Statistical uncertainty and error propagation

Martin Vermeer

March 27, 2014

Introduction

This lecture is about some of the basics of uncertainty: the use of variances and covariances to express statistical uncertainty ("error bars", error ellipses); we are concentrating on the uncertainties of geographic co-ordinates. This makes it necessary to talk about geodetic *datums* – alternative ways of fixing the starting point(s) used for fixing geodetic co-ordinates in a network solution. It also makes it desirable to discuss absolute (single point) and relative (between pairs of points) co-ordinates and their uncertainties.

We follow up by discussing some more esoteric concepts; these sections are provided as reading material, but we will discuss them only conceptually, to get the ideas across. The learning objective of this is, that you will not be completely surprised if in future work, these concepts turn up; you will be somewhat prepared to read up on these subjects and use them in your work. The subjects are (in blue):

1. Criterion matrices for modelling spatial variance structures
2. Stochastic processes as a means of modelling time series behaviour statistically; signal and noise processes
3. Some even more esoteric subjects:
   a) Statistical testing and its philosophical backgrounds
   b) Bayesian inference
   c) Inter-model comparisons by information theoretic methods, the Akaike information criterion.

Contents

1 Uncertainty expressed in variances and covariances

In this text we discuss uncertainty as approached by physical geoscientists, which differs somewhat from approaches more commonly found in geoinformatics [Devillers and Jeansoulin, 2006, e.g.]. Central concepts are variances and covariances – the variance-covariance matrix – especially of location information in the form of co-ordinates. We shall elaborate in the chapters that follow.

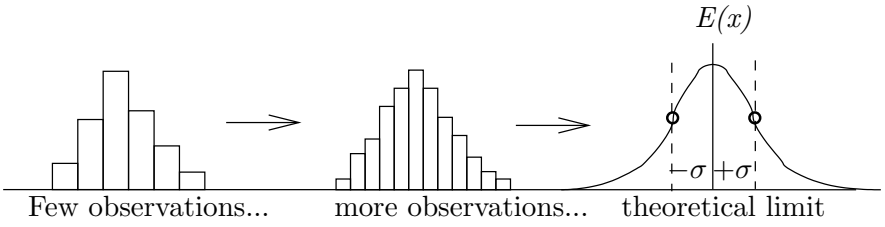
1.1 Definitions

We can describe statistical uncertainty about the value of a quantity, its random variation when it is observed again and again, by its *variance*. Similarly, we can describe the tendency of two quantities to vary randomly in somewhat the same way, as their *covariance*.

A *stochastic quantity* (random variate) is a method to produce *realizations* of a physical quantity. The number of realizations is in principle unlimited. E.g., throwing a die produces one realization of a stochastic process defined on the discrete domain $\{1, 2, 3, 4, 5, 6\}$. Throwing a coin similarly is defined on the discrete domain $\{0, 1\}$, where heads is 0, tails 1.

In spatial information, stochastic quantities typically exist on a *continuous domain*, e.g., the real numbers $\mathbb{R}$. E.g. measure a distance $d \in \mathbb{R}$. The measurements are $d_1, d_2, d_3, \ldots$ and the stochastic quantity is called $\underline{d}$ (underline).

A stochastic quantity has one more property: a *probability* (density) *distibution*. When doing a finite set of measurements, one can construct a *histogram* from those measurements.



Now when the number of measurements increases, the histogram will become more and more detailed, and in the limit become a smooth function. This theoretical limit[1] is called the stochastic quantity $\underline{d}$'s *probability density distribution function*, distribution

[1] We cannot ever *determine* this function from the observations, only obtain *approximations* to it, which will become better, the more observations we have at our disposal. In practice we *postulate* some form for the distribution function, e.g., the Gaussian or normal distribution, in which there are two *free parameters*: the expectancy μ and the mean error or standard deviation σ. These parameters are then *estimated* from our data.

function for short. It is written as $p(x)$, where x is an element of the domain of $\underline{d}$ (i.e., in this case, a real number, a possible measurement value).

Just like the total probability of all possible discrete outcomes, $\sum_{i=1}^{m} p_i = 1$, so is also the integral $\int_{-\infty}^{+\infty} p(x)\,dx = 1$. The integral

$$\int_a^b p(x)\,dx$$

again describes the probability of $\underline{x}$ lying within the interval $[a, b]$. If this probability is 0, we say that such an outcome is *impossible*; if it is 1, we say that it is *certain*.

A very common density distribution often found[2] when measured quantities contain a large number of small, independent error contributions, is the *normal* or *Gaussian* distribution ("bell curve"). It is the one depicted above titled "theoretical limit". The central axis of the curve corrsponds to the expectation $E(x)$, the two inflection points on the left and right slope (see Fig. 3) are at locations $E(x) \pm \sigma$, where σ is called the *mean error* or *standard deviation* associated with this distribution curve. The broader the curve, the larger the mean error.

1.2 Variances and covariances

Just like "distribution function" is the theoretical idealization of "histogram", we also define *expectancy* as the theoretical counterpart of mean/average value. If we have a set of measurements $x_i, i = 1, \ldots, n$, the mean is computed as

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Within this set, $1/n$ is the empirical probability, p_i, for measurement value x_i to occur if picked at random out of the set of n measurements (and note that $\sum_{i=1}^{n} p_i = 1$). So we may write

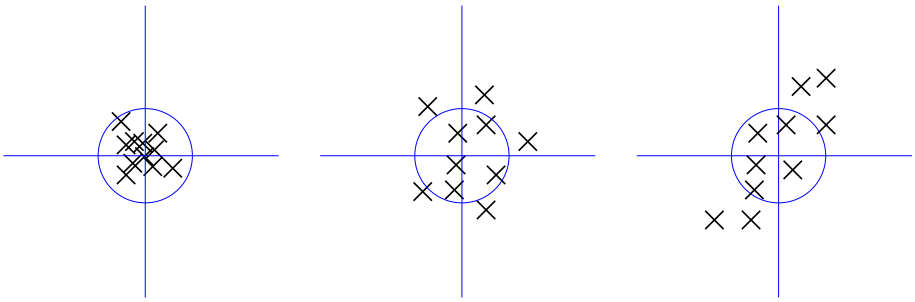
$$\overline{x} = \sum_{i=1}^{n} p_i x_i.$$

Now the theoretical idealization of this is an integral:

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

This is the *expectancy*, or expected value, of $\underline{x}$; the centre of gravity of the distribution function. It is the value to which the average will tend for larger and larger numbers of measurements.

[2] In fact, often when the amount of data is too small to clearly establish what the distribution function is, a normal distribution is routinely assumed, because it occurs so commonly (and is thus likely to be a correct guess), and because it has nice mathematical properties.



(a) Small random errors (b) Large random errors (c) Correlated errors

Figure 1: Different types of random error in a stochastic variable on $\mathbb{R}^2$. Left, small random error; middle, large random error. The picture on the right shows *correlation* between the horizontal and vertical random variables, both on $\mathbb{R}$.

Now defining variances and covariances is easy. Let us have a real stochastic variable $\underline{x}$ with distribution $p(x)$. Then its *variance* is

$$Var\left\{\underline{x}\right\} = E\left\{\left(\underline{x} - E\left\{\underline{x}\right\}\right)^2\right\}.$$

It describes, to first order, the amount of "spread", or *dispersion*, of the quantity around its own expected value, how much it is expected to deviate from this value.

Covariances are defined for *two* stochastic variables, $\underline{x}$ and $\underline{y}$:

$$Cov\left(\underline{x},\underline{y}\right) = E\left\{\left(\underline{x} - E\left\{\underline{x}\right\}\right)\left(\underline{y} - E\left\{\underline{y}\right\}\right)\right\}.$$

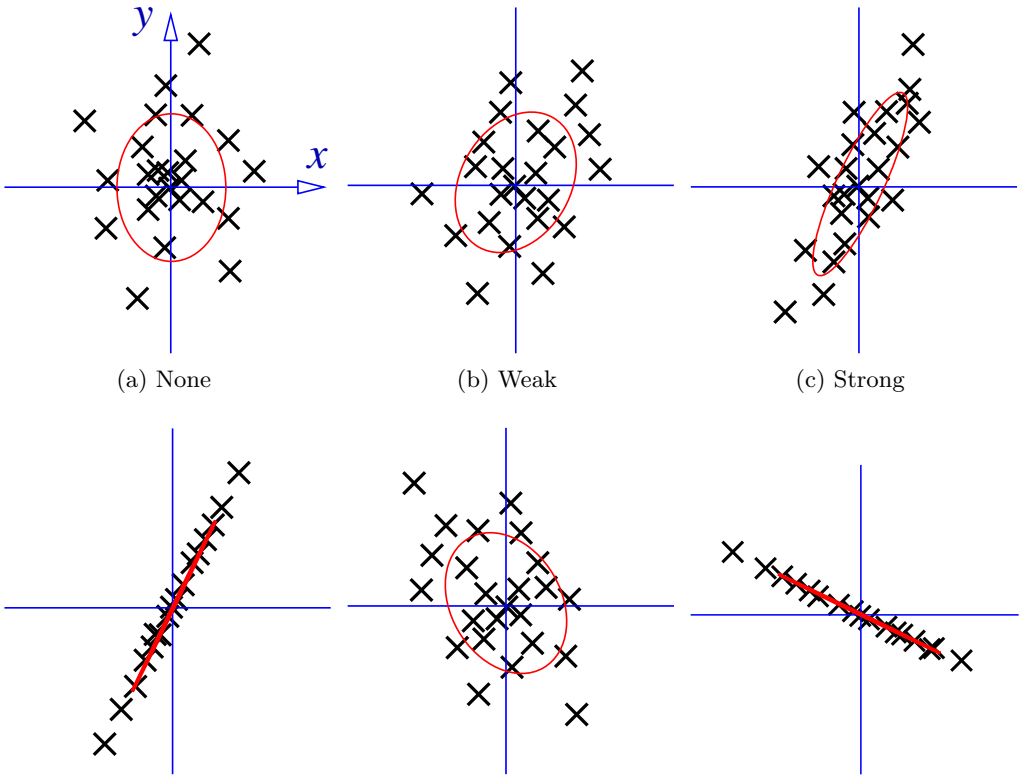
It describes to what extent variables $\underline{x}$ and $\underline{y}$ "co-vary" randomly, in other words, how likely it is, when $\underline{x}$ is bigger (or smaller) than its expected value, that then also the corresponding realization of $\underline{y}$ will be.

Once we have the covariance, we can also define the *correlation*:

$$Corr\left\{\underline{x},\underline{y}\right\} = \frac{Cov\left\{\underline{x},\underline{y}\right\}}{\sqrt{Var\left\{\underline{x}\right\}\ Var\left\{\underline{y}\right\}}}$$

Correlation is covariance *scaled* relative to the variances of the two stochastic variables considered. It is always in the range of $[-1,1]$, or $[-100\%,+100\%]$. If the correlation is negative, we say that $\underline{x}$ and $\underline{y}$ are *anticorrelated*.

Warning: *correlation doesn't prove causation!* Or more precisely, correlation doesn't tell us anything about what is the cause and what is the effect. Typically, when two stochastioc quantities correlate, it may be that one causes the other, that the second causes the first, or that both have a common third cause. The correlation as such proves nothing about this. However, sufficiently strong correlation (as established by statistical testing) is accepted in science a proof of the *existence* of a causal relationship.



(a) None (b) Weak (c) Strong

(d) 100% (e) Anticorrelation (f) -100%

Figure 2: Examples of correlations and anticorrelations

1.3 Error propagation

The expectancy operator $E[\cdot]$ is *linear*. This means that, if $\underline{u} = a\underline{x}$, then $E\{\underline{u}\} = aE\{\underline{x}\}$. This follows directly from the definition:

$$E\{\underline{u}\} = \int_{-\infty}^{\infty} up(u)\,du = \int_{-\infty}^{+\infty} axp(x)\,dx = aE\{\underline{x}\},$$

because $p(u)\,du = p(x)\,dx = dp$ refers to the same infinitesimal probability.

This propagation law can be extended to variances and covariances (with $\underline{v} = b\underline{y}$):

$$\begin{aligned} Var\{\underline{u}\} &= a^2\,Var\{\underline{x}\}, \\ Cov\{\underline{u},\underline{v}\} &= ab\,Cov\{\underline{x},\underline{y}\}. \end{aligned} \tag{2}$$

If a stochastic variable is a linear combination of two variables, say, $\underline{u} = a\underline{x} + b\underline{y}$, we get similarly (we also give the matric version):

$$Var\{\underline{u}\} = a^2\,Var\{\underline{x}\} + b^2\,Var\{\underline{y}\} + 2ab\,Cov\{\underline{x},\underline{y}\} =$$

$$
\begin{aligned}
&= \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} Var\left\{\underline{x}\right\} & Cov\left\{\underline{x},\underline{y}\right\} \\ Cov\left\{\underline{x},\underline{y}\right\} & Var\left\{\underline{y}\right\} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}, \\
Cov\left\{\underline{u},\underline{v}\right\} &= a\, Cov\left\{\underline{x},\underline{v}\right\} + b\, Cov\left\{\underline{y},\underline{v}\right\} = \\
&= \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} Cov\left\{\underline{x},\underline{v}\right\} \\ Cov\left\{\underline{y},\underline{v}\right\} \end{bmatrix}. \qquad (3)
\end{aligned}
$$

1.4 Co-ordinates, point errors, error ellipses

In spatial data, the quantities most often studied are *co-ordinates* of points on the Earth's surface. Spatial co-ordinates are typically three-dimensional; there is however a large body of theory and geodetic practice connected with treating location or map co-ordinates, which are two-dimensional. In this most common case we have co-ordinates modelled as stochastic (random) variables on the domain $\mathbb{R}^2$.

We give a point location as a pair of co-ordinates (x, y). Typically these are map co-ordinates, e.g., *kkj*, in some datum or co-ordinate reference system. As they are uncertain, we write them as $\left(\underline{x},\underline{y}\right)$ For brevity we may also write:

$$\mathbf{x} = (x, y)$$

indicating in one symbol both (generally: all three/one/two) co-ordinates. $\mathbf{x} \in \mathbb{R}^2$.

Now a given real world point co-ordinate pair will be *uncertain*. This means, it is describable by a stochastic variable on the domain $\mathbb{R}^2$. Also, the probability density distribution function p will then be a function of two real arguments, $p\left(x, y\right)$. See Fig. 3, depicting the two-dimensional Gaussian distribution.

Computing the variance of this two-dimensional quantity goes formally as follows:

$$\mathrm{Var}\left\{\underline{\mathbf{x}}\right\} = \begin{bmatrix} Var\left\{\underline{x}\right\} & Cov\left\{\underline{x},\underline{y}\right\} \\ Cov\left\{\underline{x},\underline{y}\right\} & Var\left\{\underline{y}\right\} \end{bmatrix} = \Sigma_{\mathbf{xx}} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}.$$

This is a symmetric matrix called the *point variance matrix* of the point $\mathbf{x}$.

This variance matrix can be depicted by an *ellipse* in the plane: the *error ellipse*. In Fig. 3 this error ellipse is depicted as the *horizontal* intersection curve of the Gaussian probability surface with a plane at the level of its inflection point. In one dimension, this corresponds to the two points at one standard deviation σ away from the central value, also at the inflection points of the curve.

The matrix $\Sigma_{\mathbf{xx}}$ defines an *eigenvalue problem*

$$\left(\Sigma_{\mathbf{xx}} - \lambda I\right)\mathbf{v} = 0;$$

this problem has two eigenvalues with associated eigenvectors. We find the eigenvalues by

$$\det\left[\Sigma_{\mathbf{xx}} - \lambda I\right] = 0,$$

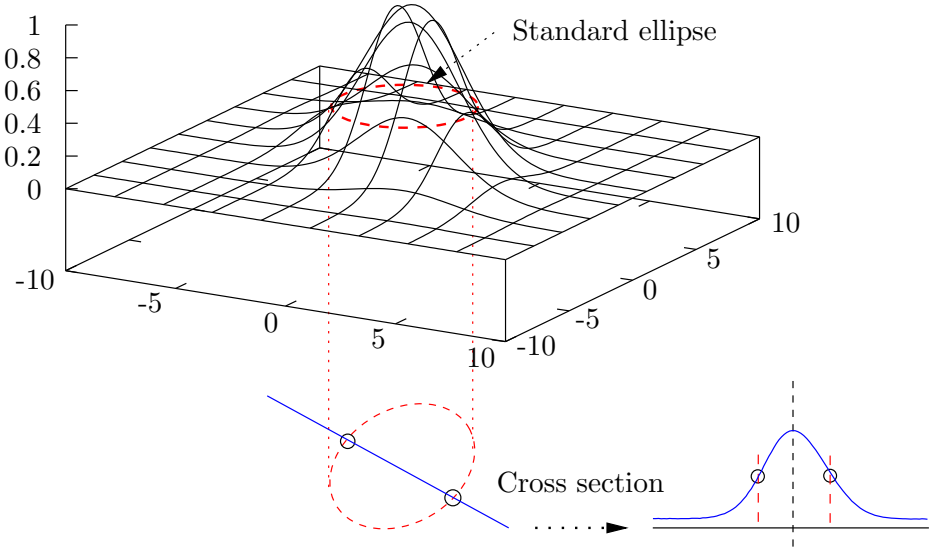


Figure 3: A two-dimensional Gaussian distribution function ("tropical hat"). A cross-section gives the uncertainty in a certain direction, its standard deviation corresponding to the intersections with the (one-sigma) error ellipse.

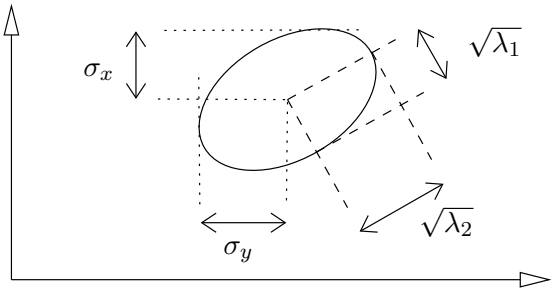


Figure 4: Parameters of the error ellipse.

or written out:

$$(\sigma_{11} - \lambda)(\sigma_{22} - \lambda) - \sigma_{12}^2 = 0,$$

a quadratic equation with two solutions $\lambda_{1,2}$. These solutions are the *principal axes* of the error ellipse, see the figure. The long semi-axis depicts the direction (eigenvector) of greatest uncertainty; the short semi-axis, that of the smallest uncertainty[3].

[3]For curiosity, we obtain for the eigenvalues

$$\begin{aligned}
\lambda_{1,2} &= \frac{1}{2}\left[\sigma_{11} + \sigma_{22} \pm \sqrt{(\sigma_{11} + \sigma_{22})^2 - 4(\sigma_{11}\sigma_{22} - \sigma_{12}^2)}\right] = \\
&= \frac{1}{2}\left[\sigma_{11} + \sigma_{22} \pm \sqrt{(\sigma_{11} - \sigma_{22})^2 + 4\sigma_{12}^2}\right] = \\
&= \frac{1}{2}(\sigma_{11} + \sigma_{22}) \pm \sqrt{\left[\frac{1}{2}(\sigma_{11} - \sigma_{22})\right]^2 + \sigma_{12}^2},
\end{aligned}$$

We have for the sum of the eigenvalues:

$$\lambda_1 + \lambda_2 = \sigma_{11} + \sigma_{22} = \sigma_x^2 + \sigma_y^2 = \sigma_P^2,$$

the *point variance*, a quantity describing the location precision of the point $\mathbf{x}$ in a co-ordinate-independent (*invariant*) fashion. In older literature the symbol m_P^2 is used for this. See also the figure. We denote the mean error of x by $\sigma_x = \sqrt{\sigma_{11}}$, that of y by $\sigma_y = \sqrt{\sigma_{22}}$. Also here, older literature has m_x and m_y, respectively.

1.5 Absolute and relative variances

In a situation where we have several points in the plane, also the variance-covariance matrix will become bigger. Also co-ordinates belonging to different points may have nonzero covariances between them. E.g., if we have two points $\mathbf{x}_1 = (x_1, y_1)$ and $\mathbf{x}_2 = (x_2, y_2)$, we have a variance-covariance matrix

$$\begin{aligned}\Sigma_{\mathrm{VV}} &= \\ &= \left[\begin{array}{cc|cc} Var\left\{\underline{x_1}\right\} & Cov\left\{\underline{x_1}, \underline{y_1}\right\} & Cov\left\{\underline{x_1}, \underline{x_2}\right\} & Cov\left\{\underline{x_1}, \underline{y_2}\right\} \\ Cov\left\{\underline{x_1}, \underline{y_1}\right\} & Var\left\{\underline{y_1}\right\} & Cov\left\{\underline{y_1}, \underline{x_2}\right\} & Cov\left\{\underline{y_1}, \underline{y_2}\right\} \\ \hline Cov\left\{\underline{x_2}, \underline{x_1}\right\} & Cov\left\{\underline{x_2}, \underline{y_1}\right\} & Var\left\{\underline{x_2}\right\} & Cov\left\{\underline{x_2}, \underline{y_2}\right\} \\ Cov\left\{\underline{y_2}, \underline{x_1}\right\} & Cov\left\{\underline{y_2}, \underline{y_1}\right\} & Cov\left\{\underline{x_2}, \underline{y_2}\right\} & Var\left\{\underline{y_2}\right\} \end{array}\right] =\end{aligned}$$

and the long and short semi-axes of the ellipse are $\sqrt{\lambda_1}, \sqrt{\lambda_2}$.

Also the *directions* of the axes can be obtained: study the linear combination of co-ordinates

$$z(\theta) = x \sin\theta + y \cos\theta,$$

which is a function of the direction angle θ.

According to propagation of variances we obtain

$$\mathrm{Var}(z(\theta)) = \sigma_{11} \sin^2\theta + \sigma_{22} \cos^2\theta + 2 \sin\theta \cos\theta \sigma_{12};$$

the axes of the ellipse are the *stationary values* of this function of θ,

$$\frac{d}{d\theta} \mathrm{Var}(z) = 0.$$

Differentiate:

$$2 \sin\theta \cos\theta (\sigma_{11} - \sigma_{22}) + 2\left(\cos^2\theta - \sin^2\theta\right)\Sigma_{12} = 0$$

i.e.,

$$\sin 2\theta (\sigma_{11} - \sigma_{22}) + 2 \cos 2\theta \cdot \sigma_{12} = 0$$

and

$$\begin{aligned}\theta &= \frac{1}{2} \arctan\left(-\frac{2\sigma_{12}}{\sigma_{11} - \sigma_{22}}\right) + k \cdot 90^\circ = \\ &= \arctan\left(-\frac{\sigma_{12}}{\sigma_{12} + \sqrt{\left[\frac{1}{2}(\sigma_{11} - \sigma_{22})\right]^2 + \sigma_{12}^2}}\right) + k \cdot 90^\circ,\end{aligned}$$

by using the *half angle formula* for the arc tangent. The integer k gives all the axes directions, which are 90° apart.

$$= \begin{bmatrix} \mathrm{Var}\left\{\underline{\mathbf{x}}_1\right\} & \mathrm{Cov}\left\{\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2\right\} \\ \mathrm{Cov}\left\{\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2\right\} & \mathrm{Var}\left\{\underline{\mathbf{x}}_2\right\} \end{bmatrix} = \begin{bmatrix} \Sigma_{\mathbf{x}_1\mathbf{x}_1} & \Sigma_{\mathbf{x}_1\mathbf{x}_2} \\ \Sigma_{\mathbf{x}_1\mathbf{x}_2} & \Sigma_{\mathbf{x}_2\mathbf{x}_2} \end{bmatrix}$$

for the four-dimensional abstract vector

$$\mathrm{v} = \left[\begin{array}{c} \underline{x}_1 \\ \underline{y}_1 \\ \hline \underline{x}_2 \\ \underline{y}_2 \end{array}\right] = \begin{bmatrix} \underline{\mathbf{x}}_1 \\ \underline{\mathbf{x}}_2 \end{bmatrix}.$$

This matrix is still symmetric. Note the "partitioned" way of writing this matrix.

1.6 Lifting point variances from a big variance matrix

Now looking at such a big matrix – and it can be *much* bigger! – we can simply lift the 2×2 matrix corresponding to a single point out of it, and it will be the point variance matrix of that point.

E.g., lifting the top left sub-matrix out of the above matrix gives

$$\mathrm{Var}\left\{\underline{\mathbf{x}}_1\right\} = \begin{bmatrix} Var\left\{\underline{x}_1\right\} & Cov\left\{\underline{x}_1, \underline{y}_1\right\} \\ Cov\left\{\underline{x}_1, \underline{y}_1\right\} & Var\left\{\underline{y}_1\right\} \end{bmatrix},$$

the variance-covariance matrix of point $\mathbf{x}_1$, corresponding to that point's error ellipse.

2 Relative location error by error propagation

2.1 Description

Rather often we are interested, not in the *absolute* precision of the location of a point $\mathbf{x}_1$, but rather in the *relative* location of that point in relation to another point $\mathbf{x}_2$. In other words, the precision of the *co-ordinate difference vector*

$$\underline{\mathbf{x}}_1 - \underline{\mathbf{x}}_2 = \begin{bmatrix} \underline{x}_1 - \underline{x}_2 \\ \underline{y}_1 - \underline{y}_2 \end{bmatrix}.$$

If we have the big matrix containing the two points $\mathbf{x}_1$ and $\mathbf{x}_2$, this problem is easy to solve, using a generalized (matric) version of the law of propagation of variances:

$$\mathrm{Var}\left\{\underline{\mathbf{x}}_1 - \underline{\mathbf{x}}_2\right\} = \mathrm{Var}\left\{\underline{\mathbf{x}}_1\right\} + \mathrm{Var}\left\{\underline{\mathbf{x}}_2\right\} - 2\mathrm{Cov}\left\{\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2\right\}.$$

(If we write $Var\left\{\underline{x}\right\} \rightarrow Cov\left\{\underline{x}, \underline{x}\right\}$, the logic of the equation is even more obvious: we have

$$\begin{aligned} Var\left\{\underline{x}_1 - \underline{x}_2\right\} &= Cov\left\{\underline{x}_1 - \underline{x}_2, \underline{x}_1 - \underline{x}_2\right\} = \\ &= Cov\left\{\underline{x}_1, \underline{x}_1\right\} - Cov\left\{\underline{x}_1, \underline{x}_2\right\} - Cov\left\{\underline{x}_2, \underline{x}_1\right\} + Cov\left\{\underline{x}_2, \underline{x}_2\right\}, \end{aligned}$$

as the covariance is by its definition linear in both arguments. This holds for $\underline{x_1}, \underline{x_2}$ being simple real-valued variables but just as well for them being vectors. In that case the variances and covariances become matrices.)

In this case we have to lift from the big matrix only *three* 2×2 sub-matrices. The two covariance sub-matrices off the diagonal are identical (or rather, each other's transpose), so we need only one of them (in fact, it is wise to store only the upper triangle of the big matrix, saving approx. half the storage space!)

2.2 Example (1)

Let us have a set of (many) points and associated co-ordinate variance-covariance matrix:

$$\mathrm{Var}\left\{\underline{\mathrm{v}}\right\} = \begin{bmatrix} \ddots & \cdot & \cdot & \cdot & \cdot \\ \cdot & \Sigma_{\mathbf{x}_i\mathbf{x}_i} & \cdots & \Sigma_{\mathbf{x}_i\mathbf{x}_j} & \cdot \\ \cdot & \vdots & \ddots & \vdots & \cdot \\ \cdot & \Sigma_{\mathbf{x}_i\mathbf{x}_j} & \cdots & \Sigma_{\mathbf{x}_j\mathbf{x}_j} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \ddots \end{bmatrix}.$$

Now we are interested in the *relative* variance-covariance matrix, and the corresponding relative error ellipse, *between* the two points $\mathbf{x}_i$ and $\mathbf{x}_j$. We obtain it as

$$\mathrm{Var}\left\{\underline{\mathbf{x}_i} - \underline{\mathbf{x}_j}\right\} = \mathrm{Var}\left\{\underline{\mathbf{x}_i}\right\} + \mathrm{Var}\left\{\underline{\mathbf{x}_j}\right\} - 2\mathrm{Cov}\left\{\underline{\mathbf{x}_i}, \underline{\mathbf{x}_j}\right\}.$$

2.3 Example (2): uncertainty in surface area from uncertain edge co-ordinates

This example is based on [van Oort, 2005] section 5.1.2, where the uncertainty in surface area of a parcel of land (described by a closed polygon) is expressed in the uncertainties of the edge node locations, as part of estimating the uncertainty in the value of the parcel.

We approach the problem as follows: express the surface area in *node co-ordinates* using the Gauss triangular equation:

$$A = \frac{1}{2}\sum_{i=1}^{n} x_i \left[y_{i+1} - y_{i-1}\right].$$

Here, it is assumed that $i = 0, 1, 2, \ldots, n-1, n$ is circular, so that point n is the same as point 0.

In preparation for formal error propagation, we must now *linearize*:

$$\begin{aligned} dA &= \frac{1}{2}\sum_{i=1}^{n}\left(\left[y_{i+1}-y_{i-1}\right]dx_i + x_i dy_{i+1} - x_i dy_{i-1}\right) = \\ &= \frac{1}{2}\sum_{i=1}^{n}\left(\left[y_{i+1}-y_{i-1}\right]dx_i + \left[x_{i-1}-x_{i+1}\right]dy_i\right), \end{aligned}$$

cleverly circularly renumbering the indices in the final two terms: $(x_{i+1} \rightarrow x_i, dy_i \rightarrow dy_{i-1})$ and $(x_i \rightarrow x_{i+1}, dy_{i-1} \rightarrow dy_i)$, which doesn't change the sum.

We now have to decide what uncertainties to assume for the co-ordinates x_i, y_i and especially what *correlations*, or covariances, between x_i and y_i for the same node point, and between x_i and x_j, and x_i and y_j, for different nodes.

In the dissertation referenced [van Oort, 2005], the following assumptions made are:

$$\begin{aligned} Var\left(x_i\right) &= \sigma_x^2; \\ Var\left(y_i\right) &= \sigma_y^2; \\ Cov\left(x_i, y_i\right) &= 0; \\ Cov\left(x_i, x_j\right) &= 0 \text{ if } i \neq j; \\ Cov\left(x_i, y_j\right) &= 0 \text{ if } i \neq j. \end{aligned}$$

Then, doing standard *propagation of variances* yields:

$$Var\left(A\right) = \frac{1}{4}\sum_{i=1}^{n}\left[y_{i+1}-y_{i-1}\right]^2\sigma_x^2 + \frac{1}{4}\sum_{i=1}^{n}\left[x_{i+1}-x_{i-1}\right]^2\sigma_y^2.$$

This equation is a little simpler than Eq. (9) in Section 5 of [van Oort, 2005] (page 62), because we have assumed (as we always do in geodesy!) that our co-ordinate uncertainties are *small* compared to the co-ordinate values – and even co-ordinate differences between the nodes of the polygon – themselves.

This approach can easily be generalized. Van Oort presents the case of multiple polygons with common points, e.g., adjoining lots with a common border. Then one finds for polygons p and q

$$Cov\left(A^p, A^q\right) = \frac{1}{8}\sum_{i=1}^{n_p}\sum_{j=1}^{n_q}\delta_{p_i q_j}\left[\begin{array}{l} \sigma_x^2\left(y_{i+1}^p - y_{i-1}^p\right)\left(y_{j+1}^q - y_{j-1}^q\right) + \\ +\sigma_y^2\left(x_{i+1}^p - x_{i-1}^p\right)\left(x_{j+1}^q - x_{j-1}^q\right) \end{array}\right],$$

where

$$\delta_{p_i q_j} = \left\{\begin{array}{l} 1 \text{ if point } p_i \text{ is the same as point } q_j \\ 0 \text{ otherwise} \end{array}\right. ,$$

1. Note that in this formula, the same point occurs twice, first as p_i, then as q_j. That explains the division by 8 instead of 4.

2. The above assumptions that

   a) σ_x^2, σ_y^2 are constants, and

   b) $Cov\,(x_i, x_j) = Cov\,(y_i, y_j) = 0$

   are probably unrealistic for points that are positioned during a single network measurement session, as is likely the case. They will contain *correlated co-ordinate errors*. A good approach to modelling these would be to construct a realistic looking, synthetic *criterion matrix*, as we shall discuss in Section 4.

3 Co-ordinate uncertainty, datum and S-transformation

3.1 What is a datum? Levelling network example

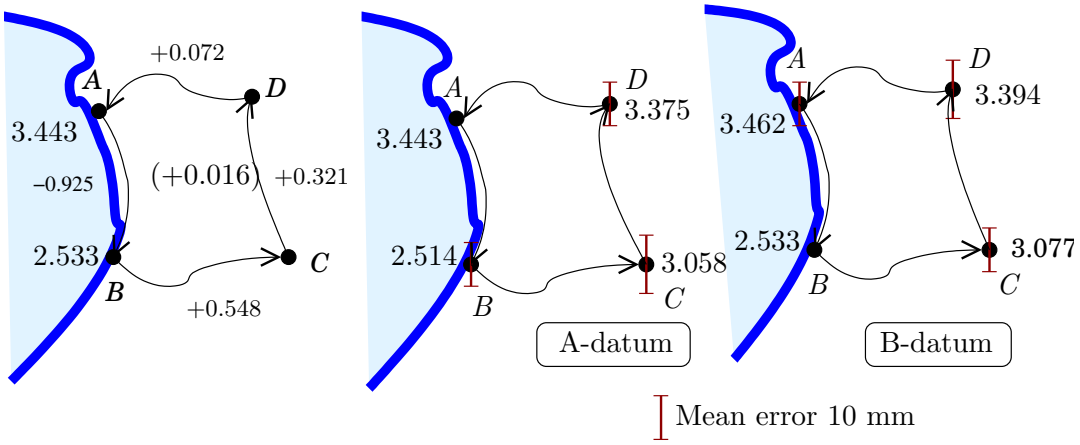
The old Finnish official height system N60 – now replaced by the newer N2000 – is based on the height of a fixed point in the garden of the Astronomical Observatory in Helsinki. This is a historical accident: Helsinki is the capital. It could just as well have been Turku.

Though the national precise levelling networks, the heights of N60 have been extended over all of Finland. Point height precisions therefore obviously will become poorer (i.e, uncertainties, like mean errors, will increase) with increasing distance to Helsinki. Clearly we know the height of Utsjoki poorer than that of Jyväskylä – relative to Helsinki. The height of Turku is also somewhat uncertain, because the measurements between Helsinki and Turku were somewhat imprecise.

On the other hand are the heights of points close to Helsinki all very precisely known, because the *datum point*, Helsinki Observatory, is so near.

Imagine for a moment that not Helsinki, but Turku were the capital, and that the Finnish height datum point was the mark in the wall of Turku's Cathedral. Then, all points close to Turku would be known very precisely in height, but those points close to Helsinki would be just as uncertain as in the present system, points close to Turku are: measurements from Turku to Helsinki are not absolutely precise.

Precision depends on your point of view: the chosen *datum*.



The picture shows a levelling network of four points. Given are the levelled height differences AB, BC, CD ja DA. Also given are the heights of coastal points A ja B from mean sea level, as measured by a *tide gauge*.

Firstly, we adjust the loop:

Interval	**Observed**	**Correction**	**Adjusted**
AB	-0.925	-0.004	-0.929
BC	+0.548	-0.004	+0.544
CD	+0.321	-0.004	+0.317
DA	+0.072	-0.004	+0.068
Closure error	+0.016		-

Next, two alternative procedures:

1. Use point A as the fixed or starting point and calculate point heights (the mean errors are made up, though realistic) :

Point	**Height**	**Mean error**
A	3.443	±0.000
B	2.514	±0.010
C	3.058	±0.014
D	3.375	±0.010

2. Same way but using point B for starting:

Point	**Height**	**Mean error**
B	2.533	±0.000
C	3.077	±0.010
D	3.394	±0.014
A	3.462	±0.010

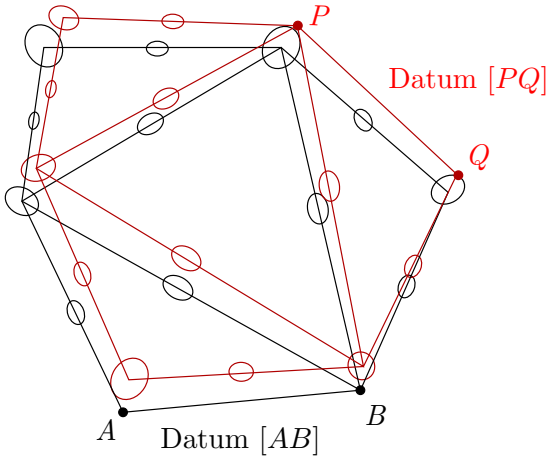


Figure 5: A horizontal control network calculated in two different datums, the AB- and the PQ-datum.

We see that in the latter case, all calculated heights are 0.019 m larger. The height *differences* are the same in both cases.

The *datum shift* between the A datum and the B-datum is 0.019 m.

3.2 Datum transformations

3.3 2D example

Different datums exist also in two dimensions, in the plane for location co-ordinates. In Fig. 5 are depicted an AB-datum and a PQ-datum, created by, in the calculation of the network, keeping either the points A and B, or the points P and Q fixed to conventionally adopted co-ordinate values. After that, the co-ordinates of the other points are calculated relative to these (The calculation may involve a network adjustment).

As we see, we get different results for different adopted starting point pairs or *datums*. Nevertheless the *shapes* of the network solutions look the same (and one can show that the two are connected by a Helmert similarity transformation).

Again, the choice of the co-ordinates of the starting points is somewhat arbitrary and conventional, a "formal truth". Making such a choice means defining a new geodetic datum.

Looking at the figure, you also see that the *error ellipses* change when choosing a different datum. The datum points themselves are exact, because they were chose so; other points are the more uncertain, the farther away they are from the datum points.

Also *relative* error ellipses are shown, on the lines between neighbouring points. As you see, these are a lot less sensitive to the choice of datum points! We sometimes say that

relative variances (the quantities depicted by these ellipses) are (quasi-)*invariant* for S-transformations.

3.4 What is an S-transformation?

An S-transformation is nothing but the similarity transformation that convert points in one datum to another datum, based on other fixed points in the same network; i.e., a *datum transformation*. As the two datums are usually very close together, the S-transformation *for the co-ordinates* will be close to unity. However, transforming point variances and covariances is a little more complicated: it is a transformation of small difference quantities and requires *linearization*. This transformation will not be close to unity. The formulas involved are quite complex, even if we describe co-ordinates as complex numbers as is common in flat geometry. We will therefore not present them.

4 Modelling location uncertainty by criterion matrices

4.1 About absolute and relative precision

Above we discussed absolute and relative variance-covariance matrices and error ellipses. When modelling location uncertainty in spatial point sets, we are usually interested most in the *relative* precision between *neighbouring* points. We want especially for this precision to become better (smaller) the closer the points are together.

The reason for this is that, if points are close together and their relative location is uncertain, they introduce *deformations* into any map or co-ordinate data set based on them. Say, a relative position uncertainty of $\pm 10\,\mathrm{cm}$ is not too bad if points are 1000 km apart; but if their distance is only 1 km, and they are both being used as the geometric basis for a bridge construction project, bad things may happen...

4.2 Precision in ppm and in $^{\mathrm{mm}}/\sqrt{\mathrm{km}}$

One can describe relative accuracy in a number of ways. Simplest is ppm (parts per million, or $^{\mathrm{mm}}/_{\mathrm{km}}$).

Especially for levelling, but also, e.g., for GPS networks, the practice has been to express precision in $^{\mathrm{mm}}/\sqrt{\mathrm{km}}$. This makes sense if the precision, or mean error, of a co-ordinate difference between two points in a network grows in proportion to the square root of the distance between the points.

For al levelling line, this holds: the variance of the height difference between start and end points grows linearly with the distance, and thus the mean error with its square root. So a four times longer line will have a twice poorer precision between end points.

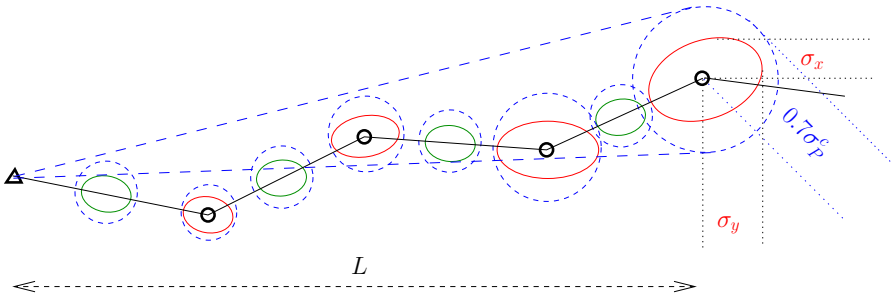


Figure 6: Criterion error ellipses for a polygon.

4.3 Modelling spatial uncertainty behaviour by criterion matrices

Often we will want to model the uncertainty behaviour or variance structure of a point field on the Earth's surface, or a set of network points, in a simple and geometrically realistic way, without however having access to – or wishing to use – the true variance-covariance matrix of these points. In that case, one often uses *criterion matrices*, synthetic covariance matrices designed to behave in this way.

In the figure is shown such a criterion variance structure of the ppm type in the case of a polygon. Note that criterion matrices behave in the same way as "real" variance-covariance matrices: they have error ellipses (in the figure's example, circles), etc. They also transform the same way under datum or S-transformations.

Often we use criterion matrices for *testing* or *replacing* real variance matrices. In both cases one should demand that the real variance matrix lies completely inside the criterion matrix, in some mathematically sensible sense[4]. Then also the true error ellipses will lie completely inside the criterion ellipses/circles (and due to the above mentioned transformation property, this property is *invariant* under datum transformations!)

This invariance is part of the generality which makes criterion matrices attractive for describing the "goodness" of a network or empirically positioned point set on the Earth surface. Co-ordinate precision measures will depend on the datum chosen, often quite arbitrarily, to describe them in; satisfying a given criterion matrix model will be independent of this choice.

5 The modelling of signals

5.1 Introduction

In the above, we have concentrated on the modelling of uncertainty as a property of stochastic quantities. The stochastic quantities considered were either geodetic obser-

[4] A condition on the maximum eigenvalue of an eigenvalue problem; this means that the n-dimensional hyper-ellipsoid of the criterion matrix completely encloses the corresponding hyper-ellipsoid of the true variance-covariance matrix. Her, n is the total number of point co-ordinates.

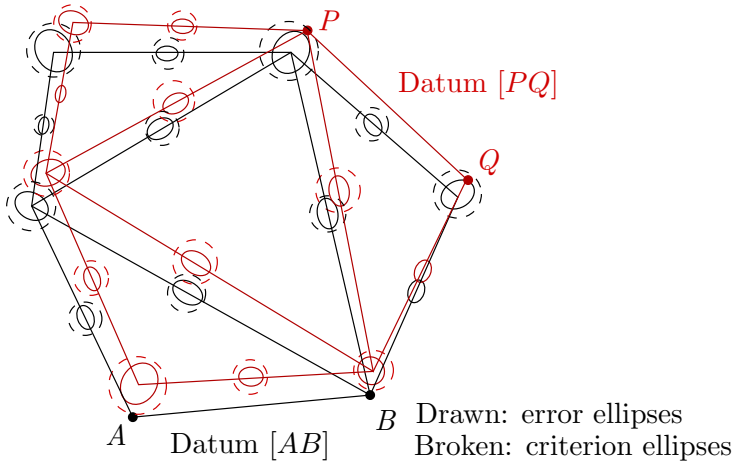


Figure 7: True and criterion variance-covariance matrices: both are S-transformation invariant: if the variance-covariance matrix lies inside the criterin matrix, than *all error ellipses*, absolute and relative, will lie inside their corresponding criterion ellipses *in all possible datums*!

vations, or quantities derived from geodetic observations, such as point co-ordinates on the Earth's surface. The stochasticity in this case is that of the uncertain *measurement process*: it is the stochasticity of observation *noise*, contamininating the quantities we really wish to study.

However, there are in spatial information science also quantities that are *intrinsically* stochastic, i.e., they can themselves be described as stochastic processes. As an example, we may mention gravity anomalies on the Earth's surface, who display a random-like behaviour with location when depicted on a map. This is *signal stochasticity*. We don't want to eliminate or minimise it; we wish to study and model it as well as we can.

Another situation where signal stochasticity occurs, is in *navigation*, where we model the randomness in the trajectory of a navigating vehicle or satellite by a stochastic process of time. Here, concepts like Markov processes and linear predictive filtering come into play. These techniques, while primarily designed for stochastic processes of time, can often be generated to the spatial information case of stochastic processes on the Earth surface or in the map plane.

For a more extensive discussion, see [Vermeer, 2008].

5.2 Stochastic processes

A *stochastic process* is a stochastic quantity the *domain*, or space of possible values, of which is a *function space*. An example of a stochastic process is the position in space $\underline{\mathbf{x}}(t)$ of a satellite; the stochasticity exists in that, next time around, the satellite, although

starting from the same position with the same velocity, will follow a slightly different trajectory due to variations in air drag etc.

A stochastic process typically is a function of time; this doesn't have to be the case, however. Gravity anomalies $\underline{\Delta g}(\varphi, \lambda)$ are a textbook example of a stochastic process that is a function of location on the Earth's surface.

Just as for stochastic quantites we have variance, covariance and correlation, so we have corresponding quantities here.

Variance:

$$C_{xx}(t) = Var\{\underline{x}(t)\}.$$

Autocovariance:

$$C_{xx}(t_1, t_2) = Cov\{\underline{x}(t_1), \underline{x}(t_2)\}.$$

Autocorrelation (always between -1 and +1):

$$\gamma_{xx}(t_1, t_2) = \frac{Cov\{\underline{x}(t_1), \underline{x}(t_2)\}}{\sqrt{Var\{\underline{x}(t_1)\} Var\{\underline{x}(t_2)\}}}.$$

Cross-covariance:

$$C_{xy}(t) = Cov\{\underline{x}(t), \underline{y}(t)\}$$

or

$$C_{xy}(t_1, t_2) = Cov(\underline{x}(t_1), \underline{y}(t_2)).$$

5.3 Covariance function

On the Earth surface, we may write the covariance function — in reality, the autocovariance function – of a stochastic process $\underline{x}(\varphi, \lambda)$ as follows:

$$C_{xx}(\varphi, \lambda, \varphi', \lambda'),$$

which represents the covariance between the gravity anomalies in two points: $\Delta g(\varphi, \lambda)$ and $\Delta g(\varphi', \lambda')$.

Now usually we assume that the gravity field of the Earth, like the Earth itself, is approximately spherically symmetric. This means we can postulate the following two properties.

homogeneity: this means that the statistical behaviour, specifically the covariance function, does not depend on absolute location.

isotropy: this means that the statistical behaviour, specifically the coveriance function, does not depend on the *direction* or *azimut* between the two points under consideration. It only depends on the geocentric spherical distance ψ.

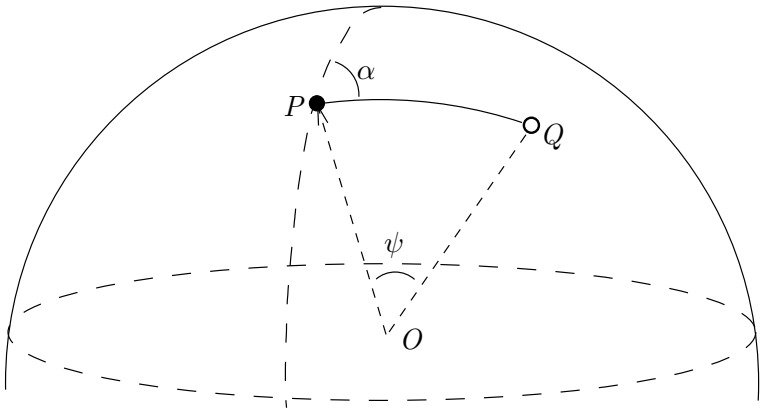


Figure 8: Spherical distance and azimut between two points P and Q.

$$C_{xx}(\psi)$$

where ψ is the geocentric angular distance between the points (φ, λ) and (φ', λ'). For the sphere we have, e.g.,

$$\cos\psi = \sin\varphi \sin\varphi' + \cos\varphi \cos\varphi' \cos\left(\lambda - \lambda'\right).$$

Now, for a quantity like gravity anomalies $\underline{\Delta g}(\varphi, \lambda)$, we may have, according to R. A. Hirvonen,

$$C_{\Delta g \Delta g} = \frac{C_0}{1 + \left(\psi/\psi_0\right)^2}, \tag{4}$$

where ψ_0 is called the *correlation length* and $C_0 = C_{\Delta g \Delta g}(0)$ the *variance* of the gravity anomalies. This is a computationally practical and in many areas of the Earth, rather realistic formula for describing the statistical behaviour of gravity.

For gravity anomalies, because of the way they are defined, we have furthermore

$$E\left\{\underline{\Delta g}\right\} = 0,$$

i.e., they are centered on zero.

5.4 Least-squares collocation

Measurements on the Earth surface are typically done in discrete points. As examples we may mention temperature or air pressure measurements at weather stations, sea level observations at tide gauges, gravimetric measurements in the field, sonar depth soundings at sea, etc. etc.

Our subject of interest, however, is what the value of the field being sampled is *everywhere*, not just in the finite number of points that have been sampled. This seemingly impossible problem can be solved, thanks to the fact that all these fields have a *covariance function*, which describes the statistical relationship of the field values in points that are close to each other. The values thus obtained are *estimates*, and contain uncertainty; but if the average separation of the measurement points is small compared to the field's correlation length, the uncertainty will be small.

As an example, given, e.g., gravity anomaly values $\underline{\Delta g_i}$ measured in a number of points $i = 1, 2, \ldots, n$, we may write up the variance-covariance matrix of these measured values as

$$C_{ij} = C(\psi_{ij}) = \begin{bmatrix} C_0 & C(\psi_{12}) & \cdots & C(\psi_{1n}) \\ C(\psi_{21}) & C_0 & \cdots & C(\psi_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ C(\psi_{n1}) & C(\psi_{n2}) & \cdots & C_0 \end{bmatrix}.$$

Note that this matrix describes the *physical behaviour* of the field of gravity anomalies in an area, the natural variation of gravity when moving from point to neighbouring point.

Additionally we also have to consider that the gravity values $\underline{\Delta g_i}$ were the result of *measurement*, and that this measurement was *uncertain*. We can describe this measurement uncertainty, or *dispersion*, or *variance*, by the measurement variance or dispersion matrix

$$D_{ij} = \mathrm{Var}\left\{\underline{\Delta g_i}, \underline{\Delta g_j}\right\}.$$

Typically

$$D_{ij} = \sigma^2 I = \sigma^2 \delta_{ij},$$

where $I = \delta_{ij}$ is the unit matrix or Kronecker delta, and σ^2 the variance of measurement of gravity anomalies (itself consisting of a number of error contributions). This is the common assumption that the measurements are 1) of equal precision for all points, and 2) uncorrelated because statistically independent between any two points.

If we now want to *predict* the gravity anomaly in a point P, we can do so in the following way. First, consider the variance matrix between the unknown point P and the known points i:

$$C_{Pi} = C(\psi_{Pi}) = \begin{bmatrix} C(\psi_{P1}) \\ C(\psi_{P2}) \\ \vdots \\ C(\psi_{Pn}) \end{bmatrix}.$$

This preparatory work having been done, we can now write the *estimator*:

$$\widehat{\Delta g_P} = C_{Pi} \left(C_{ij} + D_{ij}\right)^{-1} \underline{\mathrm{g}}_j,$$

where the *data vector*

$$\underline{g}_j = \begin{bmatrix} \underline{\Delta g_1} \\ \underline{\Delta g_2} \\ \vdots \\ \underline{\Delta g_n} \end{bmatrix}$$

is an abstract vector containing the measured anomaly values.

We can also get a value for the *error variance* of this estimate: the error variance of P is

$$\Sigma_{PP} = C_0 - C_{Pi} \left(C_{ij} + D_{ij} \right)^{-1} C_{jP}.$$

More extensive treatment of the situation where there are several unknown points P can be found in the literature.

This estimation technique is called *least-squares collocation* and it produces statistically optimal results. It is also a suitable interpolation technique in situations where spatial data is given only on a sparse, irregularly spaced set of points but would be needed, e.g., on a regular grid. A computational drawback is, that a matrix has to be inverted which is the same size as the number of data points given.

5.5 Semi-variances and kriging

Starting from the above HIRVONEN covariance function (4), we can compute the variance of the *difference* between two gravity anomalies in points $P$and Q, as follows:

$$\begin{aligned} Var\left\{\underline{\Delta g_P} - \underline{\Delta g_Q}\right\} &= Var\left\{\underline{\Delta g_P}\right\} + Var\left\{\underline{\Delta g_Q}\right\} - 2Cov\left\{\underline{\Delta g_P}, \underline{\Delta g_Q}\right\} = \\ &= 2C_0 - 2\frac{C_0}{1 + \left(\psi/\psi_0\right)^2} = \frac{2C_0 \left(\psi/\psi_0\right)^2}{1 + \left(\psi/\psi_0\right)^2} = \\ &= \frac{2C_0\psi^2}{\psi^2 + \psi_0^2}. \end{aligned}$$

In the situation where $\psi \ll \psi_0$, we get

$$Var\left\{\underline{\Delta g_P} - \underline{\Delta g_Q}\right\} \approx 2C_0 \left(\psi/\psi_0\right)^2.$$

On the other hand, for $\psi \gg \psi_0$, we get

$$Var\left\{\underline{\Delta g_P} - \underline{\Delta g_Q}\right\} \approx 2C_0.$$

We can identify *half* of this expression, $\frac{1}{2}Var\left\{\underline{\Delta g_P} - \underline{\Delta g_Q}\right\}$, with the *semi-variance* of $\underline{\Delta g}$. We also recognize ψ_0, or perhaps a few times ψ_0, as the "sill" at which the semi-variance levels off to the constant value C_0.

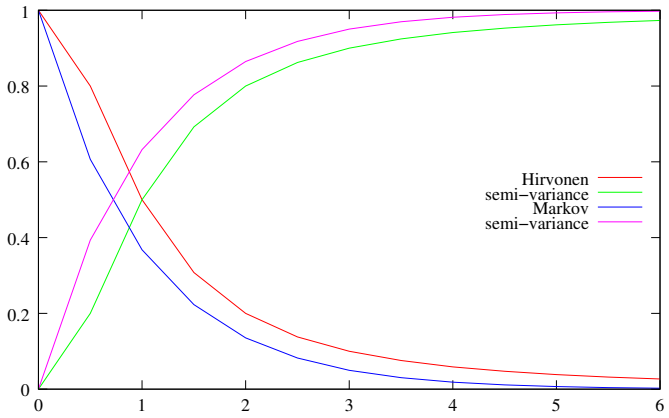


Figure 9: HIRVONEN's covariance function (for parameter values $C_0 = \psi_0 = 1$) and the associated semi-variance function; MARKOV's covariance function and associated semi-variance

For the alternative MARKOV covariance function (see below):

$$C(\psi) = C_0 e^{-\psi/\psi_0},$$

we get

$$\begin{aligned} \frac{1}{2} Var\left\{\underline{\Delta g_P} - \underline{\Delta g_Q}\right\} &= C_0 - C_0 e^{-\psi/\psi_0} = \\ &= C_0\left(1 - e^{-\psi/\psi_0}\right). \end{aligned}$$

Now, for $\psi \ll \psi_0$ this becomes $C_0\psi/\psi_0$, while for $\psi \gg \psi_0$ we obtain again C_0. Note the linear behaviour for small ψ, which differs from the quadratic behaviour of the HIRVONEN function and is typical for a "random walk" type process.

Kriging is a form of least-squares collocation described within this semi-variance formalism.

5.6 Markov processes

We often have a situation, where we want to generate, e.g., for simulation purposes, a *coloured noise* process of time t. A standard method for doing so, starting from an available white noise process $\underline{n}(t)$, is to use a MARKOV (or GAUSS-MARKOV) *process*[5].

First, let us explain what we mean by a white noise process. It is a process in which two function values for different epochs t_1 and t_2 are free of correlation. More precisely, it has an autocovariance function

$$C_{nn}(t_1, t_2) = Q(t_1)\,\delta(t_2 - t_1),$$

[5] A.A. MARKOV, Russian mathematician, 1856 – 1922

where Q is called the variance of $\underline{n}$. $\delta(\cdot)$ is the *Delta function*[6], for which holds:

$$\delta(\tau) = \begin{cases} \infty & \text{for } \tau = 0 \\ 0 & \text{for } \tau \neq 0 \end{cases}$$

$$\int_{-\infty}^{+\infty} \delta(\tau)\, d\tau = 1.$$

Now we can construct a (first-order) Markov process $\underline{x}$ by the following first-order differential equation:

$$\frac{d\underline{x}}{dt} = -k\underline{x} + \underline{n}. \tag{6}$$

The beauty of this is that the process can be generated sequentially, from a starting value $\underline{x}(t_0)$, integrating forward in time taking in values $\underline{n}(t)$ "on the fly".

One can show that the autocovariance function of such a MARKOV process, considered as a stationary process, looks like

$$C_{xx}(t_1, t_2) = \frac{Q}{2k} e^{-k|t_2 - t_1|}. \tag{7}$$

This function is depicted in Figure 11 for some values of k.

Figure 10: Andrey Andreyevich MARKOV

We can discretize the differential equation to describe the MARKOV process as a MARKOV *chain*:

$$\underline{x}_i = (1 - k\Delta t)\, \underline{x}_{i-1} + \underline{n}_i.$$

Like the process, also this chain can be computed sequentially though in index order i, where in every step, *only* knowledge of the previous index value is needed. One could say that a Markov chain is "memory free". The time step Δt here stands for $t_i - t_{i-1}$.

[6] Intuitively we can have a mental picture of how such a "function" is built.

First the following block function is defined:

$$\delta_b(\tau) = \begin{cases} 0 & \text{if } \tau > \frac{b}{2} \text{ or } \tau < -\frac{b}{2} \\ \frac{1}{b} & \text{if } -\frac{b}{2} \leq \tau \leq \frac{b}{2} \end{cases}$$

Obviously the integral of this function

$$\int_{-\infty}^{+\infty} \delta_b(\tau)\, d\tau = 1 \tag{5}$$

and $\delta_b(\tau) = 0$ if $|\tau|$ is large enough.

Now let in the limit $b \to 0$. Then $\delta_b(0) \to \infty$, and to every τ value $\tau \neq 0$ there is always a corresponding bounding value for b under which $\delta_b(\tau) = 0$.

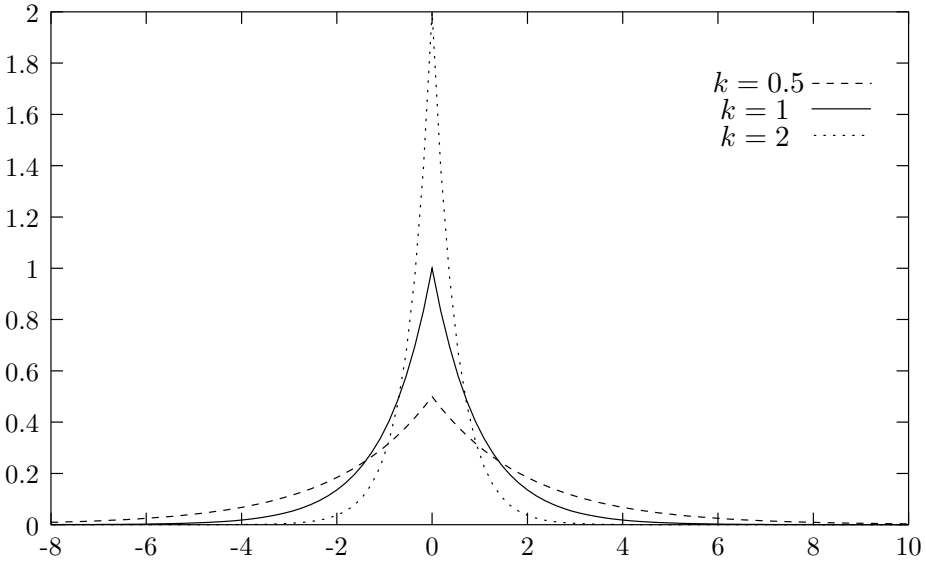


Figure 11: The autocovariance of a first-order MARKOV process.

Instead of sequentially, we can compute this in one go by putting all the $\underline{x}_i$ into one vector:

$$\underline{\mathrm{x}} = \begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \\ \vdots \\ \underline{x}_n \end{bmatrix}.$$

Now we can write the matric equation

$$\underline{\mathrm{x}} = (1 - k\Delta t)\, V\underline{\mathrm{x}} + \underline{\mathrm{n}},$$

where $\underline{\mathrm{n}}$ is an abstract n-vector of uncorrelated random values of known variance Q, and V is the "predecessor matrix" which looks like

$$V_{ij} = \begin{cases} 1 & \text{for } j = i - 1 \\ 0 & \text{otherwise} \end{cases}$$

In other words, it has one sub-diagonal of ones, shifted one place to the left from the main diagonal:

$$V = \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ & \ddots & \ddots & & \\ & & 1 & 0 & \\ & & & 1 & 0 \end{bmatrix}.$$

Looking at the differential equation (6), we can write it just as well in the inverted time

direction. Then we get

$$\frac{d\underline{x}}{d\,(-t)} = -k\underline{x} + \underline{n} \Rightarrow \frac{d\underline{x}}{dt} = k\underline{x} - \underline{n}.$$

The statistical properties, like the autocovariance function, are precisely the same as in the original. And as this function (7) is *symmetric*, we can add the two versions together and have still essentially the same autocovariance function!

The corresponding discrete matric equation is

$$\underline{\mathrm{x}} = (1 - k\Delta t)\, V^T \underline{\mathrm{x}} + \underline{\mathrm{n}},$$

where now V^T, the transpose of V, has ones only on the sub-diagonal one place to the right of the main diagonal.

Adding together the original and the time-reversed equation yields

$$\underline{\mathrm{x}} = (1 - k\Delta t)\, \frac{1}{2} W \underline{\mathrm{x}} + \underline{\mathrm{n}},$$

where now $W = V + V^T$, the *proximity matrix*, looks like

$$W = \begin{bmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & 1 & 0 & 1 & \\ & & 1 & 0 & 1 \\ & & & 1 & 0 \end{bmatrix},$$

i.e, it has 1 for neighbours, and 0 for non-neighbours. Luckily for numerical work, this matrix is very sparse.

Inverting this equation gives

$$\underline{\mathrm{x}} = \left[I - \frac{1}{2}\,(1 - k\Delta t)\, W \right]^{-1} \underline{\mathrm{n}}.$$

We can identify here the subexpression $(1 - k\Delta t)$ with ρ, the parameter describing the correlation strength.

This description was given in one dimension, time t; the whole procedure can be easily generalized to two dimensions, described, e.g., by two co-ordinates x and y. In that case, instead of two neighbours, the *predecessor* and the *successor*, every element x_i of the discretized process will have *four* direct neighbours. In this multidimensional context it is perhaps better to just forget about sequential evaluation.

Also this explanation was only about *first-order* Gauss-Markov processes, described by a first order differential equation. It can be generalized to second-order Gauss-Markov processes. Then many image-processing methods such as *Gaussian blur* can be described by it as well.

6 Modern approaches in statistics

For a more extensive and somewhat different discussion, see [Vermeer, 2008].

6.1 Statistical testing

You will all know how standard statistical testing is done: you formulate a hypothesis, you construct a confidence interval, and you test whether the variate of study lies within the interval ("the hypothesis is rejected") or outside it ("the hypothesis is accepted").

Typically, in the social sciences and humanities, the *significance level* of the test will be chosen, conventionally, equal to 95% – a value often called α. This is purely conventional, and rarely justified in the literature.

We will use as a textbook example a triangle in the plane where three angles α, β and γ have been measured. The hypothesis we want to test is "one or more of these angle measurements contains a gross error", and the testing variate is $\underline{\Delta} = \underline{\alpha} + \underline{\beta} + \underline{\gamma} - 180°$, which has a standard deviation of σ and will deviate significantly from zero if the tested hypothesis is true.

Some pertinent remarks:

1. You need, in addition to the hypothesis you are testing (the "alternative hypothesis" H_a), a *null hypothesis* you are testing against. This hypothesis has to make physical sense[7], i.e., it should be reasonably possible that it is true. In our example, the case that all three angles were measured without gross error is the null hypothesis, and it is certainly a reasonable one.
2. You need to set a *significance level* for testing. If we assume that Δ is normally distributed (it will be if the measurements $\underline{\alpha}, \underline{\beta}$ and $\underline{\gamma}$ are, and are statistically independent), then a 95% significance level corresponds to a (two-sided) *significance interval* of $\Delta \pm 1.96\sigma$. So, if $\|\Delta\| > 1.96\sigma$, we accept the alternative hypothesis ("there is a gross error in one or more of the angle measurements") and reject the null.
3. A 95% significance level corresponds to a p value of $p = 0.05$. Often the significance level of a test is stated as its p value, which can be extracted from the tabulated cumulative probability distribution (e.g., the normal distribution).

It is important to understand what the p value means: it states

> the probability that the testing variate Δ is as large or larger than the testing limit *in case the null hypothesis is true.*

What it does *not* state is

> ~~the probability that the alternative hypothesis is false.~~

Please take note!

[7] I.e., it should not be a "silly null". `http://www.ets.org/Media/Research/pdf/RR-01-24-Wainer.pdf`.

6.2 Philosophical background of statistical testing

The problem with classical statistical testing (R.A. Fisher, 1890-1962) as described above is, that it doesn't really give us what we are interested in. What we would really, *really* like to know, is "how probable is it that our alternative hypothesis is true?". And the test doesn't give us that. What it gives us is the probability, *given* that the null hypothesis is true, of seeing a value of Δ that leads to the null being rejected and the alternative hypothesis being accepted.

The latter reasoning ("the null is true $\rightarrow$ the probability of seeing a testing variate Δ at least this big is 5%") is called *forward inference*. The former reasoning, which we would like to do but cannot ("we see a testing variate Δ that is this big $\rightarrow$ we may conclude, with x% probability, that the alternative hypothesis is true"), is called *reverse inference*.

Why 95%? Please remember that this value is *choosable*. It is conventionally set to 95% in the social sciences, but this really is convention only.

In statistical testing we will make two different kinds of errors: errors of the first kind, and errors of the second kind.

1. Error of the first kind (type I error): we reject the null, although in reality it is true. The probability of a type I error occurring is $1 - \alpha = 1 - 0.95 = 5\%$.
2. Error of the second kind (type II error): we accept the null, although in reality it is false (i.e., in our example there is a gross error in one or more of the angle measurements). The probability of a type II error *depends on the size of the gross error considered*. If the gross error is small, it is likely to slip through in testing; if it is large, it is likely to get caught.

Figure 12: R.A. Fisher

The probability of catching a gross error *of given size* will *depend on the significance level chosen*. E.g.,

- For a significance level $\alpha = 95\%$, the probability of a gross error of size 3σ being caught is $\beta = 85.1\%$.
- For a significance level $\alpha = 99\%$, the probability of catching *the same* gross error is only $\beta = 66.4\%$. To reach the same detection probability of 85.1%, the gross error would have to be of size 3.615σ.

We thus conclude that the choice of significance level is an *optimization problem*, weighing the "cost" of type I errors – throwing away perfactly good observation data – against the "cost" of type II errors – delivering a result that is contaminated by undetected gross errors of a certain size, which will increase with increasing α.

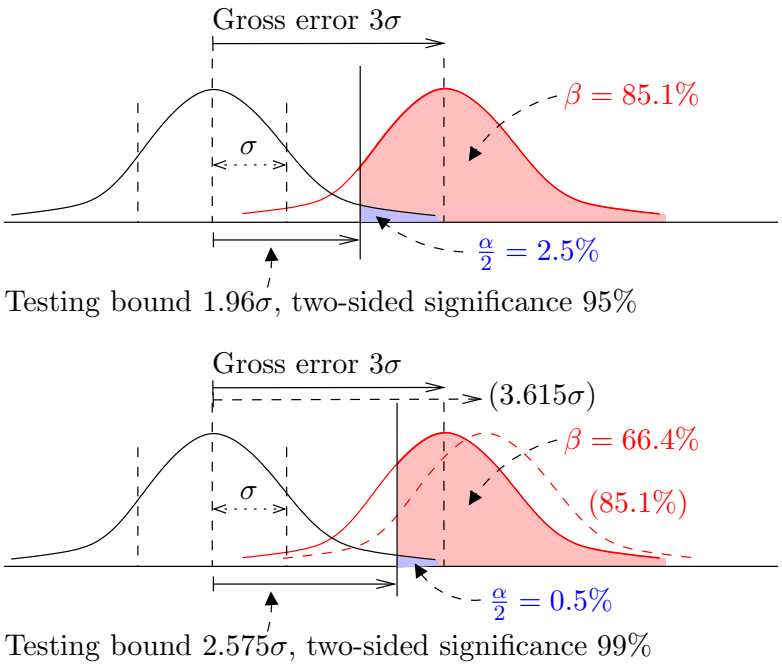


Figure 13: Statistical testing: significance level α and power β

In order to properly solve this optimization problem, we would have to know *how probable it is that gross errors of a certain size will occur*. We rarely if ever have this information, so-called *prior information*.

We can consider also other situations in which statistical testing is applied: e.g., detection of Earth crustal motions from geodetic measurements around the San Andreas fault line. In this case we have:

H_0**:** there have been no crustal motions along the fault line during the period of measurement

H_a**:** the Earth's crust has moved along the fault line.

Again, the test to be applied will *not* give us the probability that crustal motion took place; it will only tell us the probability that the test will give us a false alarm, i.e., detect crustal motion which in reality didn't happen, a type I error. Of course we want this probability to be small: false alarms are embarrassing and costly and undermine confidence. But we also do not want to miss a real crustal motion. In order to judge to probability of this happening, we must know the probability of such crustal motions in general, i.e., *prior information*.

6.3 Bayesian inference

Bayesian inference is a technique that allows us to do *reverse inference*, i.e., conclude from observations on the relative probabilities of various hypotheses that can explain the observational data. As we saw above, such reverse inference is only possible if we have *prior information* on the probabilities of null and alternative hypotheses.

The Bayesian approach to inference does not consider hypotheses at all. Rather, it estimates the joint *probability distribution* of the parameters of interest, given the observational data, and given the *prior information* we have, also in the form of a joint probability distribution, on what values these parameters are likely to assume. The source of this prior information may be, e.g., previously obtained and processed measurements, general physical considerations, the considered opinions of experts in the field, etc. etc.

The method goes back to the famous equation by the Reverend Thomas Bayes[8]:

$$P(\theta \mid x) = \frac{P(x \mid \theta)}{P(x)} P(\theta),$$

where θ is the set of unknown parameters and x is the set of observations.

What we are looking for is the distribution $P(\theta \mid x)$, the probability distribution of the parameters θ, given (on condition of) the observations x. What we have access to is $P(x \mid \theta)$, the probability distribution of the observations given (i.e., conditional upon) certain values of the parameters θ. This probability distribution is usually given in a form that contains the θ as unknowns, and can be computed for an assumed θ and the given observation set x.

Figure 14: Thomas Bayes (probably not!)

$P(x)$, the *a priori* probability distribution of the observations, acts as a normalizing constant here, just guaranteeing that all probabilities add up to 1.

$P(\theta)$ is the interesting thing here: it is the *prior*, or prior distribution assumed for the unknown parameters to be estimated.

So, to summarize:

1. Bayesian inference doesn't estimate parameter values or variances etc. at all; instead, it obtains a *posterior probability distribution* for the parameter set – from which, if so desired, expected values, most likely values, variances etc. can be obtained.

[8] England, 1702-1761

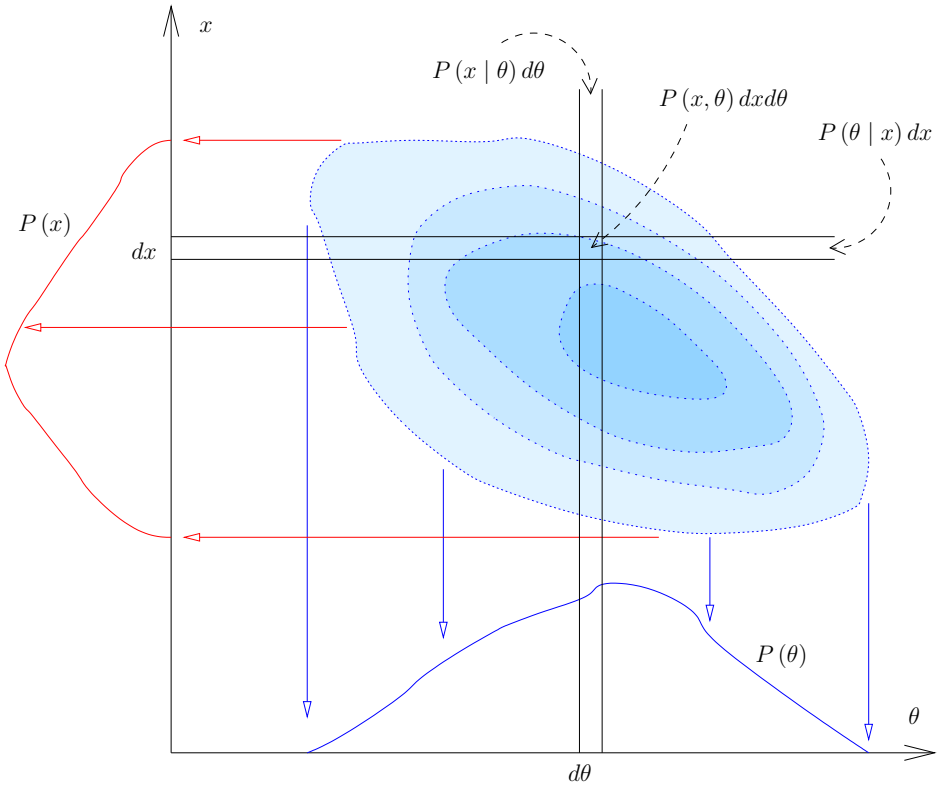


Figure 15: The continuous case of applying the BAYES theorem. This shows also how the integrated or "marginal" probabilities $P(x)$ and $P(\theta)$ are defined.

2. Bayesian inference does *reverse inference*, i.e., it actually provides actionable information on the parameter set of interest, given the observations. But
3. Bayesian inference requires that we have *prior information* on the parameter set of interest.

If it is not clear what would be appropriate prior information, and it appears that the result of Bayesian estimation depends sensitively upon the prior, we have a problem: *not enough observational information*. In that case, more observations should be obtained if possible; if not, the sensitivity of the result to the assumed prior should be studied and honestly reported.

6.4 Information theoretical methods

Often we may want to establish which one of a set of alternative models is the most appropriate one for describing a given set of observational data. Also here, often people reflexively turn to hypothesis testing, treating one of the alternative models as the null

and testing the others against it; or even worse, test each model against "nothing"[9]. More appropriate in this case would be to test models against each other without assuming one of them to be "privileged" as being the null.

Such methods are called *information theoretical methods of model selection*; the oldest, classical one is the AKAIKE Information Criterion (AIC), cf. [Burnham and Anderson, 2002].

$$AIC = 2k - 2\ln L,$$

where k is the number of model parameters or unknowns, and L is the value of the *likelihood function* for the model to be maximized.

Under the model assumption of normally and independently distributed observations, we may substitute the square sum of residuals for this:

$$AIC = 2k + n\left[\ln\frac{2\pi\sum_{i=1}^{n} v_i^2}{n} + 1\right],$$

where v_i are the residuals, and n the number of observations.

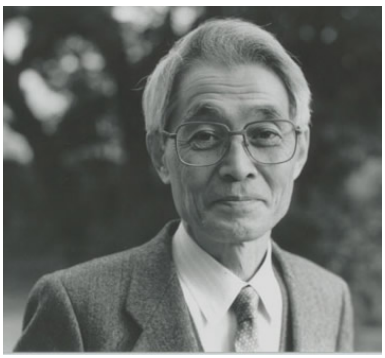


Figure 16: Hirotugu AKAIKE 1927-2009

Note that the information criterion imposes a *penalty* for introducing extra unknowns (parameter k): every extra unknown introduced makes it easier to achieve a good fit to the data[10]. The proper question to ask is, if it achieves a better fit *above and beyond* this trivial improvement due to simply having more free parameters to fit. The Akaike criterion is a was to more honestly answer this question, and to avoid "fitting an elephant".

Note that for small sample sizes there is a "corrected" version of the AKAIKE criterion, which in the limit for large sample sizes is equivalent to the above one:

$$AIC_C = 2k + n\ln\frac{\sum_{i=1}^{n} v_i^2}{n} + \frac{2k\,(k+1)}{n-k-1}.$$

This is the recommended formula for general use [Burnham and Anderson, 2002].

[9] Note that JAYNES [Jaynes, 2003, p. 504] quotes JEFFREYS: "Jeffreys (1939, p. 321) notes that there has never been a time in the history of gravitational theory when an orthodox significance test, which takes no note of alternatives, would not have rejected Newton's law and left us with no law at all. ..."

[10] Note that this corresponds to "OCCAM's Razor" or the Law of Parsimony: "entities must not be multiplied beyond necessity"

References

[Burnham and Anderson, 2002] Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*. Springer, Berlin.

[Devillers and Jeansoulin, 2006] Devillers, R. and Jeansoulin, R., editors (2006). *Fundamentals of Spatial Data Quality*. ISTE, London.

[Jaynes, 2003] Jaynes, E. (2003). *Probability theory: the logic of science*. Cambridge University press. ISBN 0-521-59271-2.

[van Oort, 2005] van Oort, P. (2005). *Spatial data quality: from description to application*. PhD thesis, Wageningen University, Wageningen, The Netherlands. URL: `http://library.wur.nl/wda/dissertations/dis3888.pdf`.

[Vermeer, 2008] Vermeer, M. (2008). Covariance analysis for geodesy and geophysics. Lecture notes, NKG Summer School, Nesjavellir, Iceland, 25-28 Aug. Landmaelingar Islands, URL: `http://www.lmi.is/Files/Skra_0029262.pdf`.