TMA4250 Spatial Statistics Exercise 1, Spring 2019

Group members: Henrik Syversveen Lie, Øyvind Auestad

18.02.2019

Problem 1: Gaussian RF - model characteristics

We consider the continuous spatial variable $\{r(x) : x \in D : [1, 50] \subset \mathbb{R}^1$, and assume that it is modeled as a stationary 1D Gaussian RF with the following model parameters:

$$\begin{aligned} \mathrm{E}\{r(x)\} &= \mu_r = 0 \\ \mathrm{Var}\{r(x)\} &= \sigma_r^2 \\ \mathrm{Corr}\{r(x), r(x')\} &= \rho_r(\tau), \end{aligned}$$

where $\rho_r(\tau)$; $\tau = |x - x'|/10$ is the spatial correlation function. Let $D : [1, 50]$ be discretised in $L \in \{1, 2, \ldots, 50\}$ and define the discretised Gaussian RF $\{r(x); x \in L\}$, represented by the n-vector $\mathbf{r} \in \mathbb{R}^n$.

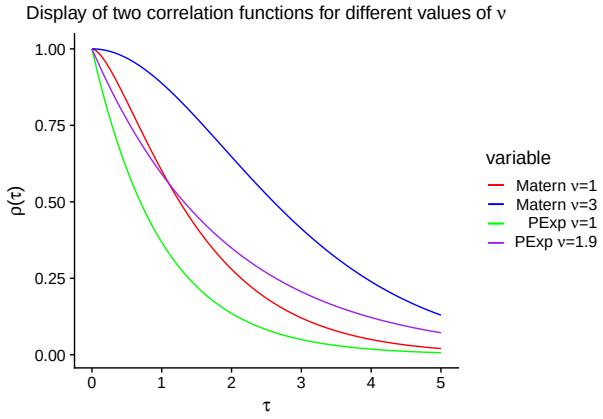
Let the spatial correlation function $\rho_r(\tau)$, be either Powered exponential with parameter $\nu_r \in [1, 1.9]$ or Matern with parameter $\nu_r \in [1, 3]$. Let the variance take the values $\sigma_r^2 \in [1, 5]$.

a)

We restrict the spatial correlation functions to be only positive definite functions. Covariance matrices need to be positive definite, and a positive definite correlation function ensures that all covariances matrices in our stationary Gaussian RF are positive definite for all configurations and dimensions. Further, we define a function $\rho(\tau) : \mathbb{R}^q \to \mathbb{R}$ to be positive definite if

$$\begin{aligned} &\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \rho(\mathbf{x}_i - \mathbf{x}_j) \geq 0, \\ &\text{for all configuations } [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{q \times n}, \\ &\text{for all weights } \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n) \in \mathbb{R}^n, \\ &\text{for all } n \in \mathbb{N}_+ \setminus \{1\}. \end{aligned}$$

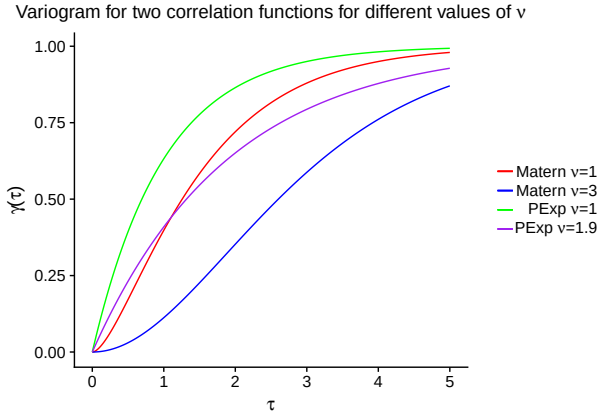
To further investigate different correlation functions, we display two types of correlation functions, the Matern: $\rho(\tau) = 2^{1-\nu}/\Gamma(\nu) \cdot \tau^\nu \mathcal{B}_\nu(\tau)$, and Powered exponential: $\rho(\tau) = \exp(-\tau^\nu)$, for different values of ν_r.



For all correlation functions $\rho_r(0) = 1$, and $\rho_r(\tau) \in [-1, 1]$; $\tau \in \mathbb{R}_+$, because the function represents correlation between two random variables. The correlation functions are continuous everywhere, except at $\tau = 0$ where a step may occur. If the correlation function is continuous at $\tau = 0$, then the random field is continuous almost everywhere. Away from $\tau = 0$, the correlation function must be smooth. From the displayed correlation functions, we see that $\rho_r(\tau) \to 0$ as $\tau \to \infty$. This will be the case for all correlation functions, implying that two points $\mathbf{x}$ and $\mathbf{x}'$ will tend towards being uncorrelated as $|\mathbf{x} - \mathbf{x}'| \to \infty$. Uncorrelated Gaussian random fields means that they will be independent, and by extension asymptotics Gaussian random fields are ergodic.

The correlation functions define correlation between points in our random field. This means that larger values of the correlation function implies a smoother random field. From the plotted correlation function, we see that the function values increases for larger values of ν_r. Consequently, larger values of ν_r implies smoother random fields for both these correlation functions.

We now define the variogram function, which for stationary Gaussian random fields is defined as $\gamma_r(\tau) = \sigma_r^2[1 - \rho_r(\tau)]$. Then, we display the variogram functions associated with the previously plotted correlation functions for $\sigma_r^2 = 1$.



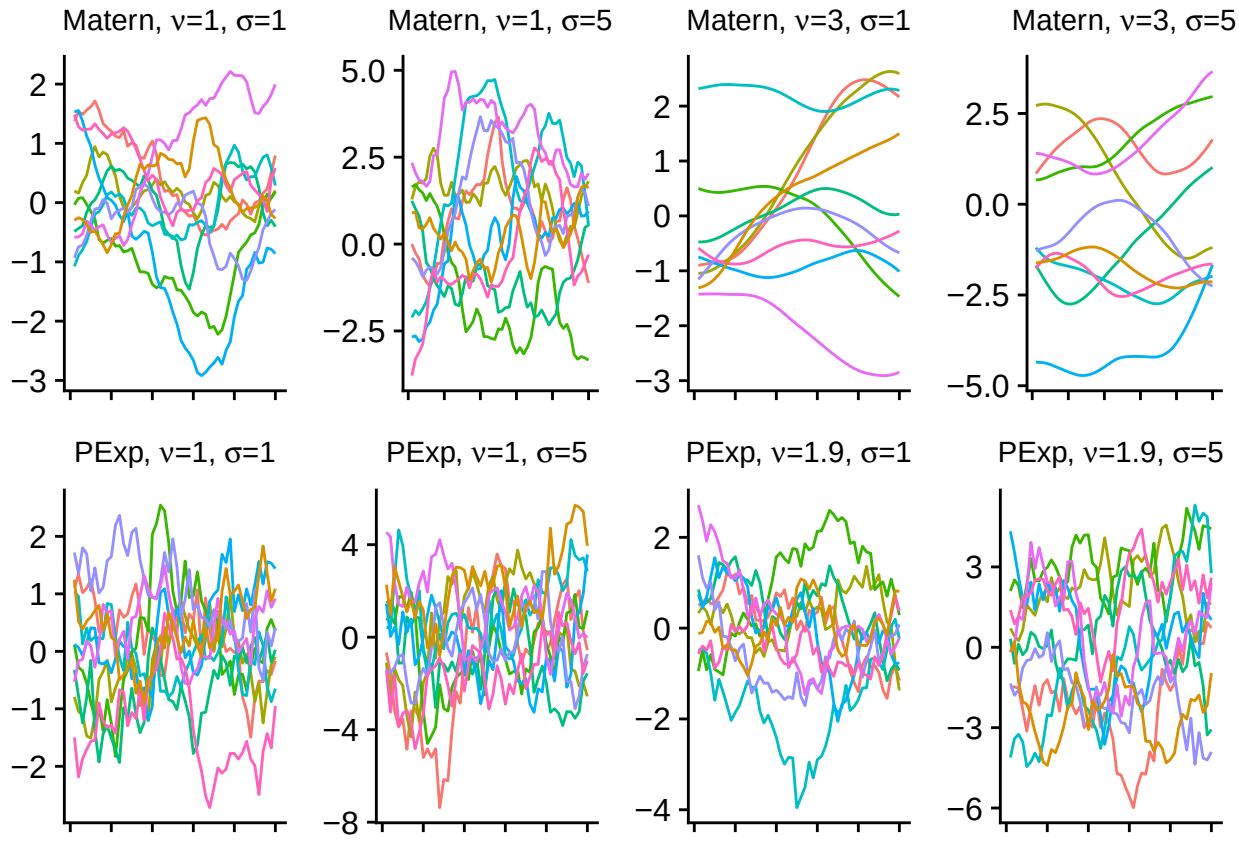
b)

The prior Gaussion random field is defined on the discretized representation $L \in \{1, 2, \ldots, 50\}$ by

$$\mathbf{r} \sim p(\mathbf{r}) = \phi_n(\mathbf{r}; \mu_r \mathbf{i}_n, \sigma_r^2 \mathbf{\Sigma}_r^\rho),$$

which is a discretized stationary Gaussian random field with expectation $\mu(\mathbf{x}) = \mu_r$, variance $\sigma^2(\mathbf{x}) = \sigma_r^2$ and correlation function $\rho(\mathbf{x}, \mathbf{x}') = \rho_r(|\mathbf{x} - \mathbf{x}'|/10)$.

We now want to simulate ten realizations of the Gaussian random field on L for all the previously displayed correlation functions and $\sigma_r^2 \in [1, 5]$.



From the display, we see that a greater value of ν_r leads to smoother random fields. Also, a higher value of σ_r^2 leads to larger variations around the expected level $\mu_r = 0$. Also, the Matern correlation function leads to smoother random fields than the Powered exponential function.

c)

We now observe the spatial variable as $\{d(x); x \in [10, 25, 30] \subset L\}$ according to the acquisition model,

$$d(x) = r(x) + \epsilon(x) \qquad x \in [10, 25, 30],$$

with measurement errors $\epsilon(\cdot)$ centred, i.i.d Gaussian with variance σ_ϵ^2. Further, we assume that $r(x)$ and $\epsilon(x')$ are independent for all x, x'.

The likelihood model $p(\mathbf{d}|\mathbf{r})$ links the observations to the spatial variable. The observations $\mathbf{d}$ are known, while $\mathbf{r}$ is the spatial variable. This means that the likelihood $p(\mathbf{d}|\mathbf{r})$ is not a pdf w.r.t. $\mathbf{r}$, and we do not need to normalize the likelihood.

We can define a Gauss-linear likelihood model relative to the discretized spatial variable $\mathbf{r}$, and observe $\mathbf{d} \in \mathbb{R}^m$ according to

$$[\mathbf{d}|\mathbf{r}] = \mathbf{H}\mathbf{r} + \boldsymbol{\epsilon}_{d|r} \sim p(\mathbf{d}|\mathbf{r}) = \phi_m(\mathbf{d}; \mathbf{H}\mathbf{r}, \boldsymbol{\Sigma}_{d|r}),$$

where $\mathbf{H}$ is a $(m \times n)$ observation matrix. In our case, because we have three observations, $m = 3$. Also, the observations are assumed to have i.i.d. errors, which implies $\boldsymbol{\Sigma}_{d|r} = \sigma_\epsilon^2 I_{m \times m}$.

d)

The pdf for the discretised posterior Gaussian random field, given the observations is defined in the following way,

$$[\mathbf{r}|\mathbf{d}] \sim p(\mathbf{r}|\mathbf{d}) = \phi_n(\mathbf{r}; \boldsymbol{\mu}_{r|d}, \boldsymbol{\Sigma}_{r|d}),$$

with

$$\boldsymbol{\mu}_{r|d} = \mu_r \mathbf{i}_n + \sigma_r^2 \boldsymbol{\Sigma}_r^\rho \mathbf{H}^T \left[\sigma_r^2 \mathbf{H} \boldsymbol{\Sigma}_r^\rho \mathbf{H} + \boldsymbol{\Sigma}_{d|r}\right]^{-1} [\mathbf{d} - \mu_r \mathbf{H} \mathbf{i}_n],$$
$$\boldsymbol{\Sigma}_{r|d} = \boldsymbol{\Sigma}_{r|d}^\sigma \boldsymbol{\Sigma}_{r|d}^\rho \boldsymbol{\Sigma}_{r|d}^\sigma = \sigma_r^2 \boldsymbol{\Sigma}_r^\rho - \sigma_r^2 \boldsymbol{\Sigma}_r^\rho \mathbf{H}^T \left[\sigma_r^2 \mathbf{H} \boldsymbol{\Sigma}_r^\rho \mathbf{H} + \boldsymbol{\Sigma}_{d|r}\right]^{-1} \sigma_r^2 \mathbf{H} \boldsymbol{\Sigma}_r^\rho.$$

We use as prior model one of the realizations with $\sigma_r^2 = 5$, $\nu = 1$ and Matern correlation function. Then, a prediction of the spatial variable $\{\hat{r}(\mathbf{x}); \mathbf{x} \in L\}$ represented by the vector $\hat{\mathbf{r}}$ is taken to minimize squared error, yielding

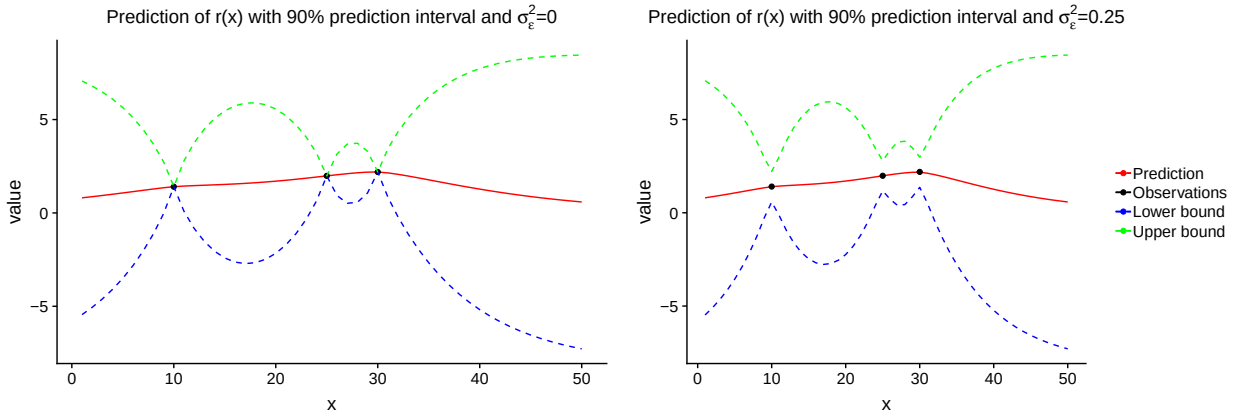
$$\hat{\mathbf{r}} = \mathbf{E}[\mathbf{r}|\mathbf{d}] = \boldsymbol{\mu}_{r|d}. \tag{1}$$

The associated $(1 - \alpha)$ prediction intervals are,

$$PI_\alpha = \boldsymbol{\mu}_{r|d} \pm z_{\alpha/2} \boldsymbol{\sigma}_{r|d},$$

where $\boldsymbol{\sigma}_{r|d}$ is a n-vector containing the diagonal elements of the standard deviation matrix $\boldsymbol{\Sigma}_{r|d}^\sigma$.

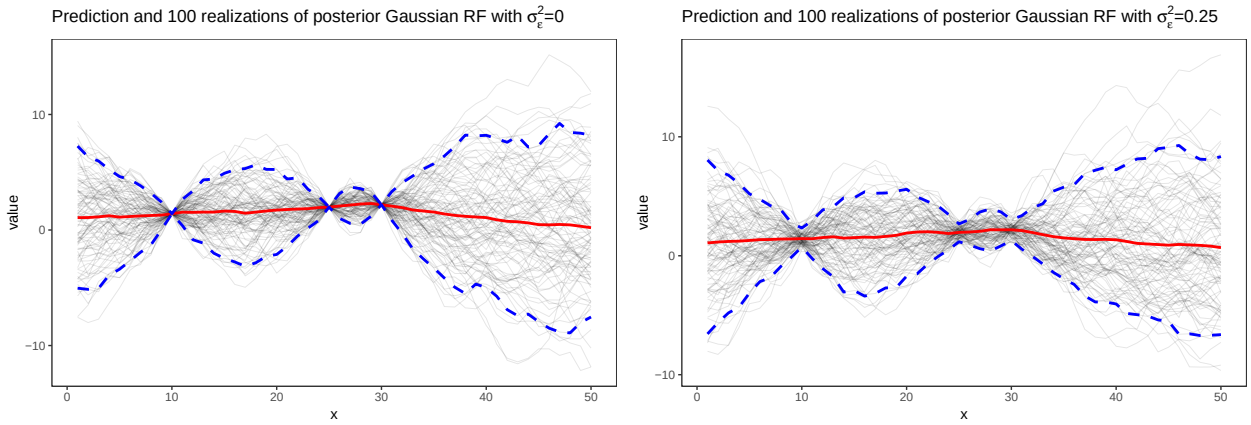
We compute and plot both the prediction and the interval.



From the plots, we see that with no observation error, the prediction coincides with the observations in the points where we have an observation. Also, in these points there is no uncertainty, so the lower and upper bounds are also equal the observations. When observation error is present, we get uncertainty in the observed points. In addition, adding an observation error leads to larger uncertainty, and thus variance, over the entire prediction.

e)

We now go on to simulate 100 realizations from the posterior distribution, using the mean and covariance matrix previously computed. Then, we make a prediction based on the 100 samples and create a 90% empirical prediction interval by ordering the realizations and plotting the 5th smallest and largest value for each $x \in L$.



From the plots, we again see that with no observation error, the prediction coincides with the observations in the points where we have an observation, and we get zero prediction variance. However, because we here only have simulated values (and only 100 simulations), the variance with observation error is not larger for all other points than the variance without observation error.

Also, we see that the variance and prediction based on the 100 samples coincides well with the theoretical prediction and variance. Still, we observe lower degree of smoothness in the lines, which is as expected when comparing simulated values with theoretical quantities.

f)

We use the previously generated $n = 100$ realizations $\{r_i(x)\}, i \in 1, \ldots, n$ with $\sigma_\epsilon^2 = 0$ to provide a prediction $\hat{A}_r$ for the non-linear function on $\{r(x); x \in D\}$,

$$A_r = \int_D I(r(x) > 2)dx.$$

The predictor $\hat{A}_r$ will be

$$\hat{A}_r = \frac{1}{n}\sum_{i=1}^{n} A_{r_i}.$$

We also compute the prediction variance, which will be the sample variance, given as

$$\text{Var}(\hat{A}_r) = \frac{1}{n-1}\sum_{i=1}^{n}(A_{r_i} - \hat{A}_r)^2.$$

An alternative predictor is,

$$\tilde{A}_r = \sum_{x \in L} I(\hat{r}(x) > 2),$$

with $\hat{r}(x)$ given as in equation (1). This value is also computed.

```
## A hat:  20.61
```

```
## A tilde:  9
```

```
## Predicted variance of A hat:  70.58374
```

From the print-out, we see that $\hat{A}_r > \tilde{A}_r$. To shed some light on this result, we make use of Jensen's inequality, which states that for a random variable X and a convex function ψ,

$$\psi(E[X]) \leq E[\psi(X)].$$

We then note that $\hat{r}(x) = E[r_i(x)]$, and with $\psi(\xi) = I(\xi > 2)$, we get after integrating over D (summing over L),

$$\sum_{x \in L} I(\hat{r}(x) > 2) \leq \int_D E\big[I(r_i(x) > 2)\big] dx.$$

Now, $\tilde{A}_r$ is equal to the left hand side, and $\hat{A}_r$ is an unbiased prediction of the right hand side, and one would therefore expect $\tilde{A}_r < \hat{A}_r$.

g)

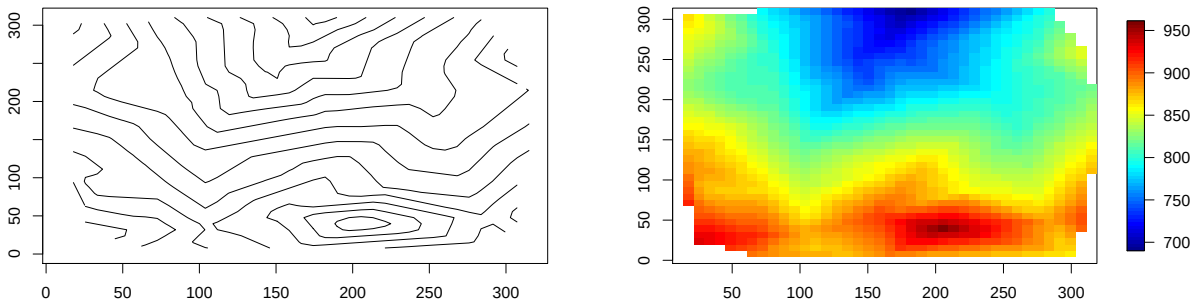
We now want to sum up our experiences made during this exercise. Firstly, we have investigated and achieved greater understanding as to why correlation functions need to be positive definite. Furhtermore, the choice of correlation function and correlation function parameters has great implications for the structure of the Gaussian RF. By choosing a suitable correlation function, we can impact both smoothness and overall variance of the resulting RF. Including an observation error on the observations also has great impact on the prediction variance in those points, and smaller impact on points far away from the observations. We have also gotten a greater understanding in making predictions on Gaussian RFs, and experienced an application of Jensen's inequality.

Problem 2: Gaussian RF - real data

In this problem, we consider observations of terrain elevation from a provided data file. The 52 data observations are located on a domain $D = (0, 315) \times (0, 315) \subset \mathbb{R}^2$. We let the 52-vector of exact observations be $\mathbf{d} = (r(\mathbf{x}_1^0), \ldots, r(\mathbf{x}_{52}^0))^T$.

a)

We then display the observations in two different ways.



A stationary Gaussian RF is a reasonable model for the terrain elevation in domain D provided the correlation function is such that the dependence between different points is decreasing with distance. This RF is ergodic as well, which is natural for the model.

b)

We let the terrain elevation on the domain D be modeled by the Gaussian RF $\{r(\mathbf{x}); \mathbf{x} \in D \subset \mathbb{R}^2\}$ with

$$E\{r(\mathbf{x})\} = \mathbf{g}(\mathbf{x})^T \boldsymbol{\beta}_r,$$
$$Var\{r(\mathbf{x})\} = \sigma_r^2,$$
$$Corr\{r(\mathbf{x}), r(\mathbf{x}')\} = \rho_r(\tau/\xi),$$

where $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_{n_g}(\mathbf{x}))^T$ is a n_g-vector of known explanatory variables on $\mathbf{x} \in D$, and $\boldsymbol{\beta}_r = (\beta_1, \ldots, \beta_{n_g})^T$ is a n_g-vector of unknown parameters. Moreover, we let the variance $\sigma_r^2 = 2500$ and the spatial correlation function have $\xi = 100$ with $\tau = |\mathbf{x} - \mathbf{x}'|$.

We now want to develop the expression for the minimization problem to be solved for the universal Kriging predictor. We assume the model parameters $\boldsymbol{\beta}_r^+$ to be unknown, while the parameters (σ_r^2, η_r) are known. Then, we need the Kriging predictor $\hat{r}_0$ to be unbiased, which requires

$$\mathbf{E}[\hat{r}_0 - r_0] = \boldsymbol{\alpha}^T \mathbf{E}[\mathbf{r}^d] - \mathbf{E}[r_0] = 0,$$

implying

$$\boldsymbol{\alpha}^T \mathbf{G}_d \boldsymbol{\beta}_r^+ = \mathbf{g}_{\mathbf{x_0}} \boldsymbol{\beta}_r^+,$$

which finally implies

$$\mathbf{G}_d^T \boldsymbol{\alpha} = \mathbf{g}_{\mathbf{x_0}}.$$

In addition to being unbiased, we want the predictor to have minimum variance. We ensure minimum variance through the minimum squared-error criterion, which leads to the following minimization problem

$$\begin{aligned}
\hat{\boldsymbol{\alpha}} &= \arg\min_{\boldsymbol{\alpha}} \{\mathbf{Var}[\hat{r}_0 - r_0]\} \\
&= \arg\min_{\boldsymbol{\alpha}} \{\sigma_r^2 - 2\boldsymbol{\alpha}^T \sigma_r^2 \boldsymbol{\rho}_0 + \boldsymbol{\alpha}^T \sigma_r^2 \boldsymbol{\Sigma}_d^\rho \boldsymbol{\alpha}\} \\
&\text{constrained by: } \mathbf{G}_d^T \boldsymbol{\alpha} = \mathbf{g}_{\mathbf{x_0}}.
\end{aligned}$$

This is a quadratic optimization problem with linear constraints. Thus, it is possible to solve it analytically, giving the universal Kriging predictor and associated prediction variance,

$$\begin{aligned}
\hat{r}_0 &= \hat{\boldsymbol{\alpha}}^T \mathbf{r}^d \\
\sigma_{\hat{r}}^2 &= \sigma_r^2 [1 - 2\hat{\boldsymbol{\alpha}}^T \boldsymbol{\rho}_0 + \hat{\boldsymbol{\alpha}}^T \boldsymbol{\Sigma}_d^\rho \hat{\boldsymbol{\alpha}}],
\end{aligned}$$

with

$$\hat{\boldsymbol{\alpha}} = [\boldsymbol{\Sigma}_d^\rho]^{-1} \Big[\boldsymbol{\rho}_0 - \mathbf{G}_d^T \big[\mathbf{G}_d [\boldsymbol{\Sigma}_d^\rho]^{-1} \mathbf{G}_d^T \big]^{-1} \big[\mathbf{G}_d [\boldsymbol{\Sigma}_d^\rho]^{-1} \boldsymbol{\rho}_0 - \mathbf{g}_{\mathbf{x_0}} \big] \Big].$$

c)

We now let the reference variable $\mathbf{x} \in D \subset \mathbb{R}^2$ be denoted $\mathbf{x} = (x_v, x_h)$ and set $n_g = 6$. Then we define the set of known polynomial functions $\mathbf{g}(\mathbf{x})$ to be all polynomial $x_v^k x_h^l$ for $(k, l) \in \{(0,0), (1,0), (0,1), (1,1), (2,0), (0,2)\}$. The resulting n_g-vector $\mathbf{g}(\mathbf{x})$ will be

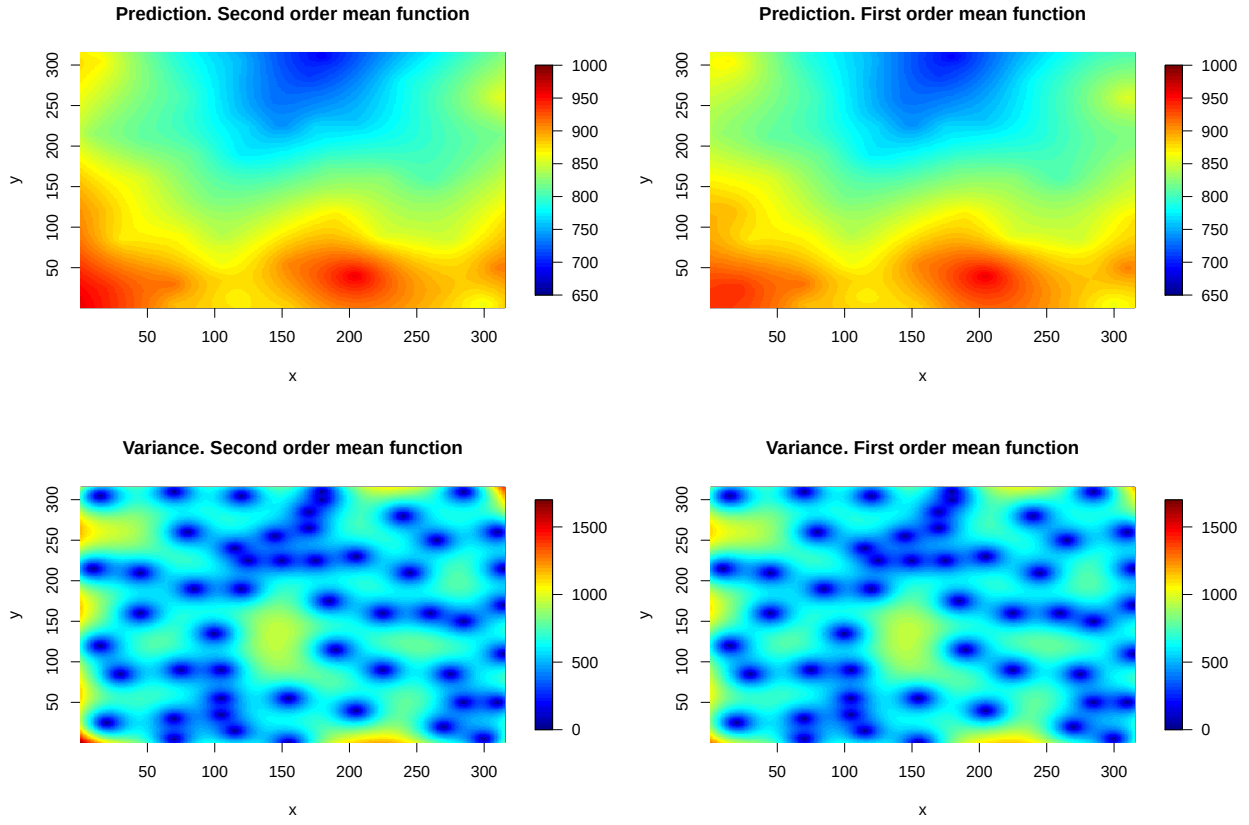
$$\mathbf{g}(\mathbf{x}) = \{1, x_v, x_h, x_v x_h, x_v^2, x_h^2\}.$$

This gives the following expectation value for $r(\mathbf{x})$

$$\mathbf{E}(r(\mathbf{x})) = \mathbf{g}(\mathbf{x})^T \boldsymbol{\beta}_r = \beta_1 + x_v \beta_2 + x_h \beta_3 + x_v x_h \beta_4 + x_v^2 \beta_5 + x_h^2 \beta_6.$$

We also present a model without the second order terms.

We can interpret the model such that mean is the level around which the elevation fluctuates, while the variance is the degree of fluctuation. The level is here a second order polynomial in the coordinates, while the variance is constant over the coordinates.



If we change the parametrization of the expectation function, it could be natural also to change the value of the model variance, σ_r^2. For example when we reduce the degree of the polynomial in the expectation function it could be that we might have to compensate with a larger model variance to get bigger fluctutations around this new 'less flexible' mean. But in practice it would be more natural to estimate the model variance using maximum likelihood, with our data points.

The Kriging predictions look very similar, but there are some differences, which become most clear in the corners of the field. The same is true for the Kriging variances. Adding more parameters in the expectation function, we get more constraints on the minimization problem for α, and so a larger or equal Kriging variance. From the plots we see that the model with second order expectation has the largest Kriging variances, as expected.

From these plots it is clear where the observations are located. Moving further away from the observations we see that the Kriging variance increases. It always stays below the model variance (2500), with the maximum value being just above 1600, obtained in the bottom left corner of the first variance plot.

d)

Since the random field is Gaussian, the best linear predictor (our kriging predictor) coincides with the conditional expectation, $E(r_0|\mathbf{r^d})$. The Gaussian distribution is closed under conditioning, and it follows that

$$[r_0|\mathbf{r^d}] \sim \phi(r_0; \hat{r}_0, \sigma^2_{\hat{r}}),$$

where $\sigma^2_{\hat{r}}$ is as previously defined.

Then we get

$$P(r((100,100)) < 700) = P\Big(\frac{r((100,100)) - \hat{r}_0}{\sigma_{\hat{r}}} < \frac{700 - \hat{r}_0}{\sigma_{\hat{r}}}\Big) = \Phi\Big(\frac{700 - \hat{r}_0}{\sigma_{\hat{r}}}\Big),$$

and the elevation for which it is 0.9 probability that the true elevation is below it will be the solution, $r_{0.9}$, to

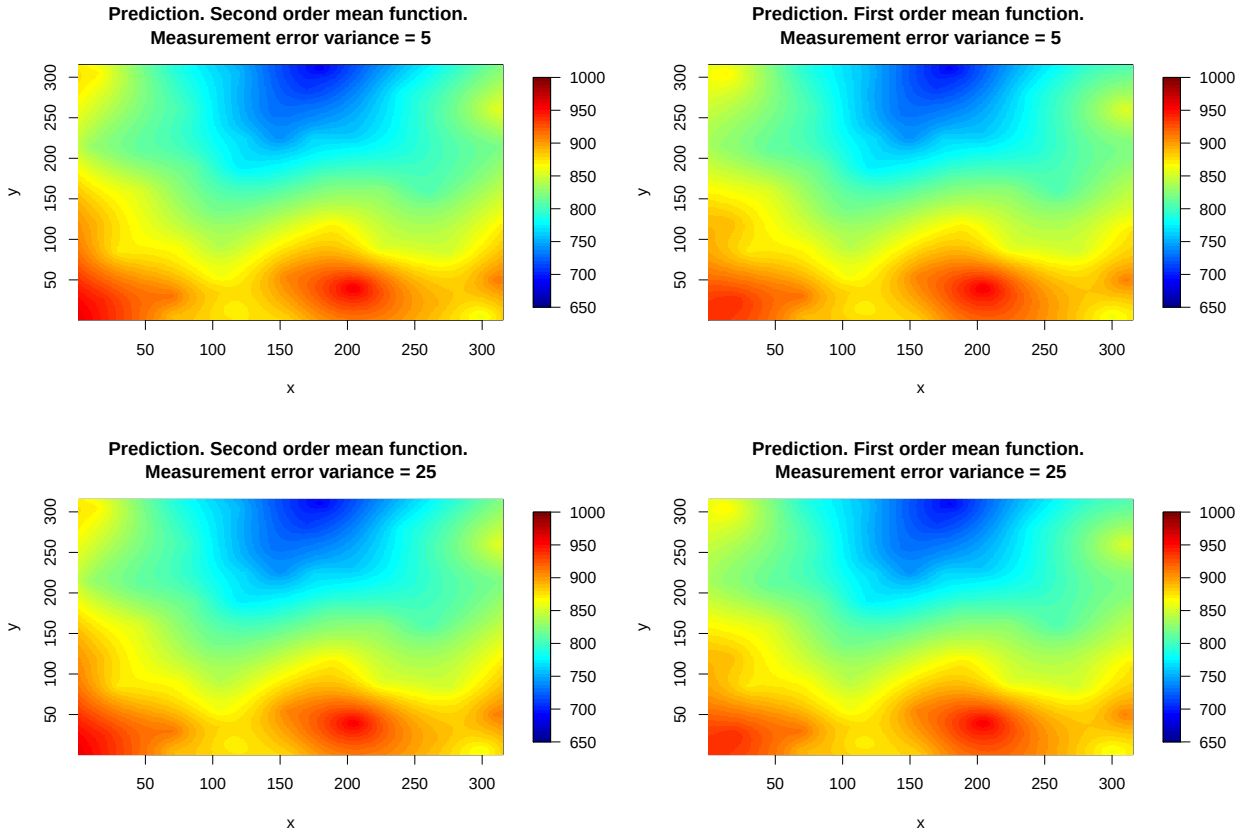
$$P(r((100,100)) < r_{0.9}) = 0.9 \Rightarrow \Phi\Big(\frac{r_{0.9} - \hat{r}_0}{\sigma_{\hat{r}}}\Big) = 0.9.$$

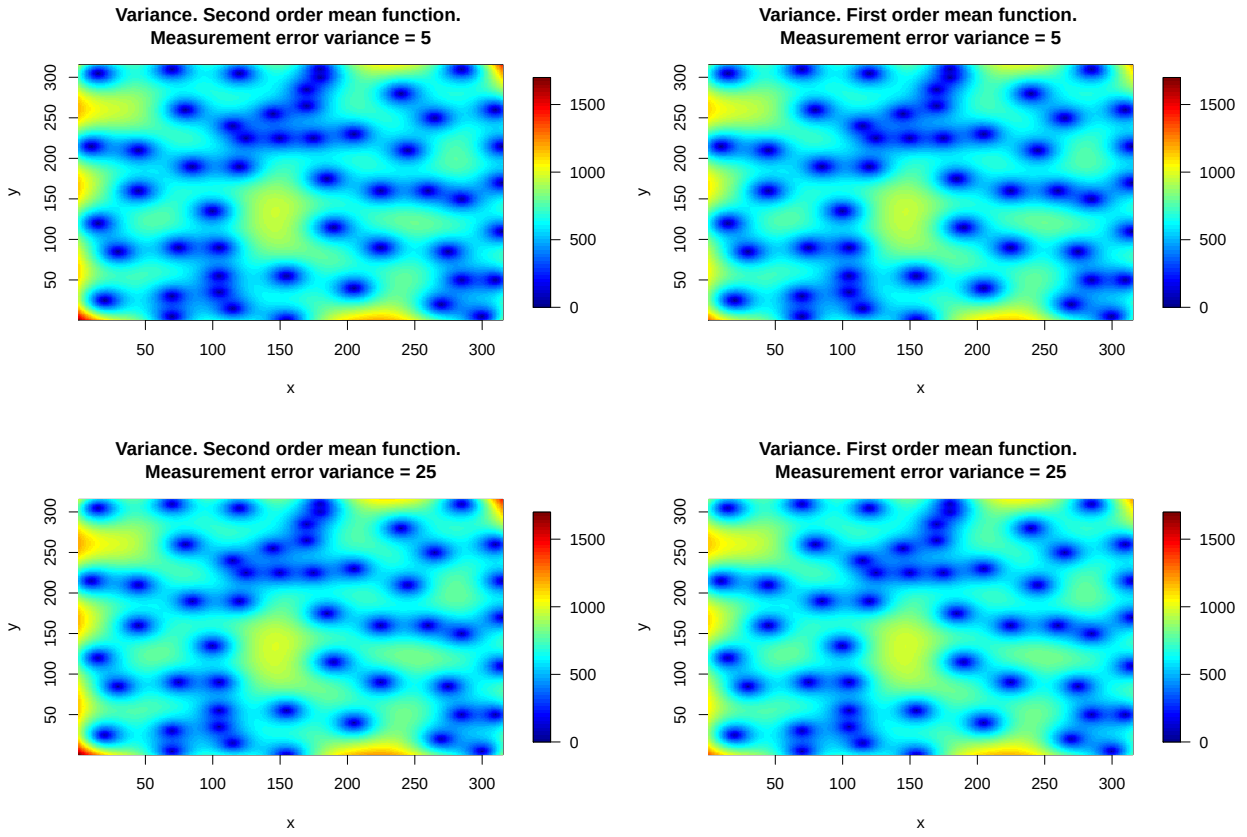
```
## Probability of r((100, 100)) being larger than 700m:  1
```

```
## Elevation at which the probability of r((100, 100)) being smaller is 0.9:  865.9738
```

e)

We now assume that the observations in the n_g-vector $\mathbf{d}$ are associated with observation errors being centered Gaussian and independent from each other and the terrain elevation, with error variance σ^2_ϵ. We calculate the Kriging predictions with associated prediction variance for the two values of the observation error variance, $\sigma^2_\epsilon \in \{5, 25\}$. The results are displayed below.





It is hard to see a difference in the plots of the Kriging predictions. The same goes for the Kriging variances. The added measurement error is not large enough to be visible on the plot. But if we look at the Kriging variance values we see that a larger measurement error variance gives a larger Kriging variance, which we would expect. If we increase this variance even more we might be able to see the difference in the plots.

f)

Our experiences using real data in this exercise has been useful. It has been helpful to visualize the data in order to get a better feeling with the models and methods used. But it could be useful to get introduced to some of the algorithms used for making the predictions, instead of just putting the data in a black box.

Problem 3: Parameter estimation

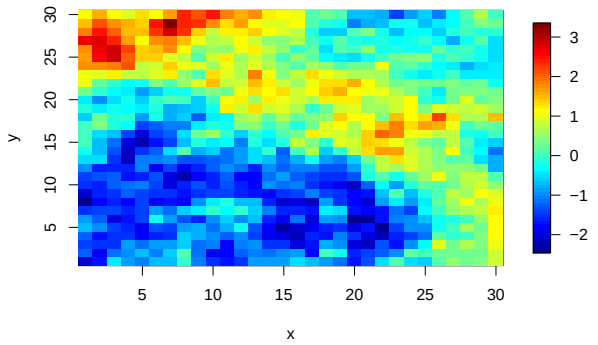
We consider the stationary Gaussian RF $\{r(\mathbf{x}); \mathbf{x} \in D \subset \mathbb{R}^2\}$ with $D : [(1, 30), (1, 30)]$, with

$$
\begin{aligned}
E\{r(\mathbf{x})\} &= \mu_r = 0 \\
Var\{r(\mathbf{x})\} &= \sigma_r^2 \\
Coor\{r(\mathbf{x}), r(\mathbf{x}')\} &= \exp\{-\tau/\xi_r\},
\end{aligned}
$$

with $\tau = |\mathbf{x} - \mathbf{x}'|$.

a)

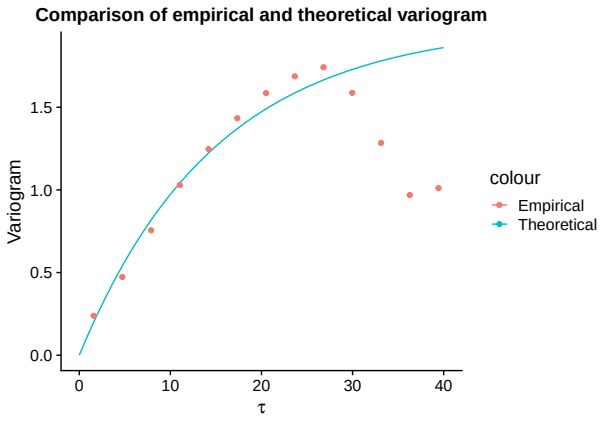
We discretize the Gaussian RF to get $\{r(x); x \in L\}$ on a grid $L : [30 \times 30] \in D$. The model parameters are set to $\sigma_r^2 = 2$ and $\xi_r = 15$. Then we generate one realization of the discretized Gaussian RF and display it.



From the display, it is reasonable to assume that the generated realization comes from a Gaussian RF. This is as expected because the realization is generated from a Gaussian RF.

b)

We now compute the empirical variogram based on the exact observations of the full realization displayed above. Then, the estimated variogram is displayed along with the correct variogram function $\gamma(\tau) = \sigma_r^2[1 - \rho(\tau)]$.

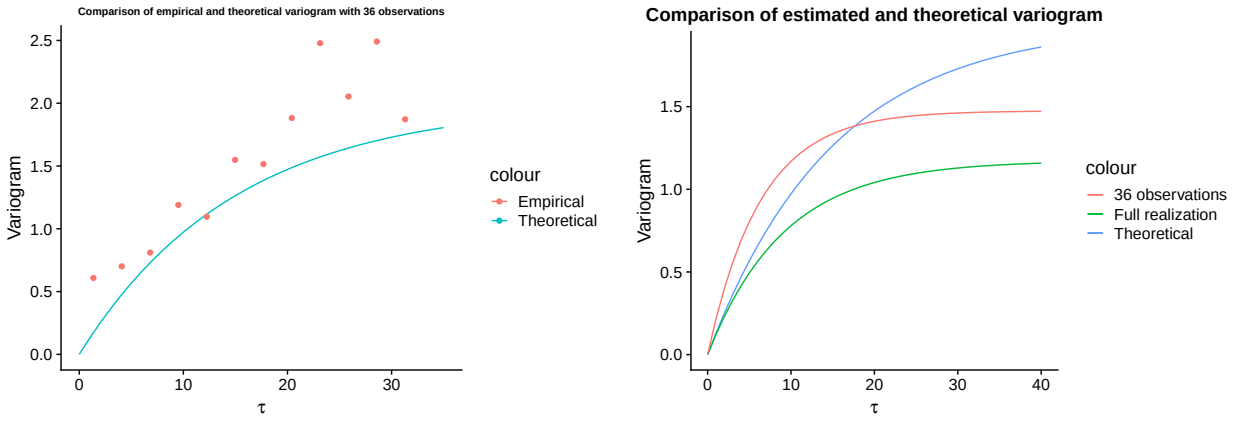


From the plot, we see that the empirical variogram differs somewhat from the theoretical variogram. The two variograms coincide well for small values of τ, but with greater τ the variograms differ significantly. This is as to be expected, because the empirical variogram is generated on only one realization and relatively few observations, especialy for large values of τ.

c)

We now generate 36 locations uniformly randomly on the grid L. Then we compute the empirical variogram estimate based on the corresponding 36 exact observations. The estimate is then displayed jointly with the theoretical variogram function.

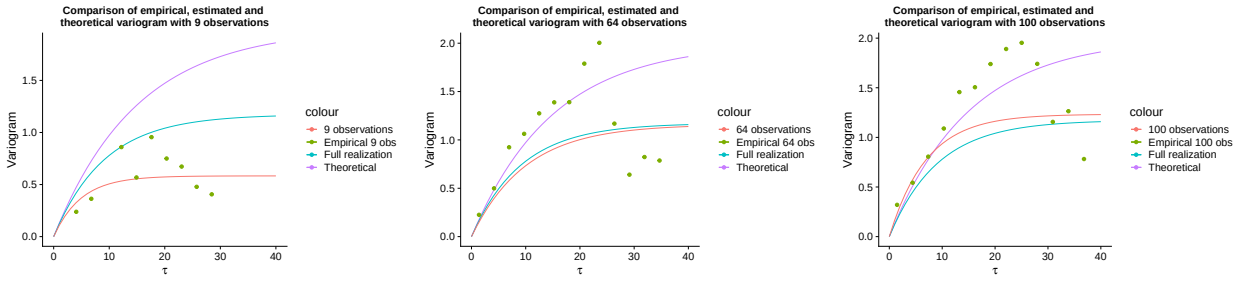
Then we consider the model parameters σ_r^2 and ξ_r to be unknown. These are then estimated by a maximum likelihood criterion based on exact observation of the full realization and based on the 36 observations. The two estimated variogram functions are then jointly displayed with the correct variogram function.



From the plots, we see that the empirical variogram based on only 36 observations fits the correct variogram worse than the one based on all 900 observations. However, the estimated variogram based on the 36 observations fits the correct variogram better than the estimated variogram based on the full realization. This result seems spurious, as adding more observations should in theory provide a better estimate of the true parameters.

d)

We then repeat the process with 9, 64 and 100 exact observations from the realization, and present the results.



From the plots, we see that including more observations gives a better fit to the correct variogram for both the empirical and the estimated variogram. This is what we would expect, and seems like a more reasonable result than the one achieved for 36 data points.

e)

In this problem, we have used the variogram function for the first time, and used maximum likelihood to find covariance parameters for the first time. The achieved results seemed bizarre, as 36 observations provided a better fit than 900 observations. We suspect this result to be due to chance, and would not expect to get the same result if the experiment was repeated. Again, some insight into the `likfit()` function would be interesting to understand what is happening, instead of treating the function as a black box.