VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

KLASIFIKACE TESTOVACÍCH MANÉVRŮ Z LETOVÝCH DAT

CLASSIFICATION OF TESTING MANEUVERS FROM FLIGHT DATA

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE   Bc. MARTIN FUNIAK
AUTHOR

VEDOUCÍ PRÁCE   Ing. PETER CHUDÝ, Ph.D. MBA
SUPERVISOR

BRNO 2015

Abstrakt

Zapisovač letových údajů je zařízení určené pro zaznamenávání letových dat z různých senzorů v letadlech. Analýza letových údajů hraje důležitou roli ve vývoji a testování avioniky. Testování a hodnocení charakteristik letadla se často provádí pomocí testovacích manévrů. Naměřená data z jednoho letu jsou uložena v jednom letovém záznamu, který může obsahovat několik testovacích manévrů. Cílem této práce je identifikovat základní testovací manévry s pomocí naměřených letových dat. Teoretická část popisuje letové manévry a formát měřených letových dat. Analytická část popisuje výzkum v oblasti klasifikace založené na statistice a teorii pravděpodobnosti potřebnou pro pochopení složitých Gaussovských směšovacích modelů. Práce uvádí implementaci, kde jsou Gaussovy směšovací modely použité pro klasifikaci testovacích manévrů. Navržené řešení bylo testováno pro data získána z letového simulátoru a ze skutečného letadla. Ukázalo se, že Gaussovy směšovací modely poskytují vhodné řešení pro tento úkol. Další možný vývoj práce je popsán v závěrečné kapitole.

Abstract

A flight data recorder is a device dedicated for recording flight data from a wide range of aircraft sensors. The analysis of the flight data has an important role in the development and testing of avionics. The testing and evaluation of the aircraft responses is often performed using the testing maneuvers. The sensor data from one flight are stored within one flight record, which may contain several testing maneuvers. The goal of this thesis is to identify basic testing maneuvers using measured flight data. The theoretical part describes the flight maneuvers and a related format of the flight data. A research in the field of statistics driven data classification and the probability theory, necessary to understand complex Gaussian Mixture Models, is covered in the analytical part. The thesis also introduces an implementation where the Gaussian Mixture Models are used for the classification of the testing maneuvers in the flight data. The testing of the proposed solution for the data from a flight simulator and a real aircraft showed that the Gaussian Mixture Models provide a suitable solution for this task. A possible future research direction is described in the final chapter.

Klíčová slova

analýza letových dat, klasifikace, učení bez učitele, identifikace letových manévrů, EM algoritmus, Gaussove směšovací modely

Keywords

flight data analysis, classification, unsupervised learning, identification of flight maneuvers, EM algorithm, Gaussian Mixture Models

Citace

Classification of Testing Maneuvers from Flight Data

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením pana Ing., Petra Chudého, Ph.D. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

....
Martin Funiak
May 21, 2015

Poděkování

Chcel by som poďakovať môjmu vedúcemu Petrovi Chudému za ochotu a čas, ktorý mi venoval pri mnohých konzultáciách.

Contents

1 Introduction

The main principles of the flight instruments have remained the same over a long period of time. The flight instruments are displays or sensors and they help the pilot to observe the state of an aircraft. There are six basic flight instruments in the cockpit, the airspeed indicator, the attitude indicator, the altimeter, the vertical speed indicator, the heading indicator and the turn coordinator. These flight instruments offer the observation of the aircraft state in real time. They are based on the mechanical, analogue and barometric devices which offer low failure rates. Besides the flight instruments, modern aircraft often use a Flight Data Recorder (FDR), which is connected to the Inertial Measurement Unit (IMU).

The Flight Data Recorder (FDR) gathers the flight parameters such as the static pressure and the dynamic pressure, angular rates, Global Positioning System (GPS) coordinates, Angle of Attack or Angle of Sideslip. The measuring instruments are the gyroscopes, accelerometers and others.

We could assume gathering of flight parameters as a simple task. The challenging task comes afterwards. The analysis and the correction of the measured data is a very difficult task, due to a great deal of potential errors in the measuring instruments. Noise, insufficient calibration or inappropriate location of a sensor, could lead to inaccurate measurements of flight data. However, there is always a way to perform the correction of the measurement to achieve a better accuracy.

Measured flight parameters give us the possibility to evaluate the flight in many different ways. For example, if we consider a deflection of the rudder and the smoothness of motion of an aircraft during a flight maneuver, then we could evaluate the quality of the pilot's ability to control the aircraft. Measurement of the flight parameters is to be applied in case of failure prevention, pilot capabilities observation and monitoring of the physical condition of the aircraft body or in a case of an accident, it helps in providing evidence for the estimation of the cause of a failure.

There are many reasons to use the identification of the testing maneuvers. For example, if we want to analyse or examine the characteristics or the aerodynamic coefficients of the aircraft, such as lift or drag coefficients. The feedback of the aircraft to the testing maneuvers can be evaluated using an impulse response of a system, when we examine reaction of some system to the unit step. The common test flight maneuvers include, for example, the doublets and pulses in the elevator, aileron or rudder. These testing maneuvers are characterized by a significant change in the deflection of the elevator, aileron or rudder. Only an experienced pilot can perform testing maneuvers without any major shortcomings. A pilot must hold the aircraft in constant airspeed and constant altitude before executing the testing maneuvers. These are the basic conditions to be met.

The aim of this work is to process measured data from a flight data recorder and then identify the testing maneuvers as a result of the classification process. The final

framework for the classification should provide an option to the configuration settings and the visualization of the classification results. Results should be available in the form of plots with highlighted time segments with the testing maneuvers for each flight record. Another approach to the visualization task should provide an option to analyse selected flight parameters during the identified testing maneuvers.

1.1 Structure of thesis

The text of this thesis is divided into six chapters. Chapter 2 deals with the required theoretical basis such as the aircraft reference frames, the format of measured data from a FDR, description of the flight maneuvers and the error characteristics of the inertial sensors used to measure the flight data. Chapter 3 covers the research and analysis in the field of the classification. The main part of this chapter is formed by the probability theory, which is necessary to introduce complex Gaussian Mixture Models. The fourth chapter demonstrates the application of the Gaussian Mixture Models in practice and a description of the implementation part of the proposed system. The experiments and testing of the implemented classifier on the data coming from the simulation and real aircraft are covered in Chapter 5. Lastly, Chapter 6 sums up the work and discusses its potential future development.

2 Theoretical basis

The automation in the process of the identification of the flight events involves an advanced knowledge built on the comprehension of the basics of the aircraft dynamics. The explanation of the aircraft frames of reference includes an introduction to the flight dynamics. Furthermore, the format of the measured data, their meaning and physical units are introduced. This is followed by a list of basic flight maneuvers.

This chapter describes the theory as listed above. The last section of this chapter contains an overview of the error characteristics of the inertial sensors.

The description of the coordinate systems, flight data and flight maneuvers is intended to give the reader a broader view of the issue of flight data analysis. Detailed description of the flight data and flight maneuvers prepares the ground for more complicated task of the identification of all flight maneuvers during whole flight. It could be a direction for future research.

2.1 Coordinate systems and reference frames

We use the coordinate system [5] to keep the track of an aircraft orientation and motion in space. A coordinate system determines the position variable and the way one describes or observes motion in a specific reference frame. The reference frame serves for an effective description and observation of motion. Reference frames are divided into ground reference frames and vehicle-carried reference frames. Standard convention in the description of the attitude requires axes of the reference system and the axis of rigid aircraft body. The transformation between different reference frames has an important role in the description of an aircraft's motion and position.

2.1.1 ECEF coordinate system (Earth Centred Earth Fixed)

The ECEF coordinate system is a Cartesian coordinate system. Its origin O_e is located in the centre of the Earth (Earth Centred) and its axes are fixed to the Earth surface (Earth fixed) and are aligned with an International Reference Pole (IRP) and International Reference Meridian (IRM) as illustrated in the Figure 2.1. The ECEF coordinates are the ones used by most satellite system to designate an Earth position. The X_e axis lies in the equatorial plane and points through the Greenwich Meridian. The Y_e axis lies in the equatorial plane to form a right hand system with the X_e axis and Z_e axis. Z_e axis is the rotation axis of the Earth [8].
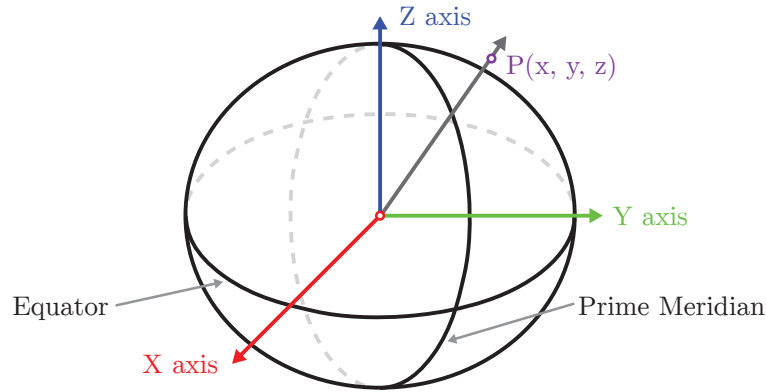
Figure 2.1: ECEF Coordinate system

2.1.2   WGS-84 Reference Frame

The Global Positioning System (GPS) uses World Geodetic System reference frame from 1984 (WGS-84). There are three angular or curvilinear coordinates that are in everyday use in maps. First coordinate is the latitude ϕ; it is the angle up from the Equator. Zero value of the angle is on the Equator and we distinguish South latitude and North latitude. Usual range for the latitude is from 0 degrees to 90 as we go from the Equator towards the Poles. The longitude λ is the second coordinate and it represents the angle on the equatorial plane. Zero value of the angle is on the Prime Meridian (Greenwich, England) and we distinguish West and East longitude. The Figure 2.2 represents the use of the coordinates in WGS-84 reference frame. Rotational effects cause flattening of the Earth and the cross-section of a Meridian is an ellipse, not a circle. The WGS-84 uses the *geodetic latitude*, the angle that makes the equatorial plane with the line perpendicular to the ellipsoid. The height determination requires the application of reference surface of the Earth [21].
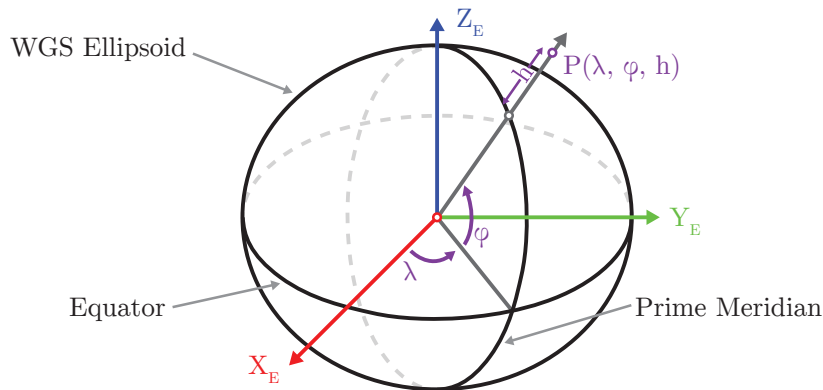

Figure 2.2: WGS84 Reference frame

The ellipsoid is used by the WGS-84 frame as a reference surface [21] and the resulting height is called the *geodetic height* or *ellipsoidal height*. The height used on maps is called the *orthometric height* or *Mean Sea Level* (MSL) height. The geoid is the best fit

approximation of the Earth's real surface. The geoid approximates the surface at the mean sea level and it can be below or above the ellipsoid. Still, we need to approximate geoid by mathematically simpler reference surface. An ellipsoid is more accurate than sphere and it has the variation from the real Earth surface less than 100 meters. An ellipsoid is defined by the semi-major axis (on the equatorial plane) and the semi-minor axis (on the North-South pole axis). The flattening is the amount by which the ellipsoid deviates from an ideal sphere. The Figure 2.3 demonstrates the difference between the geoid, the ellipsoid and the Earth surface.

H - orthometric height
h - ellipsoidal height
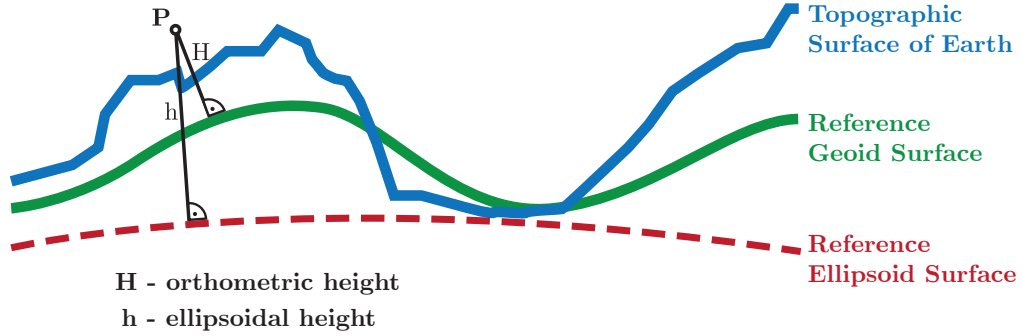
Figure 2.3: Difference between the Geoid, the Ellipsoid and the Earth surface

There are several important parameters in WGS-84 [21]:

- the semi-major axis $R_{Ea} = 6378137.0m$,

- the flattening factor $f = 1/298,257223436$,

- the semi-minor axis $R_{Eb} = R_{Ea}(1 - f) = 6356752.0m$,

- the first eccentricity $e = \sqrt{\frac{R_{Ea}^2 - R_{Eb}^2}{R_{Ea}}} = 0.08181919$,

- the prime vertical radius of curvature $N(\phi) = \frac{R_{Ea}}{\sqrt{1 - e^2 sin^2(\phi)}}$.

2.1.3 NED Frame (North-East-Down System)

The NED frame is used to define the position of the aircraft. It's origin is located in a reference point of the aircraft. The X_o axis points towards the geographic North Pole and it's parallel to the local geoid surface. The Y_o axis is pointing East to form a right hand system with X_o axis and it's parallel to the local geoid surface, too. The Z_o axis is perpendicular to the local geoid surface and it points downwards. Generally speaking, the movement of an aircraft influences the axis directions in the NED frame.
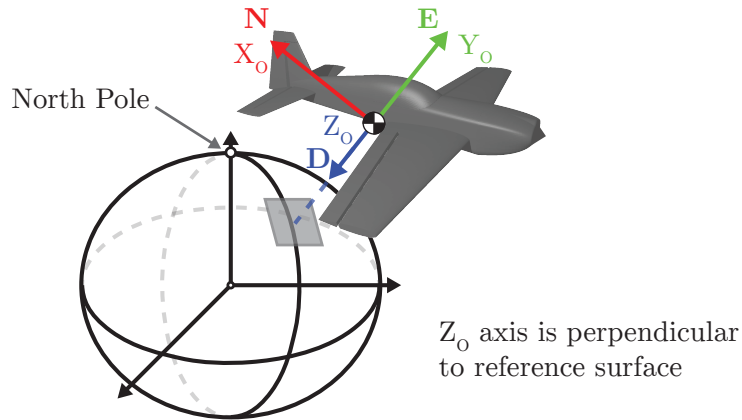
Figure 2.4: NED frame

2.1.4 BFF frame (Body fixed frame)

The body fixed frame is a vehicle-carried frame and it is exactly defined on the body of an aircraft. This reference frame is often used to describe the forces and moments acting within the body of an aircraft. The origin O_b is located in the aircraft centre of gravity. The X_b axis points towards the aircraft front in the symmetry plane. The Y_b axis points to the right wing and forms an orthogonal right hand system. The Z_b axis is perpendicular to X_b and Y_b axis and points downwards in the symmetry plane of the aircraft [8].
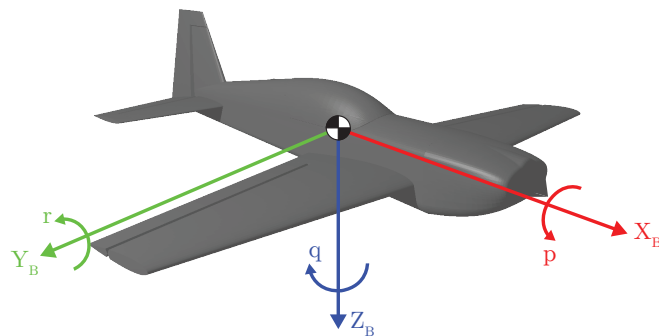
Figure 2.5: Body fixed frame

2.1.5 Body structural frame

The Body structural frame is used to define positions of parts within the aircraft. The origin O_{bs} is situated at the front of an aircraft. The X_{bs} axis points backward from the front of an aircraft in the symmetry plane. The Y_{bs} axis points towards the right wing to form an orthogonal right hand system. The Z_{bs} axis points upward.

Figure 2.6: Body structural frame

2.1.6 Transformation between reference frames

Transformations between reference frames can be very useful in dealing with the description of an aircraft motion. There are two main methods how to do this transformation [5]. The Euler angles ϕ, ψ, Θ denote rotations around the X_O axis, Y_O axis and Z_O axis, respectively. The final position after transformation depends on the order of the transformations. Changing the order will give a different final position.
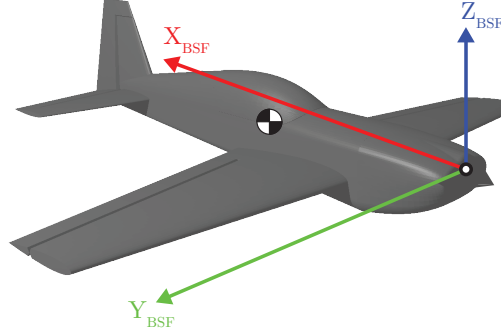
Transformation from WGS-84 coordinates to Cartesian ECEF coordinates
Cartesian ECEF coordinates are results of following equations:

- $x = (N(\phi) + h)cos(\phi)cos(\lambda)$

- $y = (N(\phi) + h)cos(\phi)sin(\lambda)$

- $z = [N(\phi)(1 - e^2) + h]sin(\phi)$

The transformation from ECEF coordinates to WGS-84 coordinates is more difficult and possible by using iterative methods.

Transformation between NED and ECEF frame
When transforming from NED reference frame to ECEF reference frame two rotations are needed. First, we make rotation around the axis Z_e with an angle of geodetic longitude λ. This single rotation is represented by a transformation matrix:

$$M_{OE,1}(\lambda) = \begin{pmatrix} \cos\lambda & \sin\lambda & 0 \\ -\sin\lambda & \cos\lambda & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.1}$$

Second, there is a rotation around the axis y_0 with a negative angle of geodetic latitude $\phi - \frac{\pi}{2}$ represented by the transformation matrix:

$$M_{OE,2}(\phi) = \begin{pmatrix} \cos(-\phi - \frac{\pi}{2}) & 0 & \sin(-\phi - \frac{\pi}{2}) \\ 0 & 1 & 0 \\ -\sin(-\phi - \frac{\pi}{2}) & 0 & \cos(-\phi - \frac{\pi}{2}) \end{pmatrix} \tag{2.2}$$

Final transformation matrix M_{OE} is a result of following multiplication

$$M_{OE} = M_{OE,2}(\lambda) \cdot M_{OE,1}(\phi) = \begin{pmatrix} -\sin\phi\cos\lambda & -\sin\phi\sin\lambda & \cos\phi \\ -\sin\lambda & \cos\lambda & 0 \\ -\cos\phi\cos\lambda & -\cos\phi\sin\lambda & -\sin\phi \end{pmatrix} \quad (2.3)$$

By using transformation matrix, which is orthogonal, we take the advantage of a reduced effort with backward transformation. The transformation from ECEF reference frame to NED reference frame needs simple transposition of the transformation matrix [5].

$$M_{EO} = M_{OE}^T \quad (2.4)$$

Transformation between BFF and NED frame

A transformation matrix is represented by an Euler angle and hence a sequence of matrix multiplications denotes sequence of rotations by an Euler angle. Rotation of a vector from the NED reference frame to the Body Fixed Frame can be performed by multiplying the vector by the transformation matrix M_{BO}. First rotation around the z_0 axis (Z axis in NED reference frame) by the azimuth angle Ψ (or the yaw Euler angle) produces new reference frame, where the Z axis is aligned with NED reference frame and the X axis and Y axis are rotated by the azimuth angle Ψ.

$$M_{BO,1}(\Psi) = \begin{pmatrix} \cos\Psi & \sin\Psi & 0 \\ -\sin\Psi & \cos\Psi & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (2.5)$$

Second rotation around the k_2 axis (Y axis in Body fixed frame) by a pitch angle θ (or a inclination angle) produces new reference frame, where the X axis and the Z axis are rotated by a pitch angle θ. The transformation matrix from the second step is shown below.

$$M_{BO,2}(\theta) = \begin{pmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{pmatrix} \quad (2.6)$$

Lastly, third step includes rotation around the X_B axis (in Body Fixed Frame) by the roll angle Φ (or the bank angle). This step leads to the creation of the third transformation matrix $M_{BO,3}$.

$$M_{BO,3}(\Phi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\Phi & \sin\Phi \\ 0 & -\sin\Phi & \cos\Phi \end{pmatrix} \quad (2.7)$$

Resulting transform matrix is given by the multiplication of partial transformation matrices, $M_{B0} = M_{BO,3}(\Phi) \cdot M_{BO,2}(\theta) \cdot M_{BO,1}(\Psi)$. The order of the multiplication is very important.

$$M_{BO} = \begin{pmatrix} \cos\Psi\cos\theta & \sin\Psi\cos\theta & -\sin\theta \\ \cos\Psi\sin\theta\sin\Phi - \sin\Psi\cos\Phi & \sin\Psi\sin\theta\sin\Phi + \cos\Psi\cos\Phi & \cos\theta\sin\Phi \\ \cos\Psi\sin\theta\cos\Phi + \sin\Psi\sin\Phi & \sin\Psi\sin\theta\cos\Phi - \cos\Psi\sin\Phi & \cos\theta\cos\Phi \end{pmatrix} \quad (2.8)$$

An advantage of reduced effort with backward transformation remains the same. The transformation matrix is orthogonal, and so simple transposition is sufficient.

2.2 Flight data

Flight data are indispensable in the process of flight evaluation or flight event identification. Data can be obtained from real measurements during flight tests on a flight simulator, too. The data are recorded using various aircraft sensors. Flight data in civil aviation are recorded by a *Flight Data Recorder* (FDR). The FDR is often referred to the "black box" and it is built to withstand high temperatures, pressure and other extreme conditions which occur during an aircraft crash. Improvements of the avionic systems such as the incidence avoidance were developed mainly because of aircraft crash investigations based on the flight data from the FDR [15].

There are several important issues to deal with, e.g. the way how the flight data are being recorded, the content of the flight data set or the flight data sampling frequency. The list of a signals in measured flight data is contained in the Appendix A. Probably the most important problem connected with the flight data analysis is the corruption of this data with noise and errors. Therefore Section 2.5 is dedicated to the calibration of sensors and contains a list of typical errors that affect sensors used in an aircraft.

First, the list of flight data is being divided into particular groups, followed by the discussion over the meaning and significance of each group.

2.2.1 Acceleration

The acceleration along the axes of a Body fixed frame is measured by accelerometers in meters per second $(m \cdot s^{-2})$. An accelerometer is a compact instrument, which measures non-gravitational linear acceleration [19]. The accelerometers are widely used for example in the inertial navigation.

The location of an accelerometer is very important when measuring the acceleration in BFF. The exact location of the accelerometer is needed and then the measured acceleration is being compensated due to the off Centre of Gravity (C.G.) location. The required compensation will be discussed later.

2.2.2 Angular velocity

The angular velocity denotes the axis and angular speed of rotating object. The unit of the angular velocity is radian per second. The angular velocity of aircraft in body fixed reference frame is measured in three spatial dimensions by a gyroscope. A gyroscope is a device which measures or maintains the rotational motion of an object [19].

2.2.3 Euler angles

Three Euler angles are $\Phi(Phi)$ - representing roll angle, $\theta(Theta)$ - representing pitch angle and $\Psi(Psi)$ - representing yaw angle [5]. The angles are measured in radians.

Measurements are performed by a gyroscope, as well. The gyroscopes are included in the Attitude and Heading Reference System (AHRS). This system measures position angles and angular velocities in all three axes. The attitude heading reference system provides roll, pitch and yaw information, that are necessary for guidance, state estimation, control and navigation in military, spacecraft, commercial or civil applications. Traditional gimballed inertial navigation systems based on mechanical gyroscopes were replaced by Microelectromechanical System (MEMS) technology. The MEMS technology has the advantage of lower size, complexity, weight and price. Hardware structure of a MEMS attitude heading
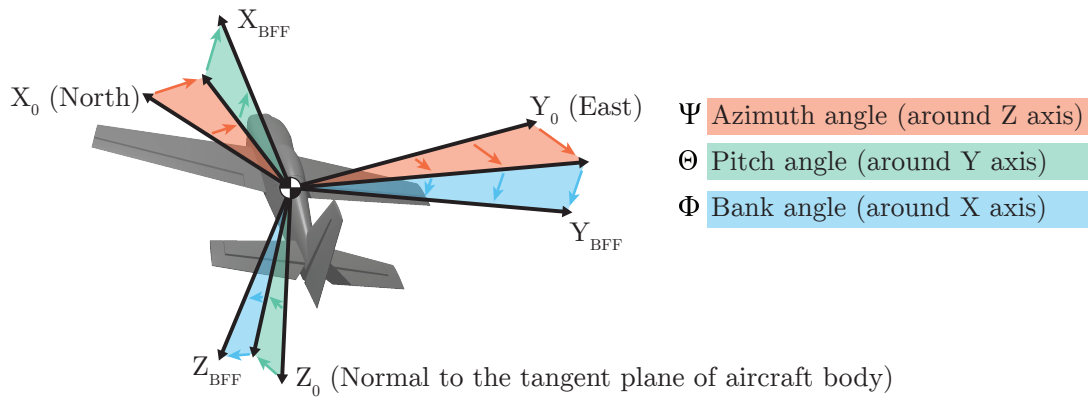
Figure 2.7: Illustration of Euler angles

reference system contains magnetic and inertial sensors, or rather MEMS integrated circuits of rate gyroscopes, magnetometers and accelerometers for all three Cartesian axes. Rotation rates are measured directly by rate gyroscopes and then integrated to get the rotation Euler angles. The MEMS sensors compared with fibre-optic or mechanical gyroscopes are affected by drifts and measurement errors.

The Euler angles and angular velocities are the required fundamentals when defining the position of an aircraft. With a given attitude and acceleration we can determine the trajectory of an aircraft.

2.2.4 Aerodynamic angles

The Angle of Attack (denoted as α) and the Angle of Sideslip (denoted as β) are the aerodynamic angles measured in radians [5]. The Angle of Attack is defined as the angle measured between the longitudinal axis of an aircraft and the relative motion of the air (measured in the $Y - Z$ plane of Body Fixed Frame). Measurement of the Angle of Attack is performed by the Angle of Attack probe.

Figure 2.8: Angle of Attack

The Angle of Sideslip is defined as the angle measured between the longitudinal axis of an aircraft and the relative motion of the air (measured in the $X - Z$ plane of body fixed frame). The Angle of Sideslip probe is used to measure the angle.

Figure 2.9: Angle of Sideslip

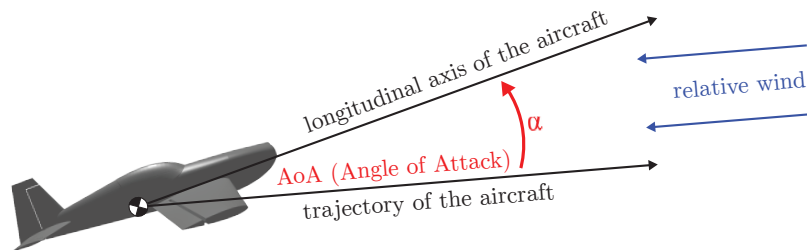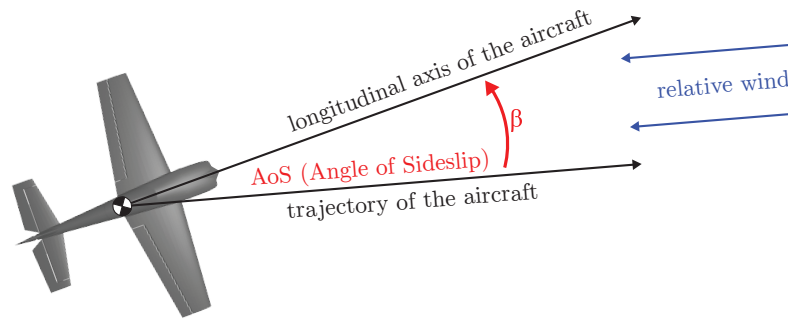Both probes are not usually located at the Centre of Gravity of an aircraft. Common location is at the front of an aircraft or under the wing of an aircraft. The compensation of the aerodynamic angles has to be taken into consideration, too.

2.2.5 Force controls

The group of force control parameters in flight data includes elevator control force, aileron control force, right pedal control force and left pedal control force. All are measured in Newtons. These control parameters give us the option to trace back the actions of the pilot during maneuvering.

2.2.6 Deflection of control surfaces

The deflection of the rudder, aileron and elevator belong into the group of deflections of aerodynamic control surfaces, which are represented by the angles measured in radians [2]. Light, sport category aircraft still use mechanical flight control systems because aerodynamic forces acting on the aircraft body are not excessive. More sophisticated approaches use hydro-mechanical designs with mechanical and hydraulic circuits or „fly-by-wire" system where the physical connection between the pilot and the control surfaces is replaced by an electrical interface.

An elevator and an horizontal stabilizer are for the majority of aircraft located at the rear of the fuselage. Stability of the aircraft is provided by the horizontal stabilizer. It prevents the pitching motion (up-and-down) of the aircraft nose. The elevator is a hinged section located at the rear of the horizontal stabilizer. There are usually two elevators situated at each side of the horizontal stabilizer. Both elevators work together, so when the left elevator goes up, the right elevator also goes up as well. The movement of the elevator influences the amount of force generated by the tail surface of the fuselage. The elevators are used during a banked turn to perform tighter turn by increasing the lift force. Moreover the elevators are useful during a take off when they are necessary to initiate a climb.

The ailerons are hinged sections located at the trailing edge of each wing. A rolling motion of an aircraft is generated by the ailerons. The ailerons usually operate in difference mode, i.e. if one aileron is deflected up the other aileron is deflected down by the different amount. When lift on one wing decreases, the lift on the opposite wing increases. The torque about the Center of Gravity is created by the lift forces, which are not equal. The ailerons are used to rotate about the roll axis of the aircraft and so perform the turns. The

resultant rolling moment is a function of aileron deflection, aileron size, and its distance from the centerline of aircraft fuselage.

The rudder and vertical stabilizer are located at the rear end of the fuselage. The vertical stabilizer provides lateral directional stability of the aircraft by preventing yawing motion of the aircraft nose. The rudder is a movable surface section located at the rear of the vertical stabilizer. Movement of the rudder influences the amount of the force generated by tail surface. The rudder is used to generate and control yawing motion of the aircraft. Main function of the rudder is not turning but compensating unbalanced side force component created by the banking of the aircraft.

The Figure 2.10 illustrates the location of the aerodynamic control surfaces.
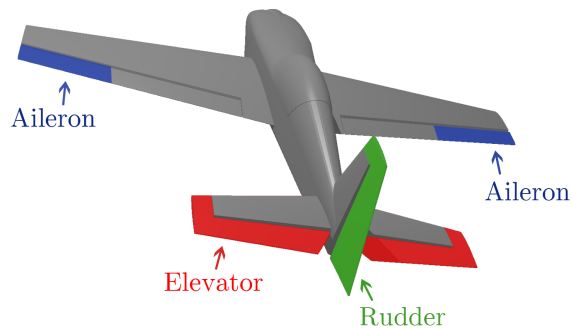
Figure 2.10: Aerodynamic control surfaces of the aircraft.

2.2.7 Barometric data

The static pressure and dynamic pressure measured in Pascals belong to barometric data. Among the barometric data, we can find the outside air temperature and the engine inlet air temperature measured in Kelvins as well.

The Pitot static tube [10] is a pressure measurement instrument widely used to determine the airspeed of an aircraft. The Pitot static tube is shown in the Figure 2.11.
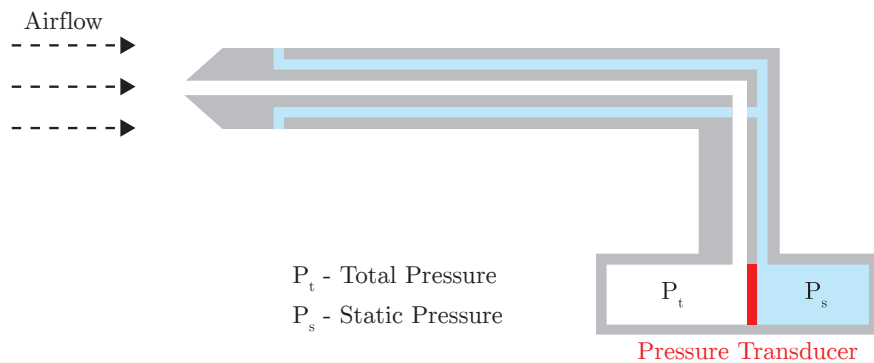
Figure 2.11: Pitot static tube

The centre hole is located in the axis of a tube and several holes are drilled around the outer tube surface. The central hole is pointing in the direction of the flow and outside holes are perpendicular to the centre tube. The outside holes are connected to one pressure chamber; the centre hole is kept separate from outer holes and connected to another pressure chamber. In between these pressure chambers is a pressure transducer. The transducer measures the difference between pressures in chambers by measuring the strain on a thin element. The pressure in the outside holes is the static pressure (denoted as P_s) because of the position of these holes. They are perpendicular to the direction of the flow. The centre tube is pressurized by both components of the air velocity, the random and the ordered and so the centre tube measures the total pressure (denoted as P_t) [10].

2.2.8 Magnetic field parameters

The magnetic field is measured in three axes by magnetometers. The output of the magnetometers is the magnetic heading information and in a combination with the output from the inertial sensors we can estimate a 3D position of the aircraft. In order to obtain the assumed position of an aircraft magnetometers have to be calibrated for the magnetic distortions and sensor errors.

2.2.9 GPS data

The data received from the Global Positioning System belong to this group. Global Navigation Satellite System (GNSS) is used in Satellite Navigation to determine the position and the time anywhere on Earth precisely. The Global Position System (GPS) was developed and operated by the United States Department of Defence (DoD). Apart from the GPS, there are some other alternatives; in Europe, there is system GALILEO and in Russia, there is GLONASS system. These alternatives have been developed, because the GPS system is dependent on one country, the United States.

The full name of the GPS is NAVSTAR-GPS (*NAVigation System with Timing and Ranging Global Position System*). The three main goals while creating the GPS [21] were:

1. provision of the capability of determining the position, speed and time, whether in motion or at rest,

2. global, continuous and all-weather three dimensional positioning with a high degree of accuracy,

3. potential for civilian use.

To determine the coordinates in a three dimensional space (longitude, latitude, altitude) and an exact time we need signal from four satellites (time-signal transmitters). The distance in GPS is computed from the travel time between satellite and our position and the fourth satellite assist with the clock synchronization.

The format of GPS coordinates and the complication of the altitude specification is described in Section 2.1.2.

2.2.10 Engine parameters

The engine parameters are engine revolutions per second, fuel flow to engine and manifold pressure. The engine revolutions help us to estimate the thrust generated by the propulsion

system. There is also an efficiency factor related to fuel flow to engine, called the *thrust specific fuel consumption* which characterizes the fuel efficiency of an engine. The absolute pressure in the engine intake manifold is indicated by the manifold pressure gauge. The torque produced by the aircraft engine and the value of the intake manifold pressure are in correlation, when the engine operates at normal conditions.

2.3   Forces acting on the aircraft

This section describes the forces that act upon aircraft during flight. Comprehension of how these forces work is essential to understanding flight maneuvers.

Airmass flowing past an aircraft must be diverted from its original path, and such deflections lead to changes in the speed of the air [16]. There are only two ways that the airmass and an aircraft can exert upon each other. While aircraft moves forward the molecules of the air slide over its surface. The molecules that are closest to the surface of the aircraft fuselage act as they are stuck to the surface and they move with the aircraft. There is a slip between non-moving molecules away from the aircraft and the molecules closest to the aircraft, which are moving with it. The *viscosity* is the tendency of air to resist shear deformation and so additional air near the surface of aircraft is dragged along the flightpath. An aircraft experiences a resultant aerodynamic forces and moments as a result of these processes.

The Figure 2.12 illustrates the forces acting on an aircraft.
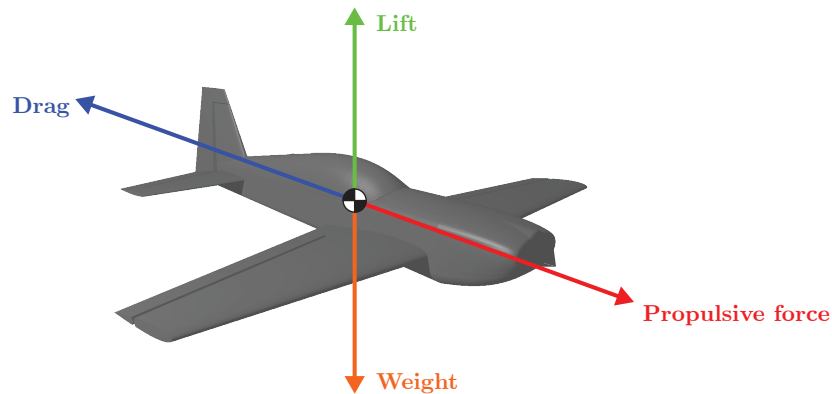
Figure 2.12: Forces acting on the aircraft.

2.3.1   Thrust

The trust is the forward propulsive force that is invariably obtained by increasing the rearward momentum of gas quantity. Usually it acts in parallel with longitudinal axis of an aircraft. Forward movement of aircraft is caused by thrust force that is larger than drag force. If thrust force and drag force are equal, the aircraft moves with constant airspeed. If the engine power is decreased during the level flight, the aircraft slows down. The aircraft decelerates as long as the thrust is less than the drag until the moment when its airspeed is insufficient to support it in the air. There are various types of straight-and-level flight conditions according to airspeed of an aircraft. For example, one type of straight-and-level

flight is possible at low airspeed, but the Angle of Attack (AOA) must be relatively high to sustain balance between lift and Gravitational force. If there is no compensation on AOA with an increase of thrust force, the aircraft starts to climb. It is important to coordinate the AOA when we want to perform straight- and-level flight.

2.3.2 Drag

The drag is a force that exerts in opposite direction than trust. The airspeed of an aircraft is reduced by the drag. There are many types of drag and we can define these types based upon the origin of the drag force. Drag force, which is not strongly related to lift, is usually denoted as parasite drag [10].

Parasite drag includes the turbulence generate in airmass, hindrance of air that moves over the surface of aircraft fuselage or displacement of the air by an aircraft. The parasite drag has three forms: skin friction drag, interference drag and form drag.

Skin friction drag is formed by the aerodynamic resistance of the surface of an aircraft. The air molecules that flow in the air around the surface of an aircraft have free-stream velocity. The boundary layer is located closer to the surface of an aircraft in between the wing and free-stream velocity level. An effective shape of an object is created by the boundary layer and it can be different from the real physical shape. Aircraft designers try to remove any irregularities from the surface of the wings of an aircraft to reduce skin friction drag.

Interference drag arises in the intersections of airmass, for example in the intersection of the fuselage with the wing. Perpendicular angles create the most of interference drag.

Form drag is also denoted as viscous separation drag and depends on the location of the separation point of the aircraft fuselage. When the separation point is located nearer to the front of fuselage the drag is much higher as this point is located at the rear of the fuselage.

Induced drag is rearward component of the lift and it is related with effectiveness of system. Required work of system is always supplemented with some additional work which is lost or dissipated.

2.3.3 Gravitational Force

Gravitational force is caused by gravity and it is concentrated in the Center of Gravity (CG). The gravitational force follows the direction of gravitational acceleration. When an aircraft is in a state of equilibrium and flies in straight-and-level flight the lift force is equal to the gravitational force and there is no change of altitude.

2.3.4 Lift

Lift force is controlled by a pilot and is related to the Angle of Attack (AOA). The aerofoil shape of wings is effective when an aircraft is moving. The lift force is proportional to the spare of velocity of an aircraft, so when the velocity is increased the lift force is increased, too. Lift and drag are dependent on the density of the air. The pressure, humidity and temperature affect density of the air and there is less air density in higher altitude than it is in a lower altitude. If pilots want to maintain same lift force on aircraft in a higher altitude, they must fly at greater airspeed. The amount of the lift force and the velocity of an aircraft are two main aerodynamics factors, which are to be controlled by pilots. Typical situation

arises in straight-and-level flight when pilot maintains constant altitude by adjusting lift to sustain required velocity.

2.3.5   Aerodynamic Forces in Flight Maneuvers

Aerodynamic forces change during the flight maneuvers. The Newton First Law of Motion states that an object at rest stays at rest and an object in motion stays in motion with the same speed and in the same direction unless acted upon by an external force [10]. If we apply this law to the steady motion (straight-level flight) of the aircraft we can expect that some additional side force is required for the aircraft to execute the turn maneuver. The lift force can be for clarity divided into two components during the turn, the horizontal and the vertical component of the lift force. The horizontal component of the lift force exerts in horizontal direction towards the center of the turn. The aircraft is in yawing motion and executes the turn maneuver. The centrifugal force acts in opposite direction to the horizontal component of the lift force. This force is created as a reaction to the change of the aircraft flight direction and it pushes the aircraft out of the turn. The vertical component of the lift force acts in opposite direction to the Gravitational force.

In the well coordinated turn, the horizontal and the centrifugal force are equal. The nose and the tail of an aircraft follow the same path. If there is unbalance between forces, that act on an aircraft during the turn, then a skidding or a slipping motion is present. Skidding and slipping turns are described below in Subsection 2.4.2.

2.4   Flight events

In this section we introduce two types of flight events, the events suitable for tests and the events represented as a phase of flight. The following paragraphs are intended to give the reader brief concept about flight test and the typical testing maneuvers.

2.4.1   Flight tests

This section of thesis describes flight tests and testing maneuvers performed during flight tests. The main objective of the flight test is to obtain or record the response of the aircraft during and after a control system induced pulses or doublets [17]. This approach could be compared with performing of step response test on a system under investigation. Pulses and doublets of the elevator, aileron or rudder are indicated as *testing maneuvers*. It is important to examine flight data during and after the test to receive specific and complete response of an investigated aircraft. It is difficult to perform suitable testing maneuvers that meet required conditions. The testing maneuver can be initiated only during a steady level flight, where the airspeed, height, vertical speed are constant and there is no pitching, yawing, or rolling motion. Another condition that must be met is the speed and the shape of the intended pulse or doublet. It is important to perform a pulse or a doublet only through one control effector. For example, when performing a testing maneuver using the elevator, the pilot can not command the aileron or rudder during the flight test. When considering all above mentioned conditions, one must agree with the fact that performing the test events is very difficult task, requiring an experienced pilot.

The flight test are performed under certain airspeed and altitude. Usually, there are several tests performed at different altitude and different airspeed conditions during one

flight. The flight test can take place in the aircraft under real flight conditions, in flight simulator or using a scaled aircraft in the wind-tunnel.

Pulses

Figure 2.13 illustrates an idealised pulse in the deflection of a elevator, aileron or rudder. Actual part of the flight data with pulse in the deflection of the elevator is shown on the right side of the figure for comparison. Real flight tests usually examine the behaviour for a complex set of observed signals. For example, the object of the examination could be a roll rate response to an aileron pulse.
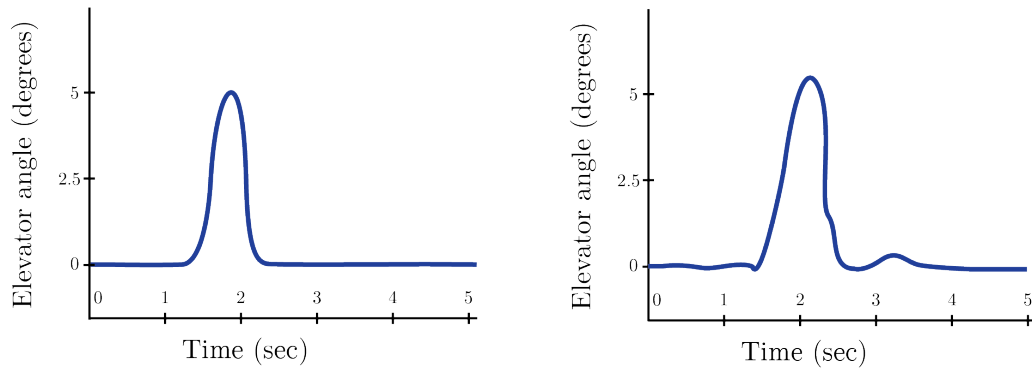
Figure 2.13: Pulse in the deflection of the elevator.

Doublets

Figure 2.14 illustrates an idealised doublet in the deflection of the elevator, aileron or rudder. Part of the flight data with doublet in elevator is shown for comparison on the right side of the figure. We can observe differences between the idealised and the actual records in the shape and the duration of the impulse.
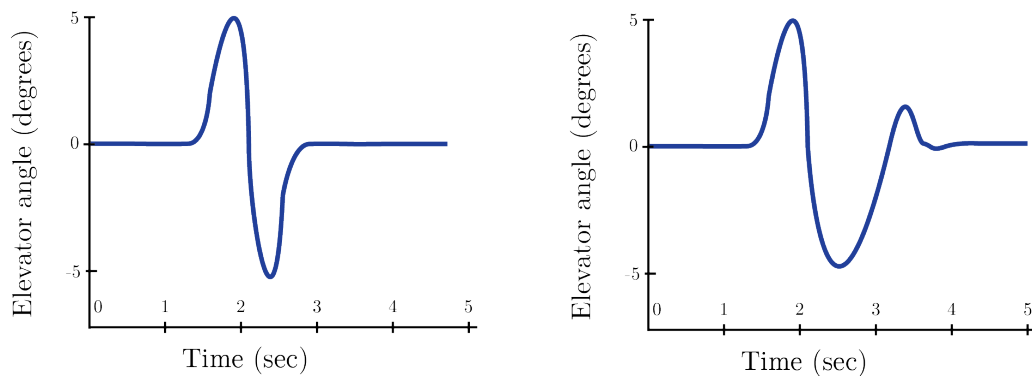

Figure 2.14: Doublet in the deflection of the elevator.

19

2.4.2   Phases of flight

Before starting the analysis and identification of testing maneuvers, we have a specification of the main phases of flight and basic flight maneuvers. Basic maneuvers include the climb and descent, which represent the change in the longitudinal motion of an aircraft. There are also the turns, which represent the change in the lateral motion of an aircraft. The following description is divided to paragraphs according to the main phases of flight. The list containing all phases of a flight could be the output of a flight data analysis tool. The identification of particular flight maneuvers can be supplemented with continuous list of the flight phases and the flight maneuvers. In this thesis we focus on the identification of the testing maneuvers and following list of the flight phases is attended to give the reader an overall picture of the flight course.

The en route phase of the flight includes the specification of basic flight maneuvers. For each phase of flight there is a brief description based on [12] and presumable estimation of flight parameters.

Standing

This is the initial or terminal phase of a flight. The engine of an aircraft is not in operation. Standing phase occurs after the arrival at the parking area and it could be prior to taxi or pushback phase. The body of an aircraft is stationary during this phase. We examine engine and acceleration flight parameters to identify this phase of flight.

Pushback / Towing

An aircraft is moving from the runway to the parking area, or gate, assisted by a tow vehicle, the pushback tractor or tug. There is also an option for the aircraft to move on the ground by using reverse thrust, but the resulting jet blast may cause damage to the surrounding objects and buildings.

Taxi

This phase refers to movement of an aircraft on the ground, under its own power. The power back subphase takes place when the aircraft reverses from the parking position. The following is taxi to runway (when the aircraft is moving on its own power); leaving the gate and finish at the runway. Taxi to take-off position starts with entering the runway and lasts until reaching the take-off position. The final taxi phase takes place after landing, when the aircraft goes back from the runway to the parking area. The speed of an aircraft when moving on the ground is limited to approximately $18,5$ kilometres per hour.

Take-off

The take-off phase starts upon taxiing onto the runway and ends when climb-out on the runway is being established. During the take-off phase the engine operates at its full performance. The end of the take-off phase is determined to an altitude of 35 feet above the runway elevation [2]. Two important speed measurements are made on the runway. First is the highest ground speed at which an aborted take-off is possible, second is the airspeed at which the front of an aircraft is being lifted. Moreover, the barometric altitude rate or GPS-derived altitude rate is monitored and measured.

Initial Climb

The takeoff phase is followed by the initial climb until the specific altitude is being reached. The lift force on the wings of an aircraft exceeds weight of aircraft and allows a climb. The important parameters that induce a climb are the angle of attack and the engine thrust.

En route

The en route phase leads from initial airport to the destination airport. In case of commercial flight, this phase usually takes the majority of the travel and it is the fuel most efficient phase of the flight. Some of the basic flight maneuvers that occur during the flight are listed below. We can classify flight maneuvers in terms of speed, climbs, descents and turns. Important parameters to specify particular flight maneuver are the yaw rate, the vertical speed, the angle of sideslip and the airspeed.

- Shallow turns are those in which the roll angle is less than 20 degrees and the aileron has to be applied to maintain the roll angle.

- Medium turns results from a roll angle in range from 20 degrees to 45 degrees. An aircraft remains at same roll angle during medium turns.

- Steep turns are those resulting from a roll angle equal to 45 degrees or more. At the steep turn the roll angle disrupt stability and the aileron is applied to prevent it.

- In a coordinated turn there is no skidding or slipping [1]. The skidding turn is characterized by excess of yaw in the direction of turn. The nose of an aircraft heads to the middle of the turn and the air flows crosswise over the fuselage. There is an excess of the centrifugal force over the horizontal lift component and the aircraft is pulled out of the turn. The pilot has to apply the rudder to correct a skidding turn.
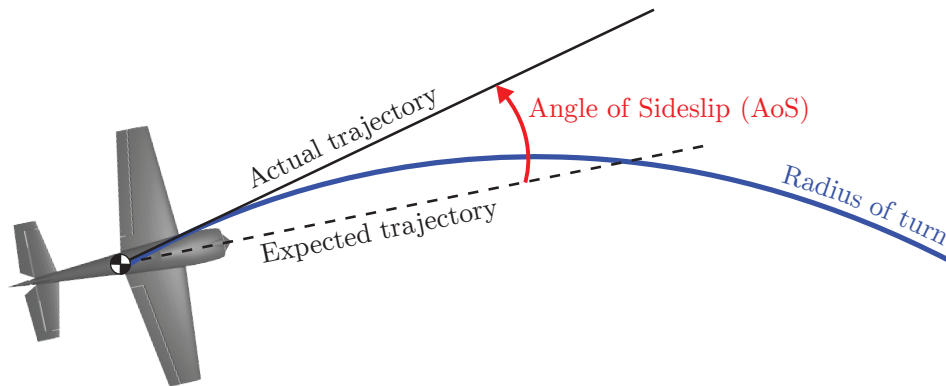

Figure 2.15: Illustration of skidding turn

- In a slipping turn the aircraft falls to the inside of the turn. This effect is induced by a horizontal lift component which is larger than the centrifugal force. The aircraft is banked too much for the rate of the turn and thus the slipping turn is not a coordinated turn.
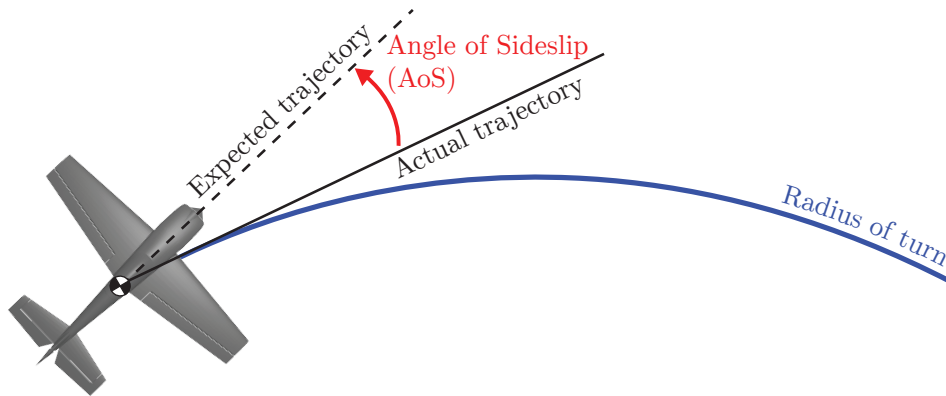
Figure 2.16: Illustration of slipping turn

Approach

The approach phase begins, when the specific altitude (typically *15 metres* above the ground) is reached, the flaps are fully extended and the landing gear is down. During this phase of flight, the constant airspeed and constant angle of descent are ensured. The angle of descent is controlled by the pilot varying the engine power and pitch angle to keep the constant airspeed. The meteorological conditions are important during the approach phase, there is the crosswind limitation (from 25 to 35 knots) and its value could be even reduced by low visibility or runway conditions (wet, dry).

Landing

At the landing, one of the last phases of flight, the aircraft lands on the runway. This phase begins when the runway is in sight and ends when the aircraft leaves the runway. The airspeed and the rate of descent have to allow smooth landing. The reduction of thrust and increase of drag using flaps cause that the aircraft constantly slows down. Wheel braking, reverse thrust and ground spoilers are used immediately after touching the runway to decelerate the desired speed and leave the runway. The reverse thrust is applied in higher speeds and the wheel brakes are used at low speeds.

2.5 Data Analysis

The goal of this thesis is to develop system for identification of testing maneuvers. In previous sections we have defined theoretical basis that we can use to design the proposed system. The complexity of proposed system has to deal with the fact, that system is proposed for research and prevention purposes. Input of the system, the flight data will be available to analysis only after the flight. Resulting system should offer the user quick and brief overview of the testing maneuvers. Before we move to the classification process there is the stage of the data analysis. This section includes the error characteristics of the inertial sensors.

2.5.1 Error Characteristics of Inertial Sensors

The inertial navigation is a technique for determining an aircraft velocity and position by measuring its acceleration and processing this measured value in a computer [7].

The basis of the inertial navigation are the accelerometers and angular rate sensors which can be found in different physical construction to measure specific angular rate or force. All types of gyroscopes and accelerometers are exposed to cross-coupling, biases, scale factor errors and random noise to a certain extent [8]. Each systematic error has four components: a fixed contribution, a temperature- dependent variation, a run-to-run variation and in-run variation. The fixed contribution error is present with every use of the sensor and it could be corrected by the Inertial Measurement Unit (IMU) using laboratory calibration data. The temperature-dependent error component can be corrected by the IMU. The run-to-run variation error component is different each time the sensor is used but remains constant within any run. It is possible to correct the run-to- ran variation using integration algorithms. The in-run variation contribution to the error source changes slowly during the flight and to correct it, the integration with other navigation sensors is used. Besides contribution errors listed above, there are also residual fixed contributions left from the sensor calibration process.

Constant Bias

The *bias* is a constant error exhibited by all gyroscopes and accelerometers. In most cases, the bias produces dominant contribution to the overall error of an inertial sensor. Vectors $b_a = (b_{a,x}, b_{a,y}, b_{a,z})$ and $b_g = (b_{g,x}, b_{g,y}, b_{g,z})$ denote the accelerometer and gyroscopes biases, respectively. The errors of the inertial sensors are expressed in the axes of the Body Fixed Frame. It is useful to split the biases into dynamic and static components. The dynamic component of bias comprises in- run variation and the static component of bias comprise run-to-run variation. Bias of the accelerometer is quoted in terms of the acceleration due to gravity (denoted as g). Gravitational acceleration [3]: $1g = 9.80665 m.s^{-2}$ is usually expressed in units of mili-g (mg) or micro-g (μg). Gyroscope biases are expressed in degrees per hour (deg/hr or $°hr^{-1}$), where $1°hr^{-1} = 4.848 \times 10^{-6} rads^{-1}$

Cross-Coupling and Scale Factor Errors

The *scale factor* is a calibration error which causes the variation of the input-output gradient of the instrument from a unity following unit conversion by the IMU. The gyroscope output error due to the scale factor is proportional to the angular rate around the sensitive axis, while the acceleration output error is proportional to the specific force along the sensitive axis. Vectors $s_a = (s_{a,x}, s_{a,y}, s_{a,z})$ and $s_g = (s_{g,x}, s_{g,y}, s_{g,z})$ denote the accelerometer and the gyroscopes scale factor errors, respectively. The *cross coupling* errors arise from the misalignment of the sensitive axis of the inertial sensors with axes of the Body Fixed Frame. Scale factor and cross coupling errors are between 10^{-4} and 10^{-3} (*100-1000 parts per milion - ppm*).

Random Noise

Random noise from a number of sources influences all inertial sensors. The resolution of an inertial sensor is limited by the electrical noise. Vectors $r_a = (r_{a,x}, r_{a,y}, r_{a,z})$ and $r_g = (r_{g,x}, r_{g,y}, r_{g,z})$ denote the accelerometer and the gyroscopes random noise errors, respectively. The noise of an inertial sensor is usually expressed in terms of the root *Power Spectral*

Density (PSD). Accelerometer random noise units is $\mu g/\sqrt{Hz}$, where $1\mu g/\sqrt{Hz} = 9.80665 \times 10^{-5} ms^{-1.5}$. Gyroscope random noise unit is $°/\sqrt{hr}$ or $°/hr/\sqrt{Hz}$ where $1°/\sqrt{hr} = 2.909 \times 10^{-4} rad\ s^{-0.5}$ and $1°/hr/\sqrt{Hz} = 4.848 \times 10^{-6} rad\ s^{-0.5}$. We obtain standard deviation of the random noise samples (denoted as SD or σ) by multiplying the sampling rate with corresponding root PSDs.

Error effects caused by the environment [18]

- **Temperature** - An accelerometer is designated to give the highest accuracy factor in a specific given range of temperature. Otherwise if environment does not satisfy the temperature conditions the accelerometers should work with a lower accuracy. Special treatment is necessary when situating the accelerometer on board of aircraft. The heat required to maintain the accelerometer operating temperature has to be supplied in symmetrical manner, if not then improper application can cause temperature changes leading to errors. Typical example of described situation occurs when aircraft starts and sensors including accelerometers are warmed up rapidly to its operating temperature and just after a few minutes a stable output suitable for navigator alignment is necessary.

- **Vibration.** When dealing with an aircraft, a vibration is oscillating or any other periodic motion of an aircraft body forced from a state of equilibrium. The harmonic vibration has constant frequency and magnitude. The accelerometers are often used in environment where the vibration conditions may be severe, circular and angular vibration may be present in aircraft. It is difficult to calculate the exact effects of such vibrations.

3 Data classification

We meet the process of classification in everyday life. In [14] Michi, Spiegelhalter and Taylor proposed, that the term of classification could cover any context in which some decision or forecast is made on the basis of currently available information. The classification procedure is then some formal method for repeatedly making such judgements in new situations. More specific definition of classification from [14] says: „Classification is the problem that concerns the construction of a procedure that will be applied to a continuing sequence of cases, in which each new case must be assigned to one of a set of predefined classes on the basis of observed attributes or features."

We distinguish two different approaches to the task of classification. First, the pattern recognition or supervised learning, where a set of training data and its classes are known and available before a classification procedure. A classification procedure is build above provided training data and classes. Second, there is clustering or unsupervised learning, when a classification procedure is derived from the data.

The concept of supervised flight data classification is shown in the Figure 3.1. There are two databases of flight events which provide data to classifier. The classifier assign label to new flight data. The result of classification is used in further application and also to update the databases of flight events. Further application represents, for example, the visualization of the classification results or further study of an aircraft behaviour during the flight events.
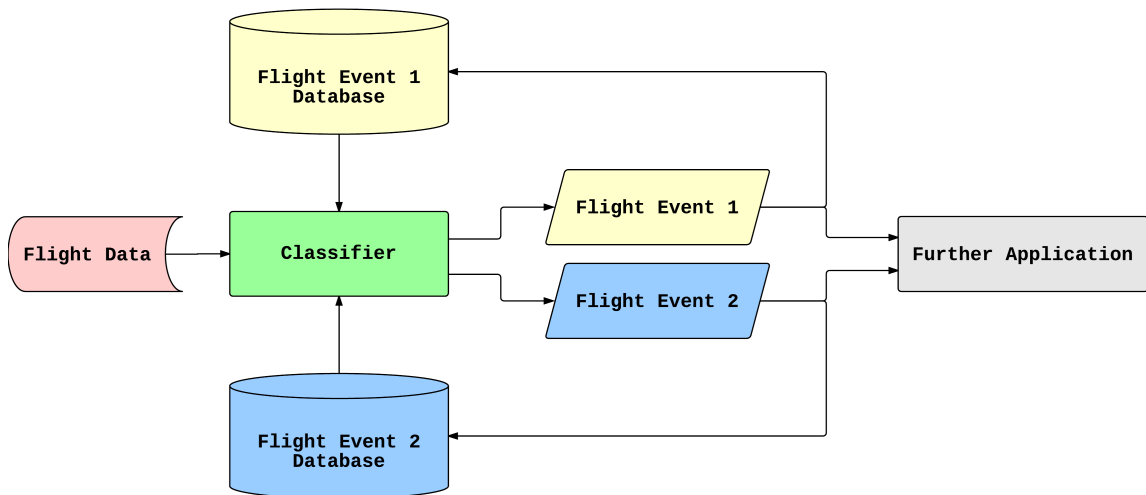

Figure 3.1: The concept of supervised flight data classification.

The conceptual scheme for unsupervised learning would be similar to the one for supervised learning except for the databases for flight events. Usually, the classifier estimates the parameters of probability density function of the data being classified and make a decision based on the estimated parameters.

In this thesis, the objective is to design a classifier that will assign appropriate classes to measured flight data. Three major issues have to be solved, the approach to the classification, the definition of classes for classification and the selection of features with highest discriminative power for the classification. Nowadays, the most favoured approach would be one of these two: statistical or neural- network. The thesis focuses mainly on statistical analysis and leaves the other approach to the possible future research. This work requires approach with unsupervised learning due to the format of flight data, when it is difficult task to prepare training data set for the classification of the test maneuvers. Gaussian Mixture Models (GMMs) with EM algorithm are among the most statistically mature method used for unsupervised learning. Our decision to choose GMM to classify the flight data was approved by statistical analysis of flight data performed in Matlab. Statistical analysis proved that measured flight data has the normal distribution. Besides that, density estimation with GMMs is a popular approach which often produces good results and it is computationally manageable.

Next sections provide the fundamental basis in terminology used within the process of classification, the overview of probability theory and then introduce the Gaussian Mixture Models with all definitions required to understand this concept.

3.1 Pattern Recognition

The pattern recognition has application in classification of geometric figures or in waveform classification [6]. In our case we want to classify a flight data into two classes, normal flight and testing maneuver. Simple approach reduces this task to discrimination of flight data from normal flight and testing maneuvers. We also need the observation of flight data over a period of time.

At the beginning, we have time-sampled values of waveform $x(t_1), \ldots, x(t_n)$. A vector X consist of n measurements. The value of each measurement $x(t_i)$ is *random variable* and so X is *random vector*. Many waveforms form a distribution of X in the n-dimensional space. If the distributions of two classes of X are known from past experience, we can find a boundary $g(x_1, x_2) = 0$ between two classes. This boundary divides the two dimensional space into two regions. New samples without a class label could be classified on the basis of selected boundary. The decision is dependent on the sign of *discriminant function* $g(x_1, x_2)$. A network that detects a sign of *discriminant* function is denoted as *pattern recognition network* or a *classifier*. The process of *training* or *learning* includes study of the characteristics of the distribution of X for each class. Hence, the *pattern recognition* is a process of estimating density function in a n-dimensional space and then dividing that space into smaller regions of classes.

3.1.1 Terminology

In this section we introduce terminology of pattern recognition in more detail:

- The term **pattern** denotes the n-dimensional data vector $x(t_1), \ldots, x(t_n)$ of measurements or time-sampled values. Each component $x_{(t_i)}$ is measurements of features of

an object.

- **Classes** or groups are denoted as $\omega_1, \ldots, \omega_C$, where C is number of classes. Each pattern has associated class and categorical variable z denotes this class so, if $z = i$, then the pattern belong to ω_i, $i \in \{1, \ldots, C\}$.

- **Training set** is large set of N patterns and it is used to determine the parameters of an adaptive model.

- **Target vector** represents the class of corresponding pattern and it could be created by hand-labelling the data in *training set*. There is one target vector for each pattern.

- **Test set** contains new data that will be used after training of the classifier. The term *generalization* refers to the ability of the classifier to categorize correctly new data that differs from data in the training set. Generalization is usually the main objective of pattern recognition.

- **Feature extraction** is a method of preprocessing that is applied to training data. The goal of feature extraction is to reduce the variability and size complexity of the training data while maintaining its discriminatory information.

- **Supervised learning** or machine learning takes a known set of training data and corresponding classes to the training data, and then design an adaptive model that will correctly classify new data.

- One goal of **unsupervised learning** is to discover the classes (similarities) of training data without any corresponding classes. This approach is called *clustering*. Another approach is known as *density estimation* and it determines probability distribution of the training data.

3.2 Probability Theory

This section presents a fundamental basis required for pattern recognition. Probability has significant impact in the field of pattern recognition due to finite set of training data and noise present in the measurements as well. [4] The section starts with definition of basic concepts in the probability theory and then move further to advanced theory. Extra emphasis is placed on definition of *Gaussian Distribution* and *Mixture of Gaussian*

At the beginning we start with definition of the sum and product rules for two random variables, X and Y. X takes values $\{x_i\}$ where $i = 1, \ldots, M$ and Y takes values $\{y_j\}$ where $j = 1, \ldots, L$. The Figure 2.12 illustrates the table with 4 columns and 3 rows, hence $M = 4$ and $L = 3$. N is the total number of values of variables X and Y. The number of values in particular cell where $X = x_i$ and $Y = y_j$ is denoted as n_{ij}. Similarly c_i denotes the number of values in column i where $X = x_i$ and r_j denotes the number of values in row j where $Y = y_j$.

The probability that random variable X will take the value x_i and Y will take the value y_j is defined as $p(X = x_i, Y = y_j)$ and denoted as *joint* probability of $X = x_i$ and $Y = y_j$. The following formula define joint probability as a fraction where in numerator is number of values in cell i, j and in denominator is total number of values N.

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \tag{3.1}$$

The probability that X will take value x_i independently from value of Y is denoted as $p(X = x_i)$.

$$p(X = x_i) = \frac{c_i}{N} \tag{3.2}$$

This probability is given by fraction of the total number of values to number of values in column i, which can be represented by the sum $c_i = \sum_j n_{ij}$. From Equations 3.1 and 3.2 we can derive sum rule of probability:

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j) \tag{3.3}$$

The *conditional probability* logically follows foregoing definition, where we take only such values that $X = x_i$ and $Y = y_j$. This probability is denoted as $p(Y = y_j | X = x_i)$ and is called the *conditional probability* of $Y = y_j$ given $X = x_i$.

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i} \tag{3.4}$$

We can derive the *product rule* of probability from Equations 3.4, 3.2 and 3.4:

$$p(X = x_i, Y = y_j) = p(Y = y_j | X = x_i)p(X = x_i) \tag{3.5}$$

Bayes' theorem could be derived from the Product Rule 3.5 using symmetry property $p(X = x_i, Y = y_j) = p(Y = y_j, X = x_i)$:

$$p(Y = y_j | X = x_i) = \frac{p(X = x_i | Y = y_j)p(Y = y_j)}{p(X = x_i)} \tag{3.6}$$

Denominator in Bayes' theorem 3.6 has function as the normalization constant which ensures that sum of the conditional probability over all values of $Y = y_j$ will be equal to one. With application the Sum Rule to 3.6 we get:

$$p(X = x_i) = \sum_{y_j} p(X = x_i | Y = y_j)p(Y = y_j) \tag{3.7}$$

At this point the we should define *prior probability* and *posterior probability* as the interpretation of Bayes' theorem is. A *prior probability* is available before some measurements or observation are taken into account. It is a description of what is known a priori about the parameter to be estimated. After estimation of parameter we can use Bayes' theorem to compute the *posterior probability*.

3.2.1 Probability densities

The probability that continuous random variable x will lie in an interval (a, b) is called *probability density* and it is defined as:

$$p(x \in (a, b)) = \int_a^b p(x)dx \tag{3.8}$$

The probability density must satisfy two conditions:

$$p(x) \geq 0 \tag{3.9}$$

$$\int_{-\infty}^{\infty} p(x)dx = 1. \tag{3.10}$$

First conditions arises because of probabilities are non-negative and second is connected with the fact that value of continuous random variable x must lie on the real axis.

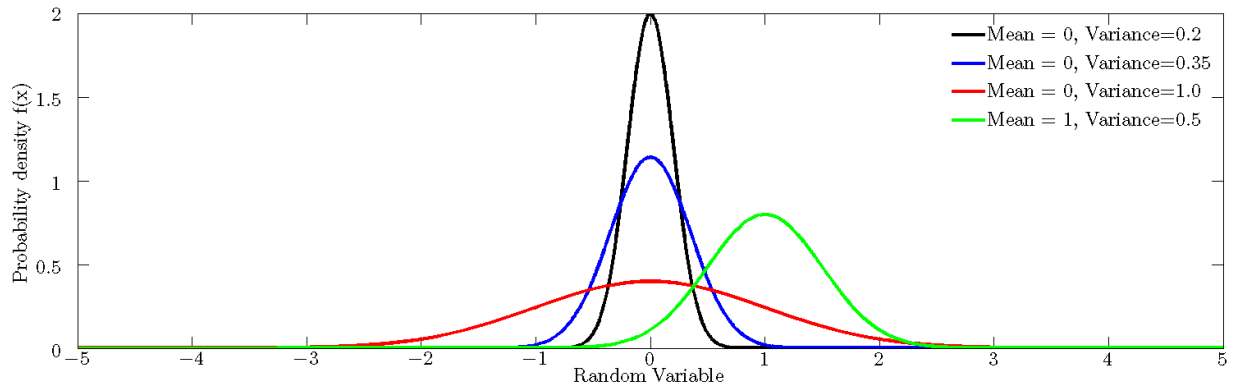Figure 3.2: Probability density function for Gaussian distribution.

The probability density can be expressed as the derivative of a cumulative distribution function $P'(x) = p(x)$.

The *cumulative distribution function* express probability that x lies in the interval $(-\infty, z)$:
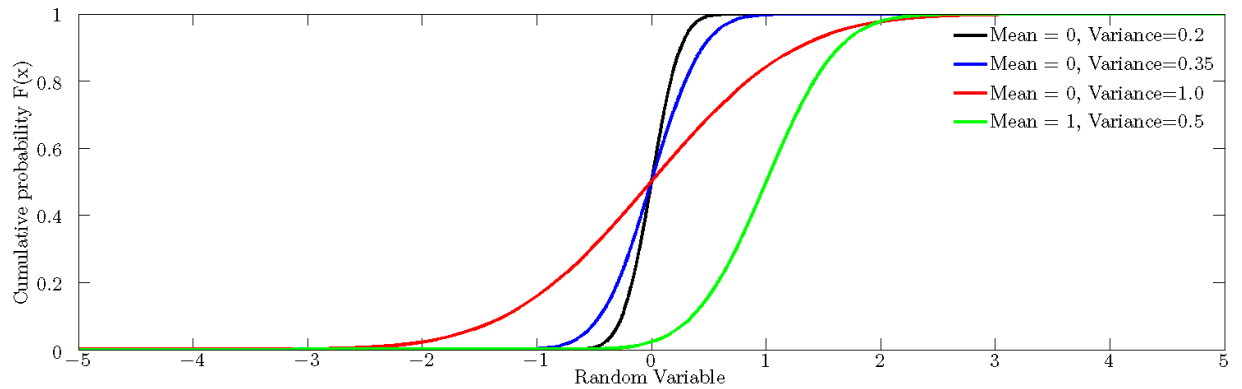
$$P(z) = \int_{-\infty}^{z} p(x)dx \tag{3.11}$$

Figure 3.3: Cumulative distribution function for Gaussian distribution.

3.2.2 Expectations and covariances

The *expectation*, denoted as $\mathbb{E}[f]$ represents the average value of function $f(x)$ under probability distribution $p(x)$. The average value is weighted by relative probabilities of the different values of random variable x. For a discrete random distribution the expectation is defined by:

$$\mathbb{E}[f] = \sum_x p(x)f(x) \tag{3.12}$$

The expectation of continuous variable includes integration and it is given by:

$$\mathbb{E}[f] = \int p(x)f(x)dx \tag{3.13}$$

If we know just N values of probability density (N is finite number), the expectation can be defined by a sum:

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n). \tag{3.14}$$

Conditional expectation is related to a conditional distribution and given by:

$$\mathbb{E}_x[f|y] = \sum_x p(x|y)f(x) \tag{3.15}$$

The *variance* represents a measure of how much the values of $f(x)$ varies from its expectation $\mathbb{E}[f]$ and is defined by:

$$var[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] \tag{3.16}$$

The *covariance* for two variables x and y is defined by:

$$\begin{aligned} cov[x,y] &= \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned} \tag{3.17}$$

The extent in which x and y are different is represented by the covariance. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors of random variables, then the covariance is matrix such:

$$\begin{aligned} cov[\boldsymbol{x}, \boldsymbol{y}] &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\{\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]\}\{\boldsymbol{y}^T - \mathbb{E}[\boldsymbol{y}^T]\}] \\ &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\boldsymbol{x}\boldsymbol{y}^T] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}^T] \end{aligned} \tag{3.18}$$

3.2.3 Bayesian probabilities

Bayesian interpretation of probability allows us to make revisions of uncertain probability using new observations and then make optimal decisions. Bayes' theorem is defined by:

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})}{p(\mathcal{D})} \tag{3.19}$$

where $\mathcal{D} = \{t_1, \ldots, t_N\}$ stands for data observed with the conditional probability $p(\mathcal{D}|\boldsymbol{w})$. This probability is a function of parameter vector $\boldsymbol{w}$ and is denoted as *likelihood function*. Likelihood function describes how probable the observed data set is for different values of the vector $\boldsymbol{w}$. Posterior probability is $p(\boldsymbol{w}|\mathcal{D})$.

3.2.4 Gaussian Distribution

The *Gaussian* or *normal* distribution is very important in statistics, which is confirmed by the *Central Limit Theorem*. The *Central Limit Theorem* states that the sum of a set of multiple random variables, which is a random variable itself, has a probability distribution that becomes Gaussian as the number of terms in the sum increases [6]. It means that the distribution of random sampled data will be normal or close to normal, if the size of random data set is large enough.

The Gaussian distribution for single variable x is defined by two parameters, the *mean* denoted as μ and the *variance* denoted as σ^2. The *standard deviation* is square root of the variance denoted as σ. The formula for Gaussian is given by:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{3.20}$$

Which satisfy the required conditions for a valid probability density function:

$$\mathcal{N}(x|\mu, \sigma^2) > 0 \tag{3.21}$$

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2)dx = 1 \tag{3.22}$$

The *expectation* for a Gaussian distribution is defined by:

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2)x\,dx = \mu \tag{3.23}$$

The definition of second order central moment follows:

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2)x^2\,dx = \mu^2 + \sigma^2 \tag{3.24}$$

The *variance* is derived from 3.23 and 3.24:

$$var[x] = \mathbb{E}[x^2] - \mathbb{E}[x^2] = \sigma^2 \tag{3.25}$$

The Gaussian distribution can be defined over a P-dimensional vector $\boldsymbol{x}$. The mean μ is P-dimensional vector and the covariance Σ is $P \times P$ matrix. Symbol $|\Sigma|$ denotes the determinant of Σ.

$$\mathcal{N}(\boldsymbol{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{P}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\} \tag{3.26}$$

Likelihood function for Gaussian distribution is given by:

$$p(\boldsymbol{x}|\mu, \sigma^2) = \prod_{p=1}^{P} \mathcal{N}(x|\mu, \sigma^2) \tag{3.27}$$

It is usual to find the parameters that maximize the likelihood function when determining the parameters of probability density. There is a problem called *underflow* of the numerical precision of the computer when dealing with maximization of likelihood function in practice. This problem arises in connection with multiplication of large number of small

probabilities. This is the reason to use the log likelihood function derived from 3.20 and 3.27.

$$\ln p(\boldsymbol{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) \tag{3.28}$$

If we maximize 3.28 with the respect to μ we receive *sample mean* or the mean of observed values $\{x_n\}$:

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{3.29}$$

Furthermore if we maximize 3.28 with the respect to σ^2 we receive *sample variance* expressed with respect to the μ_{ML}:

$$\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2 \tag{3.30}$$

3.2.5 Mixture of Gaussian

The data obtained from real measurements typically cannot be described by one simple Gaussian distribution. It is appropriate to define *Mixture of Gaussians* if we want to apply the Gaussian distribution in practice. We can take the advantage and define required number of Gaussians with corresponding means and covariances. Mixture of required amount of Gaussians could describe almost any continuous random distribution with sufficient accuracy. Mixture of K Gaussians is defined by:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\mu_k, \Sigma_k) \tag{3.31}$$

A *component* of the mixture is one Gaussian density distribution $\mathcal{N}(\boldsymbol{x}|\mu_k, \Sigma_k)$ with mean μ_k and covariance Σ_k. The parameters π_k are denoted as *mixing coefficients*.

In the Figure 3.4 we see an example of a Gaussian mixture distribution, where are two Gaussians (blue and green) combined into one mixture distribution (red).
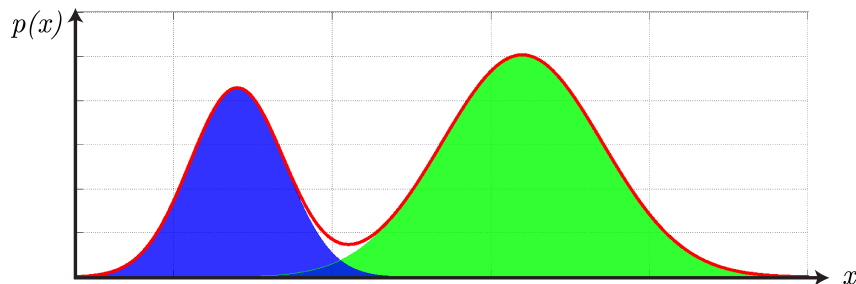

Figure 3.4: Example of Gaussian mixture distribution.

3.3 Gaussian Mixture Models

There are two basic limitations of the Gaussian distribution. First, it is not possible to define too many parameters of this distribution and second, there is limited range of distributions that could be represented sufficiently. [9] These issues are addressed by *hidden* or *latent* variables which allows us to create mixture of Gaussians or multimodal distributions. Multimodal distribution has multiple modes or maximum values. Hidden variables improve quality of model proposed to probability density estimation. These models add possibility to control number of free parameters independently from dimensionality of the training data while concede capture of correlations in the training data.

This section begins with a definition of *K-means* algorithm which is suitable to split the data into clusters. *K-means* algorithm gives us good starting point to introduce *expectation-maximization* algorithm (EM). Latent variables that were mentioned above are used for assigning values from training data to particular components of the Gaussian mixture.

3.3.1 K-Means

The objective of clustering algorithm is to split the observations into clusters and thus reduce dissimilarities between samples assigned to one cluster and samples which are in different clusters [9]. There is one important assumption to satisfy, the observed data must be *independent and identically distributed* (i.i.d.) which means that each random variable has the same density distribution as the others and all are mutually independent. Clustering algorithms applied in *Mixture modelling* design a parametrized model that characterize the density function of random variable. This model describes a mixture of component density functions where each component define one of the cluster.

The result of K-Means algorithm for randomly generated data coming from two Gaussian distribution is illustrated in the Figure 3.5.
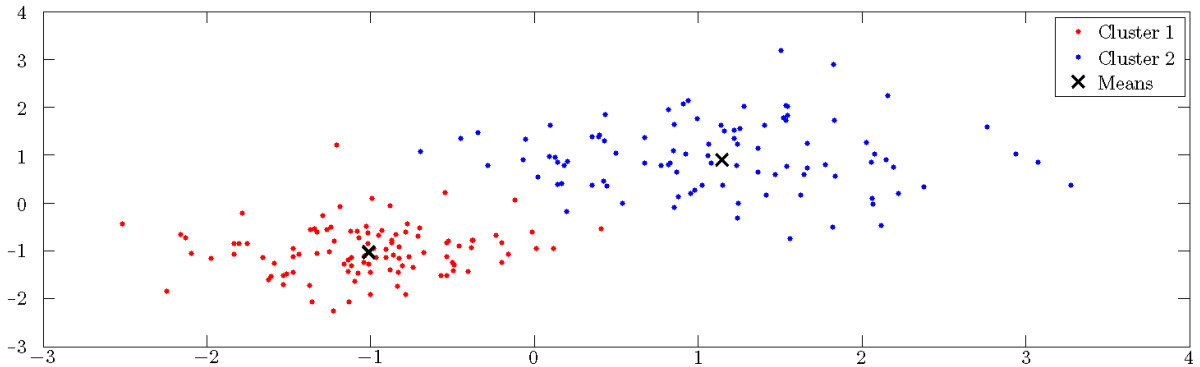

Figure 3.5: Exemplary result of K-Means algorithm.

The K-means algorithm is popular iterative descent clustering method. Letter K stands for K clusters. The objective is to group data values in multidimensional space. The data belong to data set $x_1, \ldots, x_N$ where N represents number of observations of a random D-dimensional variable x. We start with a set of D-dimensional vector μ_k for $k = 1, \ldots, K$. Centres of cluster are described with μ_k. The goal of K- means algorithm is to find set of vectors $\{\mu_k\}$ and allocation of data values to clusters, such that there is minimal Euclidean distance of each data value to its closest vector μ_k.

Formal description of K-means starts with definition of binary variable which indicates assignment of data values to cluster. Each data value has its set of binary variables $r_{nk} \in \{0, 1\}$, $k = 1, \ldots, K$, that determine cluster where is the data value x_n assigned to. *Distortion measure* is objective function that illustrates the distance between each data value and its assign vector μ_k:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2 \tag{3.32}$$

The aim is to discover values of $\{r_{nk}\}$ and $\{\mu_k\}$ which will minimize value of J. Iterative algorithm with two successive optimizations of the r_{nk} and the μ_k is suitable for this task. There are two phases, where we once minimize J with respect to r_{nk} (in second phase to μ_k), while keeping the μ_k fixed (in second phase r_{nk}). These optimizations are iterated until convergence. Minimization of J with respect to r_{nk} while keeping μ_k fixed is done by assigning the n-th data point to the nearest centre of cluster:

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg \min_j ||x_n - \mu_j||^2 \\ 0 & \text{otherwise} \end{cases} \tag{3.33}$$

Minimization of J with respect to μ_k while keeping r_{nk} fixed is done by setting derivative of quadratic function of μ_k to zero:

$$2 \sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0 \tag{3.34}$$

The Equation 3.34 solved for μ_k:

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}} \tag{3.35}$$

As result of K-means algorithm we set μ_k equal to the mean of all data values x_n assigned to cluster k. The algorithms ends after maximum number of iterations or in moment when there is no change in assignments to K clusters. Each iteration decreases value of J so that convergence is guaranteed. The problem may arise when this algorithm converges to local rather than to global minimum of J. K-means algorithm is widely used to initialize parameters of Gaussian mixture model before application of *EM* Algorithm.

3.3.2 Expectation-maximization Algorithm (EM)

Mixture models have received wide application because they provide more adaptability to the density distribution of random variable than the simpler normal-based models [20]. It is appropriate to consider several issues connected to mixture models. Principal question concerns about the parameter estimation of mixture model. There is also question related to number of mixture components and question whether there are some kind of bindings or groupings in the data.

The *EM* algorithm applies a probabilistic approach rather than hard assignments as for K-means algorithm. Probabilistic approach allows us to make soft assignments and gives us more suitable way to reflect uncertainty in these assignments.

We start with a K-dimensional binary variable z that use 1-hot encoding when only one value z_k is set to 1 and all other elements are equal to 0. The marginal distribution over z is connected to mixing coefficients π_k:

$$p(z_k = 1) = \pi_k \tag{3.36}$$

In order to be valid probabilities, the parameters π_k must satisfy two conditions:

$$0 \leq \pi_k \leq 1 \tag{3.37}$$

$$\sum_{k=1}^{K} \pi_k = 1 \tag{3.38}$$

Distribution of z that use 1-hot encoding can be written in form:

$$p(z) = \prod_{k=1}^{K} \pi_k^{zk} \tag{3.39}$$

The conditional distribution of random variable x given z is defined by Gaussian distribution:

$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{zk} \tag{3.40}$$

Marginal distribution of x is obtained from 3.39 and 3.40 and calculated by summing the joint distribution $p(z)p(x|z)$ over all potential values of z:

$$p(x) = \sum_{z} p(z)p(x|z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{3.41}$$

After above listed definitions we can use joint distribution $p(x, z)$ rather than marginal distribution $p(x)$ which leads to simplifications. With use of Bayes' theorem we can define the conditional probability of z given x as:

$$
\begin{aligned}
\gamma(z_k) \equiv p(z_k = 1|x) &= \frac{p(z_k = 1)p(x|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x|z_j = 1)} \\
&= \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}
\end{aligned}
\tag{3.42}
$$

where $\gamma(z_k)$ denotes $p(z_k = 1|x)$. Parameter vector π_k represents the prior probability and the quantity $\gamma(z_k)$ corresponding posterior probability after observation of x. It is often to denote $\gamma(z_k)$ as the *responsibility* that mixture component k takes for 'describing' the observed value x.

Data set of N observed values $\{x_1, \ldots, x_N\}$ is represented as $N \times D$ matrix where n-th row is formed by x_n^{T}. The set of hidden variables is represented as $N \times K$ matrix with z_n^{T} in rows. The function to be maximized is the log likelihood function defined by:

$$\ln p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\} \tag{3.43}$$

The *expectation-maximization* algorithm (EM) provides possibility for finding maximum likelihood solutions for Gaussian Mixture Models (GMM) with hidden variables. If the derivatives $\ln p(X|\pi, \mu, \Sigma)$ from 3.43 are set to 0 and from 3.26 we obtain:

$$0 = -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} \Sigma_k (x_n - \mu_k) \tag{3.44}$$

The minus on the right side of equation appears because the value of probability density function is constrained to being positive and summing up to one. We can see that the denominator is equal to $\gamma(z_{nk})$. After multiplying by Σ_k^{-1} and rearranging we retrieve:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n \tag{3.45}$$

while N_k, the number of data values allocated to cluster k is defined by:

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{3.46}$$

Maximum likelihood solution for covariance of a single Gaussian is defined after setting the derivative $\ln p(X|\pi, \mu, \Sigma) = 0$ (with respect to Σ_k):

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^{\mathrm{T}} \tag{3.47}$$

We can see that the denominator is defined by the effective number of data values in specific component and each data value is weighed by the corresponding posterior probability. Lastly, we can maximize $\ln p(X|\pi, \mu, \Sigma)$ with regard to the mixing coefficients π_k. Mixing coefficients must satisfy the condition 3.38. This condition is satisfied if we apply Langrange multiplier and maximize the subsequent quantity:

$$\ln p(X|\pi, \mu, \Sigma) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{3.48}$$

which gives

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} + \lambda \tag{3.49}$$

Now we multiply both sides by π_k and sum over k, we obtain $\lambda = -N$. We can apply this to remove λ and after rearranging remains:

$$\pi_k = \frac{N_k}{N} \tag{3.50}$$

It should be appointed that mixing coefficient for k-th component is obtained from the average responsibility which that component takes for definition of the data values.

EM Algorithm for Gaussian Mixtures

The EM algorithm starts with choosing initial values for the means, covariances and mixing coefficients [13]. Initialization is followed by iterative approach with two main update steps. First is the *expectation step*, where present values of parameters are used to compute the posterior probabilities on the basis of 3.42. These probabilities are applied in the second step called *maximization step* to re-estimate the means, covariances and mixture coefficients according to 3.45, 3.47 and 3.50.

The connection between K-means algorithm and EM algorithm is, that we use K-means to find an appropriate initialization for a Gaussian mixture model that is adapted by EM algorithm. It is practical use of K-means which reduces the complexity of EM algorithm. Recapitulation of EM algorithm is shown below:

1. Initialize the means μ_k, covariances Σ_k and mixing coefficients π_k, then calculate the initial value of the log likelihood.

2. **Expectation step.** With the current parameter values use:

$$\gamma(z_{nk}) = \frac{\mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(x_n|\mu_j,\Sigma_j)} \qquad (3.51)$$

to evaluate the responsibilities.

3. **Maximization step**. Re-estimate the parameters using the present responsibilities:

$$\mu_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n \qquad (3.52)$$

$$\Sigma_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k^{new})(x_n-\mu_k^{new})^{\mathrm{T}} \qquad (3.53)$$

$$\pi_k^{new} = \frac{N_k}{N} \qquad (3.54)$$

where

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk}). \qquad (3.55)$$

4. Compute the log likelihood:

$$\ln p(X|\mu,\Sigma,\pi) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k)\right\} \qquad (3.56)$$

Perform the check for the convergence criterion after step 4. The convergence criterion is represented by a minimal value of the difference between the estimated values of the parameters between two successive steps of EM algorithm. If the convergence criterion is not satisfied then return to step 2.

4 Implementation details

This chapter describes implementation of flight data classification based on *Gaussian Mixture Model* that was described in previous chapter. The design and structure of implementation part is introduced step by step. The main part of the implementation is programmed in `Python` programming language. First part of this section introduces basic structure of modules contained in the implementation part. Second part explains the function of particular modules and contribution of library modules being used. In the end of this section, there is to be found pseudo-code and flow diagram of main script, which can be controlled by configuration file.

The diagram, which describes main structure of implementation, is illustrated in the Figure 4.1. There are modules and connections between them visualized in logical order.
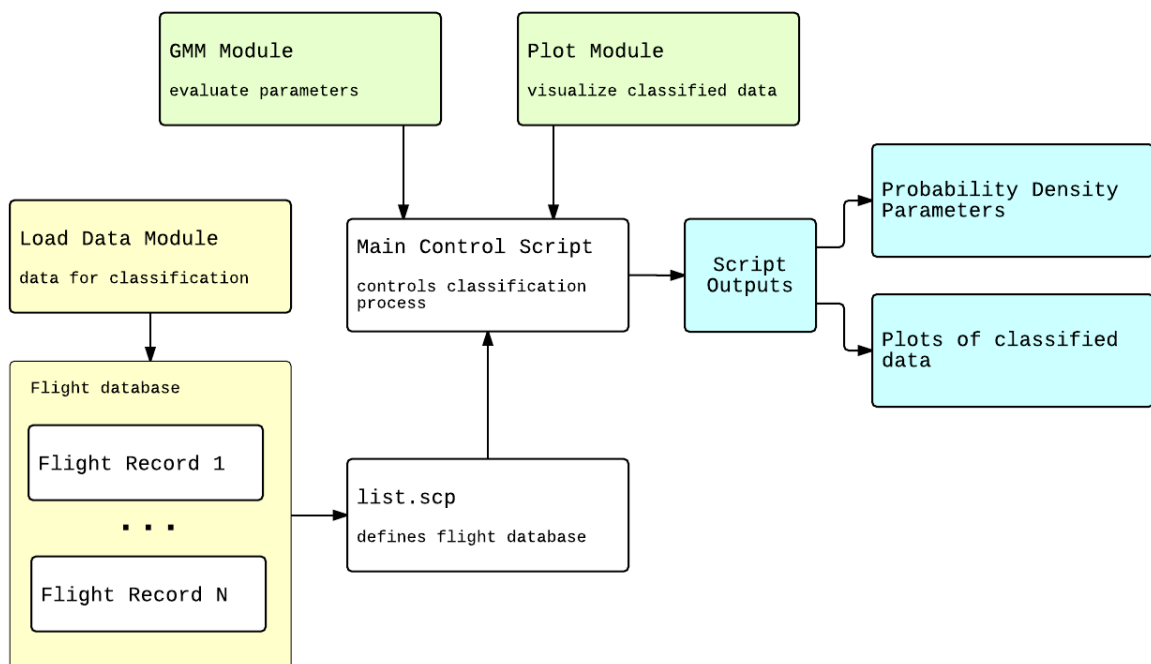

Figure 4.1: Comprehensive structure of implementation part.

The `Load Data` module prepares input data for classifier. It takes the flight data from flight database and then extracts selected features. The list of input files is saved in `.scp` files which represent standard way to store the training data and possible corresponding label files. The most important module of implementation is the `GMM` module that is intended to create *Gaussian Mixture Model* and train it with *Expectation-Maximization* algorithm

to classify flight data. The `Main Script` initiates the process of classification and calls all required functions included in the modules. The `Plot` module visualizes the results of classification process.

4.1 Modules

The implementation part of this thesis is divided into modules according to distinct functions of particular modules. This division allows to make easier comprehension of whole implementation part and it prepares the ground for possible future extension.

4.1.1 Data preprocessing

Since we are processing flight data, the classification process is attended by data preprocessing according to study provided by Section 2.5. Module `Load Data` processes database of flight records. Main objective of this module is to prepare training files that contain features for classification process. This module is written in `Matlab` programming language. Main script `process_db.m` creates flight database from flight records. This script calls functions for calculation of derived parameters such as height instrument or true airspeed (TAS). Basic compensations due to sensor off C.G. location are performed by functions `compensation_acc.m` and `compensation_hi.m`. `Load Data` module allows user to visualize input flight data and flight parameters before and after compensation, or for example the visualization of the difference between height derived from static pressure and GPS height, as was described in Section 2.1.2. Besides the data processing, there is also possibility to perform statics analysis of flight data. Function `statics_analysis` uses Matlab function `fitdist` which fits the distribution of input random data to get parameter values. This could be a preliminary step before application of *Gaussian Mixture Model* to sampled training data, because it can make clear an idea about probability density function of training data.

4.1.2 GMM

The task of `GMM` module is to prepare and iteratively update *Gaussian Mixture Model* that is most appropriate for estimation of density function for the training data. *Gaussian Mixture Models* are estimated from data using Expectation Maximization algorithm.

This module uses two known libraries for scientific and mathematical calculations. The `NumPy` library for storing data in N-dimensional arrays and `SciPy` library for scientific calculations.

Key functions in GMM module

The GMM module offers required functionality to build and estimate Gaussian Mixture Models 3.3 with EM algorithm which was described in 3.3.2.

- Function `gmm_eval_prep` is responsible for initial configuration or set-up of the Gaussian Mixture Model. It takes three arguments, the weights, the means and the covariances, which were set to the initial values. This function sets GMM properties on the basis of proposed parameters. For full covariance matrix it sets corresponding vectors for `gconst`, `invCovs`, `invCovMeans` and `utr` with `utc`, where `gconst` is precomputed constant for log probability of a Gaussian, `invCovs` is vector of invariant covariance

and `invCovMeans` is invariant covariance multiplied by value of mean. Values `utr` and `utc` represents indices into upper triangular part of covariance matrix.

- The vector of log-likelihoods is computed in function `gmm_eval`. Log-likelihoods are evaluated with parameters stored in GMM dictionary object for each frame of data matrix. This function must follow after call of `gmm_eval_prep` function that initializes GMM dictionary object. Depending on specified parameters, the statistics of zero, first and second order are returned in additional variables from this function. The data are expanded quadratically and then log-likelihoods for each frame and each Gaussian component of the mixture are calculated. New value of responsibility denoted as `gamma` is calculated from parameters of GMM object and then zero order statistics are sum of `gamma` values in each frame, first order statics are result of dot product of responsibilities and the data, and finally second order statistics are result of dot product of responsibilities and squared data. Call of this function is equivalent to Expectation step in EM algorithm.

- Maximization step from EM algorithm is represented by function `gmm_update`. Basic parameters of GMM are results of GMM update. New values of parameters are recalculated with new return values from function `gmm_eval`. Weights, means and covariances are re-estimated according to Maximization step of EM algorithm given in 3.3.2. To finish one complete iteration of EM algorithm another one call of function `gmm_eval_prep` is required. This function update GMM model with new parameters.

- After number of iterations in EM algorithm function `gmm_llhs` computes final vector of log-likelihoods. The vector of log-likelihoods is evaluated for each frame of data matrix and for all components of Gaussian mixture. The GMM object that were recently updated with `gmm_eval_prep` function is used to compute final value of log-likelihoods vector. Values of log-likelihoods are used later to determine the class of the data.

4.1.3 Plot module

The `Plot` module uses library `Matplotlib` to visualize results of classification process. Core function of this module is to mark data classified as flight events and visualize it together with the source data. Additional functionality gathered in plot module allows user to choose flight parameters that will be included in the set of resultant plots. These plots are selected only from time intervals where the flight maneuvers were performed. Time intervals are chosen on the basis of the classification process.

Matplotlib
The `matplotlib` is a 2D plotting library which produces high quality plots in a various formats. This library has its origin in emulation of MATLAB graphics commands but is programmed in Python programming language. It is connected with `NumPy` library and so it allows users to plot and visualize large arrays of data. [11]

4.2 Main script

Main script called `main.py` controls the process of classification. It calls function from included modules that were described above. Diagram illustrated in 4.2 shows the control
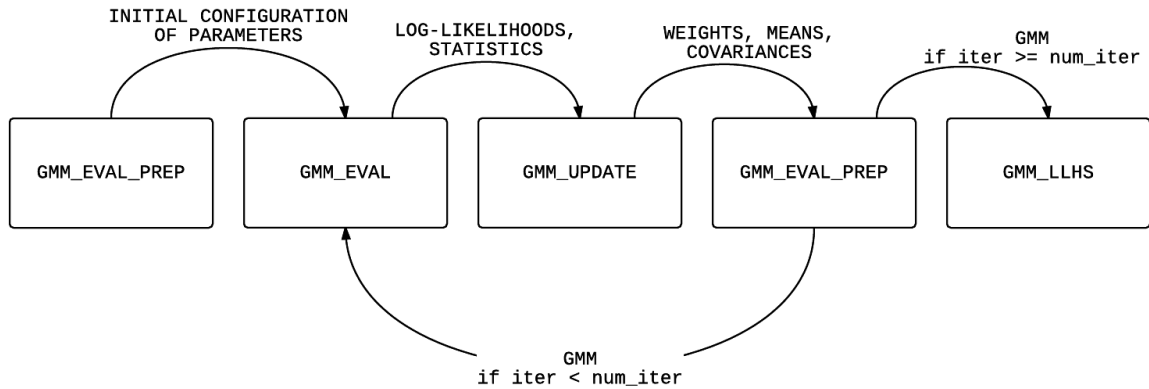
flow in the main script.

Figure 4.2: Control flow in main script that runs estimation of the GMM parameters.

For better illustration, we attach below the pseudo-code which contains list of commands and function calls in the main script.

```
load flight_database
for flight_record in flight_database
    data_file = load_flight_record
    data = extract_features
    apply_running_mean(data)
    framing(data)
    get_energy(data)
    normalize_energy(data)
    initialize_parameters(w,m,c) // weights, means, covariances
    GMM = gmm_eval_prep(w,m,c)
    for i in num_iterations
        llh, statistics = gmm_eval(data, GMM) // computation of log-likelihoods
        w, m, c = gmm_update(statistics)      // update of parameters
        GMM = gmm_eval_prep(w,m,c)            // update of GMM

    gmm_llhs(data, GMM)          // evaluate log-likelihoods

    for llh in llhs:
        if (llh < flight_event_threshold):
            data_out=flight_event_marker // classify data
    visualize_results(data_out)
```

Above listed pseudo-code shows that main script loads the flight data prepared by `load` module and extracts the features for each flight record from the flight database. Then applies EM algorithm for Gaussian Mixture Models from `gmm` module to obtain maximum log-likelihoods. According to resultant log-likelihoods it classifies the data and then calls function from `plot` module to visualize the results.

41

5 Testing and experiments

This chapter describes results and experiments achieved during classification process of the flight data. Theory that was introduced in Chapter 3 created the basis for the implementation part and in next sections we will demonstrate the application of the implementation for measured flight data. In first part of experiments we will try to apply classification based on Gaussian Mixture Models to flight data from flight simulator. Next part of experiments includes utilization of implemented classifier to real flight data as they were described in Section 2.2. For the first experiment there is the plot of resultant Gaussian components and most important, the plots of classified data with corresponding description.

5.1   Flight data from flight simulator

Flight data recorded from flight simulator are produced by simulation model of aircraft which consists of the differential equations called equations of motion. In first experiments we try to find and classify flight events that occur in the data from simulation. We performed flight tests with testing maneuver, the doublet in elevator. The segment of 40 seconds from measured data is shown in the Figure 5.1.
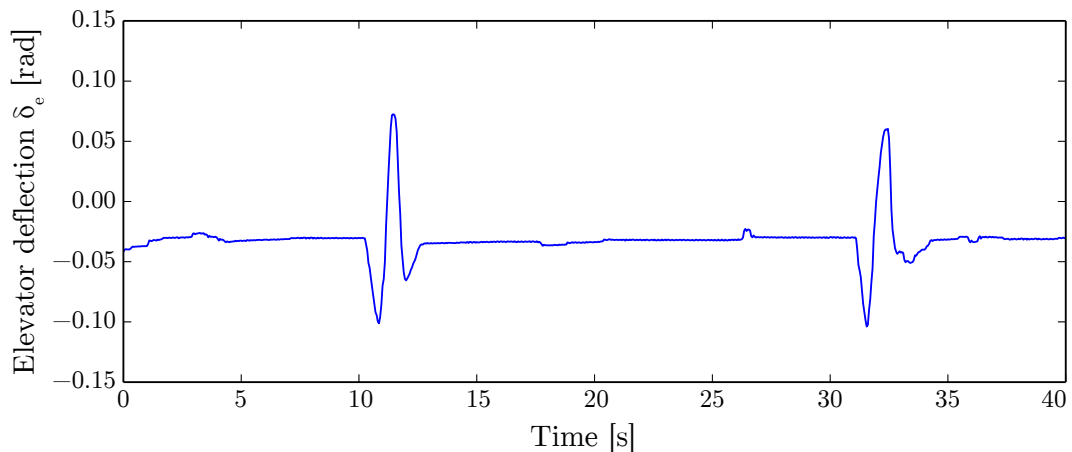

Figure 5.1: Segment of flight data measured on flight simulator.

In picture shown above, there is a waveform of flight parameter called deflection of elevator. This flight parameter represents the angle of deflection of elevator and it is measured in radians. There are two impulses with spacing visible in this segment. The goal of classification is to find and mark the flight events, in this case testing maneuver called doublet in the elevator.

We start the classification process with a definition of the basic parameters. We choose Gaussian Mixture with 3 components, that could sufficiently describe the measured flight data. At first, we set initial configuration for the Gaussian Mixture components means, covariances and weights:

```
means = [-1.0, 0.0, 1.0]
covariances = [1.0, 1.0, 1.0]
weights = [0.33, 0.33, 0.33]
```

After several iterations, the change of the parameters of GMM is insignificant and the text output of the main control script is shown bellow:

```
means = [-0.4522, 0.0175, 1.4538]
covariances = [ 0.0002212, 0.1388, 1.9407]
weights [ 0.6749, 0.1165, 0.2085]
```

Components of the GMM with these parameters are illustrated in the Figure 5.2 to give a clearer picture of the probability density function that describes flight data from the simulator.
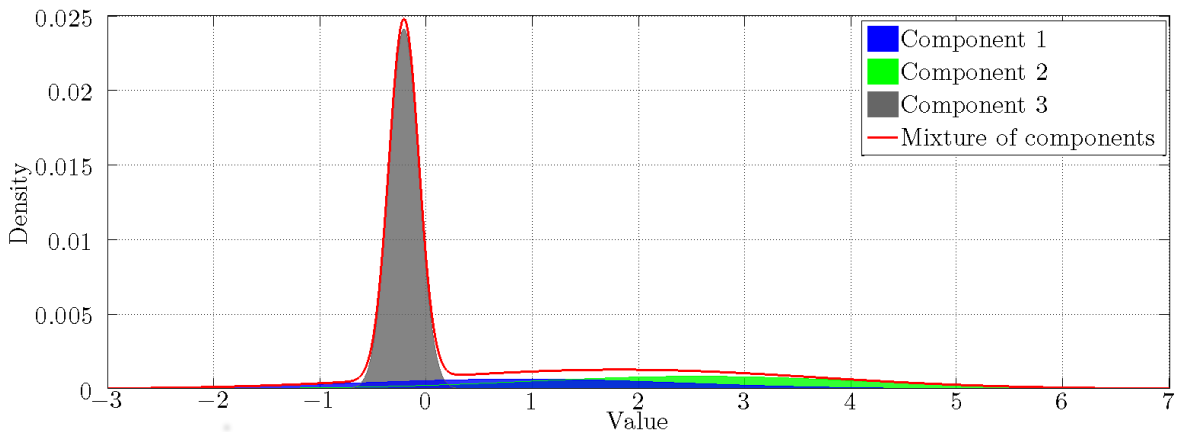

Figure 5.2: Components of GMM with parameters estimated by an EM algorithm.

After explaining the text output of the main script we can proceed to the step of the results visualization. There are two main options available in the configuration of the main script. The user can visualize the whole waveform for the flight parameter with the highlighted segments which are classified as particular flight maneuver. The second option involves the generation of plots for each occurrence of the flight maneuver in the flight parameter. In this case it is possible to define intervals (in samples) which indicate the time period of the flight data shown before and after occurrence of the flight maneuver. The main advantage of the second approach is an option to define the list of flight parameters that will be illustrated in the output plots. So this option provides good basis for the examination of important flight parameters during the flight maneuvers.

The flight data with corresponding markers are illustrated in the Figure 5.3. Each marked segment represents one classified flight event, in this particular case it is the doublet in the elevator. This figure is the output of the main control script and its configuration is located in configuration file.
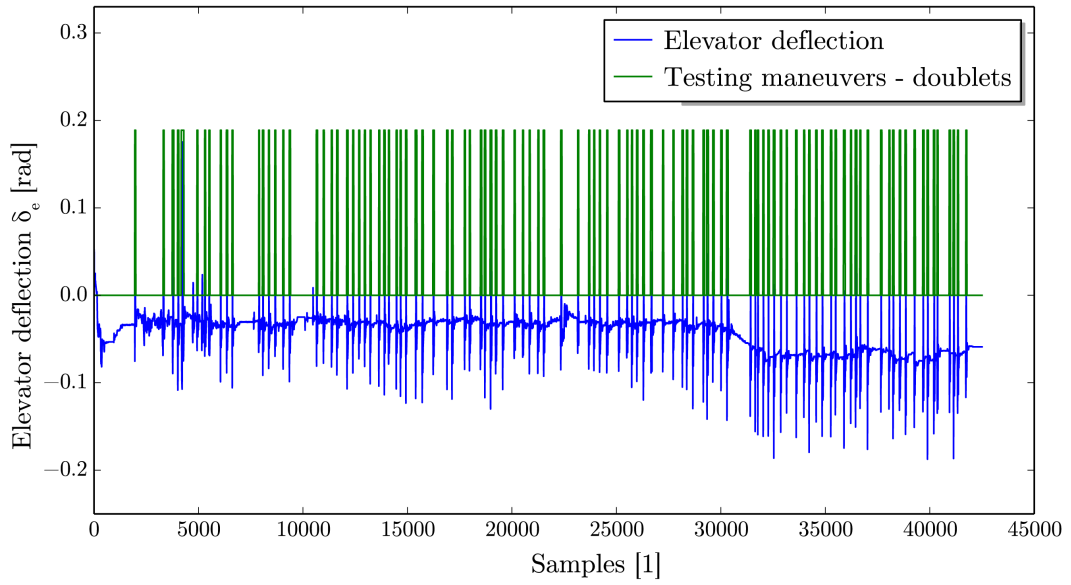
Figure 5.3: Deflection of elevator rudder with highlighted flight events.

In the Figure 5.4 we can see one chosen flight maneuver corresponding with the theoretical basis given in 2.4.1. The plot is divided into 4 subplots. First subplot shows the deflection of the elevator, where we can see the testing maneuver called doublet. Second subplot contains the deflection of the aileron, where we observe small changes caused by the pilot input on the control stick. The third subplot illustrates the dependency between the deflection of the elevator and angle rate about Y_B axis in Body Fixed Frame denoted as q. In the last subplot we can see the change in the measured height of the aircraft during the flight maneuver.
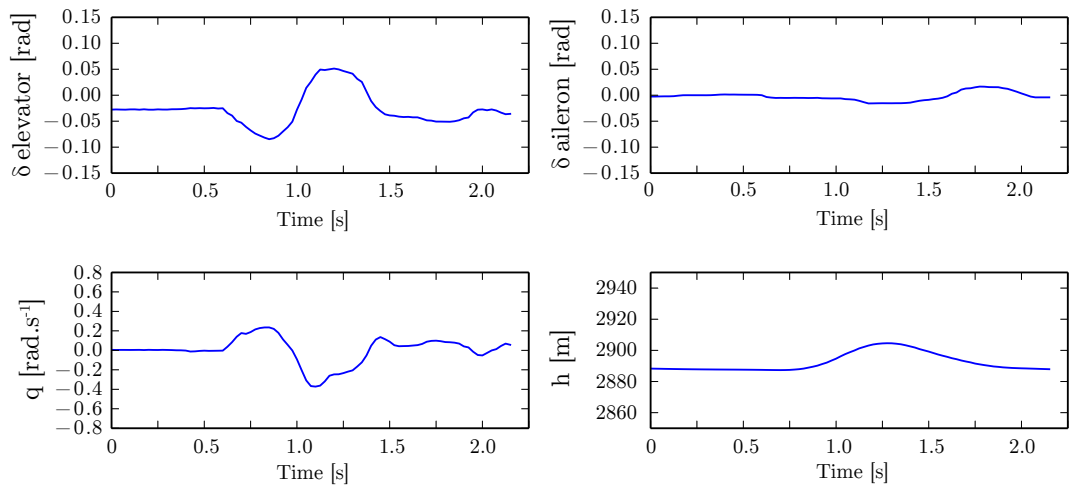

Figure 5.4: Output of the main script for single classified doublet in the elevator rudder.

44

5.2 Flight data from aircraft

Flight simulator approximates real aircraft with the model based on differential equations. The following tests investigate the classification process performed on the flight data from a real aircraft. We obtain the flight data from flight experiments and in the Figure 5.5 we can see the segment of 60 seconds from the measured flight data.
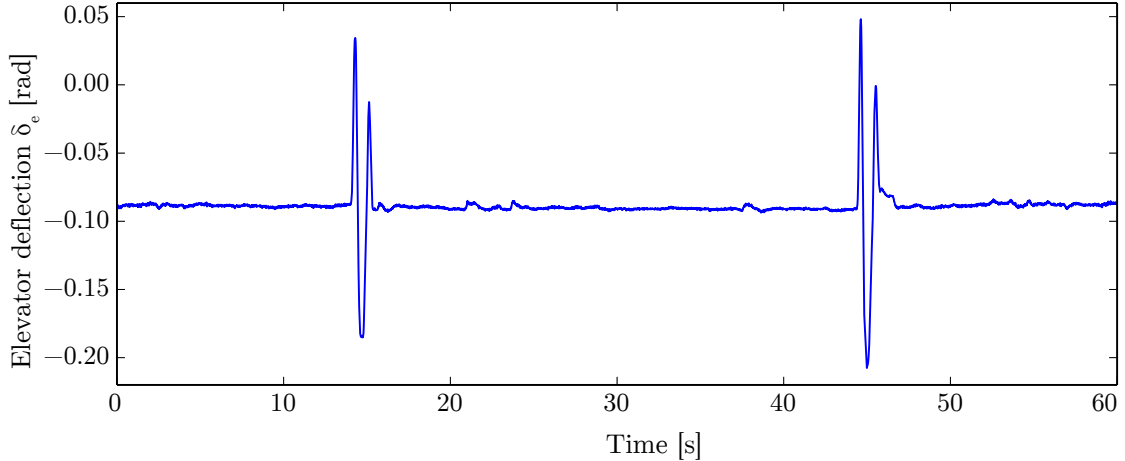

Figure 5.5: Segment of flight data measured on aircraft.

The segment of the flight data comes from the flight parameter deflection of the elevator. We can observe two impulses in interval of 100 seconds, where the sampling frequency is $100Hz$ and so 100 samples represent 1 second. Similar to the classification of the flight data from the flight simulator, the aim of the classification is to separate the flight data which comes from normal flight, with no occurrence of the flight maneuver and the flight data which comes from testing maneuvers such as a doublet in the elevator.

As described in Section 2.5, the flight data measured by the inertial sensors in real aircraft suffers from various kind of errors. In the Figure 5.6 we can see results of a finite impulse response (FIR) filtering called *running mean* or *moving average*. Source data with mean that is moving up and down is colored blue and the filtered data with mean aligned to zero are colored green.

Classification process starts with the definition of the basic parameters. We set Gaussian Mixture with 3 components, that could sufficiently describe measured flight data. We initialize the configuration for the Gaussian Mixture component means, covariances and weights:

```
means = [-1.0, 0.0, 1.0]
covariances = [1.0, 1.0, 1.0]
weights = [0.33, 0.33, 0.33]
```

After several iterations, the change of the parameters of the GMM is insignificant and the text output of the main control script is shown bellow:

```
means = [-0.2457, -0.1416, 2.8296]
covariances = [ 0.0001318, 0.007031, 4.7005]
weights [ 0.7712, 0.1541, 0.0746]
```
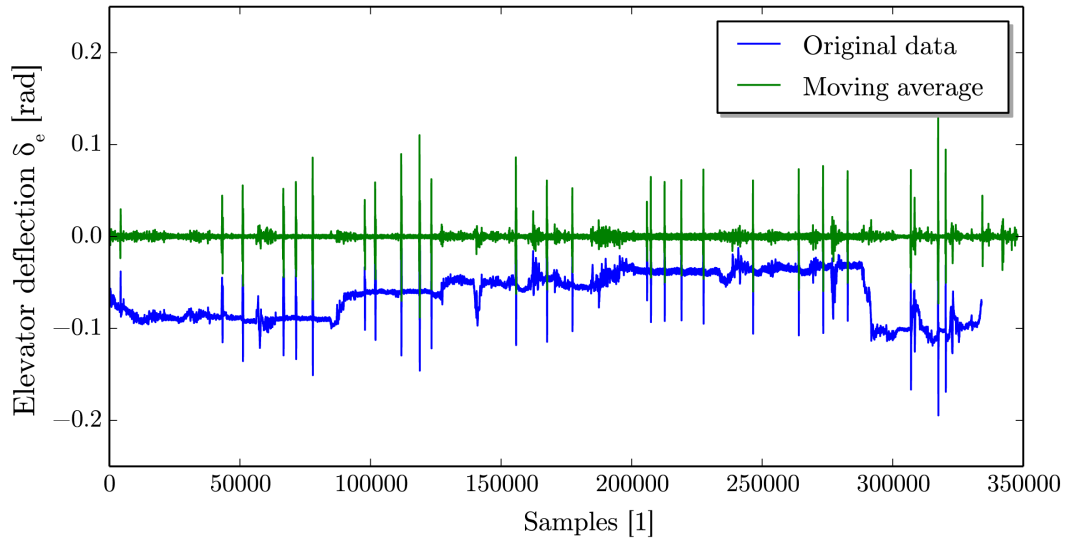
Figure 5.6: Example of running mean filtering of the flight parameter Deflection of elevator.

The visualization of the results is a key role of `Plot module` and in the Figure 5.7 we can see the flight with corresponding markers. Each marked segment of the flight data represents one performance of testing maneuver called doublet in the elevator.
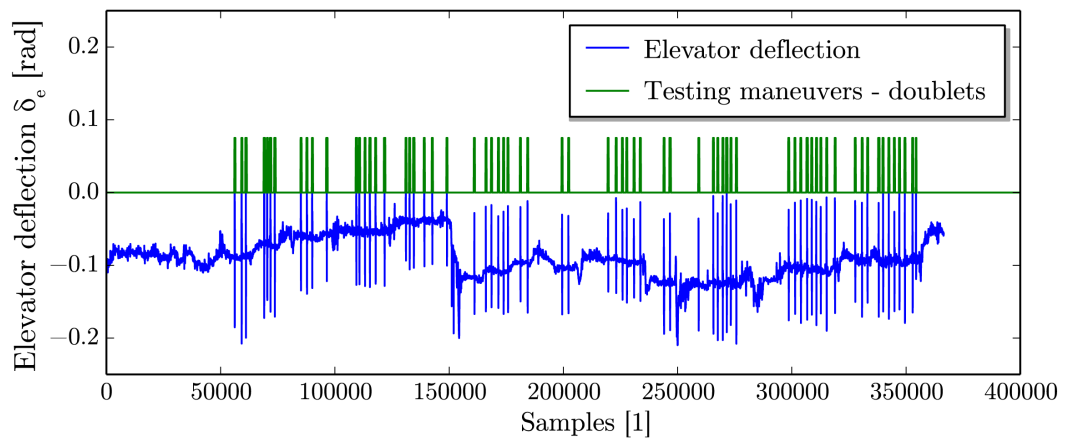

Figure 5.7: Deflection of the elevator with identified testing maneuvers.

In the Figure 5.8 we can observe one chosen testing maneuver, doublet in elevator. We can confirm the fact that the classified flight maneuver corresponds with the desired output, which was shown in the Figure 2.14. The plot is divided into 4 subplots. First subplot shows parameter deflection of the elevator, which is in this case the feature used for the classification. Second subplot contains deflection of the aileron, where we can observe insignificant changes caused by pilot control of the control stick. If these changes were more significant, this flight maneuver would not be suitable for identification. The term identification represents the application of the classification results, when we examine plots of flight maneuvers for different flight parameters and then we try to estimate the values

46

of the coefficients for the differential equations, which describe the aircraft. Third subplot illustrates the dependency between the deflection of the elevator and pitch rate about Y_B axis of the Body Fixed Frame denoted as q. In the last subplot we can see the change in pitch angle denoted as *Theta* (θ). We see again, that there is time dependent relation between the pitch rate q and the pitch angle θ.
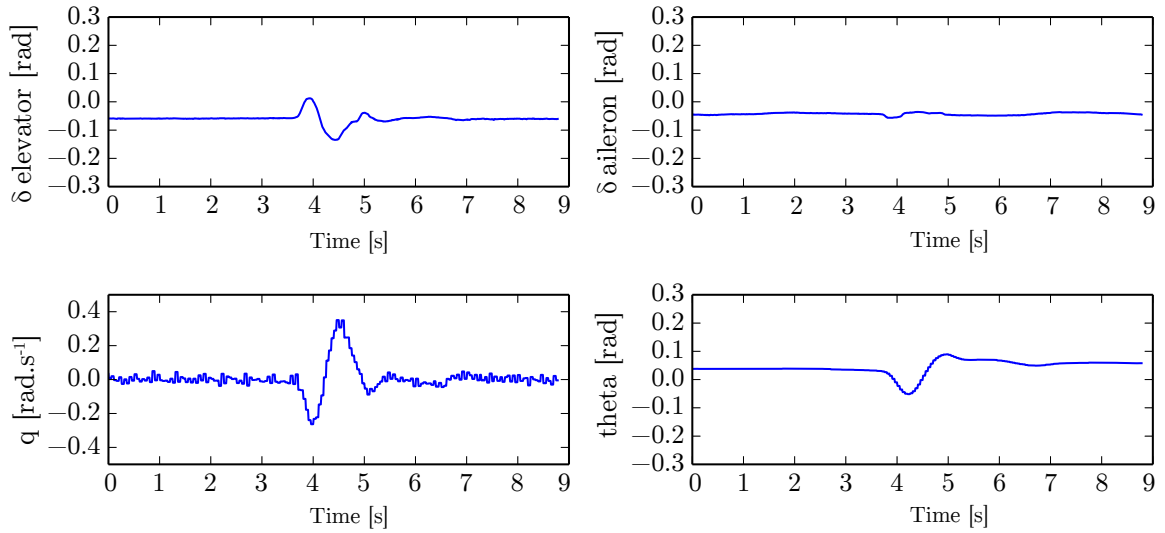

Figure 5.8: Various flight parameters for single classified testing maneuver.

The segment of the measured deflection of the elevator with a length of 60 seconds is shown in the Figure 5.5. The following Figure 5.9 illustrates the same segment of the measured flight data with the testing maneuvers being highlighted. We can see identified testing maneuvers, which meet the requirements according to the specification given in 2.4.1.
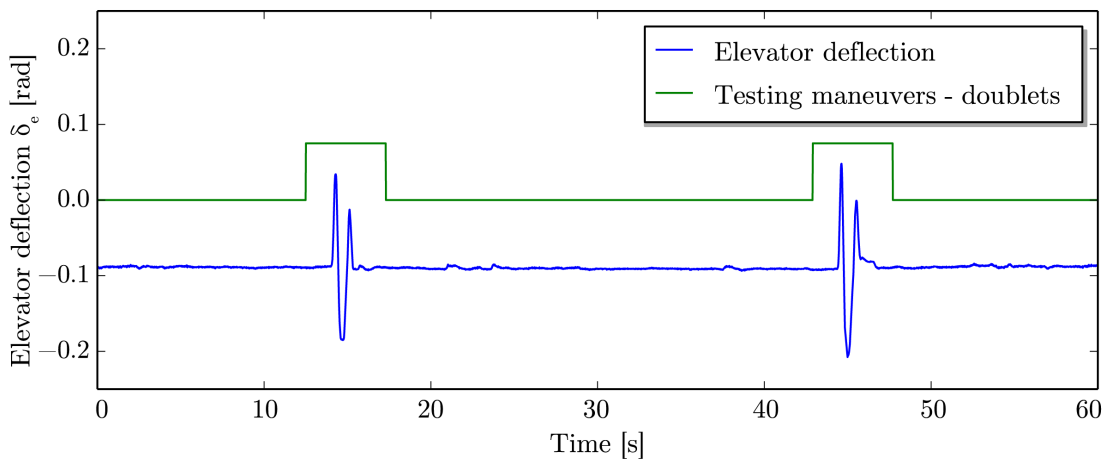

Figure 5.9: Segment of flight data with two highlighted testing maneuvers.

6 Conclusion

The identification of the flight events in the flight data has an important role in prevention, testing and analysis of the flight data. It is appropriate to prevent a routine when we create records with large amount of the flight data which will no longer be used. Important use case of the flight maneuvers identification arises in the field of development and testing of avionics systems, where the flight data is used in the evaluation and investigation of such systems. The testing and investigation of the aircraft behaviour and characteristics are often performed by a testing maneuvers.

The classification of the testing maneuvers was the main objective of this thesis. There are three main approaches in the classification task, the machine learning, the statistical and the neural network approach. This work focuses on statistical methods and leaves other methods to possible future research.

This work is based on the probability theory and theory of the Gaussian Mixture Models and proposes the implementation written in Python programming language for the classification of the testing maneuvers in the flight data. First chapters give an overview required to understand the format and values of various flight parameters. Current approaches to the data classification were considered and according to the gained insight the Gaussian Mixture Models were studied in depth. The Gaussian Mixture Models offer a powerful way to probability density estimation of random data, in the case of this work, the data from the flight recorder. The unsupervised learning without the need for training data is the advantage of the Gaussian Mixture Models. The parameters of the Gaussian Mixture Model are estimated by the Expectation Maximization algorithm which forms the core of the implementation part. The classification procedure makes decision on the basis of the parameters being estimated. The classified data are processed to the further application. Such application includes the visualization and analysis of the results as shown during the testing and experiments. The performed experiments confirm that the Gaussian Mixture Models are suitable tool for the task of the identification of the testing maneuvers in flight data. A successful application of a classifier on the flight data and flight simulation shows that the Gaussian Mixture Models have a wider applicability. The results of the classification process include plots with identified testing maneuvers and plots with selected flight parameters during the testing maneuvers.

6.1 Future work

The analysis of the flight data offers wide area to exploration and research. The work proposed in this thesis focuses on identification of testing maneuvers using classier based on the Gaussian Mixture Models. There are many other approaches and extension that can be used in the analysis of the flight data. The unsupervised learning can be replaced

with the supervised learning based on the training dataset. The training set would contain marked flight maneuvers which brings us to the next possible improvement. The set of the classified flight maneuvers can be expanded to include all of the basic flight maneuvers. The results of classification of all flight maneuvers would give an overview of the whole flight.

Possible future improvement includes application that is deployed in real-time during the flight. After the end of flight the application should submit the report containing the list of performed flight maneuvers.

Bibliography

[1] F.A. Administration and U.S.D. Transportation. *Instrument Flying Handbook: FAA-H-8083-15B*. 2014.

[2] F.A. Administration and U.S.D. Transportation. *Pilot's Handbook of Aeronautical Knowledge: FAA-H-8083-25A*. CreateSpace Independent Publishing Platform, 2014.

[3] J. Anderson. *Introduction to Flight*. McGraw-Hill Education, 2007.

[4] C.M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[5] W. Durham. *Aircraft Flight Dynamics and Control*. Aerospace Series. Wiley, 2013.

[6] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Computer science and scientific computing. Elsevier Science, 2013.

[7] M.S. Grewal, L.R. Weill, and A.P. Andrews. *Global Positioning Systems, Inertial Navigation, and Integration*. Wiley, 2001.

[8] P.D. Groves. *Principles of GNSS, Inertial, and Multisensor Integrated Navigation Systems, Second Edition:*. GNSS/GPS. Artech House, 2013.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer, 2009.

[10] E.L. Houghton and P.W. Carpenter. *Aerodynamics for Engineering Students*. Referex Engineering. Butterworth-Heinemann, 2003.

[11] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[12] M. Kayton and W.R. Fried. *Avionics Navigation Systems*. A Wiley-Interscience publication. John Wiley & Sons, 1997.

[13] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley, 2007.

[14] Donald Michie, D. J. Spiegelhalter, C. C. Taylor, and John Campbell, editors. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, Upper Saddle River, NJ, USA, 1994.

[15] Iossif B. Mugtussidis, M. R. Anderson, E. M. Cliff, W. C. Durham, C. D. Hall, F. H. Lutze, and Iossif B. Mugtussidis. Flight data processing techniques to identify unusual events, 2000.

[16] R.C. Nelson. *Flight Stability and Automatic Control.* Aerospace Science & Technology. McGraw-Hill International Editions, 1998.

[17] R. Rose, Aeronautical Research Council, O.P. Nicholas, and G. Vorley. *Flight Measurements of the Elevator and Aileron Hinge-moment Derivatives of the Fairey Delta 2 Aircraft Up to a Mach Number of 1.6 and Comparisons with Wind-tunnel Results.* Reports and memoranda / Aeronautical Research Council. H.M. Stationery Office, 1967.

[18] I.L. Thomas and R.H. Evans. *Performance Characteristics and Methods of Testing of Force-feedback Accelerometers.* Reports and memoranda. H.M. Stationery Office, 1969.

[19] D. Titterton, J.L. Weston, and Institution of Electrical Engineers. *Strapdown Inertial Navigation Technology, 2nd Edition.* IEE radar, sonar, navigation, and avionics series. Institution of Engineering and Technology, 2004.

[20] A.R. Webb. *Statistical Pattern Recognition.* A Hodder Arnold Publication. Arnold, 1999.

[21] J.M. Zogg and U-Blox (Firma). *GPS: Essentials of Satellite Navigation : Compendium : Theorie and Principles of Satellite Navigation, Overview of GPS/GNSS Systems and Applications.* U-Blox, 2009.

Appendix A - Flight Data

Table A.1: List of measured flight parameters.

Short name	Long name	Unit
a_x	Longitudinal acceleration	$m.s^{-2}$
a_y	Lateral acceleration	$m.s^{-2}$
a_z	Normal acceleration	$m.s^{-2}$
p	Roll speed	$rad.s^{-1}$
q	Pitch speed	$rad.s^{-1}$
r	Yaw speed	$rad.s^{-1}$
ϕ	Roll angle	rad
θ	Pitch angle	rad
ψ	Yaw angle	rad
δ_e	Deflection of elevator	rad
δ_f	Deflection of flaps	rad
δ_r	Deflection of rudder	rad
δ_a	Deflection of aileron	rad
f_e	Elevator control force	N
f_a	Aileron control force	N
f_{rr}	Right pedal control force	N
f_{rl}	Left pedal control force	N
q_{bar}	Differential pressure	Pa
p_{stat}	Static pressure	Pa
$temp$	Outside air temperature	K
GPS_{lon}	GPS longitude	deg
GPS_{lat}	GPS latitude	deg
GPS_h	GPS height	m

Appendix B - Content of CD

Root directory

`/tech_report.pdf`	– the text of this thesis in pdf format
`/tech_report/`	– source files of the text of this thesis
`/src/`	– source files of the implementation part

Main script:

`/src/main.py`

Modules:

`/src/modules/load/`	– Load module
`/src/modules/gmm.py`	– GMM module
`/src/modules/plot.py`	– Plot module

Directory `/src/config/` contains the configuration scripts.

Directory `/src/lists/` contains the list of input files with the flight data.

There is `README.txt` file in `/src/` directory, which contains detailed instructions required to run the classification process.