Image and Model Fusion: Unexpected Counterintuitive Behavior of Traditional Statistical Techniques and Resulting Need for Expert Knowledge

Omar Ochoa[1] , Aaron Velasco[2] and Vladik Kreinovich*[1]

[1]Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA
[2]Department of Geological Sciences, University of Texas at El Paso, El Paso, TX 79968, USA

Email: Omar Ochoa - omar@miners.utep.edu; Aaron Velasco - velasco@geo.utep.edu; Vladik Kreinovich*- vladik@utep.edu;

*Corresponding author

Abstract

In many real-life situations, we have different types of data. For example, in geosciences, we have seismic data, gravity data, magnetic data, etc. Ideally, we should jointly process all this data, but often, such a joint processing is not yet practically possible. In such situations, it is desirable to "fuse" models (images) corresponding to different types of data: e.g., to fuse an image corresponding to seismic data and an image corresponding to gravity data. At first glance, if we assume that all the approximation errors are independent and normally distributed, then we get a reasonably standard statistical problem which can be solved by the traditional statistical techniques such as the Maximum Likelihood method. Surprisingly, it turns out that for this seemingly simple and natural problem, the traditional Maximum Likelihood approach leads to non-physical results. To make the fusion results physically meaningful, it is therefore necessary to take into account expert knowledge.

Model (and Image) Fusion: Formulation of a Problem

Need to combine data from different sources

In many areas of science and engineering, we have different sources of data. For example, in geophysics, there are many sources of data for Earth models:

- first-arrival passive seismic data (from actual earthquakes); see, e.g., [8];

- first-arrival active seismic data (from seismic experiments using man-made sources); see, e.g., [2, 5];
- gravity data; and
- surface waves; see, e.g., [9].

Datasets coming from different sources provide complimentary information. For example, different geophysical datasets contain different information on earth structure:

- each measured gravity anomaly at a point is the result of the density distribution over a relatively large region of the earth, so estimates based on gravity measurements provide information about densities at both low and high depths;
- in contrast, each seismic data point (arrival time) comes from a trajectory (ray) a seismic wave travels within the earth, so the resulting values only cover areas above the Moho surface, where these rays propagate.

Usually, there are several different geophysical datasets available. At present, each of these datasets is often processed separately, resulting in several different "models" – 2-D or 3-D images reflecting different aspects of the studied phenomena. It is therefore desirable to combine data from different datasets.

Joint inversion: an ideal future approach

The ideal approach would be to use all the datasets to produce a single model (= image). At present, however, in many research areas – including geophysics – there are no efficient algorithms for simultaneously processing all the different datasets.

Designing such joint inversion techniques presents an important theoretical and practical challenge.

Model (image) fusion: main idea

While joint inversion methods are being developed, as a first step, we propose a practical solution: to fuse all the *models* (images) coming from processing different datasets; see, e.g., [11–13, 16].

How to fuse images: towards a precise formulation of the problem

To fuse the images, we need, for each spatial location $j = 1, \ldots, N$, to fuse the intensities of different images corresponding to this location. Each of these intensities estimates the actual (unknown) value of the desired quantity (such as density) at the selected location. Let $\widetilde{x}_j^{(1)}, \ldots, \widetilde{x}_j^{(n)}$ be the values from different images

corresponding to the same spatial location. Our objective is to merge these values into a single more accurate estimate for the actual value x_j of the desired quantity at this location; see, e.g., [15].

In many practical situations, the approximation errors $\Delta x_j^{(i)} \stackrel{\text{def}}{=} \widetilde{x}_j^{(i)} - x_j$ corresponding to different images are normally distributed; see, e.g., [15,17]. This fact that the approximation errors are normally distributed can be justified by the Central Limit Theorem, according to which, under certain reasonable conditions, the joint effect of many relatively small errors is (approximately) normally distributed; see, e.g., [17]. For each model based on measurements of a certain type (e.g., gravity or seismic), not only the resulting error of each measurement comes from many different error sources, but also each estimate comes from several different measurements – thus further increasing the number of different error components contributing to the estimation error.

Approximation errors corresponding to different images come from different sources, so it is reasonable to assume that they are independent. Because of independence, to describe the joint distribution of all n approximation errors, it is sufficient to describe the probability distribution of each approximation error $\Delta x_j^{(i)}$.

A normal distribution is uniquely determined by its mean and standard deviation. In principle, for each image, the mean of the corresponding approximation error can be non-zero; in measurement terms, this means that the approximation errors can have a *systematic* component. However, these components can be eliminated if we appropriately calibrate the corresponding measuring instruments and data processing algorithms. Because of this, in the following text, we can safely assume that the mean value of each approximation error is 0.

Since the mean is thus fixed, to describe the probability distribution for each approximation error $\Delta x_j^{(i)}$, it is sufficient to know the corresponding standard deviation. Sometimes, values corresponding to different parts of the image are known with different accuracy. For example, in geosciences, since most measurements are performed at the Earth's surface, these measurements enable us to find the values at lower depths more accurately than the values corresponding to deeper structures. To take this into account, let us divide the image into several zones Z_k, $k = 1, \ldots, M$, and let us assume that the accuracy is the same for all the locations in each zone. In more precise terms, we assume that there are values $\sigma_k^{(i)}$, and that for each location $j \in Z_k$, the standard deviation is equal to $\sigma_k^{(i)}$.

In the following text, we will denote the total number of locations in zone k by N_k, and the total number of locations by N.

The resulting probability density for each estimation error $\Delta x_j^{(i)}$, with $j \in Z_k$, has the form

$$\frac{1}{\sqrt{2\cdot\pi}\cdot\sigma_k^{(i)}}\cdot\exp\left(-\frac{\left(\Delta x_j^{(i)}\right)^2}{2\cdot\left(\sigma_k^{(i)}\right)^2}\right)=\frac{1}{\sqrt{2\cdot\pi}\cdot\sigma_k^{(i)}}\cdot\exp\left(-\frac{\left(\widetilde{x}_j^{(i)}-x_j\right)^2}{2\cdot\left(\sigma_k^{(i)}\right)^2}\right),$$

and the probability density $\rho(x)$ corresponding to all N estimates $x=(x_1,\ldots,x_N)$ in all the zones is (due to independence) the product of these densities:

$$\rho(x)=\prod_{k=1}^{M}\prod_{j\in Z_k}\prod_{i=1}^{n}\frac{1}{\sqrt{2\cdot\pi}\cdot\sigma_k^{(i)}}\cdot\exp\left(-\frac{\left(\widetilde{x}_j^{(i)}-x_j\right)^2}{2\cdot\left(\sigma_k^{(i)}\right)^2}\right).$$

Separating the terms in front of the exponentials and the exponential terms themselves, and using the fact that $\exp(a)\cdot\exp(b)=\exp(a+b)$, we conclude that

$$\rho(x)=\left(\prod_{k=1}^{M}\prod_{j\in Z_k}\prod_{i=1}^{n}\frac{1}{\sqrt{2\cdot\pi}\cdot\sigma_k^{(i)}}\right)\cdot\exp\left(-\sum_{k=1}^{M}\sum_{j\in Z_k}\sum_{i=1}^{n}\frac{\left(\widetilde{x}_j^{(i)}-x_j\right)^2}{2\cdot\left(\sigma_k^{(i)}\right)^2}\right).$$

The expression for the first factor can be further simplified if we take into account that in this factor, the term $\sqrt{2\cdot\pi}$ is repeated as many times as there are locations, i.e., N times, and each term $\sigma_k^{(i)}$ corresponding to the k-th zone is repeated as many times as there are locations in this zone, i.e., N_k times. As a result, the expression for the probability density $\rho(x)$ takes the following form:

$$\rho(x)=\frac{1}{\left(\sqrt{2\cdot\pi}\right)^N}\cdot\prod_{k=1}^{M}\prod_{i=1}^{n}\frac{1}{\left(\sigma_k^{(i)}\right)^{N_k}}\cdot\exp\left(-\sum_{k=1}^{M}\sum_{j\in Z_k}\sum_{i=1}^{n}\frac{\left(\widetilde{x}_j^{(i)}-x_j\right)^2}{2\cdot\left(\sigma_k^{(i)}\right)^2}\right).$$

Case when we know the accuracy of different images

In some cases, we know the accuracy of different images, i.e., we know the standard deviations $\sigma^{(i)}$. In this case, the only unknown parameters are the actual value x_j of the desired quantity at different locations. A reasonable idea is to select the value for which the probability (density) $\rho(x)$ is the largest, i.e., to use the *Maximum Likelihood* method.

Since $\exp(z)$ is an increasing function, maximizing a function $A\cdot\exp(-B(x))$ is equivalent to minimizing $B(x)$, so we arrive at the following *Least Squares* approach: for each $j\in Z_k$, we must find x_j for which the sum $\sum\limits_{i=1}^{n}\dfrac{\left(\widetilde{x}_j^{(i)}-x_j\right)^2}{2\cdot\left(\sigma_k^{(i)}\right)^2}$ is the smallest possible.

Differentiating this expression with respect to x_j and equating the derivative to 0, we conclude that

$$x_j = \frac{\sum\limits_{i=1}^{n} \widetilde{x}_j^{(i)} \cdot \left(\sigma_k^{(i)}\right)^{-2}}{\sum\limits_{i=1}^{n} \left(\sigma_k^{(i)}\right)^{-2}}.$$

Comment.

The accuracy of this fused estimate can be described by the standard deviation σ_k for which

$$\sigma_k^{-2} = \sum_{i=1}^{n} \left(\sigma_k^{(i)}\right)^{-2}.$$

General case: a description

In practice, we often do not know the standard deviations $\sigma_k^{(i)}$. In this case, we need to estimate both the actual values x_j and the standard deviations $\sigma_k^{(i)}$ from the observations $\widetilde{x}_j^{(i)}$.

Comment

This is actually the most fundamental measurement situation. In reality, all we know are the results of measuring different quantities with different measuring instruments. Our estimates of the actual values of these quantities and our estimates of the accuracy of different measuring instruments must all be derived from the measurement results.

Model (and Image) Fusion: A Seemingly Reasonable Solution

General case: a seemingly reasonable approach

In the general case, we have several parameters that we need to estimate: in addition to the desired values x_j, we have the standard deviations $\sigma_k^{(i)}$ that we also need to estimate based on the measurement results.

To find the estimates of $x = (x_1, \ldots, x_N)$ and $\sigma = (\sigma_1^{(1)}, \ldots, \sigma_k^{(i)}, \ldots)$, it seems reasonable to use the same Maximum Likelihood method as before, i.e., to find the values x_j and $\sigma_k^{(i)}$ for which the above expression $\rho(x, \sigma)$ for the probability density attains its largest possible value.

General case: towards an algorithm

To find the resulting maximizing values, we can differentiate the probability density with respect to both x_j and $\sigma_k^{(k)}$ and equate the corresponding derivatives to 0.

Similarly to the case when we know the accuracies, we can simplify the computations if instead of maximizing the original function $\rho(x,\sigma)$, we consider an equivalent problem of minimizing $\psi(x,\sigma) \stackrel{\text{def}}{=} -\ln(\rho(x,\sigma))$. From the above expression for the probability density, we get the following expression for the new objective function $\psi(x,\sigma)$:

$$\psi(x) = N \cdot \ln(\sqrt{2\cdot\pi}) + \sum_{k=1}^{M}\sum_{i=1}^{n} N_k \cdot \ln\left(\sigma_k^{(i)}\right) + \sum_{k=1}^{M}\sum_{j\in Z_k}\sum_{i=1}^{n} \frac{\left(\widetilde{x}_j^{(i)} - x_j\right)^2}{2\cdot\left(\sigma_k^{(i)}\right)^2}.$$

We already know that differentiating this expression with respect to x_j and equating the derivative to 0 leads to the expression $x_j = \dfrac{\sum\limits_{i=1}^{n} \widetilde{x}_j^{(i)} \cdot \left(\sigma_k^{(i)}\right)^{-2}}{\sum\limits_{i=1}^{n} \left(\sigma_k^{(i)}\right)^{-2}}$.

Let us now show what will happen if we differentiate the objective function $\psi(x,\sigma)$ with respect to $\sigma_k^{(i)}$ and equate the derivative to 0. The only terms of the objective function $\psi(x)$ that depends on $\sigma_k^{(i)}$ are the terms

$$N_k \cdot \ln\left(\sigma_k^{(i)}\right) + \sum_{j\in Z_k} \frac{\left(\widetilde{x}_j^{(i)} - x_j\right)^2}{2\cdot\left(\sigma_k^{(i)}\right)^2},$$

i.e., the terms

$$N_k \cdot \ln\left(\sigma_k^{(i)}\right) + \frac{1}{2\cdot\left(\sigma_k^{(i)}\right)^2} \cdot \sum_{j\in Z_k} \left(\widetilde{x}_j^{(i)} - x_j\right)^2.$$

Differentiating these terms with respect to $\sigma_k^{(i)}$ and equating the result to 0, we get

$$N_k \cdot \frac{1}{\sigma_k^{(i)}} - \frac{1}{\left(\sigma_k^{(i)}\right)^3} \cdot \sum_{j\in Z_k} \left(\widetilde{x}_j^{(i)} - x_j\right)^2 = 0,$$

i.e.,

$$N_k \cdot \frac{1}{\sigma_k^{(i)}} = \frac{1}{\left(\sigma_k^{(i)}\right)^3} \cdot \sum_{j\in Z_k} \left(\widetilde{x}_j^{(i)} - x_j\right)^2.$$

We want to use this equation to find the value $\sigma_k^{(i)}$. To do that, we move all the terms containing $\sigma_k^{(i)}$ to one side, and all other terms to another side. Specifically, we multiply both sides of this equation by $\left(\sigma_k^{(i)}\right)^3$ and divide both sides by N_k. As a result, we get the following equation:

$$\left(\sigma_k^{(i)}\right)^2 = \frac{1}{N_k} \cdot \sum_{j\in Z_k} \left(\widetilde{x}_j^{(i)} - x_j\right)^2.$$

This equation makes perfect sense: it says that $\left(\sigma_k^{(i)}\right)^2$ is the mean square deviation between the actual values x_j from the k-th zone and the measurement results $\widetilde{x}_j^{(i)}$ obtained by the i-th instrument for locations from this zone.

General case: main idea of an algorithm

Thus, a seemingly reasonable idea is to do the following:

Since we have no prior information about the accuracy of different images, in the initial approximation, we assume that all the images are of the same accuracy, i.e., that all the values $\sigma_k^{(i)}$ are equal. In other words, as a first approximation $\sigma_{k,0}^{(i)}$ to the values $\sigma_k^{(i)}$, we take $\sigma_{k,0}^{(i)} = \sigma$ for some $\sigma > 0$. Based on these values $\sigma_{k0}^{(i)}$, we use the above formula for x_j to find the first approximation $x_{j,1}$ to the intensity values x_j.

One can see that the result of applying this formula does not change when we multiply all the values $\sigma_k^{(i)}$ by the same constant. This means it does not matter what initial value σ we take, we can as well take $\sigma = 1$.

Now, based on the estimates x_{j1} for x_j, we can use the above formula for $\sigma_k^{(i)}$ to find the resulting approximation $\sigma_{k,1}^{(i)}$ to the accuracies $\sigma_k^{(i)}$.

Now that we know more accurate estimates $\sigma_{k,1}^{(i)}$ for the standard deviations $\sigma_k^{(i)}$ than the initial estimates $\sigma_{k,0}^{(i)} = \sigma$, we can use these more accurate estimates and produce better approximations $x_{j,2}$ to the actual intensities x_j. Based on these better approximations for x_j, we can compute better approximations for $\sigma_k^{(i)}$. We can repeat this iterative procedure several times to get more and more accurate approximations.

Let us describe the resulting algorithm in precise terms.

Algorithm: description

This is an iterative algorithm. We start with the initial approximate values $\sigma_{k,0}^{(i)} = 1$. On each stage $p = 1, 2, \ldots$ of this algorithm:

- first, we use the values $\sigma_{k,p-1}^{(i)}$ to compute $x_{j,p-1} = \dfrac{\sum\limits_{i=1}^{n} \widetilde{x}_j^{(i)} \cdot \left(\sigma_{k,p-1}^{(i)}\right)^{-2}}{\sum\limits_{i=1}^{n} \left(\sigma_{k,p-1}^{(i)}\right)^{-2}}$;
- then, we compute the next approximation to the standard deviations as follows:

$$\left(\sigma_{k,p}^{(i)}\right)^2 = \frac{1}{N_k} \cdot \sum_{j \in Z_k} \left(\widetilde{x}_j^{(i)} - x_{j,p}\right)^2.$$

Iterations continue until these values converge, i.e., until the differences $|x_{i,p} - x_{i,p-1}|$ and $\left|\sigma_{k,p}^{(i)} - \sigma_{k,p-1}^{(i)}\right|$ become smaller than some pre-defined value $\delta > 0$.

Comment

The first two iterations of this algorithm are actively used in astrometry, when we determine the coordinates x_j of stars and other distant astronomical objects; see, e.g., [14]. For many objects, we have measurements performed by different telescopes; the accuracies $\sigma_k^{(i)}$ with which different telescopes i measures coordinates are not exactly known.

Since these accuracies $\sigma_k^{(i)}$ are not known, at first, we assume that all the telescopes are equally accurate, i.e., take $\sigma_{k,0}^{(i)} = \sigma$ for some $\sigma > 0$. Based on these approximate values, we find an approximate coordinates $x_{j,1}$ of all the stars j. Since $\sigma_{k,0}^{(i)} = \text{const}$, the corresponding formula for $x_{j,1}$ means that for each star, we simply take the arithmetic average of all the observations.

Once we have these approximate coordinates, we can compute a better approximation $\sigma_{k,1}^{(i)}$ to the accuracies of different telescopes. Now that we know these accuracies, we can get a better approximation $x_{j,2}$ to the actual coordinates – by giving more weight to more accurate observations.

Image and Model Fusion: Unexpected Counterintuitive Behavior of Traditional Statistical Techniques

Surprisingly, the results of applying the above algorithm do not make any physical sense

In astrometry, computations usually stop after the second iteration. Intuitively, as we have mentioned, the more iterations we perform, the more accurate the resulting estimates. So, to achieve the best possible accuracy, we decided to continue the above iterations.

The result was completely unexpected: the process did converge, but instead of converging to physically meaningful values of x_j and $\sigma_k^{(i)}$, our iterations converged to a set of values for which one of the standard deviations $\sigma_k^{(i)}$ is 0, and each value x_j is equal to the corresponding measurement result $\widetilde{x}_j^{(i)}$. In other words, no matter what we started with, our conclusion was that one of the measuring instruments is absolutely accurate. Instead of fusing the measurement results, we simply select one of these results.

This conclusion does not make any physical sense – we know that none of the measuring instruments is perfect.

Toy example

To get a better understanding of what is going on, we decided to apply the algorithm to a simple (toy) example. Let us assume that we are measuring a single quantity x_1 by using three measuring instruments, and we get three different values $\widetilde{x}_1^{(i)}$. Without losing generality, let us sort these three values in increasing

order: $\widetilde{x}_1^{(1)} < \widetilde{x}_1^{(2)} < \widetilde{x}_1^{(3)}$. Let us denote, by x_0, the midpoint $x_0 = \dfrac{\widetilde{x}_1^{(1)} + \widetilde{x}_1^{(3)}}{2}$ of the resulting range of values, and by Δ, the radius (half-width) $\Delta = \dfrac{\widetilde{x}_1^{(3)} + \widetilde{x}_1^{(1)}}{2}$ of this range. In these notations, we have $\widetilde{x}_1^{(1)} = x_0 - \Delta$ and $\widetilde{x}_1^{(3)} = x_0 + \Delta$.

Let us first consider the simplest case when the three values are equally spaced, i.e., when the middle value $\widetilde{x}_1^{(2)}$ coincides with the midpoint x_0. On the first iteration, we compute the arithmetic mean of all three measurement results $x_0 - \Delta$, x_0, and $x_0 + \Delta$. This mean is equal to $x_{11} = x_0$. Now, we compute estimates $\sigma_{k,1}^{(i)}$. In this case, the above formula leads to $\sigma_{k,1}^{(i)} = \left|\widetilde{x}_j^{(i)} - x_{j,1}\right|$, so we get $\sigma_{k,1}^{(1)} = \Delta$, $\sigma_{k,1}^{(2)} = 0$, and $\sigma_{k,1}^{(3)} = \Delta$. Since $\sigma_{k,1}^{(1)} > 0$, $\sigma_{k,1}^{(2)} = 0$, and $\sigma_{k,1}^{(3)} > 0$, and on the next iteration $x_{1,2}$, the weights with which we add different measurement results $\widetilde{x}_1^{(i)}$ are proportional to $\left(\sigma_{k,1}^{(i)}\right)^{-2}$, we only take into account the second measurement result, i.e., we get $x_{1,2} = \widetilde{x}_1^{(2)} = x_0$. One can see that now, the process converges – the next iteration does not change anything. Thus, we indeed conclude that one of the accuracies becomes 0 and the actual value coincides with the result of the corresponding measurement.

Maybe this conclusion was caused by the fact that we assumed that middle value $\widetilde{x}_1^{(2)}$ exactly coincides with the midpoint x_0? Let us try a more general case, when this middle value can be anywhere between $x_0 - \Delta$ and $x_0 + \Delta$, i.e., when it is equal to $x_0 + \theta \cdot \Delta$, for some $\theta \in (-1, 1)$. In this case, the arithmetic means of these three values is equal to

$$\frac{(x_0 - \Delta) + (x_0 + \theta \cdot \Delta) + (x_0 + \Delta)}{3} = x_0 + \frac{\theta}{3} \cdot \Delta.$$

We thus get $\sigma_{1,1}^{(i)} = \Delta \cdot \left(1 + \dfrac{1}{3} \cdot \theta\right)$, $\sigma_{2,1}^{(i)} = \Delta \cdot \dfrac{2}{3} \cdot \theta$, and $\sigma_{3,1}^{(i)} = \Delta \cdot \left(1 - \dfrac{1}{3} \cdot \theta\right)$. As a result, we get

$$x_{1,2} = x_0 + \Delta \cdot \frac{-\dfrac{1}{\left(1 + \dfrac{1}{3} \cdot \theta\right)^2} + \dfrac{\theta}{\left(\dfrac{2}{3} \cdot \theta\right)^2} + \dfrac{1}{\left(1 - \dfrac{1}{3} \cdot \theta\right)^2}}{\dfrac{1}{\left(1 + \dfrac{1}{3} \cdot \theta\right)^2} + \dfrac{1}{\left(\dfrac{2}{3} \cdot \theta\right)^2} + \dfrac{1}{\left(1 - \dfrac{1}{3} \cdot \theta\right)^2}}.$$

For small θ, in the first approximation, the main term in the numerator is equal to $\dfrac{\theta}{\left(\dfrac{2}{3} \cdot \theta\right)^2}$, and the main term in the denominator is equal to $\dfrac{1}{\left(\dfrac{2}{3} \cdot \theta\right)^2}$, so we get $x_{1,2} \approx x_0 + \theta \cdot \Delta$, i.e., $x_{1,2} \approx \widetilde{x}_1^{(2)}$. Thus, the fused value (almost) coincides with the result of the second measurement, and on the next iteration, we conclude that the corresponding standard deviation $\sigma_{1,2}^{(2)} = \left|\widetilde{x}_1^{(2)} - x_{1,2}\right|$ is (almost) 0.

On the next iteration, the difference between the estimate $x_{1,2}$ and the second measurement result $\widetilde{x}_1^{(2)}$ becomes even smaller, and in the limit, we get $x_{1,2} = \widetilde{x}_1^{(2)}$.

General case

Maybe the problem is with the iterative algorithm, and the actual maximum of the likelihood function is attained at a more physically meaningful solution? Alas, as we will show, the maximum of the likelihood function is attained exactly at the above physically meaningless solution.

Indeed, let us pick a real number $\varepsilon > 0$. Let us select some i_0 (e.g., $i_0 = 1$), and, for all locations j, take $x_j = \widetilde{x}_j^{(i_0)} + \varepsilon$ and $\sigma_k^{(i_0)} = \varepsilon$. For all other $i \neq i_0$, we take $\sigma_k^{(i_0)} = 1$. In this case,

$$\frac{(\widetilde{x}_j^{(i_0)} - x_j)^2}{2 \cdot (\sigma_k^{(i_0)})^2} = \frac{\varepsilon^2}{2\varepsilon^2} = \frac{1}{2},$$

and for $i \neq i_0$, we get

$$\frac{(\widetilde{x}_j^{(i)} - x_j)^2}{2 \cdot (\sigma_k^{(i)})^2} = \frac{(\widetilde{x}_j^{(i)} - (\widetilde{x}_j^{(i_0)} + \varepsilon))^2}{2} \to \frac{(\widetilde{x}_j^{(i)} - \widetilde{x}_j^{(i_0)})^2}{2}$$

as $\varepsilon \to 0$.

So, when $\varepsilon \to 0$, the exponential term

$$\exp\left(-\sum_{k=1}^{M} \sum_{j \in Z_k} \sum_{i=1}^{n} \frac{(\widetilde{x}_j^{(i)} - x_j)^2}{2 \cdot (\sigma_k^{(i)})^2}\right)$$

tends to a finite (and non-zero) value

$$\exp\left(-\sum_{k=1}^{M} \sum_{j \in Z_k} \left(\frac{1}{2} + \sum_{i \neq i_0} \frac{(\widetilde{x}_j^{(i)} - \widetilde{x}_j^{(i_0)})^2}{2}\right)\right).$$

On the other hand, since $\sigma_k^{(i_0)} = \varepsilon \to 0$, the first factor in the above expression for $\rho(x)$, i.e., the expression

$$\prod_{k=1}^{M} \prod_{j \in Z_k} \prod_{i=1}^{n} \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma_k^{(i)}},$$

tends to infinity. Thus, the largest possible – infinite – value of $\rho(x)$ is attained when x_j coincides with one of the measurement results: $x_j = \widetilde{x}_j^{(i_0)}$.

So, if we use the Maximum Likelihood method, then, instead of fusing different images – as we wanted – we now select one of these images. This is not what we wanted. In other words, in the general case, the Maximum Likelihood method does not lead to a physically meaningful result.

How to Make Fused Images Physically Meaningful: Heuristic Solutions

Let us first describe two heuristic solutions that can help us arrive at physically meaningful image fusion.

Limiting number of iterations

The first idea is to do what researchers do in astrometry: only perform the first two iterations.

This idea is similar to what often happens in *asymptotic series* (see, e.g., [4]) when we have a divergent series in which the sum of the first few terms is a very good approximation to the desired computations. Such asymptotic series are ubiquitous in quantum field theory; see, e.g., [1].

Maximum Entropy Approach: idea

When the above iterations converge, we get the values x_j and $\sigma_k^{(i)}$ for which $x_j = \dfrac{\sum\limits_{i=1}^{n} \widetilde{x}_j^{(i)} \cdot \left(\sigma_k^{(i)}\right)^{-2}}{\sum\limits_{i=1}^{n} \left(\sigma_k^{(i)}\right)^{-2}}$ and $\left(\sigma_k^{(i)}\right)^2 = \dfrac{1}{N_k} \cdot \sum\limits_{j \in Z_k} \left(\widetilde{x}_j^{(i)} - x_j\right)^2$.

As we have mentioned, in general, this system has several solutions: for example, we can select any measuring instrument i_0 and take $x_j = \widetilde{x}_j^{(i_0)}$ and $\sigma_k^{(i_0)} = 0$ for all j and k. As we will see, in addition to these solutions, we can also have physically meaningful solutions, for which $\sigma_k^{(i)} > 0$ for all i and k.

The idea is to select, among all such solutions, the one for which the entropy of the corresponding probability distribution $S \stackrel{\text{def}}{=} -\int \rho(x) \cdot \ln(\rho(x))\, dx$ is the largest. The idea that in the case of uncertainty, we should select a distribution with the largest entropy is actively used in probability and statistics; see, e.g., [6].

Why Maximum Entropy idea works here

Let us explain why the Maximum Entropy idea always leads to a solution for which $\sigma_k^{(i)} > 0$ for all i and k. Indeed, here, all the measurement errors are independent random variables. It is known that for independent variables, the entropy of a joint distribution is equal to the sum of the entropies of individual distributions. Each individual distribution is Gaussian, with the probability distribution

$$\rho(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

By definition of the expected value $E[f(x)] \stackrel{\text{def}}{=} \int \rho(x) \cdot f(x)\, dx$, the entropy S of this distribution is the expected value $E[\psi(x)]$ of the function

$$\psi(x) = -\ln(\rho(x)) = \ln\left(\sqrt{2\pi}\right) + \ln(\sigma) + \frac{(x-\mu)^2}{2\sigma^2}.$$

Here, $\ln(\sqrt{z}) = \frac{1}{2} \cdot \ln(z)$, and the expected value of $(x-\mu)^2$ is, by definition, equal to the variance σ^2. Thus, we get

$$E\left[\frac{(x-\mu)^2}{2\sigma^2}\right] = \frac{\sigma^2}{2\sigma^2} = \frac{1}{2},$$

and so,

$$S = E[\psi(x)] = \frac{1}{2} \cdot \ln(2\pi) + \ln(\sigma) + \frac{1}{2}.$$

When $\sigma \to 0$, we get $\ln(\sigma) \to -\infty$. Thus, if one of the values $\sigma_k^{(i)}$ is equal to 0, the corresponding entropy is equal to $-\infty$ and therefore, the whole sum is equal to $-\infty$, i.e., to the smallest possible value. So, when the entropy is the largest, we have to have $S > -\infty$, and hence, we have $\sigma_k^{(i)} > 0$ for all i and k.

Example of a physically meaningful solution

Let us show, on the above toy example, that the above system of equations has a physically meaningful solution, i.e., a solution for which $\sigma_k^{(i)} > 0$ for all i and k. In this toy example, we have measurement results $\widetilde{x}_1^{(i)} = x_0 + y_i \cdot \Delta$, where $y_1 = -1$, $y_2 = \theta$ for some $\theta \in (-1, 1)$, and $y_3 = 1$. In other words, we have $\widetilde{x}_1^{(1)} = x_0 - \Delta$, $\widetilde{x}_1^{(2)} = x_0 + \theta \cdot \Delta$, and $\widetilde{x}_1^{(3)} = x_0 + \Delta$. To make computations easier, let us represent the resulting fused value x_1 in a similar way, as $x_1 = x_0 + y \cdot \Delta$; we can do it if we take $y \stackrel{\text{def}}{=} \dfrac{x_1 - x_0}{\Delta}$. For these values $\widetilde{x}_1^{(i)}$, we have $\sigma_1^{(i)} = \left|\widetilde{x}_1^{(i)} - x_1\right|$, i.e., $\sigma_1^{(1)} = \Delta \cdot |1 + y|$, $\sigma_1^{(2)} = \Delta \cdot |y - \theta|$, and $\sigma_1^{(3)} = \Delta \cdot |1 - y|$.

The formula for x_1 has the form

$$x_1 \cdot \sum_{i=1}^{2} \left(\sigma_1^{(i)}\right)^{-2} = \sum_{i=1}^{2} \widetilde{x}_1^{(i)} \left(\sigma_1^{(i)}\right)^{-2}.$$

Substituting the above expressions for x_1 and $\widetilde{x}_1^{(i)}$ into this formula, we get

$$(x_0 + y \cdot \Delta) \cdot \sum_{i=1}^{2} \left(\sigma_1^{(i)}\right)^{-2} = \sum_{i=1}^{2} (x_0 + y_i \cdot \Delta) \cdot \left(\sigma_1^{(i)}\right)^{-2}.$$

The terms proportional to x_0 in both parts are the same, so we can cancel them from both parts, and get a simplified equation

$$y \cdot \Delta \cdot \sum_{i=1}^{2} \left(\sigma_1^{(i)}\right)^{-2} = \sum_{i=1}^{2} y_i \cdot \Delta \cdot \left(\sigma_1^{(i)}\right)^{-2}.$$

Dividing both sides by Δ, we get:

$$y \cdot \sum_{i=1}^{2} \left(\sigma_1^{(i)}\right)^{-2} = \sum_{i=1}^{2} y_i \cdot \left(\sigma_1^{(i)}\right)^{-2}.$$

Substituting the above expressions for $\sigma_1^{(i)}$ and multiplying both sides by Δ^2, we conclude that

$$y \cdot \left(\frac{1}{(1+y)^2} + \frac{1}{(y-\theta)^2} + \frac{1}{(1-y)^2}\right) = -\frac{1}{(1+y)^2} + \frac{\theta}{(y-\theta)^2} + \frac{1}{(1-y)^2}.$$

Here, by adding and subtracting two fractions, we get

$$\frac{1}{(1+y)^2}+\frac{1}{(1-y)^2}=\frac{1-2y+y^2+1+2y+y^2}{(1-y^2)^2}=\frac{2\cdot(1+y^2)}{(1-y^2)^2}$$

and

$$\frac{1}{(1-y)^2}-\frac{1}{(1+y)^2}=\frac{1+2y+y^2-(1-2y+y^2)}{(1-y^2)^2}=\frac{4y}{(1-y^2)^2}.$$

Thus, the above equation has the form

$$y\cdot\left(\frac{2\cdot(1+y^2)}{(1-y^2)^2}+\frac{1}{(y-\theta)^2}\right)=\frac{4y}{(1-y^2)^2}+\frac{\theta}{(y-\theta)^2}.$$

Bringing both sides to the common denominator, we get

$$2y\cdot(1+y^2)\cdot(y-\theta)^2+y\cdot(1-y^2)^2=4y\cdot(y-\theta)^2+\theta\cdot(1-y^2)^2.$$

Moving all the terms to the left-hand side and using the fact that the terms $y\cdot(1-y^2)^2$ and $\theta\cdot(1-y^2)^2$ have a common factor $(1-y^2)^2$, we conclude that

$$2y\cdot(1+y^2)\cdot(y-\theta)^2+(y-\theta)\cdot(1-y^2)^2-4y\cdot(y-\theta)^2=0.$$

We are looking for physically meaningful solutions, i.e., solutions for which $\sigma_k^{(i)}>0$ for all i and k. In particular, this means that $\sigma_1^{(2)}=\Delta\cdot|y-\theta|>0$ and thus, that $y-\theta\neq 0$. So, we can divide both sides of the above equality by $y-\theta$ and get the following simplified equation

$$2y\cdot(1+y^2)\cdot(y-\theta)+(1-y^2)^2-4y\cdot(y-\theta)=0.$$

The two terms proportional to $y-\theta$ can be combined into a single term

$$2y\cdot(1+y^2)\cdot(y-\theta)-4y\cdot(y-\theta)=2y\cdot(1+y^2-2)\cdot(y-\theta)=-2y\cdot(1-y^2)\cdot(y-\theta).$$

Thus, the above equation takes the form

$$-2y\cdot(1-y^2)\cdot(y-\theta)+(1-y^2)^2=0.$$

Since we assume that $\sigma_1^{(1)}=\Delta\cdot|1+y|>0$ and $\sigma_1^{(3)}=\Delta\cdot|1-y|>0$, we conclude that $y\neq -1$ and $y\neq 1$ and thus, $1-y^2\neq 0$. By dividing both sides of the above equation by $1-y^2$, we get

$$-2y\cdot(y-\theta)+(1-y^2)=0.$$

If we open parentheses and change the sign, we get the following quadratic equation:

$$2y^2-2\theta\cdot y-1+y^2=3y^2-2\theta\cdot y-1=0.$$

The discriminant $D = 4\theta^2 + 12$ of this quadratic equation is positive, so it has two solutions

$$y = \frac{2\theta \pm \sqrt{4\theta^2 + 12}}{6} = \frac{\theta \pm \sqrt{\theta^2 + 3}}{3}.$$

Comment.

In particular, for $\theta = 0$, we get two solutions $y = \frac{\sqrt{3}}{3}$ and $y = -\frac{\sqrt{3}}{3}$ corresponding to $x_1 = x_0 + \frac{\sqrt{3}}{3} \cdot \Delta$ and $x_1 = x_0 - \frac{\sqrt{3}}{3} \cdot \Delta$.

Neither of these two solutions reproduces the symmetry of this situation, since the three values $\widetilde{x}_1^{(i)}$ are, in this case, symmetric around $\widetilde{x}_1^{(2)} = x_0$. However, this is OK, since the only value x_0 which is symmetric under this symmetry transformation $x \to x_0 - (x - x_0)$ is the value $x_1 = x_0$ itself, which corresponds to the physically meaningless case $\sigma_1^{(2)} = \left|\widetilde{x}_1^{(2)} - x_1\right| = 0$.

A better solution: taking expert knowledge into account

The information that all the values $\sigma_k^{(i)}$ should be positive (and not too small) constitutes the additional expert knowledge. It is therefore desirable to explicitly take this additional expert knowledge into account.

Bayesian approach: idea

From the statistical viewpoint, a natural idea is to use *Bayesian* approach, i.e., to assume some prior distribution $\rho_\sigma(z)$ on the set of all possible values of $\sigma_k^{(i)}$ that would make very small values improbable. In this case, instead of maximizing the above-described probability density ρ, we maximize the posterior likelihood

$$L(x, \sigma) = \rho(x, \sigma) \cdot \prod_{i,k} \rho_\sigma\left(\sigma_k^{(i)}\right).$$

Since the value $\sigma_k^{(i)}$ is always positive, according to the usual statistics practice, a natural distribution for this value is *lognormal*, i.e., a distribution in which the logarithm $\ln\left(\sigma_k^{(i)}\right)$ is normally distributed.

Bayesian approach: limitations

The main limitation is that the result of using the Bayesian approach depends on the selection of the prior distribution. There are many other possible prior distributions for a positive value $\sigma_k^{(i)}$, and different distributions lead to different estimates for x_j.

The second limitation is that it is not a panacea. We tried the lognormal distribution with seemingly reasonable parameters, and we got value of $\sigma_k^{(i)}$ which were positive but still un-physically small. Of course, in

principle, we can manipulate the parameters of the distribution until we start getting physically meaningful results, but if we do that, we get, in effect, a new heuristic method – and we already have two heuristic methods that work reasonably well.

Fuzzy approach

The main problem with the Bayesian approach is that the experts describe their additional knowledge not in terms of precise numbers – like numbers that form a prior distribution – but rather is terms of words from a natural language. For example, an expert may say that the standard deviation $\sigma_k^{(i)}$ (describing the accuracy of the measuring instrument) cannot be too small. To take this knowledge into account, it is therefore reasonable to use *fuzzy logic*, technique that have been invented explicitly for transforming such "fuzzy" natural language terms and rules into precise numbers – which can then be used by a computer; see, e.g., [7, 10].

In fuzzy logic, the degree to which a certain value x satisfies a given property is described by a membership function $\mu(x)$. We can describe the degree to which the values x and σ are consistent with observations as proportional to the corresponding likelihood function ρ, and we can interpret "and" (as in "the first value $\sigma_1^{(1)}$ satisfies the property described by an expert, *and* the second value $\sigma_1^{(2)}$ satisfies the property described by an expert, etc"). In this case, the degree $d(x, \sigma)$ to which a given pair of tuples (x, σ) is consistent both with the observations and with the expert's knowledge is proportional to the product

$$d(x, \sigma) = \rho(x, \sigma) \cdot \prod_{i,k} \mu\left(\sigma_k^{(i)}\right).$$

It is then reasonable to select a pair (x, σ) for which this degree is the largest.

It is worth mentioning that from the purely mathematical viewpoint, the resulting optimization problem is exactly the same as in the Bayesian approach, but from the practical viewpoint, we have a big advantage: instead of the reasonably arbitrary difficult-to-get prior probabilities $\rho_0(z)$, we now have membership functions $\mu(z)$ which can be determined by one of several knowledge elicitation procedures [7, 10].

Acknowledgements

This work was supported in part by the National Science Foundation grants HRD-0734825 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by Grant 1 T36 GM078000-01 from the National Institutes of Health. The authors are thankful to Dr. Peter Moschopoulos for helpful discussions.

References

1. Akhiezer AI, Berestetskii VB: *Quantum Electrodynamics*, New York: Interscience Publishers, 1st edition 1965; last edition 2008.
2. Averill MG: *A Lithospheric Investigation of the Southern Rio Grande Rift*, *PhD Dissertation*, University of Texas at El Paso, Department of Geological Sciences, 2007.
3. Averill MG, Miller KC, Keller GR, Kreinovich V, Araiza R, Starks SA: **Using expert knowledge in solving the seismic inverse problem**, *International Journal of Approximate Reasoning*. 2007, **45**(3):564–587.
4. Gradshteyn IS, Ryzhik, IM: **Asymptotic Series**, Section 0.33 in *Tables of Integrals, Series, and Products*, 6th ed., San Diego, California: Academic Press 2000.
5. Hole JA: **Nonlinear high-resolution three-dimensional seismic travel time tomography**, *Journal of Geophysical Research*. 1992, **97**:6553–6562.
6. Jaynes ET: *Probability Theory: The Logic of Science*, Cambridge: Cambridge University Press, Cambridge 2003.
7. Klir G, Yuan B: *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, Upper Saddle River, New Jersey: Prentice Hall 1995.
8. Lees JM, Crosson RS: **Tomographic inversion for three-dimensional velocity structure at Mount St. Helens using earthquake data**, *Journal of Geophysical Research*. 1989, **94**:5716–5728.
9. Maceira M, Taylor SR, Ammon CJ, Yang X, Velasco AA: **High-resolution Rayleigh wave slowness tomography of Central Asia**, *Journal of Geophysical Research*, 2005, **110**:Paper B06304.
10. Nguyen HT, Walker EA: *First Course in Fuzzy Logic* Boca Raton, Florida: CRC Press 2006.
11. Ochoa O: **Towards a fast practical alternative to joint inversion of multiple datasets: model fusion**, *Abstracts of the 2009 Annual Conference of the Computing Alliance of Hispanic-Serving Institutions CAHSI*, Mountain View, California, January 15–18, 2009.
12. Ochoa O, Velasco AA, Kreinovich V, Servin C: **Model fusion: a fast, practical alternative towards joint inversion of multiple datasets**, *Abstracts of the Annual Fall Meeting of the American Geophysical Union AGU'08*, San Francisco, California, December 15–19, 2008.
13. Ochoa O, Velasco AA, Servin C, Kreinovich V: **Model Fusion under Probabilistic and Interval Uncertainty, with Application to Earth Sciences**, *International Journal of Reliability and Safety*. 2012, **6**:167–187.
14. Owen WM Jr: **On combining data from two complete star catalogs**, *Astronomical Journal*. 1990, **99**:1014–1015.
15. Rabinovich S: *Measurement Errors and Uncertainties: Theory and Practice*. New York: Springer Verlag, 2005.
16. Servin C, Ochoa O, Velasco AA: **Probabilistic and interval uncertainty of the results of data fusion, with application to geosciences**, *Abstracts of 13th International Symposium on Scientific Computing, Computer Arithmetic, and Verified Numerical Computations SCAN'2008*, El Paso, Texas, September 29 – October 3, 2008:128.
17. Sheskin DG: *Handbook of Parametric and Nonparametric Statistical Procedures*, Boca Raton, Florida: Chapman & Hall/CRC Press 2007.