Ubiquity of Data and Model Fusion: from Geophysics and Environmental Sciences to Estimating Individual Risk During an Epidemic

Omar Ochoa, Aline Jaimes, Christian Servin,
Craig Tweedie, Aaron Velasco, Martine Ceberio,
and Vladik Kreinovich
Cyber-ShARE Center, University of Texas at El Paso
500 W. University, El Paso, TX 79968, USA
Contact email: vladik@utep.edu

***Abstract*—In many practical situations, we need to combine the results of measuring a local value of a certain quantity with results of measuring average values of this same quantity. For example, in geosciences, we need to combine the seismic models (which describe density at different locations and depths) with gravity models which describe density averaged over certain regions. Similarly, in estimating the risk of an epidemic to an individual, we need to combine probabilities describe the risk to people of the corresponding age group, to people of the corresponding geographical region, etc. In this paper, we provide general techniques for solving such *model fusion* problems.**

To properly perform data and model fusion, we need to know the accuracy of different data points. Sometimes, this accuracy is not given. For such situations, we describe how this accuracy can be estimated based on the available data.

I. Formulation of the General Problem

Need for data fusion: reminder. In many real-life situations, we have several measurements and/or expert estimates $\widetilde{x}^{(1)}, \ldots, \widetilde{x}^{(n)}$ of the same quantity x.

- These values may come from the actual (direct) measurements of the quantity x.
- Alternatively, these values may come from *indirect* measurements of x, i.e., from different models, in which, based on the corresponding measurement results, the i-th model leads to an estimate $\widetilde{x}_i$ for x.

In such situations, it is desirable to fuse these estimates into a single more accurate estimate for x; see, e.g., [6].

A typical situation in measurement practice is when each estimation error $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x$ is normally distributed with 0 mean and known standard deviation σ_i, and estimation errors Δx_i corresponding to different models are independent.

Comment. In practice, the estimation errors are indeed often normally distributed. This empirical fact 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., [8]. 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.

Data fusion: formulas. For the normal distribution, the probability density for each estimation error Δx_i has the form

$$\frac{1}{\sqrt{2\cdot\pi}\cdot\sigma_i}\cdot\exp\left(-\frac{(\Delta x_i)^2}{2\cdot(\sigma_i)^2}\right) = \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma_i}\cdot\exp\left(-\frac{(\widetilde{x}_i - x)^2}{2\cdot(\sigma_i)^2}\right),$$

and the probability density $\rho(x)$ corresponding to all n estimates is (due to independence) the product of these densities:

$$\rho(x) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\cdot\pi}\cdot\sigma_i}\cdot\exp\left(-\frac{(\widetilde{x}_i - x)^2}{2\cdot(\sigma_i)^2}\right) = \left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\cdot\pi}\cdot\sigma_i}\right)\cdot\exp\left(-\sum_{i=1}^{n}\frac{(\widetilde{x}_i - x)^2}{2\cdot(\sigma_i)^2}\right).$$

As a single estimate x for the desired quantity, it is reasonable to select the value for which this 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: find x for which the sum $\sum\limits_{i=1}^{n}\frac{(\widetilde{x}_i - x)^2}{2\cdot(\sigma_i)^2}$ is the smallest possible.

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

$$x = \frac{\sum\limits_{i=1}^{n}\widetilde{x}_i\cdot(\sigma_i)^{-2}}{\sum\limits_{i=1}^{n}(\sigma_i)^{-2}}.$$

The accuracy of this fused estimate can be described by its standard deviation σ:

$$\sigma^{-2} = \sum_{i=1}^{n}(\sigma_i)^{-2}.$$

Need to go from data fusion to model fusion. In the case of data fusion, we fuse several results of measuring the same quantity. In practice, often, some of the results measure the current value of quantity, while other results measure the *average* of this quantity – over time, over a spatial region, or over a group. In this case, we cannot simply fuse the values one-by-one, we need to take into account the values of the given quantity at different points. In other words, instead of fusing *data points*, we need to fuse *models* that describe how data changes with location and/or with time.

Let us give a few examples where such *model fusion* is needed.

First example: geophysics. One of the main objectives of geophysics is to find out the density at different depths at different locations. There are many different sources of such information. For example, we can perform active seismic experiments: make an explosion, and measure the travel time during which the resulting seismic waves propagate to different sensors. Based on the observed travel-times, we can determine the velocity with which the seismic waves propagate, and this will enables us to predict the corresponding density values; see, e.g., [2].

Alternatively, we can measure the exact values of the gravitation field at different locations, and try to reconstruct density at different depths and different locations based on these measurement results. The problem here is that gravitation is a reasonably weak field, in the sense that it is very difficult to measure the effect of a small part of the Earth. Our measuring instruments can only detect the joint effect of reasonably big-size areas. Thus, each gravity measurement does not measure the density in a single point, it measures the *average* density at a big area.

Combining seismic measurements (which reflect local values) with gravity measurements (which reflect averages) is therefore a typical problem of model fusion. Mathematically, the problem is challenging, but the results are promising; see, e.g., [3], [4], [5], [7].

Second example: predicting voters' behavior. We have statistics describing how people from different social groups vote. The problem is that a person usually belongs to several different voting groups, groups with different voting behavior. For example, to predict how a female Hispanic professor with a certain annual salary will vote, we can look at how females vote, how Hispanics vote, how people of this level of income vote, etc. We need to combine these probabilities into a single estimate.

Each of these probabilities corresponds to an average over a group, so, in our terminology, this is a clear problem of model fusion.

Comment. In geophysics, the need for model fusion is clear: it is difficult to get direct information about the deep Earth structure, so we have to use all the information we can, and it so happens that a significant part of this information is about averages.

For voting, this need may be less clear: why not take a homogeneous group? The main reason is that all people are different. If we try to narrow our sample down to a very homogeneous group, we will end up with a sample which is so small that it is impossible to make statistically meaningful predictions. To make predictions, we thus need to consider larger sample; these samples are heterogeneous, and thus, the corresponding frequencies do not fully reflect the preferences of the voter of interest. Therefore, we need to fuse these probabilities.

Estimating the threat of an epidemic. Each individual faces the same problem when estimating the threat of an epidemic – and thus, e.g., deciding on whether appropriate prophylactic measures are in order. We usually have several types of data about this threat:

- we have anecdotal evidence about our close friends, how many of them got, e.g., flu; this is probably the most appropriate group – but this group is usually too small to produce statistically meaningful estimates of the probability;
- we also have probability with which people of a given age group get this disease, probability with which people from this geographic area get this disease, etc.

To make a meaningful decision, we need to combine all these estimates into a single one. These estimates describe the behavior of different groups containing a given individual, so this is a typical problem of model fusion.

What we do in this paper. In this paper, we extend our previous work [3], [4], [5], [7] – which was based on *specific* geophysical applications of data fusion – and provide *general* techniques for model fusion.

II. Solving the Problem: General Idea

Model fusion: reminder. In model fusion, to estimate the value v of a certain quantity in a given situation, we use several estimates v_i of averages of v over different groups that contain this situation.

General idea. The very fact that we can use the estimations v_i of the averages to gauge the desired value v means that these averages v_i are approximations to v. So, a reasonable idea is to estimate the standard deviation σ_i of the corresponding approximation.

Once we know these standard deviations σ_i, we can use the above data fusion formulas to fuse these estimates v_i into a single estimate for the desired quantity v.

What remains to be done. To use this idea, we need to be able to estimate the standard deviations σ_i. Let us analyze how this can be done.

Outline.

- We will start our analysis with the simplest case, when we can ignore the measurement inaccuracy and assume that all the averages are exactly known.

- After that, we will consider the case when the measurement inaccuracy is known. As a particular case, we will consider estimates based on observed frequencies – like in the voting and epidemic situations.
- Finally, we will discuss what can be done if the measurement inaccuracy is not known.

III. Case When Measurements are Exact

Description of the case. Let us start with the case when measurement errors can be ignored, i.e., when we can safely assume that the measurements are exact.

Natural solution. In this case, a natural measure of the difference between the average v_i over the group and individual values within this group is the within-group standard deviation σ_i. In other words, if we have a sample of individual values $v^{(1)}, \ldots, v^{(M)}$ within the i-th group, then, as the desired accuracy, we can take the sample standard deviation: $\sigma_i = \sigma_{\text{samp},i}$, where

$$\sigma_{\text{samp},i} \stackrel{\text{def}}{=} \sqrt{\frac{1}{M} \cdot \sum_{k=1}^{M} \left(v^{(k)} - E_{\text{samp},i}\right)^2},$$

and

$$E_{\text{samp},i} \stackrel{\text{def}}{=} \frac{1}{M} \cdot \sum_{k=1}^{M} v^{(k)}.$$

IV. Case When We Know Measurement Accuracy

Description of the case. Let us now consider the case when we know the accuracy σ_{ind} with which we measure individual values, and we also know the accuracy $\sigma_{\text{meas},i}$ with which we measure the i-th average.

Analysis of the problem. In this case, there are two independent reasons why the observed values $v^{(k)}$ differ from the average:

- first, due to measurement errors, and
- second, due to diversity within the group.

In precise terms, each observed value $v^{(k)}$ can be represented as

$$v^{(k)} = v_{\text{act}}^{(k)} + \Delta v^{(k)},$$

where $v_{\text{act}}^{(k)}$ is the actual (unknown) value of the corresponding quantity and $\Delta v^{(k)} \stackrel{\text{def}}{=} v^{(k)} - v_{\text{act}}^{(k)}$ is the measurement error. The actual value $v_{\text{act}}^{(k)}$, in its turn, can be represented as the sum

$$v_{\text{act}}^{(k)} = v_{\text{act},i} + (v_{\text{act}}^{(k)} - v_{\text{act},i})$$

of the actual average value $v_{\text{act},i}$ and the deviation

$$v_{\text{act}}^{(k)} - v_{\text{act},i}$$

of the actual individual values from the actual average. Combining these two formulas, we get

$$v^{(k)} = v_{\text{act},i} + (v_{\text{act}}^{(k)} - v_{\text{act},i}) + \Delta v^{(k)}.$$

In other words, the difference $v^{(k)} - v_{\text{act},i}$ between the observed individual values $v^{(k)}$ and the actual average $v_{\text{act},i}$ is caused by two independent factors:

$$v^{(k)} - v_{\text{act},i} = (v_{\text{act}}^{(k)} - v_{\text{act},i}) + \Delta v^{(k)}.$$

Since these factors are independent, the total within-sample variance $\sigma_{\text{samp},i}^2$ is equal to the sum of the two variances corresponding to individual factors:

- the variance $\sigma_{\text{act},i}^2$ which describes the deviation of the actual individual values from the average (and from each other), and
- the variance σ_{ind}^2 which describes the measurement error of individual measurements.

In other words, $\sigma_{\text{samp},i}^2 = \sigma_{\text{act},i}^2 + \sigma_{\text{ind}}^2$. We can estimate the within-sample variance $\sigma_{\text{samp},i}^2$, and, in the current case, we know the measurement accuracy σ_{ind}^2. Thus, we can estimate the variance $\sigma_{\text{act},i}^2$ (which describes the deviation of the actual individual values from the average) as

$$\sigma_{\text{act},i}^2 = \sigma_{\text{samp},i}^2 - \sigma_{\text{ind}}^2.$$

How does this accuracy translates into the accuracy with which the observed average v_i approximates the desired value v? Similarly to the above case, the difference $v - v_i$ is caused by two factors:

- first, due to diversity, the value v, in general, differs from the actual average $v_{\text{act},i}$;
- second, due to measurement errors, the observed value v_i of the average is, in general, different from the actual average $v_{\text{act},i}$.

In precise terms, we have

$$v - v_i = (v - v_{\text{act},i}) + (v_{\text{act},i} - v_i).$$

So, the difference $v - v_i$ is the sum of two independent random terms:

- the first term $v - v_{\text{act},i}$ has standard deviation $\sigma_{\text{act},i}$, and
- the second term $v_{\text{act},i} - v_i$ has standard deviation $\sigma_{\text{meas},i}$.

Since these factors are independent, the total variance σ_i^2 of the difference $v - v_i$ is equal to the sum of the two variances corresponding to these two terms:

$$\sigma_i^2 = \sigma_{\text{act},i}^2 + \sigma_{\text{meas},i}^2.$$

Substituting the above expression for $\sigma_{\text{act},i}^2$ into this formula, we arrive at the following formula:

$$\sigma_i^2 = \sigma_{\text{samp},i}^2 - \sigma_{\text{ind}}^2 + \sigma_{\text{meas},i}^2.$$

Resulting estimate. To find the standard deviation σ_i of the difference $v - v_i$, we first compute the sample standard deviation

$$\sigma_{\text{samp},i}^2 = \frac{1}{M} \cdot \sum_{k=1}^{M} \left(v^{(k)} - E_{\text{samp},i}\right)^2,$$

where

$$E_{\text{samp},i} = \frac{1}{M} \cdot \sum_{k=1}^{M} v^{(k)},$$

and then take

$$\sigma_i^2 = \sigma_{\text{samp},i}^2 - \sigma_{\text{ind}}^2 + \sigma_{\text{meas},i}^2.$$

V. Important Particular Case: Estimates Based on Observed Frequencies

Description of the situation. In situations like epidemic estimates or voting predictions, the value v_i are frequencies based on the i-th group.

How to estimate accuracy of frequency estimates. According to statistics (see, e.g., [8]), the accuracy $\sigma_{\text{meas},i}^2$ with which the observed frequency v_i estimates the actual probability is equal to

$$\sigma_{\text{meas},i}^2 = \frac{v_i \cdot (1 - v_i)}{n_i},$$

where n_i is the number of elements in the i-th group.

To estimate the within-sample variance, we can subdivide the sample of n_i elements into several (M) subgroups of smaller size n_i/M, and estimate the frequency $v_i^{(k)}$ within each subgroup k, $k = 1, \ldots, M$. Here, the accuracy of σ_{ind}^2 of individual measurements is equal to

$$\sigma_{\text{ind}}^2 = \frac{v_i \cdot (1 - v_i)}{n_i/M}.$$

Thus, the above formula $\sigma_i^2 = \sigma_{\text{samp},i}^2 - \sigma_{\text{ind}}^2 + \sigma_{\text{meas},i}^2$ takes the following form:

$$\sigma_i^2 = \sigma_{\text{samp},i}^2 - \frac{v_i \cdot (1 - v_i)}{n_i/M} + \frac{v_i \cdot (1 - v_i)}{n_i}.$$

VI. What If We Do Not Know the Measurement Accuracy

Formulation of the problem. In the previous sections, we assumed that we have a good description of the uncertainty of the original data. In practice, often, we do not have this information, we need to extract it from the data.

Extracting uncertainty from data: traditional approach. The usual way to gauge of the uncertainty of the measuring instrument is to compare the result $\widetilde{x}$ produced by this measuring instruments with the result $\widetilde{x}_s$ of measuring the same quantity x by a much more accurate ("standard") measuring instrument.

Since the "standard" measuring instrument is much more accurate than the instrument that we are trying to calibrate, we can safely ignore the inaccuracy of its measurements and take $\widetilde{x}_s$ as a good approximation to the actual value x. In this case, the difference $\widetilde{x} - \widetilde{x}_s$ between the measurement results can serve as a good approximation to the desired measurement accuracy $\Delta x = \widetilde{x} - x$.

Traditional approach cannot be applied for calibrating state-of-the-art measuring instruments. The above traditional approach works well for many measuring instruments. However, we cannot apply this approach for calibrating state-of-the-art instrument, because these instruments are the best we have. There are no other instruments which are much more accurate than these ones – and which can therefore serve as standard measuring instruments for our calibration.

Such situations are ubiquitous; for example:

- in the environmental sciences, we want to gauge the accuracy with which the Eddy covariance tower measure the Carbon and heat fluxes; see, e.g., [1]
- in the geosciences, we want to gauge how accurately seismic, gravity, and other techniques reconstruct the density at different depths and different locations.

How state-of-the-art measuring instruments are calibrated: case of normally distributed measurement errors. Calibration of state-of-the-art measuring instruments is possible if we make a usual assumption that the measurement errors are normally distributed with mean 0. Under this assumption, to fully describe the distribution of the measurement errors, it is sufficient to estimate the standard deviation σ of this distribution.

There are two possible approaches for estimating this standard deviation. The first approach is applicable when we have several similar measuring instruments. For example, we can have two nearby towers, or we can bring additional sensors to the existing tower. In such a situation, instead of a single measurement result $\widetilde{x}$, we have two different results $\widetilde{x}^{(1)}$ and $\widetilde{x}^{(2)}$ of measuring the same quantity x. Here, by definition of the measurement error, $\widetilde{x}^{(1)} = x + \Delta x^{(1)}$ and $\widetilde{x}^{(2)} = x + \Delta x^{(2)}$ and therefore,

$$\widetilde{x}^{(1)} - \widetilde{x}^{(2)} = \Delta x^{(1)} - \Delta x^{(2)}.$$

Each of the random variables $\Delta x^{(1)}$ and $\Delta x^{(2)}$ is normally distributed with mean 0 and (unknown) standard deviation σ (i.e., variance σ^2). Since the two measuring instruments are independence, the corresponding random variables $\Delta x^{(1)}$ and $\Delta x^{(2)}$ are also independent, and so, the variance of their difference is equal to the sum of their variances $\sigma^2 + \sigma^2 = 2\sigma^2$. Thus, the standard deviation σ' of this difference is equal to $\sqrt{2} \cdot \sigma$. We can estimate this standard deviation σ' based on the observed differences $\widetilde{x}^{(1)} - \widetilde{x}^{(2)}$ and therefore, we can estimate σ as $\dfrac{\sigma'}{\sqrt{2}}$.

This approach is not applicable in the geosciences applications, when we usually have only one seismic map, only one gravity map, etc. In such situations, we have several measurement results $\widetilde{x}^{(i)}$ with, in general, different standard deviations $\sigma^{(i)}$. For every two measuring instruments i and j, the difference $\widetilde{x}^{(i)} - \widetilde{x}^{(j)}$ is normally distributed with the variance $\left(\sigma^{(i)}\right)^2 + \left(\sigma^{(j)}\right)^2$. By comparing actual measurement results, we can estimate this variance and thus, get an estimate e_{ij} for the sum. As a result, e.g., for the case when we have three different measuring instruments, we get three values e_{ij} for which:

$$e_{12} = \left(\sigma^{(1)}\right)^2 + \left(\sigma^{(2)}\right)^2;$$

$$e_{13} = \left(\sigma^{(1)}\right)^2 + \left(\sigma^{(3)}\right)^2;$$

$$e_{23} = \left(\sigma^{(2)}\right)^2 + \left(\sigma^{(3)}\right)^2.$$

Here, we have a system of three linear equations with three unknowns, from which we can uniquely determined all three desired variances $\left(\sigma^{(i)}\right)^2$:

$$\left(\sigma^{(1)}\right)^2 = \frac{e_{12} + e_{13} - e_{23}}{2};$$

$$\left(\sigma^{(2)}\right)^2 = \frac{e_{12} + e_{23} - e_{13}}{2};$$

$$\left(\sigma^{(3)}\right)^2 = \frac{e_{13} + e_{23} - e_{12}}{2}.$$

Need to go beyond normal distributions, and resulting problem. In practice, the distribution of measurement errors is often different from normal; this is the case, e.g., in measuring fluxes [1]. In such cases, we can still use the same techniques to find the standard deviation of the measurement error. However, in general, it is not enough to know the standard deviation to uniquely determine the distribution: e.g., we may have (and we sometimes do have) an asymmetric distribution, for which the skewness is different from 0 (i.e., equivalently, the expected value of $(\Delta x)^3$ is different from 0).

It is known that in this case, in contrast to the case of the normal distribution, we cannot uniquely reconstruct the distribution of Δx from the known distribution of the difference $\Delta x^{(1)} - \Delta x^{(2)}$. Indeed, if we have an asymmetric distribution for Δx, i.e., a distribution which is not invariant under the transformation $\Delta x \to -\Delta x$, this means that the distribution for $\Delta y \stackrel{\text{def}}{=} -\Delta x$ is different from the distribution for Δx. However, since

$$\Delta y^{(1)} - \Delta y^{(2)} = \Delta x^{(2)} - \Delta x^{(1)},$$

the y-difference is also equal to the difference between two independent variables with the distribution Δx and thus, distribution for the difference $\Delta y^{(1)} - \Delta y^{(2)}$ is exactly the same as for the difference $\Delta x^{(1)} - \Delta x^{(2)}$. In other words, if we know the distribution for the difference $\Delta x^{(1)} - \Delta x^{(2)}$, we cannot uniquely reconstruct the distribution for Δx, because, in addition to the original distribution for Δx, all the observations are also consistent with the distribution for $\Delta y = -\Delta x$.

This known non-uniqueness naturally leads to the following questions:

- first, a theoretical question: since we cannot uniquely reconstruct the distribution for Δx, what information about this distribution can we reconstruct?
- second, a practical question: for those characteristics of Δx which can be theoretically reconstructed, we need to design computationally efficient algorithms for reconstructing these characteristics.

Techniques to use. To solve these questions, let us use the Fourier analysis technique.

What we want to find is the probability density $\rho(z)$ describing the distribution of the measurement error $z \stackrel{\text{def}}{=} \Delta x$. In order to find the unknown probability density, we will first find its Fourier transform

$$F(\omega) = \int \rho(z) \cdot e^{\mathrm{i}\cdot\omega\cdot z}\, dz.$$

By definition, this Fourier transform is equal to the mathematical expectation of the function $e^{\mathrm{i}\cdot\omega\cdot z}$:

$$F(\omega) = E\left[e^{\mathrm{i}\cdot\omega\cdot z}\right].$$

Such a mathematical expectation is also known as a *characteristic function* of the random variable z.

Based on the observed values of the difference $z^{(1)} - z^{(2)}$, we can estimate the characteristic function $D(\omega)$ of this difference:

$$D(\omega) = E\left[e^{\mathrm{i}\cdot\omega\cdot(z^{(1)} - z^{(2)})}\right].$$

Here,

$$e^{\mathrm{i}\cdot\omega\cdot(z^{(1)} - z^{(2)})} = e^{(\mathrm{i}\cdot\omega\cdot z^{(1)}) + (-\mathrm{i}\cdot\omega\cdot z^{(2)})} = e^{\mathrm{i}\cdot\omega\cdot z^{(1)}} \cdot e^{-\mathrm{i}\cdot\omega\cdot z^{(2)}}.$$

Measurement errors $z^{(1)}$ and $z^{(2)}$ corresponding to two measuring instruments are usually assumed to be independent. Thus, the variables $e^{\mathrm{i}\cdot\omega\cdot z^{(1)}}$ and $e^{-\mathrm{i}\cdot\omega\cdot z^{(2)}}$ are also independent. It is known that the expected value of the product of two independent variables is equal to the product of their expected values, thus,

$$D(\omega) = E\left[e^{\mathrm{i}\cdot\omega\cdot z^{(1)}}\right] \cdot E\left[e^{-\mathrm{i}\cdot\omega\cdot z^{(2)}}\right],$$

i.e.,

$$D(\omega) = F(\omega) \cdot F(-\omega).$$

Here,

$$F(-\omega) = E\left[e^{-\mathrm{i}\cdot\omega\cdot z}\right] = E\left[\left(e^{\mathrm{i}\cdot\omega\cdot z}\right)^*\right],$$

where t^* means complex conjugation, i.e., an operation that transforms $t = a + b\cdot\mathrm{i}$ into $t^* = a - b\cdot\mathrm{i}$. Thus, $F(-\omega) = F^*(\omega)$, and the above formula takes the form

$$D(\omega) = F(\omega) \cdot F^*(\omega) = |F(\omega)|^2.$$

In other words, the fact that we know $D(\omega)$ means that we know the absolute value (modulus) of the complex-valued function $F(\omega)$.

In these terms, the problems becomes: how can we reconstruct the complex-valued function $F(\omega)$ if we only know its absolute value?

How to use Fourier techniques to solve the theoretical question. First, let us address the theoretical question: since, in general, we cannot reconstruct $\rho(z)$ (or, equivalently, $F(\omega)$) uniquely, what information about $\rho(z)$ (and, correspondingly, about $F(\omega)$) can we reconstruct?

To solve this theoretical question, let us take into account the practical features of this problem. First, it needs to be mentioned that, from the practical viewpoint, we need to take into account that the situation in, e.g., Eddy covariance tower measurements is more complex that we described, because the

tower does not measure *one* single quantity, it simultaneously measuring *several* quantities: carbon flux, heat flux, etc. Since these different measurements are based on data from the same sensors, it is reasonable to expect that the resulting measurement errors are correlated. Thus, to fully describe the measurement uncertainty, it is not enough to describe the distribution of each 1-D measurement error, we need to describe a joint distribution of all the measurement errors $z = (z_1, z_2, \ldots)$. In this multi-D case, we can use the multi-D Fourier transforms and characteristic functions, where for $\omega = (\omega_1, \omega_2, \ldots)$, we define

$$F(\omega) = E\left[e^{\mathrm{i}\cdot\omega\cdot z}\right],$$

with

$$\omega \cdot z \stackrel{\text{def}}{=} \omega_1 \cdot z_1 + \omega_2 \cdot z_2 + \ldots$$

Second, we need to take into account that while theoretically, we can consider all possible values of the difference $z^{(1)} - z^{(2)}$, in practice, we can only get values which are proportional to the smallest measuring unit h. For example, if we measure distance and the smallest distance we can measure is centimeters, then the measuring instrument can only return values 0 cm, 1 cm, 2 cm, etc. In other words, in reality, the value z can only take discrete values. If we take the smallest value of z as the new starting point (i.e., as 0), then the possible values of z take the form $z = 0$, $z = h$, $z = 2h$, …, until we reach the upper bound $z = N \cdot h$ for some integer N. For these values, in the 1-D case, the Fourier transform takes the form

$$F(\omega) = E\left[e^{\mathrm{i}\cdot\omega\cdot z}\right] = \sum_{k=0}^{N} p_k \cdot e^{\mathrm{i}\cdot\omega\cdot k\cdot h},$$

where p_k is the probability of the value $z = k\cdot h$. This formula can be equivalently rewritten as

$$F(\omega) = \sum_{k=0}^{N} p_k \cdot s^k,$$

where $s \stackrel{\text{def}}{=} e^{\mathrm{i}\cdot\omega\cdot h}$. Similarly, in the multi-D case, we have $z = (k_1 \cdot h_1, k_2 \cdot h_2, \ldots)$, and thus,

$$e^{\mathrm{i}\cdot\omega\cdot k\cdot h} = e^{\mathrm{i}\cdot\omega\cdot(k_1\cdot h_1 + k_2\cdot h_2 + \ldots)} = e^{\mathrm{i}\cdot\omega_1\cdot k_1\cdot h_1} \cdot e^{\mathrm{i}\cdot\omega\cdot k_2\cdot h_2} \cdot \ldots,$$

so we have

$$F(\omega) = \sum_{k_1=0}^{N_1} \sum_{k_2=0}^{N_2} \ldots p_k \cdot s_1^{k_1} \cdot s_2^{k_2} \cdot \ldots,$$

where $s_k \stackrel{\text{def}}{=} e^{\mathrm{i}\cdot\omega_k\cdot h_k}$. In other words, we have a polynomial of the variables $s_1, s_2, \ldots$:

$$P(s_1, s_2, \ldots) = \sum_{k_1=0}^{N_1} \sum_{k_2=0}^{N_2} \ldots p_k \cdot s_1^{k_1} \cdot s_2^{k_2} \cdot \ldots$$

Different values of ω correspond to different values of $s = (s_1, s_2, \ldots)$. Thus, the fact that we know the values of $|F(\omega)|^2$ for different ω is equivalent to knowing the values of $|P(s)|^2$ for all possible values $s = (s_1, s_2, \ldots)$.

In these terms, the theoretical question takes the following form: we know the values $D(s) = |P(s)|^2 = P(s) \cdot P^*(s)$ for some polynomial $P(s)$, we need to reconstruct this polynomial. In the 1-D case, each complex-valued polynomial of degree N has, in general, N complex roots $s^{(1)}$, $s^{(2)}$, etc., and can, therefore, be represented as

$$|P(s)|^2 = \text{const} \cdot (s - s^{(1)}) \cdot (s - s^{(2)}) \cdot \ldots$$

In this case, there are many factors, so there are many ways to represent it as a product – which explains the above-described non-uniqueness of representing $D(s)$ as the product of two polynomials $P(s)$ and $P^*(s)$

Interestingly, in contrast to the 1-D case, in which each polynomial can be represented as a product of polynomials of 1st order, in the multi-D case, a generic polynomial *cannot* be represented as a product of polynomials of smaller degrees. This fact can be easily illustrated on the example of polynomials of two variables.

To describe a general polynomial of two variables $\sum_{k=0}^{n} \sum_{l=1}^{n} c_{kl} \cdot s_1^k \cdot s_2^l$ in which each of the variables has a degree $\le n$, we need to describe all possible coefficients c_{kl}. Each of the indices k and l can take $n+1$ possible values 0, 1, …, n, so overall, we need to describe $(n+1)^2$ coefficients.

When two polynomials multiply, the degrees add: $s^m \cdot s^{m'} = s^{m+m'}$. Thus, if we represent $P(s)$ as a product of two polynomials, one of them must have a degree $m < n$, and the other one degree $n - m$. In general:

- we need $(m+1)^2$ coefficients to describe a polynomial of degree m and
- we need $(n-m+1)^2$ coefficients to describe a polynomial of degree $n - m$,
- so to describe arbitrary products of such polynomials, we need $(m+1)^2 + (n-m+1)^2$ coefficients.

To be more precise, in such a product, we can always multiply one of the polynomials by a constant and divide another one by the same constant, without changing the product. Thus, we can always assume that, e.g., in the first polynomial, the free term c_{00} is equal to 1. As a result, we need one fewer coefficient to describe a general product: $(m+1)^2 + (n-m+1)^2 - 1$.

To be able to represent a generic polynomial $P(s)$ of degree n as such a product

$$P(s) = P_m(s) \cdot P_{n-m}(s),$$

we need to make sure that the coefficients at all all $(n+1)^2$ possible degrees $s_1^k \cdot s_2^l$ are the same on both sides of this equation. This requirement leads to $(n+1)^2$ equations with $(m+1)^2 + (n-m+1)^2 - 1$ unknowns.

In general, a system of equations is solvable if the number of equations does not exceed the number of unknowns. Thus, we must have

$$(n+1)^2 \le (m+1)^2 + (n-m+1)^2 - 1.$$

Opening parentheses, we get

$$n^2 + 2n + 1 \le m^2 + 2m + 1 + (n-m)^2 + 2\cdot(n-m) + 1 - 1.$$

The constant terms in both sides cancel each other, as well as the terms $2n$ in the left-hand side and $2m + 2 \cdot (n - m) = 2n$ in the right-hand side, so we get an equivalent inequality

$$n^2 \le m^2 + (n - m)^2.$$

Opening parentheses, we get

$$n^2 \le m^2 + n^2 - 2 \cdot n \cdot m + m^2.$$

Canceling n^2 in both sides, we get

$$0 \le 2m^2 - 2 \cdot n \cdot m.$$

Dividing both sides by $2m$, we get an equivalent inequality $0 \le m - n$, which clearly contradicts to our assumption that $m < n$.

Let us go back to our problem. We know the product $D(s) = P(s) \cdot P^*(s)$, and we want to reconstruct the polynomial $P(s)$. We know that this problem is not uniquely solvable, i.e., that there exist other polynomials $Q(s) \ne P(s)$ for which $D(s) = P(s) \cdot P^*(s) = Q(s) \cdot Q^*(s)$. Since, in general, a polynomial $P(s)$ of several variables cannot be represented as a product – i.e., is "prime" in terms of factorization the same way prime numbers are – the fact that the two products coincide means that $Q(s)$ must be equal to one of the two prime factors in the decomposition $D(s) = P(s) \cdot P^*(s)$. Since we know that $Q(s)$ is different from $P(s)$, we thus conclude that $Q(s) = P^*(s)$.

By going back to the definitions, one can see that for the distribution $\rho'(x) = \rho(-x)$, the corresponding polynomial has exactly the form $Q(s) = P^*(s)$. Thus, in general, this is the *only* non-uniqueness that we have: each distribution which is consistent with the observation of differences coincides either with the original distribution $\rho(x)$ or with the distribution $\rho'(x) = \rho(-x)$. In other words, we arrive at the following result.

Answer to the theoretical question. We have proven that, in general, each distribution which is consistent with the observation of differences $\Delta x^{(1)} - \Delta x^{(2)}$ coincides either with the original distribution $\rho(x)$ or with the distribution

$$\rho'(x) \stackrel{\text{def}}{=} \rho(-x).$$

How to use Fourier techniques to solve the practical question: idea. We want to find a probability distribution $\rho(z)$ which is consistent with the observed characteristic function $D(\omega)$ for the difference. In precise terms, we want to find a function $\rho(z)$ which satisfies the following two conditions:

- $\rho(z) \ge 0$ for all z, and
- $|F(\omega)|^2 = D(\omega)$, where $F(\omega)$ denotes the Fourier transform of the function $\rho(x)$.

One way to find the unknown function that satisfies two conditions is to use the method of successive projections. In this method, we start with an arbitrary function $\rho^{(0)}(z)$. On the k-th iteration, we start with the result $\rho^{(k-1)}(z)$ of the previous iteration, and we do the following:

- first, we project this function $\rho^{(k-1)}(z)$ onto the set of all functions which satisfy the first condition; to be more precise, among all the functions which satisfy the first condition, we find the function $\rho'(x)$ which is the closest to $\rho^{(k-1)}(z)$;
- then, we project the function $\rho'(z)$ onto the set of all functions which satisfy the second condition; to be more precise, among all the functions which satisfy the second condition, we find the function $\rho^{(k)}(x)$ which is the closest to $\rho'(z)$.

We continue this process until it converges.

As the distance between the two functions $f(z)$ and $g(z)$ – describing how close they are – it is reasonable to take the natural analog of the Euclidean distance:

$$d(f, g) \stackrel{\text{def}}{=} \sqrt{\int (f(z) - g(z))^2 \, dz}.$$

One can check that for this distance function:

- the closest function in the first part of the iteration is the function $\rho'(z) = \max(0, \rho^{(k-1)}(z))$, and
- on the second part, the function whose Fourier transform is equal to

$$F^{(k)}(\omega) = \frac{\sqrt{|D(\omega)|}}{|F'(\omega)|} \cdot F'(\omega).$$

Thus, we arrive at the following algorithm.

How to use Fourier techniques to solve the practical question: algorithm. We start with an arbitrary function $\rho^{(0)}(z)$. On the k-th iteration, we start with the function $\rho^{(k-1)}(z)$ obtained on the previous iteration, and we do the following:

- first, we compute $\rho'(z) = \max(0, \rho^{(k-1)}(z))$;
- then, we apply Fourier transform to $\rho'(z)$ and get $F'(z)$;
- after that, we compute

$$F^{(k)}(\omega) = \frac{\sqrt{|D(\omega)|}}{|F'(\omega)|} \cdot F'(\omega);$$

- finally, as the next approximation $\rho^{(k)}(z)$, we take the result of applying the inverse Fourier transform to $F^{(k)}(\omega)$.

We continue this process until it converges.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, by Grant 1 T36 GM078000-01 from the National Institutes of Health, and by a grant on F-transforms from the Office of Naval Research.

The authors are thankful to all the members of the Interdisciplinary Working Group on Decision making, especially to Larry Cohn, for valuable discussions.

REFERENCES

[1] M. Aubinet, T. Vesala, and D. Papale (eds.), *Eddy Covariance – A Practical Guide to Measurement and Data Analysis*, Springer, Dordrecht, Hiedelberg, London, New York, 2012.

[2] J. A. Hole, "Nonlinear high-resolution three-dimensional seismic travel time tomography", *Journal of Geophysical Research*, 1992, Vol. 97, pp. 6553–6562.

[3] O. Ochoa, A. A. Velasco, V. Kreinovich, and C. Servin, "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.

[4] O. Ochoa, "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.

[5] O. Ochoa, A. A. Velasco, C. Servin, and V. Kreinovich, "Model Fusion under Probabilistic and Interval Uncertainty, with Application to Earth Sciences", *International Journal of Reliability and Safety*, 2012, Vol. 6, No. 1–3, pp. 167–187.

[6] S. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, American Institute of Physics, New York, 2005.

[7] C. Servin, O. Ochoa, and A. A. Velasco, "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, p. 128.

[8] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC Press, Boca Raton, Florida, 2011.