Bayesian Retrievals

Note: This follows the discussion in Chapter 2 of Rogers (2000)

As we have seen, the problem with the nadir viewing emission measurements is they do not contain sufficient information for "stand-alone" retrievals of vertical atmospheric structure without smoothing which increases accuracy but limits our ability to determine the vertical structure of the atmosphere. Optimum utilization of their information is achieved by combining them with other "*apriori*" information about the atmospheric state. Examples of such *apriori* information are climatology and weather forecasts.

Use of *apriori* information suggests a Bayesian approach or framework, we know something about the atmospheric state, x, before we make a set of measurements, y, and then refine our knowledge of x based on the measurements, y. We need to define some probabilities.

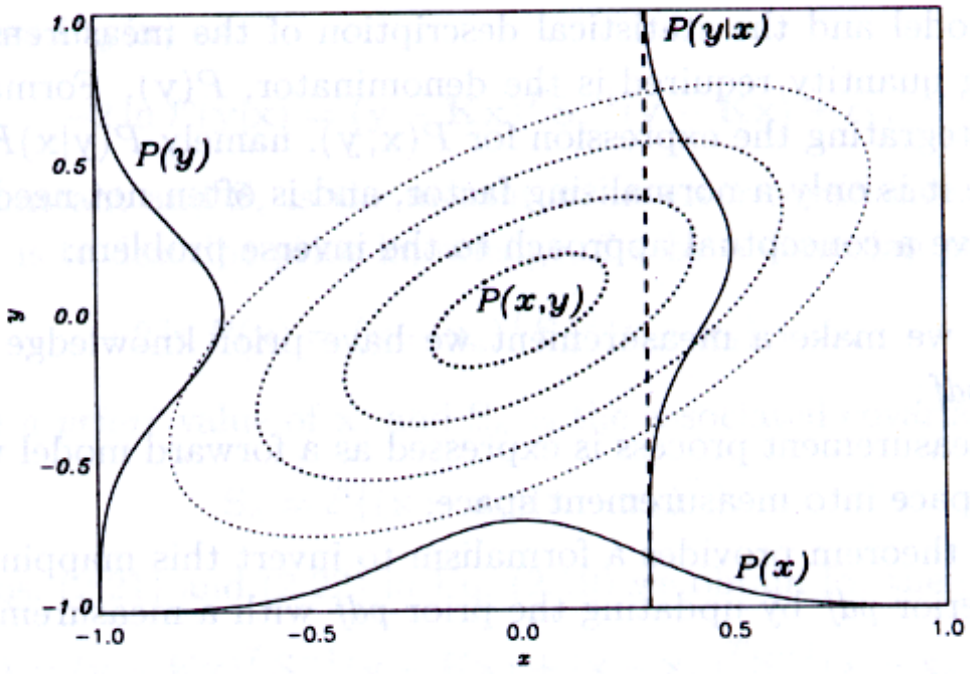
$P(x)$ is the *apriori* pdf of the state, x.

$P(y)$ is the *apriori* pdf of the measurement, y.

$P(x,y)$ is the joint pdf of x and y meaning that $P(x,y)\,dx\,dy$ is the probability that x lies in the interval $(x, x+dx)$ ***and*** y lies in $(y,y+dy)$

$P(y|x)$ is the conditional pdf of y given x meaning that $P(y|x)\,dy$ is the probability that y lies in $(y, y+dy)$ when x has a given value

$P(x|y)$ is the conditional pdf of x given y meaning that $P(x|y)\,dx$ is the probability that x lies in $(x, x+dx)$ when measurement y has a given value



Consider the joint probability, $P(x,y)$, shown as the contours in the figure above. $P(x)$ is given by the integral of $P(x,y)$ over all values of y.

$$P(x) = \int_{-\infty}^{\infty} P(x,y)\,dy \qquad (1)$$

Similarly

$$P(y) = \int_{-\infty}^{\infty} P(x,y)dx \tag{2}$$

The conditional probability, $P(y|x=x_1)$, is proportional to $P(x,y)$ as a function of y for a given value of $x = x_1$. This is defined as the $P(x,y)$ along a particular vertical line of $x = x_1$. Now $P(y|x_1)$ must be normalized such that

$$\int_{-\infty}^{\infty} P(y|x_1)dy = 1 \tag{3}$$

To normalize, we divide $P(x=x_1,y)$ by the integral of $P(x=x_1,y)$ over all y.

$$P(y|x_1) = \frac{P(x_1,y)}{\int_{-\infty}^{\infty} P(x_1,y)dy} \tag{4}$$

Now we substitute (1) into (4)

$$P(y|x_1) = \frac{P(x_1,y)}{P(x_1)} \tag{5}$$

We can do the same for $P(x|y=y_1)$

$$P(x|y_1) = \frac{P(x,y_1)}{P(y_1)} \tag{6}$$

We can combine (5) and (6) to get

$$P(x|y)P(y) = P(y|x)P(x) \tag{7}$$

For the present context where we are interested in the best estimate of the state, x, given measurements, y, we write (7) as

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \tag{8}$$

In this form, we see that on the left side we have the *posterior* probability density of x given a particular set of measurements, y. On the right side we have the pdf of the *apriori* knowledge of state, x, and the dependence of the measurements, y, on the state, x. The term, $P(y)$, is usually just viewed as a normalization factor for P(x|y) such that

$$\int_{-\infty}^{\infty} P(x|y)dx = 1 \tag{9}$$

With these we can update the *apriori* probability of the state, x_a, based on the actual observations and form the refined *posterior* probability density function, $P(x|y)$.

Note that (8) tells us the probability of x but not x itself.

Now consider a linear problem with Gaussian *pdf*s.

In the ***linear*** case, $y = \boldsymbol{K}x$ where $\boldsymbol{K}$ is a matrix that represents dy/dx. If this were all that there were to the situation, then knowing x and $\boldsymbol{K}$, we would know y and the pdf, $P(y|x)$, would simply be a delta function, $d(y=Kx)$. However, in a more realistic case, $y = \boldsymbol{K}x + \varepsilon$ where ε is the set of measurement errors. Therefore we don't know y exactly if we know x because y is a bit blurry due to its random errors.

We assume the y errors are Gaussian and generalize the Gaussian pdf for a scalar, y,

$$P(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \tag{10}$$

to a vector of measurements, y, where μ is the mean of random variable, y, and σ is the standard deviation of y defined as $<(y-\mu)^2>^{1/2}$.

The natural log of the pdf is quite useful when working with Gaussian pdfs creating a linear relation between the conditional pdfs. From (8) we can write

$$\ln\left[P(x|y)\right] = \ln\left[\frac{P(y|x)P(x)}{P(y)}\right] = \ln\left[P(y|x)\right] + \ln\left[P(x)\right] - \ln\left[P(y)\right] \tag{ln 8}$$

So with y and x now being measurement and state vectors respectively, we can write $P(y|x)$ as

$$-2\ln P(y|x) = (y - Kx)^T S_\varepsilon^{-1} (y - Kx) + c_1 \tag{11}$$

where c_1 is some constant and we have assumed the errors in y have zero mean (generally not strictly true), such that the mean y is Kx. Note that $\boldsymbol{K}x$ is the expected measurement if x is the state vector. The covariance of the measurement errors is S_e.

The error covariance is defined as follows. Given a set of measurements, $y_1, y_2, \ldots y_n$, each with an error, $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n$, then the covariance of the errors is shown in (12) for a set of 4 measurements

$$S_\varepsilon = \begin{matrix} \overline{\varepsilon_1\varepsilon_1} & \overline{\varepsilon_2\varepsilon_1} & \overline{\varepsilon_3\varepsilon_1} & \overline{\varepsilon_4\varepsilon_1} \\ \overline{\varepsilon_1\varepsilon_2} & \overline{\varepsilon_2\varepsilon_2} & \overline{\varepsilon_3\varepsilon_2} & \overline{\varepsilon_4\varepsilon_2} \\ \overline{\varepsilon_1\varepsilon_3} & \overline{\varepsilon_2\varepsilon_3} & \overline{\varepsilon_3\varepsilon_3} & \overline{\varepsilon_4\varepsilon_3} \\ \overline{\varepsilon_1\varepsilon_4} & \overline{\varepsilon_2\varepsilon_4} & \overline{\varepsilon_3\varepsilon_4} & \overline{\varepsilon_4\varepsilon_4} \end{matrix} \tag{12}$$

where $\overline{\varepsilon_1\varepsilon_2}$ represent the expected value of $\varepsilon_1\varepsilon_2$. For a Gaussian *pdf*, a mean and covariance are all that are required to define the *pdf*.

pdf of the apriori state

Now we look at the *pdf* of the *apriori* estimate of the state, x_a. To keep things simple, we also assume a Gaussian distribution, an assumption that is generally less realistic than the measurement error. The resulting *pdf* is given by

$$-2\ln P(x) = (x - x_a)^T S_a^{-1} (x - x_a) + c_2 \tag{13}$$

where S_a is the associated covariance matrix

$$S_a = \overline{\left\{(x - x_a)(x - x_a)^T\right\}} \tag{14}$$

Now we plug (11) and (13) into (ln 8) to get the *posterior pdf*, $P(x|y)$

$$-2\ln P(x|y) = (y-Kx)^T S_\varepsilon^{-1}(y-Kx) + (x-x_a)^T S_a^{-1}(x-x_a) + c_3 \tag{15}$$

Now we recognize that (15) is quadratic in x and must therefore be writeable as a Gaussian distribution. So we match it with a Gaussian solution

$$-2\ln P(x|y) = (x-\hat{x})^T \hat{S}^{-1}(x-\hat{x}) + c_4 \tag{16}$$

where $\hat{x}$ is the optimum solution and $\hat{S}$ is the *posterior* covariance representing the Gaussian distribution of uncertainty in the optimum solution. Equating the terms that are quadratic in x in (15) and (16):

$$x^T K^T S_\varepsilon^{-1} K x + x^T S_a^{-1} x = x^T \hat{S}^{-1} x \tag{17}$$

so that the inverse of the *posterior* covariance, $\hat{S}^{-1}$, is

$$\hat{S}^{-1} = K^T S_\varepsilon^{-1} K + S_a^{-1} \tag{18}$$

Now we also equate the terms in (15) and (16) that are linear in x^T

$$(-Kx)^T S_\varepsilon^{-1}(y) + x^T S_a^{-1}(-x_a) = x^T \hat{S}^{-1}(-\hat{x}) \tag{19}$$

Substituting (18) into (19) yields

$$(-x)^T K^T S_\varepsilon^{-1}(y) + x^T S_a^{-1}(-x_a) = x^T \left(K^T S_\varepsilon^{-1} K + S_a^{-1}\right)(-\hat{x}) \tag{20}$$

Now this must be true for any x and x^T so

$$K^T S_\varepsilon^{-1} y + S_a^{-1} x_a = \left(K^T S_\varepsilon^{-1} K + S_a^{-1}\right)\hat{x} \tag{21}$$

So

$$\hat{x} = \left(K^T S_\varepsilon^{-1} K + S_a^{-1}\right)^{-1}\left(K^T S_\varepsilon^{-1} y + S_a^{-1} x_a\right) \tag{22}$$

(22) shows that the optimum state, $\hat{x}$, is a weighted average of the apriori guess and the measurements.

We get the form of (22) that I have shown previously by isolating and manipulating the x_a term…

$$\begin{aligned}\left(K^T S_\varepsilon^{-1} K + S_a^{-1}\right)^{-1} S_a^{-1} x_a &= \left(K^T S_\varepsilon^{-1} K + S_a^{-1}\right)^{-1} S_a^{-1} x_a + \left(K^T S_\varepsilon^{-1} K + S_a^{-1}\right)^{-1}\left(K^T S_\varepsilon^{-1} K - K^T S_\varepsilon^{-1} K\right) x_a \\ &= \left(K^T S_\varepsilon^{-1} K + S_a^{-1}\right)^{-1}\left(K^T S_\varepsilon^{-1} K + S_a^{-1}\right) x_a + \left(K^T S_\varepsilon^{-1} K + S_a^{-1}\right)^{-1}\left(-K^T S_\varepsilon^{-1} K\right) x_a \\ &= x_a + \left(K^T S_\varepsilon^{-1} K + S_a^{-1}\right)^{-1}\left(-K^T S_\varepsilon^{-1} K\right) x_a \end{aligned} \tag{23}$$

Plug (23) back into (22) to get

$$\hat{x} = x_a + \left(K^T S_\varepsilon^{-1} K + S_a^{-1}\right)^{-1} K^T S_\varepsilon^{-1}(y - K x_a) \tag{24}$$

which is indeed the form I showed previously. As I said in class and in the Atmo seminar, (24) shows that the optimum *posterior* solution, $\hat{x}$, for the atmospheric state given an *apriori* estimate

of the atmospheric state, x_a, and its covariance, S_a, and a set of measurements, y, and its error covariance, S_e, is the *apriori* guess plus a weighted version of the difference between the expected measurement, Kx_a, and the actual measurement, y.

Note that while (8) is true in general independent of the *pdf*s, (22) and (24) are true as long as the uncertainty in the apriori state estimate and measurements can be accurately described in terms of Gaussian *pdf*s.

Note also that the unique solution, , depends on the apriori estimate, the measurements and their respective covariances. Change the covariances and the optimum state changes. Also note that it is assumed that there is not bias in the measurement errors or the apriori estimate. This is not true in general.

Consider the analogous situation with two estimates, x_1 and x_2, of the same variable, x.
We want the best estimate of x, $\hat{x}$, given the two estimates. To create this estimate, we need to know the uncertainty in each of the two estimates. If we know the uncertainty in each, s_1 and s_2, then we can weight the two estimates to create the best estimate. We will take the best estimate to be the estimate with the smallest uncertainty, s. The uncertainty or error in $\hat{x}$ is the expected value of $(\hat{x} - x_T)^2$ where x_T is the true value of x.

$$\hat{x} = A\, x_1 + (1\text{-}A)\, x_2. \tag{25}$$

where A is a weight between 0 and 1. So the variance of the error in $\hat{x}$ is

$$\overline{(\hat{x} - x_T)^2} = \overline{\left[Ax_1 + (1-A)x_2 - x_T\right]^2} \tag{26}$$

$$= \overline{\left[Ax_1 - Ax_T + (1-A)x_2 - (1-A)x_T\right]^2} = \overline{\left[A(x_1 - x_T) + (1-A)(x_2 - x_T)\right]^2}$$

$$= \overline{\left[\{A(x_1 - x_T)\}^2 + \{(1-A)(x_2 - x_T)\}^2 + \{A(x_1 - x_T)\}\{(1-A)(x_2 - x_T)\}\right]}$$

$$= A^2\overline{(x_1 - x_T)^2} + (1-A)^2\overline{(x_2 - x_T)^2} + A(1-A)\overline{(x_1 - x_T)(x_2 - x_T)}$$

$$\sigma_{\hat{x}}^2 = A^2\sigma_1^2 + (1-A)^2\sigma_2^2 + A(1-A)\overline{(x_1 - x_T)(x_2 - x_T)} \tag{27}$$

where we have used the definitions of s_1 and s_2. The next question is the cross term. If the errors in x_1 and x_2 are uncorrelated then the cross term is 0. So let's keep things simple and assume this to be the case where the two estimates come from two different measurement systems.
However, if this is not the case then the cross term must be known. This term is equivalent to the off diagonal terms in the covariance in (12).

$$\sigma_{\hat{x}}^2 = A^2\sigma_1^2 + (1-A)^2\sigma_2^2 \tag{28}$$

We want the solution that minimizes $\sigma_{\hat{x}}^2$. So we take the derivative of $\sigma_{\hat{x}}^2$ with respect to A and set it to 0.

$$\frac{d\sigma_{\hat{x}}^2}{dA} = 2A\sigma_1^2 - 2(1-A)\sigma_2^2 = 0 \tag{29}$$

and the solution for A is

$$A = \frac{\sigma_2^2}{\left(\sigma_1^2 + \sigma_2^2\right)} \tag{30}$$

so the optimum solution for x is

$$\hat{x} = \frac{\sigma_2^2}{\left(\sigma_1^2 + \sigma_2^2\right)} x_1 + \frac{\sigma_1^2}{\left(\sigma_1^2 + \sigma_2^2\right)} x_2 \tag{31}$$

Manipulate this a bit…

$$\hat{x} = \frac{1}{\left(\frac{\sigma_1^2}{\sigma_2^2} + 1\right)} x_1 + \frac{1}{\left(1 + \frac{\sigma_2^2}{\sigma_1^2}\right)} x_2 = \frac{1}{\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_1^2}\right)} x_1 + \frac{1}{\left(\frac{\sigma_2^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2}\right)} x_2 = \frac{\sigma_1^{-2}}{\left(\sigma_2^{-2} + \sigma_1^{-2}\right)} x_1 + \frac{\sigma_2^{-2}}{\left(\sigma_2^{-2} + \sigma_1^{-2}\right)} x_2$$

$$\hat{x} = \frac{\sigma_1^{-2}}{\left(\sigma_2^{-2} + \sigma_1^{-2}\right)} x_1 + \frac{\sigma_2^{-2}}{\left(\sigma_2^{-2} + \sigma_1^{-2}\right)} x_2 \tag{32}$$

Note the similarity of (32) with

$$\hat{x} = \left(K^T S_\varepsilon^{-1} K + S_a^{-1}\right)^{-1} \left(K^T S_\varepsilon^{-1} y + S_a^{-1} x_a\right) \tag{22}$$

The variance of the error in x is

$$\sigma_{\hat{x}}^2 = \frac{\sigma_2^4}{\left(\sigma_1^2 + \sigma_2^2\right)^2} \sigma_1^2 + \frac{\sigma_1^4}{\left(\sigma_1^2 + \sigma_2^2\right)^2} \sigma_2^2 = \frac{\sigma_2^4 \sigma_1^2 + \sigma_1^4 \sigma_2^2}{\left(\sigma_1^2 + \sigma_2^2\right)^2} = \frac{\sigma_2^2 \sigma_1^2 \left(\sigma_1^2 + \sigma_2^2\right)}{\left(\sigma_1^2 + \sigma_2^2\right)^2}$$

$$\sigma_{\hat{x}}^2 = \frac{\sigma_2^2 \sigma_1^2}{\left(\sigma_1^2 + \sigma_2^2\right)} = \frac{1}{\left(\frac{1}{\sigma_2^2} + \frac{1}{\sigma_1^2}\right)} \tag{33}$$

Take the inverse of (33) and note the similarity between it and (18).

$$\left(\sigma_{\hat{x}}^2\right)^{-1} = \left(\sigma_2^2\right)^{-1} + \left(\sigma_1^2\right)^{-1} \tag{34}$$

$$\hat{S}^{-1} = K^T S_\varepsilon^{-1} K + S_a^{-1} \tag{18}$$

So the vector/matrix form is indeed a generalization of the scalar form (as it must be).

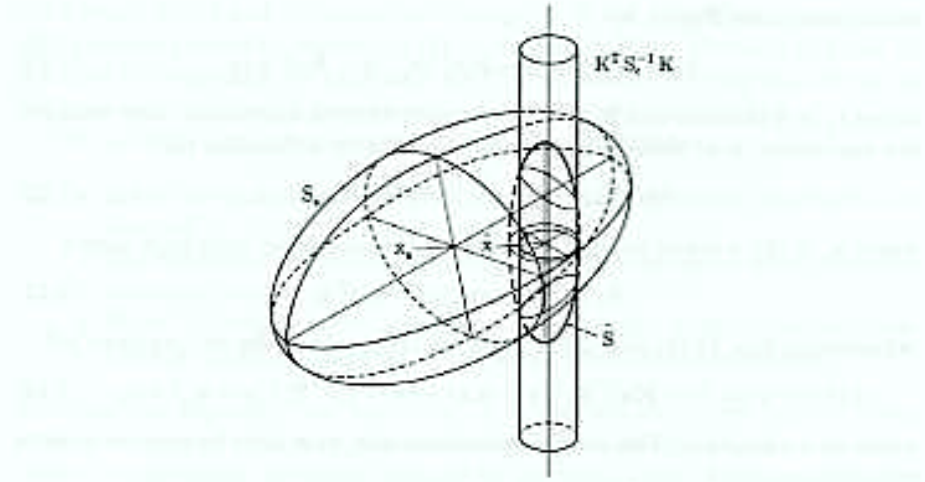


Fig. 2.4 Illustrating the relationship between the prior state estimate, the measurement mapped into state space, and the posterior estimate, for a three-dimensional state space and a two-dimensional measurement space. The large ellipsoid is a contour of the prior *pdf*, the cylinder is a contour of the *pdf* of the state given only the measurement, and the small ellipsoid is a contour of the posterior *pdf*.

Data Assimilation

So what is x_a and where does it come from? In numerical weather prediction (NWP) systems, x_a is the short term weather forecast over the next forecast update cycle, typically 1 to 12 hours. As such, x_a is a state estimate produced by a combination of atmospheric model and past observations. This is a very powerful way to assimilate and utilize observations n the weather forecasting business. The resulting updated state estimate, $\hat{x}$, is called the analysis or the analyzed state. The analysis is used as the initial atmospheric state to start the next model forecast run. The weather model needs an initial state to propagate forward in time.

Using analyses to study climate

These analyses are used as the best estimate of the atmospheric state to study climate. In theory these analyses are the best possible state estimate, optimally using the available atmospheric model and observational information. The problem, from a climate standpoint, is that the atmospheric models contain unknown errors including biases which are built right into the analyses via the x_a term. As such, it is problematic to use the analyses to evaluate climate model performance because they are based in part on the model. The degree of any particular atmospheric state variable depends on the degree to which observations constrain that variable. The less the variable is constrained by the observations, the more the behavior of that variable as represented in the analysis is the result of the forecast model.

It is this incestuous problem for determining climate, climate evolution and evaluating climate models that has driven us to make ATOMMS completely independent of atmospheric weather and climate models.