REFERENCES


References


References


[37] B. Ramachandran, C. Justice, and A. Abrams, Eds., *Land Remote Sensing and Global Envi-


References


References


References


[94] C. Shah, M. Arora, and P. Corresponding, “Unsupervised classification of hyperspectral data:
References


References


References


[122] S. Myint, “A robust texture analysis and classification approach for urban land-use and land-
References


[123] G. Okin, D. Roberts, B. Murray, and W. Okin, “Practical limits on hyperspectral vegetation
discrimination in arid and semiarid environments,” *Remote Sensing of Environment*, vol. 77,

1205, 1995.

[125] M. Ulfarsson, J. Benediktsson, and J. Sveinsson, “Data fusion and feature extraction in the
2003.

mixture analysis of multispectral satellite imagery: A case study of the Greater Cairo region,

identification,” *Photogrammetric Engineering and Remote Sensing*, vol. 65, no. 11, pp. 1303–
1310, 1999.


Difference Vegetation Index at continental scales,” *Geoscience and Remote Sensing, IEEE

[130] M. Hall-Beyer, “Comparison of single-year and multiyear NDVI time series principal com-
ponents in cold temperate biomes,” *Geoscience and Remote Sensing, IEEE Transactions on*,

course spatial resolution data to derive continuous and discrete maps for complex regions,”
References


References


References


References


References


References


pp. 6236–6239.


APPENDIX A

MATHEMATICAL BACKGROUND

This appendix contains the mathematical background that is needed to understand Chapter 3, Chapter 4 and Chapter 5. In Section A.1 an overview is given of stochastic calculus. In the last five sections of the appendix the following mathematical concepts or ideas are summarised: Gaussian quadrature, Cholesky factorisation, the method of Langragian multipliers and KDE.

A.1 STOCHASTIC CALCULUS

The basic definitions used in Chapter 3 are defined in this section and can also be found in [48, 59, 82].

Definition 1 (A σ-field \( \mathcal{F} \)) Let \( \Omega \) be a non-empty set. A \( \sigma \)-field \( \mathcal{F} \) on \( \Omega \) is a family of subsets of \( \Omega \) such that

1. the empty set \( \emptyset \) belongs to \( \mathcal{F} \);

2. if \( A \) belongs to \( \mathcal{F} \), then so does the complement \( \Omega \setminus A \);

3. if \( A_1, A_2, \ldots \) is a sequence of sets in \( \mathcal{F} \), then their union \( A_1 \cup A_2 \cup \cdots \) also belongs to \( \mathcal{F} \).

Definition 2 (Family of Borel sets) The family of Borel sets \( \mathcal{B} = \mathcal{B}(\mathbb{R}) \) is a \( \sigma \)-field on \( \mathbb{R} \). The Borel \( \sigma \)-field in \( \mathbb{R} \) is the smallest \( \sigma \)-field containing all intervals in \( \mathbb{R} \).

Definition 3 (Probability measure) Let \( \mathcal{F} \) be a \( \sigma \)-field on \( \Omega \). A probability measure \( P \) is a function

\[ P : \mathcal{F} \to [0, 1] \]
such that

1. \( P(\Omega) = 1; \)

2. if \( A_1, A_2, \cdots \) are pairwise disjoint sets (that is, \( A_i \cap A_j = \emptyset \) for \( i \neq j \)) belonging to \( \mathcal{F} \), then

\[
P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots.
\]

The triple \((\Omega, \mathcal{F}, P)\) is known as a probability space. Each set \( A \) contained in \( \mathcal{F} \) is referred to as an event. Whenever the \( P(A) = 1 \) then event \( A \) is said to occur almost surely.

**Definition 4 (Random variable)** If \( \mathcal{F} \) is a \( \sigma \)-field on \( \Omega \), then a function \( \xi : \Omega \to \mathbb{R} \) is said to be \( \mathcal{F} \)-measurable if

\[
\{ \xi \in B \} \in \mathcal{F}
\]

for every Borel set \( B \in \mathcal{B}(\mathbb{R}) \). If \((\Omega, \mathcal{F}, P)\) is a probability space, then such a function \( \xi \) is called a random variable.

The short-hand notation \( \{ \xi \in B \} \) in the above definition represents the inverse image \( \xi^{-1}(B) \). In expanded form \( \{ \xi \in B \} \) is written as

\[
\{ \omega \in \Omega : \xi(\omega) \in B \}.
\]

**Definition 5 (The \( \sigma \)-field generated by a random variable)** The \( \sigma \)-field \( \mathcal{F}(\xi) \) generated by a random variable \( \xi : \Omega \to \mathbb{R} \) consists of all sets of the form \( \{ \xi \in B \} \), where \( B \) is a Borel set in \( \mathbb{R} \).

**Definition 6 (The \( \sigma \)-field generated by a family of random variables)** The \( \sigma \)-field \( \{ \xi_i : i \in I \} \) generated by a family \( \{ \xi_i : i \in I \} \) of random variables is defined to be the smallest \( \sigma \)-field containing all the events of the form \( \{ \xi_i \in B \} \), where \( B \) is a Borel set in \( \mathbb{R} \) and \( i \in I \) (\( I \) is an index set).

**Definition 7 (The distribution of a random variable and the cumulative distribution function)**

Every random variable \( \xi : \Omega \to \mathbb{R} \) gives rise to a probability measure

\[
P_\xi(B) = P\{ \xi \in B \}
\]

on \( \mathbb{R} \) defined on the \( \sigma \)-field of Borel sets \( B \in \mathcal{B}(\mathbb{R}) \). The probability measure \( P_\xi \) is called the distribution of \( \xi \). The function \( F_\xi : \mathbb{R} \to [0, 1] \) defined by

\[
F_\xi(x) = P\{ \xi \leq x \} = P_\xi((-\infty, x]) , \ x \in \mathbb{R}
\]
Appendix A Mathematical Background

is called the cumulative distribution function of $\xi$.

Definition 7 implies that the measurability of $\xi$ ensures that $\xi$ generates a probability measure $P_\xi$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, such that $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_\xi)$ is also a probability space. If $P_\xi$ is known, then the structure of $(\Omega, \mathcal{F}, P)$ is no longer needed to describe the behavior of $\xi$.

**Definition 8 (Probability density function)** If there is a Borel function $f_\xi : \mathbb{R} \to \mathbb{R}$ such that for any Borel set $B \subset \mathbb{R}$

$$P\{\xi \in B\} = \int_B f_\xi(x)dx,$$

then $\xi$ is said to be a random variable with absolutely continuous distribution and $f_\xi$ is called the density of $\xi$. If there is a (finite or infinite) sequence of pairwise distinct real number $x_1, x_2, \cdots$ such that for any Borel set $B \subset \mathbb{R}$

$$P\{\xi \in B\} = \sum_{x_i \in B} P\{\xi = x_i\},$$

then $\xi$ is said to have discrete distribution with values $x_1, x_2, \cdots$ and mass $P\{\xi = x_i\}$ at $x_i$.

**Definition 9 (Essential supremum)** Let $(\xi_i)_{i \in I}$ (where $I$ is an index set) be a family of real-valued random variables on $(\Omega, \mathcal{F}, P)$, bounded by another variable. The essential supremum of $(\xi_i)_{i \in I}$ is $\Xi$ (denoted by $\Xi = \text{ess sup}_I \xi_i$) if

$$(\forall i \in I) \xi_i \leq \mathcal{R}^c P - \text{almost surely} \iff \Xi \leq \mathcal{R}^c P - \text{almost surely}.$$ 

**Definition 10 (Expectation)** A random variable $\xi : \Omega \to \mathbb{R}$ is said to be integrable if

$$\int_{\Omega} |\xi|dP < \infty.$$ 

Then

$$\mathbb{E}[\xi] = \int_{\Omega} \xi dP$$

exists and is called the expectation of $\xi$.

The expectation $\mathbb{E}[\xi]$ can also be expressed as a Riemann integral as follows [48]:

$$\mathbb{E}[\xi] = \int_{\Omega} \xi dP = \int_{\mathbb{R}} xdP_\xi = \int_{-\infty}^{\infty} xdF_\xi(x) = \int_{-\infty}^{\infty} xf_\xi(x)dx.$$
It is also important to note that \( \frac{d}{dx} F_\xi(x) = f_\xi(x) \).

**Definition 11 (Conditioning on an event)** For an integrable random variable \( \xi \) and any event \( B \in \mathcal{F} \) such that \( P(B) \neq 0 \) the conditional expectation of \( \xi \) given \( B \) is defined by

\[
E[\xi | B] = \frac{1}{P(B)} \int_B \xi dP.
\]

**Definition 12 (Conditioning on a discrete random variable)** Let \( \xi \) be an integrable random variable and let \( \eta \) be a discrete random variable. Then the conditional expectation of \( \xi \) given \( \eta \) is defined to be a random variable \( E[\xi | \eta] \) such that

\[
E[\xi | \eta](\omega) = E[\xi | \{ \eta = y_n \}] \text{ if } \eta(\omega) = y_n
\]

for any \( n = 1, 2, \cdots \).

By using Definition 12 the following proposition can be derived (stated here without proof) [82]:

**Proposition 1** If \( \xi \) is an integrable random variable and \( \eta \) is a discrete random variable, then

1. \( E[\xi | \eta] \) is \( \sigma(\eta) \)-measurable;

2. For any \( A \in \sigma(\eta) \)

\[
\int_A E[\xi | \eta] dP = \int_A \xi dP. \tag{A.1}
\]

**Definition 13 (Conditioning on an arbitrary random variable)** Let \( \xi \) be an integrable random variable and let \( \eta \) be an arbitrary random variable. Then the conditional expectation of \( \xi \) given \( \eta \) is defined to be a random variable \( E[\xi | \eta] \) such that

1. \( E[\xi | \eta] \) is \( \sigma(\eta) \)-measurable;

2. For any \( A \in \sigma(\eta) \)

\[
\int_A E[\xi | \eta] dP = \int_A \xi dP. \tag{A.2}
\]

**Definition 14 (Conditioning on a \( \sigma \)-field)** Let \( \xi \) be an integrable random variable on a probability space \( (\Omega, \mathcal{F}, P) \), and let \( \mathcal{G} \) be a \( \sigma \)-field contained in \( \mathcal{F} \). Then the conditional expectation of \( \xi \) given \( \mathcal{G} \) is defined to be a random variable \( E[\xi | \mathcal{G}] \) such that
Appendix A Mathematical Background

1. \( \mathbb{E}[\xi | \mathcal{G}] \) is \( \mathcal{G} \)-measurable

2. For any \( A \in \mathcal{G} \)

\[
\int_A \mathbb{E}[\xi | \mathcal{G}] \, dP = \int_A \xi \, dP.
\]

**Definition 15 (Sample path)** The sequence of numbers \( \xi_1(\omega), \xi_2(\omega), \cdots \) for any fixed \( \omega \in \Omega \) is called a sample path.

**Definition 16 (Filtration)** A sequence of \( \sigma \)-fields \( \mathcal{F}_1, \mathcal{F}_2, \cdots \) on \( \Omega \) such that

\[
\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}
\]

is called a filtration.

**Definition 17 (Adapted to a filtration)** A sequence of random variables \( \xi_1, \xi_2, \cdots \) is adapted to a filtration \( \mathcal{F}_1, \mathcal{F}_2, \cdots \) if \( \xi_n \) is \( \mathcal{F}_n \)-measurable for each \( n = 1, 2, \cdots \).

**Definition 18 (Stopping time)** A random variable \( T \) with values in the set \( \{1, 2, \cdots\} \cup \{\infty\} \) is called a stopping time (with respect to a filtration \( \mathcal{F}_n \)) if for each \( n = 1, 2, \cdots \)

\[
\{T = n\} \in \mathcal{F}_n.
\]

At this point Wald’s identities can be presented (and is stated here without proof) [48]. Wald’s identities were used in Chapter 3 to derive the OC and ASN functions.

**Theorem 6 (Wald’s identities)** Suppose \( \{s_k; k = 1, 2, \cdots\} \) is an i.i.d. sequence adapted to the filtration \( \{\mathcal{F}_k\} \), and let \( S_k \) denote the sequence of cumulative sums, \( S_k = \sum_{i=1}^{k} s_i \). Then the following statements are true:

1. Suppose \( \mathbb{E}[s_1] \) is finite, then for every stopping time \( T \) satisfying \( \mathbb{E}[T] < \infty \), \( \mathbb{E}[S_T] = \mathbb{E}[s_1] \mathbb{E}[T] \implies \mathbb{E}[T] = \frac{\mathbb{E}[S_T]}{\mathbb{E}[s_1]} \) when \( \mathbb{E}[s_1] \neq 0 \).

2. Suppose \( \mathbb{E}[s_1^2] \) is finite, then for every stopping time \( T \) satisfying \( \mathbb{E}[T] < \infty \), \( \mathbb{E}[S_T - T \mathbb{E}[s_1]]^2 = \mathbb{E}[T] \mathbb{E}[s_1 - \mathbb{E}[s_1]]^2 \implies \mathbb{E}[T] = \frac{\mathbb{E}[S_T]}{\mathbb{E}[s_1]} \) when \( \mathbb{E}[s_1] = 0 \).

3. For scalars \( a, h > 0 \) define the stopping time \( T^h_{-a} = \inf\{k | S_k \notin (-b, a)\} \). Suppose \( \omega \neq 0 \) is such that \( \mathbb{E}[e^{-\alpha s_1}] < \infty \), then \( \mathbb{E}[e^{\alpha T} \mathbb{E}[e^{-\alpha s_1}])^{-T}] = 1 \), holds for any stopping time \( T \) such that \( P(T \leq T^h_{-a}) = 1 \).
Appendix A Mathematical Background

**Definition 19 (Stochastic process)** A stochastic process is a family of random variables $\xi(t)$ parameterised by $t \in \mathcal{T}$, where $\mathcal{\mathcal{T}} \subset \mathbb{R}$. When $\mathcal{T} = \{1, 2, \cdots\}$ then $\xi(t)$ is a stochastic process in discrete time (a sequence of random variables). When $\mathcal{T}$ is an interval in $\mathbb{R}$ (typically $\mathcal{T} = [0, \infty)$) then $\xi(t)$ is a stochastic process in continuous time. Moreover, for every $\omega \in \Omega$ the function $\mathcal{T} \ni t \rightarrow \xi(t, \omega)$ is called a sample path of $\xi(t)$.

**Definition 20 (Brownian motion)** The Wiener process (or Brownian motion) is a stochastic process $W(t)$ with values in $\mathbb{R}$ defined for $t \in [0, \infty)$ such that

1. $W(0) = 0$ almost surely;

2. the sample paths $t \rightarrow W(t)$ are almost surely continuous;

3. for any finite sequence of times $0 < t_1 < \cdots < t_n$ and Borel sets $A_1, \cdots, A_n \subset \mathbb{R}$

$$P\{W(t_1) \in A_1, \cdots, W(t_n) \in A_n\} = \int_{A_1} \cdots \int_{A_n} p(t_1, 0, x_1) p(t_2 - t_1, x_1, x_2) \cdots p(t_n - t_{n-1}, x_{n-1}, x_n) dx_1 \cdots dx_2,$$

where

$$p(t, w, y) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-w)^2}{2t}}$$

defined for any $x, y \in \mathbb{R}$ and $t > 0$ is called the transition density.

From Definition 20 the following theorem can be derived (stated here without proof) [82]:

**Theorem 7** A stochastic process $W(t)$, $t \geq 0$, is a Wiener process if and only if the following conditions hold:

1. $W(0) = 0$ almost surely;

2. the sample paths $t \rightarrow W(t)$ are almost surely continuous;

3. $W(t)$ has stationary (the distribution of $X(s+t) - X(s)$ does not depend on $s$ for all $s,t > 0$) independent ($\mathbb{E}[(W(u) - W(t))(W(s) - W(r))] = 0$ for any $0 \leq r \leq s \leq t \leq u$) increments;

4. the increment $W(t) - W(s)$ has the normal distribution with mean 0 and variance $t - s$ for any $0 \leq s \leq t$. 

Department of Electrical, Electronic and Computer Engineering
University of Pretoria
Definition 21 (Random step process) A process \( f(t), t \geq 0 \) is a random step process if there is a finite sequence of numbers \( 0 = t_0 < t_1 < \cdots < t_n \) and square integrable random variables \( \eta_0, \eta_1, \cdots, \eta_{n-1} \) such that

\[
f(t) = \sum_{j=0}^{n-1} \eta_j 1_{[t_j, t_{j+1})}(t),
\]

where \( \eta_j \) is \( \mathcal{F}_{t_j} \)-measurable for \( j = 0, 1, \cdots, n-1 \). The set of random step processes will be denoted by \( M_{\text{step}}^2 \). In Equation A.3, \( 1(t) \) represents the indicator function.

Definition 22 (Stochastic integral of a random step process) The stochastic integral of a random step process \( f \in M_{\text{step}}^2 \) is defined by

\[
I(f) = \sum_{j=0}^{n-1} \eta_j (W(t_{j+1}) - W(t_j)).
\]

Definition 23 \( M^2 \) is the class of stochastic processes \( f(t), t \geq 0 \) that satisfy

\[
\mathbb{E}\left[ \int_0^\infty |f(t)|^2 \right] < \infty
\]

and

\[
\lim_{n \to \infty} \mathbb{E}\left[ \int_0^\infty |f(t) - f_n(t)|^2 \right] = 0. \tag{A.3}
\]

In this case the sequence of step processes \( f_1, f_2, \cdots \) approximates \( f \) in \( M^2 \).

Definition 24 (Itô stochastic integral) \( I(f) \in L^2 \) (\( L^2 \) is the space of square integrable random variables) is called the Itô stochastic integral [from 0 to \( \infty \)] of \( f \in M^2 \) if

\[
\lim_{n \to \infty} \mathbb{E}[|I(f) - I(f_n)|^2] = 0
\]

for any sequence \( f_1, f_2, \cdots \in M_{\text{step}}^2 \) of random step processes that approximates \( f \) in \( M^2 \) (i.e. such that Equation A.3 is satisfied). \( I(f) \) and \( \int_0^\infty f(t)dW(t) \) are interchangeable.

A.2 GAUSSIAN QUADRATURE

The Gauss-Legendre quadrature rule is formally expressed as [40]

\[
\int_a^b f(x) \, dx \approx \frac{b-a}{2} \sum_{i=1}^n w_i f \left( \frac{b-a}{2} z_i + \frac{a+b}{2} \right),
\]

where \( z_i \) is the \( i \)-th root of the Legendre polynomial \( P_n(z) = \frac{1}{2^n n!} \frac{d^n}{dz^n} \left[ (z^2 - 1)^n \right] \) and

\[
w_i = \frac{2}{(1-z_i^2) [P_n'(z_i)]^2}.
\]
A.3 CHOLESKY FACTORISATION

If $A$ has real entries and is symmetric ($A = A^T$) and positive definite ($z^T A z$ is positive, for any column vector $z$), then $A$ can be written as

$$A = LL^*,$$

where $L$ has positive diagonal entries and is a lower triangular matrix, and $L^*$ is equal to the conjugate transpose of $L$. Writing $A$ as the product $LL^*$ is known as Cholesky decomposition [40].

The Cholesky algorithm, used to calculate the decomposition matrix $L$ is described next. The recursive Cholesky algorithm starts by setting $i = 1$ and $A(1) = A$. At step $i$, the matrix $A(i)$ then has the following form:

$$A(i) = \begin{pmatrix}
I_{i-1} & 0 & 0 \\
0 & a_{i,i} & b_i^* \\
0 & b_i & B(i)
\end{pmatrix},$$

where $I_{i-1}$ is the identity matrix of dimension $i-1$.

If matrix $L_i$ is defined as

$$L_i = \begin{pmatrix}
I_{i-1} & 0 & 0 \\
0 & \sqrt{a_{i,i}} & 0 \\
0 & \sqrt{b_{i,i}} b_i & I_{n-i}
\end{pmatrix},$$

then $A(i)$ can be written as

$$A(i) = L_i A(i+1) L_i^*,$$

where

$$A(i+1) = \begin{pmatrix}
I_{i-1} & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & B(i) - \frac{1}{a_{i,i}} b_i b_i^*
\end{pmatrix}.$$

Note that $b_i b_i^*$ is an outer product. If the above is repeated enough times then at step $n$ (which is also the dimension of the matrix $A$), $A(n+1) = I$. Hence, the lower triangular matrix $L$ is equal to

$$L = L_1 L_2 \ldots L_n.$$
A.4 LAGRANGE MULTIPLIERS

The method of Lagrange multipliers is employed to solve the following optimization problem:

\[
\begin{align*}
\max_{x,y} & \quad f(x,y) \\
\text{s.t.} & \quad g(x,y) = c,
\end{align*}
\]  
(A.4)

where \( f \) and \( g \) are functions that have continuous first order partial derivatives and \( c \) is a constant. The Lagrange function is derived from \( f(x,y), g(x,y), c \) and a new variable \( \lambda \) (the Lagrange multiplier) and is defined as

\[
\mathcal{L}(x,y,\lambda) = f(x,y) + \lambda \cdot \left( g(x,y) - c \right).
\]

If \( x_0 \) and \( y_0 \) are solutions of Equation A.4, then there exists an \( \lambda_0 \) such that \( (x_0, y_0, \lambda_0) \) is a stationary point of \( \mathcal{L}(x,y,\lambda) \) [40].

A.5 KERNEL DENSITY ESTIMATION

If \( (x_1, x_2, \ldots, x_n) \) are i.i.d. samples drawn from a distribution with an unknown density \( f \), then the kernel density estimator of \( f \) is

\[
\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - x_i}{h} \right),
\]

where \( K(x) \) is a symmetric but not necessarily positive function that integrates to one (and is known as the kernel), \( h > 0 \) is called the bandwidth (functions as a smoothing parameter) and \( K_h(x) = 1/hK(x/h) \) (and is known as the scaled kernel). If Gaussian kernels are used to approximate univariate data, and the underlying density being estimated is itself a Gaussian then Silverman’s rule of thumb is the optimal choice of \( h \). Silverman’s rule of thumb is:

\[
h = \left( \frac{4\hat{\sigma}^5}{3n} \right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5},
\]

where \( \hat{\sigma} \) is the standard deviation of the samples \( (x_1, x_2, \ldots, x_n) \). The multivariate case is approached in a similar way [195].