1 Introduction and Necessary Tools

Spatial statistics is a very broad subset of statistics, whose history and development spans astronomy, probability theory, mining engineering, forestry, agriculture and ecology. The field can be broadly broken into three subcategories:

- Continuous spatial variation [contains geostatistics]
  - Geophysical processes [e.g., temperature, precipitation, pressure]
  - Ores and sediments [e.g., along a stream or in a mining area]
  - Pollution [e.g., monitoring]
  - Parameterized ODEs/PDEs
  - Health data [e.g., blood pressure across time, cell counts across time]
  - Time series

- Discrete spatial variation [lattice, areal units]
  - County or state level census data
  - Disease mapping
  - Images
  - Agricultural fields
  - Spatial econometrics

- Spatial point processes

\[^1\text{Department of Applied Mathematics, University of Colorado, Boulder, CO. Author e-mail: william.kleiber@colorado.edu}\]
- Tree or species locations
- Earthquake epicenters
- Spatial epidemiology [e.g., disease morbidity and mortality]
- Terrorist attacks and crime patterns
- Eye movement behavior

[R examples (Intro.R): introductory examples]

Motivating examples:

- Estimating lapse rate
- Simulating terrain or making a DEM
- Spatial prediction
- Characterize and quantify the differences between functions/surfaces
- Estimate local extrema

1.1 Simple linear regression

Suppose we have a set of \( n \) observations \( y_1, \ldots, y_n \) that we wish to relate to a set of predictors or covariates \( x_1, \ldots, x_n \). We posit a relationship between the variable \( Y_i \) [that is, before making any observations, \( Y_i \) is random] and its covariate \( X_i \) as

\[
Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1}
\]

where \( \varepsilon_i \) is set of independent mean zero error terms each with variance \( \sigma^2 \). This is known as simple linear regression. (1) is an example of a statistical model.

If \( \varepsilon \) is assumed to be normal, given \( X_i = x_i \), this is equivalent to assuming \( [Y_i | X_i = x_i] \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \). The corresponding equation for an observed \( y_i \) and covariate \( x_i \) is

\[
y_i = \beta_0 + \beta_1 x_i + \varepsilon_i
\]

where now \( \varepsilon_i \) is an observed residual that is no longer random.
Example 1. In the lapse rate example we have \( n = 213 \) observations of average March/April/May temperature along with elevations \( x_1, \ldots, x_{213} \) ranging from 811 to 3448 meters. The model is

\[
\text{temp at location } i = \beta_0 + \beta_1(\text{elevation at location } i) + \text{error at location } i
\]

where \( \beta_0 \) can be interpreted as the average temperature at sea level and \( \beta_1 \) can be interpreted as the lapse rate, i.e., the change in temperature for a unit increase in elevation.

There are differing levels of assumptions that are typically made in practice regarding (1).

A1 1. The relationship (1) holds

2. \( \varepsilon_i \) are iid (independent and identically distributed) \( N(0, \sigma^2) \) for all \( i = 1, \ldots, n \)

A2 1. The relationship (1) holds

2. \( \mathbb{E}\varepsilon_i = 0 \)

3. \( \text{Var}\varepsilon_i = \sigma^2 \) (homoskedasticity)

4. \( \varepsilon_i \) and \( \varepsilon_j \) are iid

A3 1. The relationship (1) holds

2. \( \mathbb{E}\varepsilon_i = 0 \)

3. \( \text{Var}\varepsilon_i = \sigma^2 \) (homoskedasticity)

4. \( \varepsilon_i \) and \( \varepsilon_j \) are uncorrelated for \( i \neq j \)

What is common to all three levels of assumptions is that the posited relationship (1) holds, and that the residual random variables are mean zero and all have the same variance (a property known as homoskedasticity). The differences between the three assumptions are about the assumed behavior of the residuals. A1 is the strongest in that we explicitly assume the residuals are normally distributed, whereas A2 relaxes this assumption to being iid from some unspecified probability distribution. Statistical independence is a stronger statement than correlation, and A3 is the weakest in that we only put assumptions on the
first two moments of the variables. For instance, $E\varepsilon_1^3$ may be different than $E\varepsilon_2^3$ in A3, but not for A2 or A1.

Throughout the following development, we will work under assumptions A3, explicitly pointing out when A1 may be assessed. Our goals are usually two- to three-fold

1. Estimate parameters $\beta_0, \beta_1$ and $\sigma^2$ (and quantify uncertainty in our estimates)

2. Assess the validity of our assumed model (1)

3. Predict $Y$ at a new covariate value $X = x$.

1.2 OLS and diagnostics

The heuristic for estimating the regression parameters $\beta_0$ and $\beta_1$ is that we should minimize the distance between $y_i$ and its fitted value $\beta_0 + \beta_1 x_i$. [Draw picture on board]. Thus, the ordinary least squares (OLS) estimators for $\beta_0$ and $\beta_1$ are found by minimizing

$$
\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.
$$

Note that other types of distance can be used, but least squares is the most popular, in part due to the simplicity of the solution (and also a connection to maximum likelihood). Doing this minimization results in the estimators

$$
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
$$

$$
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
$$

where $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ and $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$. Notice that the estimators are linear in the observations. It can be shown that the OLS estimators are the best linear unbiased estimators (BLUEs) under squared loss.

Based on our estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we have estimated residuals

$$
\hat{\varepsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).
$$

If our assumptions are correct, then the estimated residuals $\{\hat{\varepsilon}_i\}_i$ should not have any structure and can be used to judge the quality of assumptions. One important diagnostic plot
is the residuals vs. fitted values scatterplot. If there are definite trends in the plot then the
assumed relationship (1) may be violated. If the variability changes with fitted values, then
homoskedasticity may be violated; this latter case is when the errors exhibit heteroskedas-
ticity. [Picture of second case] The strongest set of assumptions, A1, suggest the residual
terms $\varepsilon_i$ are normally distributed. To assess normality of the estimated residuals $\{\hat{\varepsilon}_i\}_i$, the
quantile-quantile plot (Q-Q plot) is often used. The Q-Q plot plots the theoretical quantiles
of a standard normal versus the estimated quantiles of the standardized observed residuals.
If the plot falls along the identity line, it may be reasonable to assume the errors arose
from a normal distribution. Q-Q plots can be used to assess heavy-tailedness and skewness
compared to the normal distribution. [Picture of standard quantile-quantile plot] Later in
the course we will explore other types of diagnostic plots that are particular to spatial data.

Lastly, we must estimate $\sigma^2$. An unbiased estimate of $\sigma^2$ can be found by scaling the
residual sum of squares by its degrees of freedom,

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

where the $-2$ occurs since we are estimating two parameters $\beta_0$ and $\beta_1$ in the mean function.
Recall the usual unbiased estimator for variance divides by $n - 1$, which is when only $\beta_0$ is
being estimated.

There’s nothing special about $X$ in (1), we can just as easily use some transformation of
$X$, such as

$$Y = \beta_0 + \beta_1 X^2 + \varepsilon$$

which might be appropriate for the lapse rate data. Or, we could use a transformation of $Y$,
such as

$$\log Y = \beta_0 + \beta_1 X + \varepsilon.$$ 

This latter case is equivalent to assuming

$$Y = \alpha \exp(\beta_1 X) \exp(\varepsilon),$$

that $Y$ depends exponentially on $X$ and the errors are multiplicative. The difference between
each of these models is the posited relationship between the variable $Y$ and the covariate
X. Typically the form of the assumed model will be based on expert understanding of the problem at hand, or can also be motivated by exploratory analysis. There are methods to statistically test for differences between modeling assumptions, but we will not cover these here. In the next example, we will see a case where assumption (4) is violated, in that the residuals display correlation.

We now have point estimates of \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), but can we quantify the uncertainty in our estimates? Note that \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) both involve \( y \), so before the experiment is performed these can be considered random variables. The variances of the OLS estimators are

\[
\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
\text{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.
\]

However, we don’t know the true \( \sigma^2 \), so the associated standard errors are \( SE = \sqrt{\text{Var}} \) where

\[
\text{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
\text{Var}(\hat{\beta}_0) = \frac{\hat{\sigma}^2}{n} + \frac{\hat{\sigma}^2 \bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.
\]

Thus, 95% approximate confidence intervals for our estimated parameters are

\[
\hat{\beta}_0 \pm 1.96SE(\hat{\beta}_0) \\
\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1).
\]

Note that if \( \varepsilon \sim N(0, \sigma^2) \) is exact and \( \hat{\sigma}^2 \) is used in \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), then we should use \( t \)-distribution quantiles. However, for \( n \gg 0 \), the normal approximation is ok.

### 1.3 Multiple regression and OLS

Sometimes it makes sense to regress \( Y \) on multiple covariates \( X_1, X_2, \ldots, X_p \), for instance elevation and latitude in the lapse rate example. If we have \( n \) variables that we observe, \( Y_1, \ldots, Y_n \), where the \( i \)th has \( p \) corresponding covariates \( X_{i1}, \ldots, X_{ip} \), then the multiple linear regression model assumes

\[
Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i
\]
with the only difference here being the introduction of extra covariates.

**Example 2.** Average temperatures depend on elevation as well as latitude, thus a possible model is

\[
\text{temp at location } i = \beta_0 + \beta_1 (\text{elevation at location } i) + \beta_2 (\text{latitude at location } i) + \text{error at location } i
\]

where now \( \beta_2 \) is the change in average temperature for a unit change in latitude with elevation held fixed.

It is convenient to use matrix notation to describe the model and resulting OLS estimators. In particular, define

\[
Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix},
\]

Note \( \beta \) is \( (p+1) \times 1 \) and \( X \) is \( n \times (p+1) \). We can then write the model for all observations \( Y_1, \ldots, Y_n \) succinctly as

\[
Y = X\beta + \varepsilon.
\]

[Example for lapse rate]

The OLS estimator for \( \beta \) is then found by minimizing

\[
\sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}))^2 = (Y - X\beta)'(Y - X\beta),
\]

which results in

\[
\hat{\beta}_{OLS} = (X'X)^{-1}X'Y.
\]

[Check dimensions on board]

Standard errors for the OLS estimator \( \hat{\beta}_{OLS} \) are found as the square roots of the diagonal of

\[
\text{Var}(\hat{\beta}_{OLS}) = \hat{\sigma}^2 (X'X)^{-1}
\]
where the residual variance is estimated by

\[ \hat{\sigma}^2 = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n - (p + 1)}, \]

this is an unbiased estimator for \( \sigma^2 \).

[R example (LapseRate.R)]

### 1.4 Pause: Crash Course in Matrix Algebra

A matrix \( A \) with \( n \) rows and \( m \) columns (i.e., an \( n \times m \) matrix) is an element of \( \mathbb{R}^n \times \mathbb{R}^m \). For example, a \( 3 \times 2 \) matrix has real-valued elements and

\[
A = \begin{pmatrix}
a_{11} & a_{12} \\
a_{21} & a_{22} \\
a_{31} & a_{32}
\end{pmatrix}
\]

where, generally \( a_{ij} \) will refer to the \((i, j)\)th element, that is, the row \( i \) column \( j \) element. Sometimes we shorthand this as

\[
A = (a_{ij})^{3,2}_{i=1,j=1} = (a_{ij})
\]

A vector is just a matrix with one column, (known as a column vector), e.g.,

\[
x = \begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{pmatrix}
\]

The transpose of a matrix is \( A^T = (a_{ji}) \), switching column and row places. E.g.,

\[
A = \begin{pmatrix}
2 & \pi \\
0 & 10 \\
-3 & 3
\end{pmatrix} \quad \text{has} \quad A^T = \begin{pmatrix}
2 & 0 & -3 \\
\pi & 10 & 3
\end{pmatrix}.
\]

Note that \((A^T)^T = A\). A matrix is square if its number of rows is the number of columns. A square matrix is symmetric if \( A = A^T \), e.g.,

\[
A = \begin{pmatrix}
1 & 3 \\
3 & 10
\end{pmatrix}
\]
The elements \( \text{diag}(\mathbf{A}) = (a_{11}, a_{22}, \ldots, a_{nn}) \) define the diagonal of a matrix. The \( n \)-dimensional identity matrix \( \mathbf{I} \) is the square \( n \times n \) matrix with 1s along the diagonal:

\[
\mathbf{I}_n = \begin{pmatrix}
1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1
\end{pmatrix}
\]

An upper triangular matrix is a matrix with zero entries below the main diagonal:

\[
\begin{pmatrix}
2 & 8 & 3 & 0 \\
0 & 0 & 10 & 7 \\
0 & 0 & 4 & 7 \\
0 & 0 & 0 & 5
\end{pmatrix}
\]

and a lower triangular matrix has zeros above the diagonal. If we wanted to sum all elements of

\[
\mathbf{A} = \begin{pmatrix}
a_{11} & a_{12} \\
a_{21} & a_{22} \\
a_{31} & a_{32}
\end{pmatrix}
\]

we can write this as

\[
\sum_{i=1}^{3} \sum_{j=1}^{2} a_{ij} = a_{11} + a_{12} + a_{21} + a_{22} + a_{31} + a_{32}
\]

Any two \( n \times p \) matrices \( \mathbf{A} \) and \( \mathbf{B} \) may be added by taking elementwise sums,

\[
\begin{pmatrix}
2 & \pi \\
0 & 10 \\
-3 & 3
\end{pmatrix} + \begin{pmatrix}
3 & 1 \\
4 & 2 \\
5 & 0
\end{pmatrix} = \begin{pmatrix}
5 & \pi + 1 \\
4 & 12 \\
2 & 3
\end{pmatrix}.
\]

Multiplication by a constant \( b \mathbf{A} = b(a_{ij}) = (b a_{ij}) \) multiplies elementwise. Some properties of matrix addition: (both \( \mathbf{A} \) and \( \mathbf{B} \) must be \( n \times m \)):

- \( \mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \)
- \( (\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T \)
- \( (\mathbf{A} - \mathbf{B})^T = \mathbf{A}^T - \mathbf{B}^T \)
- \( (\mathbf{x} + \mathbf{y})^T = \mathbf{x}^T + \mathbf{y}^T \)
• \((x - y)^T = x^T - y^T\)

Matrices \(A(n \times m)\) and \(B(m \times p)\) multiply to form a \(n \times p\) matrix \(C\) with

\[c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj},\]

that is, take the sum of elementwise products of the \(i\)th row of \(A\) with the \(j\)th column of \(B\).

E.g.,

\[
\begin{pmatrix}
2 & 0 \\
0 & 10 \\
-3 & 3
\end{pmatrix}
\begin{pmatrix}
3 & 1 \\
4 & 2
\end{pmatrix}
= \begin{pmatrix}
6 & 2 \\
40 & 20 \\
3 & 3
\end{pmatrix}.
\]

Two matrices can multiply if the number of columns of the first equals the number of rows of the second matrix, e.g., \(AB\) makes sense, but then \(BA\) may not conform, and, even if they do,

\(AB \neq BA\).

Some properties of matrix multiplication:

- \(A(B + C) = AB + AC\)
- \(A(B - C) = AB - AC\)
- \((A + B)C = AC + BC\)
- \((A - B)C = AC - BC\)
- \((A + B)(C + D) = AC + BC + AD + BD\)
- \((AB)C = A(BC)\)

Transposes work as

\((AB)^T = B^T A^T\).

Note, then, that if \(x \in \mathbb{R}^n\) and \(y \in \mathbb{R}^n\),

\[x^T y = \sum_{i=1}^{n} x_i y_i \in \mathbb{R}.\]
which is sometimes called the dot product of $x$ and $y$. Moreover the squared Euclidean norm of $x$ is defined as

$$\|x\|_2^2 \equiv x^Tx = \sum_{i=1}^n x_i^2$$

where $\| \cdot \|_2^2$ denotes the squared $L_2$ length of the vector $x$. More properties:

- $\sum_{i=1}^n a^Tx_i = a^T\sum_{i=1}^n x_i$
- $\sum_{i=1}^n Ax_i = A\sum_{i=1}^n x_i$
- $\sum_{i=1}^n (Ax_i)(Ax_i)^T = A\left(\sum_{i=1}^n x_ix_i^T\right)A^T$

If $A$ is square $(n \times n)$ and $x$ and $y$ are vectors,

$$y^TAy = \sum_{i=1}^n \sum_{j=1}^n y_iy_ja_{ij}$$

is called a \textit{quadratic form}. Note it is a scalar.

The \textit{inverse} of a square matrix $A$, denoted $A^{-1}$, is the matrix that satisfies

$$AA^{-1} = A^{-1}A = I,$$

if it exists. We have

$$(AB)^{-1} = B^{-1}A^{-1},$$

if all inverses exist.

A square $n \times n$ matrix $A$ is \textit{nonnegative definite} if, for any vector $a$,

$$a^TAa \geq 0.$$  

Positive definite matrices (replacing $\geq$ with $>$) \textit{always admit an inverse}. Additionally, $A$ has a “square root” called the \textit{Cholesky factor} or \textit{Cholesky decomposition}

$$A = T^TT$$

where $T$ is an invertible upper triangular matrix. The Cholesky factor $T$ is unique, but there are many other “square root” matrices.

The \textit{determinant} of a matrix is a number and is written $\det(A) = |A|$, and has the following properties:
\begin{itemize}
  \item \(\det I = 1\)
  \item \(\det(A^T) = \det A\)
  \item \(\det A^{-1} = 1/(\det A)\)
  \item \(\det(AB) = \det(A)\det(B)\) for square matrices
  \item \(\det(cA) = c^n\det A\) where \(A\) is \(n \times n\).
  \item The determinant of a triangular or diagonal matrix is the product of its diagonal.
\end{itemize}

The \textit{trace} of a matrix is the sum of its diagonals, \(\text{tr}A = \sum_{i=1}^{n} a_{ii}\). We have

\begin{itemize}
  \item \(\text{tr}(A + B) = \text{tr}A + \text{tr}B\)
  \item \(\text{tr}(AB) = \text{tr}(BA)\) (even if \(AB \neq BA\))
\end{itemize}

Two vectors \(a\) and \(b\) \textit{orthogonal} if

\[ a^T b = 0. \]

If \(\|a\|_2 = 1\) then \(a\) is said to be \textit{normalized}. Any vector can be \textit{normalized} by

\[ b = \frac{a}{\|a\|_2}. \]

A square matrix \(C = (c_1, c_2, \ldots, c_p)\) is said to be \textit{orthogonal} if its columns and rows are normalized and mutually orthogonal, whence

\[ C^T C = CC^T = I. \]

Thus, \(C^{-1} = C^T\). Multiplication by an orthogonal matrix has the effect of \textit{rotating the axes}, that is, if \(z = Cx\) then note

\[ z^T z = x^T C^T C x = x^T x \]

maintains the same length.

An \textit{eigenvalue} \(\lambda\) of a matrix \(A\) is a number where

\[ Ax = \lambda x \]
and $\mathbf{x}$ is called the corresponding eigenvector. They can be calculated by finding solutions to $\det(\mathbf{A} - \lambda \mathbf{I}) = 0$. Note that, if $\mathbf{x}$ is an eigenvector, and $k$ is a number, then

$$\mathbf{A}(k\mathbf{x}) = k\lambda(\mathbf{x}) = \lambda(k\mathbf{x})$$

so $k\mathbf{x}$ is also an eigenvector, and thus we usually scale eigenvectors to be unit length.

If $\lambda$ is an eigenvalue of $\mathbf{A}$ and $\mathbf{x}$ is the corresponding eigenvector, then $1\pm\lambda$ is an eigenvalue of $\mathbf{I} \pm \mathbf{A}$, and $\mathbf{x}$ is still the corresponding eigenvector.

If $\mathbf{A}$ is a square matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$,

- $\text{tr}\mathbf{A} = \sum_{i=1}^{n} \lambda_i$
- $\det\mathbf{A} = \prod_{i=1}^{n} \lambda_i.$

Importantly,

- If $\mathbf{A}$ is positive definite, then its eigenvalues are all positive
- If $\mathbf{A}$ is nonnegative definite, then its eigenvalues are either positive or zero.
- If $\mathbf{A}$ is symmetric, its eigenvectors are all mutually orthogonal.

By the last property, if $\mathbf{A}$ has eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, form

$$\mathbf{C} = [\mathbf{x}_1 \mathbf{x}_2 \cdots \mathbf{x}_n]$$

and note $\mathbf{C}$ is orthogonal. Then

$$\mathbf{A} = \mathbf{A}\mathbf{I}$$
$$= \mathbf{A}\mathbf{C}\mathbf{C}^T$$
$$= [\mathbf{A}\mathbf{x}_1 \mathbf{A}\mathbf{x}_2 \cdots \mathbf{A}\mathbf{x}_n]\mathbf{C}^T$$
$$= [\lambda_1\mathbf{x}_1 \lambda_2\mathbf{x}_2 \cdots \lambda_n\mathbf{x}_n]\mathbf{C}^T$$
$$= \mathbf{C}\mathbf{D}\mathbf{C}^T$$
where
\[ D = \begin{pmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_n
\end{pmatrix} \]
is diagonal with eigenvalues on the diagonal. This is known as the \textit{spectral decomposition} or \textit{eigendecomposition} of \( A \). We can also diagonalize \( A \) by
\[ C^T A C = D. \]

A \textit{square root matrix} is \( A^{1/2} = CD^{1/2}C^T \), where \( D^{1/2} = \text{diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2}) \). The square of \( A \) can be calculated
\[ A^2 = CD^2C^T \]
and the inverse
\[ A^{-1} = CD^{-1}C^T \]
all of which only involve applying these functions to the eigenvalues in \( D \), which is one reason the eigendecomposition is nice.

The \textit{rank} of a matrix is the number of linearly independent column vectors (or row vectors). If \( A \) is \( n \times p \), then

- \( \text{rank}(A) \leq \min(n, p) \)
- \( \text{rank}(A^T A) = \text{rank}(A A^T) = \text{rank}(A) \)
- If \( A \) is \( n \times n \) and has full rank \( (n) \) then \( A \) is invertible.

We will mostly be only dealing with full rank matrices.

Let \( A \) be an \( n \times p \) matrix of rank \( k \). The \textit{singular value decomposition} of \( A \) is
\[ A = U D V^T \]
where \( U \) is \( n \times k \), \( D \) is \( k \times k \) and \( V \) is \( p \times k \). \( D = \text{diag}(\lambda_1, \ldots, \lambda_k) \) contains the positive square roots of the \( \lambda_1^2, \ldots, \lambda_k^2 \) nonzero eigenvalues of \( A A^T \) or \( A^T A \). The \( k \) columns of \( U \) are
normalized eigenvectors of $\mathbf{A}\mathbf{A}^T$ corresponding to eigenvalues $\lambda_1^2, \ldots, \lambda_k^2$. The $k$ columns of $\mathbf{V}$ are the normalized eigenvectors of $\mathbf{A}^T\mathbf{A}$ corresponding to eigenvalues $\lambda_1^2, \ldots, \lambda_k^2$. Then,

$$\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}.$$ 

If $\mathbf{A}$ is positive definite, then the singular value decomposition is the same as the spectral decomposition.

### 1.5 Multivariate normal

Recall the definition of covariance and correlation: if $X$ and $Y$ are random variables with mean and standard deviation $\mu_X, \sigma_X$ and $\mu_Y, \sigma_Y$, respectively, then

$$\text{Cov}(X, Y) := \mathbb{E}((X - \mu_X)(Y - \mu_Y))$$

and

$$\text{Cor}(X, Y) := \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} (= \rho).$$

In particular, $\text{Cor}(X, Y) \in [-1, 1]$ is unitless and is a measure of the linear dependence between $X$ and $Y$, and $\text{Cov}(X, Y)$ is in the units of $X$ times the units of $Y$. Covariances enjoy the following properties:

- $\text{Var}X = \text{Cov}(X, X)$
- $\text{Cov}(aX + b, Y) = a \text{Cov}(X, Y)$ for any $a, b \in \mathbb{R}$ [prove]
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$.

How are covariances defined, and how do they act, for vectors of random variables? Suppose $\mathbf{X} = (X_1, \ldots, X_n)'$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)'$ are two random vectors, where $'$ denotes the transpose. Define $\text{Cov}(\mathbf{X}, \mathbf{Y})$ to be the $n \times m$ matrix with $(i, j)$-th entry $\text{Cov}(X_i, Y_j)$, that is,

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = \begin{pmatrix}
\text{Cov}(X_1, Y_1) & \text{Cov}(X_1, Y_2) & \cdots & \text{Cov}(X_1, Y_m) \\
\text{Cov}(X_2, Y_1) & \text{Cov}(X_2, Y_2) & \cdots & \text{Cov}(X_2, Y_m) \\
\vdots & \vdots & \ddots & \vdots \\
\text{Cov}(X_n, Y_1) & \text{Cov}(X_n, Y_2) & \cdots & \text{Cov}(X_n, Y_m)
\end{pmatrix}$$

The following properties also hold for covariances of random vectors:
• \( \text{Var} \mathbf{X} = \text{Cov} (\mathbf{X}, \mathbf{X}) \) [This is how we define the variance of a random vector]

• \( \text{Cov} (A \mathbf{X} + \mathbf{\mu}, B \mathbf{Y} + \mathbf{\nu}) = A \text{Cov} (\mathbf{X}, \mathbf{Y}) B' \) for any \( k \times n \) matrix \( A \), \( j \times m \) matrix \( B \), \( \mathbf{\mu} \in \mathbb{R}^k \) and \( \mathbf{\nu} \in \mathbb{R}^j \)

• \( \text{Cov} (\mathbf{X}, \mathbf{Y}) = \text{Cov} (\mathbf{Y}, \mathbf{X})' \).

If \( \mathbf{X} \) is a random vector such that \( \text{Var} X_i = \sigma_i^2 \), then the covariance matrix \( \text{Cov} (\mathbf{X}, \mathbf{X}) = (\text{Cov} (X_i, X_j))_{i,j=1}^n \) can be transformed into a correlation matrix

\[
\text{Cor} (\mathbf{X}, \mathbf{X}) = \left( \frac{\text{Cov} (X_i, X_j)}{\sigma_i \sigma_j} \right)_{i,j=1}^n.
\]

The correlation matrix is usually easier to interpret since it consists of all pairwise correlations between the component random variables.

**Definition 3.** We say the random vector \((X_1, X_2)\) has a bivariate normal distribution if

\[
f(x_1, x_2) = \frac{1}{\sqrt{(2\pi)^2 \det \left( \begin{array}{cc} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{array} \right)}} \exp \left( -\frac{1}{2} \left( \begin{array}{c} x_1 - \mu_1 \\ x_2 - \mu_2 \end{array} \right)' \left( \begin{array}{cc} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{array} \right)^{-1} \left( \begin{array}{c} x_1 - \mu_1 \\ x_2 - \mu_2 \end{array} \right) \right)
\]

for \((x_1, x_2) \in \mathbb{R}^2\).

Here, parameters have the following interpretations,

• \( \mathbb{E} X_1 = \mu_1 \)

• \( \mathbb{E} X_1 = \mu_2 \)

• \( \text{Var} X_1 = \sigma_1^2 \)

• \( \text{Var} X_2 = \sigma_2^2 \)

• \( \text{Cov} (X_1, X_2) = \rho \sigma_1 \sigma_2 \)

• \( \text{Cor} (X_1, X_2) = \rho \).

Note that the marginals are also normally distributed, e.g., \( X_1 \sim N(\mu_1, \sigma_1^2) \).

[Bivariate density with projected marginals]
Definition 4. We say the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ has a multivariate normal distribution if

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp \left( -\frac{1}{2} (\mathbf{x} - \mu)' \Sigma^{-1} (\mathbf{x} - \mu) \right)$$

for $\mathbf{x} \in \mathbb{R}^n$. We often write $\mathbf{X} \sim MVN(\mu, \Sigma)$, $\mathbf{X} \sim N_n(\mu, \Sigma)$ or $\mathbf{X} \sim N(\mu, \Sigma)$.

Here, $\mathbb{E}\mathbf{X} = \mu$ is the vector of means (i.e., $\mathbb{E}X_i = \mu_i$), and $\text{Cov}(\mathbf{X}, \mathbf{X}) = \Sigma$ is the covariance matrix (sometimes called variance-covariance matrix). The $(i, j)$th entry of $\Sigma$ is $\text{Cov}(X_i, X_j)$. [Imagine this as two parameters, $\mu$ tells us the average behavior of the vector, and $\Sigma$ tells us how all elements relate to each other]

The covariance matrix is symmetric since $\Sigma_{ij} = \text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i) = \Sigma_{ji}$. The diagonal of $\Sigma$ contains the variances of each component. Additionally, we have

- Each $X_i$ is marginally normally distributed: $X_i \sim N(\mu_i, \sigma_i^2)$
- If $A$ is a $k \times n$ real matrix and $\mathbf{b} \in \mathbb{R}^k$, then $A\mathbf{X} + \mathbf{b} \sim MVN(A\mu + \mathbf{b}, A\Sigma A')$.

The connection to spatial statistics is to think of $X_i$ as being a process, say temperature, at a particular location, so $\mathbf{X}$ is all of our temperatures at a bunch of places. Then two locations that are nearby will have similar temperatures, which should be endowed in the covariance matrix $\Sigma$.

[R example (MultivariateNormal.R)]

1.6 GLS

In the lapse rate example we saw that the errors $\varepsilon_i$ showed evidence of correlation. In this case, we don’t have $n$ uncorrelated samples, the effective number of data points is less than $n$ due to correlation. Thus, our standard errors for $\hat{\beta}_0$ and $\hat{\beta}_1$ are too optimistic (i.e., too narrow).

We require one more relaxation to our assumptions, in particular

A4 1. The relationship $\mathbf{Y} = \mathbf{X} \beta + \varepsilon$ holds
The key difference here is that the residuals are allowed to be correlated, i.e., $\text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$.

Define the covariance matrix of the vector $\varepsilon$ to be

$$\Sigma = \text{Cov}(\varepsilon, \varepsilon)$$

Then the generalized least squares estimator for $\beta$ is

$$\hat{\beta}_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y.$$ 

Note the connection to OLS: if the errors are uncorrelated and have the same variance then $\Sigma = \sigma^2 I$, and

$$\hat{\beta}_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$$
$$= (\sigma^{-2}X'X)^{-1}X'\sigma^{-2}Y$$
$$= (X'X)^{-1}X'Y$$
$$= \hat{\beta}_{OLS}.$$ 

Standard errors for the GLS estimator $\hat{\beta}_{GLS}$ are from

$$\text{Var}(\hat{\beta}_{GLS}) = (X'\Sigma^{-1}X)^{-1}.$$ 

[which is the same as the variance for OLS when $\Sigma = \sigma^2 I$]

Both the OLS and GLS estimators are unbiased:

$$E\hat{\beta}_{OLS} = E(X'X)^{-1}X'Y$$
$$= (X'X)^{-1}X'EX$$
$$= (X'X)^{-1}X'E(X\beta + \varepsilon)$$
$$= (X'X)^{-1}X'X\beta$$
$$= \beta$$
\[ \mathbb{E} \hat{\beta}_{GLS} = \mathbb{E}(X^\prime \Sigma^{-1} X)^{-1} X^\prime \Sigma^{-1} Y \]
\[ = (X^\prime \Sigma^{-1} X)^{-1} X^\prime \Sigma^{-1} \mathbb{E} Y \]
\[ = (X^\prime \Sigma^{-1} X)^{-1} X^\prime \Sigma^{-1} (X \beta + \varepsilon) \]
\[ = (X^\prime \Sigma^{-1} X)^{-1} X^\prime \Sigma^{-1} X \beta \]
\[ = \beta, \]

but it can be shown that

\[ \text{Var}(\mathbf{a}^\prime \hat{\beta}_{GLS}) \leq \text{Var}(\mathbf{a}^\prime \hat{\beta}_{OLS}) \]

for any choice of \( \mathbf{a} \). In fact, this holds for any other unbiased estimator: \( \hat{\beta}_{GLS} \) is the best (under squared loss) linear unbiased estimator for \( \beta \).

### 1.7 Prediction based on a linear model

Suppose \( y_1, \ldots, y_n \) are a set of data from \( n \) spatial locations with coordinates \((\text{lon}_i, \text{lat}_i)\). We might consider a model of the form

\[ Y = \beta_0 + \beta_1 \text{lon} + \beta_2 \text{lat} + \varepsilon \]

or

\[ Y = \beta_0 + \beta_1 \text{lon} + \beta_2 \text{lat} + \beta_3 (\text{lon} \cdot \text{lat}) + \beta_4 \text{lon}^2 + \beta_5 \text{lat}^2 + \varepsilon \]

or something similar. In other words, we are describing spatial variation as a polynomial in the coordinates (in this example up to degree 2). In either case, let’s put the data equations in matrix notation:

\[ \mathbf{Y} = \mathbf{X} \beta + \varepsilon. \]

Suppose we want to know \( Y_* \) at a new location \((\text{lon}_*, \text{lat}_*)\). Equivalently, predict \( Y_* \) for new set of covariates \( \mathbf{x}_* = (1, x_{s1}, x_{s2}, \ldots, x_{sp})' \) (which might just be \((1, \text{lon}_*, \text{lat}_*)^T\) or \((1, \text{lon}_*, \text{lat}_*, \text{lon}_* \text{lat}_*)^T\)). There are two we might want to predict with uncertainty at \( \mathbf{x}_* \):

- The mean value \( \mathbb{E}(y_*|\mathbf{x}_*) = \mathbf{x}_*^' \beta \) (fit)
• A single new observation \( y_* = x'_* \beta + \varepsilon_* \) (prediction).

[In the first case we want to quantify our uncertainty about the mean function, whereas in the second case we want to include our uncertainty about what the residual will be]

The natural point predictor for both cases is

\[
\hat{y}_* = x'_* \hat{\beta} = x'_* (X'X)^{-1} XY.
\]

[Note predictor is linear in the observations \( Y \)!

Our uncertainty depends on which case we’re looking at. Recall we define \( SE(\cdot) = \sqrt{\text{Var}(\cdot)} \).

The standard error of prediction is then

\[
\text{Var}(x'_* \hat{\beta}) = \hat{\sigma}^2 x'_* (X'X)^{-1} x_* \quad \text{(fit)}
\]
\[
\text{Var}(x'_* \hat{\beta} + \varepsilon_*) = \hat{\sigma}^2(1 + x'_* (X'X)^{-1} x_*) \quad \text{(prediction)}
\]

and our 95% predictive confidence intervals are then

\[
\hat{y}_* \pm 1.96 \times SE(x'_* \hat{\beta}) \quad \text{(fit)}
\]
\[
\hat{y}_* \pm 1.96 \times SE(x'_* \hat{\beta} + \varepsilon_*) \quad \text{(prediction)}
\]

[Note the standard error of prediction is greater than just \( \hat{\sigma} \), since there is uncertainty in estimating \( \beta \)]. If the residuals \( \varepsilon \) are correlated, then we can improve the point and interval estimates by taking account of this extra structure – this is what kriging is designed for.

[R example (LapseRate2.R)]
2 Competing Frameworks for Spatial Smoothing

Given a set of data, say in time, \( Y(s_1), \ldots, Y(s_n) \) where \( s_i \in \mathbb{R} \), we might like to interpolate, smooth or predict a value at a new location \( s_0 \) (where \( s_0 \) could be outside of the domain of the data for extrapolation, or in the domain of the data for interpolation or smoothing).

There are at least four popular viewpoints for tackling this problem:

1. Use a locally-weighted combination of the \( Y \)s:

\[
\hat{Y}(s_0) = \frac{\sum_{i=1}^{n} w_i Y(s_i)}{\sum_{i=1}^{n} w_i}
\]

where \( w_i \) is the weight on the \( i \)th data point (and the denominator ensures \( \sum_{i=1}^{n} w_i = 1 \)). Inverse distance weighting (IDW) is one way of doing this where we rig \( w_i = 1/|s_0 - s_i|^p \) for some positive power parameter \( p \).

2. Consider a linear model of the form

\[
Y(s) = f(s) + \varepsilon(s) = \beta_0 + \beta_1 k_1(s) + \beta_2 k_2(s) + \cdots + \beta_p k_p(s) + \varepsilon(s)
\]

where \( k_1, \ldots, k_p \) are some functions of space (a simple example is polynomial regression \( k_i(s) = s^i \)) and \( \varepsilon(s) \) is an error at a particular location. There are lots of ways we could estimate the \( \beta_i \)s, an easy way is OLS where we minimize

\[
\sum_{i=1}^{n} (Y(s_i) - f(s_i))^2
\]

in which case our predictor would be

\[
\hat{Y}(s_0) = \hat{f}(s_0) = \hat{\beta}_0 + \hat{\beta}_1 k_1(s_0) + \hat{\beta}_2 k_2(s_0) + \cdots + \hat{\beta}_p k_p(s_0).
\]

3. Instead of using a polynomial function for \( f(s) \) in the previous model, we could just say something general like \( f \) is any function that is close to the data. An estimator for
\( f \) could minimize
\[
\sum_{i=1}^{n} (Y(s_i) - f(s_i))^2
\]
except that there are a lot of functions that will make this minimization exactly zero (and are not good fits). So we might consider a penalty that enforces some degree of smoothness on \( f \) while being close to the data:
\[
\sum_{i=1}^{n} (Y(s_i) - f(s_i))^2 + \lambda \int f''(x)^2 dx
\]
where \( \lambda > 0 \) is a smoothing parameters and the integral penalty ensures that the second derivative does not get too large (i.e., the function is not too wiggly). Whatever \( f \) minimizes the penalized least squares problem is the predictor:
\[
\hat{Y}(s_0) = \hat{f}(s_0).
\]

4. Think of \( Y(s_1), \ldots, Y(s_n) \) as a realization of a stochastic process where we define means and covariances as functions of \( s \). The choice of how we choose models for covariances will turn out to be a prior on a class of functions \( f \). If we consider linear combinations of the data values as predictors we might want to minimize
\[
\mathbb{E}((Y(s_0) - (w_1 Y(s_1) + \cdots + w_n Y(s_n)))^2)
\]
in which case we get the so-called kriging predictor
\[
\hat{Y}(s_0) = \hat{w}_1 Y(s_1) + \cdots + \hat{w}_n Y(s_n).
\]
It will turn out that all four of these approaches are equivalent under a certain lens.
3 Basic Stochastic Process Theory

Stochastic processes play the fundamental role in describing and modeling the behavior of spatial data. We call \( Z(s) \) as stochastic process where \( s \in D \subset \mathbb{R}^d \) [think of \( Z \) as temperature at a set of locations \( s \) like Colorado, \( D \) indicating CO] if \( Z(s) \) is a random variable for all \( s \). You can think of \( Z(s) \) as a random function or a random surface. Here, \( d \) is the dimension of the spatial support of the process; usually \( d = 1 \) (time series), \( d = 2 \) (planar processes) or \( d = 3 \) (e.g., pressure at different altitudes over the USA).

[Note we should really write \( \{Z(s)\}_s \) as describing the whole process, since \( Z(s) \) is the random variable at location \( s \). In class we will abuse notation and usually write \( Z \) and \( Z(s) \) to mean the process, but will try to make it clear when \( Z(s) \) refers to the variable at a specific location].

Restricting attention to a set of “observation locations” \( s_1, \ldots, s_n \), we have that \( Z(s_1), \ldots, Z(s_n) \) is just a set of \( n \) random variables that may or may not be related. We denote the joint probability density function (pdf) of \( (Z(s_1), \ldots, Z(s_n))' \) as \( f(z_1, \ldots, z_n) \) and its cumulative distribution function (cdf) as

\[
F(z_1, \ldots, z_n) = P(Z(s_1) \leq z_1, \ldots, Z(s_n) \leq z_n).
\]

**Definition 5.** The stochastic process \( Z(s) \) is a Gaussian process if all finite-dimensional distributions are multivariate normal, that is, given any \( s_1, \ldots, s_n \), the vector \( (Z(s_1), \ldots, Z(s_n))' \) is multivariate normal.

Note that if \( Z(s) \) is a Gaussian process then its entirely probability distribution is described by a mean vector and covariance matrix.

3.1 Examples

**Example 6.** Let \( d = 1 \) and \( D = \mathbb{Z} \). Define a stochastic process by

\[
Z(i) = \begin{cases} 
U(0,1) & \text{if } i \text{ odd} \\
\text{Gamma}(1,2) & \text{if } i \text{ even} \quad [1,2 \text{ are shape and rate}]
\end{cases}
\]


where $Z(i)$ independent of $Z(j)$ for $i \neq j$. Then the joint pdf of $(Z(1), Z(2), Z(3))'$ is

$$f(z_1, z_2, z_3) = 1 \cdot 1 \cdot 2 \exp(-2z_2), \quad z_1 \in [0, 1], z_2 > 0, z_3 \in [0, 1].$$

**Example 7.** Let $d = 1$ and $s \in D = [0, 1] \subset \mathbb{R}$. Define a stochastic process $Z(s)$ by

$$Z(s) = U \cos(2\pi s) + V \sin(2\pi s)$$

where $U$ and $V$ are independent $U(0, 1)$ random variables.

**Example 8.** Let $d = 1$ and $s \in D = [0, 1] \subset \mathbb{R}$. Define a stochastic process $Z(s)$ by

$$Z(s) = W_0 + W_1 s + W_2 s^2 + W_3 s^3$$

where $W_i$ are independent $N(0, 1)$ random variables for $i = 0, \ldots, 3$.

[On homework: simulate these three examples, grad students find marginal and joint distributions for Examples 7 and 8].

**Example 9.** Let $d = 1$ and $s \in \mathbb{Z}$. Let $Z(s)$ be a Gaussian process with

$$\text{Cov}(Z(s), Z(t)) = \begin{cases} 1, & s = t \\ 0.5, & |s - t| = 1 \\ 0, & \text{otherwise.} \end{cases}$$

So, for example, the covariance matrix for $Z = (Z(1), Z(2), Z(3), Z(4))^T$ is

$$\text{Var} Z = \begin{pmatrix} 1 & 0.5 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0 \\ 0 & 0.5 & 1 & 0.5 \\ 0 & 0 & 0.5 & 1 \end{pmatrix}.$$  

### 3.2 Mean and Covariance Functions

Generally, the two most important parts of a spatial process are its mean and covariance functions.

**Definition 10.** The mean function of a stochastic process $Z(s)$ is

$$\mu(s) = \mathbb{E}Z(s).$$
Definition 11. The covariance function of a stochastic process $Z(s)$ is

$$C(s_1, s_2) = \text{Cov}(Z(s_1), Z(s_2)).$$

The mean function describes the average evolution of the process across space, while the covariance function describes the variability about this mean curve. Notice the general mean function depends on the location of $Z$, and the general covariance function depends on the pair of locations $(s_1, s_2)$. Technically we should state that we are working with processes whose covariance function exists – for example, this rules out the case where $Z(s)$ is a Cauchy random variable. [We won’t state this assumption again, but will tacitly assume throughout the rest of the class the existence of $Z$’s first and second moments].

3.3 Stationarity and Isotropy

Although we have the general definition of a mean and covariance function, in practice we typically require simplifying assumptions to reduce the classes of possible models. The two usual simplifications are when $Z(s)$ is assumed to be stationary or isotropic.

Definition 12. A stochastic process $Z(s)$ is strictly stationary or strongly stationary if all finite-dimensional distributions are invariant under shifts, that is,

$$P(Z(s_1) \leq z_1, \ldots, Z(s_n) \leq z_n) = P(Z(s_1 + h) \leq z_1, \ldots, Z(s_n + h) \leq z_n)$$

for any vector $h \in \mathbb{R}^d$ and any choices of spatial locations.

Note that strict stationarity is a statement about the entire probability structure of $Z$, whereas the following type relaxes this assumption.

Definition 13. A stochastic process $Z(s)$ is weakly stationary or second-order stationary if the mean is spatially constant,

$$\mathbb{E}Z(s) = \mathbb{E}Z(s + h) \ (= \mu)$$

and the covariance is a function of only lag,

$$\text{Cov}(Z(s), Z(s + h)) = C(h)$$
for any choice of \( h \in \mathbb{R}^d \). [Second-order stationarity is named this way since assumptions are only made up to the second moment].

All strictly stationary processes are weakly stationary, but not all weakly stationary processes are strictly stationary. Weak and strong stationarity are equivalent for Gaussian processes – however this is a very special case, in general the two are not equivalent. Throughout the rest of the course we will use the term stationary to mean weakly stationary.

For a stationary process, the mean is constant and the covariance only depends on the lag vector between two locations. A special subclass of stationary processes are isotropic processes.

**Definition 14.** A stochastic process \( Z(s) \) is isotropic if it is weakly stationary and the covariance is a function of only distance

\[
\text{Cov}(Z(s), Z(s + h)) = C(\|h\|)
\]

for any choice of \( h \in \mathbb{R}^d \).

Isotropy means that the process is rotation invariant, and note that \( \mathbb{E}Z(s) \) is spatially constant for isotropic processes. Notationally, we will sometimes write

\[
\text{Cov}(Z(s), Z(s + h)) = C(\|h\|) \\
\text{Cov}(Z(s_1), Z(s_2)) = C(\|s_1 - s_2\|) \\
\text{Cov}(Z(s_1), Z(s_2)) = C(r) \text{ where } r = \|s_1 - s_2\|
\]

all of which are equivalent.

What do weak stationarity (and thus isotropy) imply about the variance at at two locations \( s_1 \) and \( s_2 \)?

\[
\text{Var}Z(s_1) = \text{Cov}(Z(s_1), Z(s_1)) = C(s_1 - s_1) = C(0) \\
= C(s_2 - s_2) = \text{Cov}(Z(s_2), Z(s_2)) = \text{Var}Z(s_2)
\]

Stationary (also isotropic) processes have spatially constant mean and variance.

[Example (ThreeSims.png)]
3.4 Parametric Covariance Functions

We will mostly be focusing on parametric spatial models, that is, mean and covariance functions that can be described with a few parameters. [For instance, if \( X \sim N(\mu, \sigma^2) \), we say that \( \mu \) and \( \sigma^2 \) are the parameters defining the probability structure of \( X \)].

Parametric covariance functions are especially useful as they specify spatial dependence of the process between arbitrary location pairs, not only those making up a set of observations. [Plot 2 locations, supposing we can estimate covariance between those two but want to know the correlation with a third location]

[Some intuition: what should we expect covariance functions to look like?]

Some of the most common parametric models are:

- Exponential:
  \[
  C(r) = \exp \left( -\frac{r}{a} \right), \quad a > 0
  \]

- Powered exponential:
  \[
  C(r) = \exp \left( -\left( \frac{r}{a} \right)^\alpha \right), \quad a > 0, 0 < \alpha \leq 2
  \]

- Matérn:
  \[
  C(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{r}{a} \right)^\nu K_\nu \left( \frac{r}{a} \right), \quad a > 0, \nu > 0,
  \]
  where \( K_\nu \) is a modified Bessel function of the second kind of order \( \nu \).

- Cauchy:
  \[
  C(r) = \left( 1 + \left( \frac{r}{a} \right)^\alpha \right)^{-\beta/\alpha}, \quad a > 0, \beta > 0, 0 < \alpha \leq 2
  \]

- Spherical:
  \[
  1 - \frac{3}{2} \frac{r}{a} + \frac{1}{2} \left( \frac{r}{a} \right)^3, \quad a > 0, 0 \leq r < a,
  \]
  only valid for \( d \leq 3 \).

- Exponentially damped cosine:
  \[
  C(r) = \exp \left( -b \frac{r}{a} \right) \cos \left( \frac{r}{a} \right), \quad a > 0, b \geq \left( \tan \frac{\pi}{2d} \right)^{-1}
  \]
In all of the above, $a$ is called a range (or sometimes scale) parameter that determines the rate of decay of the covariance function. The $\nu$ in the Matérn covariance is a smoothness parameter; $\alpha$ in the powered exponential model also controls smoothness, but is not as flexible. The Cauchy covariance has a long-range dependence parameter $\beta$. The spherical covariance is compactly supported, and finally the exponentially damped cosine takes on negative values (note it depends on the dimension $d$, this will be important shortly). The Matérn contains the exponential and squared exponential as special cases, specifically when $\nu = 0.5$ and $\nu \rightarrow \infty$, respectively. Sometimes the powered exponential with $\alpha = 2$ is called a Gaussian covariance, but powered exponential is preferred.

[R example (Covariances.R)]

If the process $Z(s)$ has isotropic covariance function $C(\cdot)$, then at a set of locations $s_1, \ldots, s_n$, the vector $Z = (Z(s_1), \ldots, Z(s_n))^\prime$ has covariance matrix

$$
\Sigma = \text{Cov}(Z, Z) = \begin{pmatrix}
C(0) & C(\|s_1 - s_2\|) & \cdots & C(\|s_1 - s_n\|) \\
C(\|s_2 - s_1\|) & C(0) & \cdots & C(\|s_2 - s_n\|) \\
\vdots & \vdots & \ddots & \vdots \\
C(\|s_n - s_1\|) & C(\|s_n - s_2\|) & \cdots & C(0)
\end{pmatrix}.
$$

The matrix $\Sigma$ is the covariance matrix, and is sometimes called the spatial covariance matrix.

[Point out how this would differ for a stationary and a general covariance]

What properties of this matrix do we know must hold?

Any covariance function must satisfy the following properties:

1. $C(s, s) \geq 0$ [variances must be nonnegative; how would you show this?]

2. $C(s_1, s_2) = C(s_2, s_1)$ [symmetry; how would you show this?]

3. $C$ is a nonnegative definite function in that, for any choices of locations $s_1, \ldots, s_n$ and arbitrary real numbers $a_1, \ldots, a_n \in \mathbb{R}$, we have

$$
\sum_{i=1}^{n} \sum_{j=1}^{n} a_i C(s_i, s_j) a_j \geq 0
$$

which is equivalent to

$$
\text{Var}(a^\prime Z) = a^\prime \text{Cov}(Z, Z)a \geq 0.
$$
[Note definition of positive definite matrix].

4. If \( C \) is isotropic, then \( C(0) \geq C(r) \) for any \( r \geq 0 \).

All covariance functions are positive definite, and all positive definite functions are covariance functions. Testing whether a given candidate function is positive definite is not trivial, but we will later find some useful tools for doing so.

One technical difficulty is that it is possible for covariance functions that are valid in \( d \) dimensions to be invalid in \( d + 1 \) dimensions. We will only focus on those that are valid in all dimensions. In particular we have the following theorem:

**Theorem 15** (Schoenberg (1938)). A function \( C : [0, \infty) \rightarrow \mathbb{R} \) is an isotropic covariance function valid in all dimensions if and only if

\[
C(r) = \int_0^\infty \exp(-t^2 r^2) dF(t)
\]

where \( F \) is a probability measure on \([0, \infty)\).

[Think of this as \( \int \exp(-t^2 r^2) f(t) dt \) for some \( f(t) \geq 0 \)]

Theorem 15 has three important implications: 1) all isotropic covariance functions valid in all dimensions are scale mixtures of a squared exponential kernel; 2) we now have at least one roadmap to building more covariance models [e.g., specify some \( f(t) > 0 \)]; and 3) no covariance model that is valid in all dimensions can take on negative values.

Some processes exhibit negative correlations, i.e., at certain spatial lags they are anti-correlated. The exponentially damped cosine model takes on negative values, and is only valid up to \( d \) dimensions.

### 3.5 Empirical Covariance Functions

Suppose we have data \( z(s_1), \ldots, z(s_n) \) from an isotropic process, we would like to develop an empirical estimator of a covariance, call it \( \hat{C}(r) \), (in much the same way that \( \bar{z} \) is an estimator of the mean). Note that we can calculate

\[
(z(s_i) - \bar{z})(z(s_j) - \bar{z})
\]
for all pairs $i, j = 1, \ldots, n$, resulting in $n^2$ points. In expectation, each of these is a really variable estimator whose mean is close to the true covariance:

$$\mathbb{E}(Z(s_i) - \bar{Z})(Z(s_j) - \bar{Z}) \approx C(\|s_i - s_j\|).$$

So, if we plotted

$$(z(s_i) - \bar{z})(z(s_j) - \bar{z}) \sim \|s_i - s_j\| \quad \text{for all } i, j = 1, \ldots, n$$

then, on average, the point cloud would look like $C(r)$ for values of $r \in [0, \max\{\|s_i - s_j\|\}]$.

It turns out that individual products are highly variable, and so we’d like to average these to produce a more stable estimator. Thus, one empirical covariance estimator is:

$$\hat{C}(r) = \frac{1}{|N(r)|} \sum_{(s_i, s_j) \in N(r)} (z(s_i) - \bar{z})(z(s_j) - \bar{z})$$

where $N(r)$ denotes a set of pairs of locations whose distance is approximately $r$ and $|N(r)|$ is the number of pairs in the set. This is a binned covariance estimator. The number of bins is chosen by the practitioner, but a typical rule of thumb is to include at least 30 pairs within any bin. Additionally, for large distances $r \gg 0$ so few data pairs are included that $\hat{C}(r)$ should be interpreted with care; usually we only examine the covariance at distances up to half the maximal interpoint distance.

[Picture]

Empirical covariance estimators for stationary but anisotropic processes follow the same idea but split $N(h)$ by direction (for example, cardinal directions). Nonstationary empirical covariances can be derived assuming the process is locally stationary, or whose covariance does not vary too much over space.

[R example (EmpiricalCovariances.R)]

### 3.6 Nugget Effect

[Example (Nugget.png)]

**Definition 16.** A stochastic process $\varepsilon(s)$ is a white noise process if
• $\mathbb{E}\varepsilon(s) = \mu$, i.e., $\varepsilon$ has a spatially constant mean

• $\text{Var}\varepsilon(s) = \text{Cov}(\varepsilon(s), \varepsilon(s)) = \tau^2$, and

• $\text{Cov}(\varepsilon(s_1), \varepsilon(s_2)) = 0$ for $s_1 \neq s_2$.

### 3.7 Continuity and Smoothness

Choice of the covariance function has important implications on continuity and smoothness of the resulting random field. First, it is not exactly clear what continuity and smoothness mean for a random function.

**Definition 17.** We say a stochastic process $Z(s)$ is mean square continuous if

$$\mathbb{E}(Z(s) - Z(s + h))^2 \to 0$$

as $\|h\| \to 0$.

For a stationary process,

$$\mathbb{E}(Z(s) - Z(s + h))^2 = \text{Var}(Z(s) - Z(s + h))$$

$$= \text{Cov}(Z(s) - Z(s + h), Z(s) - Z(s + h))$$

$$= \text{Cov}(Z(s), Z(s)) - 2\text{Cov}(Z(s), Z(s + h)) + \text{Cov}(Z(s + h), Z(s + h))$$

$$= 2(C(0) - C(h)).$$

This goes to zero as $\|h\| \to 0$ when $C(\cdot)$ is continuous at the origin. For example, a Gaussian process with exponential covariance function is mean square continuous, whereas the process plus a nugget effect is no longer mean square continuous.

There is another type of continuity called almost sure continuity, which happens at a point $s_0$ if $Z(s) \to Z(s_0)$ almost surely; $Z(s)$ is then almost surely continuous if it is almost surely continuous at every point. In general, one form of continuity does not necessarily imply the other.

If $Z(s)$ is an isotropic Gaussian process, the behavior of $C(r)$ at the origin determines its smoothness properties (i.e., differentiability). Specifically, if $C(r)$ is $2m$ times differentiable at the origin then the sample paths of the associated process are $m$ times differentiable.
Example 18. A Gaussian process with Matérn covariance with smoothness \( \nu \) has \( m \)-times differentiable sample paths if and only if \( m < \nu \).

Example 19. A Gaussian process with powered exponential covariance with power of \( \alpha < 2 \) has sample paths that are differentiable nowhere, whereas the case \( \alpha = 2 \) implies infinitely differentiable sample paths.

This distinction between the Matérn and powered exponential covariances has led more authors to prefer the Matérn due to its ability to smoothly transition between differing smoothnesses.

3.8 Existence: Kolmogorov’s Theorem

Until now we have assumed the existence of random fields, and have specified their behavior through finite dimensional properties. It turns out these finite dimensional descriptions determine the full stochastic process under one critical condition.

If we have a collection of random variables \( Z(s), s \in \mathbb{R}^d \) together with a collection of finite dimensional measures \( P_{s_1, \ldots, s_n} \) that is consistent in that

\[
P_{s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n} (Z(s_1) \in B_1, \ldots, Z(s_{i-1}) \in B_{i-1}, Z(s_{i+1}) \in B_{i+1}, \ldots, Z(s_n) \in B_n)
= P_{s_1, \ldots, s_n} (Z(s_1) \in B_1, \ldots, Z(s_{i-1}) \in B_{i-1}, Z(s_i) \in \mathbb{R}, Z(s_{i+1}) \in B_{i+1}, \ldots, Z(s_n) \in B_n)
\]

for every \( n \geq 1 \) and Borel sets \( B_1, \ldots, B_n \), then Kolmogorov’s Theorem states that there is necessarily an extension to a probability measure on an infinite dimensional space defining \( Z \).
4 Classical Geostatistical Methods

The basic setup in a typical spatial statistics problem is that a variable \( Y(s) \) has been observed over a region of interest \( D \subset \mathbb{R}^d \) at a set of locations \( s_1, \ldots, s_n \). Our goal is to make inferences about the process that governs the spatial distribution of this variable as well as estimating or predicting the variable at unobserved locations. The key geostatistical assumption is that \( \{Y(s_i)\}_{i=1}^n \) form a sample from one realization of a random field \( Y(s) \).

4.1 Geostatistical Model

As only one partially observed realization of \( Y(s) \) is available, some structure must be imposed in order to perform any reasonable inferences. The classical geostatistical model imposes structure by assuming

\[
Y(s) = \mu(s) + Z(s) + \varepsilon(s)
\]

where \( \mu(s) = \mathbb{E}Y(s) \) is the mean function, assumed to be deterministic and continuous, \( Z(s) \) is a stationary random process centered about zero and \( \varepsilon(s) \) is a mean zero white noise process. The \( \varepsilon(s) \) represents measurement error, noise or microscale variation depending on the dataset.

We will also write this as

\[
Y(s) = \mu(s) + W(s) = \text{fixed part} + \text{random part}
\]

where \( W(s) = Z(s) + \varepsilon(s) \) represents the random variation about the mean trend. In the geostatistical approach the only assumptions are about the first two moments of the process — we do not assume any structure on \( W(\cdot) \) (i.e., \( Z(\cdot) \) or \( \varepsilon(\cdot) \)) at higher moments.

[R example: (ModelDecomposition.R)]

Our goals are to:

- Estimate \( \mu \) and describe the stochastic structure of \( Z \) and \( \varepsilon \)
- Predict \( \mu(s_0) + Z(s_0) \) at a (new) location \( s_0 \). If \( s_0 \) is a ungauged location we call this prediction, if \( s_0 = s \) is a gauged location this is sometimes called filtering. [Difference between predicting the process at a new location and filtering the noise \( \varepsilon(s) \)]

Usually we specify \( \mu(s) = x(s)'\beta \) as a linear model on some observed covariates that depend on the location \( s \). Typical covariates are elevation, lat/lon coordinates, some power transformation of these, or other covariates that are relevant to a particular scientific application. Note that semiparametric or nonparametric methods for estimating \( \mu(s) \) may also be used, but tradition seems to favor linear specifications. [Part of the motivation for this is that disentangling spatial structure due to the mean vs. due to the spatially correlated variation \( Z(s) \) is not straightforward, and it is thus recommended to use mean functions that can be justified]

### 4.1.1 The Effect of the Nugget Effect

In geostatistics, \( \varepsilon(s) \) (and sometimes \( \text{Var} \varepsilon(s) = \tau^2 \)) is referred to as the nugget effect, and is often interpreted as representing observational error or small scale variability that cannot be directly estimated from the data due to restrictions in spatial sampling (i.e., if the minimal interpoint distance in the data is 5 km then we cannot say anything about how the process is correlated at 3 km). The name nugget comes from its mining history, where it was used to represent small scale variability in ore deposits beneath the ground.

What is the effect of including \( \varepsilon \) on the covariance function for \( Y \)?

\[
\text{Cov}(Y(s_1), Y(s_2)) = \text{Cov}(\mu(s_1) + Z(s_1) + \varepsilon(s_1), \mu(s_2) + Z(s_2) + \varepsilon(s_2)) \\
= \text{Cov}(Z(s_1), Z(s_2)) + \text{Cov}(\varepsilon(s_1), \varepsilon(s_2)) \\
= C(s_1, s_2) + \tau^2 \mathbb{1}_{[s_1 = s_2]}
\]

where \( \mathbb{1}_A \) is an indicator function that is 1 if the set condition \( A \) holds and is zero otherwise, that is,

\[
\mathbb{1}_{[s_1 = s_2]} = \begin{cases} 
0 & \text{if } s_1 \neq s_2 \\
1 & \text{if } s_1 = s_2.
\end{cases}
\]
Thus if everything is nice and isotropic we can write

\[
\text{Cov}(Y(s_1), Y(s_2)) = \begin{cases} 
C(0) + \tau^2 & \text{if } s_1 = s_2 \\
C(||s_1 - s_2||) & \text{if } s_1 = s_2.
\end{cases}
\]

We will see that the nugget effect has implications on spatial prediction.

4.2 Variograms

The classical tool for assessing spatial structure is the variogram. First, there is a relaxed form of stationarity that is sometimes useful.

**Definition 20.** A process \( Z(s) \) is intrinsically stationary if \( \text{Var}(Z(s+h) - Z(s)) \) is a function of only the lag vector \( h \in \mathbb{R}^d \).

This class is mostly concerned with weakly stationary processes – all weakly stationary processes are intrinsically stationary, but not necessarily vice versa.

**Definition 21.** The semivariogram for a weakly stationary or intrinsically stationary process is

\[
\gamma(h) = \frac{1}{2} \text{Var}(Z(s + h) - Z(s))
\]

for any \( h \in \mathbb{R}^d \). Or, if everything is nice and isotropic then we have

\[
\gamma(r) = \gamma(||s_1 - s_2||) = \frac{1}{2} \text{Var}(Z(s_1) - Z(s_2))
\]

where \( r \) is the distance between \( s_1 \) and \( s_2 \).

The function \( 2\gamma(h) \) is called the variogram, although often semivariogram and variogram are used interchangeably in the literature.
If $Z(s)$ is stationary with covariance function $C(h)$, then

\[
\gamma(h) = \frac{1}{2} \text{Var} (Z(s + h) - Z(s)) \\
= \frac{1}{2} \text{Cov} (Z(s + h) - Z(s), Z(s + h) - Z(s)) \\
= \frac{1}{2} (\text{Cov}(Z(s + h), Z(s + h)) - 2\text{Cov}(Z(s + h), Z(s)) + \text{Cov}(Z(s), Z(s))) \\
= \frac{1}{2} (C(0) - 2C(h) + C(0)) \\
= C(0) - C(h).
\]

Note that the semivariogram is in the same units as the process variance.

What would we expect typical variograms to look like?

Any semivariogram must satisfy the following properties (which are necessary and sufficient conditions):

1. Vanishes at zero, $\gamma(0) = 0$

2. Is an even function, $\gamma(-h) = \gamma(h)$ for all $h \in \mathbb{R}^d$

3. $\gamma$ is conditionally nonnegative definite in that, for any choices of locations $s_1, \ldots, s_n$ and real numbers $a_1, \ldots, a_n \in \mathbb{R}$ such that $\sum_{i=1}^{n} a_i = 0$, we have

\[
\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \gamma(s_i - s_j) \leq 0.
\]

To see this last requirement, if $Z(s)$ is mean zero and has semivariogram $\gamma(\cdot)$, take any $a_1, \ldots, a_n$ such that $\sum_{i=1}^{n} a_i = 0$ and any $s_1, \ldots, s_n$. Then,

\[
-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j (Z(s_i) - Z(s_j))^2 = \left( \sum_{i=1}^{n} a_i Z(s_i) \right)^2.
\]

Take an expectation on both sides,

\[
-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \gamma(s_i - s_j) = \text{Var} \left( \sum_{i=1}^{n} Z(s_i) \right) \geq 0.
\]

We will discuss empirical estimation of the semivariogram in the next section, but typically parametric variograms are preferred due to the inherent variability of empirical estimators,
that empirical estimators often fail one of the above requirements, and that empirical estimators are only available at certain lags. The following theorem gives a nice relationship between valid variogram models and positive definite functions.

**Theorem 22.** If $\gamma(\cdot)$ is a continuous function on $\mathbb{R}^d$ satisfying $\gamma(0) = 0$, then the following are equivalent:

1. $\gamma(\cdot)$ is conditionally negative definite
2. For all $a > 0$, $\exp(-a\gamma(\cdot))$ is positive definite.

The basic variogram models are isotropic, in that $\gamma(h) = \gamma(||h||) = \gamma(r)$ depends only on the distance between locations. Some of the most common isotropic parametric models that are valid in $d = 2, 3$ dimensions are:

- **Spherical:**
  \[
  \gamma(r) = \begin{cases} 
  \sigma^2 \left( \frac{3}{2} r - \frac{1}{2} \left( \frac{r}{a} \right)^3 \right) & \text{for } 0 \leq r \leq a \\
  \sigma^2 & \text{for } r > a 
  \end{cases}
  \]
- **Exponential:**
  \[
  \gamma(r) = \sigma^2 \left( 1 - \exp \left( -\frac{r}{a} \right) \right), \quad a > 0
  \]
- **Squared exponential [some call it Gaussian, but you won’t]:**
  \[
  \gamma(r) = \sigma^2 \left( 1 - \exp \left( -\frac{r^2}{a^2} \right) \right), \quad a > 0
  \]
- **Matérn:**
  \[
  \gamma(r) = \sigma^2 \left( 1 - \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{r}{a} \right)^\nu K_\nu \left( \frac{r}{a} \right) \right), \quad a > 0, \nu > 0,
  \]
  where $K_\nu$ is a modified Bessel function of the second kind of order $\nu$.
- **Power:**
  \[
  \gamma(r) = \sigma^2 r^\alpha, \quad 0 \leq \alpha < 2
  \]
Analogously to covariances, the Matérn includes the exponential and squared exponential case when the smoothness is set as \( \nu = 0.5 \) and \( \nu \to \infty \), respectively. Also, for all models, \( a \) is called the range parameter, and \( \sigma^2 \) is often termed the marginal variance (or just variance). The power model is the semivariogram for a fractional isotropic Brownian motion in \( \mathbb{R}^d \).

If \( \varepsilon(s) \) is a white noise process with variance \( \tau^2 \) then

\[
\gamma(r) = \begin{cases} 
0 & \text{if } r = 0 \\
\tau^2 & \text{if } r > 0.
\end{cases}
\]

[Picture of the difference between semivariogram for correlated/noise processes]

The above variogram models are designed for the continuous variation \( Z(s) \) process. How does including a nugget effect affect the variogram for the observed process \( Y(s) = \mu(s) + Z(s) + \varepsilon(s) \), where \( \text{Var}(\varepsilon(s)) = \tau^2? \) Call \( \gamma_Z(r) \) the semivariogram for \( Z(s) \); then for \( r = \|h\| > 0 \),

\[
2\gamma_Y(r) = \text{Var}(Y(s+h) - Y(s)) = \text{Var}(\mu(s+h) + Z(s+h) + \varepsilon(s+h) - \mu(s) - Z(s) - \varepsilon(s)) = \text{Var}(Z(s+h) - Z(s)) + \text{Var}(\varepsilon(s+h) - \varepsilon(s)) = 2\gamma_Z(r) + 2\tau^2,
\]

so in particular,

\[
\gamma_Y(r) = \tau^2 + \gamma_Z(r).
\]

Note that \( \gamma_Y(0) = \text{Var}(Y(s+0) - Y(s)) = \text{Var}(0) = 0 \). For example, if \( Z(s) \) has an exponential semivariogram (or covariance function), the observation process has semivariogram

\[
\gamma(r) = \begin{cases} 
0 & \text{if } r = 0 \\
\tau^2 + \sigma^2 \left(1 - \exp\left(-\frac{r}{a}\right)\right) & \text{if } r > 0.
\end{cases}
\]

Some attributes of the semivariogram are often referred to by special names.

**Definition 23.** The **sill** of a semivariogram \( \gamma(r) \) is \( \lim_{r \to \infty} \gamma(r) \) if this limit is finite.

If the sill is finite, it is \( \sigma^2 \) if there is no nugget effect, or is \( \tau^2 + \sigma^2 \) with a nugget effect. If the sill exists, the process is weakly stationary, whence \( C(0) = \text{sill} \).
Definition 24. The effective range of a semivariogram is the smallest distance $r$ at which $\gamma(r)$ is 95% of its sill.

[Picture of a variogram with and without nugget effect, indicate sill and effective range]

Finally, although sometimes used by some authors, the squared exponential model should be used with caution – it implies process realizations that are infinitely differentiable, which is a very strong assumption for natural processes. Additionally, the spherical model is popular among domain scientists, [i.e., not statisticians] as the correlation decays to exactly zero after a certain range, but its process realizations are extremely rough and indeed never differentiable. Thus, the Matérn is a favored flexible model between these two extremes.

4.3 Estimation

Given a set of data $Y(s_1), \ldots, Y(s_n)$, how do we empirically estimate the semivariogram? The first step is to remove the mean trend $\mu(s)$, say by specifying a linear regression that is estimated by OLS. Form the estimated residuals $\hat{W}(s_i) = Y(s_i) - \hat{\mu}(s_i)$. We will suppose the process is known to be isotropic.

Note that distances $r$ are only available at the available observation subsets $r_{ij} = \|s_i - s_j\|$. We could define the empirical semivariogram as

$$\hat{\gamma}(r_{ij}) = \frac{1}{2}(\hat{W}(s_i) - \hat{W}(s_j))^2$$

but this results in a point cloud that is extremely noisy and difficult to read. Thus, we bin pairwise locations that are of similar distance and average to produce the empirical semivariogram, defined as

$$\hat{\gamma}(r) = \frac{1}{2|N(r)|} \sum_{(s_i, s_j) \in N(r)} (\hat{W}(s_i) - \hat{W}(s_j))^2$$

where $N(r)$ is a neighborhood of pairs of points whose distance is about $r$ and $|N(r)|$ is the total number of pairs within this neighborhood. The number of bins is chosen by the practitioner, but a typical rule of thumb is to include at least 30 pairs within any bin. Additionally, for large distances $r \gg 0$ so few data pairs are included that $\hat{\gamma}(r)$ should be
interpreted with care; usually we only examine the semivariogram at distances up to half the maximal distance.

The empirical semivariogram is not very robust against outlying data points; Cressie and Hawkins (1980) proposed a robust empirical semivariogram of the form

$$
\hat{\gamma}_{\text{robust}}(r) = \left( \frac{1}{|N(r)|} \sum_{(s_i, s_j) \in N(r)} \left| \hat{W}(s_i) - \hat{W}(s_j) \right|^{1/2} \right)^4 \left( 0.914 + \frac{0.988}{|N(r)|} \right). 
$$

Two main procedures for estimating the parameters of a chosen parametric covariance or variogram model are weighted least squares (WLS) or maximum likelihood (ML). We review ML later.

The WLS estimator of a parametric variogram $\gamma(r) = \gamma(r|\theta)$ where $\theta = (a, \sigma^2, \tau^2)$ usually (sometimes including $\nu$ if a Matérn is used, e.g.) is

$$
\hat{\theta} = \arg\min_{\theta} \frac{1}{K} \sum_{k=1}^{K} \frac{|N(r_k)|}{\gamma(r_k|\theta)} \frac{1}{2} \gamma(r_k|\theta) - \gamma(r_k|\hat{\theta})^2
$$

where $k = 1, \ldots, K$ bins have been chosen. Note that if a bin has few observations, or if the semivariogram is large then less weight is put on matching the empirical semivariogram at that lag.

There is not a strong statistical argument for estimating variogram parameters in this way, and indeed there seems to be an overemphasis in the literature on matching theoretical variograms to empirical versions.

[Although there are theoretical results stating that, in certain circumstances, WLS estimators are consistent and asymptotically normal, they are suboptimal and do not rest on a formal probabilistic model for the spatial process]

An outline of a geostatistical analysis given observations $Y(s_1), \ldots, Y(s_n)$ is:

1. Choose a parametric mean function $\mu(s)$
2. Estimate parameters of $\mu(s)$ by ordinary least squares
3. Form the estimated residuals $\hat{W}(s_i) = Y(s_i) - \hat{\mu}(s_i)$
4. Choose and fit a parametric variogram model to the empirical variogram.

5. Re-estimate the mean function by generalized least squares.

6. Krige the observations to a new location $s_0$. [Details next!]

[R example: (GeostatisticalAnalysis.R)]
5 Simulation

We have already seen that random field simulations can play a useful role in understanding properties of a given spatial model. Later we will additionally see that simulations can also play an important role statistically. There are many simulation techniques, we review some of the most important here, and focus exclusively on Gaussian process simulation.

There are two ways to think about building (and thus simulating) processes, the first is to explicitly decompose a process into weighted sums of random variables, and the second is to specify a process via its mean and covariance function. For instance, to simulate

\[ Z(t) = U \cos(2\pi t) + V \sin(2\pi t) \]

where \( U \) and \( V \) are iid \( N(0,1) \) random variables, we only need to simulate two normal variables and then weight them by the cos and sin harmonics. On the other hand, if we suppose \( Z(t) \) is a time series with exponential covariance function, how would we go about simulating it? We focus on the latter problem.

5.1 Direct: Cholesky Decomposition

A generic approach to simulating a Gaussian stochastic process is using the Cholesky decomposition. This does not require stationarity or isotropy of the covariance, and works for arbitrary multivariate normal vectors.

**Definition 25.** If \( \Sigma \) is a positive definite matrix, its Cholesky decomposition is

\[ \Sigma = LL^T \]

where \( L \) is lower triangular. Note \( L \) is sometimes called the Cholesky factor.

If \( Z(s) \) is a mean zero Gaussian process with covariance function \( C(\cdot, \cdot) \), to simulate \( Z \) at locations \( s_1, \ldots, s_n \), take

\[ L\varepsilon = L(\varepsilon_1, \ldots, \varepsilon_n)^T \]
where $L$ is the Cholesky factor of the covariance matrix $\Sigma = (C(s_i, s_j))_{i,j=1}^n$ and $\varepsilon_1, \ldots, \varepsilon_n$ are iid $N(0,1)$ random variables (thus, $\varepsilon \sim N(0, I)$). To see that this produces an exact simulation, note that $L\varepsilon$ is a multivariate normal vector with mean and covariance matrix
\[
E(L\varepsilon) = LE\varepsilon = 0
\]
\[
\text{Cov}(L\varepsilon, L\varepsilon) = LCov(\varepsilon, \varepsilon)L^T = LIL^T = LL^T = \Sigma.
\]
While this produces exact simulations of a Gaussian process, calculating the Cholesky factor can be computationally expensive, where the whole simulation requires $6m^5$ flops for simulation on a $m \times m$ lattice.

5.2 Random Coins

The random coin, or dilution, method, requires that the stationary covariance function can be represented as a convolution,
\[
C(h) = \int g(s)g(s + h)ds.
\]
The approximate simulation is $\sum_{h \in \Pi} g(s - h)$ where $\Pi$ is a stationary Poisson point process with unit intensity; in practice one sums multiple simulations and then relies on the central limit theorem to generate approximate stationary Gaussian process realizations. [Random coins typically refer to when $g$ is random]

5.3 Turning Bands

The turning bands method relies on simulating a multidimensional ($d > 1$) isotropic random field by summing simulations from one-dimensional processes that have been embedded in more than one dimension. In particular,
\[
Z(s) = W(s \cdot e)
\]
where $W(t)$ is a one-dimensional (Gaussian) process, $e$ is a randomly oriented unit vector and $\cdot$ indicates the dot product. Suppose $W(t)$ is mean zero and has covariance $\text{Cov}(W(t), W(t + h)) = C(h)$. Then we have
\[
EZ(s) = EW(s \cdot e) = E(E(W(s \cdot e)|e)) = 0
\]
\[ \text{Cov}(Z(s), Z(s + h)) = \mathbb{E}(\mathbb{E}(W(s \cdot e)W((s + h) \cdot e)|e)) \]
\[ = \mathbb{E}(C(h \cdot e)) \]
\[ = \int_{\partial b_d} C(h \cdot r)U(dr) \]

where \( U \) is the uniform probability measure on the boundary of the unit sphere \( \partial b_d \).

This original formulation results in random fields that are constant along \( s \cdot e = c \), so in practice we rely on the central limit theorem and take simulations by

\[ Z(s) = \frac{1}{\sqrt{L}} \sum_{i=1}^{L} W_i(s \cdot e_i) \]

where \( \{W_i(\cdot)\}_{i=1}^{L} \) are mutually independent one-dimensional Gaussian processes, \( \{e_i\}_{i=1}^{L} \) are randomly oriented unit vectors and \( \cdot \) indicates the dot product. To specify the one-dimensional covariance function that corresponds to a desired multidimensional covariance structure, one must solve an Abel integral equation (in \( d = 2 \)).

[R example (SimulationMethods.R)]

### 5.4 Circulant Embedding

The most popular modern method for stationary random field simulation is circulant embedding, in part due to the fact that it produces exact Gaussian simulations (i.e., does not rely on the central limit theorem) and is fast for high dimensional gridded simulation.

**Definition 26.** A \( n \times n \) matrix \( M \) is circulant if

\[ M = \begin{pmatrix}
  c_0 & c_1 & c_2 & \cdots & c_{n-1} \\
  c_{n-1} & c_0 & c_1 & \cdots & c_{n-2} \\
  c_{n-2} & c_{n-1} & c_0 & \cdots & c_{n-3} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  c_1 & c_2 & c_3 & \cdots & c_0
\end{pmatrix} \]

for \( c_0, \ldots, c_{n-1} \in \mathbb{R} \).
Define \( r_j = \exp(2\pi ij/n) \) for \( j = 0, \ldots, n - 1 \) where \( i \) is the imaginary number, and define the matrix \( U \) as

\[
U = [u_0 \ u_1 \ \cdots \ u_{n-1}]
\]

where \( u_j = n^{-1/2}(1, r_j, r_j^2, \ldots, r_j^{n-1})' \). Note that \( U \) is a unitary matrix in that \( UU^* = U^*U = I \), where * denotes the complex conjugate transpose. Also note that \( U^*a \) is the discrete Fourier transform of a vector \( a \), and \( Ua \) is the inverse discrete Fourier transform.

**Lemma 27.** If \( M \) is a circulant matrix, it admits an eigendecomposition

\[
M = UDU^*
\]

such that \( D \) contains eigenvalues that are the discrete Fourier transform (DFT) of \( M \)'s first row vector. In particular, we have that the eigenvalues \( \text{diag}(D) = (d_0, \ldots, d_{n-1}) \) are

\[
d_j = \sum_{k=0}^{n-1} c_k r_j^{-k}; \quad j = 0, \ldots, n - 1.
\]

The key to circulant embedding’s success is that the DFT can be quickly calculated using the fast Fourier transform (FFT). Suppose we want to simulate a one-dimensional process \( Z(t) \) at time points \( t = 1, 2, \ldots, m \). If \( Z(t) \) has stationary covariance function \( C(\cdot) \), i.e., \( \text{Cov}(Z(i), Z(j)) = C(i - j) \), then we seek to embed the covariance matrix in a circulant matrix. Call \( \Sigma \) the covariance matrix of \( (Z(1), \ldots, Z(m))' \). Then form the matrix

\[
M = \begin{pmatrix} \Sigma & M_1 \\ M_2 & M_3 \end{pmatrix}
\]

where \( M_1, M_2 \) and \( M_3 \) are specified so that \( M \) is circulant. Specifically, \( M \) has dimension \((2m - 2) \times (2m - 2)\) and the first row of \( M \) is

\[
[C(0) \ C(1) \ C(2) \ \cdots \ C(m - 2) \ C(m - 1) \ C(m - 2) \ \cdots \ C(2) \ C(1)].
\]

Simulations are then formed by:

1. Calculate the eigenvalues of \( M \) using the FFT
2. Simulate a set of \( 2m - 2 \) iid normal variables \( w \)
3. Compute the FFT of \( w \), \( \mathcal{F}(w) \) [this is a vector]

4. Form \( D^{1/2}\mathcal{F}(w) \)

5. Compute the inverse FFT \( \mathcal{F}^{-1}(D^{1/2}\mathcal{F}(w)) \)

6. The first \( m \) entries of the resulting vector are an exact mean zero Gaussian simulation whose covariance matrix is \( \Sigma \).

Succinctly, we can write this procedure as forming \( W = UD^{1/2}U^*w \). This new vector has the correct covariance structure in its first \( n \) components:

\[
\begin{align*}
\text{Var}(W) &= \text{Cov}(W, W) \\
&= \text{Cov}(UD^{1/2}U^*w, UD^{1/2}U^*w) \\
&= UD^{1/2}U^*\text{Cov}(w, w)UD^{1/2}U^* \\
&= UD^{1/2}U^*UD^{1/2}U^* \\
&= UDU^* \\
&= M.
\end{align*}
\]

[Note \( M \) contains \( \Sigma \) in the first \( m \times m \) entries, the covariance matrix we’re really interested in]

For two-dimensional simulation, if simulation points are on a regular grid then the corresponding covariance matrix is block Toeplitz, which can also be embedded in a circulant matrix, although details become more complicated. The key to circulant embedding is that simulation based on it is much faster than via the Cholesky factor – \( 40n^2\log_2 2n \) flops. There are many other approaches to simulation, for example the spectral method relies on the spectral representation for a stationary process (stay tuned), or via Gibbs sampling.

[R example (CirculantEmbedding.R)]
6 Kriging

A major goal in a spatial statistical analysis is smoothing or prediction to estimate the underlying process at locations where observations are not necessarily available. The geostatistical method of doing this is called kriging, named after Danie G. Krige, a South African mining engineer who pioneered the original ordinary kriging method.

Throughout this section we will focus on kriging under various models. In particular, let $\mu$ be a constant mean, $Z(s)$ a spatially correlated mean zero process, $\varepsilon(s)$ a mean zero white noise process and let $Y(s)$ represent the observations. We will entertain models summarized in the following table.

Table 1: Possible models and corresponding kriging predictor names

<table>
<thead>
<tr>
<th>Model number</th>
<th>Model</th>
<th>Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$Y(s) = Z(s)$</td>
<td>$\mu$ known</td>
</tr>
<tr>
<td>1</td>
<td>$Y(s) = \mu + Z(s)$</td>
<td>$\mu$ unknown</td>
</tr>
<tr>
<td>2</td>
<td>$Y(s) = \mu + Z(s)$</td>
<td>$\mu$ unknown</td>
</tr>
<tr>
<td>3</td>
<td>$Y(s) = \beta_0 + \beta_1 X_1(s) + \cdots + \beta_p X_p(s) + Z(s)$</td>
<td>${\beta_i}$ unknown</td>
</tr>
<tr>
<td>4</td>
<td>$Y(s) = Z(s) + \varepsilon(s)$</td>
<td>${\beta_i}$ unknown</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model number</th>
<th>What to predict</th>
<th>Kriging type</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$Z(s)$</td>
<td>Simple</td>
</tr>
<tr>
<td>1</td>
<td>$\mu + Z(s)$</td>
<td>Simple</td>
</tr>
<tr>
<td>2</td>
<td>$\mu + Z(s)$</td>
<td>Ordinary</td>
</tr>
<tr>
<td>3</td>
<td>$\beta_0 + \beta_1 X_1(s) + \cdots + \beta_p X_p(s) + Z(s)$</td>
<td>Universal</td>
</tr>
<tr>
<td>4</td>
<td>$Z(s)$ (or $Y(s)$?)</td>
<td>Simple</td>
</tr>
</tbody>
</table>

In model 3, the $X_i(s)$ are known predictors that are functions of spatial locations, e.g., longitude, latitude and elevation are typical regressors (although they may involve other predictors, such as solutions from a weather or pollution prediction model). The corresponding $\beta_i$s are unknown linear regression parameters that must be estimated from the data.

Models 0-3 can be written

$$Y(s) = f(s)$$

and model 4 can be written

$$Y(s) = f(s) + \varepsilon(s).$$
In each case, the goal is to estimate the function \( f(s) \) based on a weighted linear combination of observations. This is known as kriging, and kriging breaks up into three types: simple, ordinary and universal. Simple kriging is used when the mean is known; ordinary assumes a spatially constant, but unknown, mean, while universal kriging supposes the mean function is essentially a regression model.

Until we reach model 4, we will assume the observations are perfect, and so prediction at a location \( s_0 \) involves predicting \( Y(s_0) = f(s_0) \). Thus, we will represent the kriging predictor by \( \hat{f}(s_0) \). In model 4, we will either predict \( f \) or \( Y \), depending on the goal.

### 6.1 Best Linear Unbiased Prediction

Kriging is the best linear unbiased predictor (BLUP; sometimes BLUE for estimator) for predicting at a location \( s_0 \). That is, the kriging predictor minimizes

\[
\mathbb{E}(Y(s_0) - \hat{f}(s_0))^2
\]

over (a) the class of all linear predictors

\[
\hat{f}(s_0) = \sum_{i=1}^{n} w_i Y(s_i)
\]

that are (b) unbiased,

\[
\mathbb{E}\hat{f}(s_0) = \mathbb{E}Y(s_0) = \mathbb{E}f(s_0).
\]

We call \( w_1, \ldots, w_n \) the kriging weights. Note that the unbiasedness condition implies we are really minimizing the predictive error variance

\[
\text{Var}(Y(s_0) - \hat{f}(s_0)).
\]

There are many methods to choose the weights \( w_1, \ldots, w_n \), for example inverse distance weighting, but these are ad hoc methods that are not based on formal optimization criteria.

It is convenient to introduce matrix notation for the kriging section. The goal will be to predict the process at a location \( s_0 \) (not necessarily an observation site). Define the vectors

\[
Z = (Z(s_1), \ldots, Z(s_n))^T, \quad Y = (Y(s_1), \ldots, Y(s_n))^T, \quad \varepsilon = (\varepsilon(s_1), \ldots, \varepsilon(s_n))^T
\]
and the covariance vector

\[ \Sigma_0 = \text{Cov}(Z(s_0), Y) = (C(s_0, s_1), C(s_0, s_1), \ldots, C(s_0, s_n)), \]

that is, the row vector of covariances between \( Z \) at the new location \( s_0 \) and \( Y \) all of the observation locations. Similarly, define

\[ \Sigma = \text{Cov}(Z, Z) = (C(s_i, s_j))_{i,j=1}^n \]

as the \( n \times n \) matrix of covariances between \( Z \) at all pairwise data locations.

### 6.2 Simple Kriging

In the simple kriging case, we suppose the mean function is spatially constant and known, that is, \( Y(s) = \mu + Z(s) \), where \( \mu \) is known. Then we can work with the centered observations \( Y(s_i) - \mu \), which are direct observations of the residuals \( Z(s) = Y(s) - \mu \). Thus, predicting \( Y(s_0) \) is equivalent to predicting \( Z(s_0) \) and adding \( \mu \).

If the mean \( \mu \) is zero, then we claim the simple kriging predictor is

\[ \hat{f}(s_0) = \Sigma_0 \Sigma^{-1} Y \]

or with the known mean \( \mu \),

\[ \hat{f}(s_0) = \Sigma_0 \Sigma^{-1} (Y - \mu) + \mu \]

where \( \mu = (\mu, \ldots, \mu)' \).

There are multiple ways to derive the kriging predictor \( \hat{f}(s_0) \), the most common of which is to use Lagrange multipliers to enforce the unbiasedness condition. We can motivate the solution in another way, however.

Let \( w = (w_1, \ldots, w_n)^T \) be the optimal kriging weights and consider any other set of weights \( v = w + u \). Then we have

\[
\text{Var}(v^T Z - Z(s_0)) = \text{Var}(u^T Z + (w^T Z - Z(s_0))) \\
= \text{Var}(u^T Z) + \text{Var}(w^T Z - Z(s_0)) + 2\text{Cov}(u^T Z, w^T Z - Z(s_0)) \\
\geq \text{Var}(w^T Z - Z(s_0)) + 2\text{Cov}(u^T Z, w^T Z - Z(s_0)),
\]

49
Now the second term is
\[
\text{Cov}(u^T Z, w^T Z - Z(s_0)) = \text{Cov}(u^T Z, w^T Z) - \text{Cov}(Z(s_0), u^T Z)
\]
\[
= u^T \Sigma w - \Sigma_0 u
\]
\[
= w^T \Sigma u - \Sigma_0 u
\]
which is zero for any choice of \( u \) when \( w^T = \Sigma_0 \Sigma^{-1} \). We have shown that this choice of kriging weights minimizes the predictive variance (and thus mean squared error).

Note that \( w_i \) depends on \( \Sigma^{-1} \), and thus depends on the entire spatial distribution of sample locations \( s_1, \ldots, s_n \), not just the distances \( ||s_0 - s_i|| \). If we removed one location, say \( s_n \), all kriging weights would need to be recalculated.

**Example 28.** Suppose \( Z(t) \) is a time series \( t \in \mathbb{R} \) with covariance function
\[
C(h) = \begin{cases} 
1 - |h|, & |h| \leq 1 \\
0, & |h| > 1.
\end{cases}
\]
Note this covariance is valid in \( d = 1 \), but not \( d \geq 2 \). Suppose \( Z \) is observed at \( t = \pm 1/2 \) and we seek to predict \( Z \) at \( t_0 = 0 \). Then
\[
\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \Sigma_0 = (1/2 \ 1/2)
\]
The simple kriging predictor is
\[
\hat{f}(0) = \Sigma_0 \Sigma^{-1} Z = Z(-1/2)/2 + Z(1/2)/2,
\]
just an evenly weighted average between the two observations where \( w_1 = w_2 = 1/2 \).

Now suppose a third observation is added at \( 1 < t_3 \leq 3/2 \); we have the covariance of \((Z(-1/2), Z(1/2), Z(t_3))^T \) being
\[
\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 3/2 - t_3 \\ 0 & 3/2 - t_3 & 1 \end{pmatrix} \quad \text{and} \quad \Sigma_0 = (1/2 \ 1/2 \ 0).
\]
Then the simple kriging predictor is
\[
\hat{f}(0) = \frac{1}{2} Z(-1/2) + \frac{1}{2(1 - (3/2 - t_3)^2)} Z(1/2) + \frac{t_3 - 3/2}{2(1 - (3/2 - t_3)^2)} Z(t_3).
\]
In particular, note that \( \hat{Z}(0) \) depends on \( Z(t_3) \), even though \( \text{Cov}(Z(0), Z(t_3)) = 0 \), and the weight on \( Z(1/2) \) depends on \( t_3 \). If \( t_3 \geq 3/2 \) this reduces to the same estimator as if \( Z(t_3) \) was unobserved.

[R example: (SimpleKriging.R)]

### 6.3 Ordinary Kriging

Ordinary kriging refers to the case where the mean \( \mu(s) = \mu \) is spatially constant, but unknown. The ordinary kriging predictor is

\[
\hat{f}(s_0) = \mathbf{w}^T \mathbf{Y} = \left( \Sigma_0 + \frac{1 - \Sigma_0 \Sigma^{-1}_0 \mathbf{1}\mathbf{1}^T}{\mathbf{1}^T \Sigma^{-1}_0 \mathbf{1}} \right) \Sigma^{-1} \mathbf{Y}
\]

where \( \mathbf{1} \) is a vector of 1’s of length \( n \). Recall the two key requirements for the kriging predictor: (a) linear in \( \mathbf{Y} \) and (b) unbiasedness. For ordinary kriging, these conditions imply

\[
\sum_{i=1}^{n} w_i = \mathbf{w}^T \mathbf{1} = 1,
\]

since

\[
\mu = \mathbb{E}Y(s_0) = \mathbb{E} \sum_{i=1}^{n} w_i Y(s_i) = \sum_{i=1}^{n} w_i \mathbb{E}Y(s_i) = \mu \sum_{i=1}^{n} w_i.
\]

To derive the ordinary kriging predictor, we use a Lagrange multiplier to enforce the unbiasedness condition:

\[
\mathbb{E}(\hat{f}(s_0) - Y(s_0))^2 + 2\lambda(1 - \mathbf{w}^T \mathbf{1}) = \text{Var}(\hat{f}(s_0) - Y(s_0)) + 2\lambda(1 - \mathbf{w}^T \mathbf{1}) = \text{Cov}(\mathbf{w}^T \mathbf{Y} - Y(s_0), \mathbf{w}^T \mathbf{Y} - Y(s_0)) + 2\lambda(1 - \mathbf{w}^T \mathbf{1}) = \mathbf{w}^T \Sigma \mathbf{w} - \Sigma_0 \mathbf{w} + C(s_0, s_0) + 2\lambda(1 - \mathbf{w}^T \mathbf{1}).
\]

Taking the derivative with respect to \( \mathbf{w} \) (using the useful linear algebra facts: \( (d/d\mathbf{w})\mathbf{w}^T \mathbf{u} = \mathbf{u} \) and \( (d/d\mathbf{w})\mathbf{w}^T \mathbf{A} \mathbf{w} = 2\mathbf{A} \mathbf{w} \)) yields

\[
\Sigma \mathbf{w} - \Sigma_0^T - \lambda \mathbf{1} = 0 \tag{2}
\]

and taking the derivative with respect to \( \lambda \) yields

\[
\mathbf{1}^T \mathbf{w} = 1. \tag{3}
\]
Equation (2) yields
\[ w = \Sigma^{-1}(\Sigma^T_0 + \lambda \mathbf{1}), \]
which plugging into (3) yields
\[ 1^T \Sigma^{-1}(\Sigma^T_0 + \lambda \mathbf{1}) = 1, \]
and rearranging yields
\[ \lambda = \frac{1 - \Sigma_0 \Sigma^{-1} \mathbf{1}}{1^T \Sigma^{-1} \mathbf{1}}. \]
Plugging back in yields the ordinary kriging weights.

Notice the connection between ordinary and simple kriging,
\[ \hat{f}(s_0) = \Sigma_0 \Sigma^{-1} \mathbf{Y} + \frac{1 - \Sigma_0 \Sigma^{-1} \mathbf{1}}{1^T \Sigma^{-1} \mathbf{1}} 1^T \Sigma^{-1} \mathbf{Y} \]
\[ = \text{simple kriging predictor + correction}. \]

### 6.4 Universal Kriging

Universal kriging is when the mean function \( \mu(s) = \mathbf{x}^T \beta \) is a general linear function, usually including covariates that are polynomial terms of coordinates. [e.g., if \( s = (x, y)^T \) then \( \mathbf{x} = (1, x, y, xy, x^2, y^2)^T \)]. The universal kriging predictor is just a slightly generalized form of the ordinary kriging predictor.

For observations \( \mathbf{Y} \) if the mean is modeled as \( \mathbb{E} \mathbf{Y} = \mathbf{X} \beta \) and the goal is to predict the process at a new location \( s_0 \) with corresponding covariates \( \mathbf{x}_0 \), then the universal kriging predictor is
\[ \hat{f}(s_0) = \mathbf{w}^T \mathbf{Y} = (\Sigma_0 + (\mathbf{x}_0^T - \Sigma_0 \Sigma^{-1} \mathbf{X})(\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^T) \Sigma^{-1} \mathbf{Y}. \]
Note that this reduces to ordinary kriging when the mean function is only a constant, \( \mathbf{x}_0 = 1, \mathbf{X}^T = (1, \ldots, 1)^T \). The derivation for the universal kriging predictor is analogous to the the derivation for ordinary kriging and is left as homework.

### 6.5 Kriging Uncertainty

A key benefit of kriging (vs. other smoothing/interpolation methods) is that predictive uncertainty can be readily quantified. Standard errors for kriging are based on the predictive variance \( \text{Var}(\hat{f}(s_0) - Y(s_0)) \) (which, note, is the quantity kriging minimizes).
Generically, for any linear predictor we have

\[ \sigma_K(s_0)^2 = \text{Var}(\hat{f}(s_0) - Y(s_0)) = \text{Cov}(w^T Y - Y(s_0), w^T Y - Y(s_0)) \]

\[ = \text{Cov}(w^T Y, w^T Y) - \text{Cov}(w^T Y, Y(s_0)) - \text{Cov}(Y(s_0), w^T Y) \]

\[ + \text{Cov}(Y(s_0), Y(s_0)) \]

\[ = w^T \Sigma w - 2\Sigma_0 w + C(s_0, s_0) \]

where \( \sigma_K(s_0)^2 \) stands for the kriging predictive variance. Plugging in our known weights for the various types of kriging yields the following predictive variances:

- Simple kriging:
  \[ \sigma_K(s_0)^2 = C(s_0, s_0) - \Sigma_0 \Sigma^{-1} \Sigma_0^T \]

- Universal kriging:
  \[ \sigma_K(s_0)^2 = C(s_0, s_0) - \Sigma_0 \Sigma^{-1} \Sigma_0^T + (x_0 - X \Sigma^{-1} \Sigma_0)^T (X \Sigma^{-1} X)^{-1} (x_0 - X \Sigma^{-1} \Sigma_0). \]

Thus, a full kriging solution will produce prediction intervals

\[ \hat{Z}(s_0) \pm 1.96 \sigma_K(s_0). \]

Calculation of these can be laborious, especially for large systems of observations and prediction locations, but simulation allows for one way out of direct computation.

[R example: (KrigingUncertainty.R)]

### 6.6 Screening Effect

It can happen that nearby data locations reduce influence of further locations when kriging. This is known as the screening effect.

**Example 29.** Suppose \( Z(t), t = \ldots, -1, 0, 1, \ldots \) is a mean zero process with exponential covariance \( C(i) = \exp(-i) \equiv \rho^i \) observed at \( Z = (Z(1), Z(2), \ldots, Z(n))^T \). Then the covariance
matrix $\Sigma = \text{Cov}(Z,Z)$ has inverse

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix}
1 & -\rho & 0 & \cdots & 0 & 0 & 0 \\
-\rho & 1 + \rho^2 & -\rho & \cdots & 0 & 0 & 0 \\
0 & -\rho & 1 + \rho^2 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 + \rho^2 & -\rho & 0 \\
0 & 0 & 0 & \cdots & -\rho & 1 + \rho^2 & -\rho \\
0 & 0 & 0 & \cdots & 0 & -\rho & 1 \\
\end{pmatrix}$$

Homework will ask you to verify that prediction of $Z(0)$ reduces to $\hat{f}(0) = \rho Z(1)$, i.e., all other observations have been screened out by the nearest one.

Conditions on the covariance function that (don’t) result in a screening effect are complicated, and are an active area of current research.

6.7 Conditional Simulation

Suppose $Z(s)$ is a mean zero stationary process with covariance $C(\cdot)$ that has been observed at locations $s_1, \ldots, s_n$. *Conditional simulation* is a method to generate realizations of the random field at new locations that are consistent with the observations $\{Z(s_i)\}_{i=1}^n$ (i.e., realizations that pass through the observed values and that has unconditional covariance function $C(\cdot)$). Another way to say this is that a conditional simulation is a possible realization of the random function that could have given rise to the observed data. [Note the following development supposes $Z$ is stationary, but this isn’t necessary and will work for nonstationary processes].

Call $\hat{Z}(s)$ the simple kriging predictor at location $s$. A conditional simulation at (new) location $s_0$ takes the following steps:

1. *Unconditionally* simulate a vector

\[ (W(s_1), W(s_2), \ldots, W(s_n), W(s_0)) \]

with the same covariance function as $Z$.

2. Krige $W = (W(s_1), W(s_2), \ldots, W(s_n))^T$ to $s_0$, call this $\hat{W}(s_0) = \Sigma_0 \Sigma^{-1} W$. 

54
(3) Form the conditional simulation

\[ Z_{CS}(s_0) = \hat{Z}(s_0) + (W(s_0) - \hat{W}(s_0)). \]

We claim

\[ (Z(s_1), Z(s_2), \ldots, Z(s_n), Z_{CS}(s_0)) = (Z^T, Z_{CS}(s_0)) \]

is a realization with the correct covariance structure in that

\[ \text{Var}(Z_{CS}(s_0)) = C(0) \]

and

\[ \text{Cov}(Z_{CS}(s_0), Z) = (C(s_0 - s_1), \ldots, C(s_0 - s_n)) = \Sigma_0. \]

Note that \( EZ_{CS}(s) = 0 \). We have

\[
\text{Var}(Z_{CS}(s_0)) = \text{Var}(\Sigma_0 \Sigma^{-1} Z + W(s_0) - \Sigma_0 \Sigma^{-1} W) \\
= \Sigma_0 \Sigma^{-1} \Sigma_0^T + C(0) - \Sigma_0 \Sigma^{-1} \Sigma_0^T \\
= C(0).
\]

[Check the covariance: homework].

At data locations \( s_0 = s_i \), we have

\[ Z_{CS}(s_i) = \hat{Z}(s_i) - (W(s_i) - \hat{W}(s_i)) = Z(s_i) + W(s_i) - W(s_i) = Z(s_i), \]

so that the conditional simulation passes through the observations when there is no nugget effect.

Moreover, for a fixed set of data \( Z \), the conditional simulation \( Z_{CS}(s_0) \) is only a function of \( W(s_0) - \hat{W}(s_0) \), so

\[ \text{Var}(Z_{CS}(s_0)|Z) = \text{Var}(W(s_0) - \hat{W}(s_0)) = \sigma_k(s_0)^2 \]

is the kriging variance. Sometimes it is easier to generate conditional samples \( Z_{CS} \) than it is to explicitly calculate \( \sigma_k(s_0)^2 \): if that is the case, then \( \sigma_k(s_0)^2 \) can be empirically estimated by Monte Carlo methods (that is, simulate many \( Z_{CS} \)’s and take their empirical variance).

[R example: (ConditionalSimulation.R)]
6.8 The Effect of Including a Nugget Effect

Kriging is often referred to as an interpolator, that is if \( Y(s_i) \) is an observation, then \( \hat{Y}(s_i) = Y(s_i) \). This is not always true, and it strongly depends on modeling assumptions. Let’s examine the simple kriging estimator under two models:

- **Model 0**: \( Y(s) = Z(s) \) and
- **Model 4**: \( Y(s) = Z(s) + \varepsilon(s) \)

that is, both mean zero, and without/with measurement error, respectively. Suppose \( Z(s) \) is stationary with mean zero and has covariance \( C(h) \) where \( C(0) = \text{Var}Z(s) = \sigma^2 \), and the process is observed only at location \( s_0 \).

- **Kriging under Model 0**

  We would krig using

  \[
  \Sigma_0 = \text{Cov}(Z(s_0), Y(s_0)) = \sigma^2 \quad \Sigma = \text{Cov}(Y(s_0), Y(s_0)) = \sigma^2
  \]

  so that

  \[
  \hat{Z}(s_0) = \Sigma_0 \Sigma^{-1} Y(s_0) = \sigma^2 \sigma^{-2} Y(s_0) = Y(s_0).
  \]

- **Kriging under Model 4**

  Now, if \( \varepsilon(\cdot) \) is a mean zero white noise process with \( \text{Var}\varepsilon(s) = \tau^2 \), we have

  \[
  \Sigma_0 = \text{Cov}(Z(s_0), Y(s_0)) = \sigma^2 \quad \Sigma = \text{Cov}(Y(s_0), Y(s_0)) = \sigma^2 + \tau^2
  \]

  so that

  \[
  \hat{Z}(s_0) = \Sigma_0 \Sigma^{-1} Y(s_0) = \frac{\sigma^2}{\sigma^2 + \tau^2} Y(s_0).
  \]

  In particular, note that kriging with a nugget has the effect of shrinking \( Y(s_0) \) toward the mean of the process, zero, and no longer interpolates, but rather smooths the data.

  However, we can still have an interpolator even under Model 4, depending on what we are predicting. It all comes down to whether we want to predict the process \( Z(s_0) \) or an
observation $Y(s_0)$ under the model with measurement error. The two cases are distinguished by minimizing $E(Z(s_0) - \hat{Z}(s_0))^2$ and $E(Y(s_0) - \hat{Y}(s_0))^2$, respectively, (noting that $\hat{Z}(s_0) = wY(s_0)$ and $\hat{Y}(s_0) = wY(s_0)$ are still linear predictors) and lead to predictors

$$
\hat{Z}(s_0) = \text{Cov}(Z(s_0), Y(s_0)) \Sigma^{-1} Y(s_0) = \frac{\sigma^2}{\sigma^2 + \tau^2} Y(s_0)
$$

$$
\hat{Y}(s_0) = \text{Cov}(Y(s_0), Y(s_0)) \Sigma^{-1} Y(s_0) = \frac{\sigma^2 + \tau^2}{\sigma^2 + \tau^2} Y(s_0) = Y(s_0).
$$

Thus, kriging is a smoother only in the case where we seek to estimate an underlying process $Z$ based on noisy observational data; however, this is a common goal in many applications.

Now, suppose

$$
Y(s) = Z(s) + \varepsilon(s)
$$

where $\text{Cov}(Z(s + h), Z(s)) = C(h) = \sigma^2 R(h)$ where $R$ is the correlation function for $Z$. Note that $R$ is a correlation function in that $R(0) = 1$ and $C$ is the covariance function in that $C(0) = \sigma^2$. $R$ can be defined as $R(h) = \text{Cor}(Z(s), Z(s + h))$.

Given observations $Y(s_1), \ldots, Y(s_n)$ if we seek to predict $Z(s_0)$, note that

$$
\text{Cov}(Z(s_i), Y(s_0)) = C(s_i - s_0) = \sigma^2 R(s_i - s_0)
$$

$$
\text{Cov}(Y(s_i), Y(s_j)) = C(s_i - s_j) + \tau^2 \mathbb{I}_{[s_i = s_j]} = \sigma^2 R(s_i - s_0) + \tau^2 \mathbb{I}_{[s_i = s_j]}
$$

so

$$
\Sigma_0 = \text{Cov}(Z(s_0), Y) = \sigma^2 (R(s_0 - s_1), \ldots, R(s_0 - s_n)) = \sigma^2 \Sigma_{R0}
$$

and

$$
\Sigma = \text{Cov}(Y, Y) = \sigma^2 (R(s_i - s_j))_{i,j=1}^n + \tau^2 I = \sigma^2 \Sigma_R + \tau^2 I
$$

where $\Sigma_{R0}$ and $\Sigma_R$ are the same matrices as $\Sigma_0$ and $\Sigma$, but evaluated at $R$ instead of $C$. Then the kriging predictor can be written

$$
\hat{Z}(s_0) = \Sigma_0 \Sigma^{-1} Y
$$

$$
= \sigma^2 \Sigma_{R0} (\sigma^2 \Sigma_R + \tau^2 I)^{-1} Y
$$

$$
= \Sigma_{R0} \left( \Sigma_R + \frac{\tau^2}{\sigma^2} I \right)^{-1} Y
$$

57
so that smoothing is a function of the *nugget-to-marginal variance ratio* $\tau^2/\sigma^2$. If $\tau^2 \gg \sigma^2$ then kriging will shrink all data and surface estimates toward the mean, whereas $\sigma^2 \gg \tau^2$ will be close to an interpolation of the surface (but not strict interpolation).

[R example: (KrigingWithNugget.R)]

### 6.9 Kriging Weight Function as a Local Smoother

Recall that the kriging estimator is

$$\hat{Z}(s_0) = \frac{1}{n} \sum_{i=1}^{n} w_i Y(s_i).$$

The weights change with the location we want to predict, so we might want to write $w_i = w(s_0, s_i)$ for the weight put on the $i$th location when predicting at $s_0$. The weight function also depends on the exact distribution of the remaining locations $s_1, \ldots, s_n$, and typically looks like a kernel function.

[R example: (KrigingWeightFunction.R)]
7 Likelihood-Based Methods

The estimation techniques of the classical geostatistical section are not known to be optimal by any current metrics. Maximum likelihood (ML) is a preferred method for estimation given the optimality of these estimators, as well as desirable asymptotic properties. We review some simple applications of maximum likelihood, and recall important properties of ML estimators.

**Definition 30.** If \( X_1, \ldots, X_n \) are a set of random variables with joint probability density function \( L(\theta; x_1, \ldots, x_n) \) that depends on a set of parameters \( \theta \in \mathbb{R}^p \), then \( \hat{\theta} \) is a maximum likelihood estimator (MLE) for \( \theta \) if

\[
\hat{\theta} = \arg \max_\theta L(\theta; x_1, \ldots, x_n).
\]

In practice, likelihood values for large numbers of random variables are very small, so the MLE is a value that maximizes the log-likelihood

\[
\ell(\theta; x_1, \ldots, x_n) = \log L(\theta; x_1, \ldots, x_n).
\]

**Example 31.** Suppose \( X_1, \ldots, X_n \) are iid Bernoulli(p) random variables (that is, with pmf \( p(x) = p^x(1-p)^{1-x} \) for \( x = 0, 1 \)). Then the likelihood for observed values \( x_1, \ldots, x_n \in \{0, 1\} \) is

\[
L(x_1, \ldots, x_n) = L(p; x_1, \ldots, x_n) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}
\]

with log-likelihood

\[
\ell(p; x_1, \ldots, x_n) = \sum_{i=1}^n x_i \log p + (n - \sum_{i=1}^n x_i) \log(1-p)
\]

\[
= n\bar{x} \log p + n(1-\bar{x}) \log(1-p).
\]

Maximizing over \( p \) yields \( \hat{p} = \bar{x} \), i.e., the sample mean is a MLE for \( p \).

**Example 32.** Suppose \( Y_i = x_i^T \beta + \varepsilon_i, i = 1, \ldots, n \) where \( \varepsilon_i \) are correlated normal errors with known covariance matrix \( \text{Cov}(\varepsilon, \varepsilon) = \Sigma \) (i.e., \( Y = X\beta + \varepsilon \)). Then the likelihood function for \( Y = y \) is

\[
L(y) = L(\beta; y) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp \left( -\frac{1}{2} (y - X\beta)^T \Sigma^{-1} (y - X\beta) \right).
\]

The negative log-likelihood is then

\[
-\ell(\beta; y) = (1/2) \log((2\pi)^n \det \Sigma) + (1/2)(y - X\beta)^T \Sigma^{-1}(y - X\beta).
\]
Minimizing $-\ell(\beta; y)$ over $\beta$ yields the generalized least squares estimator for $\beta$, $\hat{\beta}_{\text{GLS}} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y$. Thus, the GLS estimator is a maximum likelihood estimator under a joint normality assumption if the covariance structure is known.

Under certain regularity conditions, ML estimators are consistent and asymptotically normal.

**Theorem 33.** Suppose $X_1, \ldots, X_n$ are iid with density $L_{\theta_0}(x), \theta \in \mathbb{R}^p$. If $\hat{\theta}_n$ is a maximum likelihood estimator for $\theta_0$, then under some conditions on the likelihood function,

- $\hat{\theta}_n \rightarrow_P \theta_0$
- $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, I^{-1}(\theta_0))$.

Here, $I(\theta_0)$ is the *Fisher information matrix*,

$$I(\theta_0) = \left( E \left( -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log L_{\theta_0}(X) \right) \right)_{i,j=1}^p$$

where the likelihood is evaluated at a single random variable. Note that this theorem refers to a sequence of iid random variables; in the correlated case it is not trivial to show that maximum likelihood estimators are consistent, and in fact sometimes they are not, depending on the type of asymptotics in use.

**Example 34.** Suppose $X_1, \ldots, X_n$ are iid Bernoulli($p$) random variables with

$$\ell(p; x_1, \ldots, x_n) = n\bar{x} \log p + n(1 - \bar{x}) \log (1 - p).$$

Then

$$\frac{\partial^2}{\partial p^2} \ell(p; X) = -\frac{X}{p^2} - \frac{1 - X}{(1 - p)^2}$$

so

$$I(\theta_0) = -E \left( -\frac{X}{p^2} - \frac{1 - X}{(1 - p)^2} \right)$$

$$= \frac{1}{p} + \frac{1}{(1 - p)}$$

$$= \frac{1}{p(1 - p)}.$$

Thus, the ML estimator $\hat{p} = \bar{X}$ is consistent and has asymptotic distribution

$$\sqrt{n}(\hat{p} - p) \rightarrow N(0, p(1 - p)).$$
7.1 Maximum Likelihood for Spatial Processes

Suppose \( Y(s) = x(s)^T \beta + W(s) \) where \( W(s) \) is a Gaussian process (either with or without a nugget) is observed at locations \( s_1, \ldots, s_n \). If \( W = (W(s_1), \ldots, W(s_n))^T \) has covariance matrix \( \Sigma = \Sigma(\theta) \) where \( \theta \) is a vector of spatial parameters (e.g., \( \theta = (\tau^2, \sigma^2, a, \nu)^T \)) and \( EY = X\beta \), then the negative log-likelihood function for \( Y \) is

\[
-\ell(\beta, \theta; y) = \frac{n}{2} \log(2\pi) + \frac{1}{2} \log \det \Sigma(\theta) + \frac{1}{2} (y - X\beta)^T \Sigma(\theta)^{-1} (y - X\beta).
\]

In practice none of the parameters are known, and closed form solutions maximizing \( \ell \) are unavailable. Thus, we rely on numerically maximizing \( \ell \) (or minimizing \(-\ell\)).

For any fixed value \( \theta \), the value of \( \beta \) that maximizes \( \ell \) is \( \hat{\beta}(\theta, y) = (X^T \Sigma(\theta)^{-1} X)^{-1} X^T \Sigma(\theta)^{-1} y \). So to maximize \( \ell \) over both \( \theta \) and \( \beta \), we can maximize

\[
\log \det \Sigma(\theta) + (y - X\hat{\beta}(\theta, y))^T \Sigma(\theta)^{-1} (y - X\hat{\beta}(\theta, y)).
\]

Plugging in \( \hat{\beta}(\theta, y) \) yields

\[
-\ell(\theta; y) = \frac{n}{2} \log(2\pi) + \frac{1}{2} \log \det \Sigma + \frac{1}{2} y^T (\Sigma^{-1} - \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}) y \tag{4}
\]

which is a function of only \( \theta \). Eliminating \( \beta \) is called profiling and \( \ell(\theta; y) \) is the profile log likelihood for \( \theta \).

Closed form solutions for \( \hat{\theta} \) are usually unavailable, so maximizing \( \ell(\theta; y) \) over \( \theta \) requires nonlinear optimization techniques. If \( \theta \in \mathbb{R}^p \) where \( p \) is small, then sometimes a grid search for \( \theta \) can be used. Otherwise, gradient algorithms are typically used.

Gradient algorithms are iterative, where the \((k + 1)\)st iterate \( \theta^{(k+1)} \) is computed from the \( k \)th \( \theta^{(k)} \) via

\[
\theta^{(k+1)} = \theta^{(k)} + \rho^{(k)} M^{(k)} g^{(k)}
\]

where \( \rho^{(k)} \) is scalar, \( M^{(k)} \) is \( p \times p \) and \( g^{(k)} \) is the gradient of \( \ell(\theta; y) \) evaluated at \( \theta^{(k)} \). [The second term in the gradient algorithm defines the size of the step \( \rho \) to be taken in a particular direction \( (Mg) \)]

The two most common algorithms to maximize a log-likelihood are the Newton-Raphson and Fisher scoring procedures. In Newton-Raphson,

\[
(M^{(k)})^{-1} = \left( -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\theta; y) \right)_{i,j=1}^p \Bigg|_{\theta=\theta^{(k)}}
\]

61
In the Fisher scoring algorithm,

\[(M^{(k)})^{-1} = B^{(k)}\]

where \(B^{(k)}\) is the Fisher information matrix for \(\ell(\theta; Y)\) evaluated at \(\theta^{(k)}\). Thus, the difference between the two is that the second-order partials are replaced by their expectations. In both cases, \(\rho^{(k)} = 1\). In practice, the search starting value can be informed by a traditional geostatistical estimate.

In many standard statistics problems, a unique MLE exists. For spatial data there is no guarantee that a unique MLE exists, and often the (log)-likelihood surface appears relatively flat so that finding a maximum can be difficult. Although simple cases of the likelihood having multiple modes have been published, in our experience these are rare in practice. [To determine if the likelihood is multi-modal one might start iterative procedures from differing starting values]

Finally, the Cholesky decomposition can be used to alleviate some computational issues with solving for likelihood values. In particular, if \(\Sigma = MM^T\) is the Cholesky decomposition where \(M\) is lower triangular, then

\[z^T\Sigma^{-1}z = z^T(MM^T)^{-1}z = z^TM^{-T}M^{-1}z = (M^{-1}z)^T(M^{-1}z).\]

A forward solve can be used to calculate \(M^{-1}(z)\) efficiently. Additionally recall that, for two square matrices \(A\) and \(B\), \(\det AB = \det A \det B\) and if \(A\) is triangular (lower or upper), \(\det A\) is the product of its diagonal elements. Thus,

\[\det \Sigma = \det MM^T = \det M \det M^T\]

can be found by taking the product of \(M\)’s diagonal, squared.

Under some regularity conditions, MLEs for spatial processes are asymptotically normal.

**Theorem 35.** If \(Y = X\beta + W\) where \(W(s)\) has covariance \(C_\theta\), then under some conditions, if \(\phi = (\beta^T, \theta^T)^T\) has MLE \(\hat{\phi} = (\hat{\beta}^T, \hat{\theta}^T)^T\), for large sample size \(n\),

\[\hat{\phi} \sim N(\phi, B_n^{-1})\]

where

\[B_n = -E\left(\frac{\partial^2}{\partial \phi_i \partial \phi_j} \ell(\phi; y)\right)_{i,j}\]
with \( \ell(\phi; y) \) being the log-likelihood function for \( Y \).

[R example (MaximumLikelihood.R)]

### 7.2 REML

In standard ML estimation for spatial processes, the MLEs \( \hat{\theta} \) are biased due to simultaneous estimation of \( \beta \). For example, if \( \beta = \beta_0 \) were the true parameter but we used the MLE \( \hat{\beta} \), we have by construction

\[
(Y - X\hat{\beta})^T \Sigma^{-1}(Y - X\hat{\beta}) \leq (Y - X\beta_0)^T \Sigma^{-1}(Y - X\beta_0)
\]

so that \( X\hat{\beta} \) is “closer” to \( Y \) than the truth \( X\beta_0 \). Thus, the MLE for \( \theta \) will tend to underestimate the variation of the process than if we knew the true \( \beta_0 \).

This bias can be substantial for even moderate sample sizes if the spatial correlation is strong or \( \beta \) is high dimensional. The bias can sometimes be reduced using a variant of ML known as restricted (or residual) maximum likelihood (REML).

Under the model for observations \( Y = X\beta + W \) where \( \text{Var} W = \Sigma \) and \( X \) is of full rank \( p + 1 \), the REML estimators are those found by maximizing the (log)-likelihood of any set of \( n - (p + 1) \) linearly independent elements of the vector of contrasts

\[
Y - X\hat{\beta}_{\text{OLS}} = (I - X(X^TX)^{-1}X^T)Y.
\]

We say two error contrasts \( a^TY \) and \( b^TY \) are linearly independent if \( a \) and \( b \) are linearly independent vectors.

Up to a constant \( K \), the REML estimator for \( \theta \) is the value minimizing

\[
-\ell_R(\theta; y) = K + \frac{1}{2} \log \det \Sigma + \frac{1}{2} \log \det(X^T\Sigma^{-1}X) + \frac{1}{2}y^T(\Sigma^{-1} - \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1})y.
\]

Deriving this log-likelihood is not trivial. As with MLEs in the spatial case, \( \ell_R(\theta; y) \) must be maximized numerically, after which the estimator for \( \beta \) is the usual GLS estimator with \( \hat{\theta} \) plugged in.
7.3 Model Comparisons

A common goal is to select a model by comparing competing options. For example, whether or not to include elevation as a covariate, or whether a Matérn or Cauchy covariance better describes the covariance structure of a given dataset. In usual likelihood theory, likelihood ratio tests can be performed to compare full parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ to reduced dimensional versions.

If $\beta \in \mathbb{R}^{p+1}$ and $\theta \in \mathbb{R}^{m}$, non-nested model may be informally compared using information criteria such as Akaike’s information criterion,

$$AIC = -2 \log L(\hat{\beta}, \hat{\theta}) + 2(p + 1 + m)$$

or Schwarz’s Bayesian information criterion,

$$BIC = -2 \log L(\hat{\beta}, \hat{\theta}) + (p + 1 + m) \log n$$

where $n$ samples are available, and $\hat{\beta}$ and $\hat{\theta}$ are the MLEs. When comparing models, the model with the smaller value of AIC or BIC is preferred. BIC tends to favor simpler models than AIC. Note that both criteria penalize the mean parameters $\beta$ and covariance parameters $\theta$ in an equal way, but it is not clear that this is appropriate.

Within a Bayesian framework, the deviance information criterion (DIC) is common,

$$DIC = \bar{D} + p_D$$

where

$$\bar{D} = \mathbb{E}(D(\beta, \theta)|y)$$

and

$$D(\beta, \theta) = -2 \log L(\beta, \theta; y).$$

The final term is

$$p_D = \bar{D} - D(\bar{\beta}, \bar{\theta})$$

where $(\bar{\beta}, \bar{\theta})$ are the posterior means,

$$(\bar{\beta}^T, \bar{\theta}^T)^T = \mathbb{E}((\beta^T, \theta^T)^T|y).$$
The first term of \( DIC \) is a measure of model fit, while the second term \( p_D \) is a measure of effective number of parameters. \( DIC \) is usually calculated by using Markov chain Monte Carlo samples from the posterior.

[R example (ModelComparison.R)]

### 7.4 Computational Issues

Likelihood-based estimation generally requires substantial computations, where generally calculating the determinant of \( \Sigma \) and quadratic forms involving \( \Sigma^{-1} \) both require \( O(n^3) \) calculations. Additionally, for each proposed value of \( \theta \) these quantities must be re-calculated. Thus, for large sample sizes ML estimation becomes infeasible.

Often the covariance function can be written \( C(h) = \sigma^2 R(h) \) for a correlation function \( R \), we have that \( \Sigma(\theta) = \sigma^2 \Sigma_R \) where \( \Sigma_R \) is a correlation matrix only involving remaining spatial parameters (note \( (\sigma^2 \Sigma_R)^{-1} = \sigma^{-2} \Sigma_R^{-1} \)). For example, this occurs when \( Y = X\beta + Z \) and \( Z \) is a stationary process with covariance \( C \). Then we have an expression for \( \text{Var} Y \) as

\[
\Sigma = (C(s_i - s_j))_{i,j=1}^n = \sigma^2 (R(s_i - s_j))_{i,j=1}^n = \sigma^2 \Sigma_R.
\]

In this case, the log-likelihood function can be written, using the previous profiling of \( \hat{\beta}_{GLS} \),

\[
-\ell(\theta; y) = \frac{n}{2} \log(2\pi) + \frac{1}{2} \log \det \Sigma + \frac{1}{2} y^T (\Sigma^{-1} - \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}) y
= \frac{n}{2} \log(2\pi) + \frac{n}{2} \log \sigma^2 + \frac{1}{2} \log \det \Sigma_R + \frac{1}{2\sigma^2} y^T Q_R y
\]

where

\[
Q_R = (\Sigma_R^{-1} - \Sigma_R^{-1} X (X^T \Sigma_R^{-1} X)^{-1} X^T \Sigma_R^{-1}).
\]

Taking the derivative with respect to \( \sigma^2 \) and minimizing yields

\[
\hat{\sigma}^2 = \frac{1}{n} y^T Q_R y.
\]

Plugging this back into \( \ell(\theta; y) \) implies that remaining parameters are found by minimizing

\[
-\ell_{-1}(\theta_{-1}; y) = \log \det \Sigma_R + n \log(y^T Q_R y).
\]
In this instance we have profiled over $\sigma^2$. Note the presence of $\det \Sigma_R$ and $\Sigma_R^{-1}$ may still yield obstacles for $n \gg 0$.

Consider breaking up the observation vector into $b$ shorter components, and define

$$y = (y_1^T, \ldots, y_b^T)^T$$

$$y_{(j)} = (y_1^T, \ldots, y_j^T)^T, \quad j = 1, \ldots, b.$$  

Then if $y$ has density $L(y; \beta, \theta)$, we can write

$$L(y; \beta, \theta) = L(y_1; \beta, \theta) \prod_{j=2}^b L(y_j|y_{(j-1)}; \beta, \theta)$$

as an exact decomposition. *Approximate likelihood* involves replacing $y_{(j-1)}$ by some subvector, $w_{(j-1)}$ of $y_{(j-1)}$, i.e.,

$$L(y_1; \beta, \theta) \prod_{j=2}^b L(y_j|y_{(j-1)}; \beta, \theta) \rightarrow L(y_1; \beta, \theta) \prod_{j=2}^b L(y_j|w_{(j-1)}; \beta, \theta)$$

which yields smaller matrix inverses and determinants.

*Composite likelihood* is a similar idea in which the full likelihood is approximated by a product of marginal likelihoods,

$$L(y; \beta, \theta) \approx \prod_{j=1}^b L(y_j; \beta, \theta).$$

By judiciously choosing the subcomponents $y_j$, no $n \times n$ matrix calculations are required. The key is to balance enough observations within $y_j$ to accurately describe the spatial behavior while keeping the sample small enough to afford feasible calculations. In both the approximate and composite likelihood cases, the price for reduced computational cost is loss of efficiency in estimation.

*Covariance tapering* is based on the idea that for distance pairwise locations $\|s_i - s_j\| \gg 0$, we have $C(s_i, s_j) \approx 0$. If these small values can be thresholded to *exactly* zero, then sparse matrix methods can be used to perform calculations involving the covariance matrix. In particular, if $Z(s)$ has covariance $C(h; \theta)$, then we swap in the *tapered covariance*

$$C_1(h; \theta, \gamma) = C(h; \theta)C_T(h; \gamma)$$

where $C_T$ is a compactly supported covariance parameterized by $\gamma$, such as the spherical or Wendland covariance. Properties of positive definite functions imply $C_1$ is a valid covariance.