Forecasting Research

Forecasting Research Division
Technical Report No. 117

Bayes' Theorem in assimilation and quality control

by

Andrew Lorenc

September 1994

Meteorological Office
London Road
Bracknell
Berkshire
RG12 2SZ
United Kingdom

Forecasting Research Division
Technical Report No. 117

Bayes' Theorem in assimilation and quality control

Andrew Lorenc[1]

September 1994

Forecasting Research
Meteorological Office
London Road
Bracknell
RG12 2SZ

This paper was prepared for the Sixth BMRC Modelling Workshop "Data Assimilation in Meteorology and Oceanography" BMRC, Melbourne, 5-7 October 1994
It is based on lectures given at the International Summer School on *Assimilation of Meteorological and Oceanographic Observations*. La Garde, France August 1993

ABSTRACT

Probabilities can be used to quantify the extent of our knowledge about past events. Bayes' Theorem tells us how to combine these probabilities. It enables us to calculate weights for the information sources, as a function of their error distributions.
The statistical methods used to combine observations, such as the "OI" method used in meteorology, were derived as minimum variance best estimates. Variational analysis methods, like those solved using adjoints in some recent schemes, minimise a quadratic measure of the deviations from the data. With the aid of very simple examples, it is explained how both these methods can be derived from Bayes' Theorem, assuming Gaussian probability distributions.
It is usual to reject observations which differ too far from our prior estimate of what they should be. It is shown for a simple example how this is consistent with a long-tailed (non-Gaussian) observational error probability distribution. Even for the simplest non-Gaussian error models, full Bayesian analysis is impractical for the number of observations used in NWP analysis. Three different practical approaches are discussed.

---

[1] aclorenc@email.meto.govt.uk

Bayesian Probabilities

Nothing is certain in life, especially in weather forecasting. Forecasters are used to using probabilities to express this. "There is a 25% chance of rain tomorrow" does not necessarily imply that the atmosphere is random, but rather that the forecaster is uncertain. The use of probabilities to quantify the extent of our knowledge, about future or past events, is the key to the Bayesian approach.

Discrete Events

Suppose we have discrete events A and B, then we can use $P(A)$ and $P(B)$ to describe the probability of them occurring in the future, or the extent of our knowledge about them, if they have occurred in the past. Similarly we use $P(A \cap B)$ to denote the probability that A and B both occur, and $P(A|B)$ to denote the conditional probability of A given that B has occurred.

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

$$\begin{aligned} P(A \cap B) &= P(B)\ P(A|B) \\ &= P(A)\ P(B|A) \end{aligned} \qquad \textbf{(1)}$$

These lead directly to Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)} \qquad \textbf{(2)}$$

When we apply this, we often use relative probabilities (i.e. we do not bother with the factor $P(B)$), or else we calculate $P(B)$ from:

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) \qquad \textbf{(3)}$$

Example: Bayesian Dice

I have two dice. One is weighted towards throwing sixes. I have performed some experiments with them, and have the prior statistics that:

for the weighted (W) die, $P(6|W) = 58/60$

for the good (G) die, $P(6|G) = 10/60$

I choose one at random: $P(W) = P(G) = 1/2$

I throw this die, and it shows a six. Now:-

$$\begin{aligned} P(6) &= P(6|W)\ P(W) + P(6|G)\ P(G) \\ &= 58/60\ \ 1/2\ \ + 10/60\ \ 1/2 \\ &= 34/60 \end{aligned}$$

We can now apply Bayes' Theorem:

$$P(G|6) = P(6|G)\ P(G)\ /\ P(6)$$
$$= 10/60\ \ 1/2\ \ /\ 34/60\ \ =\ \ 5/34$$

$$P(W|6) = P(6|W)\ P(W)\ /\ P(6)$$
$$= 58/60\ \ 1/2\ \ /\ 34/60\ \ =\ 29/34$$

The information that I have thrown a six has added to my knowledge, so that the posterior probability that the chosen die is weighted has increased. If I were to throw again, and get another six, the probability would increase again.

Continuous Variables

We can go from discrete events to continuous variables by defining events such as:
X: the true value x_t is such that $x \le x_t < x+\delta x$

Then $$P(\mathrm{X}) = p(x)\ \delta x$$

Taking the limit $\delta x \to 0$, $p(x)$ is a probability density function (pdf).

Zero-dimensional Bayesian Analysis
A Gaussian pdf about a prior value x_b with background error variance V_b is:

$$p(x) = N(x|x_b,V_b) = (2\pi\ V_b)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\frac{(x-x_b)^2}{V_b}\right) \qquad \textbf{(4)}$$

where $N(x|m,V)$ denoted a normal distribution with mean m and variance V.

$P(y_o|x) = p(y_o|x)\ dy_o$, is the probability of getting an observation y_o, given that the true value is x.[2] For example a Gaussian pdf with observational error variance V_o is:

$$p(y_o|x) = N(y_o|x,V_o) = (2\pi\ V_o)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\frac{(y-x)^2}{V_o}\right) \qquad \textbf{(5)}$$

We can get $p(y_o)$ by integrating over all x:

$$p(y_o) = \int p(y_o|x)p(x)dx \qquad \textbf{(6)}$$

Since the convolution of two Gaussians is another Gaussian, for our examples this gives:

---

[2] Note that all pdfs are conditional on knowing the prior value x_b. To simplify, we do not show this explicitly in the notation.

$$p(y_o) = N(y_o|x_b, V_o+V_b) \tag{7}$$

Bayes' Theorem in continuous form is:

$$p(x|y_o) = \frac{p(y_o|x)p(x)}{p(y_o)} \tag{8}$$

$p(x|y_o)$ is called the posterior distribution, $p(x)$ the prior distribution, and $p(y_o|x)$ is the likelihood function for x [3].

Substituting the pdfs from our example, and noting that the product of two Gaussians is another Gaussian, gives:

$$p(x|y_o) = N(x|x_a, V_a) \tag{9}$$

where

$$\begin{aligned} \frac{x_a}{V_a} &= \frac{y_o}{V_o} + \frac{x_b}{V_b} \\ \frac{1}{V_a} &= \frac{1}{V_o} + \frac{1}{V_b} \end{aligned} \tag{10}$$

This is the standard formula for the combination of observations, studied by Gauss (1823). It is illustrated in figure 1. Note that, due to the unique properties of Gaussian distributions, the posterior distribution has the same shape for all values of y and x_b.

One-dimensional Bayesian Analysis

We will consider the simplest possible example that contains the essentials of problem, to illustrate the more abstract approach of Lorenc (1986). Our model consists of two grid points, at which the values are x_1 and x_2. For ease of manipulation we combine these into a vector $\boldsymbol{x}$:

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{11}$$

[3] It does not integrate to one over x, so it is not a probability.

We have one observed value y_o midway between points 1 and 2. So we can interpolate an estimate y of the observed value:

$$\begin{aligned} y &= K(\boldsymbol{x}) = \tfrac{1}{2}x_1 + \tfrac{1}{2}x_2 \\ &= \mathbf{K}\,\boldsymbol{x} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \end{aligned} \tag{12}$$

We have a prior estimate x_{b1} for the value at point 1. The probability that the true value x_{t1} [4] lies between x_1 and $x_1+\delta x_1$ is $p(x_1)\delta x_1$, where

$$p(x_1) = N(x|x_{b1},V_b) = (2\pi\ V_b)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(x_1-x_{b1})^2}{V_b} \right) \tag{13}$$

and similarly for x_2. But the errors in x_1 and x_2 are correlated:

$$\left\langle (x_{b1}-x_{t1})(x_{b2}-x_{t2}) \right\rangle = \mu\ V_b \tag{14}$$

So we cannot get the joint probability by multiplying $p(x_1)\times p(x_2)$. Instead we must use a multi-dimensional Gaussian:

$$\begin{aligned} p(x_1 \cap x_2) = p(\boldsymbol{x}) &= N(\boldsymbol{x}|\boldsymbol{x}_b,\mathbf{B}) \\ &= \left((2\pi\ )^2|\mathbf{B}|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_b)^T\mathbf{B}^{-1}(\boldsymbol{x}-\boldsymbol{x}_b) \right) \end{aligned} \tag{15}$$

where **B** is the covariance matrix:

$$\mathbf{B} = V_b \begin{pmatrix} 1 & \mu \\ \mu & 1 \end{pmatrix} \tag{16}$$

Since our example $\boldsymbol{x}$ only has two elements, we can plot $p(\boldsymbol{x})$, as shown in figure 2.

---

[4] When we set up our grid-point representation, we have to define how the real world is projected onto it. So $\boldsymbol{x}_t$ is the projection onto our grid of the real truth.

The instrumental error is described by the probability that the observed value lies between y_o and $y_o+\delta y_o$, given the true value y_t [5]:

$$p(y_o|y_t) = N(y_o|y_t,V_o) = (2\pi\ V_o)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\frac{(y_o-y_t)^2}{V_o}\right) \qquad (17)$$

Because we have only a finite representation of reality, in our example only two values, knowing $\boldsymbol{x}_t$ does not give us precise knowledge of y_t. This error of representativeness has the pdf:

$$p_t(y_o|\boldsymbol{x}_t) = N(y_o|K(\boldsymbol{x}_t),V_f) = (2\pi\ V_f)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\frac{(y_o-K(\boldsymbol{x}_t))^2}{V_f}\right) \qquad (18)$$

Note that the error of representativeness is a function of the resolution of our model $\boldsymbol{x}$. Figure 3 illustrates this for wind.

The observational error is the sum of these two effects. Its pdf is obtained by integrating over all y which might be y_t:

$$\begin{aligned} p(y_o|\boldsymbol{x}_t) &= \int p(y_o|y)p_t(y|\boldsymbol{x}_t)dy \\ &= N(y_o|K(\boldsymbol{x}_t),V_o+V_f) \\ &= N(y_o|K(\boldsymbol{x}_t),\mathbf{E}) \\ &= (2\pi\ |\mathbf{E}|)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(y_o-K(\boldsymbol{x}_t)^T\mathbf{E}^{-1}(y_o-K(\boldsymbol{x}_t))\right) \end{aligned} \qquad (19)$$

where we have replaced V_o+V_f by a one-by-one matrix $\mathbf{E}$ so that the equations below can be easily generalised to more than one observation. We can plot $p(y_o|\boldsymbol{x})$ as a function of x_1 and x_2 (figure 4).

Substituting into the Bayesian analysis equation, we have to multiply the functions shown in

[5] We define y_t to be what would be observed by a perfect instrument, with the same observing footprint as the real one.

figures 2 and 4:

$$p(\boldsymbol{x}|y_o) = \frac{p(y_o|\boldsymbol{x})p(\boldsymbol{x})}{p(y_o)}$$
$$= \frac{N(y_o|K(\boldsymbol{x}),\mathbf{E})\ N(\boldsymbol{x}|\boldsymbol{x}_b,\mathbf{B})}{\int [\ N(y_o|K(\boldsymbol{x}),\mathbf{E})\ N(\boldsymbol{x}|\boldsymbol{x}_b,\mathbf{B})\ ]d\boldsymbol{x}} \tag{20}$$

The product of two Gaussians can (as long as K is linearisable) be reorganised to collect the $\boldsymbol{x}$ terms into a single Gaussian:

$$N(y_o|K(\boldsymbol{x}),\mathbf{E})\ N(\boldsymbol{x}|\boldsymbol{x}_b,\mathbf{B}) = N(y_o|K(\boldsymbol{x}_b),\mathbf{E}+\mathbf{K}\mathbf{B}\mathbf{K}^T)\ N(\boldsymbol{x}|\boldsymbol{x}_a,\mathbf{A}) \tag{21}$$

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

$$\mathbf{A} = \mathbf{B}-\mathbf{B}\mathbf{K}^T(\mathbf{K}\mathbf{B}\mathbf{K}^T+\mathbf{E})^{-1}\mathbf{K}\mathbf{B}$$
$$\boldsymbol{x}_a = \boldsymbol{x}_b+\mathbf{B}\mathbf{K}^T(\mathbf{K}\mathbf{B}\mathbf{K}^T+\mathbf{E})^{-1}(y_o-K(\boldsymbol{x}_b)) \tag{22}$$

Substituting (21) in the denominator of (20), the Gaussian in $\boldsymbol{x}$ integrates to one. Cancelling the other Gaussian top and bottom gives:

$$p(\boldsymbol{x}|y_o) = N(\boldsymbol{x}|\boldsymbol{x}_a,\mathbf{A}) \tag{23}$$

Substituting our expresions for $\mathbf{K}$ and $\boldsymbol{x}$, we get:

$$\mathbf{B}\mathbf{K}^T(\mathbf{K}\mathbf{B}\mathbf{K}^T+\mathbf{E})^{-1} = \begin{pmatrix} V_b & \mu\ V_b \\ \mu\ V_b & V_b \end{pmatrix}\begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}\left[\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}\begin{pmatrix} V_b & \mu\ V_b \\ \mu\ V_b & V_b \end{pmatrix}\begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}+V_o+V_f\right]^{-1}$$
$$= \frac{V_b\left(\frac{1+\mu}{2}\right)}{V_o+V_f+V_b\left(\frac{1+\mu}{2}\right)}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{24}$$

$$\mathbf{A} = V_b\begin{pmatrix} 1 & \mu \\ \mu & 1 \end{pmatrix} - \frac{\left(V_b\left(\frac{1+\mu}{2}\right)\right)^2}{V_o+V_f+V_b\left(\frac{1+\mu}{2}\right)}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{25}$$

$$\boldsymbol{x}_a = \begin{pmatrix} x_{a1} \\ x_{a2} \end{pmatrix} = \begin{pmatrix} x_{b1} \\ x_{b2} \end{pmatrix} + \frac{\left(V_b\left(\frac{1+\mu}{2}\right)\right)^2}{V_o+V_f+V_b\left(\frac{1+\mu}{2}\right)}\left[y_o - \frac{x_{b1}+x_{b2}}{2}\right]\begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad (26)$$

If this problem is solved using optimal interpolation (OI), the approximation is made that $\mathbf{KBK}^T = V_b$, so the $\left(\frac{1+\mu}{2}\right)$ do not appear.

Log(probabilities) - Penalty Functions

The Bayesian analysis equation, with its product of probabilities, is hard to handle. For variational algorithms, it is more convenient to take minus the logarithm. The equation then becomes:

$$-\ln[p(x|y_o)] = -\ln[p(y_o|x)] - \ln[p(x)] + constant \qquad (27)$$

The simple zero-dimensional analyses shown in figure 1 become quadratics:

$$-\ln[p(x|y_o)] = \frac{1}{2}\frac{(x_b-x)^2}{V_b} + \frac{1}{2}\frac{(y_o-x)^2}{V_o} + constant. \qquad (28)$$

as shown in figure 5.

The full Bayesian analysis equation becomes:

$$-\ln[p(\boldsymbol{x}|\boldsymbol{y}_o)] = \frac{1}{2}(\boldsymbol{x}_b-\boldsymbol{x})^T\mathbf{B}^{-1}(\boldsymbol{x}_b-\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y}_o-K(\boldsymbol{x}))^T(\mathbf{O}+\mathbf{F})^{-1}(\boldsymbol{y}_o-K(\boldsymbol{x})) + constant. \qquad (29)$$

So we have provided a Bayesian justification for the form of penalty function used in many variational analysis schemes.

Non-Gaussian PDFs

It is satisfying to provide a justification for the quadratic penalty function. However the main purpose of this lecture is to provide a foundation for coping with non-Gaussian errors.

Model for Observational Errors

The simplest model that allows for the observed fact that observational errors are not in practice Gaussian, is to assume that a small fraction of the observations are corrupted, and hence worthless. The others have Gaussian errors. For each observation we have:

$$p(y_o|\boldsymbol{x}) = p(y_o|G\cap\boldsymbol{x})P(G) + p(y_o|\bar{G}\cap\boldsymbol{x})P(\bar{G}) \tag{30}$$

G is the event "there is a gross error".[6]
where $\bar{G}$ means *not* G.

$$p(y_o|\bar{G}\cap\boldsymbol{x}) = N(y_o|K(\boldsymbol{x}),\mathbf{O}+\mathbf{F})$$

$$p(y_o|G\cap\boldsymbol{x}) = \begin{cases} k & \textit{over the range of plausible values} \\ 0 & \textit{elsewhere} \end{cases} \tag{31}$$

The results of assuming such a pdf can be quite dramatic, even if $P(G)$ is small. Figure 6 shows the equivalent of figure 1, with errors appropriate for pressure observations from ships, which have about 5% gross errors. When the observation and the background agree, there is little difference from figure 1. But when they disagree, the posterior distribution becomes bi-modal.

Figure 7 shows the same examples in the log(probability) form of figure 5. The observational penalty is not quadratic; it has plateaus away from the observed value. Adding this to a quadratic background penalty can give multiple minima.

Practical Methods of Quality Control

The Bayesian methodology can be used applied in various ways to get practical schemes. In the remainder of this lecture I will describe three, before going on to discuss more general philosophy of quality control.

Individual Quality Control

If we accept the model for observational errors introduced in (30) and (31), it is straightforward to apply the discrete Bayes' theorem to the event G "there is a gross error":

$$P(G|y_o) = \frac{P(y_o|G)\ P(G)}{P(y_o)} \tag{32}$$

We can derive the formula for the posterior pdf $p(x|y_o)$ either directly from the continuous Bayes theorem, or else from:

$$p(\boldsymbol{x}|y_o) = p(\boldsymbol{x}|y_o\cap G)P(G|y_o) + p(\boldsymbol{x}|y_o\cap\bar{G})P(\bar{G}|y_o) \tag{33}$$

---

[6] We assume that G is independent of $\boldsymbol{x}$, so that $P(G|\boldsymbol{x})=P(G)$.

We saw the curves this gives in figure 6. The posterior pdf is the weighted sum of two Gaussian, corresponding to accepting or rejecting the observation. The weights given to each are the posterior probabilities of G and $\bar{G}$. When the peaks are distinct, these correspond to the areas under each.

Lorenc and Hammon (1988) extended this analysis to two observations with independent gross errors, and background values. This can be treated exactly by first calculating $P(G_i|y_o)$ for each, using the formula above, and then modifying them by a buddy check factor:

$$P(G_1|y_{o1}\cap y_{o2}) = P(G_1|y_{o1})\ \frac{P(y_{o1})P(y_{o2})}{P(y_{o1}\cap y_{o2})} \tag{34}$$

$$P(G_2|y_{o1}\cap y_{o2}) = P(G_2|y_{o2})\ \frac{P(y_{o1})P(y_{o2})}{P(y_{o1}\cap y_{o2})} \tag{35}$$

where

$$\frac{P(y_{o1})P(y_{o2})}{P(y_{o1}\cap y_{o2})} = \left[1-P(\bar{G}_1|y_{o1})P(\bar{G}_2|y_{o2})\left\{1-\frac{P(y_{o1}\cap y_{o2}|\bar{G}_1\cap\bar{G}_2)}{P(y_{o1}|\bar{G}_1)P(y_{o2}|\bar{G}_2)}\right\}\right]^{-1} \tag{36}$$

The algebra, and computation, to extend this exact calculation to n observation goes as 2^n. However it has been found in practice that sequentially applying the two observation buddy check is a reasonable approximation. (See Lorenc and Hammon 1993 appendix C for more details of the pairwise buddy check).

Simultaneous Quality Control

Lorenc (1981) introduced the Optimal Interpolation (OI) analysis method used operationally at ECMWF (until it is replaced soon by 3DVAR). This performs an explicit solution of a quadratic variation problem. The solution is calculated in boxes for as many observations as we can afford to handle at once.

A key feature of the ECMWF system is the use of the same methodology for quality control. Lorenc (1981) shows how, once the inverse of the OI matrix $\mathbf{M}$ ($=\mathbf{KBK}+\mathbf{O}+\mathbf{F}$ in our current notation) has been calculated, then it is possible with relatively few operations to solve the system of equations involving a smaller matrix omitting one (or a few) observations. He used this to check each observation in turn against a value analysed using all the other observations. An observation fails if:

$$(y_o-y_a)^2 > T^2(V_o+V_a) \tag{37}$$

where y_a, with error variance V_a, is the analysis obtained using the OI equations, at the position of the observation being checked, omitting the observation being checked and other

rejected observations.

In the Lorenc (1981) paper the tolerance (T) was set in a somewhat empirical manner to 4.0. When, as the scheme developed, we tried to account for the better quality of weather ship observations by reducing their observational error V_o, we found that this resulted in more being rejected:- not what we wanted. (It was this behavior, and the subjective tolerance in what was otherwise an objective analysis, that induced me to study the Bayesian approach.) It is shown in Lorenc and Hammon that the tolerance T should be given by:

$$T^2 = 2\ln\left[\frac{P(\bar{G})}{P(G)}\right] + \ln\left[\frac{k^{-2}}{2\pi\ (V_o + V_a)}\right] \qquad \textbf{(38)}$$

where k is the probability density of observations with gross error, as defined in (31). T is shown in figure 8.

In applying this method, observations have to be either included in, or excluded from, the analysis. While an observation is checked, the decisions on other observations are frozen. The ECMWF scheme follows a pragmatic approach of rejecting the worst, then rechecking the others, iteratively until no more fail.

Ingleby and Lorenc (1993) present equations for extending the Bayesian approach of Lorenc and Hammon. From the n gross error events G_i, they define 2^n new combined events C_α each corresponding to a particular set of rejections:

$$\begin{aligned} C_0 &= G_n \cap G_{n-1} \ldots \cap G_2 \cap G_1 \\ C_1 &= G_n \cap G_{n-1} \ldots \cap G_2 \cap \bar{G}_1 \\ C_2 &= G_n \cap G_{n-1} \ldots \cap \bar{G}_2 \cap G_1 \\ &\vdots \\ C_{2^n-1} &= \bar{G}_n \cap \bar{G}_{n-1} \ldots \cap \bar{G}_2 \cap \bar{G}_1 \end{aligned} \qquad \textbf{(39)}$$

Bayes theorem can be applied to each of these combined events:

$$P(C_\alpha \,|\, y_o) = \frac{P(y_o|C_\alpha)P(C_\alpha)}{\sum_{\alpha'=0}^{2^n-1} P(y_o|C_{\alpha'})P(C_{\alpha'})} \qquad \textbf{(40)}$$

Note that the denominator is the same in all the expressions; if we only want to find the most likely C_α it need not be evaluated. Evaluation of one $P(y_o|C_\alpha)$ involves evaluating only a single multi-variate Gaussian. In fact the ECMWF OI method (with Bayesian tolerance T) is deciding which is more likely out of two C_α which differ just by the observation being tested. Because it is judging between sets of quality control decisions, we call this approach *Simultaneous Quality Control*.

The C_α can be regarded as being the vertices of an n-dimensional hyper-cube. The method

starts from a first-guess set of rejections, and tests each observation in turn. This is equivalent to searching for the most probable of the adjacent vertices. It then iterates, re-testing vertices adjacent to the new C_α. This strategy is like the SIMPLEX algorithm of linear programming, but applied to a non-linear problem. It is not guaranteed to find the absolute maximum.

Variational Analysis with non-Gaussian Errors

It is possible to use our model of observational errors directly in a variational algorithm. Dharssi *et al.* (1992) did this for simulated windlidar observations. Instead of the quadratic $\frac{1}{2}(\mathbf{y}_o-K(\mathbf{x}))^T(\mathbf{O}+\mathbf{F})^{-1}(\mathbf{y}_o-K(\mathbf{x}))$, the observational penalty becomes (for diagonal $\mathbf{O}+\mathbf{F}$):

$$J_o = \sum_i -\ln\left[N(y_{o_i}|K(\mathbf{x}),V_{e_i})P(\bar{G}_i)+k_iP(G_i)\right] \qquad (41)$$

where V_{ei} ($=O_{ii}+F_{ii}$) is the observational error of i if it has not a gross error.
Differentiating this gives:

$$\frac{\partial J_o}{\partial y_i} = \frac{(y_{o_i}-y_i)}{V_{e_i}}\left\{\frac{N(y_{o_i}|y_i,V_{o_i})P(\bar{G}_i)}{N(y_{o_i}|y_i,V_{o_i})P(\bar{G}_i)+k_iP(G_i)}\right\} \qquad (42)$$

where y_i is the element of $K(\mathbf{x})$ corresponding to $\mathbf{x}$ interpolated to the ith observation position. The first term is just what we get if the observation error is Gaussian, as in variational methods with a quadratic penalty function. The term in braces is equal to the probability that observation i has not a gross error, given that $\mathbf{x}$ is exactly correct. So for each iteration u of the descent algorithm, all one has to do to allow for gross errors is to replace the observational error for each observation in the formulae for the standard quadratic penalty by:

$$E_{o\,[u]} = \frac{V_e}{P(\bar{G}|y_o\cap K(x_{[u]}))} \qquad (43)$$

i.e. the observational error variance should be inflated by one over the probability (given the current best estimate $\mathbf{x}_{[u]}$) that the observation has not a gross error.
Note that this only gives the correct gradient of J_o; it does not give the correct penalty (for which we need (41)) nor the correct second derivative. In general, analysis error estimates based on the second derivative, valid for Gaussian distributions (e.g. in (22)) will be over-optimistic for long-tailed distributions.

We saw in figure 7 that if errors are non-Gaussian, the penalty function is non-quadratic and can have multiple minima. So the end point of a descent algorithm iteration will depend on the first-guess. In a set of variational assimilation experiments with a one-dimensional shallow water model and its adjoint, I found that (for the example studied) the first-guess had to be very good to get convergence to the best solution (Lorenc 1988, figure 13).

Dharssi *et al.* (1992) studied various approaches for overcoming this problem for a simple example with two observations, so that the penalty function can be plotted as contours. Ordinary descent algorithms did not always find the lowest minimum (figure 9). Better results could be obtained by artificially increasing the assumed observational error for early iterations, slowly reducing it to its true value. Alternatively, one can artificially decrease the prior $P(G)$ to zero for early iterations. Neither approach always worked. However in simulations using rather dense observations with large probabilities of gross error (up to 50%), the method with increased observational error worked satisfactorily.

Comparison

Ingleby and Lorenc (1993) compared Individual Quality Control (IQC), Simultaneous Quality Control (SQC), and non-Gaussian Variational Analysis (VAN) for some simple examples. One is shown in figure 10. There are two observations and thus four combinations C_α, each corresponding to a dotted Gaussian curve in figure 10. The table below shows their posterior probabilities.

	G_2	$\bar{G}_2$	$G_2 \cup \bar{G}_2$
G_1	.393	.191	.584
$\bar{G}_1$	.003	.413	.416
$G_1 \cup \bar{G}_1$	.396	.604	1.0

The most likely combination is for them both to be correct: the table shows $P(\bar{G}_1 \cap \bar{G}_2) = .413$. This is the SQC result. Note that the simplex-like search algorithm would not work in this case: if we start at $G_1 \cap G_2$, both $\bar{G}_1 \cap G_2$ and $G_1 \cap \bar{G}_2$ are less likely, so we do not get to $\bar{G}_1 \cap \bar{G}_2$.
To get the probabilities for IQC we have to sum rows and columns: the table show $P(\bar{G}_1) = .416$, $P(\bar{G}_2) = .604$. So observation 1 probably has a gross error, and observation 2 is probably correct. Note however that if we make these decisions individually, and then draw the analysis assuming $G_1 \cap \bar{G}_2$, then we are choosing a rather unlikely combination.
The variational analysis method choses the highest point on the solid curve in figure 3. Note that a descent algorithm starting from the background would not have found this.

Monitoring

To the manager of a manufacturing company, Quality Control has a different meaning to that we have implied so far; he wants to know about and prevent errors. In NWP we call the equivalent process "monitoring". The purpose is to collect statistics on the performance of observing and processing systems, to detect systems that are not performing as expected, and to feed this information back so the deficiency is corrected at source. To do this we need:

- a comprehensive database of basic and processed observed values, independent estimates of the same quantities, and parameters affecting the processing
- software for catagorising, sorting, and analysing the database
- effort to try catagorisations and look for "unexpected" behavior
- communications, willpower and persistence, to get errors from stages out of your direct control rectified.

Design of the monitoring system is as important as design of the data-assimilation scheme; it should not be added on as an afterthought.

Another important product of monitoring is a good description of the observational error characteristics. If we are using the gross-error model, we need to know the prior probability of error, and the error distributions of gross-error and of "good" observations. Without these, the quality control is not objective [7]. Lorenc and Hammon (1988) showed how the statistics of observations processed by their quality-control scheme could be used (with some added "judgement") to "bootstrap" the assumed prior distributions. For some observation types, more complicated error models are called for. For instance there are many different ways that a radiosonde temperature and geopotential report can be corrupted. Gandin *et al.* (1993) have devised a "Comprehensive Quality Control" scheme which looks for sixteen.

Why Quality Control?

The quality control we do in data-assimilation has two reasons:

1. We have physical reasons for believing certain events may occur which affect the observed value. We wish to detect these events.
2. The distribution of errors associated with a datum is such that there is a non-negligable probability of errors that would be unacceptably large *for the use we are making of the datum*.

Note that 2 depends on our use of the observation. If we are using an analysis method based on a quadratic penalty function, it is linear in the observed values. A single large error can then be disasterous (figure 11a). However if instead we minimise the mean absolute deviation (this is the correct norm if the pdf is proportional to an exponential of the absolute deviation), the bad datum is ignored (figure 11b).

What is the Best Analysis?

The Bayesian approach can, in theory, give us the posterior pdf, but to start a forecast we need a single best analysis. In our simple example of figure 5, is the best analysis in the tallest peak, or in the peak with the largest area? To approach this objectively, we have to define how much it costs us to be wrong, or alternatively how much we benefit from being nearly right. If we have a quadratic benefit function, then the best analysis is the mean of the posterior pdf. If we have a spike (delta function) benefit function, the best analysis is at the maximium of the posterior pdf. Probably the best simple model is between the two: a Gaussian shaped benefit function such that analyses close to the truth are useful, while those a long way from correct are equally worthless. However we are a long way from being able to quantify this, and even further from being able to compute the full posterior pdf and then maximise it, for an NWP analysis with non-Gaussian observational errors.

---

[7] by "objective" I mean more than the automatic application of *ad hoc* rules, rather that the rules themselves have some statistical foundation.

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

Gandin Lev S., L.L.Morone and W G Collins 1993
"Two years of operational comprehensive quality control at the National Meteorological Center" *Weather and Forecasting*, **8**, 57-72

Gauss 1823
"Theoria combinationis observationum erroribus minimis obnoxiae".

Ingleby, N.B., and Lorenc, A.C. 1993
"Bayesian quality control using multivariate normal distributions". *Quart.J.Roy.Met.Soc.*, to appear July 1993.

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.

Tarantola,A. 1987
"Inverse Problem Theory - Methods for data fitting and model parameter estimation". publ. Elsevier. 613pp. ISBN 0-444-42765-1

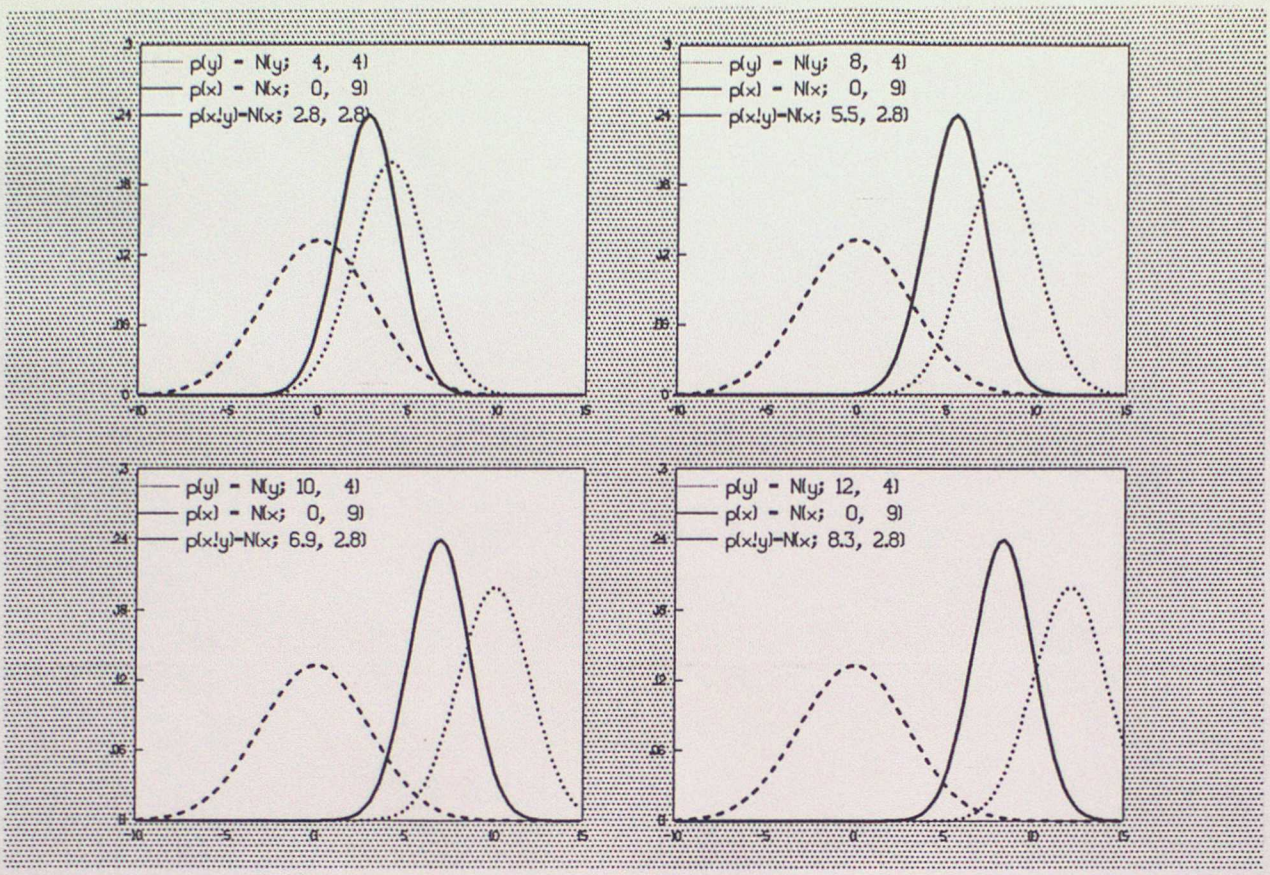

Figure 1. Prior pdf $p(x)$ *(dashed line)*, posterior pdf $p(x|y_o)$ *(solid line)*, and likelihood of observation $p(y_o|x)$ *(dotted line)*, plotted against x for various values of y_o. (Adapted from Lorenc and Hammon 1988).

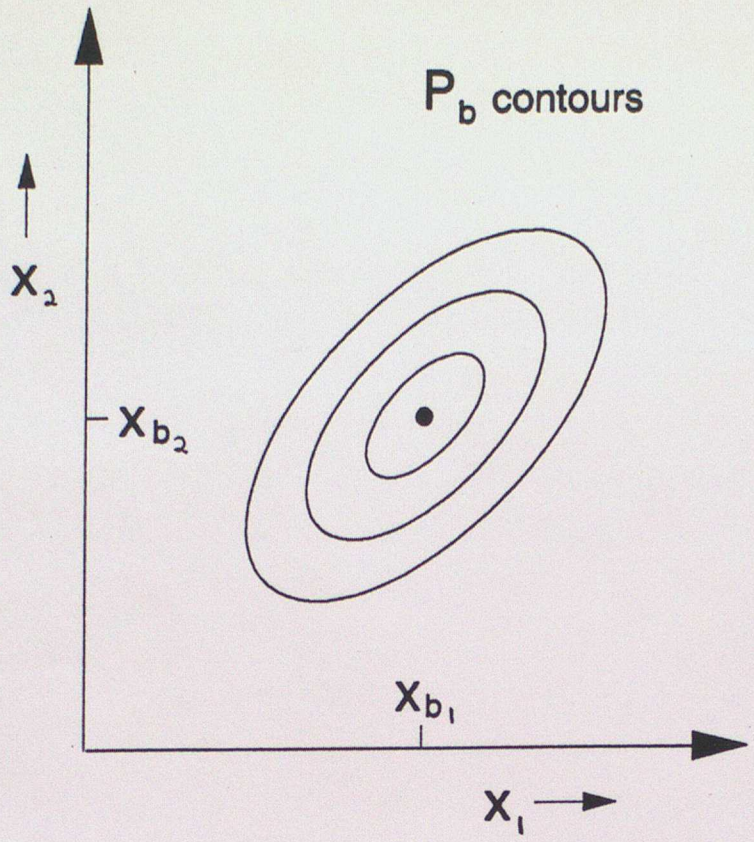

Figure 2. Contour plot of the prior pdf $p(x_1,x_2)$, for the simple example with a positive correlation between the background errors of x_{b1} and x_{b2}.

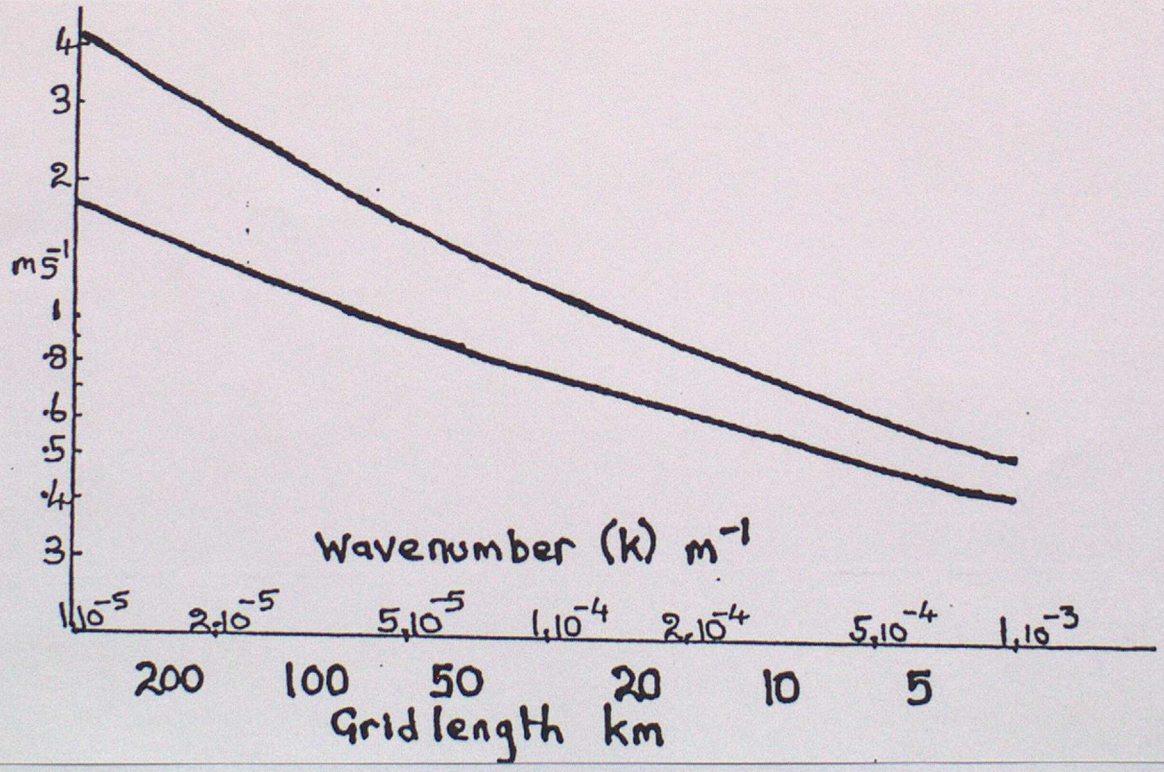

Figure 3. Estimated horizontal representativeness error for a spot wind observation, plotted against the grid-length used to represent the field. The values were calculated from the upper and lower limits of observed wind spectra, assuming a $k^{-5/3}$ law.

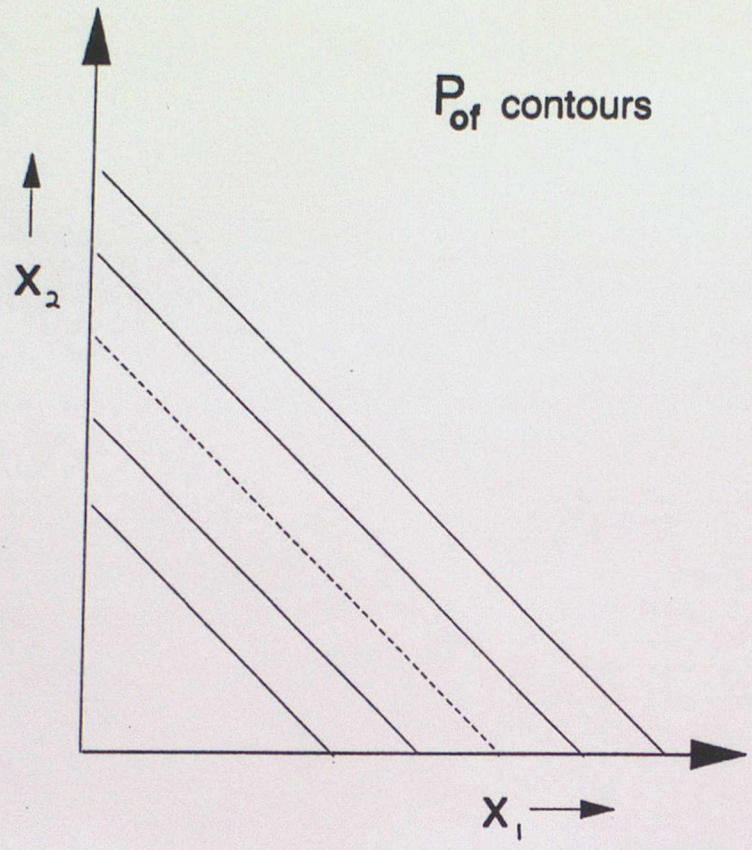

Figure 4. Likelihood function: the probability ($p(y_o|x)$) of getting the observation, plotted as a function of x.

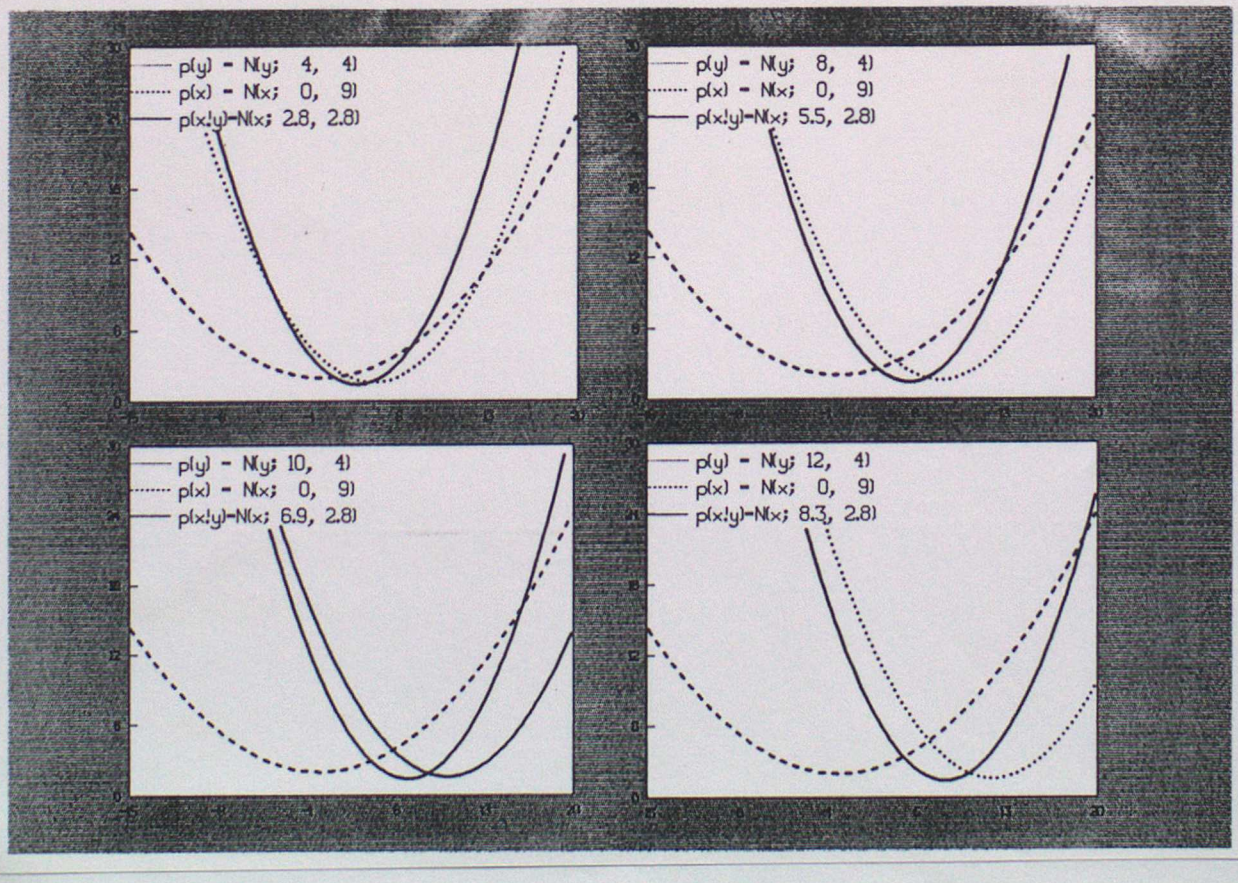

Figure 5. As figure 1 for -log(probabilities).

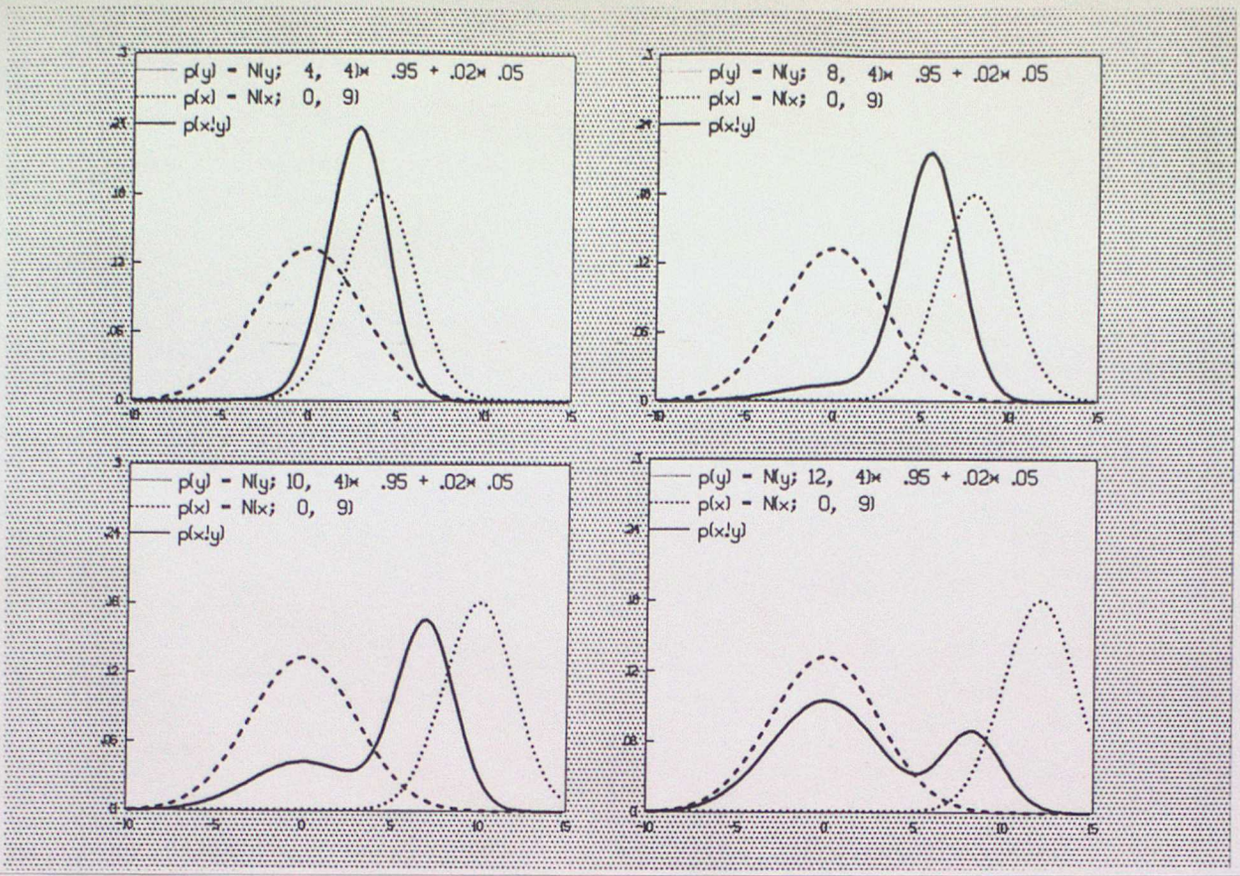

Figure 6. As figure 1 for an observation with a 5% chance of gross error.

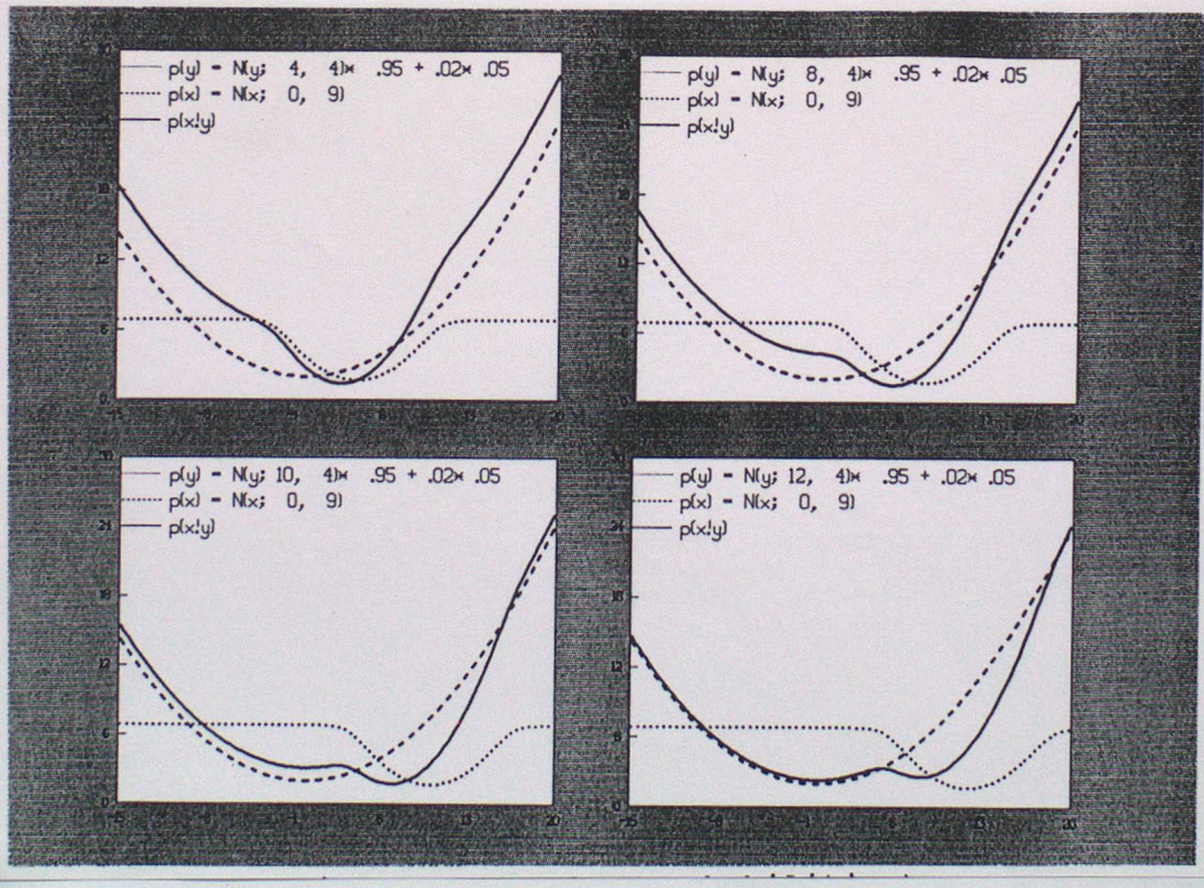

Figure 7. As figure 6 for -log(probabilities).

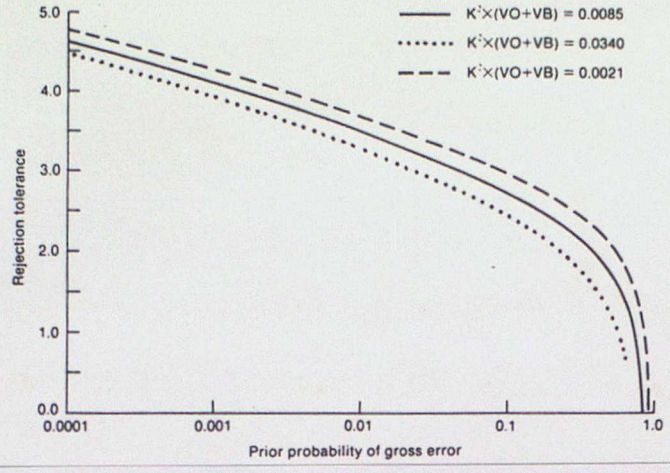

Figure 8. Rejection tolerance T, plotted against prior probability of gross error (from Lorenc and Hammon 1988).

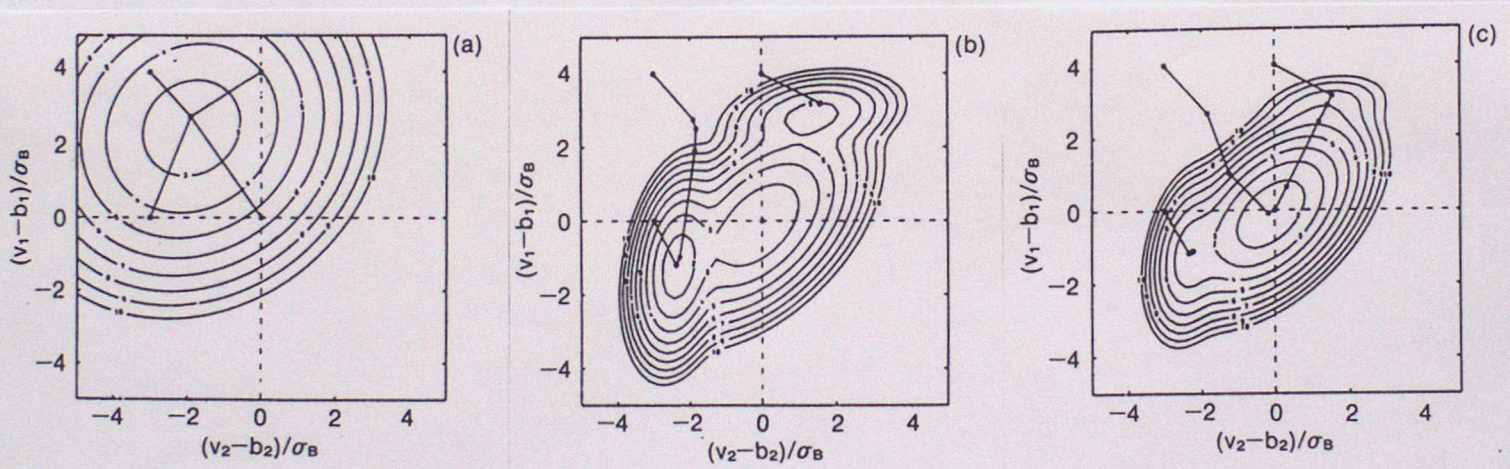

Figure 7. Contours of the penalty function (11) when only two observations are present. The observation values are given by $o_1 - b_1 = 4\,\sigma_B$, $o_2 - b_2 = -3\,\sigma_B$ and the background error correlation between the two observation points is set to 0.5. The observation error standard deviation $\sigma_0 = 0.5\,\sigma_B$ and $a_i = b_i$. (a) displays the contours when the initial probability of gross error $P_g = 0.0$. (b) and (c) display the contour maps when P_g is 0.1 and 0.5 respectively. The tracks superimposed (dots connected by solid line) represent the path taken by the iterative scheme (12) through this space, for four different starting points which are at $(0, 0)$, $(-3, 0)$, $(-3, 4)$ and $(0, 4)$.

Figure 9. Figure 7 from Dharssi *et al.* (1992).

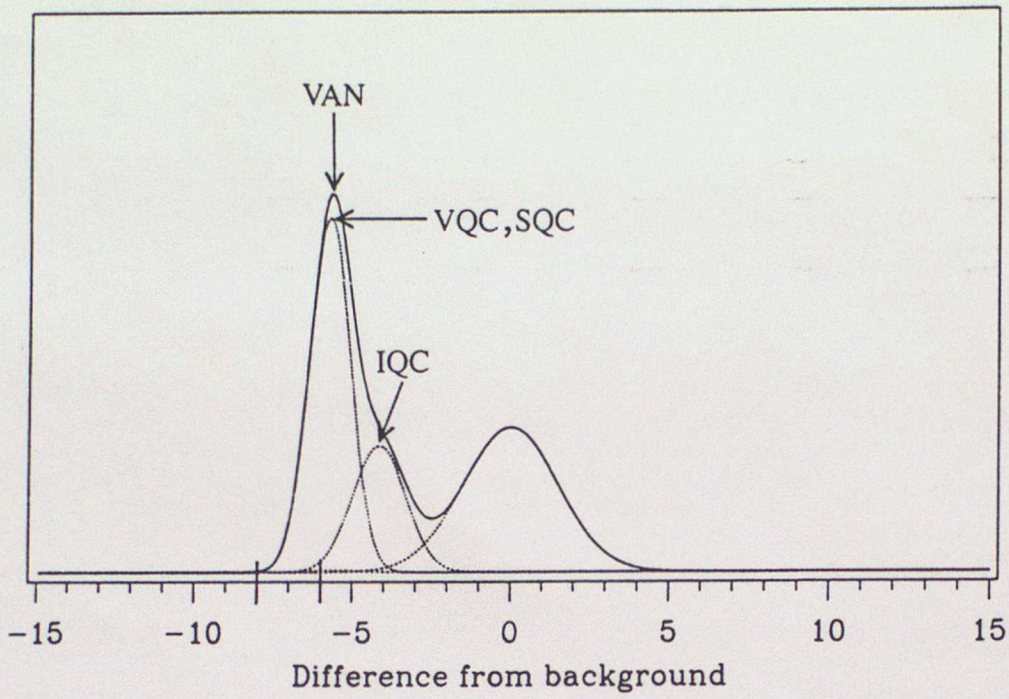

Figure 10. Posterior pdf for collocated observations differing by -8mb and -6mb from the background. The prior $P(G)=.05$ for each, the error variance of good observations is $V_e=1.0\text{mb}^2$, the background error variance is $V_b=(1.5\text{mb})^2$, and the probability density of observations with gross errors is $k=.043\text{mb}^{-1}$. (from Ingleby and Lorenc 1993).

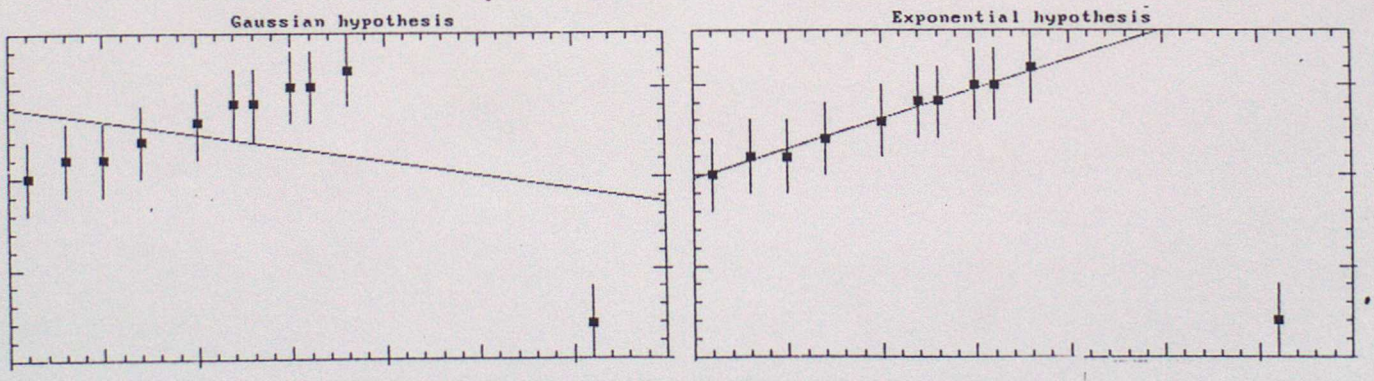

Figure 1.26: Two variables y (ordinate) and t (abscissa) are related by the relationship y = a t + b , where a and b are unknown parameters. In order to estimate a and b , an experiment has been performed which has furnished the 11 experimental points shown in the figure. The exact meaning of the "error bars" is not indicated.

Figure 11. Best fit straight line to data including a gross error, (a) using a quadratic (L2) norm, (b) using a mean absolute (L1) norm (from Tarantola 1987).