The method of least squares used to invert an orbit problem

Ross Bannister

Data Assimilation Research Centre,

Department of Meteorology, University of Reading,

Reading, RG6 6BB, United Kingdom

Abstract

Six parameters uniquely describe the orbit of a body about the Sun. Given these parameters it is possible to make predictions of the body's position by solving its equation of motion. The parameters cannot be directly measured, so they must be inferred indirectly by an inversion method which uses measurements of other quantities in combination with the equation of motion. Inverse techniques are valuable tools in many applications where only noisy, incomplete and indirect observations are available for estimating parameter values. The methodology of the approach is introduced and the Kepler problem is used as a real-world example.

I. INTRODUCTION

Central to science is its ability to make predictions about the behavior of a system. The system could be as simple as the motion of a pendulum, or as complex as the dynamics of the Earth's atmosphere, but the process of predicting outcomes is the same in each case. The process requires a a model that embodies the interactions between the system's components.

Although there is no such thing as an exact model, a useful model is one that is consistent with existing observations and that suggests new ones. Given enough information, that is, the physical laws and the values of the model's parameters, the model should determine what the observations will find if the model represents the system adequately. A model that works in this order is often termed a "forward model." It may, for example, involve the calculation of the position of a pendulum bob at a specified time, or the thermal radiation emitted by a column of the atmosphere at the time that a satellite passes overhead. We contrast a forward model to an "inverse model" which works in the reverse sense and asks what laws or model parameters are needed to agree with a given set of observations.

There are two broad classes of inverse problems. The first class consists of exact mathematical inversion techniques that require perfect and complete input data. A familiar example is the inversion of the Laplacian operator to solve Poisson's equation, which requires the source term and boundary conditions to be known exactly[1]. The second class consists of techniques that account for noisy and incomplete observations.[2] The latter kind of problem is the focus of this paper. Problems in both classes often reduce to a matrix equation, which require the inversion of a matrix[3] to solve the problem.

Inverse modeling techniques are used in many problems where the desired parameters cannot be measured directly. Instead, the information must be inferred by using a combination of models and measurements. A well-known technique in atmospheric science is known as "four-dimensional variational data assimilation"[4] (three space and one time dimension) and is used by weather forecasters to compute a realistic set of initial conditions that allow the best possible weather forecast to be made. This inverse method can exploit remotely-sensed measurements made by satellites. Solar scientists use inversion techniques to determine the interior properties of the Sun, such as the temperature, density, and rotation rates, by observing seismic waves on the Sun's surface.[5] Geologists use measurements of seismic waves in a similar way to investigate the structure of the Earth's crust.[6] These are only a few applications of inverse modeling.

As shown in this paper, the formulation of the inverse problem itself requires use of the forward model. From a first guess of the model parameters the observations are predicted and compared to the actual observations. This allows the parameters to be adjusted in an iterative fashion for consistency between the predicted and actual observations within the bounds of observational uncertainty.

Some of the earliest and most celebrated examples of inverse methods are in astronomy. Newton, Halley, and Laplace contributed to the problem of orbit determination.[7] In 1705 Halley applied the method described by Newton to comet sightings made in preceding centuries. Perhaps the inverse method that is best suited to the problem of orbit determination is the "method of least squares" due to Gauss.[8,9] It is sometimes called the "method of differential correction,"[10] because it uses observations to refine an initial estimate of the parameters needed to define an orbit. The parameters cannot be known exactly because all observations have inaccuracies as do all models of planetary motion. However, the accuracy to which the parameters can be deduced can be improved with more and better observations and better models.

The usefulness of inverse methods lies in their ability to infer information that is not measurable directly – information that is confirmable by subsequent measurements (as demonstrated famously by Halley[11] when he predicted the return of his comet in 1758 and by Gauss[12] in 1802 when he was able to predict the position of the minor planet Ceres using only a minimum of prior observations.

Inverse problems can be particularly difficult to treat if there are many unknown parameters, or if the forward model is non-linear, especially if no a priori knowledge of the parameters is available. The aim of this paper is to demonstrate the method of least squares to determine the six orbital parameters needed to describe the orbit of Venus using a set of simple line of sight observations. The method of least squares is described in Sec. II. Our non-linear forward model is discussed in Sec. III, and we make use of only limited prior knowledge (that Venus's orbit is elliptical). However, the problem is manageable because the number of parameters is relatively low (six). The results are compared to the known parameters[13] in Secs. IV and V.

This example is educationally valuable as it combines a range of skills in a single project. Such skills include the physics of the forward model (the mechanics of Keplerian orbits, astronomy, and coordinate systems), linear algebra, statistics, error analysis, measurements and uncertainties, as well as the inversion technique itself.

II. THE METHOD OF LEAST SQUARES

The method of least squares was invented by Gauss as a means of fitting observations to a model in an optimal fashion. This inversion method involves minimizing a function of the model parameters which penalizes the misfit to the observation set.

A. Probability distributions

Our description starts with a statistical representation of the observations. If we assume that the observation errors obey Gaussian statistics, the probability distribution function, P, for a single observable, y, is a Gaussian with standard deviation σ and mean $\langle y \rangle$,

$$P(s_y) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\big(-\frac{s_y^2}{2\sigma^2}\big), \tag{1}$$

where the parameter s_y is the departure from the mean, $s_y = y - \langle y \rangle$. $P(s_y)\mathrm{d}s_y$ is the probability of making a measurement of y in the interval $\langle y \rangle + s_y$ to $\langle y \rangle + s_y + \mathrm{d}s_y$. Usually more than one variable is measured. The observations can be assembled into a vector, $\mathbf{y}$, with n components (for n measurements). The probability distribution function relevant to this case is akin to Eq. (1),

$$P(\mathbf{s_y}) = \frac{1}{(2\pi)^{n/2}|\mathbf{R}|^{1/2}} \exp\big(-\frac{1}{2}\mathbf{s_y}^T\mathbf{R}^{-1}\mathbf{s_y}\big), \tag{2}$$

where the vector $\mathbf{s_y} = \mathbf{y} - \langle \mathbf{y} \rangle$. $P(\mathbf{s_y})\mathrm{d}\mathbf{s_y}$ is the probability of measuring the set of variables whose values are inside the n-dimensional box at $\langle \mathbf{y} \rangle + \mathbf{s_y}$ and with volume $\mathrm{d}\mathbf{s_y}$. The superscript ^{T} applied to vectors (and matrices) means "transpose" and makes rows into columns and vice versa. $\mathbf{R}$ is an $n \times n$ error covariance matrix, and is akin to the standard deviation in the single variable case. The combination $\mathbf{s_y}^T\mathbf{R}^{-1}\mathbf{s_y}$ is an inner product and yields a scalar.

If the errors in the observations are uncorrelated (independent), then $\mathbf{R}$ is diagonal and all elements are zero except for those along the diagonal; these elements are the variances, v_i, of the probability distribution functions of each measurement. Thus, if the parameter y_i is measured with standard error σ_i, then the element $R_{ii} = v_i = \sigma_i^2$. The contours of constant P are ellipsoids, and have principle axes, in the uncorrelated case, along the observation axes.[14] In the uncorrelated case the inner product can be written as simply,

$$-\frac{1}{2}\mathbf{s_y}^T\mathbf{R}^{-1}\mathbf{s_y} = -\frac{(y_1 - \langle y_1\rangle)^2}{2\sigma_1^2} - \frac{(y_2 - \langle y_2\rangle)^2}{2\sigma_2^2} - \ldots - \frac{(y_n - \langle y_n\rangle)^2}{2\sigma_n^2}. \tag{3}$$

All measurements are taken to have uncorrelated errors and so this simple form holds, although the compact notation (left-hand side of Eq. (3)) is maintained.

Each observation of the planet consists of an altitude/azimuth pair and $\mathbf{y}$ comprises $n/2$ such pairs as follows,

$$\mathbf{y} = \left\{ \begin{array}{c} \text{altitude of observable 1} \\ \text{azimuth of observable 1} \\ \text{altitude of observable 2} \\ \text{azimuth of observable 2} \\ \ldots \\ \text{altitude of observable } n/2 \\ \text{azimuth of observable } n/2 \end{array} \right\}, \tag{4}$$

where the altitude is the angle of the planet in the sky above the horizon at the time of each measurement, and the azimuth is its horizontal compass bearing. Other arrangements of the observations in $\mathbf{y}$ are possible, but must correspond with the diagonal elements of $\mathbf{R}$.

We let the set of model parameters (which are to be determined) be contained in the vector $\mathbf{x}$. For a Keplerian orbit, $\mathbf{x}$ has six components (see Sec. III for details), namely a, the length of the semi-major axis, e, the eccentricity, i, the inclination to the ellipse, Ω, the longitude of the ascending node, ϖ, the longitude of the perihelion, and ε, the mean longitude at a specific time called the epoch. We introduce the forward model symbolically as H. H represents the mathematical relations between the model parameters, $\mathbf{x}$ and the predicted observation values, $\mathbf{y^x}$. The input to H is a six-element vector, and its output is an n-element vector having the same structure as Eq. (4),

$$\mathbf{y^x} = H[\mathbf{x}]. \tag{5}$$

We assume that model errors are negligible. In some applications, this assumption is known formally as a "strong constraint."[15] The actual observations, $\mathbf{y}$, will be different from the predicted set, $\mathbf{y^x}$. Let the difference be the vector $\mathbf{s_y}$ (called the residual),

$$\mathbf{s_y} = \mathbf{y} - H[\mathbf{x}]. \tag{6}$$

Another probability distribution function is the conditional probability distribution function $P(\mathbf{y}^{\mathbf{x}}|\mathbf{x})$, that is, the probability of predicting the set of observations given the model state $\mathbf{x}$. This probability is found by substituting Eq. (6) into Eq. (2),

$$P(\mathbf{y}^{\mathbf{x}}|\mathbf{x}) \propto \exp\big(-\frac{1}{2}(\mathbf{y} - H[\mathbf{x}])^T \mathbf{R}^{-1}(\mathbf{y} - H[\mathbf{x}])\big), \tag{7}$$

which is called the likelihood function.

B. The cost function

The desired model parameters (the as yet unknown $\mathbf{x}$) are those that have maximum likelihood in Eq. (7). $P(\mathbf{y}^{\mathbf{x}}|\mathbf{x})$ is maximized when the scalar $\frac{1}{2}(\mathbf{y} - H[\mathbf{x}])^T \mathbf{R}^{-1}(\mathbf{y} - H[\mathbf{x}])$ is minimized. The inversion can then be posed as the variational problem of minimizing this scalar, called the cost function, $J[\mathbf{x}]$,

$$J[\mathbf{x}] = \frac{1}{2}(\mathbf{y} - H[\mathbf{x}])^T \mathbf{R}^{-1}(\mathbf{y} - H[\mathbf{x}]). \tag{8}$$

Minimization of the cost function is the rationale behind the method of least squares. It is called "least squares" because Eq. (8) is quadratic in $\mathbf{y} - H[\mathbf{x}]$, which can be seen easily for uncorrelated errors by expanding the matrix notation of Eq. (8) in a similar fashion to Eq. (3). The size of each error in R is important as each weights the impact of its associated observation, so a measurement with a large uncertainty contributes less to the cost of the misfit than one with a small error. The method can be extended to include explicit a priori knowledge of $\mathbf{x}$ by the use of Bayes' theorem[16] (we use no a priori information here).

$J[\mathbf{x}]$ may have a complicated structure and may, for example, have multiple minima. Minimization is made simpler if the forward model is linearized, and then the approximate forward model becomes an $n \times 6$ matrix. Let this matrix be $\mathcal{H}$, which has matrix elements of partial derivatives $\mathcal{H}_{ij} = \partial y_i^x / \partial x_j$. Such an element is the sensitivity of the predicted observation i to changes in the model parameter j, and can be evaluated either analytically or by finite differences. This matrix $\mathcal{H}$ is often wrtten in vector notation as $\mathcal{H} = \partial \mathbf{y}^{\mathbf{x}} / \partial \mathbf{x}$, which is known as a Fréchet derivative or as a Jacobian. In general, we can make the forward model linear by the approximation (linearizing about $\mathbf{x} = \mathbf{x}^0$),

$$H[\mathbf{x}] \approx H[\mathbf{x}^0] + \mathcal{H}(\mathbf{x} - \mathbf{x}^0), \tag{9}$$

where $H[\mathbf{x}^0]$ invokes the non-linear model acting on the linearization, $\mathbf{x}^0$. If we substitute Eq. (9) into the cost function, Eq. (8) and set $\delta\mathbf{y} = \mathbf{y} - H[\mathbf{x}^0]$ and $\delta\mathbf{x} = \mathbf{x} - \mathbf{x}^0$, we obtain an approximation, J_L, to the full cost function, J,

$$J_L[\delta\mathbf{x}] = \frac{1}{2}(\delta\mathbf{y} - \mathcal{H}\delta\mathbf{x})^T\mathbf{R}^{-1}(\delta\mathbf{y} - \mathcal{H}\delta\mathbf{x}) \approx J[\mathbf{x}^0 + \delta\mathbf{x}]. \tag{10}$$

This approximate cost function is quadratic, and so, unlike J, J_L will not possess multiple minima, but will have a shape similar to J near $\mathbf{x}^0$.

C. Minimizing the cost function

In general, minimizing functions in many dimensions is not easy, but provided that the problem is only moderately non-linear and has only a few unknown parameters, it can be broken down into many linear steps. To start the process, a guess of the solution is needed. In the absence of initial knowledge, a trial guess can be found by plotting some cross sections of J, Eq. (8), through six dimensional parameter space and choosing a value of each parameter close to the apparent global minimum of J.

In each iteration of the minimization process, the task is to find the vector $\delta\mathbf{x}$ that minimizes J_L. This is the point where the gradient of J_L with respect to every component of $\delta\mathbf{x}$ is zero. The derivative of J_L with respect to $\delta\mathbf{x}$ is written as $\nabla_{\delta x} J_L$ and is a six-component vector. The component j is the partial derivative $\partial J_L/\partial\delta x_j$. It is shown in the appendix that the gradient vector can be written as,

$$\nabla_{\delta x} J_L = -\mathcal{H}^T\mathbf{R}^{-1}(\delta\mathbf{y} - \mathcal{H}\delta\mathbf{x}). \tag{11}$$

Care must be taken when finding the residuals $\delta\mathbf{y}$ $(= \mathbf{y} - H[\mathbf{x}^0])$, components of which are angles. Because the domain of the angles is periodic, it is possible to obtain very large residuals even when the true residuals are small. For example, if, for one component $y_i = 358^\circ$ and $(H[\mathbf{x}^0])_i = 1^\circ$, the residual, $\delta y_i = y_i - (H[\mathbf{x}^0])_i$, should be -3° and not 357°. Thus residuals must be modified by adding $n360^\circ$ (integer n) where appropriate, such that all residuals lie between -180° and 180°.

The turning point of J_L is where $\nabla_{\delta x} J_L = 0$. It might be expected that the turning point is found by setting $\delta\mathbf{y} - \mathcal{H}\delta\mathbf{x} = 0$ in Eq. (11), and hence $\delta\mathbf{x} = \mathcal{H}^{-1}\delta\mathbf{y}$. This solution does not exist in general as the matrix $\mathcal{H}$ is not square (there are more observations than parameters). A sequence of matrices that *is* square, and can be inverted (as long as it is non-singular) is the Hessian matrix, $\mathbf{A}$,

$$\mathbf{A} = \mathcal{H}^T \mathbf{R}^{-1} \mathcal{H}, \tag{12}$$

which is the 6×6 matrix of second derivatives of J_L (see the appendix). This matrix arises naturally by setting the entire right-hand side of Eq. (11) to zero. Doing this and rearranging, we obtain the value of $\delta\mathbf{x}$ needed to minimize J_L:

$$\delta\mathbf{x} = \mathbf{A}^{-1} \mathcal{H}^T \mathbf{R}^{-1} \delta\mathbf{y}, \tag{13}$$

$$= -\mathbf{A}^{-1} \nabla_{\delta x = 0} J_L, \tag{14}$$

where in the last line, Eq. (11) has been used to write $\delta\mathbf{x}$ in terms of the gradient evaluated at the point of linearization, $\delta\mathbf{x} = 0$.

Note that, due to the non-existence of the inverse of $\mathcal{H}$ and $\mathcal{H}^T$, the substitution of Eq. (12) into Eq. (13) does not reduce Eq. (13) to $\delta\mathbf{x} = \mathcal{H}^{-1}\delta\mathbf{y}$. If this reduction were possible, the solution would be insensitive to the observation errors, and we would not need to construct a cost function. Equation (13) is essentially Newton's method,[17] approximated because the forward model has been linearized.

It is unlikely that the minimum of J_L, Eq. (10), is that of the true cost function J, Eq. (8), and so the process of linearizing and minimizing is repeated, but now about the new point in parameter space, $\mathbf{x}^0 \rightarrow \mathbf{x}^0 + \delta\mathbf{x}$. This cycle is terminated when the values of parameters have converged suitably. The vector that gives rise to $J_{\min}$, the minimum of the cost function, will be known as the analysis vector $\mathbf{x}^a$.

Although Eq. (13) gives $\delta\mathbf{x}$ explicitly for a particular iteration, $\mathbf{A}$ will not normally be inverted, as indicated here. We would instead treat Eq. (13) with $\mathbf{A}$ on the left-hand side and solve for $\delta\mathbf{x}$ by, for example, Gaussian elimination.

D. Some properties of the analysis

If the observation errors have been specified correctly in $\mathbf{R}$, then the value of J (Eq. (8)) evaluated at the minimum, $J_{\min}$. should have a value close to,

$$J_{\min} \sim (n - 6)/2, \tag{15}$$

for $n/2$ observation pairs and 6 parameters. This result requires that the observation errors are Gaussian and well characterized. If $J_{\min}$ is too large (small), then this is an indication that some or all of the errors in $\mathbf{R}$ are too small (large). This result arises from the fact that $2J_{\min}$ follows a χ^2 distribution which, in our case, has a mean value of $n-6$.[18]

Just as the observations have uncertainties, so will the analyzed parameters. The inverse Hessian, $\mathbf{A}^{-1}$, evaluated at $\mathbf{x}^a$ gives the covariance of the analysis errors.[19] The variances of each parameter are denoted by the expectation, $\langle(x_i^a - x_i^t)^2\rangle$ (where x_i^a is the ith parameter of $\mathbf{x}^a$ and x_i^t is the ith parameter of the known set of parameters[13], $\mathbf{x}^t$, where the superscript t stands for "truth"). These variances are contained in the diagonal elements of $\mathbf{A}^{-1}$. Our interest is in the standard deviations of the analyzed parameters, found from the square roots of these variances. The more observations that are included in the inversion, the smaller the analysis errors will be. Care must be taken because this interpretation of $\mathbf{A}^{-1}$ is only an estimate of the errors, and the following should be noted: (1) $\mathbf{A}^{-1}$ is a linear construct and is applicable to the error analysis only if the errors are within the range where the linearization, Eq. (9), is a good approximation. Otherwise the non-linear model will distort appreciably the Gaussian nature of the probability distribution function in the analysis. (2) The forward model error is ignored in this error analysis (we have assumed a perfect model throughout, which is not strictly valid).

We examine the validity of this use of $\mathbf{A}^{-1}$ in Secs. IV and V.

III. THE FORWARD MODEL

With an estimate of the orbital parameters for the planet in question, $\mathbf{x} = (a, e, i, \Omega, \varpi, \varepsilon)$, the forward model emulates the observations. The model requires further information such as the longitude, λ, and latitude, ϕ, of the observer on Earth, and the time, t, that each observation is made. All of our observations consist of pairs of altitude (h) and azimuth (A) measurements, although it is common to use right ascension (RA) and declination (dec). The forward model, Eq. (5), predicts each h/A pair ($\mathbf{x} \rightarrow (h, A)$, where (h, A) are two elements of $\mathbf{y}^{\mathbf{x}}$ – see Eq. (4)). It is a step-by-step algorithm which is now described (more details can be found elsewhere.[20]).

A. Position in the orbital plane

In the two body problem (assumed to be valid for planets), an orbit is a conic section, and we concentrate on the elliptical solutions. An orbit is shown in Fig. 1 to illustrate some of the orbital parameters. The position of the planet at a particular time in the plane of the ellipse (Q in Fig. 1) is a distance r from the Sun (the focus F), and an angle w from the perihelion ($w = \angle$PFQ is called the true anomaly). The perihelion P is the point on the orbit closest to the Sun.

The physics of planetary motion is described by Kepler's equation,[20] which derives from Newton's second law and an inverse-square central force. The parameters r and w are found by first solving Kepler's equation for the intermediate variable, E, called the eccentric anomaly,[21]

$$M = E - e\sin E. \tag{16}$$

M is called the mean anomaly (found below), and e is the eccentricity of the orbit. M is the angle swept out by a imaginary body in a circular orbit of radius a, the same as the semi-major axis length of the planet's orbit. The most useful expression for M for our purposes is,[22]

$$M = \sqrt{GM_s/a^3}(t - t_0) - \varpi + \varepsilon, \tag{17}$$

where G is the universal gravitational constant, M_s is the mass of the Sun (we assume $M_s \gg M_{\text{planet}}$), $t - t_0$ is the time elapsed because the epoch of midnight on 1st January 2000 UT (in our convention), ϖ is the longitude of the perihelion (defined as $\varpi \equiv \omega + \Omega$, even though ω and Ω are in different planes – see Fig. 1), and ε is the mean longitude (defined such that $-\varpi + \varepsilon$ is the orbit's mean anomaly at the epoch).

E can be found by solving Eq. (16) iteratively (for example, by Newton-Raphson iterations) and r and w follow from the trigometric identities,[21]

$$\tan\frac{w}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{E}{2}, \tag{18}$$

$$r = a(1 - e\cos E). \tag{19}$$

The heliocentric (Sun centered) position vector in the Cartesian system of the planet's orbit (thick arrows in Fig. 1), with right-handed unit vectors $\hat{\mathbf{x}}_p$ (defined in the direction FP), $\hat{\mathbf{y}}_p$ (defined orthogonal to $\hat{\mathbf{x}}_p$, but still in the orbital plane), and $\hat{\mathbf{z}}_p$ (perpendicular to the orbital plane) is,

$$\mathbf{r}_p = \begin{pmatrix} r\cos w \\ r\sin w \\ 0 \end{pmatrix}. \tag{20}$$

So far, four parameters $(a, e, \varpi, \varepsilon)$ have been used. These define the size and shape of the ellipse, the position of the planet in its orbital track, and partly define the orientation of the orbit in the orbital plane.

B. The position in the ecliptic plane

The vector $\mathbf{r}_p$ is given in the coordinate system of the orbital plane of the planet. A standard reference system used in astronomy is defined with respect to the ecliptic plane (the orbital plane of the Earth), which is still centered on the Sun. Three rotations of the axes give the planet's position in ecliptic co-ordinates, $\mathbf{r}_e$. The first is of angle ω (clockwise) about $\hat{\mathbf{z}}_p$ so that the new, intermediate x-axis lies along FN (Fig. 1). The second is of the inclination angle i about this x-axis so that the new z-axis points perpendicularly to the ecliptic plane, and the third is a clockwise rotation of Ω about this new z-axis,

$$\mathbf{r}_e = \begin{pmatrix} \cos\Omega & -\sin\Omega & 0 \\ \sin\Omega & \cos\Omega & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos i & -\sin i \\ 0 & \sin i & \cos i \end{pmatrix} \begin{pmatrix} \cos\omega & -\sin\omega & 0 \\ \sin\omega & \cos\omega & 0 \\ 0 & 0 & 1 \end{pmatrix} \mathbf{r}_p. \tag{21}$$

The ecliptic basis vectors $\hat{\mathbf{x}}_e$, $\hat{\mathbf{y}}_e$ and $\hat{\mathbf{z}}_e$ are defined in Fig. 1 (thin arrows), where $\hat{\mathbf{z}}_e$ is perpendicular to the ecliptic plane and $\hat{\mathbf{x}}_e$ is special as it points in the direction of the vernal equinox, which is a standard reference point for astronomical observations.

C. Conversion to geocentric equatorial co-ordinates

All six orbital parameters have now been used, but further transformations are required in order to predict observations made from the surface of the Earth. It is necessary to specify the planet's position relative to the Earth, and in the plane of the Earth's equator. To do this we first calculate, in a similar way as for the planet, the Earth's position in ecliptic co-ordinates, $\mathbf{r}_e^E$. (We repeat the steps in Secs. 3.1 and 3.2, this time using the known orbital parameters of the Earth.[13]) The position of the planet relative to the Earth in geocentric co-ordinates is,

$$\mathbf{r}_g = \mathbf{r}_e - \mathbf{r}_e^E. \tag{22}$$

This position vector is then projected onto another set of co-ordinates whose xy plane is coincident with the Earth's equator. The equatorial plane is the angle, $\bar{\varepsilon}$ (the obliquity of the ecliptic) from the ecliptic plane. The axis rotation of $\bar{\varepsilon}$ about the direction of the vernal equinox (the x-direction used to represent $\mathbf{r}_g$ and $\mathbf{r}_e$) gives the planet's position in geocentric equatorial co-ordinates, $\mathbf{r}_{ge}$,

$$\mathbf{r}_{ge} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\bar{\varepsilon} & -\sin\bar{\varepsilon} \\ 0 & \sin\bar{\varepsilon} & \cos\bar{\varepsilon} \end{pmatrix} \mathbf{r}_g, \tag{23}$$

where the basis vectors are $\hat{\mathbf{x}}_{ge}$, $\hat{\mathbf{y}}_{ge}$, and $\hat{\mathbf{z}}_{ge}$: $\hat{\mathbf{x}}_{ge}(=\hat{\mathbf{x}}_e)$ and $\hat{\mathbf{y}}_{ge}$ are in the Earth's equatorial plane, and $\hat{\mathbf{z}}_{ge}$ points along the north pole. At this stage it is possible to calculate directly, if required, the planet's RA (the angle from $\hat{\mathbf{x}}_{ge}$ of the planet's position projected onto equatorial plane), the dec. (the angle from the equatorial plane) and the distance from Earth ($|\mathbf{r}_{ge}|$).

D. Local horizontal co-ordinates

The measurements used in this article are altitude and azimuth (alt/azi) pairs and so a final transformation is required to co-ordinates convenient to the observer's horizon. The local axes are: $\hat{\mathbf{x}}_l$ (the direction of local east), $\hat{\mathbf{y}}_l$ (north) and $\hat{\mathbf{z}}_l$ (zenith) which are dependent on the observer's longitude, latitude and time. The relationship between these local axes and the geocentric equatorial co-ordinates is shown graphically in Fig. 2. The position vector of the planet in the local co-ordinates is $\mathbf{r}_l$,

$$\mathbf{r}_l = \begin{pmatrix} x_l \\ y_l \\ z_l \end{pmatrix} = \begin{pmatrix} \sin\lambda' & -\cos\lambda' & 0 \\ \sin\phi\cos\lambda' & \sin\phi\sin\lambda' & \cos\phi \\ -\cos\phi\cos\lambda' & -\cos\phi\sin\lambda' & \sin\phi \end{pmatrix} \mathbf{r}_{ge}. \tag{24}$$

Here, ϕ is the observer's latitude and λ' is the effective longitude (allowing for the Earth's rotation),

$$\lambda' = \lambda + 360^\circ \frac{t - t_{\mathrm{ref}}}{\Delta t_s}. \tag{25}$$

In Eq. (25), λ is the observer's longitude, $t - t_{\rm ref}$ is the time elapsed between some reference time, and Δt_s is the rotation time of the Earth in the same time units. Δt_s is the sidereal day length and is 23 hours, 56 minutes and 4.09 seconds. A convenient choice of $t_{\rm ref}$ is midnight on the day of the vernal equinox (of any year).

The altitude h and azimuthal A angles follow from the components of $\mathbf{r}_l$ by simple trigonometry,

$$h = \tan^{-1} \frac{z_l}{\sqrt{x_l^2 + y_l^2}}, \tag{26}$$

$$A = \tan^{-1} \frac{x_l}{y_l}, \tag{27}$$

where we need to add 180° to A if $y_l < 0$.

IV. TESTS WITH SYNTHETIC OBSERVATIONS

Before using real data, we test the inversion process by assimilating synthetic h/A observations of Venus. Synthetic observations are simulated from an application of the forward model with the known orbital parameters of Venus[13] (see Table I, row 1). The synthetic observations span a prescribed time period and random noise (according to a Gaussian distribution function) is added to the "observations." When fed back into the inversion, they provide a controlled self-consistency test of the inversion algorithm. Synthetic observations are made every month for two years (48 pieces of information). For the purposes of the inversion we try to recover the true parameters from the synthetic observations.

The results are shown in Table I, rows 2–5. The inversion process starts with a guess, $\mathbf{x}^0$ (row 2), which is found by plotting cross sections through the full cost function, Eq. (8), and choosing parameters close to the apparent global minimum. Although this is not the guaranteed global minimum, it is the simplest means of starting the iterations in the absence of a priori information. After refining this guess iteratively (using the procedure outlined in Sec. II), we obtain the results in row 3, which has uncertainty (row 4) estimated from the inverse Hessian. Row 5 gives the departure of the analysis from the truth (row 1) in units of standard deviations.

All of the parameters have converged, within error, to the true values, taking less than 10 iterations. The convergence of this inversion is shown in Fig. 3. The process terminates when $|\delta\mathbf{x}|$ is smaller than a defined cut-off, and the value of $J_{\rm min}$ is 23. The latter value is sufficiently close to the expected value of $21 = (48 - 6)/2$ (Eq. (15)). As explained in Sec. II D, this agreement indicates that the inversion is consistent with the synthetic observations and that the analyzed parameters are near optimal.

Figure 3 shows that the procedure does not reduce the value of the cost function monotonically. The increment $\delta\mathbf{x}$ prescribed by Eq. (13) will act to minimize a quadratic cost function. The true cost function is not exactly quadratic (due to the non-linear forward model), which has two consequences. The first, of course, is that the minimum must be obtained iteratively, and the second is that there is the possibility of overshooting the true minimum. In this case, and if the cost function is very steep at the new position, then it is possible to actually increase J from one iteration to the next. Presumably, such an increase has happened between iterations 4 and 5, although this increase is more than compensated for in the remaining steps.

V. INVERSION WITH REAL OBSERVATIONS

We have collected 14 h/A observations of Venus spanning 24 months and spaced irregularly in time, using a simple angular scale attached to an h/A telescope mount. At least three observations (six pieces of information) are required to estimate the six orbital parameters, and so, with 28 pieces of information, we expect the inversion to be well-behaved. Excellent convergence of J is found after nine iterations (not shown). In practice, convergence to $J_{\min} \sim 11 = (28 - 6)/2$ (see Eq. (15)) is achieved only if the observation errors are well characterized. We find that we can measure angles to accuracies $\Delta A \approx 1.5^\circ$ and $\Delta h \approx 1.2^\circ$ which yield a final value of $J = 13.4$, which is close to $J_{\min}$.

The last 4 rows of Table I summarize the results. Starting from the initial guess, $\mathbf{x}^0$ (row 6), running the inverse model gives a reasonable estimate, $\mathbf{x}^a$, (row 7) of the true parameters of Venus, $\mathbf{x}^t$ (row 1). Taking into account the errors in the analysis (row 8 found from the diagonal elements of the inverse Hessian, we find that all elements of $\mathbf{x}^a$ agree with $\mathbf{x}^t$ within two standard deviations (row 9).

The usefulness of the inverse model is better shown by plotting planetary trajectories in celestial co-ordinates (right ascension and declination, RA/dec). In Fig. 4a the trajectories over the period in which observations were accumulated (13 February 2001 to 15 February 2003) are plotted. The paths are projected onto celestial co-ordinates using the forward model (Sec. III) according to the analyzed parameters $\mathbf{x}^a$ (continuous line) and the true parameters $\mathbf{x}^t$ (dashed line). Also shown are the observations (diamonds) converted from h/A to RA/dec for comparison.

The closeness of the two lines in Fig. 4a demonstrates that our inverse model, if fitted to very simple observations, is useful and can do a reasonable job at capturing the true trajectory, including the two "retrograde loops" within this period. This result is not surprising because this is the same period of time that the observations were taken, and the parameters have been optimized to fit them. Another test of the usefulness of the technique is to use the analyzed parameters to make predictions. Similar trajectories based on $\mathbf{x}^a$ and $\mathbf{x}^t$, but for the period of 24 months following the last observation are plotted in Fig. 4b. This plot also shows reasonable agreement between our trajectory and true trajectory. The retrograde loop in this forecast is present, broadly at the right time, but this loop is less well captured than those in the observation period. If this plot were being used to plan future observations, it suggests that the single most important period of time to take observations is during the retrograde loop, because there seems to be a greater sensitivity of the planet's projected position to its orbital parameters at this time than at any other.

Although our analyzed values are reasonable, they do highlight some pitfalls with our inverse approach. The inverse technique assumes that the forward model is exact and linear, the measurements are unbiased, and all errors are known and obey Gaussian statistics. The last condition should then apply, for example, to the deviation $\mathbf{x}^a - \mathbf{x}^t$. In particular, it is probable that two thirds of the parameters will agree with the true parameters within one standard deviation. This requirement is not satisfied by our results (only one third agrees within one standard deviation), which means that either the parameters are sub-optimal or there is a breakdown of the assumptions listed above. In fact all the assumptions are violated to some extent. This is also the case in most real inverse applications, but as we have demonstrated, this does not mean that such inverse methods are not useful. A better test of our technique has already been presented in Sec. IV where the forward model is exact, the measurements are unbiased, and the measurement errors are known and made to obey Gaussian statistics. In this case the results are better.

VI. SUMMARY

An inverse modeling study such as one that we have presented provides training in a variety of practical and theoretical skills. The motion of a planet about the Sun is sufficiently easy to model and to measure, allowing real data to be used in combination with models. We have demonstrated that simple concepts and assumptions about the statistical nature of the measurements allow the values of the parameters to be inferred from the least squares-based inversion technique. The parameters would otherwise be difficult or impossible to measure directly.

There is much scope for developing this system for use as a student project.[23] Examples may include corrections for systematic biases in the measurements (for example, those caused by atmospheric refraction) or examining how the non-linearities in the forward model distort the Gaussian observation errors when propagated to the analysis error. Te latter idea could be studied by running an ensemble of inverse models (as a Monte-Carlo simulation), each with different perturbed observations that is consistent with the Gaussian probability density functions of the observation errors.

When used with real observations, the method requires a fairly long observational record, which would need to be of decades in duration for the outer planets. A different application of Kepler's equation for shorter period orbits might be for the orbital determination of the moon or even artificial satellites about the Earth. (The application of the two-body approximation of Kepler's equation is more restricted in these examples due to many-body perturbations and dipole components in the Earth's gravitational field.)

APPENDIX A: THE FIRST AND SECOND DERIVATIVES OF THE COST FUNCTION

A scalar such as J_L of Eq. (10) is a smooth function of each component of its argument $\delta\mathbf{x}$. Therefore, a gradient of J_L exists with respect to each component of $\delta\mathbf{x}$. This is Eq. (11), which we shall now demonstrate. If we expand Eq. (10) so that it is an explicit function of all vector components, we find,

$$J_L[\mathbf{x}] = \frac{1}{2}\sum_{mn}\big(\delta y_m - \sum_q \mathcal{H}_{mq}\delta x_q\big)(\mathbf{R}^{-1})_{mn}\big(\delta y_n - \sum_p \mathcal{H}_{np}\delta x_p\big). \tag{A1}$$

We differentiate Eq. (A1) with respect to an arbitrary element δx_i and obtain, by the product rule for differentiation,

$$\begin{aligned}\frac{\partial J_L}{\partial \delta x_i} &= -\frac{1}{2}\Big\{\sum_{mn}\big(\sum_q \mathcal{H}_{mq}\delta_{qi}\big)(\mathbf{R}^{-1})_{mn}\big(\delta y_n - \sum_p \mathcal{H}_{np}\delta x_p\big) \\ &\qquad -\sum_{mn}\big(\delta y_m - \sum_q \mathcal{H}_{mq}\delta x_q\big)(\mathbf{R}^{-1})_{mn}\big(\sum_p \mathcal{H}_{np}\delta_{pi}\big)\Big\}, \end{aligned} \tag{A2}$$

where the kronecker δ-functions arise from $\partial\delta x_q/\partial\delta x_i = \delta_{qi}$, etc. In the observation system discussed in this paper, $\mathbf{R}$ is diagonal, but this is not assumed in the above, making this proof more general. A covariance matrix, $\mathbf{R}$ and its inverse is always symmetrical, which means that both terms in Eq. (A2) are identical. Thus,

$$\frac{\partial J_L}{\partial \delta x_i} = -\sum_{mn}\mathcal{H}^T_{im}(\mathbf{R}^{-1})_{mn}\big(\delta y_n - \sum_p \mathcal{H}_{np}\delta x_p\big), \tag{A3}$$

where $\mathcal{H}^T_{im} = \mathcal{H}_{mi}$. This is the ith component of the vector of Eq. (11), thus proving this result.

The matrix element i, j of the Hessian is the second derivative $\partial^2 J_L/\partial\delta x_i\partial\delta x_j$. If we differentiate Eq. (A3) further with respect to δx_j, we obtain,

$$\frac{\partial^2 J_L}{\partial\delta x_i\partial\delta x_j} = \sum_{mn}\mathcal{H}^T_{im}(\mathbf{R}^{-1})_{mn}\sum_p \mathcal{H}_{np}\delta_{pj} = \sum_{mn}\mathcal{H}^T_{im}(\mathbf{R}^{-1})_{mn}\mathcal{H}_{nj}. \tag{A4}$$

If we associate each pair i, j with a matrix element, we find Eq. (12), thus showing this result for a linear operator $\mathcal{H}$.

ACKNOWLEDGMENTS

I am grateful to Profs. Alan O'Neill and Nancy Nichols of the University of Reading and Dr. Mark Dixon of the United Kingdom's Meteorological Office for advice with the manuscript. I would also like to thank two anonymous referees and Prof. Harvey Gould of the American Journal of Physics for suggesting many improvements.

---

[1] Mary L. Boas, *Mathematical Methods in the Physical Sciences* (John Wiley & Sons, New York, 1966), chapter 13.

[2] Albert Tarantola, *Inverse Problem Theory* (Elsevier, Amsterdam, 1994), 2nd ed., pp. XIII–XV.

[3] Cornelius Lanczos, *Applied Analysis* (Dover Publications, New York, 1988), chapter II.

[4] Thomas W. Schlatter, "Variational assimilation of meteorological observations in the lower atmosphere: a tutorial on how it works," J. Atmos. Solar Terr. Phys. **62**, 1057–1070 (2000).

[5] Douglas O. Gough, J. W. Leibacher, and P. H. Scherrer, "Perspectives in helioseismology," Science **272**, 1281–1283 (1996).

[6] D. W. Oldenburg, "An introduction to linear inverse theory," IEEE Transactions Geoscience Remote Sensing, GE-**22** (6), 665–674 (1984).

[7] Roger R. Bate, Donald D. Mueller, and Jerry E. White, *Fundamentals of Astrodynamics* (Dover Publications, New York, 1971), Sec. 2.1.

[8] Ref. 2, Sec. 4.1.

[9] L. G. Taff, *Celestial Mechanics: A Computational Guide for the Practitioner* (Wiley, 1985).

[10] Ref. 7, Sec. 2.12.

[11] A. Armitage, *Edmond Halley* (Thomas Nelson & Sons, London, 1966).

[12] E. T. Bell, *The Prince Of Mathematicians in The world of Mathematics* (Simon and Schuster, New York, 1956), Vol. 1.

[13] Parameters for the planets and other bodies are provided by NASA at `<http://ssd.jpl.nasa.gov/>`.

[14] If the observations are not independent, then the correlations are represented by additional off-diagonal elements in $\mathbf{R}$. Off-diagonal elements are called covariances and are a generalization of the standard deviation. See e.g., Roger J. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences* (John Wiley and Sons, Chichester, 1989), Sec. 2.6.3.

[15] Yoshikazu Sasaki, "Some basic formalisms in numerical variational analysis," Monthly Weather Review **98**, 875–883 (1970).

[16] Clive D. Rodgers, *Inverse Methods for Atmospheric Sounding: Theory and Practice* (World Scientific, 2000), Vol. 2, Sec. 2.3.2.

[17] William H. Press, Saul A. Teukolsky, William T. Vettering, and Brian P. Flannery, *Numerical Recipes:- The Art of Scientific Computing* (Cambridge University Press, 1986), Sec. 10.7.

[18] Ref. 14, Sec. 6.4.

[19] Ref. 2, Sec. 4.3.

[20] T. E. Sterne, *An Introduction to Celestial Mechanics* (Interscience Publishers, New York, 1960),

Chaps. 1 and 3.

[21] Ref. 20, Sec. 1.5.

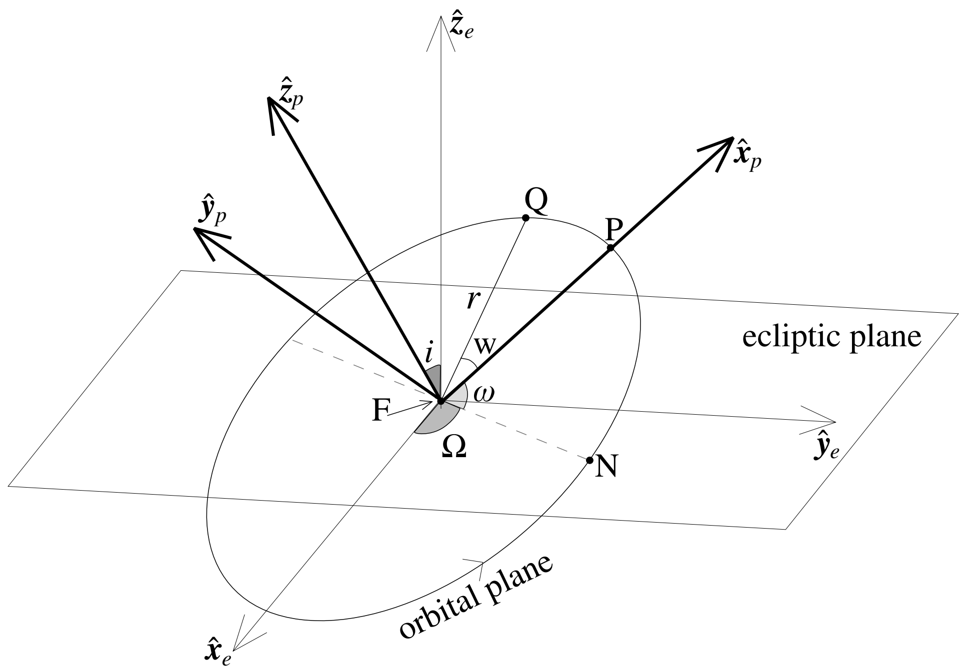
[22] Ref. 20, Sec. 3.2.

[23] Our code (in Fortran 77) and user documentation is available at <http://darc.nerc.ac.uk/models>.

		a (A.U.)	e	$i(^\circ)$	$\Omega(^\circ)$	$\varpi(^\circ)$	$\varepsilon(^\circ)$
Known parameters							
	1 $\mathbf{x}^t$	0.7233	0.0067	3.39	76.7	131.5	182.0
Synthetic observations							
	2 $\mathbf{x}^0$	0.7000	0.1000	5.00	60.0	210.0	150.0
	3 $\mathbf{x}^a$	0.7240	0.0073	3.54	75.0	153.4	182.6
	4 σ^a	0.0010	0.0042	0.29	4.2	35.7	0.8
	5 $\|\delta/\sigma^a\|$	0.7	0.1	0.5	0.4	0.6	0.8
Real observations							
	6 $\mathbf{x}^0$	0.7000	0.1000	5.00	90.0	100.0	110.0
	7 $\mathbf{x}^a$	0.7210	0.0201	4.23	88.9	110.0	176.6
	8 σ^a	0.0020	0.0078	0.70	8.1	50.3	6.8
	9 $\|\delta/\sigma^a\|$	1.2	1.7	1.2	1.5	0.4	0.8

TABLE I: Parameters pertaining to the synthetic and real inversions. In row 1 are the parameters that are regarded as the truth[13] which are compared to our results. The analysed parameters, optimized from synthetic alt/azi data of Venus, are in row 3. The first guess parameters, starting the iterations are shown beforehand in row 2. The uncertainties of the analysis, row 4, are for 1 standard deviation, and the relative error with respect to the true parameters are in row 5 ($\delta \equiv (\mathbf{x}^a - \mathbf{x}^t)$). Rows 6-9 are as rows 2-5, but for the inversion with real observations.

FIGURE CAPTIONS



Bannister, Fig. 1

FIG. 1: An elliptical orbit, and some of the parameters used to describe it. The natural coordinate systems of the planetary plane (thick unit vectors) and the ecliptic plane (thin unit vectors) are shown. F is the occupied focus (the Sun), P is the perihelion, N is the node where the ascending part of the orbit intersects the ecliptic plane, and Q is an arbitrary point on the orbit. Other symbols are described in the text.

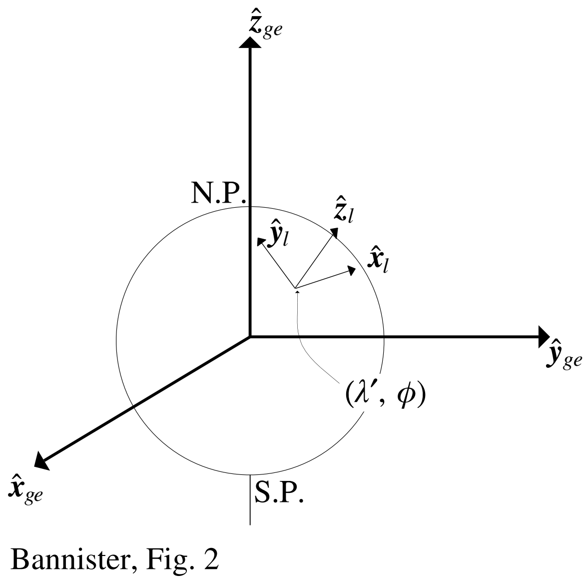


FIG. 2: The relationship between the geocentric equatorial coordinate system (thick unit vectors) and a local system (thin unit vectors centered on the surface of the Earth (disk) at effective longitude λ' and latitude ϕ).

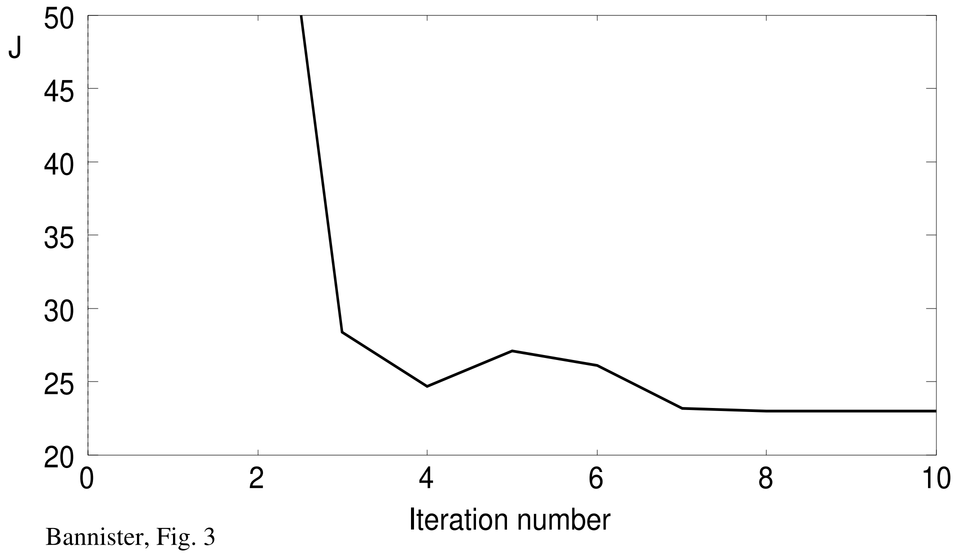


FIG. 3: The value of the cost function with each iteration of the inverse model used with synthetic "observations." The process shows good convergence after 10 iterations.

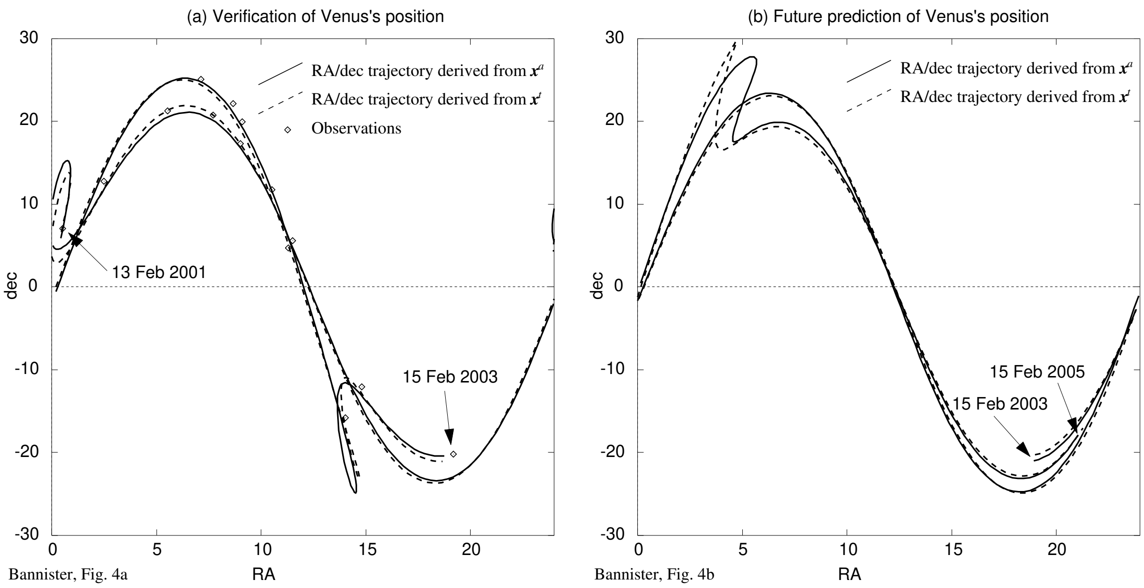


FIG. 4: Trajectories in celestial co-ordinates (RA/dec) derived from the analysed parameters (solid lines, $\mathbf{x}^a$) and "truth" (dashed lines, $\mathbf{x}^t$). Panel a is for the time period where observations were taken (February 2001 to February 2003) and panel b is for the subsequent 24-month period (February 2003 to February 2005). The observations shown in panel a (diamonds) have been converted from alt/azi to RA/dec. co-ordinates.