Least-Squares, Sentinels and Substractive Optimally Localized Average

Guy Chavent

b

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Least-Squares, Sentinels and Substractive Optimally Localized Average

Guy Chavent

N° 3332

January 1998

THÈME 4b

ISSN 0249-6399

INRIA
ROCQUENCOURT

Least-Squares, Sentinels and Substractive Optimally Localized Average

Guy Chavent

Thème 4b — Simulation et optimisation
de systèmes complexes
Projet Estime

Rapport de recherche n°3332 — January 1998 — 14 pages

Abstract: We present with unified notations three approaches to linear parameter estimation: least-squares, sentinels, and Substrative Optimally Localized Average (SOLA). It becomes then obvious that the two last approaches correspond to the very same mathematical problem. This brings a new interpretation to sentinels, new computational tools to SOLA, and makes clear their link to the classical least-squares approach.

Key-words: parameter estimation, inverse problem, sentinels, localized averages, least squares

(Résumé : tsvp)

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
Téléphone : 01 39 63 55 11 - International : +33 1 39 63 55 11
Télécopie : (33) 01 39 63 53 30 - International : +33 1 39 63 53 30

Moindres Carrés, Sentinelles et Moyennes Localisées Optimales Soustractives

Résumé : Nous présentons, avec des notations unifiées, trois approches de l'estimation linéaires de paramètres : les moindres carrés, les sentinelles, et les moyennes localisées soustractives optimales (SOLA). Il devient alors clair que les deux dernières approches correspondent exactement au même problème mathématique. On obtient ainsi une nouvelle interprétation du sentinelle, de nouveaux outils de calcul pour les moyennes localisées, et on précise leurs liens avec l'approche moindres carrés.

Mots-clé : estimation de paramètres, problème inverse, sentinelles, moyennes localisées, moindres carrés

1 Introduction

We consider the problem of estimating a parameter x from a date d in a linear model A: knowing

$$d = Ax_{true} + \varepsilon,$$

one wants an estimate of x_{true} together with an error estimate. In order to eliminate all technical difficulties, we shall consider only the finite dimensional case where x is in $\mathbb{R}^n$ and d in $\mathbb{R}^m$.
The oldest solution to this problem is the least-squares approach, where one searches for the "optimal" parameter $\hat{x}$ which minimizes a least-squares misfit between d and Ax. For Gaussian noises, this approach produces also an error estimate on $\hat{x}$. However, when x is made of the values of an unknown parameter function on a grid (we shall say in that case that x is a "distributed parameter"), this error estimates can be so large that it becomes useless when the size of the grid is small.

In 1967, Backus and Gilbert [1] [2] tried to solve simultaneously the estimation and uncertainity problem specifically for the case of a distributed parameter: their idea was to determine which "spatially localized average" cx of x could be determined from the data with a reasonnable level of uncertainty (the vector c is called the averaging kernel). Their approach was very successful in earth sciences, and developped later by Pijpers and Thomson into a Substractive Optimally Localized Average (SOLA) method [7] [8]: given a target averaging kernel $\hat{c}$, a kernel c is searched, which achieves a compromize between c being close to $\hat{c}$ and the uncertainty on cx_{true} being not too large. We refer to [4] for examples of application of the SOLA approach to earth sciences.

Then in 1988, Lions [5] [6] introduced the notion of sentinels: his original idea was to estimate only one component x_i of the parameter vector x by performing a scalar product of the data with a vector w which he called a sentinel. It became quickly apparent that a sentinel w could be designed to estimate any linear combination $\hat{c}x$ of x, and not only the combination $x_i = e_i x$, and that the sentinel estimate wd of $\hat{c}x$ coincided with its least squares estimate $\hat{c}\hat{x}$ [3].

We give in this paper an elementary and hopefully pedagogical presentation, with unified notations, of the three above approaches: least squares, sentinels, SOLA. It becomes then apparent that the last two approaches lead to the same mathematical problem, albeit the motivations are quite different. This makes precise the link between the SOLA and Least Squares estimates, shows that all the machinery developped for the computation of sentinels (primal and dual formulations) can be used to solve one SOLA problem, and brings a new interpretation to the sentinels in term of resolving power of the data.

2 The linear inverse problem

Let us suppose that we have recorded data $d = (d_1...d_m) \in I\!R^m$ generated by a linear model from an unknown parameter $x = (x_1...x_n) \in I\!R^n$:

$$d = Ax_{true} + \varepsilon \tag{1}$$

where A is a known $m \times n$ rectangular matrix, $x_{true} \in I\!R^n$ is the vector of parameters which have generated the data, and where $\varepsilon = (\varepsilon_1...\varepsilon_m)$ is the vector of measurement errors.

In order to evaluate quantitatively how well a vector Ax fits to the data d, one needs to choose a norm on the data space $I\!R^m$: for any vector $y \in I\!R^m$ we set

$$\|y\|^2_{E^{-1}} =< y, E^{-1}y >_{I\!R^m}, \tag{2}$$

where $<, >$ denotes the usual scalar product in $I\!R^m$, and where E is a given symmetric positive definite matrix. The choice of E will depend on wether we have or not statistical information on the error vector $\varepsilon = (\varepsilon_1...\varepsilon_m)$:

the probabilistic framework corresponds to the case where the error vector ε is made of zero-mean random variables with known covariance matrix. The natural norm in the data space corresponds then to

$$E = \text{ error covariance matrix } \mathcal{E}\{\varepsilon\varepsilon^t\}, \tag{3}$$

where $\mathcal{E}$ denotes the expectation of a random variable.
In the case where the error ε is Gaussian, the norm $\| \ \|_{E^{-1}}$ on the data space has a nice probabilistic interpretation: for any vector $y \in I\!R^m$ one has

$$e^{-\frac{1}{2}\|y\|^2_{E^{-1}}} = \text{ probability density function of } \varepsilon : \tag{4}$$

the smallest the norm of y, the likeliest y ! if moreover the errors ε_i are Gaussian and independant with covariance $\sigma_i^2 > 0$, then

$$E = \ diag\ (\sigma_1^2, ..., \sigma_m^2) \tag{5}$$

and, if all errors have the same covariance $\sigma^2 > 0$:

$$E = \sigma^2 I \tag{6}$$

the deterministic framework has to be used when no information on the statistical properties of the errors is available. In that case one can simply choose

$$E = I \tag{7}$$

or better, chosse a matrix E such that the norm of the misfit $d - Ax$ measures the relative error on the observations, for example

$$E = \ diag\ (\max\{d_i^2, \sigma^2\}) \tag{8}$$

where $\sigma > 0$ is a threshold level to be chosen.

In the sequel, we shall suppose that E is given by (3) (in the probabilistic case) or by (7) or (8) (in the deterministic case) and satisfies

$$E \text{ symetric positive definite.} \tag{9}$$

We shall also restrict ourselves, for the sake of simplicity, to the case where

$$A \text{ is injective, ie } m \geq n = \text{ Rank } A \tag{10}$$

This covers in the applications many overdetermined ($m \geq n$) linear inverse problems, and implies that the inverse problem has, at least theoretically, a unique solution. However, the n non-zero singular values of A will often have very different order of magnitudes, so that the numerical determination of x will tend to be unstable, and produce x's with unrealistic large norms.
The classical cure to this problem is regularization, where one searches for an x which realizes a compromize between explaining the data and having a small norm. This will be made precise in the next paragraph where we define the regularized least-squares setting of the inverse problem. Although a probabilistic interpretation of regularization is possible if an a-priori probability law with known covariance matrix is available in the parameter space, we shall not pursue further along this line, as such covariance matrix is rarely known. So we shall use for the regularization the norm

$$\|x\|^2_{R^{-1}} = < R^{-1}x, x >_{I\!R^n} \tag{11}$$

with

$$R^{-1} = n \times x \text{ symmetric positive definite.} \tag{12}$$

The matrix R^{-1} will be chosen from deterministic considerations:

- $R^{-1} = I$ if one wants to control the energy of x,
- R^{-1} = matrix associated to first or second derivatives if one wants to control the smoothness of x.

Of course, if some statistical information is available on x, one can also choose for R the covariance matrix of the a-priori uncertainty of x.

3 The regularized least-squares (RLS) approach

As we just explained, one searches in the this approach for the parameter $\hat{x}$ which realizes a compromise between the output $A\hat{x}$ being close to data d (in the norm chosen on the data space) and the norm of $\hat{x}$ (as chosen on the parameter space):

Definition 3.1 *The regularized least-squares (RLS) formulation of the linear inverse problem is:*

$$\textit{Find } \hat{x} \in I\!R^n \textit{ which minimizes } J(x) \textit{ over } I\!R^n \tag{13}$$

with

$$J(x) = \frac{1}{2}\|d - Ax\|^2_{E^{-1}} + \frac{\alpha}{2}\|x\|^2_{R^{-1}} \tag{14}$$

where α is the regularization parameter, chosen such that

$$\alpha > 0 \tag{15}$$

The coefficient α sets the compromise between the two competing parts of the objective function: a small α will tend to generate a good fit but a large parameter norm, a long α will ensure a small parameter norm, but produce a poor fit !

We shall denote by

$$H_0 = A^T E^{-1} A \tag{16}$$

the Hessian of the unregularized objective function J given by (14) with α set to zero. Under hypothesis (9) (10), H_0 is positive definite, but it can be extremely illconditioned - and it actually is in many applications.

Then we shall denote by

$$H = A^T E^{-1} A + \alpha R^{-1} = H_0 + \alpha R^{-1} \tag{17}$$

the Hessian of the regularized objective function, which of course, as $\alpha > 0$, is also positive definite.
Hence the function J is strictly convex, and goes to $+\infty$ when the norm of x tends to infinity. This ensures the existence and uniqueness of $\hat{x}$:

Proposition 3.1 *the RLS formulation (13) (14) has a unique solution $\hat{x}$, given by the normal equation:*

$$H\hat{x} = A^T E^{-1} d \tag{18}$$

A classical result of regularization theory is that, when $\alpha \to 0$, the regularized solution $\hat{x}$ converges to the solution $\hat{x}_0$ of the unregularized problem (minimization of J given by (14) where one has set α to zero). But letting α go to zero is numerically unfeasible, as the conditioning of H tends to that of H_0, which is often extremely poor, so that the numerical resolution of (18) becomes more and more difficult.
So in this paper we shall consider that one has chosen one $\alpha > 0$ - the difficulty being now of course to choose this α ! There is a huge literrature on this problem, which we do not intend to cover systematically, but we shall give some clues on the subject during the course of the paper.

We turn now to the problem of the stability of the RLS solution to the inverse problem.

In the deterministic case, where nothing is known about the statistical properties of the error vector ε, there is not much to say beyond (18): a perturbation δd on the data induces a perturbation $\delta\hat{x}$ of the optimal estimate given by:

$$H\delta\hat{x} = A^T E^{-1}\delta d, \tag{19}$$

and the classical results of linear algebra on the conditioning of a matrix can be used.

But in the probabilistic case where E is the covariance matrix of the observation error, we can, given a time parameter x_{true}, consider the data d given by (1) as a vector of random variables. Then the optimal estimate $\hat{x}$ given by (18) becomes itself a vector of random variables. Its expected value is given by:

$$\begin{aligned}
\mathcal{E}(\hat{x}) = &\quad \mathcal{E}\{H^{-1}A^T E^{-1}(Ax_{true} + \varepsilon)\} \\
\mathcal{E}(\hat{x}) = &\quad H^{-1}H_0 x_{true} + H^{-1}A^T E^{-1}\mathcal{E}(\varepsilon) \\
\mathcal{E}(\hat{x}) = &\quad H^{-1}H_0 x_{true} = (I - \alpha H^{-1}R^{-1})x_{true}
\end{aligned}$$

which shows that the regularization introduces a bias in the least square estimate $\hat{x}$, which is not really surprising as $\hat{x}$ results of a compromise !

The covariance of $\hat{x}$ can also easily be calculated:

$$\begin{aligned}
\text{Cov }\{\hat{x}\} = &\quad \mathcal{E}\{H^{-1}A^T E^{-1}\varepsilon\varepsilon^T E^{-1}AH^{-1}\} \\
&\quad H^{-1}A^T E^{-1}\mathcal{E}\{\varepsilon\varepsilon^T\}E^{-1}AH^{-1}
\end{aligned}$$

which, as $\mathcal{E}(\varepsilon\varepsilon^T) = E$, reduces to

$$\begin{aligned}
\text{Cov }\{\hat{x}\} &= H^{-1}A^T E^{-1}AH^{-1} \\
&= H^{-1}H_0 H^{-1}.
\end{aligned}$$

Proposition 3.2 *In the probabilistic case, the RLS estimate $\hat{x}$ of x_{true} satisfies:*

$$\mathcal{E}\{\hat{x}\} = H^{-1}H_0 x_{true} = (I - \alpha H^{-1}R^{-1})x_{true} \tag{20}$$

$$Cov\,\{\hat{x}\} = H^{-1}H_0 H^{-1} = (I - \alpha H^{-1}R^{-1})H^{-1} \tag{21}$$

where H_0 and $H = H_0 + \alpha R^{-1}$ one the Hessians of the unregularized and regularized objective functions.

So when the regularization parameter is small, the covariance matrix of $\hat{x}$ is close to $H_0^{-1} = (A^T E^{-1}A)^{-1}$. As H_0 is usually poorly conditioned, H_0^{-1} will have large elements on its diagonal, so that the variance of the individual $\hat{x}_i$'s will be very large.

When $\alpha \to +\infty$, the covariance matrix behaves like $R^{-1}H_0 R^{-1}/\alpha^2$ and hence tends to zero. This shows that the regularization has the desired effect of reducing the covariance of $\hat{x}$ - but at the same time drives its mean value from x_{true} towards zero (see (20)) ! However,

in many applications, especially those when $x_1 x_2 ... x_n$ represents the value of an unknown function at the node of a grid, it is impossible to find a regularization level α which produces both a satisfying fit to the data, and a pactically usable uncertainity level on each of the $\hat{x}_i$: either α is taken small enough to produce a reasonnable fit to the data, but the uncertainity of each $\hat{x}_i$ is huge, or α is increased enough for the uncertainity on the $\hat{x}_i$'s to become reasonnable, but then the fir to the data is awful. This happens in particular when the unknown function has been discretized on a grid with a number of node larger than the number of significantly non-zero eigenvalues of H_0.
This leads to the idea that, if one cannot estimate in a statistically meaningful way the point values x_i of the unknown function, there may exist certain linear functions of $x_1 ... x_n$ which can be retrieved with a good level of confidence: given a vector $\hat{c} = (\hat{c}_1 ... \hat{c}_n)$, the covariance of the linear function $\hat{c}^T \hat{x}$ is

$$\text{Cov } \{\hat{c}^T \hat{x}\} = c^T \text{ Cov } \{\hat{x}\} c$$

$$\text{Cov } \{\hat{c}^T \hat{x}\} = c^T H^{-1} H_0 H^{-1} c \tag{22}$$

This point of view will be developped in the next two paragraph, where we will explain how to compute $\hat{c}^T \hat{x}$ (section 3 on sentinels) and choose $\hat{c}$ and α (section 4 on optimally localized average methods).

4 The sentinel approach

The sentinels were originaly introduced by Lions as vectors w in the data space $\mathbb{R}^m$ such that $w^T Ax$ has derivatives equal to zero with respect to the "non interesting components" of x, and a derivative equal to 1 with respect to one "interesting component" of x. Such a w provides a "sentinel" for the monitoring of the "interesting component": each time a data vector d is recorded, the simple scalar product $w^T d$ is computed; variations in the successive values of $w^T d$ reflect the variations of the "interesting component" of x, and allow its monitoring.
It appeared rapidly that the sentinel approach was not limited to the monitoring of one component of x, but could be used to monitor any linear combination of the x_i's for example $\hat{c}^T x$, where

$$\hat{c} \in \mathbb{R}^m \text{ is a given monitoring vector} \tag{23}$$

(the monitoring of, say, x_1 is then obtained by chosing $\hat{c} = (1, 0...0)$).
A natural way to monitor $\hat{c}^T x$ would be to choose for sentinel a vector w such that $w^T d$ is the scalar product of $\hat{c}$ with the best practically available estimate of x, ie with the solution $\hat{x}$ of the RLS formulation (13) (14) !

Definition 4.1 *Given a monitoring vector $\hat{c} \in \mathbb{R}^n$, the sentinel for the monitoring of $\hat{c}^T x$ is a vector $w \in \mathbb{R}^m$, one has*

$$w^T d = \hat{c}^T \hat{x}, \tag{24}$$

where $\hat{x}$ is the RLS estimate of section 2.

As we have seen in proposition 3.2, the RLS formulation (13) (14) has a unique solution, given by (18), so that:

$$\hat{c}^T \hat{x} = \hat{c} H^{-1} A^T E^{-1} d.$$

Hence the sentinel w is given by the transposed of the coefficient of d:

$$w = E^{-1} A H^{-1} \hat{c}. \tag{25}$$

Determination of w by (25) corresponds to the so-called "primal resolution" of the sentinel problem: one first determines $r = H^{-1}\hat{c}$ by solving the $n \times n$ equation

$$(A^T E^{-1} A + \alpha R^{-1}) r = \hat{c}, \tag{26}$$

then the sentinel $w \in I\!R^m$ is given by

$$w = E^{-1} A r. \tag{27}$$

It is also possible to write an equation which gives directly $w \in I\!R^m$, this is the so-called "dual resolution" of the sentinel problem: multiplying first (25) by ARA^T we obtain, as $A^T E^{-1} A = H - \alpha R^{-1}$:

$$ARA^T w = AR\hat{c} - \alpha A H^{-1} \hat{c},$$

and, using again (25) to replace $AH^{-1}\hat{c}$ by Ew:

$$ARA^T w = AR\hat{c} - \alpha E w.$$

Reordering the terms, we obtain

$$(ARA^T + \alpha E) w = AR\hat{c} \tag{28}$$

which is the sought equation for w.
As R and E are positive definite and $\alpha > 0$, equations (26) (27) as well as equation (28) detrmine uniquely the sentinel vector w.
Noticing that (26) and (28) are the normal equations of quadratical minimization problem, we can summarize the above results as follows:

Proposition 4.1 *Given any monitoring vector $\hat{c} \in I\!R^n$, the sentinel problem (24) admits a unique solution $w \in I\!R^m$, given either by:*

- *the solution of the "primal sentinel control problem":*

$$\left\{ \begin{array}{l} \textit{find } r \in I\!R^n \textit{ which minimizes} \\ \frac{1}{2}\|Ar\|^2_{E^{-1}} - < \hat{c}, r >_{I\!R^n} + \frac{\alpha}{2}\|r\|^2_{R^{-1}} \end{array} \right. \tag{29}$$

  (or equivalently: solve the normal equation (26)) followed by

$$w = E^{-1} A r, \tag{30}$$

- *the solution of the "dual sentinel control problem":*

$$\left\{ \begin{array}{l} \textit{find } w \in I\!R^m \textit{ which minimizes} \\ \frac{1}{2}\|\hat{c} - A^T w\|_R^2 + \frac{\alpha}{2}\|w\|_E^2 \end{array} \right. \tag{31}$$

(or equivalently: solve the normal equation (28)).
The residual $\hat{c} - A^T w$ of the dual problem and the solution r of the primal problem are related by:

$$\hat{c} - A^T w = \alpha R^{-1} r \tag{32}$$

The theory of regularization, which we already mentioned after proposition 3.1, shows that, when $\alpha \to 0$, the solution w of (31) converges to the solution w_0 of

$$\hat{c} = A^T w_0 \tag{33}$$

which has the smallest $\| \ \|_E$ norm. Remember that A^T is not in general injective, so (33) alone does not determine w_0 uniquely. This w_0 is called the Minimum $\| \ \|_E$ -Norm Solution of (33), but, despite its name, the norm $\|w_0\|_E$ can be very large when the matrix $A^T E^{-1} A$ is poorly conditioned, which is a typical situation.

Proposition 4.2 *For a given monitoring vector $\hat{c} \in I\!R^m$, the norm $\|w\|_t$ of the sentinel decreases from $\|w_0\|_E$ (where w_0 is the Minimum $\| \ \|_E$ Norm solution of (33)) to zero when the regularization parameter α increases from zero to infinity. Hence α can be used to control the norm $\|w\|_E$ of the sentinel.*

In the probabilistic case, we can calculate the expected value of $w^T d = \hat{c}^T \hat{x}$. If we start from $w^T d$ we find

$$\mathcal{E}\{w^T d\} = \mathcal{E}\{w^T (A x_{true} + \varepsilon)\} = w^T A x_{true}. \tag{34}$$

But one could also start from $\hat{c}^T \hat{x}$:

$$\begin{array}{rl} \mathcal{E}\{\hat{c}^T \hat{x}\} & = \hat{c}^T \mathcal{E}\{\hat{x}\} \\ & = \hat{c}^T H^{-1} H_0 x_{true} = c^T (I - \alpha H^{-1} R^{-1}) x_{true} \end{array} \tag{35}$$

As $w^T d = \hat{c}^T \hat{x}$, the right-hand sides of (34) and (35) shoould be the same, which is the cases as we see from (25) that $A^T w = H_0 H^{-1} \hat{c}$!
The covariance (it is in fact a variance as the quantity under consideration is a scalar) is then given by

$$\begin{array}{rl} \text{Cov } \{w^T d\} & = \mathcal{E}\{(w^T d - w^T A x_{true})^2\} \\ & = \mathcal{E}\{w^T \varepsilon \varepsilon^T w\} \\ & = w^T \mathcal{E}\{\varepsilon \varepsilon^T\} w \\ & = w^T E w = \|w\|_E^2 \end{array} \tag{36}$$

or:

$$\begin{array}{rl} \text{Cov } \{\hat{c}^T \hat{x}\} & = \mathcal{E}\{\hat{c}^T (\hat{x} - \mathcal{E}\{\hat{x}\})(\hat{x} - \mathcal{E}(\hat{x})^T \hat{c}\} \\ & = \hat{c}^T \text{ Cov } \hat{x} c \\ & = \hat{c}^T H^{-1} H_0 H^{-1} c \end{array} \tag{37}$$

Once again, the right-hand sides of (36) and (37) should be - and are - the same, as one checks easily using (25).

Proposition 4.3 *in the probabilistic case, the sentinel estimate $w^T d$ of $\hat{c}^T x_{true}$ satisfies*

$$\mathcal{E}\{w^T d\} = w^T A x_{true} = \hat{c}^T H^{-1} H_0 x_{true} = \hat{c}^T (I - \alpha H^{-1} R^{-1}) x_{true} \tag{38}$$

$$Cov\,\{w^T d\} = \|w\|_E^2 = \hat{c}^T H^{-1} H_0 H^{-1} \hat{c} = \hat{c}^T (I - \alpha H^{-1} R^{-1}) H^{-1} \hat{c} \tag{39}$$

So given a "target" monitoring vector $\hat{c}$, we see (last equality in (38)) that the sentinel estimate $w^T d$ gives only a biased estimation of $\hat{c}^T x_{true}$, but gives (first equality in (38)) an unbiased estimation of cx_{true} where we have set

$$c = A^T w \tag{40}$$

More over, the variance of this estimate (first equality in (39)) is nothing but the covariance weighted norm $\|w\|_E^2$ of the sentinel vector, which as we have seen in proposition 4.1 cam be controlled by the level α of regularization.
We shall see in the next paragraph that the above machinery is very close, and in most cases coincides with that introduced for the search of localized averaging kernels by Backus and Gilbert and their followers.

5 Localized averaging kernels

This approach was introduced by Backus and Gilbert for the estimation of spatially distributed parameters in the presence of noise in the observations. So we are in this section in the probabilistic case of section 1. Backus and Gilbert's idea was to find a linear combination $w^T d$ of the data, where $w \in I\!R^n$ is a vector of inversion coefficients to be determined, such that $w^T d$ gives a statistically sound estimation of some localized average $c^T x_{true}$ around a given spatial location. In Backus and Gilbert terminology, $c \in I\!R^m$ is called an averaging kernel.
So, given a spatial location around which one want to estimate the unknown parameter x_{true}, the inversion coefficient vector $w \in I\!R^m$ and the averaging kernel $\hat{c} \in I\!R^n$ have to be chosen along the following lines:

$$\begin{cases} w^T d \text{ is an estimate of } c^T x_{true}, \text{ ie :} \\ \mathcal{E}\{w^T d\} = c^T x_{true} \end{cases} \tag{41}$$

$$\begin{cases} \text{Cov } \{w^T d\} \text{ is a small enough so that} \\ w^T d \text{ is not dominated by noise} \end{cases} \tag{42}$$

$$\begin{cases} \text{the averaging kernel } c \text{ is localized} \\ \text{around a given spatial location,} \\ \text{with less possible (especially negative) side-} \\ \text{-lobes, and with a total mass } \sum_{i=1}^n c_i \text{ close to one.} \end{cases} \tag{43}$$

Conditions (42) and (43) are only qualitative, so we can expect more than one way of constructing localized averaging kernels. But all should have in common that they satisfy the

quantitative condition (41). From the definition (3.1) of d we see that $\mathcal{E}\{w^T d\} = w^T A x_{true}$ so that (41) rewrites

$$w^T A x_{true} = c^T x_{true}$$

which, as we do not know x_{true}, can be satisfied in all instances only if

$$c = A^T w. \tag{44}$$

The condition is necessary and sufficient to ensure that (41) holds, and we shall suppose from now on that it is satisfied. Then:

$$\begin{array}{ll} \text{Cov } \{w^T d\} & = \mathcal{E}\{(w^T d - w^T A x_{true})(w^T d - w^T A x_{true})^T\} \\ & = \mathcal{E}\{(w^T \varepsilon \varepsilon^T w\} \\ & = w^T \mathcal{E}\{\varepsilon \varepsilon^T\} w \\ \text{Cov } \{w^T d\} & = \|w\|_E^2. \end{array} \tag{45}$$

Hence satisfying (42) amounts to control the covariance weighted norm $\|w\|_E^2$ of the inversion coefficient vector w.

Before we indicate how to satisfy, both (42) and (43), notice that they correspond to adversely competing objectives: their physical intuition told Backus and Gilbert that a high noise level on the estimator of $c^T x_{true}$ was the price to pay for c^T to achieve a high spatial resolution. So conditions (42) and (43) express that one has to chose a compromise between noise level and spatial resolution of the estimator.

Rather than describing the original solution to (41) (42) (43) proposed by Backus and Gilbert, we shall present a variant introduced by Oldenberg and generalized by Pijpers and Thompson, the Substractive Optimally Localized Average (SOLA) approach, which is now in wide use.

In this approach, one choses first a target averaging kernel $\hat{c}$, localized at the spatial location to be investigated, with $\sum_{i=1}^n \hat{c}_i = 1$, and, unless one is a masochist, with no sidelobes. The spatial resolution of $\hat{c}$ is first guessed, but it will be later adjusted by a trial and error process as we shall see below. Then the inversion coefficient vector w is chosen in such a way that $\|w\|_E^2$ is not too large (which, because of (45), will ensure that (42) is satisfied) and that some norm of $\hat{c} - A^T w$ is not too large (which, because of (44) will ensure that (43) is satisfied):

Definition 5.1 *Given a target averaging kernel $\hat{c} \in I\!R^n$, the Substractive Optimally Localized Average (SOLA) approach defines an inversion coefficient vector $w \in I\!R^m$ as the solution to*

$$\begin{array}{l} \textit{find } w \in I\!R^m \textit{ which minimizes} \\ \frac{1}{2}\|\hat{c} - A^T w\|_R^2 + \frac{\alpha}{2}\|w\|_E^2, \end{array} \tag{46}$$

and an averaging kernel $c \in I\!R^n$ by

$$c = A^T w, \tag{47}$$

where R is a positive definite matrix on the parameter space, which defines the norm used to measure the distance of c to $\hat{c}$, E is the covariance matrix of the noise on data, $\alpha > 0$ is a coefficient used to set the compromise between spatial resolution and estimation error.

In the practice of SOLA, the optimization problem (46) is solved for various values of $\alpha > 0$ and of the width (spatial resolution) of $\hat{c}$, until a $c = A^T w$ with as little as possible side lobes and with $\sum_{i=1}^{n} c_i$ close enough to 1, and a w with acceptable $\|w\|_E$ are obtained. Then the simple computation of $w^T d$ gives an estimate of $c^T x_{true}$ with standard deviation $\|w\|_E$, and the shape of c visualizes the spatial resolution achieved at the error level $\|w\|_E$.

It is remarkable that, although the SOLA approach was developped completely independantly of the RLS approach and the sentinel approach (and much earlier than this latter), there are very strong ties between these approaches:

Proposition 5.1 *Let $\hat{x}$ be the solution of the RLS problem (13) (14), and w and c be the inversion coefficients and averaging kernels constructed by the SOLA kernel-tayloring approach (46) (47) from a given target kernel $\hat{c}$.*
Then:

- *w coincides with the sentinel associated to the monitoring vector $\hat{c}$, as the SOLA optimization problem (46) coincides with the sentinel dual problem (31),*
- *the vector $w \in \mathbb{R}^m$ of the SOLA approach can be equivalently determined by (46) or by the sentinel primal problem (29) (30), whose unknown is $r \in \mathbb{R}^n$, which may be an advantage as $n \leq m$,*
- *the SOLA estimate $w^T d$ coincides with target average $\hat{c}^T \hat{x}$ of the RLS solution $\hat{x}$.*

This results from the interpretation of the SOLA approach in term of sentinels. Notice that the regularization term $\|x\|^2_{R^{-1}}$ used in the RLS approach and the kernel misfit term $\|\hat{c} - A^T w\|^2_R$ used in the SOLS approach correspond to inverse matrices R^{-1} and R.

References

[1] G. Backus and F. Gilbert. Numerical applications of a formalism for geophysical inverse problems. *Geophys, J. R. ast. Soc.*, 13:247–276, 1967.

[2] G. Backus and F. Gilbert. The resolving power of gross earth data. *Geophys, J. R. ast. Soc.*, 16:169–240, 1968.

[3] G. Chavent. Generalized sentinels defined via least squares. *Appl. Math. Optim.*, 31:189–218, 1995.

[4] BH Jacobson and K. Moosegard. Inverse methods: interdisciplinary elements of methodology, computation and applications. In P. Siborni, editor, *Lecture Notes in Earth Sciences*, volume 63. Springer, 1996.

[5] J. L. Lions. *Controlabilité exacte, perturbation et stabilisation de systèmes distribués*, volume 1 et 2. Masson, 1988.

[6] J. L. Lions. *Sur les sentinelles des systèmes distribués*, volume 307. CRAS, 1988.

[7] FP. Pijpers and MJ Thompson. Faster formulations of the optimally localized averages method for helioseismic inversion. *Astron. Astrophys*, 262:L33–L36, 1992.

[8] FP. Pijpers and MJ Thompson. The SOLA method for helioseismic inversion. *Astron. Astrophys*, 281:231–240, 1994.

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399