# Statistical Data Science - Home assignment 2

Prof. Dr. Philipp Otto

> **Note**
>
> issue date :
> Submission date :
> Name, matriculation number:
> Evaluation:

## Problem - Simulation of dependent random processes

Simulate spatial lattice data $\{Z(s) : s = (s_1, s_2)' \in Z^2, 1 \leq s_1, s_2 \leq d\}$ from a gaussian distribution, where the covariance is

$$Cov(Z(s_i), Z(s_j)) = a \exp(-b||h||) \qquad \text{for all} \quad i, j = 1, \ldots, n \tag{1}$$

with $h = s_i - s_j$, $a > 0$, $b \geq 0$. Choose initially $d = 25$, $a = 0.2$, $b = 0.15$, and $||\cdot||$ as Euclidean norm.

1. Compare the resulting random fields $\{z(s)\}$ for different choices of $a$ and $b$. How $a$ and $b$ can be interpreted?
2. Compare the results for this so-called exponential covariance model to
   a) a spherical model,
   b) a linear model,
   c) and a power model.
3. Compare the results for different norms $||\cdot||$, e.g. $||\cdot||_p$ for $p \in \{1, 2, \infty\}$. Do the norms have an impact on the spatial dependence? In order to compare the results, the same random seed should be specified.

## Problem - Monte Carlo simulation study - computation time

Simulate the above described spatial model for increasing sizes of the spatial random field. Perform a Monte Carlo simulation study with $m = 100$ replications to evaluate the computation time, if the number of observations $n$ is increasing ($n \in \{16, 100, 1024, 4900\}$). Visualize your results graphically and shortly explain them.

## Problem: Covariance tapering

For covariance tapering, zeros are introduced into $C$ in order to make it sparse The tapered covariance function is then given by the product

$$C_{tap}(s_j - s_i) = C(s_j - s_i)C_\theta(s_j - s_i)$$

Explain why the tapering matrix $C_\theta$ must be chosen as valid covariance function (i.e., positive definite).

## Problem: Simulation of sparse matrices

Simulate binary matrices with 60, 80, 95 per cent zero elements. The matrices should be of dimension $100 \times 100$ and the random number generator should be initialized by a random seed.

1. Explain the difference between sparse matrices and band matrices.
2. Permute the simulated matrix using the Cuthill-McKee algorithm. Compute the bandwidth of the permuted matrices.
3. Why band matrices are (sometimes) preferred in computational statistics?
4. Compute the determinant!

## Problem: Monte Carlo simulation study - minimal bandwidth

Simulate the above described binary matrices and perform a Monte Carlo simulation study with $m = 1000$ replications. Compute the average minimal bandwidth, which can be achieved by the Cuthill-McKee algorithm. Visualize your results graphically and shortly explain them.

## Problem: Monte Carlo simulation study - sparse matrices

Simulate the above described binary matrices and perform a Monte Carlo simulation study with $m = 1000$ replications to evaluate the computation time and required memory (RAM) for computing the determinant. Assess the computational advantages

1. if a class for sparse matrices is used (e.g., C++ Eigen::SparseMatrix, Python scipy.sparse, R Matrix, ...),
2. if the matrices are permuted by the Cuthill-McKee algorithm.

Furthermore, how the computation time changes if the dimension of the matrices increases ($20 \times 20$, 50, $100 \times 100$) Visualize all results graphically and shortly explain them.