4.8 Introduction of statistics

Let us recall that the present course treats of variational data assimilation with a deterministic point of view mainly; it is based on minimization processes (variational approach). Filtering is the second option in the data assimilation science, it is based on statistical estimators (e.g. Kalman's filter). We present below very roughly the most simple estimator: the Best Linear Unbiased Estimate (BLUE). The goal of the present section is to show the link we have between the variational approach and the stochastic approach.

In the linear case, we show below that the variational approach and the BLUE are equivalent (one can refer for example to [4] for more details).

4.8.1 A first simple example

Let us consider the following basic and instructive example (this example has been communicated to the author by E. Blayo, [4]. We have two measurements of the scalar quantity x: $x_1 = 1$ and $x_2 = 2$.
'Naturally', we seek x such that it minimizes the cost function: $j(x) = (x - x_1)^2 + (x - x_2)^2$. It is a standard least-square problem (in the present case, x does not have to satisfy an underlying model). The solution is: $x^* = 3/2$.

Now, let us assume that the second measurement measured the quantity $2x$ (and not x). The two measurements we have are: $y_1 = 1$ and $y_2 = 4$.
Again, 'naturally', we seek x such that it minimizes the cost function: $g(x) = (x - y_1)^2 + (2x - y_2)^2$. The solution is: $x^* = 9/5$. It differs from the previous one....

This simple example illustrates that the solution of the minimization problem (here a simple least-square problem) depends on the normalization of the cost function and/or of the measurements.

In addition, with the present straightforward least-square approach, the solution does take into account the accuracy of the measurements. In other words it does not take into account the confidence we have on measurements

In order to take into account the measurements accuracy (in others words, the confidence we have on), we will *introduce a-priori statistical information in the norm definition of the cost function.*

We will present shortly how to do so below. But, first we present (recall) the Best Linear Unbiased Estimate (BLUE).

4.8.2 The Best Linear Unbiased Estimate (BLUE)

Basic recalls of probability and statistics

We recall some basic definitions and notations of probability and statistics (the explanations below mainly come from public documents published on the web, and wikipedia webpages).

Let $\hat{X}$ be an aleatory (random) variable.

Its average, also called the expected value, expectation or mean, is defined by:

$$\mu \equiv E[\hat{X}] = \int_a^b x f(x) dx$$

It is the probability-weighted average over all samples, $f(x)$ being the probability density function (pdf).

The covariance between two random variables $\hat{X}$ and $\hat{Y}$ is denoted by $cov(\hat{X}, \hat{Y})$ or $\sigma_{\hat{X}\hat{Y}}$. It measures how much the two random variables change together.
If the two variables $\hat{X}$ and $\hat{Y}$ are independent then $cov(\hat{X}, \hat{Y}) = 0$. The reciprocal is not true. Random variables whose covariance is zero are called uncorrelated.
The definition is:

$$\sigma_{\hat{X}\hat{Y}} \equiv cov(\hat{X}, \hat{Y}) = E[(\hat{X} - E[\hat{X}])(\hat{Y} - E[\hat{Y}])]$$

We have the property:

$$cov(\hat{X}, \hat{Y}) = E[\hat{X}\hat{Y}] - E[\hat{X}]E[\hat{Y}]$$

The form $cov(.,.)$ is bilinear symmetrical positive define. Its quadratic form associated is the variance (see below).

Its normalized version is the correlation coefficient $\rho(\hat{X}, \hat{Y})$. Its expressions is:

$$\rho(\hat{X}, \hat{Y}) = \frac{\sigma_{\hat{X}\hat{Y}}}{\sigma_{\hat{X}} \sigma_{\hat{Y}}}$$

It measures the strength of the linear relation between $\hat{X}$ and $\hat{Y}$. It is a dimensionless number. If $\rho = 0$ the two variables are uncorrolated; if $\rho = +/-1$, the two are linearly dependent.

The variance of a random variable or distribution, denoted by $Var(\hat{X})$ (or $\sigma^2_{\hat{X}}$), is a measure of how far the numbers lie from the average.
In other words, the variance is a measure of the amount of variation of the values of that variable, taking account of all possible values and their probabilities or weightings (not just the extremes which give the range).
It is the expected value of the squared difference between the variable's realization and the variable's mean:

$$\begin{aligned} \sigma^2_{\hat{X}} \equiv Var(\hat{X}) &= cov(\hat{X}, \hat{X}) \\ &= E[(\hat{X} - \mu)((\hat{X} - \mu)] \\ &= E[(\hat{X} - \mu)^2] \end{aligned}$$

The standard deviation ("ecart-type" in french) is the square root of the variance. It is denoted $\sigma_{\hat{X}}$. It has the same dimension as the random variable $\hat{X}$.
Equivalently, it is the square root of the average value of $(\hat{X} - E(\hat{X})^2)$.

Roughly, the standard deviation measures the variation (dispersion) from the average. A low standard deviation indicates that the data points tend to be very close to the mean; high standard deviation indicates that the data points are spread out over a large range of values, Fig. 4.15.

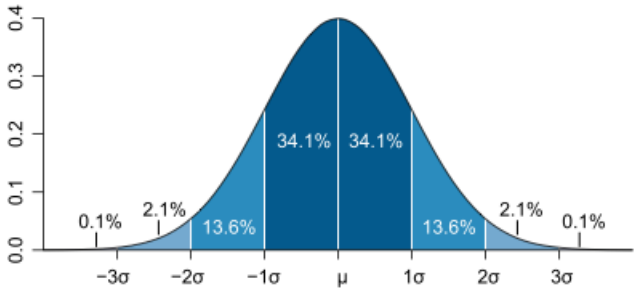


Figure 4.15: A plot of a normal (Gaussian) distribution. Each band has a width of 1 standard deviation. From Wikipedia webpage.

In statistics, an estimator is a rule for calculating an estimate of a given quantity based on observed data; its result is called the estimate.

The Mean Square Error (MSE) of an estimator quantifies the difference between values implied by an estimator and the true values of the quantity being estimated.
The mean-square error measures how far, on average, the collection of estimates are from a parameter x being estimated. It is defined by:

$$MSE(\hat{X}) = E((\hat{X} - x)^2)$$

We have:

$$MSE(\hat{X}) = Var(\hat{X}) + (Bias(\hat{X}, x))^2$$

Hence, if the random variable $\hat{X}$ is unbiased, the variance equals the MSE.

The BLUE

BLUE means Best Linear Unbiased Estimate. The method is an estimator.

Definition 4.8.1. *The Best Linear Unbiased Estimate of an aleatory variable $\hat{X}$ based on data Y is:*

a) *a linear function of Y.*
(Comment: it is more simple).

b) *unbiased:* $E(\hat{X}) = x$
(Comment: it is desirable).

c) *of minimal variance* $Var(\hat{X})$ *among all unbiased linear estimators.*
(Comment: with optimal accuracy).

Back to the simple example In the previous example, now let us denote the two measurements with errors by:

$$y_i = x + e_i,\ i = 1, 2$$

The errors of measurements e_i are supposed to be:
– unbiased: $E(e_i) = 0$.
(Comment: the sensors are unbiased, the mean of errors is null).
– with a variance given: $Var(e_i) = \sigma_i,\ i = 1, 2$.
(Comment: the accuracy of sensors are known...).
– uncorrelated : $E(e_1 e_2) = 0$.
(Comment: measurements are independent hence the covariance vanish; furthermore their means are null, hence the relation).

Let us seek the Best Linear Unbiased Estimate (BLUE) denoted by x^* (here "best" means giving the lowest mean squared error).

By definition, the BLUE satisfies: $x^* = a_1 y_1 + a_2 y_2$ (linear relaionship), with the coefficients a_i to be determined.

Since the errors are unbiased, we have $E(x^* - x) = 0$. Since:

$$E(x^*) = (a_1 + a_2)x + a_1 E(e_1) + a_2 E(e_2) = (a_1 + a_2)x$$

then: $a_2 = 1 - a_1$.

Next, let us calculate the variance of x^*:

$$\begin{aligned} Var(x^*) &= E((x^* - x)^2) = E((a_1 y_1 + a_2 y_2)^2) \\ &= a_1^2 E(e_1^2) + 2a_1 a_2 E(e_1 e_2) + a_2^2 E(e_2^2) \\ &= a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 \\ &= a_1^2 \sigma_1^2 + (1 - a_1)^2 \sigma_2^2 \end{aligned}$$

The estimate x^* we seek must minimize the variance. Hence, it is minimal if its derivative with respect to a_1 vanishes i.e. if:

$$a_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

Therefore, the BLUE writes:

$$\boxed{x^* = \frac{1}{(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2})}(\frac{1}{\sigma_1^2}y_1 + \frac{1}{\sigma_2^2}y_2)} \tag{4.20}$$

and we have:

$$Var(x^*) = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$$

or equivalently:

$$\frac{1}{Var(x^*)} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}$$

4.8.3 Equivalence with a variational solution

It is straightforward to verify that the BLUE x^* defined by (4.17) minimizes the following quadratic cost function:

$$\boxed{j(x) = \frac{1}{2}\frac{(x-y_1)^2}{\sigma_1^2} + \frac{1}{2}\frac{(x-y_2)^2}{\sigma_2^2}} \tag{4.21}$$

And we have:

$$j''(x) = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} = \frac{1}{Var(x^*)}$$

In others words, *the convexity of j equals the accuracy of the estimate.*

We illustrate this property of relationship between the Hessian and the quality of the analysis in 1d in Figure 4.16.

In summary, the introduction of statistic in the measurements lead to a natural and best definition of norms of the cost function $j(x)$.
This remark remains true for non-linear and large dimensional cases.

Furthermore let us remark that the BLUE calculated above (and under the assumption that measurements are uncorrolated) minimizes the following cost function too:

$$j(x) = \frac{1}{2}(x-y_1, x-y_2)\begin{pmatrix}\sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}^{-1}\begin{pmatrix}x-y_1 \\ x-y_2\end{pmatrix} \tag{4.22}$$

where in the norm definition we have assumed that the errors are a-priori corrolated (see the extra diagonal terms of the matrix norm).

Extension to m observations. The extension to m observations is straigthforward. We denote the m measurements with errors by: $y_i = x + e_i,\ i = 1, .., m$. We assume that the errors are:
– unbiased: $E(e_i) = 0,\ i = 1, .., m$.
– with a variance given: $Var(e_i) = \sigma_i,\ i = 1, .., m$.

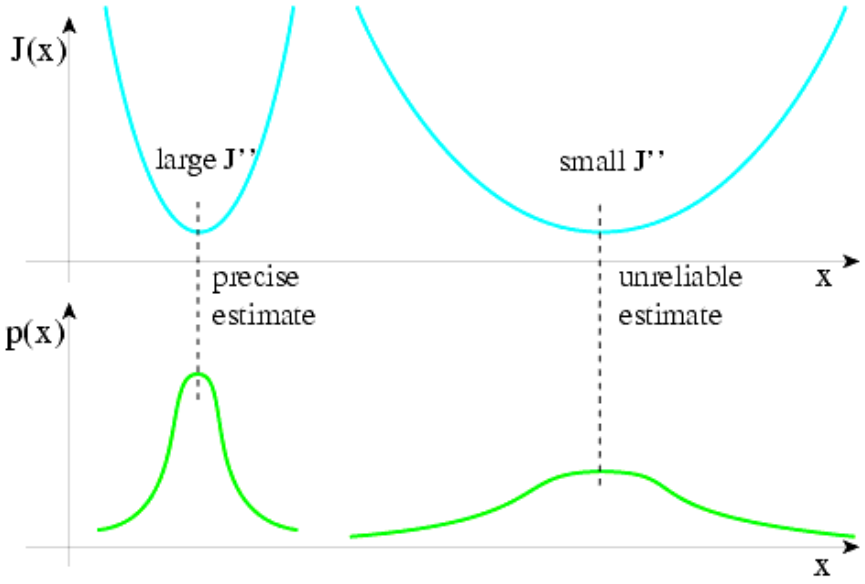


Figure 4.16: In the linear quadratic case: convexity=reliability. In 1d, the Hessian is the second derivative hence the convexity; it is also the measure of the sharpness of the pdf of the analysis. Figure from the ECMWF data assimilation course (www.ecmwf.int) by F. Bouttier, P. Courtier.

– uncorrelated : $E(e_i e_j) = 0$ for all $(i,j),\ 1 \leq i,j \leq m,\ i \neq j$.

The Best Linear Unbiased Estimate writes:

$$x^* = \frac{1}{\sum_{i=1}^m \frac{1}{\sigma_i^2}} (\sum_{i=1}^m \frac{1}{\sigma_i^2} y_i) \text{ with } \frac{1}{Var(x^*)} = \sum_{i=1}^m \frac{1}{\sigma_i^2}$$

which is equivalent to minimize:

$$j(x) = \frac{1}{2} \sum_{i=1}^m \frac{(x - y_i)^2}{\sigma_i^2} \tag{4.23}$$

Hence: $j''(x) = \dfrac{1}{Var(x^*)}$.

4.8.4 Link with a filtering point of view.

Filters are stochastic algorithms which operates recursively on streams of noisy input data (measurements observed over time) to produce a statistically optimal estimate of the underlying state.

Let go back to the simple example with two observations. The BLUE writes:

$$x^* = \frac{\sigma_2^2 y_1 + \sigma_1^2 y_2}{\sigma_1^2 + \sigma_2^2} = y_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(y_2 - y_1)$$

Let us consider:
a) y_1 be a first estimate (it is what we call a *background* or a *first guess*).
b) y_2 be an independent observation.

We denote: $x_b = y_1$ and $y = y_2$. We obtain:

$$\boxed{x^* = x_b + (\frac{\sigma_b^2}{\sigma_b^2 + \sigma_0^2})(y - x_b)}$$

Using the terminology fo data assimilation, in a filtering point of view(e.g. the Kalman filter), this equation reads as follows.
The best estimate x^* equals the first guess + the *gain* times the *innovation* $(y - x_b)$.

4.8.5 Extension to the vectorial case.

We seek to estimate the vectorial quantity $x = (x_1...x_n)^T \in \mathbb{R}^n$. The measurements / observations are: $y = (y_1...y_m)^T \in \mathbb{R}^m$. We assume that the *observation operator* H is linear: $y = Hx$. Hence H is a $m \times n$ matrix.

We set: $Y = Hx + e$ with $e \in \mathbb{R}^m$ the error vector (aleatory variables).

We assume that:
– $E(e) = 0$. *(The sensors are unbiased).*
– $Cov(e) = E(ee^T) = \Sigma$. *(The covariances, hence the accuracy, are known).*
Let us point out that the observations are not supposed independent anymore, hence the matrix Σ may be non-diagonal.

We write the BLUE, then using the *Gauss-Markov theorem*, one can prove that the BLUE is equivalent to the minimization of the following cost function:

$$\boxed{j(x) = \frac{1}{2}\|Hx - y\|^2_{\Sigma^{-1}} = \frac{1}{2}(Hx - y)^T\Sigma^{-1}(Hx - y)} \tag{4.24}$$

The matrix Σ being the errors covariance matrix.

4.8.6 In conclusion

. In the linear case (ie linear observation operator, quadratic cost function), we have an equivalence between the BLUE (stochastic) and the minimization approach (deterministic). And, this point of view allow to define the 'best' norms in the cost function to be minimized. The equivalence show that norms must be defined by the inverse of the covariance errors matrix, see (4.21).

. The convexity of j is the accuracy of the estimate ! As a matter of fact:

$$Hess(j) = H^T\Sigma^{-1}H = [Cov(x^*)]^{-1}$$

. *What about in non-linear cases ?* If the underlying model is linear and the observation operator is linear than the cost function is quadratic, and it admits a unique (global) minimum. At contrary, either if the underlying model is non-linear and/or the observation operator is non-linear than the cost function is non quadratic anymore. In such a case, the minimization approach can be local only (and more or less well conditioned...).
In other respect, one do not have equivalence anymore between a statistical method and a variational method.
Nevertheless, the knowledge of the statistical errors is a crucial point to "well" define the cost function to be minimized. This remains true for non-linear cases; even if this knowledge is generally difficult to obtain or approximate.

. Generally, the errors of observations are supposed to be uncorrelated: $E(\varepsilon_i \varepsilon_j) = 0$ for $i \neq j$.
Then the matrix $R = \Sigma^{-1}$ is diagonal and *the diagonal coefficients of the matrix are the a-priori confidence we have in the observation.*

Remark 4.8.1. *A last remark on filtering methods. Filtering methods can be extended to non-linear systems, by linearizing for example the model at each time step (e.g. Extended Kalman Filter). In such a case, the filter is not optimal anymore. Many other extension exists, let us cite the Ensemble Kalman filter (Monte-Carlo method to estimate covariance error matrices). In other respect, method of reduction order have been developed to make decrease the cpu time required to compute the gain matrices (e.g. the filter SEEK).*

4.8.7 A second simple example

The Best Linear Unbiased Estimate (BLUE) in image and for a simple example: What Time Is It ?

An example due to Prof. O. Thual
Univ. Toulouse INPT & CERFACS
See O. Thual webpage



Horloges de précisions égales

- Horloge 1 : il est 14h23mn à 5mn près
- Horloge 2 : il est 14h19mn à 5mn près

The BLUE is: $14H21\text{mn}$ at $+/-$ $5\sqrt{2}$ mn$(\approx 3,5$ mn$)$
ie. the mean at +/- the variance

Horloges de précisions inégales

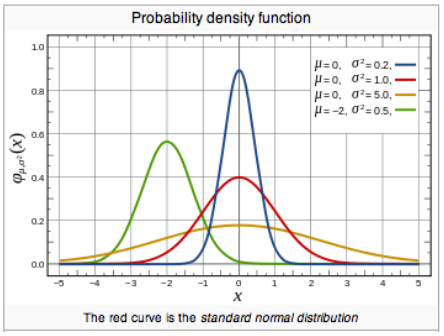
- Horloge 1 : il est 14h23mn à 2 mn près
- Horloge 2 : il est 14h19mn à 5 mn près

The BLUE is: $14H19\text{mn}32s$ at $+/-$ $10\sqrt{29}$ mn$(\approx 1,85$ mn$)$
the mean at +/- the variance

Basic recalls of Probability & Statistics
(Random variables, continuous & discrete)

The probability density function (pdf) in case of a Normal (or Gaussian) distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\bar{x})^2}{2\sigma^2})$$

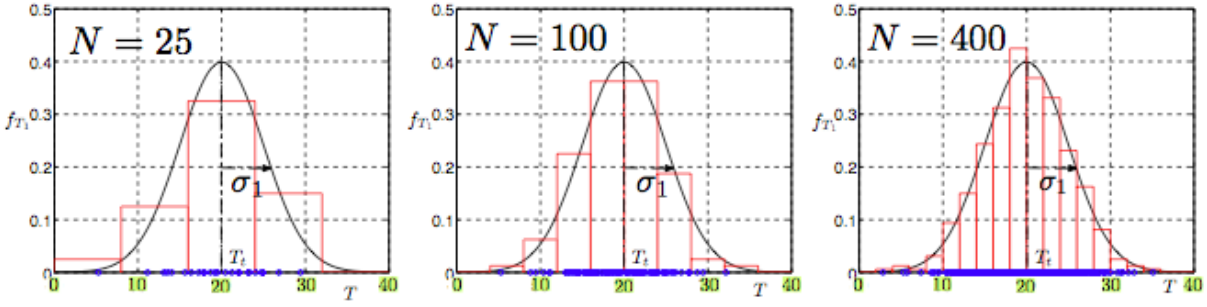


The mean (expected value):

$$\mu \equiv E(\hat{X}) = \int_a^b x f(x) dx \approx \frac{1}{N} \sum_{i=1}^{N} \hat{X}_n$$

The variance: $Var(\hat{X}) = E[(\hat{X} - \mu)^2]$

The standard deviation: $\sigma = (Var(\hat{X}))^2$



Plot: $\mu = 20$ mn $\sigma = 5^2$ mn

Calculation of the BLUE

1) The BLUE (called the estimate or the « analysis ») satisfies a linear relationship:

$$T_k = (1-k)T_1 + kT_2$$

2) Errors are supposed unbiased and uncorrelated, hence the variance writes:

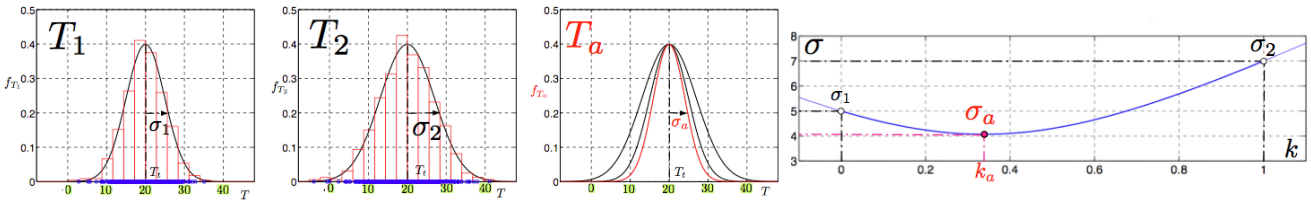
$$\sigma_a^2 = (1-k)^2\sigma_1^2 + k^2\sigma_2^2$$

3) The estimate minimizes the variance:

$$k_a = \frac{\sigma_2^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}}$$

Finally, we obtain the BLUE / estimate / analysis:

$$\boxed{T_a = \frac{\alpha_1 T_1 + \alpha_2 T_2}{\alpha_1 + \alpha_2}} \quad \text{with} \quad \alpha_i = \frac{1}{\sigma_i^2}\ ,\ i = 1, 2. \quad \text{and} \quad \alpha_a = \frac{1}{\sigma_a^2} = \alpha_1 + \alpha_2$$



Equivalency between the BLUE and an optimal solution (the variational one)

*) The BLUE: $T_a = \dfrac{\alpha_1 T_1 + \alpha_2 T_2}{\alpha_1 + \alpha_2}$ minimizes the following cost function:

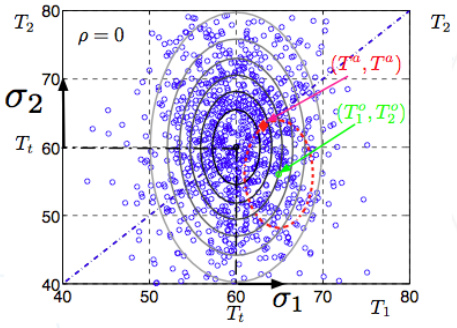
$$\boxed{j(T) = \frac{1}{2}\frac{(T-T_1)^2}{\sigma_1^2} + \frac{1}{2}\frac{(T-T_2)^2}{\sigma_2^2}}$$

*) One remark that the BLUE (obtained under the assumption of uncorrelated errors), minimizes the following cost function too:

$$j(T) = \frac{1}{2}(T-T_1, T-T_2)\begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}^{-1}\begin{pmatrix} T-T_1 \\ T-T_2 \end{pmatrix}$$

i.e. with a-priori correlated errors ! (see the correlation coefficients in the extra diagonal terms).

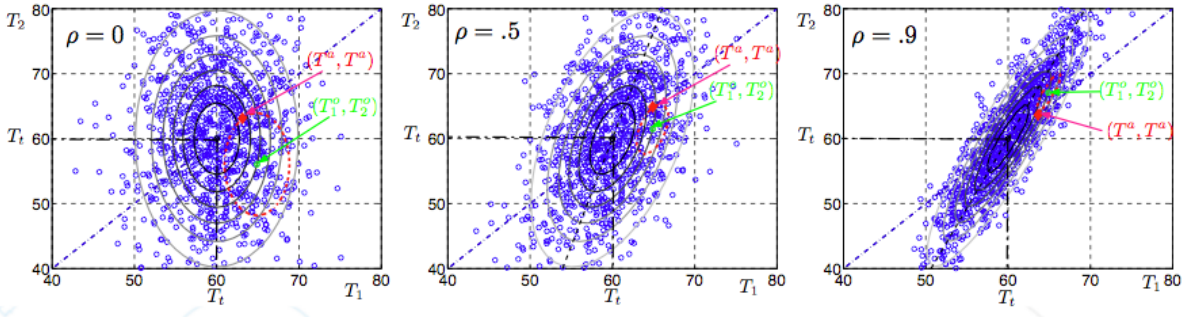
*) T1 and T2 are two measurements uncorrolated: the correlation coefficient vanishs.



$$\rho(T_1, T_2) = \frac{cov(T_1, T_2)}{\sigma_1 \sigma_2} = 0$$

*) T1 and T2 are two measurements corrolated:

$$\rho(T_1, T_2) = \frac{cov(T_1, T_2)}{\sigma_1 \sigma_2} \neq 0$$



How to improve the estimate ?
How to assimilate extra measurements ?

Assumption.

One knowns the « accuracy » of each measurement
i.e. each variance is known (or equivalently the standard deviation):

$$\sigma_i \ , \ i = 1, .., m \ (\text{for } m \text{ measurements}) \text{ given.}$$

Next, each extra measurement may « enhance the accuracy « of the estimate in the sense:

$$\frac{1}{\sigma_a^2} = \frac{1}{\sigma_1^2} + ... + \frac{1}{\sigma_m^2}$$

In fact, each extra measurement makes decrease the dispersion from the mean,
but not necessarily the accuracy of the estimate if the extra measurements are not accurate !...

The BLUEstimate writes:

$$T_a = \frac{1}{\sum_{i=1}^m \frac{1}{\sigma_i^2}} (\sum_{i=1}^m \frac{1}{\sigma_i^2} T_i) \text{ with } \frac{1}{Var(T_a)} = \sum_{i=1}^m \frac{1}{\sigma_i^2}$$

Finally, what time is it ?