Aalto University

OPEN ACCESS

This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Vermeer, Martinus

Covariance analysis for geodesy and geophysics

Published: 01/01/2008

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Vermeer, M. (2008). *Covariance analysis for geodesy and geophysics*. (Lecture notes, NKG Summer School, Nesjavellir, Iceland, 25-28 Aug. Landmaelingar Islands).

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Covariance analysis for geodesy and geophysics

Martin Vermeer

Contribution for the NKG Summer School in Nesjavellir, Iceland, August 25-28, 2008

Contents

1 Introduction **3**

2 Error propagation and criterion matrices for GNSS networks **4**
2.1 Geocentric variance structure of a GPS network 4
2.2 Combining spatial and temporal dependency 6

3 Autocorrelation in time series studies **7**
3.1 Time series and white noise 7
3.2 "Coloured noise", Gauß-Markov (autoregressive) process 9
3.2.1 Power spectral density of a Gauß-Markov process 11
3.3 AR(1), linear regression and variance 12
3.3.1 Least squares regression in absence of correlation 12
3.3.2 AR(1) process 14
3.4 Various autoregressive models 15
3.4.1 Classification 16
3.4.2 Fractional Gaussian noise 17
3.4.3 Durbin-Watkins (DW) test 18
3.4.4 The r^2 test and its limitations 19
3.5 The Hurst exponent and Hurst rescaling 21
3.5.1 Computing the Hurst exponent 21
3.5.2 Hurst rescaling 22
3.6 Pink noise 24
3.7 Data points per degree of freedom 25
3.7.1 Maximum entropy method 27
3.7.2 Bayes, Akaike info criteria 28
3.8 CASE I: computing climate sensitivity from the oceans' heat capacity 28
3.9 AR(1) vectorial case 30
3.10 CASE II: computing climate sensitivity from temperature and radiative balance . . . 32

4 Slings and arrows of Bayesian inference **36**
4.1 An example 36
4.2 Some philosophy 37
4.3 Continuous-valued parameter and observation spaces 37
4.4 Where do you get your prior? 38
4.5 Some more philosophy 39
4.6 Another example 40
4.7 Modelling a simple observation process 42

1 Introduction

In these lecture notes we will investigate some of the lesser trodden paths of variance-covariance modelling, which are often used both in- and outside geodesy and geophysics, but which may be of more relevance and use to us than many may realise.

1. The use of formal covariance structures in the form of *criterion matrices* for describing the covariance behaviour of satellite positioning based geodetic networks
2. The modelling and parameter estimation of time series characterised by strong, or long-range, temporal autocorrelation
3. Bayesian estimation and inference, which in many ways is counterintuitive.

I hope to present some interesting items that may inspire the reader to deeper investigation. For this purpose a bibliography is provided. The material presented is rather heterogeneous, due to the generality of the given commission. Note also that much of the subject matter and examples are taken from climatology, which is not my main specialism. This is not bad, as a scientist is supposed to continue learning all his life, and not only within his own little acre.

So this will be a discussion opportunity as much as a teaching event.

There is no PowerPoint$^{\text{TM}}$ version of these lecture notes. There has been some discussion[1] on the suitability of this kind of presentation software for technical reports, and one may similarly question its suitability for teaching.

[1] As the Columbia Accident Investigation Board found after the loss of Space Shuttle Columbia:

1. PP is an inappropriate tool for engineering reports, presentations, documentation; and
2. the technical report is superior to PP. Matched up against alternative tools, PowerPoint loses.

(Source: Edward Tufte, [Tuf05].)

2 Error propagation and criterion matrices for GNSS networks

It is well known that if we collect geodetic GNSS data over long time periods, that thinning out the data from the typical 30 s to even several minutes does not appear to compromise the actual content of valuable information of the data. This is due to the strong *autocorrelation* between GNSS observations collected at successive intervals.

Another practical experience is that the *spatial* covariance structure is not influenced by the presence of temporal autocorrelation. The effect is limited to producing too optimistic point and inter-point variances and covariances, which are "scaled" by a constant factor.

As to this spatial covariance structure, a fruitful approach to modelling, or approximating, it in a general way, abstracting from the details of network structure and measurements, is by a *criterion matrix*.

As to the practical uses of criterion matrices, there are two of them:

1. If the true variance-covariance matrix of an old network is not available, a criterion matrix generated for realistic precision assumptions may often be a good replacement; and
2. one may use a criterion matrix as a precision *requirement* for a measurement project, requiring the variance-covariance matrix to be enclosed by it; after meeting the requirement, again the criterion matrix may be used in its stead, as its summary, as it were.

2.1 Geocentric variance structure of a GPS network

Let us first derive a rough but plausible, geocentric expression for the variance-covariance structure of a typical geodetic network. The true error propagation of GPS measurements is an extremely complex subject. Here, we try to represent the bulk co-ordinate precision behaviour in a simple but plausible way.

Also the full theory of criterion matrices and datum transformations is complicated [Baa73]. Here we shall cut some corners. We assume that the inter-point position variance between two network points A and B, co-ordinates (X_A, Y_A, Z_A) and (X_B, Y_B, Z_B), is of the form

$$\begin{aligned} &\mathrm{Var}\,(\mathbf{r}_B - \mathbf{r}_A) = \\ &= Q_0 \left((X_B - X_A)^2 + (Y_B - Y_A)^2 + (Z_B - Z_A)^2 \right)^{\frac{k}{2}} \\ &= Q_0 d_{AB}^k, \end{aligned} \tag{2.1}$$

with k and Q_0 as the free parameters (assumed constant for now), and $d_{AB} = \|\mathbf{r}_B - \mathbf{r}_A\|$ the $A-B$ inter-point distance.

For this to be meaningful, we must know what is meant by the variance or covariance of vectors. In three dimensions, we interpret this as:

$$\mathrm{Cov}\left(\mathbf{r}_A, \mathbf{r}_B\right) = \mathrm{Cov}\left(\begin{bmatrix} X_A \\ Y_A \\ Z_A \end{bmatrix}, \begin{bmatrix} X_B \\ Y_B \\ Z_B \end{bmatrix}\right) =$$

$$= \begin{bmatrix} \mathrm{cov}\left(X_A, X_B\right) & & \\ & \ddots & \\ & & \mathrm{cov}\left(Z_A, Z_B\right) \end{bmatrix},$$

i.e., a 3×3 elements tensorial function. Also Q_0 is in this case a 3×3 tensor. The approach is not restricted to three dimensions, however.

Eq. 2.1 is fairly realistic for a broad range of geodetic networks: for (one-dimensional) levelling networks we know that $k = 1$ gives good results. In this case $\sqrt{Q_0} = \sigma_0$, a scalar called the *kilometre precision* is expressed in $\mathrm{mm}/\sqrt{\mathrm{km}}$. For two-dimensional networks on the Earth's surface, we have due to isotropy $Q_0 = \sigma_0^2 I_2$, with I_2 the 2×2 unit matrix. This is valid in a small enough area for the Earth's curvature to be negligible, so that map projection co-ordinates (x, y) can be used.

Also for GPS networks an exponent of $k = 1$ has been found appropriate (e.g., [BBB$^+$89]). The 3×3 matrix Q_0 contains the component variances and will, in a local horizon system (x, y, H) in a small enough area, typically be diagonal:

$$Q_{0,hor} = \begin{bmatrix} \sigma_h^2 & & \\ & \sigma_h^2 & \\ & & \sigma_v^2 \end{bmatrix},$$

where σ_h^2 and σ_v^2 are the separate horizontal and vertical standard variances. In a geocentric system we get then the location-dependent expression

$$Q_0\left(\mathbf{r}\right) = R\left(\mathbf{r}\right) Q_{0,hor} R^T\left(\mathbf{r}\right),$$

with $R\left(\mathbf{r}\right)$ the rotation matrix from geocentric to local horizon orientation for location $\mathbf{r}$.

Now if we choose the following expressions for the variance and covariance of absolute (geocentric) position vectors:

$$\begin{aligned} \mathrm{Var}\left(\mathbf{r}_A\right) &= Q_0\left(\mathbf{r}_A\right) R^k, \\ \mathrm{Var}\left(\mathbf{r}_B\right) &= Q_0\left(\mathbf{r}_B\right) R^k \\ \mathrm{Cov}\left(\mathbf{r}_A, \mathbf{r}_B\right) &= \overline{Q}_{0,AB}\left[R^k - \frac{1}{2} d_{AB}^k\right], \end{aligned} \tag{2.2}$$

with R the Earth's mean radius, then we obtain the following, generalized expression for the difference vector:

$$\mathrm{Var}\left(\mathbf{r}_B - \mathbf{r}_A\right) = \overline{Q}_{0,AB} d_{AB}^k, \tag{2.3}$$

with $\overline{Q}_{0,AB} \equiv \frac{1}{2}\left[Q_0\left(\mathbf{r}_A\right) + Q_0\left(\mathbf{r}_B\right)\right]$. This yields a consistent variance structure.

In practice, the transformation to a common geocentric frame will be done using known parameters found in the literature [BA07] for a number of combinations ITRFxx/ETRFyy, where xx/yy are year numbers. Our concern here is only the precision of the co-ordinates thus obtained. We need to know this precision when combining GPS data sets from domains having different canonical WGS84 realizations, requiring their transformation to a suitable common frame.

2.2 Combining spatial and temporal dependency

There is some discussion in an old paper [Ver99]. Essentially combination is done by writing the precision of a vector as a power of both vector length an measurement time:

$$\sigma^2 = C\ell^p t^q,$$

with ℓ vector length and t measurement duration.

We can easily generalize this by substituting any of the above covariance structures (2.1, 2.2, 2.3) for $C\ell^p$, and take for t the measurement duration of a session (taken constant). Combining sessions then becomes more complicated, but no more so than the law of propagation of variances.

Alternatively, one could say that, in the above expressions, the various constants of type Q_0 should be replaced by time dependent functions:

$$Q_0\left(t\right) = Q_0 t^q.$$

Based on practical experience the exponents p and q are typically close to 1.

3 Autocorrelation in time series studies

In real life, we work with *time series*, which are stochastic variables the value space of which are series typically connected to time arguments:

$$x_i = x\left(t_i\right), i = 1, \ldots$$

Often the times t_i are equi-spaced. A more abstract way of looking at time series is to consider them stochastic variables on the value space of *functions* (usually of time). These variables are then called *stochastic processes*.

In the sequel we shall use both approaches somewhat interchangably.

3.1 Time series and white noise

*Noise i*s a stochastic process with an expected value of 0:

$$E\left\{\underline{n}\left(t\right)\right\} = 0.$$

White noise is noise that consists of all possible frequencies. The mathematical way of describing this is saying that the autocovariance

$$A_n\left(\Delta t\right) = 0,\ \Delta t \neq 0.$$

In other words, the process values $\underline{n}\left(t_1\right)$ and $\underline{n}\left(t_2\right)$ do not correlate at all, no matter how close $t_2 - t_1$ is to zero.

Nevertheless we would have

$$A_n\left(0\right) = \infty.$$

And furthermore it holds that

$$\int_{-\infty}^{+\infty} A_n\left(\tau\right) d\tau = Q.$$

Here we assume all the time stationarity.

Perhaps you may want to stare at the above formulas for a while. Here we have a function $A_n\left(\tau\right)$ which is "almost everywhere" zero (namely if $\tau \neq 0$) but in the only point where it isn't zero (namely if $\tau = 0$) it is infinite! And furthermore, the integral function over the τ domain produces exactly Q!

Such a function does not actually exist. It is a mathematical auxiliary device called *distribution*. It is the *delta-function,* named after the quantum physicist Paul Dirac:

$$A_n\left(\tau\right) = Q\delta\left(\tau\right). \tag{3.1}$$

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

$$\int_{-\infty}^{+\infty} \delta_b(\tau)\, d\tau = 1 \tag{3.2}$$

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 $\delta_b(\tau) = 0$.

The handling rule of distributions is simply, that first we integrate, and then in the result obtained we let $b \to 0$.

"Random walk" is obtained if white noise is integrated over time. Let the autocovariance of the noise $\underline{n}$ be

$$A_n(\Delta t) = Q\delta(\Delta t).$$

Then we integrate this function:

$$\underline{x}(t) = \int_{t_0}^{t} \underline{n}(\tau)\, d\tau.$$

Note that

$$E\{\underline{x}(t)\} = \int_{t_0}^{t} E\{\underline{n}(\tau)\}\, d\tau = 0.$$

The autocovariance function is obtained as:

$$\begin{aligned} A_x(t_1,t_2) = &\quad E\{(\underline{x}(t_2) - E\{\underline{x}(t_2)\})(\underline{x}(t_1) - E\{\underline{x}(t_1)\})\} = \\ = &\quad E\{\underline{x}(t_2)\,\underline{x}(t_1)\} = \\ = &\quad E\left\{\int_{t_0}^{t_2} \underline{n}(\tau_2)\, d\tau_2 \int_{t_0}^{t_1} \underline{n}(\tau_1)\, d\tau_1\right\} = \\ = &\quad \int_{t_0}^{t_2}\left[\int_{t_0}^{t_1} E(\underline{n}(\tau_1)\,\underline{n}(\tau_2))\, d\tau_1\right] d\tau_2. \end{aligned}$$

Here

$$\begin{aligned} &\int_{t_0}^{t_1} E\{\underline{n}(\tau_1)\,\underline{n}(\tau_2)\}\, d\tau_1 = \\ &= \int_{t_0}^{t_1} A_n(\tau_2 - \tau_1)\, d\tau_1 = \\ &= Q\int_{t_0}^{t_1} \delta(\tau_2 - \tau_1)\, d\tau_1 = \begin{cases} Q & \text{jos/if } t_1 > \tau_2 \\ 0 & \text{jos /if } t_1 < \tau_2 \end{cases} \end{aligned}$$

From this it follows that

$$\begin{aligned} A_x(t_1,t_2) &= Q\int_{t_0}^{t_2}\left[\int_{t_0}^{t_1} \delta(\tau_2 - \tau_1)\, d\tau_1\right] d\tau_2 = \\ &= Q(t_1 - t_0) + 0.(t_2 - t_1) = \\ &= Q(t_1 - t_0). \end{aligned} \tag{3.3}$$

In this derivation it has been assumed that the autocovariance of the noise function $\underline{n}$ is *stationary*, in other words, that Q is a constant. This can easily be generalized to the case where $Q(t)$ is a function of time:

$$A_x(t_1,t_2) = \int_{t_0}^{t_1} Q(t)\,dt. \tag{3.4}$$

In both equations (3.3, 3.4) it is assumed that $t_1 \le t_2$.

3.2 "Coloured noise", Gauß-Markov (autoregressive) process

Let us study the simple dynamic equation

$$\frac{d\underline{x}}{dt} = -k\underline{x} + \underline{n}, \tag{3.5}$$

where $\underline{n}$ is white noise, the autocovariance function of which is $Q\delta(t_1,t_2)$, and k is a constant. The solution of this differential equation is

$$\underline{x}(t) = e^{-kt}\left\{\underline{x}(t_0)\,e^{kt_0} + \int_{t_0}^{t} \underline{n}(\tau)\,e^{k\tau}d\tau\right\}.$$

The solution satisfies also the initial condition $\underline{x}(t_0)$, which is assumed given.

If we assume that this initial value is errorless, and that the autocovariance function of $\underline{n}$ is

$$A_n(t_1,t_2) = Q(t_1)\,\delta(t_1-t_2),$$

we obtain the autocovariance function of $\underline{x}$:

$$\begin{aligned} A_x(t_1,t_2) &= \\ &= e^{-k(t_1+t_2)}E\left\{\int_{t_0}^{t_1}\underline{n}(\tau_1)\,e^{k\tau_1}d\tau_1\int_{t_0}^{t_2}\underline{n}(\tau_2)\,e^{k\tau_2}d\tau_2\right\} = \\ &= e^{-k(t_1+t_2)}\int_{t_0}^{t_1} e^{k\tau_1}\left[\int_{t_0}^{t_2} E\left\{\underline{n}(\tau_1)\,\underline{n}(\tau_2)\right\}e^{k\tau_2}d\tau_2\right]d\tau_1. \end{aligned}$$

Here

$$\begin{aligned} &\int_{t_0}^{t_2} E\left\{\underline{n}(\tau_1)\,\underline{n}(\tau_2)\right\}e^{k\tau_2}d\tau_2 = \\ &= \int_{t_0}^{t_2} A_n(\tau_2-\tau_1)\,e^{k\tau_2}d\tau_2 = \\ &\quad Q\int_{t_0}^{t_2}\delta(\tau_2-\tau_1)\,e^{k\tau_2}d\tau_2 = \begin{cases} Qe^{k\tau_1} & \text{jos } t_2 > \tau_1 \\ 0 & \text{jos } t_2 < \tau_1 \end{cases} \end{aligned}$$

So assuming that $t_2 < t_1$:

$$\begin{aligned} A_x(t_1,t_2) &= Qe^{-k(t_1+t_2)}\left[\int_{t_0}^{t_2} e^{2k\tau_1}d\tau_1 + \int_{t_2}^{t_1} 0\,d\tau_1\right] = \\ &= \frac{Q}{2k}e^{-k(t_1+t_2)}\left[e^{2kt_2} - e^{2kt_0}\right]. \end{aligned}$$

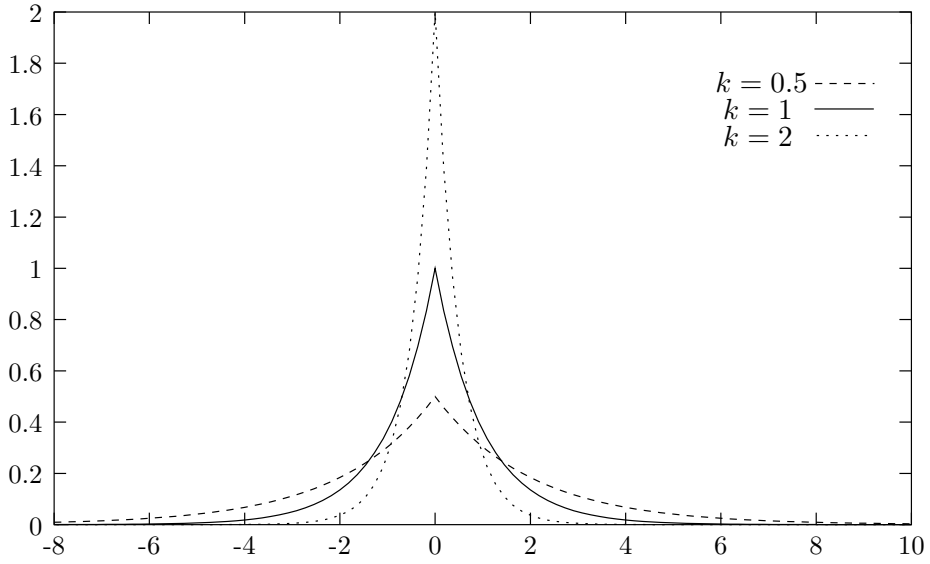

Figure 3.1: Gauß-Markov process autocovariance function

In this case where $t_2 > t_1$ this gives:

$$\begin{aligned} A_x(t_1, t_2) &= Qe^{-k(t_1+t_2)} \int_{t_0}^{t_1} e^{2k\tau_1} d\tau_1 = \\ &= \frac{Q}{2k} e^{-k(t_1+t_2)} \left[e^{2kt_1} - e^{2kt_0} \right]. \end{aligned}$$

In both cases we get

$$A_x(t_1, t_2) = \frac{Q}{2k} \left[e^{-k|t_1-t_2|} - e^{-k(t_1+t_2-2t_0)} \right]. \tag{3.6}$$

In the situation where $t_1, t_2 \gg t_0$ (stationary state long after starting) we obtain

$$A_x(t_2 - t_1) \equiv A_x(t_1, t_2) \approx \frac{Q}{2k} e^{-k|t_2-t_1|}. \tag{3.7}$$

In this (stationary) case we talk about coloured noiseand the process above is called a (first order) *Gauß-Markov* process, also an autoregressive of order 1 (AR(1)) process.

Let us also write

$$Q \equiv qk^2.$$

Then the surface area under the $A_x(t_2 - t_1)$ curve is

$$\int_{-\infty}^{+\infty} A_x(\tau)\, d\tau = \frac{qk}{2} \cdot 2 \int_0^{\infty} e^{k\tau} d\tau = q,$$

a constant if q is constant.

The extreme case $k \to \infty$ leads to the autocovariance function $A_x(t_2 - t_1)$ becoming infinitely narrow, but the surface area under the curve of the function does not change.In other words:

$$A_x(t_2 - t_1) = q\delta(t_2 - t_1).$$

This corresponds to the formula's (3.5) degeneration, where not only $k \to \infty$, but also the noise's $\underline{n}$ variance $Q \to \infty$. So:

$$0 = k\underline{x} - k\underline{\nu} \;\Rightarrow\; \underline{x} = \underline{\nu},$$

where $\underline{\nu} \equiv -\frac{\underline{n}}{k}$:s variance is $q = Qk^{-2}$.

The other borderline case case, where $k \to 0$, is the same as the case presented above (section 3.3). So "random walk" is a Gauß-Markov process the time constant of which is infinitely long. In that case we have to use the whole formula (3.6):

$$A_x(t_1, t_2) = \frac{Q}{2k}\left[e^{-k|t_1-t_2|} - e^{-k(t_1+t_2-2t_0)}\right].$$

In this case, if $t_2 \approx t_1 \equiv t$, we get

$$\begin{aligned} A_x(t) &= \frac{Q}{2k}\left[1 - e^{-2k(t-t_0)}\right] \approx \\ &\approx Q(t - t_0), \end{aligned}$$

which is in practice the same as equation (3.3).

The corresponding dynamic equation is obtained from Eq. (3.5) by substituting $k = 0$:

$$\frac{d\underline{x}}{dt} = \underline{n},$$

so $\underline{x}$ is the time-integral of the white noise $\underline{n}$ as it should be.

Summary	k	dynamic model	autocovariance
Random walk	0	$\frac{d\underline{x}}{dt} = \underline{n}$	$Q(t - t_0)$
Gauß-Markov process	$\in (0, \infty)$	$\frac{d\underline{x}}{dt} = -k\underline{x} + \underline{n}$	$\frac{Q}{2k}e^{-k\lvert t_1-t_2\rvert}$
White noise	∞	$\underline{x} = \frac{\underline{n}}{k}$	$Qk^{-2}\delta(t_1 - t_2)$

Often the model used to generate the "coloured" noise (3.5) or the process – in case where we know beforehand that the properties of the process are of that type. This is easily done by adding one unknown x to the state vector and one equation to the dynamic model of the Kalman filter.

3.2.1 Power spectral density of a Gauß-Markov process

We have the auto-covariance function as Eq. (3.7):

$$A_x(t) = \frac{Q}{2k}e^{-k|t|}.$$

From this follows the Power Spectral Density (PSD) by integration:

$$\begin{aligned} \widetilde{A_x}(f) &= \int_{-\infty}^{+\infty} A_x(t)\exp(-2\pi i f t)\,dt = \\ &= \frac{Q}{2k}\int_{-\infty}^{+\infty} \exp(-k|t|)\exp(-2\pi i f t)\,dt. \end{aligned}$$

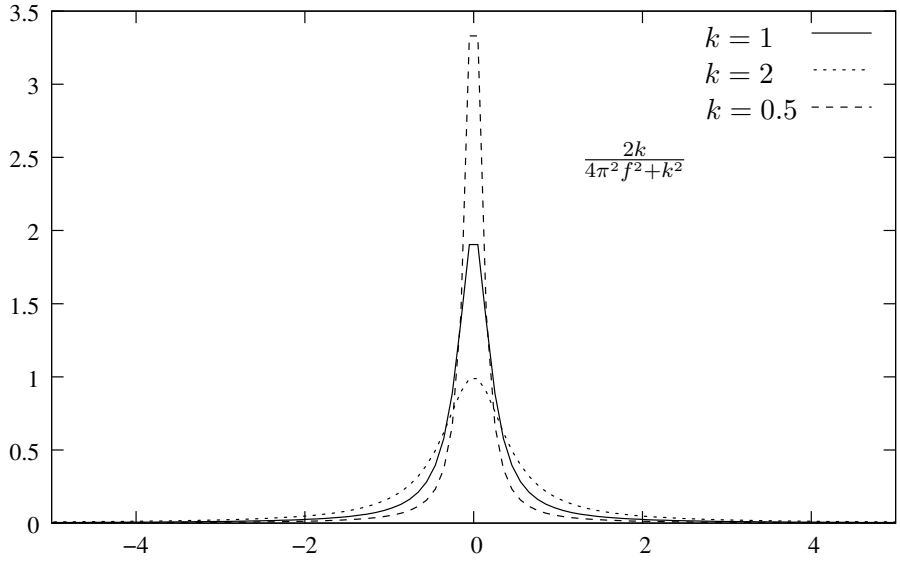

Figure 3.2: Power Spectral Density of a Gauß-Markov process

This integral isn't quite easy to evaluate; it is found in tabulations of integrals and can also be done using computer algebra systems, e.g., [Wol08]. The result is[1]

$$\widetilde{A_x}(f) = \frac{Q}{4\pi^2 f^2 + k^2} = \frac{2kA_x(0)}{4\pi^2 f^2 + k^2}.$$

cf. [Jek01] Eq. (6.75). In Figure 3.2 are plotted values of this function for $Q = 2k$ (i.e., we keep the *variance* of $\underline{x}$, which is equal to $A_x(0) = Q/2k$, at unity) with $k = 0.5, 1, 2$.

3.3 AR(1), linear regression and variance

3.3.1 Least squares regression in absence of correlation

Linear regression starts from the well known equation

$$y = a + bx$$

where we have given many point pairs $(x_i, y_i), i = 1, \ldots n$. This is more precisely an *observation equation*

$$\underline{y}_i = a + bx_i + \underline{n}_i,$$

where stochastic process $\underline{n}_i$ models the *stochastic uncertainty of the measurement process,* i.e., the *noise.*

We assume the noise to behave so, that the variance is a constant independent of i, and the covariance vanishes identically ("white noise"):

$$\begin{aligned} \operatorname{var}\{n_i\} &= \sigma^2, \\ \operatorname{cov}\{n_i, n_j\} &= 0,\ i \neq j. \end{aligned}$$

[1] Note that for $k \to 0$ (random walk) we obtain that $\widetilde{A_x}(f) \propto f^{-2}$.

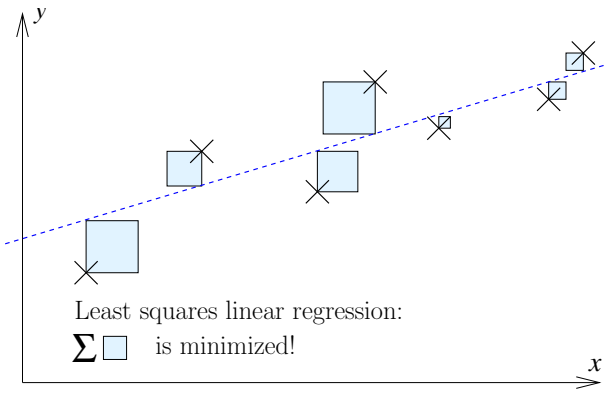

Figure 3.3: The idea of linear regression

This is called the *statistical model.*

We may write the observation equations into the form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & \vdots \\ 1 & x_{n-1} \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix},$$

where now $\mathrm{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^T$ is the vector of observations (in an n-dimensional abstract vector space), $\mathrm{x} = \begin{bmatrix} a & b \end{bmatrix}^T$ is the vector of unknowns (parameters), and

$$A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{n-1} \\ 1 & x_n \end{bmatrix}$$

is the (second order) *design matrix.* This way of presentation is referred to as the *functional model.*

Based on the assumed statistical model we may compute the least squares solution with the help of the *normal equations*:

$$\left(A^T A\right) \hat{\mathrm{x}} = A^T \underline{\mathrm{y}}.$$

More concretely:

$$A^T A = \begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix},$$

or (Cramer's rule):

$$\left(A^T A\right)^{-1} = \frac{1}{n \sum x^2 - \left(\sum x\right)^2} \begin{bmatrix} \sum x^2 & -\sum x \\ -\sum x & n \end{bmatrix},$$

from which

$$\begin{aligned}\hat{a} &= \frac{\sum x^2 \sum \underline{y} - \sum x \sum x\underline{y}}{n\sum x^2 - \left(\sum x\right)^2},\\ \hat{b} &= \frac{-\sum x \sum \underline{y} + n\sum x\underline{y}}{n\sum x^2 - \left(\sum x\right)^2}\end{aligned}$$

are the least squares *estimators* of the unknowns. Their precision (uncertainty, mean error) is given (formal error propagation) by the diagonal elements of the inverted normal matrix $\left(A^T A\right)^{-1}$, scaled by σ:

$$\sigma_a = \sigma\sqrt{\frac{\sum x^2}{n\sum x^2 - \left(\sum x\right)^2}},\ \sigma_b = \sigma\sqrt{\frac{n}{n\sum x^2 - \left(\sum x\right)^2}}.$$

Most often we are particularly interested in the trend b, meaning that we should compare the value $\hat{b}$ obtained with its own mean error σ_b. If σ is not known *a priori*, it should be evaluated from the *residuals*: the square sum of residuals

$$\sum \underline{v}^2 = \sum \left(\underline{y} - \hat{a} - \hat{b}x\right)^2$$

has the expected value of $(n-2)\,\sigma^2$, where $n-2$ is the number of *degrees of freedom* (overdetermination), 2 being the number of unknowns estimated. Or

$$\widehat{\sigma^2} = \frac{\sum v^2}{n-2}.$$

3.3.2 AR(1) process

The assumption made above, that the observational errors $\underline{n}_i$ are uncorrelated between themselves, is often *wrong*. Nevertheless least squares regression is such a simple method – available, e.g., in popular spreadsheets and pocket calculators – that it is often used even though the zero correlation requirement is not fulfilled.

If the autocorrelation of the noise process $\underline{n}_i$ does not vanish, we can often model it as a so-called AR(1) (auto-regressive first-order) or Gauß-Markov process. Such a process is described as a *Markov chain*:

$$\underline{n}_{i+1} = \rho\underline{n}_i + \widetilde{n}_i, \tag{3.8}$$

where ρ is a suitable parameter, $0 < \rho < 1$, and $\tilde{n}$ is a truly non-correlating "white noise" process:

$$\begin{aligned}\mathrm{var}\left\{\widetilde{n}_i\right\} &= \widetilde{\sigma}^2,\\ \mathrm{cov}\left\{\widetilde{n}_i, \widetilde{n}_j\right\} &= 0,\ i \neq j.\end{aligned}$$

Write now the observation equation two times, multiplied the second time around by $-\rho$:

$$\begin{aligned}y_{i+1} &= a + bx_{i+1} + n_{i+1},\\ -\rho y_i &= -\rho a - \rho b x_i - \rho n_i\,,\end{aligned}$$

... and sum together:

$$y_{i+1} - \rho y_i = (a - \rho a) + b\,(x_{i+1} - \rho x_i) + (n_{i+1} - \rho n_i)\,.$$

This equation is of the form

$$Y_i = A + bX_i + \tilde{n}_i,$$

where

$$\begin{aligned} \widetilde{n}_i &\quad \text{as described above,} \\ A &= (1-\rho)\, a, \\ X_i &= x_{i+1} - \rho x_i, \\ Y_i &= y_{i+1} - \rho y_i. \end{aligned}$$

i.e., the formula for the non-correlated linear regression.

The recipe now is:

1. Compute X_i and Y_i according to above formulae;
2. Solve $\widehat{A}$ and $\widehat{b}$ according to non-correlated linear regression;
3. Compute $\hat{a} = (1-\rho)^{-1}\, \hat{A}$;
4. The ratio between $\widetilde{\sigma}^2$ and σ^2: from equation 3.8 it follows, that based on stationarity

   $$\sigma^2 = \rho^2 \sigma^2 + \widetilde{\sigma}^2,$$

   in other words,

   $$\left(1-\rho^2\right) \sigma^2 = \widetilde{\sigma}^2.$$

   So, one either computes an empirical $\widetilde{\sigma}^2$ and transforms it into a σ^2 of the original observations, or the given σ^2 of the original observations is transformed to $\widetilde{\sigma}^2$ in order to evaluate the precisions of the estimators of A and b.
5. From point 4 we may also conclude that

   $$\sigma^2_{b,\mathrm{AR}(1)} = \frac{\sigma^2_{b,\mathrm{nocorr}}}{1-\rho^2},$$

   where $\sigma^2_{b,\mathrm{nocorr}}$ is the "naively" calculated variance of the trend parameter.

Conclusion: if there is autocorrelation in the data, a simple linear regression will give a too optimistic picture of the trend parameter b's mean error, and thus also of its significance (difference from zero).

If the data is given as an equi-spaced function of time, i.e., $x_i = x_0 + (i-1)\,\Delta t$, we may connect the parameter ρ of the AR(1) process in a simple way to its *correlation length*: the solution of equation 3.8 (without noise) is

$$n_j = \rho^{j-i} n_i = e^{(j-i)\ln\rho} = \exp\left(-\frac{(j-i)\,\Delta t}{\tau}\right),$$

where τ is the correlation length in units of time.

For consideration of non-strictly-AR(1) processes, see [Tam08d].

3.4 Various autoregressive models

Remark: much work on this extensive body of theory has been done by econometrists, who obviously work extensively with time series.

3.4.1 Classification

We can write [Ano08b] a general autoregressive model in the form

$$\underline{x}_i = \sum_{j=1}^{p} \rho_j \underline{x}_{i-j} + \underline{n}_i \tag{3.9}$$

where $\underline{n}_i$ is a white noise process. It is commonly assumed that $|\rho_j| < 1$. The case $p = 1, \rho_1 = 1$ produces *random walk*, not strictly an autoregressive model as it is not variance stationary. We call the above an AR(p) model, as the next value in the time series will depend on the p previous values[2].

Starting from the above equation (3.9) we may write

$$\begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i-1} & x_{i-2} & x_{i-3} & \cdots & x_{i-p} \\ x_i & x_{i-1} & x_{i-2} & \cdots & x_{i+1-p} \\ x_{i+1} & x_i & x_{i-1} & \cdots & x_{i+2-p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \rho_p \end{bmatrix} = \begin{bmatrix} \vdots \\ x_i \\ x_{i+1} \\ x_{i+2} \\ \vdots \end{bmatrix} - \begin{bmatrix} \vdots \\ n_i \\ n_{i+1} \\ n_{i+2} \\ \vdots \end{bmatrix},$$

which is a set of *observation equations* in the unknowns ρ_j. We can solve this by constructing *normal equations*[3]; if we define

$$A = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i-1} & x_{i-2} & x_{i-3} & \cdots & x_{i-p} \\ x_i & x_{i-1} & x_{i-2} & \cdots & x_{i+1-p} \\ x_{i+1} & x_i & x_{i-1} & \cdots & x_{i+2-p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \text{ and } \mathrm{b} = \begin{bmatrix} \vdots \\ x_i \\ x_{i+1} \\ x_{i+2} \\ \vdots \end{bmatrix},$$

the normal equation system will be

$$A^T A \mathrm{r} = A^T \mathrm{b},$$

from which the unknowns

$$\mathrm{r} = \begin{bmatrix} \rho_1 & \rho_2 & \rho_3 & \cdots & \rho_p \end{bmatrix}$$

can be solved. The underlying assumption is of course that the noise vector $\underline{n}$ has equal variances and vanishing covariances, as always.

Note that at the top and bottom of the A matrix there will be values $\underline{x}_i$ that are not necessarily known, because they are before the start or after the end of the registered time series. One can substitute zeroes here, but this will change the result for r. It is clear that only in case the number of elements N in the time series is substantially greater than p, that these changes will be minor.

Example: in the case $p = 1$ we have the following observation equations:

[2] When writing the model as an ordinary differential equation, it will be of order p.
[3] In the literature other approaches are found, e.g., through the Yule-Walker equations.

$$\begin{bmatrix} \vdots \\ x_{i-1} \\ x_i \\ x_{i+1} \\ \vdots \end{bmatrix} \rho = \begin{bmatrix} \vdots \\ x_i \\ x_{i+1} \\ x_{i+2} \\ \vdots \end{bmatrix} - \begin{bmatrix} \vdots \\ n_i \\ n_{i+1} \\ n_{i+2} \\ \vdots \end{bmatrix},$$

from which the normal equation

$$\sum_{i=2}^{N} x_{i-1}^2 \cdot \rho = \sum_{i=2}^{N} x_{i-1} x_i,$$

from which directly

$$\rho = \frac{\sum_{i=2}^{N} x_{i-1} x_i}{\sum_{i=2}^{N} x_{i-1}^2}.$$

In this expression we recognize the numerator as an estimator for the covariance over a lag of one time step, and the denominator as one for the variance. Both are constant for a stationary time series. (The observant reader will notice that there would be one more square available for inclusion in the denominator, x_N^2. Then however we must multiply by $N/N-1$ to make the result for ρ unbiased.)

There are a large number of variations on the above theme. In the literature we find abbreviations like

- ARMA: autoregressive moving average (sometimes referred to as Box-Jenkins)
- ARIMA autoregressive integrated moving average
- FARIMA fractional autoregressive integrated moving average
- etc...

In the ARMA model, we have the model equation

$$\underline{x}_i = \sum_{j=1}^{p} \rho_j \underline{x}_{i-j} + \sum_{k=1}^{q} \theta_k \underline{n}_{i-k} + \underline{n}_i. \tag{3.10}$$

As one can see, this is as if the white noise driving term has been replaced by a moving average of several (q+1) values n_{i-k}, effectively producing coloured noise. This is referred to as ARMA(p, q). Now we have also the unknowns θ_k to contend with...

In the ARIMA model, we apply the ARMA methodology on a time series derived by *differencing* from the original one, and from there infer the model for the original time series ("integrated"). This is done if the original time series is not stationary [Ano08d].

3.4.2 Fractional Gaussian noise

The FARIMA model is based on *fractional Gaussian noise*, a phenomenon relating to *long term persistence* that we shall also discuss below in connection with the Hurst phenomenon.

The easiest way to begin is from a *random walk* process:

$$\underline{x}_i = \underline{x}_{i-1} + \underline{n}_i,$$

a single integral of white noise, in the discrete version. We can write this using the difference operator:

$$\nabla \underline{x}_i = \underline{n}_i,$$

or, with $\nabla = 1 - L$, the *lag operator* L defined by $Lx_i = x_{i-1}$:

$$(1 - L)\,\underline{x}_i = \underline{n}_i.$$

Now we generalize this to the following form:

$$(1 - L)^d\,\underline{x}_i = \underline{n}_i,$$

where now the exponent d is allowed to also have non-integer values. A process of this kind is referred to as *fractional Gaussian noise,* cf. [Kou02].

For $d = 1$, we get back as a special case the random walk. For $d = 0$, we get a white noise process (all the time assuming that $\underline{n}_i$ is white, uncorrelated Gaussian noise).

The interesting range for d is $-1/2 < d < 1/2$; for positive d in this range the process has positive long range correlations and persistence. One can show that there is a relationship with the Hurst exponent: $d = H - 1/2$.

We formally write out the above expression using the binomial expansion, which yields:

$$(1 - L)^d = 1 - dL + \frac{d\,(d-1)}{1 \cdot 2} L^2 - \cdots,$$

and applying the operator:

$$\underline{x}_i - d\underline{x}_{i-1} + \frac{d\,(d-1)}{1 \cdot 2}\underline{x}_{i-2} - \cdots = \underline{n}_i$$

or abstractly:

$$\sum_{k=0}^{\infty} \binom{d}{k} (-1)^k\,\underline{x}_{i-k} = \underline{n}_i.$$

For non-integer d, the binomial expansion never stops, so the next value in the process is generated using *all* previous values. It can be shown that the coefficients in this expansion, and the autocovariance as a function of lag, fall off hyperbolically rather than exponentially as is the case for a true AR process.

3.4.3 Durbin-Watkins (DW) test

This test is used to verify that the residuals after any model fit are indistinguishable from white noise. It works like this: consider the *residuals* produced by an attempt at modelling an observed time series (possibly, but not necessarily, by a regression analysis). Then, compute

$$DW = \frac{\sum_{i=2}^{N} \left(\underline{v}_i - \underline{v}_{i-1}\right)^2}{\sum_{i=1}^{N} \underline{v}_i^2}.$$

This value will always lie in the range $[0, 4]$. In the absence of autocorrelation in the *residuals* – meaning that the model appears to capture the covariance behaviour of the time series correctly –, the value will be 2. A good test for autocorrelation is then, to see if DW is closer to 2 than some specified rejection bounds. These bounds are tabulated in the literature, based on number of observations, number of unknowns and desired significance level of the test.

3.4.4 The r^2 test and its limitations

This test is used to verify that two stochastic variables significantly correlate with each other. It is defined as follows:

$$r = \frac{\operatorname{cov}\{x, y\}}{\sqrt{\operatorname{var}\{x\} \operatorname{var}\{y\}}},$$

where $\underline{x}, \underline{y}$ are the stochastic variables considered.

An estimator of r can be computed as follows (overbar denotes average):

$$\underline{r} = \frac{\sum_{i=1}^{n} \left(\underline{x}_i - \overline{x}\right) \left(\underline{y}_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n} \left(\underline{x}_i - \overline{x}\right) \sum_{i=1}^{n} \left(\underline{y}_i - \overline{y}\right)}}.$$

This is itself a stochastic variable, the distribution of which depends on

- the number of data points n
- the *expected value* of r.

Based on this, a statistical test can be designed to detect if r – or r^2– differs significantly from 0.

It is important to be aware what r^2 is really testing. It is testing the existence, and significance, of a *linear functional relationship* between variables x and y :

$$\underline{y} = a + b\underline{x} + \underline{n}. \tag{3.11}$$

It doesn't matter what a and b are, all the test says is that x and y "co-vary": if x goes up, so does y, if x goes down, so does y. It makes no difference if the average *level* of y is completely different from that of x, or if the *size* of the variations in y is very different from that of the variations in x.

In fact, you may apply an arbitrary linear transformation to y. Write

$$\underline{Y} = c\underline{y} + d,$$

and you will have

$$\underline{Y} = A + B\underline{x} + \underline{N},$$

with A, B computable from a, b, c, d.

Now, the r^2 value of Y against x will be *identical* to that of y against x. The r^2 test is invariant to what the linear relationship exactly looks like.

Sometimes, testing relationship (3.11) is what we want; sometimes it is not. One situation where it is not is, if we try to determine *not* whether y is a linear function of x, but whether y *is* (an estimator or reconstruction of) x:

$$\underline{y} = \underline{x} + \underline{n}. \tag{3.12}$$

An example of this occurs in climate field reconstruction, as described in [WA07].

This calls for an entirely different kind of test. Using the r^2 test blindly in this case is not just a blunt instrument, it is the *wrong* instrument. You are not testing the conjecture you're supposed to test. *Ibid.* give some nice examples with plots how with the r^2 test you can both reject perfectly satisfactory reconstructions and swallow junk...

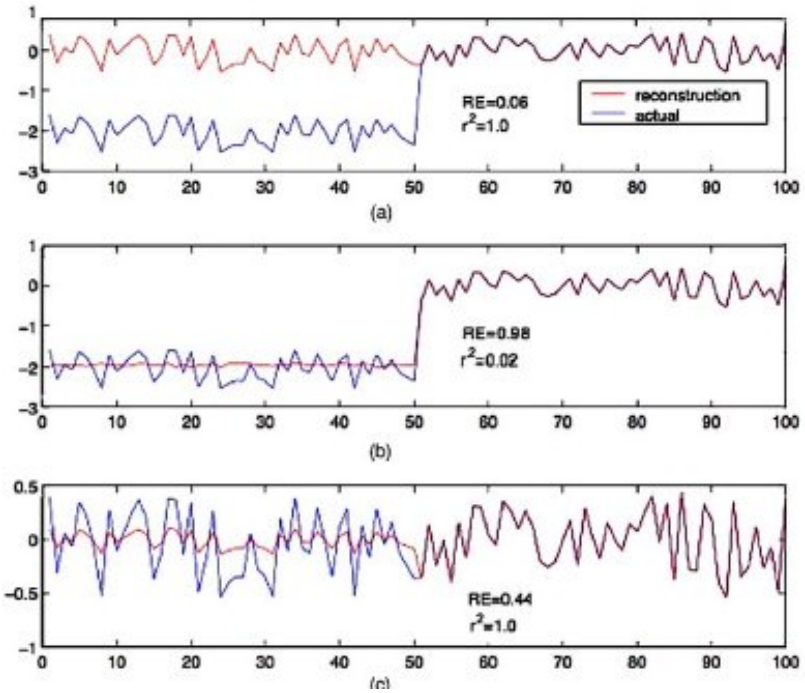

Figure 3.4: Example figure from [WA07], demonstrating cases where r^2 is a poor measure of reconstruction fidelity. (a) accepting a large offset in the mean; (b) rejecting good fidelity in the mean, but failure to reproduce high-frequency details; (c) accepting failure to reproduce amplitude. RE is an alternative measure of reconstruction fidelity often used.

Intermezzo. In the statistical literature we find, in addition to the above definition of r^2 also called Pearson's product moment correlation, a more abstract and broadly valid form often referred to as R^2 and called the "coefficient of determination[4]": how much the RMS of the residuals is less than the RMS of the modelled quantity, i.e., how well the model "explains" the variance of the modelled quantity. In a formula:

$$\underline{R}^2 = 1 - \frac{\sum_{i=1}^{n} \left(\underline{y}_i - \widehat{y}\left(x_i\right) \right)^2}{\sum_{i=1}^{n} \underline{y}_i^2}.$$

For linear regression, Eq. (3.11), r^2 and R^2 are the same. For a reconstruction, however, they are quite different, and in fact R^2 is equivalent to RE. Note that the definition of R^2 is sensitive to the choice of origin for y. Replace y by $y - c$, and R^2 changes. See [WA07] for details.

The r^2 variable is related to the linear regression coefficient in the following way: if b is the regression trend coefficient of y w.r.t. x, and b' that of x w.r.t. y, then $r^2 = bb'$. A good alternative to r^2, in the linear regression case, is to just compute the regression trend b and its standard deviation σ_b, construct a confidence interval, and see if 0 – or whatever your null hypothesis is – lies inside it.

More generally then, one formulates a realistic error model for one's input data, propagate this through the computation to obtain an error model for the unknowns, and judge that against one's null hypothesis.

[4] Or: "variance explained"

3.5 The Hurst exponent and Hurst rescaling

3.5.1 Computing the Hurst exponent

The American hydrological engineer H.E. Hurst already in the 1950's observed a remarkable behaviour of the Nile's water run-off [Hur51] which in today's terminology would be called "fractal" or "self-similar". See [Cle06] or [NJ05].

A simple illustration is given by the standard random walk process. As shown by Eq. (3.3), it has (under assumed stationarity) an autocovariance function of

$$A_x(t_1, t_2) = Q(t_1 - t_0),$$

where t_0 is the start of the time series and $t_1 \leq t_2$. This means that the variance at time t is proportional to $t - t_0$, and *grows without bound* with the length of the time series. If the variance grows without bound, than also the actual values will likely deviate without bound from any computed average, as time proceeds. From his studies, Hurst found that the Nile's run-off actually appeared to behave in this way, as many after him have found for other geophysical processes.

This differs from the behaviour of some other geophysical processes, however. E.g., for the case of temperature time series, we *know* that they are of bounded variance. There are negative feedbacks at work which tend to bring temperatures back to their equilibrium value. Such processes are better described by a Gauß-Markov like Eq. (3.5), containing this feedback term. Its autocovariance is given by Eq. (3.7):

$$A_x(t_1, t_2) \approx \frac{Q}{2k} e^{-k|t_2 - t_1|},$$

which is very much bounded... and of limited "memory": if $t_2 - t_1 \gg k^{-1}$, the autocorrelation vanishes and the value of the process at time t_2 owes nothing to that at time t_1. Contrary to what is the case for random walk, which "remembers" indefinitely the value it had at any time in the past.

Continuing with the Hurst exponent, assume having a time series

$$x_i = x(t_i),\ i = 1, 2, \ldots, N.$$

where the times are equi-spaced:

$$t_i = t_0 + (i - 1)\Delta t.$$

Now we perform a logaritmic transform producing the logaritms of *successive ratios* (i.e., the differences of successive logarithms)

$$y_i = \ln\left(\frac{x_{i+1}}{x_i}\right).$$

Now we study the statistics of this quantity on various-sized subsets of the original time series. We divide the time series into M sub-series of length K so that $M \times K = N$. Label each sub-period with index $m = 1, \ldots, M$, and each value within it by $k = 1, \ldots, K$. Now we can write

$$y_{mk} \equiv y_i.$$

For each sub-period we calculate a *mean*:

$$\mu_m = \frac{1}{K}\sum_{k=1}^{K} y_{mk}.$$

Next we calculate a time series of *accumulated differences from this mean*:

$$X_{mk} = \sum_{k'=1}^{k} \left(y_{mk'} - \mu_m\right),\ k = 1, \ldots, K.$$

We study the *range* spanned by the values in this new time series, defined as

$$R_m = \max_{k=1,\ldots,K}\left(X_{mk}\right) - \min_{k=1,\ldots,K}\left(X_{mk}\right).$$

Next, we compute the standard deviation for each sub-period:

$$\sigma_m = \sqrt{\frac{1}{K}\sum_{k'=1}^{K}\left(y_{mk'} - \mu_m\right)^2}.$$

Now we are positioned to compute the *rescaled range* for *all* sub-series of length K – remember there were M of those – as follows:

$$\left(\frac{R}{S}\right)_K = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{R_m}{\sigma_m}\right).$$

In this way we obtain a set of rescaled ranges, one for every possible value pair (K, M) into which the time series length N can be split (for some of the calculations this means discarding some values if N is not well divisible).

Now we *plot* $\ln\left(\frac{R}{S}\right)_K$ against $\ln K$. The *slope* of the regression line in this plot is called the *Hurst exponent* or *parameter* H.

As a practical matter, for a random walk, $H = 0.5$, as follows from the standard deviation being proportional to the square root of the time span considered (Eq. (3.3)). Conversely however, it cannot be said that a Hurst exponent of 0.5 means that the time series is a random walk. However, if H is found to be in the range $[0.5, 1]$, this is diagnostic of the presence of *long-range dependence* (LRD) or *persistence*.

Note that there can be no guarantee that a well defined Hurst exponent even exists... the derived regression slope may be almost meaningless if the plotted points do not clearly form a straight line. About the difficulties of estimating the Hurst exponent, see [Cle06].

3.5.2 Hurst rescaling

This technique is sometimes used for the detection of "regime changes" i.e. skips or discontinuities in the data, like the calibration of a measurement instrument suddenly changing. Yes, you could call this "long term memory" too.....

We present the example case described in [RO03]. See Figure 3.5.

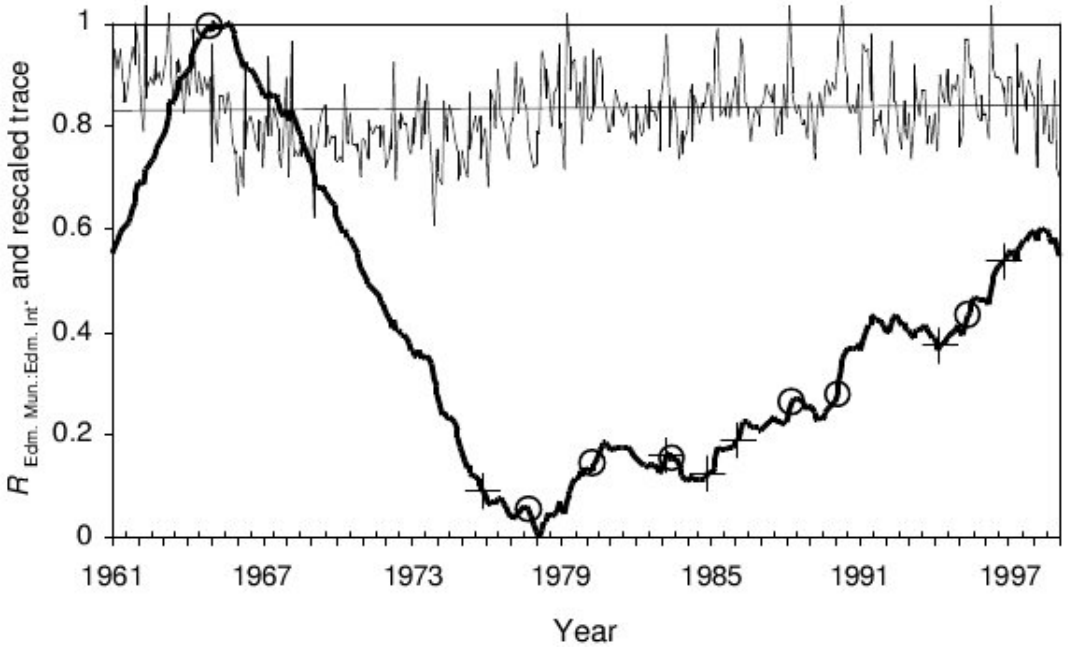

Figure 3.5: $R_{\text{Edmonton Municipal - Edmonton International}}$ and the long-term mean of the series (upper). Hurst retraced scale (lower), with symbols denoting station history events at Edmonton Municipal (+) and Edmonton International (o). The Hurst exponent for the series is 0.76 [from [RO03]].

What Runnals and Oke are doing is to cleverly construct a time series of "cooling ratios" R between meteorological stations. Then, the cumulative values are plotted and rescaled to the interval [0,1].

Now the "skips" in the time series show up as changes in slope, and appear to correlate with known changes in the stations' environments.

Two remarks:

1. The original time series is not even remotely a random walking process, rather it is clearly equilibrium seeking. Only, the equilibrium position changes significantly over time.
2. In this case the phenomenon causing the high Hurst exponent is not in the studied, geophysically interesting quantity, but rather in the measurement process, more precisely its changing calibration in time.

This suggests that whenever a high-H process is found in nature, it would be wise to first have a hard look at the measurement process before assuming that the geophysics has long range persistence. Of course, both may be the case.

Third remark: the actual skip detection process has nothing to do with Hurst. The Hurst exponent is only used as a diagnostic, in this case for the presence of changes in the microclimate regime around the stations. This is often referred to as *change point analysis*.

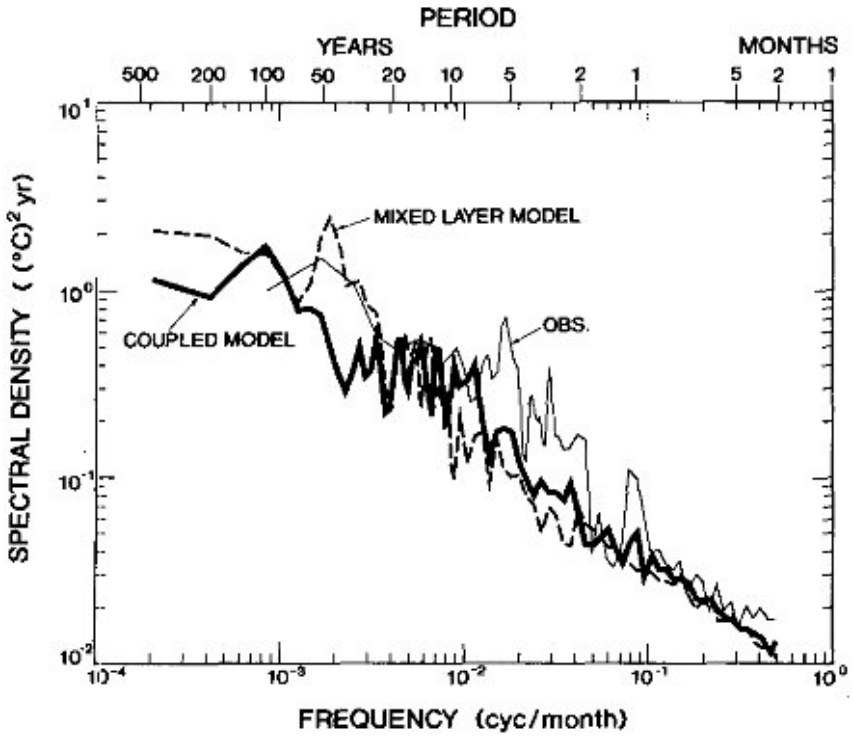

Figure 3.6: Power spectral density of detrended global temperatures [MS96]

3.6 Pink noise

"Pink noise" or "flicker noise" is the kind of time series where the power spectral density is proportional to the inverse of the frequency, $1/f$. This means that the amount of power in a frequency interval df, or in a corresponding time scale interval $d\tau = -f^{-2}df$, is

$$dP = f^{-1}df = -\tau^{-1}d\tau = d(\ln f) = -d(\ln \tau)\,.$$

One could say that such a process contains a fixed amount of power for every "octave" in frequency – or equivalently, in time scale.

Many processes in nature are like this, or very close. We may refer to Figure 17 in [MS96], reproduced here, depicting the spectrum of *detrended* global mean temperature, both from observations and model computations.

Similar pictures are found in the IPCC reports, like [IPC07] Figures 9.7 and 9.8 on pages 686 and 687. Note that in those graphs, the unit on the vertical axis is wrong! It should be $(\,^\circ\mathrm{C})^2\,\mathrm{yr}$.

The exponent of f appears to be here -0.75 rather than -1. The power present is around $(0.1^\circ\mathrm{C})^2$ for every power of ten in the time scale.

It can be seen that the total power integral,

$$\int_0^\infty f^{-1}df = \ln(\infty) - \ln(0)\,,$$

is two-sided divergent, but only "mildly" so – this contrary to white noise, that diverges on the high end only, but strongly.

In real life, the time series we have will always be finite in both duration and sampling density, so the total power will in practice be finite.

According to the literature, pink noise may have a Hurst exponent of between 0 and 0.5; the power spectrum alone does not prescribe this, or even guarantee its unique existence. Being < 0.5 makes it "anti-persistent" or mean-seeking. For global mean temperatures, this is certainly true, although in the presence of trends this may be masked. Or put differently, trend removal makes mean-seeking behaviour visible.

See [WG07].

3.7 Data points per degree of freedom

We have seen that in the presence of long-range correlation, i.e., a non-zero autocovariance between successive values in a time series, variance estimation on, e.g., the trend parameter gives way too optimistic results. The longer the time constant $\tau = k^{-1}$ of autocorrelation is compared to the data point spacing Δt, the worse it gets.

In the extreme limit, when $k \to 0$, i.e., $\tau \to \infty$, we arrive at the case of random walk, where this problem is really bad.

We can reformulate this problem as one of "data points per degree of freedom": the data points are not independent of each other, and each "effective data point" or *degree of freedom* represents effectively as many data points as will fit into the covariance time scale τ.

We can make this consideration more quantitative, following [Tam08a]. For stationary, autocorrelated time series we may define the autocorrelation as

$$\rho_j = \frac{\operatorname{cov}\left\{\underline{x}\left(t_{i+j}\right), \underline{x}\left(t_i\right)\right\}}{\sqrt{\operatorname{var}\left\{\underline{x}\left(t_{i+j}\right)\right\} \operatorname{var}\left\{\underline{x}\left(t_i\right)\right\}}}$$

for "lag" j. Then the number of data points per degree of freedom becomes

$$\nu = 1 + 2\sum_{j=1}^{\infty} \rho_j. \tag{3.13}$$

Now we just divide the number of data points N by this number and obtain an "effective number of data points"

$$N_{\mathit{eff}} = \frac{N}{\nu}.$$

For an AR(1) process, we have

$$\rho_j = \alpha^j = e^{-jk\Delta t},$$

with $\alpha = \exp\left(-k\Delta t\right) \approx 1 - k\Delta t$ being the autoregressive parameter. In this special case we may estimate

$$\nu = 1 + 2\sum_{j=1}^{\infty} \alpha^j = 1 + \frac{2\alpha}{1-\alpha} = \frac{1+\alpha}{1-\alpha}.$$

So, we can compensate for autocorrelation in the data by using an effective number of points

$$N_{\mathit{eff}} = N\frac{1-\alpha}{1+\alpha}.$$

However, be aware that Eq. (3.13) is only approximate. The exact formula is

$$\nu = 1 + 2\sum_{j=1}^{\infty} \rho_j f_j,$$

where f_j is a factor between -1 and $+1$ that depends on the precise nature of the time series. It is known that for small lags j, $f_j \approx 1$, and that for large lags j and not-too-long range autocorrelation, α^j will tend to zero quickly. This justifies the above approximation.

There is one problem situation however, and that is the situation where the sum

$$\sum_{j=1}^{\infty} \rho_j \tag{3.14}$$

fails to converge at all. This is precisely the situation where we have *long range dependence*. Of course, in practice already slow convergence may trip us up.

Very often, Tamino remarks, a LRD type process has an autocovariance that may be asymptotically modelled as

$$\rho_j \approx C j^{-\beta},$$

with β being a suitable exponent. Note that if $\beta > 1$, Eq. (3.14) will neatly converge. For $\beta = 1$, we will have *logarithmic* (i.e., slow but certain) *divergence*; for $\beta < 1$, we will have solid divergence.

The β parameter is related to the Hurst parameter through

$$H = 1 - \frac{\beta}{2}.$$

We may say

$$H \begin{cases} \in (0.5, 1] & \text{long range dependence or persistence} \\ = 0.5 & \text{long range dependence (barely; e.g., random walk)} \\ \in [0, 0.5) & \text{no long range dependence or anti-persistence.} \end{cases}$$

One lesson to take home is also that the often routinely used method of extracting statistical parameters from residuals "as if" they applied to original process, is no longer allowed here. It *is* allowed if the number of data points is much greater than the number of parameters: e.g., 100 data points to estimate 2 parameters a, b of a linear regression fit. But if there is long-range correlation over a significant part of the time range of the time series, *this no longer holds*. For random walk, where correlation length is infinite, this is *a fortiori* the case.

In these cases, one will have to use "out of sample" statistics. E.g., when trying to estimate a trend from a time series contaminated by a random walk -style noise process, you'll have to do the statistical testing using variances derived from similar data in which the sought-after trend is not present. Alternatively, it *is* possible to derive the statistics required from the original time series, but the math involved is not simple.

To complicate matters, often one will have time series of finite length where one gets the impression that there are long range correlations – but of course there is no way to verify this from the data. *In practice* one may have to assume that the effective number of data points is very small.

3.7.1 Maximum entropy method

The concept of entropy originated from thermodynamics (statistical mechanics), where it describes, roughly, the state of disorder a system is in. More precisely, it is the (log of the number of micro-states corresponding to the macro-state (e.g., characterized by temperature and pressure) the system is in.

Characteristic for entropy is, that in a system left to itself, it can only increase. In a given system it also increases with increasing temperature, but not proportionally. (One would think that this means that temperature will rise indefinitely in a closed system, but it does not: the temperature is prescribed by the free energy available in the system. however, temperature *differences* will tend to even out, increasing the entropy.)

Stating that entropy is related to the amount of disorder in a system, is a negative statement. A positive way to state the same is saying that entropy represents (minus the log of) the *information content* of the system; more precisely, the number of bits needed to describe what micro-state, of all those available for this macro-state, the system is in.

This link between statistical mechanics and information theory was clearly seen by Claude Shannon. Then, in a seminal 1957 paper, E. T. Jaynes proposed the statistical method of *Maximum Entropy inference* [Jay57].

Maximum entropy has some interesting consequences. It tells us which probability distribution we should assume in case we know only very little, and don't want to imply more knowledge than we have.

- In the discrete case, if all we know is that a system has six states (a die), and we know nothing suggesting that any of those states would be preferred over any other, then the Maximum Entropy probability distribution is: $1/6$ for every state. This is in fact a restatement of Laplace's *principle of indifference.*
- For a real-valued random variable known to be within a bounded interval $[a, b]$, and nothing further is known: also here the ME distribution is the *uniform* one:

$$p(x) = \begin{cases} 1/b-a & a < x < b \\ 0 & \text{otherwise.} \end{cases}$$

  Note that we assume here that every subinterval dx within $[a, b]$ is "equivalent", i.e. the continuous version of the principle of indifference presented above for the discrete case.
- If we know of a real-valued random variable its *mean* and its *standard deviation*, the ME distribution is the normal or Gaussian distribution with those parameters. This corresponds to the experiental fact that adding together lots of little random effects tends in the limit to produce a normal distribution; the "pajazzo effect".

  Note again that, when working with real-valued variables, there is an implied assumption of *translational symmetry*, i.e., the assumption that all places on the real line are equivalent. This property is invariant under affine, but not under general, transformations.
- If we know that the value is real and *positive,* and we have its *mean* μ, the ME distribution is exponential:

$$p(x) = \frac{1}{\mu} \exp \frac{x}{\mu}$$

  This, and related distributions, are often found in thermodynamics.

There are some good articles on the Maximum Entropy method and theory on Wikipedia.

3.7.2 Bayes, Akaike info criteria

Above we saw that strongly autocorrelated time series contain significantly less *independent* information than they on the surface, based on the number of data points, appear to contain. This begs the question, how one can judge the amount of meaningful information a data set really contains, i.e., especially in the context of statistical inference, how well one can *model*, i.e., describe, the data using a minimum of free parameters.

This contains two issues:

1. goodness of fit of the model to the data, typically using the metric of sum of squares of residuals of fit;
2. number of free parameters needed.

There are different ways of combining these two aspects in building criteria for *model selection*.

The Akaike information criterion is [Ano08a]

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

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.

In the common case of normally and independently distributed observations, this becomes

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

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

The alternative Schwarz or *Bayesian* information criterion [Ano08c] is

$$BIC = k\ln n - 2\ln L,$$

and again in the normally and independently distributed case

$$BIC = k\ln n + n\ln\frac{\sum_{i=1}^{n} v_i^2}{n}.$$

The idea with all of this is, that the parameter should be minimized, leading to as small as possible residuals, but not at the expense of using a large number of free parameters.

3.8 CASE I: computing climate sensitivity from the oceans' heat capacity

Determining the sensitivity of the Earth's climate system, i.e., the number of degrees of global mean temperature increase ofter the concentration of CO_2 has doubled and the system has been allowed

to return to a state of equilibrium, has been an interesting challenge. Current best values cluster around 3°C, but with substantial uncertainty[5] [IPC07].

We may model the response of the climate system as an AR(1) process as follows:

$$C\frac{dT}{dt} = F(t) - \sigma T^4,$$

where T is temperature, t time, $F(t)$ the external forcing, and σT^4 the loss of heat due to outgoing longwave radiation (Stefan-Boltzmann). We can linearize this relative to an equilibrium temperature T_0, and an equilibrium forcing F_0, as follows[6]:

$$C\frac{d}{dt}\Delta T = \Delta F(t) - 4\sigma T_0^3 \Delta T. \tag{3.15}$$

Putting the external forcing anomaly ΔF to zero and adding a noise term produces:

$$\frac{d}{dt}\underline{\Delta T} = -\frac{4\sigma T_0^3}{C}\underline{\Delta T} + \underline{n}(t), \tag{3.16}$$

where $\underline{n}$ is the noise forcing assumed to have the characteristics of white noise.

Eq. (3.16) is essentially the same as Eq. (3.5), i.e., it describes a Gauß-Markov process. According to Eq. (3.7), the autocovariance function should be proportional to

$$e^{-k|t_2-t_1|},$$

where

$$k = \frac{4\sigma T_0^3}{C}.$$

In the latter expression, *only the heat capacity C is unknown.*

This gives us a possibility to estimate the heat capacity of the climate system, and from that, the climate sensitivity, by empirically determining the temporal correlation length k^{-1} from observed Earth temperature time series. This approach was taken by Stephen E. Schwartz in 2007 [Sch07], finding a value of $k^{-1}$of 5 ± 1 years, and a corresponding equilibrium doubling sensitivity of only 1.1 ± 0.5 K, much smaller than the currently accepted best value of 3 K.

This result drew quite some attention, and it was pointed out that the paper contained a number of weaknesses. One obvious weakness is the assumption that global temperatures behave like an AR(1) type stochastic process. We actually know that there is not just one time scale k^{-1}at which the ocean responds to external forcings, some of which are much longer than mere decades. There is especially a difference in the response time scale of surface layers and deeper layers.

A comment (one of several) was published in JGR pointing out these problems: [FAMS08].

A response by Schwartz to the comments: [Sch08].

The comments led to an upward revision of both climate system response time and doubling sensitivity by 70%, to 8.5 ± 2.5 years and 1.9 ± 1.0 K, respectively, mostly removing the apparent contradiction.

Note that this re-analysis is still based on the assumption of a single time constant, i.e., a scalar AR(1) model. Below we show how to refine the approach to take into account the separate roles of surface and deep ocean.

[5] Also the definition contains some potential ambiguity: traditionally the slow effect of albedo feedback by continental ice sheets has not been included in it.

[6] In this formula, we have assumed the Earth to be a black body for outgoing radiation. This is not a necessary assumption.

3.9 AR(1) vectorial case

We start from the above equation (3.16), rewritten as

$$\frac{d}{dt}T_1 = -\frac{4\sigma T_0^3}{C_1}T_1 + n,$$

where the index 1 refers to the Earth's atmosphere and ocean surface layers. This equation describes (in linearized form) the processes of various forcings adding energy, outward long wavelength radiation removing it, and a source of random variation.

We describe the process of transport of heat energy between surface and deep ocean (index 2) by a transport coefficient γ, modifying the above equation to

$$\frac{d}{dt}T_1 = -\frac{4\sigma T_0^3}{C_1}T_1 + \frac{\gamma}{C_1}\left(T_2 - T_1\right) + n$$

and adding an equation for the deep ocean:

$$\frac{d}{dt}T_2 = -\frac{\gamma}{C_2}\left(T_2 - T_1\right),$$

where now C_2 stands for the heat capacity of the deep ocean.

Defining the formal vector

$$\mathrm{T} = \begin{bmatrix} T_1 \\ T_2 \end{bmatrix}$$

we may write this as

$$\frac{d}{dt}\begin{bmatrix} T_1 \\ T_2 \end{bmatrix} = \begin{bmatrix} -\frac{1}{C_1}\left(4\sigma T_0^3 + \gamma\right) & \frac{\gamma}{C_1} \\ \frac{\gamma}{C_2} & -\frac{\gamma}{C_2} \end{bmatrix}\begin{bmatrix} T_1 \\ T_2 \end{bmatrix} + \begin{bmatrix} n \\ 0 \end{bmatrix}. \tag{3.17}$$

The coefficient matrix of this equation has two eigenvalues. As the determinant is positive:

$$\det\begin{bmatrix} -\frac{1}{C_1}\left(4\sigma T_0^3 + \gamma\right) & \frac{\gamma}{C_1} \\ \frac{\gamma}{C_2} & -\frac{\gamma}{C_2} \end{bmatrix} = \frac{4\sigma T_0^3\gamma}{C_1 C_2},$$

it follows that both are real and of the same algebraic sign. That sign is negative (why?). As we also know that $C_2 \gg C_1$, and the numerically larger eigenvalue is likely to be close to the earlier derived

$$-\lambda_1 = k_1 \approx \frac{4\sigma T_0^3}{C_1},$$

it follows that the second, much smaller, eigenvalue (associated with a much longer time scale) will be

$$-\lambda_2 = k_2 \approx \frac{\gamma}{C_2}.$$

With this, the general solutions for Eq. (3.17) in the absence of the noise term, for the temperature anomalies will be of form

$$\begin{aligned} T_1 &= \alpha_1 \exp\left(-k_1 t\right) + \alpha_2 \exp\left(-k_2 t\right), \\ T_2 &= \beta_1 \exp\left(-k_1 t\right) + \beta_2 \exp\left(-k_2 t\right). \end{aligned}$$

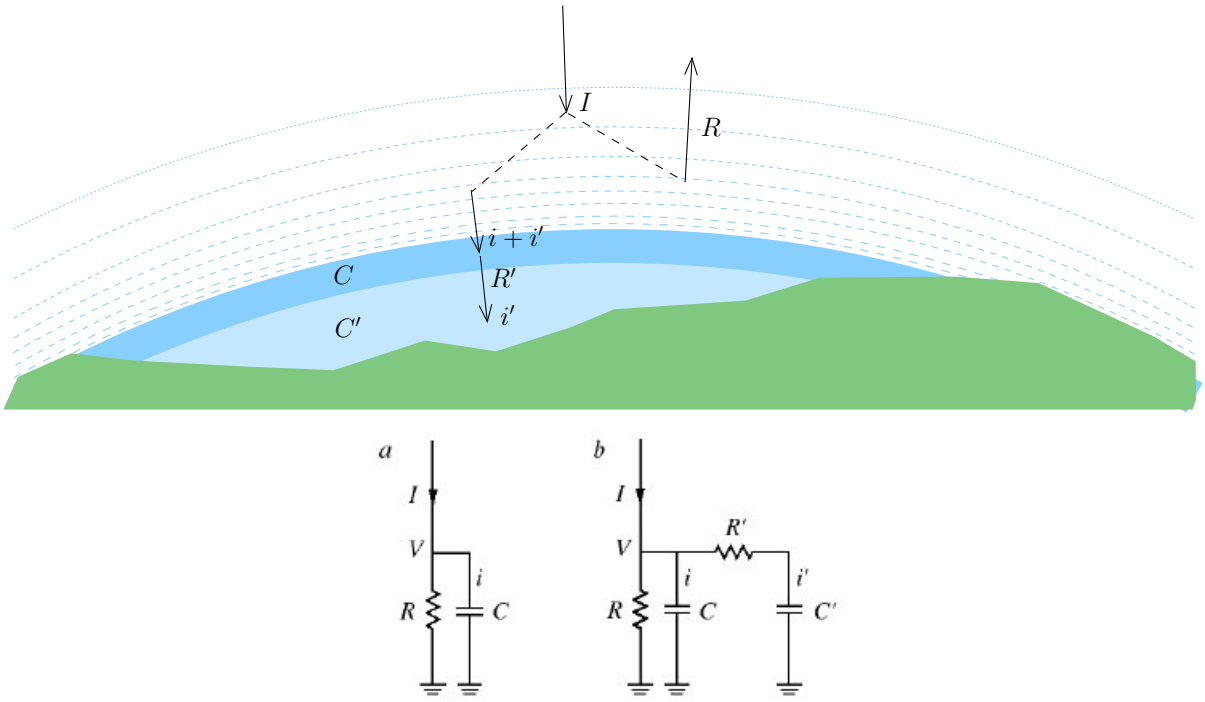

Figure 3.7: Equivalent electrical circuits for determination of climate sensitivity. Subfigure a: single capacitance, single time constant; subfigure b: two capacitances and two time constants. From [Sch08].

Here the alphas and betas depend completely on the initial conditions. Also the autocovariance function will be a mix of two different exponents, and will have much thicker "tails" than the function $\exp(-k_1 |t_2 - t_1|)$ of k_1 alone. This behaviour is somewhat similar to "long range dependence" discussed elsewhere. It complicates empirical determination from not-very-long time series.

In [Sch08] a figure is given for the electrical analogues (RC circuits) for the scalar and vectorial AR(1) processes as used in determining the climate sensitivity.

It is worthwhile especially to read pages 5 and 6 of [Sch08], where he argues that especially R' (transport to the deep ocean) in the figure is so relatively small, that the climate system including ocean surface waters is essentially decoupled from the deep ocean, at least where determination of climate sensitivity from relatively short time series is concerned. The effect of the big deep ocean reservoir C' is mainly to vastly extend the time (to as much as 3000 years) for reaching equilibrium. This also implies that part of the equilibrium warming will be "in the pipeline" for a long time.

Note that it can be shown [WG07] that for a superposition of AR(1) processes with time constants λ uniformly distributed between λ_1 and λ_2, the power spectral density will be roughly of $1/f$ type for $\lambda_1 < f < \lambda_2$, cf. section 3.6. And note: "Even the sum of as few as three AR(1) processes with widely distributed coefficients (e.g., 0.1, 0.5, 0.9) gives a reasonable approximation to a $1/f$ power spectrum" [WG07]. This may well be the situation for the climate system.

3.10 CASE II: computing climate sensitivity from temperature and radiative balance

We can, with the existence of satellites observing both ingoing and outgoing radiation at the top of atmosphere, try to determine sensitivity by studying the relationship between variations in tropospheric (or Earth surface) temperature and in the balance between incoming (Solar, short wavelength, "visual") and outgoing (long wavelength, "calorific") radiation energy.

A number of authors have done this. We start from Eq. (3.15), writing it as

$$C\frac{dT}{dt} = F - \lambda T, \tag{3.18}$$

with

$$\lambda = 4\sigma T_0^3 + \lambda_w + \cdots$$

representing a number of so-called feedback terms. The first, $\lambda_0 = 4\sigma T_0^3$, is the radiative cooling "feedback[7]" according to Stefan-Boltzmann. λ_w stands for the water vapour feedback; there are several more.

Thus, we have

$$C\frac{dT}{dt} = F - \{\lambda_0 + \lambda_w + \cdots\}\, T.$$

We see that for a stable (equilibrium) state, $\frac{dT}{dt} = 0$ and thus

$$\frac{T}{F} = \frac{1}{\lambda} = \frac{1}{\lambda_0 + \lambda_w + \cdots},$$

the *equilibrium sensitivity* to a specified forcing F.

Now we go for a slightly different approach. we look at the top-of-atmosphere (TOA) radiation balance N, and write it as the sum of two terms:

$$N = Q - \lambda T. \tag{3.19}$$

Here, anything that changes the TOA radiation balance due to a change in surface temperature T – and that includes increased outgoing radiation and the increased effect on water vapour – is called *climate feedback* and is included in the second term. everyting else that is not temperature dependent – carbon dioxide forcing, solar variability – is included in the *radiative forcing* term Q.

Note that *if we knew* Q, we would be able to solve for λ, and thus the sensitivity λ^{-1}, from the observed co-variations of N and T. Here we have assumed that λ is a simple number, independent of the time scale of these variations. This is however not the case.

Note also that if the forcing $F(t)$ (or $Q(t)$) is given, Eq. (3.18) allows us to determine $T(t)$. Then again Eq. (3.19) gives us the top-of-atmosphere radiation balance N, *provided* we know λ. If, on the other hand we also have *observations* of N, we may use this to get a better handle on λ.

Several researchers have used this approach rather successfully, e.g., [FG06]. Here we refer to a blog post [Tam08c] addressing this technique, and especially how to handle the presence of feedbacks with different response time scales.

[7] This is not commonly called a feedback by most authors.

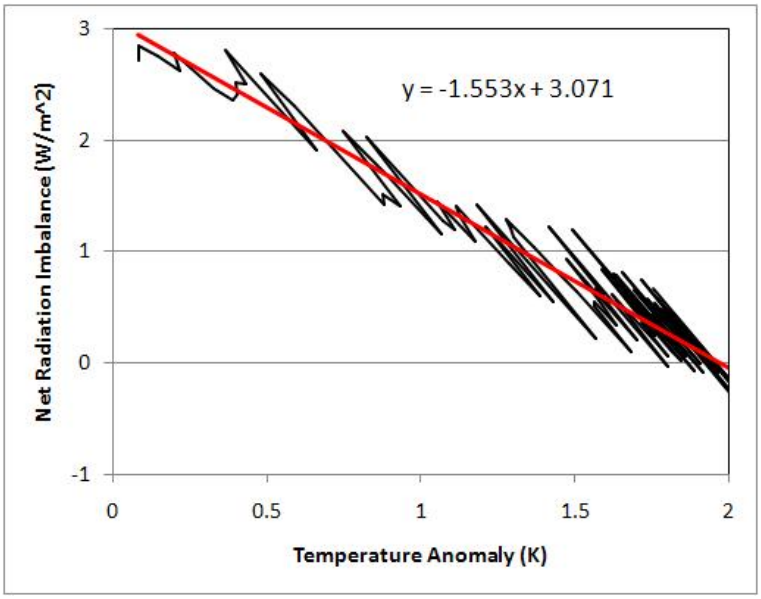

Figure 3.8: Temperature anomaly vs. top-of-atmosphere radiative unbalance in the presence of two different feedback time scales; simulation result. From [Tam08c]

The radiative response of the Earth to temperature change, as characterized by λ_0, is pretty immediate: it comes into full effect as soon as the temperature change has spread throughout the troposphere, typically a matter of hours or days.

The water vapour reponse however takes a longer time: some ten days, which we may compute by dividing global average precipitation, 2.6 mm/day, on the global average water vapour column, 25 mm water equivalent. The same result can be obtained by removing water vapour from a global circulation model and observing how long it takes to be restored [Sch05].

Figure 3.8 shows how the relationship between temperature anomaly and top-of-atmosphere radiation unbalance looks when we assume two different feedbacks with different time scales:

1. Radiative feedback of $3.3\,\mathrm{W\ m^{-2}K^{-1}}$, instantaneous;
2. Water vapour feedback at $-1.5\,\mathrm{W\ m^{-2}K^{-1}}$, slow (in reality this is ten days).

This becomes interesting when looking at a slide show presentation by Roy Spencer [SB08, slide 5]. The figures look very similar to the theoretical plot by Tamino.

We can compute theoretically what the "effective" feedback coefficient λ will be for a sinusoidal forcing variation of a given period: clearly it will be in between the extremes of 3.3 and $3.3-1.5 = 1.8$. We write the feedback g as a function of forcing f:

$$g\left(T\right) = \int_{-\infty}^{T} e^{\frac{t-T}{\tau}} f\left(t\right) dt = \int_{-\infty}^{T} e^{\frac{t-T}{\tau}} \cos\left(\frac{2\pi t}{\tau'}\right) dt,$$

where τ is the decay time scale of the feedback – like 10 days for water vapour – and τ' the sinusoidal period under consideration.

We substitute $x = {}^{t}/_{\tau}$, $X = {}^{T}/_{\tau}$ and $\sigma = {}^{\tau}/_{\tau'}$, yielding

$$g\left(X\right) = \int_{-\infty}^{X} e^{x-X} \cos\left(\frac{2\pi x}{\sigma}\right) dx = e^{-X} \int_{-\infty}^{X} e^{x} \cos\left(\frac{2\pi x}{\sigma}\right).$$

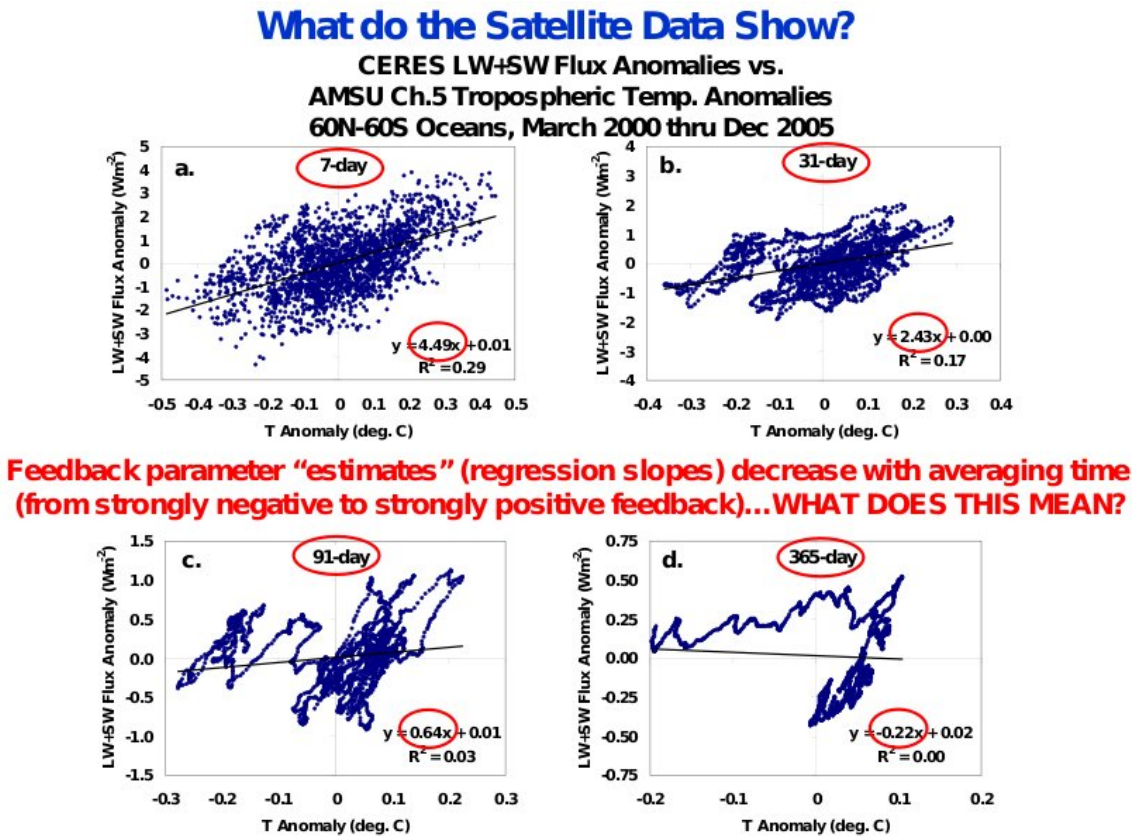

Figure 3.9: Roy Spencer's graphs, slide 5 in [SB08]. See text.

This is a standard integral. According to Wolfram's on-line integrator [Wol08] it is

$$\begin{aligned} g(X) &= e^{-X}\left[\frac{e^{x}\left\{\sigma\cos\left(\frac{2\pi x}{\sigma}\right)+2\pi\sin\left(\frac{2\pi x}{\sigma}\right)\right\}}{\sigma^{2}+4\pi^{2}}\right]_{-\infty}^{X} = \\ &= e^{-X}\left[e^{X}\frac{\left\{\sigma\cos\left(\frac{2\pi X}{\sigma}\right)+2\pi\sin\left(\frac{2\pi X}{\sigma}\right)\right\}}{\sigma^{2}+4\pi^{2}}\right] = \\ &= \frac{\left\{\sigma\cos\left(\frac{2\pi X}{\sigma}\right)+2\pi\sin\left(\frac{2\pi X}{\sigma}\right)\right\}}{\sigma^{2}+4\pi^{2}}. \end{aligned}$$

What we are interested in here is the *amplitude*. The amplitude of $f(t)=\cos\left(\frac{2\pi t}{\tau'}\right)$ is trivially $A_f=1$. As for the amplitude of $g(X)$, observe that the sine and the cosine terms are mutually orthogonal in phase. This means that we can compute the amplitude by root-mean-squaring these two amplitudes, yielding

$$A_g=\frac{\sqrt{\sigma^{2}+4\pi^{2}}}{\sigma^{2}+4\pi^{2}}=\frac{1}{\sqrt{\sigma^{2}+4\pi^{2}}}.$$

With this, we have obtained the *attenuation factor* for this feedback, which is

$$\rho(\sigma)=\rho\left(\tau,\tau'\right)=\frac{1}{\sqrt{\sigma^{2}+4\pi^{2}}}=\frac{1}{\sqrt{\left(\tau/\tau'\right)^{2}+4\pi^{2}}}.$$

For water vapour, $\tau = 10$ days. We can tabulate $\rho(\tau')$, and the total feedback

$$\lambda(\tau') = \lambda_0 + \lambda_w(\tau') = 3.3\,\mathrm{Wm^{-2}K^{-1}} - \rho(\tau') \cdot 1.5\,\mathrm{Wm^{-2}K^{-1}},$$

as a function of τ':

τ'	ρ	$\lambda(\tau')$	Spencer coeff.
7 days	0.111	3.13	4.49
31 days	0.442	2.64	2.43
91 days	0.823	2.07	0.64
365 days	0.986	1.82	-0.22

In this table we have also listed for comparison the coefficients of linear regression derived by Spencer from his graphs. Note that a direct comparison would not be fair, as our feedback coefficients were derived for one sinusoidal time scale τ', while Spencer applied the time scale as a moving average, i.e., a kind of low pass filter. Still it can be seen that the patterns look rather similar.

Looking again at Spencer's graphs, it must be obvious that the data used is rather noisy, and the coefficients obtained correspondingly uncertain. However, the presence of the diagonal "striations" associated with rapid response without water vapour feedback, on top of the weaker trend for longer time scales where water vapour feedback is present, is a powerful demonstration of the reality of this feedback, at least on the monthly to interannual time scales. This was probably not Spencer and Braswell's intention...

4 Slings and arrows of Bayesian inference

4.1 An example

This example is from [Yud03]:

- 1.0% of women contract breast cancer.
- 80% of women with breast cancer test positive.
- 9.6% of women without breast cancer also test positive.

What is the probability that a woman who tested positive, has breast cancer?

In this case, Bayesian analysis looks at *frequencies*[1]. Say, we have 1000 women. Let the parameter be P, having two possible values, $P = 0$ no cancer, $P = 1$ cancer. Let the observation be the test Q, 0 meaning testing negative, 1 testing positive. Then we can draw the following PQ diagram of frequencies:

	$Q = 0$	$Q = 1$
$P = 0$	895	95
$P = 1$	2	8

From this we see that of the 95+8 women who test positive, 8, or slightly under 8%, actually have breast cancer.

We can abstract this from the size of the population by dividing by it, yielding percentages:

	$Q = 0$	$Q = 1$
$P = 0$	89.5	9.5
$P = 1$	0.2	0.8

We can now define the following probabilities: $p(P)$ the probability of having ($p(P = 1) = 1\%$) or not having ($p(P = 0) = 99\%$) cancer $p(Q)$ the probability of testing positive ($p(Q = 1) = 10.3\%$) or negative ($p(Q = 0) = 89.7\%$). $p(Q|P)$ conditional probability of Q given P: e.g., 9.5%/(89.5%+9.5%) = 9.6% for getting $Q = 1$ if $P = 0$, i.e., getting a false positive. $p(P|Q)$ conditional probability of P given Q: e.g. 0.8%/(0.8%+9.5%) = 7.7% for getting $P = 1$ when $Q = 1$, i.e. having cancer if testing positive.

Now, Bayes' theorem says (and this is easy to prove in this case where we have complete frequency population data):

$p(P|Q) = {}^{p(Q|P)p(P)}/_{p(Q)}$.

[1] In the literature, you will often see Bayesian opposed to "frequentist" approaches. There is a substantial body of underlying philosophy connected with this apparent contradiction.

The interesting case arises where we don't have access to such complete data. E.g., we have observations Q and knowledge of which distribution of observations will be produced by any given parameter value P; and we want to know, or infer, what the probability distribution is of P given our observations Q. This is called *reverse inference*, and the above theorem allows us to do that... provided we have access to the distribution $p(P)$, the so-called *prior* distribution.

4.2 Some philosophy

A lot of philosophizing has been going on around Bayesian inference. Understanding it is fairly straightforward in the case where probabilities can be directly related to frequencies within a population. And even then, conclusions may seem counterintuitive.

The common error intuitively reasoning people make, is using the probabilities given for forward inference directly for reverse inference, without considering *a priori* probabilities. (Like: 80% of women with breast cancer test positive, *so* a positive test result means 80% probability of heaving breast cancer...)

However, the frequentist interpretation is only one of the possibilities. Another often used interpretation is that those p values represent *subjective plausibility judgements*, e.g., by experts in a field.

In that case the question arises, what is a plausible probability distribution in the absence of observations (in the form of population statistics on the incidence of breast cancer)? 50%/50%? Some value taken from a similar population elsewhere? This is the conundrum of the *"ignorant prior"*.

In simple (discrete) cases the answer is clear: for coin-tossing, if all we know about the coin is that it is a coin, a reasonable prior is 50%/50%. For a die, if all we know is that it is a die with six faces, it is six times 1/6 for every face. For the continuous-valued case it gets tricky.

4.3 Continuous-valued parameter and observation spaces

For continuous-valued parameter P and observation Q we have the corresponding version:

$p(P|Q)\,dP = {}^{p(Q|P)dQp(P)dP}/_{p(Q)dQ}$,

where now the $p(\cdot)$ functions describe probability densities. We can divide out the differentials dP,dQ yielding the same form as before

$p(P|Q) = {}^{p(Q|P)p(P)}/_{p(Q)}$

but now with probability densities instead of probabilities.

In the PQ plane it looks like Figure 4.1, where $p(P,Q)$ is the two-dimensional probability density.

We can derive that

$p(P) = \int_Q p(P,Q)\,dQ$

$p(Q) = \int_P p(P,Q)\,dP$

$p(P|Q) = {}^{p(P,Q)}/_{\int_P p(P,Q)dP}$

$p(Q|P) = {}^{p(P,Q)}/_{\int_Q p(P,Q)dQ}$

... from which we can again read the form of the Bayes theorem.

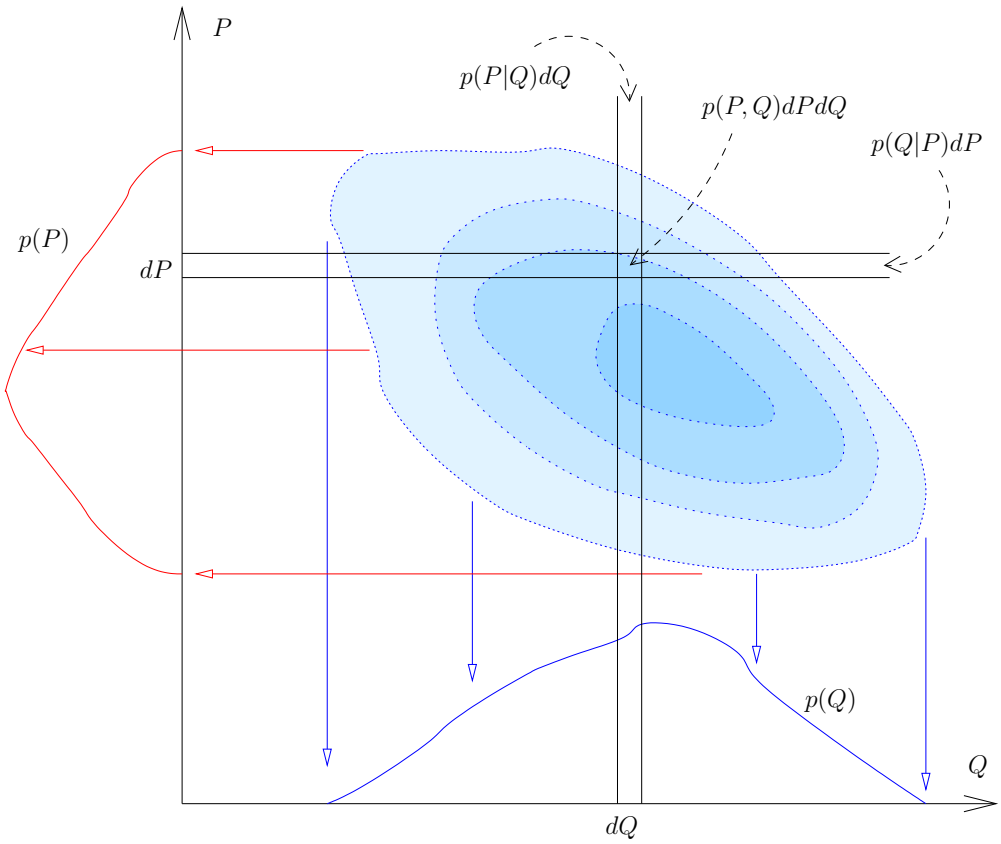

Figure 4.1: The two-dimensional probability density. In this figure, also conditional densities and integrated densities (curves) are given. All densities are scaled to produce a total probability of unity.

4.4 Where do you get your prior?

In the breast cancer case, getting prior probabilities $p(P)$, which is needed to apply the Bayes theorem to obtain probabilities for Q depending on the observations P – i.e., *posterior* probabilities $p(Q|P)$ – is simple. *We know all the frequencies involved*, as we have access to either the complete population, or a well-drawn sample from it.

In the real life of a scientist, it is often not so simple. What if the example's screening test is the only information we have, and we don't really know how likely it is for women to develop breast cancer? This could be realistic when screening a population in a country where this kind of statistics is not being kept.

An example from physical science is the following blog post: [Ann08]. The issue is estimating the equilibrium sensitivity of the Earth's climate system to a doubling of atmopheric CO_2 content, a value believed to be around 3°C. Bayesian inference is the commonly used technique for establishing a likely range of values for this quantity; the observations P in this case are the various determinations of of the sensitivity made by diverse techniques, such as the response to short term volcanic aerosol forcings, palaeoclimate behaviour during ice ages and interglacials, etc.

The blog entry referenced takes issue with the practice of using a "uniform prior" for the sensitivity parameter in this context. It is supposed to represent the absence of *a priori* knowledge, but Annan

argues eloquently that it does not actually do that.

An alternative, and probably more sensible, approach is to take the prior from atmospheric circulation model results. These are supposed to be based purely on physics and not include any of the observational material directly.

4.5 Some more philosophy

The case of a uniform prior is interesting in that it allows the use of forward inference statistics directly and naively for reverse inference. The question however arises, does a uniform prior really represent "observational ignorance", i.e., is it an *ignorant prior*?

In the case of estimating the sensitivity parameter S, this is related to this sensitivity being a highy nonlinear function of the more physical *feedback parameter* f. We may write:

$$S = \frac{S_0}{1-f}.$$

For $f \approx 0$ this gives $S \approx S_0$, in the neighbourhood of which the relationship between f and S will be roughly linear. For $f \to 1$ however we see that $S \to \infty$, so the overall relationship is highly nonlinear.

In a recent paper linked to in [Ann08], our observational knowledge is described in the f domain, as a simple Gaussian (normal distribution) curve centred on zero. This curve will be non-zero when going to $f \to 1$. The curve is then re-written as a function of S, and combined with a uniform prior *on* S. Which Annan and Hargreaves then argue, produces high probabilities for large values of S, like $S > 10^{\circ}\mathrm{C}$, which are believed to be physically completely unrealistic.

The point of the paper is, that this reasoning has been actually applied by some, apparently flawed as it is. One can however make another point, which is that one should not in this way cut the link between observational probabilities and a prior *properly* associated with it.

In the frequentist philosophy this would be obvious: the Gaussian likelihood curve in the f domain *represents probability densities* and should therefore only be used together with a uniform prior in the same domain. In the S domain then, this prior then transforms to the function

$$S_0/S^2$$

(through the relationship $dS = S^2/S_0 df$.) With this, the probabilities for high S values are cut down to size.

Intermezzo. As a matter of theoretical interest, one can show that any probability density $p(\alpha)$ of an argument $\alpha \in (-\infty, \infty)$ can be transformed to a probability density $p(\beta)$of $\beta(\alpha) \in (0,1)$, as follows:

$$\beta = \int_{-\infty}^{\alpha} p(\alpha)\, d\alpha,$$

leading to

$$p(\beta)\, d\beta = p(\alpha)\, d\alpha$$

with $p(\beta) = 1$. For $\alpha = \infty$ we get $\beta = 1$, i.e., the full real line for α maps onto the unit interval for β.

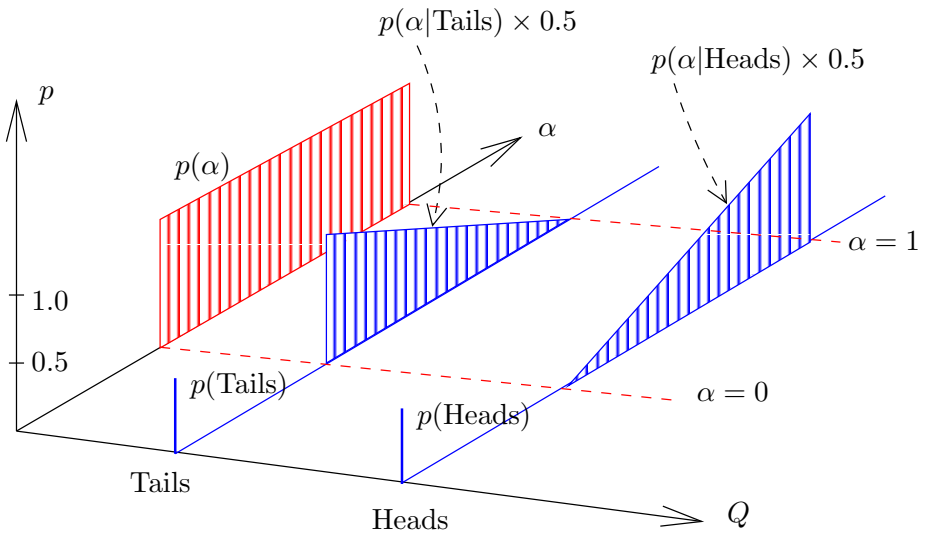

Figure 4.2: Throwing a coin; probabilities under uniform prior

What this means is that the specification of a uniform prior as "ignorant" makes no sense in general, as any prior can be made uniform by a suitable parameter transformation.

It only makes sense if the parameter on which the prior is uniform *makes physical sense*, i.e., there exists an (approximate or exact) translational symmetry in this parameter for the physical situation studied.

Some more interesting philosophical considerations, and useful examples, are to be found in a blog post: [Tam08b]. Do remember to also read the comments.

4.6 Another example

We study the example of modelling the throw of a coin. If a coin is *fair*, it will have the same probability of throwing heads as throwing tails, both 50%. We try to draw a similar diagram of probabilities in the (P, Q) space. Now, throwing a coin is usually depicted as a *discrete process*; this is correct where the outcome of the throw is concerned. This outcome, Q, is the observation and can have one of two values, *heads* or *tails*, or 1 or 0.

However, the quantity we are interested in (on a slightly more abstract level than what the next throw will be), is the probability α that the next throw will be heads (or equivalently, the probability $1 - \alpha$ that the next throw will be tails).

Looking at the picture, you see that, for a value of $\alpha = 0$, the probability of throwing heads is 0 (impossible), and of throwing tails, 1 (certain). The converse is true for $\alpha = 1$. For $\alpha = 0.5$ we see that the two probabilities are equal, as expected for a fair coin, and something any correct model should produce.

However, note that the prior distribution of α is *uniform.* Is this reasonable? It means that it is more likely than unlikely that the coin is unfair ($\alpha \neq 0.5$), and even somewhat likely that the coin is highly unfair.

There is another way to arrive at an unreasonable result (thanks to James Annan for the example). Before doing any coin throw, the expectation of the first throw is

$$E\{\alpha\} = \int_0^1 \alpha p(\alpha)\, d\alpha = \int_0^1 \alpha d\alpha = \frac{1}{2}.$$

Now say that this first throw produces *heads*. Then, by the Bayes theorem,

$$\begin{aligned} p(\alpha|\text{heads}) &= p(\text{heads}|\alpha)\, p(\alpha)\, /p(\text{heads}) = \\ &= \alpha \cdot 1/0.5 = 2\alpha. \end{aligned}$$

From this we get the expectation by

$$\begin{aligned} E\{\alpha\} &= \int_0^1 \alpha p(\alpha|\text{heads})\, d\alpha = \\ &= \int_0^1 2\alpha^2 d\alpha = \frac{2}{3}! \end{aligned}$$

So, suddenly, because the first throw gave heads, we would be prepared to bet two euros against one that the next one will be heads too!

Obviously this is not reasonable for coins coming out of the general circulation (as opposed to a stage conjuror's pocket). What is being ignored here is the tacit knowledge that we all have, that coins will be close to fair. There is even a mechanical rationale for this: a coin is typically thin compared to its diameter. This makes it difficult (short of magic tricks) to make it deviate very much from fairness.

(If this were not a coin, but an electronic device producing randomly ones and zeroes, and a "black box" otherwise, then the above behaviour could be reasonable.)

A better model for the behaviour of a real coin assumes approximate fairness. It is the Gaussian or normal function centred on 0.5, with a small standard deviation parameter σ:

$$p(\alpha) = N(\alpha - 0.5, \sigma).$$

In this case we will have

$$\begin{aligned} p(\alpha|\text{Heads}) &= \alpha \cdot N(\alpha - 0.5, \sigma), \\ p(\alpha|\text{Tails}) &= (1-\alpha)\, N(\alpha - 0.5, \sigma). \end{aligned}$$

For the first throw we have again the expectation

$$E\{\alpha\} = \int_0^1 \alpha N(\alpha - 0.5, \sigma)\, d\alpha = 0.5$$

like before (due to symmetry); but for the second throw we find

$$\begin{aligned} p(\alpha|\text{Heads}) &= p(\text{Heads}|\alpha)\, p(\alpha)\, /p(\text{Heads}) = \\ &= \alpha N(\alpha - 0.5, \sigma)\, /0.5, \end{aligned}$$

and from this the expectation

$$\begin{aligned} E\{\alpha\} &= \int_0^1 \alpha p(\alpha|\text{Heads})\, d\alpha = \\ &= \int_0^1 2\alpha^2 N(\alpha - 0.5, \sigma)\, d\alpha \approx \frac{1}{2}, \end{aligned}$$

for small values of σ when we may substitute

$$N(\alpha - 0.5, \sigma) \approx \delta(\alpha - 0.5).$$

So now we see that the prior knowledge of the fairness of the coin has an influence also on the second throw. For $\sigma = 0$ this knowledge is absolute, and we are in the classical case where, no matter how often we see heads come up, the odds for the next throw are always 0.5. That isn't quite real life either.

4.7 Modelling a simple observation process

This example is taken from [HLKV87]. It refers to horizontal angle measurements by theodolite. Typical for this measurement situation is that the measurement accuracy, i.e., the range of values in which the "true value" can reasonably be expected to lie once the "reading" or measured value is given, is *very much smaller* than the range of values that these horizontal angles can *a priori* belong to, which is a significant part of the circle $[0, 360°)$

The consequence of this is that the *a priori* probability distribution of the unknown angle within the small range about the measured value is very close to uniform.

Discussion. What this means is the following. Suppose we understand the angle measurement process executed by the theodolite well enough to be able to say that, if the "true value" of an angle is α, then the "readings" $\underline{\theta} = \alpha + \underline{n}$ will be distributed normally about the true value, i.e., their probability distribution $p(\theta|\alpha)$ is $N(\alpha, \sigma)$ with standard deviation σ describing the measurement uncertainty. Now, if the prior distribution $p(\alpha)$ of α – the set of values we know α may assume before doing any measurement – is locally uniform within a very narrow range of values about the measured value, we may apply Bayes as follows:

$$p(\alpha|\theta) = {}^{p(\theta|\alpha)p(\alpha)}/_{p(\theta)} \propto p(\theta|\alpha),$$

as also the $\underline{\theta}$ values are "all over the place" and thus $p(\theta)$ locally uniform.

This means that, for a given reading $\underline{\theta}$, the Bayesian estimate $\underline{\alpha} = \underline{\theta}$ for the true angle has the probability distribution $N(\theta, \sigma)$ – which of course we knew intuitively all along. Note that $\underline{\theta}$ may be a "constructed" reading obtained by averaging many raw readings, which preserves normality.

We decided to test this assumption of "local uniformity" empirically. We use the distribution of horizontal angles in the SW Finland Test Network. We find:

$30-40$	$40-50$	$50-60$	$60-70$	$70-80$	$80-90$	$90-100$	$100-110$	$110-120$	>120
2	18	38	27	17	12	7	1	1	1

It is clear that 60° is a very popular value for a horizontal angle in this network, as it tends to be for geometrically well designed networks.

Next we look at the decimal fractions of the measured angles: the first decimal behind the point, and the second decimal behind the point, respectively. Both decimals divide the set of angle measurements into ten equivalence classes; if the distribution, within this sub-degree interval, is uniform, there should be an equal number of values in each equivalence class, equal to the average of ${}^{124}/_{10} = 12.4$. We test this using the χ^2 test for 9 degrees of freedom.

On the first decimal of degrees:

Decimal	0	1	2	3	4	5	6	7	8	9
Count	10	10	8	12	10	15	10	17	10	22
χ^2 contrib	0.4645	0.4645	1.5613	0.0129	0.4645	0.5452	0.4645	1.7065	0.4645	7.4322

The sum of χ^2_9 is 13.5806. This corresponds to a significance level (i.e., the probability of exceeding purely by chance this value) of $p = 13.81\%$. At the 90% probability level the hypothesis of uniformity is accepted.

On the second decimal of degrees:

Decimal	0	1	2	3	4	5	6	7	8	9
Count	19	11	15	14	11	11	14	10	10	0
χ^2 contrib	3.5129	0.0129	0.5451	0.2065	0.1581	0.1581	0.2065	0.4645	0.4645	1.5612

The sum χ^2_9 is 7.2903. This corresponds to a significance level of $p = 60.7\%$. Even at the 67% probability level the hypothesis of uniformity is accepted.

Acknowledgements

Much of the material for these lecture notes was found in raw form on the Internet. Especially "Tamino"'s blog "Open Mind" deserves a mention. It is worth visiting as a new series on autoregressive modelling has just started.

The writing up of these notes was done in part from the author's summer home on a Nokia Internet Tablet type 810, kindly donated by Nokia. The document was edited remotely using the LyX document processor and X11 tunnelled over ssh; arguably this is the smallest form factor fully capable Unix work station on the market today.

Bibliography

[Ann08] James Annan. Once more unto the breach dear friends, once more... URL: http://julesandjames.blogspot.com/2008/05/once-more-into-breech-dear-friends-once.html, May 2008. Accessed July 2, 2008.

[Ano08a] Anon. Akaike information criterion. URL: http://en.wikipedia.org/wiki/Akaike_information_criterion, 2008. Accessed July 3, 2008.

[Ano08b] Anon. Autoregressive moving average model. URL: http://en.wikipedia.org/wiki/Autoregressive_moving_average_model, 2008. Accessed July 3, 2008.

[Ano08c] Anon. Bayesian information criterion. URL: http://en.wikipedia.org/wiki/Bayesian_information_criterion, 2008. Accessed July 3, 2008.

[Ano08d] Anon. *Engineering Statistics Handbook*, chapter Box-Jenkins Models. National Institute of Standards and Time (NIST), 2008. URL: http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc445.htm. Accessed July 3, 2008.

[BA07] Claude Boucher and Zuhair Altamimi. Specifications for reference frame fixing in the analysis of a EUREF GPS campaign. memo, 2007. URL: http://users.auth.gr/kvek/20070327-MEMO-ver6.pdf, accessed August 19, 2008.

[Baa73] Willem Baarda. S-transformations and criterion matrices. Publications on Geodesy, Netherlands Geodetic Commission, Delft, 1973. New Series, Vol. 5 No. 1.

[BBB+89] G. Beutler, I. Bauersima, S. Botton, C. Boucher, W. Gurtner, M. Rothacher, and T. Schildknecht. Accuracy and biases in the geodetic application of the Global Positioning System. *manuscripta geodaetica*, 14(1):28–35, 1989.

[Cle06] Richard G. Clegg. A practical guide to measuring the Hurst parameter. ArXiv, October 2006. URL: http://arxiv.org/abs/math/0610756v1, accessed August 19, 2008.

[FAMS08] Grant Foster, James Annan, Michael Mann, and Gavin Schmidt. Comment on "Heat capacity, time constant and sensitivity of Earth's climate system" by S. Schwartz. *J. Geoph. Res. D (Atmospheres)*, 2008. doi:10.1029/2007JD009373, in press. On line: http://www.jamstec.go.jp/frsgc/research/d5/jdannan/comment_on_schwartz.pdf, accessed August 19, 2008.

[FG06] P. M. D. F. Forster and J. M. Gregory. The Climate Sensitivity and Its Components Diagnosed from Earth Radiation Budget Data. *Journal of Climate*, 19:39–52, January 2006.

[HLKV87] Günter W. Hein, Herbert Landau, Juhani Kakkuri, and Martin Vermeer. Integrated 3-D adjustment of the SW Finland test net with the FAF Munich OPERA 2.3 software. Report 87:3, Finnish Geodetic Institute, Helsinki, 1987.

[Hur51] H.E. Hurst. Long-term storage capacity of reservoirs. *Trans. Am. Soc. Civil Engineers*, pages 770–808, 1951.

[IPC07] IPCC. Climate Change 2007: The Physical Science Basis. Contribution of Working Group I. In S. Solomon, D. Qin, M. Mannin, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller, editors, *Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2007. URL: `http://www.ipcc.ch/ipccreports/ar4-wg1.htm`, accessed July 24, 2008.

[Jay57] E. T. Jaynes. Information Theory and Statistical Mechanics. *Physics Reviews*, 1957. URL: `http://bayes.wustl.edu/etj/articles/theory.1.pdf`, accessed July 24, 2008.

[Jek01] Christopher Jekeli. *Inertial Navigation Systems with Geodetic Applications*. Walter de Gruyter, Berlin – New York, 2001.

[Kou02] Demetris Koutsoyiannis. The Hurst phenomenon and fractional Gaussian noise made easy. *Hydrological Sciences Journal*, (47), 2002. URL: `http://www.cig.ensmp.fr/~iahs/hsj/470/hysj_47_04_0573.pdf`, accessed July 24, 2008.

[MS96] Syukuro Manabe and Ronald J. Stouffer. Low-Frequency Variability of Surface Air Temperature in a 1000-Year Integration of a Coupled Atmosphere-Ocean-Land Surface Model. *Journal of Climate*, pages 376–393, 1996. DOI: 10.1175/1520-0442(1996)009<0376:LFVOSA>2.0.CO;2. On line: `http://ams.allenpress.com/perlserv/?request=res-loc&uri=urn%3Aap%3Apdf%3Adoi%3A10.1175%2F1520-0442%281996%29009%3C0376%3ALFVOSA%3E2.0.CO%3B2`, accessed August 15, 2008.

[NJ05] P. Norouzzadeh and G.R. Jafari. Application of multifractal measures to Tehran price index. *Physica A: Statistical Mechanics and its Applications*, 356(2-4):609–627, October 2005. URL: `http://dx.doi.org/10.1016/j.physa.2005.02.046`, accessed August 19, 2008.

[RO03] K.E. Runnals and T.R. Oke. A technique to detect microclimatic inhomogeneities in historical temperature records. EGS - AGU - EUG Joint Assembly, Abstracts from the meeting held in Nice, France, 6 - 11 April 2003, abstract #3306, April 2003. Full article: `http://www.geo.uni.lodz.pl/~icuc5/text/O_34_4.pdf`, accessed August 19, 2008.

[SB08] Roy W. Spencer and William D. Braswell. Feedback vs. Chaotic Radiative Forcing: "Smoking Gun" Evidence for an Insensitive Climate System?, 2008. URL: `http://climatesci.org/wp-content/uploads/spencer-ppt.pdf`, accessed August 5, 2008.

[Sch05] Gavin Schmidt. Water vapour: feedback or forcing? URL: `http://realclimate.org/index.php?p=142`, April 2005. Accessed August 8, 2008.

[Sch07] Stephen E. Schwartz. Heat capacity, time constant, and sensitivity of Earth's climate system. *J. Geophys. Res.*, 112(D24S05), 2007. Online URL: `http://www.ecd.bnl.gov/steve/pubs/HeatCapacity.pdf`, doi:10.1029/2007JD008746. Accessed July 2, 2008.

[Sch08] Stephen E. Schwartz. Reply to comments by G. Foster et al., R. Knutti et al., and N. Scafetta on "Heat capacity, time constant, and sensitivity of Earth's climate system". *J. Geophys. Res*, 2008. doi:10.1029/2008JD009872, in press. On line: `http://www.ecd.bnl.gov/pubs/BNL-80226-2008-JA.pdf`, accessed August 19, 2008.

[Tam08a] Tamino. Hurst. URL: `http://tamino.wordpress.com/2008/06/10/hurst/`, June 2008. Accessed July 3, 2008.

[Tam08b] Tamino. Reverend Bayes. URL: `http://tamino.wordpress.com/2008/07/10/reverend-bayes/`, June 2008. Accessed July 16, 2008.

[Tam08c] Tamino. Spencer's folly. URL: `http://tamino.wordpress.com/2008/08/01/spencers-folly-3/`, August 2008. Accessed August 5, 2008.

[Tam08d] Tamino. To AR1 or not to AR1. URL: `http://tamino.wordpress.com/2008/08/04/to-ar1-or-not-to-ar1`, August 2008. Accessed August 5, 2008.

[Tuf05] Edward Tufte. PowerPoint Does Rocket Science – and Better Techniques for Technical Reports. URL: `http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0001yB&topic_id=1&topic=Ask+E%2eT%2e`, October 2005. Accessed August 19, 2008.

[Ver99] Martin Vermeer. Geodetic science and the tools of the trade. In Friedhelm Krumm and Volker S. Schwarze, editors, *Quo vadis geodesia...? Festschrift for Erik W. Grafarend on the occasion of his* 60^{th} *birthday. Part 2*, number 1999.6-2 in Department of Geodesy and Geoinformatics, pages 497–503. Universität Stuttgart, 1999. ISSN 0933-2829. Online at `http://www.uni-stuttgart.de/gi/research/schriftenreihe/quo_vadis/pdf/vermeer.pdf`, accessed July 3, 2008.

[WA07] Eugene R. Wahl and Caspar M. Ammann. Robustness of the Mann, Bradley, Hughes reconstruction of Northern Hemisphere surface temperatures: Examination of criticisms based on the nature and processing of proxy climate evidence. *Journal of Climate*, pages 33–67, 2007. DOI: DOI 10.1007/s10584-006-9105-7. On line: `http://www.cgd.ucar.edu/ccr/ammann/millennium/refs/Wahl_ClimChange2007.pdf`, accessed August 19, 2008.

[WG07] Lawrence M. Ward and Priscilla E Greenwood. 1/f noise. Scholarpedia. URL: `http://www.scholarpedia.org/article/1/f_noise`, 2007.

[Wol08] Wolfram Research. Wolfram Mathematica® Online Integrator. URL: `http://integrals.wolfram.com/index.jsp`, 2008. Accessed August 8, 2008.

[Yud03] Eliezer Yudkowsky. An Intuitive Explanation of Bayesian Reasoning, 2003. URL: `http://yudkowsky.net/bayes/bayes.html`. Accessed July 2, 2008.