Chapter 4

Inverse Methods and Retrieval Techniques

In this chapter we will discuss inverse methods and retrieval techniques using the example of temperature profile retrieval from the idealized nadir infrared sounder, introduced in the previous chapter.

4.1 Example: Retrieval of atmospheric temperature profiles

In the idealized example obove, we have seen that the measured brightness temperature at frequency ν is given by the vertical integral over the atmospheric temperatures, weighted by the weighting functions K_ν:

$$T_{b,\nu}(\infty) = \int_0^\infty T(z) K_\nu(z)\, dz \tag{4.1}$$

(here we have neglected the surface contribution for simplicity, see the corresponding equation above).

To solve this numerically, we will replace the integral by a sum over some finite layers:

$$T_{b,j}(\infty) = \sum_i T_i K_{j,i} \Delta z_i. \tag{4.2}$$

Here T_i is the Temperature at level i, Δz_i is the thickness of layer i, $T_{b,j}(\infty)$ is the measured brightness temperature at frequency j, and $K_{j,i}$ is the weighting function at layer i and frequency j.

This can be written in more compact matrix notation as

$$\mathbf{y} = \mathbf{K} \cdot \mathbf{x}, \tag{4.3}$$

where $\mathbf{y}$ is a vector containing the measured brightness temperatures ($y_j = T_{b,j}(\infty)$), $\mathbf{x}$ is a vector containg the atmospheric temperature profile ($x_i = T_i$) and $\mathbf{K}$ is the weighting function matrix (with elements $K_{j,i}\Delta z_i$).

This equation thus provides us with a way to calculate the measured brightness temperature from an atmospheric temperature profile (and is sometimes called the 'forward' equation. In remote sensing we want to go the oposite way: retrieving the atmospheric temperature profile from the measured brightness temperature. This can be done by the so-called 'inversion' of the forward equation. Naively, this may be done by direct inversion of the weighting function matrix:

$$\mathbf{x} = \mathbf{K}^{-1} \cdot \mathbf{y}. \tag{4.4}$$

However, not only is the inverse of $\mathbf{K}$ only defined if $i = j$, it is also generally very sensitive to small errors in $\mathbf{y}$ and will not give any meaningfull result. This is an example of a so called ill-posed problem.

4.2 Introduction to estimation theory

Here we search for a solution of equation

$$\mathbf{A}\mathbf{x} = \mathbf{d} \tag{4.5}$$

where $\mathbf{d}$ are our measured data and $\mathbf{x}$ is the unknown parameter vector, in our example the unknown atmospheric temperature profile.

4.2.1 Vectors and matrices

The equation above can be written in components as

$$d_i = \sum_j A_{ij} x_j \tag{4.6}$$

where $i = 1, \ldots, m$ and $j = 1, \ldots, n$. The product of two matrizes is defined as

$$\mathbf{AB} = \sum_k A_{ik} B_{kj} \tag{4.7}$$

Vectors can thus be seen as a special case of an $n \times 1$-matrix; in our notation vectors are *column*-vectors. *Row*-vectors (the transpose $\mathbf{x}^T$ of a column vector $\mathbf{x}$) are linear forms, i.e. linear maps of a vector onto a scalar.

We define the *null space* of a matrix (german: Kern der Matrix) $\mathbf{A}$ as the set of all $\mathbf{x}$ for which

$$\mathbf{Ax} = 0 \tag{4.8}$$

The null space itself is a vector space.

We define the *rank* of a Matrix $\mathbf{A}$ as the number of linearly independent rows of $\mathbf{A}$.

$$\text{dim}(\text{null space}(\mathbf{A})) + \text{rank}(\mathbf{A}) = n \tag{4.9}$$

If an inverse to $\mathbf{A}$ exists, then $\mathbf{A}$ is called *regular*, if not $\mathbf{A}$ is *singular*. For regular matrices

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{d} \tag{4.10}$$

However, for remote sensing, the weighting function $\mathbf{A}$ is almost always singular or nearly singular, so that we cannot simply invert $\mathbf{A}$.

4.3 The overdetermined case

Let us call the difference between the data $\mathbf{d}$ and the linear model $\mathbf{A}x$ with

$$\mathbf{e} = \mathbf{Ax} - \mathbf{d} \tag{4.11}$$

We now search for a solution $\mathbf{x}$ for which the norm of $\mathbf{e}$ is minimum:

$$|\mathbf{e}| = \sqrt{\sum_i e_i^2} = \sqrt{\mathbf{e}^T\mathbf{e}} \tag{4.12}$$

or

$$|\mathbf{e}|^2 = \mathbf{e}^T\mathbf{e} \tag{4.13}$$

$$\begin{aligned} |\mathbf{e}|^2 &= (\mathbf{Ax} - \mathbf{d})^T(\mathbf{Ax} - \mathbf{d}) && (4.14)\\ &= (\mathbf{x}^T\mathbf{A}^T - \mathbf{d}^T)(\mathbf{Ax} - \mathbf{d}) && (4.15)\\ &= \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{A}^T\mathbf{d} - \mathbf{d}^T\mathbf{Ax} + \mathbf{d}^T\mathbf{d} && (4.16) \end{aligned}$$

At the minimum

$$\frac{\partial|\mathbf{e}|^2}{\partial x_k} = 0 \tag{4.17}$$

$$
\begin{aligned}
\frac{\partial |\mathbf{e}|^2}{\partial x_k} &= \mathbf{I}_k^T \mathbf{A}^T \mathbf{A} \mathbf{x} + \mathbf{x}^T \mathbf{A}^T \mathbf{A} \mathbf{I}_k - \mathbf{I}_k^T \mathbf{A}^T \mathbf{d} - \mathbf{d}^T \mathbf{A} \mathbf{I}_k && (4.18)\\
&= \mathbf{I}_k^T (\mathbf{A}^T \mathbf{A} \mathbf{x} - \mathbf{A}^T \mathbf{d}) + (\mathbf{x}^T \mathbf{A}^T \mathbf{A} - \mathbf{d}^T \mathbf{A}) \mathbf{I}_k && (4.19)\\
&= 2\mathbf{I}_k^T (\mathbf{A}^T \mathbf{A} \mathbf{x} - \mathbf{A}^T \mathbf{d}) && (4.20)
\end{aligned}
$$

Thus the minimum difference between data $\mathbf{d}$ and model $\mathbf{A}x$ is at

$$\mathbf{A}^T \mathbf{A} x - \mathbf{A}^T \mathbf{d} = 0 \tag{4.21}$$

and thus the so called *least squares* solution to the equation $\mathbf{A}x = \mathbf{d}$ is given as

$$\hat{\mathbf{x}} = \left(\mathbf{A}^T \mathbf{A}\right)^{-1} \mathbf{A}^T \mathbf{d} \tag{4.22}$$

4.3.1 Statistical basics

Consider a random variable x. Individual measurements of x will lead to the actual value x_i. The *expectation value* $E(x)$ is given by

$$E(x) \approx \bar{x} = \frac{1}{N} \sum_i^N x_i \tag{4.23}$$

$$E(x) = \lim_{N \to \infty} \frac{1}{N} \sum_i^N x_i \tag{4.24}$$

The *covariance* between two (scalar) random variables u and v is defined as

$$
\begin{aligned}
\sigma_{uv}^2 &= \lim_{N \to \infty} \frac{1}{N} \sum_i \left(u_i - E(u)\right)\left(v_i - E(v)\right) && (4.25)\\
&= E\left(\left(u - E(u)\right)\left(v - E(v)\right)\right) && (4.26)
\end{aligned}
$$

More generally for a random vector $\mathbf{x}$:

$$\sigma_{jk}^2(\mathbf{x}) = lim_{N \to \infty} \frac{1}{N} \sum_i \left(x_{ij} - E(x_j)\right)\left(x_{ik} - E(x_k)\right) \tag{4.27}$$

The covariances thus for a matrix, the covariance matrix $\mathrm{cov}(\mathbf{x})$. The empirical covarainvce matrix kann be calculated as

$$S_{jk}^2 = \frac{1}{N-1} \sum_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \tag{4.28}$$

with

$$\bar{x}_j = \frac{1}{N}\sum_i x_{ij} \tag{4.29}$$

The empirical correlation coefficient is defined as

$$r_{jk} = \frac{S_{jk}^2}{S_j S_k} \tag{4.30}$$

with

$$S_j = \sqrt{S_{jj}^2} \tag{4.31}$$

Note that constants can be extracted from expectation values, i.e.

$$E(\mathbf{Ax}) = \mathbf{A}E(\mathbf{x}) \tag{4.32}$$

It thus follows:

$$\mathrm{cov}(\mathbf{Ax}) = \mathbf{A}\mathrm{cov}(\mathbf{x})\mathbf{A}^T \tag{4.33}$$

4.4 Singular value decomposition

Every real $m \times n$ matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T \tag{4.34}$$

with

$\mathbf{U}$ an orthogonal $m \times m$ matrix

$\mathbf{V}$ an orthogonal $n \times n$ matrix

$\mathbf{S}$ an $m \times n$ diagonal matrix with $s_i = S_{ii} > 0$ and $S_{ij} = 0$.

The s_i are called the singular values of $\mathbf{A}$. The matrices $\mathbf{U}$ and $\mathbf{V}$ form an orthonormal basis of the m-dimensional data space and the n-dimensional parameter space, respectively. Base vectors $\mathbf{V}_i$ that belong to vanishing singular values ($s_i = 0$) are in the null space of $\mathbf{A}$. The number of non-zero singular vectors corresponds to the rank of $\mathbf{A}$.

Let p be the number of non-zero singular values:

$$\mathbf{A} = [\mathbf{U}_1 \ldots \mathbf{U}_p \mathbf{U}_{p+1} \ldots \mathbf{U}_m] \begin{bmatrix} s_1 & & & & & \\ & \ddots & & & & \\ & & s_p & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \vdots \\ \mathbf{V}_p^T \\ \mathbf{V}_{p+1}^T \\ \vdots \\ \mathbf{V}_m^T \end{bmatrix} \tag{4.35}$$

$$[\mathbf{U}_1 \ldots \mathbf{U}_p] \begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_p \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \vdots \\ \mathbf{V}_p^T \end{bmatrix} \tag{4.36}$$

$$\mathbf{A} = \mathbf{U}_p \mathbf{S}_p \mathbf{V}_p^T \tag{4.37}$$

The singular value decomposition can be used to find the least squares solution

$$\mathbf{A}^T \mathbf{A} \hat{\mathbf{x}} = \mathbf{A}^T \mathbf{d} : \tag{4.38}$$

$$\mathbf{V}_p \mathbf{S}_p \mathbf{U}_p^T \mathbf{U}_p \mathbf{S}_p \mathbf{V}_p^T \hat{\mathbf{x}} = \mathbf{V}_p \mathbf{S}_p \mathbf{U}_p^T \mathbf{d} \tag{4.39}$$

$$\mathbf{V}_p \mathbf{S}_p^2 \mathbf{V}_p^T \hat{\mathbf{x}} = \mathbf{V}_p \mathbf{S}_p \mathbf{U}_p^T \mathbf{d} \tag{4.40}$$

$$\hat{\mathbf{x}} = \mathbf{V}_p \mathbf{S}_p^{-1} \mathbf{U}_p^T \mathbf{d} \tag{4.41}$$

This is easy to calculate once the singular value decomposition (svd) has been calculated, because

$$\mathbf{S}_p^{-1} = \begin{bmatrix} 1/s_1 & & \\ & \ddots & \\ & & 1/s_p \end{bmatrix} \tag{4.42}$$

The matrix $\mathbf{V}_p \mathbf{S}_p^{-1} \mathbf{U}_p^T$ is called the pseudo inverse of $\mathbf{A}$.

If a matrix has vanishing singular values (i.e. some $s_i = 0$), the matrix is singular and no inverse exists. However, in many cases the singular values are not exactly zero, but decrease exponentially; this leads to large numbers in the calculation of the pseudo inserve with large amplification of measurement noise. So call ill-posed problems.

To solve this, the ‘method of truncated singular values can be used. For this, singular values below a certain threshold will be set exactly to zero. This is one example of a *regularization* method.

4.5 Tikhonov-Phillips Regularization

So far we have searched for the smallest $\hat{\mathbf{x}}$ (in the sense of a quadratic norm) that minimizes the difference to the data. However, we may add arbitrary vectors from the null space to the solution that agree as good with the data. To make the soilution unique, other conditions or constraints can be considered.

We consider the two functionals

$$\mathcal{A}[\mathbf{x}] = ||\mathbf{Ax} - \mathbf{d}||^2 \tag{4.43}$$

and

$$\mathcal{B}[\mathbf{x}] = ||\mathbf{Bx} - \mathbf{b}||^2 \tag{4.44}$$

I.e., we want to solve $\mathbf{Ax} = \mathbf{d}$ under the additional condition that $\mathbf{B}x = \mathbf{b}$. One example could be that we require $\mathbf{x}$ to be close to some climatology $\mathbf{b}$. We weight both conditions with some parameter γ and minimize

$$||\mathbf{A}x - \mathbf{d}||^2 + \gamma^2 ||\mathbf{Bx} - \mathbf{b}||^2. \tag{4.45}$$

This leads to (similar to the derivation of the ordinary least squares solution):

$$(\mathbf{A}^T\mathbf{A} + \gamma^2\mathbf{B}^T\mathbf{B})\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{d} + \gamma^2\mathbf{B}^T\mathbf{b} \tag{4.46}$$

or

$$\hat{\mathbf{x}} = (\mathbf{A}^T\mathbf{A} + \gamma^2\mathbf{B}^T\mathbf{B})^{-1}(\mathbf{A}^T\mathbf{d} + \gamma^2\mathbf{B}^T\mathbf{b}) \tag{4.47}$$

One common condition is to require $\hat{\mathbf{x}}$ to be close to some a priori value $\mathbf{x}_0$, so that $\mathbf{B} = \mathbf{I}$ and $\mathbf{b} = \mathbf{x}_0$. Then:

$$\hat{\mathbf{x}} = \mathbf{x}_0 + (\mathbf{A}^T\mathbf{A} + \gamma^2\mathbf{I})^{-1}\mathbf{A}^T(\mathbf{d} - \mathbf{A}x_0) \tag{4.48}$$

4.6 Optimal Estimation

We look for a solution $\hat{\mathbf{x}}$ that minimizes the functional

$$(\mathbf{Ax} - \mathbf{d})^T \mathrm{cov}(\mathbf{d})^{-1}(\mathbf{Ax} - \mathbf{d}) + (\mathbf{x} - \mathbf{x}_0)^T \mathrm{cov}(\mathbf{x})^{-1}(\mathbf{x} - \mathbf{x}_0) \tag{4.49}$$

It follows

$$\hat{\mathbf{x}} = \mathrm{cov}(\mathbf{x})\mathbf{A}^T(\mathbf{A}\mathrm{cov}(\mathbf{x})\mathbf{A}^T + \mathrm{cov}(\mathbf{d}))^{-1}(\mathbf{d} - \mathbf{Ax}_0) + \mathbf{x}_0 \tag{4.50}$$

This is the so called optimal estimation solution.

One way to derive the optimal estimation solution is to start from the theorem of Gauss-Markov:

Let $\mathbf{x}$ and $\mathbf{d}$ be random variables with expectation values

$$E(\mathbf{x}) = E(\mathbf{d}) = 0 \tag{4.51}$$

and given covariance matrices

$$\mathbf{R}_{\mathbf{x}} = E(\mathbf{x}\mathbf{x}^T) \tag{4.52}$$

$$\mathbf{R}_{\mathbf{x}d} = E(\mathbf{x}\mathbf{d}^T) \tag{4.53}$$

$$\mathbf{R}_{\mathbf{d}} = E(\mathbf{d}\mathbf{d}^T) \tag{4.54}$$

The best linear estimator for $\mathbf{x}$ is then given as

$$\hat{\mathbf{x}} = \mathbf{R}_{\mathbf{x}d}\mathbf{R}_{\mathbf{d}}^{-1}\mathbf{d} \tag{4.55}$$

The corresponding covariance matrix for the error

$$\mathbf{e} = \hat{\mathbf{x}} - \mathbf{x} \tag{4.56}$$

is given by

$$\mathbf{R}_{\mathbf{e}} = \mathbf{R}_{\mathbf{x}} - \mathbf{R}_{\mathbf{x}d}\mathbf{R}_{\mathbf{d}}^{-1}\mathbf{R}_{\mathbf{x}d}^T \tag{4.57}$$

Note that if $\mathbf{R}_{\mathbf{d}}$ and $\mathbf{R}_{\mathbf{x}d}$ are known, e.g. through statistical ananlyses of external data, a retrieval is possible even without knowledge of the physical model $\mathbf{A}$.

However, if the physical model is known (in our case the weighting functions) and if we write our retrieval problem as

$$\mathbf{A}\mathbf{x} + \mathbf{n} = \mathbf{d} \tag{4.58}$$

($\mathbf{n}$ the instrument noise) it follows that

$$\hat{\mathbf{x}} = \mathbf{R}_{\mathbf{x}}\mathbf{A}^T(\mathbf{A}\mathbf{R}_{\mathbf{x}}\mathbf{A}^T + \mathbf{R}_{\mathbf{n}})^{-1}\mathbf{d} \tag{4.59}$$

because if the noise $\mathbf{n}$ is uncorrelated

$$\mathbf{R}_{\mathbf{x}n} = E(\mathbf{x}\mathbf{n}^T) = 0 \tag{4.60}$$

$$\mathbf{R}_{\mathbf{d}n} = E(\mathbf{d}\mathbf{n}^T) = 0 \tag{4.61}$$

$$E(\mathbf{n}) = 0 \tag{4.62}$$

$$\mathbf{R}_{\mathbf{n}} = E(\mathbf{n}\mathbf{n}^T) \neq 0 \tag{4.63}$$

$$\begin{aligned}\mathbf{R}_{\mathbf{x}d} &= E(\mathbf{x}\mathbf{d}^T) &(4.64)\\ &= E(\mathbf{x}(\mathbf{A}\mathbf{x}-\mathbf{n})^T) &(4.65)\\ &= E(\mathbf{x}\mathbf{x}^T\mathbf{A}^T-\mathbf{x}\mathbf{n}^T) &(4.66)\\ &= E(\mathbf{x}\mathbf{x}^T)\mathbf{A}^T-E(\mathbf{x}\mathbf{n}^T) &(4.67)\\ &= \mathbf{R}_{\mathbf{x}}\mathbf{A}^T &(4.68)\end{aligned}$$

$$\begin{aligned}\mathbf{R}_{\mathbf{d}} &= E((\mathbf{A}\mathbf{x}-\mathbf{n})(\mathbf{A}\mathbf{x}-\mathbf{n})^T) &(4.69)\\ &= E(\mathbf{A}\mathbf{x}\mathbf{x}^T\mathbf{A}^T-\mathbf{A}\mathbf{x}\mathbf{n}^T-\mathbf{n}\mathbf{x}^T\mathbf{A}^T+\mathbf{n}\mathbf{n}^T) &(4.70)\\ &= \mathbf{A}\mathbf{R}_{\mathbf{x}}\mathbf{A}^T+\mathbf{R}_{\mathbf{n}} &(4.71)\end{aligned}$$

and thus

$$\hat{\mathbf{x}} = \mathbf{R}_{\mathbf{x}}\mathbf{A}^T(\mathbf{A}\mathbf{R}_{\mathbf{x}}\mathbf{A}^T+\mathbf{R}_{\mathbf{n}})^{-1}\mathbf{d} \tag{4.72}$$

as stated above.

More generally when

$$E(\mathbf{x}) = \mathbf{x}_0 \tag{4.73}$$

and

$$\begin{aligned}E(\mathbf{d}) &= E(\mathbf{A}\mathbf{x}+\mathbf{n}) &(4.74)\\ &= E(\mathbf{A}\mathbf{x})+E(\mathbf{n}) &(4.75)\\ &= \mathbf{A}\mathbf{x}_0 &(4.76)\end{aligned}$$

the optimal estimation solution can be written as

$$\hat{\mathbf{x}} = \mathbf{R}_{\mathbf{x}}\mathbf{A}^T(\mathbf{A}\mathbf{R}_{\mathbf{x}}\mathbf{A}^T+\mathbf{R}_{\mathbf{n}})^{-1}(\mathbf{d}-\mathbf{A}\mathbf{x}_0)+\mathbf{x}_0 \tag{4.77}$$

Note that for the special case that the covariance matrices $\mathbf{R}_{\mathbf{x}}$ and $\mathbf{R}_{\mathbf{n}}$ are diagonal the optimal estimation solution takes the same form as the Tikhonov-Phillips regularization, with the paramter γ given by the ratio between $\mathbf{R}_{\mathbf{x}}$ and $\mathbf{R}_{\mathbf{n}}$ (‘signal to noise ratio).

4.7 Averaging kernels

The linear retrieval methods we have discussed (truncated singular values, Tikhonov-Phillis, Optimal Estimation) all can be written in the form

$$\hat{\mathbf{x}} = \mathbf{x}_0 + \mathbf{G}(\mathbf{d} - \mathbf{A}\mathbf{x}_0) \tag{4.78}$$

Insert $\mathbf{A}\mathbf{x} = \mathbf{d}$:

$$\hat{\mathbf{x}} = \mathbf{x}_0 + \mathbf{G}\mathbf{A}(\mathbf{x} - \mathbf{x}_0) \tag{4.79}$$

The matrix $\mathbf{GA}$ is the so called averaging kernel matrix and relates the true, but unknown profile $\mathbf{x}$ to the retrieval solution $\hat{\mathbf{x}}$. Rearranging the equation above gives

$$\hat{\mathbf{x}} = \mathbf{G}\mathbf{A}\mathbf{x} + (\mathbf{I} - \mathbf{G}\mathbf{A})\mathbf{x}_0 \tag{4.80}$$

This tells us, how the retrieved profile $\hat{\mathbf{x}}$ depends on the true but unknown profile $\mathbf{x}$ and the a priori (or climatological) profile $\mathbf{x}_0$. The practical interpretation is, that the retrieval of the remote sensing observations is a smoothed version of the real profile plus some contribution of the a priori profile. Inspection of the averaging kernel matrix gives information on the vertical resolution of the retrieval.