NON-GAUSSIAN PROBABILITIES IN DATA ASSIMILATION AND QUALITY CONTROL DECISIONS

Andrew C Lorenc
Forecasting Research, Meteorological Office,
Bracknell, England.

Summary: The Bayesian justification for the penalty function used in variational methods is reviewed. The formalism is applied first to Gaussian distributions, then manageable approximations for handling non-Gaussian distributions are discussed. The Bayesian approach provides a sound basis both for variational analysis and the quality control of observations.
The normal variational approach, using a descent algorithm, is not necessarily robust if the possibility of gross observational errors is taken into account. Examples are presented where a satisfactory solution can be found, and where a variational descent algorithm fails to find the correct minimum. The accuracy of the first-guess is crucial.
The same Bayesian formalism can be applied to the decision taking algorithms used in traditional quality control methods. The relationships between non-Gaussian variational methods and the operational quality control algorithms at the Met Office and at ECMWF are discussed.

1. BAYESIAN DERIVATION OF ANALYSIS EQUATION

This derivation mainly follows Lorenc (1986).

1.1 Notation

$\mathbf{x}$ atmosphere as represented in model

$\mathbf{x}_t$ model representation of the true state of the atmosphere

$\mathbf{x}_b$ prior estimate of $\mathbf{x}_t$ (e.g. from forecast)

$\mathbf{y}$ observations

$\mathbf{y}_t$ observations that would be given by error-free instruments

$K(\mathbf{x})$ forward operator for calculating $\mathbf{y}$ from $\mathbf{x}$

K tangent linear operator of K,
such that $K(\mathbf{x}+\delta\mathbf{x})=K(\mathbf{x})+\mathrm{K}\delta\mathbf{x}+O(\delta\mathbf{x}^2)$.

P probability

p probability distribution function

$\mathrm{P}(\mathbf{x})$ = probability that $\mathbf{x}\le\mathbf{x}_t<\mathbf{x}+\mathbf{dx}$
= $p(\mathbf{x})\mathbf{dx}$

N.B. We use $\mathbf{x}$ both for the vector of values, and for the event $\mathbf{x}\le\mathbf{x}_t<\mathbf{x}+\mathbf{dx}$.

P(A|B) is the conditional probability of A, given B.

1.2 Probability Equations

Probabilities are used in a Bayesian way to describe the state of information. We have some prior information about **x**. We add to this information from observations **y**. We need to know the posterior knowledge about **x**. Operator K does not have a normal inverse.

From now on all probabilities are conditional on knowing $\mathbf{x}_b$. To simplify notation we write $P(\cdot)$ instead of $P(\cdot|\mathbf{x}_b)$.

The basis of the derivation is the identity:

$$\begin{aligned} P(\mathbf{x}\cap\mathbf{y}) &= P(\mathbf{x}|\mathbf{y}) \quad P(\mathbf{y}) &&= P(\mathbf{y}|\mathbf{x}) \quad P(\mathbf{x}) \\ &= p(\mathbf{x}|\mathbf{y})d\mathbf{x}\ p(\mathbf{y})d\mathbf{y} &&= p(\mathbf{y}|\mathbf{x})d\mathbf{y}\ p(\mathbf{x})d\mathbf{x} \end{aligned} \tag{1}$$

What we want an expression for is:

$P(\mathbf{x}|\mathbf{y})$ $=p(\mathbf{x}|\mathbf{y})d\mathbf{x}$, the analysis probability, i.e. the probability that $\mathbf{x}\le\mathbf{x}_t<\mathbf{x}+d\mathbf{x}$, given the background $\mathbf{x}_b$ and the observations **y**.

We assume we know certain distributions, based on our prior experience and our knowledge of the physics:

$P(\mathbf{x})$ $=p(\mathbf{x})d\mathbf{x}$, is the probability that $\mathbf{x}\le\mathbf{x}_t<\mathbf{x}+d\mathbf{x}$, given only the prior knowledge of $\mathbf{x}_b$.

$p(\mathbf{y}|\mathbf{y}_t\cap\mathbf{x})$ is the instrumental error distribution.

$p(\mathbf{y}_t|\mathbf{x})$ is the forward operator error distribution.

From the last two distributions, we can find:

$P(\mathbf{y}|\mathbf{x})$ $=$ $p(\mathbf{y}|\mathbf{x})d\mathbf{y}$, the probability of getting observations **y** given $\mathbf{x}=\mathbf{x}_t$.

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{y}_t\cap\mathbf{x})\ p(\mathbf{y}_t|\mathbf{x})\ d\mathbf{y}_t \tag{2}$$

From this, and our prior knowledge of **x**, we can find:

$P(\mathbf{y})$ $=p(\mathbf{y})d\mathbf{y}$, the probability of getting observations **y**.

$$\begin{aligned} p(\mathbf{y}) &= \int p(\mathbf{y}|\mathbf{x})\ p(\mathbf{x})\ d\mathbf{x} \\ &= \int\int p(\mathbf{y}|\mathbf{y}_t\cap\mathbf{x})\ p(\mathbf{y}_t|\mathbf{x})\ d\mathbf{y}_t\ p(\mathbf{x})\ d\mathbf{x} \end{aligned} \tag{3}$$

Bayes' Theorem, which follows from the basic identity (1), is:

$$p(\mathbf{x}|\mathbf{y}) = p(\mathbf{y}|\mathbf{x})\ p(\mathbf{x})\ /\ p(\mathbf{y}) \tag{4}$$

We can substitute the expressions derived above to give:

$$p(\mathbf{x}|\mathbf{y}) = \frac{\int p(\mathbf{y}|\mathbf{y}_t \cap \mathbf{x})\ p(\mathbf{y}_t|\mathbf{x})\ \mathbf{dy}_t\ p(\mathbf{x})}{\int\int p(\mathbf{y}|\mathbf{y}_t \cap \mathbf{x})\ p(\mathbf{y}_t|\mathbf{x})\ \mathbf{dy}_t\ p(\mathbf{x})\ \mathbf{dx}} \tag{5}$$

This p.d.f. describes our total posterior information about $\mathbf{x}$, given $\mathbf{x}_b$ and $\mathbf{y}$.

1.3 Solution Using Gaussian Probability Distributions

We assume K can be linearized in the region of $\mathbf{x}_b$ and $\mathbf{x}_a$ such that

$$K(\mathbf{x}_a) = K(\mathbf{x}_b) + \mathbf{K}\ (\mathbf{x}_a - \mathbf{x}_b), \tag{6}$$

We assume all the p.d.f.s are Gaussian, and use the notation

$$\mathrm{N}(\mathbf{x}|\mathbf{m},\mathbf{B}) = ((2\pi)^{N}|\mathbf{B}|)^{-1/2}\exp(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{T}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{m})) \tag{7}$$

where $\mathbf{B}$ is an N×N positive definite matrix, and $|\mathbf{B}|$ is its determinant.

We assume that we know:

the background error distribution:

$$p(\mathbf{x}) = \mathrm{N}(\mathbf{x}|\mathbf{x}_b,\mathbf{B}),$$

the instrumental error distribution:

$$p(\mathbf{y}|\mathbf{y}_t \cap \mathbf{x}) = \mathrm{N}(\mathbf{y}|\mathbf{y}_t,\mathbf{O}),$$

the forward operator error distribution:

$$p_f(\mathbf{y}_t|\mathbf{x}) = \mathrm{N}(\mathbf{y}_t|K(\mathbf{x}),\mathbf{F}),$$

where $\mathbf{B}$ $\mathbf{O}$ and $\mathbf{F}$ are covariances.

Then, using the properties of Gaussians, the observational error distribution is given by the convolution:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \int p(\mathbf{y}|\mathbf{y}_t \cap \mathbf{x})\ p(\mathbf{y}_t|\mathbf{x})\ \mathbf{dy}_t \\ &= \mathrm{N}(\mathbf{y}|K(\mathbf{x}_t),\mathbf{O}+\mathbf{F}) \end{aligned} \tag{8}$$

where $\mathbf{O}+\mathbf{F}$ (=$\mathbf{E}$) is the observational error covariance.

The observation distribution, only knowing $\mathbf{x}_b$, is given by:

$$p(\mathbf{y}) = \mathrm{N}(\mathbf{y}|K(\mathbf{x}_b),\mathbf{O}+\mathbf{F}+\mathbf{KBK}^{T}). \tag{9}$$

Substituting these into Bayes' Theorem (4) gives:

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &= \mathrm{N}(\mathbf{y}|K(\mathbf{x}_t),\mathrm{O+F})\ \mathrm{N}(\mathbf{x}|\mathbf{x}_b,\mathbf{B})\ /\ \mathrm{N}(\mathbf{y}|K(\mathbf{x}_b),\mathrm{O+F+KBK^T}) \\ &= \mathrm{N}(\mathbf{x}|\mathbf{x}_a,\mathbf{A}). \end{aligned} \tag{10}$$

where $\mathbf{x}_a$ and $\mathbf{A}$ are defined by

$$\begin{aligned} \mathbf{A} &= \mathbf{B} - \mathbf{BK^T(KBK^T+O+F)^{-1}KB} \\ \mathbf{x}_a &= \mathbf{x}_b + \mathbf{BK^T(KBK^T+O+F)^{-1}}(\mathbf{y}-K(\mathbf{x}_b)). \end{aligned} \tag{11}$$

It is normal to assume that the "best" estimate of $\mathbf{x}_t$ is given by the mean $\mathbf{x}_a$ of the Gaussian posterior distribution. Thus using the above equation we can calculate $\mathbf{x}_a$ directly. This is the "OI" equation.

1.4 Variational Solution

If K is more nonlinear, or the p.d.f.s are non-Gaussian, then the direct solution derived above cannot be used. Although the Bayes' Theorem (4) for the analysis p.d.f. is still valid, the expression for p which results is usually too complicated to be very useful in describing our knowledge about $\mathbf{x}$; we want an estimate of the "best" $\mathbf{x}$, without evaluating the full p.d.f.. First, to define "best", we define a *loss function* $L(\mathbf{x}_l,\mathbf{x})$ giving the cost to us of making an estimate $\mathbf{x}_l$ when the true value is $\mathbf{x}$. The expected loss R is

$$R(\mathbf{x}_l) = \int L(\mathbf{x}_l,\mathbf{x})\ p(\mathbf{x}|\mathbf{y})\ \mathbf{dx} \tag{12}$$

The best estimate is the $\mathbf{x}_l$ which minimizes this. In general this require evaluating all of $p(\mathbf{x}|\mathbf{y})$. This can be avoided by making L a negative delta function, so that there is a gain from getting exactly the $\mathbf{x}_t$. With this spike loss

$$L(\mathbf{x}_l,\mathbf{x}) = -\delta(\mathbf{x}_l-\mathbf{x}) \tag{13}$$

$$R(\mathbf{x}) = -p(\mathbf{x}|\mathbf{y}) \tag{14}$$

Substituting in the Bayesian expression for $p(\mathbf{x}|\mathbf{y})$, and since $p(\mathbf{y})$ is independent of $\mathbf{x}$, the $\mathbf{x}$ which minimizes $R(\mathbf{x})$ is the same as the $\mathbf{x}$ which minimizes a penalty function $\mathcal{J}$ given by

$$\mathcal{J} = -\ln(p(\mathbf{y}|\mathbf{x})) - \ln(p(\mathbf{x})). \tag{15}$$

If we substitute the Gaussian p.d.f.s of the last section into this, we get:

$$\mathcal{J} = \tfrac{1}{2}(\mathbf{y}-K(\mathbf{x}))^{\mathrm{T}}(\mathrm{O+F})^{-1}(\mathbf{y}-K(\mathbf{x})) + \tfrac{1}{2}(\mathbf{x}_b-\mathbf{x})^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}_b-\mathbf{x}) + \text{constant}. \tag{16}$$

If furthermore we make K linearizable, we see why the linear problem with Gaussians is easier to solve: $\mathcal{J}$ becomes a quadratic in $\mathbf{x}$. Using the same algebraic manipulations as are needed to establish the properties of Gaussians used in the last section, and the same definitions (11) of $\mathbf{x}_a$ and $\mathbf{A}$, gives:

$$\mathcal{J} = \frac{1}{2}(\mathbf{x}_a-\mathbf{x})^T\mathbf{A}^{-1}(\mathbf{x}_a-\mathbf{x}) + \text{constant}. \tag{17}$$

For large problems it is easier to find $\mathbf{x}_a$ iteratively, even if $\mathcal{J}$ is quadratic. If K cannot be linearized over the whole range containing $\mathbf{x}_b$ and possible $\mathbf{x}_a$s, then an explicit solution is not possible. If K is still differentiable, so that

$$K(\mathbf{x}+\delta\mathbf{x}) = K(\mathbf{x})+\mathbf{K}_\mathbf{x}\delta\mathbf{x}, \quad \text{as } \delta\mathbf{x}\rightarrow 0 \tag{18}$$

then we can look for the minimum of $\mathcal{J}$ using a descent algorithm. At the minimum, the gradient of $\mathcal{J}$ with respect to the components of $\mathbf{x}$ is zero:

$$\mathcal{J}' = - \mathbf{K}_\mathbf{x}{}^T(\mathbf{O}+\mathbf{F})^{-1}(\mathbf{y}-K(\mathbf{x})) - \mathbf{B}^{-1}(\mathbf{x}_b-\mathbf{x}) = 0. \tag{19}$$

This formula is exact; we can find the most probable $\mathbf{x}$. The next stage of generalization is to allow the p.d.f.s to be weakly non-Gaussian. That is, we use the Gaussian formulae with $\mathbf{O}_\mathbf{x}$, $\mathbf{F}_\mathbf{x}$ and $\mathbf{B}_\mathbf{x}$ being slowly varying functions of $\mathbf{x}$, whose derivatives we can neglect. We also neglect derivatives of $\mathbf{K}_\mathbf{x}$. Then if we define $\mathbf{x}_a$ as the $\mathbf{x}$ which minimizes $\mathcal{J}$, i.e.

$$\mathcal{J}' = - \mathbf{K}_{\mathbf{x}a}{}^T(\mathbf{O}_{\mathbf{x}a}+\mathbf{F}_{\mathbf{x}a})^{-1}(\mathbf{y}-K(\mathbf{x}_a)) - \mathbf{B}_{\mathbf{x}a}{}^{-1}(\mathbf{x}_b-\mathbf{x}_a) = 0. \tag{20}$$

Then

$$\mathcal{J}'' \cong \mathbf{K}_{\mathbf{x}a}{}^T(\mathbf{O}_{\mathbf{x}a}+\mathbf{F}_{\mathbf{x}a})^{-1}\mathbf{K}_{\mathbf{x}a} + \mathbf{B}_{\mathbf{x}a}{}^{-1} = \mathbf{A}^{-1}. \tag{21}$$

Then, in the neighbourhood of $\mathbf{x}_a$,

$$p_a(\mathbf{x}|\mathbf{y}) \propto \mathrm{N}(\mathbf{x}|\mathbf{x}_a,\mathcal{J}''^{-1}). \tag{22}$$

If K is sufficiently nonlinear, or the p.d.f.s are sufficiently non-Gaussian, $p_a(\mathbf{x}|\mathbf{y})$ may have multiple maxima. We have then to consider how to decide which is best. We can generalize on the spike loss, by allowing the loss function to be a Gaussian:

$$L(\mathbf{x}_l,\mathbf{x}) = - \mathrm{N}(\mathbf{x}_l|\mathbf{x},\mathbf{L}). \tag{23}$$

As L tends to zero this gives us the spike loss. For the Gaussian analysis problem we can evaluate the convolution explicitly:

$$R(\mathbf{x}_l) = - \mathrm{N}(\mathbf{x}_l|\mathbf{x}_a,\mathbf{A}+\mathbf{L}) \tag{24}$$

Thus the loss is minimum when $\mathbf{x}_l=\mathbf{x}_a$, as we would expect. We can use this expression to help us in deciding between peaks in a non-Gaussian posterior p.d.f., by assuming that the peaks can be approximated by a local Gaussian. We assume the spread of the entire posterior p.d.f. can be characterized by **S** (i.e. **S** describes the distance between peaks). If **L**>>**S** then the loss function is quadratic over the range of significant probabilities, and the best estimate is the mean of the full p.d.f. (which may fall between two peaks). But if **L**<<**S** then we may consider the peaks separately. Then if in the vicinity of the ith local maximum the p.d.f. is

$$p(\mathbf{x}|\mathbf{y}) \cong \mathrm{P}_i\ \mathrm{N}(\mathbf{x}|\mathbf{x}_i,\mathbf{A}_i) \tag{25}$$

Then the loss associated with choosing the analysis to be at this maximum of $p(\mathbf{x}|\mathbf{y})$ is given by

$$R(\mathbf{x}_i) = -\mathrm{P}_i\ \mathrm{N}(\mathbf{x}_l|\mathbf{x}_i,\mathbf{A}_i\mathrm{+L}) \tag{26}$$

If $\mathbf{S}>>\mathbf{A}_i>>\mathbf{L}$ then $R(\mathbf{x}_i) \cong -p(\mathbf{x}_i|\mathbf{y})$, and the best peak is the highest.
If $\mathbf{A}_i<<\mathbf{L}<<\mathbf{S}$ then $R(\mathbf{x}_i) \cong -\mathrm{P}_i \times$ constant, and the best peak is that with the largest area.

2. VARIATIONAL METHODS WITH NON-GAUSSIAN OBSERVATIONAL ERRORS

Lorenc and Hammon (1988) introduced a simple model of observational errors: They are uncorrelated, so that each observation can be considered separately. For each, either the observation is good, in which case its error comes from a Gaussian, or it has a gross error, in which all observed values over a range of plausible values are equally likely. Thus we have (for "plausible" y)

$$p(\mathrm{y}|\mathbf{x}) = (1-\mathrm{P(G)})\ \mathrm{N}(\mathrm{y}|K(\mathbf{x}),\mathrm{E}) + \mathrm{P(G)\ k} \tag{27}$$

where E is the observational error variance (=O+F), P(G) is the probability of gross error, **x** is the true value, and k is given by

$$\int_{\substack{\text{plausible}\\\text{values}}} \mathrm{k\ dy} = 1 \tag{28}$$

It is instructive to look at some simple posterior p.d.f.s resulting from this model, before going on to the full multivariate analysis problem. The simplest case is for a single observation of one parameter, and a prior (background) estimate $\mathrm{y_b}$ $(=K(\mathrm{x_b}))$ from a Gaussian distribution. Because P(y|x) is non-Gaussian, the shape of the posterior p.d.f. depends on the difference between y and $\mathrm{y_b}$, as illustrated in figure 1. Even in this, the simplest case, there are multiple maxima, and there are configurations in which a variational search, starting from the prior estimate y_b, will not find the best value.

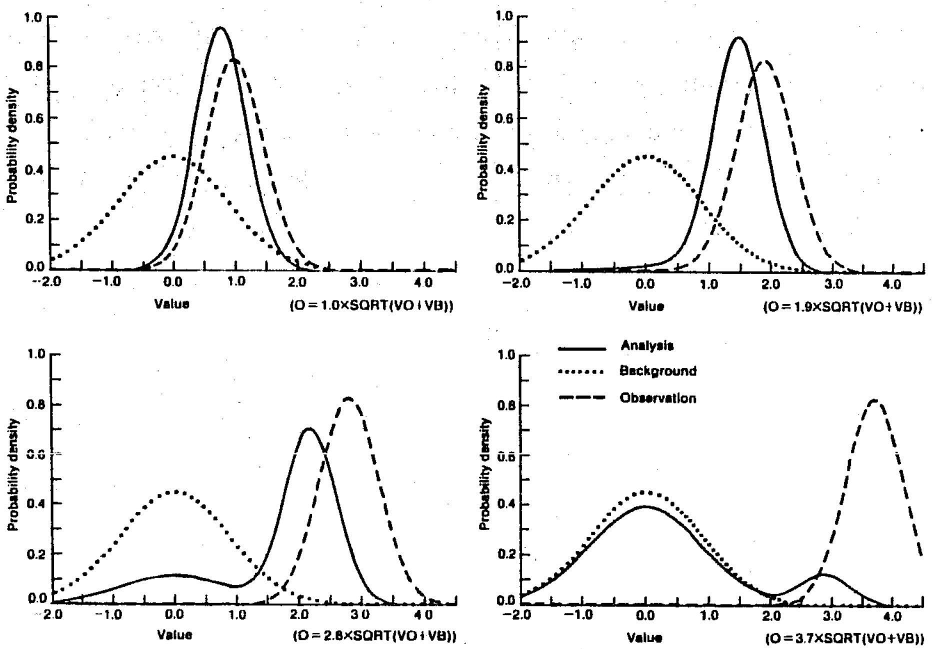

Fig 1. p.d.f.s for an observation, background, and Bayesian analysis, for a selection of observation-background differences o. p.d.f.s are appropriate for ship observations of surface pressure, with P(G)=0·5. (Lorenc and Hammon 1988).

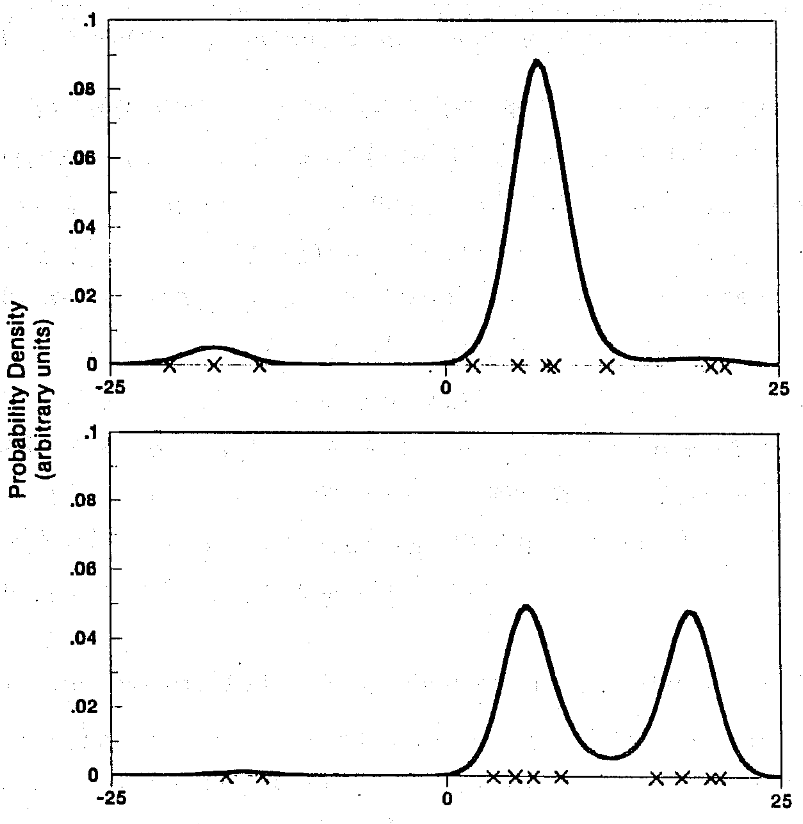

Fig 2. Two examples of p.d.f.s from simulated doppler wind observations, with x_t=7, good observations having E=9 and P(G)=0·5. (Dharssi *et al.* 1992).

Figure 2 shows a similar error model applied to two realizations, each of ten observations, from an idealized doppler observing system. With poor signal to noise ratio, P(G) may be large for such an instrument; we have used P(G)=0·5. In the lower example, it is not clear which is the "best" estimate; no method can consistently find it. In the top example there are multiple maxima, which become more obvious minima if we convert to a ln(p) penalty function $\mathcal{J}$, so even in this case a descent algorithm must start near the correct value.

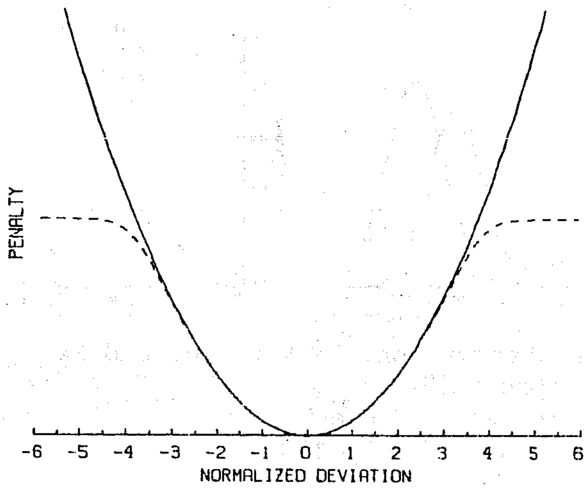

Fig 3. Solid line: quadratic penalty function for a single observation, Dashed line: penalty function assuming a P(G)=0·05. (Lorenc 1988)

Lorenc (1988) used an observational error distribution like (27) in a variational analysis based on minimizing (15). The possibility of gross errors converts the quadratic penalty function of (16) into one with plateaus (Figure 3). If the current estimate in an iterative algorithm is on one of these, the gradient does not well indicate which way to adjust towards the minimum. Note that the width of the minimum depends on E, while the spread of the deviations between initial estimate and observations depends on B+E. So if B is large, the iteration may not move towards the absolute minimum. This was the case in the experiments of Lorenc (1988). He tried various methods to improve the first-guess of the iteration, for instance by first setting P(G)=0, but with limited success.

Dharssi *et al.* (1992) had more success in their examples. In simple single value problems like those shown in figure 2, they found that increasing the observational error E in early iterations helped the iterative estimate move towards the best value. In a two-dimensional simulation of winds from a scanning lidar, they found that for relatively dense but unreliable

(P(G)=0·5) observations, the iteration did converge. It is an open question whether descent algorithms, suitably modified in early iterations, will be sufficient for practical applications, or whether we will still need the decision algorithms described in the next section.

3. Quality Control

3.1 Posterior probability of gross error

The posterior p.d.f.s shown in figure 1 are each the sum of two Gaussians, one corresponding to there being a gross error (G), one corresponding to the observation being correct ($\bar{G}$). Lorenc and Hammon (1988) proposed applying Bayes' theorem directly to the gross error event G:

$$\begin{aligned} P(G|y) &= p(y|G)\ P(G)\ /\ p(y) \\ &= p(y|G)\ P(G)\ /\ (p(y|G)\ P(G) + p(y|\bar{G})\ P(\bar{G})) \end{aligned} \tag{29}$$

Using (2) (27) and (28), we have

$$p(y|G) = k \tag{30}$$

Using (2) (27) and (9), we have

$$p(y|\bar{G}) = N(y|K(\mathbf{x}_b),\mathbf{E+KBK}^T) \tag{31}$$

so (29) can be readily evaluated. The two Gaussians in figure 1 are weighted by P(G|y) and P($\bar{G}$|y) respectively. Thus accepting or rejecting an observation depending on whether P(G|y) is greater than or less than 0·5 is consistent with the "best" analysis in terms of a Gaussian loss function, as discussed in relation to (26), as long as S>>L>>A. This is the basis of the decision taking algorithms used in Bayesian quality control schemes.

Dharssi *et al.* (1992) pointed out an interesting relationship between the variational method and the posterior probability of gross error. If we calculate $\mathcal{J}'$ using the error model (27), then we get:

$$\mathcal{J}' = -\,\mathbf{K}_{\mathbf{x}}^{\,T}(\mathbf{E}_{\mathbf{x}})^{-1}(\mathbf{y}-K(\mathbf{x})) - \mathbf{B}^{-1}(\mathbf{x}_b-\mathbf{x}) = 0. \tag{32}$$

where the diagonal element of $\mathbf{E}_{\mathbf{x}}$, for observation i, is given by

$$(\mathbf{E}_{\mathbf{x}})_{ii} = E_i\ /\ P(\bar{G}_i|\mathbf{x}). \tag{33}$$

E_i is the observational error variance of observation i if it does not have a gross error, and $P(\bar{G}_i|\mathbf{x})$ is the posterior probability that it does not have a gross error, given that $\mathbf{x}=\mathbf{x}_t$. We are effectively increasing the assumed error variance of observations which are unlikely to be correct. (This is not the same as the artificial increase discussed in section 2, where E_i is increased when calculating $E_i/P(\bar{G}_i|\mathbf{x})$ in early iterations, to aid convergence towards the global minimum). (32) has the same form as (19), so by using (33) each iteration, a variational method for Gaussian errors is converted to one for non-Gaussian errors.

At convergence, there will exist a final estimate of $P(\bar{G}_i|\mathbf{x})$ for each observation. It can be considered to be a Variational Quality Control (VQC) decision about the observations quality.

3.2 Individual Quality Control (IQC)

(29) can be extended to consider more than one observation. Lorenc and Hammon (1988) give the derivation for two observations:

$$P(G_1|\mathbf{y}) = P(G_1|y_1) \;/\; (p(\mathbf{y})/p(y_1)p(y_2)) \tag{34}$$

$$p(\mathbf{y})/p(y_1)p(y_2) = 1-P(\bar{G}_1|y_1)P(\bar{G}_2|y_2)\{1-p(\mathbf{y}|\bar{G}_1\cap\bar{G}_2)/(p(y_1|\bar{G}_1)p(y_2|\bar{G}_2))\} \tag{35}$$

Ingleby and Lorenc (1992) give a more general derivation. The number of terms to be considered in the extended equation goes as 2^n, where n is the number of observations, so evaluation of the exact equation rapidly becomes impractical. Lorenc and Hammon (1988) suggest sequential application of the "buddy check" equation for two observations as an approximation. This is the method used operationally at the Met Office. The decision about whether to use each observation i is made individually, based on an approximation to its posterior probability of gross error $P(G_i|\mathbf{y})$. The analysis is then made using the accepted observations, assuming they have Gaussian errors.

3.3 Simultaneous Quality Control (SQC)

The 2^n terms in the full expression for $P(G_i|\mathbf{y})$ come from the various combinations of accepted and rejected observations. Each combination C_α is associated with a multivariate normal distribution, each individually calculated using (10), so that the total p.d.f. is given by (Ingleby and Lorenc 1992):

$$p(\mathbf{x}|\mathbf{y}) = \sum_{\alpha=0}^{2^n-1} p(\mathbf{x}|\mathbf{y}\cap C_\alpha)\ P(C_\alpha|\mathbf{y}). \tag{36}$$

The posterior probability for each combination of gross errors can be found using Bayes' theorem:

$$P(C_\alpha|\mathbf{y}) = p(\mathbf{y}|C_\alpha)\ P(C_\alpha)\ /\ p(\mathbf{y}). \tag{37}$$

If we assume that each of the Gaussians which makes a significant contribution to (36) has a distinct peak, then we can apply (26) to decide which gives the best estimate of **x**. If **S>>L>>A** it is the one with the maximum $P(C_\alpha|\mathbf{y})$.

Evaluating all 2^n probabilities is impossible for large n. Since $p(\mathbf{y})$ is the same for each C_α, we can instead search for the combination with the maximum $p(\mathbf{y}|C_\alpha)\ P(C_\alpha)$. The states C_α correspond to the vertices of an n-dimensional hypercube. One possible algorithm for searching only a small subset of possible combinations is related to the SIMPLEX algorithm in integer linear programming. We start with an estimate of the best, and then search to see if any of its neighbours is more likely. Moving from one C_α to a neighbour corresponds to changing the quality control decision on one observation, while keeping those on other observations the same. If one of the neighbouring combinations is more likely, we can then search its neighbours, and so on. This is the basis of the OI quality control algorithm of Lorenc (1981), which is used at ECMWF. Rather like the variational descent algorithms, this search algorithm relies on having a good first guess of the best C_α, since there will in general be multiple local maxima.

4. Comparison of quality control criteria

Figure 4 shows an example chosen to illustrate the differences between the approaches. The solid line shows the posterior p.d.f. given by (36), while the dotted lines are the constituent Gaussians. Variational analysis, using a spike loss function, will pick the highest peak (VAN). Note however that a simple descent algorithm would have to start quite close to $\mathbf{x}_{\mathrm{VAN}}$ if it is to converge to the correct value; starting from $\mathbf{x}_b$ will not do.

Assuming this $\mathbf{x}_{VAN}$ is correct, all the observations have $P(\bar{G}_i|\mathbf{x}_{VAN})>0{\cdot}5$, so if we were to use this as an acceptance criterion, and do a Gaussian analysis using the observations, we would get the value corresponding to the peak VQC.

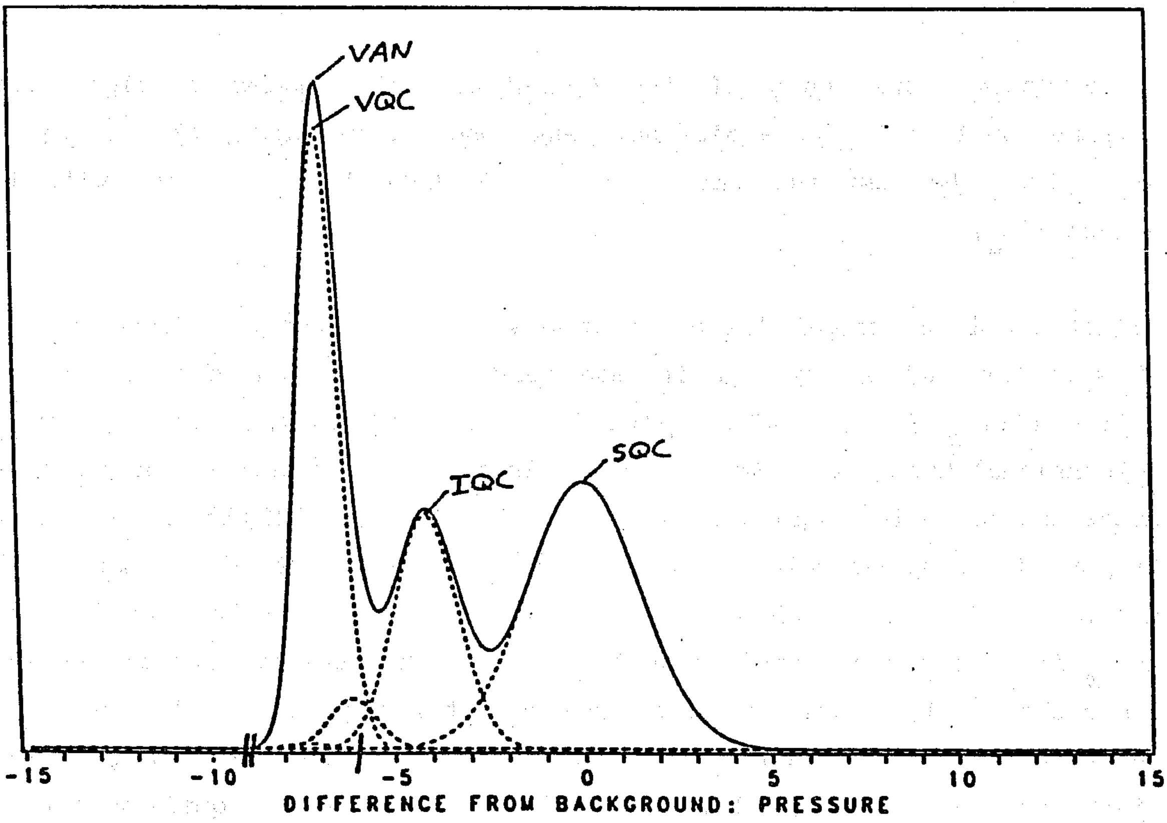

Fig 4 Solid curve: $P(x|y)$, dashed curves: $P(x|y\cap C_\alpha)$, for y_i=-9, -9 and -6, x_b=0, and other values appropriate for sea-level-pressure observations (E=1, B=2·25, k=0·043, and P(G)=0·04), from Ingleby and Lorenc (1992). For meaning of annotations, see text.

Calculating the $P(G_i|y)$ for each observation (IQC), the two observations of -9 both have posterior probabilities less than 0·5 (i.e. they fail) while the observation of -6 just passes. This pass is in part due to contributions from the possibility that the other observations were actually correct; IQC can give inconsistent decisions.

Simultaneous Quality Control does look for a consistent decision; in this case the Gaussian with the largest area is that labelled SQC. It corresponds to rejection of all the observations, i.e. it is the background distribution. Note however that the SIMPLEX algorithm will not work well in this case. There is one local maximum for the combinations accepting both observations of -9, and another for the combinations rejecting them both. The SIMPLEX algorithm will converge to one of these; it cannot get from one to the other because intermediate combinations (accepting one and rejecting the other) are less likely.

5. Concluding remarks

We have shown that the Bayesian approach provides a sound justification for the quadratic penalty function used in variational analysis, if error distributions are Gaussian. It also indicates how the method can be extended to observations with non-Gaussian distributions.

The proper "best" analysis depends on an appropriately defined loss function. Finding it requires convolutions over the posterior probability density function, which for non-Gaussian distributions is impractical. Variational analysis (VAN and VQC), and quality control algorithms (IQC and SQC) are making approximations to the ideal loss function. In NWP, we have a background $\mathbf{x}_b$ which usually would lead to a forecast that is not too bad. Large improvements on this accuracy are not required, so $\mathbf{L} \simeq \mathbf{B}$. Individual peaks in the p.d.f. have $\mathbf{A}_i < \mathbf{B}$. So the assumption that $\mathbf{S} >> \mathbf{L} >> \mathbf{A}_i$ may not be too bad for NWP assimilation.

There have also to be approximations in implementation; none of the methods can be implemented perfectly in practical NWP problems. In the approximate forms discussed here:

VAN and VQC use a descent algorithm, with a modified penalty function in early iterations to try to get convergence to the best $\mathbf{x}$ from as wide a range as possible of first-guesses. This has been tried on simulated data by Dharssi *et al.* (1992) and is an attractive candidate for future variational NWP assimilation systems.

IQC, as used at the Met Office (Lorenc and Hammon 1988), uses a sequential pairwise buddy check to approximate the method for >2 close observations. Some tuning of this has been found to be necessary.

SQC, with a SIMPLEX search, does not necessarily correctly handle close observations which agree with each other, but might both be wrong. The method used at ECMWF (Lorenc 1981) is similar to this (although the rejection tolerances are set directly, rather than via P(G)).

The Bayesian approach has allowed us to understand the relationship between these different methods.

Acknowledgements

I am grateful to Imtiaz Dharssi and Bruce Ingleby for their contributions to our joint papers quoted here, and to Jim Purser for initiating me in Bayesian analysis and quality control theory.

References

Dharssi,I., Lorenc,A.C. and Ingleby,N.B. 1992
"Treatment of gross errors using maximum probability theory"
Quart.J.Roy.Met.Soc., **118**, 1017-1036

Ingleby, N.B., and Lorenc, A.C. 1992
"Bayesian quality control using multivariate normal distributions".
MetO(S) Sci.Paper No.10. submitted to QJRMS

Lorenc, A.C. 1981
"A global three-dimensional multivariate statistical analysis scheme."
Mon. Wea. Rev. **109**, 701-721.

Lorenc, A.C. 1986
"Analysis methods for numerical weather prediction." Quart. J. Roy. Met. Soc. **112**, 1177-1194.

Lorenc, A.C. and Hammon, O., 1988
"Objective quality control of observations using Bayesian methods - Theory, and a practical implementation." Quart. J. Roy. Met. Soc. **114**, 515-543

Lorenc, A.C. 1988
"Optimal nonlinear Objective Analysis." Quart. J. Roy. Met. Soc. **114**, 205-240.