

Inverse Theory week 6 and 7: *A priori* information

How to get it and how to avoid it.

Hugh C. Pumphrey

November 5, 2008

1 The *a priori* profile: where can you get one?

We have seen in the previous lectures that the Maximum A-posteriori Probability (MAP) solution is a good way of estimating atmospheric profiles from remotely sensed data. The formula is essentially a combination of two different estimates of the profile: an exact solution (which contains all the information in the measurements) and an *a priori* (which specifies what we knew before the measurements were made). It is therefore an obvious “feature” of the Maximum A-posteriori Probability solution that you need to have an *a priori* before you use it. In this section, we will look at where you might get an *a priori* profile from.

1.1 Guess.

If you know nothing about your profile before you make the measurements, you might have to do this. This is what I did for the nadir temperature-sounder example I have been using for this course. One should guess sensibly. We know that the temperature at the surface is about 300 K. We know it gets colder as you go up. It seems unlikely that the temperature will be less than 170 K or greater than 330 K so I chose a constant value of 250 K with a standard deviation of 80 K at all heights.

We will look later at a formula called the Twomey-Tikhonov formula which enables you to do without *a priori*. If you really know nothing, this might be a better idea than using the Maximum A-posteriori Probability solution with a badly guessed *a priori*.

1.2 Use a Climatology.

This is a name used for the average value of the profile over the last few years. Depending on what your profile

is a profile of, you may find climatologies of various complexities. At the simplest level, you might use a global mean of all the profiles you can find, as measured by various other instruments. A more sophisticated climatology might have different profiles for different seasons and latitudes.

1.3 Use a forecast.

You might find that a meteorological service provides forecasts of the quantity that you are trying to measure.

1.4 Use a profile you retrieved earlier - a Kalman Filter.

This one is more subtle, and worth looking at in some more detail. Suppose that your temperature sounder measures a profile which is typical for a heat wave. You would not expect the next profile along to be typical of a hard freeze. It therefore makes some sense to use the retrieved profile from one scan as the *a priori* for the next. For the *a priori* covariance, we need to estimate how quickly the atmosphere is likely to change from one pixel to the next. Let's use the subscript i to indicate which profile along the measurement track we are talking about. Our model of the situation is therefore like this:

$$\mathbf{y}_i = \mathbf{K}_i \mathbf{x}_i + \varepsilon_i \quad (1)$$

$$\mathbf{x}_i = \mathbf{M}_i \mathbf{x}_{i-1} + \mathbf{z}_i \quad (2)$$

Equation 1 is the linearised forward model, relating the i th set of measurements \mathbf{y}_i to the i th profile \mathbf{x}_i ; ε is the measurement error with covariance matrix \mathbf{S}_y . Equation 2 is the equation used to predict the i th profile from the $i - 1$ th profile. In the simple case where you predict that the next profile will be just like this one,

\mathbf{M} is simply a unit matrix. One can, however, imagine more sophisticated predictors which might perhaps vary with latitude. The prediction error, \mathbf{z} , has a covariance matrix \mathbf{S}_z . We generate an *a priori* for the i th profile thus:

$$\mathbf{x}_{ai} = \mathbf{M}_i \hat{\mathbf{x}}_{i-1}$$

$$\mathbf{S}_{ai} = \mathbf{M}_i \hat{\mathbf{S}}_{i-1} \mathbf{M}_i^T + \mathbf{S}_{zi}.$$

We then use these in the usual Maximum A-posteriori Probability formula (which I have written in the m -form) to estimate the i th profile:

$$\mathbf{D}_i = \mathbf{S}_{ai} \mathbf{K}_i^T (\mathbf{K}_i \mathbf{S}_{ai} \mathbf{K}_i^T + \mathbf{S}_y)^{-1}$$

$$\hat{\mathbf{x}}_i = \mathbf{x}_{ai} + \mathbf{D}_i (\mathbf{y}_i - \mathbf{K}_i \mathbf{x}_{ai})$$

$$\hat{\mathbf{S}}_i = \mathbf{S}_{ai} - \mathbf{D}_i \mathbf{K}_i \mathbf{S}_{ai}$$

This approach is called a Kalman Filter and is found in many other areas of engineering. Of course, we need to provide an *a priori* for the first profile to set the Kalman filter going, but this can be a very rough guess with large error bars. Once the filter has run past a few profiles, \mathbf{x}_{ai} and \mathbf{S}_{ai} will settle down to sensible values, provided that you have made sensible choices for \mathbf{S}_z and \mathbf{M} .

1.5 Direct Regression

We have considered what we might do if we have limited *a priori* information. Now we consider what we could do if we had no forward model, that is, no matrix \mathbf{K} . Imagine that we had an instrument which we knew was sensitive to the temperature but that we had forgotten to calibrate it properly or found that the radiative transfer modelling was far too difficult to do accurately. Now suppose, however, that in a very limited sub-set of the places where we make satellite measurements, we also have measurements from another source: weather balloons, perhaps. We can do without a physical forward model now, because we can compare the satellite and balloon measurements where we have both to establish a relationship between the radiances \mathbf{y} and the profile \mathbf{x} . This approach is similar in essence to supervised classification, because we are using a “training set” of profiles where we know the answer in order to estimate what the atmosphere is like in places where we have the remote sensing data only.

Suppose we have a reasonably large ensemble of N cases where we have a set of radiances \mathbf{y}_i and a profile

\mathbf{x}_i . As usual, we let the mean value of all the sets of radiances be defined by

$$\bar{\mathbf{y}} = \frac{1}{N} \sum_{j=0}^N \mathbf{y}_j$$

and the mean of all the measured profiles be

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{j=0}^N \mathbf{x}_j.$$

We will let our estimate of a profile where we don't have measurements be of the form

$$\hat{\mathbf{x}} = \bar{\mathbf{x}} + \mathbf{D}(\mathbf{y} - \bar{\mathbf{y}}). \quad (3)$$

We want to find \mathbf{D} such that if we apply Equation 3 to the cases where we know the answer, then, on average, the mean distance between our estimates and the true profiles will be as small as possible. This in turn means that we want the quantity:

$$C = \frac{1}{N} \sum_{j=1}^N (\mathbf{x}_j - \hat{\mathbf{x}}_j)^T (\mathbf{x}_j - \hat{\mathbf{x}}_j) \quad (4)$$

to be as small as possible. We substitute Equation 3 into Equation 4 to give

$$C = \frac{1}{N} \sum_{j=1}^N (\mathbf{x}_j - \bar{\mathbf{x}} - \mathbf{D}(\mathbf{y}_j - \bar{\mathbf{y}}))^T (\mathbf{x}_j - \bar{\mathbf{x}} - \mathbf{D}(\mathbf{y}_j - \bar{\mathbf{y}})) \quad (5)$$

and then look for the matrix \mathbf{D} minimises C . This means that we will have to differentiate the scalar C with respect to a matrix \mathbf{D} . The result will be a matrix of the same size as \mathbf{D} with elements $(\frac{dC}{d\mathbf{D}})_{ij} = \frac{\partial C}{\partial D_{ij}}$ which we will set equal to a matrix of zeros, \mathbf{O} . We will need two results which you can prove using suffix notation: $\frac{d}{d\mathbf{A}}(\mathbf{w}^T \mathbf{A} \mathbf{z}) = \frac{d}{d\mathbf{A}}(\mathbf{z}^T \mathbf{A}^T \mathbf{w}) = \mathbf{w} \mathbf{z}^T$ and $\frac{d}{d\mathbf{A}}(\mathbf{z}^T \mathbf{A}^T \mathbf{A} \mathbf{z}) = 2\mathbf{A} \mathbf{z} \mathbf{z}^T$. Letting $\mathbf{x}_j - \bar{\mathbf{x}} = \mathbf{w}$ and $\mathbf{y}_j - \bar{\mathbf{y}} = \mathbf{z}$ we can write Equation 5 as:

$$C = \frac{1}{N} \sum_{j=1}^N (\mathbf{w} - \mathbf{D} \mathbf{z})^T (\mathbf{w} - \mathbf{D} \mathbf{z}). \quad (6)$$

Expanding the brackets gives:

$$C = \frac{1}{N} \sum_{j=1}^N \mathbf{w}^T \mathbf{w} - \mathbf{w}^T \mathbf{D} \mathbf{z} - \mathbf{z}^T \mathbf{D}^T \mathbf{w} + \mathbf{z}^T \mathbf{D}^T \mathbf{D} \mathbf{z}$$

and then differentiating and setting equal to \mathbf{O} gives:

$$\mathbf{O} = \frac{2}{N} \sum_{j=1}^N \mathbf{D} \mathbf{z} \mathbf{z}^T - \mathbf{w} \mathbf{z}^T.$$

and hence

$$\sum_{j=1}^N \mathbf{D} \mathbf{z} \mathbf{z}^T = \sum_{j=1}^N \mathbf{w} \mathbf{z}^T.$$

Now, \mathbf{D} is a constant as far as the sums go, so we can take it out of the bracket and obtain

$$\mathbf{D} = \left[\sum_{j=1}^N \mathbf{w} \mathbf{z}^T \right] \left[\sum_{j=1}^N \mathbf{z} \mathbf{z}^T \right]^{-1}.$$

Substituting in for \mathbf{z} and \mathbf{w} gives:

$$\mathbf{D} = \left[\sum_{j=1}^N (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{y}_j - \bar{\mathbf{y}})^T \right] \left[\sum_{j=1}^N (\mathbf{y}_j - \bar{\mathbf{y}})(\mathbf{y}_j - \bar{\mathbf{y}})^T \right]^{-1} \quad (7)$$

and that is our answer. We can take the list of profiles measured by the balloons, \mathbf{x}_j , the coincident sets of satellite measurements, \mathbf{y}_j , put them into Equation 7 and calculate a set of contribution functions \mathbf{D} which we can then use in Equation 3 to obtain profiles for the places where we have satellite measurements but no balloon measurements. So, if we have a sufficient “training set” we can do without a forward model.

It is instructive to see what happens if we imagine using this method in a case where we do have a forward model. We take a representative set of profiles \mathbf{x}_i but instead of using real measurements from some other instrument, we assume that the radiances \mathbf{y}_i would be given by $\mathbf{y}_i = \mathbf{K} \mathbf{x}_i + \varepsilon_i$ where ε_i is a vector of measurement noise with covariance matrix \mathbf{S}_y . The second bracket of Equation 7 now becomes:

$$\begin{aligned} & \sum_{j=1}^N (\mathbf{K} \mathbf{x}_j - \mathbf{K} \bar{\mathbf{x}} + \varepsilon_i)(\mathbf{K} \mathbf{x}_j - \mathbf{K} \bar{\mathbf{x}} + \varepsilon_i)^T \\ &= \mathbf{K} \left[\sum_{j=1}^N (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^T \right] \mathbf{K}^T + \sum_{j=1}^N \varepsilon_i \varepsilon_i^T \end{aligned}$$

where we have used the fact that the mean value of ε is zero and thrown away terms with one power of ε . The quantity in the square brackets is just the covariance matrix of the training set of profiles about their mean which

we will call \mathbf{S}_x and so the whole of the second bracket from Equation 7 is $\mathbf{K} \mathbf{S}_x \mathbf{K}^T + \mathbf{S}_y$. We can expand the first bracket of Equation 7 in a similar way to give $\mathbf{S}_x \mathbf{K}^T$ so that \mathbf{D} is given by

$$\mathbf{D} = \mathbf{S}_x \mathbf{K}^T (\mathbf{K} \mathbf{S}_x \mathbf{K}^T + \mathbf{S}_y)^{-1}.$$

and the retrieval formula, Equation 3, becomes

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

But this is exactly the same as the m -form of the Maximum A-posteriori Probability solution, with the training set used as the *a priori*. What we have learned from this is that a regression against another set of measurements gives us exactly the same formula as we would have got if we had a forward model and used the Maximum A-posteriori Probability formula, with that set of measurements providing the *a priori*.

2 Other retrieval formulae.

The Maximum A-posteriori Probability formula is justifiably popular on account of its statistical rigour, among other reasons. You do need an *a priori* profile and its covariance matrix before you can apply the formula. Many other formulae have been used for retrieval problems and many of these require less in the way of *a priori* information. We look at some of these in this section. Some of these are of historical interest only, others are used for specific types of instrument.

2.1 Over-constrained problems: Weighted Least Squares.

Most nadir-sounding problems are shockingly under-constrained, as we have seen. This is not true of every remote sensing problem, however. Limb sounding instruments (especially solar occultation instruments) often measure radiances at tangent heights less than 1 km apart and may measure at several frequencies for each tangent height. There may therefore be several times *more* measurements than there are elements in the profile. There is nothing in the derivation of the Maximum A-posteriori Probability formula which insists that the problem must be under-constrained. Applying Bayes' theorem to such a problem still leads to the minimisation of the function

$$\begin{aligned} -2 \ln P(\mathbf{x}|\mathbf{y}) + c_4 &= (\mathbf{x} - \mathbf{x}_a)^T \mathbf{S}_a^{-1} (\mathbf{x} - \mathbf{x}_a) \\ &+ (\mathbf{y} - \mathbf{K} \mathbf{x})^T \mathbf{S}_y^{-1} (\mathbf{y} - \mathbf{K} \mathbf{x}) \end{aligned} \quad (8)$$

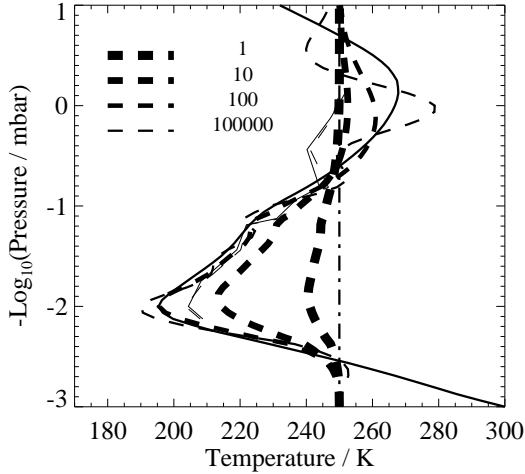


Figure 1: This figure shows the result of applying Equation 13 to the nadir-sounding example we have used before. The dashed lines of various thicknesses are solutions for different values of γ . The dot-dash line is the *a priori*.

2.3 The Backus-Gilbert method.

(Not on syllabus for 2005-6 or after. Notes left here for interest.) We have seen how a lot of solutions are a trade-off of some sort - both the Maximum A-posteriori Probability and Twomey-Tikhonov formulae trade off agreement with the measurements against agreement with the *a priori*. If we have no *a priori* knowledge we might consider that we should admit that we cannot retrieve the profile but that we might have enough information in the measurements alone to get a smoothed version of the profile. We can do this by considering the averaging kernels and trying to find a set of contribution functions which give you averaging kernels with the smallest possible spread. We defined the spread of a function $a(z)$ about a height z_0 as:

$$s(z_0) = 12 \int (z - z_0)^2 a(z)^2 dz \quad (14)$$

where the integrals are taken over the whole range of z which we are interested in. The definition requires that

$$\int a(z) dz = 1 \quad (15)$$

Because the spread is defined in terms of an integral, we will consider the profile \mathbf{x} , the rows of \mathbf{K} and the columns

of \mathbf{D} to be continuous functions of height $x(z)$, $K_i(z)$ and $D_i(z)$. The subscripts i refer to the measurements. We will consider a single height z and try to find the set of i numbers $D_i(z)$ which gives the averaging kernel for that height which has the smallest possible spread. Since we are using continuous functions, the forward model for the i th channel becomes:

$$y_i = \int K_i(z') x(z') dz' + e_i \quad (16)$$

and the retrieval becomes:

$$\hat{x}(z) = \sum_{i=1}^m D_i(z) y_i \quad (17)$$

Inserting Equation 16 into Equation 17 gives

$$\hat{x}(z) = \int A(z, z') x(z') dz' + \sum_{i=1}^m D_i(z) e_i$$

with the averaging kernel for altitude z given by $A(z, z') = \sum_{i=1}^m D_i(z) K_i(z')$. The spread of the averaging kernel for height z is given by:

$$s(z) = 12 \int \left[\sum_{i=1}^m D_i(z) K_i(z') \right]^2 (z - z')^2 dz'$$

which can be re-written as

$$s(z) = \sum_{i=1}^m \sum_{j=1}^m D_i(z) Q_{ij}(z) D_j(z)$$

where the matrix of functions Q depends only on the weighting functions K :

$$Q_{ij}(z) = \int (z - z')^2 K_i(z') K_j(z') dz'$$

We require A to have unit area:

$$1 = \int A(z, z') dz' = \int \sum_{i=1}^m D_i(z) K_i(z') dz'$$

$$= \sum_{i=1}^m k_i D_i(z)$$

where $k_i = \int K_i(z') dz'$. We now need to minimise the spread, with the unit area requirement as a constraint. We do this by the method of Lagrange multipliers:

$$\frac{\partial}{\partial D_p(z)} \left[\sum_{i=1}^m \sum_{j=1}^m D_i(z) Q_{ij}(z) D_j(z) + \lambda(z) \left\{ \sum_{i=1}^m D_i(z) k_i - 1 \right\} \right] = 0 \quad (18)$$

At this point we make a change of notation and let the m values of $D_i(z)$ be an m -element vector \mathbf{d} , the m values k_i be the vector \mathbf{k} and the m^2 numbers Q_{ij} be the matrix \mathbf{Q} . Equation 18 now becomes

$$\frac{\partial}{\partial \mathbf{d}} [\mathbf{d}^T \mathbf{Q} \mathbf{d} + \lambda \{\mathbf{d}^T \mathbf{k} - 1\}] = 0 \quad (19)$$

which we now have to solve. Differentiating gives

$$2\mathbf{Q}\mathbf{d} + \lambda\mathbf{k} = 0 \quad (20)$$

which is m equations in $m+1$ unknowns: the m elements of \mathbf{d} plus λ . The constraint $\mathbf{k}^T \mathbf{d} = \mathbf{d}^T \mathbf{k} = 1$ gives us the one extra equation needed to find a solution. We multiply Equation 20 by \mathbf{Q}^{-1} and re-arrange to give

$$\mathbf{d} = -\frac{\lambda}{2} \mathbf{Q}^{-1} \mathbf{k}. \quad (21)$$

We then multiply on the left by \mathbf{k}^T and use the constraint in the form $\mathbf{k}^T \mathbf{d} = 1$ to give:

$$1 = -\frac{\lambda}{2} \mathbf{k}^T \mathbf{Q}^{-1} \mathbf{k}. \quad (22)$$

Eliminating λ between Equations 21 and 22 gives the final answer: .

$$\mathbf{d} = \frac{\mathbf{Q}^{-1} \mathbf{k}}{\mathbf{k}^T \mathbf{Q}^{-1} \mathbf{k}}. \quad (23)$$

We now have a formula for the m values of the contribution functions at the altitude z . We have to calculate this for each of the altitudes in the profile. We can then make up a contribution function matrix \mathbf{D} whose rows are the \mathbf{d} s we calculated for each altitude. Note that it is only at this point that we discretise the profile and decide how many elements we will use to represent it. We then calculate the retrieved profile using

$$\hat{\mathbf{x}} = \mathbf{D}\mathbf{y}. \quad (24)$$

The retrieval has a noise level given by: $\sigma^2(z) = \mathbf{d}^T \mathbf{S}_y \mathbf{d}$. Since we have not given any thought to making the noise

small it is likely to be large if we apply Equation 23 as it stands. We can fix this by minimising

$$\frac{\partial}{\partial \mathbf{d}} [\mathbf{d}^T \mathbf{Q} \mathbf{d} + \lambda \mathbf{d}^T \mathbf{k} + \mu \mathbf{d}^T \mathbf{S}_y \mathbf{d}] = 0;$$

since this equation is the same as 19 with \mathbf{Q} replaced by $\mathbf{Q} + \mu \mathbf{S}_y$ the solution will be

$$\mathbf{d} = \frac{(\mathbf{Q} + \mu \mathbf{S})^{-1} \mathbf{k}}{\mathbf{k}^T (\mathbf{Q} + \mu \mathbf{S})^{-1} \mathbf{k}}. \quad (25)$$

Here, μ is a tradeoff term between vertical resolution and noise.

How well does this work for our test case? As before, we take a profile \mathbf{x}_t which we regard as true, calculate some measurements from it with $\mathbf{y} = \mathbf{K}\mathbf{x} + \varepsilon$ and then apply Equation 24 to see how good an estimate can be obtained. We first try this with the noise level set to zero. The true and retrieved profiles are shown in Figure 2. Notice how, even though we have set $\mu = 0$

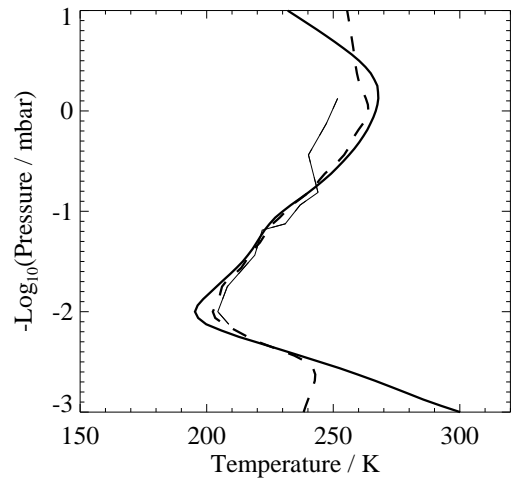


Figure 2: Backus-Gilbert retrieval for our usual test case, with no measurement noise. We have set $\mu = 0$.

the retrieved profile is smoother than the Maximum A-posteriori Probability solution. The averaging kernels are shown in Figure 3. It is useful to compare them to those obtained from the Maximum A-posteriori Probability formula. In Figure 4 we plot the Backus-Gilbert spread of these kernels, along with the Full Width at half height. Note that the B-G spread is actually less than the FWHH, in contrast to the averaging kernels we got with the Maximum A-posteriori Probability method.

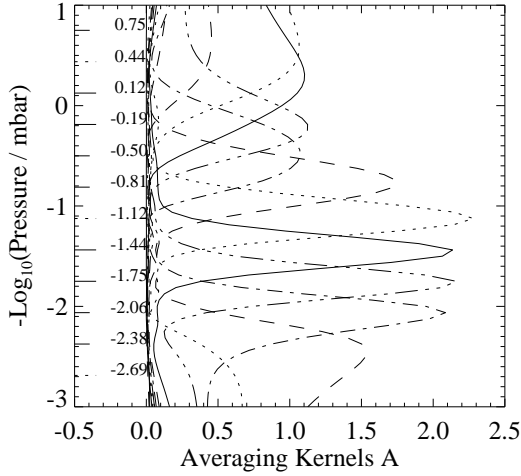


Figure 3: Averaging kernels for a Backus-Gilbert retrieval with $\mu = 0$. Notice how the main peak of each kernel is broader than in the Maximum A-posteriori Probability retrieval but that the wiggles to either side of the main peak are much reduced.

Now let's add 1 K of noise to our measurements and see what happens. Recall that this much noise made the exact solution go haywire and caused errors on the order of 10 K in the maximum A-posteriori Probability solution. The profile is noisier now, as we would expect, but the noise seems of a reasonable size. The contribution functions, shown in Figure 6, confirm that we should expect good noise characteristics from this retrieval. Now that we have a noisy retrieval, we should see what effect the tradeoff parameter μ has. Figure 7 is the same as Figure 5 but with $\mu = 0.03$ instead of $\mu = 0$. Note how the profile is less noisy but more smoothed. In order to find out what is a good choice for μ we consider the averaging kernel for a height in the middle of the range and calculate its spread for various values of μ . The averaging kernel itself is shown in Figure 8. In Figure 9 we plot the B-G spread of this averaging kernel against the noise on the retrieved profile.

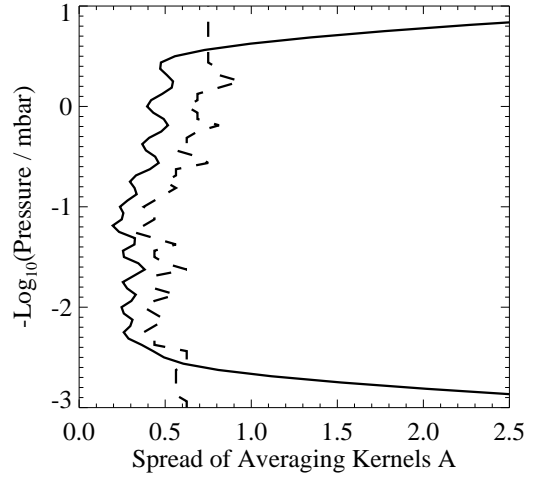


Figure 4: Backus Gilbert spread (solid line) and Full width at half height (dashed line) for the Backus-Gilbert retrieval (with $\mu = 0$).

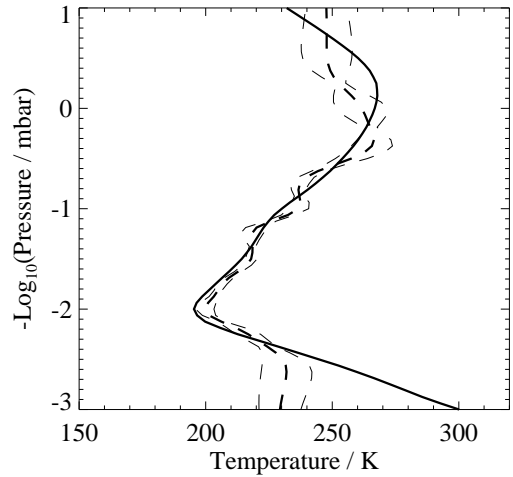


Figure 5: Backus-Gilbert retrieval with 1 K measurement noise. The thin dashed lines are the error in the retrieval due to measurement noise - they don't include smoothing error.

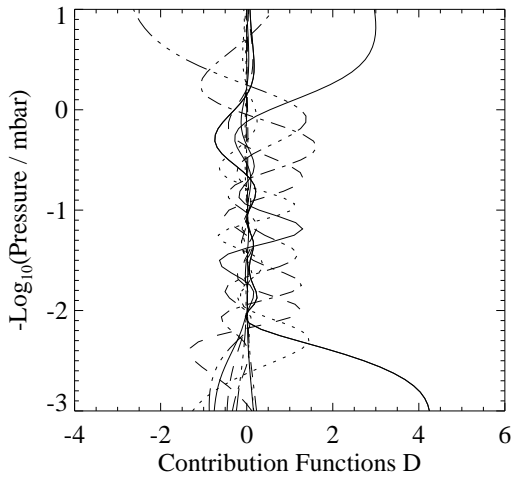


Figure 6: Contribution functions for the Backus-Gilbert retrieval. They are less than 10 everywhere so we should expect less than 10 K of noise in the solution if the measurement noise is 1 K.

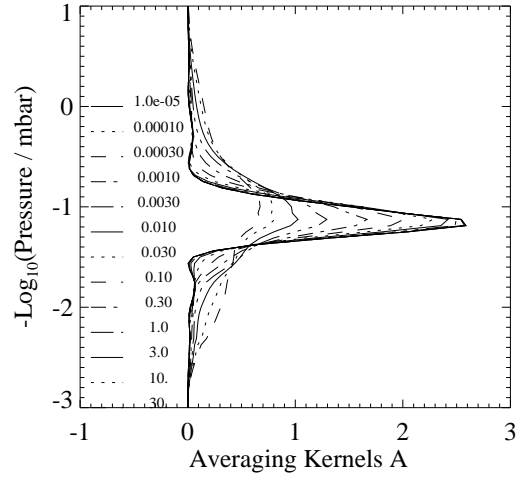


Figure 8: The averaging kernel for one altitude, for various values of μ .

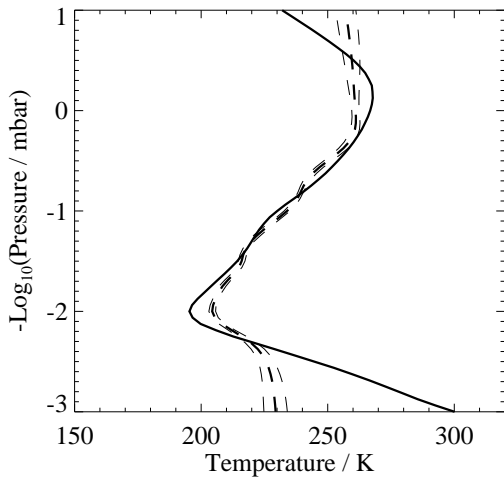


Figure 7: Figure 5 but with $\mu = 0.03$ instead of $\mu = 0$.

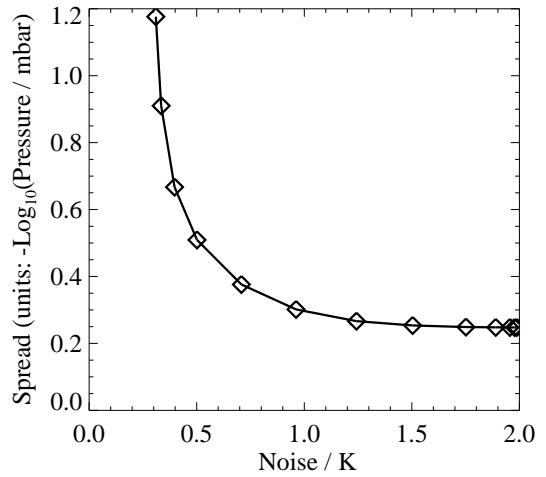


Figure 9: Backus-Gilbert spread plotted against retrieval noise for various values of μ . Clearly, a value of μ which corresponds to a point near the ‘knee’ of the curve will give a good tradeoff.