ST3454 Stochastic processes in Space and Time

Rozenn Dahyot

School of Computer Science & Statistics
Trinity College Dublin, Ireland
https://www.scss.tcd.ie/Rozenn.Dahyot/
Rozenn.Dahyot@scss.tcd.ie

Hilary term 2017
This course is concerned by the study of quantities (stochastic processes) that varies over the spacetime domain.

The spacetime of our universe is usually interpreted as an Euclidean space with space consisting of three dimensions, and time as consisting of one dimension (the 4\textsuperscript{th} dimension).
Space and time II

In this course:

- space $x$ will be limited to $\mathbb{R}^2$ (2D Euclidian space) as a local planar approximation of the earth surface

- time $t$ will be limited to $\mathbb{R}^+$. 
we note $s$ the stochastic process of interest, that may vary over the time domain $s(t)$, the space domain $s(x)$, or spacetime $s(x, t)$. For instance, the stochastic process of interest may be the temperature in Ireland (space) over the next few days (time).
Sometimes, \( s \) is not measurable directly (at any location in spacetime) but instead we can observe another related stochastic process \( o \).

A simple example is \( o \) the temperature as read on a thermometer, while \( s \) is the true temperature with the relation \( o = s + \epsilon \) such that \( \epsilon \) is the noise associated with the sensor (thermometer).
Many techniques seen in this class are Linear Smoothing methods:

\[ s(x, t) = \sum_{i=1}^{n} \lambda_i(x, t) \ s^{(i)} + \epsilon(x, t) \]

with having observed the stochastic process \( s \) at \( n \) (spacetime) locations \( \{(x^{(i)}, t^{(i)}, s^{(i)})\}_{i=1,...,n} \).

The goal is to estimate the quantity \( s \) at a new given location \( (x, t) \) and if possible getting a measure for its uncertainty.
Estimation depends on the model chosen for the weights $\lambda$ and hypotheses about $\epsilon$.

$$\hat{s}(x, t) = \sum_{i=1}^{n} \hat{\lambda}_i(x, t) s^{(i)} \quad \text{or for short} \quad \hat{s} = \sum_{i=1}^{n} \hat{\lambda}_i s^{(i)} = \langle \vec{\lambda} | \vec{s} \rangle$$

Note $\sum_{i=1}^{n} \hat{\lambda}_i s^{(i)} = \vec{\lambda}^T \vec{s} = \langle \vec{\lambda} | \vec{s} \rangle$.

Examples: Regression, ARIMA, Filtering, Kriging, Nadaraya-Watson estimator.
Content & Exam

We will attempt to study all the following:

- Regression as a Linear Smoothing method
- ARIMA models
- Multivariate State-Space Model & Kalman Filter,
- Kriging,
- Functional data analysis

For illustration, some R labs may be organised.

Exam (100%):
References

- Geostatistics for Environmental Scientists, R. Webster & M. A. Oliver.
Stochastic processes I

Definition (stochastic process)

A Spatio-temporal stochastic process $s(x, t)$ is defined such that:

- $x$ is a spatial location,
- $t$ is the time,
- $s$ is a random quantity of interest. $s$ takes values in the state space $S$.

For a two state model, $S = \{S_1, S_2\}$, a probability $\mathbb{P}(s_{xt} = S_1)$ (or $\mathbb{P}(s_{xt} = S_2)$), such that $\mathbb{P}(s_{xt} = S_1) + \mathbb{P}(s_{xt} = S_2) = 1$ is associated with $s$ at the spatiotemporal position $(x, t)$ (noted $s_{xt}$).
Like the state space, the variables $x$ and $t$ can take values in a **discrete** (e.g. $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{N}^2$ etc.) or **continuous** spaces (e.g. $\mathbb{R}$, $\mathbb{R}^2$, or $\mathbb{R}^3$ or $S^2$ (unit sphere)).

When no direct observations is available for $s$ (hidden random variable), another stochastic process $o(x, t)$ linked to $s(x, t)$ and for which observations are available, is defined.
Many applications use stochastic processes to estimate $s$ with its associated uncertainty at a new (unobserved) location in spacetime.

For most examples in this course, we look at stochastic processes $s$ that are 1-dimensional (it is a random variable and not a random vector).
Stochastic processes IV

- Several cost functions (e.g. probability density functions) will be presented in the course to compute a prediction (estimate) and its associated uncertainty.

- The Normal distribution is often used as a hypothesis in these modellings.
Imagine the following model for stochastic process $s$:

$$s(t) = c + \alpha t + \epsilon(t)$$

Let assume:

- $\mathbb{E}[\epsilon(t)] = 0$, $\forall t$.
- $\mathbb{E}[(\epsilon(t))^2] = \sigma^2$, $\forall t$. 

Questions. Assuming that we have collected \( n \) observations \( \{(t^{(i)}, s^{(i)})\}_{i=1}^{n} \)

- propose an estimate \( \hat{s} \) for \( s \) at a new location \( t \).
- show that that estimate can be written as

\[
\hat{s}(t) = \sum_{i=1}^{n} \hat{\lambda}_i(t) \ s^{(i)}
\]
Regression as a Linear Smoothing method III

**Answers.** Defining $\vec{s} = [s^{(1)} \ldots s^{(n)}]^T$ and matrix

$$U = \begin{bmatrix} 1 & t^{(1)} \\ \vdots & \vdots \\ 1 & t^{(n)} \end{bmatrix}$$

then the least squares estimate minimising the sum of square errors

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (s^{(i)} - c - \alpha t^{(i)})^2$$

is

$$\begin{bmatrix} \hat{c} \\ \hat{\alpha} \end{bmatrix} = (U^T U)^{-1} U^T \vec{s}$$
Having compute estimates $\hat{c}$ and $\hat{\alpha}$ an estimate for $s$ at new temporal location $t$ can be proposed:

$$\hat{s} = \hat{c} + \hat{\alpha} t = [1 \ t] \begin{bmatrix} \hat{c} \\ \hat{\alpha} \end{bmatrix} = [1 \ t] \ (U^T U)^{-1} U^T \bar{s} = \sum_{i=1}^{n} \hat{\lambda}_i \ s^{(i)}$$
We suggested the following model for stochastic process $s$:

$$s(t) = c + \alpha t + \epsilon(t) \quad \text{with} \quad \epsilon(t) \sim \mathcal{N}(0; \sigma^2), \forall t.$$
Exercise: Consider the model

\[ s(x, t) = c + \alpha t + \beta x + \gamma x t + \epsilon(x, t) \quad \text{with} \quad \epsilon(x, t) \sim \mathcal{N}(0; \sigma^2), \forall t \]

\( n \) observations \( \{(x^{(i)}, t^{(i)}, s^{(i)})\}_{i=1,\ldots,n} \) are available.

- Propose an estimate \( \hat{s} \) at a new location \( (x, t) \).
- Show that this solution can also be written \( \hat{s}(x, t) = \sum_{i=1}^{n} \hat{\lambda}_i(x, t) s^{(i)} \).
### Example time series: Beer production

The Monthly Australian beer production has been observed between Jan 1991 to Aug 1995.

<table>
<thead>
<tr>
<th></th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>164</td>
<td>148</td>
<td>152</td>
<td>144</td>
<td>155</td>
<td>125</td>
<td>153</td>
<td>146</td>
<td>138</td>
<td>190</td>
<td>192</td>
<td>192</td>
</tr>
<tr>
<td>1992</td>
<td>147</td>
<td>133</td>
<td>163</td>
<td>150</td>
<td>129</td>
<td>131</td>
<td>145</td>
<td>137</td>
<td>138</td>
<td>168</td>
<td>176</td>
<td>188</td>
</tr>
<tr>
<td>1993</td>
<td>139</td>
<td>143</td>
<td>150</td>
<td>154</td>
<td>137</td>
<td>129</td>
<td>128</td>
<td>140</td>
<td>143</td>
<td>151</td>
<td>177</td>
<td>184</td>
</tr>
<tr>
<td>1994</td>
<td>151</td>
<td>134</td>
<td>164</td>
<td>126</td>
<td>131</td>
<td>125</td>
<td>127</td>
<td>143</td>
<td>143</td>
<td>160</td>
<td>190</td>
<td>182</td>
</tr>
<tr>
<td>1995</td>
<td>138</td>
<td>136</td>
<td>152</td>
<td>127</td>
<td>151</td>
<td>130</td>
<td>119</td>
<td>153</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Questions:

- What are the patterns in the beer data?
- What model can you propose to model this stochastic process? (what explanatory variable would you use?)
ARIMA models I

Stochastic process $s$ in the time domain

- **AR(1)**

\[
\{ s(t + \Delta) = c + \phi_1 s(t) + \epsilon(t + \Delta) \} \quad \equiv \quad \{ s_{t+1} = c + \phi_1 s_t + \epsilon_{t+1} \}
\]

- **MA(1)**

\[
\{ s(t + \Delta) = c + \phi_1 \epsilon(t) + \epsilon(t + \Delta) \} \quad \equiv \quad \{ s_{t+1} = c + \phi_1 \epsilon_t + \epsilon_{t+1} \}
\]

with $\Delta$ a fixed interval of time.
Exercise: Assume a simple AR(1) model:

\[ s_t = \alpha s_{t-1} + \epsilon_t \quad \text{with} \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \quad \forall t, \ |\alpha| < 1 \]

with observations \( \{s^{(1)}_t\}_{t=1,\ldots,n} \). Draw a probabilistic graph to model this time series.

Exercise: Assume a simple AR(1) model:

\[ s_t = \alpha s_{t-1} + \epsilon_t \quad \text{with} \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \forall t, \ |\alpha| < 1 \]

Show

\[ p(s_t|s_{t-1}, \ldots, s_1) = p(s_t|s_{t-1}) = \mathcal{N}(s_t; \alpha s_{t-1}, \sigma^2) \]
Exercise: Given a temporal stochastic process $s$ with the following Markov property

$$p(s_t | s_{t-1}) = \mathcal{N}(s_t; \alpha s_{t-1}, \sigma^2) \quad \forall t$$

with observations $\{s_t^{(1)}\}_{t=1,\ldots,n}$.

- Estimate $s$ at a new time $n + 1$ with its confidence interval.
- Give an estimate for $s$ at a new time $n + 2$ with its confidence interval.
ARIMA models V

\[
\begin{align*}
&s_{t-1} \\
&\quad \downarrow \\
&s_{t} \\
&\quad \uparrow \\
&s^{(1)}_{t-1} \\
\end{align*}
\]

\[
\begin{align*}
&s_{n} \\
&\quad \downarrow \\
&s_{n+1} \\
&\quad \uparrow \\
&s^{(1)}_{n+1} \\
\end{align*}
\]

\[
\begin{align*}
&s_{n} \\
&\quad \downarrow \\
&s_{n+2} \\
&\quad \uparrow \\
&s^{(1)}_{n+2} \\
\end{align*}
\]
ARIMA models VI

We know

\[ p(s_t|s_{t-1}) = \mathcal{N}(s_t; \alpha s_{t-1}, \sigma^2) \quad \forall t \]

so at \( n + 1 \):

\[ p(s_{n+1}|s_n) = \mathcal{N}(s_{n+1}; \alpha s_n, \sigma^2) \]

and we use the observation \( s_n^{(1)} \) that is available for \( s_n \):

\[ p(s_{n+1}|s_n^{(1)}) = \mathcal{N}(s_{n+1}; \alpha s_n^{(1)}, \sigma^2) \]

A good estimate is the expectation \( \hat{s}_{n+1} = \alpha s_n^{(1)} \) and the 95% prediction interval is \( \hat{s}_{n+1} \pm 2\sigma \).
ARIMA models VII

At $n + 2$:

$$ p(s_{n+2} | s_{n+1}) = \mathcal{N}(s_{n+2}; \alpha s_{n+1}, \sigma^2) $$

But we don't have observation for $s_{n+1}$! so instead we use:

$$ p(s_{n+2} | s_n) = \int p(s_{n+2}, s_{n+1} | s_n) \, ds_{n+1} $$

$$ = \int p(s_{n+2} | s_{n+1}, s_n) \, p(s_{n+1} | s_n) \, ds_{n+1} $$

$$ = \int p(s_{n+2} | s_{n+1}) \, p(s_{n+1} | s_n) \, ds_{n+1} $$

$$ = \int \mathcal{N}(s_{n+2}; \alpha s_{n+1}, \sigma^2) \, \mathcal{N}(s_{n+1}; \alpha s_n, \sigma^2) \, ds_{n+1} $$

$$ = \mathcal{N}(s_{n+2}; \alpha^2 s_n, \sigma^2 (1 + \alpha^2)) $$

A good estimate is the expectation $\hat{s}_{n+2} = \alpha^2 s_n^{(1)}$ using the observation $s_n^{(1)}$, and the 95% prediction interval is $\hat{s}_{n+2} \pm 2\sigma \sqrt{1 + \alpha^2}$. 
ARIMA models VIII
Filters I

\[ \hat{s}_1 \rightarrow s_2 \rightarrow \ldots \rightarrow \hat{s}_{j-2} \rightarrow s_{j-1} \rightarrow \hat{s}_j \]

\[ s_1 \rightarrow o_1 \rightarrow o_1^{(1)} \]
\[ s_2 \rightarrow o_2 \rightarrow o_2^{(1)} \]
\[ s_{j-2} \rightarrow o_{j-2} \rightarrow o_{j-2}^{(1)} \]
\[ s_{j-1} \rightarrow o_{j-1} \rightarrow o_{j-1}^{(1)} \]
\[ s_j \rightarrow o_j \rightarrow o_j^{(1)} \]
Given a hidden temporal stochastic process \((s_1, \cdots, s_J)\) and observed one \((o_1, \cdots, o_J)\) linked by

\[
\begin{align*}
  &p_{s_1}(s_1) &\text{Initial state} \\
  &p_{s_j|s_{j-1}}(s_j|s_{j-1}) &\text{transition (order 1)} \\
  &p_{o_j|s_j}(o_j|s_j) &\text{link between observed variable and hidden state}
\end{align*}
\]

Estimate the posterior \(p_{s_j|o_1:j}\). Hint: Express \(p_{s_j|o_1:j}(s_j|o_1:j)\) w.r.t. \(p_{s_{j-1}|o_1:j-1}(s_{j-1}|o_1:j-1)\) (notation \(o_{1:j-1} = (o_1, \cdots, o_{j-1})\)).
Filters III

\[ p_{s_j|o_{1:j}}(s_j|o_{1:j}) = \int p(s_j, s_{j-1}|o_{1:j}) \, ds_{j-1} \quad \text{(Total probability theorem)} \]

\[ = \frac{1}{p(o_{1:j})} \int p(s_t, s_{t-1}, o_{1:j}) \, ds_{j-1} \quad \text{(Bayes)} \]

\[ = \frac{1}{p(o_{1:j})} \int p(s_j, s_{j-1}, o_j, o_{1:j-1}) \, ds_{j-1} \]

\[ = \frac{p(o_{1:j-1})}{p(o_{1:j})} \int p(o_j|s_j, s_{j-1}, o_{1:j-1}) \, p(s_j|s_{j-1}, o_{1:j-1}) \, p(s_{j-1}|o_{1:j-1}) \, ds_{j-1} \quad \text{(Bayes)} \]

\[ = \frac{p(o_{1:j-1})}{p(o_{1:j})} \int p(o_j|s_j) \, p(s_j|s_{j-1}) \, p(s_{j-1}|o_{1:j-1}) \, ds_{j-1} \]
Hence

\[ p(s_j|o_{1:j}) = \frac{p(o_{1:j-1})}{p(o_{1:j})} \cdot p(o_j|s_j) \int p(s_j|s_{j-1}) \cdot p(s_{j-1}|o_{1:j-1}) \, ds_{j-1} \]

given a hidden temporal stochastic process \((s_1, \cdots, s_J)\) and observed one \((o_1, \cdots, o_J)\) linked by

\[
\begin{align*}
    p_{s_1}(s_1) & \quad \text{Initial state} \\
    p_{s_j|s_{j-1}}(s_j|s_{j-1}) & \quad \text{transition (order 1)} \\
    p_{o_j|s_j}(o_j|s_j) & \quad \text{link between observed variable and hidden state}
\end{align*}
\]
In the general case, we have the following integral to solve:

\[ p(s_j | o_{1:j}) = \frac{p(o_{1:j-1})}{p(o_{1:j})} p(o_j | s_j) \int p(s_j | s_{j-1}) p(s_{j-1} | o_{1:j-1}) \, ds_{j-1} \]

1. Prediction:

\[ p(s_j | o_{1:j-1}) = \int p(s_j | s_{j-1}) p(s_{j-1} | o_{1:j-1}) \, ds_{j-1} \]

2. Update

\[ p(s_j | o_{1:j}) \propto p(o_j | s_j) p(s_j | o_{1:j-1}) \]
Having collected observations \( \{o_t^{(1)}\}_{t=1,\ldots,j-1} \) (also noted \( o_{1:j-1}^{(1)} \)), the probability density function to predict is \( s_j \) is

\[
p(s_j \mid o_{1:j-1}^{(1)})
\]

Once the new observation has been collected at time \( j \), \( o_j^{(1)} \), the probability density function taking this update into account is

\[
p(s_j \mid o_{1:j}^{(1)})
\]
Filters VII

\[ \hat{S}_1 \rightarrow S_1 \rightarrow O_1 \rightarrow O_1^{(1)} \]
\[ \hat{S}_2 \rightarrow S_2 \rightarrow O_2 \rightarrow O_2^{(1)} \]
\[ \ldots \]
\[ \hat{S}_{j-2} \rightarrow S_{j-2} \rightarrow O_{j-2} \rightarrow O_{j-2}^{(1)} \]
\[ \hat{S}_{j-1} \rightarrow S_{j-1} \rightarrow O_{j-1} \rightarrow O_{j-1}^{(1)} \]
\[ \hat{S}_j \rightarrow S_j \rightarrow O_j \rightarrow O_j^{(1)} \]
We have illustrated the idea behind Filtering with a transition model of order 1 (This can be extended to any order). In practice, we need to be able to calculate or compute $p(s_j | o_{1:j})$. Kalman filtering is one elegant way of doing that.

In 1960, R.E. Kalman published his famous paper describing a recursive solution to the discrete-data linear filtering problem.

The Kalman filter uses hypotheses of linearity (in the state transition equation, and the observation equation) and Normal distributions.
Exercise: Given

\[
\begin{align*}
p(s_1) &= \mathcal{N}(s_1; \mu_1, \sigma^2_1) \quad \text{(initial state pdf)} \\
p(s_j | s_{j-1}) &= \mathcal{N}(s_j; \alpha s_{j-1}, \sigma^2) \quad \text{(state transition pdf)} \\
p(o_j | s_j) &= \mathcal{N}(o_j; s_j, \sigma^2_o) \quad \text{(observation pdf)}
\end{align*}
\]

with the notation \(\mathcal{N}(x; \mu, \sigma^2)\) corresponding to the normal pdf for r.v. \(x\) with mean \(\mu\) and variance \(\sigma^2\). Show by induction that \(p(s_j | o_{1:j})\) is a normal distribution \(\mathcal{N}(s_j; \mu_j, \sigma^2_j)\) and give the recursive formula to compute parameters \((\mu_j, \sigma^2_j)\) assuming \(p(s_{j-1} | o_{1:j-1}) = \mathcal{N}(s_{j-1}; \mu_{j-1}, \sigma^2_{j-1})\).
Kalman Filter III

Answer.

1. at \( j = 1 \), we know the initial state \( s_1 \) is normally distributed (given in the hypothesis).

2. assuming at \( j - 1 \) that \( p(s_{j-1}|o_{1:j-1}) = \mathcal{N}(s_{j-1}; \mu_{j-1}, \sigma^2_{j-1}) \), then solving the prediction and update steps of filtering, we find that:

\[
\begin{align*}
\mu_j &= \alpha (1 - K_j) \mu_{j-1} + K_j o_j \\
\sigma^2_j &= \sigma^2_o K_j \\
K_j &= \frac{\sigma^2 + \alpha^2 \sigma^2_{j-1}}{\sigma^2 + \sigma^2_o + \alpha^2 \sigma^2_{j-1}}
\end{align*}
\]
the prediction step corresponds to (using result $\mathcal{N}_A|\mathcal{N}_B$):

$$p(s_j|o_{1:j-1}) = \int \mathcal{N}(s_j; \alpha s_{j-1}, \sigma^2) \mathcal{N}(s_{j-1}; \mu_{j-1}, \sigma^2_{j-1}) \, ds_{j-1}$$

$$= \frac{1}{|\alpha|} \int \mathcal{N} \left( s_{j-1}; \frac{s_j}{\alpha}, \frac{\sigma^2}{\alpha^2} \right) \mathcal{N}(s_{j-1}; \mu_{j-1}, \sigma^2_{j-1}) \, ds_{j-1}$$

$$= \frac{1}{|\alpha|} \mathcal{N} \left( 0; \frac{s_j}{\alpha} - \mu_{j-1}, \frac{\sigma^2}{\alpha^2} + \sigma^2_{j-1} \right)$$

$$= \mathcal{N}(s_j; \alpha \mu_{j-1}, \sigma^2 + \alpha^2 \sigma^2_{j-1})$$

update (rewriting is required to be able to identify $(\mu_j, \sigma^2_j)$):

$$p(s_j|o_{1:j}) \propto \mathcal{N}(o_j; s_j, \sigma^2_o) \mathcal{N}(s_j; \alpha \mu_{j-1}, \sigma^2 + \alpha^2 \sigma^2_{j-1})$$

$$= \mathcal{N}(s_j; \mu_j, \sigma^2_j)$$
These two parameters \((\mu_j, \sigma^2_j)\) are enough to characterise fully the Normal distribution \(p(s_j|o_{1:j})\).

The observations available \(\{o_j^{(1)}\}\) are plugged in the expression \((\mu_j, \sigma^2_j)\) so that they can be computed iteratively.
State-Space Model I

In the Multivariate case:

- **State equation** with $\vec{s}_j$ a random vector
  \[ \vec{s}_j = F \vec{s}_{j-1} + G \vec{\epsilon}_j \]
  
  $F$ is a square matrix, $G$ is a matrix, and $\vec{\epsilon}_j \sim \mathcal{N}(0, \Sigma)$ is a random vector.

- **Observation equation**
  \[ \vec{o}_j = H \vec{s}_j + \vec{\epsilon}_j \]
  
  with $H$ a matrix and $\vec{\epsilon}_j \sim \mathcal{N}(0, \Sigma_o)$. 
State-Space Model II

Given \( p(\vec{s}_{j-1}|\vec{o}_{1:j-1}) = \mathcal{N}(\vec{s}_{j-1}; \mu_{j-1}, \Sigma_{j-1}) \)

- **prediction**
  \[
p(\vec{s}_{j}|\vec{o}_{1:j-1}) = \mathcal{N}(\vec{s}_{j}; F\mu_{j-1}, R = F\Sigma_{j-1}F^T + G\Sigma G^T)\]

- **update** \( p(\vec{s}_{j}|\vec{o}_{1:j}) = \mathcal{N}(\vec{s}_{j}; \mu_{j}, \Sigma_{j}) \) with
  \[
  \Sigma_{j} = (H^T\Sigma_o^{-1}H + R^{-1})^{-1}
  \]
  and
  \[
  \mu_{j} = \Sigma_{j}(H^T\Sigma_{o}^{-1}\vec{o}_{j} + R^{-1}F\mu_{j-1})
  \]
Consider the following model

1. Observation equation:

\[ o_j = s_j + e_j \]

2. State equation

\[ AR(3) : \quad s_j = \phi_1 s_{j-1} + \phi_2 s_{j-2} + \phi_3 s_{j-3} + \epsilon_j \]

Define \( s_j \), \( o_j \), \( H \), \( F \) and \( G \).
State-Space Model IV

**Answer:** Using Matrix notation, the State equation is:

\[
\begin{pmatrix}
    s_j \\
    s_{j-1} \\
    s_{j-2} \\
\end{pmatrix} =
\begin{pmatrix}
    \phi_1 & \phi_2 & \phi_3 \\
    1 & 0 & 0 \\
    0 & 1 & 0 \\
\end{pmatrix}
\begin{pmatrix}
    s_{j-1} \\
    s_{j-2} \\
    s_{j-3} \\
\end{pmatrix} +
\begin{pmatrix}
    \epsilon_j \\
    0 \\
    0 \\
\end{pmatrix}
\]

and the observation equation is:

\[
o_j = (1, 0, 0)
\begin{pmatrix}
    s_j \\
    s_{j-1} \\
    s_{j-2} \\
\end{pmatrix} + e_j
\]

```
State-Space Model V

**Exercises:** Find $F$ and $G$ when $s_j$ follows

1. an MA(2)
2. an ARMA(2,2)
State-Space Model VI

- F, G and H are assumed to be known and constant over time.
- More general state space models allow these matrices to depend on time.
Founders: D. G. Krige and G. Matheron formulated the theory of geostatistics and Kriging in the 1950’s.

Application to mining industry

We want to predict at unobserved locations.

Geostatistics is the study of spatial data where the spatial correlation is modeled through the variogram. Focus of geostatistical analyses is on predicting responses at unsampled sites.
Introduction to geostatistics II

Applications:

- hydrological data,
- mining applications,
- air quality studies
- soil science data
- biological applications
- economic housing data
- etc.

Introduction to geostatistics III

Let's consider:

- $J$ physical locations $\{x_j\}_{j=1,\ldots,J}$,

- Some information of interest (e.g. radioactivity levels) is modeled as a stochastic process at these locations $\{s_j = s(x_j)\}_{j=1,\ldots,J}$,

- One observation (or measurement) is available for each site $\{s_j^{(1)}\}_{j=1,\ldots,J}$.

- At a new location $x_0$, we want to predict $s_0$. 
Content:

1. Adhoc methods: spatial interpolation

2. Statistical modeling with Kriging
   1. simple, ordinary and universal Kriging
   2. modelling with covariance and/or variogram
   3. Stationarity assumptions
Spatial interpolation I

Prediction of $s$ at a new site $x_0$ can be expressed as a weighted averages of data:

$$s(x_0) = \sum_{j=1}^{J} \lambda_j \ s(x_j)$$

with the constraints:

$$(\lambda_j \geq 0, \ \forall j) \ \land \ \left( \sum_{j=1}^{J} \lambda_j = 1 \right)$$
Spatial interpolation II

1. Thiessen polygons (Voronoi polygons, Dirichlet tessellations)
2. Triangulation
3. Natural Neighbour interpolation
4. Inverse function of distance
5. Trend surface
1 Thiessen polygons (Voronoi polygons, Dirichlet tessellations)

\[ s(x_0) = s(x_j) \quad \text{with} \quad x_j = \arg \min_{i=1, \ldots, J} \| x_i - x_0 \| \]

so we have binary weights:

\[ \lambda_j = 1, \quad \text{and} \quad \lambda_i = 0 \quad \forall i \neq j \]
Spatial interpolation IV

Figure: Voronoi polygons
Spatial interpolation V

**Triangulation.** Sampling points are linked to their neighbours by straight lines to create triangles that do not contain any of the points. Having a new position \( x_0 = (u_0, v_0) \) in one of the triangle, let says the one defined by \((x_1, x_2, x_3)\), then

\[
\lambda_1 = \frac{|x_0 - x_3 ; x_2 - x_3|}{|x_1 - x_3 ; x_2 - x_3|}
\]

with the notation

\[
|x_0 - x_3 ; x_2 - x_3| = \begin{vmatrix} u_0 - u_3 & u_2 - u_3 \\ v_0 - v_3 & v_2 - v_3 \end{vmatrix} = \det \begin{bmatrix} u_0 - u_3 & u_2 - u_3 \\ v_0 - v_3 & v_2 - v_3 \end{bmatrix}
\]

\(\lambda_2\) and \(\lambda_3\) are defined in a similar fashion and all the other \(\lambda\)s are 0s. Unlike Thiessen method, the resulting surface is continuous but yet has abrupt changes in gradient at the margins of the triangles.
Spatial interpolation VI

Figure: Triangulation: The weights $\lambda_1$ corresponds to the blue area divided by the area of the triangle $(x_1, x_2, x_3)$. Similarly $\lambda_2$ corresponds to the green area divided by the area of the triangle $(x_1, x_2, x_3)$ and $\lambda_3$ corresponds to the pink area divided by the area of the triangle $(x_1, x_2, x_3)$. 
Spatial interpolation VII

3 Natural Neighbour interpolation

![Graph showing Natural Neighbour interpolation](image-url)
Spatial interpolation VIII
Spatial interpolation IX
Spatial interpolation X

The diagram illustrates a spatial interpolation process with multiple regions and labeled vertices. The regions are marked as A1, A2, A3, A4, and A5. The vertices are labeled as a1, a2, a3, a4, and a5. The coordinates are indicated along the x and y axes.
Spatial interpolation XI

\[ \lambda_j = \frac{a_j}{\sum_{j=1}^{J} a_j} \]

with \( a_j = 0 \) if \( x_j \) is not a natural neighbour to \( x_0 \).
Inverse function of distance

\[ \lambda_j \propto \frac{1}{\|x_j - x_0\|^\beta}, \quad \beta > 0 \]

- The weights \( \{\lambda_j\}_{j=1,\ldots,J} \) are scaled such that they sum up to 1.
- Usually, \( \beta = 2 \) (Euclidian distance).
- If \( x_0 = x_j \), then \( s(x_0) = s(x_j) \).
- There are no discontinuities in the map \( s \).
- There is no measure of the error.
Trend Surface. This method proposes to do regression:

\[ s(x) = \mu(x) + \epsilon \]

with the error term \( \epsilon \sim \mathcal{N}(0, \sigma^2) \). The function \( \mu \) is a parametric function such as planes or quadratics e.g.

\[ \mu(x = (u, v)) = b_0 + b_1 u + b_2 v \]

Coefficients \( \mathbf{b} = (b_0, b_1, b_2)^T \) can then be estimated by Least Squares using the \( J \) observations.

Once \( \hat{\mathbf{b}} \) is estimated, the prediction at the new location \( x_0 \) is computed by:

\[ \hat{s}(x_0) = \hat{b}_0 + \hat{b}_1 u_0 + \hat{b}_2 v_0 \]
Spatial interpolation XIV

Limits of Interpolation for prediction:

- Some interpolators give a crude prediction and the spatial variation is displayed poorly.
- The interpolators fail to provide any estimates of the error on the prediction.
- With the exception of trend surface, these methods were deterministic. However the processes are stochastic by nature.
- In practice the modelling with trend surface is too simplistic to perform well and the uncertainty is the same everywhere.
Introduction: Kriging I

- The aim of Kriging is to estimate the value of a random variable \( s \) at one or more unsampled points or locations, from more or less sparse sample data on a given support say \( \{s(x_1), \cdots, s(x_J)\} \) at \( \{x_1, \cdots, x_J\} \).

- Different kinds of kriging methods exist, which pertains to the assumptions about the mean structure of the model:

\[
\mathbb{E}[s(x)] = \mu(x) \quad \text{or} \quad \mathbb{E}[s(x) - \mu(x)] = 0
\]

\[\tilde{s}(x)\]
Introduction: Kriging II

Different Kriging methods:

- **Ordinary Kriging** :
  \[ \mathbb{E}[s(x)] = \mu \quad (\mu \text{ is unknown}) \]

- **Simple Kriging** :
  \[ \mathbb{E}[s(x)] = \mu \quad (\mu \text{ is known}) \]

- **Universal Kriging** : the mean is unknown and depends on a linear model:
  \[ \mu(x) = \sum_{p=0}^{P} \beta_p \phi_p(x) \]

  and coefficients \( \{\beta_p\} \) need to be estimated.
Ordinary kriging is the most common type of kriging.

The underlying model assumption in ordinary kriging is:

$$\mathbb{E}[s(x)] = \mu$$

with $\mu$ unknown.

The stochastic process $s$ has been observed at $J$ sites (the r.v. $s(x_j) = s_j$ has one observation $s_j^{(1)}$ associated with it).
Ordinary kriging II

The model for $s(x_0)$ is:

$$s(x_0) - \mu = \sum_{j=1}^{J} \lambda_j (s(x_j) - \mu) + \epsilon(x_0)$$

or

$$s(x_0) = \sum_{j=1}^{J} \lambda_j s(x_j) + \mu (1 - \sum_{j=1}^{J} \lambda_j) + \epsilon(x_0)$$

We filter the unknown mean by requiring that the kriging weights sum to 1, leading to the ordinary kriging estimator:

$$s(x_0) = \sum_{j=1}^{J} \lambda_j s(x_j) + \epsilon(x_0) \quad \text{subject to } \sum_{j=1}^{J} \lambda_j = 1$$
Ordinary kriging III

- $\epsilon(x_0)$ is the noise at position $x_0$ such that:

$$\mathbb{E}[\epsilon(x_0)] = 0$$

- We want to estimate $\hat{s}(x_0)$. In other words we need to get the appropriate $\{\lambda_j\}_{j=1, \cdots, J}$.

- Estimation by Mean square errors subject to a constraint:

$$(\hat{\lambda}_1, \cdots, \hat{\lambda}_J) = \arg \min_{\lambda_1, \cdots, \lambda_J} \{ \mathbb{E}[\epsilon(x_0)^2] \} \quad \text{subject to} \quad \sum_{j=1}^J \lambda_j = 1$$
Ordinary kriging IV

This is solved using Lagrange multipliers. We define the energy \( J \) that depends both on \( \{\lambda_j\}_{j=1}^J \) and \( \psi \):

\[
(\{\hat{\lambda}_j\}_{j=1}^J, \psi) = \arg \min_{\psi, \lambda_1, \ldots, \lambda_J} \left\{ J(\lambda_1, \ldots, \lambda_J, \psi) = \mathbb{E}[\epsilon(x_0)^2] + 2\psi \left( \sum_{j=1}^J \lambda_j - 1 \right) \right\}
\]
First we express the expectation of the error:

\[ \mathbb{E}[\epsilon(x_0)^2] = \mathbb{E} \left[ \left( s(x_0) - \sum_{j=1}^{J} \lambda_j s(x_j) \right)^2 \right] \]

\[ = \mathbb{E} \left[ \left( s(x_0) - \mu + \mu - \sum_{j=1}^{J} \lambda_j s(x_j) \right)^2 \right] \]

\[ = \mathbb{E} \left[ (s(x_0) - \mu)^2 \right] - 2 \sum_{j=1}^{J} \lambda_j \mathbb{E} [(s(x_0) - \mu)(s(x_j) - \mu)] + \sum_{i=1}^{J} \sum_{j=1}^{J} \lambda_i \lambda_j \mathbb{E} [(s(x_j) - \mu)(s(x_i) - \mu)] \]

Remember that the covariance is defined as

\[ \text{Cov}(s(x_j); s(x_i)) = c_{ij} = \mathbb{E} [(s(x_j) - \mu)(s(x_i) - \mu)] \]

So the energy to minimize:

\[ J(\lambda_1, \cdots, \lambda_J, \psi) = c_{00} - 2 \sum_{j=1}^{J} \lambda_j c_{0j} + \sum_{i=1}^{J} \sum_{j=1}^{J} \lambda_i \lambda_j c_{ij} + 2 \psi \left( \sum_{j=1}^{J} \lambda_j - 1 \right) \]
Second, we differentiate $J$ w.r.t. $\lambda_k$, $k = 1, \cdots, J$ and $\psi$, and the minimum of $J$ is found when all the derivatives are equal to zeros.

$$\begin{cases}
\frac{\partial J}{\partial \psi} = 0 \\
\frac{\partial J}{\partial \lambda_k} = 0, \quad \forall k = 1, \cdots, J
\end{cases}$$

The derivative w.r.t. $\psi$ is:

$$\frac{\partial J}{\partial \psi} = \sum_{j=1}^{J} \lambda_j - 1 = 0$$

The derivative w.r.t. $\lambda_k$ is:

$$\frac{\partial J}{\partial \lambda_k} = 2 \psi - 2 c_{0k} + 2 \sum_{j=1}^{J} \lambda_j c_{ik} = 0$$
The solution is:

\[
\begin{bmatrix}
c_{11} & \cdots & c_{1J} & 1 \\
\vdots & & 1 & \\
c_{J1} & \cdots & c_{JJ} & 1 \\
1 & \cdots & 1 & 0
\end{bmatrix}
\begin{pmatrix}
\hat{\lambda}_1 \\
\vdots \\
\hat{\lambda}_J \\
\hat{\psi}
\end{pmatrix}
= 
\begin{pmatrix}
c_{10} \\
\vdots \\
c_{J0} \\
1
\end{pmatrix}
\]

\[A^{-1}b\]

or

\[\hat{\lambda} = A^{-1}b\]
Once you have the estimate $\hat{\lambda}$, then you can predict (using the observations):

$$\hat{s}(x_0) = \sum_{j=1}^{J} \hat{\lambda}_j s_j^{(1)}$$
Simple Kriging I

Assumption for Simple Kriging:

- The mean $\mathbb{E}[s(x)] = \mu$ is known.

We estimate $s(x_0)$ using the relation (same as Ordinary Kriging):

\[
s(x_0) - \mu = \sum_{j=1}^{J} \lambda_j (s(x_j) - \mu) + \epsilon(x_0)
\]

or

\[
s(x_0) = \sum_{j=1}^{J} \lambda_j s(x_j) + \mu \left(1 - \sum_{j=1}^{J} \lambda_j \right) + \epsilon(x_0)
\]

where $\mu$ is known. The $\lambda_j$ do not need to be constrained to sum to 1 anymore and the second term insured that $\mathbb{E}[s(x)] = \mu$, $\forall x$. 
Simple Kriging II

The hypothesis for the error is $\mathbb{E} [\epsilon(x_0)] = 0$ and we estimate $\{\lambda_j\}_{j=1,\ldots,J}$ such that the Mean Square Error $\mathbb{E}[\epsilon^2(x_0)]$ is minimised.

The solution is then:

$$
\begin{pmatrix}
\hat{\lambda}_1 \\
\vdots \\
\hat{\lambda}_J
\end{pmatrix} = 
\begin{bmatrix}
c_{11} & \cdots & c_{1J} \\
\vdots & \ddots & \vdots \\
c_{J1} & \cdots & c_{JJ}
\end{bmatrix}^{-1}
\begin{pmatrix}
c_{10} \\
\vdots \\
c_{J0}
\end{pmatrix}
$$

Once you have the estimate $\hat{\lambda}$, then you can predict (using the observations):

$$
\hat{s}(x_0) = \sum_{j=1}^{J} \hat{\lambda}_j s_j^{(1)} + \mu \left( 1 - \sum_{j=1}^{J} \hat{\lambda}_j \right)
$$
Universal Kriging I

For a new location $x_0$, we have the following model

$$s(x_0) - \mu(x_0) = \sum_{j=1}^{J} \lambda_j \left( s(x_j) - \mu(x_j) \right) + \epsilon_{x_0}$$

or

$$s(x_0) = \sum_{j=1}^{J} \lambda_j \hspace{1mm} s(x_j) + \mu(x_0) - \sum_{j=1}^{J} \lambda_j \hspace{1mm} \mu(x_j) + \epsilon_{x_0}$$

In the Universal Kriging, the mean of $s$ depends on the position $x$:

$$\mu(x) = \sum_{p=0}^{P} \beta_p \hspace{1mm} \phi_p(x)$$
Example of choice of the functions \( \{\phi_p\}_{p=1,\ldots,P} \) of \( x = (u, v) \in \mathbb{R}^2 \):

- Linear trend \((P = 2)\):
  \[
  \phi_0(x) = 1, \quad \phi_1(x) = u, \quad \phi_2(x) = v
  \]

- Quadratic trend \((P = 5)\):
  \[
  \phi_3(x) = u^2, \quad \phi_4(x) = uv, \quad \phi_5(x) = v^2
  \]
In a similar fashion as ordinary kriging (we don't know the $\beta_p$, so $\mu$ is unknown), we rewrite:

$$s(x_0) = \sum_{j=1}^{J} \lambda_j s(x_j) + \mu(x_0) - \sum_{j=1}^{J} \lambda_j \mu(x_j) + \epsilon_{x_0}$$

as

$$s(x_0) = \sum_{j=1}^{J} \lambda_j s(x_j) + \epsilon_{x_0} \quad \text{subject to} \quad \mu(x_0) - \sum_{j=1}^{J} \lambda_j \mu(x_j) = 0$$

constraint
Universal Kriging IV

Having $\mu(x) = \sum_{p=0}^{P} \beta_p \phi_p(x)$, the constraint is equivalent to:

$$\sum_{p=0}^{P} \beta_p \phi_p(x_0) = \sum_{p=0}^{P} \beta_p \sum_{j=1}^{J} \lambda_j \phi_p(x_j)$$

This is true for any combination of $\beta_p$. Hence we have in fact $P + 1$ constraints:

$$\left( \phi_p(x_0) = \sum_{j=1}^{J} \lambda_j \phi_p(x_j) \right) \quad \forall p = 0, \cdots, P$$
Note that at $p = 0$, using $\phi_0(x) = 1$, we recover the constraint $\sum_{j=1}^J \lambda_j = 1$.

This minimisation is solved by introducing $P + 1$ Lagrange multipliers:

$$(\hat{\lambda}_1, \ldots, \hat{\lambda}_J, \hat{m}_0, \ldots, \hat{m}_P) = \text{arg min} \left\{ \mathbb{E}[\epsilon_{x_0}^2] + \sum_{p=0}^P m_p \left( \phi_p(x_0) - \sum_{j=1}^J \lambda_j \phi_p(x_j) \right) \right\}$$
The solution of Universal Kriging is:

\[
\begin{bmatrix}
C & F^T \\
F & 0
\end{bmatrix}
\begin{pmatrix}
\hat{\lambda}_1 \\
\vdots \\
\hat{\lambda}_J \\
\hat{m}_0 \\
\hat{m}_1 \\
\vdots \\
\hat{m}_P
\end{pmatrix}
= 
\begin{pmatrix}
c_{01} \\
\vdots \\
c_{0J} \\
\phi_0(x_0) \\
\phi_1(x_0) \\
\vdots \\
\phi_P(x_0)
\end{pmatrix}
\]
Universal Kriging VII

with

\[
F = \begin{bmatrix}
\phi_0(x_1) & \phi_0(x_2) & \cdots & \phi_0(x_J) \\
\phi_1(x_1) & \phi_1(x_2) & \cdots & \phi_1(x_J) \\
\phi_2(x_1) & \phi_2(x_2) & \cdots & \phi_2(x_J) \\
\vdots & & & \\
\phi_P(x_1) & \phi_P(x_2) & \cdots & \phi_P(x_J)
\end{bmatrix}
\]

and

\[
C = \begin{bmatrix}
c_{11} & \cdots & c_{1J} \\
\vdots & & \vdots \\
c_{J1} & \cdots & c_{JJ}
\end{bmatrix}
\]

The prediction at \( x_0 \) is computed by:

\[
\hat{s}_0 = \sum_{j=1}^{J} \hat{\lambda}_j \, s_j^{(1)}
\]
Remarks about Kriging I

- The covariance function $C[s(x), s(x')]$ is a function assumed to be known in all the solutions proposed here for Kriging.

- In practice such a function is not known and need to be estimated.

- Some hypotheses will be used about the process to ease this estimation.

- We define the concept of variogram as an alternative to covariance and we see next what are the assumptions about the process that will be used for Kriging.
Remarks about Kriging II

Definition (variogram)

The variogram is defined as:

\[
\gamma(s(x_i), s(x_j)) = \gamma_{ij} = \frac{1}{2} \mathbb{E}[(s(x_i) - s(x_j))^2]
\]

When \( \mathbb{E}[(s(x_i) - s(x_j))] = 0 \) then the variogram is linked to the covariance as follow:

\[
\gamma_{ij} = \frac{1}{2} (c_{ii} + c_{jj} - 2 c_{ij})
\]
Definition (Strict Stationarity)

The distribution of the random process has certain attributes that are the same everywhere. **Strict stationarity** indicates that for any number \( k \) of any sites \( x_1, x_2, \cdots, x_k \), the joint cumulative distribution of \((s(x_1), \cdots, s(x_k))\) remains the same under an arbitrary translation \( h \):

\[
P(s(x_1), \cdots, s(x_k)) = P(s(x_1 + h), \cdots, s(x_k + h))
\]

If the random process is strictly stationary, its moments if they exist are also invariant under translations.
Stationarity II

Definition (weak stationarity)

A process $s(x)$ is said second order stationary (or weakly stationary, or wide-sense stationary) when

1. the mean of the process does not depend on $x$: $\mathbb{E}[s(x)] = \mu$

2. the variance of the process does not depend on $x$: $\mathbb{E}[(s(x) - \mu)^2] = \sigma^2$

3. the covariance between $s(x)$ and $s(x+h)$ only depends on $h$:

$$\text{Cov} [s(x), s(x+h)] = \mathbb{E} [(s(x) - \mathbb{E}[s(x)]) (s(x+h) - \mathbb{E}[s(x+h)])]$$

$$= \mathbb{E} [(s(x) - \mu) (s(x+h) - \mu)]$$

$$= C(h)$$

Note that $C(0) = \sigma^2$. 
For a weakly stationary process, the autocorrelation can also be defined as:

$$\rho(h) = \frac{C(h)}{C(0)}$$

with $C(0)$ is the covariance at lag 0, i.e. $\sigma^2$.

If a random field with the function $C(h)$ only dependent on the distance $||h||$ and not on its orientation, it is said to be isotropic.
Definition (Intrinsic Stationarity)

$s(x)$ is said to be an intrinsic random function such that:

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

and

$$\text{Var}[s(x + h) - s(x)] = 2 \gamma(h)$$

The function $\gamma(h)$ is called the variogram.
Stationarity V

**Exercises:** For second-order stationary processes,

1. Express $\gamma(h)$ w.r.t. $C(h)$.

2. Express $\gamma(h)$ w.r.t. $\rho(h)$.

3. Show that a process that is weakly stationary is also intrinsically stationary.
Brownian Motion & Ornstein-Uhlenbeck Processes

- A second order stationary process is also an intrinsic stationary process.
- But an intrinsic stationary process is not always a second order stationary process or a strictly stationary process.
- To illustrate the weakly stationary and intrinsic stationary, we look at the following processes:
  - Brownian Motion,
  - Ornstein-Uhlenbeck Process.
Brownian Motion I

Definition (diffusion & Brownian Motion)

A diffusion is a continuous time Stochastic Process \( s(t) \) with the following properties:

1. \( s(0) = 0 \),
2. \( s(t) \) has independent increments,
3. \( P(s(t_2) \mid s(t_1)) \) has a density function \( f(s(t_2) \mid t_1, t_2, s(t_1)) \).

Standard Brownian motion is a diffusion \( s(t) \), \( t \geq 0 \) satisfying the following:

1. \( s(0) = 0 \).
2. \( s(t) \) has independent increments.
3. For \( t_2 > t_1 \), \( s(t_2) - s(t_1) \sim \mathcal{N}(0, \sigma^2(t_2 - t_1)) \).
Brownian Motion II

- $s(t)$ has independent increments means that for all times $0 \leq t_1 \leq t_2 \cdots \leq t_n$ the increments $s(t_n) - s(t_{n-1})$, $s(t_{n-1}) - s(t_{n-2})$, $\cdots$, $s(t_2) - s(t_1)$ are independent random variables.

- For $t_2 > t_1$, $s(t_2) - s(t_1) \sim \mathcal{N}(0, \sigma^2(t_2 - t_1))$ is equivalent to

$$s(t_2) \mid s(t_1) \sim \mathcal{N}(s(t_1), \sigma^2(t_2 - t_1))$$

$s(t_2)$ given $s(t_1)$ is normally distributed with mean $s(t_1)$ and variance $\sigma^2(t_2 - t_1)$. 
Brownian Motion III

Exercises:

1. Show that a standard brownian motion is intrinsically stationary

2. Show that a standard brownian motion is not weakly stationary
Ornstein-Uhlenbeck Process I

Definition (Ornstein-Uhlenbeck)

Let \( s(t) \) be standard Brownian motion. The process

\[
V(t) = e^{-t} s(e^{2t})
\]

is called the Ornstein-Uhlenbeck process.
Show that $V(t)$ is intrinsically stationary and weakly stationary.
Brownian Motion and Ornstein-Uhlenbeck processes

Historical remarks:

- Robert Brown observes ceaseless irregular motion of small particles in a fluid. Motion explained by believing particles to be alive (1827-1829).
- Goul puts forward a kinetic theory to explain the motion; it is due to rapid bombardment of a huge number of fluid molecules (1860).
- Einstein presents the theory of “Brownian motion” (c. 1900). At the same time (c. 1900), Bachelier defines it to model stock options.
- The Ornstein-Uhlenbeck process was proposed by Uhlenbeck and Ornstein (1930) in a physical modelling context, as an alternative to Brownian Motion. The model has been used since in a wide variety of applications areas e.g. in finance (see Vasicek (1977) interest rate model).
Kriging formulation with the variogram I

In Kriging, we have

\[
\begin{align*}
\underbrace{s(x_0) - \mu(x_0)}_{\tilde{s}(x_0)} &= \sum_{j=1}^{J} \lambda_j \underbrace{(s(x_j) - \mu(x_j))}_{\tilde{s}(x_j)} + \epsilon(x_0) \\
\epsilon(x_0) &= \tilde{s}(x_0) - \sum_{j=1}^{J} \lambda_j \tilde{s}(x_j)
\end{align*}
\]

with the assumption \( \mathbb{E}[\epsilon(x_0)] = 0. \)
Kriging formulation with the variogram II

The estimation is performed by:

\[(\hat{\lambda}_1, \cdots, \hat{\lambda}_J) = \arg \min E[\epsilon(x_0)^2]\]

solved

- without constraint: Simple Kriging
- with constraints: Universal/Ordinary Kriging

and the estimate is computed using the observations:

\[
\hat{s}_0 = \sum_{j=1}^{J} \hat{\lambda}_j s_j^{(1)}
\]

Ordinary/Universal Kriging

or

\[
\hat{s}_0 = \sum_{j=1}^{J} \hat{\lambda}_j s_j^{(1)} + \mu \left(1 - \sum_{j=1}^{J} \hat{\lambda}_j \right)
\]

Simple Kriging
Exercise: With the constraint \( \sum_{j=1}^{J} \lambda_j = 1 \) (true in Universal/Ordinary Kriging):

1. Show that:

\[
\left( \tilde{s}_0 - \sum_{j=1}^{J} \lambda_j \tilde{s}_j \right)^2 = -\frac{1}{2} \sum_{j=1}^{J} \sum_{i=1}^{J} \lambda_i \lambda_j (\tilde{s}_i - \tilde{s}_j)^2 + \sum_{j=1}^{J} \lambda_j (\tilde{s}_0 - \tilde{s}_j)^2
\]

2. Show that

\[
\mathbb{E}[\epsilon(x_0)^2] = -\sum_{j=1}^{J} \sum_{i=1}^{J} \lambda_i \lambda_j \gamma_{ij} + 2 \sum_{j=1}^{J} \lambda_j \gamma_{0j}
\]

3. Deduce what are the assumptions of stationarity (weak or intrinsic) that will be used for Simple, Ordinary and Universal Kriging.
Kriging formulation with the variogram IV

**Definition (Kriging variance)**

The minimised mean-squared prediction error $\mathbb{E}[\epsilon(x_0)^2]$ is sometimes called the Kriging variance:

$$
\sigma_0^2 = 2 \sum_{j=1}^{J} \hat{\lambda}_j \gamma_{0j} - \sum_{j=1}^{J} \sum_{i=1}^{J} \hat{\lambda}_i \hat{\lambda}_j \gamma_{ij} \quad \text{(Or./Un. Kriging)}
$$

or

$$
\sigma_0^2 = c_{00} - 2 \sum_{j=1}^{J} \hat{\lambda}_j c_{0j} + \sum_{i=1}^{J} \sum_{j=1}^{J} \hat{\lambda}_i \hat{\lambda}_j c_{ij} \quad \text{(Simple Kriging)}
$$

so prediction interval can be constructed: $\hat{s}_0 \pm 2\sigma_0$. 

Kriging formulation with the variogram $V$

<table>
<thead>
<tr>
<th>Kriging</th>
<th>Hypotheses $\mu(x)$</th>
<th>Constraints</th>
<th>Prediction at $x_0$</th>
<th>Variance</th>
<th>Stationarity?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ordinary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Universal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We have presented the Kriging approaches to perform prediction. These approaches require the knowledge of the covariance function or variogram of the stochastic process $\tilde{s}(x) = s(x) - \mu(x)$.

**Definition (variogram cloud)**

For any set of data we can compute the semivariance for each pair of points $x_i$ and $x_j$ individually as

$$\hat{\gamma}(x_i, x_j) = \hat{\gamma}_{ij} = \frac{1}{2} (\tilde{s}(x_i) - \tilde{s}(x_j))^2$$

These values are plotted against the difference $h = x_i - x_j$ between the sites. When a variogram is isotropic, it depends only on the distance $||h||$. A scatterplot of this form is called the **variogram cloud**.
Exercises:

1. How many points does the variogram cloud contain if the separation distance \( h \) is not limited?

2. Consider \( x_1 = (2, 3), x_2 = (4, 4) \) and \( x_3 = (4, 3) \) where we have measured \( s_1 = 2, s_2 = 3 \) and \( s_3 = 2.4 \). Assuming the variogram is isotropic, plot the variogram cloud.
Definition (Average semi variance)

The definition of the semivariance is

\[ \gamma(h) = \frac{1}{2} \mathbb{E}[(s(x) - s(x + h))^2] \]

which can be estimated by:

\[ \hat{\gamma}(h) = \frac{1}{2m(h)} \sum_{i=1}^{m(h)} (s(x_i) - s(x_i + h))^2 \]

\(m(h)\) is the number of pairs of sites separated by the particular lag vector \(h\).

The way it is computed depends on the sampling design.
Example: In labs, we explore the first tutorial given with the R-package *gstat*. This tutorial looks at the Meuse data set that gives locations and topsoil heavy metal concentrations, collected in a flood plain of the river Meuse, near the village of Stein (NL). Heavy metal concentrations are from composite samples of an area of approximately 15 m x 15 m.
Modelling variogram V

Figure: Variogram cloud of $\log(zinc)$ in the Meuse data set.
Figure: Average semiVariogram of $\log(zinc)$ in the Meuse data set.
Consider the region $D \subset \mathbb{R}^2$, the location \( \{x_j\}_{j=1,\ldots,J} \) can be chosen by:

1. Systematic Sampling (e.g. grid)
2. Pure random sampling
3. Stratified sampling
Conclusion

- We have seen how to estimate a variogram using the observation we have for stochastic processes $\{s(x_j)\}_{j=1,\ldots,J}$ that are measured at each site $x_j$.

- Can we estimate the variogram $\gamma$ in this fashion when Universal Kriging will be used to perform prediction at a new position $x_0$?

- What is the stochastic process for which the variogram is used in Universal Kriging? Ordinary Kriging? Simple Kriging?
Variogram modelling for Universal Kriging I

- If \( \tilde{s}(x_i) - \tilde{s}(x_j) = s(x_i) - s(x_j) \) for simple and ordinary Kriging, this is not the case anymore for universal Kriging where \( \mathbb{E}[s(x)] = \mu(x) \).
- We have collected observations for \( s \) but not for \( \tilde{s} \) and we don't know the form of \( \mu(x) \).
- For instance, if we assume that \( s(x) = a \, x + b \) (linear drift), then the difference

\[
s(x + h) - s(x) = a \, h
\]

so the estimated variogram (variogram plot) will be affected by this drift and looks like a parabola:

\[
\hat{\gamma}(h) = \frac{1}{2} (a \, h)^2
\]
The first method proposes to get a quick estimate of the mean to remove it:

1. Specify a linear model \( \mu(x) = \sum_{p=0}^{P} \beta_p \phi_p(x) \) on a chosen basis of functions \( \{\phi_p\}_{p=0,...,P} \).

2. Estimate \( \{\beta_p\}_{p=0,...,P} \) by Least Squares to give an estimate of the mean \( \hat{\mu}(x) = \sum_{p=0}^{P} \hat{\beta}_p \phi_p(x) \).

3. Compute the variogram of \( \tilde{s}(x) = s(x) - \hat{\mu}(x) \).
The other approach to estimate the variogram works as follow:

1. Choose a theoretic variogram function.

2. Specify a linear model \( \mu(x) = \sum_{p=0}^{P} \beta_p \phi_p(x) \) on a chosen basis of functions \( \{\phi_p\}_{p=0,\ldots,P} \).

3. Estimate \( \hat{\mu}(x) \) (i.e. find \( \{\hat{\beta}_p\}_{p=0,\ldots,P} \)) such that the residuals \( \tilde{s}(x) = s(x) - \hat{\mu}(x) \) have a variogram that matches the chosen theoretical one.
Several analytical functions have been proposed in the literature to model a variogram. They all follow a few basic rules.

**Limits on variogram function.**

- **Behaviour near the origin.** The way in which the variogram approaches the origin is determined by the continuity (or lack of continuity) of the variable $s(x)$ itself. The semivariance at $|h| = 0$ is by definition 0. It can happen however that experimental values give a positive value (positive intercept, nugget effect).

Near 0, the variogram can have a linear approach $\gamma(h) \simeq b|h|$ or quadratic $\gamma(h) \simeq b|h|^2$ as $|h| \to 0$. 
Variogram functions II

- **Behaviour towards infinity.** The variogram is constrained by:

  \[ \lim_{|h| \to \infty} \frac{\gamma(h)}{|h|^2} = 0 \quad \text{as} \quad |h| \to \infty \]

  If it does not then the process is not entirely random (and is not compatible with the intrinsic hypothesis).
Variogram functions III

Examples of models:

▶ Unbounded models.

\[ \gamma(h) = \omega |h|^\alpha \] \text{ for } 0 < \alpha < 2

\( \alpha = 1 \) the variogram is linear.

▶ Bounded models. Based on experience, bounded variation is more common than unbounded one.

▶ Bounded linear
▶ Spherical
▶ Exponential
▶ ...

...
Once we have computed the variogram plot, we select the best analytical (theoretic) variogram function by:

1. fitting the models to the observed variogram plot.
2. choosing the theoretic model that has the smallest RSS (Residual sum of square) and/or AIC (Akaike Information Criterion).

**Exercise:** Assuming that $s$ is weakly stationary, explain why the variogram cannot be unbounded.
We consider a time series where observations have been collected at regular interval of time. One of the earliest general types of models for time series was that they were a combination of

1. a **trend** describing a long term behaviour of the series (deterministic)
2. a **periodic** or **seasonal** component describing variation within a cycle (deterministic)
3. an **error** term describing random fluctuations of the series about the trend and seasonal components (stochastic).

\[
(s(t_j) = \mu(t_j) + \epsilon(t_j)) \equiv (s_j = \mu_j + \epsilon_j)
\]

where \( \mu \) captures the deterministic behaviour (while both \( s \) and \( \epsilon \) are stochastic).
**Introduction Spectral Analysis II**

The purpose of the trend term, is to describe the long-term features of the series (e.g. an overall tendency), it would usually correspond to some smooth slowly changing function.

By differencing the time series the trend can easily be removed. So now we assume that we have only a remaining seasonal pattern.

For certain series we can specify the periodicity of the seasonal component very accurately (e.g. in the case of economic or geophysical series which contain a strict 12-month cycle).

However in other series it may not be possible to specify the period of the cyclical term *a priori* or there may be several (non harmonically related) periodic terms present with unknown periods.
We need special techniques to detect these hidden periodicities. We can represent the seasonal component by a sum of sine and cosine terms (well suited explanatory variables to capture periodic information):

\[ s_j = \sum_p (A_p \cos(\omega_p j) + B_p \sin(\omega_p j)) + \epsilon_j \]

How to estimate \( \{(A_p, B_p, \omega_p)\} \)?
Periodogram I

Definition (Periodogram)

Given a time series \( s_1, s_2, \cdots, s_J \), the periodogram is defined by:

\[
I(\omega) = \frac{2}{J} \left\{ \left( \sum_{j=1}^{J} s_j \cos(\omega \times j) \right)^2 + \left( \sum_{j=1}^{J} s_j \sin(\omega \times j) \right)^2 \right\}
\]

The function \( I(\omega) \) is then computed for \( \omega_p = \frac{2\pi p}{J} \) with \( p = 0, 1, \cdots, \frac{J}{2} \).

The periodogram is also sometimes represented w.r.t. the frequency \( f = \frac{\omega}{2\pi} \).
This implies \( f \in \{ 0, \frac{1}{J}, \frac{2}{J}, \cdots, \frac{1}{2} \} \).
## Example (Pure sine time series)

Consider a noiseless time series:

\[ s_j = \cos\left(\frac{2 \times \pi \times 5}{J} \times j\right), \quad j = 1, \cdots, J = 100 \]

1. Compute numerically \( \omega = \frac{2\pi \times 5}{J} \).

2. Compute numerically the corresponding frequency \( f = \frac{\omega}{2\pi} \).

3. Look at the periodogram and conclude.
Figure: Pure sine time series $s_j = \cos \left( \frac{2\times\pi\times5}{J} \times j \right), \quad j = 1, \cdots, J = 100.$
Figure: Periodogram $I(f)$.
Example (Combination of pure sine time series)

Consider the time series:

\[ s_j = 2 \cos \left( \frac{2\pi \times 5}{J} \times j \right) + \sin \left( \frac{2\pi \times 10}{J} \times j \right), \quad j = 1, \ldots, J = 100 \]

1. Compute numerically \( \omega_1 = \frac{2\pi \times 5}{J} \) and \( \omega_2 = \frac{2\pi \times 10}{J} \).
2. Compute numerically the corresponding frequencies \( f_1 \) and \( f_2 \).
3. Look at the periodogram and conclude.
Figure: Combination of two pure sine time series.
Figure: Periodogram $I(f)$. 

Periodogram VII
Periodogram VIII

For a time series with a cyclical patterns:

\[ s_j = \sum_p (A_p \cos(\omega_p j) + B_p \sin(\omega_p j)) + \epsilon_j, \quad j = 1, \ldots, J \]

- By computing the periodogram, you can identify the most prevalent frequencies \( \{\omega_p\} \).
- How would you propose to estimate the corresponding coefficients \( \{A_p, B_p\} \) ?
Periodogram IX

- Let's define the complex vector:

\[
V_p = \begin{pmatrix}
\exp(-j\omega_p 1) \\
\exp(-j\omega_p 2) \\
\vdots \\
\exp(-j\omega_p J)
\end{pmatrix} = \begin{pmatrix}
\cos(-\omega_p 1) + j\sin(-\omega_p 1) \\
\cos(-\omega_p 2) + j\sin(-\omega_p 2) \\
\vdots \\
\cos(-\omega_p J) + j\sin(-\omega_p J)
\end{pmatrix}
\]

- You can check that the periodogram \( I(\omega_p) \) (with \( \bar{s} = [s_1, \cdots, s_J]^T \)):

\[
I(\omega_p) \propto \| \langle \bar{s} | V_p \rangle \|^2
\]

with \( \langle \cdot | \cdot \rangle \) the dot product and \( \| \cdot \| \) the euclidian distance.
Periodogram X

**Definition (Inner product)**

For complex vectors $x$ and $y$ of length $J$, the inner product is:

$$< x, y > = \sum_{n=1}^{J} x_n \overline{y}_n$$

where $\overline{y}_n$ is the complex conjugate of $y_n$.

- You can check also that ($\omega_p = \frac{2\pi p}{J}$):

$$< V_{p_1} | V_{p_2} >= J \cdot \delta_{p_1 p_2} = J \times \begin{cases} 0 & \text{if } p_1 \neq p_2 \\ 1 & \text{otherwise} \end{cases}$$

($\delta_{p_1 p_2}$ indicates the Kronecker delta).
The set of vectors \( \{ V_p \}_{p=0, \ldots, J-1} \) forms a basis of the space \( \mathbb{C}^J \). So any \( \vec{s} \in \mathbb{C}^J \) can be written exactly (note \( V_0 = V_J \))

\[
\vec{s} = \sum_{p=0}^{J-1} \theta_p V_p \quad \text{or} \quad \vec{s} = \sum_{p=1}^{J} \theta_p V_p
\]

Because here \( s \) is a time series of real numbers, then it can be shown that

\[
\theta_p = \overline{\theta}_{J-p} \quad \text{(complex conjugate)}
\]

so only half of the coefficients \( \{\theta_p\}_{p=0, \ldots, J/2} \) are needed.
Periodogram XII

Definition (Discrete Fourier Transform and Power spectrum)

The **discrete Fourier Transform** (DFT) of the time series $s_1, \cdots, s_J$ is defined:

$$\theta_p = \langle \bar{s} | V_p \rangle = \sum_{j=1}^{J} s_j \exp(-\omega_p j)$$

with $\omega_p = \frac{2\pi p}{J}$. \{\theta_p\}_{p=0,\ldots,J-1} are the Fourier coefficients of the time series $s$ and:

$$\bar{s} = \sum_{p=0}^{J} \theta_p V_p$$

The **power spectrum** is defined as $P(\omega_p) = \theta_p^2$ and the **periodogram** is therefore proportional to the power spectral density of a signal.
Periodogram XIII

Reading and using a periodogram

- High values of the periodogram indicate periodic patterns.

- Detect the main maximum (or maxima) and identify the corresponding frequency(ies) $f$.

- The period of the cyclical pattern is recovered by computing $1/f$. If the time series have been recorded every months, then the period is $1/f$ months.
Case study: Variable Star Brightness I

- The brightness or the magnitude of a particular star at midnight is observed on 600 consecutive nights
Two prominent peaks appear for $f_1 = 0.035$ and $f_2 = 0.04167$. 
Exercises:

1. What are the periods in days associated with $f_1$ and $f_2$?

2. What model would you suggest to fit to this time series?
Case study: Variable Star Brightness IV

Answers:

1. \( f_1 \) corresponds to 29 day period and \( f_2 \) corresponds to 24 day period.
2. model:

\[
s_t = \beta_0 + \beta_1 \cos(2\pi f_1 t) + \beta_2 \sin(2\pi f_1 t) + \beta_3 \cos(2\pi f_2 t) + \beta_4 \sin(2\pi f_2 t) + \epsilon_t
\]
Conclusion

- Remember that spectral analysis by discrete Fourier transform can be applied when observations have been collected at regular interval.

- Frequency domain analysis or spectral analysis has been found to be useful in (for instance):
  - Acoustic
  - communication engineering
  - geophysical science
  - biomedical science
The Fourier transform is an important tool to analyse signals, discrete or continuous.

Definition (Fourier transform)

The Fourier integral measures how much oscillations at the frequency $\omega$ there is the function $f$:

$$F(\omega) = \int_{-\infty}^{\infty} f(t) \exp(-\omega t) \, dt$$
Fourier transform II

Definition (Inverse Fourier transform)

The Inverse Fourier transform can be computed:

\[ f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega) \exp(i\omega t) \, d\omega \]

Exercises: Fourier transform

Compute the Fourier transforms of:

1. \( f(t) = \begin{cases} 1 & -T < t < T \\ 0 & \text{otherwise} \end{cases} \)

2. \( f(t) = \delta(t) \)

3. \( f(t) = \delta(t - t_0) \)
Fourier transform III

4. \( f(t) = \exp(-at) \, h(t) \) (\( h \) is the heaviside step function)

5. \( f(t) = \cos(\omega_0 t) \)

6. Show

   1. Linearity:
      \[
      a \, x(t) + b \, y(t) \leftrightarrow a \, X(\omega) + b \, Y(\omega)
      \]

   2. Time shifting:
      \[
      x(t - t_0) \leftrightarrow \exp(-i\omega t_0) \, X(\omega)
      \]
Fourier transform IV

Definition (Convolution)

The convolution of two functions $x(t)$ and $h(t)$ gives a function $y(t)$ such that:

$$y(t) = x(t) * h(t) = \int_{-\infty}^{+\infty} x(\tau) h(t - \tau) \, d\tau$$

Theorem (Convolution)

If $y(t) = x(t) * h(t)$ then:

$$Y(\omega) = X(\omega) \cdot H(\omega)$$

Prove this theorem.
Two-dimensional Fourier transform I

Definition (2D Fourier Transform)

\[ F(\omega_x, \omega_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \exp(-\omega_x x) \exp(-\omega_y y) \, dx \, dy \]

Remember that Spectral analysis is well suited for signal that is observed at regular interval. In particular, 2D Fourier Transform is suitable to analyse images:

Definition (image)

An image can be understood as a stochastic process \( s(x) \) that has been observed at \( J \) locations \( \{x_j\}_{j=1, \ldots, J} \) lying on a regular 2D grid. For grey level images, the state space of \( s \) is all integers between 0 (black) and 255 (white).
Two-dimensional Fourier transform II

Low pass filtering
Two-dimensional Fourier transform III

High pass filtering
Definition (functional data)

Functional data indicate that the collected (observed) data are observations of functions varying over a continuum.

Functional Data Analysis (FDA) aims at:

- representing data in ways that help further analysis.
- to display the data so as to highlight various characteristics
- studying important sources of patterns and variations among the data
Functional Data Analysis: Introduction III

Example: Height of Girls

Figure 1.1. The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement.
10 subjects took part in the study.

Each record has 31 discrete values of height that are not equally spaced.

There are uncertainty or noise in the collected height values (about 3mm).

The underlying process can be assumed to be a smooth function $\mu_j(t)$ for each subject $j$ taking part in the study and observations have been collected for the stochastic process

$$s_j(t) = \mu_j(t) + \epsilon_j(t)$$
For each record $j$, it is interesting to calculate an estimate of the function $\mu_j(t)$ (Linear Smoothing methods).

When having a family of functions $\{\mu_j(t)\}_{j=1,\ldots,J}$, it is interesting to investigate the variations of this family of functions (functional Principal Component Analysis).

For simplicity here, we are considering temporal stochastic processes. All methods can easily be extended to spatio-temporal stochastic processes.
We consider that we have observations \( \{(t^{(i)}, s^{(i)})\}_{i=1, \ldots, n} \)

**Definition (Linear smoothing)**

A **linear smoother** estimates the function value \( s(t) \) by a linear combination of the discrete observations:

\[
\hat{s}(t) = \sum_{i=1}^{n} \lambda_i(t) s^{(i)} = <\tilde{\lambda}(t)|\tilde{s}>
\]

The behaviour of the smoother at \( t \) is controlled by the weights \( \lambda_i(t) \).
Example of linear smoothing methods:

1. Kriging ✓
2. Regression on basis of functions
3. Nadaraya-Watson estimator

Note that these methods have different hypotheses concerning the nature of the noise $\epsilon$. 
A model can be proposed for the deterministic part of the process as follow:

$$\mu(t) = \sum_{k=0}^{K} \theta_k \phi_k(t) = \langle \vec{\phi}(t) | \vec{\theta} \rangle$$

The standard model for the error $\epsilon$ is to be $N(0, \sigma^2)$ and independent on time (or $\mathbb{E}[s(t)] = \mu(t)$ and $\mathbb{V}ar[s(t)] = \sigma^2$). Observations collected are $\{(t^{(i)}, s^{(i)})\}_{i=1,\ldots,n}$.
The coefficients \( \{\theta_k\}_{k=0, \ldots, K} \) are estimated using least squares:

\[
\hat{\theta} = (\Phi^T \Phi)^{-1} \Phi^T \bar{s}
\]

where \( \bar{s} = [s^{(i)}, s^{(2)}, \ldots, s^{(n)}]^T \), \( \Phi \) is the \( n \times (K + 1) \) matrix collecting the values \( \{\phi_k(t^{(i)})\} \). So the estimate for \( \mathbb{E}(s(t)) \) is:

\[
\hat{s}(t) = \langle \hat{\phi}(t) | \hat{\theta} \rangle = \underbrace{\hat{\phi}(t)^T (\Phi^T \Phi)^{-1} \Phi^T}_{\bar{S}(t)^T} \bar{s}
\]

**Exercise:** What are the basis functions \( \{\phi_k\}_{k=0, \ldots, K} \) that can be used?
Example: Vancouver precipitation data with 13 Fourier Bases.
Choice of Basis Expansions.

When performing least squares fitting of basis expansions \( \phi_0(t), \phi_1(t), \cdots \) is a basis system for \( s \). The choice of this basis system is important in particular if you want to explore time derivatives of \( s(t) \).
Linear Smoothing: Regression on basis of functions V

- Monomial Basis $\phi_0(t) = 1, \phi_1(t) = t, \phi_2(t) = t^2, \ldots, \phi_K(t) = t^K$

Properties

- difficult for $K > 6$
- Derivatives of $\phi_k(t)$ get simpler but this is often not a desirable property when fitting real-world data.
Linear Smoothing: Regression on basis of functions VI

$K = 0$  

$K = 1$

$K = 2$  

$K = 3$
Linear Smoothing: Regression on basis of functions

VII

- **Fourier Basis**

\[ \{1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t), \sin(3\omega t), \cos(3\omega t), \ldots, \sin(K\omega t), \cos(K\omega t)\} \]

**Properties**

- natural basis for periodic data
- Derivatives retain complexities.
Linear Smoothing: Regression on basis of functions

$K = 3$

$K = 4$

$K = 5$

$K = 6$
Splines. Splines are polynomial segments joined end-to-end. They are constrained to be smooth at the joints (called knots). The order (order = degree+1) of the polynomial segments and the location of the knots define the system.
Linear Smoothing: Regression on basis of functions $X$
Other basis:
- wavelet
- ...

How to choose the number $K$ of functions?
- Information Criterion: AIC, BIC.
- Cross-Validation
The model is $s(t) = \mu(t) + \epsilon(t)$ with $\mathbb{E}(\epsilon(t)) = 0$.

The estimate $\hat{s}(t)$ is computed as an expectation of $s$ at time $t$:

$$\mathbb{E}[s|t] = \mathbb{E}[s(t)] = \mathbb{E}[\mu(t) + \epsilon(t)] = \mu(t)$$

By definition of expectation:

$$\mathbb{E}[s|t] = \int s \ p_{s|t}(s|t) \ ds = \frac{\int s \ p_{st}(s,t) \ ds}{p_t(t)}$$

Note how Bayes formula is used.
Since we don't know the true densities $p_{st}(s, t)$ and $p_t(t)$, the idea of the Nadaraya-Watson estimator is to replace them by density estimates such that the integral is easy to compute and the solution provides a smooth estimate of $s(t)$.

Consequently the Nadaraya-Watson estimator can be written:

$$
\mathbb{E}[s(t)] \simeq \int \frac{s \hat{p}_{st}(s, t)}{\hat{p}_t(t)} \, ds = \frac{\sum_{i=1}^{n} \frac{1}{h} k \left( \frac{t-t^{(i)}}{h} \right) s^{(i)}}{\sum_{i=1}^{n} \frac{1}{h} k \left( \frac{t-t^{(i)}}{h} \right)} = \sum_{i=1}^{n} S_i(t) s^{(i)} = \hat{s}(t)
$$
Choosing the kernel \( k \) as the Dirac function is equivalent as using the empirical density estimates for \( p_{st}(s, t) \) and \( p_t(t) \).

The resulting estimate for \( \hat{s}(t) \) would not be smooth with the Dirac kernel and other kernels can be used (e.g. Gaussian, uniform, quadratic, triangle, Epanechnikov, etc.).
Consider a random variable $x$ for which observations \( \{x^{(i)}\}_{i=1,\ldots,n} \) have been collected. Choosing a even function $k$ such that

$$\int k(x) \, dx = 1 \quad \text{and} \quad k(x) \geq 0 \quad \forall x$$

the a Kernel Density Estimate (KDE) for the density function $p_x$ is:

$$\hat{p}_x(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - x^{(i)}}{h}\right)$$

$h$ is called the bandwidth and is a parameter controlling the level of smoothing.
Conclusion:

- **Nadaraya Watson estimator**: the bandwidth needs to be chosen/estimated.

- **Regression on basis of functions**: \( \{\theta_k\} \) needs to be estimated, the basis \( \{\phi_k(t)\} \) needs to be chosen, and the number \( K \) of functions needs to be selected.
Extension: Regularised basis approach

- When fitting the model $\mu(t) = \sum_{k=1}^{K} \theta_k \phi_k(t)$, one may add a smoothness constraint to ensure that the estimated function is smooth.
- If the basis of functions is differentiable then $D^2\mu(t)$ can be computed and a constraint (prior) can be added to the cost function $C$ to estimate the coefficients:

$$C(\theta_1, \ldots, \theta_K) = \sum_{i=1}^{n} (s^{(i)} - \mu(t^{(i)}))^2 + \lambda \int [D^2\mu(t)]^2 \, dt$$
Another example, consider growth, exponential growth or decay (e.g. population, radioactive decay, economic indicators), we know that such system follows an ODE:

$$Lx = -\gamma x + Dx = 0$$

then fitting the model $$\mu(t) = \sum_{k=1}^{K} \theta_k \phi_k(t)$$ can be constrained with

$$C(\theta_1, \cdots, \theta_K) = \sum_{j=1}^{n} (s^{(i)} - \mu(t^{(i)})^2 + \lambda \int [L\mu(t)]^2 dt$$

where $$\gamma$$ is unknown it can be estimated.
Depending of the problem, derivatives can be used for regularisation, adding prior for the type of solutions we are looking for (e.g. smoothness, or subject to a known ODE).

Derivatives can also be the subject of interest in the study i.e. For instance in mechanics, if you collect spatial positions over time (trajectories), one may be interested in studying the velocity (first derivative) or acceleration (second derivatives).
Example: Height of Girls

Figure 1.1. The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement.
Figure 1.2. The estimated accelerations of height for 10 girls, measured in centimeters per year. The heavy dashed line is the cross-sectional mean, and is a rather poor summary of the curves.
Functional PCA I

Definition (Mean and variance functions)

Consider the functions \{\mu_j(t)\}_{j=1,\ldots,J} that are observed instances of the random function \(s(t)\), the mean function is the average of the functions:

\[
\mathbb{E}[s(t)] = \bar{\mu}(t) = \frac{1}{J} \sum_{j=1}^{J} \mu_j(t)
\]

and the variance function is:

\[
\text{Var}[s(t)] = \frac{1}{J-1} \sum_{j=1}^{J} (\mu_j(t) - \bar{\mu}(t))^2
\]
Example (Height of Girls)

$$\text{Height}(t) = \frac{1}{10} \sum_{i=1}^{10} \text{Height}_i(t)$$

$$\text{Var}_{\text{Height}}(t) = \frac{1}{10-1} \sum_{i=1}^{10} (\text{Height}_i(t) - \overline{\text{Height}}(t))^2$$
FPCA, Principal Component Analysis for functions:

- We compute the mean $\mu(t)$ and subtract it to each curve:

$$\tilde{\mu}_j(t) = \mu_j(t) - \bar{\mu}(t)$$

- Let $\nu(t_1, t_2) = \mathbb{E}[\tilde{\mu}(t_1) \tilde{\mu}(t_2)]$ be the sample covariance function, it can be estimated by:

$$\hat{\nu}(t_1, t_2) = \frac{1}{J} \sum_{j=1}^{J} \tilde{\mu}_j(t_1) \tilde{\mu}_j(t_2)$$

- An eigencurve $u(t)$ is then computed such that:

$$\forall t_1, \int \hat{\nu}(t_1, t) u(t) \, dt = \lambda u(t_1) \quad \text{subject to} \quad \int u(t)^2 \, dt = 1$$
A simple approach is to discretise the functions $\tilde{\mu}_j(t)$ to a fine grid of $N$ sites equally spaced on the time line. This allows to treat these functions as finite dimensional vectors and standard PCA can be applied to estimate the eigenvectors.

The continuous principal component $u(t)$ is recovered by interpolating the discrete eigenvector.

The choice of interpolation method does not matter when recovering the eigencurve from the eigenvector when the space between the $N$ sampling sites is small.

This approach is the earliest approach to functional principal component analysis.
Transforms I

Definition (Integral transform)

An integral transform $\mathcal{T}$ is defined as:

$$\mathcal{T}[f(t)] \equiv F(u) = \int_{t_1}^{t_2} K(t, u) f(t) \, dt$$

where $K$ is the kernel function. This can also be understood as a linear combination ('continuous sum') over a basis of functions $K$.

Example

- The Laplace transform is an example of integral transform with $K(t, u) = \exp(-u t)$.
- The Fourier transform is another example with $K(t, u) = \exp(-ju t)$. 
Another useful example is the Radon transform:

**Definition (Radon transform)**

Having a function $f$ defined on a domain $x = (x, y) \in \mathbb{R}^2$, the Radon transform of $f$ is its integral along the line of equation $\rho - x \cos \theta - y \sin \theta = 0$ i.e.:

$$Rf(\rho, \theta) = \int_\mathbb{R} \int_\mathbb{R} \delta(\rho - x \cos \theta - y \sin \theta) \, f(x, y) \, dx \, dy$$

$$= \int_{\mathbb{R}^2} \delta(\rho - x^T n) \, f(x) \, dx$$

with $\delta(\cdot)$ the Dirac delta function, $\rho \in \mathbb{R}$, $\theta \in [-\frac{\pi}{2}; \frac{\pi}{2}]$ and $n = (\cos \theta, \sin \theta)^T$. 

Transforms II
Transforms III

\[ \rho - x^T n = 0 \]
Exercises:

1. Assuming that $f(x) = p_x(x)$ is the probability density function of $x$, what is $Rf(\rho, \theta)$ when $\theta = 0$ and $\theta = \frac{\pi}{2}$?

2. Assuming that $f(x) = p_x(x)$ is the probability density function of $x$, is $Rf(\rho, \theta)$ probability density function?
Consider the r.v. $x = (x, y)$ for which only one observation $x^{(1)}$ is available. Then the empirical density function is:

$$
\hat{p}_x(x) = \delta(x - x^{(1)}) = \delta(x - x^{(1)}) \delta(y - y^{(1)})
$$

What is the Radon transform of $p_x(x)$?

With two observations

$$
\hat{p}_x(x) = \frac{1}{2} \left[ \delta(x - x^{(1)}) + \delta(x - x^{(2)}) \right]
$$

What is the Radon transform of $p_x(x)$?
Application Radon Transform II
Application Radon Transform III
Application Radon Transform IV
Conclude on how the Radon transform can be used to detect perfect lineament.
Remarks:

- The *Hough transform* is a very closely related method to the Radon transform to find shapes (line, circles, etc.).

- The formulation of the Radon transform itself has been generalised to consider quadratic shapes for instance:
Application Radon Transform VIII

Example: crater detection on Mars -