# Retrieval Algorithm

Sébastien Payan – LATMOS –May 2022

Sorbonne Université & CNRS

## Summaries

# 1 Introduction

In most of the case, the signal measured by an optical sensor cannot be translate directly into atmospheric parameter.

For each instruments, it is necessary to develop an **algorithm** allowing to infer **target parameters** (e.g. vertical profiles of temperature or gas concentrations) from spectral measurements (radiances).

To distinguish the principal algorithm classes used, the term of **retrieval method** is usually used. In all the cases it is necessary to solve the **direct problem**: to have an algorithm allowing to simulate/calculate signal that should receive the instrument for a given state of the atmosphere and for a given observation geometry.

It is easy to understand that this calculation will be so much reliable that **instrument model** itself will be the more accurate, as for the **radiative transfer** calculation.

Then it is necessary to build an accurate and efficient **retrieval method** (often iterative, often with dumping because “ill posed”) able to better account for information content in the measured atmospheric signal.

In addition, it is necessary for operational processing purposes, to automate processes and to **optimize calculation time** (and then sometimes apply approximations) which is not always fitted the ultimate accuracy that could be reach with a more meticulous algorithm.

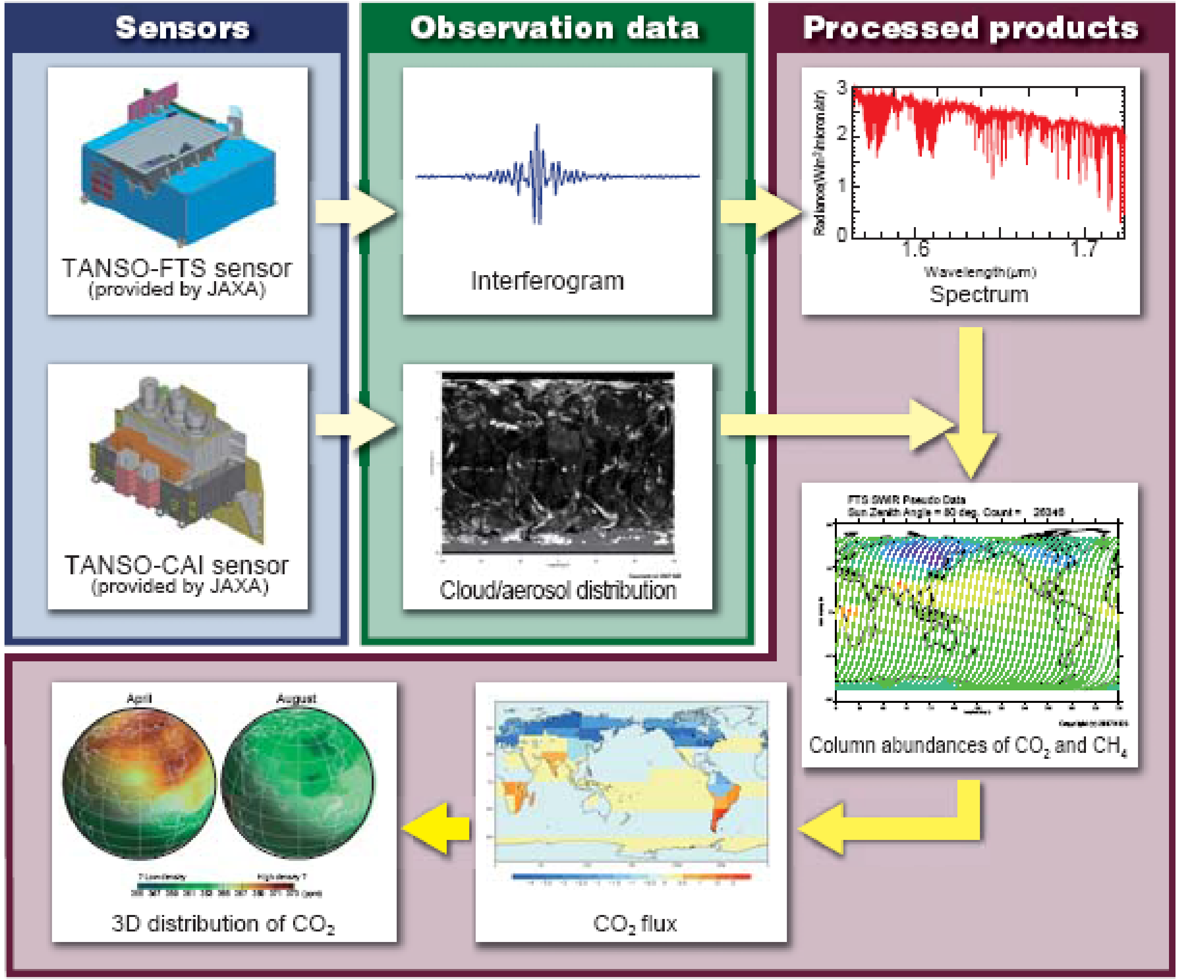
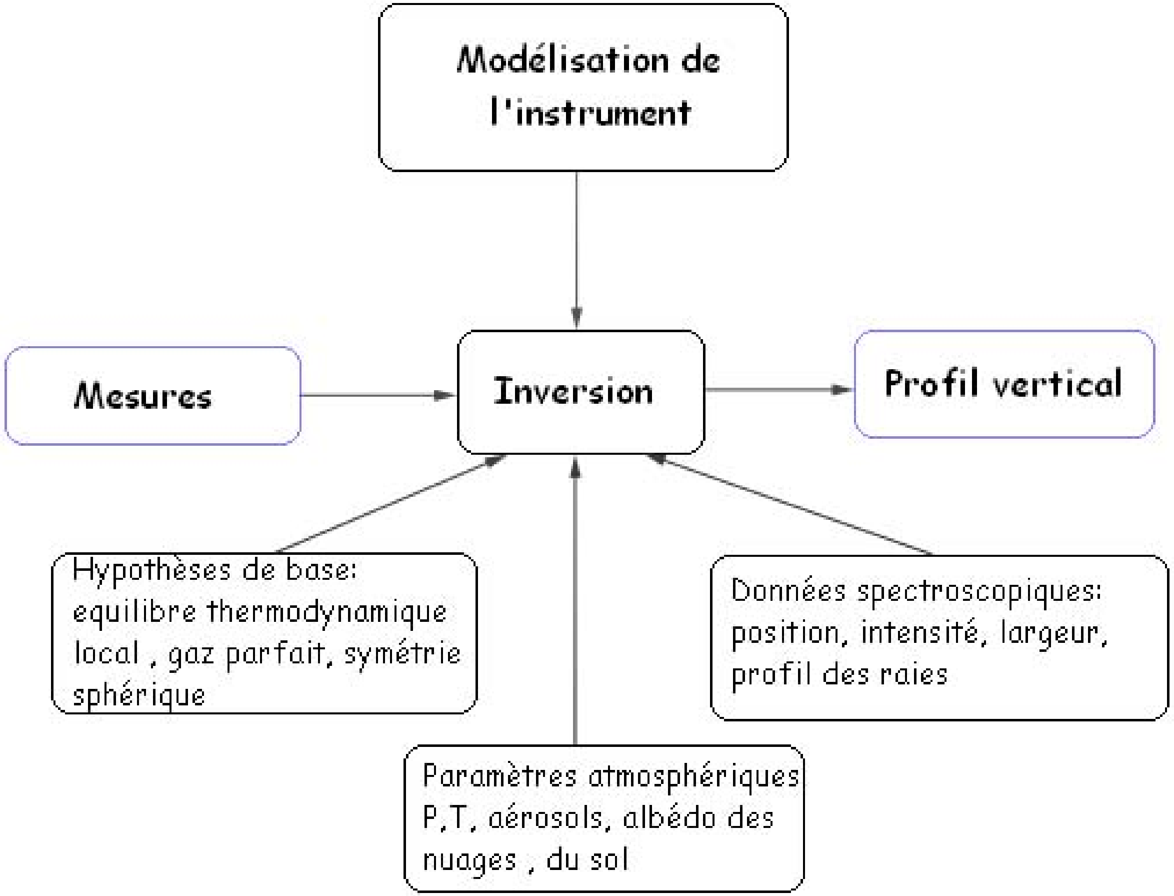


Figure 7. Outline of GOSAT data processing[1]

[1] GOSAT Research announcement, Appendix A, 2008

# 2 Spectra retrieval

## *2.1 Retrieval principle*

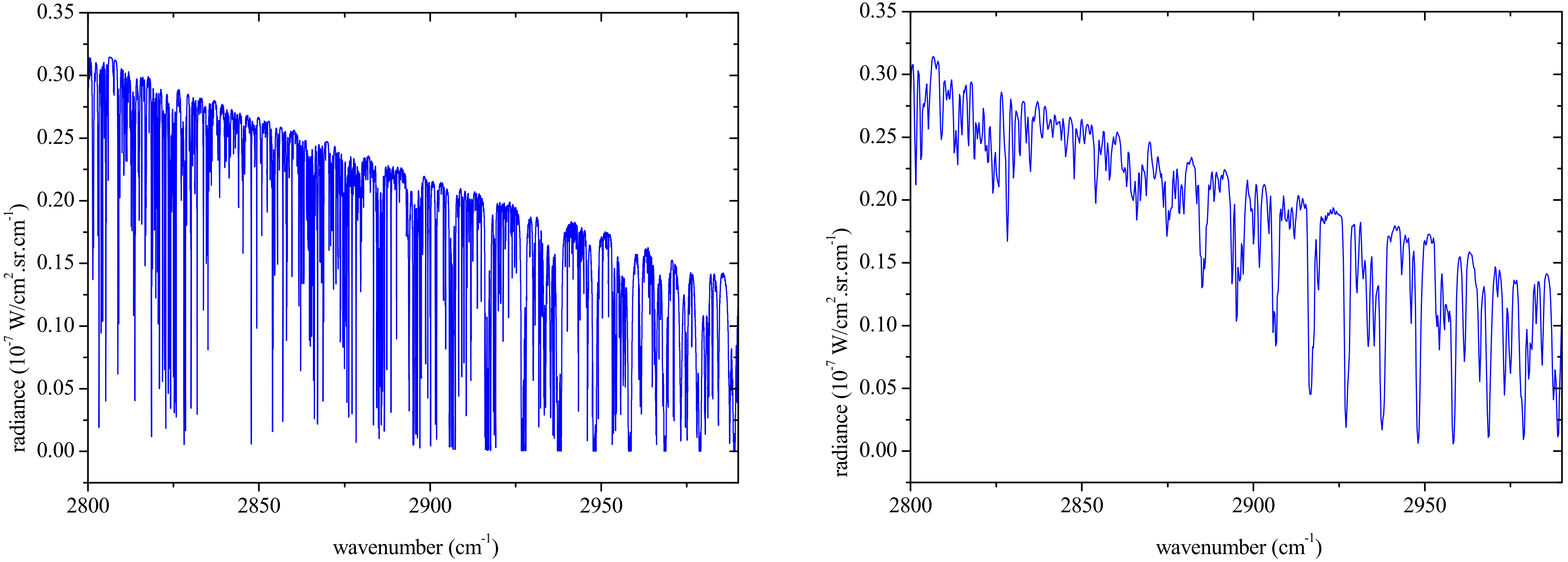

Modélisation de l'instrument
Mesures
Inversion
Profil vertical
Hypothèses de base: equilibre thermodynamique local , gaz parfait, symétrie sphérique
Données spectroscopiques: position, intensité, largeur, profil des raies
Paramètres atmosphériques P,T, aérosols, albédo des nuages , du sol

The problem we have to consider is to use available information at the best: spectrum or set of atmospheric spectra corresponding to one or more situations (line of sight) from which it is possible to "draw parallels" one or more parameters to be determined.

But in most cases, received signal by an onboard sensor is not directly interpretable in terms of atmospheric parameters. This requires for each instrument to develop an algorithm allowing, from spectral measurements, to infer the desired parameters (concentration profiles for example).

⇨ **Inverse model**

Instrumental effects : example of IASI instrument line shape (ILS) :

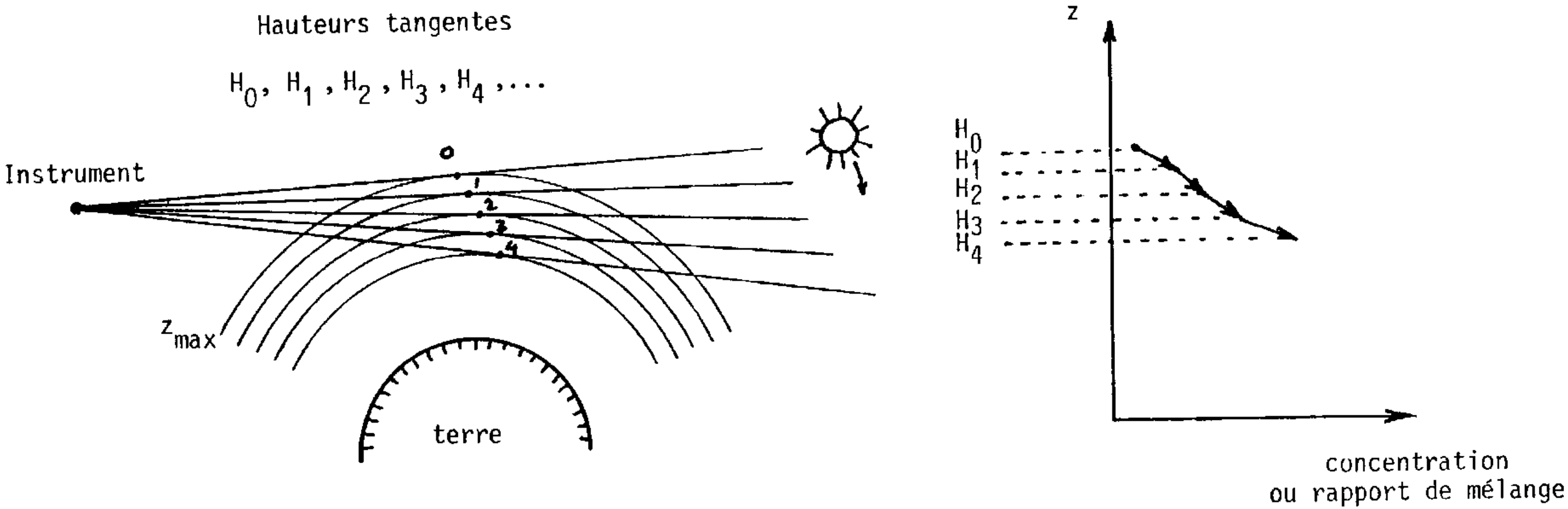


### *2.2 Some approaches for multi-spectrum setup.*

We present here briefly 3 methods :

- Two methods that can be implement at low or medium spectral resolution
  - onion-peeling technique
  - Mill’s Method

- Global fit method implemented for high resolution spectra. It consists to fit simultaneously (least square meaning) all parameters (vertical profiles, but also solar spectrum, ILS, aerosol optical depth, …) for one to several species, for all the Line of sight, and for all the selected spectral windows.



**Onion peeling Method**

### *2.3 Least square fitting of the spectra.*

We have a set of spectral measurements (a spectrum) which are representatives of a set of parameters that we want to determine values.

Let use the following notations :

- index *i* correspond to a given spectral element.
- We call *npt* the total number of considered spectral points
- index *j* (or *k*) corresponds to one of the *np* parameters to be determined.

Parameter determination will be of course possible only if criteria $npt \geq np$ is verified.

Let call :

- $\mathrm{Y}$ the measured quantity (e.g. spectral radiance)
- $\hat{\mathrm{Y}}$ (or $\mathrm{F(X)}$) the radiance calculated with a model
- $\mathrm{X}$ the parameters to be determine
- $\mathrm{E}$ the difference between observed and calculated spectrum

General relation between measured quantities and model parameters can be written as :

$$y_i = \hat{y}_i + e_i = f_{ij}(x_j) + e_i \tag{1}$$

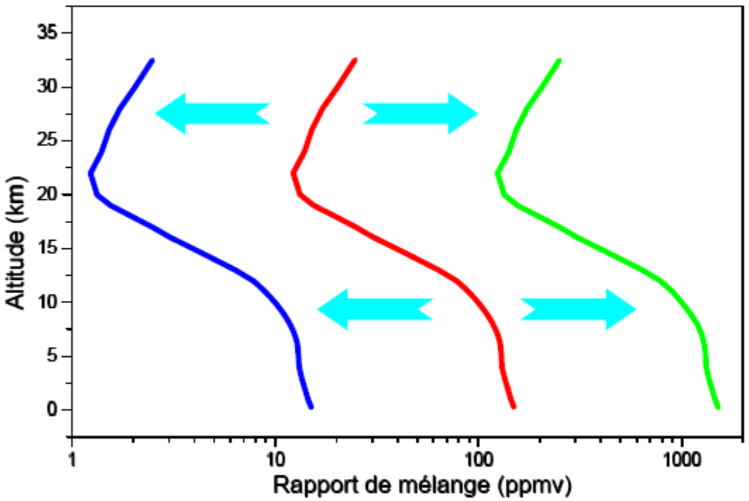
### 2.3.1 Linear least square

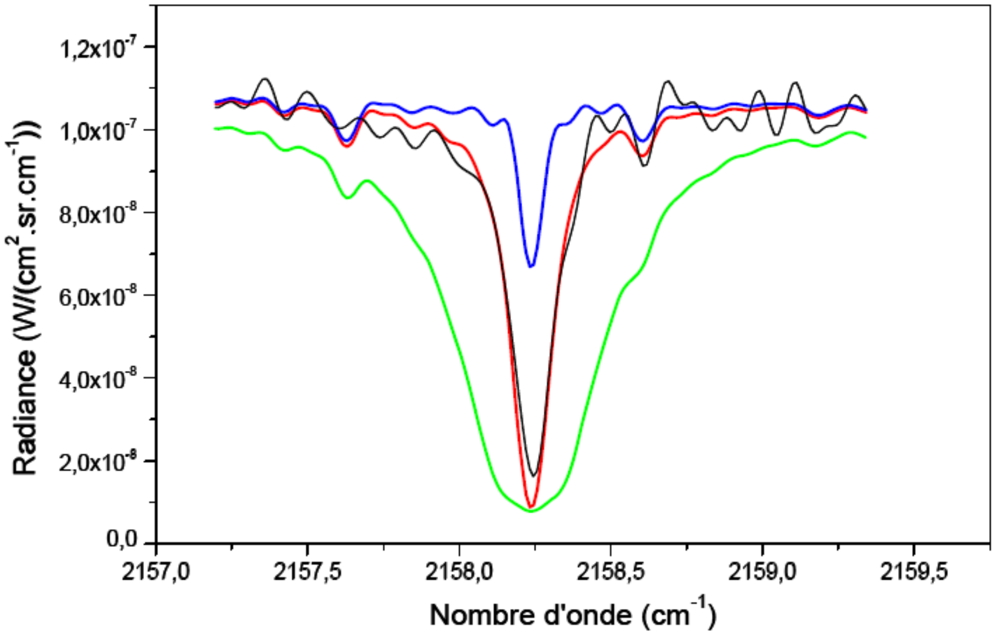
Let consider a simple case where quantity measured can be expressed using a linear model. Eq. (1) can then be written using a matrix form :

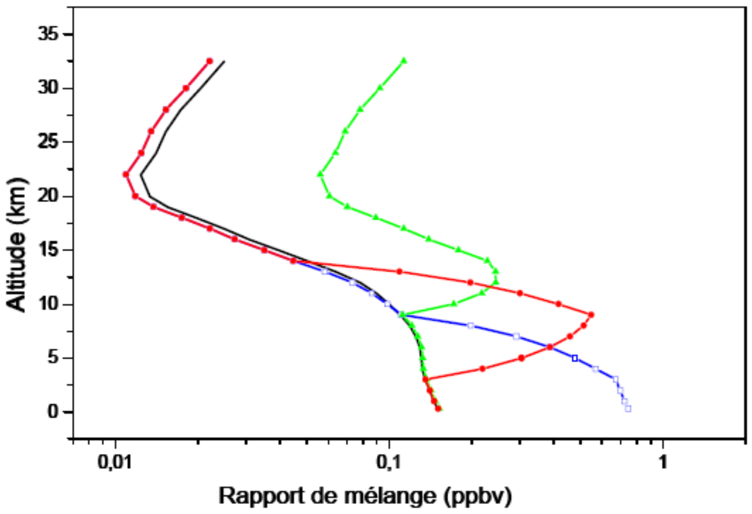
$$Y = \hat{Y} + E = FX + E \tag{2}$$

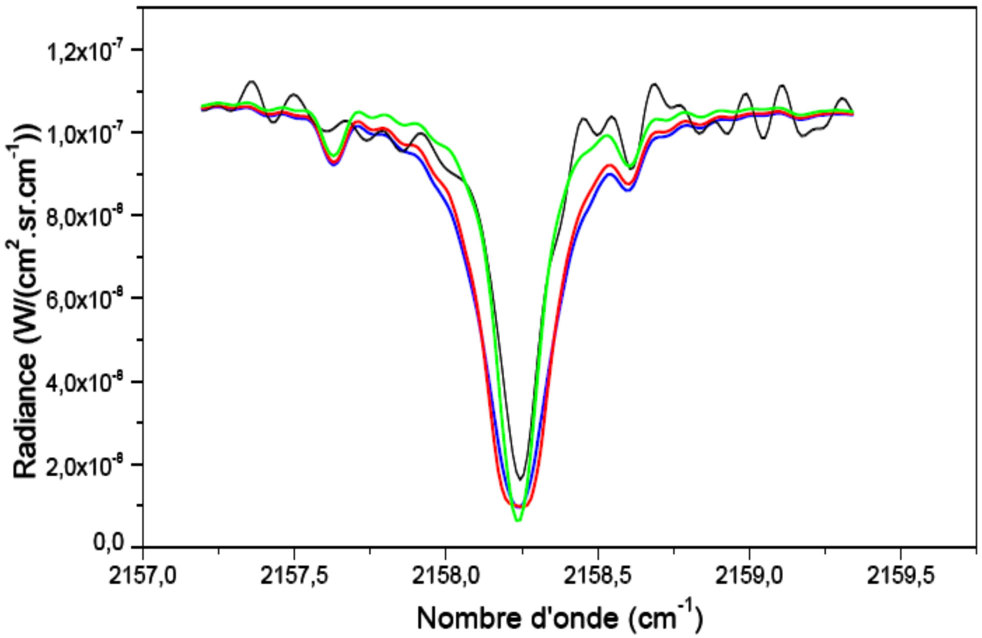
Spectrum $\hat{Y}$ is sometimes call *synthetic spectrum*.

We are looking to determine parameters $x_j$ that give the best reproduction of measurements, that is to say the ones for which the differences $e_i$ $\left(= y_i - f_{ij}(x_j)\right)$ are the lowers possible. Differences $e_i$ are often called *residuals*.


Altitude (km)
0
5
10
15
20
25
30
35
1
10
100
1000
Rapport de mélange (ppmv)


Radiance (W/(cm$^2$.sr.cm$^{-1}$))
0,0
2,0x10$^{-8}$
4,0x10$^{-8}$
6,0x10$^{-8}$
8,0x10$^{-8}$
1,0x10$^{-7}$
1,2x10$^{-7}$
2157,0
2157,5
2158,0
2158,5
2159,0
2159,5
Nombre d'onde (cm$^{-1}$)


Altitude (km)
0
5
10
15
20
25
30
35
0,01
0,1
1
Rapport de mélange (ppbv)


Radiance (W/(cm$^2$.sr.cm$^{-1}$))
0,0
2,0x10$^{-8}$
4,0x10$^{-8}$
6,0x10$^{-8}$
8,0x10$^{-8}$
1,0x10$^{-7}$
1,2x10$^{-7}$
2157,0
2157,5
2158,0
2158,5
2159,0
2159,5
Nombre d'onde (cm$^{-1}$)

Then, we are looking to determine parameters $x_j$ (that give the best fit of measurements) for which the residuals $e_i$ are the lowers as possible. The corresponding parameters minimizing $\chi^2$ function defined by :

$$\chi^2 = \sum_{i=1}^{npt} \frac{(y_i - \hat{y}_i)^2}{(\Delta y_i)^2} = \sum_{i=1}^{npt} (y_i - \hat{y}_i)^2 w_i \tag{3}$$

Where $\Delta y_i$ is representing the standard deviation on $y_i$, and where $w_i = 1/\Delta y_i^2$ is the corresponding weight.

We define in addition reduced $\chi^2$ defined as :

$$\chi_r^2 = \frac{\chi^2}{npt - np} \tag{4}$$

Then we define a quality criteria of the fit that would be statistically so much better that it would be close to 1 (residuals equivalent to measurement noise).

When $\chi_r^2$ is $> 1$, retrieval is not satisfactory.

When $\chi_r^2$ is < 1, values of measurement errors are likelihood overestimated.

We mentioned previously that target parameters $x_j$ are the one minimizing $\chi^2$ function. We can rewrite this function using matrix notation :

$$\chi^2 = E^T W E \tag{5}$$

Where $E^T$ represents transpose of $E = Y - \hat{Y}$ vector, and where W is a diagonal matrix having weights $w_i$ as diagonal elements (inverse of variances). Non-diagonal elements of W can be different from zero when there are some correlations between the spectral points.

One necessary condition to have minimum of $\chi^2$ is that derivatives with respect to each parameter be null :

$$\left(\frac{\partial\chi^2}{\partial x}\right) = 0 \quad \Leftrightarrow \quad \left(\frac{\partial\chi^2}{\partial x_j}\right) = 0 \quad \text{pour } j \in [1, np],\ j \in |\mathbf{N} \tag{6}$$

Using matrix notation (5) and using the expression of E derived from equation (2), the above equation can be written :

$$\frac{\partial}{\partial x}\left\{ (Y - Fx)^T W (Y - Fx) \right\} = 0 \tag{7}$$

Using matrix theory it is shown that the above equation is equivalent to :

$$\left(F^T WF\right) x = F^T WY \tag{8}$$

Equation (8), called normal equation has the solution :

$$x = \left(F^T WF\right)^{-1} F^T W\ Y \tag{9}$$

Insofar as *npt-np* >> 1, an estimate of the parameter error (confidence interval 68%) is given by :

$$\left[x_j - \sqrt{\left(F^T W F\right)_{jj}^{-1} \chi^2}, x_j + \sqrt{\left(F^T W F\right)_{jj}^{-1} \chi^2}\right] \tag{10}$$

Where $x_j$ is estimation derived from (9) for the *j*<sup>th</sup> target parameter.

### 2.3.2 Non-linear least square

The model used to reproduce measurements is usually non-linear, that is to say that there is no matrix relationship Y = F X. However, in some special cases it is possible to linearize equation (1) using a suitable change of variable :

$$y_i = f_{ij}\left(x_j\right) + e_i \xrightarrow{g} y_i' = f_{ij}' x_j + e_i' \tag{11}$$

With :

$$y_i' = g\left(y_i\right) \tag{12}$$

$$f_{ij}' x_j + e_i' = g\left(f_{ij}\left(x_j\right) + e_i\right) \tag{13}$$

This linearization requires modifying the weights. Indeed, putting $y_i = g^{-1}(y_i') \text{ et } \hat{y}_i = g^{-1}(\hat{y}_i')$, we can write in the vicinity of the solution :

$$y_i - \hat{y}_i = (y_i' - \hat{y}_i') \frac{\partial g^{-1}}{\partial y_i'} \tag{14}$$

We must therefore minimize :

$$\sum_{i=1}^{npt} (y_i - \hat{y}_i)^2 w_i = \sum_{i=1}^{npt} (y_i' - \hat{y}_i')^2 \left( \frac{\partial g^{-1}}{\partial y_i'} \right)^2 w_i \tag{15}$$

either :

$$\sum_{i=1}^{npt} (y_i' - \hat{y}_i')^2 w_i' \tag{16}$$

Where the new weights $w_i'$ have the following expression :

$$w_i' = \frac{w_i}{\left( \frac{\partial g}{\partial y_i} \right)^2} \tag{17}$$

Since the partial derivatives of the variable change function and its inverse are connected by :

$$\left(\frac{\partial g^{-1}}{\partial y_i'}\right) = \frac{1}{\left(\frac{\partial g}{\partial y_i}\right)} \tag{18}$$

It then returns the minimizing to the linear case since by hypothesis :

$$y_i' = f_{ij}'x_j + e_i' \tag{19}$$

### 2.3.3 Linearization of the forward model

When equation (1) linking the model parameters and measured values is inherently nonlinear, we can try to linearize the model by a Taylor expansion in the vicinity of the initial solution *X⁰*. We can then write :

$$y_i = f_i\left(x^0\right) + \sum_{j=1}^{np}\left(x_j - x_j^0\right)\left(\frac{\partial f_i}{\partial x_j}\right)_{x=x^0} + \sum_{j=1}^{np}\sum_{k=1}^{np}\frac{1}{2}\left(x_j - x_j^0\right)\left(x_k - x_k^0\right)\left(\frac{\partial^2 f_i}{\partial x_j \partial x_k}\right)_{x=x^0} + \ldots \tag{20}$$

By retaining in (20) that the linear terms in X, we obtain in place of equation (2) :

$$Y = F\left(x^0\right) + K\left(x^0\right)\left(x - x^0\right) + E_2 \tag{21}$$

In which :

- $F(x^0)$ represents the value of the model for initial parameters $X^0$,
- $K(x^0)$ represents derivative matrix (Jacobian) of the model with respect to parameters, in the vicinity of the initial solution $x^0$ with :

$$k_{i,j} = \frac{\partial f_i(x)}{\partial x_j} \tag{22}$$

- $E_2$ represents the difference vector (or deviation vector) which may have different statistical properties of the E vector introduced in (2), because of the linearization powered

Then, equation (21) can be rewritten :

$$Y = F(x^0) + K(x^0)\Delta x + E_2 \qquad \text{with} : \Delta x = x - x^0 \tag{23}$$

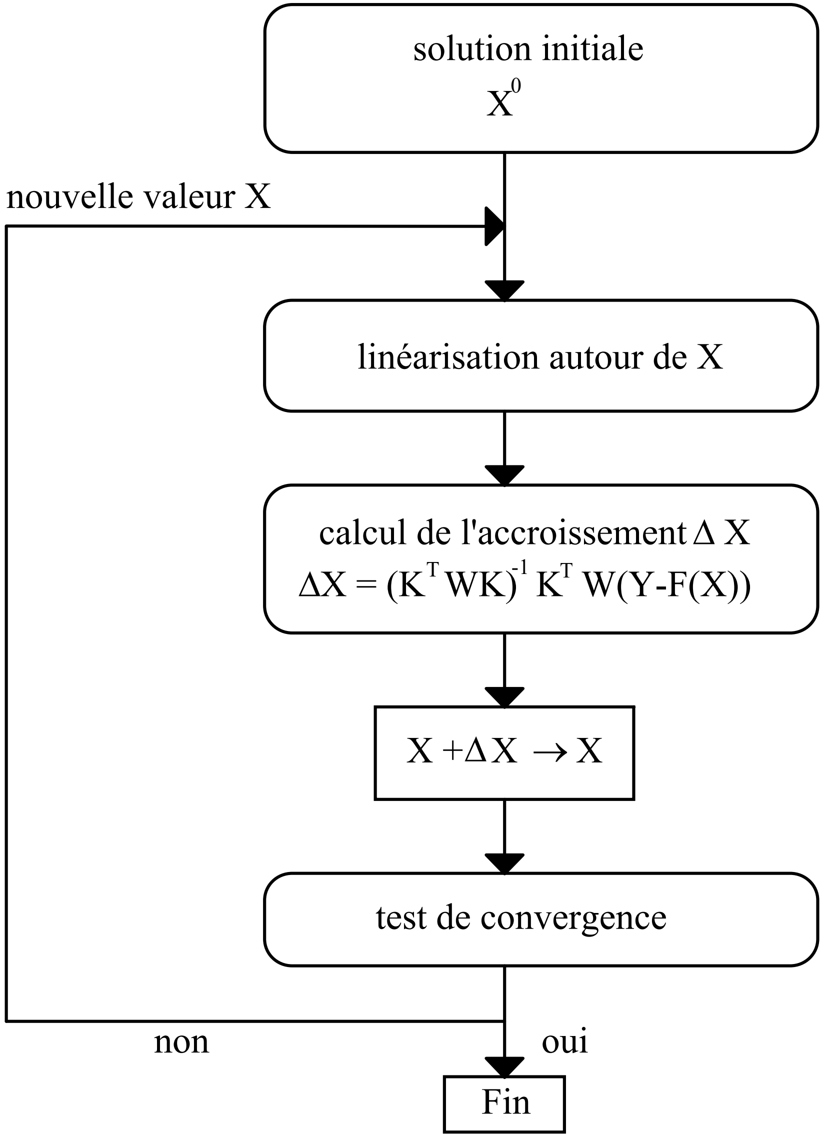
One can then apply linear least squares method on $\Delta x$, leading to the following normal equation :

$$\left(K^T W K\right)\Delta x = K^T W\left(Y - F(x)\right) \tag{24}$$

That is to say :

$$\Delta x = \left(K^T W K\right)^{-1} K^T W\left(Y - F(x)\right) \tag{25}$$

In practice, the inversion algorithm is iterative as in the flow chart shown in the following figure.


solution initiale
$X^0$
nouvelle valeur X
linéarisation autour de X
calcul de l'accroissement Δ X
$\Delta X = (K^T WK)^{-1} K^T W(Y-F(X))$
$X + \Delta X \rightarrow X$
test de convergence
non
oui
Fin

Flowchart of the inversion algorithm of the least squares method with damping

### 2.3.4 Damped least square

Insofar as the Taylor expansion limited to the first order (20) is not strictly correct, the value of parameter increment $\Delta x$ obtained by solving the equation (24) may be unrealistic and be particularly sensitive to measurement noise (uncertainty on observations Y). Under these conditions, it may be useful during the iterative process to limit the variations of parameters. This is called *damped least squares*, what is achieved by replacing the $\chi^2$ function by the new Z function (cost function) defined by :

$$Z = E^T W E + \lambda \Delta x^T S_a^{-1} \Delta x \tag{26}$$

with :

$$E = Y - \hat{Y} = Y - F(x) \tag{27}$$

where λ is the damping parameter and $s_a$ is an *a priori* estimate (or prior estimate) of the variance-covariance matrix of the parameters (assumed diagonal here). Thus, during the iterative process one penalizes too large variations of parameters with respect to their errors as they can be estimated *a priori*. Indeed, the diagonal elements of the variance-covariance matrix provide a statistical estimate of the error in the determination of the parameters. The introduction of a damping implies a constraint, on the variation of these parameters during the fit, in a domain limited approximately by the value of their error.

After linearization, the minimization of the new cost function Z defined in (26) leads to a new increment calculable as :

$$\Delta x = \left(K^{T}WK + \lambda S_{a}^{-1}\right)^{-1} K^{T}W\left(Y - F(x)\right) \tag{28}$$

We find under this form the expression used by Rodgers.

Two pitfalls must be avoided :

- if λ is too large (high damping), the solution will tend to remain confined to the vicinity of the initial solution $X^{0}$, although it is not the optimal solution;
- if λ is too small, one are practically reduced to the case of least squares without damping (this is the case if we make λ = 0 in equation (26).

In fact the λ parameter may be changed during the iteration process and may tend to zero in principle when the method converges. In practice, when a solution is obtained which appears to be satisfactory, one starts again from this solution as an initial value with zero damping and one verifies that the new solution does not differ from the initial solution (deviations lower than errors on the parameters). This means that the judicious choice of the parameter λ will more or less accelerate convergence, but does not affect, in principle, the solution found.

### 2.3.5 Levenberg-Marquardt method.

We discuss in this section a method that does not seek to come down to a linear relationship between the measured variables and parameters that have to be determined, but which is a procedure allowing to improve, at each step of an iterative process, the initial values of a set of given parameters, and this by minimizing the merit function $\chi^2$ defined above (equation (3) and (4)). This section has been writing using Numerical Recipes in Fortran[2].

Supposing that the function $\chi^2$ (expression (3) and (5)) can be approximated by a quadratic form :

$$\chi^2(x) = \chi^2\left(x^0\right) + \left(x - x^0\right)^T B + \frac{1}{2}\left(x - x^0\right)^T D\left(x - x^0\right) \tag{29}$$

where B is the gradient vector and D the Hessian matrix whose elements are defined respectively by :

$$b_j = \frac{\partial \chi^2\left(x^0\right)}{\partial x_j} = -2\sum_{i=1}^{npt} w_i (y_i - \hat{y}_i)\frac{\partial \hat{y}_i}{\partial x_j} \qquad j=1, np \tag{30}$$

$$d_{j,k} = \frac{\partial^2 \chi^2\left(x^0\right)}{\partial x_j \partial x_k} = 2\sum_{i=1}^{npt} w_i \left(\frac{\partial \hat{y}_i}{\partial x_j}\frac{\partial \hat{y}_i}{\partial x_k} - (y_i - \hat{y}_i)\frac{\partial^2 \hat{y}_i}{\partial x_j \partial x_k}\right) \qquad j,k=1, np \tag{31}$$

[2] Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P., Numerical Recipies in fortran (second edition)., Cambridge University Press, Cambridge, (1992),

In order to come down to a matrix writing more widespread of the problem, we define the matrix K and the matrix L as follows :

$$K_{i,j} = \frac{\partial \hat{y}_i}{\partial x_j} \qquad j=1,np \quad i=1,npt \tag{32}$$

$$L_{i,jk} = \frac{\partial^2 \hat{y}_i}{\partial x_j \partial x_k} \qquad j,k=1,np \quad i=1,npt \tag{33}$$

Then one can express the vector B and the matrix D in the following manner :

$$B = -2K^T WE \tag{34}$$

$$D = 2\left(K^T WK + L^T WE\right) \tag{35}$$

If $\Delta X = X - X^0$, we can rewrite equation (29) as:

$$\chi^2(x) = \chi^2\left(x^0\right) - 2\Delta x^T K^T WE + \Delta x^T\left(K^T WK + L^T WE\right)\Delta x \tag{36}$$

Deriving equation (36) with respect to $x$, one obtains:

$$\nabla\chi^2(x) = -2K^T WE + 2\left(K^T WK + L^T WE\right)\Delta x \tag{37}$$

if $\mathrm{x}$ minimize $\chi^2(\mathrm{x})$ then $\nabla\chi^2(\mathrm{x})=0$ and from above we have :

$$\Delta\mathrm{x} = \left(\mathrm{K}^{\mathrm{T}}\mathrm{WK}+\mathrm{L}^{\mathrm{T}}\mathrm{WE}\right)^{-1}\left(\mathrm{K}^{\mathrm{T}}\mathrm{WE}\right) \tag{38}$$

This set of equations is solved for the increment $\Delta\mathrm{x}$ that gives us the new parameter values for the next iteration (if they actually improve $\chi^2$) :

$$\mathrm{x}\rightarrow\mathrm{x}^0+\Delta\mathrm{x} \tag{39}$$

Using equation (38) implies that one is able to calculate the gradient and the Hessian of $\chi^2$.

This poses no problem in principle because we know to write the forward model.

Indeed, $\chi^2$ given by the expression (3) depends only on the values of the measured quantity, their estimation by the forward model $\left(\hat{\mathrm{Y}}=\mathrm{F}(\mathrm{x})\right)$, and uncertainty between measured value and estimation of the model. When derivate with respect to the target parameters, there remains only the term depending on the model and calculation of derivatives is therefore possible.

We can see that the elements of the matrices K and L, needed to calculate the Hessian matrix (expression (35)), depend on the first derivatives and second derivatives of $\chi^2$ (expression (32) and (33)) with respect to parameters. Some conventional treatments ignore the second derivatives, but do not always justify this approximation. We will also adopt this approximation after some explanations.

The second derivative term can be removed when nil (case of linear dependence) or small enough to be neglected compare the terms involving the first derivatives. But the term multiplying the second derivative is $\left(y_i - \hat{y}_i\right)$. For a suitable model, this term should be close to random measurement error at each point (of any sign) and should generally not be correlated with the model. Therefore, the second derivative terms tend to cancel when one summarizes on measurement points. In addition, it should be noted that the presence of the second derivative terms can be destabilizing if the model reproduces poorly measurement points or if it is contaminated by outliers that will be difficult to "compensated" by points for which differences have opposite signs.

In conclusion, we will use as definition of the elements of the Hessian matrix :

$$D = 2K^{T}WK \tag{40}$$

and we rewrite the expression (38) in the form :

$$\Delta x = \left(K^{T}WK\right)^{-1}\left(K^{T}WE\right) \tag{41}$$

If now the expression (36) is a poor local approximation of the function we seek to minimize, the gradient method of steepest descent with predetermined step (steepest descent method) consists to estimate a new value of the target parameters from the gradient vector multiplied by a positive damping coefficient μ and small enough that this estimate does not give a value too far from the starting solution $X^0$ :

$$\Delta x = -\mu B = -\mu 2K^{T}WE \tag{42}$$

and a new value of the parameters is determined in this way : $x \rightarrow x^0 + \Delta x$

Two methods, corresponding to the case where the model is good or bad accounted for measurements respectively, are considered when using an order 2 development of the function to be minimized. The Levenberg-Marquardt brings together these two methods in a single method.

The method of steepest gradient with predetermined step (42) is used far from the minimum, gradually giving way to the method of the inverse Hessian (41) when approaching the minimum. This method is of course iterative and determines a solution step by step changing at each step the relative importance of the two processes for the reduction of $\chi^2$.

$\chi^2$ is dimensionless while elements B have the dimension of $1/x_j$ (each element of B can be of different dimension). The multiplicative constant μ in equation (42) must therefore have the dimension of $x_j^2$. Thus the inverse of the diagonal elements of the Hessian matrix $1/\left(K^T WK\right)_{j,j}$ (having the dimension of $x_j^2$) allow us to obtain an estimate of the "bounds" of the constant μ, but this interval may be unrealistic, it is then divided by a factor $\lambda$ dimensionless with the opportunity to take it definitely below 1 allowing to increase the range in which is the new solution. We therefore replace equation (42) by integrating the factor of 2 in the $\lambda$ factor by :

$$\Delta x_j = \frac{1}{\lambda\left(K^T WK\right)_{j,j}}\left(K^T WE\right)_j \;\Rightarrow\; \lambda\left(K^T WK\right)_{j,j}\Delta x_j = \left(K^T WE\right)_j \qquad (43)$$

and using matrix notation, we can write :

$$\gamma \, \mathrm{I} \, \Delta \mathrm{x} = \mathrm{K}^{\mathrm{T}} \mathrm{WE} \tag{44}$$

where *I* is the identity matrix and $\gamma$ is the matrix whose elements are given by :

$$\begin{aligned} (\gamma)_{\mathrm{j,j}} &= \lambda \left( \mathrm{K}^{\mathrm{T}} \mathrm{WK} \right)_{\mathrm{j,j}} \qquad \mathrm{j} = 1, \mathrm{np} \\ (\gamma)_{\mathrm{j,k}} &= 0 \qquad \mathrm{j,k} = 1, \mathrm{np} \; ; \; \mathrm{j} \neq \mathrm{k} \end{aligned} \tag{45}$$

The relation (43) or (44) is true only if $1/\left( \mathrm{K}^{\mathrm{T}} \mathrm{WK} \right)_{\mathrm{j,j}}$ is positive for all k, which is true given its definition (equation (31)) for j = k and neglecting the second derivatives.

The Levenberg-Marquardt combines equations (44) and (41) for determining a new increment $\Delta \mathrm{x}$ :

$$\Delta \mathrm{x} = \left( \mathrm{K}^{\mathrm{T}} \mathrm{WK} + \gamma \mathrm{I} \right)^{-1} \left( \mathrm{K}^{\mathrm{T}} \mathrm{WE} \right) \tag{46}$$

which is equivalent to define a new Hessian matrix $\mathrm{D'} = \mathrm{K}^{\mathrm{T}} \mathrm{WK} + \gamma \, \mathrm{I}$

When $\lambda$ is very large, the matrix $D'$ is "forced" by its diagonal elements and equation (36) tends to be identical to equation (44).

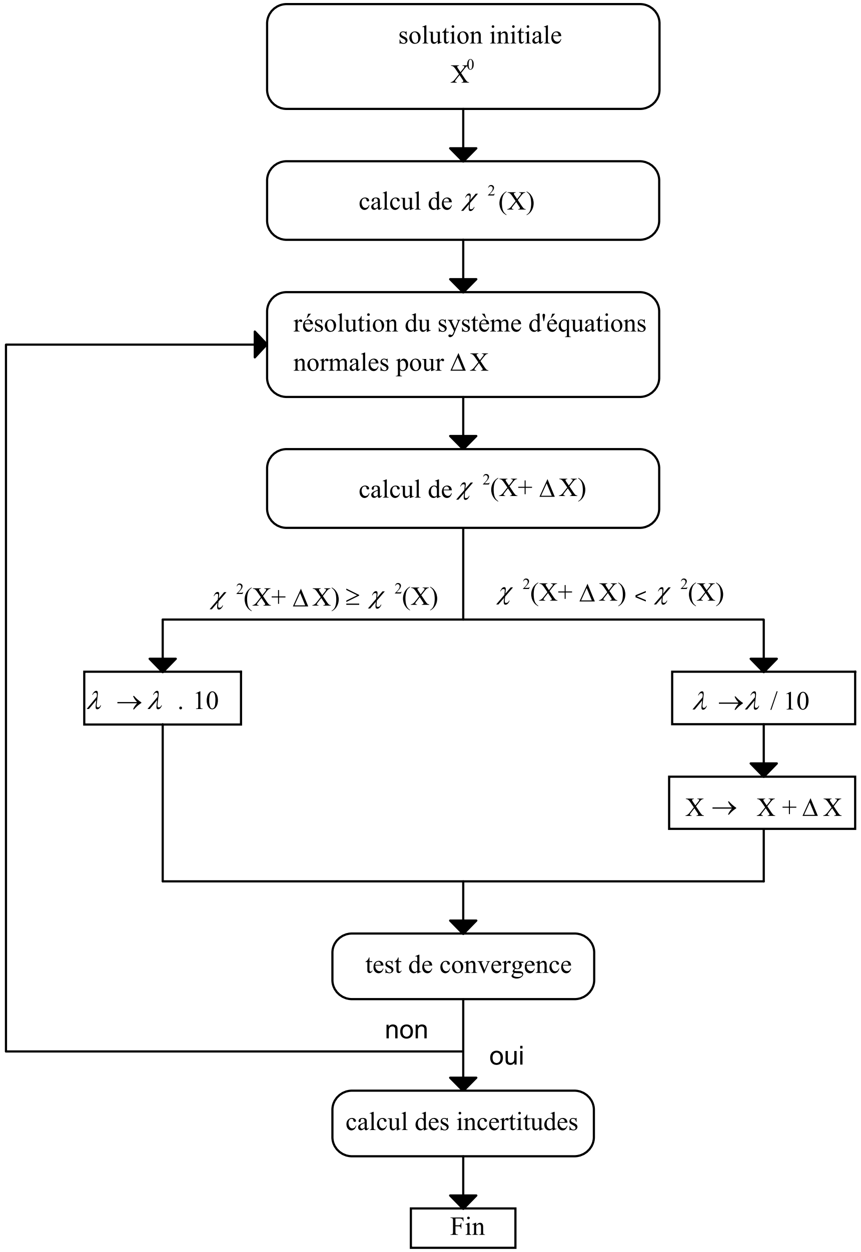
In contrast, when $\lambda$ is close to zero, equation (36) tends to be identical to equation (41). A set of input data being fixed and for a selected parameter X to retrieve, the inversion process is iterative.

It is necessary to establish a criteria to stop the iteration because in practice it is rarely to reach $\chi^2 = 1$. In fact, the sought minimum represents only the best statistical estimate of the parameters *X* and a parameter change that modify the $\chi^2$ of an amount lower than 1 is rarely significant.

When acceptable minimum has been found, we recompute the Hessian matrix $D'$ for λ = 0. When inversed, this matrix gives an estimate of the variance-covariance matrix of the errors on the determination of the retrieved parameters (the diagonal elements of the matrix corresponding to the variances). We can then obtain a confidence interval (confidence level of 68%) for each retrieved parameter.

$$\left[ x_j - \sqrt{\chi_r^2 \left(K^T W K\right)_{jj}^{-1}} \; , x_j + \sqrt{\chi_r^2 \left(K^T W K\right)_{jj}^{-1}} \right] \tag{47}$$

where $\chi_r^2$ is the reduced $\chi^2$ (equation (4)) and where $\left(K^T W K\right)_{jj}^{-1}$ is the *j*^*th*^ diagonal element of the inverse Hessian matrix.


solution initiale
$X^0$
calcul de $\chi^2(X)$
résolution du système d'équations normales pour $\Delta X$
calcul de $\chi^2(X+\Delta X)$
$\chi^2(X+\Delta X) \geq \chi^2(X)$
$\chi^2(X+\Delta X) < \chi^2(X)$
$\lambda \rightarrow \lambda \,.\, 10$
$\lambda \rightarrow \lambda \,/\, 10$
$X \rightarrow X + \Delta X$
test de convergence
non
oui
calcul des incertitudes
Fin

Flowchart of the inversion algorithm of Levenberg-Marquardt

### *2.4 A priori information*

Until then (except in 2.3.4), in the process of calculation, any **constraints on the physical reality** of the solutions have been taken into account during the retrieval. Then It could be possible that solution determined by the Levenberg-Marquardt corresponds to a **minimum of the cost function but has no physical reality**. We therefore introduced an additional constraint to take into account the consistency of the solutions via *a priori* information on the target parameters.

The ***a priori*** **information** can, for example, be the vertical profile of concentration of a chemical species you want to measure. In general, it is the **state of the atmosphere known before the measurement**. The *a prior*i profile can be used for simplicity as starting profile (first guest) but it is not an obligation.

This additional constraint is added to the cost function defined above ($\chi^2$ or Z) but $\Delta X = x - x_0$ becomes $x - x_a$ where $x_a$ is the vector of *a priori values* of the target parameters (before measurement), it is also a starting solution having a physical meaning.

We can also introduce the inverse of the error matrix, $S_e^{-1}$ whose values correspond to the diagonal elements $w_i$. Then we have :

$$Z = E^T S_e^{-1} E + \lambda \left( x - x_a \right)^T S_a^{-1} \left( x - x_a \right) \tag{48}$$

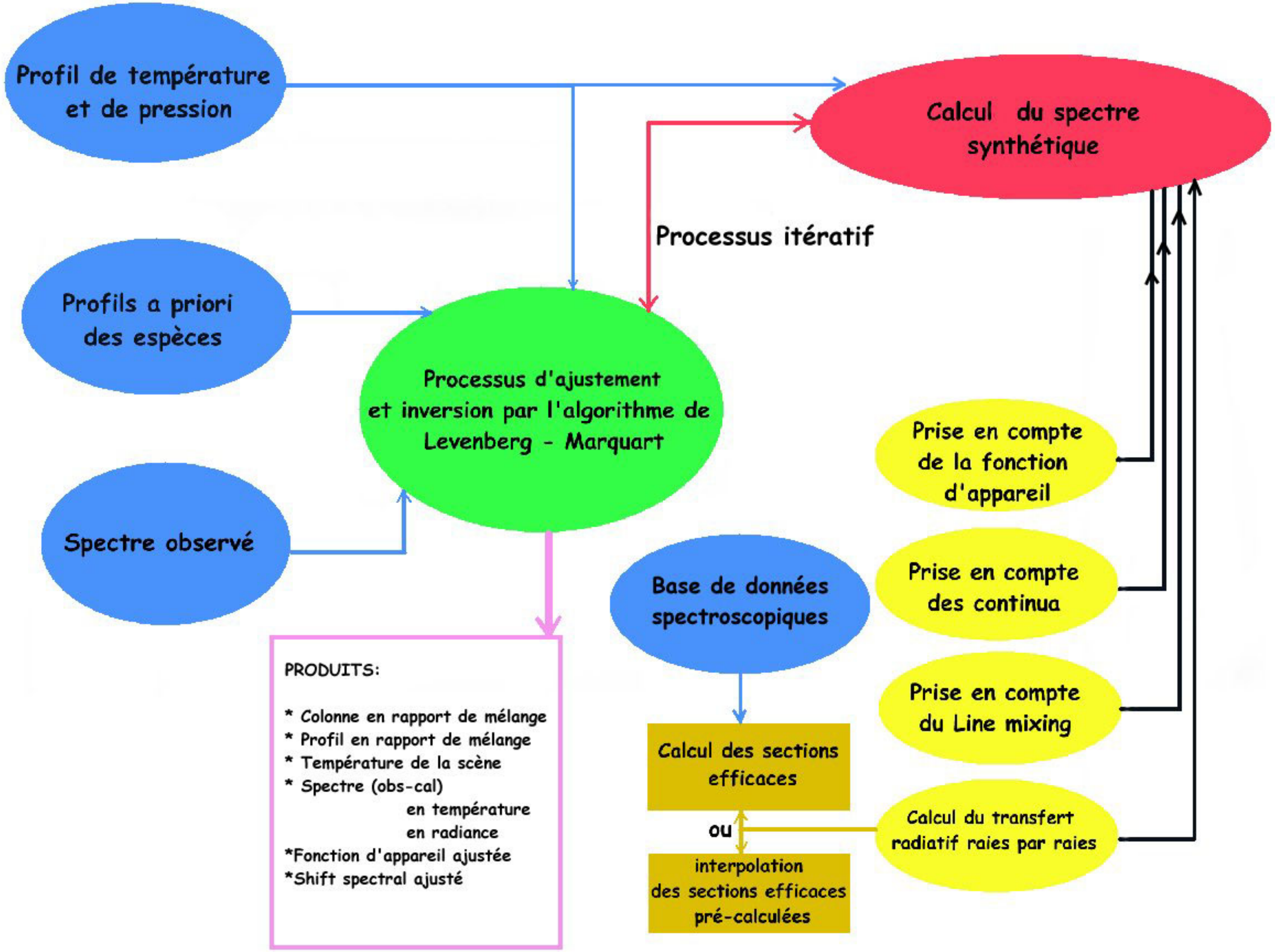
$S_a$ is an *a priori* estimate of the variance-covariance matrix of the parameters. Thus, one penalizes during the iterative process too large variations of parameters with respect to the *a priori* solution, taking into account the satisfactory "fluctuations" around this solution based on errors, also estimated *a priori*, the diagonal elements of the variance matrix covariance giving a **statistical estimate of the error** in the determination of the *a priori*.

Combined with the gradient method of steepest descent with undetermined step (as in the Levenberg-Marquardt described before) the increment of iteration has for new term:

$$\Delta x = \left(K^T S_e^{-1} K + S_a^{-1} + \gamma I\right)^{-1} \left(\left(K^T S_e^{-1} E\right) - S_a^{-1}\left(x - x_a\right)\right) \tag{49}$$

Thus, a new expression of the confidence interval (confidence level of 68%) for each retrieved parameter that takes into account the *a priori* information and those forces the solution in a field of physical validity, is given by :

$$\left[x_j - \sqrt{\chi_r^2 \left(K^T S_e^{-1} K + S_a^{-1}\right)_{jj}^{-1}}, x_j + \sqrt{\chi_r^2 \left(K^T S_e^{-1} K + S_a^{-1}\right)_{jj}^{-1}}\right] \tag{50}$$



Profil de température et de pression

Calcul du spectre synthétique

Processus itératif

Profils a priori des espèces

Processus d'ajustement et inversion par l'algorithme de Levenberg - Marquart

Spectre observé

Prise en compte de la fonction d'appareil

Prise en compte des continua

Base de données spectroscopiques

Prise en compte du Line mixing

Calcul des sections efficaces

ou

interpolation des sections efficaces pré-calculées

Calcul du transfert radiatif raies par raies

PRODUITS:

* Colonne en rapport de mélange
* Profil en rapport de mélange
* Température de la scène
* Spectre (obs-cal)
en température
en radiance
*Fonction d'appareil ajustée
*Shift spectral ajusté

# 3 Information content

Consider the case of remote sensing measurement of a vertical concentration profile. The efficiency and resolution (vertical) of a remote sensing sounding can be expressed in two different ways depending on whether one considers only the quality of the forward model (**weighting function**) or that takes into account the "power of the inverse model to resolve the fine vertical structures of the atmosphere "(**averaging kernels**). These two variables are used to represent each spectrum, the contribution to the absorption of molecules of the studied species according to their distribution in altitude.

## *3.1 Weighting functions/ Jacobians*

To show the weighting functions, one must use the expression between the observed spectrum of the state vector and the forward model :

$$\mathrm{y} = \mathrm{F}(\mathrm{x},\mathrm{b}) + \varepsilon \qquad (51)$$

In this expression $\mathrm{x}$ is the state vector of $np$ parameters, b represents the model parameters which are not fully known to the user, such as the spectroscopic parameters, the instrument line shape, weather data type temperature profiles and pressure. The term ε is the experimental error.

During the retrieval, we are looking to determine the state vector which allows better simulation of the observed spectrum through a radiative transfer algorithm, which is to write the vector :

$$\hat{x} = R[y, b, x_a] = R[F(x, b) + \varepsilon, b, x_a] \tag{52}$$

Where :

- $\hat{x}$ is the retrieved state vector.
- *R* is the function of the inverse model (transfer function).
- $x_a$ is the state vector with the *a priori* values of the desired parameters.

By linearizing the forward model in the vincinity of $x_a$ we have:

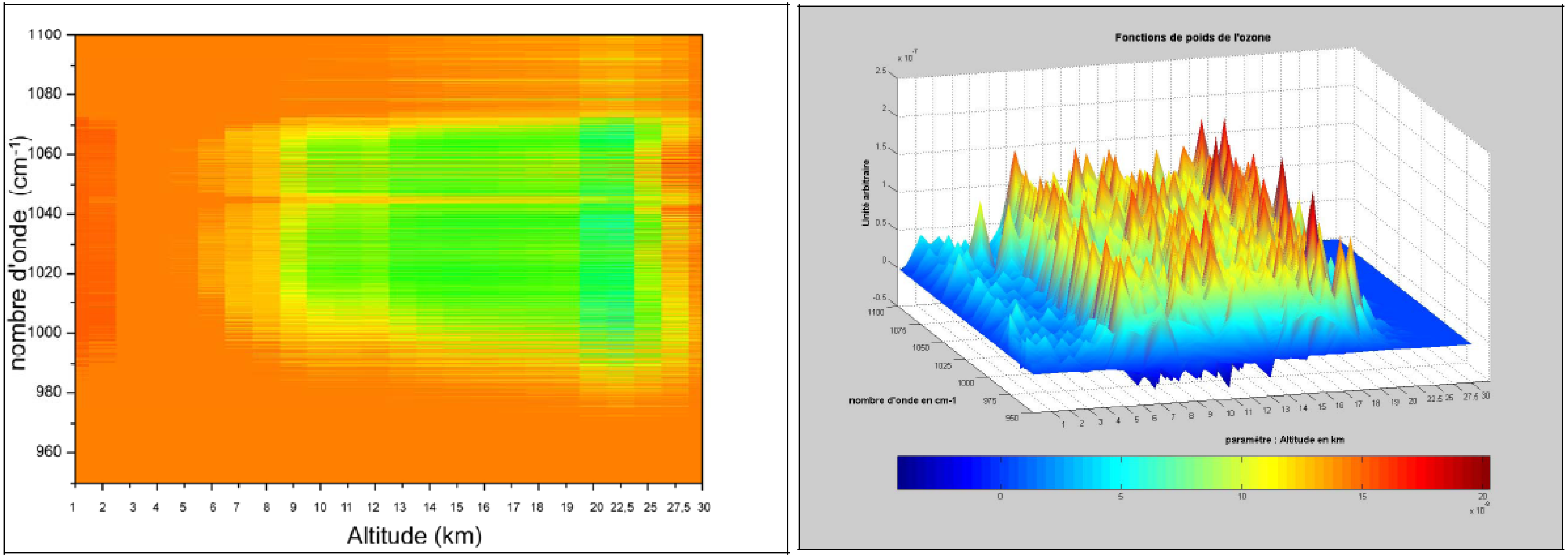
$$\hat{x} = R[F(x_a, b) + K_x(x - x_a) + \varepsilon, b, x_a] \tag{53}$$

It thus shows the weighting function matrix or Jacobian $K_x$ :

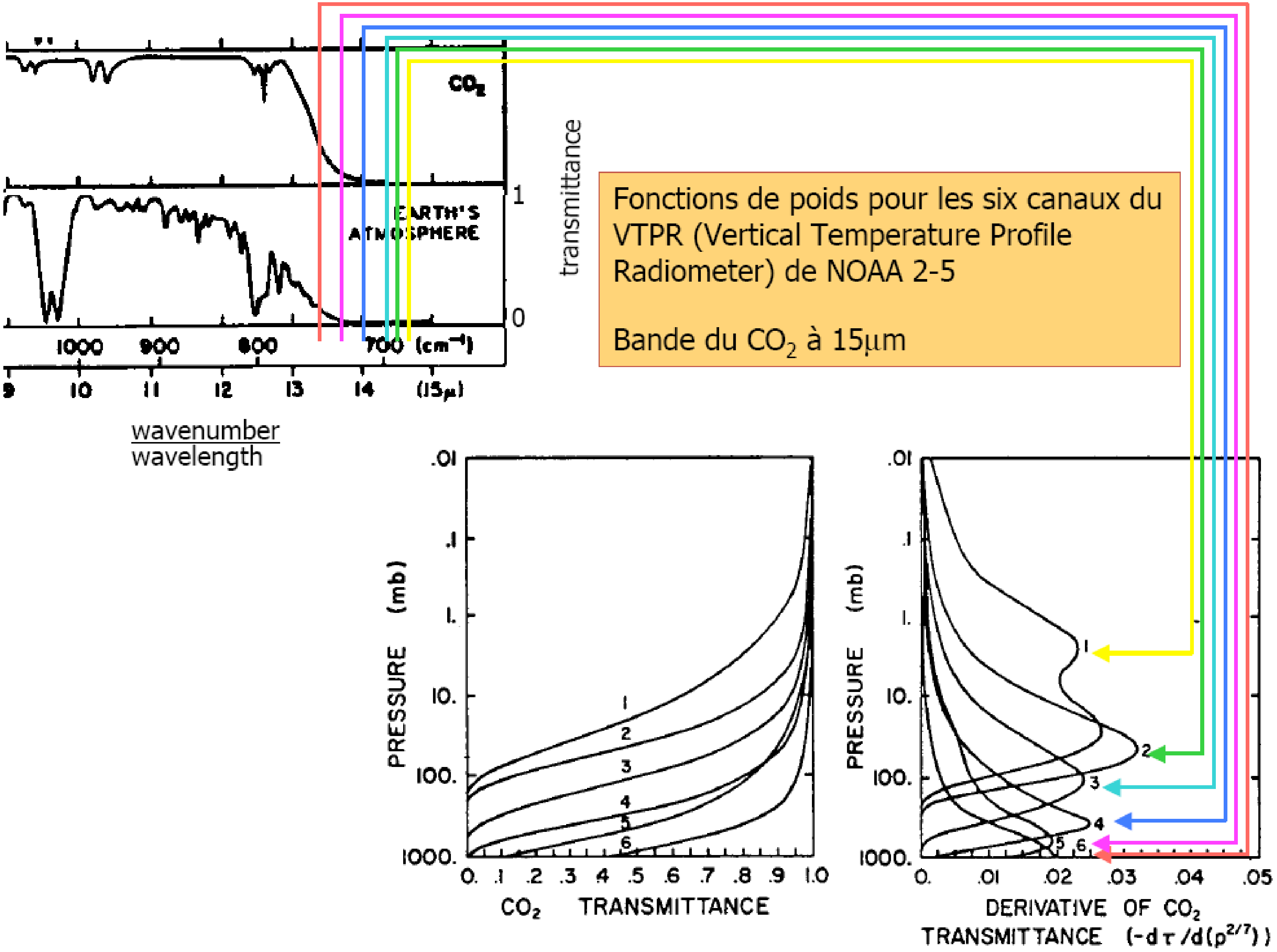
$$k_{i,j} = \frac{\partial f_i(x)}{\partial x_j} \tag{54}$$

It is a matrix K (*m* spectral points × *n retrieved parameters*)

The weight functions represent the sensitivity of the model parameters that are to be retrieved. These functions usually have the form of a peak for limb line of sight and are broadened for line of sight from the ground or nadir. The width at half maximum of weighting function peaks gives information on the vertical resolution of the measurement and on the sensitivity of the spectrum with respect to each retrieved parameter (sensitivity according to the altitude, for example).



**Graph of weighting functions in 2 and 3 dimensions for each altitude of the ozone profile (IASI balloon experience)**



Fonctions de poids pour les six canaux du VTPR (Vertical Temperature Profile Radiometer) de NOAA 2-5

Bande du CO$_2$ à 15μm [3]

[3] *www.lmd.ens.fr/jpduvel/Cours*

### 3.2 *Averaging kernels*

By linearizing the inverse model R with respect to y we obtain :

$$\hat{x} = R[F(x_a, b), b, x_a] + G_y[K_x(x - x_a) + \varepsilon, b, x_a] \tag{55}$$

where $G_y = \frac{\partial R}{\partial y}$ is the sensitivity of the inversion to the measurement. We can rewrite the previous expression as follows :

$$\begin{aligned} \hat{x} - x_a &= R[F(x_a, b)] - x_a && \ldots \text{ bias} \\ &+ A(x - x_a) && \ldots \text{ smoothing} \\ &+ G_y\,\varepsilon && \ldots \text{ retrieval error} \end{aligned} \tag{56}$$

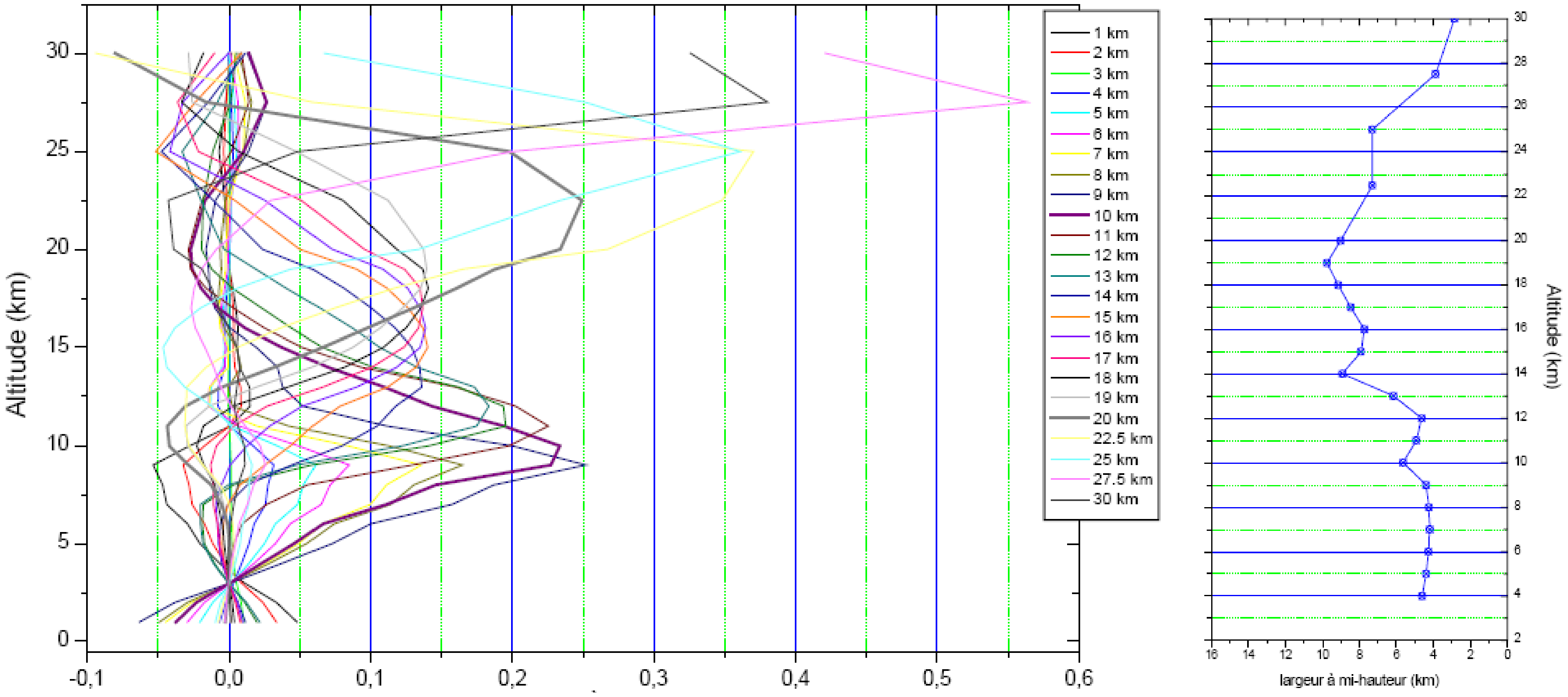
Where :

$$A = G_y K_x = \frac{\partial \hat{x}}{\partial X} \tag{57}$$

The rows of the matrix *A (n x n)* are the averaging kernels ("noyaux moyens") which can be considered one by one as the functions responsible for smoothing (averaging function) of the corresponding retrieved parameters.

**Each element of the result which comes from the retrieval appears as the product of the true values or retrieved by the corresponding averaging function.**

Under favorable conditions, the averaging kernel are "peaks" functions whose width at half maximum gives an estimate of the vertical resolution of the observing system. Ideally these averaging kernels are Dirac peaks. The integral of the averaging kernel area, usually close to unity, determines the contribution to the measurement retrieved parameters returned.



Averaging kernels for ozone profile retrieval from IASI-balloon experiment

Using averaging kernels and weighting functions it is possible to establish the **sensitivity of the retrieval** for each retrieved parameter and then to make an **optimal selection of the parameters** by keeping first those who have the greatest sensitivity spectrum and thus which are the most “exploitable” by the radiative transfer model. This saves computation time by limiting the number of parameters that are to be retrieved, but also quality of spectrum fit and of retrieval. With fewer parameters to retrieve while maintaining good flexibility in the fit avoids too much correlation between parameters, which generally introduces compensating effects that can lead to non-physical solutions.

### *3.3 Informational entropy / degrees of freedom*

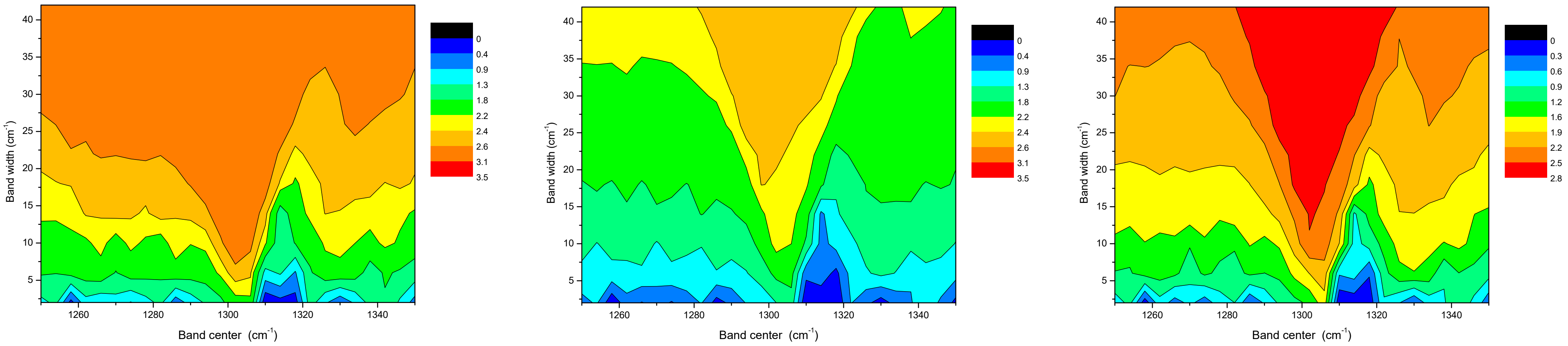
The number of degrees of freedom of a given signal can be considered as a measure of information. Rodgers takes the concept of information in the formulation of Shannon and apply them to inverse problems in atmospheric soundings. From the information entropy and the concept of degrees of freedom it is possible to estimate the information contained in a spectrum for a given chemical species

For more details, you can consult the Clive Rodgers book. We just give here the expression for the number of degrees of freedom :

$$\mathrm{d_s} = \sum_{\mathrm{i}} \mathrm{d_{s_i}} = \sum_{\mathrm{i}} \frac{\lambda_{\mathrm{i}}^2}{1+\lambda_{\mathrm{i}}^2} = \mathrm{tr}(\mathrm{A}) \tag{58}$$

where $\lambda_{\mathrm{i}}$ are singular values of the matrix $\widetilde{\mathrm{K}}$ defined by :

$$\widetilde{\mathrm{K}} = \mathrm{S}_{\varepsilon}^{-1/2}\mathrm{K}\,\mathrm{S}_{\mathrm{a}}^{1/2} \tag{59}$$



FWHM = 0.125 cm$^{-1}$

NEΔT = 0.1 K

FWHM = 0.125 cm$^{-1}$

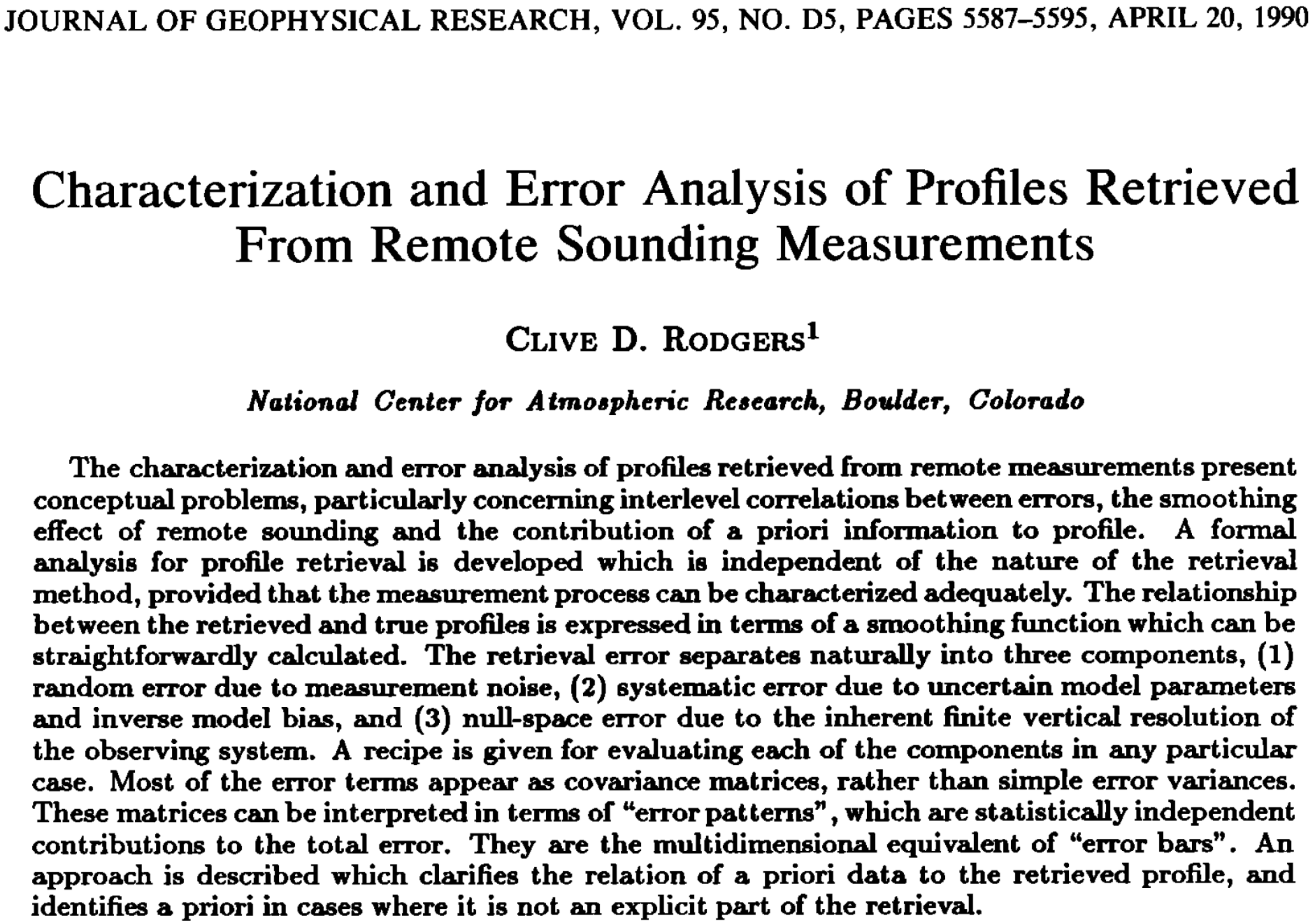
NEΔT = 0.20 K

FWHM = 0.250 cm$^{-1}$

NEΔT = 0.10 K]

**Example: DOFS for nadir looking CH$_4$ retrieval**

# 4 Analysis and characterization of errors



JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 95, NO. D5, PAGES 5587–5595, APRIL 20, 1990

## Characterization and Error Analysis of Profiles Retrieved From Remote Sounding Measurements

CLIVE D. RODGERS[1]

*National Center for Atmospheric Research, Boulder, Colorado*

The characterization and error analysis of profiles retrieved from remote measurements present conceptual problems, particularly concerning interlevel correlations between errors, the smoothing effect of remote sounding and the contribution of a priori information to profile. A formal analysis for profile retrieval is developed which is independent of the nature of the retrieval method, provided that the measurement process can be characterized adequately. The relationship between the retrieved and true profiles is expressed in terms of a smoothing function which can be straightforwardly calculated. The retrieval error separates naturally into three components, (1) random error due to measurement noise, (2) systematic error due to uncertain model parameters and inverse model bias, and (3) null-space error due to the inherent finite vertical resolution of the observing system. A recipe is given for evaluating each of the components in any particular case. Most of the error terms appear as covariance matrices, rather than simple error variances. These matrices can be interpreted in terms of "error patterns", which are statistically independent contributions to the total error. They are the multidimensional equivalent of "error bars". An approach is described which clarifies the relation of a priori data to the retrieved profile, and identifies a priori in cases where it is not an explicit part of the retrieval.

In the linear approximation of the retrieval method presented above, we can write the total error as the sum of four terms.

### *4.1 The different sources of error*

Resuming the transfer function. The state vector is a reversed function of the  :

$$\hat{\mathrm{x}} = \mathrm{R}(\mathrm{y}, \hat{\mathrm{b}}, \mathrm{c})$$

The relation with the real state X is the following :

$$\hat{\mathrm{x}} = \mathrm{R}(\mathrm{f}(\mathrm{x}, \mathrm{b}) + \epsilon, \hat{\mathrm{b}}, \mathrm{c})$$

That we can represent as follows :

$$\hat{\mathrm{x}} = \mathrm{T}(\mathrm{x}, \mathrm{b}, \epsilon, \hat{\mathrm{b}}, \mathrm{c})$$

$$\hat{\mathbf{x}} = T(\bar{\mathbf{x}}, \hat{\mathbf{b}}, \hat{\mathbf{c}}) + \mathbf{D}_y \mathbf{K}_x (\mathbf{x} - \bar{\mathbf{x}}) + \mathbf{D}_y \mathbf{K}_b (\mathbf{b} - \hat{\mathbf{b}}) + \mathbf{D}_y \epsilon_y$$

$$\hat{\mathbf{x}} - \mathbf{x} = [T(\bar{\mathbf{x}}, \hat{\mathbf{b}}, \hat{\mathbf{c}}) - \bar{\mathbf{x}}] + (\mathbf{A} - \mathbf{I})(\mathbf{x} - \bar{\mathbf{x}}) + \mathbf{D}_y \mathbf{K}_b \epsilon_b + \mathbf{D}_y \epsilon_y$$

By linearizing the transfer function, Clive Rodgers shows that we can write the error of retrieval in the following form :

$$\begin{array}{rcl} \hat{\mathbf{x}} - \mathbf{x} = & (\mathbf{A} - \mathbf{I})(\mathbf{x} - \mathbf{x}_a) & \textit{smoothing} \\ & +\mathbf{G}_y\mathbf{K}_b(\mathbf{b} - \hat{\mathbf{b}}) & \textit{model parameters} \\ & +\mathbf{G}_y\Delta\mathbf{f}(\mathbf{x}, \mathbf{b}, \mathbf{b}') & \textit{modelling error} \\ & +\mathbf{G}_y\epsilon & \textit{measurement noise} \end{array}$$

| | |
|---|---|
| $\partial\hat{\mathbf{x}}/\partial\mathbf{x} = \mathbf{A}$ | Sensitivity to the real state: averaging kernels. |
| $\partial\hat{\mathbf{x}}/\partial\epsilon = \mathbf{G}_y$ | Sensitivity to noise measurement |
| $\partial\hat{\mathbf{x}}/\partial\mathbf{b} = \mathbf{G}_b$ | Sensitivity to non-retrieved parameters |
| $\partial\hat{\mathbf{x}}/\partial\mathbf{c} = \mathbf{G}_c$ | Sensitivity to the inverse method |

The sum of the first two terms (smoothing error or smoothing, and measurement error) may be associated with the **internal errors**, which are specific to the viewing geometry and performance of the instrument, while the sum the other two terms, associated to the atmosphere knowledge is the **external error**.

Some of these terms are easy to estimate, and the other not.

### 4.2 Measurement noise

The measurement error is due to the instrumental noise. Its covariance matrix is given by:

$$\mathbf{S}_n = \mathbf{G}_y \mathbf{S}_\epsilon \mathbf{G}_y^T$$

### 4.3 Smoothing error

This error considers smoothing true profile (or the true state vector) by means nuclei. The covariance matrix of the error smoothing is given by:

$$\mathbf{S_s = (A - I)\, S_a\, (A - I)^T}$$

To determine properly this error it is necessary to have a good knowledge of the climatology covariance of the target parameters.

### 4.4 Forward model parameters

This is the error on the non-retrieved parameters of the forward model. Its covariance is written :

$$\mathbf{S}_p = \mathbf{GK}_b \mathbf{S}_b (\mathbf{GK}_b)^T$$

Where $\mathbf{S}_b$ is the covariance matrix representing the uncertainty on the fixed parameters of the forward model.

### *4.5 Forward model*

$$\mathit{modelling\ error} = \mathbf{G}_y \Delta \mathbf{f} = \mathbf{G}_y(\mathbf{f}(\mathbf{x}, \mathbf{b}, \mathbf{b}') - \mathbf{F}(\mathbf{x}, \mathbf{b}))$$

Often very difficult to determine because it requires a model *f* to quantify the differences between the forward model and physical reality ...

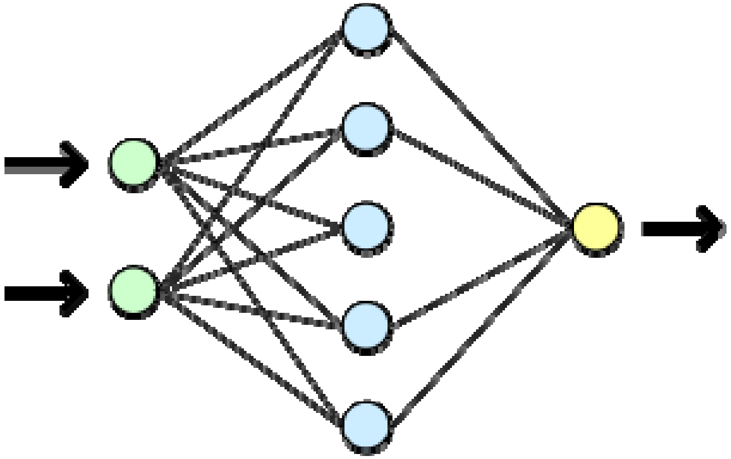
### *4.6 Covariance error matrix*

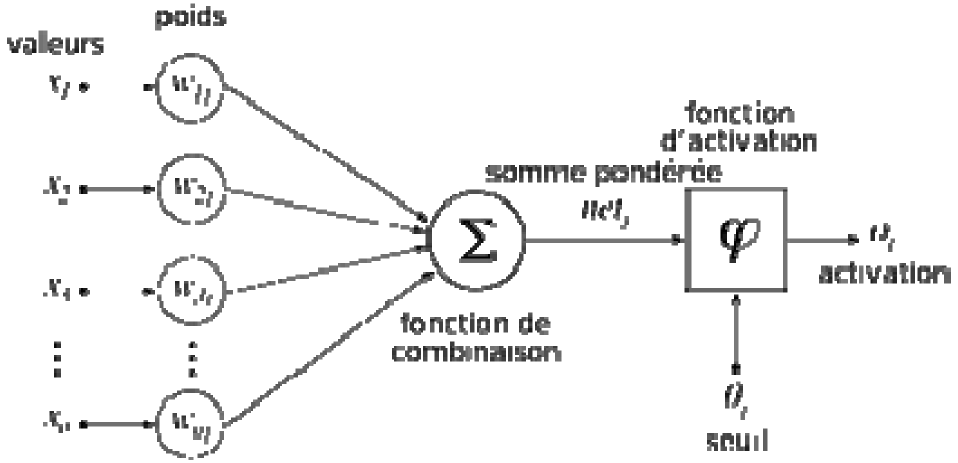
The total error covariance matrix is then given by:

$$\mathbf{S}_T = (\mathbf{S}_s + \mathbf{S}_m) + (\mathbf{S}_p + \mathbf{S}_{cs})$$

## 5 Other methods

There are other methods to "inverse" the atmospheric spectral measurements. One of them is based on artificial neural networks. It is a computational model whose design is inspired schematically from the operation of biological neurons. Neural networks are generally optimized by probabilistic learning methods, especially Bayesian.





Neural network has generally a "training" algorithm that consists in modifying the synaptic weights according to a set of data presented to the input of the network. The purpose of this training is to enable the neural network to "learn" from examples. If the training is properly performed, the network is able to provide output results very close to original values of the set of training data.

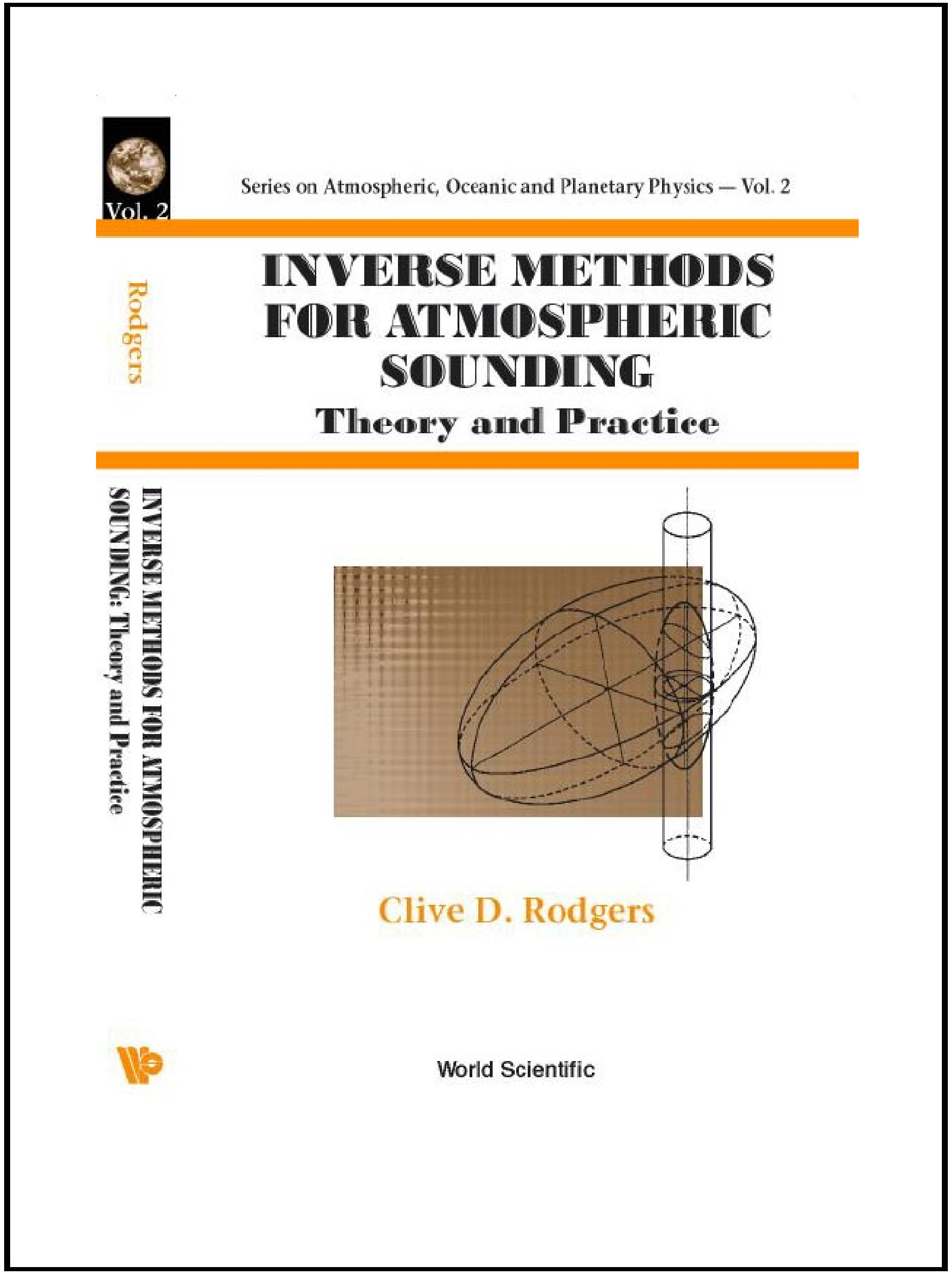
But the whole point of neural networks lies in their ability to generalize from the test set.

To go further **:**

C.D. Rodgers, Inverse methods for atmospheric sounding: theory and practice, World Scientific, Singapore (2000).

Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P., Numerical Recipies in fortran (second edition)., Cambridge University Press, Cambridge, (1992) .

Physique moléculaire Physique de l'atmosphère - Molecular Physics, Atmospheric Physics, Editions du CNRS 1982, Montfoulon (preès d'Alençon), 1-10 decembre 1980, C. Camy-Peyret editor, isbn 1-222-03104-4



Vol. 2

Series on Atmospheric, Oceanic and Planetary Physics — Vol. 2

Rodgers

INVERSE METHODS FOR ATMOSPHERIC SOUNDING

Theory and Practice

INVERSE METHODS FOR ATMOSPHERIC SOUNDING: Theory and Practice

Clive D. Rodgers

World Scientific