Chapter 2
Orbit Determination from Observations

2.1 Position of the Problem

In general terms, the determination of an orbit is an iterative process meant to know and predict the position and velocity (or the orbital elements) of a space object, with respect to a primary celestial body, from observations of that object. For example, the orbit determination for an artificial satellite revolving about the Earth is a series of operations aimed at determining the motion (that is, the six orbital elements or the six Cartesian components of the position and velocity vectors at some given epoch) of the satellite with respect to a reference system having its origin in the centre of mass of the Earth. Likewise, for a natural celestial body revolving about the Sun, an astronomer computes the orbital elements of that body with respect to the Sun on the basis of observations which have been performed at some place on the surface of the Earth.

The methods used to this end may be classified into two broad categories. The first category includes the classical (or deterministic) methods, which consider the measurements as free from errors, and use therefore the minimum number of measurements required to determine an orbit. The second category includes the modern (or statistical) methods, which consider the measurements as affected by errors, and use therefore more measurements than those which would be strictly necessary to determine an orbit, with the view of reducing the influence of such errors by means of a suitable mathematical treatment of the data gathered.

Let $\boldsymbol{x} = \boldsymbol{x}(t)$ be the state vector, at a given time t, of an artificial satellite orbiting around the Earth, that is, the vector whose six components are the three components of the position vector, $\boldsymbol{r}$, and the three components of the velocity vector, $\boldsymbol{v}$, of the satellite at time t with respect to either the true-of-date or the J2000.0 geocentric-equatorial reference system XYZ (defined in Sect. 1.9). Let $\boldsymbol{x}_0 = \boldsymbol{x}(t_0)$ be the known state vector of the same satellite at a given initial time t_0 with respect to the same reference system. When the forces (the central force and its perturbations) acting upon the satellite are known exactly, then it is possible to determine its state

A. de Iaco Veris, *Practical Astrodynamics*, Springer Aerospace Technology,
https://doi.org/10.1007/978-3-319-62220-0_2

vector $\boldsymbol{x}$ at any given time t by integrating the equations of motion. However, the state vector $\boldsymbol{x}$ of the satellite at the initial time t_0 is known approximately, not exactly. In addition, the force model used to compute the forces acting upon the satellite at a time t can only provide approximate values of the true forces, because the physical constants (e.g. the gravitational parameter of the Earth) are known approximately and also because the mathematical model used to compute the forces can never be exact. Consequently, the orbit determination of a satellite requires that the observations of the satellite (which are affected by random and systematic errors due to the non-exactness of the force model used) should be mathematically treated in such a way as to obtain the best estimate of the orbital elements representing the satellite motion at any time. To this end, it was customary to operate in two stages. The first stage, called preliminary orbit determination, is the approximate determination of an orbit by means of a minimum number of observations of the orbiting object.

The second stage, called differential correction of orbits or orbit improvement, comprises:

1. the collection of more observations than those strictly required, and
2. the fitting of the data gathered to an orbit by means of some mathematical algorithm, based usually on the method of the least squares.

The results of the second stage are differential corrections to the preliminary orbital elements, which have been determined in the first stage.

In the first stage, both the main attracting body and the orbiting body are considered as isolated particles subject only to their mutual gravitational forces. Therefore, the motion is governed by the second-order differential equation shown in Sect. 1.1, that is, by

$$\boldsymbol{r}'' + \left(\frac{\mu}{r^3}\right)\boldsymbol{r} = \boldsymbol{0}$$

where $\boldsymbol{r}''$ is the acceleration vector of the orbiting body, μ is the gravitational parameter of the attracting body (resulting from the product of the mass, M, of the attracting body by the gravitational constant, G), and $\boldsymbol{r}$ is the position vector of the orbiting body with respect to an inertial reference system having its origin in the centre of mass of the attracting body, whose mass M is supposed to be much greater than the mass m of the orbiting body. As also shown in Sect. 1.1, the constants of integration of this differential equation are six, that is, the orbital elements of the orbiting body.

In the second stage, the perturbations to the primary gravitational force are taken into account, so that the differential equation of motion is expressed as follows

$$\boldsymbol{r}'' + \left(\frac{\mu}{r^3}\right)\boldsymbol{r} = \boldsymbol{a}$$

where $\boldsymbol{a}$ is the vector sum of the perturbing accelerations acting on the orbiting body. In case of an artificial satellite of the Earth, as will be shown at length in Chap. 3, these accelerations are primarily due to:

- the non-spherical shape and the non-homogeneous mass density of the Earth;
- the gravitational attraction exerted on the satellite by other celestial bodies than the Earth (in particular, by the Sun and the Moon);
- solid Earth and ocean tides;
- the aerodynamic drag exerted by the Earth atmosphere, particularly important in case of artificial satellites orbiting at low altitudes;
- solar radiation pressure; and
- thrusters used for orbital manoeuvres.

By the way, solid Earth tides are similar to ocean tides, both of them being due to the gravitational forces exerted upon the Earth by the Sun and the Moon. The difference between the two types of tide resides in the much smaller tidal distortion of the Earth (about 30 cm a day) in comparison with that of the ocean, because of the resistance opposed by the rocks.

As a result of the perturbing accelerations, the orbital elements of an artificial satellite revolving about the Earth are not constant, but vary slowly with time.

Consequently, the second-order differential equation $\boldsymbol{r}'' + (\mu/r^3)\boldsymbol{r} = \boldsymbol{a}$ is expressed as a system of m first-order differential equations

$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x}, t)$$

with the initial conditions

$$\boldsymbol{x}(t_0) = \boldsymbol{x}_0$$

where $\boldsymbol{x} = \boldsymbol{x}(t)$ is the augmented m-dimensional, time-dependent state vector, that is, the state vector comprising not only the six position (x, y, z) and velocity (v_x, v_y, v_z) co-ordinates of the orbiting body but also the physical constants provided by the force model used (that is, the physical quantities which are independent of time); $\boldsymbol{f}(\boldsymbol{x}, t)$ is a vector-valued function (the integrand function) which depends on the state vector and on time; and $\boldsymbol{x}_0$ is the known state vector at some epoch t_0.

On the other hand, the observations can be represented by the following system of n nonlinear algebraic equations

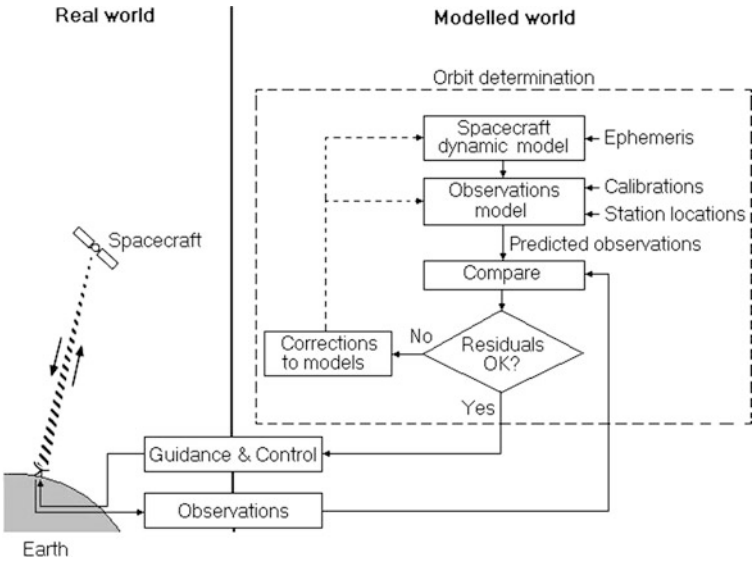
$$\boldsymbol{z}_i = \boldsymbol{g}(\boldsymbol{x}_i, t_i) + \boldsymbol{\varepsilon}_i$$

($i = 1, 2, \ldots, n$), where $\boldsymbol{z}_i$ are the actual observations made at epochs $t_1, t_2, \ldots, t_n$, $\boldsymbol{g}(\boldsymbol{x}_i, t_i)$ are the predicted values, and $\boldsymbol{\varepsilon}_i$ are random errors of measurement affecting the observations. Such equations, solved for $\boldsymbol{x}_i$, yield the state vector

$$\boldsymbol{x}(t_i) = \boldsymbol{\Theta}(\boldsymbol{x}_0, t_0, t_i)$$

($i = 1, 2, \ldots, n$) at times, respectively, $t_1, t_2, \ldots, t_n$. This way of proceeding, comprising the two stages described above, is called batch estimation processing and is based on the least-squares method.

Recently, as a result of the mathematical theory due to Kalman and Bucy [44, 45] and also of the advent of advanced computing means, the two stages are no more separated from each other, and each epoch of observations is processed individually by means of a sequential estimation algorithm, which is just the Kalman filter. The preliminary orbit determination comprises several methods meant to determine the orbital elements. These methods will be shown in the following paragraphs. The determination of an orbit is part of a broader process called integrated guidance, navigation, and control. The whole process of spacecraft guidance, navigation, and control is shown in the following scheme, which is redrawn from Thornton and Border [70].



Guidance is the actual steering of a spacecraft travelling through space. This steering may come from sources placed either inside (a human crew or an on-board computer) or outside (a ground station) the spacecraft. An example is provided by the commands given to the rocket motor of a spacecraft to control the thrust. Navigation is the measurement of the motion (position, orientation, and velocity) of a spacecraft in space, obtained by means of observations performed by the crew, or by automatic on-board sensors, or by tracking equipment located on the ground.

Control is the alignment and stabilisation of a spacecraft while the guidance and navigation functions are being performed. These functions make it necessary not only to determine the orbital elements (or the position and velocity vectors) of a spacecraft, but also to adjust its path and attitude by means of control forces and moments obtained by firing the main rocket motor or by routing commands to other on-board devices (such as thrusters, reaction wheels, control moment gyroscopes, or aerodynamic surfaces), which produce the desired forces on the spacecraft. The functions of guidance and control are meant to compute and send a series of commands to the propulsion system to alter the spacecraft velocity or attitude.

Following Thornton and Border [70], the orbit determination process requires an a priori estimate (prediction) of the spacecraft trajectory, referred to as the nominal orbit. In this estimate, expected values of the observable quantities are computed on the basis of nominal values for the trajectory and models of such observable quantities. The computed quantities differ from the correspondent observed quantities coming from the tracking system. Such differences form the residuals (computed minus observed quantities), which are due not only to random uncorrelated measurement errors (such as thermal noise in the tracking receiver), but also to errors in the trajectory and the observable models. The errors of the latter type introduce distinctive signatures, that is, characteristic marks, in the residuals. Such signatures make it possible to adjust the model parameters by means of a procedure, called weighted linear least-squares estimation, which yields the set of parameter values corresponding to the minimum weighted sum of the squares of the residuals.

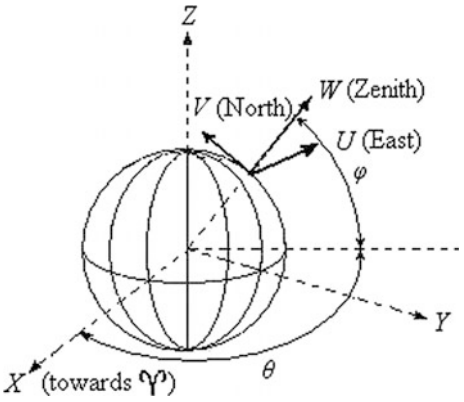
This procedure, when the data are weighted by the inverse of their error covariance, yields a minimum-variance estimator. Since this estimator provides a linear solution to a nonlinear problem, then the estimation procedure comprises more iterative steps, each of which uses the parameter estimation of the previous step, until convergence is reached.

After the orbit determination process has been completed, the orbital elements are compared with those required by the project. In case of discrepancies, trajectory correction manoeuvres must be planned and executed. For example, when the actual flight path differs from that which is required by the planned mission, it is necessary to compute the magnitude and direction of the $\Delta \boldsymbol{v}$ vector required to correct to the desired trajectory. In addition, suitable times must be computed, at which the corrective manoeuvres will be executed. At the proper time, the spacecraft attitude will change the direction of the rocket motor axis to the desired one and the thrusters will be fired for a determined interval of time. The magnitude of $\Delta \boldsymbol{v}$ is usually small (metres or tens of metres per second), because of the limited amount of propellant which can be carried on board.

Orbit trim manoeuvres are sometimes necessary in case of a satellite revolving around the Earth, in order to prevent orbital decay due to air drag, and also to perform orbit changes required by the mission objectives.

2.2 Topocentric Co-ordinate Systems

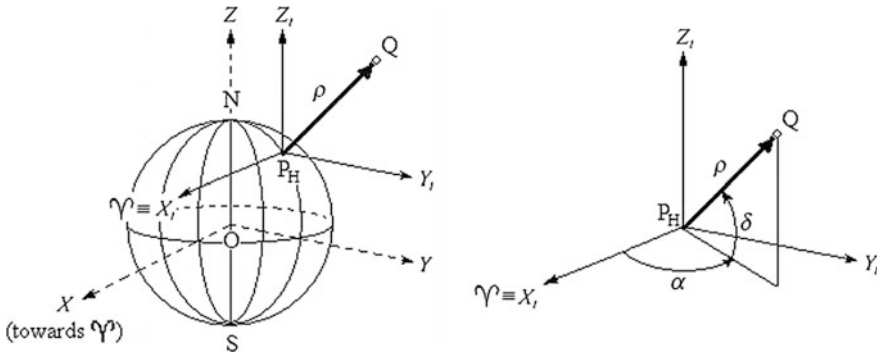
A radar station located on the surface of the Earth can measure the position and velocity of an object, for example, of an artificial satellite, with respect to the place where the station is located. However, the position vector $\boldsymbol{r}$ of interest to the observer has its origin in another place, that is, in the centre of mass of the Earth. In addition, the velocity vector $\boldsymbol{v} \equiv \boldsymbol{r}'$ of interest to the observer relates to the geocentric-equatorial reference system *XYZ*, which moves with respect to the radar station, because the Earth rotates about its axis. Thus, a series of measurements relating to reference system located in the radar station must be converted so as to provide the correspondent measurements with respect to the geocentric-equatorial system. The measurements performed at a radar station are related to the topocentric-horizon system *UVW*, which is defined below.



The topocentric-horizon reference system *UVW*, shown in the preceding figure, has its origin in the point on the surface of the Earth where the radar station is located, and its three Cartesian axes *U*, *V*, and *W* are directed so that its fundamental plane *UV* is the horizon, with the *U*-axis pointing towards east, the *V*-axis pointing towards north, and the *W*-axis perpendicular to the horizon and pointing upwards (that is, towards the zenith). As to the directions forming the horizon, the agreement of the authors is not unanimous. Some of them, for example Bate et al. [5] and Vallado [77], choose the south and east directions. Other authors, for example Montenbruck and Gill [53] and Curtis [20], choose the east and north directions. We follow the latter convention.

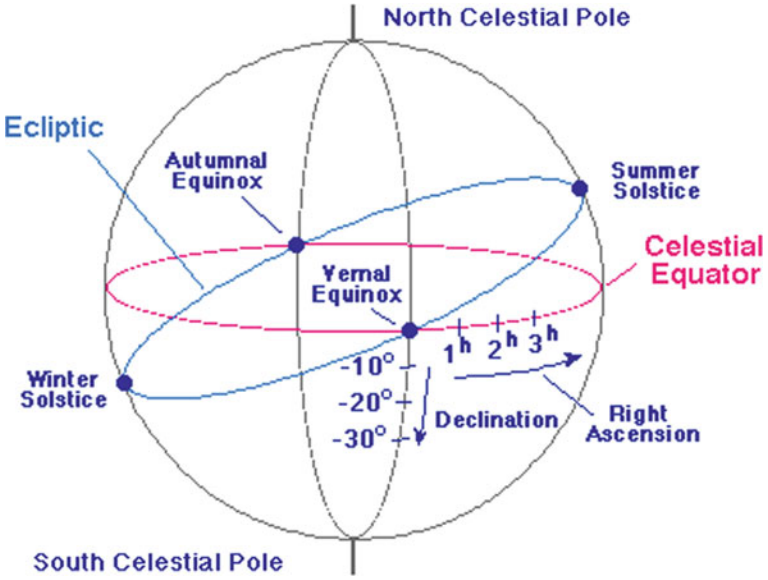
Let the unit vectors of the three axes U, V, and W be, respectively, $\boldsymbol{u}_U$, $\boldsymbol{u}_V$, and $\boldsymbol{u}_W$. Unlike the geocentric-equatorial reference system XYZ, which is fixed in space with respect to the stars and does not rotate with the Earth, the topocentric-horizon reference system UVW is evidently a non-inertial system, because it is fixed to the Earth and does rotate with it.

Another reference system of the topocentric type is the topocentric-equatorial co-ordinate system, shown on the left-hand side of the following figure.



The origin P_H of this reference system is the point on the surface of the Earth where the observer (or the radar station) is located. The axes X_t, Y_t, and Z_t of this co-ordinate system are parallel to the axes, respectively, X, Y, and Z of the geocentric-equatorial co-ordinate system having its origin O in the centre of the Earth. The fundamental plane X_tY_t of the topocentric-equatorial system $X_tY_tZ_t$ is parallel to, but not coincident with, the fundamental plane XY of the geocentric-equatorial system XYZ. Both X and X_t point towards the vernal equinox ♈. The position vector $P_HQ \equiv \boldsymbol{\rho}$, of magnitude ρ, goes from the point of observation P_H to the space object Q. The position of Q is usually expressed by means of the three co-ordinates ρ, δ, and α, where the two angles δ (declination) and α (right ascension) are shown on the right-hand side of the preceding figure, which is an enlarged view of the topocentric-equatorial co-ordinate system.

In the topocentric-equatorial system the North and South celestial poles are determined by intersecting the rotation axis of the Earth with the celestial sphere, that is, with the sky as seen from the Earth. The centre of the Earth is also the centre of the celestial sphere, as shown in the following figure, which is due to the courtesy of the University of Tennessee [76].



The celestial poles and equator are the geographic poles and equator projected up onto the celestial sphere. The equivalent of the geographic latitude in the topocentric-equatorial system is called declination and is measured in degrees North (positive numbers) or South (negative numbers) of the celestial equator. The equivalent of the geographic longitude in the topocentric-equatorial system is called right ascension, and may also be measured in degrees, but for historical reasons it is more common to measure it in time (hours, minutes, seconds), 1 hour of right ascension being equivalent to 15 degrees of apparent sky rotation.

In addition to the celestial equator, another important plane intersecting the celestial sphere is the ecliptic plane, which is the plane containing the orbit of the Earth about the Sun, inclined of about 23°.4 with respect to the celestial equator, as shown in the preceding figure. The inclination angle $\varepsilon \approx 23°.4$ is called the obliquity of the ecliptic.

The position of a space object Q in the topocentric-equatorial system is expressed by the following vector

$$\boldsymbol{\rho} = (\rho \cos\delta \cos\alpha)\boldsymbol{u}_X + (\rho \cos\delta \sin\alpha)\boldsymbol{u}_Y + (\rho \sin\delta)\boldsymbol{u}_Z$$

where $\boldsymbol{u}_X$, $\boldsymbol{u}_Y$, and $\boldsymbol{u}_Z$ are the unit vectors of, respectively, X, Y, and Z.

Let us express the position vector $\boldsymbol{\rho}$ as follows

$$\boldsymbol{\rho} = \rho\left(\frac{\boldsymbol{\rho}}{\rho}\right) = \rho\boldsymbol{u}_\rho$$

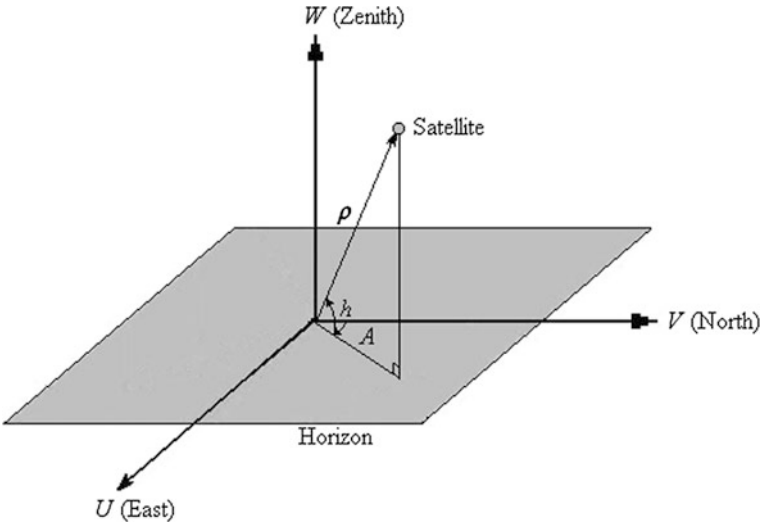
where ρ is the magnitude of $\boldsymbol{\rho}$ and $\boldsymbol{u}_\rho = \boldsymbol{\rho}/\rho$ is the unit vector having the direction of $\boldsymbol{\rho}$. By so doing, it is possible to write

$$\boldsymbol{u}_\rho = (\cos\delta\ \cos\alpha)\boldsymbol{u}_X + (\cos\delta\ \sin\alpha)\boldsymbol{u}_Y + (\sin\delta)\boldsymbol{u}_Z$$

As shown above, the geocentric-equatorial system and the topocentric-equatorial system have two different origins. Consequently, the direction cosines of the position vectors $\boldsymbol{r}$ and $\boldsymbol{\rho}$ are not the same. Thus, the topocentric declination and right ascension of a space object are not equal to the geocentric declination and right ascension of the same object. However, when the magnitude of $\boldsymbol{r}$ is much greater than the equatorial radius of the Earth, as is the case with stars and distant planets, then the differences in topocentric and geocentric angles can be neglected.

2.3 Orbit Determination from a Single Radar Observation

With reference to the following figure, let us consider an artificial satellite inside the field of view of a radar station located on the surface of the Earth.



The radar station measures the range, that is, the magnitude of the position vector going from the radar station to the satellite and the direction of this position vector with respect to a reference system having its origin in the place where the radar station is located. Reference systems of this type are called topocentric, because their origin is in the place of observation. One of these is the

topocentric-horizon reference system *UVW*, defined in Sect. 2.2 and also shown in the preceding figure.

Let $\boldsymbol{\rho}$ be the position vector of an artificial satellite with respect to the topocentric-horizon system indicated above. The direction of $\boldsymbol{\rho}$ is measured by two angles, namely azimuth (A) and altitude (h), which result from the gimbal axes on which the radar antenna is mounted.

The azimuth angle A is measured clockwise, for an observer in the direction of the positive W-axis, from the north–south direction to the projection of $\boldsymbol{\rho}$ onto the horizon (U, V) plane, so that $0 \leq A \leq 360°$; the altitude angle h is measured counterclockwise, for an observer in the direction of the positive U-axis, from the projection of $\boldsymbol{\rho}$ onto the horizon plane to $\boldsymbol{\rho}$ itself, so that $-90° \leq h \leq 90°$.

If the station is equipped with a Doppler radar, that is, of a radar capable of detecting a shift in frequency in the returning echo, then the rate of change of $\boldsymbol{\rho}$ can also be measured. On the other hand, sensors on the gimbal axes can measure the rate of change of the azimuth and altitude angles. Thus, the radar apparatus can measure six quantities related to a satellite in its field of view, namely range (ρ), azimuth angle (A), altitude angle (h), rate of change of range (ρ'), rate of change of azimuth angle (A'), and rate of change of altitude angle (h').

Let ℓ_U, ℓ_V, and ℓ_W be direction cosines of $\boldsymbol{u}_\rho = \boldsymbol{\rho}/\rho$ with respect to the topocentric-horizon system *UVW*, which system has its origin in P and its axes pointing to, respectively, east, north, and zenith. The unit vector $\boldsymbol{u}_\rho = \boldsymbol{\rho}/\rho$ is expressible as follows

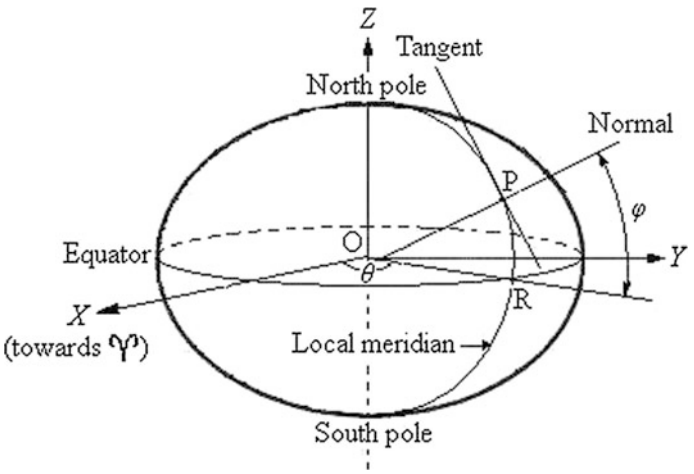
$$\boldsymbol{u}_\rho = \ell_U \boldsymbol{u}_U + \ell_V \boldsymbol{u}_V + \ell_W \boldsymbol{u}_W$$

where $\boldsymbol{u}_U$, $\boldsymbol{u}_V$, and $\boldsymbol{u}_W$ are the unit vectors pointing to, respectively, east, north, and zenith. By projecting $\boldsymbol{u}_\rho$ onto, respectively, U, V, and W, there results

$$\boldsymbol{u}_\rho = (\cos h \, \sin A)\, \boldsymbol{u}_U + (\cos h \, \cos A)\, \boldsymbol{u}_V + (\sin h)\, \boldsymbol{u}_W$$

Having obtained the components of the position vector $\boldsymbol{\rho}$ in the topocentric-horizon reference system *UVW*, it remains to transform the same components, for the purpose of computing the components of the position vector in the geocentric-equatorial reference system *XYZ*. To this end, it is necessary to determine the projections of the unit vectors $\boldsymbol{u}_U$, $\boldsymbol{u}_V$, and $\boldsymbol{u}_W$ of the topocentric-horizon system *UVW* onto the unit vectors $\boldsymbol{u}_X$, $\boldsymbol{u}_Y$, and $\boldsymbol{u}_Z$ of the geocentric-equatorial system *XYZ*.

With reference to the following figure, let φ be the geodetic latitude (defined below) of the point P where the observer or the radar station is located, and let θ be the angular distance, measured in the equatorial plane, between the X-axis (pointing towards the vernal equinox) and the local meridian passing through P.



In the preceding figure, the Earth is represented as an oblate ellipsoid, whose equatorial bulge is exaggerated for the sake of clarity. The geodetic latitude φ is the angle between the equatorial plane XY and the normal in P to the surface of the ellipsoid, which approximates the Earth. Since the Earth is not perfectly spherical, then φ does not coincide with the geocentric latitude, which is the angle between the equatorial plane XY and the normal in P to the surface of the sphere.

Let $\boldsymbol{u}_X{}^*$ be the unit vector of OR, where OR lies in the plane of the meridian passing through P and is perpendicular to the Z-axis. Let $\boldsymbol{u}_W$ be the unit vector of the zenith direction in P. By projecting $\boldsymbol{u}_W$ onto OR and the Z-axis, there results

$$\boldsymbol{u}_W = (\cos\varphi)\,\boldsymbol{u}_X^* + (\sin\varphi)\boldsymbol{u}_Z$$

where $\boldsymbol{u}_Z$ is the unit vector of the Z-axis.

On the other hand, by projecting $\boldsymbol{u}_X{}^*$ onto X and Y, there results

$$\boldsymbol{u}_X^* = (\cos\theta)\,\boldsymbol{u}_X + (\sin\theta)\,\boldsymbol{u}_Y$$

where $\boldsymbol{u}_X$ and $\boldsymbol{u}_Y$ are the unit vectors of, respectively, X and Y.

It follows that

$$\begin{aligned}\boldsymbol{u}_W = {} & [(\cos\theta)\boldsymbol{u}_X + (\sin\theta)\boldsymbol{u}_Y]\cos\varphi + (\sin\varphi)\,\boldsymbol{u}_Z = (\cos\varphi\,\cos\theta)\boldsymbol{u}_X \\ & + (\cos\varphi\,\sin\theta)\boldsymbol{u}_Y + (\sin\varphi)\boldsymbol{u}_Z\end{aligned}$$

The unit vector $\boldsymbol{u}_U$, directed towards east, is expressed as a function of $\boldsymbol{u}_X$ and $\boldsymbol{u}_Y$, by taking the vector product of $\boldsymbol{u}_Z$ and $\boldsymbol{u}_W$, as follows

$$\boldsymbol{u}_U = \frac{\boldsymbol{u}_Z \times \boldsymbol{u}_W}{|\boldsymbol{u}_Z \times \boldsymbol{u}_W|}$$

where $|\boldsymbol{u}_Z \times \boldsymbol{u}_W|$ is the magnitude of $\boldsymbol{u}_Z \times \boldsymbol{u}_W$. The following equality holds

$$\boldsymbol{a} \times \boldsymbol{b} = \begin{bmatrix} \boldsymbol{u}_X & \boldsymbol{u}_Y & \boldsymbol{u}_Z \\ a_X & a_Y & a_Z \\ b_X & b_Y & b_Z \end{bmatrix}$$

for any two vectors $\boldsymbol{a} = a_X\,\boldsymbol{u}_X + a_Y\,\boldsymbol{u}_Y + a_Z\,\boldsymbol{u}_Z$ and $\boldsymbol{b} = b_X\,\boldsymbol{u}_X + b_Y\,\boldsymbol{u}_Y + b_Z\,\boldsymbol{u}_Z$.

In the present case, there results

$$\begin{aligned} \boldsymbol{u}_Z \times \boldsymbol{u}_W &= \begin{bmatrix} \boldsymbol{u}_X & \boldsymbol{u}_Y & \boldsymbol{u}_Z \\ 0 & 0 & 1 \\ \cos\varphi\,\cos\theta & \cos\varphi\,\sin\theta & \sin\varphi \end{bmatrix} \\ &= (-\cos\varphi\,\sin\theta)\,\boldsymbol{u}_X + (\cos\varphi\,\cos\theta)\,\boldsymbol{u}_Y \end{aligned}$$

$$|\boldsymbol{u}_Z \times \boldsymbol{u}_W| = \left[(-\cos\varphi\,\sin\theta)^2 + (\cos\varphi\,\cos\theta)^2\right]^{\frac{1}{2}} = \cos\varphi$$

$$\boldsymbol{u}_U = \frac{\boldsymbol{u}_Z \times \boldsymbol{u}_W}{|\boldsymbol{u}_Z \times \boldsymbol{u}_W|} = (-\sin\theta)\,\boldsymbol{u}_X + (\cos\theta)\boldsymbol{u}_Y$$

Finally, $\boldsymbol{u}_V$ results from

$$\begin{aligned} \boldsymbol{u}_V = \boldsymbol{u}_W \times \boldsymbol{u}_U &= \begin{bmatrix} \boldsymbol{u}_X & \boldsymbol{u}_Y & \boldsymbol{u}_Z \\ \cos\varphi\,\cos\theta & \cos\varphi\,\sin\theta & \sin\varphi \\ -\sin\theta & \cos\theta & 0 \end{bmatrix} \\ &= (-\sin\varphi\,\cos\theta)\,\boldsymbol{u}_X - (\sin\varphi\,\sin\theta)\,\boldsymbol{u}_Y + (\cos\varphi\,\cos^2\theta \\ &\quad + \cos\varphi\,\sin^2\theta)\,\boldsymbol{u}_Z = (-\sin\varphi\,\cos\theta)\,\boldsymbol{u}_X - (\sin\varphi\,\sin\theta)\,\boldsymbol{u}_Y + (\cos\varphi)\,\boldsymbol{u}z \end{aligned}$$

In summary, we have obtained the following results

$$\begin{aligned} \boldsymbol{u}_U &= (-\sin\theta)\boldsymbol{u}_X + (\cos\theta)\boldsymbol{u}_Y \\ \boldsymbol{u}_V &= (-\sin\varphi\,\cos\theta)\boldsymbol{u}_X - (\sin\varphi\,\sin\theta)\boldsymbol{u}_Y + (\cos\varphi)\boldsymbol{u}_Z \\ \boldsymbol{u}_W &= (\cos\varphi\,\cos\theta)\boldsymbol{u}_X + (\cos\varphi\,\sin\theta)\boldsymbol{u}_Y + (\sin\varphi)\boldsymbol{u}_Z \end{aligned}$$

Let us consider the following scalar products

$$\begin{aligned}
\boldsymbol{u}_U \cdot \boldsymbol{u}_X &= -\sin\theta \\
\boldsymbol{u}_U \cdot \boldsymbol{u}_Y &= \cos\theta \\
\boldsymbol{u}_U \cdot \boldsymbol{u}_Z &= 0 \\
\boldsymbol{u}_V \cdot \boldsymbol{u}_X &= -\sin\varphi\,\cos\theta \\
\boldsymbol{u}_V \cdot \boldsymbol{u}_Y &= -\sin\varphi\,\sin\theta \\
\boldsymbol{u}_V \cdot \boldsymbol{u}_Z &= \cos\varphi \\
\boldsymbol{u}_W \cdot \boldsymbol{u}_X &= \cos\varphi\,\cos\theta \\
\boldsymbol{u}_W \cdot \boldsymbol{u}_Y &= \cos\varphi\,\sin\theta \\
\boldsymbol{u}_W \cdot \boldsymbol{u}_Z &= \sin\varphi
\end{aligned}$$

As shown in Sect. 1.9, in case of a transformation from perifocal co-ordinates *xyz* to geocentric-equatorial co-ordinates *XYZ*, the rotation matrix **R** is defined as follows

$$\mathbf{R} \equiv \begin{bmatrix} \boldsymbol{u}_X \cdot \boldsymbol{u}_x & \boldsymbol{u}_X \cdot \boldsymbol{u}_y & \boldsymbol{u}_X \cdot \boldsymbol{u}_z \\ \boldsymbol{u}_Y \cdot \boldsymbol{u}_x & \boldsymbol{u}_Y \cdot \boldsymbol{u}_y & \boldsymbol{u}_Y \cdot \boldsymbol{u}_z \\ \boldsymbol{u}_Z \cdot \boldsymbol{u}_x & \boldsymbol{u}_Z \cdot \boldsymbol{u}_y & \boldsymbol{u}_Z \cdot \boldsymbol{u}_z \end{bmatrix}$$

and this matrix is orthogonal, because it satisfies the condition $\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{I}$, where $\mathbf{R}^{\mathrm{T}}$ is the transpose of **R**, and **I** is the 3×3 identity matrix.

Likewise, in the present case, the rotation matrix **R**, which transforms the geocentric-equatorial co-ordinates *XYZ* into the topocentric-horizon co-ordinates *UVW*, is defined as follows

$$\mathbf{R} \equiv \begin{bmatrix} \boldsymbol{u}_U \cdot \boldsymbol{u}_X & \boldsymbol{u}_U \cdot \boldsymbol{u}_Y & \boldsymbol{u}_U \cdot \boldsymbol{u}_Z \\ \boldsymbol{u}_V \cdot \boldsymbol{u}_X & \boldsymbol{u}_V \cdot \boldsymbol{u}_Y & \boldsymbol{u}_V \cdot \boldsymbol{u}_Z \\ \boldsymbol{u}_W \cdot \boldsymbol{u}_X & \boldsymbol{u}_W \cdot \boldsymbol{u}_Y & \boldsymbol{u}_W \cdot \boldsymbol{u}_Z \end{bmatrix}$$

Thus, by replacing the scalar products by their respective values which are given above, there results

$$\begin{bmatrix} \ell_U \\ \ell_V \\ \ell_W \end{bmatrix} = \begin{bmatrix} -\sin\theta & \cos\theta & 0 \\ -\sin\varphi\,\cos\theta & -\sin\varphi\,\sin\theta & \cos\varphi \\ \cos\varphi\,\cos\theta & \cos\varphi\,\sin\theta & \sin\varphi \end{bmatrix} \begin{bmatrix} \ell_X \\ \ell_Y \\ \ell_Z \end{bmatrix}$$

where ℓ_U, ℓ_V, and ℓ_W are the direction cosines of the topocentric-horizon co-ordinates, and ℓ_X, ℓ_Y, and ℓ_Z are the direction cosines of the geocentric-equatorial co-ordinates.

The preceding expression defines the transformation from geocentric-equatorial co-ordinates *XYZ* to topocentric-horizon co-ordinates *UVW*.

As is easy to verify, the 3 × 3 square matrix written above, that is,

$$\mathbf{R} \equiv \begin{bmatrix} -\sin\theta & \cos\theta & 0 \\ -\sin\varphi\,\cos\theta & -\sin\varphi\,\sin\theta & \cos\varphi \\ \cos\varphi\,\cos\theta & \cos\varphi\,\sin\theta & \sin\varphi \end{bmatrix}$$

is an orthogonal matrix, because it satisfies the condition

$$\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{I}$$

where $\mathbf{R}^{\mathrm{T}}$ is the transpose of $\mathbf{R}$ and $\mathbf{I}$ is the 3 × 3 identity matrix.

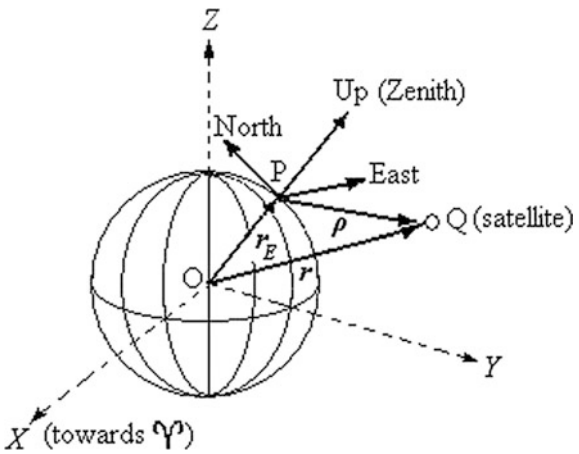
Therefore, in order to transform topocentric-horizon co-ordinates *UVW* into geocentric-equatorial co-ordinates *XYZ*, we use $\mathbf{R}^{\mathrm{T}}$, the transpose of $\mathbf{R}$, that is,

$$\begin{bmatrix} \ell_X \\ \ell_Y \\ \ell_Z \end{bmatrix} = \begin{bmatrix} -\sin\theta & -\sin\varphi\,\cos\theta & \cos\varphi\,\cos\theta \\ \cos\theta & -\sin\varphi\,\sin\theta & \cos\varphi\,\sin\theta \\ 0 & \cos\varphi & \sin\varphi \end{bmatrix} \begin{bmatrix} \ell_U \\ \ell_V \\ \ell_W \end{bmatrix}$$

The matrices $\mathbf{R}$ and $\mathbf{R}^{\mathrm{T}}$ shown above can also be used for transformations from topocentric-equatorial co-ordinates $X_tY_tZ_t$ to topocentric-horizon co-ordinates *UVW*, and from topocentric-horizon co-ordinates *UVW* to topocentric-equatorial co-ordinates $X_tY_tZ_t$, because the unit vectors $\boldsymbol{u}_X$, $\boldsymbol{u}_Y$, and $\boldsymbol{u}_Z$ of the system *XYZ* coincide with those of the system $X_tY_tZ_t$. The following relation holds between the position vector ($\boldsymbol{\rho}$) in the topocentric-horizon system *UVW* and the position vector ($\boldsymbol{r}$) in the geocentric-equatorial system *XYZ*:

$$\boldsymbol{r} = \boldsymbol{r}_E + \boldsymbol{\rho}$$

where $\boldsymbol{r}_E \equiv \mathrm{OP}$ is the vector going from the centre of mass, O, of the Earth to the position, P, of the radar station on the surface of the Earth. This can be understood by considering the following figure.

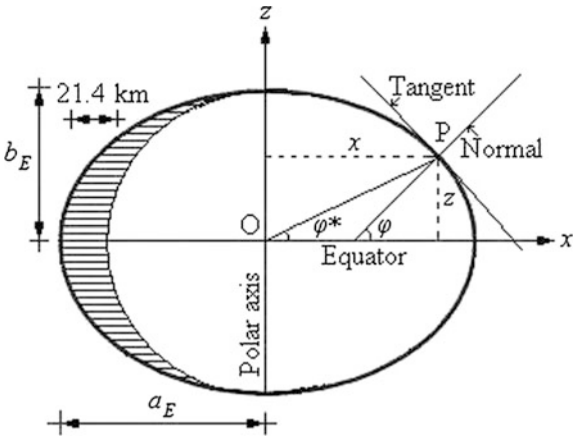


If the Earth were a perfect sphere having a radius equal to r_E, then the local vertical passing through the radar site P would join (extended downward) this site with the centre O of the Earth. In this case, $\boldsymbol{r}_E$ would be expressible as follows

$$\boldsymbol{r}_E = r_E \boldsymbol{u}_Z$$

Such is not the case in practice, because the shape of the Earth is not perfectly spherical.

The following section of this paragraph takes account of the non-spherical shape of the Earth. In the model described below, the Earth is approximated to an ellipsoid, as shown in the following figure.



Consequently, latitude cannot be used any more as a spherical co-ordinate and the radius of the Earth depends on latitude. Thus, a point on the surface of the Earth will be represented by two rectangular co-ordinates (x and z) and the longitude (λ), because the latter has the same meaning in a spherical as in an oblate Earth. The ellipsoid is an approximation to the true shape of the Earth.

The cross section of the Earth ellipsoid along a meridian is an ellipse, whose major semi-axis (a_E) is equal to the equatorial radius of the Earth and whose minor semi-axis (b_E) is equal to the polar radius of the Earth. Cross sections of the ellipsoid parallel to the equator are circles. The shaded area represents the bulge, which is maximum (21.4 km) at the equator and is zero at the poles.

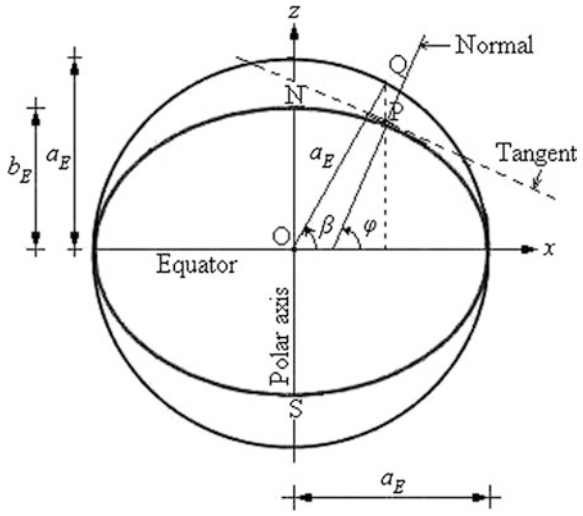
The values considered here for the major semi-axis a_E and the flattening (or oblateness) $f = (a_E - b_E)/a_E$ of the Earth ellipsoid are those given in Ref. [58], that is,

$$a_E = 6378.137\,\text{km} \qquad f = \frac{1}{298.257223563} \approx 0.0033528$$

Hence, the eccentricity (e_E) and the minor semi-axis (b_E) of the ellipsoid result, respectively, from $e_E = (2f - f^2)^{1/2} = 0.08181919$ and $b_E = a_E(1 - f) = 6356.752$ km. The reference ellipsoid defined above is an approximation to the geoid, that is, to the equipotential surface of the Earth gravity field which best fits, in the least-squares sense, the global mean sea level. In describing the non-spherical shape of the Earth, some authors, Herrick [34, 35] for one, use f instead of e_E.

The angle φ^* between the equatorial plane and the radius OP (between the centre O of the Earth and the point P of the Earth surface where the observer is located) is called geocentric latitude. The angle φ between the equatorial plane and the normal in P to the surface of the ellipsoid is called geodetic latitude and is used in the maps and charts of the Earth. The normal in P to the surface of the ellipsoid is the direction which a plumb bob would indicate, were it not for local anomalies in the gravitational field of the Earth. The angle φ_a between the equatorial plane and the actual plumb bob vertical uncorrected for these anomalies is called the astronomical latitude. In practice, since the difference between the reference ellipsoid and the mean sea level is small, then the difference between geodetic latitude φ and the astronomical latitude φ_a is negligible.

We must now calculate the station co-ordinates of a point on the surface of the reference ellipsoid, when we are given the geodetic latitude, the longitude, and the height of that point above the mean sea level, which is assumed to be the height above the reference ellipsoid. To this end, with reference to the following figure, we first determine the co-ordinates x and z of a point P on the ellipse, assuming that we know the geodetic latitude φ of P; then these co-ordinates will be adjusted for another point of which we know the height above the surface of the ellipsoid in the direction of the normal in P.



The angle β, called the reduced altitude, is introduced to express the x and z co-ordinates as functions of the equatorial radius of the Earth and of β itself.

By considering the elliptic cross section of the Earth (having a_E as its major semi-axis along the equator and b_E as its minor semi-axis along the polar axis of the Earth) and the corresponding auxiliary circumference (having a_E as its radius), the Cartesian co-ordinates of P are expressible as follows

$$x = a_E \cos\beta \qquad z = \left(\frac{b_E}{a_E}\right) a_E \sin\beta$$

In Sect. 1.3, we have shown that, for an ellipse, there results

$$b_E = a_E\left(1 - e_E^2\right)^{\frac{1}{2}}$$

where e_E is the eccentricity of the ellipse. Thus, there also results

$$x = a_E \cos\beta \qquad z = a_E\left(1 - e_E^2\right)^{\frac{1}{2}} \sin\beta$$

Now, cos β and sin β must be expressed as functions of the geodetic latitude φ and of the constant quantities a_E and b_E. To this end, since the slope of the tangent to the ellipse is $\mathrm{d}z/\mathrm{d}x$, and the slope of the normal is $-\mathrm{d}x/\mathrm{d}z = \tan\varphi$, then the differentials of the expressions written above

$$x = a_E \cos\beta \qquad z = a_E\left(1 - e_E^2\right)^{\frac{1}{2}} \sin\beta$$

are

$$\mathrm{d}x = -a_E \sin\beta\, \mathrm{d}\beta \qquad \mathrm{d}z = a_E\left(1 - e_E^2\right)^{\frac{1}{2}} \cos\beta\, \mathrm{d}\beta$$

and consequently

$$\tan\varphi = -\frac{\mathrm{d}x}{\mathrm{d}z} = \frac{\tan\beta}{\left(1 - e_E^2\right)^{\frac{1}{2}}}$$

and also

$$\tan\beta = \left(1 - e_E^2\right)^{\frac{1}{2}} \tan\varphi = \left(1 - e_E^2\right)^{\frac{1}{2}} \frac{\sin\varphi}{\cos\varphi}$$

Now, the last expression can be written in the following form

$$\tan\beta = \frac{A}{B}$$

where $A = (1 - e_E^2)^{1/2} \sin \varphi$ and $B = \cos \varphi$. In addition, since

$$\sin \alpha = \pm \frac{\tan \alpha}{(1 + \tan^2 \alpha)^{\frac{1}{2}}} \qquad \cos \alpha = \pm \frac{1}{(1 + \tan^2 \alpha)^{\frac{1}{2}}}$$

are trigonometric identities, then there results after simplification

$$\sin \beta = \pm \frac{\left(1 - e_E^2\right)^{\frac{1}{2}} \sin \varphi}{\left(1 - e_E^2 \sin^2 \varphi\right)^{\frac{1}{2}}} \qquad \cos \beta = \pm \frac{\cos \varphi}{\left(1 - e_E^2 \sin^2 \varphi\right)^{\frac{1}{2}}}$$

The preceding expressions make it possible to write the x and z co-ordinates of a point P on the surface of the ellipsoid as follows

$$x = a_E \cos \beta = \frac{a_E \cos \varphi}{\left(1 - e_E^2 \sin^2 \varphi\right)^{\frac{1}{2}}}$$

$$z = a_E \left(1 - e_E^2\right)^{\frac{1}{2}} \sin \beta = \frac{a_E \left(1 - e_E^2\right) \sin \varphi}{\left(1 - e_E^2 \sin^2 \varphi\right)^{\frac{1}{2}}}$$

Let us consider now another point P_H, placed at a height H above the ellipsoid, that is, at a height H above the mean sea level, along the normal in P to the ellipsoid. By inspection of the preceding figure, the x and z components of the height H result

$$\Delta x = H \cos \varphi$$
$$\Delta z = H \sin \varphi$$

The quantities Δx and Δy are the increments which must be added to the co-ordinates of P derived above, to obtain the co-ordinates of P_H as functions of geodetic latitude (φ), altitude (H) above the mean sea level, and equatorial radius (a_E) and eccentricity (e_E) of the ellipsoid representing the Earth:

$$x_H = \left[\frac{a_E}{\left(1 - e_E^2 \sin^2 \varphi\right)^{\frac{1}{2}}} + H\right] \cos \varphi \quad z_H = \left[\frac{a_E \left(1 - e_E^2\right)}{\left(1 - e_E^2 \sin^2 \varphi\right)^{\frac{1}{2}}} + H\right] \sin \varphi$$

When the flattening (f) of the Earth ellipsoid is used instead of e_E, that is, when e_E^2 is replaced by $2f - f^2$, then the expressions derived above become

$$x_H = \left\{\frac{a_E}{\left[1 - (2f - f^2) \sin^2 \varphi\right]^{\frac{1}{2}}} + H\right\} \cos \varphi$$

$$z_H = \left\{\frac{a_E (1 - f)^2}{\left[1 - (2f - f^2) \sin^2 \varphi\right]^{\frac{1}{2}}} + H\right\} \sin \varphi$$

The third co-ordinate of P_H is the east longitude (λ) of P_H. Thus, if we know the Greenwich sidereal time (θ_G), it can be used with the east longitude of P_H to compute the local sidereal time (θ), which is the angle between the *X*-axis (pointing towards the vernal equinox) of the geocentric-equatorial system *XYZ* and the local meridian. As will be shown below, the x_H and z_H co-ordinates plus the angle θ, in turn, are used to locate the observer site in the geocentric-equatorial system.

In the model representing the Earth as a perfect sphere of radius r_E, $\boldsymbol{r}_E$ was the vector from the centre of the Earth to the position of the radar station on the surface of the Earth, with respect to the geocentric-equatorial system *XYZ*.

Now, likewise, in the model representing the Earth as an oblate ellipsoid, let $\boldsymbol{r}_E$ be the position vector from the centre, O, of the Earth to the point P_H, with respect to the same system *XYZ*. The components of $\boldsymbol{r}_E$ are then

$$\boldsymbol{r}_E = (x_H \cos\theta)\boldsymbol{u}_X + (x_H \sin\theta)\boldsymbol{u}_Y + z_H \boldsymbol{u}_Z$$

where $\boldsymbol{u}_X$, $\boldsymbol{u}_Y$, and $\boldsymbol{u}_Z$ are the unit vectors along, respectively, *X*, *Y*, and *Z*.

Now, the vectors which have been expressed in the topocentric-horizon (or *UVW*) co-ordinates must be converted into the geocentric-equatorial (or *XYZ*) co-ordinates. The angle θ_G from the unit vector $\boldsymbol{u}_X$ (pointing towards the vernal equinox) and the Greenwich meridian is the Greenwich sidereal time. This angle, added to the geographic longitude λ of the radar site measured eastward from Greenwich, yields the local sidereal time θ, as follows

$$\theta = \theta_G + \lambda$$

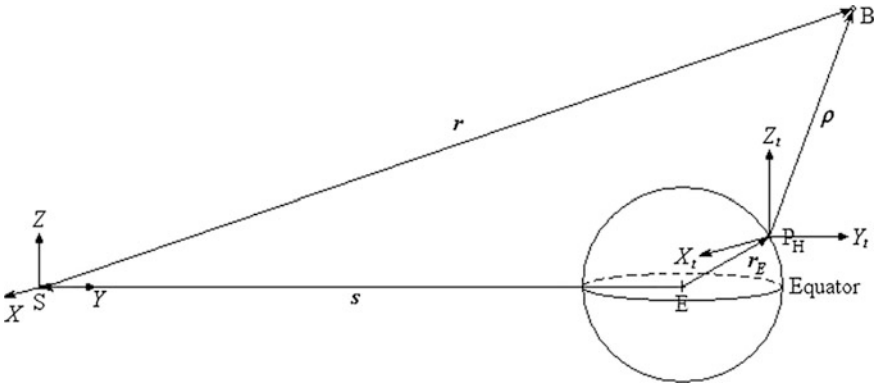
The angles φ and θ determine the relation between the topocentric-horizon system and the geocentric-equatorial system. To this end, it is necessary to compute θ_G at any given time *t*. When θ_G is known, θ results from the expression ($\theta = \theta_G + \lambda$) given above. Let θ_{G0} be the Greenwich sidereal time at some particular time t_0 (e.g. t_0 = 0 h Universal Time on the 1st of January of the year of interest). When θ_{G0} is known, then the local sidereal time θ at any given time *t* can be determined from

$$\theta = \theta_{G0} + \omega_E(t - t_0) + \lambda$$

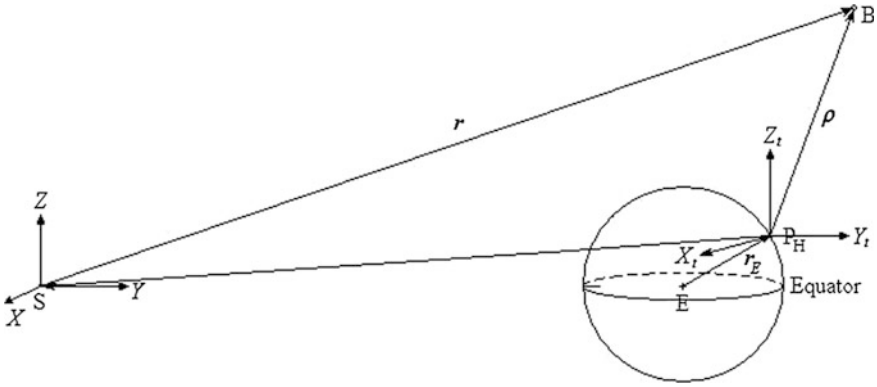
where $\omega_E = 7.292 \times 10^{-5}$ rad/s is the angular velocity of the Earth about its axis. The following paragraph will show how to compute θ_{G0}, that is, θ_G at 0 h Universal Time for every day of the year of interest. The same paragraph will also provide further information on the concept of sidereal time introduced here.

It is to be noted that the vector $\boldsymbol{r}_E \equiv \mathrm{OP_H}$ shown above is the position vector from the centre, O, of the Earth to the point P_H (where the observer is located), with respect to the geocentric-equatorial system *XYZ*. The knowledge of $\boldsymbol{r}_E$ and $\boldsymbol{\rho}$, where $\boldsymbol{\rho} \equiv \mathrm{P_HQ}$ is the position vector of a celestial body Q in the topocentric-horizon system *UVW*, suffices for orbits of bodies revolving about the Earth, because $\boldsymbol{r} = \boldsymbol{r}_E + \boldsymbol{\rho}$, where for such orbits $\boldsymbol{r}$ is the position vector from the centre of mass, O, of the Earth to the body Q revolving in a geocentric orbit.

Following Boulet [12], we consider now bodies revolving about the Sun. With reference to the following figure, let $\boldsymbol{r}$ be the position vector of a body B revolving about the Sun, with respect to a celestial reference system XYZ whose origin is the centre of mass of the Sun, and whose plane XY is the Earth equator.



Let E and S be the centres of mass of, respectively, the Earth and the Sun. Let $\boldsymbol{\rho}$ be the topocentric position vector of the body B. As shown below, the knowledge of $\boldsymbol{r}_E \equiv \mathrm{EP_H}$ and $\boldsymbol{s} \equiv \mathrm{ES}$ is necessary to compute $\mathrm{P_HS} = -\mathrm{EP_H} + \boldsymbol{s}$.



The vector $\mathrm{EP_H}$ defines the position of the observer with respect to the centre of the Earth at time t. The local sidereal time θ, the latitude φ, and the east longitude λ are used to compute the rectangular components of $\mathrm{EP_H}$.

The most common unit of measurement for distances of celestial bodies within the solar system is the astronomical unit (AU), which has been defined in Sect. 1.13
.

Let $a_E = 4.263523 \times 10^{-5}$ AU (from Ref. [12]) be the radius of the Earth expressed in astronomical units (AU). The rectangular components of $\boldsymbol{r}_E \equiv \mathrm{EP_H}$ at time t are

$$\begin{aligned} r_{EX} &= a_E \cos\varphi \cos\theta \\ r_{EY} &= a_E \cos\varphi \sin\theta \\ r_{EZ} &= a_E \sin\varphi \end{aligned}$$

As to the components X_S, Y_S, and Z_S of the solar vector $\boldsymbol{s} \equiv \mathrm{ES}$, the geocentric rectangular equatorial co-ordinates of the Sun, with respect to the mean equator and equinox of J2000.0 (the 1st of January 2000 at noon), are published each year, for every day of the year, in "The Astronomical Almanac". The tabular values of the co-ordinates of the Sun can be interpolated so as to obtain the values relating to the time of interest. Alternatively, a simple algorithm can be used for computing the angular co-ordinates of the Sun to an accuracy of about 1 arc-minute within two centuries of 2000, which is given by the U.S. Naval Observatory [72]. This and other similar algorithms require the knowledge of concepts on the measurement of time in astronomy which have been not yet introduced to the reader and will be shown at the end of Sect. 2.4.

When X_S, Y_S, and Z_S are known, the components of $\mathrm{P_HS}$ result from

$$\begin{aligned} (\mathrm{P_HS})_X &= -(a_E \cos\varphi \cos\theta) + X_S \\ (\mathrm{P_HS})_Y &= -(a_E \cos\varphi \sin\theta) + Y_S \\ (\mathrm{P_HS})_Z &= -(a_E \sin\varphi) + Z_S \end{aligned}$$

and therefore $\boldsymbol{r}$ is given, in the heliocentric system XYZ, by

$$\boldsymbol{r} = -\mathrm{P_HS} + \boldsymbol{\rho}$$

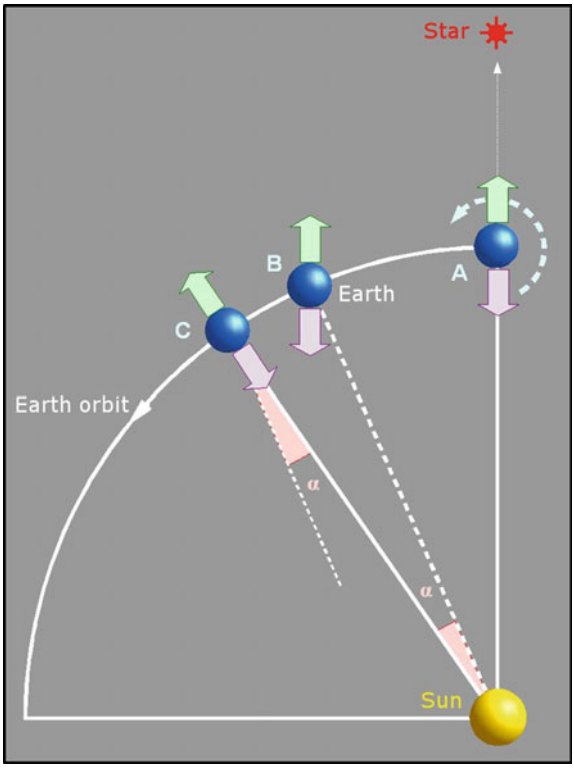
2.4 The Measurement of Time in Astronomy

As has been shown in the preceding paragraph, the determination of the orbit of either an artificial satellite or a natural celestial body by means of observations makes it necessary to record not only the observations themselves but also the times at which they have been performed.

The time commonly used in every-day life is the time indicated by ordinary clocks, that is, the mean solar time. The measurement of solar time is based on the apparent motion of the Sun through the sky, due to the rotation of the Earth about its axis, as seen by an observer placed in a fixed point of the surface of the Earth. Thus, a solar day is the interval of time taken by the Sun, in its apparent motion around the Earth, to travel an arc of 360° along the sky. In other words, a solar day

is the time elapsed between two consecutive passages of the Sun across the same meridian of the place of observation. By meridian of the place of observation we mean the great circle passing through the two celestial poles and the zenith of the site. A solar day is measured by the astronomers from noon to noon, and is divided into 24 h, or $24 \times 60 = 1440$ min, or $24 \times 60 \times 60 = 86400$ s.

However, the rotation of the Earth about its axis does not comprise exactly 360° in a solar day, because the Earth also revolves about the Sun. In order for the Earth to do one revolution about the Sun, that is, in order for the Earth to travel an arc of 360° about the Sun, 365.25 solar days are necessary.



Thus, 360° travelled in 365.25 solar days are about 0°.985626 per solar day. This means that, in the course of a solar day, the direction of the Sun seen from a point of the Earth changes by about 1°. This also means that the Earth must travel an arc of about 361° in order for the Sun to travel (in its apparent motion with respect to the Earth) an arc of 360° along the sky.

The astronomers are interested to determine the rotation time of the Earth about its axis not only with respect to the Sun, but also with respect to the distant stars (the so-called fixed stars). In other words, they want to determine how long it takes the Earth to rotate 360° around its axis with respect to the fixed stars. This period of rotation is called sidereal day and can be determined by observing the starry sky by

night. The difference between solar day and sidereal day is illustrated in the preceding figure (due to the courtesy of Wikimedia, Ref. [1]), where the amplitude of the angle $\alpha = 0°.985626$ has been exaggerated for the sake of clarity.

When the Earth is in position A along its orbit, the Sun and a given star of reference are, both of them, overhead. In position B, the Earth has performed a complete rotation about its axis, so that the star of reference (but not the Sun) is overhead again. The time interval taken by the Earth to move from A to B equals one sidereal day. Shortly after time t_B, that is, at time t_C, the Sun is overhead again. The time interval taken by the Earth to move from A to C equals one solar day. Thus, two consecutive passages of a fixed star (chosen as the star of reference) across the same meridian of the Earth measure one sidereal day, which is also divided into sidereal hours, minutes, and seconds.

Since two consecutive passages of the star of reference through the same meridian take about 23 h, 56 min and 4 s of ordinary solar time to occur, then a sidereal day is on the average 3 min and 56 s shorter than a solar day. In other terms, a solar day is longer than a sidereal day by a factor which is about

$$\frac{360 + 0.985626}{360} = 1.00273785$$

The United States Naval Observatory [75] indicates a value of 1.00273790935 for this factor. This is the value which will be used below.

In general terms, sidereal time is defined by some authors (see, e.g., Refs. [34, 53]) as the hour angle of the vernal equinox, that is, the hour angle of the ascending node of the ecliptic on the celestial equator. This is because the daily motion of the vernal equinox measures the rotation of the Earth with respect to the fixed stars, not to the Sun. By the way, the hour angle of any given point is the angle between the half plane determined by the rotation axis of the Earth and the zenith (half of the meridian plane) and the half plane determined by the rotation axis of the Earth and the given point. This angle is taken with the minus sign if the given point is eastward of the meridian plane, and with the plus sign if the given point is westward of the meridian plane. The hour angle is usually expressed in time units (hours, minutes, and seconds), where 24 h correspond to 360°. In particular, a sidereal day is the interval of time taken by the hour angle of the vernal equinox to increase by 360°.

There are no two solar days having the same duration. This is because the axis of rotation of the Earth is not perpendicular to the ecliptic (that is, to the plane containing the orbit of the Earth around the Sun), and also because the orbit of the Earth is slightly elliptical. Since the areal velocity of the Earth is constant, the Earth moves faster along its orbit at perihelion (early in January) than it does at aphelion (early in July). Thus, a mean solar day is defined by making reference to the Earth revolving about the Sun in a circular orbit placed in the same plane as that of the ecliptic and having the same period as that of the real elliptical orbit [5].

The difference between the true and the mean solar time is called the equation of time. The two causes (that is, the eccentricity of the Earth orbit and the obliquity of

the ecliptic) mentioned above produce effects which overlap with different periods of time, because the eccentricity has a period of one year, whereas the obliquity of the ecliptic has a period of half a year. Consequently, the equation of time has two minima and two maxima per year, as has been shown by Husfeld and Kronberg [38].

As we all know, the Earth comprises 24 time zones, equally spaced in longitude of about 15°, so that the mean solar time of each zone differs by ± 1 h from the mean solar times of the two contiguous time zones. The world time zones are shown in the following figure, due to the courtesy of CIA [16].

Of all these mean solar times, that which relates to the Greenwich meridian is called Greenwich Mean Time (GMT) or Universal Time (UT1) or Zulu time (Z).

There are several versions of Universal Time, the principal of which are the following:

- UT0 is the mean solar time of the Greenwich meridian obtained from direct astronomical observations.
- UT1 is UT0 corrected for the effects of small movements of the Earth relative to the axis of rotation (polar variation).
- UT2 is UT1 corrected for the effects of a small seasonal fluctuation in the rate of rotation of the Earth. UT2 is mainly of historic interest and rarely used.

Universal Time is based on the imaginary mean Sun, which takes into account the effects on the solar day of the weakly elliptic orbit of the Earth about the Sun. As shown above, UT1 is a measure of the rotation angle of the Earth around its axis as resulting from astronomical observations, account being taken of slight movements of the Earth poles of rotation. UT1 predicts the solar position with sufficient accuracy for astronomical purposes, but the duration of a second derived from UT1

varies noticeably because of variations of the Earth rotation (the velocity of the Earth rotation is variable).

As has been shown by various authors (see, e.g., McCarthy [52]), these variations may be classified into three types: secular, irregular, and periodic.

The secular variation of the Earth rotational velocity depends on the apparently linear increase in the length of the day due mainly to tidal friction. The Moon and (to a lesser extent) the Sun raise tides in the oceans. Friction carries the maximum tide ahead of the line joining the centres of the Earth and Moon. The resulting couple decreases the velocity of rotation of the Earth and increases the orbital momentum of the Moon. In other words, the Earth loses energy and slows down, whereas the Moon gains this energy and increases its orbital period and distance from the Earth. According to McCarthy, the decrease of the Earth rotational velocity results in an increase of the day duration by about 0.0005 to 0.0035 s per century. According to Espenak and Meeus [23], the secular acceleration of the Moon implies an increase in the day length of about 0.0023 s per century. The irregular variation of the Earth rotational velocity appears to be due to random accelerations, but may be correlated with physical processes occurring on or within the Earth. The resulting variation in the day length is evaluated by McCarthy to 0.001 s over the past 200 years.

Finally, the periodic variation of the Earth rotational velocity is associated with periodically repeatable physical processes affecting the Earth. According to McCarthy, tides raised in the solid Earth by the Moon and the Sun produce changes in the length of the day with amplitudes of the order of 0.00005 s and with periods of 18.6 years, 1 year, ½ year, 27.55 days, 13.66 days, and others. Due to the reasons indicated above, the rotational velocity of the Earth varies in an unpredictable manner.

In the most common civil usage, Universal Time is related to a time called Co-ordinated Universal Time (UTC), which provides the basis for Civil Time.

UTC is kept by several time laboratories around the world and is measured by high-precision atomic clocks. The length of a UTC second is defined in terms of an atomic transition of Caesium under specific conditions and does not depend on the observation of astronomical phenomena. The international standard UTC is provided by the International Bureau of Weights and Measures on the basis of the data coming from the timing laboratories and is accurate to about 1×10^{-9} s (that is, 1 ns) per day. UTC is made public by various radio stations, which also provide the difference between UTC and UT1, so that the latter can be computed as a function of the former. As mentioned above, UTC is the basis to compute Civil Time in the time zones of the Earth. By international agreements, UTC (which is measured by atomic clocks) cannot differ from UT1 (which is measured by the rotation of the Earth) by more than 0.9 s. When this limit is approached, a one-second change (called leap second) is introduced into UTC. The UTC can be easily computed starting from the local Civil Time (CT), which depends on where the observer is placed, as follows

$$\mathrm{UTC} = \mathrm{CT} + z$$

where z is the number of standard time zones by which the observer is displaced to the west of the Greenwich meridian. Some of the Earth time zones are indicated below:

- Western European Time (0 h of difference with respect to UTC);
- Central European Time (+1 h);
- USA, comprising in turn:
  - Atlantic Standard Time (−4 h);
  - Eastern Standard Time (−5 h);
  - Central Standard Time (−6 h);
  - Mountain Standard Time (−7 h); and
  - Pacific Standard Time (−8 h);
- Moscow Time (+3 h);
- Tokyo Time (+9 h).

In particular, there are four standard time zones in the conterminous USA. From east to west they are: Eastern Standard Time (EST), Central Standard Time (CST), Mountain Standard Time (MST) and Pacific Standard Time (PST), as shown in the following figure (due to the courtesy of the National Atlas of the United States [59]).

Consequently, for any given place within the conterminous USA, UTC can be computed by means of one of the following expressions

$$\begin{aligned} \text{UTC} &= \text{EST} + 5 \\ \text{UTC} &= \text{CST} + 6 \\ \text{UTC} &= \text{MST} + 7 \\ \text{UTC} &= \text{PST} + 8 \end{aligned}$$

For example, we want to compute the UTC corresponding to

$$1{:}21{:}15 \text{ pm EST}$$

First, we convert this time to a 24-h clock, as follows

$$1{:}21{:}15\,\text{pm EST} + 12^{\text{h}} = 13{:}21{:}15\,\text{EST}$$

Then, adding 5 h (see above) to 13:21:15 EST, we have

$$13{:}21{:}15 \text{ EST} + 5^{\text{h}} = 18^{\text{h}}21^{\text{m}}15^{\text{s}}\,\text{UTC}$$

The International System of Units (usually abbreviated as SI) defines the second as the duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the Caesium 133 atom. This defines, in the abstract, the Atomic Time [38]. In practice, since atomic clocks do not agree fully with one another, the weighted mean of several hundred atomic clocks, placed in various laboratories on the Earth, is used to define the so-called International Atomic Time (usually abbreviated as TAI).

As has been shown above, at about 1930, the Earth rotational period was found to be irregular and therefore, for purposes of orbital calculations, time based on Earth rotation was abandoned to choose a more uniform time scale based on the Earth orbit about the Sun. Thus, the Ephemeris Time was defined as the time scale which, together with the laws of motion, predicts correctly the positions of celestial bodies. Because of this property, the Ephemeris Time is used as the argument of the ephemerides, that is, of the tables giving the positions of the Sun, Moon, planets, and respective satellites as a function of time. For this purpose, in 1958, the International Astronomical Union (IAU) decided that "Ephemeris Time is reckoned from the instant, near the beginning of the calendar year AD 1900, when the geometric mean longitude of the Sun was 279 degrees 41 min 48.04 s, at which instant the measure of Ephemeris Time was 1900 January 0, 12 o'clock precisely". Ephemeris Time was used for ephemeris calculations for the solar system until 1979, when it was replaced by Terrestrial Dynamical Time (TDT). TDT takes into account relativistic effects and is based on International Atomic Time (TAI), which has been defined above. To ensure continuity with Ephemeris Time, TDT was defined to match ET for the date 1977 January 1. In 1991, the IAU modified the

definition of TDT to make it more precise. It was also renamed Terrestrial Time (TT); however, the old name (Terrestrial Dynamical Time) is still used. According to the United States Naval Observatory [73], Terrestrial Time is effectively equal to International Atomic Time plus 32.184 s exactly. Thus, the epoch designated "J2000.0" is specified as Julian date (see definition below) 2451545.0 TT, or 2000 January 1, 12 h TT. This epoch is also expressed as 2000 January 1, 11:59:27.816 TAI, or 2000 January 1, 11:58:55.816 UTC [74]. As a result of the decrease, at an irregular rate, of the rotational velocity of the Earth, the difference

$$\Delta T = \text{TT} - \text{UT1}$$

also decreases irregularly. The exact value of ΔT cannot be predicted and can only result from the historical record and observations.

Since this value must be known to predict the correct times of astronomical events (such as eclipses), then a series of polynomial expressions have been created by Morrison and Stephenson [55], Espenak and Meeus [23], and Islam et al. [40] to evaluate approximately ΔT during intervals of time of interest. Some examples (from Ref. [23]) of such evaluations are given below. Let the decimal year (y) be defined as follows

$$y = \text{year} + \frac{\text{month} - 0.5}{12}$$

which expression gives y for the middle of the month of interest.

Let $t = y - 2000$. Then, the approximate value of ΔT (in seconds) is given by:

$$\begin{aligned}
\Delta T &= 63.86 + 0.3345\,t - 0.060374\,t^2 + 0.0017275\,t^3 + 0.000651814\,t^4 \\
&\quad + 0.00002373599t^5 \,(\text{with year between } 1986 \text{ and } 2005) \\
\Delta T &= 62.92 + 0.32217t + 0.005589t^2 \,(\text{with year between } 2005 \text{ and } 2050) \\
\Delta T &= -20 + 32\,[(y - 1820)/100]^2 - 0.5629\,(2150 - y) \\
&\quad (\text{with year between } 2050 \text{ and } 2150)
\end{aligned}$$

Section 2.3 has shown that a change of reference system from the topocentric-horizon (or *UVW*) system to the geocentric-equatorial (or *XYZ*) system requires the latitude (φ) and longitude (λ) of the observer and the Greenwich sidereal time (θ_G). If θ_G were known on a particular day and at a particular time, then θ_G could be determined for any future day and time, because we know, as has been shown above, that the Earth turns through 1.00273790935 rotations on its axis per day.

Let θ_{G0} be the value of θ_G relating to 0 h UT1 on the 1st of January of a particular year. Let day 0 designate the 1st of January of the chosen year. Then, day 30 will designate the 31st of January, day 58 will designate the 28th of February, and so on, as shown in the following table, where the numbers in parentheses refer to the leap years.

Date	Day No.	Date	Day No.
31 January	30	31 July	211(212)
28 February	58	31 August	242(243)
31 March	89(90)	30 September	272(273)
30 April	119(120)	31 October	303(304)
31 May	150(151)	30 November	333(334)
30 June	180(181)	31 December	364(365)

In addition, any time can be expressed as a decimal fraction of a day. By so doing, any given set of values specifying a date and a time can be converted into a single floating-point number (D), such that its integral part indicates the number of days and its fractional part indicates the fraction of days elapsed from the chosen origin of times.

Consequently, θ_G is expressible in degrees as follows

$$\theta_G = \theta_{G0} + 1.00273790935 \times 360 \times D$$

or in radians as follows

$$\theta_G = \theta_{G0} + 1.00273790935 \times 2\pi \times D$$

The methods used to determine sidereal time are based on a continuous count of days and fractions of day obtained by means of the Julian Day. The Julian Day or Julian Day Number (*JDN*) is the number of days that have elapsed since an initial epoch. The initial epoch has been set at noon Universal Time, Monday, the 1st of January 4713 BC in the proleptic Julian Calendar (that is, in the Julian calendar extended earlier in time so as to include dates preceding 4 AD). The initial epoch, which is counted as Julian Day 0, corresponds to the 24th of November 4714 BC in the proleptic Gregorian calendar (that is, in the calendar obtained by extending the Gregorian calendar backwards to years earlier than 1582, using the Gregorian leap year rules). The Julian date (*JD*) is a continuous count of days and fractions thereof elapsed since the same initial epoch, as follows

$$JD = JDN + \frac{\text{UT1}}{24}$$

According to the definition given above, a Julian date comprises an integral part and a fractional part. The integral part of a Julian date is the Julian Day Number, whereas its fractional part is the time of day elapsed from noon UT1 expressed as a decimal fraction of a day.

Examples of this fractional part are given below. A fractional part equal to 0.5 indicates midnight UT1, because the Julian Day begins at noon. A fractional part of 0.1 indicates UT1/24 = 0.1, that is, UT1 = 2.4 h (or $2.4 \times 60 = 144$ min or $2.4 \times 60 \times 60 = 8640$ s) elapsed from noon. Methods and tables for computing *JDN* from a date expressed in year, month, and day of our Gregorian calendar have

been indicated by several authors and can also be found in the Internet. Curtis [20] and Boulet [12] use the following method

$$JDN = 367\,y - \mathrm{INT}\{1.75[y + \mathrm{INT}(m/12 + 0.75)]\} + \mathrm{INT}(275m/9) + d + 1721013.5$$

where the function INT(x) means truncation (that is, retaining the integral part and dropping the fractional part of the argument x of the function INT), y is the full four-digit year, m is the month and d is the day of the Gregorian calendar date to be converted. Curtis points out that y, m, and d are integers such that:

$$1901 \leq y \leq 2099$$
$$1 \leq m \leq 12$$
$$1 \leq d \leq 31$$

This means that the formula given above holds only if the year of interest falls between 1901 and 2099.

These limitations do not affect another method, proposed by Jefferys [41], which method is based on the following rules. As is the case with Boulet's method, Jefferys expresses a Gregorian calendar date as y-m-d, where y is the year, m is the month number (January = 1, February = 2, etc.), and d is the day in the month. If the month is January or February, then 1 must be subtracted from the year to get a new value of y, and 12 must be added to the month to get a new value of m. Thus, January and February of a given year are considered as being, respectively, the 13th and the 14th month of the previous year.

Then, the computation follows the following scheme

$$a = \mathrm{INT}\,(y/100)$$
$$b = \mathrm{INT}(a/4)$$
$$c = 2 - a + b$$
$$e = \mathrm{INT}[365.25\,(y + 4716)]$$
$$f = \mathrm{INT}[30.6001\,(m + 1)]$$
$$JDN = c + d + e + f - 1524.5$$

This is the Julian Day Number for the beginning of the desired date at 0 h, UT1. To convert a Julian Day Number to a Gregorian calendar date, Jefferys [41] uses the following method, assuming that the *JDN* is for 0 h, UT1 (so that it ends in 0.5). The necessary calculations are shown below. Jefferys notes that his method does not give dates accurately in the proleptic Gregorian Calendar; in particular, the method fails if y is less than 400.

$z = JDN + 0.5$
$w = \text{INT}[(z - 1867216.25)/36524.25]$
$x = \text{INT}(w/4)$
$a = z + 1 + w - x$
$b = a + 1524$
$c = \text{INT}[(b - 122.1)/365.25]$
$d = \text{INT}(365.25c)$
$e = \text{INT}[(b - d)/30.6001]$
$f = \text{INT}(30.6001e)$
day of month $= b - d - f$
month $= e - 1$ or $e - 13$ (the number obtained must be less than or equal to 12)
year $= c - 4715$ (if the month is January or February) or $c - 4716$ (otherwise)

The U.S. Naval Observatory has an Internet-based calculator [71], which makes it possible to compute automatically the Julian date.

As an example, let us convert the following Gregorian calendar date and time

3 February 2011 at 13:15:18 UT1

into the corresponding Julian date, by using Boulet's method.

Since the condition $1901 \leq y \leq 2099$ is satisfied for $y = 2011$, then we can compute the Julian Day Number as follows

$$\begin{aligned} JDN &= 367 \times 2011 - \text{INT}\{1.75 \times [2011 + \text{INT}(2/12 + 0.75)]\} + \text{INT}(275 \times 2/9) \\ &\quad + 3 + 1721013.5 \\ &= 738037 - \text{INT}\{1.75 \times [2011 + \text{INT}(0.917)]\} + \text{INT}(61.111) + 3 + 1721015.5 \\ &= 738037 - \text{INT}[1.75 \times (2011 + 0)] + 61 + 3 + 1721013.5 \\ &= 738037 - \text{INT}(3519.25) + 1721077.5 = 2455595.5 \text{ days} \end{aligned}$$

The universal time, expressed in hours, is

$$\text{UT1} = 13 + \frac{15}{60} + \frac{18}{3600} = 13.255$$

Finally, the desired Julian date is

$$JD = JDN + \frac{\text{UT1}}{24} = 2455595.5 + \frac{13.255}{24} \approx 2455596.05229 \text{ days}$$

Now, let us apply Jefferys' method to the same example. Since, in this case, the month is February, then 1 must be subtracted from the year (2011) to get a new value of y (2011 − 1 = 2010), and 12 must be added to the month (2) to get a new value of m (2 + 12 = 14). Then the Julian Day Number is computed as follows

$$a = \text{INT}(y/100) = \text{INT}(2010/100) = \text{INT}(20.10) = 20$$
$$b = \text{INT}(a/4) = \text{INT}(20/4) = \text{INT}(20/4) = 5$$
$$c = 2 - a + b = 2 - 20 + 5 = -13$$
$$e = \text{INT}[365.25\,(y + 4716)] = \text{INT}[365.25 \times (2010 + 4716)] = \text{INT}(2456671.5) = 2456671$$
$$f = \text{INT}[30.6001\,(m + 1)] = \text{INT}[30.6001 \times (14 + 1)] = \text{INT}(459.0015) = 459$$
$$JDN = c + d + e + f - 1524.5 = -13 + 3 + 2456671 + 459 - 1524.5 = 2455595.5$$

which is the same value as that computed previously by means of Boulet's method.

Now, Jefferys' method is applied to the case JDN = 2455595.5 to obtain the corresponding Gregorian calendar date. Using the rules indicated above, we have

$$z = 2455595.5 + 0.5 = 2455596$$
$$w = \text{INT}[(2455596 - 1867216.25)/36524.25] = \text{INT}(16.109) = 16$$
$$x = \text{INT}(16/4) = \text{INT}(4) = 4$$
$$a = 2455596 + 1 + 16 - 4 = 2455609$$
$$b = 2455609 + 1524 = 2457133$$
$$c = \text{INT}[(2457133 - 122.1)/365.25] = \text{INT}(6726.929) = 6726$$
$$d = \text{INT}(365.25 \times 6726) = \text{INT}(2456671.5) = 2456671$$
$$e = \text{INT}[(2457133 - 2456671)/30.6001] = \text{INT}(15.098) = 15$$
$$f = \text{INT}(30.6001 \times 15) = \text{INT}(459.0015) = 459$$
$$\text{day of month} = 2457133 - 2456671 - 459 = 3$$
$$\text{month} = 15 - 13 = 2$$
$$\text{year} = 6726 - 4715 = 2011$$

As was expected, the resulting Gregorian calendar date is 3 February 2011.

The U.S. Naval Observatory also gives algorithms in FORTRAN programming language, which are due to Fliegel and van Flandern [25]. The current Julian epoch has been set to the 1st of January 2000 at noon. This epoch is denoted by J2000.0 and corresponds to the Julian Day Number 2451545.0.

A Julian year has 365.25 days and consequently a Julian century has 36525 days.

Let T_0 be the time, expressed in Julian centuries, elapsed between a given Julian day J_0 and J2000.0. Then the time T_0 results from

$$T_0 = \frac{J_0 - 2451545.0}{36525}$$

The time θ_{G0} (that is, the Greenwich sidereal time at 0 h UT1) can be expressed (in seconds) as a function of the dimensionless time T_0 by means of the following formula given by Aoki et al. [4]:

$$\theta_{G0} = 24110^{s}.54841 + 8640184^{s}.812866\,T_0 + 0^{s}.093104{T_0}^2 - 0^{s}.0000062{10T_0}^3$$

where the superscript s stands for seconds. If, as is often the case, we want to express θ_{G0} in degrees, then the coefficients appearing in the formula given above must be multiplied by 360/(24 × 3600). The same formula expressed in degrees is

$$\theta_{G0} = 100.460618375 + 36000.7700536\,T_0 + 0.000387933\,{T_0}^2 - 2.5875 \times 10^{-8}{T_0}^3$$

Since the value computed by means of the preceding formula may be outside the interval $0° < \theta_{G0} < 360°$, then the computed value must, if necessary, be brought into that interval by adding or subtracting an integral multiple of 360°.

This done, the Greenwich sidereal time (θ_G) relating to any other universal time than 0 h can be found as follows

$$\theta_G = \theta_{G0} + 360.985647366\frac{\mathrm{UT1}}{24}$$

where 360.985647366 = 1.00273790935 × 360 is the number of degrees covered by the Earth in its rotation about its axis in 24 h (solar time). If the (local) sidereal time θ is required for a site placed at an east longitude λ, then

$$\theta = \theta_G + \lambda$$

As shown above, in case of the value of θ being greater than 360°, it is necessary to bring it into the interval $0° < \theta < 360°$ by subtracting an integral number of 360° from the computed value.

As an example of application, let us compute the local sidereal time in degrees for Kiruna, Sweden (latitude φ = 65°.85N; longitude λ = 20°.2167E) on the 13th of February 2012 at 2:30:00 UT1. First, we use Jefferys' method to compute the Julian Day Number, that is, the Julian date relating to 13th of February 2012 at 00:00:00 UT1. Since, in this case, the month is February, then 1 must be subtracted from the year (2012) to get a new value of y (2012 − 1 = 2011), and 12 must be added to the month (2) to get a new value of m (2 + 12 = 14).

Then the Julian Day Number (J_0) is computed as follows

$$a = \text{INT}(y/100) = \text{INT}(2011/100) = \text{INT}(20.11) = 20$$
$$b = \text{INT}(a/4) = \text{INT}(20/4) = \text{INT}(5) = 5$$
$$c = 2 - a + b = 2 - 20 + 5 = -13$$
$$e = \text{INT}[365.25\,(y + 4716)] = \text{INT}[365.25 \times (2011 + 4716)] = \text{INT}(2457036.75) = 2457036$$
$$f = \text{INT}[30.6001\,(m + 1)] = \text{INT}[30.6001 \times (14 + 1)] = \text{INT}(459.0015) = 459$$
$$J_0 = c + d + e + f - 1524.5 = -13 + 13 + 2457036 + 459 - 1524.5 = 2455970.5$$

Second, the dimensionless time T_0 results from the expression shown above

$$T_0 = \frac{J_0 - 2451545.0}{36525} = \frac{2455970.5 - 2451545.0}{36525} = 0.12116358658453$$

Thirdly, θ_{G0} is computed by using Aoki's formula expressed in degrees

$$\theta_{G\,0} = 100.4606184 + 36000.77005\,T_0 + 0.000387933\,T_0^2 - 2.5875 \times 10^{-8} T_0^3$$

Substituting T_0 = 0.12116358658453 in the preceding formula, there results

$$\theta_{G0} = 100.4606184 + 36000.77005 \times 0.12116358658453 + 0.000387933 \times 0.12116358658453^2 - 2.5875 \times 10^{-8} \times 0.12116358658453^3 = 4462^\circ.44304315801945$$

Since this value is outside the interval [0, 360], we bring it into that interval by subtracting a multiple of 360° from it. To this end, we observe that

$$\text{INT}\left(\frac{4462.44304315801945}{360}\right) = 12$$

It follows that

$$\theta_{G0} = 4462.44304315801945 - 12 \times 360 = 142^\circ.44304315801945$$

The universal time given in this example is 2:30:00, that is,

$$\text{UT1} = 2.5\,\text{h}$$

Thus, the Greenwich sidereal time θ_G is computed by replacing θ_{G0} with the value obtained above and UT1 with 2.5 into the expression

$$\theta_G = \theta_{G0} + 360.985647366 \frac{\text{UT1}}{24}$$

This yields

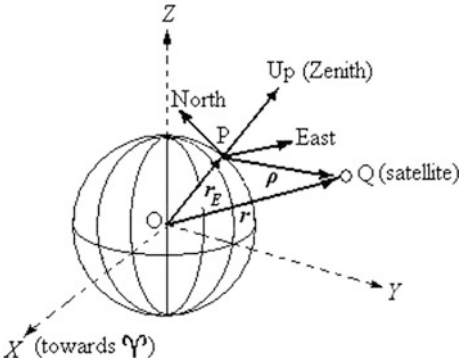
$$\begin{aligned} \theta_G &= 142^\circ.44304315801945 + 360.985647366 \times 2.5/24 \\ &= 180^\circ.04571475864445 \end{aligned}$$

Finally, the east longitude of Kiruna (λ = 20°.2167) must be added to θ_G to obtain the local sidereal time, as follows

$$\theta = \theta_G + \lambda = 180^\circ.04571475864445 + 20^\circ.2167 = 200^\circ.26241475864445$$

2.5 Orbital Elements from Angle and Range Measurements

The orbital elements of a space object Q revolving around the Earth and shown in the following figure are determined when its position ($\boldsymbol{r}$) and velocity ($\boldsymbol{r}'$) vectors, with respect to the geocentric-equatorial system *XYZ*, are known at a given time.



Section 2.3 has shown how to determine Q as a function of the line-of-sight vector $\boldsymbol{\rho}$ of Q with respect to the topocentric-horizon system *UVW* (located at the radar station P) and the position vector $\boldsymbol{r}_E$ of P with respect to the geocentric-equatorial system. The geocentric position vector $\boldsymbol{r} \equiv \text{OQ}$ of the space object Q results from

$$\boldsymbol{r} = \boldsymbol{r}_E + \boldsymbol{\rho} = \boldsymbol{r}_E + \rho \boldsymbol{u}_\rho$$

where $\boldsymbol{u}_\rho \equiv \boldsymbol{\rho}/\rho$ is the unit vector having the direction of $\boldsymbol{\rho}$. The velocity ($\boldsymbol{r}'$) and acceleration ($\boldsymbol{r}''$) vectors of Q, with respect to the geocentric-equatorial system *XYZ*, result from differentiating once and twice the preceding expression with respect to time. This yields

$$\boldsymbol{r}' = \boldsymbol{r}'_E + \rho' \boldsymbol{u}_\rho + \rho \boldsymbol{u}'_\rho$$
$$\boldsymbol{r}'' = \boldsymbol{r}''_E + \rho'' \boldsymbol{u}_\rho + \rho' \boldsymbol{u}'_\rho + \rho' \boldsymbol{u}'_\rho + \rho \boldsymbol{u}''_\rho = \boldsymbol{r}''_E + \rho'' \boldsymbol{u}_\rho + 2\rho' \boldsymbol{u}'_\rho + \rho \boldsymbol{u}''_\rho$$

Now, all vectors appearing in the preceding expressions must be expressed in the same reference system, that is, in the geocentric-equatorial reference system *XYZ*. To this end, it is to be noted that $\boldsymbol{r}_E \equiv$ OP is a vector which rotates with the Earth at a constant angular velocity $\boldsymbol{\omega}_E$ equal to

$$\boldsymbol{\omega}_E = \omega_E \boldsymbol{u}_Z$$

where ω_E is the magnitude of $\boldsymbol{\omega}_E$, and $\boldsymbol{u}_Z$ is the unit vector of the *Z*-axis. Since $\boldsymbol{r}_E$ rotates with the Earth, then its first ($\boldsymbol{r}'_E$) and second ($\boldsymbol{r}''_E$) time derivatives result from the following expressions

$$\boldsymbol{r}'_E = \boldsymbol{\omega}_E \times \boldsymbol{r}_E$$
$$\boldsymbol{r}''_E = \boldsymbol{\omega}_E \times (\boldsymbol{\omega}_E \times \boldsymbol{r}_E)$$

Let ℓ_X, ℓ_Y, and ℓ_Z be the direction cosines of PQ with respect to the topocentric-equatorial system $X_t Y_t Z_t$, having its origin in P and its axes parallel to, respectively, *X*, *Y*, and *Z*. The unit vector $\boldsymbol{u}_\rho = \boldsymbol{\rho}/\rho$ is expressible as follows

$$u_\rho = \ell_X \boldsymbol{u}_X + \ell_Y \boldsymbol{u}_Y + \ell_Z \boldsymbol{u}_Z$$

Since the unit vectors $\boldsymbol{u}_X$, $\boldsymbol{u}_Y$, and $\boldsymbol{u}_Z$ do not change with time, then the first and second time derivatives of the preceding expression are

$$\boldsymbol{u}'_\rho = \ell'_X \boldsymbol{u}_X + \ell'_Y \boldsymbol{u}_Y + \ell'_Z \boldsymbol{u}_Z$$
$$\boldsymbol{u}''_\rho = \ell''_X \boldsymbol{u}_X + \ell''_Y u_Y + \ell''_Z \boldsymbol{u}_Z$$

Now, since $\boldsymbol{u}_\rho = (\cos\delta \cos\alpha)\, \boldsymbol{u}_X + (\cos\delta \sin\alpha)\, \boldsymbol{u}_Y + (\sin\delta)\, \boldsymbol{u}_Z$, where δ and α are, respectively, the declination and right ascension of Q, then the direction cosines of PQ with respect to the topocentric-equatorial system are expressible in terms of δ and α as follows

$$\begin{aligned}\ell_X &= \cos\delta\,\cos\alpha\\ \ell_Y &= \cos\delta\,\sin\alpha\\ \ell_Z &= \sin\delta\end{aligned}$$

Differentiating the preceding expressions with respect to time yields

$$\begin{aligned}\ell'_X &= -\alpha'\sin\alpha\,\cos\delta - \delta'\cos\alpha\,\sin\delta\\ \ell'_Y &= \alpha'\cos\alpha\,\cos\delta - \delta'\sin\alpha\,\sin\delta\\ \ell'_Z &= \delta'\cos\delta\end{aligned}$$

$$\begin{aligned}\ell''_X &= -\alpha''\sin\alpha\,\cos\delta - \delta''\cos\alpha\,\sin\delta - (\alpha'^2+\delta'^2)\cos\alpha\,\cos\delta + 2\alpha'\delta'\,\sin\alpha\,\sin\delta\\ \ell''_Y &= \alpha''\,\cos\alpha\,\cos\delta - \delta''\,\sin\alpha\,\sin\delta - (\alpha'^2+\delta'^2)\sin\alpha\,\cos\delta - 2\alpha'\delta'\,\cos\alpha\,\sin\delta\\ \ell''_Z &= \delta''\,\cos\delta - \delta'^2\sin\delta\end{aligned}$$

Now, let ℓ_U, ℓ_V, and ℓ_W be the direction cosines of PQ with respect to the topocentric-horizon system *UVW*, having its origin in P and its axes pointing to, respectively, east, north, and zenith.

As shown in Sect. 2.3, the unit vector of the direction PQ is

$$\boldsymbol{u}_\rho = (\cos h\,\sin A)\boldsymbol{u}_U + (\cos h\,\cos A)\boldsymbol{u}_V + (\sin h)\,\boldsymbol{u}_W$$

where h and A are, respectively, the altitude angle and the azimuth angle of Q.

Consequently, the direction cosines of PQ with respect to the topocentric-horizon system *UVW* are

$$\begin{aligned}\ell_U &= \cos h\,\sin A\\ \ell_V &= \cos h\,\cos A\\ \ell_W &= \sin h\end{aligned}$$

The direction cosines ℓ_X, ℓ_Y, and ℓ_Z of PQ with respect to the geocentric-equatorial system *XYZ* can be obtained by means of the co-ordinate transformation shown in Sect. 2.3, that is,

$$\begin{bmatrix}\ell_X\\ \ell_Y\\ \ell_Z\end{bmatrix} = \begin{bmatrix}-\sin\theta & -\sin\varphi\,\cos\theta & \cos\varphi\,\cos\theta\\ \cos\theta & -\sin\varphi\,\sin\theta & \cos\varphi\,\sin\theta\\ 0 & \cos\varphi & \sin\varphi\end{bmatrix}\begin{bmatrix}\ell_U\\ \ell_V\\ \ell_W\end{bmatrix}.$$

replacing ℓ_X, ℓ_Y, ℓ_Z, ℓ_U, ℓ_V, and ℓ_W by their respective values yields

$$\begin{bmatrix} \cos\delta\,\cos\alpha \\ \cos\delta\,\sin\alpha \\ \sin\delta \end{bmatrix} = \begin{bmatrix} -\sin\theta & -\sin\varphi\,\cos\theta & \cos\varphi\,\cos\theta \\ \cos\theta & -\sin\varphi\,\sin\theta & \cos\varphi\,\sin\theta \\ 0 & \cos\varphi & \sin\varphi \end{bmatrix} \begin{bmatrix} \cos h\,\sin A \\ \cos h\,\cos A \\ \sin h \end{bmatrix}$$

Expanding the matrix product in the preceding equality leads to

$$\cos\delta\,\cos\alpha = -\sin\theta\,\cos h\,\sin A - \sin\varphi\,\cos\theta\,\cos h\,\cos A + \cos\varphi\,\cos\theta\,\sin h$$

hence

$$\cos\alpha = \frac{(\cos\varphi\,\sin h - \sin\varphi\,\cos h\,\cos A)\cos\theta - \sin\theta\,\cos h\,\sin A}{\cos\delta}$$

$$\cos\delta\,\sin\alpha = \cos\theta\,\cos h\,\sin A - \sin\varphi\,\sin\theta\,\cos h\,\cos A + \cos\varphi\,\sin\theta\,\sin h$$

hence

$$\sin\alpha = \frac{(\cos\varphi\,\sin h - \sin\varphi\,\cos h\,\cos A)\sin\theta + \cos\theta\,\cos h\,\sin A}{\cos\delta}$$

$$\sin\delta = \cos\varphi\,\cos h\,\cos A + \sin\varphi\,\sin h$$

The preceding expressions of cos α, sin α, and sin δ can be simplified by using the hour angle H, defined in Sect. 2.4, which is the angular distance between the object observed and the local meridian. In terms of the variables used above, the hour angle is expressible as follows

$$H = \theta - \alpha$$

The sine and cosine of the hour angle can be expressed as follows

$$\begin{aligned} \sin(\theta - \alpha) &= \sin\theta\,\cos\alpha - \cos\theta\,\sin\alpha \\ \cos(\theta - \alpha) &= \cos\theta\,\cos\alpha + \sin\theta\,\sin\alpha \end{aligned}$$

Thus, substituting the following expressions

$$\cos\alpha = \frac{(\cos\varphi\,\sin h - \sin\varphi\,\cos h\,\cos A)\cos\theta - \sin\theta\,\cos h\,\sin A}{\cos\delta}$$

$$\sin\alpha = \frac{(\cos\varphi\,\sin h - \sin\varphi\,\cos h\,\cos A)\sin\theta + \cos\theta\,\cos h\,\sin A}{\cos\delta}$$

into

$$\sin(\theta - \alpha) \;=\; \sin\theta\,\cos\alpha - \;\cos\theta\,\sin\alpha$$

yields

$$\begin{aligned}\sin(\theta-\alpha) &= \sin\theta\left[\frac{(\cos\varphi\,\sin h - \sin\varphi\,\cos h\,\cos A)\cos\theta - \sin\theta\,\cos h\,\sin A}{\cos\delta}\right]\\ &\quad - \cos\theta\left[\frac{(\cos\varphi\,\sin h - \sin\varphi\,\cos h\,\cos A)\sin\theta + \cos\theta\,\cos h\,\sin A}{\cos\delta}\right]\\ &= -\frac{\cos h\,\sin A}{\cos\delta}\end{aligned}$$

Likewise, substituting

$$\begin{aligned}\cos\alpha &= \frac{(\cos\varphi\,\sin h - \sin\varphi\,\cos h\,\cos A)\cos\theta - \sin\theta\,\cos h\,\sin A}{\cos\delta}\\ \sin\alpha &= \frac{(\cos\varphi\,\sin h - \sin\varphi\,\cos h\,\cos A)\sin\theta + \cos\theta\,\cos h\,\sin A}{\cos\delta}\end{aligned}$$

into

$$\cos(\theta - \alpha) \;=\; \cos\theta\,\cos\alpha + \;\sin\theta\,\sin\alpha$$

yields

$$\begin{aligned}\cos(\theta-\alpha) &= \cos\theta\left[\frac{(\cos\varphi\,\sin h - \sin\varphi\,\cos h\,\cos A)\cos\theta - \sin\theta\,\cos h\,\sin A}{\cos\delta}\right]\\ &\quad + \sin\theta\left[\frac{(\cos\varphi\,\sin h - \sin\varphi\,\cos h\,\cos A)\sin\theta + \cos\theta\,\cos h\,\sin A}{\cos\delta}\right]\\ &= \frac{\cos\varphi\,\sin h - \sin\varphi\,\cos h\,\cos A}{\cos\delta}\end{aligned}$$

This expression makes it possible to compute the hour angle, as follows

$$H = \arccos\left[\frac{\cos\varphi\,\sin h - \sin\varphi\,\cos h\,\cos A}{\cos\delta}\right]$$

which holds if $\sin H > 0$. Otherwise, if $\sin H < 0$, then

$$H = 2\pi - \arccos\left[\frac{\cos\varphi\,\sin h - \sin\varphi\,\cos h\,\cos A}{\cos\delta}\right]$$

Since the altitude (h) and declination (δ) angles range from $-\pi/2$ to $\pi/2$ radians, then neither cos h nor cos δ can be negative. Consequently, the expression

$$\sin H = -\frac{\cos h \ \sin A}{\cos \delta}$$

shows that sin H is negative when sin A is positive, which happens when A ranges from 0 to π radians. In summary, let the topocentric angles azimuth (A) and altitude (h) of a space object be known at a given time. Let also the latitude (φ) of the tracking station and the sidereal time (θ) be known at the same time. Then, the topocentric declination (δ) results from the equation derived above

$$\sin \delta = \cos \varphi \ \cos h \ \cos A + \ \sin \varphi \ \sin h$$

which, solved for δ, yields

$$\delta = \arcsin (\cos \varphi \ \cos h \ \cos A + \ \sin \varphi \ \sin h)$$

This done, the hour angle (H) results from

$$H = \begin{cases} 2\pi - \arccos [(\cos \varphi \ \sin h - \sin \varphi \ \cos h \ \cos A)/ \cos \delta] & (0 < A < \pi) \\ \arccos[(\cos \varphi \ \sin h - \sin \varphi \ \cos h \ \cos A)/ \cos \delta] & (\pi \leq A \leq 2\pi) \end{cases}$$

and the right ascension (α) results from

$$\alpha = \theta - H$$

If the azimuth and altitude angles are known as functions of time, then the right ascension and declination angles can be computed as functions of time by means of the expressions given above. Then, these functions are differentiated with respect to time, and the results are introduced into the following expressions

$$\begin{aligned} \boldsymbol{u}_\rho &= \ell_X \boldsymbol{u}_X + \ell_Y \boldsymbol{u}_Y + \ell_Z \boldsymbol{u}_Z \\ \boldsymbol{u}'_\rho &= \ell'_X \boldsymbol{u}_X + \ell'_Y \boldsymbol{u}_Y + \ell'_Z \boldsymbol{u}_Z \\ \boldsymbol{u}''_\rho &= \ell''_X \boldsymbol{u}_X + \ell''_Y \boldsymbol{u}_Y + \ell''_Z \boldsymbol{u}_Z \end{aligned}$$

where

$$\begin{aligned} \ell_X &= \cos \delta \ \cos \alpha \\ \ell_Y &= \cos \delta \ \sin \alpha \\ \ell_Z &= \sin \delta \end{aligned}$$

$$\ell'_X = -\alpha' \sin\alpha \cos\delta - \delta' \cos\alpha \sin\delta$$
$$\ell'_Y = \alpha' \cos\alpha \cos\delta - \delta' \sin\alpha \sin\delta$$
$$\ell'_Z = \delta' \cos\delta$$

$$\ell''_X = -\alpha'' \sin\alpha \cos\delta - \delta'' \cos\alpha \sin\delta - (\alpha'^2 + \delta'^2) \cos\alpha \cos\delta + 2\alpha' \delta' \sin\alpha \sin\delta$$
$$\ell''_Y = \alpha'' \cos\alpha \cos\delta - \delta'' \sin\alpha \sin\delta - (\alpha'^2 + \delta'^2) \sin\alpha \cos\delta - 2\alpha' \delta' \cos\alpha \sin\delta$$
$$\ell''_Z = \delta'' \cos\delta - \delta'^2 \sin\delta$$

This makes it possible to compute the unit vector $\boldsymbol{u}_\rho$ and its time derivatives $\boldsymbol{u}'_\rho$ and $\boldsymbol{u}''_\rho$. In order to compute α' and δ' from A' and h', the expression

$$\sin\delta = \cos\varphi \cos h \cos A + \sin\varphi \sin h$$

is differentiated with respect to time, taking into account that φ = constant.

This yields

$$\delta' \cos\delta = -h' \cos\varphi \sin h \cos A - A' \cos\varphi \cos h \sin A + h' \sin\varphi \cos h$$

which in turn, solved for δ', yields

$$\delta' = \frac{-A' \cos\varphi \cos h \sin A + h'(\sin\varphi \cos h - \cos\varphi \sin h \cos A)}{\cos\delta}$$

Now, the expression $\sin H = -(\cos h \sin A)/\cos\delta$ is differentiated with respect to time. This yields

$$H' \cos H = \frac{h' \sin h \sin A}{\cos\delta} - \frac{A' \cos h \cos A}{\cos\delta} - \frac{\delta' \cos h \sin A \sin\delta}{\cos^2\delta}$$
$$= -\frac{(A' \cos h \cos A - h' \sin h \sin A)\cos\delta + \delta' \cos h \sin A \sin\delta}{\cos^2\delta}$$

Since

$$\cos H = \frac{\cos\varphi \sin h - \sin\varphi \cos h \cos A}{\cos\delta}$$

then substituting this expression of $\cos H$ into the expression of $H' \cos H$ yields

$$H' \cos H = H' \left(\frac{\cos\varphi \sin h - \sin\varphi \cos h \cos A}{\cos\delta} \right)$$
$$= -\frac{(A' \cos h \cos A - h' \sin h \sin A)\cos\delta + \delta' \cos h \sin A \sin\delta}{\cos^2\delta}$$

The preceding equation, solved for H', yields

$$H' = -\frac{A' \cos h \cos A - h' \sin h \sin A + \delta' \cos h \sin A \tan \delta}{\cos \varphi \sin h - \sin \varphi \cos h \cos A}$$

Now, since $H = \theta - \alpha$, then

$$H' = \theta' - \alpha' = \omega_E - \alpha'$$

where ω_E is the angular velocity of the Earth about its axis. It follows that

$$\alpha' = \omega_E - H'$$

that is,

$$\alpha' = \omega_E + \frac{A' \cos h \cos A - h' \sin h \sin A + \delta' \cos h \sin A \tan \delta}{\cos \varphi \sin h - \sin \varphi \cos h \cos A}$$

As an example of application, we compute the classical orbital elements of a space object, which the Catalina station of the University of Arizona, Tucson (latitude φ = 32°.417 N, longitude λ = −110°.732 E, and height H = 2509 m on the mean sea level) detected on the 5th of April 2004 at 6:00:00 UT1, obtaining the following data:

Slant range $\rho = 27148$ km
Azimuth $A = 128°$
Altitude $h = 41°$
Range rate $\rho' = 2.267$ km/s
Azimuth rate $A' = -1.86 \times 10^{-5}$ rad/s
Altitude rate $h' = 3.42 \times 10^{-5}$ rad/s

The three components of the position vector ($\boldsymbol{r}_E$) of the tracking station, with respect to the geocentric-equatorial system *XYZ*, are

$$\boldsymbol{r}_E = (x_H \cos \theta)\boldsymbol{u}_X + (x_H \sin \theta)\boldsymbol{u}_Y + z_H \boldsymbol{u}_Z$$

where $\boldsymbol{u}_X$, $\boldsymbol{u}_Y$, and $\boldsymbol{u}_Z$ are the unit vectors along, respectively, X, Y, and Z, x_H and z_H are the two co-ordinates of the point P_H representing the position of the tracking station. Let us compute the local sidereal time θ by means of the date

5 April 2004 at 6:00:00 UT1

The Julian Day Number (J_0) is computed as follows

$$
\begin{aligned}
a &= \text{INT}(y/100) = \text{INT}(2004/100) = \text{INT}(20.04) = 20\\
b &= \text{INT}(a/4) = \text{INT}(20/4) = \text{INT}(5) = 5\\
c &= 2 - a + b = 2 - 20 + 5 = -13\\
e &= \text{INT}[365.25\,(y + 4716)] = \text{INT}[365.25 \times (2004 + 4716)] = 2454480\\
f &= \text{INT}[30.6001\,(m + 1)] = \text{INT}[30.6001 \times (4 + 1)] = \text{INT}(153.0005) = 153\\
J_0 &= c + d + e + f - 1524.5 = -13 + 5 + 2454480 + 153 - 1524.5 = 2453100.5
\end{aligned}
$$

The dimensionless time T_0 results from the following expression

$$T_0 = \frac{J_0 - 2451545.0}{36525} = \frac{2453100.5 - 2451545.0}{36525} = 0.042587$$

Now, θ_{G0} is computed by using Aoki's formula

$$
\begin{aligned}
\theta_{G0} &= 100.4606184 + 36000.77005\,T_0 + 0.000387933\,T_0^2 - 2.5875 \times 10^{-8} T_0^3\\
&= 100.4606184 + 36000.77005 \times 0.042587 + 0.000387933 \times 0.042587^2\\
&\quad - 2.5875 \times 10^{-8} \times 0.042587^3 = 1633^\circ.625413
\end{aligned}
$$

Since this value is outside the interval [0, 360], we bring it into that interval by subtracting a multiple of 360° from it. To this end, we observe that

$$\text{INT}\left(\frac{1633.625}{360}\right) = 4$$

It follows that

$$\theta_{G0} = 1633.625413 - 4 \times 360 = 193^\circ.6254128$$

The universal time given in this example is 6:00:00, that is, UT1 = 6.0 h.

Thus, the Greenwich sidereal time θ_G is computed by replacing θ_{G0} with 193.6254128 and UT1 with 6.0 into the following expression

$$\theta_G = \theta_{G0} + 360.985647366\left(\frac{\text{UT1}}{24}\right)$$

This yields

$$\theta_G = 193.6254128 + 360.985647366 \times \left(\frac{6}{24}\right) = 283^\circ.8718246$$

The east longitude of Catalina ($\lambda = -110^\circ.732$ east) must be added to θ_G to obtain the local sidereal time, as follows

$$\theta = \theta_G + \lambda = 283^\circ.8718246 - 110^\circ.732 \approx 173^\circ.14$$

Now, x_H and z_H are computed as follows

$$x_H = \left\{ \frac{a_E}{\left[1 - (2f - f^2)\sin^2\varphi\right]^{\frac{1}{2}}} + H \right\} \cos\varphi$$

$$z_H = \left\{ \frac{a_E(1-f)^2}{\left[1 - (2f - f^2)\sin^2\varphi\right]^{\frac{1}{2}}} + H \right\} \sin\varphi$$

where $a_E = 6378.137$ km and $f = 0.0033528$ are, respectively, the equatorial radius and the flattening of the Earth (represented as an ellipsoid), and φ is the geodetic latitude (that is, the angle between the equatorial plane and the local vertical) measured at the tracking station. Taking $\varphi = 32^\circ.417$, x_H and z_H result from

$$x_H = \left\{6378.137/\left[1 - \left(2 \times 0.0033528 - 0.0033528^2\right) \times \sin^2 32^\circ.417\right]^{1/2} + 2.509\right\} \times \cos 32^\circ.417 = 5391.552\,\text{km}$$

$$z_H = \left\{6378.137 \times (1 - 0.0033528)^2/\left[1 - \left(2 \times 0.0033528 - 0.0033528^2\right) \times \sin^2 32^\circ.417\right]^{1/2} + 2.509\right\} \times \sin 32^\circ.417 = 3400.921\,\text{km}$$

Hence, the position vector $\boldsymbol{r}_E$ of the tracking station, with respect to the geocentric-equatorial system XYZ, is

$$\begin{aligned}\boldsymbol{r}_E &= (x_H \cos\theta)\boldsymbol{u}_X + (x_H \sin\theta)\boldsymbol{u}_Y + z_H\boldsymbol{u}_Z \\ &= (5391.552 \times \cos 173^\circ.14)\boldsymbol{u}_X + (5391.552 \times \sin 173^\circ.14)\,\boldsymbol{u}_Y + 3400.921\boldsymbol{u}_Z \\ &= -5352.954\,\boldsymbol{u}_X + 643.987\,\boldsymbol{u}_Y + 3400.921\,\boldsymbol{u}_Z\,(\text{km})\end{aligned}$$

Let us compute now the declination δ of the observed object with respect to the topocentric-equatorial system $X_tY_tZ_t$, by means of the following expression

$$\begin{aligned}\delta &= \arcsin(\cos\varphi\,\cos h\,\cos A + \sin\varphi\,\sin h) \\ &= \arcsin(\cos 32^\circ.417\,\cos 41^\circ\cos 128^\circ + \sin 32^\circ.417\,\sin 41^\circ) = -2^\circ.323\end{aligned}$$

The given azimuth ($A = 128^\circ$) lies between 0° and 180°. Consequently, the hour angle H results from the following expression

$$H = 360^\circ - \arccos\left[\frac{\cos\varphi\ \sin h - \sin\varphi\ \cos h\ \cos A}{\cos\delta}\right]$$
$$= 360^\circ - \arccos\left[\frac{\cos 32^\circ.417 \sin 41^\circ - \sin 32^\circ.417 \cos 41^\circ \cos 128^\circ}{\cos(-2^\circ.323)}\right]$$
$$= 323^\circ.472$$

Thus, the right ascension of the observed object results from

$$\alpha = \theta - H = 173^\circ.14 - 323^\circ.472 = -150^\circ.332$$

Now, we compute the unit vector $\boldsymbol{u}_\rho$ of the line joining the tracking station with the object, with respect to the topocentric-equatorial system $X_tY_tZ_t$, by means of

$$\boldsymbol{u}_\rho = \ell_X \boldsymbol{u}_X + \ell_Y \boldsymbol{u}_Y + \ell_Z \boldsymbol{u}_Z$$

where

$$\ell_X = \cos\delta\ \cos\alpha$$
$$\ell_Y = \cos\delta\ \sin\alpha$$
$$\ell_Z = \sin\delta$$

This yields

$$\boldsymbol{u}_\rho = [\cos(-2^\circ.323)\ \cos(-150^\circ.332)]\boldsymbol{u}_X + [\cos(-2^\circ.323)\ \sin(-150^\circ.332)]\boldsymbol{u}_Y$$
$$+ [\sin(-2^\circ.323)]\boldsymbol{u}_Z = -0.8682\,\boldsymbol{u}_X - 0.4946\,\boldsymbol{u}_Y - 0.0405\,\boldsymbol{u}_Z$$

The position vector $\boldsymbol{r}$ (in km) of the observed object results from

$$\boldsymbol{r} = \boldsymbol{r}_E + \rho\boldsymbol{u}_\rho = -5352.954\,\boldsymbol{u}_X + 643.987\,\boldsymbol{u}_Y + 3400.921\,\boldsymbol{u}_Z$$
$$+ 27148 \times (-0.8682\,\boldsymbol{u}_X - 0.4946\,\boldsymbol{u}_Y - 0.0405\,\boldsymbol{u}_Z)$$
$$= -28922.848\,\boldsymbol{u}_X - 12783.414\,\boldsymbol{u}_Y + 2301.427\,\boldsymbol{u}_Z$$

The velocity vector $\boldsymbol{r}'_E$ of the tracking station, with respect to the celestial geocentric-equatorial system XYZ, results from

$$\boldsymbol{r}'_E = \boldsymbol{\omega}_E \times \boldsymbol{r}_E$$

where $\boldsymbol{r}_E = -5352.954\,\boldsymbol{u}_X + 643.987\,\boldsymbol{u}_Y + 3400.921\,\boldsymbol{u}_Z$ is the position vector of the tracking station and $\boldsymbol{\omega}_E = \omega_E\,\boldsymbol{u}_Z$ is the angular velocity of the Earth around its axis with respect to XYZ. Remembering that $\omega_E = 7.292 \times 10^{-5}$ rad/s, there results

$$\boldsymbol{r}'_E = \boldsymbol{\omega}_E \times \boldsymbol{r}_E = \begin{bmatrix} \boldsymbol{u}_X & \boldsymbol{u}_Y & \boldsymbol{u}_Z \\ 0 & 0 & 7.292 \times 10^{-5} \\ -5352.954 & 643.987 & 3400.921 \end{bmatrix}$$
$$= \left(-643.987 \times 7.292 \times 10^{-5}\right)\boldsymbol{u}_X + \left(-5352.954 \times 7.292 \times 10^{-5}\right)\boldsymbol{u}_Y$$
$$= -0.047\,\boldsymbol{u}_X - 0.390\,\boldsymbol{u}_Y\ (\mathrm{km/s})$$

Now, the declination rate δ' of the observed object results from

$$\delta' = [-A' \cos\varphi \cos h \sin A + h'(\sin\varphi \cos h - \cos\varphi \sin h \cos A)]/\cos\delta$$
$$= \left[1.86 \times 10^{-5} \cos 32^\circ.417 \cos 41^\circ \sin 128^\circ + 3.42 \times 10^{-5} \times (\sin 32^\circ.417 \cos 41^\circ\right.$$
$$\left. - \cos 32^\circ.417 \sin 41^\circ \cos 128^\circ)\right]/\cos(-2^\circ.323) = 3.486 \times 10^{-5}\ (\mathrm{rad/s})$$

The right ascension rate α' of the observed object results from

$$\alpha' = \omega_E + (A' \cos h \cos A - h' \sin h \sin A + \delta' \cos h \sin A \tan\delta)/(\cos\varphi \sin h$$
$$- \sin\varphi \cos h \cos A) = 7.292 \times 10^{-5} + [-1.86 \times 10^{-5} \cos 41^\circ \cos 128^\circ$$
$$- 3.42 \times 10^{-5} \sin 41^\circ \sin 128^\circ + 3.486 \times 10^{-5} \cos 41^\circ \sin 128^\circ \tan(-2^\circ.323)]$$
$$/[\cos 32^\circ.417 \sin 41^\circ - \sin 32^\circ.417 \cos 41^\circ \cos 128^\circ] = 6.062 \times 10^{-5}\ (\mathrm{rad/s})$$

The direction cosine rate vector $\boldsymbol{u}'_\rho$ results from

$$\boldsymbol{u}'_\rho = \ell'_X \boldsymbol{u}_X + \ell'_Y \boldsymbol{u}_Y + \ell'_Z \boldsymbol{u}_Z$$

where

$$\ell'_X = -\alpha' \sin\alpha \cos\delta - \delta' \cos\alpha \sin\delta$$
$$\ell'_Y = \alpha' \cos\alpha \cos\delta - \delta' \sin\alpha \sin\delta$$
$$\ell'_Z = \delta' \cos\delta$$

Hence,

$$\ell'_X = -6.062 \times 10^{-5} \sin(-150^\circ.332) \cos(-2^\circ.323) - 3.486 \times 10^{-5}$$
$$\times \cos(-150^\circ.332) \sin(-2^\circ.323) = 2.875 \times 10^{-5}$$
$$\ell'_Y = 6.062 \times 10^{-5} \cos(-150^\circ.332) \cos(-2^\circ.323) - 3.486 \times 10^{-5}$$
$$\times \sin(-150^\circ.332) \sin(-2^\circ.323) = -5.333 \times 10^{-5}$$
$$\ell'_Z = 3.486 \times 10^{-5} \cos(-2^\circ.323) = 3.483 \times 10^{-5}$$

which, substituted into $\boldsymbol{u}'_\rho = \ell'_X\,\boldsymbol{u}_X + \ell'_Y\,\boldsymbol{u}_Y + \ell'_Z\,\boldsymbol{u}_Z$, yield

$$\boldsymbol{u}'_\rho = 2.875 \times 10^{-5}\boldsymbol{u}_X - 5.333 \times 10^{-5}\boldsymbol{u}_Y + 3.483 \times 10^{-5}\boldsymbol{u}_Z\,(\mathrm{rad/s})$$

The velocity vector $\boldsymbol{r}'$ of the observed object, with respect to the geocentric-equatorial reference system *XYZ*, results from

$$\boldsymbol{r}' = \boldsymbol{r}'_E + \rho'\boldsymbol{u}_\rho + \rho\boldsymbol{u}'_\rho$$

Hence, the velocity vector $\boldsymbol{r}'$ (in km/s) of the observed object is

$$\begin{aligned}\boldsymbol{r}' = {} & -0.047\,\boldsymbol{u}_X - 0.390\,\boldsymbol{u}_Y + 2.267 \times (-0.8682\,\boldsymbol{u}_X - 0.4946\,\boldsymbol{u}_Y - 0.0405\,\boldsymbol{u}_Z) \\ & + 27148 \times (2.875 \times 10^{-5}\boldsymbol{u}_X - 5.333 \times 10^{-5}\boldsymbol{u}_Y + 3.483 \times 10^{-5}\boldsymbol{u}_Z) \\ = {} & -1.235\,\boldsymbol{u}_X - 2.959\,\boldsymbol{u}_Y + 0.854\,\boldsymbol{u}_Z\end{aligned}$$

In summary, the position ($\boldsymbol{r}$) and velocity ($\boldsymbol{r}'$) vectors of the space object at the epoch of observation, with respect to the geocentric-equatorial system *XYZ*, are

$$\begin{aligned}\boldsymbol{r} &= -28922.848\,\boldsymbol{u}_X - 12783.414\,\boldsymbol{u}_Y + 2301.427\,\boldsymbol{u}_Z\,(\mathrm{km}) \\ \boldsymbol{r}' &= -1.235\,\boldsymbol{u}_X - 2.959\,\boldsymbol{u}_Y + 0.854\,\boldsymbol{u}_Z\,(\mathrm{km/s})\end{aligned}$$

Now that the position and velocity vectors are known, the corresponding orbital elements can be computed, as will be shown below.

The radius vector (in km) and the square of the velocity (in $\mathrm{km}^2/\mathrm{s}^2$) at epoch are

$$\begin{aligned}r_0 &= \left[(-28922.848)^2 + (-12783.414)^2 + 2301.427^2\right]^{\frac{1}{2}} = 31705.573 \\ v_0^2 &= (-1.235)^2 + (-2.959)^2 + 0.854^2 = 11.01\end{aligned}$$

The vis-viva integral

$$\frac{v^2}{\mu} = \frac{2}{r} - \frac{1}{a}$$

makes it possible to compute the major semi-axis a (in km) as follows

$$a = \frac{1}{\dfrac{2}{r_0} - \dfrac{v_0^2}{\mu}} = \frac{1}{\dfrac{2}{31705.573} - \dfrac{11.01}{398600.4}} = 28201.744$$

The angular momentum vector per unit mass (in km^2/s) is

$$\boldsymbol{h} = \boldsymbol{r} \times \boldsymbol{r}' = \begin{bmatrix} \boldsymbol{u}_X & \boldsymbol{u}_Y & \boldsymbol{u}_Z \\ -28922.848 & -12783.414 & 2301.427 \\ -1.235 & -2.959 & 0.854 \end{bmatrix}$$
$$= -4107.113\,\boldsymbol{u}_X + 21857.85\,\boldsymbol{u}_Y + 69795.191\,\boldsymbol{u}_Z$$

and its magnitude (in km^2/s) is

$$h = (\boldsymbol{h} \cdot \boldsymbol{h})^{\frac{1}{2}} = \left[(-4107.113)^2 + 21857.85^2 + 69795.191^2\right]^{\frac{1}{2}} = 73253.005$$

The semi-latus rectum p (in km) results from

$$p = \frac{h^2}{\mu} = \frac{73253.005^2}{398600.4} = 13462.084$$

The eccentricity e results from $a = p(1 - e^2)$, which, solved for e, yields

$$e = \left(1 - \frac{p}{a}\right)^{\frac{1}{2}} = \left(1 - \frac{13462.084}{28201.744}\right)^{\frac{1}{2}} = 0.72294582$$

The inclination angle i (in degrees) of the orbit with respect to the equatorial plane is

$$i = \arccos\left(\frac{h_Z}{h}\right) = \arccos\left(\frac{69795.191}{73253.005}\right) = 17^\circ.674578$$

The eccentricity vector $\boldsymbol{e}$ results from

$$\boldsymbol{e} = \frac{\boldsymbol{r}' \times \boldsymbol{h}}{\mu} - \frac{\boldsymbol{r}}{r_0}$$

$$(\boldsymbol{r}' \times \boldsymbol{h})/\mu = [(-2.959 \times 69795.191 - 21857.85 \times 0.854)\boldsymbol{u}_X + (1.235 \times 69795.191 - 4107.113 \times 0.854)\,\boldsymbol{u}_Y + (-1.235 \times 21857.85 - 4107.113 \times 2.959)\boldsymbol{u}_Z] /398600.4 = -0.5649521\boldsymbol{u}_X + 0.2074494\boldsymbol{u}_Y - 0.0982119\,\boldsymbol{u}_Z$$

$$\boldsymbol{r}/r_0 = (-28922.848\,\boldsymbol{u}_X - 12783.414\,\boldsymbol{u}_Y + 2301.427\,\boldsymbol{u}_Z)/31705.573 = -0.9122323\,\boldsymbol{u}_X - 0.4031914\,\boldsymbol{u}_Y + 0.0725875\,\boldsymbol{u}_Z$$

$$\begin{aligned}\boldsymbol{e} &= (\boldsymbol{r}' \times \boldsymbol{h})/\mu - \boldsymbol{r}/r_0 = (-0.5649521 + 0.9122323)\,\boldsymbol{u}_X \\ &\quad + (0.2074494 + 0.4031914)\,\boldsymbol{u}_Y + (-0.0982119 - 0.0725875)\,\boldsymbol{u}_Z \\ &= 0.3427113\,\boldsymbol{u}_X + 0.6106408\,\boldsymbol{u}_Y - 0.1707994\,\boldsymbol{u}_Z\end{aligned}$$

The node vector $\boldsymbol{n}$ is defined by $\boldsymbol{n} \equiv \boldsymbol{u}_Z \times \boldsymbol{h}$. In the present case, there results

$$\boldsymbol{n} = n_X\boldsymbol{u}_X + n_Y\boldsymbol{u}_Y = -h_Y\boldsymbol{u}_X + h_X\boldsymbol{u}_Y = -21857.85\,\boldsymbol{u}_X - 4107.113\,\boldsymbol{u}_Y$$

The magnitude n of the node vector is

$$n = (\boldsymbol{n}\cdot\boldsymbol{n})^{\frac{1}{2}} = \left(21857.85^2 + 4107.113^2\right)^{\frac{1}{2}} = 22240.368$$

The right ascension Ω of the ascending node (in degrees) is

$$\Omega = \arccos\left(\frac{n_X}{n}\right) \quad (n_Y \geq 0)$$

$$\Omega = 360^\circ - \arccos\left(\frac{n_X}{n}\right) \quad (n_Y < 0)$$

In the present case, $n_Y = -4107.113 < 0$; thus,

$$\Omega = 360^\circ - \arccos\left(\frac{-21857.85}{22240.368}\right) = 190^\circ.64185$$

The argument of perigee ω results from

$$\omega = \arccos\left(\frac{\boldsymbol{n}\cdot\boldsymbol{e}}{ne}\right) \quad (e_Z \geq 0)$$

$$\omega = 360^\circ - \arccos\left(\frac{\boldsymbol{n}\cdot\boldsymbol{e}}{ne}\right) \quad (e_Z < 0)$$

In the present case, $e_Z = -0.1707994 < 0$; hence,

$$\begin{aligned}\omega &= 360^\circ - \arccos\left(\frac{-21857.85 \times 0.3427113 - 4107.113 \times 0.6106408}{22240.368 \times 0.72294582}\right) \\ &= 231^\circ.54665\end{aligned}$$

The true anomaly at epoch ϕ_0 results from

$$\phi_0 = \arccos\left(\frac{\boldsymbol{e}\cdot\boldsymbol{r}}{er_0}\right) \quad (\boldsymbol{r}\cdot\boldsymbol{r}' \geq 0)$$

$$\phi_0 = 360^\circ - \arccos\left(\frac{\boldsymbol{e}\cdot\boldsymbol{r}}{er_0}\right) \quad (\boldsymbol{r}\cdot\boldsymbol{r}' < 0)$$

In the present case, there results

$$\begin{aligned}\boldsymbol{r}\cdot\boldsymbol{r}' &= -28922.848\times(-1.235) - 12783.414\times(-2.959) + 2301.427\times 0.854\\ &= 75511.258 > 0\\ \boldsymbol{e}\cdot\boldsymbol{r} &= 0.3427113\times(-28922.848) + 0.6106408\times(-12783.414) - 0.1707994\\ &\quad\times 2301.427 = -18111.34334\\ er_0 &= 0.72294582\times 31705.573 = 22921.41147\end{aligned}$$

Thus, the true anomaly (in degrees) at epoch is

$$\phi_0 = \arccos\left(\frac{-18111.34334}{22921.41147}\right) = 142^\circ.19949$$

In summary, the object detected by the given radar station at the given time revolves about the Earth in an elliptic orbit having the following elements

$$\begin{array}{ll} a = 28201.744\,\text{km} & \Omega = 190^\circ.64185\\ e = 0.72294582 & \omega = 231^\circ.54665\\ i = 17^\circ.674578 & \phi_0 = 142^\circ.19949\end{array}$$

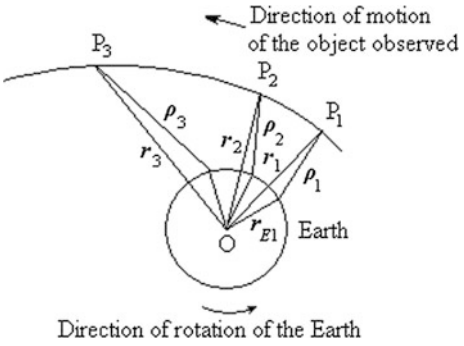
2.6 Orbital Elements from Three Measurements of Angles (Method of Gauss)

As has been shown in the preceding paragraphs, a set of six independent quantities is required to determine the motion of a celestial body. These six quantities may be the three components of the position vector and the three components of the velocity vector, or the six classical elements. A radar station like that described in Sect. 2.5 provides range and range-rate measurements. Thus, three range measurements (range, declination, and azimuth) plus three range-rate measurements (range rate, declination rate, and azimuth rate) provide the six required quantities. This implies the availability of a Doppler radar.

By contrast, the present paragraph and the following one will show how to determine the orbit of a celestial body when only angular observations are possible. This happens when only a telescope can be used as a means of observation. When only angular measurements are possible (e.g. the declination and the right ascension of the body observed), then three distinct observations are required, each of which provides the declination and the right ascension of the body.

Following Curtis [20] and Boulet [12], let t_1, t_2, and t_3 be the three distinct times at which the three single angular observations are performed. Let P_1, P_2, and P_3 be the three positions of the observed body at, respectively, t_1, t_2, and t_3, as shown in the following figure. Let $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$ be the three position vectors of the observed

body with respect to the geocentric-equatorial reference system XYZ. Let $\boldsymbol{r}_{E1}$, $\boldsymbol{r}_{E2}$, and $\boldsymbol{r}_{E3}$ be the three position vectors of the point of observation with respect to the same system of reference.



The relation between $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$ and $\boldsymbol{r}_{E1}$, $\boldsymbol{r}_{E2}$, and $\boldsymbol{r}_{E3}$ is

$$\begin{aligned}
\boldsymbol{r}_1 &= \boldsymbol{r}_{E1} + \boldsymbol{\rho}_1 = \boldsymbol{r}_{E1} + \rho_1(\boldsymbol{\rho}_1/\rho_1) &&= \boldsymbol{r}_{E1} + \rho_1\boldsymbol{u}_1 \\
\boldsymbol{r}_2 &= \boldsymbol{r}_{E2} + \boldsymbol{\rho}_2 = \boldsymbol{r}_{E2} + \rho_2(\boldsymbol{\rho}_2/\rho_2) &&= \boldsymbol{r}_{E2} + \rho_2\boldsymbol{u}_2 \\
\boldsymbol{r}_3 &= \boldsymbol{r}_{E3} + \boldsymbol{\rho}_3 = \boldsymbol{r}_{E3} + \rho_3(\boldsymbol{\rho}_3/\rho_3) &&= \boldsymbol{r}_{E3} + \rho_3\boldsymbol{u}_3
\end{aligned}$$

where $\boldsymbol{\rho}_1$, $\boldsymbol{\rho}_2$, and $\boldsymbol{\rho}_3$ are the three position vectors of the observed body, with respect to the topocentric-equatorial reference system $X_tY_tZ_t$, at, respectively, t_1, t_2, and t_3. Likewise, ρ_1, ρ_2, and ρ_3 are the magnitudes of the vectors $\boldsymbol{\rho}_1$, $\boldsymbol{\rho}_2$, and $\boldsymbol{\rho}_3$, and $\boldsymbol{u}_1 = \boldsymbol{\rho}_1/\rho_1$, $\boldsymbol{u}_2 = \boldsymbol{\rho}_2/\rho_2$, and $\boldsymbol{u}_3 = \boldsymbol{\rho}_3/\rho_3$ are the three unit vectors along, respectively, $\boldsymbol{\rho}_1$, $\boldsymbol{\rho}_2$, and $\boldsymbol{\rho}_3$. The three unit vectors $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, and $\boldsymbol{u}_3$ are determined by measuring the declination δ and the right ascension α of the observed body at, respectively, t_1, t_2, and t_3, as follows

$$\begin{aligned}
\boldsymbol{u}_1 &= (\cos\delta_1\cos\alpha_1)\,\boldsymbol{u}_X + (\cos\delta_1\sin\alpha_1)\,\boldsymbol{u}_Y + (\sin\delta_1)\,\boldsymbol{u}_Z \\
\boldsymbol{u}_2 &= (\cos\delta_2\cos\alpha_2)\,\boldsymbol{u}_X + (\cos\delta_2\sin\alpha_2)\,\boldsymbol{u}_Y + (\sin\delta_2)\,\boldsymbol{u}_Z \\
\boldsymbol{u}_3 &= (\cos\delta_3\cos\alpha_3)\,\boldsymbol{u}_X + (\cos\delta_3\sin\alpha_3)\,\boldsymbol{u}_Y + (\sin\delta_3)\,\boldsymbol{u}_Z
\end{aligned}$$

where $\boldsymbol{u}_X$, $\boldsymbol{u}_Y$, and $\boldsymbol{u}_Z$ are the three unit vectors along, respectively, X, Y, and Z.

The three vector equations written above

$$\begin{aligned}
\boldsymbol{r}_1 &= \boldsymbol{r}_{E1} + \boldsymbol{\rho}_1 = \boldsymbol{r}_{E1} + \rho_1(\boldsymbol{\rho}_1/\rho_1) &&= \boldsymbol{r}_{E1} + \rho_1\boldsymbol{u}_1 \\
\boldsymbol{r}_2 &= \boldsymbol{r}_{E2} + \boldsymbol{\rho}_2 = \boldsymbol{r}_{E2} + \rho_2(\boldsymbol{\rho}_2/\rho_2) &&= \boldsymbol{r}_{E2} + \rho_2\boldsymbol{u}_2 \\
\boldsymbol{r}_3 &= \boldsymbol{r}_{E3} + \boldsymbol{\rho}_3 = \boldsymbol{r}_{E3} + \rho_3(\boldsymbol{\rho}_3/\rho_3) &&= \boldsymbol{r}_{E3} + \rho_3\boldsymbol{u}_3
\end{aligned}$$

provide nine scalar equations in twelve unknowns, which are the three components of each of the three position vectors $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$ plus the three magnitudes ρ_1, ρ_2,

and ρ_3 ($3 \times 3 + 3 = 12$ unknowns). Three additional scalar equations are provided by the fact that the motion of the observed body is confined in a plane, because the moment of momentum per unit mass $\boldsymbol{h}$ is a constant vector (see Sect. 1.1). Consequently, the three position vectors $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$ are coplanar. This means that one of these vectors results from a linear combination of the other two. Without loss of generality, we suppose $\boldsymbol{r}_2$ to be a linear combination of $\boldsymbol{r}_1$ and $\boldsymbol{r}_3$, so that

$$\boldsymbol{r}_2 = c_1\boldsymbol{r}_1 + c_3\boldsymbol{r}_3$$

This equation, added to the $9 + 3 = 12$ scalar equations written above, introduces two new unknowns (c_1 and c_3). Thus, we have 12 scalar equations and $12 + 2 = 14$ unknowns. In addition to the constancy of $\boldsymbol{h}$, the Keplerian unperturbed motion of the observed body implies that the position vector of that body at any time can be expressed in terms of the position ($\boldsymbol{r}$) and velocity ($\boldsymbol{v}$) vectors at any other time by means of the Lagrangian coefficients (f and g). Thus, in the present case, the position vectors $\boldsymbol{r}_1$ and $\boldsymbol{r}_3$ of the observed body at times, respectively, t_1 and t_3 can be expressed in terms of $\boldsymbol{r}_2$ and $\boldsymbol{v}_2$ at time t_2, as follows

$$\begin{aligned}\boldsymbol{r}_1 &= f_1\boldsymbol{r}_2 + g_1\boldsymbol{v}_2\\ \boldsymbol{r}_3 &= f_3\boldsymbol{r}_2 + g_3\boldsymbol{v}_2\end{aligned}$$

where f_1 and g_1 are the Lagrangian coefficients computed at time t_1; likewise, f_3 and g_3 are the Lagrangian coefficients computed at time t_3.

As shown in Sect. 1.12, in case of small intervals of time between two consecutive observations, the Lagrangian coefficients f and g depend only on the distance existing at the initial time between the attracted body and its centre of attraction. In this case, designating the intermediate time t_2 as the initial time and r_2 as the distance existing at time t_2 between the two mutually attracting bodies, the Lagrangian coefficients f_1, g_1, f_3, and g_3 depend only on the distance r_2. The two vector equations

$$\begin{aligned}\boldsymbol{r}_1 &= f_1\boldsymbol{r}_2 + g_1\boldsymbol{v}_2\\ \boldsymbol{r}_3 &= f_3\boldsymbol{r}_2 + g_3\boldsymbol{v}_2\end{aligned}$$

correspond to six scalar equations. We have then $12 + 6 = 18$ equations.

On the other hand, the new unknowns introduced by these six equations are four: the three components of the velocity vector $\boldsymbol{v}_2$ and the distance r_2. We have then $14 + 4 = 18$ unknowns. Thus, under the hypothesis made above, there is only one solution to the problem of determining the vectors $\boldsymbol{r}_2$ and $\boldsymbol{v}_2$ at the initial time t_2.

To this end, we first solve the equation

$$\boldsymbol{r}_2 = c_1\boldsymbol{r}_1 + c_3\boldsymbol{r}_3$$

for c_1 and c_3. The vector product of all terms of this equation by $\boldsymbol{r}_3$ yields

$$\boldsymbol{r}_2 \times \boldsymbol{r}_3 = c_1(\boldsymbol{r}_1 \times \boldsymbol{r}_3) + c_3(\boldsymbol{r}_3 \times \boldsymbol{r}_3)$$

Since the vector product of any vector by itself is the zero vector, then $c_3(\boldsymbol{r}_3 \times \boldsymbol{r}_3)$ vanishes identically. Thus,

$$\boldsymbol{r}_2 \times \boldsymbol{r}_3 = c_1(\boldsymbol{r}_1 \times \boldsymbol{r}_3)$$

The scalar product of $(\boldsymbol{r}_1 \times \boldsymbol{r}_3)$ by all terms of the preceding expression yields

$$(\boldsymbol{r}_1 \times \boldsymbol{r}_3) \cdot (\boldsymbol{r}_2 \times \boldsymbol{r}_3) = c_1(\boldsymbol{r}_1 \times \boldsymbol{r}_3) \cdot (\boldsymbol{r}_1 \times \boldsymbol{r}_3) = c_1|\boldsymbol{r}_1 \times \boldsymbol{r}_3|^2$$

where $|\boldsymbol{r}_1 \times \boldsymbol{r}_3|^2$ is the squared magnitude of $\boldsymbol{r}_1 \times \boldsymbol{r}_3$. The preceding equation, solved for c_1, yields

$$c_1 = \frac{(\boldsymbol{r}_1 \times \boldsymbol{r}_3) \cdot (\boldsymbol{r}_2 \times \boldsymbol{r}_3)}{|\boldsymbol{r}_1 \times \boldsymbol{r}_3|^2}$$

By operating likewise, we form the vector product of all terms of the equation

$$\boldsymbol{r}_2 = c_1 \boldsymbol{r}_1 + c_3 \boldsymbol{r}_3$$

by $\boldsymbol{r}_1$. This yields

$$\boldsymbol{r}_2 \times \boldsymbol{r}_1 = c_1(\boldsymbol{r}_1 \times \boldsymbol{r}_1) + c_3(\boldsymbol{r}_3 \times \boldsymbol{r}_1) = c_3(\boldsymbol{r}_3 \times \boldsymbol{r}_1)$$

The scalar product of $(\boldsymbol{r}_3 \times \boldsymbol{r}_1)$ and all terms of the preceding expression yields

$$(\boldsymbol{r}_3 \times \boldsymbol{r}_1) \cdot (\boldsymbol{r}_2 \times \boldsymbol{r}_1) = c_3(\boldsymbol{r}_3 \times \boldsymbol{r}_1) \cdot (\boldsymbol{r}_3 \times \boldsymbol{r}_1) = c_3|\boldsymbol{r}_3 \times \boldsymbol{r}_1|^2$$

Hence,

$$c_3 = \frac{(\boldsymbol{r}_3 \times \boldsymbol{r}_1) \cdot (\boldsymbol{r}_2 \times \boldsymbol{r}_1)}{|\boldsymbol{r}_3 \times \boldsymbol{r}_1|^2}$$

Now, we form the vector product $\boldsymbol{r}_1 \times \boldsymbol{r}_3$ and introduce

$$\begin{aligned} \boldsymbol{r}_1 &= f_1 \boldsymbol{r}_2 + g_1 \boldsymbol{v}_2 \\ \boldsymbol{r}_3 &= f_3 \boldsymbol{r}_2 + g_3 \boldsymbol{v}_2 \end{aligned}$$

into the product $\boldsymbol{r}_1 \times \boldsymbol{r}_3$. This yields

$$\begin{aligned} \boldsymbol{r}_1 \times \boldsymbol{r}_3 &= (f_1 \boldsymbol{r}_2 + g_1 \boldsymbol{v}_2) \times (f_3 \boldsymbol{r}_2 + g_3 \boldsymbol{v}_2) \\ &= f_1 f_3(\boldsymbol{r}_2 \times \boldsymbol{r}_2) + f_1 g_3(\boldsymbol{r}_2 \times \boldsymbol{v}_2) + g_1 f_3(\boldsymbol{v}_2 \times \boldsymbol{r}_2) + g_1 g_3(\boldsymbol{v}_2 \times \boldsymbol{v}_2) \end{aligned}$$

Now, since the vector product of any vector by itself yields the zero vector, and

$$\begin{aligned}\boldsymbol{r}_2 \times \boldsymbol{v}_2 &= \boldsymbol{h}\\ \boldsymbol{v}_2 \times \boldsymbol{r}_2 &= -(\boldsymbol{r}_2 \times \boldsymbol{v}_2) = -\boldsymbol{h}\end{aligned}$$

where $\boldsymbol{h}$ is the constant moment of momentum, then

$$\begin{aligned}\boldsymbol{r}_1 \times \boldsymbol{r}_3 &= (f_1 g_3 - g_1 f_3)\boldsymbol{h}\\ \boldsymbol{r}_3 \times \boldsymbol{r}_1 &= -(\boldsymbol{r}_1 \times \boldsymbol{r}_3) = -(f_1 g_3 - g_1 f_3)\boldsymbol{h}\end{aligned}$$

It follows that

$$|\boldsymbol{r}_1 \times \boldsymbol{r}_3|^2 = |\boldsymbol{r}_3 \times \boldsymbol{r}_1|^2 = (f_1 g_3 - g_1 f_3)^2 h^2$$

Likewise, we form the vector product $\boldsymbol{r}_2 \times \boldsymbol{r}_3$ and introduce $\boldsymbol{r}_3 = f_3\,\boldsymbol{r}_2 + g_3\,\boldsymbol{v}_2$ into the product $\boldsymbol{r}_2 \times \boldsymbol{r}_3$. This yields

$$\boldsymbol{r}_2 \times \boldsymbol{r}_3 = \boldsymbol{r}_2 \times (f_3 \boldsymbol{r}_2 + g_3 \boldsymbol{v}_2) = f_3(\boldsymbol{r}_2 \times \boldsymbol{r}_2) + g_3(\boldsymbol{r}_2 \times \boldsymbol{v}_2) = g_3 \boldsymbol{h}$$

Again, we form the vector product $\boldsymbol{r}_2 \times \boldsymbol{r}_1$ and introduce $\boldsymbol{r}_1 = f_1\,\boldsymbol{r}_2 + g_1\,\boldsymbol{v}_2$ into the product $\boldsymbol{r}_2 \times \boldsymbol{r}_1$. This yields

$$\boldsymbol{r}_2 \times \boldsymbol{r}_1 = \boldsymbol{r}_2 \times (f_1 \boldsymbol{r}_2 + g_1 \boldsymbol{v}_2) = f_1(\boldsymbol{r}_2 \times \boldsymbol{r}_2) + g_1(\boldsymbol{r}_2 \times \boldsymbol{v}_2) = g_1 \boldsymbol{h}$$

In summary, we have obtained the following expressions

$$\begin{aligned}\boldsymbol{r}_3 \times \boldsymbol{r}_1 &= -(\boldsymbol{r}_1 \times \boldsymbol{r}_3) = -(f_1 g_3 - g_1 f_3)\boldsymbol{h}\\ |\boldsymbol{r}_1 \times \boldsymbol{r}_3|^2 &= |\boldsymbol{r}_3 \times \boldsymbol{r}_1|^2 = (f_1 g_3 - g_1 f_3)^2 h^2\\ \boldsymbol{r}_2 \times \boldsymbol{r}_1 &= g_1 \boldsymbol{h}\end{aligned}$$

which in turn, substituted into

$$c_3 = \frac{(\boldsymbol{r}_3 \times \boldsymbol{r}_1) \cdot (\boldsymbol{r}_2 \times \boldsymbol{r}_1)}{|\boldsymbol{r}_3 \times \boldsymbol{r}_1|^2}$$

yield

$$c_3 = \frac{[-(f_1 g_3 - g_1 f_3)\boldsymbol{h}] \cdot [g_1 \boldsymbol{h}]}{(f_1 g_3 - g_1 f_3)^2 h^2}$$

Since $\boldsymbol{h} \cdot \boldsymbol{h} = h^2$, then

$$c_3 = -\frac{g_1}{f_1 g_3 - g_1 f_3}$$

Let us consider now the expression

$$c_1 = \frac{(\boldsymbol{r}_1 \times \boldsymbol{r}_3) \cdot (\boldsymbol{r}_2 \times \boldsymbol{r}_3)}{|\boldsymbol{r}_1 \times \boldsymbol{r}_3|^2}$$

which has been derived above. Substituting

$$\begin{aligned}(\boldsymbol{r}_1 \times \boldsymbol{r}_3) &= (f_1 g_3 - g_1 f_3)\boldsymbol{h} \\ \boldsymbol{r}_2 \times \boldsymbol{r}_3 &= g_3 \boldsymbol{h} \\ |\boldsymbol{r}_1 \times \boldsymbol{r}_3|^2 &= |\boldsymbol{r}_3 \times \boldsymbol{r}_1|^2 = (f_1 g_3 - g_1 f_3)^2 h^2\end{aligned}$$

into this expression yields

$$c_1 = \frac{[(f_1 g_3 - g_1 f_3)\boldsymbol{h}] \cdot [g_3 \boldsymbol{h}]}{(f_1 g_3 - g_1 f_3)^2 h^2}$$

that is,

$$c_1 = \frac{g_3}{f_1 g_3 - g_1 f_3}$$

By so doing, the coefficients c_1 and c_3 appearing in the equation

$$\boldsymbol{r}_2 = c_1 \boldsymbol{r}_1 + c_3 \boldsymbol{r}_3$$

depend on the Lagrange coefficients f and g only. We set

$$\begin{aligned}\tau_1 &= t_1 - t_2 \\ \tau_3 &= t_3 - t_2\end{aligned}$$

where τ_1 and τ_3 are, by hypothesis, small intervals. Therefore, we can take the first two terms of the series expansions for the Lagrangian coefficients f and g

$$\begin{aligned}f &= 1 - \frac{1}{2}\epsilon_0 \tau^2 + \frac{1}{2}\epsilon_0 \lambda_0 \tau^3 + \frac{1}{24}\left(-2\epsilon_0 - 15\lambda_0^2 + 3\psi_0\right)\tau^4 + \cdots \\ g &= \tau - \frac{1}{6}\epsilon_0 \tau^3 + \frac{1}{4}\epsilon_0 \lambda_0 \tau^4 + \cdots\end{aligned}$$

(see Sect. 1.12). In other words, we truncate these series expansions after the first two terms, as follows

$$f_1 \approx 1 - \frac{1}{2}\left(\frac{\mu}{r_2^3}\right)\tau_1^2 \quad f_3 \approx 1 - \frac{1}{2}\left(\frac{\mu}{r_2^3}\right)\tau_3^2$$
$$g_1 \approx \tau_1 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau_1^3 \quad g_3 \approx \tau_3 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau_3^3$$

This is because, by definition, $\varepsilon_0 = \mu/r_2^3$. Thus, the quantity $f_1 g_3 - g_1 f_3$ can be approximated as follows

$$\begin{aligned} f_1 g_3 - g_1 f_3 &= \left[1 - \frac{1}{2}\left(\frac{\mu}{r_2^3}\right)\tau_1^2\right]\left[\tau_3 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau_3^3\right] - \left[\tau_1 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau_1^3\right]\left[1 - \frac{1}{2}\left(\frac{\mu}{r_2^3}\right)\tau_3^2\right] \\ &= (\tau_3 - \tau_1) - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\left(\tau_3^3 - 3\tau_1\tau_3^2 + 3\tau_1^2\tau_3 - \tau_1^3\right) \\ &\quad + \frac{1}{12}\left(\frac{\mu^2}{r_2^6}\right)\left(\tau_1^2\tau_3^3 - \tau_1^3\tau_3^2\right) \\ &= (\tau_3 - \tau_1) - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)(\tau_3 - \tau_1)^3 + \frac{1}{12}\left(\frac{\mu^2}{r_2^6}\right)\left(\tau_1^2\tau_3^3 - \tau_1^3\tau_3^2\right) \end{aligned}$$

Again, since τ_1 and τ_3 are small intervals, then the term $(\mu^2/r_2^6)\,(\tau_1^2\tau_3^3 - \tau_1^3\tau_3^2)/12$ can be neglected. Thus, setting $\tau = \tau_3 - \tau_1$ yields

$$f_1 g_3 - g_1 f_3 \approx \tau - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau^3$$

where τ is the interval of time elapsed from the first of the three observations to the last. Now, substituting

$$f_1 g_3 - g_1 f_3 \approx \tau - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau^3 \quad g_3 \approx \tau_3 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau_3^3$$

into $c_1 = g_3/(f_1\, g_3 - g_1\, f_3)$ yields

$$c_1 = \frac{\tau_3 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau_3^3}{\tau - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau^3} = \frac{\tau_3}{\tau}\frac{1 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau_3^2}{1 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau^2} = \frac{\tau_3}{\tau}\left[1 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau_3^2\right]\left[1 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau^2\right]^{-1}$$

Using the well-known binomial expansion

$$(a+b)^n = a^n + na^{n-1}b + \frac{n(n-1)}{2!}a^{n-2}b^2 + \frac{n(n-1)(n-2)}{3!}a^{n-3}b^3 + \cdots$$

with $a = 1, b = -\frac{1}{6}(\mu/r_2^3)\tau^2$ and $n = -1$, truncated after the term containing the second power of τ, yields

$$\left[1 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau^2\right]^{-1} \approx 1^{-1} + (-1)1^{1-1}\left[-\frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau^2\right] = 1 + \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau^2$$

This in turn substituted into

$$c_1 \approx \frac{\tau_3}{\tau}\left[1 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau_3^2\right]\left[1 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau^2\right]^{-1}$$

yields

$$c_1 \approx \frac{\tau_3}{\tau}\left[1 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau_3^2\right]\left[1 + \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau^2\right] = \frac{\tau_3}{\tau}\left[1 + \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)(\tau^2 - \tau_3^2)\right]$$

By operating likewise, there results

$$c_3 \approx -\frac{\tau_1}{\tau}\left[1 + \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)(\tau^2 - \tau_1^2)\right]$$

We have hitherto obtained approximate expressions of the coefficients (c_1 and c_3) appearing in the equation $\boldsymbol{r}_2 = c_1\,\boldsymbol{r}_1 + c_3\,\boldsymbol{r}_3$. These expressions depend only on the (known) time intervals between observations and the (as yet unknown) distance r_2 of the attracted body from its centre of attraction at time t_2.

The next stage of this development is the expression of the ranges ρ_1, ρ_2, and ρ_3 as functions of c_1 and c_3. To this effect, we substitute

$$\begin{aligned}\boldsymbol{r}_1 &= \boldsymbol{r}_{E1} + \rho_1\boldsymbol{r}_1\\ \boldsymbol{r}_2 &= \boldsymbol{r}_{E2} + \rho_2\boldsymbol{u}_2\\ \boldsymbol{r}_3 &= \boldsymbol{r}_{E3} + \rho_3\boldsymbol{u}_3\end{aligned}$$

into $\boldsymbol{r}_2 = c_1\,\boldsymbol{r}_1 + c_3\,\boldsymbol{r}_3$. This yields

$$\boldsymbol{r}_{E2} + \rho_2\boldsymbol{u}_2 = c_1(\boldsymbol{r}_{E2} + \rho_1\boldsymbol{u}_1) + c_3(\boldsymbol{r}_{E2} + \rho_3\boldsymbol{u}_3)$$

which can also be written as follows

$$c_1\rho_1\boldsymbol{u}_1 - \rho_2\boldsymbol{u}_2 + c_3\rho_3\boldsymbol{u}_3 = -c_1\boldsymbol{r}_{E1} + \boldsymbol{r}_{E2} - c_3\boldsymbol{r}_{E3}$$

The scalar product of each term of the preceding expression by $(\boldsymbol{u}_2 \times \boldsymbol{u}_3)$ yields

$$c_1\rho_1\boldsymbol{u}_1 \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3) - \rho_2\boldsymbol{u}_2 \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3) + c_3\rho_3\boldsymbol{u}_3 \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3)$$
$$= -c_1\boldsymbol{r}_{E1} \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3) + \boldsymbol{r}_{E2} \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3) - c_3\boldsymbol{r}_{E3} \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3)$$

Since $\boldsymbol{u}_2 \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3) = \boldsymbol{u}_3 \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3) = 0$, then the preceding expression becomes

$$c_1\rho_1\boldsymbol{u}_1 \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3) = -c_1\boldsymbol{r}_{E1} \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3) + \boldsymbol{r}_{E2} \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3) - c_3\boldsymbol{r}_{E2} \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3)$$

To simplify the notation, we set

$$D_0 = \boldsymbol{u}_1 \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3)$$
$$D_{11} = \boldsymbol{r}_{E1} \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3)$$
$$D_{21} = \boldsymbol{r}_{E2} \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3)$$
$$D_{31} = \boldsymbol{r}_{E3} \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3)$$

Assuming $D_0 \neq 0$, that is, assuming that $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, and $\boldsymbol{u}_3$ are not coplanar, we have

$$c_1\rho_1 D_0 = -c_1 D_{11} + D_{21} - c_3 D_{31}$$

which, solved for ρ_1, yields

$$\rho_1 = \left(-D_{11} + \frac{1}{c_1}D_{21} - \frac{c_3}{c_1}D_{31}\right)\frac{1}{D_0}$$

By operating likewise, we take the scalar product of each term of

$$c_1\rho_1\boldsymbol{u}_1 - \rho_2\boldsymbol{u}_2 + c_3\rho_3\boldsymbol{u}_3 = -c_1\boldsymbol{r}_{E1} + \boldsymbol{r}_{E2} - c_3\boldsymbol{r}_{E3}$$

by $(\boldsymbol{u}_1 \times \boldsymbol{u}_3)$. This yields

$$c_1\rho_1\boldsymbol{u}_1 \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) - \rho_2\boldsymbol{u}_2 \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) + c_3\rho_3\boldsymbol{u}_3 \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3)$$
$$= -c_1\boldsymbol{r}_{E1} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) + \boldsymbol{r}_{E2} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) - c_3\boldsymbol{r}_{E3} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3)$$

Since $\boldsymbol{u}_1 \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) = \boldsymbol{u}_3 \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) = 0$, then the preceding expression becomes

$$-\rho_2\boldsymbol{u}_2 \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) = -c_1\boldsymbol{r}_{E1} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) + \boldsymbol{r}_{E2} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) - c_3\boldsymbol{r}_{E3} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3)$$

Since $-\rho_2\,\boldsymbol{u}_2 \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) = \rho_2\,\boldsymbol{u}_2 \cdot (\boldsymbol{u}_3 \times \boldsymbol{u}_1) = \rho_2\,\boldsymbol{u}_1 \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3) = \rho_2\,D_0$, then the preceding expression becomes

$$\rho_2 D_0 = -c_1\boldsymbol{r}_{E1} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) + \boldsymbol{r}_{E2} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) - c_3\boldsymbol{r}_{E2} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3)$$

By setting

$$\begin{aligned} D_{12} &= \boldsymbol{r}_{E1} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) \\ D_{22} &= \boldsymbol{r}_{E2} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) \\ D_{32} &= \boldsymbol{r}_{E2} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) \end{aligned}$$

and solving the preceding expression for ρ_2, we obtain

$$\rho_2 = (-c_1 D_{12} + D_{22} - c_3 D_{32}) \frac{1}{D_0}$$

By operating likewise, we take the scalar product of each term of

$$c_1 \rho_1 \boldsymbol{u}_1 - \rho_2 \boldsymbol{u}_2 + c_3 \rho_3 \boldsymbol{u}_3 = -c_1 \boldsymbol{r}_{E1} + \boldsymbol{r}_{E2} - c_3 \boldsymbol{r}_{E3}$$

by $(\boldsymbol{u}_1 \times \boldsymbol{u}_2)$. This yields

$$\begin{aligned} &c_1 \rho_1 \boldsymbol{u}_1 \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) - \rho_2 u_2 \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) + c_3 \rho_3 \boldsymbol{u}_3 \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) \\ &\quad = -c_1 \boldsymbol{r}_{E1} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) + \boldsymbol{r}_{E2} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) - c_3 \boldsymbol{r}_{E3} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) \end{aligned}$$

Since $\boldsymbol{u}_1 \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) = \boldsymbol{u}_2 \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) = 0$, then the preceding expression becomes

$$c_3 \rho_3 \boldsymbol{u}_3 \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) = -c_1 \boldsymbol{r}_{E3} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) + \boldsymbol{r}_{E3} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) - c_3 \boldsymbol{r}_{E3} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2)$$

By noting that $\boldsymbol{u}_3 \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) = \boldsymbol{u}_1 \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3) = D_0$, setting

$$\begin{aligned} D_{13} &= \boldsymbol{r}_{E1} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) \\ D_{23} &= \boldsymbol{r}_{E2} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) \\ D_{33} &= \boldsymbol{r}_{E3} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) \end{aligned}$$

and solving the preceding expression for ρ_3, we obtain

$$\rho_3 = \left(-\frac{c_1}{c_3} D_{13} + \frac{1}{c_3} D_{23} - D_{33} \right) \frac{1}{D_0}$$

Now, substituting

$$c_1 \approx \frac{\tau_3}{\tau} \left[1 + \frac{1}{6} \left(\frac{\mu}{r_2^3} \right) (\tau^2 - \tau_3^2) \right] \quad c_3 \approx -\frac{\tau_1}{\tau} \left[1 + \frac{1}{6} \left(\frac{\mu}{r_2^3} \right) (\tau^2 - \tau_1^2) \right]$$

into $\rho_2 = (-c_1\ D_{12} + D_{22} - c_3\ D_{32})/D_0$ yields

$$\rho_2 = -D_{12}\tau_3 \frac{6 + \left(\frac{\mu}{r_2^3}\right)(\tau^2 - \tau_3^2)}{6D_0\tau} + \frac{D_{22}}{D_0} + D_{32}\tau_1 \frac{6 + \left(\frac{\mu}{r_2^3}\right)(\tau^2 - \tau_1^2)}{6D_0\tau}$$

which, after setting

$$A = \left(-D_{12}\frac{\tau_3}{\tau} + D_{22} + D_{32}\frac{\tau_1}{\tau}\right)\frac{1}{D_0}$$
$$B = \frac{-D_{12}\tau_3\left(\tau^2 - \tau_3^2\right) + D_{32}\tau_1\left(\tau^2 - \tau_1^2\right)}{6D_0\tau}$$

becomes

$$\rho_2 = A + \frac{\mu B}{r_2^3}$$

Operating the same substitution, that is,

$$c_1 \approx \frac{\tau_3}{\tau}\left[1 + \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)(\tau^2 - \tau_3^2)\right] \quad c_3 \approx -\frac{\tau_1}{\tau}\left[1 + \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)(\tau^2 - \tau_1^2)\right]$$

into

$$\rho_1 = \left(-D_{11} + \frac{1}{c_1}D_{21} - \frac{c_3}{c_1}D_{31}0\right)\frac{1}{D_0} \quad \rho_3 = \left(-\frac{c_1}{c_3}D_{13} + \frac{1}{c_3}D_{23} - D_{33}\right)\frac{1}{D_0}$$

leads to

$$\rho_1 = \left[\frac{6(D_{31}\frac{\tau_1}{\tau_3} + D_{21}\frac{\tau}{\tau_3})r_2^3 + \mu D_{31}(\tau^2 - \tau_1^2)\frac{\tau_1}{\tau_3}}{6r_2^3 + \mu(\tau^2 - \tau_3^2)} - D_{11}\right]\frac{1}{D_0}$$
$$\rho_3 = \left[\frac{6(D_{13}\frac{\tau_3}{\tau_1} - D_{23}\frac{\tau}{\tau_1})r_2^3 + \mu D_{13}(\tau^2 - \tau_3^2)\frac{\tau_3}{\tau_1}}{6r_2^3 + \mu(\tau^2 - \tau_3^2)} - D_{33}\right]\frac{1}{D_0}$$

The equation written above $\rho_2 = A + \mu B/r_2^3$ expresses the range ρ_2 as a function of the radius vector r_2 of the observed object measured from the centre of mass of the Earth at time t_2. Another relation between ρ_2 and r_2 is provided by

$$\boldsymbol{r}_2 = \boldsymbol{r}_{E2} + \rho_2\boldsymbol{u}_2$$

The scalar product of $\boldsymbol{r}_2$ by itself yields

$$\boldsymbol{r}_2 \cdot \boldsymbol{r}_2 = (\boldsymbol{r}_{E2} + \rho_2 \boldsymbol{u}_2) \cdot (\boldsymbol{r}_{E2} + \rho_2 \boldsymbol{u}_2) = r_{E2}^2 + 2\rho_2(\boldsymbol{r}_{E2} \cdot \boldsymbol{u}_2) + \rho_2^2 = r_{E2}^2 + 2E\rho_2 + \rho_2^2$$

where $E = \boldsymbol{r}_{E2} \cdot \boldsymbol{u}_2$. Substituting $\rho_2 = A + \mu B/r_2^3$ into $r_2^2 = r_{E2}^2 + 2E\rho_2 + \rho_2^2$ yields

$$\begin{aligned} r_2^2 &= r_{E2}^2 + 2E\left(A + \frac{\mu B}{r_2^3}\right) + \left(A + \frac{\mu B}{r_2^3}\right)^2 \\ &= r_{E2}^2 + 2EA + \frac{2\mu EB}{r_2^3} + A^2 + \frac{2\mu AB}{r_2^3} + \frac{\mu^2 B^2}{r_2^6} \end{aligned}$$

Multiplying all terms of the preceding expression by r_2^6 yields

$$r_2^8 - \left(r_{E2}^2 + 2EA + A^2\right) r_2^6 - 2\mu B(E + A) r_2^3 - \mu^2 B^2 = 0$$

If we set for convenience

$$\begin{aligned} x &= r_2 \\ a &= -\left(r_{E2}^2 + 2EA + A^2\right) \\ b &= -2\mu B\,(E + A) \\ c &= -\mu^2 B^2 \end{aligned}$$

the preceding expression becomes

$$x^8 + ax^6 + bx^3 + c = 0$$

which is known as Lagrange's equation. Since this polynomial has four terms, then (according to the Descartes' rule of signs, which states that the number of positive roots of a polynomial with real coefficients is equal to the number of changes of sign in the sequence of the coefficients of the polynomial, or is less than this number by a multiple of 2) the polynomial may have no more than three positive roots. When the correct root r_2 of this equation has been found, it must be substituted into

$$\rho_1 = \left[\frac{6\left(D_{31}\dfrac{\tau_1}{\tau_3} + D_{21}\dfrac{\tau}{\tau_3}\right) r_2^3 + \mu D_{31}\left(\tau^2 - \tau_1^2\right)\dfrac{\tau_1}{\tau_3}}{6r_2^3 + \mu(\tau^2 - \tau_3^2)} - D_{11}\right] \frac{1}{D_0}$$

$$\rho_2 = A + \frac{\mu B}{r_2^3}$$

$$\rho_3 = \left[\frac{6\left(D_{13}\dfrac{\tau_3}{\tau_1} - D_{23}\dfrac{\tau}{\tau_1}\right) r_2^3 + \mu D_{13}(\tau^2 - \tau_3^2)\dfrac{\tau_3}{\tau_1}}{6r_2^3 + \mu(\tau^2 - \tau_3^2)} - D_{33}\right] \frac{1}{D_0}$$

in order to compute the ranges ρ_1, ρ_2, and ρ_3. Then, the following equations

$$\begin{aligned}\boldsymbol{r}_1 &= \boldsymbol{r}_{E1} + \rho_1 \boldsymbol{u}_1 \\ \boldsymbol{r}_2 &= \boldsymbol{r}_{E2} + \rho_2 \boldsymbol{u}_2 \\ \boldsymbol{r}_3 &= \boldsymbol{r}_{E3} + \rho_3 \boldsymbol{u}_3\end{aligned}$$

yield the position vectors $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$ of the observed object at times, respectively, t_1, t_2, and t_3. It remains to compute the velocity vector $\boldsymbol{v}_2$ of the observed object at time t_2. To this end, the equation $\boldsymbol{r}_1 = f_1\, \boldsymbol{r}_2 + g_1\, \boldsymbol{v}_2$ is solved for $\boldsymbol{r}_2$. This yields

$$\boldsymbol{r}_2 = \frac{\boldsymbol{r}_1 - g_1 \boldsymbol{v}_2}{f}$$

The resulting value of $\boldsymbol{r}_2$ is substituted into $\boldsymbol{r}_3 = f_3\, \boldsymbol{r}_2 + g_3\, \boldsymbol{v}_2$. This yields

$$\boldsymbol{r}_3 = \frac{f_3}{f_1}(\boldsymbol{r}_1 - g_1 \boldsymbol{v}_2) + g_3 \boldsymbol{v}_2$$

and the preceding equation is solved for $\boldsymbol{v}_2$. This yields

$$\boldsymbol{v}_2 = \left(\frac{f_1}{f_1 g_3 - g_1 f_3}\right)\boldsymbol{r}_3 - \left(\frac{f_3}{f_1 g_3 - g_1 f_3}\right)\boldsymbol{r}_1$$

At first, the values of f_1, f_3, g_1, and g_3 to be used for computing $\boldsymbol{r}_2$ and $\boldsymbol{v}_2$ are those derived previously, which are rewritten below for convenience:

$$\begin{aligned} f_1 &\approx 1 - \frac{1}{2}\left(\frac{\mu}{r_2^3}\right)\tau_1^2 \qquad & f_3 &\approx 1 - \frac{1}{2}\left(\frac{\mu}{r_2^3}\right)\tau_3^2 \\ g_1 &\approx \tau_1 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau_1^3 \qquad & g_3 &\approx \tau_3 - \frac{1}{6}\left(\frac{\mu}{r_2^3}\right)\tau_3^3 \end{aligned}$$

Successively, improved values of fSince the condition$_1$, f_3, g_1, and g_3 are computed, as will be shown below, and new values of $\boldsymbol{r}_2$ and $\boldsymbol{v}_2$ are computed by means of these improved values, until convergence is reached. This method can be illustrated by the following example, which is based on a series of astronomical observations performed by Healy [33] and concerning the COBE artificial satellite (USSPACECOM Catalogue No. 20322; International Designation code 1989-089-A). On the 6th of November 2000, Healy made seven observations of the COBE satellite, three of which are shown in the following table.

Observed time (EST)	Right ascension (hh:mm:ss)	Declination (°)
17:31:29	21:48:00	−16.3
17:34:30	21:14:00	46.9
17:37:30	11:03:00	76.1

Since UTC = EST + 5, then (neglecting the difference between UTC and UT1) the three EST times indicated above correspond, respectively, to

$$\begin{aligned} \mathrm{UT1}_1 &= \mathrm{EST}_1 + 5 = 22^{\mathrm{h}}31^{\mathrm{m}}29^{\mathrm{s}} \\ \mathrm{UT1}_2 &= \mathrm{EST}_2 + 5 = 22^{\mathrm{h}}34^{\mathrm{m}}30^{\mathrm{s}} \\ \mathrm{UT1}_3 &= \mathrm{EST}_3 + 5 = 22^{\mathrm{h}}37^{\mathrm{m}}30^{\mathrm{s}} \end{aligned}$$

The three values of the right ascension, expressed in degrees, are as follows

$$\begin{aligned} (21 + 48/60) \times 360/24 &= 327.00 \\ (21 + 14/60) \times 360/24 &= 318.50 \\ (11 + 3/60) \times 360/24 &= 165.75 \end{aligned}$$

Healy found these values by means of the 14″ Schmidt Cassegrain telescope of the observatory of the University of Maryland, located at latitude $\varphi = 39°.00167$ North, longitude $\lambda = -76°.95667$ east and altitude $H = 53$ m.

As has been shown above, the same values can be written as follows

UT1	Right ascension (°)	Declination (°)
22:31:29	327.00	−16.3
22:34:30	318.50	46.9
22:37:30	165.75	76.1

By applying the methods shown in Sect. 2.5, the Greenwich sidereal time θ_{G0} corresponding to the 6 November 2000, at 00^{h}:00^{m}:00^{s} UT1, results from

$$\begin{aligned} a &= \mathrm{INT}(y/100) = \mathrm{INT}(2000/100) = 20 \\ b &= \mathrm{INT}(a/4) = \mathrm{INT}(20/4) = 5 \\ c &= 2 - a + b = 2 - 20 + 5 = -13 \\ e &= \mathrm{INT}[365.25(y + 4716)] = \mathrm{INT}[365.25 \times (2000 + 4716)] = 2453019 \\ f &= \mathrm{INT}[30.6001(m + 1)] = \mathrm{INT}[30.6001 \times (11 + 1)] = 367 \\ J_0 &= c + d + e + f - 1524.5 = -13 + 6 + 2453019 + 367 - 1524.5 = 2451854.5 \\ T_0 &= (J_0 - 2451545)/36525 = (2451854.5 - 2451545)/36525 = 0.0084736 \\ \theta_{G0} &= 100.4606184 + 36000.77005\,T_0 + 0.000387933\,{T_0}^2 - 2.5875 \times 10^{-8}{T_0}^3 \\ &= 100.4606184 + 36000.77005 \times 0.0084736 + 0.000387933 \times 0.0084736^2 \\ &\quad - 2.5875 \times 10^{-8} \times 0.0084736^3 = 405°.51847 \end{aligned}$$

This value is brought into the range $0° \leq \theta_{G0} \leq 360°$ by subtracting 360°, as follows

$$\theta_{G0} = 405°.51847 - 360° = 45°.51847$$

The Greenwich sidereal times θ_{G1}, θ_{G2}, and θ_{G3} corresponding to $UT1_1$ = 22:31:29, $UT1_2$ = 22:34:30 and $UT1_3$ = 22:37:30 are, respectively,

$$\theta_{G1} = \theta_{G0} + 360.985647366 \times (22 + 31/60 + 29/3600)/24 = 384°.314363$$
$$\theta_{G2} = \theta_{G0} + 360.985647366 \times (22 + 34/60 + 30/3600)/24 = 385°.070595$$
$$\theta_{G3} = \theta_{G0} + 360.985647366 \times (22 + 37/60 + 30/3600)/24 = 385°.822648$$

The local sidereal times of the three observations are, respectively,

$$\theta_1 = \theta_{G1} + \lambda = 384°.314363 - 76°.95667 = 307°.357693$$
$$\theta_2 = \theta_{G1} + \lambda = 385°.070595 - 76°.95667 = 308°.113925$$
$$\theta_3 = \theta_{G1} + \lambda = 385°.822648 - 76°.95667 = 308°.865978$$

We want to compute the position and velocity vectors of the COBE satellite at $UT1_2 = 22^h{:}34^m{:}30^s$, with an accuracy of five significant figures. The angles measured by the station and the local sidereal times are given in the following table.

Obs. No.	Time (s)	Right ascension (°)	Declination (°)	Local sidereal time (°)
1	0	327.00	−16.300	307.357693
2	181	318.50	46.900	308.113925
3	361	165.75	76.100	308.865978

First, we compute the three geocentric position vectors ($\boldsymbol{r}_{E1}$, $\boldsymbol{r}_{E2}$, and $\boldsymbol{r}_{E3}$) of the observatory at the three given times t_1, t_2, and t_3, as shown in Sect. 2.5 (that is, taking the equatorial radius a_E and the flattening f of the Earth equal, respectively, to 6378.137 km and 0.0033528). Assuming the geodetic latitude equal to the geographic latitude, the position vectors $\boldsymbol{r}_{E1}$, $\boldsymbol{r}_{E2}$, and $\boldsymbol{r}_{E3}$ are

$$x_H = \left\{a_E/\left[1 - \left(2f - f^2\right)\sin^2\varphi\right]^{1/2} + H\right\}\cos\varphi = \{6378.137/[1 - (2 \times 0.0033528 - 0.0033528^2)\sin^2 39.00167]^{1/2} + 0.053\}\cos 39.00167 = 4963.272\,\text{km}$$

$$z_H = \left\{a_E(1-f)^2/\left[1 - \left(2f - f^2\right)\sin^2\varphi\right]^{1/2} + H\right\}\sin\varphi = \left\{6378.137 \times (1 - 0.0033528)^2 / \left[1 - (2 \times 0.0033528 - 0.0033528^2)\sin^2 39.00167\right]^{1/2} + 0.053\right\}\sin 39.00167 = 3992.517\,\text{km}$$

$$\begin{aligned}\boldsymbol{r}_{E1} &= (x_H \cos\theta_1)\,\boldsymbol{u}_X + (x_H \sin\theta_1)\,\boldsymbol{u}_Y + z_H\,\boldsymbol{u}_Z = (4963.272\ \cos 307.357693)\,\boldsymbol{u}_X \\ &\quad + (4963.272\ \sin 307.357693)\,\boldsymbol{u}_Y + 3992.517\,\boldsymbol{u}_Z \\ &= 3011.7\,\boldsymbol{u}_X - 3945.1\,\boldsymbol{u}_Y + 3992.5\,\boldsymbol{u}_Z\end{aligned}$$

$$\begin{aligned}\boldsymbol{r}_{E2} &= (x_H \cos\theta_2)\boldsymbol{u}_X + (x_H \sin\theta_2)\boldsymbol{u}_Y + z_H\boldsymbol{u}_Z = (4963.272\ \cos 308.113925)\boldsymbol{u}_X \\ &\quad + (4963.272\ \sin 308.113925)\boldsymbol{u}_Y + 3992.517\,\boldsymbol{u}_Z \\ &= 3063.5\,\boldsymbol{u}_X - 3905.0\,\boldsymbol{u}_Y + 3992.5\,\boldsymbol{u}_Z\end{aligned}$$

$$\begin{aligned}\boldsymbol{r}_{E3} &= (x_H \cos\theta_3)\boldsymbol{u}_X + (x_H \sin\theta_3)\boldsymbol{u}_Y + z_H\boldsymbol{u}_Z = (4963.272\ \cos 308.865978)\boldsymbol{u}_X \\ &\quad + (4963.272\ \sin 308.865978)\boldsymbol{u}_Y + 3992.517\,\boldsymbol{u}_Z \\ &= 3114.5\,\boldsymbol{u}_X - 3864.5\,\boldsymbol{u}_Y + 3992.5\,\boldsymbol{u}_Z\end{aligned}$$

The three unit vectors $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, and $\boldsymbol{u}_3$ result from

$$\begin{aligned}\boldsymbol{u}_1 &= (\cos\delta_1 \cos\alpha_1)\boldsymbol{u}_X + (\cos\delta_1 \sin\alpha_1)\boldsymbol{u}_Y + (\sin\delta_1)\boldsymbol{u}_Z = [\cos(-16.3)\ \cos 327.0]\,\boldsymbol{u}_X \\ &\quad + [\cos(-16.3)\ \sin 327.0]\boldsymbol{u}_Y + [\sin(-16.3)]\boldsymbol{u}_Z \\ &= 0.80496\,\boldsymbol{u}_X - 0.52275\,\boldsymbol{u}_Y - 0.28067\,\boldsymbol{u}_Z\end{aligned}$$

$$\begin{aligned}\boldsymbol{u}_2 &= (\cos\delta_2 \cos\alpha_2)\,\boldsymbol{u}_X + (\cos\delta_2 \sin\alpha_2)\boldsymbol{u}_Y + (\sin\delta_2)\boldsymbol{u}_Z = (\cos 46.9\ \cos 318.5)\boldsymbol{u}_X \\ &\quad + (\cos 46.9\ \sin 318.5)\boldsymbol{u}_Y + (\sin 46.9)\boldsymbol{u}_Z \\ &= 0.51174\,\boldsymbol{u}_X - 0.45275\,\boldsymbol{u}_Y + 0.73016\,\boldsymbol{u}_Z\end{aligned}$$

$$\begin{aligned}\boldsymbol{u}_3 &= (\cos\delta_3 \cos\alpha_3)\,\boldsymbol{u}_X + (\cos\delta_3 \sin\alpha_3)\boldsymbol{u}_Y + (\sin\delta_3)\boldsymbol{u}_Z = (\cos 76.1\ \cos 165.75)\boldsymbol{u}_X \\ &\quad + (\cos 76.1\ \sin 165.75)\boldsymbol{u}_Y + (\sin 76.1)\boldsymbol{u}_Z \\ &= -0.23284\,\boldsymbol{u}_X + 0.059133\,\boldsymbol{u}_Y + 0.97072\,\boldsymbol{u}_Z\end{aligned}$$

The time intervals τ_1, τ_3, and τ are computed as follows

$$\begin{aligned}\tau_1 &= t_1 - t_2 = 0 - 181 = -181\,\text{s} \\ \tau_3 &= t_3 - t_2 = 361 - 181 = 180\,\text{s} \\ \tau &= t_3 - t_1 = 361 - 0 = 361\,\text{s}\end{aligned}$$

The vector products $(\boldsymbol{u}_2 \times \boldsymbol{u}_3)$, $(\boldsymbol{u}_1 \times \boldsymbol{u}_3)$, and $(\boldsymbol{u}_1 \times \boldsymbol{u}_2)$ result from

$$\begin{aligned}\boldsymbol{u}_2 \times \boldsymbol{u}_3 &= \begin{bmatrix} \boldsymbol{u}_X & \boldsymbol{u}_Y & \boldsymbol{u}_Z \\ 0.51174 & -0.45275 & 0.73016 \\ -0.23284 & 0.059133 & 0.97072 \end{bmatrix} \\ &= -0.48267\,\boldsymbol{u}_X - 0.66677\,\boldsymbol{u}_Y - 0.075158\,\boldsymbol{u}_Z\end{aligned}$$

$$\boldsymbol{u}_1 \times \boldsymbol{u}_3 = \begin{bmatrix} \boldsymbol{u}_X & \boldsymbol{u}_Y & \boldsymbol{u}_Z \\ 0.80496 & -0.52274 & -0.28067 \\ -0.23284 & 0.059133 & 0.97072 \end{bmatrix}$$
$$= -0.49085\,\boldsymbol{u}_X - 0.71604\,\boldsymbol{u}_Y - 0.074117\,\boldsymbol{u}_Z$$

$$\boldsymbol{u}_1 \times \boldsymbol{u}_2 = \begin{bmatrix} \boldsymbol{u}_X & \boldsymbol{u}_Y & \boldsymbol{u}_Z \\ 0.80496 & -0.52274 & -0.28067 \\ 0.51174 & -0.45275 & 0.73016 \end{bmatrix}$$
$$= -0.50876\,\boldsymbol{u}_X - 0.73138\,\boldsymbol{u}_Y - 0.096934\,\boldsymbol{u}_Z$$

The scalar products D_0, D_{11}, D_{12}, D_{13}, D_{21}, D_{22}, D_{23}, D_{31}, D_{32}, and D_{33} are

$$D_0 = \boldsymbol{u}_1 \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3) = 0.80496 \times (-0.48267) - 0.522754 \times (-0.66677) - 0.28067 \times (-0.075158) = -0.018881$$
$$D_{11} = \boldsymbol{r}_{E1} \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3) = 3011.7 \times (-0.48267) - 3945.1 \times (-0.66677) + 3992.5 \times (-0.075158) = 876.75\,\text{km}$$
$$D_{12} = \boldsymbol{r}_{E1} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) = 3011.7 \times (-0.49085) - 3945.1 \times (-0.71604) + 3992.5 \times (-0.074117) = 1050.6\ \text{km}$$
$$D_{13} = \boldsymbol{r}_{E1} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) = 3011.7 \times (-0.50876) - 3945.1 \times (-0.73138) + 3992.5 \times (-0.096934) = 966.13\ \text{km}$$
$$D_{21} = \boldsymbol{r}_{E2} \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3) = 3063.5 \times (-0.48267) - 3905.0 \times (-0.66677) + 3992.5 \times (-0.075158) = 825.01\,\text{km}$$
$$D_{22} = \boldsymbol{r}_{E2} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) = 3063.5 \times (-0.49085) - 3905.0 \times (-0.71604) + 3992.5 \times (-0.074117) = 996.51\ \text{km}$$
$$D_{23} = \boldsymbol{r}_{E2} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) = 3063.5 \times (-0.50876) - 3905.0 \times (-0.73138) + 3992.5 \times (-0.096934) = 910.44\ \text{km}$$
$$D_{31} = \boldsymbol{r}_{E3} \cdot (\boldsymbol{u}_2 \times \boldsymbol{u}_3) = 3114.5 \times (-0.48267) - 3864.5 \times (-0.66677) + 3992.5 \times (-0.075158) = 773.39\,\text{km}$$
$$D_{32} = \boldsymbol{r}_{E3} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_3) = 3114.5 \times (-0.49085) - 3864.5 \times (-0.71604) + 3992.5 \times (-0.074117) = 942.47\ \text{km}$$
$$D_{33} = \boldsymbol{r}_{E3} \cdot (\boldsymbol{u}_1 \times \boldsymbol{u}_2) = 3114.5 \times (-0.50876) - 3864.5 \times (-0.73138) + 4069.057 \times (-0.096934) = 854.88\ \text{km}$$

The quantities A and B are computed as follows

$$\tau_3/\tau = 180/361 = 0.49861$$
$$\tau_1/\tau = -181/361 = -0.50139$$
$$\tau_3\left(\tau^2 - \tau_3^2\right) = 180 \times \left(361^2 - 180^2\right) = 1.7626 \times 10^7$$
$$\tau_1\left(\tau^2 - \tau_1^2\right) = -60 \times \left(120^2 - 60^2\right) = -1.7658 \times 10^7$$
$$A = (-D_{12}\tau_3/\tau + D_{22} + D_{32}\tau_1/\tau)/D_0 = [-1050.6 \times 0.49861 + 996.51 + 942.47 \times (-0.50139)]/(-0.018881) = -6.6363\,\text{km}$$
$$B = \left[-D_{12}\tau_3\left(\tau^2 - \tau_3^2\right) + D_{32}\tau_1\left(\tau^2 - \tau_1^2\right)\right]/[6D_0\tau] = \left[-1050.6 \times 1.7626 \times 10^7 + 942.47 \times (-1.7658 \times 10^7)\right]/[6 \times (-0.018881) \times 361] = 8.5974 \times 10^8\,\text{km}\,\text{s}^2$$

The quantities E and r_{E2}^2 result from

$$E = \boldsymbol{r}_{E2} \cdot \boldsymbol{u}_2 = 3063.5 \times 0.51174 - 3905.0 \times (-0.45275) + 3992.5 \times 0.73016 = 6250.9\,\text{km}$$
$$r_{E2}^2 = \boldsymbol{r}_{E2} \cdot \boldsymbol{r}_{E2} = 3063.5^2 + 3905.0^2 + 3992.5^2 = 4.0574 \times 10^7\,\text{km}^2$$

The coefficients a, b, and c of the polynomial $x^8 + ax^6 + bx^3 + c$ result from

$$a = -\left(r_{E2}^2 + 2EA + A^2\right) = -\left[4.0574 \times 10^7 + 2 \times 6250.9 \times (-6.6363) + (-6.6363)^2\right] = -4.0491 \times 10^7\,\text{km}^2$$
$$b = -2\mu B(E + A) = -2 \times 398600.4 \times 8.5974 \times 10^8 \times (6250.9 - 6.6363) = -4.2797 \times 10^{18}\,\text{km}^5$$
$$c = -\mu^2 B^2 = -(398{,}600.4)^2 \times \left(8.5974 \times 10^8\right)^2 = -0.11744 \times 10^{30}\,\text{km}^8$$

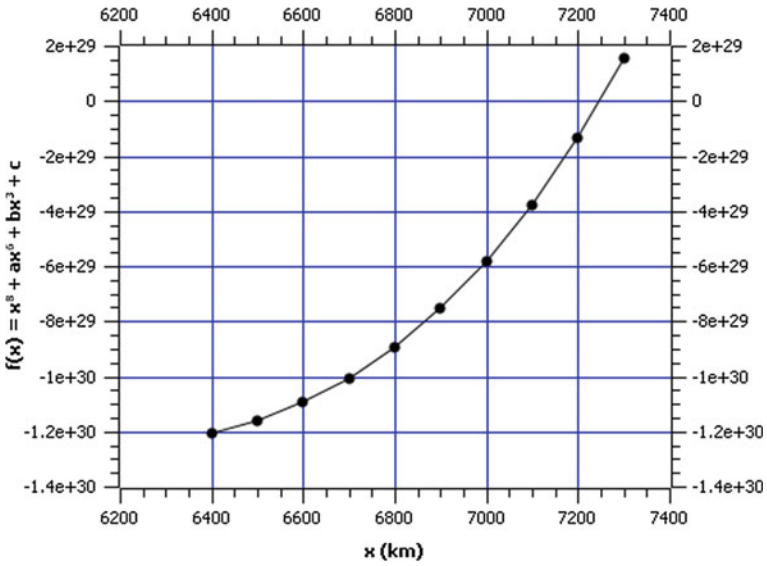
We search a value of x such that the following function

$$f(x) = x^8 + ax^6 + bx^3 + c$$

should be equal to zero in a given interval $x_{\min} \leq x \leq x_{\max}$. This search is limited to the values of x (where $x \equiv r_2$) which are physically meaningful. Consequently, x cannot be negative, or smaller than the mean radius of the Earth (6371 km). In addition, since the values computed above of the three coefficients a, b, and c are, all of them, negative, then the equation $f(x) = x^8 + ax^6 + bx^3 + c = 0$ has only one positive root. To search this root, we first evaluate $f(x)$ for x ranging from 6400 km

to 9400 km, with a step size of 100 km. The results of this evaluation are shown below.

$x \times 10^3$ (km)	$f(x) \times 10^{30}$	$x \times 10^3$ (km)	$f(x) \times 10^{30}$
6.400	−1.2071	6.900	−0.75509
6.500	−1.1601	7.000	−0.58431
6.600	−1.0941	7.100	−0.37856
6.700	−1.0067	7.200	−0.13376
6.800	−0.89472	7.300	0.15461



The plot shows that the physically meaningful root of $f(x) = 0$ is roughly the midpoint of the interval $7200 \leq x \leq 7300$ km. Therefore, the search of the root of interest is confined to this interval. As shown in the preceding table, we find

$$\begin{aligned} f(7200) &= 7200^8 - 4.0491 \times 10^7 \times 7200^6 - 4.2797 \times 10^{18} \times 7200^3 \\ &\quad - 0.11744 \times 10^{30} = -0.13376 \times 10^{30} (<0) \\ f(7300) &= 7300^8 - 4.0491 \times 10^7 \times 7300^6 - 4.2797 \times 10^{18} \times 7300^3 \\ &\quad - 0.11744 \times 10^{30} = 0.15461 \times 10^{30} (>0) \end{aligned}$$

Since the condition $f(7200)f(7300) < 0$ is satisfied, we search a zero of $f(x)$ by means of Müller's method of parabolic interpolation, which has been described in Chap. 1, Sects. 5 and 8. At the midpoint $x_0 = 7250$ km of the interval we also find

$$f(7250) = 7250^8 - 4.0491 \times 10^7 \times 7240^6 - 4.2797 \times 10^{18} \times 7250^3 - 0.11744 \times 10^{30} = 0.00469 \times 10^{30}$$

Then we set

$$\begin{aligned} x_2 &= 7200\,\text{km} & f_2 &\equiv f(x_2) = -0.13376 \times 10^{30} \\ x_0 &= 7250\,\text{km} & f_0 &\equiv f(x_0) = 0.00469 \times 10^{30} \\ x_1 &= 7300\,\text{km} & f_1 &\equiv f(x_1) = 0.15461 \times 10^{30} \end{aligned}$$

and

$$\begin{aligned} h_1 &= x_1 - x_0 = 7300 - 7250 = 50\,\text{km} \\ h_2 &= x_0 - x_2 = 7250 - 7200 = 50\,\text{km} \\ \gamma &= h_2/h_1 = 50/50 = 1 \end{aligned}$$

and compute the coefficients

$$\begin{aligned} A &= [\gamma f_1 - f_0(1+\gamma) + f_2]/[\gamma h_1^2(1+\gamma)] = [1 \times 0.15461 \times 10^{30} - 0.00469 \times 10^{30} \times (1 + 1) - 0.13376 \times 10^{30}]/[1 \times 50^2 \times (1 + 1)] = 0.000002294 \times 10^{30} \\ B &= (f_1 - f_0 - Ah_1^2)/h_1 = [0.15461 \times 10^{30} - 0.00469 \times 10^{30} - 0.000002294 \times 10^{30} \times 50^2]/50 = 0.0028837 \times 10^{30} \\ C &= f_0 = 0.00469 \times 10^{30} \end{aligned}$$

of the interpolating parabola $f(x) = A(x - x_0)^2 + B(x - x_0) + C$.

We compute the estimated root of $f(x) = 0$, as follows

$$x = x_0 - \frac{2C}{B \pm (B^2 - 4AC)^{\frac{1}{2}}}$$

where in the present case the plus sign in front of the square root takes effect, because the value of B is greater than zero. Thus,

$$x = 7250 - 2 \times 0.00469 \times 10^{30} / \left\{0.0028837 \times 10^{30} + [(0.0028837 \times 10^{30})^2 - 4 \times 0.000002294 \times 10^{30} \times 0.00469 \times 10^{30}]^{1/2}\right\} = 7248.4\,\text{km}$$

This value, substituted into $f(x) = x^8 + ax^6 + bx^3 + c$, yields

$$f(7248.4) = 7248.4^8 - 4.0491 \times 10^7 \times 7248.4^6 - 4.2797 \times 10^{18} \times 7248.4^3 - 0.11744 \times 10^{30} = 0.00009 \times 10^{30}$$

Now, since 7248.4 is less than 7250, then we take 7200, 7250 and 7248.4 for the next step. At the same time, we reset the subscripts 0, 1 and 2, as follows

$$\begin{array}{ll} x_2 = 7200.0\,\text{km} & f_2 \equiv \mathrm{f}(x_2) = -0.13376 \times 10^{30} \\ x_0 = 7248.4\,\text{km} & f_0 \equiv \mathrm{f}(x_0) = 0.00009 \times 10^{30} \\ x_1 = 7250.0\,\text{km} & f_1 \equiv \mathrm{f}(x_1) = 0.00469 \times 10^{30} \end{array}$$

Now we compute again

$$\begin{aligned} h_1 &= x_1 - x_0 = 7250.0 - 7248.4 = 1.6\,\text{km} \\ h_2 &= x_0 - x_2 = 7248.4 - 7200.0 = 48.4\,\text{km} \\ \gamma &= h_2/h_1 = 48.4/1.6 = 30.25 \end{aligned}$$

and the coefficients A, B, and C of the interpolating parabola, as follows

$$\begin{aligned} A &= [\gamma f_1 - f_0(1+\gamma) + f_2]/[\gamma h_1^2(1+\gamma)] = [30.25 \times 0.00469 \times 10^{30} - 0.00009 \times 10^{30} \\ &\quad \times (1 + 30.25) - 0.13376 \times 10^{30}]/[30.25 \times 1.6^2 \times (1 + 30.25)] = 0.00000219 \times 10^{30} \\ B &= (f_1 - f_0 - Ah_1^2)/h_1 = (0.00469 \times 10^{30} - 0.00009 \times 10^{30} \\ &\quad - 0.00000219 \times 10^{30} \times 1.6^2)/1.6 = 0.0028715 \times 10^{30} \\ C &= f_0 = 0.00009 \times 10^{30} \end{aligned}$$

Thus, the estimated root is

$$x = 7248.4 - 2 \times 0.00009 \times 10^{30} / \left\{ 0.0028715 \times 10^{30} + [(0.0028715 \times 10^{30})^2 - 4 \times 0.00000219 \times 10^{30} \times 0.00009 \times 10^{30}]^{1/2} \right\} = 7248.4\,\text{km}$$

Since this value is the same as that computed in the preceding step, then we take 7248.4 km as the correct root, within the chosen limits of accuracy, of

$$f(x) = x^8 + ax^6 + bx^3 + c = 0$$

This means that we take $x \equiv r_2$ = 7248.4 km in this preliminary approximation.

Now we compute the ranges ρ_1, ρ_2, and ρ_3, as follows

$$\rho_1 = \left[\frac{6\left(D_{31}\frac{\tau_1}{\tau_3} + D_{21}\frac{\tau}{\tau_3}\right)r_2^3 + \mu D_{31}(\tau^2 - \tau_1^2)\frac{\tau_1}{\tau_3}}{6r_2^3 + \mu(\tau^2 - \tau_3^2)} - D_{11}\right]\frac{1}{D_0}$$

$$\rho_2 = A + \frac{\mu B}{r_2^3}$$

$$\rho_3 = \left[\frac{6\left(D_{13}\frac{\tau_3}{\tau_1} - D_{23}\frac{\tau}{\tau_1}\right)r_2^3 + \mu D_{13}(\tau^2 - \tau_3^2)\frac{\tau_3}{\tau_1}}{6r_2^3 + \mu(\tau^2 - \tau_3^2)} - D_{33}\right]\frac{1}{D_0}$$

In the present case, there results

$$\begin{aligned}
&\tau_1/\tau_3 = -181/180 = -1.0056\\
&\tau/\tau_3 = 361/180 = 2.0056\\
&\tau/\tau_1 = 361/(-181) = -1.9945\\
&\left(\tau^2 - \tau_1^2\right)\tau_1/\tau_3 = [361^2 - (-181^2)] \times (-181/180) = -98102\\
&\left(\tau^2 - \tau_3^2\right)\tau_3/\tau_1 = (361^2 - 180^2) \times 180/(-181) = -97380\\
&\tau^2 - \tau_3^2 = 361^2 - 180^2 = 97921
\end{aligned}$$

and consequently

$$\begin{aligned}
\rho_1 &= \{[6 \times (773.39 \times (-1.0056) + 825.01 \times 2.0056) \times 7248.4^3 + 398600.4 \times 773.39\\
&\quad \times(-98102)]/[6 \times 7248.4^3 + 398600.4 \times 97921] - 876.75\}/(-0.018881)\\
&= 1460.3\,\text{km}\\
\rho_2 &= -6.6363 + 398600.4 \times 8.5974 \times 10^8/7248.4^3 = 893.23\,\text{km}\\
\rho_3 &= \{[6 \times (966.13/(-1.0056) - 910.44 \times (-1.9945)) \times 7248.4^3 + 398600.4 \times 966.13\\
&\quad \times(-97380)]/[6 \times 7248.4^3 + 398600.4 \times 97921] - 854.88\}/(-0.018881)\\
&= 1602.4\,\text{km}
\end{aligned}$$

We compute $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$ as follows

$$\begin{aligned}
\boldsymbol{r}_1 &= \boldsymbol{r}_{E1} + \rho_1\,\boldsymbol{u}_1 = 3011.7\,\boldsymbol{u}_X - 3945.1\,\boldsymbol{u}_Y + 3992.5\,\boldsymbol{u}_Z\\
&\quad + 1460.3 \times (0.80496\,\boldsymbol{u}_X - 0.52275\,\boldsymbol{u}_Y - 0.28067\,\boldsymbol{u}_Z)\\
&= 4187.2\,\boldsymbol{u}_X - 4708.5\,\boldsymbol{u}_Y + 3582.6\,\boldsymbol{u}_Z
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{r}_2 &= \boldsymbol{r}_{E2} + \rho_2\,\boldsymbol{u}_2 = 3063.5\,\boldsymbol{u}_X - 3905.0\,\boldsymbol{u}_Y + 3992.5\,\boldsymbol{u}_Z\\
&\quad + 893.23 \times (0.51174\,\boldsymbol{u}_X - 0.45275\,\boldsymbol{u}_Y + 0.73016\,\boldsymbol{u}_Z)\\
&= 3520.6\,\boldsymbol{u}_X - 4309.4\,\boldsymbol{u}_Y + 4644.7\,\boldsymbol{u}_Z
\end{aligned}$$

$$\begin{aligned}\boldsymbol{r}_3 = \boldsymbol{r}_{E3} + \rho_3 \boldsymbol{u}_3 = {} & 3114.5\,\boldsymbol{u}_X - 3864.5\,\boldsymbol{u}_Y + 3992.5\,\boldsymbol{u}_Z \\ & + 1602.4 \times (-0.23284\,\boldsymbol{u}_X + 0.059133\,\boldsymbol{u}_Y + 0.97072\,\boldsymbol{u}_Z) \\ = {} & 2741.4\,\boldsymbol{u}_X - 3769.7\,\boldsymbol{u}_Y + 5548.0\,\boldsymbol{u}_Z\end{aligned}$$

The Lagrangian coefficients f_1, f_3, g_1, and g_3 result from

$$f_1 \approx 1 - \frac{1}{2}\left(\mu/r_2^3\right)\tau_1^2 = 1 - \frac{1}{2} \times (398600.4/7248.4^3) \times (-181)^2 = 0.98285$$

$$f_3 \approx 1 - \frac{1}{2}\left(\mu/r_2^3\right)\tau_3^2 = 1 - \frac{1}{2} \times (398600.4/7248.4^3) \times 180^2 = 0.98304$$

$$g_1 \approx \tau_1 - \frac{1}{6}\left(\mu/r_2^3\right)\tau_1^3 = -181 - \frac{1}{6} \times 398600.4 \times (-181/7248.4)^3 = -179.97$$

$$g_3 \approx \tau_3 - \frac{1}{6}\left(\mu/r_2^3\right)\tau_3^3 = -180 - \frac{1}{6} \times 398600.4 \times (180/7248.4)^3 = 178.98$$

Now, in order to compute v_2, we evaluate

$$f_1/(f_1 g_3 - g_1 f_3) = 0.98285/[0.98285 \times 178.98 - (-179.97) \times 0.98304] = 2.7856 \times 10^{-3}$$
$$f_3/(f_1 g_3 - g_1 f_3) = 0.98304/[0.98285 \times 178.98 - (-179.97) \times 0.98304] = 2.7862 \times 10^{-3}$$

$$\begin{aligned}\boldsymbol{v}_2 = {} & [f_1/(f_1 g_3 - g_1 f_3)]\boldsymbol{r}_3 - [f_3/(f_1 g_3 - g_1 f_3)]\boldsymbol{r}_1 = 0.0027856 \times (2741.4\,\boldsymbol{u}_X - 3769.7\,\boldsymbol{u}_Y \\ & + 5548.0\,\boldsymbol{u}_Z) - 0.0027862 \times (4187.2\,\boldsymbol{u}_X - 4708.5\,\boldsymbol{u}_Y + 3582.6\,\boldsymbol{u}_Z) \\ = {} & -4.0299\,\boldsymbol{u}_X + 2.6179\,\boldsymbol{u}_Y + 5.4727\,\boldsymbol{u}_Z\end{aligned}$$

In summary, the preliminary approximation to the position and velocity vectors of the COBE satellite, observed at time t_2, is

$$\begin{aligned}\boldsymbol{r}_2 = {} & 3520.6\,\boldsymbol{u}_X - 4309.4\,\boldsymbol{u}_Y + 4644.7\,\boldsymbol{u}_Z \\ \boldsymbol{v}_2 = {} & -4.0299\,\boldsymbol{u}_X + 2.6179\,\boldsymbol{u}_Y + 5.4727\,\boldsymbol{u}_Z\end{aligned}$$

This completes the first part of the computation. The second part is meant to compute more accurate values of the vectors $\boldsymbol{r}_2$ and $\boldsymbol{v}_2$ than those computed in the first part, as will be shown below.

First iteration.

We compute the magnitudes r_2 and v_2 of the vectors $\boldsymbol{r}_2$ and $\boldsymbol{v}_2$, as follows

$$r_2 = (\boldsymbol{r}_2 \cdot \boldsymbol{r}_2)^{\frac{1}{2}} = \left[3520.6^2 + (-4309.4)^2 + 4644.7^2\right]^{\frac{1}{2}} = 7248.4\,\text{km}$$

$$v_2 = (\boldsymbol{v}_2 \cdot \boldsymbol{v}_2)^{\frac{1}{2}} = \left[(-4.0299)^2 + 2.6179^2 + 5.4727^2\right]^{\frac{1}{2}} = 53.044^{\frac{1}{2}}\,\text{km/s}$$

Now, we compute α, the inverse of the major semi-axis a of the trajectory of the observed object, by means of the vis-viva integral $v^2/\mu = 2/r - 1/a$, as follows

$$\alpha \equiv \frac{1}{a} = \frac{2}{r_2} - \frac{v_2^2}{\mu} = \frac{2}{7248.4} - \frac{53.044}{398600.4} = 1.4285 \times 10^{-4}\,\mathrm{km}^{-1}$$

We compute the radial component of v_2, as follows

$$\begin{aligned} v_{2r} &= \frac{\boldsymbol{r}_2 \cdot \boldsymbol{r}_2}{r_2} = \frac{(-4.0299) \times 3520.6 + 2.6179 \times (-4309.4) + 5.4727 \times 4644.7}{7248.4} \\ &= -0.0069249\,\mathrm{km/s} \end{aligned}$$

Then we write Kepler's equation in universal variables (see Sect. 1.7) at times t_1 and t_3 respectively, as follows

$$\begin{aligned} \mu^{\frac{1}{2}}(t_1 - t_2)\left(\frac{r_2 v_{2r}}{\mu^{\frac{1}{2}}}\right)\chi_1^2 C\left(\alpha\chi_1^2\right) + (1 - \alpha r_2)\chi_1^3 S\left(\alpha\chi_1^2\right) + r_2\chi_1 \\ \mu^{\frac{1}{2}}(t_3 - t_2)\left(\frac{r_2 v_{2r}}{\mu^{\frac{1}{2}}}\right)\chi_3^2 C\left(\alpha\chi_3^2\right) + (1 - \alpha r_2)\chi_3^3 S\left(\alpha\chi_3^2\right) + r_2\chi_3 \end{aligned}$$

where C and S are the Stumpff functions (see Sect. 1.7), which are defined as follows

$$C(z) = \begin{cases} \left[1 - \cos\left(z^{1/2}\right)\right]/z & \text{for } z > 0 \\ 1/2 & \text{for } z = 0 \\ \left[\cosh\left(-z^{1/2}\right) - 1\right]/(-z) & \text{for } z < 0 \end{cases}$$

$$S(z) = \begin{cases} \left[z^{1/2} - \sin\left(z^{1/2}\right)\right]/z^{3/2} & \text{for } z > 0 \\ 1/6 & \text{for } z = 0 \\ \left[\sinh\left(-z^{1/2}\right) - \left(-z^{1/2}\right)\right]/(-z)^{3/2} & \text{for } z < 0 \end{cases}$$

$z = \alpha\chi^2$, and χ_1 and χ_3 are the universal variables to be determined. Since

$$\begin{aligned} t_1 - t_2 = \tau_1 &= -181 \\ t_3 - t_2 = \tau_3 &= 180 \end{aligned}$$

then Kepler's equations at times t_1 and t_3 become, respectively,

$$\begin{aligned} 398600.4^{1/2} \times (-181) = &\left[7248.4 \times (-0.0069249)/398600.4^{1/2}\right]\chi_1^2 C(1.4285 \times 10^{-4}\chi_1^2) \\ &+ (1 - 1.4285 \times 10^{-4} \times 7248.4)\,\chi_1^3 S(1.4285 \times 10^{-4}\chi_1^2) \\ &+ 7248.4\,\chi_1 \end{aligned}$$

$$398600.4^{1/2} \times 180 = [7248.4 \times (-0.0069249)/398600.4^{1/2}]\chi_3^2 C(1.4285 \times 10^{-4}\chi_3^2) + (1 - 1.4285 \times 10^{-4} \times 7248.4)\chi_3^3 S(1.4285 \times 10^{-4}\chi_3^2) + 7248.4\,\chi_3$$

The preceding equations, after simplification, become, respectively,

$$-114274 = -0.079504\,\chi_1^2 C(1.4285 \times 10^{-4}\chi_1^2) - 0.035434\,\chi_1^3 \times S(1.4285 \times 10^{-4}\chi_1^2) + 7248.4\,\chi\chi_1$$

$$113643 = -0.079504\,\chi_3^2 C(1.4285 \times 10^{-4}\chi_3^2) - 0.035434\,\chi_3^3 \times S(1.4285 \times 10^{-4}\chi_1^2) + 7248.4\chi_3$$

We solve iteratively the two equations written above. To this end, an initial estimate of χ_1 and χ_3 is provided by the following formula suggested by Chobotov [18]:

$$\mu^{\frac{1}{2}}|\alpha|\Delta t$$

As to the first equation, there results

$$\mu^{\frac{1}{2}}|\alpha|\Delta t = 398600.4^{\frac{1}{2}} \times 1.4285 \times 10^{-4} \times (-181) = -16.324\,\mathrm{km}^{\frac{1}{2}}$$

Therefore, an estimate of the unknown value of χ_1 is taken tentatively in the interval $-18.0 \leq \chi_1 \leq -14.0$ around -16.324. We ascertain whether the following function

$$f(\chi) = -114274 + 0.079504\,\chi^2 C(1.4285 \times 10^{-4}\chi^2) + 0.035434\,\chi^3 \times S(1.4285 \times 10^{-4}\chi^2) - 7248.4\,\chi$$

has values of opposite signs at the endpoints of this interval. We find

$$f(-18.0) = -114274 + 0.079504 \times (-18.0)^2 \times C[1.4285 \times 10^{-4} \times (-18.0)^2] + 0.035434 \times (-18.0)^3 \times S[1.4285 \times 10^{-4} \times (-18.0)^2] - 7248.4 \times (-18.0) = 16176\,(>0)$$

$$f(-14.0) = -114274 + 0.079504 \times (-14.0)^2 \times C[1.4285 \times 10^{-4} \times (-14.0)^2] + 0.035434 \times (-14.0)^3 \times S[1.4285 \times 10^{-4} \times (-14.0)^2] - 7248.4 \times (-14.0) = -12805(<0)$$

Since the condition $f(-18.0)f(-14.0) < 0$ is satisfied, we search a zero of $f(\chi)$ by means of Müller's method of parabolic interpolation, which has been described in Sects. 1.5 and 1.8. Consequently, we choose arbitrarily a value χ_0 falling between the endpoints −18.0 and −14.0. By choosing $\chi_0 = -16.0$, the corresponding $f(\chi_0)$ is

$$\begin{aligned} f(-16.0) &= -114274 + 0.079504 \times (-16.0)^2 \times C[1.4285 \times 10^{-4} \times (-16.0)^2] \\ &\quad + 0.035434 \times (-16.0)^3 \times S[1.4285 \times 10^{-4} \times (-16.0)^2] - 7248.4 \times (-16.0) \\ &= 1686.4 \end{aligned}$$

Then we set

$$\begin{aligned} \chi_2 &= -18.0\,\mathrm{km}^{1/2} \quad f_2 \equiv f(\chi_2) = 16176 \\ \chi_0 &= -16.0\,\mathrm{km}^{1/2} \quad f_0 \equiv f(\chi_0) = 1686.4 \\ \chi_1 &= -14.0\,\mathrm{km}^{1/2} \quad f_1 \equiv f(\chi_1) = -12805 \end{aligned}$$

It is to be noted that here χ_2 and χ_1 are the endpoints of the current interval of search for the unknown χ and that they have nothing to do with the number of observations. Now we set

$$\begin{aligned} h_1 &= \chi_1 - \chi_0 = -14.0 - (-16.0) = 2.0\,\mathrm{km}^{1/2} \\ h_2 &= \chi_0 - \chi_2 = -16.0 - (-18.0) = 2.0\,\mathrm{km}^{1/2} \\ \gamma &= h_2/h_1 = 2.0/2.0 = 1.0 \end{aligned}$$

and compute the coefficients

$$\begin{aligned} A &= [\gamma f_1 - f_0(1+\gamma) + f_2]/[\gamma h_1^2(1+\gamma)] = [1.0 \times (-12805) - 1686.4 \times (1 \\ &\quad + 1.0) + 16176]/[1.0 \times 2.0^2 \times (1 + 1.0)] = -0.225 \\ B &= (f_1 - f_0 - A h_1^2)/h_1 = [-12805 - 1686.4 - (-0.225) \times 2.0^2]/2.0 = -7245.3 \\ C &= f_0 = 1686.4 \end{aligned}$$

of the interpolating parabola $f(\chi) = A(\chi - \chi_0)^2 + B(\chi - \chi_0) + C$.

This done, we compute the estimated root of $f(\chi) = 0$ as follows

$$\chi = \chi_0 - \frac{2C}{B \pm (B^2 - 4AC)^{\frac{1}{2}}}$$

where in the present case the minus sign in front of the square root takes effect, because the value of B is less than zero. Thus, there results

$$\begin{aligned} \chi &= -16.0 - 2 \times 1686.4/[-7245.3 - (7245.3^2 + 4 \times 0.225 \times 1686.4)^{1/2}] \\ &= -15.767\,\mathrm{km}^{1/2} \end{aligned}$$

This value, substituted into $f(\chi)$, yields

$$f(-15.767) = -114274 + 0.079504 \times (-15.767)^2 \times C\,[1.4285 \times 10^{-4} \times (-15.767)^2] + 0.035434 \times (-15.767)^3 \times S\,[1.4285 \times 10^{-4} \times (-15.767)^2] - 7248.4 \times (-15.767) = -1.7312$$

Since −15.767 is greater than −16.0, we take −16.0, −14.0, and −15.767 for the next step; at the same time, we reset the subscripts as follows

$$\begin{array}{ll} \chi_2 = -16.0\,\mathrm{km}^{1/2} & f_2 \equiv f(\chi_2) = 1686.4 \\ \chi_0 = -15.767\,\mathrm{km}^{1/2} & f_0 \equiv f(\chi_0) = -1.7312 \\ \chi_1 = -14.0\,\mathrm{km}^{1/2} & f_1 \equiv f(\chi_1) = -12805 \end{array}$$

Now we set

$$\begin{aligned} h_1 &= \chi_1 - \chi_0 = -14.0 - (-15.767) = 1.767\,\mathrm{km}^{1/2} \\ h_2 &= \chi_0 - \chi_2 = -15.767 - (-16.0) = 0.233\,\mathrm{km}^{1/2} \\ \gamma &= h_2/h_1 = 0.233/1.767 = 0.13186 \end{aligned}$$

and compute the coefficients

$$\begin{aligned} A &= [\gamma f_1 - f_0(1+\gamma) + f_2]/\left[\gamma h_1^2(1+\gamma)\right] = [0.13186 \times (-12805) - (-1.7312) \times (1 + 0.13186) + 1686.4]/[0.13186 \times 1.767^2 \times (1 + 0.13186)] = -0.23139 \\ B &= \left(f_1 - f_0 - Ah_1^2\right)/h_1 = [-12805 - (-1.7312) - (-0.23139) \times 1.767^2]/1.767 = -7245.4 \\ C &= f_0 = -1.7312 \end{aligned}$$

of the interpolating parabola.

The estimated root χ of $f(\chi) = 0$ results from

$$\chi = \chi_0 - 2C/\left[B - \left(B^2 - 4AC\right)^{1/2}\right] = -15.767 - 2 \times (-1.7312)/[-7245.4 - (7245.4^2 - 4 \times 0.23139 \times 1.7312)^{1/2}] = -15.767\,\mathrm{km}^{1/2}$$

Since this value is the same as that computed previously, within the chosen accuracy, we take it as correct. In other words, $\chi_1 = -15.767\ \mathrm{km}^{1/2}$ is taken as the root of Kepler's equation

$$-114274 = -0.079504\chi_1^2 C(1.4285 \times 10^{-4}\chi_1^2) - 0.035434\,\chi_1^3 \times S(1.4285 \times 10^{-4}\chi_1^2) + 7248.4\chi_1$$

By using the same iterative method, we search now the root χ_3 of the second equation, that is,

$$113643 = -0.079504\chi_3^2 \times C(1.4285 \times 10^{-4}\chi_3^2) - 0.035434\chi_3^3 \times S(1.4285 \times 10^{-4}\chi_3^2) + 7248.4\chi_3$$

To this end, it is necessary to find an interval of χ where the following function

$$f(\chi) = 113643 + 0.079504\chi^2 \times C(1.4285 \times 10^{-4}\chi^2) + 0.035434\chi^3 \times S(1.4285 \times 10^{-4}\chi^2) - 7248.4\chi$$

changes sign. We try $14.0 \leq \chi \leq 18.0$ and find the following values at the endpoints:

$$f(14.0) = 113643 + 0.079504 \times 14.0^2 \times C(1.4285 \times 10^{-4} \times 14.0^2) + 0.035434 \times 14.0^3 \times S(1.4285 \times 10^{-4} \times 14.0^2) - 7248.4 \times 14.0 = 12189\ (>0)$$

$$f(18.0) = 113643 + 0.079504 \times 18.0^2 \times C(1.4285 \times 10^{-4} \times 18.0^2) + 0.035434 \times 18.0^3 \times S(1.4285 \times 10^{-4} \times 18.0^2) - 7248.4 \times 18.0 = -16781\ (<0)$$

Since the condition $f(14.0)f(18.0) < 0$ is satisfied, we choose arbitrarily a value χ_0 falling between these endpoints. Choosing $\chi_0 = 16.0$, the corresponding $f(\chi_0)$ is

$$f(16.0) = 113643 + 0.079504 \times 16.0^2 \times C(1.4285 \times 10^{-4} \times 16.0^2) + 0.035434 \times 16.0^3 \times S(1.4285 \times 10^{-4} \times 16.0^2) - 7248.4 \times 16.0 = -2297.1$$

Now we set

$$\begin{aligned} \chi_2 &= 14.0\,\mathrm{km}^{1/2} & f_2 &\equiv f(\chi_2) = 12189 \\ \chi_0 &= 16.0\,\mathrm{km}^{1/2} & f_0 &\equiv f(\chi_0) = -2297.1 \\ \chi_1 &= 18.0\,\mathrm{km}^{1/2} & f_1 &\equiv f(\chi_1) = -16781 \end{aligned}$$

Again, the subscripts used here refer only to the iterations. We also set

$$\begin{aligned} h_1 &= \chi_1 - \chi_0 = 18.0 - 16.0 = 2.0\,\mathrm{km}^{1/2} \\ h_2 &= \chi_0 - \chi_2 = 16.0 - 14.0 = 2.0\,\mathrm{km}^{1/2} \\ \gamma &= h_2/h_1 = 2.0/2.0 = 1 \end{aligned}$$

and compute the coefficients

$$A = [\gamma f_1 - f_0(1+\gamma) + f_2]/[\gamma h_1^2(1+\gamma)] = [1 \times (-16781) - (-2297.1) \times (1+1) + 12189]/[1 \times 2.0^2 \times (1+1)] = 0.275$$
$$B = (f_1 - f_0 - Ah_1^2)/h_1 = [-16781 - (-2297.1) - 0.275 \times 2.0^2]/2.0 = -7242.5$$
$$C = f_0 = -2297.1$$

of the interpolating parabola $f(\chi) = A(\chi - \chi_0)^2 + B(\chi - \chi_0) + C$.

Now, we compute the estimated root of $f(\chi) = 0$ as follows

$$\chi = \chi_0 - 2C/\left[B - \left(B^2 - 4AC\right)^{1/2}\right] = 16.0 - 2 \times (-2297.1)/[-7242.5 - (7242.5^2 + 4 \times 0.275 \times 2297.1)^{1/2}] = 15.683$$

This value, substituted into $f(\chi)$, yields

$$f(15.683) = 113643 + 0.079504 \times 15.683^2 \times C(1.4285 \times 10^{-4} \times 15.683^2) + 0.035434 \times 15.683^3 \times S(1.4285 \times 10^{-4} \times 15.683^2) - 7248.4 \times 15.683 = -1.1684$$

Since 15.683 is less than 16.0, we take 14.0, 16.0 and 15.683 for the next step. At the same time, the subscripts are reset as follows

$$\begin{array}{ll} \chi_2 = 14.0\,\mathrm{km}^{1/2} & f_2 \equiv f(\chi_2) = 12189 \\ \chi_0 = 15.683\,\mathrm{km}^{1/2} & f_0 \equiv f(\chi_0) = -1.1684 \\ \chi_1 = 16.0\,\mathrm{km}^{1/2} & f_1 \equiv f(\chi_1) = -2297.1 \end{array}$$

Now we set

$$h_1 = \chi_1 - \chi_0 = 16.0 - 15.683 = 0.317\,\mathrm{km}^{1/2}$$
$$h_2 = \chi_0 - \chi_2 = 15.683 - 14.0 = 1.683\,\mathrm{km}^{1/2}$$
$$\gamma = h_2/h_1 = 1.683/0.317 = 5.3091$$

and compute the coefficients

$$A = [\gamma f_1 - f_0(1+\gamma) + f_2]/[\gamma h_1^2(1+\gamma)] = [5.3091 \times (-2297.1) - (-1.1684) \times (1+5.3091) + 12189]/[5.3091 \times 0.317^2 \times (1+5.3091)] = 0.24895$$
$$B = (f_1 - f_0 - Ah_1^2)/h_1 = [-2297.1 - (-1.1684) - 0.24895 \times 0.317^2]/0.317 = -7242.8$$
$$C = f_0 = -1.1684$$

of the interpolating parabola.

The estimated root χ of $f(\chi) = 0$ results from

$$\chi = \chi_0 - 2C/\left[B - \left(B^2 - 4AC\right)^{1/2}\right] = 15.683 - 2 \times (-1.1684)/[-7242.8$$
$$(7242.8^2 + 4 \times 0.24895 \times 1.1684)^{1/2}] = 15.683\,\mathrm{km}^{1/2}$$

Since this value is the same as that computed previously, within the chosen accuracy, we take it as correct. In other words, $\chi_3 = 15.683$ km$^{1/2}$ is taken as the root of Kepler's equation

$$113643 = -0.079504\chi_3^2 \times C(1.4285 \times 10^{-4}\chi_3^2) - 0.035434\chi_3^3 \times S(1.4285 \times 10^{-4}\chi_3^2) + 7248.4\chi_3$$

Now, $\chi_1 = -15.767$ km$^{1/2}$ and $\chi_3 = 15.683$ km$^{1/2}$ are used to compute again the Lagrangian coefficients f_1, f_3, g_1 and g_3 as follows

$$f_1 = 1 - [\chi_1^2/r_2]C(\alpha\chi_1^2) = 1 - \left[(-15.767)^2/7248.4\right] \times C\,[1.4285 \times 10^{-4} \times (-15.767)^2] = 0.98290$$

$$f_3 = 1 - [\chi_3^2/\mathrm{r}_2]C(\alpha\chi_3^2) = 1 - \left(15.683^2/7248.4\right) \times C(1.4285 \times 10^{-4} \times 15.683^2) = 0.98308$$

$$g_1 = \tau_1 - [\chi_1^3/\mu^{1/2}]S(\alpha\chi_1^2) = -181 - [(-15.767)^3/398600.4^{1/2}] \times S[1.4285 \times 10^{-4} \times (-15.767)^2] = -179.97\,\mathrm{s}$$

$$g_3 = \tau_3 - [\chi_3^3/\mu^{1/2}]S(\alpha\chi_3^2) = 180 - (15.683^3/398600.4^{1/2}) \times S(1.4285 \times 10^{-4} \times 15.683^2) = 178.98\,\mathrm{s}$$

These values of the Lagrangian coefficients are to be compared with those resulting from the preliminary approximation. The two sets are shown below.

Preliminary-approximation values	First-iteration values
$f_1 = 0.98285$	$f_1 = 0.98290$
$f_3 = 0.98304$	$f_3 = 0.98308$
$g_1 = -179.97$	$g_1 = -179.97$
$g_3 = 178.98$	$g_3 = 178.98$

In order to improve the convergence of the process, Curtis [20] suggests to replace the first-iteration set of values by an arithmetic average of the two sets. This leads to the following values

$$f_1 = (0.98285 + 0.98290)/2 = 0.98288$$
$$f_3 = (0.98304 + 0.98308)/2 = 0.98306$$
$$g_1 = -179.97$$
$$g_3 = 178.98$$

which are used to compute again c_1 and c_3 by means of

$$c_1 = \frac{g_3}{f_1 g_3 - g_1 f_3} \qquad c_3 = -\frac{g_1}{f_1 g_3 - g_1 f_3}$$

Thus,

$$c_1 = 178.98/[0.98288 \times 178.98 - (-179.97) \times 0.98306] = 0.50726$$
$$c_3 = -(-179.97)/[0.98288 \times 178.98 - (-179.97) \times 0.98306] = 0.51007$$

These values of c_1 and c_3 are used to compute new values of ρ_1, ρ_2, and ρ_3, as follows

$$\rho_1 = (-D_{11} + D_{21}/c_1 - c_3 D_{31}/c_1)/D_0 = (-876.75 + 825.01/0.50726 - 0.51007 \times 773.39/0.50726)/(-0.018881) = 1484.0\,\text{km}$$
$$\rho_2 = (-c_1 D_{12} + D_{22} - c_3 D_{32})/D_0 = (-0.50726 \times 1050.6 + 996.51 - 0.51007 \times 942.47)/(-0.018881) = 907.95\,\text{km}$$
$$\rho_3 = (-c_1 D_{13}/c_3 + D_{23}/c_3 - D_{33})/D_0 = (-0.50726 \times 966.13/0.51007 + 910.44/0.51007 - 854.88)/(-0.018881) = 1628.9\,\text{km}$$

These values of ρ_1, ρ_2, and ρ_3 are used to compute new values of $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$, as follows

$$\boldsymbol{r}_1 = \boldsymbol{r}_{E1} + \rho_1 \boldsymbol{u}_1 = 3011.7\,\boldsymbol{u}_X - 3945.1\,\boldsymbol{u}_Y + 3992.5\,\boldsymbol{u}_Z + 1484.0 \times (0.80496\,\boldsymbol{u}_X - 0.52275\,\boldsymbol{u}_Y - 0.28067\,\boldsymbol{u}_Z) = 4206.3\,\boldsymbol{u}_X - 4720.9\,\boldsymbol{u}_Y + 3576.0\,\boldsymbol{u}_Z$$

$$\boldsymbol{r}_2 = \boldsymbol{r}_{E2} + \rho_2 \boldsymbol{u}_2 = 3063.5\,\boldsymbol{u}_X - 3905.0\,\boldsymbol{u}_Y + 3992.5\,\boldsymbol{u}_Z + 907.95 \times (0.51174\,\boldsymbol{u}_X - 0.45275\,\boldsymbol{u}_Y + 0.73016\,\boldsymbol{u}_Z) = 3528.1\,\boldsymbol{u}_X - 4316.1\,\boldsymbol{u}_Y + 4655.4\,\boldsymbol{u}_Z$$

$$\boldsymbol{r}_3 = \boldsymbol{r}_{E3} + \rho_3 \boldsymbol{u}_3 = 3114.5\,\boldsymbol{u}_X - 3864.5\,\boldsymbol{u}_Y + 3992.5\,\boldsymbol{u}_Z + 1628.9 \times (-0.23284\,\boldsymbol{u}_X + 0.059133\,\boldsymbol{u}_Y + 0.97072\,\boldsymbol{u}_Z) = 2735.2\,\boldsymbol{u}_X - 3768.2\,\boldsymbol{u}_Y + 5573.7\,\boldsymbol{u}_Z$$

The values of $\boldsymbol{r}_2$ and $\boldsymbol{r}_3$ obtained above are used to compute a new value of $\boldsymbol{v}_2$, as follows

$$\boldsymbol{v}_2 = \left(\frac{f_1}{f_1 g_3 - g_1 f_3}\right)\boldsymbol{r}_3 - \left(\frac{f_3}{f_1 g_3 - g_1 f_3}\right)\boldsymbol{r}_1$$

Since

$$f_1/(f_1 g_3 - g_1 f_3) = 0.98288/[0.98288 \times 178.98 - (-179.97) \times 0.98306] = 0.0027856$$
$$f_3/(f_1 g_3 - g_1 f_3) = 0.98306/[0.98288 \times 178.98 - (-179.97) \times 0.98306] = 0.0027862$$

then

$$\begin{aligned}\boldsymbol{v}_2 &= [f_1/(f_1 g_3 - g_1 f_3)]\boldsymbol{r}_3 - [f_3/(f_1 g_3 - g_1 f_3)]\boldsymbol{r}_1 = 0.0027856 \times (2735.2\,\boldsymbol{u}_X - 3768.2\,\boldsymbol{u}_Y \\ &\quad + 5573.7\,\boldsymbol{u}_Z) - 0.0027862 \times (4206.3\,\boldsymbol{u}_X - 4720.9\,\boldsymbol{u}_Y + 3576.0\,\boldsymbol{u}_Z) \\ &= -4.1004\,\boldsymbol{u}_X + 2.6567\,\boldsymbol{u}_Y + 5.5626\,\boldsymbol{u}_Z\end{aligned}$$

In summary, the values of the position and velocity vectors of the COBE satellite at time t_2, at this stage of the iterative procedure, are

$$\begin{aligned}\boldsymbol{r}_2 &= 3528.1\,\boldsymbol{u}_X - 4316.1\,\boldsymbol{u}_Y + 4655.4\,\boldsymbol{u}_Z \\ \boldsymbol{v}_2 &= -4.1004\,\boldsymbol{u}_X + 2.6567\,\boldsymbol{u}_Y + 5.5626\,\boldsymbol{u}_Z\end{aligned}$$

The remaining part of the computation is straightforward, but long. Therefore, it is not given here. Suffice it to say that, after the fifth iteration, the values of the four Lagrangian coefficients converge to those given below:

$$\begin{aligned}f_1 &= 0.98297 \\ f_3 &= 0.98315 \\ g_1 &= -179.97 \\ g_3 &= 178.99\end{aligned}$$

Consequently, the iterative process of refinement terminates after the fifth iteration. The position ($\boldsymbol{r}_2$) and velocity ($\boldsymbol{v}_2$) vectors of the COBE satellite at time t_2 can be computed by introducing these values in the expressions shown above. The computed values of these vectors are given below.

$$\begin{aligned}\boldsymbol{r}_2 &= 3525.5\,\boldsymbol{u}_X - 4313.7\,\boldsymbol{u}_Y + 4651.7\,\boldsymbol{u}_Z \\ \boldsymbol{v}_2 &= -4.0755\,\boldsymbol{u}_X + 2.6425\,\boldsymbol{u}_Y + 5.5324\,\boldsymbol{u}_Z\end{aligned}$$

The computed values can be compared with the actual values of the position and velocity vectors of the COBE satellite, at the same time t_2, which are given below (from Healy [33]).

$$\begin{aligned}\boldsymbol{r}_2 &= 3528.320\,\boldsymbol{u}_X - 4313.871\,\boldsymbol{u}_Y + 4654.938\,\boldsymbol{u}_Z \\ \boldsymbol{v}_2 &= -4.103\,\boldsymbol{u}_X + 2.658\,\boldsymbol{u}_Y + 5.564\,\boldsymbol{u}_Z\end{aligned}$$

About the difference existing between computed and actual values, it is to be noted that the data used for the computation have been considered as free from errors, which cannot be true in practice; in addition, no account has been taken of the perturbations.

The method shown above is due to Gauss [26] and has been refined by Gibbs [27]. Taff [66] has strongly censured this method on the grounds of the small radius of convergence of the f and g series. According to Taff, Moulton [56] has found that the radius of convergence of these series is $T\tau/2\pi$, where T is the orbital period (of the object observed) and τ is given by

$$\tau^2 = M_0^2 + \left\{ \ln\left[1 + \left(1 - e^2\right)^{\frac{1}{2}}\right] - \ln e - \left(1 - e^2\right)^{\frac{1}{2}} \right\}^2$$

where e is the orbital eccentricity and M_0 is the value of the mean anomaly, contained in the interval $[-\pi, \pi]$, at the instant $t = t_0$ [66]. Neither Taff nor Moulton, in the cited papers, define what must be understood by "instant $t = t_0$". It may be presumed either that t_0 is the intermediate (t_2) of the three observation times (t_1, t_2, and t_3) defined above, which are such that $t_1 < t_2 < t_3$, or that t_0 is exactly the midpoint of the interval $[t_1, t_3]$, that is, $t_0 = (t_1 + t_3)/2$.

On the other hand, well before Taff and Moulton, Gibbs had noted that "The determination of an orbit from three complete observations by the solution of the equations which represent elliptic motion present so great difficulties in the general case, that in the first solution of the problem we must generally limit ourselves to the case in which the intervals between the observations are not very long" [27]. In other words, the time intervals $\tau_1 = t_1 - t_2$ and $\tau_3 = t_3 - t_2$ must be small fractions of the period T of the orbiting body observed, which period is not known a priori. However, the method of Gauss shown above computes first the components of the position and velocity vectors of the observed object from a preliminary approximation and then refines iteratively the values of such components. The knowledge of the first-approximation set of values makes it possible to compute approximately the major semi-axis $a = 1/(2/r - v^2/\mu)$ of the orbit and hence the orbital period $T = 2\pi(a^3/\mu)^{1/2}$ of the observed object.

In addition, Branham [13] notes that, according some authors, Gauss' method is restricted to low-eccentricity orbits, because the radius of convergence of the f and g series becomes smaller and smaller as the orbital eccentricity approaches unity. This is because the value of τ in Moulton's [56] formula given above also depends on eccentricity. However, Branham himself admits that this restriction can be removed by using the f and g functions rather than series.

On the same line of reasoning is Marsden, who shows that Taff's criticism of Gauss' method of initial orbit determination on the grounds of the small radius of convergence of the f and g series is completely unjustified [51].

In the example given above, the preliminary approximation to the position and velocity vectors of the COBE satellite, observed at time t_2, has given the following results

$$\boldsymbol{r}_2 = 3520.6\,\boldsymbol{u}_X - 4309.4\,\boldsymbol{u}_Y + 4644.7\,\boldsymbol{u}_Z$$
$$\boldsymbol{v}_2 = -4.0299\,\boldsymbol{u}_X + 2.6179\,\boldsymbol{u}_Y + 5.4727\,\boldsymbol{u}_Z$$

The corresponding magnitudes r_2 and v_2 of $\boldsymbol{r}_2$ and $\boldsymbol{v}_2$ have been found to be

$$r_2 = (\boldsymbol{r}_2 \cdot \boldsymbol{r}_2)^{\frac{1}{2}} = \left[3520.6^2 + (-4309.4)^2 + 4644.7^2\right]^{\frac{1}{2}} = 7248.4\,\text{km}$$
$$v_2 = (\boldsymbol{v}_2 \cdot \boldsymbol{v}_2)^{\frac{1}{2}} = \left[(-4.0299)^2 + 2.6179^2 + 5.4727^2\right]^{\frac{1}{2}} = 53.044^{\frac{1}{2}}\,\text{km/s}$$

The major semi-axis results from the vis-viva integral

$$a = \frac{1}{\dfrac{2}{r_2} - \dfrac{v_2^2}{\mu}} = \frac{1}{\dfrac{2}{7248.2} - \dfrac{53.044}{398600.4}} = 7000.5\,\text{km}$$

The corresponding orbital period of the COBE satellite results from

$$T = 2\pi\left(\frac{a^3}{\mu}\right)^{\frac{1}{2}} = 2 \times 3.1416 \times \left(\frac{7000.5^3}{398600.4}\right)^{\frac{1}{2}} = 5829.1\,\text{s}$$

and the angular momentum per unit mass is

$$\begin{aligned}\boldsymbol{h} &= \left(Y_2 Z_2' - Z_2 Y_2'\right)\boldsymbol{u}_X + \left(Z_2 X_2' - X_2 Z_2'\right)\boldsymbol{u}_Y + \left(X_2 Y_2' - Y_2 X_2'\right)\boldsymbol{u}_Z \\ &= (-4309.4 \times 5.4727 - 2.6179 \times 4644.7)\,\boldsymbol{u}_X + (-3520.6 \times 5.4727 \\ &\quad - 4.0299 \times 4644.7)\,\boldsymbol{u}_Y + (3520.6 \times 2.6179 - 4.0299 \times 4309.4)\,\boldsymbol{u}_Z \\ &= -35743\,\boldsymbol{u}_X - 37985\,\boldsymbol{u}_Y - 8149.9\,\boldsymbol{u}_Z\end{aligned}$$

The square of the magnitude of $\boldsymbol{h}$ is

$$h^2 = \boldsymbol{h} \cdot \boldsymbol{h} = (-35743)^2 + (-37985)^2 + (-8149.9)^2$$

The semi-latus rectum results from

$$p = \frac{h^2}{\mu} = \frac{(-35743)^2 + (-37985)^2 + (-8149.9)^2}{398600.4} = 6991.6\,\text{km}$$

The orbital eccentricity results from

$$p = a\left(1 - e^2\right)$$

which, solved for e, yields

$$e = \left(1 - \frac{p}{a}\right)^{\frac{1}{2}} = \left(1 - \frac{6991.6}{7000.5}\right)^{\frac{1}{2}} = 0.035656$$

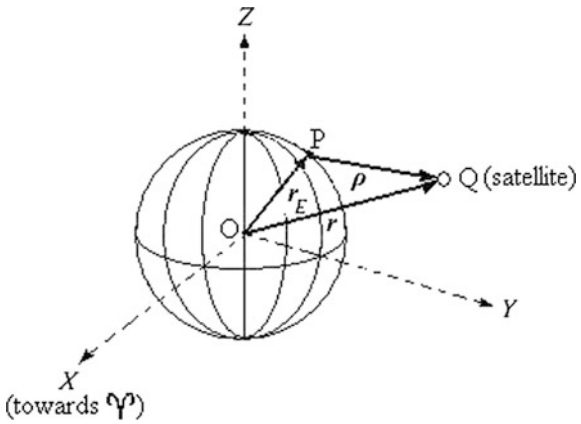
Now, since in the present case

- the time intervals $\tau_1 = -181$ s and $\tau_3 = 180$ s are equal to a small fraction (about 1/32) of the orbital period $T = 5829.1$ s; and
- the orbital eccentricity $e = 0.035656$ is much less than unity;

then the iterations converge, as has been shown above.

2.7 Orbital Elements from Three Measurements of Angles (Method of Laplace)

Let the topocentric right ascension (α) and declination (δ) of a satellite be given at three distinct times t_1, t_2, and t_3, that is, let a set of values for α_1, δ_1, α_2, δ_2, α_3, and δ_3 be given. Let $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, and $\boldsymbol{u}_3$ be the unit vectors along the line-of-sight vectors (respectively, $\boldsymbol{\rho}_1 \equiv \rho_1 \boldsymbol{u}_1$, $\boldsymbol{\rho}_2 \equiv \rho_2 \boldsymbol{u}_2$, and $\boldsymbol{\rho}_3 \equiv \rho_3 \boldsymbol{u}_3$) going from the observation site P placed on the surface of the Earth to the satellite Q observed, as shown in the following figure.



Indicating for brevity $\boldsymbol{u}_i$ ($i = 1$, 2, and 3) these unit vectors, Sect. 2.2 has shown that the components of $\boldsymbol{u}_i$ are

$$\boldsymbol{u}_i = (\cos\delta_i \cos\alpha_i)\boldsymbol{u}_X + (\cos\delta_i \sin\alpha_i)\boldsymbol{u}_Y + (\sin\delta_i)\boldsymbol{u}_Z$$

where $\boldsymbol{u}_X$, $\boldsymbol{u}_Y$, and $\boldsymbol{u}_Z$ are the unit vectors along the respective axes of the geocentric-equatorial system *XYZ*. The preceding expression, in matrix terms, is

$$\boldsymbol{u}_i = \begin{bmatrix} u_X \\ u_Y \\ u_Z \end{bmatrix}_i = \begin{bmatrix} \cos\delta\,\cos\alpha \\ \cos\delta\,\sin\alpha \\ \sin\delta \end{bmatrix}_i \qquad (i = 1,\ 2,\ 3)$$

The preceding figure also shows that

$$\boldsymbol{r}_i = \boldsymbol{r}_E + \boldsymbol{\rho}_i = \boldsymbol{r}_E + \rho_i \boldsymbol{u}_i \qquad (i = 1,\ 2,\ 3)$$

where ρ_i is the magnitude of the line-of-sight vector $\boldsymbol{\rho}_i$, and $\boldsymbol{r}_E$ is the vector from the centre of the Earth to the observation site.

As shown in Sect. 2.5, differentiating two times the preceding expression with respect to time yields

$$\begin{aligned} \boldsymbol{r}'_i &= \boldsymbol{r}'_E + \rho'_i \boldsymbol{u}_i + \rho_i \boldsymbol{u}'_i \\ \boldsymbol{r}''_i &= \boldsymbol{r}''_E + \rho''_i \boldsymbol{u}_i + \rho'_i \boldsymbol{u}'_i + \rho'_i \boldsymbol{u}'_i + \rho_i \boldsymbol{u}''_i = \boldsymbol{r}''_E + \rho''_i \boldsymbol{u}_i + 2\rho'_i \boldsymbol{u}'_i + \rho_i \boldsymbol{u}''_i \end{aligned}$$

On the other hand, the equation of motion of the satellite is

$$\boldsymbol{r}''_i = -\left(\frac{\mu}{r_i^3}\right)\boldsymbol{r}_i$$

By substituting $\boldsymbol{r}''_i = -(\mu/r_i^3)\boldsymbol{r}_i$ into $\boldsymbol{r}''_i = \boldsymbol{r}''_E + \rho''_i\ \boldsymbol{u}_i + 2\ \rho'_i\ \boldsymbol{u}'_i + \rho_i\ \boldsymbol{u}''_i$ and remembering that $\boldsymbol{r}_i = \boldsymbol{r}_E + \rho_i\ \boldsymbol{u}_i$, there results

$$\left(\frac{\mu}{r_i^3}\right)(\boldsymbol{r}_E + \rho_i \boldsymbol{u}_i) = \boldsymbol{r}''_E + \rho''_i \boldsymbol{u}_i + 2\rho'_i \boldsymbol{u}'_i + \rho_i \boldsymbol{u}''_i$$

that is,

$$\rho'' \boldsymbol{u}_i + 2\rho'_i \boldsymbol{u}'_i + \rho_i \left[\boldsymbol{u}''_i + \left(\frac{\mu}{r^3}\right)\boldsymbol{u}_i\right] = -\left[\boldsymbol{r}''_E + \left(\frac{\mu}{r^3}\right)\boldsymbol{r}_E\right]$$

(where $i = 1, 2, 3$). At a given time of observation (e.g. at the intermediate time t_2) the preceding vector equation is equivalent to 3 scalar equations in 10 unknowns ($\boldsymbol{u}'_i$, $\boldsymbol{u}''_i$, ρ_i, ρ'_i, ρ''_i, and r), where the known quantities are $\boldsymbol{u}_i$, $\boldsymbol{r}_E$, and $\boldsymbol{r}''_E$. The line-of-sight vector $\boldsymbol{u}_i$ is known at the three times t_1, t_2, and t_3. Consequently, when the time intervals $t_2 - t_1$ and $t_3 - t_2$ are small with respect to the orbital period T of the object observed, the values of $\boldsymbol{u}'_2$ and $\boldsymbol{u}''_2$ (that is, the values of $\boldsymbol{u}'_i$ and $\boldsymbol{u}''_i$ at the intermediate time t_2) can be computed by taking, respectively, the first and second time-derivative of $\boldsymbol{u}(t)$, where $\boldsymbol{u}(t)$ is the Lagrange polynomial which interpolates the three line-of-sight vectors $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, and $\boldsymbol{u}_3$. As is well known, this polynomial is

$$\boldsymbol{u}(t) = \left[\frac{(t-t_2)(t-t_3)}{(t_1-t_2)(t_1-t_3)}\right]\boldsymbol{u}_1 + \left[\frac{(t-t_1)(t-t_3)}{(t_2-t_1)(t_2-t_3)}\right]\boldsymbol{u}_2 + \left[\frac{(t-t_1)(t-t_2)}{(t_3-t_1)(t_3-t_2)}\right]\boldsymbol{u}_3$$

Differentiating once and twice this expression yields, respectively,

$$\boldsymbol{u}'(t) = \left[\frac{2t-t_2-t_3}{(t_1-t_2)(t_1-t_3)}\right]\boldsymbol{u}_1 + \left[\frac{2t-t_1-t_3}{(t_2-t_1)(t_2-t_3)}\right]\boldsymbol{u}_2 + \left[\frac{2t-t_1-t_2}{(t_3-t_1)(t_3-t_2)}\right]\boldsymbol{u}_3$$

$$\boldsymbol{u}''(t) = \left[\frac{2}{(t_1-t_2)(t_1-t_3)}\right]\boldsymbol{u}_1 + \left[\frac{2}{(t_2-t_1)(t_2-t_3)}\right]\boldsymbol{u}_2 + \left[\frac{2}{(t_3-t_1)(t_3-t_2)}\right]\boldsymbol{u}_3$$

The value of $\boldsymbol{u}'_2$ results from evaluating $\boldsymbol{u}'(t_2)$. When more than three observations are available, then $\boldsymbol{u}'_2$ and $\boldsymbol{u}''_2$ can be computed by using either Lagrange polynomials with a higher degree than two or a least-squares polynomial fit, as will be shown below.

Thus, at $t = t_2$, we have the equation

$$\rho''\boldsymbol{u} + 2\rho'\boldsymbol{u}' + \rho\left[\boldsymbol{u}'' + \left(\frac{\mu}{r^3}\right)\boldsymbol{u}\right] = -\left[\boldsymbol{r}''_E + \left(\frac{\mu}{r^3}\right)\boldsymbol{r}_E\right]$$

which, projected onto the axes X, Y, and Z, gives rise to the following scalar equations

$$\begin{aligned}
\rho'' u_X + 2\rho' u'_X + \rho[u''_X + (\mu/r^3)\ell_X] &= -[r''_{EX} + (\mu/r^3)r_{EX}] \\
\rho'' u_Y + 2\rho' u'_Y + \rho[u''_Y + (\mu/r^3)\ell_Y] &= -[r''_{EY} + (\mu/r^3)r_{EY}] \\
\rho'' u_Z + 2\rho' u'_Z + \rho[u''_Z + (\mu/r^3)\ell_Z] &= -[r''_{EZ} + (\mu/r^3)r_{EZ}]
\end{aligned}$$

for the three unknowns ρ'', ρ', and ρ. For the present, we consider r as a parameter. The matrix of the coefficients in the preceding system of equations is

$$\begin{bmatrix} u_X & 2u'_X & u''_X + (\mu/r^3)u_X \\ u_Y & 2u'_Y & u''_Y + (\mu/r^3)u_Y \\ u_Z & 2u'_Z & u''_Z + (\mu/r^3)u_Z \end{bmatrix}$$

Let D be the determinant of the preceding matrix. The value of D does not change if the first column of the matrix, multiplied by (μ/r^3), is subtracted from the third column, as follows

$$\begin{bmatrix} u_X & 2u'_X & u''_X + (\mu/r^3)u_X - u_X(\mu/r^3) \\ u_Y & 2u'_Y & u''_Y + (\mu/r^3)u_Y - u_Y(\mu/r^3) \\ u_Z & 2u'_Z & u''_Z + (\mu/r^3)u_Z - u_Z(\mu/r^3) \end{bmatrix}$$

Consequently, D is equal to

$$2\begin{bmatrix} u_X & u'_X & u''_X \\ u_Y & u'_Y & u''_Y \\ u_Z & u'_Z & u''_Z \end{bmatrix} = 2D_0$$

where D_0 is the determinant of the matrix written above. Solving the preceding system of equations according to Cramer's rule, we have

$$\rho = \frac{D_\rho}{D} = \frac{D_\rho}{2D_0}$$

where D_ρ is the determinant of the following matrix

$$-\begin{bmatrix} u_X & 2u'_X & r''_{EX} + (\mu/r^3)r_{EX} \\ u_Y & 2u'_Y & r''_{EY} + (\mu/r^3)r_{EY} \\ u_Z & 2u'_Z & r''_{EZ} + (\mu/r^3)r_{EZ} \end{bmatrix}$$

This matrix can be written as follows

$$-2\begin{bmatrix} u_X & u'_X & r''_{EX} \\ u_Y & u'_Y & r''_{EY} \\ u_Z & u'_Z & r''_{EZ} \end{bmatrix} - 2\left(\mu/r^3\right)\begin{bmatrix} u_X & u'_X & r_{EX} \\ u_Y & u'_Y & r_{EY} \\ u_Z & u'_Z & r_{EZ} \end{bmatrix}$$

Let D_1 and D_2 be the determinants of the two matrices written above. Since $\rho = D_\rho/D$, then the following equality holds

$$\rho = \frac{D_\rho}{D} = -2\frac{D_1}{D} - 2\left(\frac{\mu}{r^3}\right)\frac{D_2}{D} = -\frac{D_1}{D_0} - \left(\frac{\mu}{r^3}\right)\frac{D_2}{D_0} = -A - B\left(\frac{\mu}{r^3}\right)$$

where we have set for convenience

$$A = \frac{D_1}{D_0} = \frac{\boldsymbol{u} \cdot \left(\boldsymbol{u}' \times \boldsymbol{r}''_E\right)}{D_0}$$
$$B = \frac{D_2}{D_0} = \frac{\boldsymbol{u} \cdot (\boldsymbol{u}' \times \boldsymbol{r}_E)}{D_0}$$
$$D_0 = \boldsymbol{u} \cdot (\boldsymbol{u}' \times \boldsymbol{u}'')$$

The equation written above, that is,

$$\rho = -A - B\left(\frac{\mu}{r^3}\right)$$

holds, of course, if $D_0 \neq 0$.

On the other hand, by remembering that $\boldsymbol{r} = \boldsymbol{r}_E + \rho\,\boldsymbol{u}$ (where $\boldsymbol{r}$ and $\boldsymbol{u}$ are $\boldsymbol{r}_i$ and $\boldsymbol{u}_i$ at the intermediate time t_2) and taking the scalar product of $\boldsymbol{r}$ with itself, there results

$$\boldsymbol{r}^2 = \boldsymbol{r}\cdot\boldsymbol{r} = (\boldsymbol{r}_E + \rho\,\boldsymbol{u})\cdot(\boldsymbol{r}_E + \rho\,\boldsymbol{u}) = r_E^2 + 2\,\rho\,(\boldsymbol{r}_E\cdot\boldsymbol{u}) + \rho^2$$

By setting for convenience

$$E = -2\,\rho\,(\boldsymbol{r}_E\cdot\boldsymbol{u})$$
$$F = r_E^2$$

there results the following system of two algebraic equations

$$\rho = -A - B\left(\frac{\mu}{r^3}\right)$$

$$r^2 = \rho^2 - E\,\rho + F$$

for the two unknowns ρ and r.

Now, substituting $\rho = -A - B(\mu/r^3)$ into $r^2 = \rho^2 - E\,\rho + F$ leads to

$$r^2 = A^2 + \frac{2\mu A B}{r^3} + \frac{\mu^2 B^2}{r^6} + AE + \frac{\mu BE}{r^6} + F$$

Multiplying all terms of the preceding equation by r^6 leads to Lagrange's equation

$$r^8 + ar^6 + br^3 + c = 0$$

where

$$a = -\left(A^2 + AE + F\right)$$
$$b = -\mu(2AB + BE)$$
$$c = -\mu^2 B^2$$

This equation can be solved numerically by means of one of the methods shown in Sect. 1.5. The result is the value of r relating to the intermediate time t_2. By introducing this value into

$$\rho = -A - B\left(\frac{\mu}{r^3}\right)$$

it is possible to compute ρ. Thus, the position vector of the satellite at time t_2 results from

$$\boldsymbol{r} = \boldsymbol{r}_E + \rho\,\boldsymbol{u}$$

With the view of computing the velocity vector $\boldsymbol{v} \equiv \boldsymbol{r}'$ of the satellite at time t_2, let us consider again the system of equations

$$\begin{aligned}
\rho'' u_X + 2\rho' u'_X + \rho\left[u''_X + (\mu/r^3)\ell_X\right] &= -\left[r''_{EX} + (\mu/r^3) r_{EX}\right] \\
\rho'' u_Y + 2\rho' u'_Y + \rho\left[u''_Y + (\mu/r^3)\ell_Y\right] &= -\left[r''_{EY} + (\mu/r^3) r_{EY}\right] \\
\rho'' u_Z + 2\rho' u'_Z + \rho\left[u''_Z + \left(\mu/r^3\right)\ell_Z\right] &= -\left[r''_{EZ} + \left(\mu/r^3\right) r_{EZ}\right]
\end{aligned}$$

Using again Cramer's rule, we solve now for ρ' and write

$$\rho' = \frac{D_{\rho'}}{D}$$

where $D \neq 0$ is the determinant of the matrix of coefficients (see above) and $D_{\rho'}$ is the determinant of the following matrix

$$\begin{bmatrix}
u_X & -\left[r''_{EX} + (\mu/r^3) r_{EX}\right] & u''_X + (\mu/r^3) u_X \\
u_Y & -\left[r''_{EY} + (\mu/r^3) r_{EY}\right] & u''_Y + (\mu/r^3) u_Y \\
u_Z & -\left[r''_{EZ} + (\mu/r^3) r_{EZ}\right] & u''_Z + (\mu/r^3) u_Z
\end{bmatrix}$$

This matrix can also be written as follows

$$-\begin{bmatrix} u_X & r''_{EX} & u''_X \\ u_Y & r''_{EY} & u''_Y \\ u_Z & r''_{EZ} & u''_Z \end{bmatrix} - \left(\mu/r^3\right)\begin{bmatrix} u_X & r_{EX} & u''_X \\ u_Y & r_{EY} & u''_Y \\ u_Z & r_{EZ} & u''_Z \end{bmatrix}$$

Let D_3 be the determinant of the first and D_4 be the determinant of the second of the two matrices written above. By so doing, the previous expression $\rho' = D_{\rho'}/D$ can also be written as follows

$$\rho' = \frac{D_{\rho'}}{D} = -\frac{D_3}{D} - \left(\frac{\mu}{r^3}\right)\frac{D_4}{D}$$

where $D \neq 0$. Since we have set $D = 2D_0$, the previous equation can also be written as follows

$$\rho' = -\frac{D_3}{2D_0} - \left(\frac{\mu}{r^3}\right)\frac{D_4}{2D_0}$$

By setting for convenience

$$C = \frac{D_3}{2D_0} = \frac{\boldsymbol{u} \cdot \left(\boldsymbol{r}''_E \times \boldsymbol{u}''\right)}{2D_0}$$

$$G = \frac{D_4}{2D_0} = \frac{\boldsymbol{u} \cdot \left(\boldsymbol{r}_E \times \boldsymbol{u}''\right)}{2D_0}$$

the previous equation can be written as follows

$$\rho' = -C - G\left(\frac{\mu}{r^3}\right)$$

Since r is known, the previous expression makes it possible to compute the value of ρ'. This value, substituted into

$$\boldsymbol{v} \equiv \boldsymbol{r}' = \boldsymbol{r}'_E + \rho' \boldsymbol{u} + \rho \boldsymbol{u}'$$

yields the velocity vector $\boldsymbol{v}$ of the satellite at time t_2.

As mentioned above, this method fails when the determinant D_0 of the matrix

$$\begin{bmatrix} u_X & u'_X & u''_X \\ u_Y & u'_Y & u''_Y \\ u_Z & u'_Z & u''_Z \end{bmatrix}$$

approaches zero, which happens when the observer lies in the plane of the orbit at the intermediate time t_2.

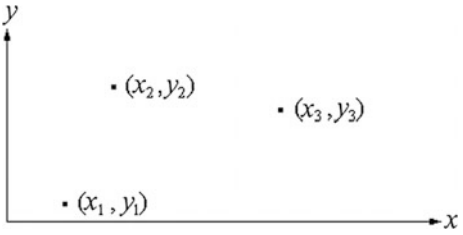
The following section of this paragraph shows how to fit the observations gathered to an orbit by using the method of least squares, which was invented by Gauss in 1795. According to Gauss (Theoria motus, Book 2, Sect. 3, paragraph 172, page 205 of the cited Ref. [26]), "Si observationes astronomicae ceterique numeri, quibus orbitarum computus innititur, absoluta praecisione gauderent, elementa quoque, sive tribus observationibus sive quatuor superstructa fuerint, absolute exacta statim prodirent (quatenus quidem motus secundum leges Kepleri exacte fieri supponitur), adeoque accitis aliis aliisque observationibus confirmari tantum possent, haud corrigi. Verum enim vero quum omnes mensurationes atque observationes nostrae nihil sint nisi approximationes ad veritatem, idemque de omnibus calculis illis innitentibus valere debent, scopum summum omnium computorum circa phaenomena concreta institutorum in eo ponere oportebit, ut ad veritatem quam proxime fieri potest accedamus. Hoc autem aliter fieri nequit, nisi per idoneam combinationem observationum plurium, quam quot ad determinationem quantitatum incognitarum absolute requiruntur. Hoc negotium tunc demum suscipere licebit, quando orbitae cognitio approximata iam innotuit, quae dein ita rectificanda est, ut omnibus observationibus quam exactissime satisfaciat".

In plain English, "If the astronomical observations and other numbers, on which the computation of orbit is based, were absolutely precise, the elements also, deduced by means of three or four observations, would be absolutely exact (within the limits of validity of Kepler's laws). Hence, the computed elements could only be confirmed, but never corrected, by further observations. In practice, since all our measurements and observations are nothing more than approximations to the truth, the same holds for all calculations based on them. Thus, the principal purpose of all computations concerning concrete phenomena must needs be that of drawing us near the truth to the maximum possible extent. This can be done in no other way

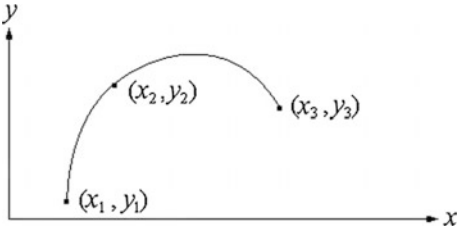
than by using a suitable combination of more observations than would be strictly necessary to determine the unknown quantities. This task can only be undertaken, when an approximate knowledge of the orbit has already been reached, which is then to be corrected in such a way as to satisfy all the observations to the maximum extent of exactness".

In a few words, since all actual observations are affected by random errors, then more observations than those which are strictly necessary are needed, the usefulness of the redundant measurements being the mutual cancellation of the errors to the maximum possible extent.

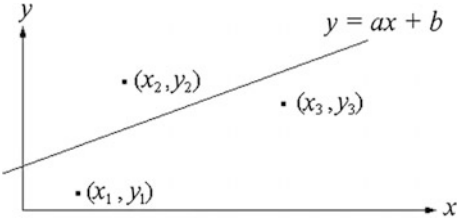
With reference to the following figure, suppose we have a set of approximate values y_i which correspond to discrete values x_i (where $i = 1, 2, \ldots, n$) of the independent variable x.



If the three points (x_1, y_1), (x_2, y_2), and (x_3, y_3) were joined by an interpolating polynomial, as shown in the following figure, then the resulting curve would pass exactly through each of the three points but would also have an oscillating behaviour, which makes the curve unfit to represent the overall trend of our data.



If, on the contrary, the same data were represented by a straight line obtained by means of the least-squares method, then the result would be like that shown below.



The least-squares method produces a functional form $f(x, c_0, c_1, \ldots, c_m) \equiv f_m(x)$, which depends not only on x but also on $m + 1$ parameters $c_0, c_1, \ldots, c_m$ to be determined in such a way as to minimise the squares of the residuals, that is, the squares of the differences between the functional form $f_m(x)$ and each data point.

A type of functional form $f_m(x)$ frequently used is an algebraic polynomial:

$$f_m(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_m x^m$$

The degree m of this polynomial must be chosen by the solver. However, m must be less than $n - 1$, where n is the number of the given points. Otherwise, if m were equal to $n - 1$, then $f_m(x)$ would be just the interpolating polynomial.

For the purpose of illustrating the method, suppose that we are given a set of n points (x_i, y_i) with $i = 1, 2, \ldots, n$, and that we search the least-squares fit of these points by means of a second-degree polynomial $f_2(x) = c_0 + c_1 x + c_2 x^2$. In this case, according to the least-squares method, the unknown coefficients c_0, c_1, and c_2 must be chosen in such a way as to correspond to the least squares of the residuals, the residuals being the components $\rho_1, \rho_2, \ldots, \rho_n$ of the following $n \times 1$ residual vector

$$\boldsymbol{\rho} \equiv \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_n \end{bmatrix}$$

Such components are

$$\begin{aligned} \rho_1 = y_1 - f_2(x_1) &= y_1 - c_0 - c_1 x_1 - c_2 x_1^2 \\ \rho_2 = y_2 - f_2(x_2) &= y_2 - c_0 - c_1 x_2 - c_2 x_2^2 \\ &\vdots \\ \rho_n = y_n - f_2(x_n) &= y_n - c_0 - c_1 x_n - c_2 x_n^2 \end{aligned}$$

This is because the Euclidean length ρ of the residual vector $\boldsymbol{\rho}$ is just the square root of the sum of the squares of its components: $\rho = (\rho_1^2 + \rho_2^2 + \ldots + \rho_n^2)^{1/2}$.

Consequently, in order for the residual vector $\boldsymbol{\rho}$ to have the minimum Euclidean length, ρ^2 must have the minimum value, where

$$\begin{aligned} \rho^2 = \rho_1^2 + \rho_2^2 + \cdots + \rho_n^2 = {} & \left(y_1 - c_0 - c_1 x_1 - c_2 x_1^2\right)^2 \\ & + \left(y_2 - c_0 - c_1 x_2 - c_2 x_2^2\right)^2 + \cdots + \left(y_n - c_0 - c_1 x_n - c_2 x_n^2\right)^2 \end{aligned}$$

This condition is satisfied by those values of $c_0, c_1, \ldots, c_m$ which cause the first derivative of ρ^2 to vanish. If $m = 2$, this leads to the following three conditions:

$$\frac{\partial \rho^2}{\partial c_0} = \frac{\partial \rho_1^2}{\partial c_0} + \frac{\partial \rho_2^2}{\partial c_0} + \cdots + \frac{\partial \rho_n^2}{\partial c_0} = -2[(y_1 - c_0 - c_1 x_1 - c_2 x_1^2) + (y_2 - c_0 - c_1 x_2 - c_2 x_2^2) + \cdots + (y_n - c_0 - c_1 x_n - c_2 x_n^2)] = 0$$

$$\frac{\partial \rho^2}{\partial c_1} = \frac{\partial \rho_1^2}{\partial c_1} + \frac{\partial \rho_2^2}{\partial c_1} + \cdots + \frac{\partial \rho_n^2}{\partial c_1} = -2[x_1(y_1 - c_0 - c_1 x_1 - c_2 x_1^2) + x_2(y_2 - c_0 - c_1 x_2 - c_2 x_2^2) + \cdots + x_n(y_n - c_0 - c_1 x_n - c_2 x_n^2)] = 0$$

$$\frac{\partial \rho^2}{\partial c_2} = \frac{\partial \rho_1^2}{\partial c_2} + \frac{\partial \rho_2^2}{\partial c_2} + \cdots + \frac{\partial \rho_n^2}{\partial c_2} = -2[x_1^2(y_1 - c_0 - c_1 x_1 - c_2 x_1^2) + x_2^2(y_2 - c_0 - c_1 x_2 - c_2 x_2^2) + \cdots + x_n^2(y_n - c_0 - c_1 x_n - c_2 x_n^2)] = 0$$

Since each of the three partial derivatives must be equal to zero, then the coefficient in front of the square brackets (that is, −2) can be dropped. Thus, the previous three conditions reduce to

$$nc_0 + (x_1 + x_2 + \ldots + x_n)c_1 + (x_1^2 + x_2^2 + \ldots + x_n^2)c_2 = y_1 + y_2 + \ldots + y_n$$

$$(x_1 + x_2 + \ldots + x_n)c_0 + (x_1^2 + x_2^2 + \ldots + x_n^2)c_1 + (x_1^3 + x_2^3 + \ldots + x_n^3)c_2 = x_1(y_1 + y_2 + \ldots + y_n)$$

$$(x_1^2 + x_2^2 + \ldots + x_n^2)c_0 + (x_1^3 + x_2^3 + \ldots + x_n^3)c_1 + (x_1^4 + x_2^4 + \ldots + x_n^4)c_2 = x_1^2(y_1 + y_2 + \ldots + y_n)$$

This is a system of three linear algebraic equations for the three unknowns c_0, c_1, and c_2. The values of c_0, c_1, and c_2 determine the least-squares fit parabola

$$f_2(x) = c_0 + c_1 x + c_2 x^2$$

Let us consider now the differentiation of a general (m-degree) least-squares fit polynomial. Let

$$f_m(x = c_0 + c_1 x + c_2 x^2 + \ldots + c_n x^m$$

be this polynomial. Differentiating once and twice the polynomial with respect to x yields

$$\frac{\mathrm{d}f_m}{\mathrm{d}x} = c_1 + 2c_2 x + 3c_3 x^2 + \cdots$$

$$\frac{d^2 f_m}{\mathrm{d}x^2} = 2c_2 + 6c_3 x + \cdots$$

The expressions shown above make it possible to compute $f_m(x)$ and its first and second derivative at any point x of interest. If that point be chosen as the origin of the series expansions, then there results $x = 0$, $f_m(0) = c_0$, $(\mathrm{d}f_m/\mathrm{d}x)_0 = c_1$, and $(\mathrm{d}^2 f_m/\mathrm{d}x^2)_0 = 2c_2$.

2.8 Improvement in Orbit Determination by Differential Correction

The preceding paragraphs have shown some methods for the preliminary determination of the orbit followed by a space object. These methods use only the minimum number of observations necessary to compute an orbit. Since six independent quantities are strictly necessary to determine the motion of a body, then these methods use a set of six independent quantities.

However, as has been shown numerically in an example given in Sect. 2.6, the six orbital elements (or the 3 + 3 components of the position and velocity vectors at a given time t_0) computed by means of these methods differ to some extent from the actual orbital elements. For the purpose of reducing, as far as possible, the influence of the errors, the residuals (that is, the differences between the computed and observed quantities) must be determined at each time of observation. In some cases, it is possible to refine a preliminary orbit without taking account of the perturbations, that is, considering a purely Keplerian orbit. As shown in Sect. 2.6, this happens when there are at least three reliable observations which cover a significant part of the orbit but are not too distant in time with respect to the epoch of the preliminary determination of the orbital elements.

The differential correction is a numerical procedure based on the principle of least squares, which is meant to correct the computed elements by minimising the residuals. The quantities which concur to determine the observed values of the right ascension (α) and declination (δ) of a given body are substantially the orbital elements of that body and the position vector $\boldsymbol{r}_E$ which specifies the location of the observer with respect to the geocentric-equatorial system. Of these quantities, $\boldsymbol{r}_E$ can be supposed to be known accurately enough to require no improvement in accuracy make it useless to improve $\boldsymbol{r}_E$. Consequently, the uncertainties in the values of α and δ can be attributed to the orbital elements, that is, to the components of the position ($\boldsymbol{r}$) and velocity ($\boldsymbol{r}'$) vectors of the space object. Let t_0 be an epoch chosen arbitrarily and let x_0, y_0, z_0, x'_0, y'_0, and z'_0 be the components of such vectors at t_0. The dependency of α and δ on x_0, y_0, z_0, x'_0, y'_0, and z'_0 can be written as follows

$$\begin{aligned} \alpha &= \alpha\left(x_0, y_0, z_0, x'_0, y'_0, z'_0\right) \\ \delta &= \delta\left(x_0, y_0, z_0, x'_0, y'_0, z'_0\right) \end{aligned}$$

The total differentials of α and δ are

$$\begin{aligned} \mathrm{d}\alpha &= \frac{\partial \alpha}{\partial x_0}\mathrm{d}x_0 + \frac{\partial \alpha}{\partial y_0}\mathrm{d}y_0 + \cdots + \frac{\partial \alpha}{\partial z'_0}\mathrm{d}z'_0 \\ \mathrm{d}\delta &= \frac{\partial \delta}{\partial x_0}\mathrm{d}x_0 + \frac{\partial \delta}{\partial y_0}\mathrm{d}y_0 + \cdots + \frac{\partial \delta}{\partial z'_0}\mathrm{d}z'_0 \end{aligned}$$

and express the changes in α and δ resulting from the independent changes in x_0, y_0, …, z'_0. Replacing the differentials by the corresponding finite differences leads to

$$\Delta\alpha = \frac{\partial\alpha}{\partial x_0}\Delta x_0 + \frac{\partial\alpha}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\alpha}{\partial z'_0}\Delta z'_0$$
$$\Delta\delta = \frac{\partial\delta}{\partial x_0}\Delta x_0 + \frac{\partial\delta}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\delta}{\partial z'_0}\Delta z'_0$$

where $\Delta\alpha$ and $\Delta\delta$ are the residuals in, respectively, right ascension and declination, and Δx_0, Δy_0, …, $\Delta z'_0$ are the changes to be made in, respectively, x_0, y_0, …, z'_0 for the purpose of reducing such residuals to zero. Now, $\Delta\alpha$ and $\Delta\delta$ are known quantities, which result from measuring the position of the observed body on the celestial sphere. When three couples of values (α, δ) are measured at three different times t_1, t_2, and t_3, that is, when we have a set of six values $\{\alpha_1, \delta_1, \alpha_2, \delta_2, \alpha_3, \delta_3\}$, then 2×3 independent equations of conditions can be written as follows

$$\Delta\alpha_1 = \frac{\partial\alpha_1}{\partial x_0}\Delta x_0 + \frac{\partial\alpha_1}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\alpha_1}{\partial z'_0}\Delta z'_0$$
$$\Delta\delta_1 = \frac{\partial\delta_1}{\partial x_0}\Delta x_0 + \frac{\partial\delta_1}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\delta_1}{\partial z'_0}\Delta z'_0$$
$$\Delta\alpha_2 = \frac{\partial\alpha_2}{\partial x_0}\Delta x_0 + \frac{\partial\alpha_2}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\alpha_2}{\partial z'_0}\Delta z'_0$$
$$\Delta\delta_2 = \frac{\partial\delta_2}{\partial x_0}\Delta x_0 + \frac{\partial\delta_2}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\delta_2}{\partial z'_0}\Delta z'_0$$
$$\Delta\alpha_3 = \frac{\partial\alpha_3}{\partial x_0}\Delta x_0 + \frac{\partial\alpha_3}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\alpha_3}{\partial z'_0}\Delta z'_0$$
$$\Delta\delta_3 = \frac{\partial\delta_3}{\partial x_0}\Delta x_0 + \frac{\partial\delta_3}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\delta_3}{\partial z'_0}\Delta z'_0$$

When we use more measurements than those strictly necessary, that is, when a set of n measurements $\{\alpha_1, \delta_1, \alpha_2, \delta_2, \ldots, \alpha_n, \delta_n\}$ is available, where $n \geq 3$, then $2n$ independent equations can be written as follows

$$\Delta\alpha_1 = \frac{\partial\alpha_1}{\partial x_0}\Delta x_0 + \frac{\partial\alpha_1}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\alpha_1}{\partial z'_0}\Delta z'_0$$
$$\Delta\delta_1 = \frac{\partial\delta_1}{\partial x_0}\Delta x_0 + \frac{\partial\delta_1}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\delta_1}{\partial z'_0}\Delta z'_0$$
$$\Delta\alpha_2 = \frac{\partial\alpha_2}{\partial x_0}\Delta x_0 + \frac{\partial\alpha_2}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\alpha_2}{\partial z'_0}\Delta z'_0$$
$$\vdots$$
$$\Delta\alpha_n = \frac{\partial\alpha_n}{\partial x_0}\Delta x_0 + \frac{\partial\alpha_n}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\alpha_n}{\partial z'_0}\Delta z'_0$$
$$\Delta\delta_n = \frac{\partial\delta_n}{\partial x_0}\Delta x_0 + \frac{\partial\delta_n}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\delta_n}{\partial z'_0}\Delta z'_0$$

Thus, the problem reduces to obtaining suitable numerical values of the partial derivatives, because these values, introduced into the preceding equations, make it possible to obtain the values of the corrections Δx_0, Δy_0, …, $\Delta z'_0$ which best fit the available measurements α_1, δ_1, α_2, δ_2, …, α_n, δ_n.

To this end, Escobal [22] suggests the following procedure. Let ε be any one of the six variables x_0, y_0, z_0, x'_0, y'_0, and z'_0. Let $\Delta\varepsilon$ be some small change introduced in that variable. The partial derivatives of α_i and δ_i ($i = 1, 2, \ldots, n$) with respect to ε can be approximated by means of the following $2n$ expressions

$$\frac{\partial \alpha_i}{\partial \epsilon} \approx \frac{\alpha_i\left(x_0, \ldots, \epsilon_0 + \Delta\epsilon, \ldots, z'_0\right) - \alpha_i\left(x_0, \ldots, \epsilon_0, \ldots, z'_0\right)}{\Delta\epsilon}$$

$$\frac{\partial \delta_i}{\partial \epsilon} \approx \frac{\delta_i\left(x_0, \ldots, \epsilon_0 + \Delta\epsilon, \ldots, z'_0\right) - \delta_i\left(x_0, \ldots, \epsilon_0, \ldots, z'_0\right)}{\Delta\epsilon}$$

where ε_0 is the value of ε at the epoch chosen (t_0). Each of the variables is incremented in turn, while the others maintain their original values. By so doing, the two equations written above approximate the needed $6n$ partial derivatives. Usually an increment $\Delta\varepsilon$ equal to a few units per cent (Bate et al. [5], suggest 1 or 2%) suffices to provide a satisfactory approximation of the partial derivatives $\partial\alpha_i/\partial\varepsilon$ and $\partial\delta_i/\partial\varepsilon$ ($i = 1, 2, \ldots, n$). Now the values of these partial derivatives, obtained as has been shown above, are introduced into

$$\begin{aligned}
\Delta\alpha_1 &= \frac{\partial\alpha_1}{\partial x_0}\Delta x_0 + \frac{\partial\alpha_1}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\alpha_1}{\partial z'_0}\Delta z'_0 \\
\Delta\delta_1 &= \frac{\partial\delta_1}{\partial x_0}\Delta x_0 + \frac{\partial\delta_1}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\delta_1}{\partial z'_0}\Delta z'_0 \\
\Delta\alpha_2 &= \frac{\partial\alpha_2}{\partial x_0}\Delta x_0 + \frac{\partial\alpha_2}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\alpha_2}{\partial z'_0}\Delta z'_0 \\
&\vdots \\
\Delta\alpha_n &= \frac{\partial\alpha_n}{\partial x_0}\Delta x_0 + \frac{\partial\alpha_n}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\alpha_n}{\partial z'_0}\Delta z'_0 \\
\Delta\delta_n &= \frac{\partial\delta_n}{\partial x_0}\Delta x_0 + \frac{\partial\delta_n}{\partial y_0}\Delta y_0 + \cdots + \frac{\partial\delta_n}{\partial z'_0}\Delta z'_0
\end{aligned}$$

and these equations are solved for Δx_0, Δy_0, …, $\Delta z'_0$. Now the variations Δx_0, Δy_0, …, $\Delta z'_0$ are added to the respective quantities x_0, y_0, …, z'_0. This makes it possible to compute a new couple of position ($\boldsymbol{r}$) and velocity ($\boldsymbol{r}'$) vectors

$$\boldsymbol{r}_0 \equiv \begin{bmatrix} x_0 + \Delta x_0 \\ y_0 + \Delta y_0 \\ z_0 + \Delta z_0 \end{bmatrix} \qquad \boldsymbol{r}'_0 \equiv \begin{bmatrix} x'_0 + \Delta x'_0 \\ y'_0 + \Delta y'_0 \\ z'_0 + \Delta z'_0 \end{bmatrix}$$

at the same epoch (t_0). Hence, new values of right ascension and declination are computed and compared with the observed values. The differences between the two sets (computed quantities minus observed quantities) are the residuals. If these residuals are greater than a chosen tolerance, then the refinement process is repeated; otherwise it is stopped.

2.9 Improvement in Orbit Determination by Weighted Least-Squares Estimation

The preceding paragraph has shown how to determine the components of the position and velocity vectors of an observed object so that the residuals should be as small as possible. In practice, since different measurements have different units and degrees of reliability, then a weighting factor is applied to each residual, and consequently the quantity to be minimised is the square of the weighted residuals. In addition, as shown in Sect. 2.1, the force model used to compute the forces acting upon the observed object at a time t can only provide approximate values of the true forces, because the physical constants are known approximately and also because the mathematical model used to compute the forces can never be exact.

The mathematical formulation of this general problem will be shown below following Montenbruck and Gill [53]. Let us consider the m-dimensional column vector $\boldsymbol{x}$, whose components $x_1, x_2, \ldots, x_m$ are the time-dependent components of the position ($\boldsymbol{r}$) and velocity ($\boldsymbol{r}'$) vectors of the observed object, and the components of two vectors $\boldsymbol{p}$ and $\boldsymbol{q}$, which contain the free parameters characterising the force and measurement model, as follows

$$\boldsymbol{x} \equiv \begin{bmatrix} \boldsymbol{r} \\ \boldsymbol{r}' \\ \boldsymbol{p} \\ \boldsymbol{q} \end{bmatrix}$$

Let

$$\begin{aligned} \boldsymbol{x}' &= \boldsymbol{f}(t, \boldsymbol{x}) \\ \boldsymbol{x}_0 &= \boldsymbol{x}(t_0) \end{aligned}$$

be the differential equation and the initial condition which describe the evolution in time of the augmented state vector $\boldsymbol{x}$. Let

$$\boldsymbol{z} \equiv \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}$$

be an n-dimensional column vector, whose components z_1, z_2, …, z_n are the observations at times t_1, t_2, …, t_n. These observations can be expressed as follows

$$z_i(t_i) = g_i(t_i, \boldsymbol{x}_i) + \varepsilon_i = h_i(t_i, \boldsymbol{x}_0) + \varepsilon_i$$

(with $i = 1, 2, \ldots, n$) or in vector terms

$$\boldsymbol{z} = \boldsymbol{h}(x_0) + \boldsymbol{\varepsilon}$$

where g_i indicates the computed value of the i^{th} observation as a function of time t_i and the instantaneous state vector $\boldsymbol{x}_i$, and h_i indicates the same value as a function of the state vector $\boldsymbol{x}_0$ at the reference epoch t_0. The presence of the quantities ε_i is due to the difference between computed and actual values because of measurement errors, which are assumed to be randomly distributed with zero mean value. Using the least-squares method, we search the state vector $\boldsymbol{x}_0^{\text{lsq}}$ corresponding to the minimum value of the loss function

$$J(\boldsymbol{x}_0) = \boldsymbol{\rho} \cdot \boldsymbol{\rho} \equiv \boldsymbol{\rho}^{\mathrm{T}} \boldsymbol{\rho} = [\boldsymbol{z} - \boldsymbol{h}(\boldsymbol{x}_0)]^{\mathrm{T}} [\boldsymbol{z} - \boldsymbol{h}(\boldsymbol{x}_0)]$$

that is, corresponding to the minimum value of the squared sums of the residuals ρ_1, ρ_2, …, ρ_n, for a given set of observations z_1, z_2, …, z_n.

The expression of the loss function given above holds for observations of equal type and quality. This assumption will be removed in the following section of this paragraph. In addition, the number (n) of observations is assumed to be greater than or equal to the number (m) of unknowns.

The vector-valued function $\boldsymbol{h}(\boldsymbol{x}_0)$ appearing in the equation $\boldsymbol{z} = \boldsymbol{h}(\boldsymbol{x}_0) + \boldsymbol{\varepsilon}$ is a nonlinear function. Since an approximate value ($\boldsymbol{x}_0^{\text{appr}}$) of the actual state vector ($\boldsymbol{x}_0$) at epoch is known, the problem can be simplified by linearising the function $\boldsymbol{h}$.

Let $\boldsymbol{x}_0^{\text{ref}}$ be a reference state vector, which is initially set equal to $\boldsymbol{x}_0^{\text{appr}}$. The residual vector $\boldsymbol{\rho} = \boldsymbol{z} - \boldsymbol{h}(\boldsymbol{x}_0)$ is expressed approximately as follows

$$\boldsymbol{\rho} = \boldsymbol{z} - \boldsymbol{h}(\boldsymbol{x}_0) \approx \boldsymbol{z} - \boldsymbol{h}\left(\boldsymbol{x}_0^{\text{ref}}\right) - \left(\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}_0}\right)_{\text{ref}} \left(\boldsymbol{x}_0 - \boldsymbol{x}_0^{\text{ref}}\right) = \Delta \boldsymbol{z} - \mathbf{H} \Delta \boldsymbol{x}_0$$

where $\Delta \boldsymbol{x}_0 = \boldsymbol{x}_0 - \boldsymbol{x}_0^{\text{ref}}$ is the difference between $\boldsymbol{x}_0$ and the reference state vector, $\Delta \boldsymbol{z} = \boldsymbol{z} - \boldsymbol{h}(\boldsymbol{x}_0^{\text{ref}})$ is the difference between the actual observations and those resulting from the computed reference orbit, and the Jacobian $\mathbf{H} = (\partial \boldsymbol{h} / \partial \boldsymbol{x}_0)_{\text{ref}}$ is the matrix of the partial derivatives of the computed values with respect to the state vector (here the subscript ref indicates that the partial derivatives are to be evaluated at $\boldsymbol{x}_0 = \boldsymbol{x}_0^{\text{ref}}$) at the reference epoch t_0. Thus, by means of

$$\boldsymbol{\rho} = \Delta \boldsymbol{z} - \mathbf{H}\, \Delta \boldsymbol{x}_0$$

we obtain the residual vector $\boldsymbol{\rho}$ after applying a correction $\Delta \boldsymbol{x}_0$ to the reference state vector and recomputing the observations.

This makes it possible to reduce the original nonlinear least-squares problem of finding the minimum value of the loss function

$$J(\boldsymbol{x}_0) = [z - \boldsymbol{h}(\boldsymbol{x}_0)]^{\mathrm{T}}[z - \boldsymbol{h}(\boldsymbol{x}_0)]$$

to the simpler linear least-squares problem of searching $\Delta\boldsymbol{x}_0^{\mathrm{lsq}}$ corresponding to the minimum value of the loss function

$$J(\boldsymbol{x}_0) = [\Delta z - \mathbf{H}\,\Delta\boldsymbol{x}_0]^{\mathrm{T}}[\Delta z - \mathbf{H}\,\Delta\boldsymbol{x}_0]$$

If the Jacobian $\mathbf{H}$ has full rank m (the rank of a matrix $\mathbf{A}$ is the maximum number of linearly independent rows or columns of $\mathbf{A}$), that is, if the columns of $\mathbf{H}$ are linearly independent, then the minimum value of the loss function is attained when the partial derivatives of $J(\boldsymbol{x}_0)$ with respect to $\Delta\boldsymbol{x}_0$ vanish, that is, when

$$\frac{\partial\left\{[\Delta z - \mathbf{H}\Delta\boldsymbol{x}_0]^{\mathrm{T}}[\Delta z - \mathbf{H}\Delta\boldsymbol{x}_0]\right\}}{\partial\Delta\boldsymbol{x}_0} = 0$$

where the partial derivatives are to be evaluated at $\Delta\boldsymbol{x}_0 = \Delta\boldsymbol{x}_0^{\mathrm{lsq}}$.

The partial derivatives of the scalar product $\boldsymbol{\rho}\cdot\boldsymbol{\rho} \equiv \boldsymbol{\rho}^{\mathrm{T}}\boldsymbol{\rho}$ with respect to $\Delta\boldsymbol{x}_0$ can be computed by means of the following identity

$$\frac{\partial(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b})}{\partial\boldsymbol{c}} = \boldsymbol{a}^{\mathrm{T}}\left(\frac{\partial\boldsymbol{b}}{\partial\boldsymbol{c}}\right) + \boldsymbol{b}^{\mathrm{T}}\left(\frac{\partial\boldsymbol{a}}{\partial\boldsymbol{c}}\right)$$

Thus, the general solution of the linear least-squares problem may be obtained by solving the following system of m algebraic equations

$$\left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)\Delta\boldsymbol{x}_0^{\mathrm{lsq}} = \left(\mathbf{H}^{\mathrm{T}}\,\Delta z\right)$$

The solution is

$$\Delta\boldsymbol{x}_0^{\mathrm{lsq}} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\left(\mathbf{H}^{\mathrm{T}}\Delta z\right)$$

where $\mathbf{H}^{\mathrm{T}}\mathbf{H}$ is an $m \times m$ symmetric matrix. In order to compute the solution, Montenbruck and Gill [53] recommend the use of standard techniques for positive definite linear systems of equations. One of such techniques is the Cholesky decomposition, which is shown below.

Let $\mathbf{A}$ be an $m \times m$ symmetric ($\mathbf{A} = \mathbf{A}^{\mathrm{T}}$, where $\mathbf{A}^{\mathrm{T}}$ is the transpose of $\mathbf{A}$), positive definite ($\langle\mathbf{A}\boldsymbol{x}, \boldsymbol{x}\rangle > 0$ for any real vector $\boldsymbol{x}$ other than the zero vector) matrix with real entries a_{ij}, as follows

$$\mathbf{A} \equiv \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \dots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{bmatrix}$$

Then $\mathbf{A}$ has a unique decomposition of the form $\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$, where $\mathbf{L}$ is a lower triangular matrix with positive diagonal entries ($\ell_{ii} > 0$), and $\mathbf{L}^{\mathrm{T}}$ is the transpose of $\mathbf{L}$. The two matrices $\mathbf{L}$ and $\mathbf{L}^{\mathrm{T}}$ have the following entries

$$\mathbf{L} \equiv \begin{bmatrix} \ell_{11} & 0 & \dots & 0 \\ \ell_{21} & \ell_{22} & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ \ell_{m1} & \ell_{m2} & \dots & \ell_{mm} \end{bmatrix} \qquad \mathbf{L}^{\mathrm{T}} \equiv \begin{bmatrix} \ell_{11} & \ell_{21} & \dots & \ell_{m1} \\ 0 & \ell_{22} & \dots & \ell_{m2} \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \ell_{mm} \end{bmatrix}$$

For example, we want to decompose the following 3×3 symmetric matrix

$$\mathbf{A} \equiv \begin{bmatrix} 4 & 2 & -6 \\ 2 & 10 & 9 \\ -6 & 9 & 26 \end{bmatrix}$$

into the product of two matrices $\mathbf{L}$ and $\mathbf{L}^{\mathrm{T}}$ such that

$$\mathbf{L} \equiv \begin{bmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{bmatrix} \qquad \mathbf{L}^{\mathrm{T}} \equiv \begin{bmatrix} \ell_{11} & \ell_{21} & \ell_{31} \\ 0 & \ell_{22} & \ell_{32} \\ 0 & 0 & \ell_{33} \end{bmatrix}$$

By applying the rules of matrix multiplication, we have the following six equations for the six unknowns ℓ_{11}, ℓ_{21}, ℓ_{22}, ℓ_{31}, ℓ_{32}, and ℓ_{33}:

$$\begin{array}{llll} \ell_{11}^2 = 4 & \text{hence } \ell_{11} = 2 & \ell_{21}^2 + \ell_{22}^2 = 10 & \text{hence } \ell_{22} = 3 \\ \ell_{11}\ell_{21} = 2 & \text{hence } \ell_{21} = 1 & \ell_{21}\ell_{31} + \ell_{22}\ell_{32} = 9 & \text{hence } \ell_{32} = 4 \\ \ell_{11}\ell_{31} = -6 & \text{hence } \ell_{31} = -3 & \ell_{31}^2 + \ell_{32}^2 + \ell_{33}^2 = 26 & \text{hence } \ell_{33} = 1 \end{array}$$

Consequently, the Cholesky decomposition for $\mathbf{A}$ is $\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$, where

$$\mathbf{L} \equiv \begin{bmatrix} 2 & 0 & 0 \\ 1 & 3 & 0 \\ -3 & 4 & 1 \end{bmatrix} \qquad \mathbf{L}^{\mathrm{T}} \equiv \begin{bmatrix} 2 & 1 & -3 \\ 0 & 3 & 4 \\ 0 & 0 & 1 \end{bmatrix}$$

For a 3×3 matrix $\mathbf{A}$, the nonzero entries of $\mathbf{L}$ can be computed, as a function of the entries of $\mathbf{A}$, by means of the following sequence of operations:

$$(1)\quad \ell_{11} = (a_{11})^{\frac{1}{2}} \qquad (2)\quad \ell_{21} = \frac{a_{21}}{\ell_{11}} \qquad (3)\quad \ell_{31} = \frac{a_{31}}{\ell_{11}}$$

$$(4)\quad \ell_{22} = \left(a_{22} - \ell_{21}^2\right)^{\frac{1}{2}} \qquad (5)\quad \ell_{32} = \frac{a_{32} - \ell_{21}\ell_{31}}{\ell_{22}} \qquad (6)\quad \ell_{33} = \left(a_{33} - \ell_{31}^2 - \ell_{32}^2\right)^{\frac{1}{2}}$$

In the general case, for the Cholesky decomposition of an $m \times m$ symmetric, positive definite matrix $\mathbf{A}$, the following procedure can be used.

For $i = 1, 2, \ldots, m$ and $j = i + 1, i + 2, \ldots, m$:

$$\ell_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} \ell_{ik}^2\right)^{\frac{1}{2}} \qquad \ell_{ji} = \frac{1}{\ell_{ii}}\left(a_{ji} - \sum_{k=1}^{i-1} \ell_{jk}\ell_{ik}\right)$$

Since $\mathbf{A}$ is symmetric and positive definite, then the expression under square root is always positive and the entries ℓ_{ij} of $\mathbf{L}$ are all real. Therefore, to solve a system

$$\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$$

of linear algebraic equations such that $\mathbf{A}$ is a square, symmetric and positive definite matrix, we put the preceding expression in the form

$$\mathbf{L}\mathbf{L}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{b}$$

This done, the equations

$$\mathbf{L}\boldsymbol{z} = \boldsymbol{b}$$

are solved for $\boldsymbol{z}$ by forward substitution; then the equations

$$\mathbf{L}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{z}$$

are solved for $\boldsymbol{x}$ by backward substitution.

Another form of the Cholesky decomposition, which eliminates the need to compute square roots, is the following

$$\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^{\mathrm{T}}$$

where $\mathbf{A}$ is an $m \times m$ symmetric, positive definite matrix with real entries a_{ij}, $\mathbf{D}$ is a diagonal matrix with all positive nonzero entries d_{jj}, and $\mathbf{L}$ and $\mathbf{L}^{\mathrm{T}}$ are, respectively, a unit lower triangular and a unit upper triangular matrix.

It is to be noted that the matrix $\mathbf{L}^* \equiv \{\ell^*_{ij}\}$ of the decomposition $\mathbf{A} = \mathbf{L}^*\mathbf{L}^{*\mathrm{T}}$ and the matrix $\mathbf{L} \equiv \{\ell_{ij}\}$ of the decomposition $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^{\mathrm{T}}$ are not the same matrix.

For example, for $m = 3$, we have

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{bmatrix} \begin{bmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & 0 \\ 0 & 0 & d_{33} \end{bmatrix} \begin{bmatrix} 1 & \ell_{21} & \ell_{31} \\ 0 & 1 & \ell_{32} \\ 0 & 0 & 1 \end{bmatrix}$$

The entries d_{jj} of $\mathbf{D}$ and the entries ℓ_{ij} of $\mathbf{L}$ (with $i > j$) result from

$$d_{jj} = a_{jj} - \sum_{k=1}^{j-1} \ell_{jk}^2 d_{kk} \ell_{ij} = \frac{1}{d_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} \ell_{ik} \ell_{jk} d_{kk} \right)$$

For example, we want to decompose the same 3×3 symmetric positive definite matrix as that considered previously, that is,

$$\mathbf{A} \equiv \begin{bmatrix} 4 & 2 & -6 \\ 2 & 10 & 9 \\ -6 & 9 & 26 \end{bmatrix}$$

into the product $\mathbf{LDL}^{\mathrm{T}}$. There results

$$\begin{aligned} d_{11} &= a_{11} = 4 \\ \ell_{21} &= a_{21}/d_{11} = 2/4 = 1/2 \\ d_{22} &= a_{22} - \ell_{21}^2 d_{11} = 10 - (1/2)^2 \times 4 = 9 \\ \ell_{31} &= a_{31}/d_{11} = -6/4 = -1.5 \\ \ell_{32} &= (1/d_{22})(a_{32} - \ell_{31}\ell_{21}d_{11}) = 1/9 \times [9 - (-3/2) \times 1/2 \times 4) = 1.333 \\ d_{33} &= a_{33} - \ell_{31}^2 d_{11} - \ell_{32}^2 d_{22} = 26 - (-6/4)^2 \times 4 - (4/3)^2 \times 9 = 1 \end{aligned}$$

Therefore, the $\mathbf{LDL}^{\mathrm{T}}$ Cholesky decomposition of the given matrix $\mathbf{A}$ is

$$\begin{bmatrix} 4 & 2 & -6 \\ 2 & 10 & 9 \\ -6 & 9 & 26 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ -1.5 & 1.333 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0.5 & -1.5 \\ 0 & 1 & 1.333 \\ 0 & 0 & 1 \end{bmatrix}$$

The $\mathbf{L}^*\mathbf{L}^{*\mathrm{T}}$ and $\mathbf{LDL}^{\mathrm{T}}$ Cholesky decompositions of the same matrix $\mathbf{A}$ are related to each other as shown below

$$\mathbf{A} = \mathbf{LDL}^{\mathrm{T}} = \mathbf{L}\mathbf{D}^{\frac{1}{2}}\mathbf{D}^{\frac{1}{2}}\mathbf{L}^{\mathrm{T}} = \mathbf{L}\mathbf{D}^{\frac{1}{2}}\left(\mathbf{D}^{\frac{1}{2}}\right)^{\mathrm{T}}\mathbf{L}^{\mathrm{T}} = \mathbf{L}\mathbf{D}^{\frac{1}{2}}\left(\mathbf{L}\mathbf{D}^{\frac{1}{2}}\right)^{\mathrm{T}}$$

where $\mathbf{D}^{1/2}$ is also a diagonal matrix, whose nonzero entries are the square roots of the corresponding entries of $\mathbf{D}$, that is, $\mathbf{D}^{1/2} = \mathrm{diag}(d_{11}^{1/2}, d_{22}^{1/2}, \ldots, d_{mm}^{1/2})$. Hence,

$$\mathbf{A} = \mathbf{LDL}^{\mathrm{T}} = \mathbf{L}^*\mathbf{L}^{*\mathrm{T}}$$

where $\mathbf{L}^* = \mathbf{L}\,\mathbf{D}^{1/2}$.

Now, let us come back to the linear least-squares problem. Since $\boldsymbol{h}(\boldsymbol{x}_0)$ is actually a nonlinear function, the value $\Delta\boldsymbol{x}_0^{\mathrm{lsq}}$ computed as shown above (and consequently

$\boldsymbol{x}_0^{\text{lsq}} = \boldsymbol{x}_0^{\text{ref}} + \Delta\boldsymbol{x}_0^{\text{lsq}}$) is only an approximate solution of the orbit determination problem. A better solution than this can be obtained by replacing $\boldsymbol{x}_0^{\text{ref}}$ with the value of $\boldsymbol{x}_0^{\text{lsq}}$ computed previously and performing a new iteration. Let the superscripts $[i]$ and $[i + 1]$ denote the value of $\boldsymbol{x}_0^{\text{lsq}}$ resulting from, respectively, the i^{th} and the $(i + 1)^{\text{th}}$ iteration. The iterative process of refinement can be expressed as follows

$$\boldsymbol{x}_0^{[i+1]} = \boldsymbol{x}_0^{[i]} + \left(\mathbf{H}^{[i]\text{T}}\mathbf{H}^{[i]}\right)^{-1}\mathbf{H}^{[i]\text{T}}\left[\boldsymbol{z} - \boldsymbol{h}\left(\boldsymbol{x}_0^{[i]}\right)\right]$$

which formula has $\boldsymbol{x}_0^{\text{appr}}$ as its starting point, that is, $\boldsymbol{x}_0^{[0]} = \boldsymbol{x}_0^{\text{appr}}$, and continues till the difference $\boldsymbol{x}_0^{[i+1]} - \boldsymbol{x}_0^{[i]}$ is greater in magnitude than a desired tolerance.

For the best convergence of the iterative process, the Jacobian $\mathbf{H}^{[i]}$ must be computed at each iteration; however, Montenbruck and Gill [53] observe that $\mathbf{H}^{[i]}$ can be replaced by a constant $\mathbf{H}^{[0]}$. This replacement implies a higher number of iterations, but results often in less computational work than would be required if $\mathbf{H}^{[i]}$ were computed at each step. The method shown above does not take into account the different errors by which the measurements may be affected. However, this method can be made general by weighting each observation z_i with the inverse of its mean measurement error σ_i (where $i = 1, 2, \ldots, n$).

By so doing, each residual ρ_i is replaced by the corresponding normalised residual ρ^*_i defined as follows

$$\rho_i^* = \frac{\rho_i}{\sigma_i} = \frac{[\boldsymbol{z} - \boldsymbol{h}(\boldsymbol{x}_0)]_i}{\sigma_i}$$

The mean measurement error σ_i includes both the random and systematic errors. A typical example of the latter type of error is provided by the refraction of light rays due to the Earth atmosphere. By using the normalised residuals ρ^*_i instead of the ordinary residuals ρ_i, the expression $\Delta\boldsymbol{x}_0^{\text{lsq}} = (\mathbf{H}^{\text{T}}\,\mathbf{H})^{-1}\,(\mathbf{H}^{\text{T}}\,\Delta\boldsymbol{z})$ shown above remains formally identical, the only difference being that in the present case (weighted observations) the Jacobian $\mathbf{H}$ and the difference vector $\Delta\boldsymbol{z}$ must be replaced by their respective counterparts $\mathbf{H}^*$ and $\Delta\boldsymbol{z}^*$, as follows

$$\Delta\boldsymbol{x}_0^{\text{lsq}} = \left(\mathbf{H}^{*\text{T}}\mathbf{H}^*\right)^{-1}\left(\mathbf{H}^{*\text{T}}\Delta\boldsymbol{z}^*\right)$$

where $\mathbf{H}^* = \boldsymbol{\Sigma}\,\mathbf{H}$ and $\Delta\boldsymbol{z}^* = \boldsymbol{\Sigma}\,\Delta\boldsymbol{z}$; $\boldsymbol{\Sigma}$ is the $n \times n$ diagonal matrix filled with zeros, except the elements placed along its main diagonal, which are $\sigma_1^{-1}, \ldots, \sigma_n^{-1}$.

In other words, $\boldsymbol{\Sigma} \equiv \text{diag}(\sigma_1^{-1}, \ldots, \sigma_n^{-1})$ or, which is the same,

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} 1/\sigma_1 & 0 & \ldots & 0 \\ 0 & 1/\sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1/\sigma_n \end{bmatrix}$$

The solution given above

$$\Delta \boldsymbol{x}_0^{\mathrm{lsq}} = \left(\mathbf{H}^{*\mathrm{T}}\mathbf{H}^{*}\right)^{-1}\left(\mathbf{H}^{*\mathrm{T}}\Delta \boldsymbol{z}^{*}\right)$$

of the problem of weighted least-squares estimation may also be expressed in terms of the original Jacobian matrix $\mathbf{H}$ and the original difference vector $\Delta \boldsymbol{z}$ (instead of $\mathbf{H}^* = \boldsymbol{\Sigma}\,\mathbf{H}$ and $\Delta \boldsymbol{z}^* = \boldsymbol{\Sigma}\,\Delta \boldsymbol{z}$) as follows

$$\Delta \boldsymbol{x}_0^{\mathrm{lsq}} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\mathbf{H}\right)^{-1}\left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\Delta \boldsymbol{z}\right)$$

where $\mathbf{W}$ is the weighting matrix defined as $\mathbf{W} \equiv \boldsymbol{\Sigma}^2 \equiv \mathrm{diag}(\sigma_1^{-2}, \ldots, \sigma_n^{-2})$, or, which is the same,

$$\mathbf{W} \equiv \boldsymbol{\Sigma}^2 \equiv \begin{bmatrix} 1/\sigma_1^2 & 0 & \ldots & 0 \\ 0 & 1/\sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1/\sigma_n^2 \end{bmatrix}$$

The preceding definition of $\mathbf{W}$ as a diagonal matrix holds in case of uncorrelated errors of measurement. Should such errors be correlated, then $\mathbf{W}$ would become a non-diagonal matrix. In order to understand the correlation of errors, it is necessary to have some concepts of probability theory, which are given below.

As has been shown above, if $\boldsymbol{x}_0$ and $\boldsymbol{\varepsilon}$ designate, respectively, the actual (augmented) state vector at epoch and the vector containing the measurement errors, then the observation vector $\boldsymbol{z}$ results from

$$\boldsymbol{z} = \boldsymbol{h}(\boldsymbol{x}_0) + \boldsymbol{\varepsilon}$$

The preceding expression, after linearisation, becomes

$$\Delta \boldsymbol{z} = \mathbf{H}\left(\boldsymbol{x}_0 - \boldsymbol{x}_0^{\mathrm{ref}}\right) + \boldsymbol{\varepsilon}$$

where $\boldsymbol{x}_0^{\mathrm{ref}}$ is a reference state vector sufficiently close to $\boldsymbol{x}_0$. The solution of this least-squares problem has been shown to be

$$\begin{aligned} \boldsymbol{x}_0^{\mathrm{lsq}} &= \boldsymbol{x}_0^{\mathrm{ref}} + \left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H}\right)^{-1}\left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\Delta \boldsymbol{z}\right) \\ &= \boldsymbol{x}_0 + \left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H}\right)^{-1}\left(\mathbf{H}^{\mathrm{T}}\,\mathbf{W}\boldsymbol{\varepsilon}\right) \end{aligned}$$

The preceding expression shows that, when measurement errors are committed, the computed state vector $\boldsymbol{x}_0^{\mathrm{lsq}}$ differs from the actual state vector $\boldsymbol{x}_0$.

In the event of systematic errors being negligible, the components of $\boldsymbol{\varepsilon}$ are only random errors. Let X be a discrete random variable which takes values in $S = \{x_1, x_2, \ldots, x_n\}$. The expected value (or mean) of X is defined as follows

$$E(X) = \sum_{i=1}^{n} x_i \, p(x_i)$$

where $p(x_i)$ is the probability of x_i. The mean of X is also denoted by $\overline{X}$ or by μ_X.

The preceding definition holds if the sum $x_1 \; p(x_1) + x_2 \; p(x_2) + \cdots + x_n \; p(x_n)$ converges to a finite value. Otherwise, the expected value of X is undefined.

For example, when a die is cast, the probability of each of the six possible events x_i (with $i = 1, 2, \ldots, 6$) is $p(x_i) = 1/6$. Thus, according to the preceding definition, the expected value is $E(X) = (1 + 2 + 3 + 4 + 5 + 6) \times (1/6) = 21/6 = 7/2 = 3.5$.

Let X be a continuous random variable in $S = [-\infty, \infty]$ with probability density function $\varphi(x)$. In this case, the expected value of X is defined as follows

$$E(X) = \int_{-\infty}^{\infty} x \, \varphi(x) \, \mathrm{d}x$$

Here, too, the value of the integral must be finite, in order for the expected value of X to be defined. Let X and $E(X)$ be, respectively, a random variable and its expected value. The variance of X is defined as follows

$$Var(X) = E\Big\{ [X - E(X)]^2] \Big\}$$

In other words, the variance of X is the expected squared distance of X from its mean $E(X)$. Let X be a discrete random variable which takes values in $S = \{x_1, x_2, \ldots, x_n\}$. Then the variance of X is defined as follows

$$Var(X) = \sum_{i=1}^{n} \left[x_i - \sum_{j=1}^{n} x_j \, p\big(x_j\big) \right]^2 p(x_i)$$

In the example of the die proposed above, $p(x_i) = 1/6$ and $E(X) = 7/2$.

Thus, according to the preceding definition, the variance of X is

$$Var(X) = \left[\left(1 - \frac{7}{2}\right)^2 + \left(2 - \frac{7}{2}\right)^2 + \left(3 - \frac{7}{2}\right)^2 + \left(4 - \frac{7}{2}\right)^2 + \left(5 - \frac{7}{2}\right)^2 + \left(6 - \frac{7}{2}\right)^2 \right] \times \frac{1}{6} = \left(\frac{25}{4} + \frac{9}{4} + \frac{1}{4} + \frac{1}{4} + \frac{9}{4} + \frac{25}{4} \right) \times \frac{1}{6} = \frac{35}{12}$$

Likewise, let X be a continuous random variable in $S = [-\infty, \infty]$ with probability density function $\varphi(x)$. Then the variance of X is

$$Var(X) = \int_{-\infty}^{\infty} \left[x - \int_{-\infty}^{\infty} \xi \varphi(\xi)\, \mathrm{d}\xi \right]^2 \varphi(x)\, \mathrm{d}x$$

In case of the expected value $E(X)$ of a random variable X being zero, that is, in case of $E(X) = 0$, then the variance of X is

$$Var(X) = E\left\{ [X - E(X)]^2] \right\} = E\left[(X - 0)^2] \right] = E\left[(X)^2 \right]$$

The variance of a random variable X is often denoted by σ_X^2, that is, $Var(X) \equiv \sigma_X^2$, and its standard deviation $[Var(X)]^{1/2}$ by σ_X.

Let X and Y be two independent random variables. Let $E(X)$ and $E(Y)$ be their respective expected values. In this case, there results

$$Var(X + Y) = Var(X) + Var(Y)$$

Instead, in case of X and Y being dependent random variables, then there results

$$Var(X + Y) = Var(X) + Var(Y) + 2\, E\{[X - E(X)][Y - E(Y)]\}$$

where the term $E\{[X - E(X)][Y - E(Y)]\}$ is called the covariance of X and Y, that is,

$$Cov(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$$

The covariance of two random variables X and Y is the average value of the deviation of X from its mean $E(X)$ and the deviation of Y from its mean $E(Y)$. The zero value of the covariance $Cov(X, Y)$ of two random variables X and Y does not imply their independence; it implies only the linear independence of X and Y.

The correlation coefficient $\rho_{XY} \equiv Corr(X, Y)$ of two random variables X and Y is defined as follows

$$Corr(X, Y) = \frac{Cov(X, Y)}{[Var(X) Var(Y)]^{\frac{1}{2}}}$$

$Corr(X, Y)$ is a dimensionless quantity having the following properties:

$$-1 \leq Corr(X, Y) \leq 1$$
$$Corr(aX + b, cY + d) = Corr(X, Y)$$

where a, b, c, and d are any real constants. This implies that $Corr(X, Y)$ is equal to unity only when a relation of a linear dependence exists between X and Y.

Let us see now the practical application of the concepts given above. If the error vector $\boldsymbol{\varepsilon}$ has only random errors as its components, then the expected values of, respectively, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}$ are such that

$$\begin{aligned} E(\boldsymbol{\varepsilon}) &= \mathbf{0} \\ E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}) &= \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2) \end{aligned}$$

In other words, the expected value $E(\varepsilon_i)$ of each component ε_i ($i = 1, 2, \ldots, n$) of the error vector $\boldsymbol{\varepsilon}$ is equal to zero; all these components are uncorrelated, in other words $\varepsilon_i\, \varepsilon_j = 0$ for $i \neq j$; and the standard deviation σ_i of ε_i is equal to $[E(\varepsilon_i^2)]^{1/2}$.

The least-squares solution $\boldsymbol{x}_0^{\mathrm{lsq}}$ of the orbit determination problem is a random variable, because it depends on $\boldsymbol{\varepsilon}$, which is a random variable by the hypothesis made above. Then, $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ implies that the expected value of $\boldsymbol{x}_0^{\mathrm{lsq}}$ is equal to the actual value of $\boldsymbol{x}_0$, because

$$E\left(\boldsymbol{x}_0^{\mathrm{lsq}}\right) = E\left[\boldsymbol{x}_0 + \left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\mathbf{H}\right)^{-1}\left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\boldsymbol{\varepsilon}\right)\right] = \boldsymbol{x}_0 + \left[\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\mathbf{H}\right]^{-1}\left[\mathbf{H}^{\mathrm{T}}\mathbf{W}\,E(\boldsymbol{\varepsilon})\right] = \boldsymbol{x}_0$$

In addition, since by definition $Cov(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$, then

$$\begin{aligned} Cov\left(\boldsymbol{x}_0^{\mathrm{lsq}}, \boldsymbol{x}_0^{\mathrm{lsq}}\right) &= E\left\{\left[\boldsymbol{x}_0^{\mathrm{lsq}} - E\left(\boldsymbol{x}_0^{\mathrm{lsq}}\right)\right]\left[\boldsymbol{x}_0^{\mathrm{lsq}} - E\left(\boldsymbol{x}_0^{\mathrm{lsq}}\right)\right]^{\mathrm{T}}\right\} \\ &= E\left\{\left[\boldsymbol{x}_0^{\mathrm{lsq}} - \boldsymbol{x}_0\right]\left[\boldsymbol{x}_0^{\mathrm{lsq}} - \boldsymbol{x}_0\right]^{\mathrm{T}}\right\} \end{aligned}$$

Substituting

$$\boldsymbol{x}_0 + \left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\mathbf{H}\right)^{-1}\left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\boldsymbol{\varepsilon}\right)$$

for $\boldsymbol{x}_0^{\mathrm{lsq}}$ into the preceding expression yields

$$\begin{aligned} Cov\left(\boldsymbol{x}_0^{\mathrm{lsq}}, \boldsymbol{x}_0^{\mathrm{lsq}}\right) &= E\left\{\left[\boldsymbol{x}_0 + \left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\mathbf{H}\right)^{-1}\left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\boldsymbol{\varepsilon}\right) - \boldsymbol{x}_0\right]\left[\boldsymbol{x}_0 + \left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\mathbf{H}\right)^{-1}\left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\boldsymbol{\varepsilon}\right) - \boldsymbol{x}_0\right]^{\mathrm{T}}\right\} \\ &= E\left\{\left[\left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\mathbf{H}\right)^{-1}\left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\boldsymbol{\varepsilon}\right)\right]\left[\left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\mathbf{H}\right)^{-1}\left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\boldsymbol{\varepsilon}\right)\right]^{\mathrm{T}}\right\} \\ &= \left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\mathbf{H}\right)^{-1}\left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\right)E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\right)\left(\mathbf{W}\,\mathbf{H}\right)\left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\mathbf{H}\right)^{-1} \end{aligned}$$

The expression written above can be simplified, if the weighting matrix $\mathbf{W}$ is chosen in accordance with the standard deviation σ_i (where $i = 1, 2, \ldots, n$) of the measurement. In this case, $\mathbf{W} = \operatorname{diag}(\sigma_1^{-2}, \ldots, \sigma_n^{-2})$ is the inverse of $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}})$, and the preceding expression reduces to

$$Cov\left(\boldsymbol{x}_0^{\mathrm{lsq}}, \boldsymbol{x}_0^{\mathrm{lsq}}\right) = \left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\mathbf{H}\right)^{-1}$$

Those elements of the covariance matrix which are placed along its main diagonal are the standard deviation

$$\sigma\left(x_{0i}^{\mathrm{lsq}}\right) = \left[Cov\left(x_{0i}^{\mathrm{lsq}}, x_{0i}^{\mathrm{lsq}}\right)\right]^{\frac{1}{2}}$$

($i = 1, 2, \ldots, n$) of the components of the $\boldsymbol{x}_0^{\mathrm{lsq}}$ vector. Likewise, the off-diagonal elements of the covariance matrix provide a measure of the correlation existing between errors of individual components.

The expected value and the covariance of $\boldsymbol{x}_0^{\mathrm{lsq}}$ define the distribution of values of $\boldsymbol{x}_0^{\mathrm{lsq}}$ which would result in an experiment of repeated orbit determinations for the same trajectory, if the measurement errors were only of the random type. If these errors have a normal distribution, then there is a probability of 67% that $\boldsymbol{x}_0^{\mathrm{lsq}}$ (resulting from the actual measurements) deviates from $\boldsymbol{x}_0$ by less than 1σ, and a probability of 99.7% that $\boldsymbol{x}_0^{\mathrm{lsq}}$ deviates from $\boldsymbol{x}_0$ by less than 3σ.

In the presence of systematic errors $\boldsymbol{\varepsilon}^*$, there is a further deviation

$$\delta \boldsymbol{x}_0^{\mathrm{lsq}} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\mathbf{H}\right)^{-1}\left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\boldsymbol{\varepsilon}^*\right)$$

of $\boldsymbol{x}_0^{\mathrm{lsq}}$ from $\mathbf{x}_0$. Montenbruck and Gill [53] point out that the measurement standard deviation $\sigma(\boldsymbol{\varepsilon})$ must be known in order to construct the weighting matrix $\mathbf{W}$.

The analysis performed so far is based on the assumption of a Gaussian distribution of errors in the observations. However, this analysis (based only on data noise errors) does not take into account the effect of model errors. It is therefore necessary to take into account the effect due to systematic errors.

To this end, the measurement equation can be rewritten as follows

$$\boldsymbol{z} = \boldsymbol{h}(\boldsymbol{x}_0, \boldsymbol{c}) + \boldsymbol{\varepsilon}$$

where $\boldsymbol{z}$ is the n-dimensional vector containing the observations at times $t_1, t_2, \ldots, t_n$; $\boldsymbol{x}_0$ is the vector of the estimated parameters; $\boldsymbol{c}$ is the vector of the so-called consider parameters (which affect the estimated parameters); $\boldsymbol{h}$ is a vector-valued function of $\boldsymbol{x}_0$ and $\boldsymbol{c}$; and $\boldsymbol{\varepsilon}$ is the vector containing the measurement noise errors.

The vector $\boldsymbol{c}$ contains the force and measurement parameters which are uncertain but are not modified as a result of the least-squares estimation process. The expected value of these parameters can be assumed, without loss of generality, equal to zero. Previously both the estimated and the consider parameters have been taken together and the expression

$$\boldsymbol{z} = \boldsymbol{h}(\boldsymbol{x}_0) + \boldsymbol{\varepsilon}$$

of the observation vector $\boldsymbol{z}$ has been linearised around the reference state vector $\boldsymbol{x}_0^{\mathrm{ref}}$ to obtain

$$\Delta \boldsymbol{z} = \mathbf{H}\left(\boldsymbol{x}_0 - \boldsymbol{x}_0^{\mathrm{ref}}\right) + \boldsymbol{\varepsilon}$$

Just in the same way, now the expression

$$z = \boldsymbol{h}(\boldsymbol{x}_0, \boldsymbol{c}) + \boldsymbol{\varepsilon}$$

is linearised around $\boldsymbol{x}_0^{\text{ref}}$ to obtain

$$\Delta z = \mathbf{H}_x\left(\boldsymbol{x}_0 - \boldsymbol{x}_0^{\text{ref}}\right) + \mathbf{H}_c\boldsymbol{c} + \boldsymbol{\varepsilon}$$

where $\mathbf{H}_x$ is the Jacobian matrix containing the partial derivatives of the vector-valued function $\boldsymbol{h}$ with respect to $\boldsymbol{x}_0$, and $\mathbf{H}_c$ is the Jacobian matrix containing the partial derivatives of $\boldsymbol{h}$ with respect to $\boldsymbol{c}$. Now the expression of the least-squares solution becomes

$$\boldsymbol{x}_0^{\text{lsq}} = \boldsymbol{x}_0 + \left(\mathbf{H}_x^{\mathrm{T}}\mathbf{W}\,\mathbf{H}_x\right)^{-1}\mathbf{H}_x^{\mathrm{T}}\,\mathbf{W}(\mathbf{H}_c\boldsymbol{c} + \boldsymbol{\varepsilon})$$

This solution differs from the true value $\boldsymbol{x}_0$ of the estimation parameters by a quantity which depends on the consider parameters (components of $\boldsymbol{c}$) and the measurement noise (components of $\boldsymbol{\varepsilon}$). The consider parameters are assumed to be random variables with zero mean and covariance matrix $\mathbf{C}$, which are uncorrelated with the measurement noise, so that the equality $E(\boldsymbol{c}\boldsymbol{\varepsilon}^{\mathrm{T}}) = \mathbf{0}$ holds. In this case, the expected value of the least-squares solution

$$E\left(\boldsymbol{x}_0^{\text{lsq}}\right) = \boldsymbol{x}_0 + \left(\mathbf{H}_x^{\mathrm{T}}\mathbf{W}\,\mathbf{H}_x\right)^{-1}\mathbf{H}_x^{\mathrm{T}}\mathbf{W}[\mathbf{H}_c\,E(\boldsymbol{c}) + E(\boldsymbol{\varepsilon})] = \boldsymbol{x}_0$$

is equal to the true state vector $\boldsymbol{x}_0$.

The consider covariance matrix $\mathbf{P}^c$ is larger than the noise-only covariance matrix $\mathbf{P} = (\mathbf{H}_x^{\mathrm{T}}\ \mathbf{W}\ \mathbf{H}_x)^{-1}$, which is also called formal or computed covariance matrix. The consider covariance matrix has the following expression

$$\begin{aligned}\mathbf{P}^c &= \left(\mathbf{H}\,\mathbf{H}_x^{\mathrm{T}}\mathbf{W}\right)\left[\mathbf{H}_c\mathbf{C}\,\mathbf{H}_c^{\mathrm{T}} + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}})\right]\left(\mathbf{P}\,\mathbf{H}_x^{\mathrm{T}}\mathbf{W}\right)^{\mathrm{T}} \\ &= \mathbf{P} + (\mathbf{P}\,\mathbf{H}_x^{\mathrm{T}}\mathbf{W})(\mathbf{H}_c\mathbf{C}\,\mathbf{H}_c^{\mathrm{T}})\left(\mathbf{P}\,\mathbf{H}_x^{\mathrm{T}}\mathbf{W}\right)^{\mathrm{T}}\end{aligned}$$

where the weighting matrix $\mathbf{W}$ has been taken as the inverse of the measurement covariance matrix.

As has been shown above, an approximate value ($\boldsymbol{x}_0^{\text{appr}}$) of the actual state vector ($\boldsymbol{x}_0$) at epoch must be known to start the least-squares orbit determination process. Some information on the accuracy of this value is often available. We want to incorporate the a priori covariance matrix $\mathbf{P}_0^{\text{apr}}$ into the least-squares estimation. To his end, the loss function shown above

$$J(\boldsymbol{x}_0) = \boldsymbol{\rho}\cdot\boldsymbol{\rho} \equiv \boldsymbol{\rho}^{\mathrm{T}}\boldsymbol{\rho} = (\Delta z - \mathbf{H}\,\Delta\boldsymbol{x}_0)^{\mathrm{T}}(\Delta z - \mathbf{H}\,\Delta\boldsymbol{x}_0)$$

is represented in another way. Remembering that

$$\Delta\boldsymbol{x}_0^{\text{lsq}} = \left(\mathbf{H}^{\mathrm{T}}\,\mathbf{H}\right)^{-1}\left(\mathbf{H}^{\mathrm{T}}\Delta z\right)$$

the loss function may be written as follows

$$J(\boldsymbol{x}_0) = \left(\Delta\boldsymbol{x}_0 - \Delta\boldsymbol{x}_0^{\text{lsq}}\right)^{\text{T}}(\mathbf{H}^{\text{T}}\,\mathbf{H})\left(\Delta\boldsymbol{x}_0 - \Delta\boldsymbol{x}_0^{\text{lsq}}\right) + \left(\Delta\boldsymbol{z}^{\text{T}}\Delta\boldsymbol{z} - \Delta\boldsymbol{x}_0^{lsq\,\text{T}}\mathbf{H}^{\text{T}}\mathbf{H}\,\Delta\boldsymbol{x}_0^{\text{lsq}}\right)$$
$$= \left(\boldsymbol{x}_0 - \boldsymbol{x}_0^{\text{lsq}}\right)^{\text{T}}\mathbf{P}_0{}^{-1}\left(\boldsymbol{x}_0 - \boldsymbol{x}_0^{\text{lsq}}\right) + \text{constant}$$

Since this expression of the loss function $J(\boldsymbol{x}_0)$ is a quadratic form of $(\boldsymbol{x}_0 - \boldsymbol{x}_0^{\text{lsq}})$ defined by the inverse covariance matrix $\mathbf{P}_0^{-1} = (\mathbf{H}^{\text{T}}\,\mathbf{H})$ of $(\boldsymbol{x}_0 - \boldsymbol{x}_0^{\text{lsq}})$, then the loss function minimum and the covariance matrix provide the same information on the least-squares estimation as that which comes from the measurement vector $\Delta\boldsymbol{z}$ and the Jacobian matrix $\mathbf{H}$. Thus, an a priori estimate

$$\boldsymbol{x}_0^{\text{apr}} = \boldsymbol{x}_0^{\text{ref}} + \Delta\boldsymbol{x}_0^{\text{apr}}$$

of the state vector $\boldsymbol{x}_0$ may come from a modified loss function

$$J = \left(\boldsymbol{x}_0 - \boldsymbol{x}_0^{\text{apr}}\right)^{\text{T}}\boldsymbol{\Lambda}\left(\boldsymbol{x}_0 - \boldsymbol{x}_0^{\text{apr}}\right) + \boldsymbol{\rho}^{\text{T}}\boldsymbol{\rho}$$

where $\boldsymbol{\Lambda} = (\mathbf{P}_0^{\text{apr}})^{-1}$, called information matrix, is used to attribute a contribution to the loss function to each deviation from $\boldsymbol{x}_0^{\text{apr}}$, and $\boldsymbol{\rho}$ is the vector containing the residuals. The information matrix $\boldsymbol{\Lambda}$ is always positive semi-definite, because it is the inverse of the covariance matrix. By the way, let $\mathbf{A}$ be a matrix with real entries a_{ij}. $\mathbf{A}$ is said to be positive semi-definite if, for any vector $\boldsymbol{x}$ with real components x_i, the dot product $\langle \mathbf{A}\boldsymbol{x}, \boldsymbol{x}\rangle$ of $\mathbf{A}\boldsymbol{x}$ and $\boldsymbol{x}$ is non-negative, that is, if

$$\langle \mathbf{A}\,\boldsymbol{x}, \boldsymbol{x}\rangle \geq 0$$

Consequently, $\boldsymbol{\Lambda}$ can be factored to form a product $\boldsymbol{\Lambda} = \mathbf{S}^{\text{T}}\mathbf{S}$, which makes it possible to determine the minimum of the loss function J. If J is written as follows

$$J = (\Delta\boldsymbol{x}_0 - \Delta\boldsymbol{x}_0^{\text{apr}})^{\text{T}}\boldsymbol{\Lambda}(\Delta\boldsymbol{x}_0 - \Delta\boldsymbol{x}_0^{\text{apr}}) + (\Delta\boldsymbol{z} - \mathbf{H}\,\Delta\boldsymbol{x}_0)^{\text{T}}(\Delta\boldsymbol{z} - \mathbf{H}\,\Delta\boldsymbol{x}_0) = \mathbf{A}^{\text{T}}\mathbf{A}$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{S}\Delta\boldsymbol{x}_0^{\text{apr}} \\ \Delta\boldsymbol{z} \end{bmatrix} - \begin{bmatrix} \mathbf{S} \\ \mathbf{H} \end{bmatrix}\Delta\boldsymbol{x}_0$$

then the information matrix $\boldsymbol{\Lambda}$ can be considered as the result of additional observations. Consequently, the minimum of the combined loss function results, after simplification, from

$$\Delta\mathbf{x}_0{}^{\text{lsq}} = \left(\boldsymbol{\Lambda} + \mathbf{H}^{\text{T}}\,\mathbf{H}\right)^{-1}\left(\boldsymbol{\Lambda}\,\Delta\boldsymbol{x}_0^{\text{apr}} + \mathbf{H}^{\text{T}}\,\Delta\boldsymbol{z}\right)$$

In case of weighted observations, the preceding expression becomes

$$\Delta \boldsymbol{x}_0^{\mathrm{lsq}} = \left(\boldsymbol{\Lambda} + \mathbf{H}^{\mathrm{T}}\mathbf{W}\,\mathbf{H}\right)^{-1}\left(\boldsymbol{\Lambda}\,\Delta \boldsymbol{x}_0^{\mathrm{apr}} + \mathbf{H}^{\mathrm{T}}\mathbf{W}\,\Delta z\right)$$

where $\mathbf{W}$ is the weighting matrix. In order for the preceding expression to be computable, the sum $\boldsymbol{\Lambda} + \mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H}$ must have a nonzero determinant, without the need for $\boldsymbol{\Lambda}$ or $\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H}$ to be non-singular (by the way, an $n \times n$ matrix $\mathbf{A}$ is said to be non-singular or invertible when there exists an $n \times n$ matrix $\mathbf{B} \equiv \mathbf{A}^{-1}$ such that $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$, where $\mathbf{I}$ is the $n \times n$ identity matrix). However, the non-singularity of the information matrix $\boldsymbol{\Lambda}$ is sufficient to ensure the resolvability of the equations independently of $\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H}$. In order to take advantage of this fact, a singularity in least-squares problems can be avoided by giving a small a priori weight to each estimation parameter and adding the corresponding diagonal matrix $\boldsymbol{\Lambda}$ to the normal equations matrix.

The expected value $\Delta \boldsymbol{x}_0^{\mathrm{lsq}} = (\boldsymbol{\Lambda} + \mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H})^{-1}\ (\boldsymbol{\Lambda}\ \Delta \boldsymbol{x}_0^{\mathrm{apr}} + \mathbf{H}^{\mathrm{T}}\mathbf{W}\ \Delta z)$ of the estimated state is equal to the actual state $\boldsymbol{x}_0$, if the a priori information $\boldsymbol{x}_0^{\mathrm{apr}}$ is also a random variable having $\boldsymbol{x}_0$ as its mean value.

The covariance matrix $\mathbf{P}_0$ of the estimate is related to the a priori covariance and the measurement information matrix by the following expression

$$(\mathbf{P}_0)^{-1} = (\mathbf{P}_0^{\mathrm{apr}})^{-1} + (\mathbf{H}^{\mathrm{T}}\mathbf{W}\mathbf{H})$$

2.10 Numerical Solution of the Least-Squares Estimation Problem

The purpose of the present paragraph is to provide the reader of this book with the basic concepts which are necessary to solve numerically the least-squares problem described in the preceding paragraph. Those among the readers who are fully conversant with such concepts can skip this paragraph and go directly to the next. For the sake of generality, the system of m normal equations of Sect. 2.9, that is, $(\mathbf{H}^{\mathrm{T}}\ \mathbf{H})\Delta \boldsymbol{x}_0^{\mathrm{lsq}} = (\mathbf{H}^{\mathrm{T}}\ \Delta z)$, is written here in the following form

$$\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)\boldsymbol{x} = \left(\mathbf{A}^{\mathrm{T}}\boldsymbol{b}\right)$$

where $\mathbf{A}$, $\boldsymbol{x}$, and $\boldsymbol{b}$ take the places of, respectively, $\mathbf{H}$, $\Delta \boldsymbol{x}_0^{\mathrm{lsq}}$, and Δz.

The numerical procedures to be described below are based on the decomposition (also called factorisation) of a given non-singular $n \times m$ matrix $\mathbf{A}$ (with $n \geq m$) into an orthogonal $n \times n$ matrix $\mathbf{Q}$ and an upper triangular $m \times m$ matrix $\mathbf{R}$. As the

bottom $(n - m)$ rows of an $n \times m$ upper triangular matrix contain only zeroes, as will be shown below, it is customary to write

$$\mathbf{A}_{n\times m} = \mathbf{Q}_{n\times n}\begin{bmatrix}\mathbf{R}_{m\times m}\\ \mathbf{0}_{(n-m)\times m}\end{bmatrix}$$

where $\mathbf{0}$ is an $(n - m) \times m$ partition, containing only zeroes, of the given matrix $\mathbf{A}$.

According to the definition given by Olver and Shakiban [60], an orthogonal (or orthonormal) matrix $\mathbf{Q}$ is a square matrix which satisfies the condition $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$, where $\mathbf{Q}^{\mathrm{T}}$ is the transpose of the given matrix and $\mathbf{I}$ is the identity matrix. Since the inverse matrix $\mathbf{Q}^{-1}$ must satisfy the condition $\mathbf{Q}^{-1}\,\mathbf{Q} = \mathbf{I}$, then a square matrix $\mathbf{Q}$ is orthogonal if and only if its transpose $\mathbf{Q}^{\mathrm{T}}$ is equal to its inverse $\mathbf{Q}^{-1}$.

In other words, a square matrix is orthogonal if and only if its columns form an orthonormal basis with respect to the Euclidean scalar (or inner or dot) product on an n-dimensional Euclidean space. This is because, if $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n$ are the columns of $\mathbf{Q}$, then $\boldsymbol{q}_1^{\mathrm{T}}, \boldsymbol{q}_2^{\mathrm{T}}, \ldots, \boldsymbol{q}_n^{\mathrm{T}}$ will be the rows of the transpose matrix $\mathbf{Q}^{\mathrm{T}}$. Now, the (i, j) entry of the product $\mathbf{Q}^{\mathrm{T}}\mathbf{Q}$ results from the product between the i^{th} row of $\mathbf{Q}^{\mathrm{T}}$ and the j^{th} column of $\mathbf{Q}$. Since $\mathbf{Q}$ is, by hypothesis, an orthogonal matrix (such that $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$), then the following relations must hold

$$\boldsymbol{q}_i \cdot \boldsymbol{q}_j = \boldsymbol{q}_i^{\mathrm{T}}\,\boldsymbol{q}_j = \begin{cases} 1 & (i = j)\\ 0 & (i \neq j)\end{cases}$$

which are just the conditions for $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n$ to form an orthonormal basis.

Guillemin [31] gives the following numerical example of an orthogonal matrix of order three

$$\mathbf{Q} \equiv \begin{bmatrix} 0.5 & 0.5 & 0.707\\ -0.707 & 0.707 & 0\\ -0.5 & -0.5 & 0.707\end{bmatrix}$$

where 0.707 approximates $\frac{1}{2}(2)^{1/2}$. As is easy to verify, the three columns

$$\boldsymbol{q}_1 \equiv \begin{bmatrix} 0.5\\ -0.707\\ -0.5\end{bmatrix} \quad \boldsymbol{q}_2 \equiv \begin{bmatrix} 0.5\\ 0.707\\ -0.5\end{bmatrix} \quad \boldsymbol{q}_3 \equiv \begin{bmatrix} 0.707\\ 0\\ 0.707\end{bmatrix}$$

of the 3×3 matrix $\mathbf{Q}$ indicated above satisfy the conditions $\boldsymbol{q}_i \cdot \boldsymbol{q}_j = 1$ (for $i = j$) and $\boldsymbol{q}_i \cdot \boldsymbol{q}_j = 0$ (for $i \neq j$). An orthogonal matrix $\mathbf{Q}$ has determinant $\det(\mathbf{Q}) = \pm 1$. This is because $\mathbf{Q}$, as an orthogonal matrix, must satisfy the condition $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$; taking the determinant of this equality yields

$$1 = \det(\mathbf{I}) = \det(\mathbf{Q}^{\mathrm{T}}\mathbf{Q}) = \det\left(\mathbf{Q}^{\mathrm{T}}\right)\det(\mathbf{Q}) = [\det(\mathbf{Q})]^2$$

If the determinant of an $n \times n$ orthogonal matrix is equal to +1, then that matrix is called proper and the corresponding orthonormal basis is a right-handed basis on

an n-dimensional Euclidean space; whereas, if the determinant of an orthogonal matrix is equal to -1, then that matrix is called improper and the corresponding orthonormal basis is a left-handed basis on the same space.

The product of two orthogonal matrices is also an orthogonal matrix. This is because, if $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are two orthogonal matrices, that is, two matrices such that

$$\mathbf{Q}_1^{\mathrm{T}}\mathbf{Q}_1 = \mathbf{I}$$
$$\mathbf{Q}_2^{\mathrm{T}}\mathbf{Q}_2 = \mathbf{I}$$

then there results $(\mathbf{Q}_1\ \mathbf{Q}_2)^{\mathrm{T}}\ (\mathbf{Q}_1\ \mathbf{Q}_2) = \mathbf{Q}_2^{\mathrm{T}}\ \mathbf{Q}_1^{\mathrm{T}}\ \mathbf{Q}_1\ \mathbf{Q}_2 = \mathbf{Q}_2^{\mathrm{T}}\ \mathbf{Q}_2 = \mathbf{I}$. Thus, the product matrix $\mathbf{Q}_1\ \mathbf{Q}_2$ is also orthogonal.

An $m \times m$ matrix $\mathbf{R}$ is said to be upper triangular, if its entries r_{ij} below the main diagonal are zero ($r_{ij} \neq 0$ for $i \leq j$; $r_{ij} = 0$ for $i > j$), as shown below:

$$\mathbf{R} \equiv \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1m} \\ 0 & r_{22} & \dots & r_{2m} \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & r_{mm} \end{bmatrix}$$

The QR decomposition applied to a given non-singular matrix $\mathbf{A}$ makes the product $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ less sensitive to small errors affecting the values of the entries a_{ij} of $\mathbf{A}$. The QR decomposition is unique, if all the diagonal elements r_{ii} ($i = 1, 2, \ldots, m$) of $\mathbf{R}$ are required to be real and positive [83].

There are several methods for computing the QR decomposition. Among them are the Gram–Schmidt procedure, the Householder transformations, and the Givens rotations. These three methods will be described below. The interested reader can find more on the matter in several books or articles, for example in Refs. [53, 29].

The numerically stable Gram–Schmidt procedure is a method to make a set of vectors orthogonal in an inner product space.

An inner product space is a vector space with an operation, which associates each pair $\boldsymbol{v}$ and $\boldsymbol{w}$ of vectors of the space with a scalar quantity known as the inner product $\langle \boldsymbol{v}, \boldsymbol{w} \rangle$ of the considered pair of vectors. A vector space V is the set of all real (column) vectors $\boldsymbol{v}$ with n components $v_1, v_2, \ldots, v_n$, this set being closed under the two operations of addition (if $\boldsymbol{v}$ and $\boldsymbol{w}$ are two vectors of V, then $\boldsymbol{v} + \boldsymbol{w}$ is also a vector of V) and multiplication by a scalar (if $\boldsymbol{v}$ is a vector of V and s is a real scalar, then $s\boldsymbol{v}$ is also a vector of V). Two vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ of a vector space V are said to be orthogonal, if their inner product $\langle \boldsymbol{v}, \boldsymbol{w} \rangle$ is zero. Vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ of an n-dimensional vector space V are linearly independent, if the equality $c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 + \cdots + c_n\boldsymbol{v}_n = \mathbf{0}$ can only hold if the coefficients $c_1, c_2, \ldots, c_n$ are all equal to zero. Vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ of an n-dimensional vector space V are a basis, if any vector $\boldsymbol{v}$ of V can be expressed as a linear combination $\boldsymbol{v} = c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 + \cdots + c_n\boldsymbol{v}_n$ with a unique choice of the coefficients $c_1, c_2, \ldots, c_n$. For an n-dimensional vector space, any nonzero linearly independent vectors form a basis. Let V and W be two sets of vectors. If each vector $\boldsymbol{v}$ contained in V can be expressed as a linear

combination of the vectors contained in W, then W is said to be the basis set (or the generating set or the spanning set) for V.

The Gram–Schmidt procedure takes a finite set $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m\}$ of linearly independent vectors and generates an orthogonal set $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m\}$ which spans the same subspace as the previous set. Let

$$\mathbf{A} \equiv \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ 0 & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & \ldots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{bmatrix}$$

be a given $n \times m$ matrix with m linearly independent columns. The same matrix can also be indicated briefly $\mathbf{A} \equiv [\boldsymbol{a}_1\, \boldsymbol{a}_2 \ldots \boldsymbol{a}_m]$, where $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m$ are the column vectors of $\mathbf{A}$. We want to construct an orthogonal $n \times n$ matrix $\mathbf{Q} \equiv \{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n\}$ and an upper triangular $m \times m$ matrix $\mathbf{R} \equiv \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_m\}$ such that

$$\mathbf{A}_{n\times m} = \mathbf{Q}_{n\times n} \begin{bmatrix} \mathbf{R}_{m\times m} \\ \mathbf{0}_{(n-m)\times m} \end{bmatrix}$$

For example, let $\mathbf{A}$ be a two-column matrix $\mathbf{A} \equiv [\boldsymbol{a}_1\, \boldsymbol{a}_2]$ and $\mathbf{Q}$ be a two-column matrix $\mathbf{Q} \equiv [\boldsymbol{q}_1\, \boldsymbol{q}_2]$. Since $\mathbf{R}$ must be a triangular matrix, then

$$\mathbf{A} \equiv [\boldsymbol{a}_1 \quad \boldsymbol{a}_2] = [\boldsymbol{q}_1 \quad \boldsymbol{q}_2] \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix} = \mathbf{QR}$$

In this case, the two column vectors $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ of $\mathbf{Q}$ and the three nonzero entries r_{11}, r_{12}, and r_{22} of $\mathbf{R}$ are the quantities to be determined as a function of the two column vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ of the given matrix $\mathbf{A}$.

Likewise, when $\mathbf{A}$ is a three-column matrix $\mathbf{A} \equiv [\boldsymbol{a}_1\, \boldsymbol{a}_2\, \boldsymbol{a}_3]$, then

$$\mathbf{A} \equiv [\boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \boldsymbol{a}_3] = [\boldsymbol{q}_1 \quad \boldsymbol{q}_2 \quad \boldsymbol{q}_3] \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix} = \mathbf{QR}$$

In this case, the three column vectors $\boldsymbol{q}_1$, $\boldsymbol{q}_2$, and $\boldsymbol{q}_3$ of $\mathbf{Q}$ and the six nonzero entries r_{11}, r_{12}, r_{13}, r_{22}, r_{23}, and r_{33} of $\mathbf{R}$ are the quantities to be determined as a function of the three column vectors $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, and $\boldsymbol{a}_3$ of the given matrix $\mathbf{A}$.

The numerically stable Gram–Schmidt procedure is described below by means of an example, due to Wikipedia [82]. Let us consider the following 3×3 matrix

$$\mathbf{A} \equiv \begin{bmatrix} 12 & -51 & 4 \\ 6 & 167 & -68 \\ -4 & 24 & -41 \end{bmatrix}$$

The three column vectors $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, and $\boldsymbol{a}_3$ of the given matrix $\mathbf{A}$ will, by degrees, be replaced by the three column vectors $\boldsymbol{q}_1$, $\boldsymbol{q}_2$, and $\boldsymbol{q}_3$ of an orthogonal matrix $\mathbf{Q}$.

At the same time, the six nonzero entries r_{11}, r_{12}, r_{13}, r_{22}, r_{23}, and r_{33} of an upper triangular matrix $\mathbf{R}$ will be computed so that $\mathbf{A} = \mathbf{QR}$.

To this end, we first compute the norm (that is, the Euclidean length) of the first (or leftmost) column vector $\boldsymbol{a}_1$ of $\mathbf{A}$. The first entry r_{11} of $\mathbf{R}$ is set equal to the norm $\|\boldsymbol{a}_1\|$. This yields

$$r_{11} = \|\boldsymbol{a}_1\| = \left(a_{11}^2 + a_{21}^2 + a_{31}^2\right)^{\frac{1}{2}} = \left[12^2 + 6^2 + (-4)^2\right]^{\frac{1}{2}} = 14$$

Now, we normalise $\boldsymbol{a}_1$ by dividing all its components by its norm $\|\boldsymbol{a}_1\|$. This yields the first column $\boldsymbol{q}_1$ of $\mathbf{Q}$, as follows

$$q_{11} = \frac{a_{11}}{\|\boldsymbol{a}_1\|} = \frac{12}{14} = \frac{6}{7}$$
$$q_{21} = \frac{a_{21}}{\|\boldsymbol{a}_1\|} = \frac{6}{14} = \frac{3}{7}$$
$$q_{31} = \frac{a_{31}}{\|\boldsymbol{a}_1\|} = -\frac{4}{14} = -\frac{2}{7}$$

At this stage of the procedure, the given matrix $\mathbf{A}$ becomes

$$\begin{bmatrix} 6/7 & -51 & 4 \\ 3/7 & 167 & -68 \\ -2/7 & 24 & -41 \end{bmatrix}$$

Now, we compute the next entries (r_{12}, r_{13}, …, r_{1m}) of $\mathbf{R}$ by taking the scalar products of the first column of the matrix shown above with each of the other columns of the same matrix. In the example considered above, this yields

$$r_{12} = (6/7) \times (-51) + (3/7) \times 167 + (-2/7) \times 24 = 21$$
$$r_{13} = (6/7) \times 4 + (3/7) \times (-68) + (-2/7) \times (-41) = -14$$

Now, for the purpose of making the second, the third, …, the m^{th} column orthogonal to the first column, we subtract r_{12} times the first column from the second column, r_{13} times the first column from the third column, …, r_{1m} times the first column from the m^{th} column. This yields

$$-51 - (6/7) \times 21 = -69 \qquad 4 - (6/7) \times (-14) = 16$$
$$167 - (3/7) \times 21 = 158 \qquad -68 - (3/7) \times (-14) = -62$$
$$24 - (-2/7) \times 21 = 30 \qquad -41 - (2/7) \times (-14) = -45$$

and the resulting matrix is

$$\begin{bmatrix} 6/7 & -69 & 16 \\ 3/7 & 158 & -62 \\ -2/7 & 30 & -45 \end{bmatrix}$$

At this stage of the procedure, the first column vector of the preceding matrix has been normalised to have unit length; in addition, the second column and the third column have been made orthogonal to the first.

Now, in order to compute $\boldsymbol{q}_2$, we normalise the second column, by dividing all its components by its norm. At the same time, the r_{22} entry of $\mathbf{R}$ is set equal to the norm (or Euclidean length) of the second column. In the example considered above, this yields

$$r_{22} = \left[(-69)^2 + 158^2 + 30^2\right]^{\frac{1}{2}} = 14$$

The resulting matrix is

$$\begin{bmatrix} 6/7 & -69/175 & 16 \\ 3/7 & 158/175 & -62 \\ -2/7 & 6/35 & -45 \end{bmatrix}$$

Now r_{23} is the scalar product of the second column and the third column:

$$r_{23} = \left(-\frac{69}{175}\right) \times 16 + \left(\frac{158}{175}\right) \times (-62) + \left(\frac{6}{35}\right) \times (-45) = -70$$

By subtracting r_{23} times the second column from the third, we obtain

$$\begin{aligned} 16 - (-70) \times (-69/175) &= -406/35 \\ -62 - (-70) \times (158/175) &= 42/35 \\ -45 - (-70) \times (6/35) &= -33 \end{aligned}$$

The resulting matrix is

$$\begin{bmatrix} 6/7 & -69/175 & -406/35 \\ 3/7 & 158/175 & 42/35 \\ -2/7 & 6/35 & -33 \end{bmatrix}$$

Now we compute r_{33} as the norm of the third vector:

$$r_{33} = \left[\left(-\frac{406}{35}\right)^2 + \left(\frac{42}{35}\right)^2 + (-33)^2\right]^{\frac{1}{2}} = 35$$

Finally, in order to compute $\boldsymbol{q}_3$, the last column is normalised by dividing all its components by its norm. This yields

$$\left(-\frac{406}{35}\right)\frac{1}{35} = -\frac{58}{175}$$

$$\left(\frac{42}{35}\right)\frac{1}{35} = \frac{6}{175}$$

$$(-33)\frac{1}{35} = -\frac{33}{35}$$

The two matrices **Q** and **R** resulting from the procedure described above are then

$$\mathbf{Q} \equiv \begin{bmatrix} 6/7 & -69/175 & -58/175 \\ 3/7 & 158/175 & 6/175 \\ -2/7 & 6/35 & -33/35 \end{bmatrix}$$

$$\mathbf{R} \equiv \begin{bmatrix} 14 & 21 & -14 \\ 0 & 175 & -70 \\ 0 & 0 & 35 \end{bmatrix}$$

where **A** = **QR**. The example shown above has described the numerically stable Gram–Schmidt procedure for a given 3 × 3 matrix **A**. The same procedure can be extended to any non-singular $m \times m$ matrix **A**, by means of the algorithm given below, which is taken from Olver and Shakiban [60]. This algorithm takes the entries a_{ij} of **A** and replaces them with the entries q_{ij} of **Q**; at the same time, it computes the nonzero entries r_{ij} of **R**, which must be stored in a separate matrix.

```
start
    for j = 1 to m
        set r_jj = (a_1j^2 + a_2j^2 + ... + a_mj^2)^½
        if r_jj = 0, stop; print "A has linearly dependent columns"
        else for i = 1 to m
            set a_ij = a_ij / r_jj
        next i
        for k = j + 1 to m
            set r_jk = a_1j a_1k + a_2j a_2k + ... + a_mj a_mk
            for i = 1 to m
                set a_ik = a_ik − a_ij r_jk
            next i
        next k
    next j
end
```

The QR decomposition based on the Householder transformations is shown below. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be any two linearly independent vectors, having the same Euclidean length $\|\boldsymbol{x}\| = \|\boldsymbol{y}\|$, of an n-dimensional vector space V.

Consider the unit vector $\boldsymbol{u} = (\boldsymbol{y} - \boldsymbol{x})/\|\boldsymbol{y} - \boldsymbol{x}\|$ and the matrix $\mathbf{H} = \mathbf{I} - 2(\boldsymbol{u}\,\boldsymbol{u}^{\mathrm{T}})$.

Then $\mathbf{H}$ is the reflection matrix such that $\mathbf{H}\,\boldsymbol{x} = \boldsymbol{y}$. It is to be noted that $\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$ is an outer (or tensor) product, which yields a matrix, not a scalar.

For example, let

$$x \equiv \begin{bmatrix} 12 \\ 6 \\ -4 \end{bmatrix} \quad y \equiv \begin{bmatrix} 14 \\ 0 \\ 0 \end{bmatrix}$$

be two column vectors of a three-dimensional vector space V.

It is easy to verify that the two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ given above are linearly independent and have the same Euclidean length $\|\boldsymbol{x}\| = \|\boldsymbol{y}\| = 14$. In this case, there results

$$\boldsymbol{y} - \boldsymbol{x} = \begin{bmatrix} 2 \\ -6 \\ 4 \end{bmatrix}$$

$$\|\boldsymbol{y} - \boldsymbol{x}\| = \left[2^2 + (-6)^2 + 4^2\right]^{\frac{1}{2}} = 2(14)^{\frac{1}{2}}$$

$$\boldsymbol{u} = \frac{\boldsymbol{y} - \boldsymbol{x}}{\|\boldsymbol{y} - \boldsymbol{x}\|} = \begin{bmatrix} 1/(14)^{\frac{1}{2}} \\ -3/(14)^{\frac{1}{2}} \\ 2/(14)^{\frac{1}{2}} \end{bmatrix}$$

$$\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}} = \begin{bmatrix} 1/(14)^{\frac{1}{2}} \\ -3/(14)^{\frac{1}{2}} \\ 2/(14)^{\frac{1}{2}} \end{bmatrix} \left[1/(14)^{\frac{1}{2}} \quad -3/(14)^{\frac{1}{2}} \quad 2/(14)^{\frac{1}{2}} \right]$$

$$= \begin{bmatrix} 1/14 & -3/14 & 2/14 \\ -3/14 & 9/14 & -6/14 \\ 2/14 & -6/14 & 4/14 \end{bmatrix}$$

$$-2\left(\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}\right) = \begin{bmatrix} -1/7 & 3/7 & -2/7 \\ 3/7 & -9/7 & 6/7 \\ -2/7 & 6/7 & -4/7 \end{bmatrix}$$

$$\mathbf{H} = \mathbf{I} - 2\left(\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}\right) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} -1/7 & 3/7 & -2/7 \\ 3/7 & -9/7 & 6/7 \\ -2/7 & 6/7 & -4/7 \end{bmatrix}$$
$$= \begin{bmatrix} 6/7 & 3/7 & -2/7 \\ 3/7 & -2/7 & 6/7 \\ -2/7 & 6/7 & 3/7 \end{bmatrix}$$

As is easy to verify, the reflection matrix $\mathbf{H}$, found as shown above, satisfies the condition $\mathbf{H}\,\boldsymbol{x} = \boldsymbol{y}$. In addition, $\mathbf{H}$ is an orthogonal matrix (that is, $\mathbf{H}^{\mathrm{T}}\mathbf{H} = \mathbf{I}$) and its determinant $\det(\mathbf{H})$ is equal to -1. Since we know how to compute the reflection matrix $\mathbf{H}$ in an n-dimensional vector space, we can apply these concepts to produce QR decompositions. To this end, we intend to introduce sub-diagonal zeros into the given matrix $\mathbf{A}$ to be decomposed, in order to gradually change $\mathbf{A}$ into an upper triangular matrix $\mathbf{R}$.

In other words, we multiply the given matrix $\mathbf{A}_{n\times m}$ on the left by a sequence of reflection matrices $\mathbf{H}_1$, $\mathbf{H}_2$, …, $\mathbf{H}_k$, so that the product $\mathbf{H}_k \cdots \mathbf{H}_2\,\mathbf{H}_1\,\mathbf{A}$ should be equal to an upper triangular matrix $\mathbf{R}$. The first step operates on the matrix $\mathbf{A}$ itself. We choose the first (that is, the leftmost) column vector $\boldsymbol{a}_1$ of the given matrix $\mathbf{A}_{n\times m} \equiv [\boldsymbol{a}_1\ \boldsymbol{a}_2\ \ldots\ \boldsymbol{a}_m]$. We compute the Euclidean length

$$\|\boldsymbol{a}_1\| = \left(a_{11}^2 + a_{21}^2 + \cdots + a_{n1}^2\right)^{\frac{1}{2}}$$

of the first column vector $\boldsymbol{a}_1$ and then find a reflection matrix $\mathbf{H}_1$ such that the first column of the product $\mathbf{H}_1\,\mathbf{A}$ should be a multiple of the vector $[1\ 0\ 0\ \ldots\ 0]^{\mathrm{T}}$. In other words, we form

$$\boldsymbol{v} = \boldsymbol{a}_1 - \|\boldsymbol{a}_1\| \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \boldsymbol{u} = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \qquad \mathbf{H}_1 = \mathbf{I} - 2(\boldsymbol{u}\,\boldsymbol{u}^{\mathrm{T}})$$

By so doing, $\mathbf{H}_1$ is a reflection matrix such that the first column of $\mathbf{H}_1\,\mathbf{A}$ is a multiple of the vector $[1\ 0\ 0\ \ldots\ 0]^{\mathrm{T}}$. In other words, the first column of $\mathbf{H}_1\,\mathbf{A}$ is a column vector having all zeros in its rows except the first.

Each of the column vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_i, \ldots, \boldsymbol{a}_m$ of $\mathbf{A}_{n\times m}$ can be reflected onto a multiple of $[1\ 0\ 0\ \ldots\ 0]^{\mathrm{T}}$ in two ways: it can be reflected onto $\|\boldsymbol{a}_i\|\ [1\ 0\ 0\ \ldots\ 0]^{\mathrm{T}}$ or onto $-\|\boldsymbol{a}_i\|\ [1\ 0\ 0\ \ldots\ 0]^{\mathrm{T}}$. The choice of the sign in front of $\|\boldsymbol{a}_i\|$ is important, because when we form the unit vector $\boldsymbol{u} = (\boldsymbol{y} - \boldsymbol{x})/\|\boldsymbol{y} - \boldsymbol{x}\|$ we divide by $\|\boldsymbol{y} - \boldsymbol{x}\|$. Consequently, a choice which makes this denominator small must be avoided. To this end, the sign in front of $\|\boldsymbol{a}_i\|$ should be the opposite of the sign in front of the

entry placed in the k^{th} row of $\|\boldsymbol{a}_i\|$, where a_{ki} is the pivot co-ordinate after which all entries are zero in the final upper triangular form of $\mathbf{A}$. Therefore, it is advisable to choose

$$\boldsymbol{u} = -\text{sgn}\,(a_{ki})\|\boldsymbol{a}_i\|[\,1 \quad 0 \quad 0 \quad \ldots \quad 0\,]^{\text{T}}$$

where the signum function sgn(x) is such that

$$\text{sgn}(x) = \begin{cases} -1 & (\text{if } x < 0) \\ 0 & (\text{if } x = 0) \\ 1 & (\text{if } x > 0) \end{cases}$$

The second step operates on the matrix $\mathbf{A}'$, which results from cancelling the first row and the first column of $\mathbf{H}_1\mathbf{A}$ and retaining the rest of $\mathbf{H}_1\mathbf{A}$. We repeat for $\mathbf{A}'$ the same operations performed in the first step, and compute the reflection matrix $\mathbf{H}'_2$. Since $\mathbf{H}'_2$ is of smaller rank than $\mathbf{H}_1$ and we want to operate with $\mathbf{H}_1\mathbf{A}$ (and not with $\mathbf{A}'$), then we expand $\mathbf{A}'$, by filling in a 1 in the upper left entry of $\mathbf{H}_1\mathbf{A}$. This means that the second reflection matrix $\mathbf{H}_2$ results from

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 0 \\ 0 & \mathbf{H}'_2 \end{bmatrix}$$

The third step operates on $\mathbf{A}''$, which results from cancelling the first row and the first column of $\mathbf{H}_2\mathbf{A}'$ and retaining the rest of $\mathbf{H}_2\mathbf{A}'$. We repeat for $\mathbf{A}''$ the same operations performed in the second step, and compute the reflection matrix $\mathbf{H}'_3$.

The third reflection matrix $\mathbf{H}_3$ results from

$$\mathbf{H}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \mathbf{H}'_3 \end{bmatrix}$$

In general, at the p^{th} step, the p^{th} reflection matrix $\mathbf{H}_p$ results from

$$\mathbf{H}_p = \begin{bmatrix} \mathbf{I}_{p-1} & 0 \\ 0 & \mathbf{H}'_p \end{bmatrix}$$

After a number $k = \min(m - 1, n)$ of steps, the result of this process will be

$$\mathbf{H}_k \cdots \mathbf{H}_2\,\mathbf{H}_1\,\mathbf{A} = \mathbf{R}$$

where $\mathbf{R}$ is an upper triangular matrix, and each of the reflection matrices ($\mathbf{H}_1$, $\mathbf{H}_2$, …, $\mathbf{H}_k$) is an orthogonal matrix. Thus, with

$$\mathbf{Q} = \mathbf{H}_1{}^{\text{T}}\,\mathbf{H}_2{}^{\text{T}} \cdots \mathbf{H}_k^{\text{T}}$$

the desired result **A = QR** will be reached. For example, we consider the matrix

$$\mathbf{A} \equiv \begin{bmatrix} 12 & -51 & 4 \\ 6 & 167 & -68 \\ -4 & 24 & -41 \end{bmatrix}$$

which was decomposed previously by means of the Gram–Schmidt procedure.

We consider the first column vector $\boldsymbol{a}_1 \equiv [12\ 6\ -4]^{\mathrm{T}}$ of the matrix $\mathbf{A}$ and compute the Euclidean length $\|\boldsymbol{a}_1\|$ of $\boldsymbol{a}_1$, as follows

$$\|\boldsymbol{a}_1\| = \left[12^2 + 6^2 + (-4)^2\right]^{\frac{1}{2}} = 14$$

Then we compute

$$\|\boldsymbol{a}_1\|[1 \quad 0 \quad 0 \quad \cdots \quad 1]^{\mathrm{T}} = [14 \quad 0 \quad 0]^{\mathrm{T}}$$

$$\boldsymbol{v} = \boldsymbol{a}_1 - \operatorname{sgn}(a_{11})\|\boldsymbol{a}_1\|[1 \quad 0 \quad 0 \quad \cdots \quad 0]^{\mathrm{T}} = [-2 \quad 6 \quad -4]^{\mathrm{T}}$$

$$\|\boldsymbol{v}\| = \left[(-2)^2 + 6^2 + (-4)^2\right]^{\frac{1}{2}} = 2(14)^{\frac{1}{2}}$$

$$\boldsymbol{u} = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} = \begin{bmatrix} -1/(14)^{\frac{1}{2}} \\ 3/(14)^{\frac{1}{2}} \\ -2/(14)^{\frac{1}{2}} \end{bmatrix}$$

$$\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}} = \begin{bmatrix} -1/(14)^{\frac{1}{2}} \\ 3/(14)^{\frac{1}{2}} \\ -2/(14)^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} -1/(14)^{\frac{1}{2}} & 3/(14)^{\frac{1}{2}} & -2/(14)^{\frac{1}{2}} \end{bmatrix}$$
$$= \begin{bmatrix} 1/14 & -3/14 & 2/14 \\ -3/14 & 9/14 & -6/14 \\ 2/14 & -6/14 & 4/14 \end{bmatrix}$$

$$\mathbf{H}_1 = \mathbf{I} - 2\left(\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}\right) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} -1/7 & 3/7 & -2/7 \\ 3/7 & -9/7 & 6/7 \\ -2/7 & 6/7 & -4/7 \end{bmatrix}$$
$$= \begin{bmatrix} 6/7 & 3/7 & -2/7 \\ 3/7 & -2/7 & 6/7 \\ -2/7 & 6/7 & 3/7 \end{bmatrix}$$

Now we consider the matrix resulting from the product $\mathbf{H}_1\mathbf{A}$, that is,

$$\mathbf{H}_1\mathbf{A} = \begin{bmatrix} 6/7 & 3/7 & -2/7 \\ 3/7 & -2/7 & 6/7 \\ -2/7 & 6/7 & 3/7 \end{bmatrix} \begin{bmatrix} 12 & -51 & 4 \\ 6 & 167 & -68 \\ -4 & 24 & -41 \end{bmatrix} = \begin{bmatrix} 14 & 21 & -14 \\ 0 & -49 & -14 \\ 0 & 168 & -77 \end{bmatrix}$$

This matrix, due to the nonzero value of its $(\mathbf{H}_1\mathbf{A})_{32}$ entry (because $168 \neq 0$), is not an upper triangular matrix. Therefore, we cancel the first row and the first column of $\mathbf{H}_1\mathbf{A}$, and consider the matrix

$$\mathbf{A}' = \begin{bmatrix} -49 & -14 \\ 168 & -77 \end{bmatrix}$$

The Euclidean length of the first column vector $\boldsymbol{a}'_1$ of $\mathbf{A}' \equiv [\boldsymbol{a}'_1\ \boldsymbol{a}'_2]$ results from

$$\left\|\boldsymbol{a}'_1\right\| = \left[(-49)^2 + 168^2\right]^{\frac{1}{2}} = 175$$

$$\boldsymbol{v} = \begin{bmatrix} -49 \\ 168 \end{bmatrix} - 175 \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -224 \\ 168 \end{bmatrix}$$

$$\|\boldsymbol{v}\| = \left[(-224)^2 + 168^2\right]^{\frac{1}{2}} = 280$$

$$\boldsymbol{u} = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} = \begin{bmatrix} -4/5 \\ 3/5 \end{bmatrix}$$

$$\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}} = \begin{bmatrix} -4/5 \\ 3/5 \end{bmatrix} [-4/5 \quad 3/5] = \begin{bmatrix} 16/25 & -12/25 \\ -12/25 & 9/25 \end{bmatrix}$$

$$\mathbf{H}_2{}' = \mathbf{I} - 2(\boldsymbol{u}\boldsymbol{u})^{\mathrm{T}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} -32/25 & 24/25 \\ 24/25 & -18/25 \end{bmatrix} = \begin{bmatrix} -7/25 & 24/25 \\ 24/25 & 7/25 \end{bmatrix}$$

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -7/25 & 24/25 \\ 0 & 24/25 & 7/25 \end{bmatrix}$$

Since the product $\mathbf{H}_2\mathbf{H}_1\mathbf{A}$ yields

$$\mathbf{H}_2\mathbf{H}_1\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -7/25 & 24/25 \\ 0 & 24/25 & 7/25 \end{bmatrix} \begin{bmatrix} 14 & 21 & -14 \\ 0 & -49 & -14 \\ 0 & 168 & -77 \end{bmatrix} = \begin{bmatrix} 14 & 21 & -14 \\ 0 & 175 & -70 \\ 0 & 0 & -35 \end{bmatrix}$$

then $\mathbf{H}_2\mathbf{H}_1\mathbf{A}$ is the desired upper triangular matrix $\mathbf{R}$ and there is no necessity of further steps. The desired matrix $\mathbf{Q}$ results from $\mathbf{Q} = \mathbf{H}_1^{\mathrm{T}}\mathbf{H}_2^{\mathrm{T}}$. Therefore

$$\mathbf{Q} = \mathbf{H_1}^{\mathrm{T}}\mathbf{H_2}^{\mathrm{T}} = \begin{bmatrix} 6/7 & 3/7 & -2/7 \\ 3/7 & -2/7 & 6/7 \\ -2/7 & 6/7 & 3/7 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -7/25 & 24/25 \\ 0 & 24/25 & 7/25 \end{bmatrix}$$
$$= \begin{bmatrix} 6/7 & -69/175 & 58/175 \\ 3/7 & 158/175 & -6/175 \\ -2/7 & 6/35 & 33/35 \end{bmatrix}$$

It is easy to verify that the product **QR** of the two matrices **Q** and **R** determined above is equal to the given matrix **A**.

The QR decomposition based on the Givens rotations is shown below.

A Givens rotation is the rotation of a column vector in the plane spanned by two co-ordinate axes. In order to perform a QR decomposition **A** = **QR**, each Givens rotation is a matrix **G** which, multiplied on the left by the matrix **A**, introduces a zero in the sub-diagonal entries of **A**, for the purpose of gradually transforming **A** into an upper triangular matrix **R**. The product of the transposes of all these Givens rotation matrices produces the orthogonal matrix **Q**.

A Givens rotation matrix is represented by the matrix shown below.

$$\mathbf{G} \equiv \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ 0 & \cdots & \cos\theta & \cdots & \sin\theta & \cdots & 0 \\ \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ 0 & \cdots & -\sin\theta & \cdots & \cos\theta & \cdots & 0 \\ \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}$$

In other words, a Givens rotation matrix **G** with entries g_{ij} results from the identity matrix **I** after operating the following substitutions:

$$g_{ii} = \cos\theta \qquad g_{ij} = \sin\theta$$
$$g_{jj} = \cos\theta \qquad g_{ji} = -\sin\theta$$

This method applies to the case of a QR decomposition as follows.

Let us consider again the matrix

$$\mathbf{A} \equiv \begin{bmatrix} 12 & -51 & 4 \\ 6 & 167 & -68 \\ -4 & 24 & -41 \end{bmatrix}$$

which was decomposed previously. We want to construct a Givens rotation matrix $\mathbf{G}_1$ for the purpose of replacing the a_{31} entry of **A** (which entry is at present $a_{31} = -4$) with a zero. By multiplying on the left the first column vector

$$\boldsymbol{a}_1 \equiv \begin{bmatrix} 12 \\ 6 \\ -4 \end{bmatrix}$$

of the matrix $\mathbf{A} \equiv [\boldsymbol{a}_1\ \boldsymbol{a}_2\ \boldsymbol{a}_3]$ by the following rotation matrix $\mathbf{G}_1$

$$\mathbf{G}_1 \equiv \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix}$$

there results the following column vector

$$\begin{bmatrix} 12 \\ 6\cos\theta - 4\sin\theta \\ -6\sin\theta - 4\cos\theta \end{bmatrix}$$

In order for the third entry ($-6 \sin \theta - 4 \cos \theta$) of this vector to be zero, θ must be

$$\theta = \arctan\left(-\frac{2}{3}\right) = -33^\circ.690$$

and consequently $\cos(-33^\circ.690) = 0.83205$ and $\sin(-33^\circ.690) = -0.55470$. Therefore, $\mathbf{G}_1$ is

$$\mathbf{G}_1 \equiv \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.83205 & -0.55470 \\ 0 & 0.55470 & 0.83205 \end{bmatrix}$$

The product $\mathbf{G}_1\ \mathbf{A}$ yields a matrix $\mathbf{A}'$, whose entry a'_{31} is equal to zero, as follows

$$\begin{aligned} \mathbf{A}' &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.83205 & -0.55470 \\ 0 & 0.55470 & 0.83205 \end{bmatrix} \begin{bmatrix} 12 & -51 & 4 \\ 6 & 167 & -68 \\ -4 & 24 & -41 \end{bmatrix} \\ &= \begin{bmatrix} 12 & -51 & 4 \\ 7.2111 & 125.64 & -33.837 \\ 0 & 112.60 & -71.834 \end{bmatrix} \end{aligned}$$

Now, starting from $\mathbf{A}'$ indicated above, we find a rotation matrix $\mathbf{G}_2$

$$\mathbf{G}_2 \equiv \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

in order to transform $a'_{21} = 7.2111$ into zero. By operating as has been shown above, we find the following condition

$$(-\sin\theta) \times 12 + (\cos\theta) \times 7.2111 = 0$$

hence $\theta = 31°.003$ and consequently $\cos\theta = 0.85714$ and $\sin\theta = 0.51508$. Then

$$\mathbf{G}_2 \equiv \begin{bmatrix} 0.85714 & 0.51508 & 0 \\ -0.51508 & 0.85714 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The product $\mathbf{A}'' = \mathbf{G}_2\mathbf{A}'$ yields a matrix such that its entries a''_{21} and a''_{31} are both of them equal to zero:

$$\mathbf{A}'' \equiv \begin{bmatrix} 14 & 21 & -14 \\ 0 & 133.96 & -31.063 \\ 0 & 112.60 & -71.834 \end{bmatrix}$$

Now, starting from $\mathbf{A}''$ indicated above, let us find a rotation matrix $\mathbf{G}_3$

$$\mathbf{G}_3 \equiv \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix}$$

which transforms $a''_{32} = 112.60$ into zero. By operating as has been shown above, we find the following condition

$$(-\sin\theta) \times 133.96 + (\cos\theta) \times 112.60 = 0$$

hence $\theta = 40°.049$, and consequently $\cos\theta = 0.76550$ and $\sin\theta = 0.64344$.

Therefore, $\mathbf{G}_3$ is

$$\mathbf{G}_3 \equiv \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.76550 & 0.64344 \\ 0 & -0.64344 & 0.76550 \end{bmatrix}$$

The product $\mathbf{G}_3\,\mathbf{A}''$ yields a matrix $\mathbf{R}$ having all of its sub-diagonal entries equal to zero, as follows

$$\mathbf{R} \equiv \begin{bmatrix} 14 & 21 & -14 \\ 0 & 175 & -70 \\ 0 & 0 & -35 \end{bmatrix}$$

As has been shown above, $\mathbf{R}$ results from $\mathbf{R} = \mathbf{G}_3\mathbf{G}_2\mathbf{G}_1\mathbf{A}$. Since $\mathbf{G}_1$, $\mathbf{G}_2$ and $\mathbf{G}_3$ are, each of them, orthogonal matrices, then the desired matrix $\mathbf{Q}$ of the decomposition $\mathbf{A} = \mathbf{QR}$ results from

$$\mathbf{Q} = \mathbf{G}_1^{\mathrm{T}}\mathbf{G}_2^{\mathrm{T}}\mathbf{G}_3^{\mathrm{T}}$$

In the example considered above, the product indicated above yields

$$\mathbf{Q} \equiv \begin{bmatrix} 6/7 & -69/175 & 58/175 \\ 3/7 & 158/175 & -6/175 \\ -2/7 & 6/35 & 33/35 \end{bmatrix}$$

It is easy to verify that the product $\mathbf{QR}$ of the two matrices $\mathbf{Q}$ and $\mathbf{R}$ determined above is equal to the given matrix $\mathbf{A}$. Generally speaking, the QR decomposition $\mathbf{A} = \mathbf{QR}$ can be used to solve systems of linear algebraic equations. This is because, by using this decomposition, the system

$$\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$$

can be written as follows

$$\mathbf{QR}\boldsymbol{x} = \boldsymbol{b}$$

Now, since $\mathbf{Q}$ is an orthogonal matrix (such that $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$), then multiplying on the left the two members of the preceding equality by $\mathbf{Q}^{\mathrm{T}}$ yields

$$\mathbf{R}\boldsymbol{x} = \mathbf{Q}^{\mathrm{T}}\boldsymbol{b}$$

Since $\mathbf{R}$ is an upper triangular matrix, then the preceding expression can be solved for $\boldsymbol{x}$ by back-substitution, as will be shown below. In the specific case of the least-squares estimation, the QR decomposition $\mathbf{A} = \mathbf{QR}$ can be used to write the loss function J described in the preceding paragraph, that is,

$$J = (\boldsymbol{b} - \mathbf{A}\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{b} - \mathbf{A}\boldsymbol{x})$$

in the way shown below. By multiplying on the left the two members of the preceding equality by $\mathbf{Q}^{\mathrm{T}}$ and remembering that $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}} = \mathbf{I}$, there results

$$J = (\mathbf{Q}^{\mathrm{T}}\boldsymbol{b} - \mathbf{Q}^{\mathrm{T}}\mathbf{A}\boldsymbol{x})^{\mathrm{T}}(\mathbf{Q}^{\mathrm{T}}\boldsymbol{b} - \mathbf{Q}^{\mathrm{T}}\mathbf{A}\boldsymbol{x}) = (\boldsymbol{d} - \mathbf{R}\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{d} - \mathbf{R}\boldsymbol{x}) + \boldsymbol{r}^{\mathrm{T}}\boldsymbol{r}$$

where the two vectors $\boldsymbol{d}$ and $\boldsymbol{r}$ are partitions of the matrix $\mathbf{Q}^{\mathrm{T}}\boldsymbol{b}$. The number of their rows is m for $\boldsymbol{d}$ and $n - m$ for $\boldsymbol{r}$. The minimum value ($\boldsymbol{r}^{\mathrm{T}}\boldsymbol{r}$) of the loss function J is reached for $\mathbf{R}\boldsymbol{x} = \boldsymbol{d}$.

If m is the rank of $\mathbf{A}$, m will also be the rank of $\mathbf{R}$. Consequently, the following system of linear algebraic equations

$$\begin{bmatrix} r_{11} & r_{12} & \dots & r_{1m} \\ 0 & r_{22} & \dots & r_{2m} \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & r_{mm} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_m \end{bmatrix}$$

has a unique solution. Since $\mathbf{R}$ is an upper triangular matrix, the components x_1, x_2, …, x_m of $\boldsymbol{x}$ can be computed by back-substitution, that is, beginning from the last (x_m) and going towards the first (x_1), as follows

$$\begin{aligned} x_m &= \frac{d_m}{r_{mm}} \\ x_{m-1} &= \frac{d_{m-1} - r_{(m-1)m} x_m}{r_{(m-1)(m-1)}} \\ &\vdots \\ x_i &= \frac{1}{r_{ii}} \left(d_i - \sum_{j=i+1}^{m} r_{ij} x_j \right) \quad (i = m-1,\, m-2, \dots, 1) \end{aligned}$$

In addition to the QR decomposition methods shown above, a singular value decomposition may be used to solve a least-squares problem. A singular value decomposition is particularly suited in case of systems of linear algebraic equations $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ having ill-conditioned coefficient matrices $\mathbf{A}$, that is, in case of matrices $\mathbf{A}$ such that small changes in their entries a_{ij} result in large changes in the solution $\boldsymbol{x}$ of the system, as will be shown below. Singular value decomposition is a means of decomposing a matrix into the product of three simpler matrices, as the sequel will show.

Any nonzero $n \times m$ matrix $\mathbf{A}$ of rank $r > 0$ can be decomposed as follows

$$\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{Q}^{\mathrm{T}}$$

that is, decomposed into the product of an $n \times r$ matrix $\mathbf{P}$ with orthonormal columns (such that $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}$), an $r \times r$ diagonal matrix $\mathbf{D} \equiv \mathrm{diag}(d_1, d_2, \dots, d_{r-1}, d_r)$, whose nonzero entries are supposed to be $d_1 \geq d_2 \geq \cdots \geq d_r \geq 0$, and an $r \times m$ matrix $\mathbf{Q}^{\mathrm{T}}$ with orthonormal columns (such that $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$). Let

$$\mathbf{D} \equiv \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & d_r \end{bmatrix}$$

be the diagonal matrix of the singular value decomposition $\mathbf{A} = \mathbf{PDQ}^{\mathrm{T}}$. Its nonzero entries $d_1, d_2, \ldots, d_r$, called the singular values of $\mathbf{A}$, are the positive square roots of the nonzero eigenvalues λ_i of the Gram matrix $\mathbf{K} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$ associated with $\mathbf{A}$, that is,

$$d_i = (\lambda_i)^{\frac{1}{2}} > 0$$

($i = 1, 2, \ldots, r$). The columns of $\mathbf{P}$, called the left singular vectors, are the normalised eigenvectors of $\mathbf{AA}^{\mathrm{T}}$. The columns of $\mathbf{Q}$, called the right singular vectors, are the normalised eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$. As has been shown by Abdi [2], singular vectors come in pairs of one left and one right singular vector corresponding to the same singular value. They can be computed either separately or as a pair.

When they are computed as a pair (by rewriting equation $\mathbf{A} = \mathbf{PDQ}^{\mathrm{T}}$ in another form, as will be shown below), then it is possible to:

- compute only one (instead of two) decomposition in eigenvectors; and
- prevent a problem which may arise because the normalised eigenvectors of a matrix are determined up to a factor equal to -1.

Since singular vectors being pairs of vectors must have compatible parities, then care must be taken of the signs in front of the components of the eigenvectors; otherwise, the matrices $\mathbf{P}$ and $\mathbf{Q}$, if computed separately, might fail to reconstruct the given matrix $\mathbf{A}$.

Since $\mathbf{P}$ and $\mathbf{Q}$ are orthogonal matrices, the preceding expression

$$\mathbf{A} = \mathbf{PDQ}^{\mathrm{T}}$$

can be rewritten as follows

$$\mathbf{P} = \mathbf{AQD}^{-1}$$

Consequently, $\mathbf{P}$ results from the product of $\mathbf{A}$, $\mathbf{Q}$ and $\mathbf{D}^{-1}$. In addition, $\mathbf{D}^{-1}$ is a diagonal matrix having its nonzero entries equal to the reciprocals of the corresponding entries of $\mathbf{D}$, because $\mathbf{D}$ is a diagonal matrix.

It is to be remembered that the eigenvalues of any $n \times n$ matrix $\mathbf{M}$ are the scalars $\lambda_1, \lambda_2, \ldots, \lambda_n$ which satisfy the following characteristic equation

$$\det(\mathbf{M} - \lambda\mathbf{I}) = 0$$

and that the eigenvectors of $\mathbf{M}$ are the corresponding nonzero vectors $\boldsymbol{v}$ such that

$$(\mathbf{M} - \lambda\mathbf{I})\boldsymbol{v} = \mathbf{0}$$

If $\boldsymbol{v}$ is an eigenvector of $\mathbf{M}$, then any other nonzero vector $\boldsymbol{w}$, resulting from the product between $\boldsymbol{v}$ and a scalar, is also an eigenvector of $\mathbf{M}$.

By the way, the number of the singular values of a matrix $\mathbf{M}$ is always equal to the rank of the matrix itself. Therefore, if an $n \times n$ matrix $\mathbf{M}$ has less than n singular values, then $\mathbf{M}$ is singular.

The condition number $k(\mathbf{M})$ of a non-singular $n \times n$ matrix $\mathbf{M}$ is defined as the ratio between the largest (d_1) and the smallest (d_n) of the singular values of $\mathbf{M}$, that is, $k(\mathbf{M}) = d_1/d_n$.

A singular matrix $\mathbf{M}$ is said to have condition number $k(\mathbf{M})$ equal to infinity; a matrix $\mathbf{M}$ having a very large condition number $k(\mathbf{M})$ is said to be ill-conditioned. In practice, this condition occurs when the condition number $k(\mathbf{M})$ of the given matrix $\mathbf{M}$ is greater than the reciprocal of the precision of the machine used. In case of single-precision computations, this occurs typically when $k(\mathbf{M})$ is greater than 10^7.

An example, taken from Leach [47], of a single value decomposition is given below. Let

$$\mathbf{A} \equiv \begin{bmatrix} 0.96 & 1.72 \\ 2.28 & 0.96 \end{bmatrix} \qquad \mathbf{A}^{\mathrm{T}} \equiv \begin{bmatrix} 0.96 & 2.28 \\ 1.72 & 0.96 \end{bmatrix}$$

be, respectively, the given matrix to be decomposed and its transpose.

The Gram matrix $\mathbf{K} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$ associated with $\mathbf{A}$ is

$$\mathbf{A}^{\mathrm{T}}\mathbf{A} = \begin{bmatrix} 0.96 & 2.28 \\ 1.72 & 0.96 \end{bmatrix} \begin{bmatrix} 0.96 & 1.72 \\ 2.28 & 0.96 \end{bmatrix} = \begin{bmatrix} 6.12 & 3.84 \\ 3.84 & 3.88 \end{bmatrix}$$

The eigenvalues λ_1 and λ_2 of $\mathbf{K} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$ satisfy the equation

$$\det(\mathbf{K} - \lambda\mathbf{I}) = 0$$

that is,

$$\det\left(\begin{bmatrix} 6.12 - \lambda & 3.84 \\ 3.84 & 3.88 - \lambda \end{bmatrix} \right) = 0$$

By expanding the determinant, there results

$$\begin{aligned} &(6.12 - \lambda)(3.88 - \lambda) - 3.84 \times 3.84 = 0 \\ &\lambda^2 - 10\lambda + 9 = 0 \end{aligned}$$

The preceding equation has two roots (sorted in descending order, in the absolute sense), which are

$$\lambda_1 = 5 + (25 - 9)^{\frac{1}{2}} = 9$$

$$\lambda_2 = 5 - (25 - 9)^{\frac{1}{2}} = 1$$

Therefore, the singular values of **A** are

$$d_1 = (\lambda_1)^{\frac{1}{2}} = 3$$

$$d_2 = (\lambda_2)^{\frac{1}{2}} = 1$$

Thus, $d_1 = 3$ and $d_2 = 1$ are the nonzero entries of the required matrix **D**, that is,

$$\mathbf{D} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$$

$\mathbf{D}^{-1}$, the inverse of **D**, results immediately from

$$\mathbf{D}^{-1} = \begin{bmatrix} 1/3 & 0 \\ 0 & 1 \end{bmatrix}$$

Now, for $\lambda_1 = 9$ and $\lambda_2 = 1$, let us compute the eigenvectors of $\mathbf{K} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$. As shown above, these eigenvectors become column vectors in a matrix (**Q**) ordered by the size of the corresponding eigenvalues. In other words, the eigenvector of the largest eigenvalue, in the absolute sense, becomes the first (that is, the leftmost) column, the eigenvector of the next largest eigenvalue becomes the second column, and so on; the eigenvector of the smallest eigenvalue becomes the last (that is, the rightmost) column of **Q**.

(For $\lambda = \lambda_1 = 9$)

$$\mathbf{A}^{\mathrm{T}}\mathbf{A} - \lambda\mathbf{I} = \begin{bmatrix} 6.12 - 9 & 3.84 \\ 3.84 & 3.88 - 9 \end{bmatrix} = \begin{bmatrix} -2.88 & 3.84 \\ 3.84 & -5.12 \end{bmatrix}$$

Let a and b be the components of the eigenvector $\boldsymbol{q}_1$ corresponding to $\lambda_1 = 9$. The components of $\boldsymbol{q}_1$ result from $(\mathbf{A}^{\mathrm{T}}\mathbf{A} - \lambda\mathbf{I})\boldsymbol{q}_1 = \mathbf{0}$. Therefore

$$\begin{bmatrix} -2.88 & 3.84 \\ 3.84 & -5.12 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$-2.88a + 3.84b = 0$$
$$3.84a - 5.12b = 0$$

Solving either equation for b yields $b = 0.75\ a$. Therefore

$$\boldsymbol{q}_1 \equiv \begin{bmatrix} a \\ 0.75\,a \end{bmatrix}$$

Dividing the two components by the Euclidean length $\|\boldsymbol{q}_1\| = 1.25\,a$ of $\boldsymbol{q}_1$ yields the normalised eigenvector $\boldsymbol{q}_1$, that is,

$$\boldsymbol{q}_1 \equiv \begin{bmatrix} 0.8 \\ 0.6 \end{bmatrix}$$

(For $\lambda = \lambda_2 = 1$)

$$\mathbf{A}^{\mathrm{T}}\mathbf{A} - \lambda\mathbf{I} = \begin{bmatrix} 6.12 - 1 & 3.84 \\ 3.84 & 3.88 - 1 \end{bmatrix} = \begin{bmatrix} 5.12 & 3.84 \\ 3.84 & 2.88 \end{bmatrix}$$

The components a and b of the corresponding eigenvector $\boldsymbol{q}_2$ result from

$$\left(\mathbf{A}^{\mathrm{T}}\mathbf{A} - \lambda\mathbf{I}\right)q_2 = \mathbf{0}$$

Therefore,

$$\begin{bmatrix} 5.12 & 3.84 \\ 3.84 & 2.88 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{aligned} 5.12\,a + 3.84\,b &= 0 \\ 3.84\,a + 2.88\,b &= 0 \end{aligned}$$

Solving either equation for b yields $b = -4a/3$. Therefore

$$\boldsymbol{q}_2 \equiv \begin{bmatrix} a \\ -\frac{4}{3}a \end{bmatrix}$$

Dividing the two components by the Euclidean length $\|\boldsymbol{q}_2\| = 5a/3$ of $\boldsymbol{q}_2$ yields the normalised eigenvector $\boldsymbol{q}_2$, that is,

$$\boldsymbol{q}_2 \equiv \begin{bmatrix} 0.6 \\ -0.8 \end{bmatrix}$$

Therefore, the required matrix $\mathbf{Q}$ is given by $\mathbf{Q} = [\boldsymbol{q}_1\boldsymbol{q}_2]$, that is,

$$\mathbf{Q} = \begin{bmatrix} 0.8 & 0.6 \\ 0.6 & -0.8 \end{bmatrix}$$

As is easy to verify, the product $\mathbf{Q}^{\mathrm{T}}\mathbf{Q}$ is equal to the 2×2 identity matrix $\mathbf{I}$. Therefore, $\mathbf{Q}$ is an orthogonal matrix.

Now, as has been shown above, we compute $\mathbf{P} = \mathbf{AQD}^{-1}$. The product $\mathbf{QD}^{-1}$ is

$$\mathbf{QD}^{-1} = \begin{bmatrix} 0.8 & 0.6 \\ 0.6 & -0.8 \end{bmatrix} \begin{bmatrix} 1/3 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0.8/3 & 0.6 \\ 0.2 & -0.8 \end{bmatrix}$$

The product $\mathbf{AQD}^{-1}$ yields $\mathbf{P}$, as follows

$$\mathbf{P} = \begin{bmatrix} 0.96 & 1.72 \\ 2.28 & 0.96 \end{bmatrix} \begin{bmatrix} 0.8/3 & 0.6 \\ 0.2 & -0.8 \end{bmatrix} = \begin{bmatrix} 0.6 & -0.8 \\ 0.8 & 0.6 \end{bmatrix}$$

It is easy to verify that $\mathbf{P}$ is an orthogonal matrix and that the product $\mathbf{PDQ}^{\mathrm{T}}$ yields the given matrix $\mathbf{A}$.

The singular value decomposition can be used for computing the pseudo-inverse of a matrix, which in turn provides a general method for solving the least-squares problem in the form $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$, as will be shown below. Following Golub and Reinsch [29], let $\mathbf{A}$ be a real $n \times m$ matrix. An $m \times n$ matrix $\mathbf{X}$ is said to be the pseudo-inverse of $\mathbf{A}$, if $\mathbf{X}$ satisfies the following four properties:

$$\begin{aligned} &\mathbf{AXA} = \mathbf{A} \\ &\mathbf{XAX} = \mathbf{X} \\ &(\mathbf{AX})^{\mathrm{T}} = \mathbf{AX} \\ &(\mathbf{XA})^{\mathrm{T}} = \mathbf{XA} \end{aligned}$$

Let $\mathbf{A}^{+}$ denote the unique solution $\mathbf{X}$. If $\mathbf{A} = \mathbf{PDQ}^{\mathrm{T}}$, then the pseudo-inverse $\mathbf{A}^{+}$ is such that

$$\mathbf{A}^{+} = \mathbf{Q}\,\mathbf{D}^{+}\mathbf{P}^{\mathrm{T}}$$

where

$$\mathbf{D}^{+} = \mathrm{diag}(d_i^{+}) \qquad d_i^{+} = \begin{cases} \dfrac{1}{d_i} & (\text{for } d_i > 0) \\ 0 & (\text{for } d_i = 0) \end{cases}$$

Consequently, the pseudo-inverse $\mathbf{A}^{+}$ of a given matrix $\mathbf{A}$ can easily be computed as a result of the singular value decomposition of $\mathbf{A}$.

This concept applies to the search for the least-squares solution of an over-determined system of n algebraic equations in m unknowns, where $n > m$. Let

$$\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$$

be such a system in matrix form. Let

$$\boldsymbol{\rho} = \boldsymbol{b} - \mathbf{A}\boldsymbol{x}$$

be the residual vector for some vector $\boldsymbol{x}$. We want to determine the vector $\boldsymbol{x}^{\text{lsq}}$ corresponding to the least-squares solution of $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$, that is, the vector corresponding to the minimum possible residual vector $\boldsymbol{\rho}$.

To this end, let us consider the squared Euclidean length

$$\|\boldsymbol{\rho}\|^2 = \boldsymbol{\rho}^{\mathrm{T}}\,\boldsymbol{\rho} = (\boldsymbol{b} - \mathbf{A}\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{b} - \mathbf{A}\boldsymbol{x}) = \boldsymbol{b}^{\mathrm{T}}\,\boldsymbol{b} - 2\,\boldsymbol{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\,\boldsymbol{b} + \boldsymbol{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\boldsymbol{x}$$

of the residual vector $\boldsymbol{\rho} = \mathbf{A}\boldsymbol{x} - \boldsymbol{b}$. To determine the minimum value of $\|\boldsymbol{\rho}\|^2$, we take the derivative with respect to $\boldsymbol{x}$ and set it to zero. This yields

$$-2\,\mathbf{A}^{\mathrm{T}}\boldsymbol{b} + 2\,\mathbf{A}^{\mathrm{T}}\mathbf{A}\,\boldsymbol{x} = \mathbf{0}$$

and hence the following $m \times m$ system of normal equations

$$\mathbf{A}^{\mathrm{T}}\mathbf{A}\,\boldsymbol{x} = \mathbf{A}^{\mathrm{T}}\boldsymbol{b}$$

Multiplying both terms of the preceding expression on the left by $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$ yields

$$\boldsymbol{x} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{b} = \mathbf{A}^{+}\boldsymbol{b}$$

Therefore, the least-squares solution ($\boldsymbol{x}$) of $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ corresponding to the minimum value of $\|\boldsymbol{\rho}\|^2$ is given by $\boldsymbol{x} = \mathbf{A}^{+}\,\boldsymbol{b}$. Weights can be applied to the original system of equations $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ by using a diagonal matrix $\mathbf{W} \equiv \boldsymbol{\Sigma}^2 \equiv \text{diag}(\sigma_1^{-2}, \ldots, \sigma_n^{-2})$.

In this case, the least-squares solution to be found is

$$\boldsymbol{x} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{W}\,\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{W}\,\boldsymbol{b}$$

Bebis [6] gives the following (over-determined) system $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ of three equations for the two unknowns x_1 and x_2:

$$\begin{aligned} -11x_1 + 2x_2 &= 0 \\ 2x_1 + 3x_2 &= 7 \\ 2x_1 - x_2 &= 5 \end{aligned}$$

In matrix form, the preceding equations are as follows

$$\begin{bmatrix} -11 & 2 \\ 2 & 3 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 7 \\ 5 \end{bmatrix}$$

As has been shown above, the least-squares solution ($\boldsymbol{x}$) results from

$$\boldsymbol{x} = \mathbf{A}^{+}\,\boldsymbol{b}$$

Therefore, the problem reduces to determining the pseudo-inverse $\mathbf{A}^+$ of the given matrix $\mathbf{A}$. Using the expression written above

$$\mathbf{A}^+ = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}$$

we compute $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ as follows

$$\mathbf{A}^{\mathrm{T}}\mathbf{A} = \begin{bmatrix} -11 & 2 & 2 \\ 2 & 3 & -1 \end{bmatrix} \begin{bmatrix} -11 & 2 \\ 2 & 3 \\ 2 & -1 \end{bmatrix} = \begin{bmatrix} 129 & -18 \\ -18 & 14 \end{bmatrix}$$

Since $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is a 2×2 matrix, it is easy to compute its inverse $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$ as follows. Let $\mathbf{M}$ be the following 2×2 matrix

$$\mathbf{M} \equiv \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Then, the inverse matrix $\mathbf{M}^{-1}$ is given by

$$\mathbf{M}^{-1} = \begin{bmatrix} d/\det(\mathbf{M}) & -b/\det(\mathbf{M}) \\ -c/\det(\mathbf{M}) & a/\det(\mathbf{M}) \end{bmatrix}$$

where $\det(\mathbf{M}) = ad - bc$ is the determinant of the given matrix $\mathbf{M}$. In this case,

$$\det\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right) = 129 \times 14 - 18 \times 18 = 1482$$

Therefore

$$\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1} = \begin{bmatrix} 14/1482 & 18/1482 \\ 18/1482 & 129/1482 \end{bmatrix}$$

The pseudo-inverse $\mathbf{A}^+ = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$ results from the product

$$\begin{bmatrix} 14/1482 & 18/1482 \\ 18/1482 & 129/1482 \end{bmatrix} \begin{bmatrix} -11 & 2 & 2 \\ 2 & 3 & -1 \end{bmatrix} = \begin{bmatrix} -0.079622 & 0.055331 & 0.0067476 \\ 0.040486 & 0.28543 & -0.062753 \end{bmatrix}$$

Consequently, the least-squares solution $\boldsymbol{x} = \mathbf{A}^+\boldsymbol{b}$ of the given system $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ is

$$\begin{bmatrix} -0.079622 & 0.055331 & 0.0067476 \\ 0.040486 & 0.28543 & -0.062753 \end{bmatrix} \begin{bmatrix} 0 \\ 7 \\ 5 \end{bmatrix} = \begin{bmatrix} 0.42106 \\ 1.6842 \end{bmatrix}$$

The same result can also be obtained by computing the singular value decomposition of the given matrix $\mathbf{A}$, as will be shown below.

The matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ has been computed above. Its eigenvalues result from

$$\det\left(\mathbf{A}^{\mathrm{T}}\mathbf{A} - \lambda\,\mathbf{I}\right) = (129 - \lambda)(14 - \lambda) - 18 \times 18 = 0$$

The preceding equation, solved for λ, yields

$$\lambda_1 = 131.75 \qquad (\lambda_1)^{\frac{1}{2}} = 11.478 \qquad \frac{1}{(\lambda_1)^{\frac{1}{2}}} = 0.087121$$

$$\lambda_2 = 11.248 \qquad (\lambda_2)^{\frac{1}{2}} = 3.3539 \qquad \frac{1}{(\lambda_2)^{\frac{1}{2}}} = 0.29816$$

For $\lambda = \lambda_1 = 131.75$, the corresponding normalised eigenvector $\boldsymbol{q}_1$ is

$$\boldsymbol{q}_1 \equiv \begin{bmatrix} -0.98852 \\ 0.15111 \end{bmatrix}$$

For $\lambda = \lambda_2 = 11.248$, the corresponding normalised eigenvector $\boldsymbol{q}_2$ is

$$\boldsymbol{q}_2 \equiv \begin{bmatrix} 0.15111 \\ 0.98852 \end{bmatrix}$$

The orthogonal matrix $\mathbf{Q}$, having $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ as its columns, is then

$$\mathbf{Q} = \begin{bmatrix} -0.98852 & 0.15111 \\ 0.15111 & 0.98852 \end{bmatrix}$$

The matrix $\mathbf{D}^{-1}$ results from

$$\mathbf{D}^{-1} = \operatorname{diag}\left(\frac{1}{\lambda_1^{\frac{1}{2}}}, \frac{1}{\lambda_2^{\frac{1}{2}}}\right) = \begin{bmatrix} 0.087121 & 0 \\ 0 & 0.29816 \end{bmatrix}$$

The product $\mathbf{QD}^{-1}$ is equal to

$$\mathbf{QD}^{-1} = \begin{bmatrix} -0.086121 & 0.045055 \\ 0.013165 & 0.29474 \end{bmatrix}$$

The orthogonal matrix $\mathbf{P}$ results from the product $\mathbf{AQD}^{-1}$ and is then

$$\mathbf{P} = \begin{bmatrix} 0.97366 & 0.093875 \\ -0.13275 & 0.97433 \\ -0.18541 & -0.20463 \end{bmatrix}$$

Finally, $\mathbf{A}^+$ results from

$$\mathbf{A}^+ = \mathbf{Q}\mathbf{D}^+\mathbf{P}^{\mathrm{T}}$$

where $\mathbf{D}^+ = \mathrm{diag}(d_i^+) = \mathrm{diag}[1/(\lambda_1)^{1/2}, 1/(\lambda_2)^{1/2}]$ has the same entries as those of $\mathbf{D}^{-1}$ computed above. The product $\mathbf{Q}\mathbf{D}^+\mathbf{P}^{\mathrm{T}}$ yields

$$\begin{bmatrix} -0.079623 & 0.055331 & 0.0067480 \\ 0.040487 & 0.28543 & -0.062752 \end{bmatrix}$$

This matrix is to be compared with the matrix which has been computed above by means of the expression $\mathbf{A}^+ = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$. The singular value decomposition applies to the least-squares problem as will be shown below. Let

$$J = (\boldsymbol{b} - \mathbf{A}\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{b} - \mathbf{A}\boldsymbol{x})$$

be the loss function. Let $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{Q}^{\mathrm{T}}$ be the matrix of the least-squares problem $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$. By defining the vectors $\boldsymbol{s}$ and $\boldsymbol{t}$ as follows

$$\boldsymbol{s} = \mathbf{Q}^{\mathrm{T}}\boldsymbol{x}$$
$$\boldsymbol{t} = \mathbf{P}^{\mathrm{T}}\boldsymbol{b}$$

the condition $(\mathbf{A}^{\mathrm{T}}\mathbf{A})\boldsymbol{x} = \mathbf{A}^{\mathrm{T}}\boldsymbol{b}$ for the minimum value of the loss function can be written as follows

$$\mathbf{D}^2\boldsymbol{s} = \mathbf{D}\,\boldsymbol{t}$$

If the matrix of the normal equations is non-singular, then the inverse $\mathbf{D}^{-1}$ of the $\mathbf{D}$ matrix exists and the solution of the least-squares problem $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ is

$$\boldsymbol{s} = \mathbf{D}^{-1}\,\boldsymbol{t}$$

Remembering the definitions of the vectors $\boldsymbol{s}$ and $\boldsymbol{t}$ and the property of orthogonality ($\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{-1}$) of the matrix $\mathbf{Q}$, the preceding expression becomes

$$\boldsymbol{x} = \mathbf{Q}\,\mathbf{D}^{-1}\mathbf{P}^{\mathrm{T}}\,\boldsymbol{b}$$

or, using the column vectors $\boldsymbol{p}_i$ and $\boldsymbol{q}_i$ of the matrices, respectively, $\mathbf{P}$ and $\mathbf{Q}$,

$$\boldsymbol{x} = \sum_{i=1}^{m} \frac{1}{d_i}\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{b}\,\boldsymbol{q}_i$$

In order to apply the mathematical methods shown above to a problem of orbit determination, let us consider again the seven observations of the COBE artificial satellite performed by Healy [33] on the 6th of November 2000. They are:

Observed time (EST)	Right ascension (hh:mm:ss)	Declination (°)	Offset
17:31:29	21:48:00	−16.3	1
17:32:30	21:41:00	−2.0	0
17:33:30	21:31:45	19.3	0
17:34:30	21:14:00	46.9	3
17:35:29	20:28:00	71.9	3
17:36:30	15:01:00	84.6	2
17:37:30	11:03:00	76.1	0

As the satellite came nearest the cross hairs, time was recorded. The offset in the last column is an estimate of how close to the cross hairs the satellite came.

We want to construct a function which best fits, in the least-squares sense and in the time interval given above, the seven observations tabulated above.

Since UTC = EST + 5, then (neglecting the difference between UTC and UT1) the seven EST times indicated above correspond, respectively, to

$$\begin{aligned}
\mathrm{UT1}_1 &= \mathrm{EST}_1 + 5 = 22^{\mathrm{h}}31^{\mathrm{m}}29^{\mathrm{s}} \\
\mathrm{UT1}_2 &= \mathrm{EST}_2 + 5 = 22^{\mathrm{h}}32^{\mathrm{m}}30^{\mathrm{s}} \\
\mathrm{UT1}_3 &= \mathrm{EST}_3 + 5 = 22^{\mathrm{h}}33^{\mathrm{m}}30^{\mathrm{s}} \\
\mathrm{UT1}_4 &= \mathrm{EST}_4 + 5 = 22^{\mathrm{h}}34^{\mathrm{m}}30^{\mathrm{s}} \\
\mathrm{UT1}_5 &= \mathrm{EST}_5 + 5 = 22^{\mathrm{h}}35^{\mathrm{m}}29^{\mathrm{s}} \\
\mathrm{UT1}_6 &= \mathrm{EST}_6 + 5 = 22^{\mathrm{h}}36^{\mathrm{m}}30^{\mathrm{s}} \\
\mathrm{UT1}_7 &= \mathrm{EST}_7 + 5 = 22^{\mathrm{h}}37^{\mathrm{m}}30^{\mathrm{s}}
\end{aligned}$$

The seven values of right ascension, expressed in degrees, are given below.

$$\begin{aligned}
&(21 + 48/60) \times 360/24 = 327.00 \\
&(21 + 41/60) \times 360/24 = 325.25 \\
&(21 + 31/60 + 45/3600) \times 360/24 = 322.9375 \\
&(21 + 14/60) \times 360/24 = 318.50 \\
&(20 + 28/60) \times 360/24 = 307.00 \\
&(15 + 1/60) \times 360/24 = 225.25 \\
&(11 + 3/60) \times 360/24 = 165.75
\end{aligned}$$

In the hypothesis of uncorrelated errors of measurement, the weighting matrix is

$$\mathbf{W} = \mathrm{diag}(w_{11}, w_{22}, \ldots, w_{77})$$

where the values of the weights w_{11}, w_{22}, …, w_{77} are preliminarily set, all of them, to unity. In other words, the starting value of the weighting matrix $\mathbf{W}$ is taken equal to $\mathbf{I}$, where $\mathbf{I}$ is the 7×7 identity matrix.

Using a method described by several authors (see, e.g., Refs. [14, 32]), called iteratively reweighted least squares, we intend to update $\mathbf{W}$ iteratively, in such a way as to give less weight to the more uncertain data points. To this end, we use a mathematical model to predict the law of variation of the right ascension and declination with time. Then we use the residuals ρ_1, ρ_2, …, ρ_7 as a measure of uncertainty. These residuals are the differences between the computed (on the basis of the mathematical model) and the observed data points. The weights chosen in each iteration are related to the magnitudes of the residuals computed in the previous iteration, so that a large residual gives rise to a small weight, as will be shown below. The preliminary values of the weights along the main diagonal of $\mathbf{W}$ are indicated in the following table.

Obs. No.	x	Right ascension (°)	Declination (°)	Preliminary weight
1	−1.0000	327.00	−16.300	1.0000
2	−0.66205	325.25	−2.0000	1.0000
3	−0.32964	322.94	19.300	1.0000
4	0.0027701	318.50	46.900	1.0000
5	0.32964	307.00	71.900	1.0000
6	0.66759	225.25	84.600	1.0000
7	1.0000	165.75	76.100	1.0000

The right ascension (α) and declination (δ) of the observed satellite vary with time t according to functions $\alpha(t)$ and $\delta(t)$, which are not known a priori.

We approximate these unknown functions, within the time interval $t_1 \leq t \leq t_7$, by means of Chebyshev polynomials T_0, T_1, …, T_n as follows

$$\begin{aligned}\alpha(x) &= \alpha_0 T_0(x) + \alpha_1 T_1(x) + \cdots + \alpha_n T_n(x)\\ \delta(x) &= \delta_0 T_0(x) + \delta_1 T_1(x) + \cdots + \delta_n T_n(x)\end{aligned}$$

where α_0, α_1, …, α_n, δ_0, δ_1, …, δ_n are constant coefficients to be determined, and

$$\begin{aligned}T_0(x) &= 1\\ T_1(x) &= x\\ T_{n+1}(x) &= 2x\,T_n(x) - T_{n-1}(x) \quad (n \geq 2)\end{aligned}$$

are Chebyshev polynomials of the first kind. Supposing that α and δ are subject to errors independently of each other, we want to determine the best (in the weighted least-squares sense) polynomial approximation for α and δ. In other words, we want to determine the unknown values of the coefficients α_0, α_1, …, α_n, δ_0, δ_1, …, δ_n corresponding to the minimum value of some norm of the residual vector ($\boldsymbol{\rho}$). In the case of the classical least-squares method, this norm is the so-called ℓ^2 norm, that is,

we search the minimum value of the sum of the squares of the residuals. However, according to Bube and Langan [14], solutions found by using this method tend to be very sensitive to data points affected by large errors. By contrast, solutions coming from searching the minimum value of the ℓ^p norm (where $1 \leq p < 2$) are less sensitive to errors. The method described in Ref. [14] is a hybrid ℓ^1/ℓ^2 technique, by means of which a ℓ^2 fit (or minimum value of the sum of the squared residuals) is used for small residuals, and a ℓ^1 fit (or minimum value of the sum of the absolute residuals) is used for large residuals.

A smooth transition from the search for the minimum ℓ^2 norm to the search for the minimum ℓ^1 norm is obtained by choosing an appropriate value for a positive parameter ε which results from an estimate of the standard deviation of the residuals, as will be shown below. In the general case of n-degree polynomials, the model matrix $\mathbf{A}$ of the system of linear algebraic equations $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ is

$$\mathbf{A} \equiv \begin{bmatrix} T_{01} & T_{11} & \dots & T_{n1} \\ T_{02} & T_{12} & \dots & T_{n2} \\ \vdots & \vdots & \dots & \vdots \\ T_{07} & T_{17} & \dots & T_{n7} \end{bmatrix}$$

where the first subscript in the entries T_{ij} of $\mathbf{A}$ indicates the degree of the Chebyshev polynomial, and the second subscript indicates the number of observation. In order for the system $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ to be over-determined with our set of seven observations, the value of n cannot be greater than five. We set $n = 5$, which corresponds to a fifth-degree Chebyshev polynomial approximation to the true unknown functions $\alpha(x)$ and $\delta(x)$. In order for the argument x of the Chebyshev polynomials to be within the interval $-1 \leq x \leq 1$, we have operated a change of variable (from t to x) in the second column of the preceding table, as follows

$$x = \frac{2t - (t_7 + t_1)}{t_7 - t_1}$$

where the subscript indicates the number of observation.

As to declination (δ), we compute the entries T_{ij} of the 7×6 matrix $\mathbf{A}$ by computing the Chebyshev polynomials up to the fifth degree for the seven observations indicated above. This yields

$$\mathbf{A} \equiv \begin{bmatrix} 1.0000 & -1.0000 & 1.0000 & -1.0000 & 1.0000 & -1.0000 \\ 1.0000 & -0.66205 & -0.12338 & 0.82542 & -0.96955 & 0.45837 \\ 1.0000 & -0.32964 & -0.78268 & 0.84564 & 0.22516 & -0.99409 \\ 1.0000 & -0.0027701 & -0.99998 & -0.0083101 & 9.99994 & 0.013850 \\ 1.0000 & 0.32964 & -0.78268 & -0.84564 & 0.22516 & 0.99409 \\ 1.0000 & 0.66759 & -0.10865 & -0.81265 & -0.97639 & -0.49101 \\ 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 \end{bmatrix}$$

The column vectors $\boldsymbol{x}$ and $\boldsymbol{b}$ contain, respectively, the six unknown coefficients δ_0, δ_1, …, δ_5 of the fifth-degree approximating polynomial

$$\delta(x) = \delta_0 T_0(x) + \delta_1 T_1(x) + \cdots + \delta_5 T_5(x)$$

and the seven observed values of declination. They are

$$\boldsymbol{x} \equiv \begin{bmatrix} \delta_0 \\ \delta_1 \\ \vdots \\ \delta_5 \end{bmatrix} \qquad \boldsymbol{b} \equiv \begin{bmatrix} -16.300 \\ -2.0000 \\ \vdots \\ 76.100 \end{bmatrix}$$

As shown above, the weighting matrix $\mathbf{W}$ is preliminarily set equal to the 7×7 identity matrix $\mathbf{I}$. Now we form the products $\mathbf{A}^{\mathrm{T}}\mathbf{W}\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}\mathbf{W}\boldsymbol{b}$. Then, the values of the coefficients δ_0, δ_1, …, δ_5 (contained in the vector $\boldsymbol{x}$) of the fifth-degree approximating polynomial indicated above result from the product of the inverse of $\mathbf{A}^{\mathrm{T}}\mathbf{W}\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}\mathbf{W}\boldsymbol{b}$, as follows

$$\boldsymbol{x} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{W}\,\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{W}\,\boldsymbol{b}$$

By so doing, according to our preliminary evaluation (with $\mathbf{W} = \mathbf{I}$) of the weights w_{11}, w_{22}, …, w_{77}, the polynomial approximation which best fits, in the weighted least-squares sense, the observed declinations of the satellite is

$$\begin{aligned}\delta(x) = {} & 39.274\,T_0(x) + 54.564\,T_1(x) - 8.3776\,T_2(x) - 9.0385\,T_3(x) \\ & - 0.99597\,T_4(x) + 0.67411\,T_5(x)\end{aligned}$$

The approximation shown above makes it possible to compute the values of declination at the given arguments x_1, x_2, …, x_7. These values are compared with the observed values of declination; then the residuals ρ_1, ρ_2, …, ρ_7 (computed minus observed values) are also computed, as will be shown below.

Then, the weighting matrix $\mathbf{W}$ is updated by means of some non-negative weighting function $f(\rho_i)$ of the residuals ρ_i. A possible way to do this is to compute the new values (which will be placed in the rightmost column of the following table) of the weights w_{ii}, as suggested by Bube and Langan [14]:

$$w_{ii} = \frac{1}{\left[1 + \left(\frac{\rho_i}{\epsilon}\right)^2\right]^{\frac{1}{4}}}$$

where $i = 1, 2, \ldots, 7$, and ε is a positive constant, called damping parameter, whose value must be chosen by the solver. By so doing, we search the minimum value of the following hybrid loss function

$$J(\boldsymbol{x}) = \sum_{i=1}^{7} \left[1 + \left(\frac{\rho_i}{\epsilon}\right)^2\right]^{\frac{1}{4}} - 1$$

Bube and Langan suggest to take ε roughly equal to $1/1.643 \approx 0.6$ times the standard deviation σ of the residuals.

In the present case, with reference to the residuals ρ_1, ρ_2, …, ρ_7 given in the following table (which are computed with $\mathbf{W} = \mathbf{I}$), we have

$$\mu = \frac{\rho_1 + \rho_2 + \cdots + \rho_7}{7} = -0.30654 \times 10^{-6}$$

$$\sigma = \left[\frac{(\rho_1 - \mu)^2 + (\rho_2 - \mu)^2 + \cdots + (\rho_7 - \mu)^2}{7}\right]^{\frac{1}{2}} = 0.0050870$$

$$\epsilon = \frac{\sigma}{1.1643} = 0.0030961$$

and use this value of ε to compute new values of the weights, as shown below.

Obs. No.	x	Decl. comp. (°)	Decl. obs. (°)	Residual	Weight
1	−1.0000	−16.300	−16.300	0.00042725	0.99530
2	−0.66205	−2.0026	−2.0000	−0.0026114	0.87430
3	−0.32964	19.307	19.300	0.0066071	0.65140
4	0.0027701	46.891	46.900	−0.0089111	0.57289
5	0.32964	71.907	71.900	0.0066452	0.64987
6	0.66759	84.597	84.600	−0.0025940	0.87551
7	1.0000	76.100	76.100	0.00043488	0.99513

When the new values of the weights w_{11}, w_{22}, …, w_{77} have been computed, the weighting matrix $\mathbf{W}$ is updated as follows $\mathbf{W} = \text{diag}(w_{11}, w_{22}, \ldots, w_{77})$, and then the least-squares problem is solved again with the new weighting matrix, updated as indicated above. In the present case, the rightmost column of the preceding table, containing the weights to be used for the next iteration, has been filled as follows

$$w_{11} = \frac{1}{\left[1 + \left(\dfrac{0.00042725}{0.0030961}\right)^2\right]^{\frac{1}{4}}} = 0.99530$$

$$w_{22} = \frac{1}{\left[1 + \left(\dfrac{-0.0026114}{0.0030961}\right)^2\right]^{\frac{1}{4}}} = 0.87430$$

$$w_{33} = \frac{1}{\left[1 + \left(\frac{0.0066071}{0.0030961}\right)^2\right]^{\frac{1}{4}}} = 0.65140$$

$$w_{44} = \frac{1}{\left[1 + \left(\frac{-0.0089111}{0.0030961}\right)^2\right]^{\frac{1}{4}}} = 0.57289$$

$$w_{55} = \frac{1}{\left[1 + \left(\frac{0.0066452}{0.0030961}\right)^2\right]^{\frac{1}{4}}} = 0.64987$$

$$w_{66} = \frac{1}{\left[1 + \left(\frac{-0.0025940}{0.0030961}\right)^2\right]^{\frac{1}{4}}} = 0.87551$$

$$w_{77} = \frac{1}{\left[1 + \left(\frac{0.00043488}{0.0030961}\right)^2\right]^{\frac{1}{4}}} = 0.99513$$

and consequently the weighting matrix $\mathbf{W}$ is updated as follows

$$\mathbf{W} = \text{diag}(0.99530,\ 0.87430,\ 0.65140,\ 0.57289,\ 0.64987,\ 0.87551,\ 0.99513)$$

The iterative process described above comes to an end when the set of weights computed in a given iteration does not differ, within a specified tolerance, from the set computed in the preceding iteration. With an accuracy of five significant figures, two further iterations lead to the following results. Since the weights shown below are the same, within the accuracy of five significant figures, as those computed in the previous iteration, we stop here the iterative process.

The final approximating polynomial for declination is then

$$\begin{aligned}\delta(x) = {} & 39.274\,T_0(x) + 54.564\,T_1(x) - 8.3772\,T_2(x) - 9.0385\,T_3(x) \\ & - 0.99662\,T_4(x) + 0.67411\,T_5(x)\end{aligned}$$

Obs. No.	x	Decl. comp. (°)	Decl. obs. (°)	Residual	Weight
1	−1.0000	−16.300	−16.300	0.00042725	0.99530
2	−0.66205	−2.0019	−2.0000	−0.0026114	0.87431
3	−0.32964	19.306	19.300	0.0066071	0.65141

(continued)

(continued)

Obs. No.	x	Decl. comp. (°)	Decl. obs. (°)	Residual	Weight
4	0.0027701	46.890	46.900	−0.0089111	0.57289
5	0.32964	71.906	71.900	0.0066452	0.64987
6	0.66759	84.598	84.600	−0.0025940	0.87552
7	1.0000	76.100	76.100	0.00043488	0.99513

The final set of weights is given in the last column of the table shown above.

Now, we use again the 7×7 identity matrix $\mathbf{I}$ as the first estimate of the weighting matrix $\mathbf{W}$, in order to compute the weighted least-squares best fit for the observed values of right ascension.

The 7×6 matrix $\mathbf{A}$ is the same as that shown above for the case of declination. The column vectors $\boldsymbol{x}$ and $\boldsymbol{b}$ contain now the six unknown coefficients $\alpha_0, \alpha_1, \ldots, \alpha_5$ of the approximating polynomial

$$\alpha(x) = \alpha_0 T_0(x) + \alpha_1 T_1(x) + \cdots + \alpha_5 T_5(x)$$

and the seven observed values of right ascension. These vectors are

$$\boldsymbol{x} \equiv \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_5 \end{bmatrix} \qquad \boldsymbol{b} \equiv \begin{bmatrix} 327.00 \\ 325.25 \\ \vdots \\ 165.75 \end{bmatrix}$$

The results found iteratively for right ascension are given below.

First iteration ($\mathbf{W} = \mathbf{I}$):

Obs. No.	x	R.A. computed (°)	R.A. observed (°)	Residual	Weight
1	−1.0000	326.73	327.00	−0.26715	0.99535
2	−0.66205	326.89	325.25	1.6442	0.87412
3	−0.32964	318.78	322.94	−4.1562	0.65137
4	0.0027701	324.10	318.50	5.6040	0.57293
5	0.32964	302.82	307.00	−4.1806	0.64981
6	0.66759	226.88	225.25	1.6297	0.87572
7	1.0000	165.48	165.75	−0.27402	0.99511

where, again, the rightmost column of the table shown above contains the weights computed for the next iteration. The value of the damping parameter ε results from the residuals $\rho_1, \rho_2, \ldots, \rho_7$ (computed minus observed values of right ascension) given above, as shown below:

$$\mu = \frac{\rho_1 + \rho_2 + \cdots + \rho_7}{7} = -0.87193 \times 10^{-5}$$

$$\sigma = \left[\frac{(\rho_1 - \mu)^2 + (\rho_2 - \mu)^2 + \cdots + (\rho_7 - \mu)^2}{7}\right]^{\frac{1}{2}} = 3.1996$$

$$\epsilon = \frac{\sigma}{1.1643} = 1.9474$$

By updating the weighting matrix $\mathbf{W}$ (set previously equal to $\mathbf{I}$) by means of the values contained in the rightmost column of the table shown above, the values of the coefficients α_0, α_1, α_2, α_3, α_4, and α_5 of the fifth-degree approximating polynomial are determined, as follows

$$\alpha(x) = 278.87\,T_0(x) - 80.293\,T_1(x) - 39.000\,T_2(x) - 10.229\,T_3(x) + 6.2309\,T_4(x) + 9.8938\,T_5(x)$$

These values make it possible to compute new values of right ascension; these, in turn, are compared with the observed values of right ascension; then the residuals ρ_1, ρ_2, …, ρ_7 (computed minus observed values) are computed again.

Third (and last) iteration:

Obs. No.	x	R.A. comp. (°)	R.A. obs. (°)	Residual	Weight
1	−1.0000	326.83	327.00	−0.26715	0.99535
2	−0.66205	326.43	325.25	1.6442	0.87412
3	−0.32964	318.95	322.94	−4.1562	0.65137
4	0.0027701	324.62	318.50	5.6040	0.57293
5	0.32964	302.97	307.00	−4.1806	0.64981
6	0.66759	226.41	225.25	1.6297	0.87573
7	1.0000	165.58	165.75	−0.27402	0.99511

Since the weights are the same, within the accuracy of five significant figures, as those computed in the previous iteration, we stop here the iterative procedure.

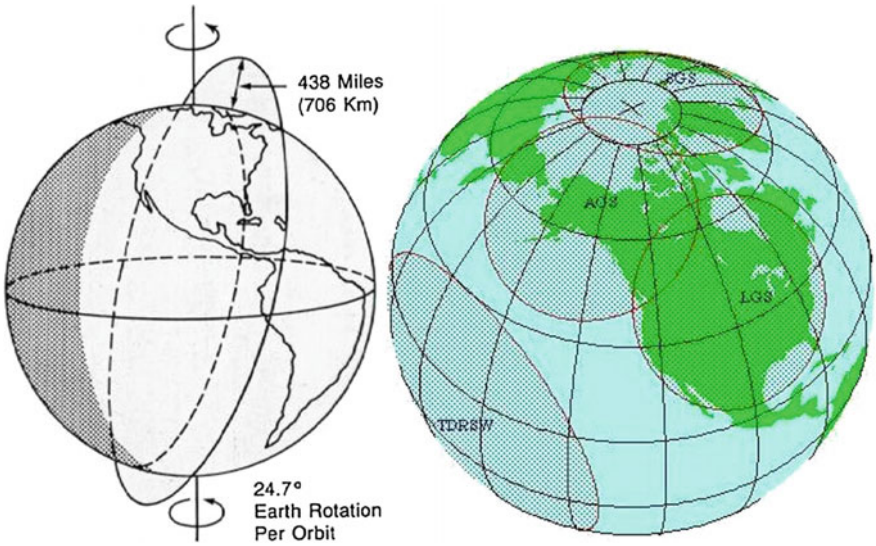
The final approximating polynomial for right ascension is then

$$\alpha(x) = 278.78\,T_0(x) - 80.289\,T_1(x) - 39.209\,T_2(x) - 10.228\,T_3(x) + 6.6361\,T_4(x) + 9.8892\,T_5(x)$$

The final set of weights is given in the last column of the table shown above.

2.11 The Kalman Filter

The batch estimation method described in the preceding paragraphs is based on the least-squares principle. It provides an estimate of the state vector $\boldsymbol{x}$ for a spacecraft at a given epoch by processing the whole amount of observations in each iteration. To this end, all the observations which are necessary to determine the spacecraft orbit must be available before the process of their improvement can take place. This requirement makes the method of batch estimation less desirable than other methods in real-time or near-real-time applications, which require a quasi-continuous update of information to produce an estimate of the state vector $\boldsymbol{x}$. As has been shown by Vergez et al. [80], one of these applications is the tracking of an Earth-orbiting satellite by means of ground stations placed on the surface of the Earth. Since the number of such stations is limited, an orbiting satellite cannot be tracked continuously. For example, the following figure, due to the courtesy of NASA [57], shows the orbit, the US ground stations and their acquisition circles used to track the Landsat 7 satellite.



A ground station can receive data from a satellite, when the satellite is within the acquisition circle of the station. When the link between the satellite and a ground station is lost, then the satellite position at some later time must be predicted, in order for another ground station to be able to establish a new link.

The tracking data result from measurements (range, azimuth angle, altitude angle, and their rates of change with time) made on an orbiting spacecraft by established ground stations. Alternative data are range measurements between two spacecraft and positions determined by using the Global Positioning System (GPS).

Such data require pre-processing to be put into a form useful for orbit state estimation. To this end, the range and angle data must be converted from a topocentric Earth-fixed co-ordinate system into a geocentric celestial co-ordinate system.

Another of these applications, according to Conway et al. [19], is the orbit control of a spacecraft, which requires accurate determination of the spacecraft position and velocity at a specified time.

One of the methods used for this purpose is the Kalman filter, which is a set of equations aimed at providing an optimal estimation of either the state vector or, equivalently, the osculating orbital elements of an orbiting spacecraft from a series of uncertain observations performed at discrete time-steps.

The Kalman filter is optimal because, in case of a linear system and a Gaussian distribution of errors, it minimises the mean square error of the estimated quantities. The Kalman filter is also recursive, because new measurements are processed as they arrive. The term "filter" comes from the theory of signal processing, where the information contained in a signal affected by noise must be extracted from the signal received, by filtering out the noise. The same term is used here, because the operation of estimating the state vector of a spacecraft from measurements affected by errors is equivalent to filtering out the errors.

The degree of goodness attained by the Kalman filter in performing this task is measured by the value of the loss function described in Sect. 2.9. In addition, the Kalman filter uses the history of the improved measurements to predict the subsequent evolution of the system states. To this end, two types of information are used by the filter:

- observations coming from appropriate measurement apparatuses (e.g. measurements of range, azimuth angle, altitude angle, and the time rate of change of the range and the angles, made from established ground stations to an orbiting spacecraft); and
- a mathematical model of the system under observation, describing how each state depends on the others and how the measurements depend on the states.

This requires the knowledge of the accuracy of both the measurements and the mathematical model of the observed system.

When the observed system is an orbiting spacecraft, the analytical model used for this purpose takes account of the central gravitational force and its perturbations. Some of the force models used for this purpose have been developed by the North American Aerospace Defence Command (NORAD), that maintains a catalogue containing the orbital parameters of about 8000 satellites and space debris computed from radar tracking data. Such models are described in Refs. [36], [54], and [78]. After computing the forces acting upon the spacecraft by means of the force model chosen, the state vector can be determined by numerically integrating the equations of motion with their initial conditions.

The result of the estimation is a predicted orbital state at the time of measurement, as well as the state error covariance matrix and the residuals (computed data

minus observed data). The magnitude of such residuals is expected to decrease with time as the filter converges to a stable estimate.

The classical Kalman filter has been created for linear systems. However, the behaviour of an orbiting spacecraft is governed by a nonlinear differential equation. The necessity of dealing with this and other nonlinear systems has given rise to the extended Kalman filter, in which the system equations are linearised about the reference trajectory. Therefore, the Kalman filter used for orbit determination is an extended Kalman filter, which estimates either the orbital elements or the Cartesian components of the state vector.

An example, taken from Rojas [62] will illustrate the operations indicated above. Let $x_1, x_2, \ldots, x_n$ be the results of measurements executed at times, respectively, t_1, $t_2, \ldots, t_n$. In case of an orbiting spacecraft, $x_1, x_2, \ldots, x_n$ are just the state vectors ($\boldsymbol{x}_1$, $\boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$) of the spacecraft resulting from n measurements. In the simple case of a one-dimensional system, $x_1, x_2, \ldots, x_n$ are scalars.

Let us consider for now the simple case. The mean μ_n of this time series is

$$\mu_n = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

When a new measurement x_{n+1} is executed at time t_{n+1}, the mean μ_{n+1} can be computed again as follows

$$\mu_{n+1} = \frac{x_1 + x_2 + \cdots + x_n + x_{n+1}}{n+1}$$

However, in order to compute the new value (μ_{n+1}) of the mean, it is more convenient to use the old value (μ_n) of the mean and make a small correction to it by means of x_{n+1}, as follows

$$\mu_{n+1} = \frac{n}{n+1}\left(\mu_n + \frac{x_{n+1}}{n}\right) = K(n\mu_n + x_{n+1})$$

where $K = 1/(n + 1)$ is called the gain factor. By adding and subtracting μ_n on the right-hand side of the equality

$$\mu_{n+1} = K(n\mu_n + x_{n+1})$$

and remembering the definition $K = 1/(n + 1)$, there results

$$\mu_{n+1} = K(n\mu_n + x_{n+1})$$

By so doing, the new value (μ_{n+1}) of the mean is expressed as a weighted mean of the old value (μ_n) of the mean and the new measurement (x_{n+1}). Since we trust more the old value (μ_n) of the mean than the new measurement (x_{n+1}), then the weight of μ_n is greater than the weight of x_{n+1}.

Not only the mean, but also the variance (also called quadratic standard deviation) of a time series can be computed recursively, as will be shown below.

Let us consider again the results x_1, x_2, …, x_n of n measurements executed at times, respectively, t_1, t_2, …, t_n. As shown in Sect. 2.9, the variance σ_n^2 of this time series is defined as follows

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_n)^2$$

When the result x_{n+1} of a new measurement comes to the filter, the new variance σ_{n+1}^2 is

$$\sigma_{n+1}^2 = \frac{1}{n+1}\sum_{i=1}^{n+1}\left(x_i - \mu_{n+1}\right)^2 = \frac{1}{n+1}\sum_{i=1}^{n+1}\left[x_i - \mu_n - K(x_{n+1} - \mu_n)\right]^2$$

where K is the gain factor defined above. By considering separately the last addend ($i = n + 1$) and carrying out the sum from 0 to n, the expression

$$\sum_{i=1}^{n+1}\left[x_i - \mu_n - K(x_{n+1} - \mu_n)\right]^2$$

can be written as follows

$$(1-K)^2(x_{n+1} - \mu_n)^2 + \sum_{i=1}^{n}\left[x_i - \mu_n - K(x_{n+1} - \mu_n)\right]^2$$

By expanding the square of the expression within square brackets, there results

$$(1-K)^2(x_{n+1} - \mu_n)^2 + nK^2(x_{n+1} - \mu_n)^2 + \sum_{i=1}^{n}(x_i - \mu_n)^2 - 2K\sum_{i=1}^{n}(x_i - \mu_n)(x_{n+1} - \mu_n)$$

The last term of the expression written above is zero, because

$$\sum_{i=1}^{n}(x_i - \mu_n) = 0$$

Therefore

$$\sum_{i=1}^{n+1}\left[x_i-\mu_n-K(x_{n+1}-\mu_n)\right]^2=(x_{n+1}-\mu_n)^2\left[(1-K)^2+nK^2\right]+\sum_{i=1}^{n}(x_i-\mu_n)^2$$

The definition $K = 1/(n + 1)$ implies that

$$(1-K)^2+nK^2=nK$$

In addition, the definition of variance

$$\sigma_n^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_n)^2$$

implies that

$$\sum_{i=1}^{n}(x_i-\mu_n)^2=n\,\sigma_n^2$$

Therefore

$$\sigma_{n+1}^2=\frac{1}{n+1}\left[n\sigma_n^2+nK(x_{n+1}-\mu_n)^2\right]=(1-K)\left[\sigma_n^2+K(x_{n+1}-\mu_n)^2\right]$$

The whole process comprises the following steps to be taken iteratively.
Let $x_1, x_2, \ldots, x_n$ be the results of n measurements.
We compute the mean

$$\mu_n=\frac{x_1+x_2+\cdots+x_n}{n}$$

and the variance

$$\sigma_n^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_n)^2$$

as has been shown above. Then we compute:

- the gain factor $K = 1/(n + 1)$ every time a new result x_{n+1} of the measurements comes to the filter;
- the new value of the mean $\mu_{n+1} = \mu_n + K(x_{n+1} - \mu_n)$;
- a preliminary estimate ${\sigma_{n+1}}^{*2}$ of the new standard deviation σ_{n+1}^2 by means of ${\sigma_{n+1}}^{*2} = \sigma_n^2 + K(x_{n+1} - \mu_n)^2$; and
- the correct value of the new standard deviation $\sigma_{n+1}^2 = (1 - K){\sigma_{n+1}}^{*2}$.

Let us consider now the general case, in which each of the measured values z is an n-dimensional vector. Let t_0 be the initial time of the process. The sequence of operations performed by the Kalman filter is described below:

(a) at time t_0, an initial estimate $\boldsymbol{x}_0^-$ of the m-dimensional state vector (including its uncertainty, expressed by the covariance matrix $\mathbf{P}_0^-$ associated with $\boldsymbol{x}_0^-$) is given to the filter;
(b) a new estimate, relating to a subsequent time t_1, is predicted by the filter on the basis of the mathematical model, and the uncertainty of this estimate is computed as a function of the initial uncertainty and the accuracy of the mathematical model;
(c) observations performed at time t_1 with a certain degree of accuracy provide the filter with new information, which is compared with the information coming from the predicted estimate and then used to compute a new updated estimate, relating to t_1, and the uncertainty of this new estimate; and
(d) still another estimate, relating to a subsequent time t_2, is predicted as in step (b), but now this estimate is based on the results of step (c).

This cycle, from step (b) to step (c), goes on until the measurements come to an end. The sequence of prediction (predict a state vector estimate and its covariance matrix to next time-step) and update (compute the Kalman gain; update the state vector estimate with measurements; and compute the new covariance matrix of the updated state vector estimate) is repeated each time new observations arrive.

Since the estimated state vector $\boldsymbol{x}_0^-$ (relating to the initial time t_0) contains m scalar random variables, its uncertainty is measured by the covariance matrix $\mathbf{P}_0^-$ associated with $\boldsymbol{x}_0^-$, that is, by the matrix having, along its main diagonal, the variances of these scalar random variables; and, in its off-diagonal terms, the covariances which represent any correlation between pairs of scalar variables. The simultaneous measurements z_1, z_2, … (taken at times, respectively, t_1, t_2, …) are also n-dimensional vectors processed sequentially in time by the recursive algorithm. At each cycle, the algorithm considers only the new measurement vector and the values coming from the previous cycle. Therefore, unlike the batch least-squares algorithm, the Kalman filter algorithm need not store in memory all past measurements.

The Kalman filter takes an initial estimated state vector and its estimated covariance matrix, and computes the weights (the Kalman gain) to be used to combine this estimate with the first state vector coming from the measurement executed. This yields an updated state vector estimate. Since the Kalman filter computes an updated state vector estimate by means of the new measurement, then the covariance matrix must also be changed, in order to take account of the new information added by the measurement. Therefore, the uncertainty of the state vector (expressed by the changed covariance matrix) decreases.

Now the Kalman filter must get ready to receive the next state vector coming from the next measurement. To this end, the Kalman filter must project ahead the

updated state vector estimate and its covariance matrix to the next measurement time. The state vector is assumed to change with time according to a linear law of transformation plus a random noise. The predicted state vector estimate can only follow this linear law of transformation, because the value of the noise is not known. The uncertainty of this predicted state vector estimate, measured by its covariance matrix, increases, because the prediction does not take the added noise into account. This is the last step of the Kalman filter cycle.

Since the state vectors coming from the measurements are recursively processed, then their uncertainty tends to decrease, because of the increasing amount of the information carried by each of them. On the other hand, since their uncertainty increases in the projection step, an equilibrium will be reached, where the uncertainty decrease, which occurs in the update step, is counter-balanced by the uncertainty increase, which occurs in the projection step.

If there were no random noise in the evolution of the observed system from one time-step to the next, the uncertainty of the state vector would reduce to zero.

Since this uncertainty changes with time, so does the Kalman gain, which must therefore be recomputed in each cycle.

We can illustrate now the equations of the extended Kalman filter in the general case, in which the measured values are n-dimensional vectors. Following Montenbruck and Gill [53], let $\boldsymbol{z}$ be a one of the batches, that is, one of the partitions, which make up the whole set of observations. Let $\boldsymbol{x}_0$ be the m-dimensional state vector at the reference epoch t_0. Let $\boldsymbol{x}_0^-$ be an a priori estimate (forecast) of $\boldsymbol{x}_0$, as indicated by a superscript minus sign, such that

$$\boldsymbol{x}_0^- = \boldsymbol{x}_0^{\text{ref}} + \Delta\boldsymbol{x}_0^-$$

where $\boldsymbol{x}_0^{\text{ref}}$ is a common state vector of reference, around which $\boldsymbol{x}_0^-$ is linearised, and $\Delta\boldsymbol{x}_0^-$ is the increment to be added to $\boldsymbol{x}_0^{\text{ref}}$ to obtain the estimate $\boldsymbol{x}_0^-$.

Let $\mathbf{P}_0^-$ be an estimate of the covariance matrix associated with $\boldsymbol{x}_0^-$. The estimates $\boldsymbol{x}_0^-$ and $\mathbf{P}_0^-$ can be obtained as a result of either the processing of previous data batches or initial information.

Let $\mathbf{H}$ be the Jacobian $\mathbf{H} = (\partial\boldsymbol{h}/\partial\boldsymbol{x}_0)_{\text{ref}}$, that is, the matrix of the partial derivatives of the computed values with respect to the state vector (here the subscript *ref* indicates that the partial derivatives are to be evaluated at $\boldsymbol{x}_0 = \boldsymbol{x}_0^{\text{ref}}$) at the reference epoch t_0. The vector-valued function $\boldsymbol{h}$ is the function appearing in the equation $\boldsymbol{z} = \boldsymbol{h}(\boldsymbol{x}_0) + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is the vector containing the measurement errors.

The measurement vector $\boldsymbol{z}$ and the a priori estimate $\boldsymbol{x}_0^-$ can be used to obtain an improved estimate $\Delta\boldsymbol{x}_0^+$ (where this improved estimate is indicated by a superscript plus sign), as follows

$$\Delta\boldsymbol{x}_0^+ = \mathbf{P}_0^+ \left((\mathbf{P}_0{}^-)^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{W}\,\Delta\boldsymbol{z}\right)$$

where $\Delta z = z - \boldsymbol{h}(\boldsymbol{x}_0^{\text{ref}})$ is the difference between the actual observations and the observations predicted on the basis of the reference trajectory, $\mathbf{W}$ is the weighting matrix, and

$$\mathbf{P}_0{}^{+} = \left((\mathbf{P}_0{}^{-})^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{H}\,\mathbf{H}\right)^{-1}$$

is the a posteriori covariance matrix, which takes into account the better knowledge of $\boldsymbol{x}_0$, resulting from the a priori information combined with the batch z of observations. The new estimate $\boldsymbol{x}_0^{+}$ is related to the previous estimate $\boldsymbol{x}_0^{-}$ by substituting the a priori information matrix $(\mathbf{P}_0^{-})^{-1}$ with the difference

$$(\mathbf{P}_0{}^{+})^{-1} - \left(\mathbf{H}^{\mathrm{T}}\mathbf{W}\,\mathbf{H}\right)$$

between the a posteriori information matrix and the measurement information matrix. This leads to

$$\Delta\boldsymbol{x}_0^{+} = \Delta\boldsymbol{x}_0^{-} + \mathbf{P}_0{}^{+}\mathbf{H}^{\mathrm{T}}\mathbf{W}\left(\Delta z - \mathbf{H}\Delta\boldsymbol{x}_0^{-}\right)$$

which estimates recursively $\Delta\boldsymbol{x}_0^{+}$. In the preceding expression, the matrix

$$\mathbf{K} = \mathbf{P}_0{}^{+}\mathbf{H}^{\mathrm{T}}\mathbf{W}$$

(which is called Kalman gain) and the residual vector

$$\boldsymbol{\rho} = \Delta z - \mathbf{H}\ \Delta\boldsymbol{x}_0^{-} = z - \boldsymbol{h}\left(\boldsymbol{x}_0^{\text{ref}}\right)\ - \mathbf{H}\,\Delta\boldsymbol{x}_0^{-}$$

are used to correct the estimates $\boldsymbol{x}_0^{-}$. The residuals contained in $\boldsymbol{\rho}$ are computed taking into account the results z of the measurements, the observations $\boldsymbol{h}(\boldsymbol{x}_0^{\text{ref}})$ computed according to the reference model, and a linear correction $\mathbf{H}\ \Delta\boldsymbol{x}_0^{-}$ which depends on the difference between the estimate $\boldsymbol{x}_0^{-}$ and the reference state $\boldsymbol{x}_0^{\text{ref}}$.

In practice, the a posteriori covariance matrix $\mathbf{P}_0^{+}$ is not computed as indicated above, that is, by using the expression

$$\mathbf{P}_0{}^{+} = \left((\mathbf{P}_0{}^{-})^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{W}\,\mathbf{H}\right)^{-1}$$

because this method requires, at each step, the inversion of an $m \times m$ matrix, where m is the dimension of the state vector. Instead, a more convenient way of computing the a posteriori covariance matrix $\mathbf{P}_0^{+}$ is given below

$$\mathbf{P}_0{}^{+} = (\mathbf{I} - \mathbf{K}\,\mathbf{H})\mathbf{P}_0{}^{-}$$

where the Kalman gain results from

$$\mathbf{K} = \mathbf{P}_0{}^{-}\mathbf{H}^{\mathrm{T}}\left(\mathbf{W}^{-1} + \mathbf{H}\,\mathbf{P}_0{}^{-}\mathbf{H}^{\mathrm{T}}\right)^{-1}$$

In summary, the recursive estimation algorithm is indicated below. Let the a priori reference state vector $\boldsymbol{x}_0^0$ be given together with the covariance matrix $\mathbf{P}_0^0$ associated with it. Let also a series of N measurement batches z_I (I = 1, 2, … N) be given. Then, recursive estimates $\boldsymbol{x}_0^I$ of the state vector $\boldsymbol{x}_0$ at epoch and their associate covariance matrices $\mathbf{P}_0^I$ are computed for each batch (I = 1, 2, …, N) by computing:

- the Kalman gain

$$\mathbf{K}_I = \mathbf{P}_0{}^{I-1}\mathbf{H}_I^{\mathrm{T}}\left(\mathbf{W}_I^{-1} + \mathbf{H}_I\mathbf{P}_0{}^{I-1}\mathbf{H}_I^{\mathrm{T}}\right)^{-1}$$

- the update of the state vector at epoch

$$\boldsymbol{x}_0^I = \boldsymbol{x}_0^{I-1} + \mathbf{K}_I\left[z_I - \boldsymbol{h}_I\left(\boldsymbol{x}_0^0\right) - \mathbf{H}_I\left(\mathbf{x}_0^{I-1} - \boldsymbol{x}_0^0\right)\right]$$

- the update of its covariance matrix at epoch

$$\mathbf{P}_0{}^I = (\mathbf{I} - \mathbf{K}_I\mathbf{H}_I)x\mathbf{P}_0{}^{I-1}$$

where $\mathbf{I}$ designates the $m \times m$ identity matrix.

The expressions given above refer to an arbitrary number of measurements in each batch. In practice, each batch comprises only a small number of measurements taken at a common epoch with possible correlations or even a single scalar observation.

In case of uncorrelated measurements, it is also possible to process them one at a time. When this happens, the measurement vector $\boldsymbol{z}_I$ is replaced by the corresponding scalar measurement z_i; in the same way, the weighting matrix $\mathbf{W}_I$ is replaced by the scalar weighting coefficient $w_i = 1/\sigma_i^2$, the Kalman gain matrix $\mathbf{K}_I$ is replaced by the gain vector $\boldsymbol{k}_i$ having the same dimension (m) as the state vector $\boldsymbol{x}_0$; the Jacobian $\mathbf{H} = (\partial \boldsymbol{h}/\partial \boldsymbol{x}_0)_{\mathrm{ref}}$ is a $1 \times m$ matrix (that is, a row vector), and the products $\mathbf{HPH}^{\mathrm{T}}$ and $\mathbf{H}\boldsymbol{x}$ are scalar quantities. Consequently, in this case, the three equations given above become, respectively,

$$\begin{aligned}
\boldsymbol{k}_i &= \mathbf{P}_0{}^{i-1}\mathbf{H}_i^{\mathrm{T}}\left(\sigma_i^2 + \mathbf{H}_i\mathbf{P}_0{}^{i-1}\mathbf{H}_i^{\mathrm{T}}\right)^{-1} \\
\boldsymbol{x}_0^i &= \boldsymbol{x}_0^{i-1} + \boldsymbol{k}_i\left[z_i - \mathrm{h}_i\left(\boldsymbol{x}_0^0\right) - \mathbf{H}_i\left(\boldsymbol{x}_0^{i-1} - \boldsymbol{x}_0^0\right)\right] \\
\mathbf{P}_0{}^i &= (\mathbf{I} - \boldsymbol{k}_i\mathbf{H}_i)\mathbf{P}_0{}^{i-1}
\end{aligned}$$

In the third of the equations written above, the tensor product $\boldsymbol{k}_i\ \mathbf{H}_i$ involves a column vector ($\boldsymbol{k}_i$) and a row vector ($\mathbf{H}_i$) and yields an $m \times m$ matrix.

This is because the tensor product of two m-dimensional vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is by definition

$$\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \dots & v_m \end{bmatrix} = \begin{bmatrix} u_1 v_1 & u_2 v_2 & \dots & u_1 v_m \\ u_2 v_1 & u_2 v_2 & \dots & u_2 v_m \\ \vdots & \vdots & \dots & \vdots \\ u_m v_1 & u_m v_2 & \dots & u_m v_m \end{bmatrix}$$

The three operations indicated above (computation of the Kalman gain, update of the state vector at epoch, and update of the covariance matrix of the state vector at epoch) provide an estimate of the state vector at epoch. In order to complete the recursion, they are followed by:

- a projection (also called propagation) of the reference state vector $\boldsymbol{x}_0^{\mathrm{ref}}$ from epoch t_0 to epoch t_1, or (which is the same) a projection of the reference state vector $\boldsymbol{x}_{i-1}^{\mathrm{ref}}$ from t_{i-1} to t_i; and
- a projection of the covariance matrix $\mathbf{P}_{i-1}$ from t_{i-1} to t_i.

To this end, let $\boldsymbol{\Phi}_i \equiv \boldsymbol{\Phi}(t_i, t_{i-1}) = \partial \boldsymbol{x}_i^{\mathrm{ref}} / \partial \boldsymbol{x}_{i-1}^{\mathrm{ref}} = \boldsymbol{\Phi}(t_i, t_0)\boldsymbol{\Phi}(t_{i-1}, t_0)^{-1}$ denote the state transition matrix from epoch t_{i-1} to epoch t_i, that is, the matrix whose product with the state vector $\boldsymbol{x}_{i-1}$ at time t_{i-1} gives the state vector $\boldsymbol{x}_i$ at time t_i, as follows

$$\boldsymbol{x}_i = \boldsymbol{\Phi}_i \boldsymbol{x}_{i-1}$$

The state vector $\boldsymbol{x}_{i-1}^{+}$ (resulting from data up to and including time t_{i-1}) is used to predict an a priori state vector $\boldsymbol{x}_i^{-}$ at time t_i, as follows

$$\boldsymbol{x}_i^{-} = \boldsymbol{x}_i^{\mathrm{ref}} + \boldsymbol{\Phi}_i \left(\boldsymbol{x}_{i-1}^{+} - \boldsymbol{x}_{i-1}^{\mathrm{ref}} \right)$$

and its covariance matrix $\mathbf{P}_i^{-}$ at time t_i, as follows

$$\begin{aligned} \mathbf{P}_i^{-} &= E\left\{ \left[\boldsymbol{x}_i^{-} - E\left(\boldsymbol{x}_i^{-} \right) \right] \left[\boldsymbol{x}_i^{-} - E\left(\boldsymbol{x}_i^{-} \right) \right]^{\mathrm{T}} \right\} \\ &= E\left\{ \boldsymbol{\Phi}_i \left[\boldsymbol{x}_{i-1}^{+} - E\left(\boldsymbol{x}_{i-1}^{+} \right) \right] \left[x_{i-1}^{+} - E\left(\boldsymbol{x}_{i-1}^{+} \right) \right]^{\mathrm{T}} \boldsymbol{\Phi}_i^{\mathrm{T}} \right\} \\ &= \boldsymbol{\Phi}_i \mathbf{P}_{i-1}^{+} \boldsymbol{\Phi}_i^{\mathrm{T}} \end{aligned}$$

where $\mathbf{P}_{i-1}^{+}$ is the covariance of the state vector $\boldsymbol{x}_{i-1}^{+}$. At this stage of the recursive algorithm, the observations $\boldsymbol{z}_i$ have not yet been taken into account. Therefore, the information contained in $\boldsymbol{x}_i^{-}$ and in $\mathbf{P}_i^{-}$ (which are, respectively, the predicted state vector at time t_i and its covariance matrix) is the same as that contained in $\boldsymbol{x}_{i-1}^{+}$ and in $\mathbf{P}_{i-1}^{+}$ (which are, respectively, the improved state vector at time t_{i-1} and its covariance matrix), except for the time to which such quantities refer.

Now, the observations z_i at time t_i have to be taken into account to update the a priori information. For this purpose, the residual vector $\boldsymbol{\rho}_i$ is expressed as a function of quantities related to time t_i instead of t_0, as follows

$$\begin{aligned}\boldsymbol{\rho}_i &= \boldsymbol{z}_i - \boldsymbol{h}_i\left(\boldsymbol{x}_0^{\text{ref}}\right) - \mathbf{H}_i\left(\boldsymbol{x}_0^{-} - \boldsymbol{x}_0^{\text{ref}}\right) \\ &= \boldsymbol{z}_i - \boldsymbol{g}_i\left(\boldsymbol{x}_i^{\text{ref}}\right) - \mathbf{G}_i\left(\boldsymbol{x}_i^{-} - \boldsymbol{x}_i^{\text{ref}}\right)\end{aligned}$$

where the function $\boldsymbol{h}_i(\boldsymbol{x}_0^{\text{ref}})$, which is used to predict the observations on the basis of the reference trajectory, has been replaced by the following function

$$\boldsymbol{g}_i[t_i, \boldsymbol{x}(t_i)] = \boldsymbol{h}_i[t_i, \boldsymbol{x}(t_0)]$$

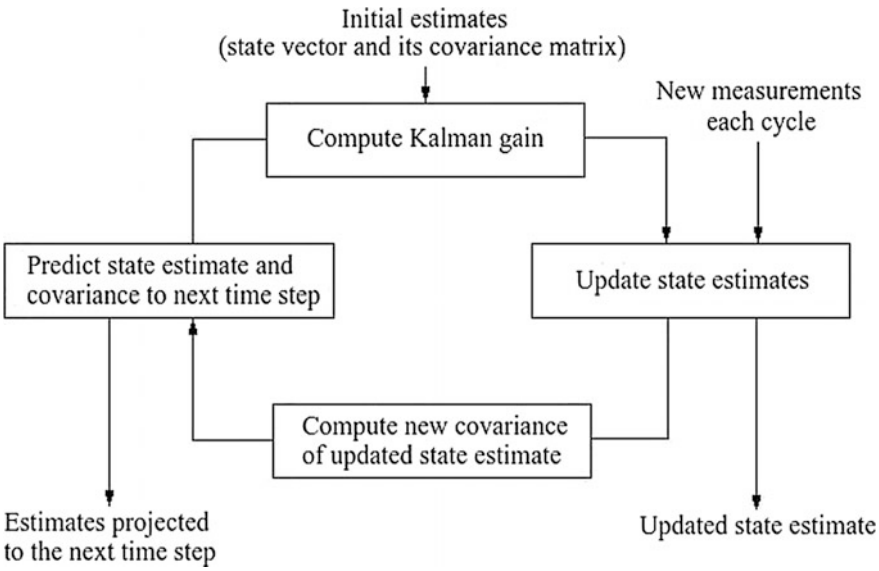
which models the observations as a function of the state vector at the time t_i of measurement. Likewise, the Jacobian $\mathbf{H}_i$ has been expressed as follows

$$\mathbf{H}_i = \frac{\partial \boldsymbol{h}_i}{\partial \boldsymbol{x}_0^{\text{ref}}} = \left(\frac{\partial \boldsymbol{g}_i}{\partial \boldsymbol{x}_i^{\text{ref}}}\right)\left(\frac{\partial \boldsymbol{x}_i^{\text{ref}}}{\partial \boldsymbol{x}_0^{\text{ref}}}\right) = \mathbf{G}_i \boldsymbol{\Phi}(t_i, t_0)$$

(where $\mathbf{G}_i = \partial \boldsymbol{g}_i / \partial \boldsymbol{x}_i^{\text{ref}}$) in the equation of the residual vector $\boldsymbol{\rho}_i$. In the same way, the Kalman gain is expressed as follows

$$\mathbf{K}_i = \mathbf{P}_i^{-} \mathbf{G}_i^{\mathrm{T}} \left(\mathbf{W}_i^{-1} + \mathbf{G}_i \mathbf{P}_i^{-} \mathbf{G}_i^{\mathrm{T}}\right)^{-1}$$

Levy [49] illustrates the recursive estimation algorithm used in the Kalman filter by means of a scheme, which is shown in the following figure.



The blocks of this scheme represent the following equations:

Compute Kalman gain	$\boldsymbol{k}_i = \mathbf{P}_0^{i-1}\,\mathbf{H}_i^{\mathrm{T}}\,(\sigma_i^2 + \mathbf{H}_i\,\mathbf{P}_0^{i-1}\,\mathbf{H}_i^{\mathrm{T}})^{-1}$
Update state estimates	$\boldsymbol{x}_0^i = \boldsymbol{x}_0^{i-1} + \boldsymbol{k}_i\,[z_i - \mathrm{h}_i\,(\boldsymbol{x}_0^0) - \mathbf{H}_i\,(\mathbf{x}_0^{i-1} - \boldsymbol{x}_0^0)]$
Compute new covariance of updated state estimate	$\mathbf{P}_0^i = (\mathbf{I} - \boldsymbol{k}_i\,\mathbf{H}_\mathrm{i})\,\mathbf{P}_0^{i-1}$
Predict state estimate and covariance to next time-step	$\boldsymbol{x}_i^- = \boldsymbol{x}_i^{\mathrm{ref}} + \boldsymbol{\Phi}_i\,(\boldsymbol{x}_{i-1}^+ - \boldsymbol{x}_{i-1}^{\mathrm{ref}})$ $\mathbf{P}_i^- = \boldsymbol{\Phi}_i\,\mathbf{P}_{i-1}^+\,\boldsymbol{\Phi}_i^{\mathrm{T}}$

The linearised Kalman filter computes estimates $\boldsymbol{x}_i^+$ (where $i = 1, 2, \ldots, n$) of the state vector $\boldsymbol{x}_i$ at measurement times t_i and the covariance matrices $\mathbf{P}_i^+$ associated with these estimates.

Then, the time update phase starts with the integration of the equation of motion and the variational equations from t_{i-1} to t_i, to obtain the reference state vector $\boldsymbol{x}_i^{\mathrm{ref}}$ and the state transition matrix $\boldsymbol{\Phi}_i = \partial \boldsymbol{x}_i^{\mathrm{ref}}/\partial \boldsymbol{x}_{i-1}^{\mathrm{ref}}$. By means of these quantities, the previous estimate $\boldsymbol{x}_{i-1}^+$ and the associated covariance matrix $\mathbf{P}_{i-1}^+$ can be projected from t_{i-1} to t_i, where t_i is the time of the current measurement, as follows

$$\begin{aligned}\boldsymbol{x}_i^- &= \boldsymbol{x}_i^{\mathrm{ref}} + \boldsymbol{\Phi}_i\left(\boldsymbol{x}_{i-1}^+ - \boldsymbol{x}_i^{\mathrm{ref}}\right)\\ \mathbf{P}_i^- &= \boldsymbol{\Phi}_i\mathbf{P}_{i-1}^+\boldsymbol{\Phi}_i^{\mathrm{T}}\end{aligned}$$

The measurement update phase starts with the computation of the Kalman gain $\mathbf{K}_i$, the state vector update $\boldsymbol{x}_i^+$, and the covariance matrix update $\mathbf{P}_i^+$, as follows

$$\begin{aligned}\mathbf{K}_i &= \mathbf{P}_i^-\mathbf{G}_i^{\mathrm{T}}\left(\mathbf{W}_i^{-1} + \mathbf{G}_i\mathbf{P}_i^-\mathbf{G}_i^{\mathrm{T}}\right)^{-1}\\ \boldsymbol{x}_i^+ &= \boldsymbol{x}_i^- + \mathbf{K}_i\left[\boldsymbol{z}_i - \boldsymbol{g}_i\left(\boldsymbol{x}_i^{\mathrm{ref}}\right) - \mathbf{G}_i\left(\boldsymbol{x}_i^- - \boldsymbol{x}_i^{\mathrm{ref}}\right)\right]\\ \mathbf{P}_i^+ &= (\mathbf{I} - \mathbf{K}_i\mathbf{G}_i)\mathbf{P}_i^-\end{aligned}$$

The starting values to be given to the filter are $\boldsymbol{x}_0^+ = \boldsymbol{x}_0^{\mathrm{ref}}$ and $\mathbf{P}_0^+ = \mathbf{P}_0^{\mathrm{ref}}$.

The ordinary Kalman filter described above can be used when the deviations between the reference state vector ($\boldsymbol{x}_{i-1}^{\mathrm{ref}}$) and the estimated state vector ($\boldsymbol{x}_{i-1}^+$) are small at any time (t_{i-1}). In order to remove this restriction, the extended Kalman filter has been developed, which derives from the ordinary Kalman filter by resetting the reference state vector to the estimated state vector at the beginning of each step. In the successive phase of time update, the current estimated state vector ($\boldsymbol{x}_{i-1}^+$) is projected ahead from the current epoch (t_{i-1}) to the next (t_i), and the variational equations for the state transition matrix are simultaneously solved. This yields the predicted state vector $\boldsymbol{x}_i^-$ at epoch t_i

$$\boldsymbol{x}_i^- = \boldsymbol{x}[t_i; \boldsymbol{x}(t_{i-1})] = \boldsymbol{x}_{i-1}^+$$

and the covariance matrix $\mathbf{P}_i^-$ associated with $\boldsymbol{x}_i^-$

$$\mathbf{P}_i^- = \boldsymbol{\Phi}_i \mathbf{P}_{i-1}^+ \boldsymbol{\Phi}_i^{\mathrm{T}}$$

The measurement update phase for the extended Kalman filter differs very little from that for the ordinary Kalman filter, the only difference being the state vector update equation, which is more simple in the former case:

$$\begin{aligned}
\mathbf{K}_i &= \mathbf{P}_i^- \mathbf{G}_i^{\mathrm{T}} \left(\mathbf{W}_i^{-1} + \mathbf{G}_i \mathbf{P}_i^- \mathbf{G}_i^{\mathrm{T}} \right)^{-1} \\
\boldsymbol{x}_i^+ &= \boldsymbol{x}_i^- + \mathbf{K}_i \left[\boldsymbol{z}_i - \boldsymbol{g}_i \left(\boldsymbol{x}_i^- \right) \right] \\
\mathbf{P}_i^+ &= (\mathbf{I} - \mathbf{K}_i \mathbf{G}_i) \mathbf{P}_i^-
\end{aligned}$$

The starting values to be given to the filter are the a priori values of the state vector ($\boldsymbol{x}_0 = \boldsymbol{x}_0^{\mathrm{apr}}$) and of the associated covariance matrix ($\mathbf{P}_0 = \mathbf{P}_0^{\mathrm{apr}}$).

The better performance of the extended Kalman filter is obtained at the cost of a heavier computational effort in the phase of projection ahead of the state vector and associated covariance matrix. This is because, when an extended Kalman filter is used, a separate initial-value problem must be solved by numerical integration for each measurement to be processed. Montenbruck and Gill [53] point out that low-order single-step methods (e.g. the fourth-order Runge-Kutta method) are employed in real-time orbit determination programs based on the extended Kalman filter.

2.12 Numerical Methods for Kalman Filtering

As is the case with the batch least-squares method, numerical errors may sometimes impair the performance of the Kalman filter, unless special care is taken in the computation. Hotop [37] has shown that this is due to the presence of a minus sign between two matrices in the update expression shown in the preceding paragraph, that is,

$$\mathbf{P}_k^+ = (\mathbf{I} - \mathbf{K}_k \mathbf{G}_k) \mathbf{P}_k^-$$

where the subscript k is the time index. Evaluating this expression on a computer with single precision (REAL*4) may give rise to negative elements along the main diagonal of the covariance matrix, which conflicts with the theory of covariance matrices [37]. As has been shown by Grewal and Kain [30], the covariance matrix must be symmetric and positive definite (in Sect. 2.9, it has been shown that an $n \times n$ real symmetric matrix $\mathbf{A}$ is positive definite if the equality $\boldsymbol{z}^{\mathrm{T}} \mathbf{A} \boldsymbol{z} > 0$ holds for all nonzero vectors $\boldsymbol{z}$ with real entries); otherwise that matrix cannot represent valid statistics for state vector components.

A new formulation of the update expression, with the view of eliminating the negative diagonal elements, is the so-called Joseph algorithm or stabilised Kalman algorithm [15]. Without going into the analytical details of the matter, suffice it to mention that Joseph uses, for the update of the covariance matrix $\mathbf{P}_k^+$, the following expression

$$\mathbf{P}_k^+ = (\mathbf{I} - \mathbf{K}_k\mathbf{G}_k)\mathbf{P}_k^-(\mathbf{I} - \mathbf{K}_k\mathbf{G}_k)^{\mathrm{T}} + \mathbf{K}_k\mathbf{W}_k^{-1}\mathbf{K}_k^{\mathrm{T}}$$

which, according to Thornton [69], is mechanised as follows:

$$\begin{aligned}\mathbf{W}_1 &= (\mathbf{I} - \mathbf{K}_k\mathbf{G}_k)\\ \mathbf{W}_2 &= \mathbf{W}_1\mathbf{P}_k^-\\ \mathbf{P}_k^+ &= \mathbf{W}_2\mathbf{W}_1^{\mathrm{T}} + \mathbf{K}_k\mathbf{W}_k^{-1}\mathbf{K}_k^{\mathrm{T}}\end{aligned}$$

According to Montenbruck and Gill [53], the Joseph algorithm ensures the positive definiteness of $\mathbf{P}_k^+$ irrespective of errors in $\mathbf{K}_k$ or in $(\mathbf{I} - \mathbf{K}_k\,\mathbf{G}_k)$.

Another method, due to Bierman [9] and meant to the same effect, computes the update of the covariance matrix $\mathbf{P}_k^+$ by means of tensor products (see Sect. 2.10), as follows

$$\begin{aligned}\mathbf{V}_1 &= \mathbf{P}_k^-\mathbf{G}_k^{\mathrm{T}}\\ \mathbf{P}_1 &= \mathbf{P}_k^- - \mathbf{K}_k\mathbf{V}_1^{\mathrm{T}}\\ \mathbf{V}_2 &= \mathbf{P}_1\mathbf{G}_k^{\mathrm{T}}\\ \mathbf{P}_k^+ &= \left(\mathbf{P}_1 - \mathbf{V}_2\mathbf{K}_k^{\mathrm{T}}\right) + \mathbf{K}_k\mathbf{W}_k^{-1}\mathbf{K}_k^{\mathrm{T}}\end{aligned}$$

However, as has been shown by Thornton [69], in either the Joseph or the Bierman formulation, this method requires more than two times the arithmetic operations required by the original Kalman method and, in addition, is still susceptible to numerical errors.

In order to avoid the degradation of the computed covariance matrix $\mathbf{P}_k^+$, Potter [61] and Schmidt [63] decompose the covariance matrix as follows

$$\mathbf{P}_k = \mathbf{S}_k\mathbf{S}_k^{\mathrm{T}}$$

and derive an algorithm, called square-root method, for recursively computing $\mathbf{S}_k$ instead of $\mathbf{P}_k$. By so doing, symmetric products of triangular factors ($\mathbf{S}_k$ and $\mathbf{S}_k^{\mathrm{T}}$) for the covariance matrix are updated instead of the covariance matrix itself.

As shown in Sect. 2.9, any $m \times m$ symmetric positive definite matrix $\mathbf{P}$ has a unique decomposition $\mathbf{P} = \mathbf{S}\mathbf{S}^{\mathrm{T}}$, where $\mathbf{S}$ is a lower triangular matrix with positive diagonal entries ($s_{ii} > 0$), and $\mathbf{S}^{\mathrm{T}}$ is the transpose of $\mathbf{S}$. The decomposition $\mathbf{P} = \mathbf{S}\mathbf{S}^{\mathrm{T}}$ can be computed by means of the Cholesky method. The square-root method guarantees positive definiteness of the computed covariance matrix $\mathbf{P}_k^+$, because the

matrix, if kept in this form, can never have a negative diagonal or become asymmetric. Therefore, a square-root filter is more stable numerically than a conventional Kalman filter.

Following Born [10], we consider the time update equation

$$\mathbf{P}_k^- = \mathbf{\Phi}_k \mathbf{P}_{k-1}^+ \mathbf{\Phi}_k^{\mathrm{T}}$$

concerning the covariance matrix $\mathbf{P}_k^-$ associated with the predicted a priori state vector $\boldsymbol{x}_k^-$ at time t_k. In order to simplify the notation, the subscripts are dropped, so that the same equation can be rewritten as follows

$$\mathbf{P}^- = \mathbf{\Phi}\,\mathbf{P}^+\,\mathbf{\Phi}^{\mathrm{T}}$$

Using the decomposition $\mathbf{P^+ = SS}^{\mathrm{T}}$, the preceding equation becomes

$$\mathbf{P}^- = \mathbf{\Phi}\,\mathbf{S}\,\mathbf{S}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}} \equiv \mathbf{S}^-\mathbf{S}^{-\mathrm{T}}$$

where $\mathbf{S}^- = \mathbf{\Phi}\ \mathbf{S}$. Assuming that the observations are processed one at a time and that their errors are uncorrelated, the expression written above for the Kalman gain, that is,

$$\mathbf{K}_k = \mathbf{P}_k^- \mathbf{G}_k^{\mathrm{T}} \left(\mathbf{W}_k^{-1} + \mathbf{G}_k \mathbf{P}_k^- \mathbf{G}_k^{\mathrm{T}}\right)^{-1}$$

becomes

$$\mathbf{K} = \mathbf{S}^-\mathbf{S}^{-\mathrm{T}}\mathbf{G}^{\mathrm{T}} \left(\sigma^2 + \mathbf{G}\mathbf{S}^-\mathbf{S}^{-\mathrm{T}}\mathbf{G}^{\mathrm{T}}\right)^{-1}$$

where σ^2 is the variance of the observation error. We set by definition

$$\begin{aligned} \alpha &= \left(\sigma^2 + \mathbf{G}\,\mathbf{S}^-\mathbf{S}^{-\mathrm{T}}\mathbf{G}^{\mathrm{T}}\right)^{-1} \\ \boldsymbol{f}^- &= \mathbf{S}^{-\mathrm{T}}\mathbf{G}^{\mathrm{T}} \end{aligned}$$

where α is a scalar and $\boldsymbol{f}^-$ is an n-dimensional column vector

$$\boldsymbol{f}^- \equiv \begin{bmatrix} f_1^- \\ \vdots \\ f_n^- \end{bmatrix}$$

By taking the transposes of the matrices on both sides of the equality

$$\boldsymbol{f}^- = \mathbf{S}^{-\mathrm{T}}\mathbf{G}^{\mathrm{T}}$$

there results

$$\boldsymbol{f}^{-\mathrm{T}} = \left(\mathbf{S}^{-\mathrm{T}}\mathbf{G}^{\mathrm{T}}\right)^{\mathrm{T}} = \left(\mathbf{G}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\mathbf{S}^{-\mathrm{T}}\right)^{\mathrm{T}} = \mathbf{G}\,\mathbf{S}^{-}$$

where $\boldsymbol{f}^{-\mathrm{T}}$ (the transpose of $\boldsymbol{f}^{-}$) is the n-dimensional row vector $\boldsymbol{f}^{-\mathrm{T}} \equiv [f_1^-, \ldots, f_n^-]$.

This follows from the identities

$$\left(\mathbf{A}^{\mathrm{T}}\right)^{\mathrm{T}} = \mathbf{A}$$
$$(\mathbf{A}\,\mathbf{B})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$$

Therefore, there results

$$\alpha = \left(\sigma^2 + \boldsymbol{f}^{-\mathrm{T}}\boldsymbol{f}^{-}\right)^{-1}$$
$$\mathbf{K} = \alpha\,\mathbf{S}^{-}\boldsymbol{f}^{-}$$

Consequently, the new covariance matrix $\mathbf{P}^{+}$ (associated with the updated state estimate $\boldsymbol{x}^{+}$) can be written as follows

$$\mathbf{P}^{+} = \mathbf{S}^{+}\,\mathbf{S}^{+\mathrm{T}} = (\mathbf{I} - \mathbf{K}\,\mathbf{G})\mathbf{P}^{-} = (\mathbf{I} - \alpha\,\mathbf{S}^{-}\boldsymbol{f}^{-}\mathbf{G})\mathbf{S}^{-}\mathbf{S}^{-\mathrm{T}} = \mathbf{S}^{-}(\mathbf{I} - \alpha\boldsymbol{f}^{-}\mathbf{G}\,\mathbf{S}^{-})\mathbf{S}^{-\mathrm{T}}$$
$$= \mathbf{S}^{-}\left(\mathbf{I} - \alpha\boldsymbol{f}^{-}\boldsymbol{f}^{-\mathrm{T}}\right)\mathbf{S}^{-\mathrm{T}}$$

Let $\mathbf{A}^{-}$ be a matrix defined as follows

$$\mathbf{A}^{-}\mathbf{A}^{-\mathrm{T}} = \left(\mathbf{I} - \alpha\boldsymbol{f}^{-}\boldsymbol{f}^{-\mathrm{T}}\right)$$

If such a matrix can be found, then the new covariance matrix $\mathbf{P}^{+}$ is expressible as

$$\mathbf{P}^{+} = \mathbf{S}^{-}\mathbf{A}^{-}\mathbf{A}^{-\mathrm{T}}\mathbf{S}^{-\mathrm{T}} = \mathbf{S}^{+}\mathbf{S}^{+\mathrm{T}}$$

To find $\mathbf{A}^{-}$, we introduce a scalar γ such that

$$\mathbf{A}^{-}\mathbf{A}^{-\mathrm{T}} = (\mathbf{I} - \gamma\,\alpha\boldsymbol{f}^{-}\boldsymbol{f}^{-T})(\mathbf{I} - \gamma\,\alpha\boldsymbol{f}^{-}\boldsymbol{f}^{-\mathrm{T}}) = \left(\mathbf{I} - \alpha\boldsymbol{f}^{-}\boldsymbol{f}^{-\mathrm{T}}\right)$$

that is,

$$\mathbf{I} - 2\,\gamma\,\alpha\boldsymbol{f}^{-}\boldsymbol{f}^{-\mathrm{T}} + \gamma^2\alpha^2\boldsymbol{f}^{-}\boldsymbol{f}^{-\mathrm{T}}\boldsymbol{f}^{-}\boldsymbol{f}^{-\mathrm{T}} = \mathbf{I} - \alpha\boldsymbol{f}^{-}\boldsymbol{f}^{-\mathrm{T}}$$

Now we define a scalar β such that

$$\beta = \boldsymbol{f}^{-\mathrm{T}}\boldsymbol{f}^{-}$$

It follows that

$$\mathbf{I} - 2\,\gamma\,\alpha \boldsymbol{f}^{-}\boldsymbol{f}^{-\mathrm{T}} + \gamma^2\alpha^2\beta \boldsymbol{f}^{-}\boldsymbol{f}^{-\mathrm{T}} = \mathbf{I} - \alpha \boldsymbol{f}^{-}\boldsymbol{f}^{-\mathrm{T}}$$

that is,

$$\begin{aligned}\left(\gamma^2\alpha^2\beta - 2\,\gamma\,\alpha + \alpha\right)\boldsymbol{f}^{-}\boldsymbol{f}^{-\mathrm{T}} &= 0\\ \left(\alpha\beta\gamma^2 - 2\gamma + 1\right)\alpha \boldsymbol{f}^{-}\boldsymbol{f}^{-\mathrm{T}} &= 0\end{aligned}$$

The equality written above is satisfied by the trivial solution $\alpha \boldsymbol{f}^{-}\boldsymbol{f}^{-\mathrm{T}} = 0$. It is also satisfied by the two values of γ for which the expression $\alpha\beta\gamma^2 - 2\gamma + 1$ vanishes.

These two values are

$$\gamma = \frac{1}{\alpha\beta} \pm \left[\frac{1}{(\alpha\beta)^2} - \frac{1}{\alpha\beta}\right]^{\frac{1}{2}}$$

Remembering the expressions written above

$$\begin{aligned}\alpha &= \left(\sigma^2 + \boldsymbol{f}^{-\mathrm{T}}\boldsymbol{f}^{-}\right)^{-1}\\ \beta &= \boldsymbol{f}^{-\mathrm{T}}\boldsymbol{f}^{-}\end{aligned}$$

there results

$$\alpha = \frac{1}{\sigma^2 + \beta}$$

hence

$$\frac{1}{\alpha} = \sigma^2 + \beta \qquad \beta = \frac{1}{\alpha} - \sigma^2 = \frac{1 - \alpha\sigma^2}{\alpha}$$

$$\frac{1}{\beta} = \frac{\alpha}{1 - \alpha\sigma^2} \qquad \frac{1}{\alpha\beta} = \frac{1}{1 - \alpha\sigma^2}$$

Therefore, the two values of γ corresponding to $\alpha\beta\gamma^2 - 2\gamma + 1 = 0$ are

$$\gamma = \frac{1}{\alpha\beta} \pm \left(\frac{1}{\alpha^2\beta^2} - \frac{1}{\alpha\beta}\right)^{\frac{1}{2}} = \frac{1 \pm (\alpha\sigma^2)^{\frac{1}{2}}}{1 - \alpha\sigma^2}$$

If we choose the upper sign (+) in the expression written above, then there results

$$\gamma_1 = \frac{1 + (\alpha\sigma^2)^{\frac{1}{2}}}{1 - \alpha\sigma^2} = \frac{1 + (\alpha\sigma^2)^{\frac{1}{2}}}{\left[1 + (\alpha\sigma^2)^{\frac{1}{2}}\right]\left[1 - (\alpha\sigma^2)^{\frac{1}{2}}\right]} = \frac{1}{1 - (\alpha\sigma^2)^{\frac{1}{2}}}$$

Otherwise, if we choose the lower sign (−) in the same expression, then there results

$$\gamma_2 = \frac{1 - (\alpha\sigma^2)^{\frac{1}{2}}}{1 - \alpha\sigma^2} = \frac{1 - (\alpha\sigma^2)^{\frac{1}{2}}}{\left[1 + (\alpha\sigma^2)^{\frac{1}{2}}\right]\left[1 - (\alpha\sigma^2)^{\frac{1}{2}}\right]} = \frac{1}{1 + (\alpha\sigma^2)^{\frac{1}{2}}}$$

In order to prevent the denominator from becoming zero in case of $\alpha\sigma^2 = 1$, we discard γ_1 and take

$$\gamma_2 = \frac{1}{1 + (\alpha\sigma^2)^{\frac{1}{2}}}$$

as the unique solution of the equation $\alpha\beta\gamma^2 - 2\gamma + 1 = 0$.

Remembering the expressions written above

$$\begin{aligned} \mathbf{P}^+ &= \mathbf{S}^-\mathbf{A}^-\mathbf{A}^{-\mathrm{T}}\mathbf{S}^{-\mathrm{T}} = \mathbf{S}^+\mathbf{S}^{+\mathrm{T}} \\ \mathbf{A}^-\mathbf{A}^{-\mathrm{T}} &= \left(\mathbf{I} - \gamma\,\alpha\boldsymbol{f}^-\boldsymbol{f}^{-\mathrm{T}}\right)\left(\mathbf{I} - \gamma\,\alpha\boldsymbol{f}^-\boldsymbol{f}^{-\mathrm{T}}\right) = \left(\mathbf{I} - \alpha\boldsymbol{f}^-\boldsymbol{f}^{-\mathrm{T}}\right) \\ \mathbf{K} &= \alpha\,\mathbf{S}^-\,\mathbf{f}^- \end{aligned}$$

there results

$$\mathbf{S}^+ = \mathbf{S}^-\mathbf{A}^- = \mathbf{S}^-\left(\mathbf{I} - \gamma\,\alpha\boldsymbol{f}^-\boldsymbol{f}^{-\mathrm{T}}\right) = \mathbf{S}^- - \gamma\,\mathbf{K}\boldsymbol{f}^{-\mathrm{T}}$$

which is the measurement update for the square-root matrix $\mathbf{S}$.

The resulting computational algorithm is indicated below. In this algorithm, we use again the subscripts k to indicate the states.

At a given state $k = 1, 2, \ldots, n$ (corresponding to a time t_k), the following quantities are to be specified: $\boldsymbol{x}_k^-$ (state vector), $\mathbf{S}_k^-$ (square-root matrix associated with $\boldsymbol{x}_k^-$, where $\mathbf{S}_k^-$ is such that $\mathbf{P}_k^- = \mathbf{S}_k^-\,\mathbf{S}_k^{-\mathrm{T}}$), z_k (measurement vector), and $\mathbf{G}_k$ (Jacobian matrix containing the partial derivatives of the measurement vector z_k with respect to the state vector $\boldsymbol{x}_k^-$).

The sequence of computation is indicated below.

(a) $\boldsymbol{f}_k^- = \mathbf{S}_k^{-\mathrm{T}}\mathbf{G}_k^{-\mathrm{T}}$
(b) $\alpha_k = 1/\left(\sigma^2 + \boldsymbol{f}_k^{-\mathrm{T}}\boldsymbol{f}_k^-\right)$
(c) $\mathbf{K}_k = \alpha_k \mathbf{S}_k^- \boldsymbol{f}_k^-$
(d) $\boldsymbol{x}_k^+ = \boldsymbol{x}_k^- + \mathbf{K}_k\left(z_k - \mathbf{G}_k \boldsymbol{x}_k^-\right)$
(e) $\gamma_k = 1/\left[1 + \left(\alpha_k \sigma^2\right)^{\frac{1}{2}}\right]$
(f) $\mathbf{S}_k^+ = \mathbf{S}_k^- - \gamma_k \mathbf{K}_k \boldsymbol{f}_k^{-\mathrm{T}}$ ($\mathbf{S}_k^+$ being such that $\mathbf{P}_k^+ = \mathbf{S}_k^+ \mathbf{S}_k^{+\mathrm{T}}$)
(g) Predict, by integrating the differential equation of the reference orbit and $\boldsymbol{\Phi}' = \mathbf{A}\,\boldsymbol{\Phi}$ forward in time, the next estimates of the state vector $\boldsymbol{x}_{k+1}^-$ and the transition matrix $\boldsymbol{\Phi}_{k+1}^-$
(h) Update the square-root matrix $\mathbf{S}_k^+$ and the state vector $\boldsymbol{x}_k^+$ to $k + 1$, as follows
(i) $\mathbf{S}_{k+1}^- = \boldsymbol{\Phi}_{k+1}^- \mathbf{S}_k^+$
(j) $\boldsymbol{x}_{k+1}^- = \boldsymbol{\Phi}_{k+1}^- \boldsymbol{x}_k^+$
(k) Increase k to $k + 1$ and go to step (a).

The sequence given above is based on the presence of a single observation (performed at the given time t_k) in the measurement vector z_k.

Bellantoni and Dodge [7] have extended this method to handle more than a single observation in z_k. At the beginning of the sequence, the initial square-root matrix $\mathbf{S}_0^-$, associated with the initial state vector $\boldsymbol{x}_0^-$, results from the Cholesky decomposition of the initial covariance matrix $\mathbf{P}_0^-$, which in turn must be symmetric and positive definite.

Throughout the Apollo missions, Potter's square-root filter has been used for lunar navigation [48].

On the basis of the square-root method, Carlson [17], Bierman [8], and Thornton [69] have derived a method which recursively computes an upper triangular covariance square-root matrix.

This method is based on the upper triangular Cholesky decomposition: any $m \times m$ symmetric positive definite matrix $\mathbf{P}$ has a unique decomposition $\mathbf{P} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$, as will be shown below. For example and without loss of generality, if $\mathbf{P}$ were a 3×3 symmetric and positive definite matrix, then the upper triangular Cholesky decomposition of $\mathbf{P}$ would be

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{12} & p_{22} & p_{23} \\ p_{13} & p_{23} & p_{33} \end{bmatrix} = \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & 0 \\ 0 & 0 & d_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ u_{12} & 1 & 0 \\ u_{13} & u_{23} & 1 \end{bmatrix}$$

where $\mathbf{U}$ is a unit upper triangular matrix and $\mathbf{D}$ is a diagonal matrix with non-negative entries d_{ii} ($i = 1, 2, \ldots, m$).

By so doing, the square-root method shown above is modified as follows

$$\mathbf{P} = \left(\mathbf{U}\mathbf{D}^{\frac{1}{2}}\right)\left(\mathbf{U}\mathbf{D}^{\frac{1}{2}}\right)^{\mathrm{T}} = \mathbf{S}\mathbf{S}^{\mathrm{T}}$$

where

$$\mathbf{D}^{\frac{1}{2}} = \text{diag}\left[(d_{11})^{\frac{1}{2}}, (d_{22})^{\frac{1}{2}}, \ldots, (d_{mm})^{\frac{1}{2}}\right]$$

The upper triangular Cholesky decomposition is based on the algorithm shown below.

Following Bierman [8], for $j = m, m - 1, \ldots, 2$, we compute the entries d_{jj} of the diagonal matrix $\mathbf{D}$, except the first entry d_{11}, as follows

$$d_{jj} = p_{jj}$$

Then, for $j = m, m - 1, \ldots, 2$, we compute the diagonal entries u_{jj} of the upper triangular matrix $\mathbf{U}$ as follows

$$u_{jj} = 1$$

Then, for $j = m, m - 1, \ldots, 2$ and $k = 1, 2, ...,j - 1$, we compute the off-diagonal entries u_{kj} of $\mathbf{U}$ as follows

$$u_{kj} = \frac{p_{kj}}{d_{jj}}$$

Then, for $k = 1, 2, \ldots, j - 1$ and $i = 1, 2, \ldots, k$, the entries p_{ik} of $\mathbf{P}$ are destroyed and replaced by the following new values

$$p_{ik} = p_{ik} - u_{ij}u_{kj}d_{jj}$$

Finally, u_{11} is set equal to unity, and d_{11} is set equal to p_{11}.

Bierman [8] has also given a Fortran codification of the algorithm indicated above. Born [11] has given the following example of $\mathbf{UDU}^{\mathrm{T}}$ decomposition. Let

$$\mathbf{P} \equiv \begin{bmatrix} 1 & 2 & 3 \\ 2 & 8 & 2 \\ 3 & 2 & 14 \end{bmatrix}$$

be the matrix to be decomposed. In case of a 3×3 matrix, we have

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{12} & p_{22} & p_{23} \\ p_{13} & p_{23} & p_{33} \end{bmatrix} = \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & 0 \\ 0 & 0 & d_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ u_{12} & 1 & 0 \\ u_{13} & u_{23} & 1 \end{bmatrix}$$

which, in the present case, becomes

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 8 & 2 \\ 3 & 2 & 14 \end{bmatrix} = \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & 0 \\ 0 & 0 & d_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ u_{12} & 1 & 0 \\ u_{13} & u_{23} & 1 \end{bmatrix}$$

By multiplying the matrices and equating to the given matrix **P**, there results

$$\begin{aligned} u_{12} &= (p_{12} - p_{13}p_{23}/p_{33})/\left(p_{22} - p_{23}^2 p_{33}\right) = (2 - 3 \times 2/14)/\left(8 - 2^2 \times 14\right) = 11/54 \\ u_{13} &= p_{13}/p_{33} = 3/14 \\ u_{23} &= p_{23}/p_{33} = 2/14 = 1/7 \\ d_{11} &= p_{11} - (p_{12} - p_{13}p_{23}/p_{33})^2/\left(p_{22} - p_{23}^2/p_{33}\right) - p_{13}^2/p_{33} \\ &= 1 - (2 - 3 \times 2/14)^2/\left(8 - 2^2/14\right) - 3^2/14 = 7/189 \\ d_{22} &= p_{22} - p_{23}^2/p_{33} = 8 - 2^2/14 = 54/7 \\ d_{33} &= p_{33} = 14 \end{aligned}$$

The matrices **U**, **D**, and $\mathbf{U}^{\mathrm{T}}$ of the decomposition $\mathbf{P} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$ are given below.

$$\begin{bmatrix} 1 & 11/54 & 3/14 \\ 0 & 1 & 1/7 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 7/189 & 0 & 0 \\ 0 & 54/7 & 0 \\ 0 & 0 & 14 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 11/54 & 1 & 0 \\ 3/14 & 1/7 & 1 \end{bmatrix}$$

As is easy to verify, the product $\mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$ reconstructs the given matrix **P**.

A further example of $\mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$ decomposition has been given by Kang [46]. Let

$$\mathbf{P} \equiv \begin{bmatrix} 130 & 186 & 152 & 20 \\ 186 & 283 & 230 & 30 \\ 152 & 230 & 249 & 35 \\ 20 & 30 & 35 & 5 \end{bmatrix}$$

be the matrix to be decomposed. As is easy to verify, **P** is a square, symmetric, and positive definite matrix.

By applying the Bierman method shown above, the matrix **P** is decomposed as follows

$$\begin{bmatrix} 130 & 186 & 152 & 20 \\ 186 & 283 & 230 & 30 \\ 152 & 230 & 249 & 35 \\ 20 & 30 & 35 & 5 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & 5 & 6 \\ 0 & 0 & 1 & 7 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 5 & 1 & 0 \\ 4 & 6 & 7 & 1 \end{bmatrix}$$

Again, the product $\mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$ on the right-hand side of the preceding equality reconstructs the given matrix **P**.

Let us consider now the application of the concepts shown above to the Kalman filter. Following Kang [46], let

$$\mathbf{P}_k^- = \mathbf{U}_k^- \mathbf{D}_k^- \mathbf{U}_k^{-\mathrm{T}}$$

be the $\mathbf{UDU}^{\mathrm{T}}$ decomposition of the predicted covariance matrix $\mathbf{P}_k^-$.

Likewise, let

$$\mathbf{P}_k^+ = \mathbf{U}_k^+ \mathbf{D}_k^+ \mathbf{U}_k^{+\mathrm{T}}$$

be the $\mathbf{UDU}^{\mathrm{T}}$ decomposition of the updated covariance matrix $\mathbf{P}_k^+$.

As shown in Sect. 2.11, the updated expression of $\mathbf{P}_k$ is

$$\mathbf{P}_k^+ = (\mathbf{I} - \mathbf{K}_k \mathbf{G}_k)\mathbf{P}_k^-$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{K}_k$ is the Kalman gain matrix, and $\mathbf{G}_k$ is the Jacobian matrix containing the partial derivatives of the measurement vector $\boldsymbol{z}_k$ with respect to the state vector $\boldsymbol{x}_k^-$. The same expression can also be written

$$\mathbf{P}_k^+ = \mathbf{P}_k^- - \mathbf{K}_k \mathbf{G}_k \mathbf{P}_k^-$$

Now, remembering the following expression (see Sect. 2.11) of the Kalman gain

$$\mathbf{K}_k = \mathbf{P}_k^- \mathbf{G}_k^{\mathrm{T}} \left(\mathbf{W}_k^{-1} + \mathbf{G}_k \mathbf{P}_k^- \mathbf{G}_k^{\mathrm{T}} \right)^{-1}$$

(where $\mathbf{W}_k^{-1}$ is the inverse of the weighting matrix $\mathbf{W}_k$) and substituting the expression of the Kalman gain into $\mathbf{P}_k^+ = \mathbf{P}_k^- - \mathbf{K}_k \mathbf{G}_k \mathbf{P}_k^-$, there results

$$\mathbf{P}_k^+ = \mathbf{P}_k^- - \mathbf{P}_k^- \mathbf{G}_k^{\mathrm{T}} \left(\mathbf{W}_k^{-1} + \mathbf{G}_k \mathbf{P}_k^- \mathbf{G}_k^{\mathrm{T}} \right)^{-1} \mathbf{G}_k \mathbf{P}_k^-$$

If the expression in parentheses (namely $\mathbf{W}_k^{-1} + \mathbf{G}_k \mathbf{P}_k^- \mathbf{G}_k^{\mathrm{T}}$) were a scalar ($s$), as the sequel will show, then the equation written above could be written as follows

$$\mathbf{P}_k^+ = \mathbf{P}_k^- - \mathbf{P}_k^- \mathbf{G}_k^{\mathrm{T}} (1/s) \mathbf{G}_k \mathbf{P}_k^-$$

Now, remembering the $\mathbf{UDU}^{\mathrm{T}}$ decomposition of the predicted covariance matrix $\mathbf{P}_k^-$, the preceding equation becomes

$$\begin{aligned}
\mathbf{P}_k^+ &= \left(\mathbf{U}_k^- \mathbf{D}_k^- \mathbf{U}_k^{-\mathrm{T}} \right) - (1/s) \left(\mathbf{U}_k^- \mathbf{D}_k^- \mathbf{U}_k^{-\mathrm{T}} \right) \mathbf{G}_k^{\mathrm{T}} \mathbf{G}_k \left(\mathbf{U}_k^- \mathbf{D}_k^- \mathbf{U}_k^{-\mathrm{T}} \right) \\
&= \mathbf{U}_k^- \left[\mathbf{D}_k^- - (1/s) \left(\mathbf{D}_k^- \mathbf{U}_k^{-\mathrm{T}} \mathbf{G}_k^{\mathrm{T}} \right) \left(\mathbf{G}_k \mathbf{U}_k^- \mathbf{D}_k^- \right) \right] \mathbf{U}_k^{-\mathrm{T}} \\
&= \mathbf{U}_k^- \left[\mathbf{D}_k^- - (1/s) \left(\mathbf{D}_k^- \mathbf{U}_k^{-\mathrm{T}} \mathbf{G}_k^{\mathrm{T}} \right) \left(\mathbf{D}_k^- \mathbf{U}_k^{-\mathrm{T}} \mathbf{G}_k^{\mathrm{T}} \right)^{\mathrm{T}} \right] \mathbf{U}_k^{-\mathrm{T}}
\end{aligned}$$

Since $(\mathbf{D}_k^- \mathbf{U}_k^{-\mathrm{T}} \mathbf{G}_k^{\mathrm{T}})\,(\mathbf{D}_k^- \mathbf{U}_k^{-\mathrm{T}} \mathbf{G}_k^{\mathrm{T}})^{\mathrm{T}}$ is a symmetric and positive definite matrix, then the term within square brackets is also a symmetric and positive definite matrix. Therefore, the $\mathbf{UDU}^{\mathrm{T}}$ decomposition also applies to this term, as follows

$$\left[\mathbf{D}_k^- - (1/s)\left(\mathbf{D}_k^- \mathbf{U}_k^{-\mathrm{T}} \mathbf{G}_k^{\mathrm{T}}\right)\left(\mathbf{D}_k^- \mathbf{U}_k^{-\mathrm{T}} \mathbf{G}_k^{\mathrm{T}}\right)^{\mathrm{T}}\right] = \mathbf{U}_k^* \mathbf{D}_k^* \mathbf{U}_k^{*\mathrm{T}}$$

Consequently, the update expression of the covariance matrix $\mathbf{P}_k$ becomes

$$\begin{aligned}\mathbf{P}_k^+ &= \mathbf{U}_k^- \left(\mathbf{U}_k^* \mathbf{D}_k^* \mathbf{U}_k^{*\mathrm{T}}\right) \mathbf{U}_k^{-\mathrm{T}} \\ &= \left(\mathbf{U}_k^- \mathbf{U}_k^*\right) \mathbf{D}_k^* \left(\mathbf{U}_k^{*\mathrm{T}} \mathbf{U}_k^{-\mathrm{T}}\right) \\ &= \left(\mathbf{U}_k^- \mathbf{U}_k^*\right) \mathbf{D}_k^* \left(\mathbf{U}_k^{-\mathrm{T}} \mathbf{U}_k^{*\mathrm{T}}\right)^{\mathrm{T}}\end{aligned}$$

Setting $\mathbf{U}_k^+ = \mathbf{U}_k^- \mathbf{U}_{\mathrm{k}}^*$ and $\mathbf{D}_k^+ = \mathbf{D}_k^*$ makes it possible to write

$$\mathbf{P}_k^+ = \mathbf{U}_k^+ \mathbf{D}_k^+ \mathbf{U}_k^{+\mathrm{T}}$$

which shows that the updated covariance matrix $\mathbf{P}_k^+$ can be decomposed as indicated above. Now it will be shown that the Kalman gain matrix

$$\mathbf{K}_k = \mathbf{P}_k^- \mathbf{G}_k^{\mathrm{T}} \left(\mathbf{W}_k^{-1} + \mathbf{G}_k \mathbf{P}_k^- \mathbf{G}_k^{\mathrm{T}}\right)^{-1}$$

can be computed without matrix inversion. To this end, let

$$\mathbf{P}_k^- \equiv \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix}_k$$

be the predicted covariance matrix. Let

$$\mathbf{G}_k \equiv \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}_k$$

be the Jacobian matrix $\mathbf{G}_k$ containing the partial derivatives of the measurement vector $\boldsymbol{z}_k$ with respect to the state vector $\boldsymbol{x}_k^-$. Let

$$\mathbf{R}_k \equiv \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}_k$$

be the matrix $\mathbf{R}_k \equiv \mathbf{W}_k^{-1}$ (that is, the inverse of the weighting matrix $\mathbf{W}_k$). Then, the expression

$$\mathbf{W}_k^{-1} + \mathbf{G}_k \mathbf{P}_k^- \mathbf{G}_k^{\mathrm{T}}$$

(which appears in the equation of the Kalman gain) becomes

$$\begin{bmatrix} r_{11}+p_{11} & r_{12}+p_{13} \\ r_{21}+p_{31} & r_{22}+p_{33} \end{bmatrix}_k$$

As shown in Sect. 2.10, let

$$\mathbf{M} \equiv \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

be a given 2 × 2 matrix. Then, the inverse matrix $\mathbf{M}^{-1}$ is given by

$$\mathbf{M}^{-1} = \begin{bmatrix} d/\det(\mathbf{M}) & -b/\det(\mathbf{M}) \\ -c/\det(\mathbf{M}) & a/\det(\mathbf{M}) \end{bmatrix}$$

where $\det(\mathbf{M}) = ad - bc$ is the determinant of the given matrix $\mathbf{M}$.

Consequently, the inverse of the matrix $\mathbf{W}_k^{-1} + \mathbf{G}_k\,\mathbf{P}_k^-\,\mathbf{G}_k^{\mathrm{T}}$ is

$$\begin{bmatrix} (r_{22}+p_{33})/\Delta & -(r_{12}+p_{13})/\Delta \\ -(r_{21}+p_{31})/\Delta & (r_{11}+p_{11})/\Delta \end{bmatrix}_k$$

where

$$\Delta = \det\left(\mathbf{W}_k^{-1} + \mathbf{G}_k\mathbf{P}_k^-\mathbf{G}_k^{\mathrm{T}}\right) = (r_{11}+p_{11})(r_{22}+p_{33}) - (r_{21}+p_{31})(r_{12}+p_{13})$$

When the predicted covariance matrix $\mathbf{P}_k^-$ is a block diagonal and the inverse weighting matrix $\mathbf{R}_k \equiv \mathbf{W}_k^{-1}$ is also diagonal, that is, when the matrices $\mathbf{P}_k^-$ and $\mathbf{R}_k$ are as follows

$$\mathbf{P}_k^- = \begin{bmatrix} p_{11}p_{12} & & \\ p_{21}p_{22} & & \\ & p_{33}p_{34} & \\ & p_{43}p_{44} & \end{bmatrix}_k \qquad \mathbf{R}_k = \begin{bmatrix} r_{11} & 0 \\ 0 & r_{22} \end{bmatrix}_k$$

then the matrix $(\mathbf{W}_k^{-1} + \mathbf{G}_k\,\mathbf{P}_k^-\,\mathbf{G}_k^{\mathrm{T}})^{-1}$ is

$$\begin{bmatrix} 1/(r_{11}+p_{11}) & 0 \\ 0 & 1/(r_{22}+p_{33}) \end{bmatrix}_k$$

which can also be written as follows

$$\begin{bmatrix} 1/s_1 & 0 \\ 0 & 1/s_2 \end{bmatrix}_k$$

This means that the nonzero entries of the inverted matrix $(\mathbf{W}_k^{-1} + \mathbf{G}_k\,\mathbf{P}_k^-\,\mathbf{G}_k^{\mathrm{T}})^{-1}$ result from an operation of scalar division. The Kalman gain matrix

$$\mathbf{K}_k = \mathbf{P}_k^- \mathbf{G}_k^{\mathrm{T}} \left(\mathbf{W}_k^{-1} + \mathbf{G}_k \mathbf{P}_k^- \mathbf{G}_k^{\mathrm{T}}\right)^{-1}$$

results from the matrix product of $\mathbf{P}_k^-\,\mathbf{G}_k^{\mathrm{T}}$ by $(\mathbf{W}_k^{-1} + \mathbf{G}_k\,\mathbf{P}_k^-\,\mathbf{G}_k^{\mathrm{T}})^{-1}$, that is,

$$\begin{bmatrix} p_{11} & p_{13} \\ p_{21} & p_{23} \\ p_{31} & p_{33} \\ p_{41} & p_{43} \end{bmatrix}_k \begin{bmatrix} 1/s_1 & 0 \\ 0 & 1/s_2 \end{bmatrix}_k$$

This yields the following expression of $\mathbf{K}_k$

$$\mathbf{K}_k = \begin{bmatrix} p_{11}/s_1 & p_{13}/s_1 \\ p_{21}/s_1 & p_{23}/s_1 \\ p_{31}/s_1 & p_{33}/s_1 \\ p_{41}/s_1 & p_{43}/s_1 \end{bmatrix}_k$$

Therefore, the Kalman gain matrix $\mathbf{K}_k$ can be computed without matrix inversion. The example given above concerns a 4×4 predicted covariance matrix $\mathbf{P}_k^-$.

In the general case, the Kalman gain matrix $\mathbf{K}_k$ results from

$$\mathbf{K}_k = \begin{bmatrix} p_{11}/s_1 & p_{13}/s_1 & \cdots & p_{1n}/s_{n-1} \\ p_{21}/s_1 & p_{23}/s_1 & \cdots & p_{2n}/s_{n-1} \\ \vdots & \vdots & \cdots & \vdots \\ p_{m1}/s_1 & p_{m3}/s_1 & \cdots & p_{mn}/s_{n-1} \end{bmatrix}_k$$

where

$$s_j = r_{jj} + \mathbf{G}_j \mathbf{P}_k^- \mathbf{G}_k^{\mathrm{T}}$$
$$p_{ij} = r_{jj} + \mathbf{G}_{ij} \mathbf{P}_k^- \mathbf{G}_{ij}^{\mathrm{T}}$$

It remains to show that the predicted covariance matrix $\mathbf{P}_k^-$ is a block diagonal.

If the updated covariance matrix $\mathbf{P}_k^+$ is a block diagonal, then the predicted covariance matrix $\mathbf{P}_k^-$ remains block diagonal, because the time update equation

$$\mathbf{P}_k^- = \mathbf{\Phi}_k \mathbf{P}_{k-1}^+ \mathbf{\Phi}_k^{\mathrm{T}}$$

(concerning the covariance matrix $\mathbf{P}_k^-$ associated with the predicted a priori state vector $\boldsymbol{x}_k^-$ at time t_k) can be written as follows

$$\mathbf{U}_k^- \mathbf{D}_k^- \mathbf{U}_k^{-\mathrm{T}} = \mathbf{\Phi}_k (\mathbf{U}_{k-1}^+ \mathbf{D}_{k-1}^+ \mathbf{U}_{k-1}^{+\mathrm{T}}) \mathbf{\Phi}_k^{\mathrm{T}}$$

Now, since (see above)

$$\begin{aligned}\mathbf{P}_{k-1}^{+} &= \mathbf{U}_{k-1}^{-}\left(\mathbf{U}_{k-1}^{*}\mathbf{D}_{k-1}^{*}\mathbf{U}_{k-1}^{*\mathrm{T}}\right)\mathbf{U}_{k-1}^{-\mathrm{T}} \\ &= (\mathbf{U}_{k-1}^{-}\mathbf{U}_{k-1}^{*})\mathbf{D}_{k-1}^{*}(\mathbf{U}_{k-1}^{*\mathrm{T}}\mathbf{U}_{k-1}^{-\mathrm{T}}) \\ &= (\mathbf{U}_{k-1}^{-}\mathbf{U}_{k-1}^{*})\mathbf{D}_{k-1}^{*}(\mathbf{U}_{k-1}^{-\mathrm{T}}\mathbf{U}_{k-1}^{*\mathrm{T}})^{\mathrm{T}}\end{aligned}$$

then

$$\begin{aligned}\mathbf{U}_{k}^{-}\mathbf{D}_{k}^{-}\mathbf{U}_{k}^{-\mathrm{T}} &= \mathbf{\Phi}_{k}[(\mathbf{U}_{k-1}^{-}\mathbf{U}_{k-1}^{*})\mathbf{D}_{k-1}^{*}(\mathbf{U}_{k-1}^{-\mathrm{T}}\mathbf{U}_{k-1}^{*\mathrm{T}})^{\mathrm{T}}]\mathbf{\Phi}_{k}^{\mathrm{T}} \\ &= (\mathbf{\Phi}_{k}\mathbf{U}_{k-1}^{-}\mathbf{U}_{k-1}^{*})\mathbf{D}_{k-1}^{*}(\mathbf{\Phi}_{k}\mathbf{U}_{k-1}^{-}\mathbf{U}_{k-1}^{*})^{\mathrm{T}}\end{aligned}$$

Since $(\mathbf{\Phi}_k\ \mathbf{U}_{k-1}^{-}\ \mathbf{U}_{k-1}^{*})^{\mathrm{T}}$ is the transpose of $(\mathbf{\Phi}_k\ \mathbf{U}_{k-1}^{-}\ \mathbf{U}_{k-1}^{*})$, then the term

$$(\mathbf{\Phi}_{k}\mathbf{U}_{k-1}^{-}\mathbf{U}_{k-1}^{*})\mathbf{D}_{k-1}^{*}(\mathbf{\Phi}_{k}\mathbf{U}_{k-1}^{-}\mathbf{U}_{k-1}^{*})^{\mathrm{T}}$$

(that is, the predicted covariance matrix $\mathbf{P}_k^{-}$) is diagonal.

In summary, the preceding analysis has shown that:

- the Kalman gain matrix $\mathbf{K}_k$ can be computed without matrix inversion; and
- both the predicted ($\mathbf{P}_k^{-}$) and updated ($\mathbf{P}_k^{+}$) covariance matrices can be decomposed as $\mathbf{P} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$.

The sequence of computation for the square-root method without matrix inversion is indicated below.

- $\mathbf{P}_k^{-} = \mathbf{\Phi}_k\mathbf{P}_{k-1}^{+}\mathbf{\Phi}_k^{\mathrm{T}} = \mathbf{U}_k^{-}\mathbf{D}_k^{-}\mathbf{U}_k^{-\mathrm{T}}$
- $\mathbf{K}_k = \mathbf{P}_k^{-}\mathbf{G}_k^{\mathrm{T}}\left(\mathbf{R}_k + \mathbf{G}_k\mathbf{P}_k^{-}\mathbf{G}_k^{\mathrm{t}}\right)^{-1}$

where

$$s_j = r_{jj} + \mathbf{G}_j\mathbf{P}_k^{-}\mathbf{G}_k^{\mathrm{t}}$$

$$p_{ij} = r_{jj} + \mathbf{G}_{ij}\mathbf{P}_k^{-}\mathbf{G}_{ij}^{\mathrm{t}}$$

$$\mathbf{K}_k = \begin{bmatrix} p_{11}/s_1 & p_{13}/s_1 & \cdots & p_{1n}/s_{n-1} \\ p_{21}/s_1 & p_{23}/s_1 & \cdots & p_{2n}/s_{n-1} \\ \vdots & \vdots & \cdots & \vdots \\ p_{m1}/s_1 & p_{m3}/s_1 & \cdots & p_{mn}/s_{n-1} \end{bmatrix}_k$$

- $\left[\mathbf{D}_k^{-} - (1/s)\,(\mathbf{D}_k^{-}\mathbf{U}_k^{-\mathrm{T}}\mathbf{G}_k^{\mathrm{T}})\,(\mathbf{D}_k^{-}\mathbf{U}_k^{-\mathrm{T}}\mathbf{G}_k^{\mathrm{T}})^{\mathrm{T}}\right] = \mathbf{U}_k^{*}\mathbf{D}_k^{*}\mathbf{U}_k^{*\mathrm{T}}$
- $\mathbf{P}_k^{+} = (\mathbf{U}_k^{-}\mathbf{U}_k^{*})\mathbf{D}_k^{*}(\mathbf{U}_k^{-\mathrm{T}}\mathbf{U}_k^{*\mathrm{T}})^{\mathrm{T}} = \mathbf{U}_k^{+}\mathbf{D}_k^{+}\mathbf{U}_k^{+\mathrm{T}}$

The updated and predicted state vectors and the transition matrix are computed as has been shown for the Potter–Schmidt method.

In addition to the numerical methods shown above, we merely mention the so-called sigma-rho filter, which takes this name from the following expression

$$p_{ij}^{-} = \sigma_i^{-} \sigma_j^{-} \rho_{ij}^{-}$$

given by Grewal and Kain [30] to the entries p_{ij}^{-} of the predicted covariance matrix $\mathbf{P}_k^{-}$. In this expression, σ_i^{-} is the standard deviation of the i^{th} component, σ_j^{-} is the standard deviation of the j^{th} component, and ρ_{ij}^{-} is the correlation coefficient between the i^{th} and the j^{th} component of the predicted state vector $\boldsymbol{x}_k^{-}$. It is to be noted that ρ is used here for the sole purpose of maintaining the nomenclature used by Grewal and Kain, and has nothing to do with the residuals. The idea on which this type of Kalman filter is based is that of updating the standard deviation σ and the correlation coefficients ρ instead of the predicted covariance matrix. By so doing, Grewal and Kain express the entries p_{ij}^{+} of the updated covariance matrix $\mathbf{P}_k^{+}$ as follows

$$\sigma_i^{+} \sigma_j^{+} \rho_{ij}^{+} = \sigma_i^{-} \sigma_j^{-} \rho_{ij}^{-} - \sum_{s=1}^{n} \mathbf{G}_s \sigma_s^{-} \sigma_j^{-} \rho_{sj}^{-}$$

The particulars of this method can be found in Ref. [30].

2.13 The Unscented Kalman Filter

As shown in Sect. 2.11, the application of the Kalman filter to nonlinear systems is sometimes difficult, especially in those of such systems which are highly nonlinear. The extended Kalman filter, described in Sect. 2.11, linearises all nonlinear models, in order for the classical Kalman filter to be applied to such nonlinear systems. The application of extended Kalman filters to nonlinear systems has the following flaws:

1. high instability of the filter, when the assumption of linearity is violated locally; and
2. difficulty in deriving the Jacobian matrices in most practical cases.

To deal with these systems, there is another class of filters, called sigma-point Kalman filters, to which class the unscented Kalman filter belongs.

This type of filter, proposed in 1992 by Julier and Uhlmann [42], has since then been dealt with by several authors (see, e.g., Refs. [70–67]).

According to Julier and Uhlmann, a general nonlinear discrete time system is represented by the following equations:

$$\begin{aligned}\boldsymbol{x}_k &= \boldsymbol{f}(\boldsymbol{x}_{k-1}, \boldsymbol{u}_{k-1}, \boldsymbol{v}_{k-1}, k-1)\\ \boldsymbol{z}_k &= \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{u}_k, k) + \boldsymbol{n}_k\end{aligned}$$

where $\boldsymbol{x}_k$ is the m-dimensional state vector at time-step k, $\boldsymbol{u}_{k-1}$ is the known input vector, $\boldsymbol{v}_{k-1}$ is the q-dimensional state noise process vector due to disturbances and modelling errors, z_k is the observation vector, $\boldsymbol{n}_k$ is the measurement noise vector, and $\boldsymbol{f}$ and $\boldsymbol{h}$ are the nonlinear vector-valued functions (supposed known) representing the system dynamic model. It is assumed that the noise vectors $\boldsymbol{v}_k$ and $\boldsymbol{n}_k$ are zero mean and that, for all i and j, the following equalities hold

$$\begin{aligned}E\left(\boldsymbol{v}_i \boldsymbol{v}_j^T\right) &= \delta_{ij}\mathbf{Q}_i\\ E\left(\boldsymbol{n}_i \boldsymbol{n}_j^T\right) &= \delta_{ij}\mathbf{R}_i\\ E\left(\boldsymbol{v}_i \boldsymbol{n}_j^T\right) &= \mathbf{0}\end{aligned}$$

where $\mathbf{Q}_i$ is the covariance matrix of the process noise $\boldsymbol{v}_k$, and $\mathbf{R}_i$ is the covariance matrix of the measurement noise $\boldsymbol{n}_k$. In most practical cases, according to van der Merwe and Wan [84], the noise terms $\boldsymbol{v}_k$ and $\boldsymbol{n}_k$ are additive, and the two equations written above can be simplified as follows

$$\begin{aligned}\boldsymbol{x}_k &= \boldsymbol{f}(\boldsymbol{x}_{k-1}, \boldsymbol{u}_{k-1}) + \boldsymbol{v}_{k-1}\\ \boldsymbol{z}_k &= \boldsymbol{h}(\boldsymbol{x}_k) + \boldsymbol{n}_k\end{aligned}$$

As shown in Sect. 2.11, the extended Kalman filter consists of:

- a first-order (in terms of Taylor-series expansion) approximation of the nonlinear functions $\boldsymbol{f}$ and $\boldsymbol{h}$ at the current state estimate; and
- the application of the classical Kalman filter to this model approximated to the first order.

By contrast, the unscented Kalman filter is based on the unscented transform, which is a numerical procedure for approximating the posterior mean and covariance of a random vector obtained from a nonlinear transformation [67]. The unscented transform determines a set of sample points, called sigma-points, around the mean. Such points are chosen so that their mean and covariance should be equal to, respectively, the mean and covariance of the augmented (see below) state vector, and then propagated through the nonlinear functions $\boldsymbol{f}$ and $\boldsymbol{h}$, so as to recover the mean and covariance of the estimate.

When the noise terms $\boldsymbol{v}_k$ and $\boldsymbol{n}_k$ are additive, there is no need to augment the state vector $\boldsymbol{x}_k$ and its covariance matrix $\mathbf{P}_k$ [68]. By contrast, this need arises when the observed dynamic system is of the general type indicated above, as will be shown in detail in Sect. 2.16.

The initial $m \times 1$ state vector $\boldsymbol{x}_0$ has known mean $\boldsymbol{\mu}_0 = E(\boldsymbol{x}_0)$ and covariance matrix $\boldsymbol{P}_0 = E[(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)^{\mathrm{T}}]$.

For time-step $k = 1, 2, \ldots, N - 1$ (N being the size of the measurement process), a set of $2\ m + 1$ sigma-point vectors $\boldsymbol{\chi}_{k-1}$ is computed as follows

$$\begin{aligned}
\boldsymbol{\chi}_{k-1}^{0} &= \boldsymbol{x}_{k-1} && i = 0 \\
\boldsymbol{\chi}_{k-1}^{i} &= \boldsymbol{x}_{k-1} + \left\{ [(m+\lambda)\mathbf{P}_{k-1}]^{\frac{1}{2}} \right\}_i && i = 1, 2, \ldots, m \\
\boldsymbol{\chi}_{k-1}^{i+m} &= \boldsymbol{x}_{k-1} - \left\{ [(m+\lambda)\mathbf{P}_{k-1}]^{\frac{1}{2}} \right\}_i && i = 1, 2, \ldots, m
\end{aligned}$$

where $\{[(m + \lambda)\mathbf{P}_{k-1}]^{1/2}\}_i$ is the i^{th} column of the matrix $\mathbf{S}_{k-1}$, which in turn is the square root of the matrix $(m + \lambda)\mathbf{P}_{k-1}$. Therefore, there results by definition

$$(m+\lambda)\mathbf{P}_{k-1} = \mathbf{S}_{k-1}\mathbf{S}_{k-1}^{\mathrm{T}}$$

The square-root matrix $\mathbf{S}_{k-1}$ is to be computed by means of some stable method, for example, by means of the Cholesky decomposition, as shown in Sect. 2.12. The set of the sigma-point vectors $\boldsymbol{\chi}_{k-1}^{i}$ ($i = 0, 1, \ldots, 2\ m$) forms the $m \times (2\ m + 1)$ sigma-point matrix $\mathbf{X}_{k-1}^{i}$, whose columns are given below

$$\mathbf{X}_{k-1}^{i} = \left[\boldsymbol{x}_{k-1} \quad \boldsymbol{x}_{k-1} + [(m+\lambda)\mathbf{P}_{k-1}]^{\frac{1}{2}} \quad \boldsymbol{x}_{k-1} - [(m+\lambda)\mathbf{P}_{k-1}]^{\frac{1}{2}} \right]$$

This matrix is such that $\boldsymbol{x}_{k-1}$ is a column vector, whereas $\boldsymbol{x}_{k-1} + [(m + \lambda)\mathbf{P}_{k-1}]^{1/2}$ and $\boldsymbol{x}_{k-1} - [(m + \lambda)\mathbf{P}_{k-1}]^{1/2}$ are, each of them, a set of m column vectors.

The sigma-point vectors $\boldsymbol{\chi}_{k-1}$ are projected ahead from time-step $k - 1$ to time-step k, by means of the nonlinear function $\boldsymbol{f}$, as follows

$$\boldsymbol{\chi}_{k}^{i} = \boldsymbol{f}\left(\boldsymbol{\chi}_{k-1}^{i}, \boldsymbol{u}_{k-1}\right) \quad i = 0, 1, \ldots, 2m$$

The projected sigma-point vectors $\boldsymbol{\chi}_{k}^{i}$ (with weights w_s^i for the state vector and w_c^i for the covariance matrix) are used to compute the predicted state vector $\boldsymbol{x}_k^-$ and the predicted covariance matrix $\mathbf{P}_k^-$ of $\boldsymbol{x}_k^-$, as follows

$$\begin{aligned}
\boldsymbol{x}_k^- &= \sum_{i=0}^{2m} w_s^i \boldsymbol{\chi}_k^i \\
\mathbf{P}_k^- &= \sum_{i=0}^{2m} w_c^i \left[\boldsymbol{\chi}_k^i - \boldsymbol{x}_k^-\right]\left[\boldsymbol{\chi}_k^i - \boldsymbol{x}_k^-\right]^{\mathrm{T}} + \mathbf{Q}_{k-1}
\end{aligned}$$

where $\mathbf{Q}_{k-1}$ is the covariance matrix of the process noise vector $\boldsymbol{v}_{k-1}$.

Again, a set of $2m + 1$ sigma-point vectors $\boldsymbol{\chi}_k^i$ ($i = 0, 1, \ldots, 2m$) is computed from the predicted state vector $\boldsymbol{x}_k^-$ and the predicted covariance matrix $\mathbf{P}_k^-$, as follows

$$\begin{aligned}
\boldsymbol{\chi}_k^0 &= \boldsymbol{x}_k^- && i = 0 \\
\boldsymbol{\chi}_k^i &= \boldsymbol{x}_k^- + \left\{ \left[(m+\lambda)\mathbf{P}_k^- \right]^{\frac{1}{2}} \right\}_i && i = 1, 2, \ldots, m \\
\boldsymbol{\chi}_k^{i+m} &= \boldsymbol{x}_k^- - \left\{ \left[(m+\lambda)\mathbf{P}_k^- \right]^{\frac{1}{2}} \right\}_i && i = 1, 2, \ldots, m
\end{aligned}$$

Now, the set of $2m + 1$ sigma-point vectors $\boldsymbol{\chi}_k^i$ is used as the argument of the nonlinear function $\boldsymbol{h}$, as follows

$$\boldsymbol{\gamma}_k^i = \boldsymbol{h}\left(\boldsymbol{\chi}_k^i\right) \quad i = 0, 1, \ldots, 2m$$

and the sigma-point vectors $\boldsymbol{\gamma}_k^i$ are used (with weights w_s^i for the state vector and w_c^i for the covariance matrix) to compute the predicted measurement vector $\boldsymbol{z}_k^-$ and the predicted covariance matrix $\mathbf{P}_{zz}^-$ of $\boldsymbol{z}_k^-$, as follows

$$\begin{aligned}
\boldsymbol{z}_k^- &= \sum_{i=0}^{2m} w_s^i \, \boldsymbol{\gamma}_k^i \\
\mathbf{P}_{zz}^- &= \sum_{i=0}^{2m} w_c^i \left[\boldsymbol{\gamma}_k^i - \boldsymbol{z}_k^- \right] \left[\boldsymbol{\gamma}_k^i - \boldsymbol{z}_k^- \right]^{\mathrm{T}} + \mathbf{R}_k
\end{aligned}$$

where $\mathbf{R}_k$ is the covariance matrix of the measurement noise vector $\boldsymbol{n}_k$.

The cross-correlation matrix $\mathbf{P}_{xz}^-$ (related to the time-step k) between $\boldsymbol{x}_k^-$ and $\boldsymbol{z}_k^-$ is computed as follows

$$\mathbf{P}_{xz}^- = \sum_{i=0}^{2m} w_c^i \left[\boldsymbol{\chi}_k^i - \boldsymbol{x}_k^- \right] \left[\boldsymbol{\gamma}_k^i - \boldsymbol{z}_k^- \right]^{\mathrm{T}} + \mathbf{R}_k$$

Finally, the Kalman gain $\mathbf{K}_k$, the updated state vector $\boldsymbol{x}_k^+$, and the updated covariance matrix $\mathbf{P}_k^+$ are computed as follows

$$\begin{aligned}
\mathbf{K}_k &= (\mathbf{P}_{xz}^-)(\mathbf{P}_{xz}^-)^{-1} \\
\boldsymbol{x}_k^+ &= \boldsymbol{x}_k^- + \mathbf{K}_k(\boldsymbol{z}_k - \boldsymbol{z}_k^-) \\
\mathbf{P}_k^+ &= \mathbf{P}_k^- - \mathbf{K}_k \mathbf{P}_{zz}^- \mathbf{K}_k^{\mathrm{T}}
\end{aligned}$$

According to Julier and Uhlmann [42], the unscented Kalman filter, due to its better properties of estimation accuracy and ease of implementation in comparison with the extended Kalman filter, is better suited than the latter in filtering applications. In quantitative terms, according to Wan and van der Merwe (Refs. [84, 81]), the posterior mean and covariance computed by means of the unscented

Kalman filter are accurate to the third order, in terms of a Taylor-series expansion, for all nonlinearities, in case of Gaussian inputs. In case of non-Gaussian inputs, the approximations are accurate to the second order. In contrast, the extended Kalman filter provides results approximated to the first order.

2.14 The Square-Root Unscented Kalman Filter

The filter described in the present paragraph has been proposed by van der Merwe and Wan [84]. It is an improvement of the standard unscented Kalman filter, which has been described in Sect. 2.13. The difference is due to the fact that the square-root unscented Kalman filter propagates (that is, projects ahead in time) the square root $\mathbf{S}_k$ of the covariance matrix $\mathbf{P}_k = \mathbf{S}_k \mathbf{S}_k^{\mathrm{T}}$, instead of the covariance matrix itself. The reasons for doing this are the same as those discussed in Sect. 2.12, namely numerical stability and positive definiteness, at each time-step k, of the covariance matrix $\mathbf{P}_k$ of the state vector $\boldsymbol{x}_k$.

Assuming again a system represented by the following equations

$$\begin{aligned} \boldsymbol{x}_k &= \boldsymbol{f}(\boldsymbol{x}_{k-1}, \boldsymbol{u}_{k-1}) + \boldsymbol{v}_{k-1} \\ \boldsymbol{z}_k &= \boldsymbol{h}(\boldsymbol{x}_k) + \boldsymbol{n}_k \end{aligned}$$

we intend to modify the expressions given in Sect. 2.13, so as to take account of the propagation of the Cholesky factor $\mathbf{S}_k$. The square-root unscented Kalman filter uses three techniques of linear algebra, namely the QR decomposition, the least-squares problem, and the Cholesky factor updating. The first two techniques have been shown in Sect. 2.10. The third will be shown in the last section of the present paragraph.

Following Terejanu [68], the initial m-dimensional state vector $\boldsymbol{x}_0$ has known mean $\boldsymbol{\mu}_0 = E(\boldsymbol{x}_0)$ and covariance matrix $\boldsymbol{P}_0 = E[(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)^{\mathrm{T}}]$, whose Cholesky factor $\mathbf{S}_0$ is found as follows

$$\mathbf{S}_0 = \mathrm{chol}\left\{ E\left[(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)^{\mathrm{T}} \right] \right\}$$

In the preceding expression, the Cholesky factorisation algorithm has been indicated by means of the function chol(), whose argument is the covariance matrix $\mathbf{P}_0$. The users of MATLAB have just this function at their disposal. Otherwise, those who have a FORTRAN compiler can use the subroutine schdc of LINPACK or LAPACK [21], or the subroutine msfd of IBM SSP [39]. Both of the subroutines named above take a symmetric positive definite matrix $\mathbf{A}$ and compute an upper triangular matrix $\mathbf{R}$ (called the Cholesky factor of $\mathbf{A}$) such that $\mathbf{A} = \mathbf{R}^{\mathrm{T}}\mathbf{R}$. For time-step k within the interval 1, 2, …, k_{end}, the sigma-point vectors $\boldsymbol{\chi}_{k-1}$ are computed as follows

$$\begin{aligned}
\boldsymbol{\chi}_{k-1}^{0} &= \boldsymbol{x}_{k-1} && i = 0 \\
\boldsymbol{\chi}_{k-1}^{i} &= \boldsymbol{x}_{k-1} + \left[(m+\lambda)^{\frac{1}{2}}\mathbf{S}_{k-1}\right]_i && i = 1, 2, \ldots, m \\
\boldsymbol{\chi}_{k-1}^{i+m} &= \boldsymbol{x}_{k-1} - \left[(m+\lambda)^{\frac{1}{2}}\mathbf{S}_{k-1}\right]_i && i = 1, 2, \ldots, m
\end{aligned}$$

These sigma-point vectors are projected ahead from time-step $k-1$ to time-step k, by means of the nonlinear function $\boldsymbol{f}$, as follows

$$\boldsymbol{\chi}_{k}^{i} = \boldsymbol{f}\left(\boldsymbol{\chi}_{k-1}^{i}, \boldsymbol{u}_{k-1}\right) \qquad i = 0, 1, \ldots, 2m$$

The projected sigma-point vectors $\boldsymbol{\chi}_k^i$ (with weights w_s^i for the state vector and w_c^i for the covariance matrix) are used to compute the predicted state vector $\boldsymbol{x}_k^-$ and the predicted Cholesky factor matrix $\mathbf{S}_k^-$, as follows

$$\begin{aligned}
\boldsymbol{x}_k^- &= \sum_{i=0}^{2m} w_s^i \boldsymbol{\chi}_k^i \\
S_k^- &= \mathrm{qr}\left\{\left[\left(w_c^i\right)^{\frac{1}{2}}\left(\boldsymbol{\chi}_k^i - x_k^i\right) \quad \left(\mathbf{Q}_k\right)^{\frac{1}{2}}\right]\right\} && i = 1, 2, \ldots, 2m \\
S_k^- &= \mathrm{cholupdate}\left[\mathbf{S}_k^-, \left(\boldsymbol{\chi}_k^0 - x_k^-\right), \mathrm{sgn}\left\{w_c^0\right\}\left(\left|w_c^0\right|\right)^{\frac{1}{2}}\right]
\end{aligned}$$

where $\mathbf{Q}_{k-1}$ is the covariance matrix of the process noise vector $\mathbf{v}_{k-1}$. The positive or negative sign of w_c^0 (that is, $w_c^0 > 0$ or $w_c^0 < 0$) determines whether the function update performs a positive or a negative rank-one update to the Cholesky factorisation, as will be shown below. In order to understand the meaning of these expressions, it is necessary to remember the expression of the predicted covariance matrix $\mathbf{P}_k^- = (\mathbf{S}_k^-)(\mathbf{S}_k^-)^{\mathrm{T}}$, given in the preceding paragraph, for the standard unscented Kalman filter, that is,

$$\mathbf{P}_k^- = \sum_{i=0}^{2m} w_c^i \left[\boldsymbol{\chi}_k^i - \boldsymbol{x}_k^-\right]\left[\boldsymbol{\chi}_k^i - \boldsymbol{x}_k^-\right]^{\mathrm{T}} + \mathbf{Q}_{k-1}$$

By extracting the first term (corresponding to $i = 0$) from the sum, this expression can be written, for $i = 1, 2, \ldots, 2m$, as follows

$$\begin{aligned}
\mathbf{P}_k^- &= \sum_{i=1}^{2m}\left\{\left(w_c^i\right)^{\frac{1}{2}}\left[\boldsymbol{\chi}_k^i - \boldsymbol{x}_k^-\right]\left(w_c^i\right)^{\frac{1}{2}}\left[\boldsymbol{\chi}_k^i - \boldsymbol{x}_k^-\right]^{\mathrm{T}} + \left[\mathbf{Q}^{\frac{1}{2}}\right]_{k-1}\left[\mathbf{Q}^{\frac{1}{2}}\right]_{k-1}^{\mathrm{T}}\right\} \\
&\quad + w_c^0\left[\boldsymbol{\chi}_k^0 - \boldsymbol{x}_k^-\right]\left[\boldsymbol{\chi}_k^0 - \boldsymbol{x}_k^-\right]^{\mathrm{T}} \\
&= \left[\left(w_c^i\right)^{\frac{1}{2}}\left(\boldsymbol{\chi}_k^i - \boldsymbol{x}_k^-\right) \quad \left(\mathbf{Q}^{\frac{1}{2}}\right)_{k-1}\right]\begin{bmatrix}\left(w_c^i\right)^{\frac{1}{2}}\left(\boldsymbol{\chi}_k^i - \boldsymbol{x}_k^-\right)^{\mathrm{T}} \\ \left(\mathbf{Q}^{\frac{1}{2}}\right)_{k-1}^{\mathrm{T}}\end{bmatrix} + w_c^0\left(\boldsymbol{\chi}_k^0 - \boldsymbol{x}_k^-\right)\left(\boldsymbol{\chi}_k^0 - \boldsymbol{x}_k^-\right)^{\mathrm{T}}
\end{aligned}$$

where $\mathbf{Q}_{k-1}^{\frac{1}{2}}$ is the square-root matrix of the process noise covariance matrix $\mathbf{Q}_{k-1}$.

The matrix $\left[\left(w_c^i\right)^{1/2}\left(\boldsymbol{\chi}_k^i - \boldsymbol{x}_k^-\right) \quad \left(\mathbf{Q}^{1/2}\right)_{k-1} \right]$ has m rows and $3m$ columns. As shown in Sect. 2.10, its 3 $m \times m$ transpose matrix $[(w_c{}^i)^{1/2}(\boldsymbol{\chi}_k^i - \boldsymbol{x}_k^-)(\mathbf{Q}^{1/2})_{\mathbf{k-1}}]^{\mathrm{T}}$ can be decomposed, by using the QR factorisation, into the product of an orthogonal $3m \times m$ matrix $\mathbf{O}_k$ and an upper triangular $m \times m$ matrix $(\mathbf{S}_k^-)^{\mathrm{T}}$, as follows

$$\left[\left(w_c^i\right)^{1/2}\left(\boldsymbol{\chi}_k^i - \boldsymbol{x}_k^-\right) \quad \left(\mathbf{Q}^{1/2}\right)_{k-1} \right]^{\mathrm{T}} = \mathbf{O}_k\left(\mathbf{S}_k^-\right)^{\mathrm{T}} \qquad (i = 1, 2, \ldots, 2m)$$

The MATLAB function which can be used for this purpose is [**Q**, **R**] = qr(**A**), where the matrices **Q**, **R**, and **A** are such that **A** = **Q** **R**.

Otherwise, with a FORTRAN compiler, it is possible to use the subroutine sqrdc of LINPACK or LAPACK [21].

Therefore, the predicted covariance matrix $\mathbf{P}_k^-$ can be expressed as follows

$$\begin{aligned}\mathbf{P}_k^- &= \left(\mathbf{S}_k^-\right)(\mathbf{O}_k)^{\mathrm{T}}(\mathbf{O}_k)\left(\mathbf{S}_k^-\right)^{\mathrm{T}} + w_c^0(\boldsymbol{\chi}_k^0 - \boldsymbol{x}_k^-)(\boldsymbol{\chi}_k^0 - \boldsymbol{x}_k^-)^{\mathrm{T}} \\ &= \left(\mathbf{S}_k^-\right)\left(\mathbf{S}_k^-\right)^{\mathrm{T}} + w_c^0(\boldsymbol{\chi}_k^0 - \boldsymbol{x}_k^-)(\boldsymbol{\chi}_k^0 - \boldsymbol{x}_k^-)^{\mathrm{T}}\end{aligned}$$

In order to include the effect of the term $w_c^0(\boldsymbol{\chi}_k^0 - \boldsymbol{x}_k^-)(\boldsymbol{\chi}_k^0 - \boldsymbol{x}_k^-)^{\mathrm{T}}$ in the square-root matrix, it is necessary to perform either a rank-one positive update (if $w_c^0 > 0$) or a rank-one negative update (if $w_c^0 < 0$) to the Cholesky factorisation, as follows

$$\mathbf{S}_k^- = \mathrm{cholupdate}\left[\mathbf{S}_k^-, \left(\boldsymbol{\chi}_k^0 - \boldsymbol{x}_k^-\right), \mathrm{sgn}\left\{w_c^0\right\}\left(\left|w_c^0\right|\right)^{\frac{1}{2}}\right]$$

where sgn(x) is the signum function, defined in Sect. 2.10 (that is, sgn(x) = −1 if $x < 0$; sgn(x) = 0 if $x = 0$; sgn(x) = 1 if $x > 0$), and cholupdate(**R**, $\boldsymbol{x}$, " + ") or cholupdate(**R**, $\boldsymbol{x}$, "−") is the MATLAB function which performs the positive rank-one update ("+") or the negative rank-one update ("**–**") to the Cholesky factorisation.

Let **R** = chol(**A**) be the Cholesky factor of a given matrix **A**, as has been shown above. Then, cholupdate(**R**, $\boldsymbol{x}$, "+") returns the upper triangular Cholesky factor of $\mathbf{A} + \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$, where $\boldsymbol{x}$ is a column vector; likewise, cholupdate(**R**, $\boldsymbol{x}$, "−") returns the upper triangular Cholesky factor of $\mathbf{A} - \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$.

In the present case, the function cholupdate[$\mathbf{S}_k^-$, $(\boldsymbol{\chi}_k^0 - \boldsymbol{x}_k^-)$, sgn$(w_c^0)(|w_c^0|)^{½}$] returns the Cholesky factor of $(\mathbf{S}_k^-)(\mathbf{S}_k^-)^{\mathrm{T}} + w_c^0(\boldsymbol{\chi}_k^0 - \boldsymbol{x}_k^-)(\boldsymbol{\chi}_k^0 - \boldsymbol{x}_k^-)^{\mathrm{T}}$.

Otherwise, with a FORTRAN compiler, it is possible to use the subroutine schud (positive update) or schdd (negative update) of LINPACK or LAPACK [21].

By so doing, the predicted covariance matrix can be written $\mathbf{P}_k^- = (\mathbf{S}_k^-)(\mathbf{S}_k^-)^{\mathrm{T}}$.

The same line of reasoning can be followed to compute the predicted covariance matrix $\mathbf{P}_{zz}^- = (\mathbf{S}_z^-)(\mathbf{S}_z^-)^{\mathrm{T}}$ of z_k^- and the updated covariance matrix $\mathbf{P}_k^+ = (\mathbf{S}_k^+)(\mathbf{S}_k^+)^{\mathrm{T}}$ of $\boldsymbol{x}_k^+$.

A summary of the equations used in the square-root unscented Kalman filter is given below. Starting from the initial m-dimensional state vector $\boldsymbol{x}_0$, whose mean $\boldsymbol{\mu}_0 = E(\boldsymbol{x}_0)$ and covariance matrix $\boldsymbol{P}_0 = E[(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)^{\mathrm{T}}]$ are known, the Cholesky factor $\mathbf{S}_0$ of $\mathbf{P}_0$ is found as follows

$$\mathbf{S}_0 = \mathrm{chol}\left\{E[(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)^{\mathrm{T}}]\right\}$$

For time-step k within the interval 1, 2, …, k_{end}, the sigma-point vectors $\boldsymbol{\chi}_{k-1}$ are computed as follows

$$\begin{aligned}
\boldsymbol{\chi}_{k-1}^{0} &= \boldsymbol{x}_{k-1} && i = 0 \\
\boldsymbol{\chi}_{k-1}^{i} &= \boldsymbol{x}_{k-1} + \left[(m+\lambda)^{\frac{1}{2}}\mathbf{S}_{k-1}\right]_i && i = 1, 2, \ldots, m \\
\boldsymbol{\chi}_{k-1}^{i+m} &= \boldsymbol{x}_{k-1} - \left[(m+\lambda)^{\frac{1}{2}}\mathbf{S}_{k-1}\right]_i && i = 1, 2, \ldots, m
\end{aligned}$$

These sigma-point vectors are projected ahead from time-step $k-1$ to time-step k, by means of the nonlinear function $\boldsymbol{f}$, as follows

$$\boldsymbol{\chi}_k^i = \boldsymbol{f}\left(\boldsymbol{\chi}_{k-1}^i, u_{k-1}\right) \qquad i = 0, 1, \ldots, 2m$$

The projected sigma-point vectors $\boldsymbol{\chi}_k^i$ (with weights w_s^i for the state vector and w_c^i for the covariance matrix) are used to compute the predicted state vector $\boldsymbol{x}_k^-$ and the predicted Cholesky factor matrix $\mathbf{S}_k^-$, as follows

$$\begin{aligned}
\boldsymbol{x}_k^- &= \sum_{i=0}^{2m} w_s^i \, \boldsymbol{\chi}_k^i \\
\mathbf{S}_k^- &= \mathrm{qr}\left\{\left[\begin{array}{cc} \left(w_c^i\right)^{\frac{1}{2}}\left(\boldsymbol{\chi}_k^i - \boldsymbol{x}_k^i\right) & \left(\mathbf{Q}_k\right)^{\frac{1}{2}} \end{array}\right]\right\} \quad i = 1, 2, \ldots, 2m \\
\mathbf{S}_k^- &= \mathrm{cholupdate}\left[\mathbf{S}_k^-, \left(\boldsymbol{\chi}_k^0 - \boldsymbol{x}_k^-\right), \mathrm{sgn}\left\{w_c^0\right\}\left(\left|w_c^0\right|\right)^{\frac{1}{2}}\right]
\end{aligned}$$

where $\mathbf{Q}_{k-1}$ is the covariance matrix of the process noise vector $\boldsymbol{v}_{k-1}$. Again, a set of $2m + 1$ sigma-point vectors $\boldsymbol{\chi}_k^i$ ($i = 0, 1, \ldots, 2m$) is computed from the predicted state vector $\boldsymbol{x}_k^-$ and the predicted Cholesky factor matrix $\mathbf{S}_k^-$, as follows

$$\begin{aligned}
\boldsymbol{\chi}_k^0 &= \boldsymbol{x}_k^- && i = 0 \\
\boldsymbol{\chi}_k^i &= \boldsymbol{x}_k^- + \left[(m+\lambda)^{\frac{1}{2}}\mathbf{S}_k^-\right]_i && i = 1, 2, \ldots, m \\
\boldsymbol{\chi}_k^{i+m} &= \boldsymbol{x}_k^- - \left[(m+\lambda)^{\frac{1}{2}}\mathbf{S}_k^-\right]_i && i = 1, 2, \ldots, m
\end{aligned}$$

Now, the set of $2m + 1$ sigma-point vectors $\boldsymbol{\chi}_k^i$ is used as the argument of the nonlinear function $\boldsymbol{h}$, as follows

$$\gamma_k^i = \boldsymbol{h}\left(\boldsymbol{\chi}_k^i\right) \qquad i = 0, 1, \ldots, 2m$$

and the sigma-point vectors γ_k^i are used (with weights w_s^i for the state vector and w_c^i for the covariance matrix) to compute the predicted measurement vector z_k^- and the predicted Cholesky factor matrix $\mathbf{S}_{zz}^-$ of z_k^-, as follows

$$
\begin{aligned}
z_k^- &= \sum_{i=0}^{2m} w_s^i \gamma_k^i \\
\mathbf{S}_{zz}^- &= \mathrm{qr}\left\{ \left[\left(w_c^i\right)^{\frac{1}{2}} \left(\gamma_k^i - \gamma_k^i\right) \quad \left(\mathbf{R}_k\right)^{\frac{1}{2}} \right] \right\} \qquad i = 1, 2, \ldots, 2m \\
\mathbf{S}_{zz}^- &= \mathrm{cholupdate}\left[\mathbf{S}_{zz}^-, \left(\gamma_k^0 - z_k^-\right), \mathrm{sgn}\left\{w_c^0\right\} \left(\left|w_c^0\right|\right)^{\frac{1}{2}} \right]
\end{aligned}
$$

where $\mathbf{R}_k$ is the covariance matrix of the measurement noise vector $\boldsymbol{n}_k$.

The cross-correlation matrix $\mathbf{P}_{\mathrm{xz}}^-$ (related to the time-step k) between $\boldsymbol{x}_k^-$ and z_k^- is computed as follows

$$
\mathbf{P}_{xz}^- = \sum_{i=0}^{2m} w_c^i \left[\boldsymbol{\chi}_k^i - \boldsymbol{x}_k^-\right] \left[\gamma_k^i - z_k^-\right]^{\mathrm{T}}
$$

Finally, the Kalman gain $\mathbf{K}_k$, the updated state vector $\boldsymbol{x}_k^+$, and the updated Cholesky factor matrix $\mathbf{S}_k^+$ are computed as follows

$$
\begin{aligned}
\mathbf{K}_k &= \left[\mathbf{P}_{xz}^- / \left(\mathbf{S}_{zz}^-\right)^{\mathrm{T}}\right] / \mathbf{S}_{zz}^- \\
\boldsymbol{x}_k^+ &= \boldsymbol{x}_k^- + \mathbf{K}_k \left(z_k - z_k^-\right) \\
\mathbf{S}_k^+ &= \mathrm{cholupdate}\left(\mathbf{S}_k^-, \mathbf{K}_k \mathbf{S}_{zz}^-, -1\right)
\end{aligned}
$$

where / indicates an operation of back-substitution, which is a better alternative to the operation of matrix inversion used in the standard unscented Kalman filter. Since the Cholesky factor $\mathbf{S}_{zz}^-$ is a lower triangular matrix, then the Kalman gain $\mathbf{K}_k$ can be computed by means of two operations of back-substitution in the following expression

$$
\mathbf{K}_k \left[\mathbf{S}_{zz}^- \left(\mathbf{S}_{zz}^-\right)^{\mathrm{T}}\right] = \mathbf{P}_{xz}^-
$$

Since the quantity $\mathbf{U} = \mathbf{K}_k \mathbf{S}_{zz}^-$, which is the middle argument of the function $\mathbf{S}_k^+ = \mathrm{cholupdate}(\mathbf{S}_k^-, \mathbf{K}_k \mathbf{S}_{zz}^-, -1)$, is an $m \times m$ matrix, then the Cholesky factor $\mathbf{S}_k^+$ is updated consecutively m times, using the m columns of the matrix $\mathbf{U}$.

As has been shown at the beginning of this paragraph, the advantages of the square-root unscented Kalman filter over the standard unscented Kalman filter are a better control of round-off errors and the assurance of positive definiteness of the covariance matrices $\mathbf{P}_k$ associated with the successive state vectors $\boldsymbol{x}_k$. In addition, as Terejanu [68] points out, the square-root unscented Kalman filter does not require matrix inversions. As a result of computational experiments performed by van der Merwe and Wan [84], the square-root unscented Kalman filter is about 20% faster than its standard counterpart.

Finally, for a better understanding of the matter shown above, the fundamental concepts on the rank-one positive or negative updates to the Cholesky factorisation are given below.

Let $\mathbf{A} \equiv \{a_{ij}\}$ be an $n \times n$ symmetric positive definite matrix. As shown in Sect. 2.9, the symmetry of $\mathbf{A}$ means that $a_{ij} = a_{ji}$, and its positive definiteness means that $\boldsymbol{v}^{\mathrm{T}}\mathbf{A}\boldsymbol{v} > 0$ for all vectors $\boldsymbol{v} \neq \mathbf{0}$. Then, there exists a unique decomposition $\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ where $\mathbf{L} \equiv \{\ell_{ij}\}$ is a lower triangular matrix with positive elements ℓ_{11}, ℓ_{22}, …, ℓ_{nn} along its main diagonal. Let

$$\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$$

be a system of linear algebraic equations in matrix form, where $\mathbf{A}$ is the matrix indicated above, and $\boldsymbol{x}$ and $\boldsymbol{b}$ are n-dimensional column vectors.

Following Gill et al. [28], when $\boldsymbol{x}$ has been computed from $\mathbf{A}$ and $\boldsymbol{b}$ (see again Sect. 2.9), it is often necessary to solve a modified system

$$\mathbf{A}^{*}\boldsymbol{x}^{*} = \boldsymbol{b}^{*}$$

We could of course form the new matrix $\mathbf{A}^*$ and compute the Cholesky decomposition $\mathbf{A}^* = \mathbf{L}^*\mathbf{L}^{*\mathrm{T}}$. However, much less labour is required when $\mathbf{L}^*$ is computed directly from $\mathbf{L}$ than is required when $\mathbf{L}^*$ is computed from $\mathbf{A}^*$. To this end, it is necessary to modify the decomposition of $\mathbf{A}$ in order to obtain the decomposition of $\mathbf{A}^*$, from which $\boldsymbol{x}^*$ can be computed. The modification of $\mathbf{A}$ considered here has the following form

$$\mathbf{A}^{*} = \mathbf{A} + \alpha \boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}$$

where α is a scalar and $\boldsymbol{z}$ is an n-dimensional vector. Since the quantity $\alpha\, \boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}$ is a matrix of rank one, then the problem indicated above is called the update of the Cholesky decomposition of a symmetric positive definite matrix ($\mathbf{A}$) following a rank-one modification. In particular, for $\alpha = \pm 1$, we consider the positive rank-one update $\mathbf{A}^* = \mathbf{A} + \boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}$ and the negative rank-one update $\mathbf{A}^* = \mathbf{A} - \boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}$.

In case of positive ($\alpha > 0$) rank-one update, $\mathbf{A}^*$ is positive definite, and therefore has a Cholesky decomposition $\mathbf{A}^* = \mathbf{L}^*\mathbf{L}^{*\mathrm{T}}$.

In case of negative ($\alpha < 0$) rank-one update, $\mathbf{A}^*$ may be not positive definite, in which case it does not possess a Cholesky decomposition. In addition, even when $\mathbf{A}^*$ is positive definite, $\mathbf{L}^*$ may be inaccurately computed, if the modification comes near to reducing the rank of $\mathbf{A}$ [21].

Gill et al. [28] give several methods for modifying matrix factorisations. In particular, some of these methods can be applied to the Cholesky factorisation. One of them (called algorithm C1 by the authors cited above) is shown below.

Let us suppose that a given $n \times n$ symmetric positive definite matrix $\mathbf{A}$ (having a Cholesky factorisation $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^{\mathrm{T}}$) has been modified by a symmetric matrix of rank one, as follows

$$\mathbf{A}^{*} = \mathbf{A} + \alpha \boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}$$

where α is a scalar and $\boldsymbol{z}$ is an n-dimensional vector.

Starting from the matrices $\mathbf{L} \equiv \{\ell_{ij}\}$ and $\mathbf{D} \equiv \{d_{jj}\}$, which are supposed to be known, we want to compute two matrices $\mathbf{L}^* \equiv \{\ell^*_{ij}\}$ and $\mathbf{D}^* \equiv \{d^*_{jj}\}$, such that the modified matrix $\mathbf{A}^*$ should be equal to

$$\mathbf{A}^* = \mathbf{L}^* \mathbf{D}^* \mathbf{L}^{*\mathrm{T}}$$

Supposing that $\mathbf{A}$ and $\mathbf{A}^*$ are, both of them, positive definite (see above), the recurrence relations for modifying $\mathbf{L}$ and $\mathbf{D}$ are given below.

1. Define $\alpha_1 = \alpha$, $\boldsymbol{w}^{(1)} = \boldsymbol{z}$
2. For $j = 1, 2, \ldots, n$, compute

$$\left.\begin{aligned} p_j &= w_j^{(j)} \\ d^*_{jj} &= d_{jj} + \alpha_j p_j^2 \\ \beta_j &= p_j \alpha_j / d^*_{jj} \\ \alpha_{j+1} &= d_{jj} \alpha_j / d^*_{jj} \\ w_r^{(j+1)} &= w_r^{(j)} - p_j \ell_{rj} \\ \ell^*_{rj} &= \ell_{rj} + \beta_j w_r^{(j+1)} \end{aligned}\right\} \quad r = j+1,\ j+2, \ldots, n$$

2.15 The Minimax Filter

The Kalman (also called H_2) filter is based on the principle of searching the minimum variance of the average estimation error for linear systems affected by a Gaussian noise. However, sometimes the statistical properties of the noise affecting the system are not known. In such cases, it is necessary to search the minimum of the worst-case, instead of the average, estimation error.

These limitations have given rise to the minimax (also called H_∞) filter. This type of filter bears this name because it searches the minimum of the maximum singular value of the transfer function from the noise to the estimation error. Unlike the Kalman filter, which requires the knowledge of the statistical properties of the noise affecting the observed process, the minimax filter does not require this knowledge. Therefore, H_∞ filters are more robust (that is, more tolerant or less sensitive to disturbances and modelling uncertainties) than are Kalman filters.

As shown in Sect. 2.9, the Kalman filter is based on the assumption of a random distribution, with zero mean value, of the measurement errors. In other words, the average value of the process noise must be zero, and the average value of the measurement noise must also be zero. This property must hold not only over the whole duration of the process, but also at each time instant. Under these conditions, the Kalman filter leads to the smallest possible quadratic standard deviation (or 2-norm) of the estimation error.

By the way, let $\varepsilon(t)$ be function representing a scalar signal in the time domain. Let $[a, b]$ be the interval of definition of the function $\varepsilon(t)$.

The 2-norm (also written ℓ^2-norm) of $\varepsilon(t)$ is defined as follows

$$\|\varepsilon\|_2 = \left(\int_a^b |\varepsilon(t)|^2 \mathrm{d}t \right)^{\frac{1}{2}}$$

If the integral of the square of the absolute value of $\varepsilon(t)$ is finite, then the function $\varepsilon(t)$ is said to be square integrable.

The infinity-norm (also written ℓ^∞) of $\varepsilon(t)$ is defined as follows

$$\|\varepsilon\|_\infty = \max_t [|\varepsilon(t)|] \qquad (a \leq t \leq b)$$

In order to understand why we seek sometimes the minimum of the infinity-norm instead of the minimum of the 2-norm, let us consider a set of measured data $f(t)$, which is smooth everywhere over the measurement interval $A \equiv a \leq t \leq b$, but has outlying values much larger (or much smaller) in a very narrow sub-interval B contained in A than those outside B. This happens, for example, when $f(t)$ has either a sharp peak or a sharp notch in B. Let us consider a polynomial $p(t)$, for example, a fifth-degree polynomial as was the case with the polynomial considered in Sect. 2.10, defined so that $p(t)$ should be the least-squares approximation to $f(t)$. In other words, let $p(t)$ be the fifth-degree approximating polynomial obtained by searching the minimum 2-norm of the standard deviation.

By defining the error as follows: $\varepsilon(t) = f(t) - p(t)$, the 2-norm of $\varepsilon(t)$ is

$$\|\varepsilon\|_2 = \left(\int_a^b |\varepsilon(t)|^2 \mathrm{d}t \right)^{\frac{1}{2}}$$

Since $p(t)$ corresponds to the minimum 2-norm of $\varepsilon(t) = f(t) - p(t)$, then the discrepancy $\varepsilon(t)$ between the measured and the computed data is expected to be in magnitude near zero outside the sub-interval B and much larger than zero inside B, just where $f(t)$ has a sharp peak (or notch).

Using the criterion of the minimum 2-norm is tantamount to rejecting outlying values from a set of measured data, on the grounds that such values appear to be inconsistent with the other data. This line of conduct is arbitrary and sometimes unjustified, because it exposes the observer to the risk of ignoring perfectly good data and the information which they carry with them.

Generally speaking, the question of whether outliers should, or should not, be retained in a set of measured values is quite difficult. It has been the subject of many studies since the times of Gauss. The interested reader can find extensive information on the matter, for example, in Refs. [3, 24]. For the purpose of the present paragraph, which describes the minimax filter, it is assumed that this problem has been solved.

Let us consider once again the problem of estimating the state vector $\boldsymbol{x}_k$ of a system which varies linearly with time. The system in question is governed by the following equations

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \boldsymbol{f}(\boldsymbol{x}_k, \boldsymbol{u}_k) + \boldsymbol{v}_k = \mathbf{A}_k \boldsymbol{x}_k + \mathbf{B}_k \boldsymbol{u}_k + \boldsymbol{v}_k \\ \boldsymbol{z}_k &= \boldsymbol{h}(\boldsymbol{x}_k) + \boldsymbol{n}_k = \mathbf{H}_k \boldsymbol{x}_k + \boldsymbol{n}_k\end{aligned}$$

where $\boldsymbol{x}_k$ is the $m \times 1$ state vector of the system, $\mathbf{A}_k$ is an $m \times m$ matrix, $\boldsymbol{u}_k$ is the $r \times 1$ vector of the known input to the system, $\mathbf{B}_k$ is an $m \times r$ matrix, $\boldsymbol{v}_k$ is the random process noise vector, $\boldsymbol{z}_k$ is the $n \times 1$ measurement vector, $\mathbf{H}_k$ is an $n \times m$ matrix, and $\boldsymbol{n}_k$ is the random measurement noise vector.

We want to estimate the state vector $\boldsymbol{x}_{k+1}$ on the basis of the measurement vector $\boldsymbol{z}_k$ and our knowledge of the system equations. Following Simon [64], the estimated state vector $\boldsymbol{x}_{k+1}$ (that is, $\boldsymbol{x}_{k+1}^-$) is expressed as follows

$$\boldsymbol{x}_{k+1}^- = \mathbf{A}_k \boldsymbol{x}_k + \mathbf{K}_k (\boldsymbol{z}_{k+1} - \mathbf{H}_k \mathbf{A}_k \boldsymbol{x}_k)$$

where $\mathbf{K}_k$ is some gain matrix which is to be determined. If the criterion of the least 2-norm of the estimation error were used, then $\mathbf{K}_k$ would be just the Kalman gain, as has been shown in the preceding paragraphs. By contrast, here we want to determine $\mathbf{K}_k$ on the basis of the least infinity-norm of the estimation error.

Among several possible solutions, Simon cites that one which determines $\mathbf{K}_k$ so that the maximum singular value of the transfer function from the noise to the estimation error should be less than a given scalar value γ. To this end, Simon defines a loss function J, which is a performance measure of the estimator, as will be shown below. Let N be the size of the measurement process (in other words, the time-step k ranges from 0 to $N-1$). Let $\mathbf{X}_k$, $\mathbf{V}_k$, and $\mathbf{N}_k$ be weighting matrices associated with, respectively, the estimation error, the process noise, and the measurement noise. Let

$$\begin{aligned}\mathrm{P} &= \sum_{k=0}^{N-1} \left\| \boldsymbol{x}_k - \boldsymbol{x}_k^- \right\|_{\mathbf{x}}^2 \\ \mathrm{Q} &= \sum_{k=0}^{N-1} \left\| \boldsymbol{v}_k \right\|_{\mathbf{v}}^2 \\ \mathrm{R} &= \sum_{k=0}^{N-1} \left\| \boldsymbol{n}_k \right\|_{\mathbf{n}}^2\end{aligned}$$

be the weighted squared two-norms of, respectively, the estimation error vector $(\boldsymbol{x}_k - \boldsymbol{x}_k^-)$, the process noise vector $\boldsymbol{v}_k$, and the measurement noise vector $\boldsymbol{n}_k$. In the expressions given above, the notation $||\boldsymbol{x}_k||_{\mathbf{x}}^2$ is used to indicate $\boldsymbol{x}_k^{\mathrm{T}} \mathbf{X}_k \boldsymbol{x}_k$, where $\mathbf{X}_k$ is a weighting matrix associated with $\boldsymbol{x}_k$.

Simon [64] defines the loss function J as follows

$$\mathrm{J} = \frac{\mathrm{P}}{\mathrm{Q} + \mathrm{R}}$$

As a result of this definition, large values of J correspond to large deviations of $\boldsymbol{x}_k$ from its estimate $\boldsymbol{x}_k^-$. The denominator Q + R of the fraction given above may be considered as the energy of the unknown noise terms; likewise, the numerator P represents the energy of the estimation error. Let the noise vectors $\boldsymbol{v}_k$ and $\boldsymbol{n}_k$ be what they may, a minimax filter is meant to provide a uniformly small estimation error $(\boldsymbol{x}_k - \boldsymbol{x}_k^-)$, so that the loss function J should be bounded by a prescribed value [50], as will be shown below.

In other words, the estimator of a minimax filter tries to determine the estimated state vector $\boldsymbol{x}_k^-$ in such a way as to make the value of J as small as possible. To this end, the filter designer chooses a value of γ such that

$$\mathrm{J} < \frac{1}{\gamma}$$

The weighting matrices $\mathbf{X}_k$, $\mathbf{V}_k$ and $\mathbf{N}_k$ must also be chosen by the filter designer so as to reach desired results.

The state vector estimate $\boldsymbol{x}_k^-$ which forces J to be less than $1/\gamma$ is given by Simon [64] in the following terms:

$$\begin{aligned}
\mathbf{L}_k &= \left(\mathbf{I} - \mathbf{X}_k\mathbf{P}_k + \mathbf{H}_k^T\mathbf{N}_k^{-1}\mathbf{H}_k\mathbf{P}_k\right)^{-1} \\
\mathbf{K}_k &= \mathbf{A}_k\mathbf{P}_k\mathbf{L}_k\mathbf{H}_k^{\mathrm{T}}\mathbf{N}_k^{-1} \\
\boldsymbol{x}_{k+1}^- &= \mathbf{A}_k\boldsymbol{x}_k^- + \mathbf{B}_k\boldsymbol{u}_k + \mathbf{K}_k\left(\boldsymbol{z}_k - \mathbf{H}_k\boldsymbol{x}_k^-\right) \\
\mathbf{P}_{k+1} &= \mathbf{A}_k\mathbf{P}_k\mathbf{L}_k\mathbf{A}_k^{\mathrm{T}} + \mathbf{V}_k
\end{aligned}$$

where $\mathbf{L}_k$ is the output matrix, $\mathbf{I}$ is the identity matrix, $\mathbf{P}_k$ is the error covariance matrix, and $\mathbf{K}_k$ is the minimax gain matrix (whose counterpart, in case of the Kalman filter, is the Kalman gain matrix).

The condition $\mathrm{J} < 1/\gamma$ shows that, in a minimax filter, the ratio of the estimation error energy to the noise energy is inversely proportional to the value chosen by the designer for γ. In practice, this value cannot be chosen arbitrarily large, because the mathematical derivation of the minimax filter equation is based on the hypothesis of an error covariance matrix $\mathbf{P}_k$ whose eigenvalues are, in magnitude, less than one. If we select a value of γ which is too large, this condition is not satisfied. In other words, it is impossible to find an estimator which makes the estimation error arbitrarily small [64].

The minimax filter equations given above require operations of matrix inversion at each time-step. In this regard, Simon [64] notes that such operations need not be performed in practice at each time-step, because the matrices $\mathbf{L}_k$, $\mathbf{P}_k$, and $\mathbf{K}_k$ can be computed off-line. In other words, when the value of γ has been properly chosen,

there is no necessity of measurements to recompute the minimax gain matrix $\mathbf{K}_k$ at each step, because the filtering problem can be solved by using a constant value ($\mathbf{K}$) of the minimax gain matrix. To determine this constant value, Simon [64] notes that the matrices $\mathbf{K}_k$, computed off-line, approach very quickly a steady-state value, that is, the matrices $\mathbf{K}_k$ converge, after a few time-steps, to a constant matrix $\mathbf{K}$. In a successive article [65], Simon proposes to compute $\mathbf{K}$ by solving the following simultaneous equations:

$$\begin{aligned}\mathbf{K} &= (\mathbf{I}+\mathbf{P}/\gamma)^{-1}\mathbf{P}\mathbf{H}^{\mathrm{T}}\\ \mathbf{P}^{-1} &= \mathbf{M}^{-1}-\mathbf{I}/\gamma+\mathbf{H}^{\mathrm{T}}\mathbf{H}\\ \mathbf{M} &= \mathbf{A}\mathbf{P}\mathbf{A}^{\mathrm{T}}+\mathbf{I}\end{aligned}$$

If the value of γ were improperly chosen by the filter designer, then no solution to these equations would be found.

Another possible method, also suggested by Simon [65], is given below.

1. Form the following $2m \times 2m$ matrix

$$\mathbf{Z} = \begin{bmatrix} \mathbf{A}^{-\mathrm{T}} & \mathbf{A}^{-\mathrm{T}}\left(\mathbf{H}^{\mathrm{T}}\mathbf{H}-\mathbf{I}/\gamma^2\right) \\ \mathbf{A}^{-\mathrm{T}} & \mathbf{A}+\mathbf{A}^{-\mathrm{T}}\left(\mathbf{H}^{\mathrm{T}}\mathbf{H}-\mathbf{I}/\gamma^2\right) \end{bmatrix}$$

2. Find the eigenvectors of $\mathbf{Z}$. Let $\boldsymbol{\chi}_1$, $\boldsymbol{\chi}_2$, …, $\boldsymbol{\chi}_m$ be those eigenvectors which correspond to eigenvalues outside the unit circle.
3. Form the following matrix

$$\begin{bmatrix}\boldsymbol{\chi}_1 & \boldsymbol{\chi}_2 & \cdots & \boldsymbol{\chi}_m\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 \\ \mathbf{X}_2\end{bmatrix}$$

   where $\mathbf{X}_1$ and $\mathbf{X}_2$ are $m \times m$ matrices.
4. Compute $\mathbf{M} = \mathbf{X}_2\,\mathbf{X}_1^{-1}$

Simon [65] notes that this method works only if $\mathbf{X}_1$ has an inverse matrix. Otherwise, the chosen value of γ is too large.

As the classical Kalman filter, so the minimax filter has been created for linear systems. For nonlinear systems (as is the case with the tracking of an orbiting spacecraft, whose behaviour is governed by a nonlinear differential equation), the system equations

$$\begin{aligned}\boldsymbol{x}_k &= \boldsymbol{f}(\boldsymbol{x}_{k-1}, \boldsymbol{u}_{k-1}) + \boldsymbol{v}_{k-1}\\ \boldsymbol{z}_k &= \boldsymbol{h}(\boldsymbol{x}_k) + \boldsymbol{n}_k\end{aligned}$$

must be are linearised (by means of a Taylor-series expansion truncated at the first order) about the reference trajectory. Therefore, the minimax filter is subject to the same limitations as those which hold with the extended Kalman filter. When one has to do with nonlinear systems and the robustness (that is, the insensitivity to the

choice of a probability model) of a filter is a concern, it is advisable to use a more robust unscented Kalman filter than that shown in Sect. 2.13. This type of filter will be described in the following paragraph.

2.16 A More Robust Unscented Kalman Filter

As shown in Sect. 2.13, a general nonlinear system is governed by the following equations

$$\begin{aligned} \boldsymbol{x}_k &= \boldsymbol{f}(\boldsymbol{x}_{k-1}, \boldsymbol{u}_{k-1}, \boldsymbol{v}_{k-1}, k-1) \\ \boldsymbol{z}_k &= \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{u}_k, k) + \boldsymbol{n}_k \end{aligned}$$

where $\boldsymbol{x}_k$ is the m-dimensional state vector at time-step k, $\boldsymbol{u}_{k-1}$ is the known input vector, $\boldsymbol{v}_{k-1}$ is the process noise vector due to disturbances and modelling errors, $\boldsymbol{z}_k$ is the observation vector, $\boldsymbol{n}_k$ is the measurement noise vector, and $\boldsymbol{f}$ and $\boldsymbol{h}$ are the nonlinear vector-valued functions (supposed known) representing the system dynamic model.

The initial state vector $\boldsymbol{x}_0$ has known mean vector $\boldsymbol{\mu}_0 = E(\boldsymbol{x}_0)$ and covariance matrix $\mathbf{P}_0 = E[(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)^{\mathrm{T}}]$. The initial state vector $\boldsymbol{x}_0$ is used to construct an augmented (denoted by the superscript a) ℓ-dimensional vector $\boldsymbol{x}_0^a$, which results from concatenating the mean of the true state vector $\boldsymbol{x}_0$ with the mean of the process noise vector $\boldsymbol{v}_0$ and the mean of the measurement noise vector $\boldsymbol{n}_0$, as shown below:

$$\boldsymbol{x}_0^a = \begin{bmatrix} E\left(\boldsymbol{x}_0^{\mathrm{T}}\right) \\ E\left(\boldsymbol{v}_0^{\mathrm{T}}\right) \\ E\left(\boldsymbol{n}_0^{\mathrm{T}}\right) \end{bmatrix}$$

Likewise, the true covariance matrix $\mathbf{P}_0$ of the true state vector $\boldsymbol{x}_0$ is used to construct the following augmented (denoted by the superscript a) covariance matrix

$$\mathbf{P}_0{}^a = \begin{bmatrix} \mathbf{P}_0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_0 \end{bmatrix}$$

where $\mathbf{Q}_0$ is the covariance matrix of the process noise vector $\boldsymbol{v}_0$, and $\mathbf{R}_0$ is the covariance matrix of the measurement noise vector $\boldsymbol{n}_0$. Van Zandt [79] has suggested the following improvement to the unscented Kalman filter: adding a fictitious process noise, even if there be none, to the system, for the purpose of taking account of the uncertainties of the dynamical model used.

By so doing, for $k = 1, 2, \ldots, N - 1$ (N being the size of the measurement process), the prediction phase and the update phase of the filter are shown below.

Prediction

By using the scaling parameter λ (see below), a set of $2\ell + 1$ sigma-point vectors $\boldsymbol{\chi}_{k-1}^{i}$ ($i = 0, 1, \ldots, 2\ell$) is computed from the augmented state vector $\boldsymbol{x}_{k-1}^{a}$ and the augmented covariance matrix $\mathbf{P}_{k-1}^{a}$, as follows

$$\begin{aligned}
\boldsymbol{\chi}_{k-1}^{0} &= \boldsymbol{x}_{k-1}^{a} && i = 0 \\
\boldsymbol{\chi}_{k-1}^{i} &= \boldsymbol{x}_{k-1}^{a} + \left\{ \left[(\ell + \lambda) \mathbf{P}_{k-1}^{a} \right]^{\frac{1}{2}} \right\}_{i} && i = 1, 2, \ldots, \ell \\
\boldsymbol{\chi}_{k-1}^{i+\ell} &= \boldsymbol{x}_{k-1}^{a} - \left\{ \left[(\ell + \lambda) \mathbf{P}_{k-1}^{a} \right]^{\frac{1}{2}} \right\}_{i} && i = 1, 2, \ldots, \ell
\end{aligned}$$

where $\{[(\ell + \lambda)\mathbf{P}_{k-1}^{a}]^{1/2}\}_{i}$ is the i^{th} column of the matrix $\mathbf{S}_{k-1}$, which is the square root of the following matrix

$$(\ell + \lambda)\mathbf{P}_{k-1}^{a}$$

so that, by definition, there results

$$(\ell + \lambda)\mathbf{P}_{k-1}^{a} = \mathbf{S}_{k-1}\mathbf{S}_{k-1}^{\mathrm{T}}$$

The square-root matrix $\mathbf{S}_{k-1}$ is to be computed by means of some stable method, for example, by means of the Cholesky decomposition, as shown in Sect. 2.12.

The set of the sigma-point vectors $\boldsymbol{\chi}_{k-1}^{i}$ ($i = 0, 1, \ldots, 2\ell$) forms the $\ell \times (2\ell + 1)$ sigma-point matrix $\mathbf{X}_{k-1}^{i}$, whose columns are given below

$$\mathbf{X}_{k-1}^{i} = \left[\boldsymbol{x}_{k-1}^{a} \quad \boldsymbol{x}_{k-1}^{a} + \left[(\ell + \lambda)\mathbf{P}_{k-1}^{a} \right]^{\frac{1}{2}} \quad \boldsymbol{x}_{k-1}^{a} - \left[(\ell + \lambda)\mathbf{P}_{k-1}^{a} \right]^{\frac{1}{2}} \right]$$

This matrix is such that $\boldsymbol{x}_{k-1}^{a}$ is a column vector, whereas $\boldsymbol{x}_{k-1}^{a} + [(\ell + \lambda)\mathbf{P}_{k-1}^{a}]^{1/2}$ and $\boldsymbol{x}_{k-1}^{a} - [(\ell + \lambda)\mathbf{P}_{k-1}^{a}]^{1/2}$ are, each of them, a set of ℓ column vectors.

Now, each of the sigma-point vectors is propagated, that is, projected ahead from time-step $k - 1$ to time-step k, by means of the nonlinear function $\boldsymbol{f}$, as follows

$$\boldsymbol{\chi}_{k}^{i} = \boldsymbol{f}\left(\boldsymbol{\chi}_{k-1}^{i}, \boldsymbol{u}_{k-1}, \boldsymbol{\chi}_{k-1}^{i}, k - 1 \right) \quad (i = 0, 1, \ldots, 2\ell)$$

where the first argument of the function $\boldsymbol{f}$ is the sigma-point vector due to the true state vector $\boldsymbol{x}_{k-1}$, and the third argument is the sigma-point vector due to the process noise vector $\boldsymbol{v}_{k-1}$.

Then, these propagated sigma-point vectors $\boldsymbol{\chi}_{k}^{i}$ are weighted (using weights w_{s}^{i} for the state vector and w_{c}^{i} for the covariance matrix) and put together in a proper manner, so as to yield the predicted state vector $\boldsymbol{x}_{k}^{-}$ and the predicted covariance matrix $\mathbf{P}_{k}^{-}$ of $\boldsymbol{x}_{k}^{-}$, as follows

$$\boldsymbol{x}_k^- = \sum_{i=0}^{2\ell} w_s^i \, \boldsymbol{\chi}_k^i$$

$$\mathbf{P}_k^- = \sum_{i=0}^{2\ell} \sum_{j=0}^{2\ell} w_c^{ij} \left(\boldsymbol{\chi}_k^i\right) \left(\boldsymbol{\chi}_k^j\right)^{\mathrm{T}}$$

where the weights w_s^i for the predicted state vector $\boldsymbol{x}_k^-$ and the weights w_c^i (i = 0, 1, …, 2ℓ) for the predicted covariance matrix $\mathbf{P}_k^-$ are as follows

$$\begin{aligned}
w_s^0 &= \tfrac{\lambda}{\ell+\lambda} && (i=0)\\
w_c^0 &= \tfrac{\lambda}{\ell+\lambda} + 1 - \alpha^2 + \beta && (i=0)\\
w_s^i &= w_c^i = \tfrac{1}{2(\ell+\lambda)} && (i \neq 0)\\
\lambda &= \alpha^2(\ell+\kappa) - \ell
\end{aligned}$$

In the expressions given above, the parameters α and κ are used to control the spread of the sigma-points around the mean of the state vector (the values of α and κ are usually set to, respectively, 1×10^{-3} and 1), and the parameter β is used to take account of previous knowledge of the distribution of the state vector around its mean (for a Gaussian distribution, the value of β is set to 2).

Update

In the update phase, each of the sigma-point vectors is used as the argument of the nonlinear function $\boldsymbol{h}$, as follows

$$\boldsymbol{\gamma}_k^i = \boldsymbol{h}\left(\boldsymbol{\chi}_k^i, \boldsymbol{u}_k, \boldsymbol{k}\right) + \boldsymbol{\chi}_k^i \qquad i = 0, 1, \ldots, 2\ell$$

where the first argument of the function $\boldsymbol{h}$ is the sigma-point vector due to the predicted state vector $\boldsymbol{x}_k^-$, and the added term is the sigma-point vector due to the measurement noise vector $\boldsymbol{n}_k$.

The sigma-point vectors $\boldsymbol{\gamma}_k^i$ are weighted (using weights w_s^i for the state vector and w_c^i for the covariance matrix) and put together in a proper manner, so as to yield the predicted measurement vector $\boldsymbol{z}_k^-$ and the predicted covariance matrix $\mathbf{P}_{zz}^-$ of $\boldsymbol{z}_k^-$, as follows

$$\boldsymbol{z}_k^- = \sum_{i=0}^{2\ell} w_s^i \, \boldsymbol{\gamma}_k^i$$

$$\mathbf{P}_{zz}^- = \sum_{i=0}^{2\ell} \sum_{j=0}^{2\ell} w_c^{ij} \left(\boldsymbol{\gamma}_k^i\right) \left(\boldsymbol{\gamma}_k^j\right)^{\mathrm{T}}$$

It is to be noted that the matrix $\mathbf{P}_{zz}^-$ (related to the time-step k) indicated above is the covariance matrix of the predicted measurement vector $\boldsymbol{z}_k^-$, as indicated above.

In order to compute the Kalman gain $\mathbf{K}_k$ for the unscented Kalman filter, it is necessary to compute first the cross-correlation matrix $\mathbf{P}_{xz}^-$ (related to the time-step k) between $\boldsymbol{x}_k^-$ and $\boldsymbol{z}_k^-$, as follows

$$\mathbf{P}_{xz}^- = \sum_{i=0}^{2\ell} \sum_{j=0}^{2\ell} w_c^{ij} \left(\boldsymbol{\chi}_k^i\right) \left(\boldsymbol{\gamma}_k^j\right)^{\mathrm{T}}$$

where $\boldsymbol{\chi}_k^i$ is the sigma-point vector due to the predicted state vector $\boldsymbol{x}_k^-$, as mentioned above. Hence, the Kalman gain $\mathbf{K}_k$ is given by

$$\mathbf{K}_k = (\mathbf{P}_{xz}^-)(\mathbf{P}_{xz}^-)^{-1}$$

As is the case with the classical Kalman filter, the updated state vector $\boldsymbol{x}_k^+$ results from the predicted state vector $\boldsymbol{x}_k^-$ plus the innovation $(z_k - z_k^-)$ weighted by the Kalman gain $\mathbf{K}_k$, as follows

$$\boldsymbol{x}_k^+ = \boldsymbol{x}_k^- + \mathbf{K}_k(z_k - z_k^-)$$

and the updated covariance matrix $\mathbf{P}_k^+$ results from the predicted covariance matrix $\mathbf{P}_k^-$ minus the predicted measurement covariance matrix $\mathbf{P}_{zz}^-$ (related to k) weighted by the Kalman gain $\mathbf{K}_k$, as follows

$$\mathbf{P}_k^+ = \mathbf{P}_k^- - \mathbf{K}_k \mathbf{P}_{zz}^- \mathbf{K}_k^{\mathrm{T}}$$

A filter like that described above has been applied by van der Merwe and Wan [85] to the problem of fusing noisy observations from the Global Positioning System (GPS), Inertial Measurement Units (IMU), and other available sensors (such as a barometric altimeter or a magnetic compass). These observations have been combined with a kinematic or dynamic model of an unmanned aerial vehicle (specifically, a remotely controlled helicopter). The results obtained by van der Merwe and Wan by using this type of filter indicate an error reduction of approximately 30% in both attitude and position estimates relative to the results coming from an extended Kalman filter [85].

References

1. http://commons.wikimedia.org/wiki/File:Sidereal_and_Synodic_Day.png
2. Abdi, H., Singular value decomposition (SVD) and generalised singular value decomposition (GSVD), 2007, article available at the web site http://www.utdallas.edu/~herve/Abdi-SVD2007-pretty.pdf
3. F.J. Anscombe, I. Guttman, Rejection of outliers. Technometrics **2**(2), 123–147 (1960)
4. S. Aoki, B. Guinot, G.H. Kaplan, H. Kinoshita, D.D. McCarty, P.K. Seidelmann, The new definition of universal time. Astron. Astrophys. **105**(2), 359–361 (1982)

5. R.R. Bate, D.D. Mueller, J.E. White, *Fundamentals of astrodynamics* (Dover Publications, New York, 1971). ISBN 0-486-60061-0
6. Bebis, G., Singular value decomposition (SVD), article available at the web site http://www.cse.unr.edu/∼bebis/MathMethods/SVD/lecture.pdf
7. J.F. Bellantoni, K.W. Dodge, A square root formulation of the Kalman-Schmidt filter. AIAA Journal **5**(7), 1309–1314 (1967)
8. G.J. Bierman, *Factorization methods for discrete sequential estimation* (Dover Publications, Mineola, N.Y., 2006). ISBN 0-486-44981-5
9. Bierman, G.J., A comparison of discrete linear filtering algorithms, IEEE Transactions on Aerospace and Electronic Systems, Vol. AES-9, Issue 1, January 1973, pp. 28–37
10. Born, G.H., Potter square root filter, ASEN 5070, Colorado Center for Astrodynamics Research, Aerospace Engineering Sciences, University of Colorado at Boulder, 30 October 2002, article available at the web site http://ccar.colorado.edu/ASEN5070/handouts/PSRF2.pdf
11. Born, G.H., Statistical orbit determination, Cholesky algorithm, ASEN 5070, Lecture 23, Colorado Center for Astrodynamics Research, Aerospace Engineering Sciences, University of Colorado at Boulder, 25 October 2006
12. D. Boulet, *Methods of orbit determination for the microcomputer* (Willmann-Bell, Richmond, 1991). ISBN 0-943396-34-4
13. Branham, R.L., Jr., Laplacian orbit determination, Astronomy in Latin America, Second Meeting on Astrometry in Latin America and Third Brazilian Meeting on Fundamental Astronomy, held on 2–5 September 2002, edited by R. Teixeira, N.V. Leister, V.A.F. Martin, and P. Benevides-Soares, ADeLA Publications Series, Vol. 1, No. 1 (2003), pp. 85-89, article also available at the web site http://www.astro.iag.usp.br/∼adelabr/Branham01_14.pdf
14. Bube, K.P. and Langan, R.T., Hybrid $\ell 1/\ell 2$ minimization with applications to tomography, Geophysics, Vol. 62, No. 4 (July-August 1997), pp. 1183–1195
15. R.S. Bucy, P.D. Joseph, *Filtering for stochastic processes with applications to guidance* (AMS Chelsea Publishing, Providence, Rhode Island, 2005). ISBN 0-8218-3782-6
16. CIA, image available at the web site https://www.cia.gov/library/publications/the-world-factbook/maps/refmap_time_zones.html
17. N.A. Carlson, Fast triangular formulation of the square root filter. AIAA Journal **11**(9), 1259–1265 (1973)
18. Chobotov, V.A. (editor), Orbital mechanics, AIAA Education Series, 3rd edition, 2002, ISBN 1-56347-537-5
19. Conway, D.J., Netreba, N.A. and Bristow, J., A self-tuning real-time orbit determination system, 1998, http://www.ai-solutions.com/library/tech.asp
20. H.D. Curtis, *Orbital mechanics for engineering students* (Butterworth-Heinemann, Oxford, 2005). ISBN 0-7506-6169-0
21. J.J. Dongarra, C.B. Moler, J.R. Bunch, G.W. Stewart, *LINPACK user" guide* (Society for Industrial and Applied Mathematics, Philadelphia, 1979). ISBN 0-89871-172-X
22. Escobal, P.R., Methods of orbit determination, Krieger Publishing Company, 1976, ISBN 0-88275-319-3
23. Espenak, F. and Meeus, J., Five millennium canon of solar eclipses: -1999 to +3000 (2000 BCE to 3000 CE), Revision 1 (2007 May 11), NASA/TP-2006-214141, document available at the web site http://eclipse.gsfc.nasa.gov/5MCSE/5MCSE-Text11.pdf
24. Ferguson, Th.S., On the rejection of outliers, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press, 1961, pp. 253–287, article available at the web site http://projecteuclid.org/euclid.bsmsp/1200512169
25. Fliegel, H. F. and van Flandern, T. C., Communications of the ACM, Vol. 11, No. 10, October 1968
26. C.F. Gauss, *Theoria motus corporum coelestium in sectionibus conicis Solem ambientium* (Perthes & Besser, Hamburg, 1809)

27. Gibbs, J.W., On the determination of elliptic orbits from three complete observations, Memoirs of the National Academy of Sciences, Vol. 4, Part 2 (1889), pp. 79–104
28. Gill, P.E., Golub, G.H., Murray, W. and Saunders, M.A., Methods for modifying matrix factorizations, Mathematics of Computation, Vol. 28, No. 126, April 1974, pp. 505–535, article also available at the web site www.stanford.edu/group/SOL/papers/ggms74.pdf
29. Golub, G.H. and Reinsch, C., Singular value decomposition and least squares solutions, Numerische Mathematik, Vol. 14, No. 5, 1970, pp. 403–420, web site http://people.duke.edu/~hpgavin/SystemID/References/Golub+Reinsch-NM-1970.pdf
30. M. Grewal, J. Kain, Kalman filter implementation with improved numerical properties. IEEE Trans. Autom. Control **55**(9), 2058–2068 (2010)
31. E.A. Guillemin, *The mathematics of circuit analysis* (John Wiley & Sons, New York, 1956)
32. Guitton, A., The iteratively reweighted least square method, http://sepwww.stanford.edu/public/docs/sep103/antoine2/paper_html/node4.html
33. Healy, L.M., Student projects for space navigation and guidance, Paper AAS 03-500, Astrodynamics 2003, Advances in the Astronautical Sciences, Volume 116, pp. 3-20, web site http://drum.lib.umd.edu/bitstream/1903/3030/2/2003-healy.pdf
34. S. Herrick, *Astrodynamics*, vol. 1 (Van Nostrand Reinhold, London, 1971). ISBN 0-442-03370-2
35. S. Herrick, *Astrodynamics*, vol. 2 (Van Nostrand Reinhold, London, 1972). ISBN 0-442-03371-0
36. Hoots, F.R. and Roehrich, R.L., Models for propagation of NORAD element sets, Spacetrack report No. 3, December 1980, available at the web site http://celestrak.com/NORAD/documentation/spacetrk.pdf
37. Hotop, H.J., Recent developments in Kalman filtering with applications in navigation, in Hawkes, P.W. (editor), Advances in electronics and electron physics, Vol. 85, Academic Press, San Diego, CA, 1993, ISBN 0–12-014727-0
38. Husfeld, D. and Kronberg, C., Astronomical time keeping, 1996, article available at the web site http://www.maa.mhn.de/Scholar/times.html
39. IBM, System/360 Scientific Subroutine Package, Version III, Programmer" Manual, Fifth Edition, August 1970
40. Islam, S., Sadiq, M. and Qureshi, M.S., Assessing polynomial approximation for ΔT, Journal of Basic and Applied Sciences, Vol. 4, No.1, 2008, pp. 1–4, article available at http://www.jbaas.com/articles/vol_4_n1_c1.php
41. Jefferys, W.H., Julian day numbers, University of Texas at Austin, web site http://quasar.as.utexas.edu/BillInfo/JulianDatesG.html
42. Julier, S.J. and Uhlmann, J.K., A new extension of the Kalman filter to nonlinear systems, 11th International Symposium on Aerospace/Defence Sensing, Simulation and Control, Orlando, Florida, 21–24 April 1997
43. Julier, S.J., The scaled unscented transformation, Proceedings of the 2002 American Control Conference, Vol. 6, 8–10 May 2002, pp. 4555-4559
44. Kalman, R.E., A new approach to linear filtering and prediction theory, Journal of Basic Engineering, Vol. 82, Series E, No. 1 (1960), pp. 35-45
45. Kalman, R.E. and Bucy, R.S., New results in linear filtering and prediction theory, Journal of Basic Engineering, Vol. 83, Series D, No. 1 (1961), pp. 95–108
46. E.W. Kang, *Radar system analysis, design, and simulation* (Artech House, Norwood, 2008). ISBN 978-1-59693-347-7
47. Leach, S., Singular value decomposition – a primer, article available at http://www.cs.brown.edu/research/ai/dynamics/tutorial/Postscript/SingularValueDecomposition.ps
48. Lemon, K. and Welch, B.W., Comparison of nonlinear filtering techniques for lunar surface roving navigation, NASA/TM-2008-215151, May 2008, pp. 1-20
49. Levy, L.J., The Kalman filter: navigation's integration workhorse, The Johns Hopkins University, Applied Physics Laboratory, 1997, available at the web site http://www.cs.unc.edu/~welch/kalman/Levy1997/Levy1997_KFWorkhorse.pdf

50. Li, X. and Zell, A., H∞ filtering for a mobile robot tracking a free rolling ball, in Lakemeyer, G., Sklar, E., Sorrenti, D.G. and Takahashi, T., ed., RoboCup 2006: Robot Soccer World Cup X, Vol. 4434, Springer, 2007, pp. 296–303
51. B.G. Marsden, Initial orbit determination: the pragmatist's point of view. The Astronomical Journal **90**(8), 1541–1547 (1985)
52. McCarthy, D.D., Astronomical time, Proceedings of the IEEE, Vol. 79, No. 7, July 1991, pp. 915–920, article also available at the web site http://www.cl.cam.ac.uk/∼mgk25/volatile/astronomical-time.pdf
53. O. Montenbruck, E. Gill, *Satellite orbits: models, methods and applications* (Springer-Verlag, Berlin, 2000). ISBN 3-540-67280-X
54. Montenbruck, O., An epoch state filter for use with analytical orbit models of low earth satellites, Aerospace Science and Technology, Volume 4, Issue 4, June 2000, pp. 277–287, http://www.weblab.dlr.de/rbrt/pdf/AST_00277.pdf
55. Morrison, L.V. and Stephenson, F.R., Historical values of the Earth's clock error ΔT and the calculation of eclipses, Journal for the History of Astronomy, Vol. 35 part 3, August 2004, No. 120, pp. 327–336, article available at the web site http://articles.adsabs.harvard.edu/full/2004JHA….35..327M
56. F.R. Moulton, The true radii of convergence of the expressions for the ratios of the triangles when developed as power-series in the time-intervals. The Astronomical Journal **23**(537–538), 93–102 (1903)
57. NASA, Landsat 7, map of US ground stations taken from the web site http://landsathandbook.gsfc.nasa.gov/handbook/handbook_htmls/chapter2/images/globe_wstations.gif
58. NIMA Technical Report TR8350.2, Third edition, 3 January 2000, web site http://earth-info.nga.mil/GandG/publications/tr8350.2/wgs84fin.pdf
59. National Atlas of the United States, March 5, 2003, http://nationalatlas.gov
60. P.J. Olver, C. Shakiban, *Applied linear algebra* (Pearson Prentice-Hall, Upper Saddle River, 2006). ISBN 0-13-147382-4
61. J.E. Potter, *New statistical formulas, Technical report, Memo 40* (Intrumental Laboratory, Massachusetts Institute of Technology, 1963)
62. Rojas, R., The Kalman filter, The Mathematical Intelligencer, Vol. 1, No. 2, 2005, pp. 1–7, web site http://robocup.mi.fu-berlin.de/buch/kalman.pdf
63. Schmidt, S.F., Computational techniques in Kalman filtering, in Leondes, C.T. (editor), Theory and applications of Kalman filtering, AGARDograph 139, February 1970
64. Simon, D., From here to infinity, Embedded Systems Programming, Vol. 14, No. 11, October 2001, pp. 20–32, article also available at the web site http://academic.csuohio.edu/simond/courses/eec641/hinfinity.pdf
65. Simon, D., Introduction to minimax filtering, EE Times Design, 21 August 2002, article available at the web site http://www.eetimes.com/design/analog-design/4017991/Introduction-to-Minimax-Filtering
66. L.G. Taff, On initial orbit determination. The Astronomical Journal **89**(9), 1426–1428 (1984)
67. Teixeira, B.O.S., Santillo, M.A., Erwin, R.S., and Bernstein, D.S., Spacecraft tracking using sampled-data Kalman filters, IEEE Control Systems Magazine, August 2008, pp. 78–94
68. Terejanu, G.A., Unscented Kalman filter tutorial, Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY 14260, web site https://cse.sc.edu/∼terejanu/files/tutorialUKF.pdf
69. C.L. Thornton, Triangular covariance factorizations for Kalman filtering. NASA Technical Memorandum **33–798**(15), 1–212 (1976)
70. Thornton, C.L. and Border, J.S., Radiometric tracking techniques for deep-space navigation, Jet propulsion Laboratory, California Institute of Technology, JPL Publication 00–11, October 2000, article available at the web site http://descanso.jpl.nasa.gov/Monograph/series1/Descanso1_all.pdf
71. United States Naval Observatory, Astronomical Applications Department, web site http://aa.usno.navy.mil/data/docs/JulianDate.php

72. United States Naval Observatory, Astronomical Applications Department, web site http://aa.usno.navy.mil/faq/docs/SunApprox.php
73. United States Naval Observatory, web site http://www.usno.navy.mil/USNO
74. United States Naval Observatory, Astronomical Applications Department, web site http://aa.usno.navy.mil/faq/docs/TT.php
75. United States Naval Observatory, Department of the Navy, Time Service Department, web site http://tycho.usno.navy.mil/sidereal2.html
76. University of Tennessee, Dept. Physics & Astronomy, web site http://csep10.phys.utk.edu/astr161/lect/time/coordinates.html
77. Vallado, D.A., Fundamentals of astrodynamics and applications, Springer-Verlag, New York, Third edition, ISBN 0-387-71831-1
78. Vallado, D.A. and Crawford, P., SGP4 orbit determination, August 2008, pp. 1–29, web site http://www.centerforspace.com/downloads/files/pubs/AIAA-2008-6770.pdf
79. Van Zandt, J.R., A more robust unscented transform, Technical report, MITRE Corporation, July 2001, article also available at the web site http://mitre.org/work/tech_papers/tech_papers_01/vanzandt_unscented/vanzandt_unscented.pdf
80. Vergez, P., Sauter, L. and Dahlke, S., An improved Kalman filter for satellite orbit predictions, The Journal of the Astronautical Sciences, Vol. 52, No. 3, July-September 2004, pp. 122, web site http://handle.dtic.mil/100.2/ADA431057
81. Wan, E.A. and van der Merwe, R., The unscented Kalman filter for nonlinear estimation, Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000, AS-SPCC, The IEEE 2000 (6 August 2002), pp. 153–158
82. Wikipedia, web site http://en.wikipedia.org/wiki/QR_decomposition
83. J.H. Wilkinson, *The algebraic eigenvalue problem* (Clarendon Press, Oxford, 1965)
84. van der Merwe, R. and Wan, E.A., The square-root unscented Kalman filter for state and parameter-estimation, Proceedings of the International Conference on Acoustic, Speech, and Signal Processing, 2001 (ICASSP "1), Salt Lake City, Utah, USA, 7–11 May 2001, Vol. 6, pp. 3461-3464
85. van der Merwe, R. and Wan, E.A., Sigma-point Kalman filters for integrated navigation, Proceedings of the 60th Annual Meeting of The Institute of Navigation, Dayton, Ohio, 7–9 June 2004, pp. 641-654, article also available at the web site http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.9.5753

http://www.springer.com/978-3-319-62219-4

Practical Astrodynamics

de Iaco Veris, A.
2018, XIV, 1309 p. 359 illus., 100 illus. in color. In 2 volumes, not available separately., Hardcover
ISBN: 978-3-319-62219-4