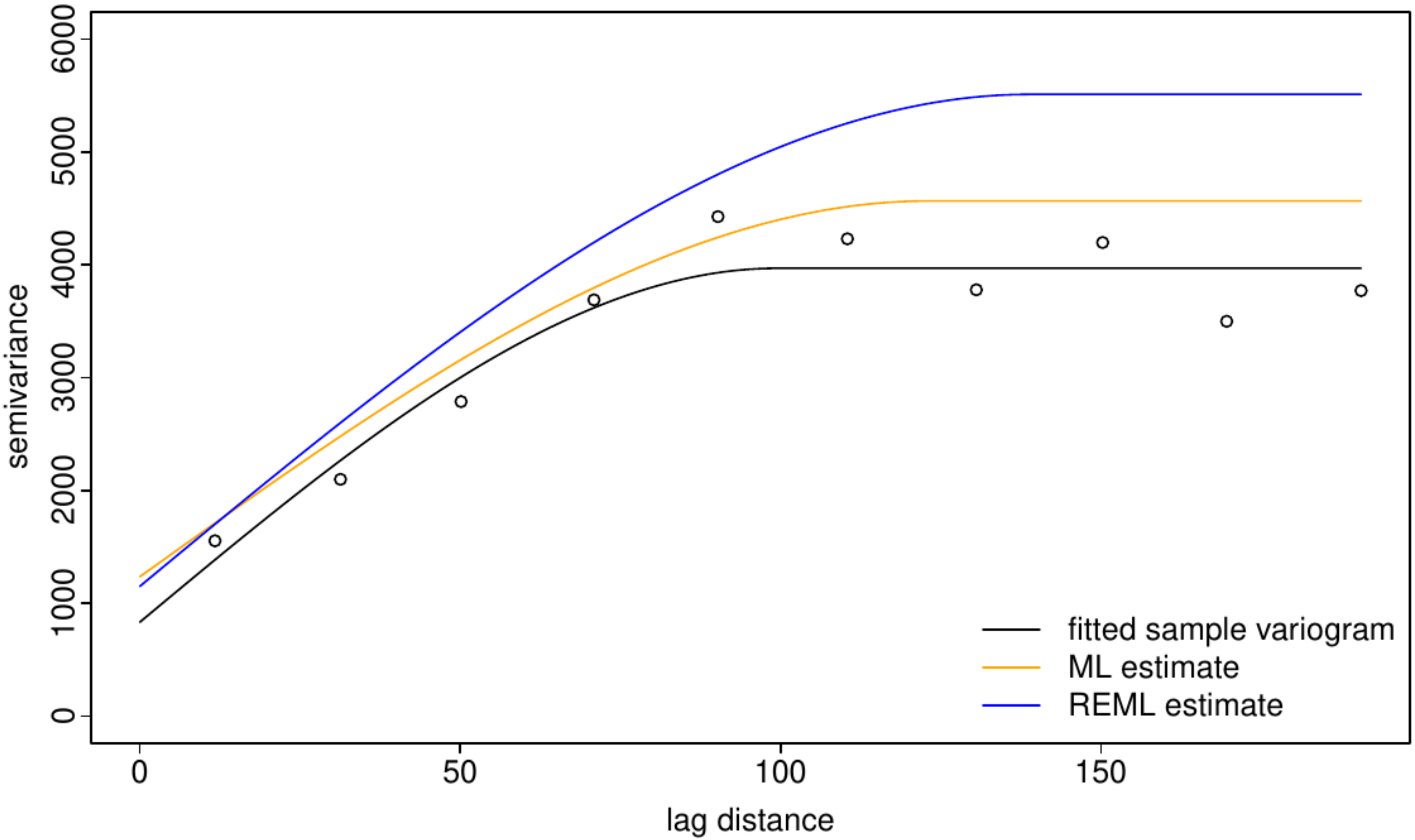
$\mathbf{A}$ is non-unique with many possibilities

2. Estimate θ by maximizing the likelihood function for $n - p$ elements of $\mathbf{Z}$.

REML estimates Wolfcamp data

```
r.georob.reml <- georob(pressure~x+y, d.w,
  locations=~x+y, variogram.model="RMspheric",
  param=c(variance=3000, nugget=1000, scale=100),
  tuning.psi=1000)
summary(r.georob.reml)
```

```
Call:georob(formula = pressure ~ x + y, data = d.w, locations = ~x +
    y, variogram.model = "RMspheric", param = c(variance = 3000,
    nugget = 1000, scale = 100), tuning.psi = 1000)

Tuning constant:  1000

Convergence in 6 function and 5 Jacobian/gradient evaluations

Estimating equations (gradient)

                                 eta         scale
  Gradient           : -2.248651e-04 -1.070402e-01

Maximized restricted log-likelihood: -456.3802

Predicted latent variable (B):
   Min     1Q Median     3Q    Max
-94.58 -60.99 -17.59  23.10 115.72

Residuals (epsilon):
    Min       1Q  Median       3Q      Max
-59.148 -18.009   6.251  15.982  54.620

Standardized residuals:
```



semivariance
0
1000
2000
3000
4000
5000
6000
0
50
100
150
lag distance
fitted sample variogram
ML estimate
REML estimate

6 Model inference

6.1 Inference, model building and assessment

Data analysis often leads to a set of equally plausible candidate models that use different set of covariates and different variograms

- compare fit of candidate models by hypothesis tests taking auto-correlation properly into account
- use established goodness-of-fit criteria (AIC, BIC) to select a “best” model, again taking auto-correlation into account
- use cross-validation to compare the power of candidate models to *predict new data*

Testing hypotheses about trend coefficients

- Likelihood ratio (LRT) test can only be used to test hypotheses and build confidence regions for θ
- LRT for regression for β in general biased (too small p-values)
- Use conditional F-tests for testing hypotheses about β:
  1. Fit covariance parameters of "largest" regression model $\Rightarrow \hat{\theta}$
  2. Compute covariance matrix $\Rightarrow \mathbf{\Gamma}_{\hat{\theta}}$
  3. Orthogonalize response vector and design matrix (using Cholesky decomposition)
  4. Conventional F-test with orthogonalized items $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{X}}$

Fit quadratic trend surface model for Wolfcamp

```r
r.georob.full <- update(r.georob.reml, .~.+I(x^2)+I(y^2)+x:y)
summary(r.georob.full)
```

```
Call:georob(formula = pressure ~ x + y + I(x^2) + I(y^2) + x:y, data = d.w,
    locations = ~x + y, variogram.model = "RMspheric", param = c(variance = 3000,
        nugget = 1000, scale = 100), tuning.psi = 1000)

Tuning constant:  1000

Convergence in 10 function and 8 Jacobian/gradient evaluations

Estimating equations (gradient)

                                 eta         scale
  Gradient           :  3.590344e-04 -4.553394e-03

Maximized restricted log-likelihood: -470.3894

Predicted latent variable (B):
   Min     1Q Median     3Q    Max 
-89.22 -46.81 -11.06  20.80  94.07 

Residuals (epsilon):
    Min      1Q  Median      3Q     Max 
-59.664 -18.086   6.783  16.245  49.986 

Standardized residuals:
```

Conditional F-test on interaction and higher-order polynomials

```r
waldtest(r.georob.full, .~.-x:y, test="F")
```

```
Wald test

Model 1: pressure ~ x + y + I(x^2) + I(y^2) + x:y
Model 2: pressure ~ x + y + I(x^2) + I(y^2)
  Res.Df Df      F Pr(>F)
1     79
2     80 -1 1.1032 0.2968
```

```r
waldtest(r.georob.full, .~.-I(x^2)-I(y^2)-x:y, test="F")
```

```
Wald test

Model 1: pressure ~ x + y + I(x^2) + I(y^2) + x:y
Model 2: pressure ~ x + y
  Res.Df Df      F Pr(>F)
1     79
2     82 -3 1.6284 0.1895
```

6.2 Model selection with stepwise

Given estimates of covariance parameters $\hat{\theta}$ and keeping them fixed, the usual stepwise procedures for selecting covariates can be used.

Selecting models based on AIC and BIC.

Stepwise selection with AIC (defaults to both directions)

```
step(r.georob.full)
```

```
Start:  AIC=922.16
pressure ~ x + y + I(x^2) + I(y^2) + x:y

         Df    AIC Converged
- I(x^2)  1 922.05         1
- I(y^2)  1 922.13         1
<none>      922.16
- x:y     1 922.49         1

Step:  AIC=922.05
pressure ~ x + y + I(y^2) + x:y

         Df    AIC Converged
<none>      922.05
+ I(x^2)  1 922.16         1
- I(y^2)  1 922.54         1
- x:y     1 924.61         1
```

Stepwise selection (defaults to both directions) with BIC

```r
step(r.georob.full, k=log(nrow(d.w)))
```

```
Start:  AIC=936.81
pressure ~ x + y + I(x^2) + I(y^2) + x:y

         Df    AIC Converged
- I(x^2)  1 934.27         1
- I(y^2)  1 934.34         1
- x:y     1 934.70         1
<none>      936.81

Step:  AIC=934.27
pressure ~ x + y + I(y^2) + x:y

         Df    AIC Converged
- I(y^2)  1 932.31         1
<none>      934.27
- x:y     1 934.38         1
+ I(x^2)  1 936.81         1

Step:  AIC=932.31
```

7 Kriging predictions

7.1 Prediction problem formulation

Observations $\mathbf{y}^{\mathrm{T}} = (y_1, \ldots, y_n)$ available for a set of n locations $\mathbf{x}_i$

Consider $\mathbf{y}$ as a realization of the random variable $\mathbf{Y}^{\mathrm{T}} = (Y_1, \ldots, Y_n)$

Model: $Y_i = S(\mathbf{x}_i) + Z_i$ with:

- Y_i: the i^{th} datum
- $S(\mathbf{x}_i)$: "signal" (the true quantity) at location $\mathbf{x}_i$
- $\{S(\mathbf{x}_i)\}$: Gaussian process, parametrized by:
  - Trend: $\mu(\mathbf{x}_i) = \sum_k d_k(\mathbf{x}_i)\beta_k = \mathbf{d}(\mathbf{x}_i)^{\mathrm{T}}\beta$
  - Covariance function $\gamma(\mathbf{h};\theta)$ or variogram $V(\mathbf{h};\theta)$
- Z_i: independent, identically distributed (iid) Gaussian measurement error with variance τ^2

Predictions: Let's say $\hat{\mathbf{S}}$ is the prediction of $\mathbf{S}^{\mathrm{T}} = (S(\mathbf{x}'_1), \ldots, S(\mathbf{x}'_m))$ for a set of m locations $\mathbf{x}'_j$ without data.

$\hat{\mathbf{S}}$ is computed from $\mathbf{Y}$, therefore $\hat{\mathbf{S}} = \hat{\mathbf{S}}(\mathbf{Y})$.

7.2 Kriging prediction at new point location

Ordinary punctual kriging

For same spatial *support* – prediction entities are assumed to have same extension as observations (i.e. sampling area, sensor size)

Prediction of signal $S(\mathbf{x}_0)$ at location $\mathbf{x}_0$ without measurement

$$\hat{S}(\mathbf{x}_0) = \sum_{i=1}^{n} \kappa_i(\mathbf{x}_0)\, y(\mathbf{x}_i)$$

with kriging weights $\kappa_i(\mathbf{x}_0)$. To ensure unbiased estimates, weights are made to sum to 1:

$$\sum_{i=i}^{N} \kappa_i(x_0) = 1$$

Expected difference $\mathbb{E}[S(\mathbf{x}_0) - \hat{S}(\mathbf{x}_0)] = 0$.

Estimate of kriging variances

Estimate of the variance of the prediction error:

$$var[\hat{S}(\mathbf{x}_0] = \mathbb{E}\left[\left\{S(\mathbf{x}_0) - \hat{S}(\mathbf{x}_0)\right\}\right]$$

$$= 2\sum_{i=1}^{N} \kappa_i(\mathbf{x}_0) V(\mathbf{x_i}, \mathbf{x_0}) - \sum_{i=1}^{N}\sum_{j=1}^{N} \kappa_i(\mathbf{x}_0)\kappa_j(\mathbf{x}_0) V(\mathbf{x_i}, \mathbf{x_j})$$

with $V(\mathbf{x_i}, \mathbf{x_0})$ being the semivariance of S between sampling point x_i and the target prediction point $\mathbf{x_0}$ and $V(\mathbf{x_i}, \mathbf{x_j})$ the semivariance between the ith and the jth sampling points.

Find kriging weights

Kriging weights are calculated by solving the system of equations that minimizes the prediction error variance subject to the constraints of the variogram model.

Find weights $\kappa_i(\mathbf{x}_0)$ that minimize kriging variances and sum to 1 by solving $N+1$ equations:

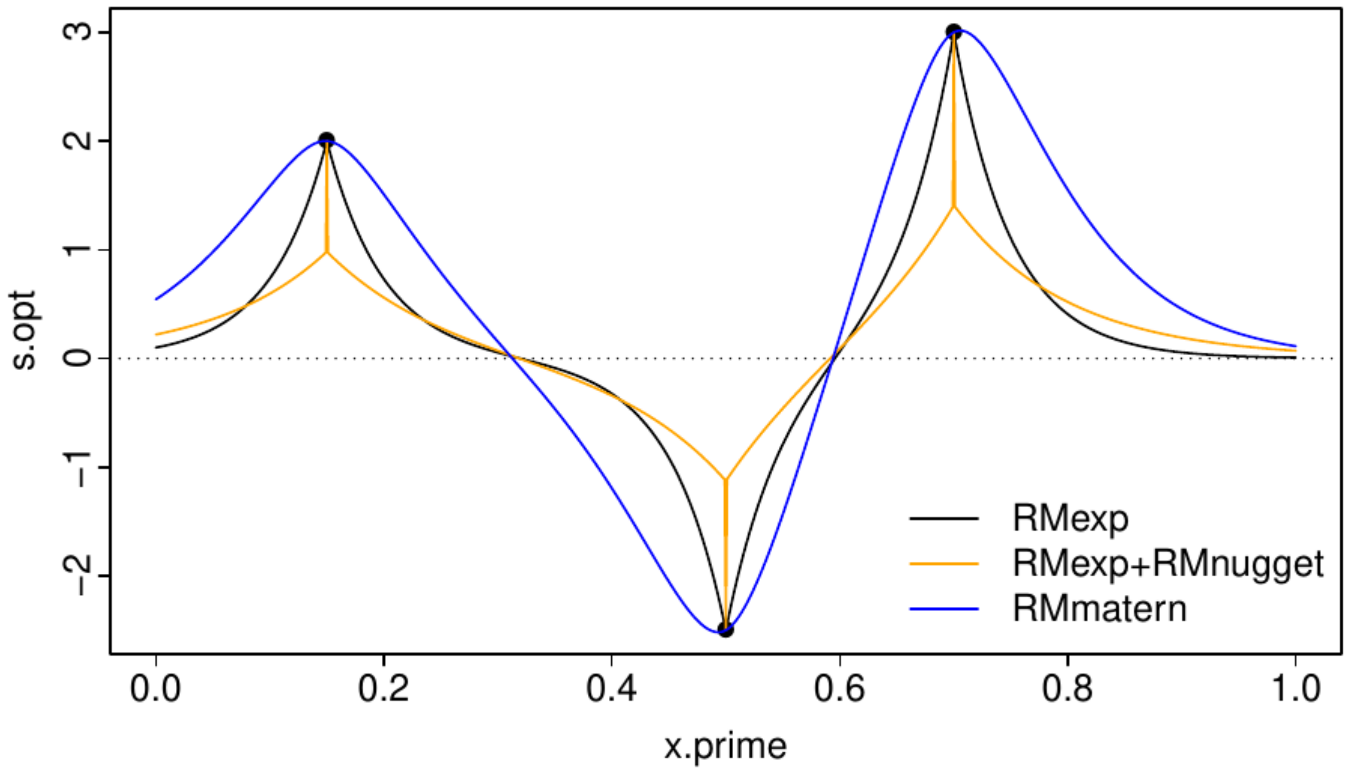
$$\sum_{i=1}^{N} \kappa_i(\mathbf{x}_0)V(\mathbf{x_i}, \mathbf{x_j}) + \psi(\mathbf{x_0}) = V(\mathbf{x_j}, \mathbf{x_0}) \quad \text{for all} \;\; j$$

$$\sum_{i=i}^{N} \kappa_i(x_0) = 1$$

The quantity $\psi(\mathbf{x_0})$ is a Lagrange multiplier introduced to achieve minimization. Computed as the inverse of the semivariance matrix multiplied by the vector of semivariances to the target point.

Properties of kriging prediction

- Shape of variogram function close to origin determine shape of prediction surface near data locations
- Continuity and diffentiability of variogram at origin control geometrical properties of simple kriging prediction surface



7.3 Universal/external drift kriging predictions

Universal kriging: often referred if trend is modelled by coordinates

External-drift kriging: trend is modelled by spatial covariates

Evaluating $\hat{\mathbf{S}}_{\mathrm{opt}}$ requires a fully specified weakly stationary model:

1. Structure of trend function is known
2. Regression coefficients β are known
3. Type of parametric covariance (variogram) function is known
4. Parameters θ and τ^2 of the covariance function are known

Relaxed assumptions: Only 1, 3, and 4 are assumed to be known, while β is implicitly estimated from the data using generalized least squares (GLS).

UK/EDK predictions

Therefore, we use the **universal kriging (UK) or external drift kriging (EDK) plug-in predictor**:

$$\hat{\mathbf{S}}_{\mathrm{k}} = \mathbf{X}_{\mathbf{S}}\hat{\beta}_{\mathrm{GLS}} + \mathbf{\Lambda}\left(\mathbf{y} - \mathbf{X}_{\mathbf{Y}}\hat{\beta}_{\mathrm{GLS}}\right)$$

with:

$$\mathbf{\Lambda} = \mathbf{\Sigma}_{\mathbf{SY}}\mathbf{\Gamma}_{\mathbf{YY}}^{-1}$$

Computing universal kriging predictor requires:

1. known structure of trend function
2. known structure and parameters of variogram

“Plug-in” predictor: uncertainty of variogram is ignored when computing predictions