

Inverse Theory Week 4 part 2: Profile retrievals - the Maximum A-posteriori Probability (MAP) Solution

Hugh C. Pumphrey

November 5, 2008

1 Introduction.

In the previous two lectures we considered the nadir sounding of a temperature profile and discovered that it was harder to do than we might have thought. In this lecture, we will find out how we can make a useful estimate of a profile. We will step back a little from the temperature example and consider a more general case, applying our results to the temperature case as an example. We will call the profile we want to measure \mathbf{x} - this is a list of n numbers at reasonably closely-spaced altitudes, perhaps 1 km apart. We will let the list of m numbers that we can measure be \mathbf{y} - in the nadir sounding of temperature example this would be the measurements of upwelling radiation at m different frequencies. We will suppose that we understand the radiative transfer so that given \mathbf{x} we can calculate \mathbf{y} . We will hide the details of this by writing

$$\mathbf{y} = F(\mathbf{x}) \quad (1)$$

where the function F is called the *forward model*. In many cases we can use a linear approximation:

$$\mathbf{y} - \mathbf{y}_L = \mathbf{K}(\mathbf{x} - \mathbf{x}_L) \quad (2)$$

where $\mathbf{y}_L = F(\mathbf{x}_L)$ with \mathbf{x}_L being a typical profile about which we have linearised. Equation 2 looks like a set of simultaneous equations but it has n unknowns and only m equations. We saw in the previous lectures that making \mathbf{x} have only m elements does not lead to satisfactory solutions. Building an instrument with n channels doesn't tend to help either, as the influence functions would overlap each other a great deal. Also, it would be impractical and expensive.

2 The Bayesian View.

Let's look at the problem from a different perspective. We can look at \mathbf{x} and \mathbf{y} as vector random variables. We describe a vector random variable by its Probability Density Function (PDF). Before we make the measurement, \mathbf{x} has a PDF which we will write as $P(\mathbf{x})$ and \mathbf{y} has one which we write as $P(\mathbf{y})$. Once we have made the measurement we have a specific value of \mathbf{y} and the PDF of \mathbf{x} is now $P(\mathbf{x}|\mathbf{y})$. Now Bayes' theorem tells us that :

$$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y})}.$$

In theory, we can calculate $P(\mathbf{y}|\mathbf{x})$ using the forward model and what we know about noise in our instrument. The formula not much use as it stands - we don't really want a PDF, we want an estimate of the profile. One obvious thing to do is try to find the profile $\hat{\mathbf{x}}$ for which $P(\mathbf{x}|\mathbf{y})$ has the largest possible value. This is the most probable value for \mathbf{x} ; a method which finds such a value is called a Maximum A-posteriori Probability (MAP) method.¹ To make useful progress, we have to assume some form for the various PDFs. The usual assumption is that they are all Normal distributions or Gaussians - for a random vector \mathbf{v} the Normal distribution takes the form:

$$P(\mathbf{v}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{S}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{v} - \bar{\mathbf{v}})^T \mathbf{S}^{-1}(\mathbf{v} - \bar{\mathbf{v}})\right)$$

where $\bar{\mathbf{v}}$ is the mean value of \mathbf{v} and \mathbf{S} is its covariance matrix. By taking the natural logarithm of this equation

¹Older text books call this a Maximum Likelihood method. Clive Rodgers, author of *Inverse Methods for Atmospheric Sounding: Theory and Practise*, claims that this usage is incorrect and Maximum A-Posteriori Probability is the correct term.

we can express $P(\mathbf{v})$ like this:

$$-2 \ln P(\mathbf{v}) = (\mathbf{v} - \bar{\mathbf{v}})^T \mathbf{S}^{-1} (\mathbf{v} - \bar{\mathbf{v}}) + c_1 \quad (3)$$

where c_1 is a constant which doesn't depend on \mathbf{v} . Note that at $\mathbf{v} = \bar{\mathbf{v}}$, $P(\mathbf{v})$ has a maximum and $-2 \ln P(\mathbf{v})$ has a minimum. We now proceed to write the various terms in Bayes' theorem in this form. $P(\mathbf{x})$ is given by:

$$-2 \ln P(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_a)^T \mathbf{S}_a^{-1} (\mathbf{x} - \mathbf{x}_a) + c_2 \quad (4)$$

where \mathbf{x}_a is the *a priori* profile, our best estimate of the profile before we make the measurement. $P(\mathbf{y}|\mathbf{x})$ is given by

$$-2 \ln P(\mathbf{y}|\mathbf{x}) = (\mathbf{y} - F(\mathbf{x}))^T \mathbf{S}_y^{-1} (\mathbf{y} - F(\mathbf{x})) + c_3 \quad (5)$$

in which the most likely value of \mathbf{y} for a given value of \mathbf{x} is the one you get by applying the forward model to \mathbf{x} . The spread of \mathbf{y} about this value is given by the experimental errors which have a covariance matrix \mathbf{S}_y . From Bayes' theorem we can see that

$$-2 \ln P(\mathbf{x}|\mathbf{y}) + c_4 = \frac{(\mathbf{x} - \mathbf{x}_a)^T \mathbf{S}_a^{-1} (\mathbf{x} - \mathbf{x}_a) + (\mathbf{y} - F(\mathbf{x}))^T \mathbf{S}_y^{-1} (\mathbf{y} - F(\mathbf{x}))}{(6)}$$

where c_4 has swallowed up c_2 , c_3 and $P(\mathbf{y})$. To find our MAP solution, we just have to find whichever \mathbf{x} makes the right-hand side of Equation 6 as small as possible². Note that this is good intuitively. A term of the form $\mathbf{z}^T \mathbf{M} \mathbf{z}$ like the right-hand side of Equation 4 is called a quadratic form. If the matrix in the middle is symmetric and positive definite (as all good covariance matrices are) then the lowest value a quadratic form can take on is zero and that only happens when \mathbf{z} is a vector of zeros. By minimising Equation 6 we are trying to achieve a balance between a solution which is like the *a priori* and one which agrees exactly with the measurements. Note that if the problem is under-constrained there are many possible \mathbf{x} s which would make the second term in Equation 6 be zero but only $\mathbf{x} = \mathbf{x}_a$ will make the first term be zero. Note also that the solution which minimises the cost function is not usually one that makes either term zero. That means that, unlike the other solutions we have tried³, it is not an exact solution — $F(\hat{\mathbf{x}}) \neq \mathbf{y}$. There is no reason that we should demand an exact solution with $F(\hat{\mathbf{x}}) - \mathbf{y} = 0$, all we

²A function like this which we minimise is often referred to as a “cost function” in the remote sensing literature.

³With the obvious exception of the least-squares solution when $m > n$.

can reasonably ask is that $F(\hat{\mathbf{x}}) - \mathbf{y}$ is of a size consistent with the experimental errors.

The sum of two quadratic forms is another quadratic form so we can write $P(\mathbf{x}|\mathbf{y})$ as

$$-2 \ln P(\mathbf{x}|\mathbf{y}) = (\mathbf{x} - \hat{\mathbf{x}})^T \hat{\mathbf{S}}^{-1} (\mathbf{x} - \hat{\mathbf{x}}) + c_4 \quad (7)$$

where $\hat{\mathbf{x}}$ is the MAP profile we are looking for and $\hat{\mathbf{S}}$ is its covariance matrix. In the case where the forward model F is linear we can find a formula for $\hat{\mathbf{x}}$. To make the equations look simpler we'll assume that \mathbf{x}_0 and \mathbf{y}_0 are both zero vectors so that Equation 2 becomes simply $\mathbf{y} = \mathbf{K}\mathbf{x}$.⁴ Now Equation 6 becomes

$$(\mathbf{x} - \hat{\mathbf{x}})^T \hat{\mathbf{S}}^{-1} (\mathbf{x} - \hat{\mathbf{x}}) + c_4 = \frac{(\mathbf{x} - \mathbf{x}_a)^T \mathbf{S}_a^{-1} (\mathbf{x} - \mathbf{x}_a) + (\mathbf{y} - \mathbf{K}\mathbf{x})^T \mathbf{S}_y^{-1} (\mathbf{y} - \mathbf{K}\mathbf{x})}{(8)}$$

If we multiply out the right-hand sides of 8 and 7 and equate the quadratic terms in \mathbf{x} , we get:

$$\mathbf{x}^T \hat{\mathbf{S}}^{-1} \mathbf{x} = \mathbf{x}^T \mathbf{S}_a^{-1} \mathbf{x} + \mathbf{x}^T \mathbf{K}^T \mathbf{S}_y^{-1} \mathbf{K} \mathbf{x}$$

and hence

$$\hat{\mathbf{S}}^{-1} = \mathbf{S}_a^{-1} + \mathbf{K}^T \mathbf{S}_y^{-1} \mathbf{K}. \quad (9)$$

We next equate the terms linear in \mathbf{x}^T to give :

$$\mathbf{x}^T \hat{\mathbf{S}}^{-1} \hat{\mathbf{x}} = \mathbf{x}^T \mathbf{S}_a^{-1} \mathbf{x}_a + (\mathbf{K}\mathbf{x})^T \mathbf{S}_y \mathbf{y}$$

which on cancelling \mathbf{x}^T and substituting for $\hat{\mathbf{S}}$ from Equation 9 gives:

$$(\mathbf{S}_a^{-1} + \mathbf{K}^T \mathbf{S}_y^{-1} \mathbf{K}) \hat{\mathbf{x}} = \mathbf{S}_a^{-1} \mathbf{x}_a + \mathbf{K}^T \mathbf{S}_y \mathbf{y}$$

and hence

$$\hat{\mathbf{x}} = (\mathbf{S}_a^{-1} + \mathbf{K}^T \mathbf{S}_y^{-1} \mathbf{K})^{-1} (\mathbf{S}_a^{-1} \mathbf{x}_a + \mathbf{K}^T \mathbf{S}_y \mathbf{y}). \quad (10)$$

and that's it. We now have, in equations 9 and 10 a straightforward recipe into which we can plug our measurements \mathbf{y} , our *a priori* profile \mathbf{x}_a and their covariance matrices and obtain our Maximum A-posteriori Probability profile $\hat{\mathbf{x}}$ and its covariance matrix. Equation 10 is similar in form to the equation we saw in the second introductory lecture of this term for combining two estimates of a random variable. We can make this more obvious as follows. If the problem is under-constrained we can find lots of exact solutions — lots of values of \mathbf{x} for which $F(\mathbf{x})$ equals our measurements \mathbf{y} exactly. We just need to find

⁴If they are not, you can always choose a new \mathbf{x} and \mathbf{y} equal to $\mathbf{x} - \mathbf{x}_0$ and $\mathbf{y} - \mathbf{y}_0$ respectively.

a matrix \mathbf{D}_e for which $\mathbf{K}\mathbf{D}_e = \mathbf{I}$ and hence $\mathbf{y} = \mathbf{K}(\mathbf{D}_e\mathbf{y})$. ($\mathbf{D}_e = \mathbf{K}^T(\mathbf{K}\mathbf{K}^T)^{-1}$ is just one possibility). Now we can rewrite Equation 10 as

$$\hat{\mathbf{x}} = (\mathbf{S}_a^{-1} + \mathbf{K}^T\mathbf{S}_y^{-1}\mathbf{K})^{-1}(\mathbf{S}_a^{-1}\mathbf{x}_a + \mathbf{K}^T\mathbf{S}_y^{-1}\mathbf{K}(\mathbf{D}_e\mathbf{y}))$$

and we can see that we are combining two estimates of \mathbf{x} : one is the *a priori* \mathbf{x}_a with covariance \mathbf{S}_a and the other is an exact solution $\mathbf{D}_e\mathbf{y}$ with a covariance matrix whose inverse is given by $\mathbf{K}^T\mathbf{S}_y^{-1}\mathbf{K}$. For an under-constrained problem this matrix will always be singular - this tells us that there are some features of the exact solution $\mathbf{D}_e\mathbf{y}$ which have infinite variance - we do not know them at all. In the MAP solution $\hat{\mathbf{x}}$ all information about these features comes from the *a priori*. We know no more about them than we did before we made the measurement. Equation 10 can be re-written in several forms of which the most useful are:

$$\hat{\mathbf{x}} = \mathbf{x}_a + (\mathbf{S}_a^{-1} + \mathbf{K}^T\mathbf{S}_y^{-1}\mathbf{K})^{-1}\mathbf{K}^T\mathbf{S}_y^{-1}(\mathbf{y} - \mathbf{K}\mathbf{x}_a) \quad (11)$$

and

$$\hat{\mathbf{x}} = \mathbf{x}_a + \mathbf{S}_a\mathbf{K}^T(\mathbf{K}\mathbf{S}_a\mathbf{K}^T + \mathbf{S}_y)^{-1}(\mathbf{y} - \mathbf{K}\mathbf{x}_a) \quad (12)$$

Equation 12 is known as the *m*-form of the equation because the matrix to be inverted is an $m \times m$ matrix - for under-constrained problems this will be easier to do than inverting the $n \times n$ matrices in Equation 11 (which is known as the *n*-form of the equation).

3 An Example of a MAP Solution.

Now let's see how it works. I use the same imaginary microwave nadir sounder as in the previous lecture. We use the same "true" profile \mathbf{x}_t and, as before, we generate some radiances \mathbf{y} using the equation $\mathbf{y} = \mathbf{K}\mathbf{x}_t + \varepsilon$. Here, ε are the experimental errors; a random vector with zero mean and covariance matrix \mathbf{S}_y . We assume that \mathbf{S}_y is diagonal i.e. there is no correlation between the channels. Instead of using a polynomial (or other) representation as we did in the previous lecture, we apply Equation 12. To do this, we need an *a priori* profile \mathbf{x}_a and its covariance matrix \mathbf{S}_a . For simplicity we'll assume that \mathbf{x}_a is an isothermal profile with a temperature of 250 K \pm 50K at all altitudes - \mathbf{S}_a therefore has diagonal elements $S_{ii} = (50 \text{ K})^2$. The off-diagonal elements of \mathbf{S}_a are not zero, we let them be

$$S_{ij} = \sqrt{S_{ii}S_{jj}} \exp\left(\frac{-(z_i - z_j)^2}{z_s^2}\right)$$

where z_i is the altitude of the *i*th element of the profile and z_s is a "smoothing length" - we use 0.2 in log pressure units - about 3km. By doing this we are effectively saying that we know in advance that bits of the profile within 3 km of each other are highly correlated and bits further apart than this are not. We can't apply the formula if we assume there is no noise on the measurements, but we can assume that it is very small, 0.0001K, say. In this case Equation 12 gives the result shown in Figure 1. Now

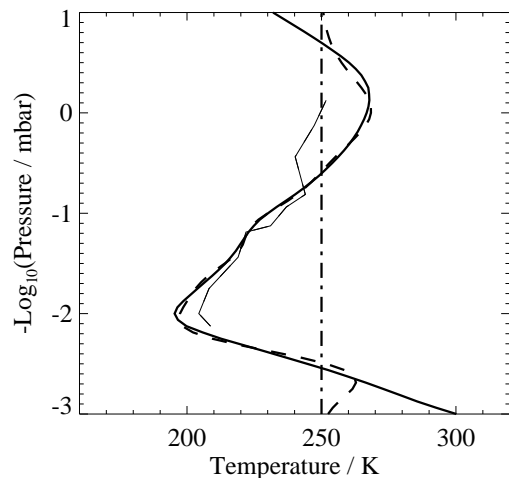


Figure 1: Maximum A-posteriori Probability (MAP) retrieval for an idealised 11-channel microwave nadir sounder. Thick solid line: true profile, thick dashed line: retrieved profile, thick dot-dash line: *a priori*. Note how the profile reverts to the *a priori* at the top and bottom of the profile where the measurements provide no information. The thin lines are the measurements, plotted at the peaks of the weighting functions.

we try adding 1 K of random noise to our measurements, the result is shown in Figure 2. The retrieved profile is worse, now, but we must remember that when we tried using polynomials as representation functions, the profile had oscillations of several hundred Kelvin. This is clearly an improvement. Another way to see this is to look at the contribution functions which are shown in Figure 3. The contribution function matrix \mathbf{D} for any retrieval method is given by $\mathbf{D} = \frac{d\hat{\mathbf{x}}}{d\mathbf{y}}$. It is essentially what you have to multiply the radiances by to get the solution. For the Maximum A-posteriori Probability solution, \mathbf{D} can be ob-

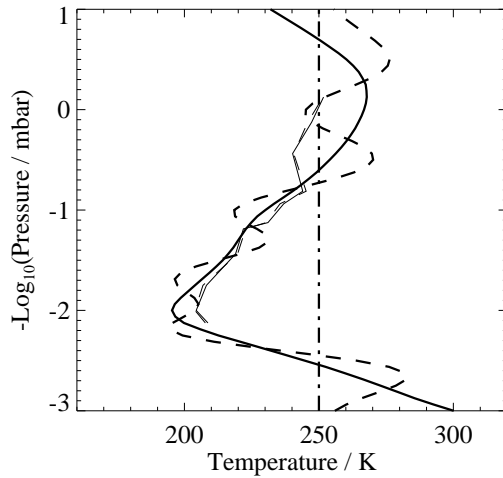


Figure 2: As Figure 1 but with 1 K noise on the measurements.

tained by re-writing Equation 12 as

$$\hat{\mathbf{x}} = \mathbf{x}_a + \mathbf{D}(\mathbf{y} - \mathbf{K}\mathbf{x}_a)$$

so that \mathbf{D} is given by $\mathbf{D} = \mathbf{S}_a \mathbf{K}^T (\mathbf{K} \mathbf{S}_a \mathbf{K}^T + \mathbf{S}_y)^{-1}$. The contribution functions for the maximum A-posteriori Probability solution are much better behaved than the ones for the exact solution using a polynomial representation. They are even somewhat better than the ones for the representation function solution using the influence functions as a representation. More importantly, we achieved this without having to make an arbitrary choice of a representation basis.

4 Summary.

We have in the MAP solution a formula with the following advantages:

1. It works.
2. It has a sound statistical foundation.
3. It makes intuitive sense as a weighted combination of the a priori with an exact solution.
4. You don't have to decide in advance on a suitable representation.

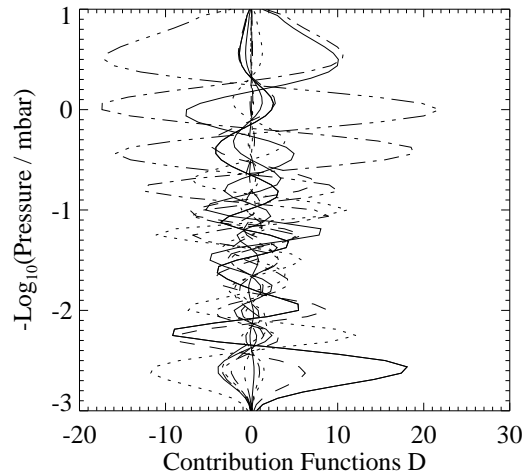


Figure 3: Contribution functions for the Maximum A-posteriori Probability Retrieval. These should be compared with the ones shown in the previous lecture for the various representation function solutions. With the Maximum A-posteriori Probability solution, we no longer get huge values for the contribution functions at the top and bottom of the profile.

5. You get an estimate of the error along with the retrieved profile.

We have the following caveat:

- You need an a priori profile and its covariance matrix before you can use the formula.

We will have to quantify exactly how good our retrieved profile is. This can be done by making a careful analysis of $\hat{\mathbf{S}}$, the covariance matrix of the retrieved profile. This is the subject of the next lecture.