A Bayesian view of Data Assimilation

JCSDA Summer School.
Andrew Lorenc, Stevenson WA. July 2009.

Thanks to Jim Purser who introduced me to the Bayesian approach.
Content

1. Bayes Theorem – adding information
   • Gaussian PDFs

2. Simplest possible Bayesian NWP analysis
   • Two gridpoints, one observation.

3. Issues in practical implementation
   • Modelling/representing prior background error covariances $B$
   • Solving the large matrix equation
   • Estimating $B$. 

4. Predicting the prior PDF
   • a Bayesian view of 4D-Var v Ensemble KF
Bayes Theorem – adding information

Gaussian PDFs
(Non-Gaussian observational errors - Quality Control will be covered in another lecture.)
Bayes' Theorem for Discrete Events

A \  B  
\text{events}

P(A)  
\text{probability of } A \text{ occurring, or knowledge about } A \text{'s past occurrence}

P(A \cap B)  
\text{probability that } A \text{ and } B \text{ both occur,}

P(A \mid B)  
\text{conditional probability of } A \text{ given } B

We have two ways of expressing \( P(A \cap B) \):

\[
P(A \cap B) = P(B) P(A \mid B) = P(A) P(B \mid A)
\]

\Rightarrow \text{Bayes' Theorem:}  
\[
P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}
\]

Can calculate \( P(B) \) from:  
\[
P(B) = P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A})
\]
Bayes theorem in continuous form, to estimate a value $x$ given an observation $y^o$:

$$p(x \mid y^o) = \frac{p(y^o \mid x)p(x)}{p(y^o)}$$

- $p(x \mid y^o)$ is the posterior distribution,
- $p(x)$ is the prior distribution,
- $p(y^o \mid x)$ is the likelihood function for $x$.

Can get $p(y^o)$ by integrating over all $x$:  
$$p(y^o) = \int p(y^o \mid x)p(x)dx$$
Assume Gaussian pdfs

Prior is Gaussian with mean $x^b$, variance $V_b$:

$$p(x) = (2\pi V_b)^{\frac{1}{2}} \exp \left( -\frac{1}{2} \frac{(x - x^b)^2}{V_b} \right)$$

Ob $y^o$, Gaussian about true value $x$ variance $V_o$:

$$p(y^o | x) = (2\pi V_o)^{\frac{1}{2}} \exp \left( -\frac{1}{2} \frac{(y^o - x)^2}{V_o} \right)$$

Substituting gives a Gaussian posterior:

$$p(x) = (2\pi V_a)^{\frac{1}{2}} \exp \left( -\frac{1}{2} \frac{(x - x^a)^2}{V_a} \right)$$
Advantages of Gaussian assumption

1. Best estimate is found by solving linear equations:

\[ \frac{1}{V_a} = \frac{1}{V_o} + \frac{1}{V_b} \quad \quad \quad \frac{1}{V_a} x^a = \frac{1}{V_o} y^o + \frac{1}{V_b} x^b \]

\[ p(x) = \left(2\pi V_a\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2} \frac{(x - x^a)^2}{V_a}\right) \]

Taking logs gives quadratic equation; differentiating to find extremum gives linear equation.

2. Best estimate is a function of values & [co-]variances only.

Often these are all we know.

3. Weights are independent of values.
Combination of Gaussian prior & observation
- Gaussian posterior,
- weights independent of values.

prior $x \sim N(0,3)$
likelihood $p(y_0|x) \sim N(3,1)$
posterior $x \sim N(2.25,0.75)$

prior $x \sim N(0,3)$
likelihood $p(y_0|x) \sim N(5,1)$
posterior $x \sim N(3.75,0.75)$

prior $x \sim N(0,3)$
likelihood $p(y_0|x) \sim N(7,1)$
posterior $x \sim N(5.25,0.75)$

prior $x \sim N(0,3)$
likelihood $p(y_0|x) \sim N(9,1)$
posterior $x \sim N(6.75,0.75)$
Variational Penalty Functions

• Finding the most probable posterior value involves maximising a product \([\text{of Gaussians}]\)

• By taking \(-\ln\) of the posterior PDF, we can instead minimise a sum \([\text{of quadratics}]\)

• This is often called the “Penalty Function” \(J\)

• Additive constants can be ignored
Penalty functions: \( J(x) = -\ln(p(x)) + c \)

\( p \) Gaussian \( \Rightarrow J \) quadratic

\[
J_b(x) = -\ln(p(x)) + c. \quad x \sim N(0,3) \\
J_o(x) = -\ln(p(yo|x)) + c. \quad p(yo|x) \sim N(3,1) \\
J(x) = J_b(x) + J_o(x)
\]

\[
J_b(x) = -\ln(p(x)) + c. \quad x \sim N(0,3) \\
J_o(x) = -\ln(p(yo|x)) + c. \quad p(yo|x) \sim N(5,1) \\
J(x) = J_b(x) + J_o(x)
\]

\[
J_b(x) = -\ln(p(x)) + c. \quad x \sim N(0,3) \\
J_o(x) = -\ln(p(yo|x)) + c. \quad p(yo|x) \sim N(7,1) \\
J(x) = J_b(x) + J_o(x)
\]

\[
J_b(x) = -\ln(p(x)) + c. \quad x \sim N(0,3) \\
J_o(x) = -\ln(p(yo|x)) + c. \quad p(yo|x) \sim N(9,1) \\
J(x) = J_b(x) + J_o(x)
\]
Simplest possible Bayesian NWP analysis
Simplest possible example – 2 grid-points, 1 observation. Standard notation:

\[
x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}
\]

Model is two grid points:

1 observed value \( y^o \) midway (but use notation for >1):

\[
y^o = (y^o)
\]

Can interpolate an estimate \( y \) of the observed value:

\[
y = H(x) = \frac{1}{2} x_1 + \frac{1}{2} x_2 = Hx = \left(\frac{1}{2} \right) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}
\]

This example \( H \) is linear, so we can use matrix notation for fields as well as increments.

We have prior estimate $x^b_1$ with error variance $V_b$:

$$p(x_1) = (2\pi V_b)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (x_1 - x^b_1)^2 / V_b\right)$$

$$p(x_2) = (2\pi V_b)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (x_2 - x^b_2)^2 / V_b\right)$$

But errors in $x_1$ and $x_2$ are usually correlated $\Rightarrow$ must use a multi-dimensional Gaussian:

$$x \sim N(x : x^b, B)$$

$$p(x_1 \cap x_2) = p(x) = (2\pi)^2 |B|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (x - x^b)^T B^{-1} (x - x^b)\right)$$

where $B$ is the covariance matrix:

$$B = V_b \begin{pmatrix} 1 & \mu \\ \mu & 1 \end{pmatrix}$$
Observational errors


**instrumental error**

\[
y^o \sim N(y^t, E)
\]

\[
p(y^o | y) = \left(2\pi |E|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (y^o - y)^T E^{-1} (y^o - y)\right)
\]

**error of representativeness**

\[
y \sim N(H(x^t), F)
\]

\[
p_t(y|x^t) = \left(2\pi |F|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (y - H(x^t))^T F^{-1} (y - H(x^t))\right)
\]

**Observational error**

combines these 2:

\[
y^o \sim N(H(x^t), E+F)
\]

\[
p(y^o | x^t) = \int p(y^o | y) p_t(y | x^t) \, dy
\]

\[
= \left(2\pi |E+F|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (y^o - H(x^t))^T (E+F)^{-1} (y^o - H(x^t))\right)
\]
background pdf
obs likelihood function
Bayesian analysis equation

\[
p(x|y^o) = \frac{p(y^o|x)p(x)}{p(y^o)}
\]

Property of Gaussians that, if \( H \) is linearisable: \( x \sim N(x^a, A) \)

where \( x^a \) and \( A \) are defined by:

\[
A^{-1} = B^{-1} + H^T (E + F)^{-1} H \\
x^a = x^b + AH^T (E + F)^{-1} (y^o - H(x^b))
\]
background pdf
obs likelihood function
posterior analysis PDF
For our simple example the algebra is easily done by hand, without manipulating matrices, giving:

\[
\mathbf{x}^a = \begin{pmatrix} x_1^a \\ x_2^a \end{pmatrix} = \begin{pmatrix} x_1^b \\ x_2^b \end{pmatrix} + \frac{\left(V^b \left(\frac{1+\mu}{2}\right)\right)^2}{E + F + V^b \left(\frac{1+\mu}{2}\right)} \begin{pmatrix} y^o - \frac{x_1^b + x_2^b}{2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}
\]
Practical implementation of the Bayesian Analysis Equation
Issues in practical implementation
“The devil is in the details”

There are significantly different choices possible for each of the following options. The combinations of these choices make up a very wide range and large number of analysis schemes, all implementing the same Bayesian equation!

• Modelling and representing prior background error covariances \( B \).
• Expressing the equations in a form amenable to solution.
• Computing the solution.
• Estimating \( B \).
Michael Ghil on OI & Kalman Filter


OPTIMAL INTERPOLATION AND THE KALMAN FILTER are similar, but the form is in the difference.

S. Cohn, M. Ghil* and E. Isaacson
Courant Institute of Mathematical Sciences
New York University, New York, New York 10012
Modelling and representing prior background error covariances B.

• Explicit point-point [multivariate] covariance functions.
• Transformed control variables to deal with inter-variable covariances.
• Vertical – horizontal split
  • EOF decomposition into modes.
  • Spectral decomposition into waves.
  • Wavelets.
• Recursive filters or diffusion operators to give local variations.
• Ensemble members.
Schlatter’s (1975) multivariate covariances

Fig. 3. Correlations among the variables $h$, $u$, and $v$ based upon the expression $\rho = 0.95 \exp(-1.24\sigma^2)$ for height-height correlation and the geostrophic relations. Diagrams centered at 110°W, 35°N, Tick marks 300 km apart.
Transformed control variable.

- Look for a “balanced” variable from which we can calculate balanced flow in all variables: streamfunction, PV.

- Define transforms from \((U)\) or to \((T)\) this variable and a residual variable, which by construction/hypothesis is uncorrelated making \(B\) block diagonal. (Compare EOFs)

- Transformed variables still need spatial covariance model, but not multivariate. (Further transforms may be used to represent these.)
Comparison of covariance models

Correction to $\theta$ based on a $u$ observation at same level

Correction to $p$ based on a $u$ observation at same level
Equations - all equivalent.

- Variational
  \[ x^a \text{ minimises } J(x) = \frac{1}{2} (x - x^b)^T B^{-1} (x - x^b) + \frac{1}{2} (y^o - H(x))^T R^{-1} (y^o - H(x)) \]
  \[ A^{-1} = \left( \frac{\partial^2 J}{\partial x^2} \right) = B^{-1} + H^T R^{-1} H \]

- Kalman Filter. Kalman Gain=K.
  \[ x^a = x^b + K(y^o - H(x^b)) \quad A = (I - KH)B \]

- Observation space
  \[ K = BH^T (HBH^T + R)^{-1} \]
  Demonstrate equivalence using Sherman–Morrison–Woodbury formula

- Model space
  \[ K = (H^T R^{-1} H + B^{-1})^{-1} H^T R^{-1} \]

- Ensemble space Square-root Filters, e.g. ETKF
  \[ B = Z^f (Z^f)^T \]
  \[ Z^a = Z^f T \]
  \[ A = Z^a (Z^a)^T \]
Estimating PDFs or covariances

- Even if we knew the “truth”, we could never run enough experiments in the lifetime of an NWP system to estimate its error PDF, or even its error covariance $\mathbf{B}$.

- Simplifying assumptions are essential (e.g. Gaussian, ...)

- Even a simplified error model has so many parameters that we cannot determine them by NWP trials to determine which give the best forecasts.

- In practice we can only measure innovations – cannot get separate estimates of $\mathbf{B}$ & $\mathbf{R}$ without assumptions (Talagrand).

- *Need to understand physics!*
How to estimate the prior PDF? How to calculate its time evolution?

i.e. 4D-Var versus Ensemble KF
Fokker-Planck Equation

Ensemble methods attempt to sample entire PDF.

chaotic growth + model error
increase spread
Gaussian Probability Distribution Functions

- Easier to fit to sampled errors.
- Quadratic optimisation problems, with linear solution methods – much more efficient.
- The Kalman filter is optimal for linear models, but
  - it is not affordable for expensive models (despite the “easy” quadratic problem)
  - it is not optimal for nonlinear models.
- Advanced methods based on the Kalman filter can be made affordable:
  - Ensemble Kalman filter (EnKF, ETKF, ...)
  - Four-dimensional variational assimilation (4D-Var)
Extended Kalman Filter

\[ x^a(t_i) = x^f(t_i) + P^f(t_i)H_i^T \left( H_iP^f(t_i)H_i^T + R_i \right)^{-1} \left( y_i - H_i(x^f(t_i)) \right) \]
\[ P^a(t_i) = P^f(t_i) - P^f(t_i)H_i^T \left( H_iP^f(t_i)H_i^T + R_i \right)^{-1} H_iP^f(t_i) \]

**Analysis step**

**Forecast step**

\[ x^f(t_{i+1}) = M_i \left( x^a(t_i) \right) \]

True discretised dynamics \( X^f \) assumed to differ by stochastic perturbations:

\[ x^f(t_{i+1}) = M_i \left( x^f(t_i) \right) + \eta(t_i) \]

where \( \eta \) is a noise process with zero mean and covariance matrix \( Q_i \).

\[ P^f(t_{i+1}) = M_i P^a(t_i) M_i^T + Q_i \]
Ensemble Kalman filter

Fit Gaussian to forecast ensemble.

chaotic growth + model error increase spread
The Ensemble Kalman Filter (EnKF)

Construct an ensemble \( \{x_i^f\}, (i = 1, \ldots, N) \) :

\[
P^f = P_e^f = \left( x^f - \bar{x}^f \right) \left( x^f - \bar{x}^f \right)^T,
\]

\[
P^f H^T = \left( x^f - \bar{x}^f \right) \left( H \left( x^f \right) - H \left( \bar{x}^f \right) \right)^T,
\]

\[
H P^f H^T = \left( H \left( x^f \right) - H \left( \bar{x}^f \right) \right) \left( H \left( x^f \right) - H \left( \bar{x}^f \right) \right)^T
\]

Use these in the standard KF equation to update the best estimate (ensemble mean):

\[
\bar{x}^a = \bar{x}^f + P^f H^T \left( HP^f H^T + R \right)^{-1} \left( y^o - H \left( \bar{x}^f \right) \right).
\]
Deterministic 4D-Var

Initial PDF is approximated by a Gaussian.

Descent algorithm only explores a small part of the PDF, on the way to a local minimum.
Simple 4D-Var, as a least-squares best fit of a deterministic model trajectory to observations

© Crown copyright Met Office
Assumptions in deriving deterministic 4D-Var

Bayes Theorem - posterior PDF:

\[ P(x|y^o) = \frac{P(y^o|x)P(x)}{P(y^o)} \]

where the obs likelihood function is given by:

\[ P(y^o|x) = f(y^o - y), \text{ where } y = H(x) \]

Impossible to evaluate the integrals necessary to find “best”.

Instead assume best x maximises PDF, and minimises -ln(PDF):

\[ J(x) = -\ln[P(y^o|x)] - \ln[P(x)] \]

Purser, R.J. 1984: "A new approach to the optimal assimilation of meteorological data by iterative Bayesian analysis". Preprints, 10th conference on weather forecasting and analysis. Am Met Soc. 102-105

The deterministic 4D-Var equations

\[ P(x | y^o) \propto P(x) P(y^o | x) \]

Bayesian posterior pdf.

\[ P(x) \propto \exp\left(-\frac{1}{2}(x - x^b)^T B^{-1}(x - x^b)\right) \]

Assume Gaussians

\[ P(y^o | x) = P(y^o | \bar{y}) \propto \exp\left(-\frac{1}{2}(y - y^o)^T R^{-1}(y - y^o)\right) \]

But nonlinear model makes pdf non-Gaussian: full pdf is too complicated to be allowed for.

\[ y = H(M(x)) \]

So seek mode of pdf by finding minimum of penalty function

\[ J(x) = \frac{1}{2}(x - x^b)^T B^{-1}(x - x^b) + \frac{1}{2}(y - y^o)^T R^{-1}(y - y^o) \]

\[ \nabla_x J(x) = B^{-1}(x - x^b) + M^*H^*R^{-1}(y - y^o) \]
Statistical, incremental 4D-Var

Statistical 4D-Var approximates entire PDF by a Gaussian.
Statistical 4D-Var - equations

Independent, Gaussian background and model errors ⇒ non-Gaussian pdf for general $y$:

Incremental linear approximations in forecasting model predictions of observed values converts this to an approximate Gaussian pdf:

The mean of this approximate pdf is identical to the mode, so it can be found by minimising:

\[
P(\delta x, \delta \eta | y^o) \propto \exp \left( -\frac{1}{2} \left( \delta x - (x^b - x^g) \right)^T B^{-1} \left( \delta x - (x^b - x^g) \right) \right. \\
\left. \exp \left( -\frac{1}{2} \left( \delta \eta + \eta^g \right)^T Q^{-1} \left( \delta \eta + \eta^g \right) \right) \right. \\
\left. \exp \left( -\frac{1}{2} \left( y - y^o \right)^T R^{-1} \left( y - y^o \right) \right) \right.
\]

\[
y = H_0 \delta x + \bar{H} \left( \bar{M} \left( x^g, \eta^g \right) \right)
\]

\[
J(\delta x, \delta \eta) = \frac{1}{2} \left( \delta x - (x^b - x^g) \right)^T B^{-1} \left( \delta x - (x^b - x^g) \right) + \frac{1}{2} \left( \delta \eta + \eta^g \right)^T Q^{-1} \left( \delta \eta + \eta^g \right) + \frac{1}{2} \left( y - y^o \right)^T R^{-1} \left( y - y^o \right)
\]
Questions and answers