EPS 237 Lesson 11: Introduction to Retrieval Theory

R. Wordsworth

March 21, 2022

The derivations here follow Rodgers (2000), *Inverse Methods for Atmospheric Sounding*, which can be referred to for more details.

1 Overview

The inverse problem exists whenever you have remote measurements and you want to use them to say something about the internal state of a system. In practice, solving the inverse problem involves setting up a set of simultaneous equations that include error terms, and solving them. To approach this in a general way, we let $\mathbf{y}$ be the measurement/signal vector and $\mathbf{x}$ the state vector. We can treat both $\mathbf{y}$ and $\mathbf{x}$ as discrete arrays of values. In practical terms, $\mathbf{y}$ could be a bunch of radiance measurements, and $\mathbf{x}$ could be temperature or composition as a function of height.

In the ideal (fantasy) situation where our model of the physics is perfect and there are no errors, we can write

$$\mathbf{y}_{ideal} = \mathbf{f}(\mathbf{x}). \tag{1}$$

i.e. the observations are some perfectly known function $\mathbf{f}$ of the state vector. In the real world,

$$\mathbf{y} = \mathbf{F}(\mathbf{x}) + \boldsymbol{\epsilon} \tag{2}$$

where $\boldsymbol{\epsilon}$ encapsulates all the errors, and $\mathbf{F}$ is our (imperfect) forward model of the physics. The fallibility of $\mathbf{F}$ comes from our incomplete knowledge of the state of the atmosphere, inaccurate representation of the physics, and so on.

In general, $\mathbf{F}$ is nonlinear, but to make progress, we can begin by linearizing the model physics around some reference state $\mathbf{x}_0$:

$$\mathbf{y} = \mathbf{F}(\mathbf{x}_0) + \frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{x}} (\mathbf{x} - \mathbf{x}_0) + \boldsymbol{\epsilon} \tag{3}$$

We define $\mathbf{K} = \frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{x}}$. This is the weighting function matrix. It tells us how much the measurement responds to little variations in the state vector. For example, if we varied the temperature at 500 mBar a little bit, how would the radiance at 10 microns vary?

If the number of measurements is the same as the dimension of the state vector, $\mathbf{K}$ can be inverted:

$$\mathbf{y} - \mathbf{F}(\mathbf{x}_0) - \boldsymbol{\epsilon} = \mathbf{K}(\mathbf{x} - \mathbf{x}_0) \tag{4}$$

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{K}^{-1}[\mathbf{y} - \mathbf{F}(\mathbf{x}_0) - \boldsymbol{\epsilon}]. \tag{5}$$

This is the 'exact' solution. In reality, though, it's impractical. The system is very ill-conditioned so big errors result from very small amounts of noise. This is a problem for many remote sensing applications, and for nadir-view temperature sounding in particular. To solve this problem, we must incorporate some prior information to help the retrieval on its way. To do this, we first need to decide what form our uncertainties will take, and we then need to derive a result called Bayes theorem.

2 Representing uncertainty

It's convenient to think in terms of probability density functions (pdfs). For some scalar y, the mean and variance of the distribution are

$$\overline{y} = \int yP(y)dy \tag{6}$$

$$\sigma^2 = \int (y-\overline{y})^2 P(y)dy \tag{7}$$

Assume a gaussian pdf (this often works in reality due to something called the central limit theorem):

$$P(y) = N(y-\overline{y},\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\overline{y})^2}{2\sigma^2}}. \tag{8}$$

Now, for a vector σ becomes the covariance matrix $\mathbf{S}$. The elements of $\mathbf{S}$ are

$$S_{ij} = \int (y_i - \overline{y}_i)(y_j - \overline{y}_j)P(\mathbf{y})d^n\mathbf{y}. \tag{9}$$

Its diagonal is simply σ_i^2. P then becomes

$$P(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{S}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}-\overline{\mathbf{y}})^T\mathbf{S}^{-1}(\mathbf{y}-\overline{\mathbf{y}})}. \tag{10}$$

3 Bayes Theorem

If $P(x,y)$ is a probability distribution that depends on two variables,

$$P(x) \equiv \int P(x,y)dy. \tag{11}$$

Now

$$P(y|x) \equiv \frac{P(x,y)}{\int P(x,y)dy} \tag{12}$$

with the division necessary for normalization of the distribution. Hence

$$P(y|x) \equiv \frac{P(x,y)}{P(x)}. \tag{13}$$

Similarly

$$P(x|y) \equiv \frac{P(x,y)}{P(y)} \tag{14}$$

This leads to Bayes theorem as stated in the slides. The vector version is

$$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y})}, \tag{15}$$

which we can also write as

$$\ln\, P(\mathbf{x}|\mathbf{y}) = \ln\, P(\mathbf{y}|\mathbf{x}) + \ln\, P(\mathbf{x}) - \ln\, P(\mathbf{y}). \tag{16}$$

$P(\mathbf{x}|\mathbf{y})$ is the probability of $\mathbf{x}$ given $\mathbf{y}$, i.e. the probability of a given T-profile for a given radiance measurement.

4 Linear Retrieval Solution

Doing a retrieval means getting the most probable value (expectation value) of $\mathbf{x}$, $\hat{\mathbf{x}}$, plus some measure of how accurate that value is, which we express as a covariance matrix $\hat{\mathbf{S}}$. If we don't obtain $\hat{\mathbf{S}}$, $\hat{\mathbf{x}}$ is pretty useless on its own! We start with the linear retrieval problem stated above:

$$\mathbf{y} = \mathbf{F}(\mathbf{x}) + \boldsymbol{\epsilon} = \mathbf{K}\mathbf{x} + \boldsymbol{\epsilon}. \tag{17}$$

Now, rather than simply trying to invert this equation, we incorporate prior information using Bayes' theorem. Specifically, we want to work out $P(\mathbf{x}|\mathbf{y})$, which means first calculating $P(\mathbf{y}|\mathbf{x})$, $P(\mathbf{x})$ and $P(\mathbf{y})$.

We write $P(\mathbf{y}|\mathbf{x})$ as

$$P(\mathbf{y}|\mathbf{x}) \propto \mathrm{e}^{-\frac{1}{2}(\mathbf{y}-\overline{\mathbf{y}})^T\mathbf{S}_\epsilon^{-1}(\mathbf{y}-\overline{\mathbf{y}})} \quad \rightarrow \quad \ln\, P(\mathbf{y}|\mathbf{x}) = -\frac{1}{2}(\mathbf{y}-\overline{\mathbf{y}})^T\mathbf{S}_\epsilon^{-1}(\mathbf{y}-\overline{\mathbf{y}}) + \text{const.} \tag{18}$$

Here $\mathbf{S}_\epsilon$ is the measurement error covariance matrix. By using the above pdf, we are assuming a gaussian error distribution. Gaussian errors are usually a reasonable assumption for measurements with small errors, unless there are systematics, in which case you might need a propagation model. We won't cover this source of complexity here.

Now $\overline{\mathbf{y}}$ is simply found by setting $\boldsymbol{\epsilon}$ to zero: $\overline{\mathbf{y}} = \mathbf{K}\mathbf{x}$. So

$$\ln\, P(\mathbf{y}|\mathbf{x}) = -\frac{1}{2}(\mathbf{y}-\mathbf{K}\mathbf{x})^T\mathbf{S}_\epsilon^{-1}(\mathbf{y}-\mathbf{K}\mathbf{x}) + \text{const.} \tag{19}$$

Next up is $P(\mathbf{x})$, the distribution of $\mathbf{x}$ before we know anything about $\mathbf{y}$, or *prior* distribution. This we write as

$$\ln\, P(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{x}_a)^T\mathbf{S}_a^{-1}(\mathbf{x}-\mathbf{x}_a) + \text{const.} \tag{20}$$

$\mathbf{x}_a$ is the prior: our best estimate of the state vector based on our previous knowledge of the system. $\mathbf{S}_a$ is the prior covariance matrix. It is often tricky to define objectively, but still has to be included. Finally we have $P(\mathbf{y})$. This one is easy: because there is no $\mathbf{x}$ dependence, we simply write

$$\ln\, P(\mathbf{y}) = \text{const.} \tag{21}$$

Using Bayes' theorem, we can now write

$$\ln\, [P(\mathbf{x}|\mathbf{y})] = -\frac{1}{2}(\mathbf{y}-\mathbf{K}\mathbf{x})^T\mathbf{S}_\epsilon^{-1}(\mathbf{y}-\mathbf{K}\mathbf{x}) - \frac{1}{2}(\mathbf{x}-\mathbf{x}_a)^T\mathbf{S}_a^{-1}(\mathbf{x}-\mathbf{x}_a) + \text{const.} \tag{22}$$

This is quadratic in $\mathbf{x}$, so we can assert that $P(\mathbf{x}|\mathbf{y})$ also has a gaussian form, such that

$$(\mathbf{x}-\hat{\mathbf{x}})^T\hat{\mathbf{S}}^{-1}(\mathbf{x}-\hat{\mathbf{x}}) = (\mathbf{y}-\mathbf{K}\mathbf{x})^T\mathbf{S}_\epsilon^{-1}(\mathbf{y}-\mathbf{K}\mathbf{x}) + (\mathbf{x}-\mathbf{x}_a)^T\mathbf{S}_a^{-1}(\mathbf{x}-\mathbf{x}_a) + \text{const.} \tag{23}$$

Now the strategy is to expand out the brackets and equate like terms. The quadratic terms in $\mathbf{x}$ give us

$$\mathbf{x}^T\hat{\mathbf{S}}^{-1} = \mathbf{x}^T\mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{K} + \mathbf{x}^T\mathbf{S}_a^{-1}. \tag{24}$$

$$\hat{\mathbf{S}}^{-1} = \mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{K} + \mathbf{S}_a^{-1} \tag{25}$$

Terms linear in $\mathbf{x}^T$ give:

$$\hat{\mathbf{S}}^{-1}\hat{\mathbf{x}} = \mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{y} + \mathbf{S}_a^{-1}\mathbf{x}_a \tag{26}$$

We could do the same for $\mathbf{x}$ of course. But we already know what $\hat{\mathbf{S}}^{-1}$ is.

$$\hat{\mathbf{x}} = \hat{\mathbf{S}}\left(\mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{y} + \mathbf{S}_a^{-1}\mathbf{x}_a\right). \tag{27}$$

From (26) we have

$$\mathbf{S}_a^{-1}\mathbf{x}_a = \left(\hat{\mathbf{S}}^{-1} - \mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{K}\right)\mathbf{x}_a \tag{28}$$

so

$$\hat{\mathbf{x}} = \mathbf{x}_a + \left(\mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{K} + \mathbf{S}_a^{-1}\right)^{-1}\mathbf{K}^T\mathbf{S}_\epsilon^{-1}(\mathbf{y}-\mathbf{K}\mathbf{x}_a). \tag{29}$$

This is our solution to the linear retrieval problem.

Note that

$$\mathbf{G} = \left(\mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{K} + \mathbf{S}_a^{-1}\right)^{-1}\mathbf{K}^T\mathbf{S}_\epsilon^{-1} \tag{30}$$

is the *gain matrix*. This can be thought of as a kind of generalised inverse of $\mathbf{K}$. It tells us the contribution of the signal (as opposed to the prior information) to $\hat{\mathbf{x}}$.

$\mathbf{A} = \mathbf{G}\mathbf{K}$ is the *averaging kernel* or *model resolution* matrix. It describes the subspace of state space where the retrieval lies. It also can be thought of as the sensitivity of the retrieval to the true state. When $\mathbf{A}$ is low, you're seeing the *a priori*, not the real data.

4.1 Scalar Linear Retrieval Solution

Making the vectors into scalars is helpful for visualization. The equation for $\hat{\mathbf{x}}$ becomes

$$\left(\sigma_\epsilon^{-2}k^2 + \sigma_a^{-2}\right)\hat{x} = \sigma_\epsilon^{-2}ky + \sigma_a^{-2}x_a \tag{31}$$

$$\hat{x} = \frac{k(y-kx_a)}{k^2+F^2} + x_a \tag{32}$$

where

$$F = \sigma_e/\sigma_a. \tag{33}$$

This equation has two limits. The first is where the measurement errors disappear and $F \to 0$:

$$\hat{x} = k^{-1}y \tag{34}$$

In this situation, we can determine x perfectly. The second is where measurement errors are huge and $F \to \infty$:

$$\hat{x} = x_a. \tag{35}$$

In this situation, we get no new information and our best estimate of x remains the prior estimate.

5 Non-linear optimal estimation

We are still looking for maximum probability state $\hat{\mathbf{x}}$. By definition, this is where

$$\nabla_{\mathbf{x}} P(\mathbf{x}|\mathbf{y}) = 0. \tag{36}$$

$\mathbf{K}$ now depends on $\mathbf{x}$

$$\mathbf{K} = \frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{x}} \tag{37}$$

We need to find the stationary point. Can use e.g., a generalized Newton's method

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \frac{\mathbf{g}(\mathbf{x}_i)}{\nabla_{\mathbf{x}} \mathbf{g}(\mathbf{x}_i)} \tag{38}$$

If we stick with gaussian errors, we can iterate the state vector according to

$$\hat{\mathbf{x}}_{i+1} = \mathbf{x}_i + \left(\mathbf{K}_i{}^T \mathbf{S}_\epsilon^{-1} \mathbf{K}_i + \mathbf{S}_a^{-1}\right)^{-1} \left[\mathbf{K}_i^T \mathbf{S}_\epsilon^{-1} (\mathbf{y} - \mathbf{F}(\mathbf{x}_i)) - \mathbf{S}_a^{-1} (\mathbf{x}_i - \mathbf{x}_a)\right] \tag{39}$$

where $\mathbf{K}_i$ is the $\mathbf{K}$-matrix for state vector $\mathbf{x}_i$. C.f. our linear retrieval expression

$$\hat{\mathbf{x}} = \mathbf{x}_a + \left(\mathbf{K}^T \mathbf{S}_\epsilon^{-1} \mathbf{K} + \mathbf{S}_a^{-1}\right)^{-1} \mathbf{K}^T \mathbf{S}_\epsilon^{-1} (\mathbf{y} - \mathbf{K}\mathbf{x}_a) \tag{40}$$

Non-linear joint retrievals of temperature and composition can sometimes be advantageous.