The Likelihood, the prior and Bayes Theorem

Douglas Nychka,

`www.image.ucar.edu/~nychka`

- Likelihoods for three examples.
- Prior, Posterior for a Normal example.
- Priors for Surface temperature and the CO_2 problem.

Supported by the National Science Foundation NCAR/IMAGe summer school Jul 2007

The big picture

The goal is to estimate ‘parameters’: θ

System States

e.g. temperature fields from Lecture 1 or the surface fluxes of CO_2

or

Statistical parameters

e.g. such as climatological means or the correlation coefficients.

We also have ...

Data

Whose distribution depends on these unknown quantities (θ).

Reversing roles in the problem: ‘inverse probability’

We start by specifying the conditional distribution of the *data* given the *parameters*.

We end by finding the conditional distribution of the *parameters* given the *data*.

Likelihood

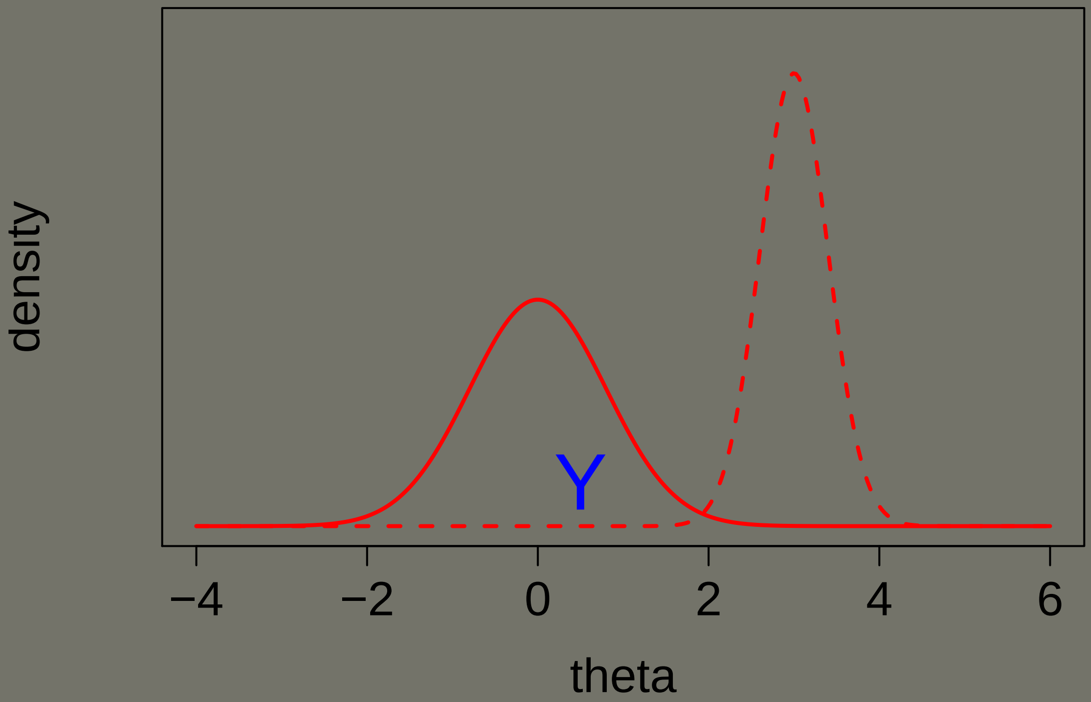
$L(Y, \theta)$ or $[Y|\theta]$

the conditional density of the data given the parameters.

Assume that you know the parameters exactly, what is the distribution of the data?

This is called a likelihood because for a given pair of data and parameters it registers how ‘likely’ is the data.

E.g.



Data is ‘unlikely’ under the dashed density.

Some likelihood examples.

It does not get easier that this!

A noisy observation of θ.

$$Y = \theta + N(0,1)$$

Likelihood:

$$L(Y,\theta) = \frac{1}{\sqrt{\pi}} e^{-\frac{(Y-\theta)^2}{2}}$$

Minus log likelihood:

$$-log(L(Y,\theta)) = \frac{(Y-\theta)^2}{2} + log(\pi)/2$$

-log likelihood usually in simpler algebraically

Note the foreshadowing of a squared term here!

Temperature station data

Observational model:

$Y_j = T(x_j) + N(0, \sigma^2)$ **for** $j = 1, ..., N$

$$L(Y,\theta) = \frac{1}{\sqrt{\pi}\sigma} e^{\frac{-(Y_1 - T(x_1))^2}{2\sigma^2}} \times \frac{1}{\sqrt{\pi}\sigma} e^{\frac{-(Y_2 - T(x_2))^2}{2\sigma^2}} \times ... \times \frac{1}{\sqrt{\pi}\sigma} e^{\frac{-(Y_N - T(x_N))^2}{2\sigma^2}}$$

combining

$$L(Y,\theta) = \frac{1}{(\sqrt{\pi}\sigma)^N} e^{-\sum_{j=1}^{N} \frac{(Y_j - T(x_j))^2}{2\sigma^2}}$$

Minus log likelihood:

$$-log(L(Y,\theta)) = \sum_{j=1}^{N} \frac{(Y_j - T(x_j))^2}{2\sigma^2} + (N/2)log(\pi) + Nlog(\sigma)$$

A simplifying complication

It will useful to express the relationship as

$$Y = HT + e$$

- Y is the data vector.
- H is an indicator matrix of ones and zeroes that maps the stations to the grid.
- T is the huge vector of all temperatures on the grid.
- e is the measurement error vector $(0, R)$.

Minus log likelihood:

$$-log(L(Y,\theta)) \propto (Y - HT)^T R^{-1} (Y - HT)/2 + N log(|R|)$$

The quadratic thing again!

CO_2 Inverse Problem

Z (data), x (concentrations) and u (sources).

$$z_j = h_j(x_i) + N(0, R)$$

for $j = 1, ..., N$
and

x_i is determined by the dynamical model:
$x_{i+1} = \Phi(x_i) + G(u)$

$$L(Z, u) = \frac{1}{(\sqrt{\pi}R)^N} e^{-\sum_{i,j} \frac{(z_j - h(x_i))^2}{2R^2}}$$

Minus log likelihood:

$$-logL(Z,u) = \sum_{i=0}^{(I-1)} \sum_{j=1}^{N} \frac{(z_j - h(x_i))^2}{2R^2} + (N/2)log(\pi) + Nlog(\sigma)$$

sources (u)

concentrations (x)

Priors

To get the conditional distribution of the parameters given the data we need the distribution of the parameters in the absence of any data. This is called the prior.

The simplest prior for θ

For the first example take θ to be $N(\mu, \sigma)$.

If this seems bizarre to put a distribution on this unknown quantity then you are probably following this lecture!

We are now ready to use Bayes theorem

Bayes Theorem again

Three ways of stating Bayes Thm:

- (parameters given data) $\propto$
  (data given parameters)$\times$ (parameters)
- $[\theta|Y] \propto [Y|\theta][\theta]$
- conditional density given the data
  $\propto$ L(Y,theta) prior(θ)

Posterior

The conditional density of the parameters given the data

Back to first example

$Y = \theta + N(0,1)$

Likelihood

$$[Y|\theta] = \frac{1}{\sqrt{\pi}} e^{-\frac{(Y-\theta)^2}{2}}$$

Prior for θ

$$[\theta] = \frac{1}{\sqrt{\pi}\sigma} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}$$

Posterior

$$[\theta|Y] \propto \frac{1}{\sqrt{\pi}} e^{-\frac{(Y-\theta)^2}{2}} \frac{1}{\sqrt{\pi}\sigma} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}$$

Simplying the posterior for Gaussian-Gaussian

$$[\theta|Y] \propto [Y|\theta][\theta] \propto e^{-\frac{(Y-\theta)^2}{2}-\frac{(\theta-\mu)^2}{2\sigma^2}} \propto e^{-\frac{(Y^*-\theta)^2}{2\sigma*}}$$

posterior mean:

$$Y^* = (Y\sigma^2 + \mu)/(1+\sigma^2)$$
$$Y^* = \mu + (Y - \mu)\sigma^2/(1+\sigma^2)$$

Remember that θ has prior mean μ

posterior variance:

$$(\sigma^*)^2 = \sigma^2/(1+\sigma^2)$$
$$(\sigma^*)^2 = \sigma^2 - 1/(1+\sigma^2)$$

Without other information Y has variance of 1 about θ.

As Jeff Anderson says:

products of Gaussians are Gaussian ...

Where do these products come from? What are statistical names for them.

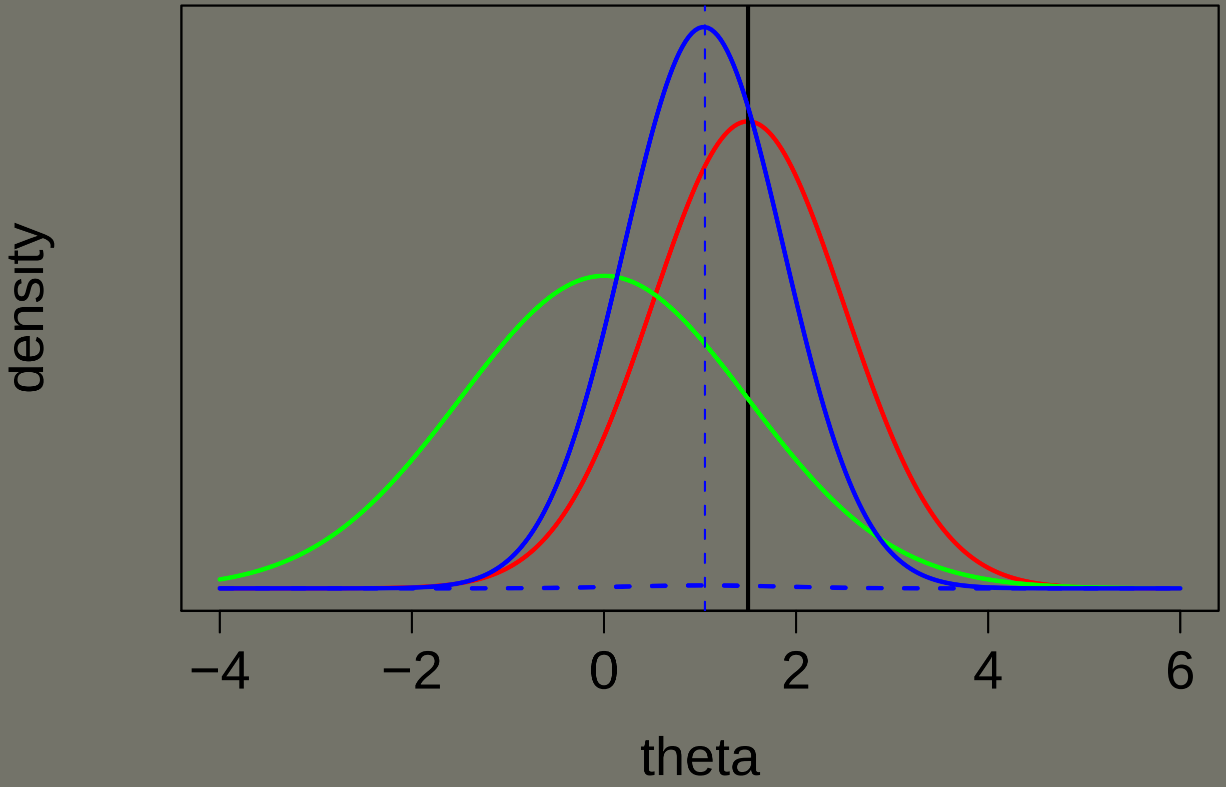
It is quite easy to consider priors where the algebra to find the posterior is impossible! Most real Bayes problems are solved numerically.

More on this topic and MCMC at the end this lecture.

Some posteriors for this example

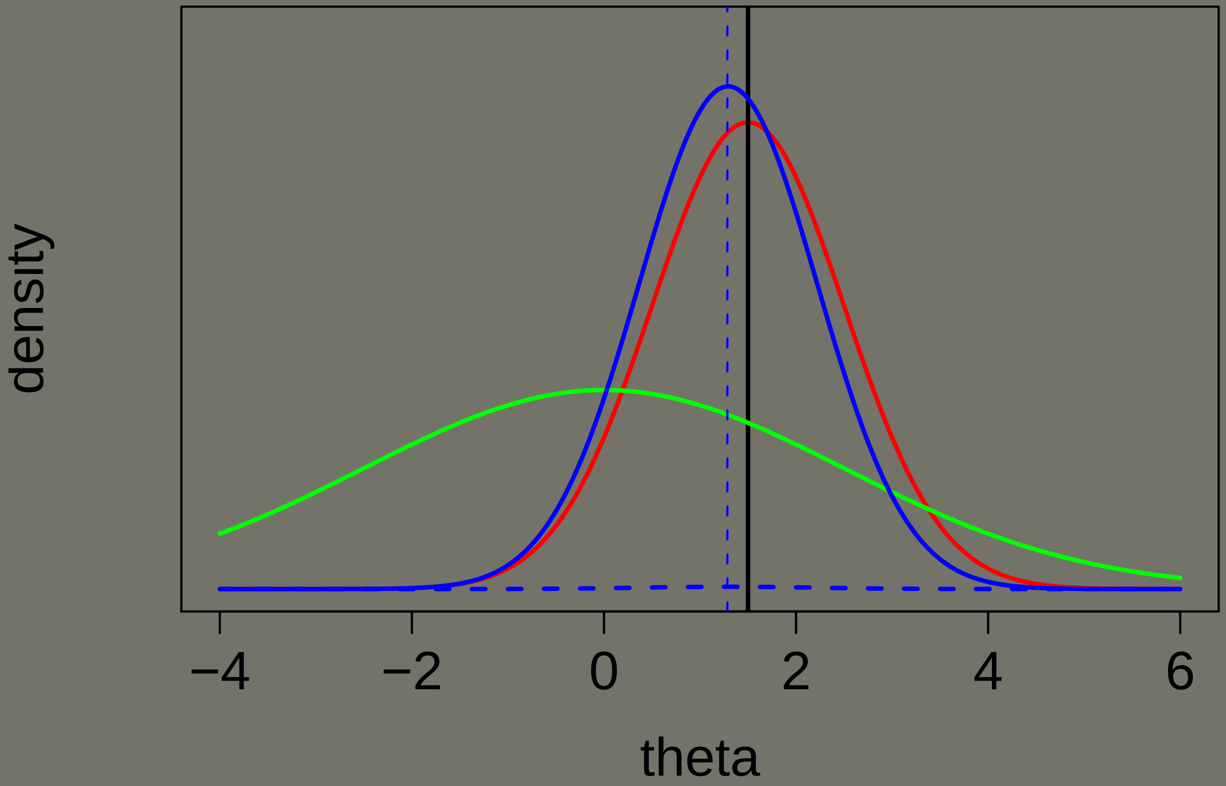
DATA $= 1.5$, **PRIOR** $N(0, (1.5)^2$

Likelihood, POSTERIOR



Prior not very informative

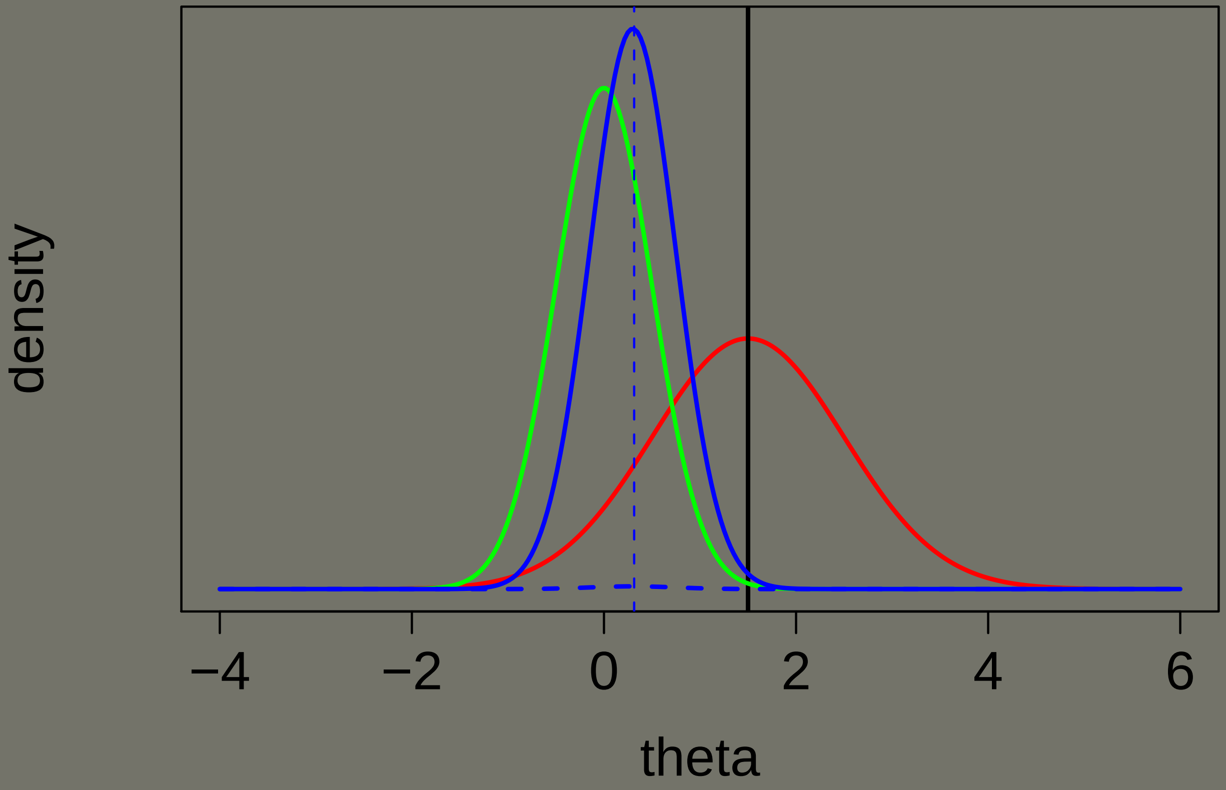
DATA **PRIOR** $N(0,(2.5)^2)$
Likelihood **POSTERIOR**



Prior is informative

DATA, PRIOR $N(0, (.5)^2)$

Likelihood POSTERIOR



-log posterior

$$\frac{(Y-\theta)^2}{2} + \frac{(\theta-\mu)^2}{2\sigma^2} + constant$$

$-$log likelihood $+$ **$-$log prior**

fit to data $+$ **control/constraints on parameter**

This is how the separate terms originate in a variational approach.

The Big Picture

It is useful to report the values where the posterior has its maximum.

This is called the posterior mode.

Variational DA techniques = finding posterior mode

Maximizing the posterior is the same as minimizing - log posterior.

Prior for the surface temperature problem

Use climatology!

The station data suggests that the temperature field is Gaussian:

$N(\mu, \Sigma)$ *and assume that* μ Σ *are known.*

We have cheated in Lecture 1 and estimated the mean field $\mu(x)$ and covariance function from the data.

There is actually some uncertainty in these choices.

Finding the posterior temperature field for a given year.

Posterior = Likelihood × Prior

Posterior $\propto N(HT, R) \times N(\mu, \Sigma)$

Jeff Anderson: *Products of Gaussian are Gaussian ...*

After some heavy lifting:

Posterior$= N(\hat{T}, P^a)$

$$\hat{T} = \mu + \Sigma H (H^T \Sigma H + R)^{-1} (Y - H\mu)$$
$$P^a = \Sigma - \Sigma H (H^T \Sigma H + R)^{-1} H^T \Sigma$$

These are the Kalman filter equations.

Another Big Picture Slide

Posterior = *Likelihood* × *Prior*

-log Posterior = *-log Likelihood* + *-log Prior*

For the temperature problem we have

-log Posterior =

$$(Y - HT)^T R^{-1}(Y - HT)/2 + (T - \mu)^T \Sigma^{-1}(T - \mu)/2$$

+ *other stuff.*

(Remember T is the free variable here.)

The CO_2 Problem

Prior

For the sources: $N(\mu_{u,k}, P_{u,k})$.
For the initial concentrations: $N(\mu_x, P_x)$

-log posterior

$$\sum_{i=0}^{(I-1)} \sum_{j=1}^{N} (z_j - h_j(x_i))^T R_j^{-1} (z_j - h_j(x_i))/2 \quad +$$

$$\sum_{k=1}^{K} (u_k - \mu_{u,k})^T P_{u,k}^{-1} (u_k - \mu_{u,k})/2 \quad +$$

$$(x_0 - \mu_x)^T P_x^{-1} (x_0 - \mu_x)/2 + constant$$

Some Comments

Dynamical constraint:

Given the time varying sources and initial concentrations all the subsequent concentrations are found by the dynamical model.

Due to linear properties of tracers

$$X = \Omega U$$

Ω: after you generate this matrix you will have used up your computing allocation!

Posterior:

Easy to write down but difficult to compute focus on finding where the posterior is maximized.

A future direction is to draw samples from the posterior.