Bayesian Principles in Data Assimilation: A Tutorial

Larry Pratt and Laura Slivinski
(Oct 21, 2017)

Bayes' Rule is commonly used in the assimilation of data into ocean and atmosphere models. We hope the following will be useful to those who are new to Bayesian inference and want to develop a little knowledge and intuition about this subject.

Example 1: A drug trial.

A Coach's Dilemma

A friend of yours, a member of a tennis team, tests positive on a mandatory test for steroids. In considering disciplinary action, the player's coach does a little research and finds out that the test turns up positive 95% of the time when given to people using steroids. He concludes that your friend is probably a user and decided to suspend her from the team. Your friend objects, telling coach that people not taking steroids may sometimes test positive. The coach agrees and decides to seek more literature on the test. He finds a new statistic, namely that the test comes up positive 2.93% when given to random tennis players. This just enforces the coach's intuition that your friend's positive test is not a false positive. However, he also realizes that if the test were given to a population with no steroid users, *all* the positives would be false! He realizes that in order to estimate the likelihood that your friend is a steroid user, he needs to know something about the percent of tennis players that are users. After a bit more research, he learns that this is 1%. He is still uncertain, however, as to how to use this information.

The coach would do well to learn a little about conditioned probability: the chance of an event happening given some condition. He wants to know the probability that a certain tennis player is taking steroids given that she has tested positive for them; but the first set of information he is given is the likelihood that a person will test positive for steroids given that she is already taking them.

To gain a bit more insight into this situation, let's look at the raw data from the drug trial that was used to produce the statistics quoted above. In this trial, the test was given to a population of 10,000 tennis players. It was first established through confidential surveys that 1% of this group are steroid users. Then from this and the information already given:

# of steroid users = 100
# of steroid users who will test positive = 95
# of total positive tests=10,000 x .0293=293
likelihood that a player testing positive is a steroid user=95/293=.324

So, the likelihood that your friend is taking steroids is only about 1 in 3.

Marginal, Conditional and Conjoint Probability

This problem can be approached using Bayes' Rule, which is the foundation for a whole branch of statistics. Before we discuss the Rule, let's introduce some formal concepts and notation.

Let p(X) denote the probability that an event X will occur, also known as the *marginal probability*. In the drug trial, the possible events were very limited. Either the drug test is positive (X=T) or negative (X=F); either the person being tested is a steroid user (X=U) or not (X=N). Thus

p(T) = probability that a random tennis player will test positive =2.93%
p(N) = probability that a random tennis player is steroid free = 99%

Next define the *conditional probability* P(X|Y) that event X will occur given that Y has occurred. For example

p(T|U)=probability that a tennis player will test positive (T), given that they are already taking steroids (U) = 95%.

p(N|F)=probability that a person is steroid free (N) given that they test negative (F) =.9995

(This last calculation takes a little more thought and the reader might want to work out the answer and check it against ours.)

To introduce Bayes' Rule, we need to consider one other type of joint probability, namely the likelihood that X and Y occur together. This is known at the *conjoint* probability and is denoted p(X,Y). For example, the probability that a random tennis player within our population of 10,000 tests positive for steroids (T) *and* that the players is using steroids (U) is

p(T,U)=95/10,000=.095.

(Contrast this with likelihood p(T|U)=.95 that a player tests positive for steroids, *given* that they are using steroids.)

Note that conjoint probability is sometimes denoted $p(X \cap Y)$.

Bayes' Rule

Thomas Bayes was an 18th Century Presbyterian minister who lived in England and who was interested in mathematics, statistics, and philosophy. To derive his Rule (which was published posthumously) suppose that we wish to compute the conjoint probability p(F,U): the likelihood that a tennis player will test negative for steroids and that the person is a user. One way to proceed would be to first identify the group that tests

negative for steroids. Inspection of the numbers given above show that this group is made up 10,000-293=9707 individuals. Next, we scan this group for steroid users, of which there are 5. So for a population of 10,000 tennis players the likelihood that a player tests negative *and* is a user is p(F,U)=5/10,000=.0005. An equivalent way of performing the same calculation is to first compute the likelihood p(F) that a person tests negative for steroids, then multiply it by the probability p(U|F) that people from within the group that tests negative is taking steroids:

$$p(F,U)= p(F)p(U|F) =(9{,}707/10{,}000) \times(5/9{,}707)=.0005 \qquad (1a)$$

Alternatively we can change the order in which we count. We can first identify all of the players that are using steroids (100) and from this group identify those who test negative (again just 5). We might denote this likelihood as p(U,F) to signify that we first identify the group of users, then select the ones from this group that test negative. Of course the answer is the same (5/10,000) as before. Equivalently, we can compute p(U,F) by multiplying the likelihood p(U) of steroid use by the probability p(F|U) of a negative test among the steroid users:

$$p(U,F)= p(U)p(F|U) =.01\times.05=.0005 \qquad (1b)$$

From (1a,b) it is apparent that

$$p(X,Y)=p(Y)p(X|Y)=p(X)p(Y|X) \qquad (2)$$

or

$$p(X|Y)= p(Y|X)\, p(X)/p(Y) \qquad \text{(Bayes' Rule)} \qquad (3)$$

In Bayesian lingo, the distribution p(X|Y) is referred to as the *posterior*, p(Y|X) as the *likelihood*, p(X) as the *prior*, and p(Y) as the *marginal distribution* or *evidence.* The intuitive connection of these terms to the problems we will address below is not always obvious. In particular, the term *likelihood* seems very general and could refer to any probability distribution. Note that this distribution is also known as *sampling distribution* or *measurement model,* both of which have a bit more meaning for the temperature measurement scenario that is presented below.

This is one form of Bayes' Rule. It is based on a principle that is self evident, almost childishly simple, namely that the odds of two simultaneous events can be computed in either order: one can first add up the number of times the first event takes place, then select from that group the number of times that the second event takes place, or *vice versa*. Had the tennis coach known about Eq. (3) he could have computed the probability p(U|T) that his tennis player was using steroids given her positive test result. The coach was given p(T|U)=.95 (the likelihood that a steroid user will test positive), p(T)=.0293 (the likelihood that a random tennis player will test positive) and p(U)=.01 (the likelihood that a random tennis player is a user). This would have given him p(U|T)=(.95×.01)/.0293=.324.

One further observation: the evidence distribution p(Y) (that Y occurs at all) is clearly the sum of all the instances of Y, divided by the total number of opportunities for Y to occur. The total number of occurrences of Y can be computed by taking the number of occurrences of Y for each possible X and then summing over all the possible X values, which we denote x_i. Thus

$$p(Y)=\sum_i p(Y|x_i)p(x_i) \qquad (4)$$

and it is only necessary, therefore, to know p(Y|X) and p(X) in order to evaluate the entire right-hand-side of (3). If X is a continuous variable then (4) is replaced by $p(Y)=\int p(Y|X)p(X)dX$. In some applications, it is easiest to regard P(Y) as it occurs in Eq. (4) as just a normalizing factor, computed to ensure that for any Y, the area under the posterior distribution p(X|Y) is unity.

Example 2: coin flips.

Suppose you make simultaneous coin tosses with 4 coins, leading to $2^4=16$ possible outcomes. You are interested in the following probabilities

1T -> Exactly one tail turns up: 4 possibilities
H -> At least one head turns up: 15 possibilities

Also note the following:

p(H)= probability of at least one head=15/16
p(1T)= probability of exactly one tail=4/16
p(H,1T)= probability of exactly one tail and at least one head=4/16
p(1T|H)= probability of exactly one tail given at least one head=4/15
p(H|1T)= probability of at least one head given exactly one tail=1

We can also verify that Bayes' Rule works for this case by noting

p(H,T)=4/16
p(T) p(H |1T)=(4/16)x1=4/16
p(H) p(1T |H)=(15/16)x(4/15)=4/16

so that

p(H |1T)= p(1T |H) p(H)/ p(1T)=(4/15)(15/16)/(4/16)=1

Example 3: Assimilation of a temperature measurement into an ocean model.

Calibration of the temperature sensor.

Suppose that you plan to measure the ocean temperature with a thermistor and that you first need to calibrate the instrument in a laboratory tank. The true temperature X of the water in the tank is considered known to a very high degree of accuracy, perhaps from an expensive laboratory thermometer. The temperature in the tank is varied over the approximate range, say 12^{o}C to 18^{o}C, that you expect to see at the location where the instrument will be deployed. We will consider two ways in which one might carry out the calibration. In each, we will divide up the temperature scale in a series of even intervals, say of size 1/10,000 ^{o}C, and we will use lower case symbols x_i or y_i to represent the temperature range of the ith interval. Any temperature reading is defined by the discrete interval that it lies in.

Case 1: The tank is maintained in the narrow temperature range x_1 corresponding, say, to 12 to 12.0001 ^{o}C, and repeated readings y_i of the thermistor are made. The resulting histogram gives an approximation of p(Y|x_1), the probability of a measurement y_i given that the true temperature x_i lies in the range 12 to 12.0001 ^{o}C. (Here Y just represents the full range of possible temperature intervals (y_1,y_2,y_3,….), which might include temperatures outside of the true range.) Next the tank temperature is reset to lie in the interval x_2=12.0001,12.0002) and the process repeated to get p(Y|x_2). Eventually the full range of expected temperatures x_i is covered. We now have a
P(Y|X) for all expected x_i.

Case 2: The temperature of the tank is allowed to vary continuously during the day in a way that mimics the expected variations of the ocean region that will be sampled. For example, the target region might be the ocean surface mixed layer in a subtropical region where the daytime/nighttime temperature varies by a few degrees. In this sense, the laboratory tank becomes a model of the region that we want to measure. To calibrate the thermistor, simultaneous readings of the thermistor and the reliable laboratory thermometer are made repeatedly over the course of the day and over many days.

For each approach, we can store the data in a table in which the rows correspond to different x_i values, the columns correspond to y_j values, and the entry (i,j) gives the number of times that x_i and y_j occur together. From this histogram we could easily construct the conjoint probability distribution p(X|Y). The entry (i,j) would then give p(x_i,y_i), the probability that true temperature x_i and measured temperature y_i occur together. Or, by examining the distribution of y_j value within a particular row i of the table, we could compile p(Y|x_i), the probability distribution of the measured temperature given the true temperature x_i. In principle this last distribution should depend only on the sensor being tested and not on the scenario under which it is tested. If the true temperature of the tank is x_i then the distribution of measurements y_j of that true temperature should not depend on how the true temperature x_i was established. So p(Y|x_i), and more generally p(Y|X), should be the same for Method 1 and Method 2,

provided that a sufficient amount of data is collected over the same true temperature range.

Although p(Y|X) should be the same for the two methods, there is no reason to expect that p(Y,X), p(Y) or P(X) will be the same. In Method 1, for example, we might have taken 1000 temperature readings when the tank temperature was set to x_5, whereas the temperature might have spent very little time in the x_5 range in the Method 2 scenario, resulting in only 100 readings. This might mean that the chance of y_5 and x_5 occurring together is much greater under Method 1.

As an illustration, suppose we have only two true temperature bins x_1 and x_2, and that all the sensor readings fall into the same two bins, y_1 and y_2. Below are hypothetical tables of (x_i,y_j) value collected under the two scenarios. In Scenario 1, where the true temperature of the tank is held constant while multiple readings are taken, there are an equal number of temperature readings, 12 to be exact, in each true temperature range. In Scenario 2, where the true temperature is forced to vary continuously, there are 8 readings corresponding to true temperature x_1 and 15 readings corresponding to x_2. Therefore the marginal probability $p(x_1)=12/24=1/2$ in the first case, but is equal to 8/23 in the second case. Likewise, $p(x_1,y_2)=3/24$ in the first case and =2/23 in the second, while $p(x_1|y_2)=3/11$ in Case 1 and =2/12 in the second. The reader can verify that only the likelihood p(Y|X) remains the same: for example, the conditional probability $p(y_1|x_2)=1/3$ in each case.

	y_1	y_2
x_1	9	3
x_2	4	8

Table 1a: A table histogram showing the number of measurements of each combination of true temperature bin x_i and measured temperature bin y_j for the Scenario 1 experiment and with the temperature range divided into only two bins. An equal number (12) of measurements have been made in each true temperature class.

	y_1	y_2
x_1	6	2
x_2	5	10

Table 1a: Same as Table 1a, but for the Scenario 2 experiment. The number of measurements made in each true temperature class is no longer uniform.

Before deploying the instrument in the ocean, we first carry out a simulated experiment using the environment established in the second method, where the tank temperature cycles through a range of values each day. This will simulate the natural diurnal temperature variation of the target ocean. Since the true temperature in the tank is now regarded as unknown, we ignore readings form the accurate thermometer. We then take a temperature reading y_j with the thermistor and ask what is the probability distribution p(X| y_j) of the true temperature X given the measurement y_j.
In the idealized case where there are only two temperature bins y_1 and y_2 we can answer this questions by looking at Table 1b. If the measured temperature is y_2, say, then the probability that the true temperature is x_1 is 5/15=1/3 and the probability that the true temperature is x_2 is 10/15=2/3.

So far we have not had to use Bayes' Rule. But suppose that we do not have access to the raw data used to construct the histogram tables. Like the tennis coach in the first example, we only have access to published statistics provided by the person or company that has carried out the testing. Again, we seek p(X|Y) because we want to know the distribution of true temperatures X that accompany any measured value $y_j \in$ Y, but we are given quantities like p(Y|X), p(Y) and p(X). Application of Bayes' Rule (3) and use of Eq. (4) yields

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)} = \frac{p(Y|X)p(X)}{\sum_i p(Y|x_i)p(x_i)} \quad . \qquad (5)$$

Thus, we really only need two distributions p(Y|X) and p(X) to find p(X|Y). As we have seen, the likelihood function (or *measurement model*) p(Y|X) depends only on the instrument and does not depend on which experiment (Case 1 or Case 2) is used to obtain it. On the other hand, the prior distribution p(X) differs between two cases, and if we are trying to interpret the meaning of a measurement made in the tank under the Case 2 scenario, it is crucial that we use the prior distribution obtained under Case 2 conditions.

The Real Ocean

If we now deploy the thermistor in the real ocean and make a single measurement y_j we are again interested in p(X| y_j), the probability distribution of the true temperature given the measurement. As long as the laboratory calibration remains valid (the sensor has not started to drift, say) we can use the likelihood distribution (measurement model) p(Y|X) already obtained. For the sake of illustration, let's suppose that the laboratory calibration shows that the instrument errors are normally distributed, with standard deviation r and are unbiased (i.e. the errors have zero mean). We also consider the limit of very small bin size, so that X and Y vary continuously. Then

$$P(Y|X) = (2\pi)^{-1/2} e^{-(Y-X)^2/2r^2} \quad . \qquad (6)$$

If there is no prior information about temperature variations in the region we are modeling, the prior distribution p(X) remains unknown. However, under the Case 2

scenario we clearly do have prior knowledge of the regional temperature variations and have, in fact, set up the laboratory tank to vary in a way that is consistent with this knowledge. We may then use the p(X) obtained in Case 2 in our application of Eq. (5). Success very much depends on how well the p(X) so obtained matches the true distribution. Had we no prior knowledge, we could have still done a calibration as in Case 1, but the p(X) so obtained would be inappropriate. Under these conditions we might choose to use a flat distribution for p(X):

$$p(X) = \begin{cases} w^{-1} & (X_{\min} < X < X_{\max}) \\ 0 & (\text{otherwise}) \end{cases}, \qquad (7)$$

where $X_{min}<X< X_{max}$ gives the possible range of temperatures that can occur in the system and $w=X_{max}-X_{min}$. The use of (7) reflects the any particular X is just as likely to occur as any other. We leave it as an exercise to show that when (7) is used (in the limit of infinite w where appropriate) then p(X|Y)= p(Y|X).

Using a model to estimate the prior: data assimilation

In the field of data assimilation, we are faced with essentially the same problem as posed above. We have a likelihood distribution p(Y|X) based on the instrument and perhaps on other uncertainties that are associated with the way in data is measured. To apply Bayes' Rule we need a prior P(X) and we now consider obtaining it from a model. The prior must reflect the probability that any particular value of X occurs in the system we are measuring. If the model is trustworthy, then we can examine its past history in order to construct an approximation for P(X), much the same as was done in the tank temperature experiment.

One challenge is that at the time the model run is initiated, the only information we have about the past is contained in the imposed initial conditions. Suppose that the model is initiated at $t=0$ and is run forward to the time t_1 of the first available observation. Then we have a limited history, in most cases too brief to construct P(X). However, we can increase the amount of available information by implementing an ensemble of model runs beginning with slightly different initial conditions. We might argue that the original initial conditions are known only within some range of uncertainty, and that any model run started from within that range yields a valid approximation of the true state of the system. Below are some common schemes that are used to implement this notion. All of these apply to *sequential* data assimilation (also known as *filtering)*.

The Particle Filter

In almost any realistic setting, we will never actually have a full continuous representation of P(X), although X will usually be a continuous variable. We therefore approximate P(X) as a discrete probability mass function (PMF), which results in the

necessity to approximate P(X|Y) as a discrete PMF as well. The particle filter is a data assimilation method which then describes how to calculate P(X|Y) via Bayes rule.

To make things simple, we continue with the case in which the model predicts a single scalar quantity, namely the temperature X at some location in the ocean. X is divided into discrete bins (or 'particles') x_i that span the full range of possible temperatures at that point. We can think of $P(x_i)$ as the probability that x_i occurs (the *marginal probability*), and represent this probability as a *weight* w_i. This distribution of weights is a discrete representation of P(X) as shown in Figure 1.

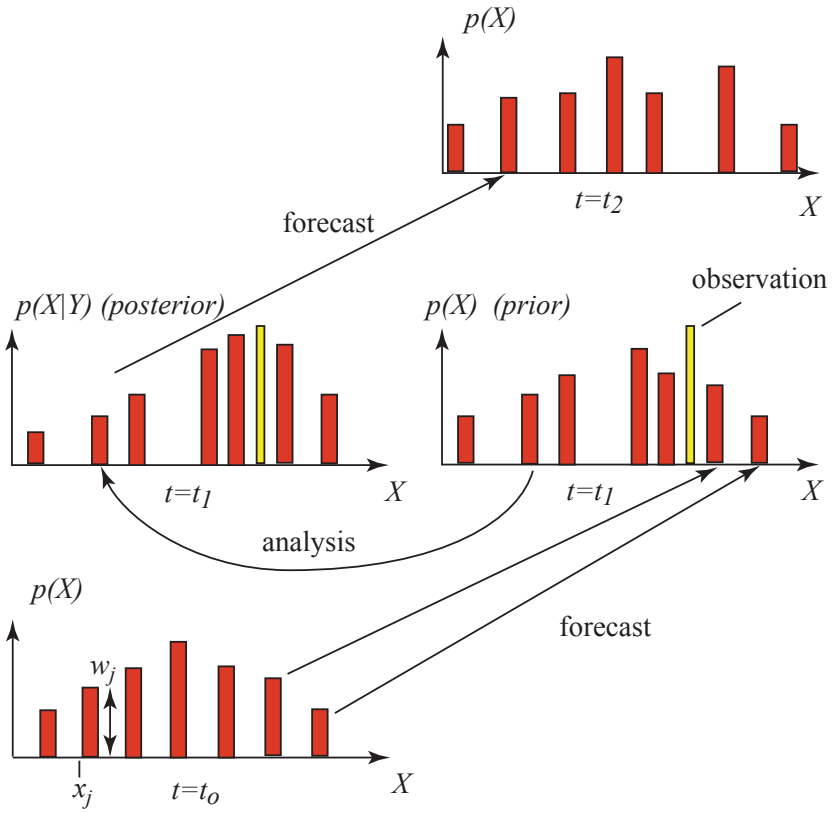


Figure 1. Schematic of the particle filter. The initial distribution (lower left) is represented as a set possible states or particles, each represented by red bar. Particle *j* is characterized by a range of X values about some central value x_j and by a probability or 'weight' w_j. In the forecast step, the *X*-value for each particle is evolved under the model from time t_o to t_1 in order to obtain the prior distribution at $t=t_1$. The weights remain unchanged during this step. Next comes the analysis stage in which Bayes' Rule is applied to each particle in order to adjust the weights, given an observation y at time t_1. During this step the X-values remain the same but the weights are updated. The posterior distribution thus obtained is then evolved forward in time in order to create a prior distribution for the analysis at time t_2.

It may be that the only prior information we have at the beginning of the model run is the initial condition. In some cases the initial condition is informed by past history; in other cases there is virtually no past history (as when the flow in a circulation model is initiated from a state of rest). In either case, we represent the initial condition using our set of weighted 'particles'. Each particle is then evolved forward in time under the model until the time t_1 of the first observation Y. The temperature (or temperature range) assigned to each particle then changes, but we require that each retain its assigned weight.
This forecast distribution is regarded as the prior P(X), to be used in the analysis stage that comes next: it contains information contained in the initial conditions and in the history of evolution of the model over $t_o<t<t_1$. Clearly this may not be a very good approximation of the true prior, but the situation will improve in time as more data are assimilated.

In the analysis step, we apply Bayes' Rule in order to update the weight of each particle in view of the data. This updated set of particles and weights is then a discrete representation of P(X|Y). To calculate the (updated) weights according to Bayes Rule, we take our prior distribution P(X), represented by the ensemble $\{x_i,w^f_i\}$ where i ranges over all the particles and "f" stands for "forecast". For each particle x_i, we calculate the likelihood p(Y|xi) evaluated at Y=y for our actual observed quantity y. given our observation y (the actual observed quantity.) In the Gaussian case described above, the likelihood is

$$P(Y\big|x_i) = (2\pi)^{-1/2} e^{-(Y-x_i)^2/2r^2}, \qquad (8)$$

and $P(x_i)=w_i^f$. Then by Bayes' rule, we have

$$w_i^a = P(x_i,Y) = \frac{(2\pi)^{-1/2} e^{-(Y-X_i)^2/2r^2} w_i^f}{P(Y)} \qquad (9)$$

which we can evaluate for each particle x_i. The *marginal* distribution P(Y) is calculated by summing over all *i*:

$$P(Y) = \sum_i (2\pi)^{-1/2} e^{-(Y-x_i)^2/2r^2} w_i^f \qquad (10)$$

The new ensemble $(x_i,w_i^a\}$ corresponding to our new set of weights is our analysis distribution. The x_i values are the same as in the prior, but weight of each is updated. Since Y in this case is just a single value (the measured temperature), P(Y) is a single value. It serves here as a normalizing factor that insures that the probability function on the right-hand side of (9) will sum to unity.

If data is available at future times t_2, t_3, etc. we repeat the above procedure, first evolving the temperature x_i of each particle forward in time while conserving its weight w_i in order

to get a new forecast distribution. The weights are then updated as in (9) to get an analysis distribution.

Note that the model may not directly predict the quantity we are measuring. In the laboratory tank we measure temperature but the model may determine density, which must then be related to temperature using an equation of state. For this reason, we introduce an *observation operator H* that projects the model state variable into observation space: Y=H(X). In this case, x_i in the above relations is replaced by $H(x_i)$. In this way the actual observation is compared to what we would expect to have observed if our model and initial conditions were perfect.

The Kalman Filter

Here the model is assumed to be linear and the model states are assumed to have a normal distribution. The governing linear equations then determine how the mean and standard deviation of this distribution evolve, so that just one calculation is needed to produce the prior distribution. The analysis stage then consists of multiplying two Gaussian distributions (P(Y|X) and P(X)) together. The product of two Gaussians is a new Gaussian with mean and standard deviation given by

$$\overline{X}_a = \overline{X}_f + \frac{\sigma_f^2}{\sigma_f^2 + \sigma_d^2}(\overline{Y} - \overline{X}_f) \quad \text{and} \quad \sigma_a^2 = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_d^2}, \qquad (11a,b)$$

where $\overline{Y}$ and $\overline{X}_f$ are the means of the observations and forecast, and σ_d^2 and σ_f^2 are the respective variances. The mean $\overline{X}_a$ is therefore intermediate between the originals, and lies closer to the mean of the distribution with the lower variance. If the variance σ_d^2 of the observational data is much smaller than the forecast variance, then $\overline{X}_a$ will lie close to the mean $\overline{Y}$ of the observation. Note further that the analysis standard deviation is smaller than the standard deviations of **both** original standard deviations. (In the Bayesian context, this can be understood as: we have more information after combining the prior and the observation, so we have less uncertainty than we had in the prior alone, or in the observation alone.)

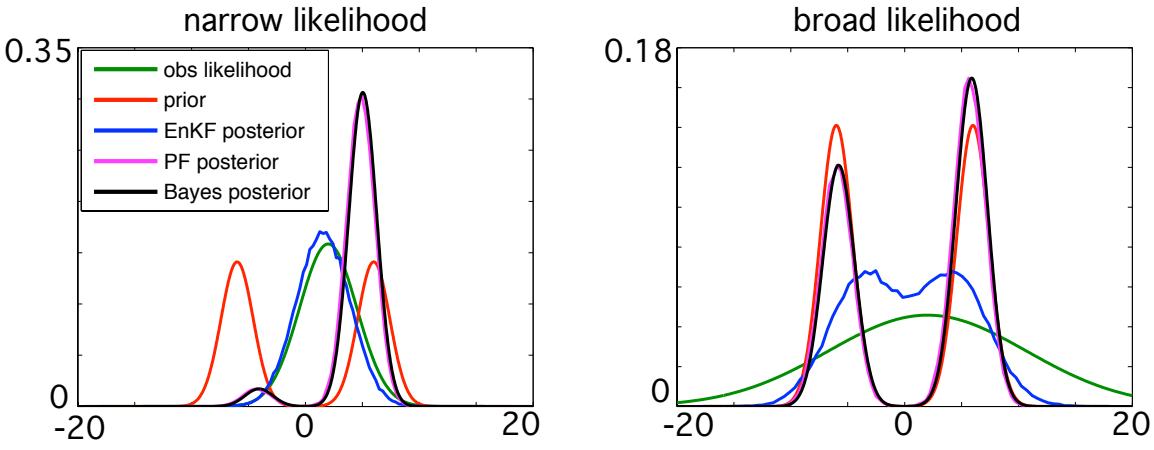
The Ensemble Kalman Filter

The ensemble Kalman filter (EnKF) attempts to relax the requirements of normal distributions and a linear model. Instead of evolving forward a Gaussian mean and standard deviation at the prediction stage, it evolves forward an ensemble of model states using the (possibly nonlinear) model. The resulting forecast ensemble, which is generally non-Gaussian, will act as a discrete approximation of the distribution of model forecast states. The analysis step then consists of updating each forecast ensemble member based on how close it is to the observation and on the uncertainties of the forecast and the data. For the prediction of a scalar quantity, the update equation for each ensemble member is just (3.4a), σ_f^2 and σ_d^2 being the sample variances of the forecast and measurement

ensembles, and replacing \bar{X}_f with x_i, and \bar{Y} with y_i. [Aside: there are several ways to implement the EnKF, but the "perturbed observation" method requires you to replace "y" above with "y_i=y+eta_i", where eta_i is a random variable with mean 0 and std dev r (observational error). This is necessary so that, when the limits are taken, the Gaussian EnKF result does have the correct Bayesian distribution. Let me know if you want to talk about this.]

When the model is linear and the distributions are Gaussian, the EnKF will provide the correct Bayesian posterior (the same as the Kalman filter) in the limit of an infinite ensemble. However, since this update is based on Gaussian assumptions, it tends to break down when the distributions are highly non-Gaussian. For instance, if the prior distribution has two distinct peaks, the posterior distribution resulting from the EnKF update will neither be Gaussian, nor will it be the true Bayesian posterior distribution.

In the two examples shown in Figure 2, the prior distribution is bimodal (two Gaussians). The observation likelihood distribution is Gaussian, with small variance (left panel) and larger variance (right panel). The posterior distributions are calculated using Bayes' rule (possible analytically, since you can write down the distribution of a two-Gaussian bimodal prior), the particle filter (PF), and the ensemble Kalman filter (EnKF.) The particle filter and the Bayesian posterior indistinguishable, because 10,000 particles were used for this 1D example. The same number of ensemble members are used in the EnKF, but clearly, it fails to correctly represent the Bayesian posterior in both cases.



The particle filter is better at handling highly non-Gaussian distributions, but the number of particles necessary for the particle filter increases exponentially as the dimension of the model state increases, and thus quickly becomes computationally infeasible. The EnKF, on the other hand, can be slightly modified to work well with 50-100 ensemble members, regardless of the state dimension. As long as the model is only weakly nonlinear (and the probability distributions are only weakly non-Gaussian), the EnKF is a relatively robust method that works well and is computationally feasible. In fact, the US

Weather Service has historically used a version of the EnKF to generate its weather forecasts.

Why is the PF more costly to implement than the EnKF?

Laura: I think it has to do with the fact that the PF is usually trying to approximate more complicated distributions than the EnKF tries to approximate. Off the top of my head, I don't know whether the PF and the EnKF can use the same number of particles to approximate a 1D Gaussian, with both getting the posterior variance right.

One practical reason that the PF might need more than the EnKF in a 1D case is that the basic, unmodified particle filter is recursive. That is, the weights are constantly updated by being multiplied by a factor derived from Bayes' rule. So, as soon as one particle starts to get more weight than other particles, it will usually just keep getting more and more weight, until all the other particles have weight 0 and your "discrete distribution" is just a single point. The EnKF update doesn't have this characteristic.

In more detail - regardless of the number of particles, that "weight collapse" is bound to happen for a finite ensemble eventually. This is why particle filters generally require "resampling." When the weight distribution hits a predetermined threshold, a new ensemble of particles is resampled from the weighted particle distribution, and the weights are reset to 1/N. (So, instead of having one particle with a very high weight and some particles with 0 weight, you throw away the particles with 0 weight and make [slightly perturbed] copies of particles with very high weight, and then continue on.)