6. **On the foundations of various approaches to improperly posed problems**

FERNANDO SANSÓ

Istituto di Topografia, Fotogrammetria e Geofisica
Politecnico di Milano, Piazza Leonardo da Vinci, 32
20133 Milano. ITALIA

1. INTRODUCTION: INVERSE PROBLEMS AND IMPROPERLY POSED PROBLEMS

The typical form of a direct problem of classical physics is to determine an unknown field \( u \) which is related to some source \( \rho \) by an action taking place within a medium, who\'s constitution is described by some function \( \sigma \); this is achieved by solving an equation that we could write

\[
L(\sigma, u) = \rho .
\]  

[1.1]

Quite often \( L \) is a differential operator, \( \sigma \) is a set of coefficients of the differential operator, \( \rho \) is the unknown term. The time dependence could be included in [1.1], or it could be already singled out by variables separation or by harmonic analysis, or also [1.1] could directly refer to a static problem.

Probably the most ancient example of a problem of the type [1.1] is the determination of the gravitational potential \( u \) of a body with known density \( \rho \), knowing that

\[
\Delta u = -\rho ;
\]  

[1.2]

as we all know the solution is given by the Newton integral.

Many times equation [1.1] is not sufficient to determine a single solution; but subsidiary conditions, on functionals of \( u \) which should be considered accessible to measurement, have to be used either in the form of asymptotic conditions (e.g. radiation conditions for the equation \( \Delta u + k^2 u = \rho \)) or in the form of conditions on the boundary of a part of space (or time) where we assume to have no information (e.g. the Dirichlet or the Neuman condition.

on a surface $S$ containing the masses that generate an exterior harmonic potential; this situation is formalized as

$$B(\sigma,u) = f \quad (f \text{ known})$$  \[1.3\]

A typical feature of the solutions of such direct problems is that relatively rough functions $(\rho, \sigma, f)$ give rise to smoother solutions $u$.

An inverse problem is the determination of $(\rho, \sigma)$, or functionals of such unknowns, when we know the whole $u$, or possibly a family of functionals of $u$, for one, more than one, or all $f$.

It is exactly this kind of problems that is most interesting in earth sciences, since in their context we usually can obtain a good deal of information on various fields on the boundary and we want to know the distribution of sources within the earth's body and different functions describing its constitution in relation to various physical phenomena (e.g. electrical conductivity, $P$ and $S$ waves velocities i.e. elastic properties, thermal conductivity, etc.).

These inverse problems have, at first sight, some ugly properties, i.e. many times they do not have exact solutions, when there are solutions these are not unique and moreover the solutions may have a wild behaviour when small perturbations are given to data, that is always the case in practice because of measurement errors. This is the unavoidable consequence of the remark that the correspondence $(\rho, \sigma, f) \rightarrow u$ is smooth, what implies that $(u,f) \rightarrow (\rho, \sigma)$ is rough.

In earth sciences the situation is sometimes even more complicated, since already at the level of building up the direct problems we are forced, for the sake of simplicity and computability, to simplify significantly the model so that equations like [1.1] and [1.3] are violated not only by observational disturbances, but also by modelling errors.

We may at this point first of all understand that the treatment of improperly posed problems is a basic tool to solve inverse problems; moreover, specially in earth-science orientated problems, we cannot rely completely on our models, so that the introduction of stochastic methods becomes natural to describe our degree of ignorance.

In this paper we shall examine the main approaches used today to solve improperly posed problems and the accent will be put on the comparison of the different underlying principles.

To simplify matters, we shall treat only linear problems which we can consider as derived from the corresponding nonlinear ones by a linearization process, which, as it can be proved, most of the times carries over the main mathematical features.

Whence we shall treat in a rather abstract form the equation

$$y = Ax$$  \[1.4\]
where $y$ represents the observable field (i.e. $f, u$ or functionals of $u$), while $x$ is the unknown field $(\rho, \sigma)$. The difference with most of the usual «direct» problems is in that the operator $A$ will be assumed to display the features of non wellposedness.

We shall further assume that there is a Hilbert space $X$ to which $x$ has to belong and another Hilbert space $Y$ to which $y$ belongs.

The space $Y$ is not completely arbitrary but it has to be chosen in such a way as to contain all disturbances $\delta y$ that are likely to relate the true field $y$ to the observed field

$$y_0 = y + \delta y$$

[1.5]

For technical reasons, that are discussed a little more in length in Appendix 1, if we like to include among disturbances a white noise, this can be more easily done by starting from $L^2$ functions and by applying the classical construction of the Wiener integral; for this reason we shall be mostly concerned with the case $Y = L^2$.

We must also warn the reader that many times the inverse problem is stated directly in a discrete finite dimensional form, so that $X, Y$, are simply euclidean spaces; this finite dimensional approach is certainly admissible since on one side $y$ is naturally discrete, as the real number of observations is always finite, on the other hand $x$ can always be projected on a finite subspace of $X$, since by [1.5] we allow for modelling errors.

Nevertheless we believe that it is always interesting and sound to be able to describe the same phenomenon from both the finite and the infinite dimensional point of view, because this gives a much better control on the approximations introduced (cfr. [9]).

Whence, though not explicitly stated, most of this paper will deal with the infinite dimensional case, but for a few cases where the transition from finite to infinite dimensions is very complicated or even not yet achieved from a theoretical point of view.

Finally we can conclude this long introduction by claiming that almost all of the material presented here is not new, but a simple discussion of existing literature, specially [1], [2], [8], [11], [12]; the only somewhat new concept is that of stable estimable functional, which as a matter of fact gives a good introduction to the Bakus approach.

2. IMPROPERLY POSED PROBLEMS ACCORDING TO HADAMARD

According to a classical definition of Hadamard the problem

$$y = Ax$$

[2.1]

can be non wellposed in three ways.
a) **Non uniqueness of the inverse** $A^{-1}$: this happens when the equation

\[ Ax = 0 \]  \hspace{1cm} [2.2]

has a set of solutions $K$, called the kernel of $A$, i.e. there is a lack of injectivity.

It is easy to see that if $A$ is a linear continuous operator, $K$ is a closed subspace of $X$.

The existence of $K \neq \emptyset$ implies that if the equation [2.1] has a solution $x$ for a certain $y$, it admits also the general solution

\[ x = \{ \overline{x} + k; k \in K \} \]  \hspace{1cm} [2.3]

This lack of uniqueness is usually overcome by choosing a particular representative of the class [2.3]. The natural choice from the geometrical point of view is the element $x^+ \equiv \{ \overline{x} + k; |x^+|_x = \min \}$ (cfr. Fig. 2.1).

![Fig. 2.1](image)

This choice, which in the literature is known as the Moore-Penrose inverse,

\[ x^+ = A^+ y \]  \hspace{1cm} [2.4]

corresponds to a decomposition

\[ X = K + K^\perp \]

and to considering the equation [2.1] only in $K^\perp$ where it has a unique solution.
Example 2.1. (The inverse gravimetric problem)

Assume that we know the gravity field \( u \) outside the body \( B \) and we want to find a density \( \rho \), \( \rho \in L^2(B) \) that generates \( u \).

We can write

\[
u (P) = \int_B \frac{\rho(Q)}{I_{PQ}} dB_Q \quad \forall \ P \in \Omega = B^C \quad [2.5]
\]

This problem in general has not a unique solution (*) and its kernel is given by the mass distributions \( k \) with zero outer gravity field, i.e.

\[
\int_B \frac{k(Q)}{I_{PQ}} dB_Q = 0 \quad \forall \ P \in \Omega .
\]

One can prove that \( K \) is given by (cfr. [5])

\[
K \equiv \{ k = \Delta u \mid u \in H^{2,2}_o(B) \} , \quad [2.6]
\]

where \( u \in H^{2,2}_o(B) \) is defined according to

\[
\begin{cases}
u, \partial_j u, \partial_{ik} u \in L^2(B) \\
u|_S = 0 \quad \Rightarrow \quad u \in H^{2,2}_o(B) \\
\frac{\partial u}{\partial k} |_S = 0
\end{cases}
\]

This yields the general solution of [2.5] in the form

\[
\rho = \rho_\perp + k ,
\]

where \( k \in K \) and \( \rho_\perp \in K_\perp \), i.e.

\[
\int_B \rho_\perp k dB = 0 , \quad \forall \ k .
\]

Taking [2.6] into account, and applying a Green's identity, we see that

\[
\int_B (\Delta \sigma_\perp) u dB = 0 \quad \forall \ u \in H^{2,2}_o(B) ,
\]

so that \( \sigma_\perp \) has to be harmonic in \( B \).

This completely characterizes the inverse gravimetric problem in \( L^2(B) \).

(*) It is to be noted that there exist also conditions for the existence of a solution, i.e. it must be (cfr. [5])

\[
u = O(r^{-1}), \ u \in L^1_w(\Omega) .
\]
Example 2.2

Assume that \( y \) is an \( N \)-dimensional vector of observations on a field \( x(t) \in X \), so that we can write

\[
< a_i, x >_X = y_i \quad i = 1, 2, \ldots, N ;
\]

our aim is to find \( x \).

As it is obvious the observations [2.7] give no information on the components of \( x \) orthogonal to \( \text{Span} \{ a_i \} \) (\(*)(\); whence the solution of minimum norm of [2.7] has to belong to \( \text{Span} \{ a_i \} \), i.e.

\[
\bar{x} = \sum \lambda_i a_i .
\]

The components \( \lambda_i \) are determined by the normal equations

\[
\sum \lambda_i < a_i, a_k >_X = y_k .
\]

Example 2.3

The following is the one dimensional version of the inverse problem of seismology. Let's consider a vibrating string, fixed at the ends, with a propagation velocity \( v(x) \) satisfying the relation

\[
v^{-2}(x) = 1 + \varepsilon U(x) .
\]

Assume further the trade-off between the constant (1 in [2.8]) and \( U(x) \) in the definition [2.8] to be fixed by the position

\[
\int_0^1 U(x) \, dx = 0
\]

\( (*) \) This is the subspace of \( X \) spanned by de linear combinations of \( a_i \), i.e. by the vectors of the form \( \sum \lambda_i a_i \).
Let's now assume that the observed quantities are the eigenfrequencies of the string, which we arrange in a decreasing sequence \( \{ \omega_n^2 \} \).

Those frequencies satisfy the eigenfunction equation

\[
\begin{align*}
\begin{cases}
\partial_x^2 u_n + \omega_n^2 [1 + \varepsilon U(x)] u_n = 0 \\
u_n(0) = u_n(1) = 0
\end{cases}
\end{align*}
\]

Upon linearización, [2.10] yields

\[
\begin{align*}
\begin{cases}
u_n = \tilde{u}_n + v_n \\
\omega_n^2 = \tilde{\omega}_n^2 (1 + n_n) \\
\partial_x^2 v_n + \tilde{\omega}_n^2 v_n + \varepsilon U(x) \tilde{\omega}_n^2 \tilde{u}_n^2 + \tilde{\omega}_n^2 n_n \tilde{u}_n = 0 \\
v_n(0) = v_n(1) = 0
\end{cases}
\end{align*}
\]

multiplying by \( \tilde{u}_n \) and integrating by parts one gets the direct relation \( U(x) \rightarrow \{ \eta_n \} \), namely

\[
\eta_n = 2 \varepsilon \int_0^1 U(x) \tilde{u}_n dx
\]

Using \( 2 \sin^2 n\pi x = 1 - \cos 2 n\pi x \) and recalling (2.9), we receive finally

\[
\eta_n = -\varepsilon \int_0^1 U(x) \cos 2 n\pi x dx
\]

Since \( \{ \sqrt{2} \cos 2 n\pi X \} \) is an orthonormal system in \( L^2(0,1) \), we see that our inversion problem can be viewed as the solution of the equation

\[
\eta = AU
\]

with

\[
\eta = \{ \eta_n \} \in L^2 \equiv Y, \quad U \in L^2(0,1) \equiv X
\]

Since on the other hand \( \{ \sqrt{2} \cos 2 n\pi X \} \) is not complete in \( L^2(0,1) \) (to achieve completeness we should add to it the set \( \{ \sqrt{2} \cos (2n + 1) \pi X \} \)) we see that our problem in an example of improperly posed problems of type

a) The minimum norm solution is indeed \( \varepsilon U^+(x) = -2 \sum \eta_n \cos 2 n\pi X \).

b) Non existence of the inverse \( A^{-1} \) this happens when \( A \) is not surjective on \( Y \), so that there are \( y \) in \( Y \) that do not correspond to any \( Ax \ (x \in X) \).

Even of \( A \) is a continuous operator we cannot be sure that the range of \( A \) in \( Y \), \( R_A \), is a closed linear subspace, but, since this case will be treated in length under the next heading, we shall simply assume that it is so. The situation is then illustrated in Fig. 2.2.
As it is apparent there is no way to use \( A \) to go back from \( Y \) to \( X \) unless \( y \) is chosen to coincide with \( \hat{y} \), a vector in \( R_A \).

Whence, if we are given as datum a vector \( y_o \not\in R_A \), because of the various errors that enter into \( y_o \), we must define a law that associates to \( y_o \) some \( \hat{y} \in R_A \), i.e. \( \hat{y} = A \hat{x} \).

The simplest geometrical construction of \( \hat{y} \) is the so-called pure least squares solution, i.e. that \( \hat{y} \) which is at minimum distance from \( y_o \), namely

\[
||y_o - \hat{y}||_2^2 = \min \quad (\hat{y} \in R_A) \quad . \quad [2.11]
\]

As we all know from simple geometrical considerations this \( \hat{y} \) exists, is unique and more precisely it is the projection of \( y_o \) on \( R_A \); note should be taken that essential for the existence of the projection \( \hat{y} \) is that \( R_A \) is a closed subspace.

The variational equation associated with [2.11] gives the solution in terms of the «parameter» \( x \), i.e.

\[
\delta ||\delta||y_o - A \hat{x}||_2^2 = 2 < -A \delta x, y_o - A \hat{x}> - 2 < \delta x, A^* (y_o - A \hat{x}) >_x = 0, \forall \delta x
\]

so that we arrive at the normal system

\[
A^* A \hat{x} = A^* y_o \quad . \quad [2.12]
\]

Here essential use has been made of the concept of the adjoint \( A^*(Y \rightarrow X) \). We notice en passant that \( A^* \) is defined on all of \( Y \), as, owing to the continuity of \( A \),

\[
|<y, Ax>_x| \leq ||A|| ||y||_y ||x||_x \quad . \quad [2.13]
\]
6. On the foundations of various approaches...

so that \(<y, Ax> = \Phi(x)\) can be considered \(\forall y \in Y\) as a linear continuous functional on \(x\) and according to Riesz' theorem we have

\[
<y, Ax> = <\xi_y, x>_X ;
\]

[2.14]

we put by definition

\[
\xi_y = A^*y .
\]

[2.15]

According to [2.13] we find

\[
||\xi_y||_X \leq ||A|| ||y||_Y .
\]

i.e. \(A^*\) is a continuous operator too, defined on all of \(Y\).

We note also that the normal operator is a continuous positive operator in \(X\), since

\[
<x, A^*Ax> = ||Ax||_Y^2 \geq 0
\]

and that \(A^*A\) is invertible if \(A\) is injective; in the contrary case one can always factorize \(X\) and find a Moore Penrose inverse for \(A^*A\).

In conclusion we arrive at the estimation formulae

\[
\begin{cases}
\hat{x} = (A^*A)^{-1}A^* y_o \\
\hat{y} = A(A^*A)^{-1} A^* y_o
\end{cases}
\]

[2.16]

This approach to the inversion of case \(b)\) is only partly satisfactory because it doesn't take into account of the stochastic structure of the disturbance \(\delta y = y_o - y\). In particular no attention is paid to the directions along which there is more probability that \(y_o\) moves away from \(y\). This is usually described by the covariance operator of \(y_o\), i.e.

\[
C_{y_o} = E \{ (y_o - y)(y_o - y)' \}
\]

[2.17]

in a vector notation, or more precisely defined by

\[
E \{ <u, y_o - y>_Y <y_o - y, v>_Y \} = <u, C_{y_o} v>_Y ,
\]

[2.18]

where the covariance operator is made to depend explicitly on the scalar product in \(Y\).

An isotropic distribution of \(y_o\) implies \(C_{y_o} = \lambda I\), so that \(C_{y_o} (\neq 0)\) is a measure of the departure of \(y_o\) from isotropy.

Here it is preferable to assume \(Y\) to be finite dimensional, in order to avoid technical difficulties on which we shall shortly comment at the end of this section.
For the isotropic case we expect the probability density of \( y_o \) (the likelihood function \( L(y_o|y) \)) to depend only on the distance \( r = ||y_o - y||_y \), so that the principle [2.11] appears as a natural extension of the maximum likelihood principle, if we further assume that \( L(r) \) is a decreasing function of \( r \) (unimodal distribution).

Accordingly we accept the estimator

\[
\begin{align*}
\hat{y} &= P_A y_o \\
P_A &= A(A^*A)^{-1} & A = \text{orthogonal projection on } R_A
\end{align*}
\]  

[2.19]

as good for the (finite dimensional) isotropic case.

To pass to the anisotropic case we invoke a principle of invariance under linear transformations; more precisely we say that if

\[
\hat{y} = S y_o
\]

[2.20]

is a good estimate of \( y \) derived from \( y_o \) and if we want to give an estimate of \( z = D y \), starting from \( z_o = B y_o \), we have to use the estimate

\[
\hat{z} = S_D z_o
\]

[2.21]

defined in such a way that the estimate of \( y \) is again [2.20], i.e.

\[
\hat{z} = D \hat{y} = D S y_o = S_D z_o = S_D D y_o
\]

This relation, taken as an identity in \( y_o \), gives

\[
S_D D S D^{-1}:
\]

[2.22]

this is the law which gives the new estimator \( S_D \) as a function of the old \( S \), when the new variable \( z = D y \) is introduced.

With this in mind we try to solve the estimator problem:

\[
\text{find } \hat{y} = A \hat{z} \text{ from } y_o,
\]

when we know that

\[
C_{y_o, y_o} = \lambda Q, \quad Q \neq I
\]

we assume that \( Q \) is strictly positive definite.

First we take any transformation \( B \) defined in such a way that

\[
BQB^* = I
\]

for instance, \( B \) could be the symmetrical square root of \( Q \), or else \( B = (T^*)^{-1} \) where \( T \) is the factor of the Cholesky decomposition \( Q = T^* T \).
Then we define

\[ z = B y \]

so that the observation \( y_o \) is transformed into \( z_o = B y_o \). Owing to the law of propagation of the covariance we have

\[ C_{z_o z_o} = B C_{y_o y_o} B^* = \lambda I \]

i.e. \( z_o \) follows an isotropic distribution. Since from

\[ \hat{y} = A \hat{x} \]

we derive

\[ \hat{z} = B \hat{y} = (BA) \hat{x} \]

from to isotropy of \( z_o \) and [2.19] we find the estimator for \( \hat{z} \)

\[ \hat{z} = S z_o = P_{BA} z_o \]

To explicate \( P_{BA} \) we notice that

\[ B Q B^* = I \Rightarrow B^* B = Q^{-1} \]

whence

\[ S = P_{BA} = BA (A^* Q^{-1} A) A^* B^* \tag{2.23} \]

Finally we consider that

\[ y = B^{-1} z \]

so that setting \( D = B^{-1} \) in [2.22] and [2.23] we get

\[ \hat{y} = S_D y_o = A (A^* Q^{-1} A)^{-1} A^* Q^{-1} y_o \tag{2.24} \]

this entails an estimate of the unknown \( x \) of the form

\[ \hat{x} = (A^* Q^{-1} A)^{-1} A^* Q^{-1} y_o \tag{2.25} \]

This is the complete and correct solution of our estimation problem, that is known as the least squares estimator and that generalizes the simple 1.s. estimate [2.16].

It can be proved that [2.24], [2.25] are the estimates of minimum variance among unbiased linear estimators.
We can also remark that the result [2.24], [2.25] can be derived from a variational principle (least squares) as well as [2.16] was derived from [2.11]: in fact it is easy to show that the principle

$$<y_o - Ax, Q^{-1} (y_o - Ax)>_y = \min$$ \[2.26\]

leads to the normal equation

$$(A^* Q^{-1} A) \hat{x} = A^* Q^{-1} y_o$$

and ultimately to [2.25]

**Remark 2.1**

Usually the solution [2.25] is completed by some information on the stochastic behaviour of $\hat{x}$, namely by its covariance operator $C_{\hat{x} \hat{x}}$.

If we assume that

$$C_{y_o y_o} = \sigma_o^2 Q$$

($\sigma_o^2$ pure multiplicative factor) we find from covariance propagation and applying the definition of covariance operator,

$$C_{\hat{x} \hat{x}} = \sigma_o^2 (A^* Q^{-1} A)^{-1}$$ \[2.27\]

when $\sigma_o^2$ is known [2.27] solves the problem; if it is unknown it is still possible to estimate it by the formula

$$\sigma_o^2 = \frac{<y_o - Ax, Q^{-1} (y_o - Ax)>_y}{n-m}$$ \[2.28\]

where $n$ is the dimension of $Y$ and $m$ the dimension of $X$.

**Remark 2.2**

We can see from [2.26] that it is not trivial to extend the above reasoning to infinite dimensional spaces. In fact let $\{u_i\}$ be the complete orthonormal basis corresponding to the spectral decomposition of the operator $Q$, so that

$$Qu_i = q_i u_i,$$

then according to the definition of $C_{y_o y_o}$,

$$E\{<y_o - y, u_i> <y_o - y, u_i>\} = \lambda <u_i, Qu_i> = \lambda q_i \delta_{ij}.$$
If we assume for the sake of simplicity that $y_o$ is normally distributed with mean $y$, we may conclude that $\langle y_o - y, u_i \rangle_y$ are independent normal variables with zero mean and variance $q_i$.

Consequently if we try to compute the quadratic form [2.26] for the true average $y = Ax$, we find

$$\langle (y_o - y), Q^{-1} (y_o - y) \rangle_y = \sum_i \frac{\langle y_o - y, u_i \rangle_y^2}{q_i},$$

which is basically an infinite sum of independent $\chi^2$ with one degree of freedom.

But then, this sum attains the value $+\infty$ with probability 1, so that the expression [2.26] is meaningless for almost all $y_o$.

### Example 2.3

In a seismic network consisting of $N (>3)$ stations located of points $(x_i, y_i)$, we observe a sequence $\{\tau_i\}$ of arrival times for a seismic event of unknown coordinates $(x, y)$:

![Fig. 2.3](image)

assuming the elastic medium to be homogeneous and isotropic, we want to find the epicenter $(x, y)$ and the propagation velocity $v$.

The observation equations are

$$\tau_{oi} = \sqrt{\left[(x - x_i)^2 + (y - y_i)^2\right]} \left(\frac{1}{v} + \nu_i\right),$$

where $\nu_i$ is the random error in the $i$-th observation.
which should be linearized around approximate values \((\tilde{x}, \tilde{y}, \tilde{v})\) to give

\[
\delta z_{\alpha} = \frac{(x - x_i)}{D_i} \delta x + \frac{(\tilde{y} - y_i)}{\tilde{D_i}} \delta y - \frac{\tilde{v}}{\tilde{v}} \delta v
\]

\[
(D_i = \sqrt{(\tilde{x} - x_i)^2 + (\tilde{y} - y_i)^2}), \quad \tilde{v} = \frac{D_i}{\tilde{v}}).
\]

As we see this is a typical situation in which we have, from the very beginning, a small number of parameters to be estimated and we try to achieve more information than that strictly necessary to determine them.

The equations [2.29], supplemented by the information on the stochastic behaviour of \(\tau_{\alpha}\) (e.g. \(\tau_{\alpha}\) are independent and of different accuracies according to the receiving apparatus) allow for a least squares estimate by means of the equation [2.25].

Example 2.4. (Reduction of the parameter space)

Many times in geophysical problems we are in the situation that our unknown \(x\) is a function \(x(t)\) and \(X\) is infinite dimensional, while the vector \(y \in Y\) is finite dimensional, i.e. we have finitely many observations on \(x\). It seems that this situation should belong to case \(a\); however it can very well be that we have some rough information on \(x(t)\) saying for instance that our function is not likely to vary too much over a characteristic interval of length \(T\).

If we agree that this variation is negligible we can substitute for instance \(x(t)\) with a staircase function

\[
x(t) = \sum_{i=1}^{n} x_i \chi(t - iT) \quad \chi(t) = \begin{cases} 1 & |t| \leq T/2 \\ 0 & |t| > T/2 \end{cases}
\]

and our infinite dimensional unknown has been seduced to a finite dimensional one \(\{x_i, i = -n, \ldots, n\}\), i.e. the parameter space \(X\) has been reduced. Quite often the experiment is designed in such a way that the number of observations is larger than the number of unknowns which we like to determine, so that we have a control on the goodness of our hypothesis [2.30], since after solving by a least squares formula we can verify whether the estimated \(\hat{\sigma}^2\) is consistent with what we consider as an acceptable error.

For instance it is usually by a similar approach that applied geophysical interpretation is performed.
6. On the foundations of various approaches...

Remark 2.3

Obviously the reduction of dimensionality performed in [2.30] is among the most simple but certainly not the unique used in practice; for instance we could have reasons for thinking that in a harmonic analysis of $x(t)$, only the first $n$ frequencies are of importance and we could write

$$x(t) = a_0 + \sum_{j=1}^{n} (a_j \cos jt + b_j \sin jt)$$

the same reasoning applies to any basis of independent functions in $X$.

c) Instability (discontinuity) of the inverse $A^{-1}$

This is the most important and difficult case, which will occupy us in the rest of this paper: we shall assume that the range of the operator $A$, $R_A$, is a dense subspace of $Y$. In other words we known that the observation $y_o$, because of various disturbances, belongs to a space $Y$, but we know that the theoretical value $y = Ax$ has to belong to a subset dense in $Y$ but not coinciding with it.

In this situation $A^{-1}$ is defined only for a $y \in R_A$ and then it cannot be a continuous operator. Whence if $y_o \not\in Y$ but $y_o \in R_A$, we can find a sequence $y_n \in R_A$ (i.e. $y_n = Ax_n$) such that $y_n \not\to y_o$ because $R_A$ is dense in $Y$, but correspondingly neither $x_n$, nor any subsequence, can converge to any vector $\xi \in X$, since otherwise, owing to the continuity of $A$, we should have $y_{on} = A\xi$, i.e $y_o \in R_A$ against our hypothesis. It is exactly this circumstance that makes his case more difficult; in fact the more we approach $y_o$ with $y \in R_A$, the less may happen that we approximate the theoretical $x$.

In order to gain a better understanding of what happens in this case we shall take advantage of the spectral representation of the operator $A$ (cfr. Appendix 1)

$$Ax = \sum \sqrt{\lambda_i} u_i \langle \nu_i, x \rangle_X$$

[2.31]

where:

$\{\nu_i\}$ is an orthonormal system in $X$
$\{u_i\}$ is an orthonormal system in $Y$

Furthermore if we assume $A$ to be injective, so that $Ax = 0 \Rightarrow x = 0$, $\{\nu_i\}$ is complete in $X$; moreover if $R_A$ has to be dense in $Y$ also $\{u_i\}$ is complete in this space (*).

In other words we never have $\sqrt{\lambda_i} = 0$.

(*) The case that $\{u_i\}$ is not complete falls under the type a) of non well posedness, while the case that $\{u_i\}$ is not complete is of the type b).
Under those conditions, and assuming to have ordered $\sqrt{\lambda_i} \to 0$, for otherways $\sqrt{\lambda_i} \geq q > 0$ and

$$||Ax||_Y^2 \geq q^2 \sum_i <v_i,x>_X^2 = q^2 ||x||_X^2$$

so that $A^{-1}$ would be a bounded (continuous) operator defined on all of $Y$. This also says to us, in mathematical terminology, that $A$ is completely continuous between $X$ and $Y$.

Now assume that we have observed

$$y_o = Ax + u$$

[2.32]

where by hypothesis $v \notin R_A$; to be definitive assume $v$ to be a normal «white noise in $Y$», so that

$$<u,v>_Y = v_i = N[0,\sigma^2_v]$$

[2.33]

are independent normal variables with zero mean and equal variance $\sigma^2_v$.

Now if we take the product of [2.32] with $\mu$, we find

$$<\mu, v>_Y = \sqrt{\lambda_i} <v_i, x>_X + v_i$$

we rearrange as

$$<\mu, v>_Y = \frac{1}{\sqrt{\lambda_i}} <u_i, y_o>_Y - \frac{v_i}{\sqrt{\lambda_i}}$$

[2.34]

and interpret this equation as the observation equation of $<v_i, x>_X$ with noise

$$\frac{v_i}{\sqrt{\lambda_i}} = N[0, \frac{\sigma^2_v}{\lambda_i}]$$

[2.35]

As we see the knowledge of $y_o$ corresponds to the observation of the components $<v_i, x>_X$ with errors of increasing variance, $\sigma^2_v / \lambda_i \to +\infty$.

This is also why if we try to sum up

$$\sum_i <v_i, x>_X v_i$$

by using the «observed»

$$\frac{<u_i, y_o>_Y}{\sqrt{\lambda_i}}$$

we get an increasing error, which eventually causes the divergence of the series.

It is worthwhile to remark that if we had some rough information on $||x||_x$ we could limit the summation so as not to exceed what we consider a reasonable bound. It is any way clear that what we must find is a suitable rule to underweight the high frequencies to avoid the inflation of the error; furthermore this should be done possibly avoiding the explicit use of the spectral decomposition, since this cannot be achieved in practice for large of $n$, without an unfeasible numerical effort. This will be the object of the next paragraphs.

**Example 2.5**

The most famous improperly posed problem is to find a function $x(t)$ from observations of its integral

$$y_o(t) = \int_0^t x(\tau) \ d\tau + v \quad (*)$$

For instance $y(t)$ could be a geoid profile derived from altimetry and $x(t)$ the corresponding deflection of the vertical along the track.

If $x \in L^2(0,1)$, we must have $y(0) = 0$ and $y \in H^{12}(0,1)$, but the observed $y_o(t)$ is in a space of less regular functions, e.g. $Y = L^2(0,1)$.

Here

$$A = \int_0^t \cdot \ d\tau$$

is well known to be a compact operator in $L^2$ and its inverse is unbounded. That $R_A$ is dense in $L^2$ comes from noting that

$$<y, Ax>_Y = \int_0^t dt \ y(t) \ \int_0^t x(\tau) \ d\tau = \int_0^t d\tau \ x(\tau) \ \int_0^t y(t) \ dt = 0$$

for every $x \in L^2$, implies $y = 0$.

(*). As usual we denote by $H^{12}(A)$ the Sobolev space of functions square integrable together with their derivatives up to the order $S$ on the set $A$, endowed with the scalar product

$$<u,v>_{\nu^{12}(A)} = \int_A \{ u(t) v(t) + u^{(5)}(t) v^{(5)}(t) \} \ dt.$$
Example 2.6

Consider a disturbing body $B$ of known density contrast, like the one in Figure 2.4, and assume that we known the gravity anomaly on its upper face $A$,

$$g_p = k\rho \int \frac{1}{r_{pq}} \frac{\partial}{\partial z_q} dB = -k\rho \int \frac{1}{r_{pq}} dB =$$

$$= -k\rho \int \frac{dS_{Q_p}}{D_{pq}} + k\rho \int \frac{dS_{Q_p}}{\sqrt{D_{pq}^2 + H^2_{Q_p}}}$$  \hspace{1cm} \text{[2.36]}

we want to derive $S$ from $g_p$.

![Diagram of Example 2.6](image)

Fig. 2.4

This is a non linear problem, but it can be linearized starting from an approximate $H_{Q_p}$, to give

$$\delta g_p = K \delta H = -k\rho \int \frac{\delta H_{Q_p}}{(D_{pq}^2 + H^2_{Q_p})^3} \delta S_{Q_p}.$$ \hspace{1cm} \text{[2.37]}

This is a Fredholm equation of the first kind and the relevant integral operator is known to be compact in $L^2$: whence if we take the observed $\delta g_p$ in $L^2$ the solution of the problem is non wellposed, of the type c).
6. On the foundations of various approaches...

That the operator \( K \) defined by [2.37] has a range dense in \( L^2(A) \) is a somewhat more difficult result of which we only sketch the proof here. Assume \( f(P) \in L^2(A) \) to be orthogonal to \( R(K) \); then the double layer potential

\[
u(Q_0, -H_{Q_0}) = -k \rho \int_A f(P) \left( \frac{\partial}{\partial H_P} \frac{1}{r_{PQ}} \bigg|_{H_P = 0} \right) dS_P =
\]

\[
= -k \rho \int_A f(P) \frac{H_{Q_0}}{(D^2_{PQ_0} + H^2_{Q_0})^3} dS_P
\]

[2.38]

in such as to verify

\[
\int_A u(Q_0, -H_{Q_0}) \delta H_{Q_0} dS_{Q_0} = 0 \, , \ \forall \delta H
\]

[2.39]

i.e. the potential \( u \) is zero on the lower face of the body \( B \). On the other hand we also have by the very definition

\[
u(Q_0, H_{Q_0} = 0) = 0 \, , \quad Q_0 \notin A
\]

[2.40]

so that \( u \) is zero on all of \( \Omega \).

On the other hand \( u \) is antisymmetric with respect to \( H_{Q_0} \), so that \( u \) has to be zero on \( \Omega \) as well. Therefore, due to the strong analytic continuation property of harmonic functions, \( u \) has to be zero everywhere, except for the disk \( A \); but in this case also the density of the double layer, namely \( f(P) \), has to be zero, as it was to be proved.

Remark 2.4

We can observe that in both examples, [2.5], [2.6] we have used the same technique to prove the dense embedding of the range of the operator \( A \) into the Hilbert space \( Y \) (actually \( L^2 \)).

Therefore it seems worth to remark its general validity. The idea is essentially that:

if \( X \perp Y \), with \( A \) bounded,

and if \( Y \perp X \), is invertible, i.e.

\[
A^*y = 0 \Rightarrow y = 0 \, ,
\]

[2.41]

than the range of \( A \) is dense in \( Y \).
In fact if \( y \in Y \) and \( \langle y, Ax \rangle_x = 0 \ \forall \ x \), we also have \( \langle A^* y, x \rangle_x = 0 \ \forall \ x \Rightarrow A^* y = 0 \); therefore by [2.41] we can claim that \( y = 0 \), i.e. \( y \perp R(A) \) implies \( y = 0 \) or \( R(A) \) is densely embedded in \( Y \).

3. ON SOME SOLUTIONS OF THE INSTABILTY PROBLEM

In this paragraph we consider several proposals to take care of the instability of the inverse \( A^{-1} \), when solving the equation \( y = Ax \).

The discussion is carried along a line running from the more straight forward to the more sophisticated solution, showing that each new step can be interpreted as a generalization of the previous one, englobing its solution as a particular case.

The last floor of the construction, namely the so called informatic approach, leads to the problem of defining the rules to construct suitable probability distributions from the physical knowledge of the process described. This my question will be rather discussed in the paragraph on the basis of a solution commonly used in geodesy.

a) Tychonof regularization

The idea is as follows: we have to solve

\[
y = Ax
\]

and we have the a priori information that \( x \in X \), however due to disturbances of the model [3.1] the \( y \) we know, let us call it \( y_0 \), does not belong to \( R(A) \). On the other hand since \( R(A) \) is dense in \( Y \) we can very well find a \( y \in R(A) \) as close as we like to \( y_0 \) and to each such \( y \) there will correspond a \( x = A^{-1} y \). Naturally the closer \( y \) goes to \( y_0 \); the larger will become \( ||x|| \) and when eventually \( y \to y_0 \), \( x \) will never converge in \( X \) as otherways it would be \( y_0 \in R(A) \) too, contrary to our former hypothesis.

The idea of Tychonof is essentially to compromise between the will of approaching \( y_0 \) by \( y \) and the need to control the blowing of \( x \): this is realized by the variational principle

\[
||y_0 - Ax||_Y^2 + \alpha^{||x||^2}_X = \min
\]

The variation equation is in this case

\[
(A^* A + \alpha I)x = A^* y_0
\]

and its unique solution

\[
x = (A^* A + \alpha I)^{-1} A^* y_0
\]
If we use the spectral representation of \( A, A^* \) (see Appendix 1) it is easy
to realize that (3.4) can be written

\[
x = \sum_i \frac{\sqrt{\lambda_i}}{\lambda_i + \alpha} v_i <u_i, y_0>.
\]  

[3.5]

which shows that \( x \) is indeed unique and exists as element of \( X \) since \( \{v_i\} \) is an
o.n. base in \( X \), \( \sqrt{\lambda_i}/(\lambda_i + \alpha) \) is bounded (remember that \( \sqrt{\lambda_i} \to 0 \)),
\[ \sum_i <u_i, y_0>^2 y = ||y_0||^2 y <+\infty. \]

It is instructive to compare [3.5] with the straightforward solution
\( \tilde{x} = A^* y_0 \), which has meaning only if \( y_0 \in R(A) \): in terms of spectral
representation we have

\[
\tilde{x} = \sum_i \frac{1}{\sqrt{\lambda_i}} v_i <u_i, y_0> y
\]  

[3.6]

The comparison is graphically represented in Fig. 3.1 for the case
\( \lambda_i = 1/i^2 \), choosing for instance \( \alpha = 0.01 \).

As we see at low frequencies the spectrum of the regularized solution
follows closely that of the pure inverse, however when \( \alpha/\lambda_i \) becomes of the
order of the unity the regularized spectrum is downwarped smoothing the
effect of the noise components present in \( y_0 \), where as the raw inverse \( 1/\lambda_i \)
would magnify them.

Naturally crucial to get an «optimal» solution is to make the right choice
of the smoothing parameter \( \alpha \).
This can be achieved either on the basis of deterministic information on the norms of the relevant vectors (hard constraints) or using a stochastic information (soft constraints).

Useful to this aim is the following lemma.

**Lemma 3.1.** Let $x$ be given by $[3.5]$, then $||y_0 - Ax||, ||x||_x$ are monotonously increasing, respectively decreasing, functions of $\alpha$.

The proof is immediately achieved by inspecting the two functions in the spectral representation, namely

$$||y_0 - Ax||^2_x = \sum \left( \frac{\alpha}{\lambda_i + \alpha} \right)^2 < u_i, y_0 >^2_y \tag{[3.7]}$$

$$||x||^2_x = \sum \left( \frac{\sqrt{\lambda_i}}{\lambda_i + \alpha} \right)^2 < u_i, y_0 >^2_y$$

More precisely looking at $[3.7]$ we find that while $\alpha$ ranges from 0 to $\infty$, we have

$$0 \leq ||y_0 - Ax||^2 \leq ||y_0||^2 \tag{[3.8]}$$

with the last bound meaning $+\infty$ whenever $y_0 \in R(A)$.

With the help of this lemma it is not difficult to solve constrained problems like

$$\text{Min } ||y_0 - Ax||^2 \quad (||x||^2 \leq \eta) \tag{[3.9]}$$

proving that its solution is equivalent to the Tychonof regularization with the constant $\alpha$ determined by the value $\eta$.

In fact first of all, if $y_0 \notin R(A)$, which is the interesting case, the functional $||y_0 - Ax||_y$ cannot achieve an unconditioned minimum for a finite $x$, as otherways we would have $y_0 \in R(A)$. Whence the minimum $[3.9]$ is achieved on the boundary and can be localized by minimizing

$$||y_0 - Ax||^2 + \alpha||x||^2_x \tag{[3.10]}$$

and then computed by imposing

$$||x||^2_x = \eta \tag{[3.11]}$$

That $[3.11]$ is satisfied for one and only one value $\alpha$ descends from the lemma $[3.1]$ and from the remark that $||x||^2_x$ is a continuous function of $\alpha$. 
In a very similar way one can see that hard constraints like,

\[ ||y_0 - Ax||_2^2 = \omega \]

or

\[ \frac{||y_0 - Ax||_2^2}{||x||_2^2} = \rho \]

can be imposed and are capable of determining uniquely the smoothing parameter \( \alpha \).

**Remark 3.1**

According to the previous discussion we see that inequality constraints can be imposed on norms or norm ratios: this «hard» approach turns out to be also very stiff, as it is only seldom that we can maintain to have such a reliable information on these norms as to impose to the estimates to satisfy them exactly.

Things become better if we rather consider some stochastic information on \( y_0, x \).

Namely \( y_0 \) is assumed to be affected by a noise \( v \) which we know \( E\{||v||_2^2\} \); this excludes the hypothesis that \( v \) be a white noise on \( Y \). Similarly we can conceive that \( x \) is some correction, to be estimated, of a previously known approximate value of our unknown function: this can come along with the knowledge of a measure of the approximation, in the form of \( E\{||x||_2^2\} \). If two averages are given, we can interpret [3,2] as a kind of generalized least squares principle and this implies the choice

\[ \alpha = \frac{E\{||v||_2^2\}}{E\{||x||_2^2\}}. \tag{3.12} \]

**Remark 3.2**

*Sometimes instead of controlling directly \( ||x||_2 \) it is preferred a more general criterion which leads to the minimum principle,*

\[ \|y_0 - Ax\|_2^2 + \alpha ||Cx||_2^2 = \min, \tag{3.13} \]

*with \( C \) a suitable (often unbounded) operator. In this case the estimation formula is modified according to*

\[ x = (A^*A + \alpha C^*C)^{-1} A^* y_0. \tag{3.14} \]
The inverse exists and is bounded on condition that the null space of \( A \), \( N(A) \), has a void intersection with \( N(C) \),

\[
N(A) \cap N(C) = \phi
\]

b) **Dual Tychonof**

We propose here an approach which can be considered as the dual point of view of the Tychonof regularization; what we gain is insight into the effect of the disturbance present in \( y_\theta \), in the case (the most common) in which this is a white noise. The first point here is to define the concept of stable estimable functional.

Let us start again from the observation equation

\[
y_\theta = ax + \nu,
\]

and let's assume that rather than looking for an estimate of all \( x \), we are searching for an estimate of one functional only \( < \phi, x >_x \).

We will say that \( < \psi, y_\theta >_y \) is an unbiased stable estimator of \( < \phi, x >_x \) if

\[
E(< \psi, y_\theta >_y) = < \phi, x >_x, \quad \sigma^2(< \psi, y_\theta >_y) = 0(\sigma^2_y) \quad [3.15]
\]

To identify the stable estimable functionals let's multiply [3.14] by any functional \( \psi \) on \( Y \), so getting the “weak form”

\[
< \psi, y_\theta >_y = < A^* \psi, x >_x + < \psi, \nu >_y. \quad [3.16]
\]

if we can put

\[
A^* \psi = \phi \quad [3.17]
\]

we see at once that a suitable estimable of

\[
< \phi, x >_x \text{ is } < \hat{\phi}, x >_x = < \psi, y_\theta >_y ;
\]

in fact from [3.16], recalling the definition of white noise,

\[
E(< \psi, y_\theta >_y) = < A^* \psi, x >_x = < \phi, x >_x, \quad \sigma^2(< \psi, y_\theta >_y) = \sigma^2(< \psi, \nu >_y) = ||\psi||^2_2 \sigma^2_\nu, \quad [3.18]
\]

i.e. the definition of unbiased stable estimator is satisfied if \( ||\psi||^2_2 < +\infty \).
We see therefore that the condition for $\phi$ to be a stably estimable functional is

$$|| \psi ||_\gamma = || (A^*)^{-1} \phi ||_\gamma < +\infty,$$  \[3.19\]

i.e. $\phi \in R(A^*)$.

In particular in spectral terms, condition [3.19] reads

$$\sum_i \frac{\langle v_i, \phi \rangle}{\lambda_i} < +\infty.$$  \[3.20\]

The problem then, at first sight, is that with the information $y_0$ affected by the white noise $v$, we can only have an unbiased estimate with finite variance of $\langle \phi, x \rangle$, if condition [3.19], or [3.20], is fulfilled: everytime that we find instability, it is because we try to estimate a forbidden functional.

**Example 3.1**

The elementary but typical example of this phenomenon is in the inversion of

$$y_0(t) = \int_0^t x(\tau) \, d\tau + v,$$

when we consider $A \int_0^t \cdot \, d\tau$ as an operator from $L^2(0,1)$ into $L^2(0,1)$.

Note that in this case $v$ is the usual white noise (namely the derivative of a Wiener process) so that the variance of a point value $v(t)$ is unbounded $\sigma^2(v(t)) = +\infty$.

Then, not only $x(\overline{t})$ for a given $\overline{t}$ is not estimable, what is very reasonable since the functional of evaluation at $t$ is itself unbounded on $L^2$, but not even a functional like

$$\int_{\overline{t}}^{\overline{t} + \Delta} x(\tau) \, d\tau$$

is estimable in a stable way. In fact this could happen only if the equation

$$A^* \psi = \int_0^t \psi(\tau) \, d\tau = \phi = \begin{cases} 0 & 0 \leq t < \overline{t} \\ 1 & \overline{t} \leq t < \overline{t} + 1 \\ 0 & \overline{t} + 1 \leq t \leq 1 \end{cases}$$

had a solution $\psi \in L^2(||\psi||_1 < +\infty)$.

But this is not so, as we know that derivatives of strongly discontinuous functions are non-square integrable distribution.
The reason why this happens is essentially that the "natural" estimator
\[ y_0(\tau + \Delta) - \bar{y}_0(t) = \int_{\tau}^{\tau+\Delta} x(\tau) \, d\tau + w(\tau + \Delta) - w(\tau) \]
has an infinite variance, as we remarked above.

So what can we do if the functional we need to estimate is unstable? The answer is that we must relax the requirement of unbiasedness.

If we go back to [3.16], we see that if \( A^*\psi \neq \phi \) the estimate is biased, namely
\[ b = E[\langle \psi, y_0 \rangle] - \langle \phi, x \rangle = \langle A^*\psi - \phi, x \rangle, \]
so that the total mean square error of \( \langle \psi, y_0 \rangle \) is
\[ \xi = E[\langle \psi, y_0 \rangle - \langle \phi, x \rangle^2] = b^2 + \sigma_x^2 \langle \psi, \nu \rangle = \langle A^*\psi - \phi, x \rangle^2 + \sigma_x^2 ||\psi||^2. \tag{3.21} \]

We cannot minimize directly [3.21] as otherwise we would get an estimator \( \hat{\psi} \) depending on \( x \): however by the Schwarz inequality we see that
\[ \xi^2 \leq ||A^*\psi - \phi||^2 ||x||^2 + \sigma_x^2 ||\psi||^2, \tag{3.22} \]
so that it becomes meaningful, though suboptimal, to minimize the second member.

This results in the variational equation
\[ (AA^* + \alpha I) \psi = A\phi \tag{3.23} \]
i.e. in the estimate
\[ \langle \phi, x \rangle = \langle (AA^* + \alpha I)^{-1} A\phi, y_0 \rangle. \tag{3.24} \]

Formula [3.24] is therefore equivalent to estimate \( x \) by the biased estimator \( \hat{x} \) given by
\[ \hat{x} = A^*(AA^* + \alpha I)^{-1} y_0 = (AA^* + \alpha)^{-1} A^* y_0. \tag{3.25} \]

The second identity in [3.25] is purely algebraic and carries us back to the Tykhonof formula with a choice of the parameter \( \alpha \) that is essentially identical with [3.12], although we have now justified it for a white noise \( \nu \), too.
6. On the foundations of various approaches...

C) *The Bayesian approach*

This approach is well known since several years and applied in both geophysical and geodetic problems. Among geophysicists many authors have contributed to it and as the most representatives we want to cite G. Bakus; among geodesists this theory has been throughly developed, specially in conjunction with the approximation theory in Hilbert Harmonic Spaces, and we want to mention the work of T. Krarup and H. Moritz.

We approach it starting from the dual formulation which we have already worked out at the point b), in a rigorous way, up to the formula of the mean square error [3.21].

Please notice in that formula that the average has been taken only on the population of the noise $v$.

In the Bayesian approach all variables are stochastic and subject to some a priori distribution and the problem is to find the a posteriori distribution, conditional to the observation $y_0$.

In this approach then $x$ is also a stochastic process which we assume to have a priori mean equal to zero: that corresponds to consider $x$ as the "correction" to the best a priori known solution. Furtheron, as minimal stochastic a priori information on $x$, we assume to know the covariance operator $C_{xx}$ of $x$.

Moreover it is assumed that $x$ and $v$ are stochastically independent variables.

We remind that by the very definition

$$E\{<\phi, x>_x|<\omega, x>_x\} = <\phi, C_{xx}\omega>_x$$  \hspace{1cm} [3.26]

and that the case

$$C_{xx} = \sigma_x^2 I$$  \hspace{1cm} [3.27]

corresponds to a white noise, $x$.

From the physical point of view [3.27] means that we want to represent a situation in which we have information (the same) on the variability of the coefficients of $x$ on an orthonormal basis,

$$\sigma^2 \{<v_x, x>_x\} = \sigma_x^2$$  \hspace{1cm} [3.28]

and we have no correlation information among them.

In this situation the proper definition of mean square error, will be, like [3.21]

$$\xi^2 = E_x, \{[<\psi, y>_y - <\phi, x>_x]^2\} =$$

$$= E_x \{<A^* \psi - \phi, x>_x^2\} + \sigma_v^2 ||\psi||^2_y$$
where however the average has now to be taken independently on both $x$ and $y$ populations.

The result, using (3.25), is

$$
\xi^2 = \langle A^* \psi - \psi, C_{xx} (A^* \psi - \phi) \rangle_x + \sigma_y^2 \|\psi\|^2_y \quad (3.29)
$$

Minimization of (3.29) is a straightforward exercise of algebra, leading to

$$
\psi = (AC_{xx}A^* + \sigma^2_y C_{xx}^{-1})^{-1} AC_{xx} \phi \quad (3.30)
$$

in the light of another algebraic identity this can be written

$$
\psi = A (A^* A + \sigma^2_y C_{xx}^{-1})^{-1} \phi \quad (3.31)
$$

It is interesting to observe that writing the estimation equation in the form

$$
\langle \phi, \tilde{x} \rangle_x = \langle \psi, y_0 \rangle_y = \langle \phi, (A^* A + \sigma^2_y C_{xx}^{-1})^{-1} A^* y_0 \rangle_x
$$

one realizes that a global estimate of $x$ is implied, with estimator

$$
\tilde{x} = (A^* A + \sigma^2_y C_{xx}^{-1})^{-1} A^* y_0. \quad (3.32)
$$

It is also instructive to compute the mean square error from (3.31): after some manipulations one gets

$$
\xi^2 = \sigma_y^2 \langle \phi, (A^* A + \sigma^2_y C_{xx}^{-1})^{-1} \phi \rangle_x \quad (3.33)
$$

Indeed (3.33), corresponding to the unique solution of the variational equation, gives a real minimum if it is finite, since a positive definite bilinear functional can have at most one minimum as stationary point.

That (3.33) is finite can be seen from the inequality (*)

$$
\sigma_y C_{xx}^{-1} \leq \sigma_y^2 C_{xx}^{-1} + A^* A
$$

which, inverted, shows that

$$
(A^* A + \sigma_y^2 C_{xx}^{-1})^{-1} \leq \frac{1}{\sigma_y^2} C_{xx} \quad (3.34)
$$

i.e. that this is a bounded operator.

(*) We are here implicitly supposing that $C_{xx}$ is invertible.
Remark 3.3

As we see from [3.32] we are back again to some kind of Tychonof regularization, more precisely to formula [3.14] if the right correspondences are set up. However in the Bayesian approach the nature of the operator $C_{\approx}^{-1}$ appearing in the estimation equation [3.32], as well as the value $\sigma^2_v$ are strictly fixed by our optimality principle, while in the deterministic approach they were left to some subjective interpretation.

This is a typical feature of stochastic approaches.

Naturally the equivalence is limited to the simplest case treated here, where all computations were performed by minimizing a quadratic operator: this corresponds to a minimum a priori knowledge on $x$ consisting in the mean ($=0$) and the covariance operator $C_{xx}$.

It is well known that our minimum principle is equivalent to a maximum likelihood, when all the relevant distributions are, e.g., normal whereas for more general distributions the estimation equations would become non linear and therefore non equivalent to the Tychonof approach.

Remark 3.4 (The infinite norm paradox problem)

The estimation error computed by [3.33] is by definition

$$\xi^2 = E\{<\phi, \hat{x} - x>^2_x\},$$

i.e., so to say, the projection of the vector error $e = \hat{x} - x$ in the direction $\phi$.

One could ask what is the total amount of error measured by $E\{||e||^2_x\}$.

We obviously have

$$E\{||e||^2_x\} = \sum E\{<v, \hat{x} - x>^2_x\} = \sigma^2_v <v, (A^*A + \sigma^2_v C_{xx}^{-1})^{-1}v>_x^2 = \sigma^2_v \text{Tr}(A^*A + \sigma^2_v C_{xx}^{-1})^{-1};$$

from [3.34] we see that whenever the covariance operator of $x$ is nuclear, $\text{Tr} C_{xx} < +\infty$, also the total error is bounded.

However if we set $C_{xx} = \sigma^2_x I$, corresponding to an isotropic priori information (white noise), we see that $E\{||e||^2_x\}$ becomes infinite as the eigenvalues of the operator of which we are to compute the trace, are

$$(\lambda_i + \frac{\sigma^2_v}{\sigma^2_x})^{-1} \geq \frac{\sigma^2_v}{\sigma^2_x} = \text{const}.$$  

On the other hand this result should not be surprising at all since $e = \hat{x} - x$ and if $x$ is white noise on $X$, the realizations of $\{x\}$ (as a stochastic process) do not belong to $X$, with probability 1,

$$P\{||x||_X < +\infty\} = 0.$$
On the contrary the regularized solution $\hat{x}$, i.e. our a posteriori knowledge of the unknown field as opposed to the a priori knowledge $x$, is in $X$. In fact using [3.20] instead of [3.31] and

$$\hat{x} = C_{xx} A^* (AC_{xx} A^* + \sigma^2_v) -1 y_0$$

instead of [3.31] and the fact that the covariance operator of $y_0 = Ax + v$ is

$$C_{x_0 y_0} = AC_{xx} A^* + \sigma^2_v I,$$

due to the stochastic hypothesis on $x, v$, we find

$$E[||\hat{x}||^2_X] = Tr C_{xx} A^* (AC_{xx} A^* + \sigma^2_v)^{-1} AC_{xx},$$

which is bounded even if $C_{xx} = \sigma^2_v I$, as far as $\sigma^2_v \geq \sigma$ and $\Sigma \lambda_i < \infty$ (*).

So the estimation process using the fresh information $y_0$ has indeed regularized the a priori white noise $\{x\}$ transforming it into a regular process in $X$; the error $e$ therefore cannot be in $X$ itself.

This is not too bad however because even if $E[||e||^2_X]$ is infinite, the estimation error of every single functional is finite according to [3.33]: this is a price one has to pay if one wants to play with white noises.

The last question one can pose is whether, under these conditions, it is of any use to supply the a priori information on $x$ in the form of a white noise.

The answer goes along with observing what happens to $\xi^2$ when we let $C_{xx} = \sigma^2_x I$ and $\sigma^2_v \to \infty$, since this seems a good model of noninformative a priori information.

From [3.33], setting $C_{xx} = \sigma^2_x I$ and going to the spectral representation, we find

$$\xi^2 = \sigma^2_v \sum \frac{<v_i, \phi>^2_X}{\lambda_i + (\sigma^2_v / \sigma^2_x)} ;$$

as it is apparent, this is a monotonous increasing function of $\sigma^2_x$, tending to

$$\xi^2 \to \sigma^2_v \sum \frac{<v_i, \phi>^2_X}{\lambda_i}$$

when $\sigma^2_x \to \infty$. This expression is in general $\infty$ if $\phi \notin R(A^{*-1})$, i.e. in all the interesting cases.

This proves the effectiveness of the a priori information introduced by assuming $x$ as a white noise with variance «density» $\sigma^2_x$.

(*): This condition means that $A$ is so «smoothing» as to send a white noise on $X$ into a regular process in $Y$. 
6. On the foundations of various approaches...

Remark 3.5

In all our formulas we have taken $v$ as a pure white noise. More generally one could consider that $v$ has its own covariance operator $C$: in this case the mean square error [3.29] is changed into

$$
\xi^2 = \langle A^* \psi - \phi, C_{xx} (A^* \psi - \phi) \rangle_x + \langle \psi, C\psi \rangle_y.
$$

Minimization of this, leads to the estimation equations

$$
\langle \phi, x \rangle_x = \langle (AC_{xx}A^* + C)^{-1} AC_{xx}\phi, y_0 \rangle_y
$$

$$
\hat{x} = C_{xx} A^* (AC_{xx}A^* + C)^{-1} y_0 = (A^* C^{-1} A + C_{xx}^{-1})^{-1} A^* C^{-1} y_0,
$$

the last step being of purely algebraic nature.

d) The informatic approach

This approach is the more recent and probably the more defined: it is related to the names of Vellette and A. Tarantola.

It can be numerized in a «quick version» through some peculiar steps. In order to avoid technical difficulties we shall assume in what follows that $a$ and $y$ are finite dimensional vectors.

1) The physical state of a system is always described by a «state space» with a measure, giving the degree of likelihood of each single state point; this measure can be or not a probability distribution. For instance an equal ignorance of states on an unbounded (*) set can be represented by a constant measure but not by a probability distribution.

2) If we call $X$ the state space and $x$ its elements, a state of perfect knowledge is represented by a distribution of the form $C \delta(x-x_0)$. On the contrary a perfectly non-informative state is given, for cartesian variables like the cartesian coordinates of a particle, by a constant measure over all the euclidean space: the same is true for all cartesian vectors.

More difficult is the situation with «non cartesian» variables. The following example is proposed for comparison.

Example 3.2

Let $\mu(x) \, dx$ be a measure of the informational content of the interval $dx$ located at the point of coordinate $x$; we understand that by changing coordinates with a translation

$$
\bar{x} = x + a
$$

we achieve a new picture, described by the new function $\bar{\mu}(\bar{x})$.

(*) More precisely of infinite euclidean measure.
If $x \rightarrow \bar{x} = x + a$ and $dx = dx$ defines the same set in the two coordinate system, we stipulate that the information content must be the same

$$\bar{\mu}(\bar{x}) \, d\bar{x} = \mu(x) \, dx = \mu(\bar{x} - a) \, d\bar{x} \quad [3.36]$$

We say that $\mu(x)$ is not informative for the cartesian variable $x$ if the action of the translation is none on $\mu$, i.e. if $\mu(x) = \mu(x)$, what entails

$$\mu(x) = \bar{\mu}(\bar{x}) = \mu(x - a),$$

i.e.

$$\mu(x) = \text{const.} \quad [3.37]$$

In this way we have clarified that $\mu(x)$ is non-informative with respect to the group of transformations of $x$ of the type translation. In an analogous way assume now that we like to consider a state for a variable which is the modulus of a vector, e.g.

$$v = \frac{|dx|}{|dt|},$$

furthermore assume we like to describe a state which is non-informative with respect to the modulus of $|dx|/|dt|$, i.e. which is the same for $v$ and for $\bar{v} = cv$.

Again we have $(v, dv) \rightarrow (\bar{v} = cv, d\bar{v} = c \, dv)$,

$$\bar{\mu}(\bar{v}) \, d\bar{v} = \mu(v) \, dv = \mu\left(\frac{\bar{v}}{c}\right) \frac{d\bar{v}}{c};$$

If we stipulate that $\bar{\mu}$ is non-informative on the scale change

$$\bar{\mu}(\bar{v}) = \mu(\bar{v})$$

we find

$$\mu(v) = \mu\left(\frac{\bar{v}}{c}\right) \frac{1}{c},$$

i.e.

$$\mu(v) = \frac{\text{const}}{v}. \quad [3.38]$$

(*) This essentially means that the average of every continuous function of the state $x$, $g(x)$ is given by $L\{g(x)\} = g(x_0)$.
This shows that, for example, from \( v \) we can construct a «cartesian» variable by defining

\[
w = \log \frac{v}{v_0} ;
\]

in fact

\[
\mu_w(w) = \mu_v(v) \frac{dv}{dw} = \frac{\text{const} / v}{1/v} = \text{const} .
\]

The above example should serve to clarify that a non-informative state, is specified by the transformation group with respect to which the state has be invariant, or non-informative.

3) The following 3 axioms define the operation of «combination of informative states» which has to be applied to represent the improved information which is obtained by combining two independent information sources.

Let \( P_1, P_2 \) represent two information states and \( M \) the non-informative state: \( P_1 \odot P_2 \) is an operation of combination of the two states if

I) \( P_1 \odot P_2 = P_2 \odot P_1 \)

II) \( P_1(A) = 0 \Rightarrow P_1 \odot P_2(A) = 0 \)

III) \( P_1 \odot M = P_1 \)

We want to represent this composition in terms of measure densities.

From II) we see that \( P_1 \odot P_2 \) is absolutely continuous with respect to \( P_1 \); so if \( P_1 \) has its own density \( f_1(x) \) we can say that, for every measurable \( A \)

\[
P_1 \odot P_2(A) = \int_A \phi_2(x) f_2(x) \, dx .
\]

Moreover, considering the property I) we have aswell.

\[
P_2 \odot P_1(A) = \int_A \phi_1(x) f_1(x) \, dx = \int_A \phi_2(x) f_1(x) \, dx ,
\]

i.e.

\[
\phi_1 f_2 = \phi_2 f_1 \Rightarrow \frac{f_1}{\phi_1} = \frac{f_2}{\phi_2} = \omega(x) .
\]

Consequently we can write

\[
P_1 \odot P_2(A) = \int_A \frac{f_1(x) f_2(x)}{\omega(x)} \, dx .
\]
non exploiting the property III) we find

\[ P_2 \otimes M(A) = \int_A \frac{f_1(x) \mu(x) \omega(x)}{\omega(x)} \, dx = P_1(A) = \int_A f_1(x) \, dx \ . \]

This implies

\[ \omega(x) = \mu(x) \ . \]

and we finally get the general form of the «combined» densities

\[ P_1 \otimes P_2(A) = \int_A \frac{f_1(x) f_2(x)}{\mu(x)} \, dx \ . \quad [3.39] \]

4) The system under observation and analysis is described by a couple of variables \((y, x)\) one of which is accessible to the observations, i.e. \(y\), the other, \(x\), is our unknown.

We want to describe here the information, that is given by the model relating \(y\) to \(x\): this is essentially described by the conditional distribution \(\theta(x|y)\), while the model itself does not give any specific information on \(x\).

Henceforth we can write for the density we are looking for:

\[ \theta(x; y) = \theta(x|y) \mu_s(x) \ . \quad [3.40] \]

The most common model, through not unique, is one where \(y\) depends on \(x\) only through its conditional mean, so that

\[ y = g(x) + s \]

with \(s\) independent from \(x\).

We shall assume that \(g(x)\) is for can be approximated by a linear relation,

\[ y = Ax + s \quad [3.41] \]

with \(s\) distributed as \(\theta_s(s)\).

In this way \(s\) appears as the «model error» of the simple linear model

\[ y = Ax \ . \]

Summarizing, the most commonly encountered distribution \([3.30]\) is of the form

\[ \theta(x; y|A) = \theta_s(x - Ax) \mu_s(x) \ . \quad [3.42] \]
where the conditioning to $A$ stresses that the information contained in [3.42] comes from the model, $A$, and the stochastic behaviour of the «model error» $s$.

5) The observational process is not a direct sampling of $y$, but rather that with a disturbance $v$ added,

\[ y_0 = y + v ; \]

$v$ is assumed to be independent from $y$ and its distribution is considered as known, $\theta_N(v)$. In many instances $v$ is modelled by a white noise.

By applying the Bayes theorem we can write

\[
\theta(y|y_0) = \frac{\theta_N(y_0-y) \theta_Y(y)}{\int \theta_N(y_0-y) \theta_Y(y) \, dy} \approx \theta_N(y_0-y) \theta_Y(y),
\]

where we neglect proportionality factors since all these will be eliminated by a final renormalization.

We include in the process of observation any possible a priori information on $x$ described by the distribution $\theta_X(x)$ and we assume this to be independent from the observation $y_0$.

Under these hypotheses the information coming from the overall observation process is described by a state of the form.

\[
\theta(x, y|y_0) = \theta_N(y_0-y) \theta_Y(y) \theta_X(x). \tag{3.44}
\]

Note that in this way any possible a priori information on $X$ has entered into the state [3.44] and not into [3.42]. Naturally in the case we have no priori information on $x$ we shall put also in [3.44].

\[
\theta_X(x) = \mu_x(x). \tag{3.45}
\]

6) The information state [3.42] is now «combined», with the information state [3.44] coming from the observation process, according to the rule [3.39], to give the density.

\[
\theta(x, y|y_0; A) = \frac{\theta_N(y_0-y) \theta_Y(y) \theta_X(x) \theta_S(y-Ax) \mu_x(x)}{\mu_Y(y) \mu_x(x)} \tag{3.46}
\]

Since our aim is to know the posterior distribution of $x, f_x(x)$ we should rather compute the marginal density of [3.45]

\[
f_x(x) = \theta_X(x) \int \frac{\theta_N(y_0-y) \theta_X(y-Ax) \theta_Y(y)}{\mu_Y(y)} \, dy. \tag{3.46}
\]
To get directly the «best estimate» of \( x \), given all the information, we can average \( x \) over [3.46]; before doing that however it is necessary to normalize \( f_r(x) \) by dividing the second member by its integral. The final result is

\[
\hat{x} = \frac{\int dx \ x \ \theta_x(x) \int dy \ \frac{\theta_N(y_0-y) \ \theta_S(y-Ax) \ \theta_Y(y)}{\mu_Y(y)}}{\int dx \ \theta_x(x) \int dy \ \frac{\theta_N(y_0-y) \ \theta_S(y-Ax) \ \theta_Y(y)}{\mu_Y(y)}}
\]

**Remark 3.6**

In the most common case, in which we assume to have no special prior information on \( y \), we can put \( \theta_Y(y) = \mu_Y(y) \) and simplify [3.46] accordingly. The result is a typical Bayesian formula for the average of \( x \) on the posterior distribution of \((x, y)\) given \( y_0 \).

Even more we want to show that if all the distributions are normal, the estimates become linear and identical with those discussed under the point c. In fact assume that

\[
\theta_N = \text{const} \ e^{-1/2 \sigma_N^2} C_{\nu_0}
\]

\[
\theta_S = \text{const} \ e^{-1/2 \sigma_S^2} C_y
\]

then

\[
\int dy \ \theta_N(y_0-y) \ \theta_S(y-Ax)
\]

can be consider as the convolution of two normal distributions, one with mean 0 and covariance \( C_{\nu_0} \), the other with mean \( Ax \) and covariance \( C_y \). The result is another normal, in the outgoing variable \( y_0 \), with mean \( Ax \) and covariance \( C = C_{\nu_0} + C_y \), i.e.

\[
\int dy \ \theta_N(y_0-y) \ \theta_S(y-Ax) = \text{const} \ e^{-1/2(y_0-Ax)^2 + 1/2(y_0-Ax)C_{\nu_0} + 1/2(y_0-Ax)C_y}
\]

Assume now that also \( x \) is normally distributed, with mean zero and covariance \( C_{xx} \); whence the posterior distribution of \( x \) has the form

\[
f_x(x) = e^{-1/2 x^T C_{xx}^{-1} x - 1/2 (y_0-Ax)^T C_{\nu_0} + 1/2 (y_0-Ax)^T C_y (y_0-Ax)}
\]

This is again a normal distribution and its mean coincides with the maximum likelihood estimate, i.e. the \( \hat{x} \) obtained by minimizing

\[
x^T C_{xx}^{-1} x + (y_0-Ax)^T C_{\nu_0} + 1/2 (y_0-Ax)^T C_y (y_0-Ax) = \min
\]
the result is

\[ \hat{x} = (A^* C^{-1} A + C_{xx}^{-1})^{-1} A^* C^{-1} y_0 \]

which proves to be identical with [3.35].

**Example 3.2**

We will not fail to discuss the solution at least of the problem formulated in Example 3.1, also because this is one of the few examples of which we are able to give the solution in closed form.

So we start with the observation equation

\[ y_0(t) = \int_0^t x(\tau) d\tau + \nu(t), \quad 0 \leq t \leq 1, \]

and we assume that \( \nu \) is a white noise of variance \( \sigma_\nu^2 \) on the interval \([0,1]\) and that following a Bayesian approach, the a priori covariance operator of \( x \) is

\[ C_{xx} = \sigma_x^2 I. \]

The estimate \( \hat{x}(t) \) we are looking for, given by [3.31], is the solution of the equation

\[ (A^* A + \alpha I) \hat{x}(t) = A^* y_0, \quad \alpha = \sigma_\nu^2 / \sigma_x^2, \quad [3.47] \]

with

\[ A = \int_0^1 \cdot dt \quad \text{and} \quad A^* = \int_1^t \cdot dt, \]

as discussed in the Example 3.1. It is not a trivial Remark that \( A^* \) depends from the scalar product of the space on which \( A \) is defined and in particular its definition depends from the interval \([0,1]\) on which our processes are defined.

Equation [3.47] is explicitly written as

\[ \int_1^t d\tau \int_0^\tau d\nu \hat{x}(\eta) + \alpha \hat{x}(t) = \int_0^1 y_0(\tau) d\tau. \quad [3.48] \]

This equation can be solved, first by assuming that \( y_0 \) is as mild a function as we need, thus getting an explicit representation formula, then extending this formula to a process \( y_0 \) of the type we are interested in.

Then, assuming suitable regularity, we differentiate twice [3.48] to get

\[ \hat{x}''(t) - \frac{1}{\alpha} \hat{x}(t) = -\frac{1}{\alpha} y_0'(t). \]
the general solution of this equation is
\[ \tilde{x}(t) = -\frac{1}{2\alpha} \int_0^t e^{-\frac{1}{\sqrt{\alpha}}|t-\tau|} y_0'(\tau) d\tau + Ae^{-\frac{1}{\sqrt{\alpha}}t} + Be^{-\frac{1}{\sqrt{\alpha}}t} = \]  
\[ -\frac{1}{2\alpha} \int_0^t \text{sgn}(t-\tau) e^{-\frac{1}{\sqrt{\alpha}}|t-\tau|} y_0'(\tau) d\tau + A e^{-\frac{1}{\sqrt{\alpha}}t} + B e^{-\frac{1}{\sqrt{\alpha}}t}. \]  

[3.49]

The second form will be used here. The two constants \( \tilde{A}, \tilde{B} \) are determined by imposing two conditions which the solution of [3.48] has to satisfy if it is regular, namely
\[ \tilde{x}(1) = 0, \quad \lim_{t \to 0} \tilde{x}'(t) = -\frac{y_0(0)}{\alpha}. \]

The final result can be written, for instance, in the form
\[ \tilde{x}(t) = -\frac{1}{2\alpha} \int_0^t \text{sgn}(t-\tau) e^{-\frac{1}{\sqrt{\alpha}}|t-\tau|} y_0'(\tau) d\tau + \]
\[ + \frac{1}{2\alpha} \frac{1}{Ch} \left[ e^{\frac{1}{\sqrt{\alpha}}} \int_0^t Ch \left( \frac{1-\tau}{\sqrt{\alpha}} \right) y_0'(\tau) d\tau + \right] \]
\[ + e^{\frac{1}{\sqrt{\alpha}}(t-1)} \int_0^t Sh \left( \frac{-\tau}{\sqrt{\alpha}} \right) y_0'(\tau) d\tau \].

[3.50]

This is one possible analytical form of the solution of [3.48]: since \( y_0(\sigma) \) appears in it only under an integral sign with square integrable (as a matter of fact continuous) functions, we can extend this formula to processes containing white noises too, like \( y_0 \).

We can remark that the «regularized» solution \( \tilde{x}(t) \) is in fact a square integrable, even a continuous, function with probability 1: this is due, as pointed out in Remark 3.4, to the fact that
\[ Av = \int_0^t v(t) dt = w(t) \]
is in fact a realization of a Wiener process, and this is even continuous with probability 1.

Another point interesting to remark, is that if we analyze the behaviour of [3.50] for \( \alpha \to 0 \), we see that the terms in square brackets tend to zero, whereas
\[ -\frac{\text{sgn}(t-\tau)}{2\alpha} e^{-\frac{1}{\sqrt{\alpha}}|t-\tau|} \to -\delta'(t-\tau), \]
so that we would get \( \tilde{x}(t) \to y_0'(t) \) for a regular \( y_0 \).
6. On the foundations of various approaches...

A final comment on the dependence of the solution on the interval of definition.

For instance if we extend it to the whole real axis, with the observation equation

\[ y_0(t) = \int_{-\infty}^{t} x(\tau) d\tau + v, \quad -\infty < t < +\infty \]  

we can represent the problem of determining a probability density from the empirical distribution function. For this specific problem however we should have \( y_0(t) \to 1 \) for \( t \to +\infty \) : to express this problem in a Hilbert space, e.g. in \( L^2(-\infty, +\infty) \), we need only to subtract from [3.51] a known relation of the type

\[ F(t) = \int_{-\infty}^{t} f(\sigma) d\sigma, \]

to get

\[ y_0 - F = \int_{-\infty}^{t} (x - f) d\tau + v. \]

For this equation we can assume for instance that both, the unknown \( x - f \) and the observable \( y_0 - F \) belong to \( L^2(-\infty, +\infty) \).

After noting that in this case

\[ A^* = \int_{-\infty}^{+\infty} d\tau, \]

one can follow the same reasoning as above, however arriving at the solution

\[ \hat{x} - f = -\frac{1}{2\alpha} \int_{-\infty}^{+\infty} \text{sgn}(t - \tau) e^{-\frac{1}{2\alpha} |t - \tau|} (y_0 - F) d\tau, \]

as this is the only function in the family (3.49) which is bounded for both \( t \to \pm \infty \).

4. The Colocation Approach: A Contribution From Geodesy

Geodesy, considered as a separate discipline from geophysics, has undergone the same (or may be a parallel) evolution as sister sciences, concerning the estimation theory for solutions of improperly posed problems.

However the specific geodetic contribution to the application of methods presented in § 3), is related mainly to the way in which the «a priori» information can be supplied.

In particular let's go back to the observation model.

\[ y_0 = Ax + v : \]

\[ [4.1] \]
in order to apply a Bayesian estimation formula, in either form
\[ \hat{x} = C_{xx} A^* (A C_{xx} A^* + \sigma^2 I)^{-1} y_0 \quad [4.2] \]

or
\[ \hat{x} = (A^* A + \sigma^2 C_{xx})^{-1} A^* y_0 \quad [4.3] \]
we must know \( C_{xx} \), beforehand.

We will show that in suitable cases under the a priori hypothesis that the
process \( \{x\} \) satisfies some invariance principle, its covariance operator
becomes accessible from the estimable covariance of \( y_0 \), \( C_{xx} \), on condition
that we know the form of \( C_{xx} \), e. g. as in [4.2], [4.3] we have \( C_{xx} = \sigma^2 I \).

The steps to achieve this result are:

1) Let \( t_u \) be a coordinate transformation group and \( T_u \) the corresponding
transformation group in \( X \), defined by
\[ T_u x(t) = x(t_u t), \quad x \in X; \quad [4.4] \]
let us further assume that \( T_u \) is unitary in \( X \), i. e. that the \( X \)-scalar product
is invariant under \( T_u \).

\[ \langle T_u u, T_u v \rangle_X = \langle u(t_u t), v(t_u t) \rangle_X = \langle u, v \rangle_X \Rightarrow T_u^* T_u = I \]

As examples one can think of the translation group for \( L^2 \) functions
defined over all the euclidean space, or the rotation group and the \( L^2 \) scalar
product for functions defined on the unit sphere, \( \sigma \).

2) Let us define a transformation group in \( Y \), more precisely in
\( R(A) \subset Y \), with the rule
\[ V_u A x \overset{\text{def}}{=} A T_u X; \quad [4.5] \]
that \( V_u \) is a representation of the same group follows from
\[ T_u T_\beta = T_\gamma \Rightarrow V_\gamma A x = A T_\gamma x = A T_\alpha T_\beta x = V_\alpha A T_\beta = V_\alpha V_\beta A x \Rightarrow V_\gamma = V_\alpha V_\beta . \]

Generaly speaking one cannot be sure that \( V_u \) is a bounded operator in \( Y \)
as it is not defined on all of \( Y \) but just on the dense subset \( R(A) \). As a mather
of fact \( \{V_u\} \) is unitary on \( R(A) \) endowed with the graphnorm of the operator
\( A^{-1} \), namely
\[ \| V_u Y \|_{R(A)} = \| A^{-1} V_u (Ax) \|_X = \| A^{-1} A T_u x \|_X = \| x \|_X = \]
\[ = \| A^{-1} Ax \|_X = \| Ax \|_{R(A)} = \| y \|_{R(A)} . \]
6. On the foundations of various approaches...

However the group \( \{ V_w \} \) is bounded and even unitary when \( Y = X \) and \( T_w \) commutes with \( A \); in this case have indeed

\[
[A, T_w] = 0 \Rightarrow V_w = T_w.
\]

Examples of this type are

\[
X = Y = L^2(-\infty, +\infty),
\]

\[
Ax = \int_{-\infty}^{\infty} A(t - \tau) X(\tau) \, d\tau,
\]

a convolution operator and \( \{ V_w \} \) the translation group, or \( X = Y = L^1(\sigma) \),

\[
Ax = \left( -\frac{\partial}{\partial r} \right)^{-1} x^{(*)}
\]

and \( T_w \) the rotation group.

3) We assume, and this is our real a priori hypothesis, that \( X \) is stationary and ergodic in the covariance with respect to the group \( \{ T_w \} \). In order to give a precise meaning to this statement also when the covariance operator \( C_{xx} \) is not represented by an ordinary function, we use the weak definition, namely:

— we say \( C_{xx} \) to be stationary with respect to \( \{ T_w \} \) if the covariance

\[
< x, \ T_w^{(*)} \phi >_x, \ < x, \ T_w^{(*)} \psi >_x \text{ is the same as that to } \ < x, \phi >_x, \ < x, \psi >_x,
\]

\( \forall \phi, \psi \in X \).

Accordingly we must have

\[
E\{<x, \ T_w^{(*)} \phi >_x < x, \ T_w^{(*)} \psi >_x\} = < T_w^{(*)} \phi, \ C_{xx} \ T_w^{(*)} \psi >_x =
\]

\[
= < \phi, \ C_{xx} \psi > = E\{<x, \phi >_x < x, \psi >_x\}.
\]

Since \( T_w^{(*)} = T_w^{-1} \), we find

\[
T_w \ C_{xx} \ T_w^{(*)} = C_{xx} \ [T_w, \ C_{xx}] = T_w \ C_{xx} - C_{xx} \ T_w = 0
\]

\((*)\) This is the inverse of the Neuman operator to the sphere, also known as the Hotine operator in geodesy.
— We say that \( X \) is ergodic in the covariance if the empirical estimate

\[
\dot{<} \phi, C_{xx} \dot{\psi} > \overset{a.s.}{\approx} \lim_{\Omega \to G} \frac{1}{\mu(\Omega)} \int_\Omega d\mu_\omega < x, T_\omega^* \phi >_X < x, T_\omega^* \psi >_X \tag{4.6}
\]

coincides almost surely with the time covariance operator \( C_{xx} \).

To explain [4.6] we must specify that the measure \( \mu \) should be the Haar invariant measure on the group \( G \), i.e. such that for any set \( \Omega \subset G \)

\[
\mu (t_\omega \Omega) = \mu (\Omega t_\omega) = \mu (\Omega) : \tag{4.7}
\]

this is also called the uniform measure on \( G \).

The limit in [4.6] is taken over a family of compact sets \( \Omega \), such that \( \Omega / G \).

This when the group \( G \) is not itself compact, as in this last case \( \mu (G) = + \infty \). For instance the translation group or \( R \) is isomorphic to \( R \) too: whence \( \mu \) is simply the Lebesgue measure, and \( \mu (G) = + \infty \). On the contrary if \( G \) is the rotation group it is well known that \( \mu (G) = 8 \pi \).

4) Furthermore we assume that \( \nu \) is stationary and ergodic with respect to \( \{ V_\omega \} \); this essentially implies that \( \{ V_\omega \} \) is unitary, so that \( V_\omega \nu \) is again a noise on \( Y \). As it has been remarked above, this happens for instance when \( \{ T_\omega \} \) communities with \( \mathcal{A} \) so that \( V_\omega = T_\omega \).

5) Finally we assume that, as a consequence of the hypothesis of independence between \( x \) and \( \nu \), the following cross-covariance estimate is zero almost surely

\[
\lim_{\Omega \to G} \frac{1}{\mu(\Omega)} \int_\Omega d\mu_\omega < x, V_\omega^* \phi >_Y < x, T_\omega^* \nu >_X = 0 \tag{4.8}
\]

Under all these conditions it is possible to estimate from one realization only, \( y_\omega \), the covariance function of the process.

In this way we find

\[
< \phi, C_{xy} \psi >_y = \lim_{\Omega \to G} \frac{1}{\mu(\Omega)} \int_\Omega d\mu_\omega < Ax + \nu, V_\omega^* \phi >_Y < Ax + \nu, V_\omega^* \psi >_Y =
\]

\[
= \lim_{\Omega \to G} \frac{1}{\mu(\Omega)} \int_\Omega d\mu_\omega \{ < AT_\omega x, \phi >_Y < \nu, \phi >_Y \} \cdot \{ < AT_\omega x, \psi >_Y < \nu, \nu >_Y \}
\]

\[
= \sigma^2 < \phi, x, A^* \psi >_x + \sigma^2 < \phi, \psi >_Y =
\]

\[
= < \phi, A C_{xx} A^* \psi >_x + \sigma^2 < \phi, \psi >_Y ,
\]
which is indeed another way of writing the well known covariance relation
\[
C_{\gamma_0\gamma_0} = A \, C_{xx} \, A^* + \sigma_v^2 \, I. \tag{4.9}
\]

However the important point here is that these quantities are now accessible by an empirical estimate.

One way of proceeding in our approach is to solve \([4.9]\) for \(C_{xx}\): first of all \(\sigma_v^2\) is estimated either from the spectral behaviour of \(C_{\gamma_0\gamma_0}\) or by representing \(C_{\gamma_0\gamma_0}\) by a Kernel \(C(t_1, t_2)\) and looking at the jump along the diagonal \(t_1 = t_2\).

Subsequently we come to the estimate of \(C_{xx}\): usually a simple parametric solution is sought, in such a way that by fixing a small number of parameters, the behaviour of \(C_{\gamma_0\gamma_0}\) is, at least roughly, reproduced.

**Remark 4.1**

It might seem contradictory to try to find a solution of \([4.1]\) with some regularization technique, avoiding a sharp inversion of \(A\), and to this aim proposing to solve \([4.9]\), what implies an inversion of both \(A\) and \(A^*\).

The point here is that we can find a «rough» solution \(\overline{C}_{xx}\) of \([4.9]\) and then use it in \([4.3]\) without affecting too much the estimate \(\overline{x}\).

In other words this method is a stepwise non linear solution of the original problem, where \(\overline{C}_{xx}\) is only roughly estimated, nevertheless providing a fine estimate of \(\overline{x}\), due to its robustness against errors in \(C_{xx}\).

In particular this is true when \(x\) is closed in such a way that it is only a «correction» to the best a priori known field.

To see this we can simply differenciate \([4.3]\) to get the linearized error relation
\[
\delta \dot{x} = -(A^* A + \sigma_v^2 \, C_{xx}^{-1})^{-1} \sigma_v^2 \, \delta \left[ C_{xx}^{-1} \right] (A^* A + \sigma_v^2 \, C_{xx}^{-1})^{-1} A^* y_0 = \\
= (A^* A + \sigma_v^2 \, C_{xx}^{-1})^{-1} \sigma_v^2 \, C_{xx}^{-1} \delta C_{xx} \, C_{xx}^{-1} \dot{x}. \tag{4.10}
\]

For the sake of simplicity, let's assume the error \(\delta C_{xx}\) to be only proportional to \(C_{xx}\).
\[
\delta C_{xx} = \eta C_{xx} : \\
\delta \dot{x} = \eta (A^* A + \sigma_v^2 \, C_{xx}^{-1})^{-1} \sigma_v^2 \, C_{xx}^{-1} \dot{x}
\]
and, since by an elementary positivity argument,
\[
\| (A^* A + \sigma_v^2 \, C_{xx}^{-1})^{-1} \sigma_v^2 \, C_{xx}^{-1} \| < 1
\]
we eventually get

$$||\delta \hat{x}|| \leq \eta ||\hat{x}|| .$$

This seems an acceptable amount of error when $\hat{x}$ is itself a small correction.

**Remark 4.2**

It might also seem fatal at this stage by hypotheses we stipulated from 1) to 5) might be extremely cumbersome and never met in practice.

We wont to stress that our a priori hypothesis of invariance and ergodicity of $x$ with respect to $T_\alpha$, reflects essentially our decision that there is no a priori evidence in the data of particular features in one part with respect to another in the area where the information is supplied.

When this basic «statistical regularity» is met, we create a stochastic process by applying to a single realization $x$ a random element of the group $T_a$. If correspondingly we find we are able to compute $C_{x \omega}$ from one realization only, we can assume there is one such process which admits $C_{x \omega}$ as covariance and with such a distribution that it can be estimated consistently. An example of this kind is met in the treatment of the global anomalous gravity field; this example cannot be worked out here and we send the interested reader to [8].

Another example however will be presented hereafter.

By the way the delicate step in this approach is in assuming that, apart from the correlated signal $Ax$, only a white noise $v$ is present in the observation model [4.1]. This hypothesis is crucial since it is through it that we can separate the identity from the compact component in [4.9] and thus reach the sought estimate of $C_{x \alpha}$.

In other words we must assume a priori that [4.1] is correct.

In this respect it is to be stressed that an isolated model error in [4.1] (i.e. an error which affects the data in a small part of their domain), might not be so dangerous in that it might not affect so much the covariance equation [4.9] because of the averaging process.

In this case the estimation of $C_{x \alpha}$ is sufficiently robust and the whole procedure can be applied.

More dangerous would be the presence, beyond $v$, of another correlated signal distributed over the whole area, as, due to mismodelling, this would be ascribed to the behaviour of $x$.

**Example 4.1**

Let us consider a simple example of very classical nature, namely the solution of the equation

$$y_0(t) = \int_{-\infty}^{+\infty} A(t-\tau) x(\tau) d\tau + v(t)$$

[4.11]
on the real line; here \( v(t) \) is a white noise, i.e. the derivative of a Wiener process.

Moreover we want to study [4.11] taking \( X = Y = L^2(-\infty, +\infty) \).

Let us assume further that by inspection of the function \( y_0(t) \) we find no evidence of unhomogeneous behaviour so that it becomes reasonable to define the invariance group \( T_w \) as

\[
T_w x(t) = x(t + \omega) ;
\]

this is obviously unitary on \( L^2 \) and commutes with \( A \), so that we have also

\[
V_w = T_w .
\]

Now let us compute an empirical covariance function of \( y_0 \), i.e.

\[
C_{y_0 y_0}(\tau) = \lim_{A \rightarrow -\infty} \frac{1}{2A} \int_A^A d\omega \int y_0(t) y_0(t + \tau) \, dt .
\]

The correspondence between [4.13] and the definition [4.6] is as follows: \( \Omega \) here is the interval \((-A, A)\) and we can write

\[
\lim_{A \rightarrow -\infty} \frac{1}{2A} \int_A^A d\omega \int y_0(t + \omega) \phi(t) \, dt \int y_0(t' + \omega) \psi(t') \, dt' = \\
\int \int dt \, dt' \phi(t) \psi(t') \lim_{A \rightarrow -\infty} \frac{1}{2A} \int_A^A d\omega \, y_0(t + \omega) y_0(t' + \omega) \]

\[
= \int \int dt \, dt' \phi(t) \psi(t') C_{y_0 y_0}(t - t') = \langle \phi, C_{y_0 y_0} \psi \rangle .
\]

This shows that [4.14] is the kernel of the interpol operator \( C_{y_0 y_0} \). It is to be stressed that being statically homogenous \( y_0 \) cannot be an \( L^2 \) function so that the limit in [4.15] is the classical limit for processes stationary and ergodic in the covariance (cfr. [8]).

If we compute [4.14] on a discrete data set, we may expect a covariance for instance like the one in Fig. 4.1, with more or less large fluctuation of the empirical values around a smooth function. We interpret this as a decomposition into a process generated by the compact operator \( A \), giving rise to the smooth part of the covariance function, plus a noise represented by the jump at the origin. It is to be noticed that the finite jump \( \sigma^2 \) is related to the discretization lag \( \Delta \) and it can be assumed to tend to infinity like \( 1/\Delta \).
The same reasoning as in [4.15] can be applied to the integral
\[ \int A(t - \tau) x(\tau) d\tau , \]
showing that the corresponding operator is
\[ \langle \phi, A c_{xx} A^* \psi \rangle = \int \int dt \, dt' \phi(t) \psi(t') \cdot \{ \int \int d\tau \, d\tau' A(t - \tau) A(t' - \tau') C_{xx}(\tau - \tau') \} . \]

Essentially [4.17] in first proved for stepwise functions \( \phi, \psi \) which are constant over intervals of length \( \delta \); in this way we can write integrals in the form
\[ \int \phi(t - \omega) dv(t) = \Sigma \phi_i \left[ W(t_{i+1} + \omega) - W(t_i - \omega) \right] . \]
6. On the foundations of various approaches...

where $W$ is the Wiener process of which $v$ is the derivative. Subsequently we can partition the interval $(-A, A)$ into $2N$ intervals of width $\delta$ and transform $\lim_{A \to \infty} \lim_{N \to \infty}$ into $\lim_{N \to \infty}$. Whence [4.17] is translated into the simpler proposition

$$(\Delta W_i = W(t_{i+1}) - W(t_i))$$

$$\lim_{N \to \infty} \frac{1}{2N} \sum_{k=1}^{N} \phi_k \Delta W_k \sum_{j=1}^{j} \psi_j = \lim_{N \to \infty} \mathbb{E}_N = \sum_{i} \phi_i \psi_i \delta \cdot \sigma^2 \quad [4.18]$$

This, in turn, is seen to hold true by exploiting the well known properties of the Wiener process, i.e.

$$E \{ \Delta W_i \Delta W_k \} = \delta \sigma^2 \delta_{i,k}$$

$$\sigma^2 \{ \Delta W_i \Delta W_k \} = \delta \sigma^4 (1 + \delta_{i,k}) ;$$

accordingly one can easily prove that

$$E \{ X_N \} = \delta \sigma^2 \sum \phi_i \psi_i$$

$$\sigma^2 \{ X_N \} \to 0 \quad (N \to \infty) ,$$

what entails almost sure convergency.

From [4.18] we can go back to [4.17] by taking advantage of the dense embedding of stepwise functions in $L^2$.

Crucial to the success of this approach is the capability of modelling covariance functions, like the one in Fig. 4.1, by a suitable choice of the covariance $C_{xx}$.

This is specific for each operator $A$ and depends from the nature of the problem under analysis.

Example 4.2

This is in a sense a particular case of the example 4.1, only it is two dimensional rather than onedimensional.
The problem is classical in gravity interpretation with two layers of constant density \( \rho_0 \) and \( \rho_1 \). The separation surface is assumed to be of the form

\[
z (P) = H + \zeta (P)
\]

where \( P \) spans the horizontal plane. The observed quantity is the gravity anomaly \( \Delta g \) on the upper surface \( S_q \), hereafter assumed to be a plane for the sake of simplicity.

The linearized relation between \( \zeta \) and \( \Delta g \) is

\[
\Delta g (P) = k \Delta \rho \int dS_q \frac{\xi_q}{[D_{pq}^2 + H^2]^{1/2}}
\]

[4.19]

\[
\Delta \rho = \rho_1 - \rho_0
\]

\[
D_{pq}^2 = (x_q - x_p)^2 + (y_q - y_p)^2
\]

Please note that the equation [4.19] is the same as [2.36]; however, a stationary behavior of \( S \) is considered, whereas in the example [2.6] the separation surface was assumed to form a closed body.

Equation [4.19] is analogous to [4.11] in that it is a convolution integral equation, it is more general in that it holds on a plane rather than on a line.

Moreover we seen that the kernel [4.19] depends on the plane distance \( D_{pq} \) only so that it naturally commutes with the rototranslation group.

In this case therefore the machinery of homogeneous-isotropic fields in \( 2D \) can be applied. In particular we can expect the covariance of \( \Delta g \) and of \( \zeta \), to be functions of \( D_{pq} \) too.

A good trick to build suitable covariance functions in this case, arises from remarking that if we put \( \zeta = h J_0 (\alpha r) \) in [4.19] we receive (*)

\[
\Delta g = h e^{-\alpha H} J_0 (\alpha r) .
\]

Accordingly if we assume \( C_{xx} \) to have a kernel

\[
C_{xx} = h^2 J_0 (\alpha r) ,
\]

we get as \( AC_{xx}A^* \) an integral operator with kernel

\[
AC_{xx}A^* = h^2 e^{-\alpha H} J_0 (\alpha r) .
\]

[4.20]

By combining different \( J_0 \) functions, like in a Fourier-Bessel series, with positive coefficients

\[
C_{xx} = \sum h^2 J_0 (\alpha r) ,
\]
we get the corresponding covariance form

$$AC_xA^* = \sum h_i^2 e^{-2\pi i n} J_0(\alpha_x r).$$  \hspace{1cm} [4.21]$$

This seems to be a fairly good model which has proven to be useful in practical cases: as a matter of fact many times even a single $J_0$ has been perfectly adequate to perform the interpolation of the empirical covariance of real gravity data.

**Remark 4.3**

In order to make the comparison between the prediction formula [4.2] and the collocation formula more straightforward, one can think of discretizing the operator $A$ in such a way that it becomes a vector of functionals applied to the unknown «signal» $x$ (**):

$$Ax = \begin{vmatrix} L_{1x} \\ L_{2x} \\ \vdots \end{vmatrix}.$$

In this way we can identify

$$C_{xx}A^* = \{ C_{x,Lx} \}$$

$$AC_{xx}A^* = \{ C_{Lxx,Lx} \}$$

so that [4.2] becomes

$$\hat{x} = \sum C_{x,Lx} \{ C_{Lxx,Lx} + \sigma_v^2 \delta_{ik} \}^{-1} y_{ik},$$

which is indeed the classical collocation formula.

## 5. CONCLUSIONS

Naturally no definitive conclusion can be known on to whether a method is «better» than another; this is even a wrong question.

It seems to the author that solving improperly posed problems is a «neverending story» like many other basic problems of mathematical physics.

(*) This is nothing but the solution of the Laplace equation in cylindrical coordinates; in fact it is easy to verify that [4.19] is essentially nothing but the solution of the Dirichlet problem for the half-space.

(**) This is in any way the natural form in which real measurements are received.
So what we can do is to summarize the state of the art and try to trace back the theoretical relation between different approaches, which was the main target of this paper.

In particular we have analyzed the transition from the classical deterministic Tychonov approach to the stochastic approach.

The first is found to be suitable to work with hard constraints but rather stiff in its conclusions; the second is more supple and can treat in a natural way a large variety of disturbances, like white noises.

In this environment the informatic approach seems to be more complete in the sense of being able to describe in a very fine way all the possible deviations from the elementary model $y = Ax$.

Apart from the theoretical basis the informatic and the Bayesian approaches are quite close one another: both of them require that some valuable a priori information be supplied, for instance in the form of one or more covariance operators.

When these are not available and only the size of the various processes is roughly known, it is customary to use covariance operators simply proportional to the identity: this leads to the same estimation formulas as with Tychonov's method.

However this choice is by no means neutral and although there is a certain stability of the estimate with respect to errors in the covariance, many times a simple «diagonal» approximation can be too rough.

In this respect a help comes from collocation theory where suggestions are made as to how to derive estimates from the data themselves and some a priori hypotheses of invariance: in this way the method becomes a mixture of a Bayesian and a non linear approach.

The drawback here is in that it seems impossible to discriminate between an a priori covariance structure of the unknown $x$ and the possible covariance structure of the mismodelling error, unless the latter is already known, for instance it is assumed to be of the white noise type.

Appendix 1

We want to justify briefly the spectral representations of $A$ and $A^*$ utilized in § 3.

To this aim we start from one of the best known theorems of functional analysis.

**Theorem A 1.1**

Let $K$ be a linear, selfadjoint, compact operator in the Hilbert space $X$, then its spectrum is real positive, it is done by isolated eigenvalues $\{\lambda_n\}$ with orthogonal finite dimensional eigenspaces, the sequence $\{\lambda_n\}$ accumulates on

(*) This would mean that the stochastic process $x$ has its realizations in $X$ almost surely.
0 and if \( 0 \notin \{\lambda_n\} \) there is an orthonormal complete sequence of eigenfunctions \( \{v_n\} \),

\[
Kv_n = \lambda_n v_n. \tag{A.1.1}
\]

If \( \{\lambda_n\} \) is suitably ordered admitting repetitions, we have \( \lambda_n \notin 0 \) and with each \( \lambda_n \) we associate only one eigenfunction \( v_n \).

Now let us assume that \( A \) is continuous and compact from \( X \) into \( Y \) with \( R(A) \) densely embedded in \( Y \); let us further assume that

\[
Ax = 0 \implies x = 0, \tag{A.1.2}
\]

i.e. \( A \) is invertible.

Under such conditions there is a sequence \( \{\lambda_n\} \) and a complete orthonormal system (in \( X \)) \( \{v_n\} \) such that

\[
(A^*A) \ v_n = \lambda_n \ v_n, \quad \lambda_n \notin 0. \tag{A.1.3}
\]

In fact if \( A \) is bounded, \( A^* \) is defined on all of \( Y^* \) and therefore bounded too; if \( A \) is compact, \( A^*A \) is compact too; moreover \( A^*A \) is clearly selfadjoint; finally \( \lambda_n = \langle v_n, A^*A v_n \rangle_x = ||A v_n||^2_x \) and \( A \) is invertible.

Therefore \( \text{[A.1.3]} \) follows from \textit{Theorem A.1.1}.

Now let us observe that under the above conditions \( A^* \) is invertible too: in fact

\[
A^*y = 0 \implies \langle A^*y, x \rangle_x = \langle y, Ax \rangle_y = 0 \quad (\forall \ x) \implies y = 0
\]

(since \( R(A) \) is dense in \( Y \)).

We can therefore define

\[
\sqrt{\lambda_n} \ u_n = A v_n; \tag{A.1.4}
\]

applying to both members the operator \( A^* \) we find the sister relation

\[
A^* u_n = \sqrt{\lambda_n} \ v_n; \tag{A.1.5}
\]

The system \( \{v_n\} \) is othonormal and complete in \( Y \). In fact

\[
\sqrt{\lambda_n} \lambda_m \ u_n, u_m \rangle_y = \langle A v_n, A v_m \rangle_x = \langle v_n, A^* A v_m \rangle_x = \lambda_m \delta_{n,m}; \tag{A.1.6}
\]

moreover, using the invertibility of \( A^* \),

\[
\sqrt{\lambda_n} \langle y, u_n \rangle_y = \langle y, A v_n \rangle_x = \langle A^* y, v_n \rangle_x = 0, \quad \forall \ n
\]

\[
\implies A^* y = 0 \implies y = 0.
\]
Accordingly we see that
\[ A x = A \sum <x, v_n> v_n = \sum \sqrt{\lambda_n} <x, v_n> u_n \]  
[A.1.7]
\[ A^* y = A^* \sum <y, u_n> y = \sum \sqrt{\lambda_n} <y, u_n> y v_n \]  
[A.1.8]
as we wanted to prove.

Appendix 2

We will use this appendix to define and construct second order stochastic processes on a Hilbert space \( X \).

Let us be given a basis \( \{u_i\} \) of \( X \) and, just to be simple, assume it is orthonormal: let's assume further that a stochastic process is defined by the sequence of variables,
\[ \omega_i = <x, u_i>; \]  
[A.2.1]
that the symbol \( <x; u_i> \) is a real scalar product (*) it is not possible to say at this stage and even false in general.

We assume \( \{ \omega_i \} \) to be a sequence of \( L^2 \) variables with zero average and we put
\[ E \{ \omega_i \} = 0 \quad \forall i \]  
[A.2.2]
\[ E \{ \omega_i \omega_j \} = C_{ij} \]

Let us now make the hypothesis that the matrix \( C \) is bounded in \( L^2 \) norm, i.e.
\[ \sum_{i,j=1}^{N} C_{i,j} u_i u_j \leq C \sum_{i=1}^{N} u_i^2, \quad \forall N, \forall \{u_i\}. \]  
[A.2.3]

Under these conditions the process can be extended to the whole Hilbert space \( X \), since we can put
\[ \omega_u = <x, u > = \lim_{N \to \infty} \sum_{i=0}^{N} \omega_i <x, u_i>; \]  
[A.2.4]
The limit in [A.2.4] exists in \( L^2 (\{\omega_i\}) \) since
\[ E \{ \sum_{i=0}^{N} \omega_i <x, u_i> \} = 0 \]
\[ E [\sum_{i=0}^{N} \omega_i <x, u_i> ]^2 = \sum_{i,j=0}^{N} C_{i,j} u_i u_j \leq C \sum_{i=0}^{N} u_i^2 \to 0. \]
6. On the foundations of various approaches...

Once we have verified the consistency of [A.2.4] we see that the span \( \{ \omega_\mu \mid u \in X \} \) coincides as a matter of fact with the original \( L^2(\{\omega_i\}) \).

In particular we can always define the covariance between two such variables, corresponding to their \( L^2(\{\omega_i\}) \) product,

\[
E \{ \omega_n \omega_i \} = \lim \sum_{i,j} C_{i,j} u_i v_j. \tag{A.2.5}
\]

Since

\[
|\sum C_{i,j} u_i v_j| \leq C \sum u_i^2 \sum v_j^2 \leq C \|u\|_X \|v\|_X
\]

we find that the limit in [A.2.5] is a continuous bilinear functional in \( u,v \), so that a bounded operator is defined, such that

\[
E \{ \omega_n \omega_i \} = E \{ <x,u> <x,v> \} = <u,C_{x} v>
\]

with

\[
\|C_{x}\| \leq C. \tag{A.2.7}
\]

Since \( \{C_{ij}\} \) is a symmetrical operator we have

\[
<u,C_{xx} v> = <v,C_{xx} u>
\]

i.e. \( C_{xx} \) is selfadjoint. Since

\[
<u,C_{xx} u> = E \{ \omega_i^2 \} \geq 0
\]

\( C_{xx} \) is positive definite. This is by definition the covariance operator of the process \( x \).

At this point one could ask whether the realizations of \( x \) belong to \( X \) or not.

The answer is contained in the following Lemma A.2.1: \( x \) belongs to \( X \) almost surely if \( C_{xx} \) is a nuclear operator, i.e.

\[
Tr C_{xx} < +\infty. \tag{A.2.8}
\]

The proof lies essentially in the following chain of identities

\[
Tr C_{xx} = \sum_i <u_i,C_{xx} u_i> = \sum_i E \{<x,u_i>^2\} = \sum_i E \{\Sigma <x,u_i>^2\} =
\]

\[
= E \{ ||x||^2 \} < +\infty,
\]

which imply essentially that the \( ||x||^2 \) is finite with probability 1.
Definition: we define a white noise on $X$ any process with zero mean and isotropic covariance function, i.e.

$$C_{x_1} = \sigma^2 I$$ \hfill [A.2.9]

Remark A.2.1

According to our Lemma A.2.1 we see that to white noise on $X$ can have realizations in $X$.

Remark A.2.2

The definition of $C_{x_1}$ and accordingly that of white noise, depend essentially on the topology of that space. For instance if $X$ is a reproducing Kernel Hilbert space, then the identity operator is associated to the Kernel

$$K(t, t') = \sum_i u_i(t) u_i(t')$$

$$\langle K(t, t'), u(t') \rangle = u(t)$$

the series in [A.2.10] being convergent in $X$ and also absolutely when $K$ is a continuous function.

Whence a white noise in such an $X$ has a covariance operator associated with the Kernel [A.2.10], which appears as the covariance function of the process.

Different spaces will give rise to completely different processes, according to the characteristics of the function [A.2.10].

Remark A.2.3

A strong warning should be given to the reader, that the definition of white noise adopted here, is not completely customary. As a matter of fact we could have rather smooth functions representing a white noise, because of the smoothness of $K(t, t')$; nevertheless these are white noises with respect to $X$ and their realizations do not belong to it.

To find the relation to the usual construction of a white noise via the Wiener process we should only look to the case in which $X = L^2$. In fact, if we make the hypothesis that our process is normal so that it is completely defined through its covariance structure, we see that for the system of characteristic functions

$$\chi_t(t) = \begin{cases} 1 & t \in I \\ 0 & t \in 0 \end{cases}$$
which is complete in $L^2$, the covariance structure gives

$$E\{ \langle x, \chi_n \rangle \langle x, \chi_n \rangle \} = \sigma^2 < \chi_n, \chi_n > =$$

$$= \sigma^2 \int \chi_n (t) \chi_n (t) dt = \sigma \mu (I_1 \cap I_2). \quad [A.2.11]$$

This is also the property characterizing Wiener processes, so that we can identify

$$\langle x, f \rangle = \frac{1}{\left( \int f (t) dW (t) \right)}.$$

It is to be stressed that it is this type of noise that usually is considered to represent experimental measuring errors, due to the martingale property of the Wiener process according to which two $\Delta W$ referring to non-overlapping intervals are independent.

We may conclude that, when it is this property that we need to represent a physical phenomenon, we are forced also to choice $C$ as $L^2$.

REFERENCES


