Towards a Fast, Practical Alternative to Joint Inversion of Multiple Datasets: Model Fusion

Omar Ochoa

Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
omar@miners.utep.edu
1. 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 the actual earthquakes);
  - first-arrival active seismic data (from the seismic experiments);
  - gravity data; and
  - surface waves.
2. Need to Combine Data (cont-d)

- Datasets coming from different sources provide complimentary information.

- Example: different geophysical datasets contain different information on earth structure.

- In general:
  - some of the datasets provide better accuracy and/or spatial resolution in some spatial areas;
  - other datasets provide a better accuracy and/or spatial resolution in other areas or depths.

- Example:
  - gravity measurements have (relatively) low resolution;
  - each seismic data point comes from a narrow trajectory of a seismic signal – so resolution is higher.
3. Joint Inversion: An Ideal Future Approach

- At present: each of the datasets is often processed separately.
- It is desirable: to combine data from different datasets.
- Ideal approach: use all the datasets to produce a single model.
- Problem: in many areas, there are no efficient algorithms for simultaneously processing all the datasets.
- Challenge: designing joint inversion techniques is an important theoretical and practical challenge.
4. Proposed Solution – Model Fusion: Main Idea

• **Reminder**: joint inversion methods are still being developed.

• **Practical solution**: to fuse the models coming from different datasets.

• **Simplest case – data fusion, probabilistic uncertainty**:  
  - we have several measurements (and/or expert estimates) $\tilde{x}_1, \ldots, \tilde{x}_n$ of the same quantity $x$.
  - each measurement error $\Delta x_i \equiv \tilde{x}_i - x$ is normally distributed with 0 mean and known st. dev. $\sigma_i$;
  - Least Squares: find $x$ that minimizes $\sum_{i=1}^{n} \frac{(\tilde{x}_i - x)^2}{2 \cdot \sigma_i^2}$;
  - solution: $x = \frac{\sum_{i=1}^{n} \tilde{x}_i \cdot \sigma_i^{-2}}{\sum_{i=1}^{n} \sigma_i^{-2}}$. 


5. **Data Fusion: Case of Interval Uncertainty**

- In some practical situations, the value $x$ is known with interval uncertainty.

- This happens, e.g., when we only know the upper bound $\Delta_i$ on each measurement error $\Delta x_i$: $|\Delta x_i| \leq \Delta_i$.

- In this case, we can conclude that $|x - \tilde{x}_i| \leq \Delta_i$, i.e., that $x \in x_i \stackrel{\text{def}}{=} [\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i]$.

- Based on each measurement result $\tilde{x}_i$, we know that the actual value $x$ belongs to the interval $x_i$.

- Thus, we know that the (unknown) actual value $x$ belongs to the intersection of these intervals:

$$x \stackrel{\text{def}}{=} \bigcap_{i=1}^{n} x_i = [\max(\tilde{x}_i - \Delta_i), \min(\tilde{x}_i + \Delta_i)].$$
6. Additional Problem: We Also Have Different Spatial Resolution

- In many situations, different measurements have not only different accuracy, but also different resolution.

- Example:
  - seismic data leads to higher-resolution estimates of the density at different locations and depths, while
  - gravity data leads to lower-estimates of the same densities.

- Towards precise formulation of the problem:
  - High-resolution measurements mean that we measure the values corresponding to small spatial cells.
  - A low-resolution measurement means that its results are affected by several neighboring spatial cells.
7. Towards Formulation of a Problem

• **What is given:**
  
  – we have high-resolution estimates \( \tilde{x}_1, \ldots, \tilde{x}_n \) of the values \( x_1, \ldots, x_n \) within several small spatial cells;
  
  – we also have low-resolution estimates \( \tilde{X}_j \) for the weighted averages

\[
X_j = \sum_{i=1}^{n} w_{j,i} \cdot x_i.
\]

• **Objective:** based on the estimates \( \tilde{x}_i \) and \( \tilde{x} \), we must provide more accurate estimates for \( x_i \).

• **Geophysical example:** we are interested in the densities \( x_i \).
8. Model Fusion: Case of Probabilistic Uncertainty

We take into account several different types of approximate equalities:

- Each high-resolution value $\tilde{x}_i$ is approximately equal to the actual value $x_i$, with the known accuracy $\sigma_{h,i}$:
  \[ \tilde{x}_i \approx x_i. \]

- Each lower-resolution value $\tilde{X}_j$ is approximately equal to the weighted average, with a known accuracy $\sigma_{l,j}$:
  \[ \tilde{X}_j \approx \sum_i w_{j,i} \cdot x_i. \]

- We usually have a prior knowledge $x_{pr,i}$ of the values $x_i$, with accuracy $\sigma_{pr,i}$: $x_i \approx x_{pr,i}$.

- Also, each lower-resolution value $\tilde{X}_j$ is approximately equal to the value within each of the smaller cells:
  \[ \tilde{X}_j \approx x_i(l,j). \]
9. Case of Probabilistic Uncertainty: Details

- Each lower-resolution value $\tilde{X}_j$ is approximately equal to the value within each of the smaller cells:

$$\tilde{X}_j \approx x_{i(l,j)}.$$ 

- The accuracy of $\tilde{X}_j \approx x_{i(l,j)}$ corresponds to the (empirical) standard deviation:

$$\sigma_{e,j}^2 \overset{\text{def}}{=} \frac{1}{k_j} \cdot \sum_{l=1}^{k_j} \left(\tilde{x}_{i(l,j)} - E_j\right)^2,$$

where

$$E_j \overset{\text{def}}{=} \frac{1}{k_j} \cdot \sum_{l=1}^{k_j} \tilde{x}_{i(l,j)}.$$
10. Model Fusion: Least Squares Approach

- **Main idea:** use the Least Squares technique to combine the approximate equalities.

- We find the desired combined values $x_i$ by minimizing the corresponding sum of weighted squared differences:

$$
\sum_{i=1}^{n} \frac{(x_i - \tilde{x}_i)^2}{\sigma_{h,i}^2} + \sum_{j=1}^{m} \frac{1}{\sigma_{l,j}^2} \cdot \left( \tilde{X}_j - \sum_{i=1}^{n} w_{j,i} \cdot x_i \right)^2 + \sum_{i=1}^{n} \frac{(x_i - x_{pr,i})^2}{\sigma_{pr,i}^2} + \sum_{j=1}^{m} \sum_{l=1}^{k_j} \frac{(\tilde{X}_j - x_{i(l,j)})^2}{\sigma_{e,j}^2}.
$$
11. Model Fusion: Solution

- To find a minimum of an expression, we:
  - differentiate it with respect to the unknowns, and
  - equate derivatives to 0.

- Differentiation with respect to $x_i$ leads to the following system of linear equations:

\[
\frac{1}{\sigma^2_{h,i}} \cdot (x_i - \tilde{x}_i) + \sum_{j: j \ni i} \frac{1}{\sigma^2_{l,j}} \cdot w_{j,i} \cdot \left( \sum_{i' = 1}^{n} w_{j,i'} \cdot x_{i'} - \tilde{X}_j \right) + \\
\frac{1}{\sigma^2_{pr,i}} \cdot (x_i - \tilde{x}_{pr,i}) + \sum_{j: j \ni i} \frac{1}{\sigma^2_{e,j}} \cdot (x_i - \tilde{X}_j) = 0,
\]

where $j \ni i$ means that the $j$-th low-resolution measurement covers $i$-th cell.
12. Simplification: Fusing High-Resolution Measurement Results and Prior Estimates

- **Idea:** fuse each high-resolution measurement result $\tilde{x}_i$ with a prior estimate $x_{pr,i}$.

- **Detail:** instead of $\frac{1}{\sigma^2_{h,i}} \cdot (x_i - \tilde{x}_i) + \frac{1}{\sigma^2_{pr,i}} \cdot (x_i - x_{pr,i})$, we have a single term $\sigma^{-2}_{f,i} \cdot (x_i - x_{f,i})$, where

  $$x_{f,i} \overset{\text{def}}{=} \tilde{x}_i \cdot \sigma^{-2}_{h,i} + x_{pr,i} \cdot \sigma^{-2}_{pr,i} \cdot \frac{1}{\sigma^{-2}_{h,i} + \sigma^{-2}_{pr,i}}, \quad \sigma^{-2}_{f,i} \overset{\text{def}}{=} \sigma^{-2}_{h,i} + \sigma^{-2}_{pr,i}.$$  

- **Resulting simplified equations:**

  $$\sigma^{-2}_{f,i} \cdot (x_i - x_{f,i}) + \sum_{j: j \ni i} \frac{1}{\sigma^2_{l,j}} \cdot w_{j,i} \cdot \left( \sum_{i'=1}^{n} w_{j,i'} \cdot x_{i'} - \tilde{X}_j \right) + \sum_{j: j \ni i} \frac{1}{\sigma^2_{e,j}} \cdot (x_i - \tilde{X}_j) = 0.$$
13. Case of a Single Low-Resolution Measurement

- **Simplest case:** we have exactly one low resolution measurement result $\tilde{X}_1$.

- **In general:** we only have the results of the high-resolution measurements for *some* of the cells.

- **In geosciences:** such a situation is typical: e.g.,
  - we have a low-resolution gravity measurement which covers a huge area in depth, and
  - we have the results of high-resolution seismic measurements which only cover depths above the Moho.

- **For convenience:** let us number the cells for which we have high-resolution measurement results first.

- Let $h$ denote the total number of such cells.

First, we compute the auxiliary value

$$
\mu \overset{\text{def}}{=} \frac{1}{\sigma_{l,1}^2} \cdot \left( \sum_{i'} w_{1,i'} \cdot x_{i'} - \tilde{X}_1 \right)
$$

as $$\mu = \frac{N}{D}$$, where

$$
N = \sum_{i=1}^{h} \frac{w_{1,i} \cdot (x_{f,i} - \tilde{X}_1)}{1 + \frac{\sigma_{f,i}^2}{\sigma_{e,1}^2}},
$$

and

$$
D = \sigma_{l,1}^2 + \sum_{i=1}^{h} \frac{w_{1,i}^2 \cdot \sigma_{f,i}^2}{1 + \frac{\sigma_{f,i}^2}{\sigma_{e,1}^2}} + \left( \sum_{i=h+1}^{n} w_{1,i}^2 \right) \cdot \sigma_{e,1}^2.
$$
15. Case of a Single Low-Resolution Measurement: Simplified Algorithm (cont-d)

• Once we know $\mu$, we compute the desired estimates for $x_i, i = 1, \ldots, h,$ as

$$x_i = \frac{x_{f,i}}{\sigma_{f,i}^2} - \frac{w_{1,i} \cdot \sigma_{f,i}^2}{\sigma_{e,1}^2} \cdot \mu + \frac{\sigma_{f,i}^2}{\sigma_{e,1}^2} \cdot \tilde{X}_1 \cdot \frac{\sigma_{f,i}^2}{\sigma_{e,1}^2} \cdot \mu.$$

• We also compute estimates $x_i$ for $i = h + 1, \ldots, n$, as

$$x_i = \tilde{X}_1 - w_{1,i} \cdot \sigma_{e,1}^2 \cdot \mu.$$
16. Numerical Example: Description

- **Objective:** to illustrate the above formulas.
- **Idea:** consider the simplest possible case, when we have
  - exactly one low resolution measurement result $\tilde{X}_1$
  - that covers all $n$ cells,
and when:
  - all the weights are all equal $w_{1,i} = 1/n$;
  - there is a high-resolution measurement corresponding to each cell ($h = n$);
  - all high-resolution measurements have the same accuracy $\sigma_{h,i} = \sigma_h$;
  - $\sigma_{l,1} \ll \sigma_h$, so $\sigma_{l,1} \approx 0$; and
  - there is no prior information, so $\sigma_{pr,i} = \infty$ and thus, $x_{f,i} = \tilde{x}_i$ and $\sigma_{f,i} = \sigma_h$. 
17. Additional Simplification

- In general: there are cells for which there are no high-resolution measurement results.

- How to deal with these cells: we added a heuristic rule that
  - each lower-resolution value is approximately equal to the value within each of the constituent cells,
  - with the accuracy corresponding to the (empirical) standard deviation $\sigma_{e,j}$.

- In our simplified example: we have high-resolution measurements in each cell.

- So, there is no need for this heuristic rule.

- The corresponding heuristic terms in the least squares approach are proportional to $\frac{1}{\sigma_{e,1}^2}$, so we take $\sigma_{e,1}^2 = \infty$. 
18. **Formulas for the Simplified Case and Numerical Example**

- **Resulting formulas:** \( x_i = \tilde{x}_i - \lambda \), where
  \[
  \lambda \overset{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} \tilde{x}_i - \tilde{X}_1. 
  \]

- **Case study:** \( n = 4 \) cells,
  - with the high-resolution accuracy \( \sigma_h = 0.5 \)
  - and the measured high-resolution values (in each of these cells)
    \[
    \tilde{x}_1 = 2.0, \quad \tilde{x}_2 = 3.0, \quad \tilde{x}_3 = 5.0, \quad \tilde{x}_4 = 6.0; 
    \]
  - the result of the corresponding low-resolution measurement is \( \tilde{X}_1 = 3.7 \).
19. High-Resolution and Low-Resolution Measurement Results: Illustration

<table>
<thead>
<tr>
<th>$\tilde{x}_1$</th>
<th>$\tilde{x}_2$</th>
<th>$\tilde{X}_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>3.0</td>
<td>3.7</td>
</tr>
<tr>
<td>5.0</td>
<td>6.0</td>
<td></td>
</tr>
</tbody>
</table>
20. Numerical Example: Discussion

- We assume that the low-resolution measurement is accurate ($\sigma_l \approx 0$).
- So, the average of the four cell values is equal to the result $\tilde{X}_1 = 3.7$ of this measurement:
  \[
  \frac{x_1 + x_2 + x_3 + x_4}{4} \approx 3.7.
  \]
- For the measured high-resolution values $\tilde{x}_i$, the average is slightly different:
  \[
  \frac{\tilde{x}_1 + \tilde{x}_2 + \tilde{x}_3 + \tilde{x}_4}{4} = \frac{2.0 + 3.0 + 5.0 + 6.0}{4} = 4.0 \neq 3.7.
  \]
- *Reason:* high-resolution measurements are much less accurate: $\sigma_h = 0.5$.
- We use the low-resolution measurements to “correct” the values of the high-resolution measurements.
21. Numerical Example: Results

- Here, the correcting term takes the form
  \[ \lambda = \frac{\tilde{x}_1 + \ldots + \tilde{x}_n}{n} - \tilde{X}_1 = \frac{2.0 + 3.0 + 5.0 + 6.0}{4} - 3.7 = 4.0 - 3.7 = 0.3. \]

- So, the corrected ("fused") values \( x_i \) take the form:
  \[ x_1 = \tilde{x}_1 - \lambda = 2.0 - 0.3 = 1.7; \quad x_2 = \tilde{x}_2 - \lambda = 3.0 - 0.3 = 2.7; \]
  \[ x_3 = \tilde{x}_3 - \lambda = 5.0 - 0.3 = 4.7; \quad x_4 = \tilde{x}_4 - \lambda = 6.0 - 0.3 = 5.7. \]

- For these corrected values, the arithmetic average is equal to the measured low-resolution value:
  \[ \frac{x_1 + x_2 + x_3 + x_4}{4} = \frac{1.7 + 2.7 + 4.7 + 5.7}{4} = 3.7. \]
22. The Result of Model Fusion: Simplified Setting

<table>
<thead>
<tr>
<th>( \tilde{x}_1 ) = 1.7</th>
<th>( \tilde{x}_2 ) = 2.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tilde{x}_3 ) = 4.7</td>
<td>( \tilde{x}_4 ) = 5.7</td>
</tr>
</tbody>
</table>
23. Taking $\sigma_{e,j}$ Into Account

- **Idea:** take into account the requirement that
  
  - the actual values in each cell are approximately equal to $\tilde{X}_1$,
  
  - with the accuracy $\sigma_{e,1}$ equal to the empirical standard deviation.

- **Resulting formulas:**
  
  $$\mu = \frac{\lambda}{\frac{1}{n} \cdot \sigma_h^2} = \frac{1}{n} \cdot \sum_{i=1}^{n} \tilde{x}_i - \tilde{X}_1$$
  
  $$\frac{x_i}{1 + \frac{\sigma_h^2}{\sigma_{e,1}^2}} + \tilde{X}_1 \cdot \frac{\sigma_h^2}{1 + \frac{\sigma_h^2}{\sigma_{e,1}^2}}.$$
24. Taking $\sigma_{e,j}$ Into Account: Numerical Example

- **General idea:** the actual values in each cell are approximately equal to $\tilde{X}_1$.

- **In our example:** $x_i \approx \tilde{X}_1$, with the accuracy

$$\sigma_{e,1}^2 = \frac{1}{4} \cdot \sum_{i=1}^{4} (\tilde{x}_i - E_1)^2,$$

where $E_1 = \frac{1}{4} \cdot \sum_{i=1}^{4} \tilde{x}_i$.

- Here, $E_1 = \frac{1}{4} \cdot \sum_{i=1}^{4} \tilde{x}_i = \frac{\tilde{x}_1 + \tilde{x}_2 + \tilde{x}_3 + \tilde{x}_4}{4} = 4.0$, thus,

$$\sigma_{e,1}^2 = \frac{(2.0 - 4.0)^2 + (3.0 - 4.0)^2 + (5.0 - 4.0)^2 + (6.0 - 4.0)^2}{4 + 1 + 1 + 4} = \frac{4}{10} = 2.5.$$

- Hence $\sigma_{e,1} \approx 1.58$. 

25. Taking $\sigma_{e,j}$ Into Account (cont-d)

- **Reminder:** $x_i = \frac{1}{\sigma_{e,1}^2} \cdot (\tilde{x}_i - \lambda) + \frac{\sigma_h^2}{\sigma_{e,1}^2} \cdot \tilde{X}_1$.

- Here, $\sigma_h = 0.5$, $\sigma_{e,1}^2 = 2.5$, $\frac{\sigma_h^2}{\sigma_{e,1}^2} = \frac{0.25}{2.5} = 0.1$, so

$$
\frac{1}{1 + \frac{\sigma_h^2}{\sigma_{e,1}^2}} = \frac{1}{1.1} \approx 0.91, \quad \text{and} \quad \frac{\sigma_h^2}{\sigma_{e,1}^2} \cdot \tilde{X}_1 = \frac{0.1}{1.1} \cdot 3.7 \approx 0.34;
$$

$$
x_1 \approx 0.91 \cdot (2.0 - 0.3) + 0.34 \approx 1.89;
$$

$$
x_2 \approx 0.91 \cdot (3.0 - 0.3) + 0.34 \approx 2.79;
$$

$$
x_3 \approx 0.91 \cdot (5.0 - 0.3) + 0.34 \approx 4.62;
$$

$$
x_4 \approx 0.91 \cdot (6.0 - 0.3) + 0.34 \approx 5.53.
$$
### 26. The Result of Model Fusion: General Setting

<table>
<thead>
<tr>
<th>( \tilde{x}_1 ) ≈ 1.89</th>
<th>( \tilde{x}_2 ) ≈ 2.79</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tilde{x}_3 ) ≈ 4.62</td>
<td>( \tilde{x}_4 ) ≈ 5.53</td>
</tr>
</tbody>
</table>

- The arithmetic average of these four values is equal to
  \[
  \frac{x_1 + x_2 + x_3 + x_4}{4} \approx \frac{1.89 + 2.79 + 4.62 + 5.53}{4} \approx 3.71.
  \]
- So, within our computation accuracy, it coincides with the measured low-resolution value \( \tilde{X}_1 = 3.7 \).
27. **Model Fusion: Case of Interval Uncertainty**

- We take into account three different types of approximate equalities:
  
  - Each high-resolution value \( \tilde{x}_i \) is approximately equal to the actual value \( x_i \):
    \[
    \tilde{x}_i - \Delta_{h,i} \leq x_i \leq \tilde{x}_i + \Delta_{h,i}.
    \]

  - Each lower-resolution value \( \tilde{X}_j \) is \( \approx \) to the average of values of all the cells \( x_i(1,j), \ldots, x_i(k,j) \):
    \[
    \tilde{X}_j - \Delta_{l,j} \leq \sum_i w_{j,i} \cdot x_i \leq \tilde{X}_j + \Delta_{l,j}.
    \]

  - Finally, we have prior bounds \( x_{pr,i} \) and \( \bar{x}_{pr,i} \) on the values \( x_i \), i.e., bounds for which
    \[
    x_{pr,i} \leq x_i \leq \bar{x}_{pr,i}.
    \]

- Our objective is to find, for each \( k = 1, \ldots, n \), the range \( [\underline{x}_k, \bar{x}_k] \) of possible values of \( x_k \).
28. Case of Interval Uncertainty: Algorithm

- The measurements lead to a system of linear inequalities for the unknown values $x_1, \ldots, x_n$.
- Thus, for each $k$, finding the endpoints $\underline{x}_k$ and $\overline{x}_k$ means optimizing the values $x_k$ under linear constraints.
- This is a particular case of a general linear programming problem.
- So, we can use Linear Programming to find these bounds:
  - the lower bound $\underline{x}_k$ can be obtained if we minimize $x_k$ under the constraints
    \[ \tilde{x}_i - \Delta_h \leq x_i \leq \tilde{x}_i + \Delta_h, \quad i = 1, \ldots, n; \]
    \[ \tilde{X}_j - \Delta_l \leq \sum_i w_{j,i} \cdot x_i \leq \tilde{X}_j + \Delta_l; \quad x_{pr,i} \leq x_i \leq \overline{x}_{pr,i}. \]
  - the upper bound $\overline{x}_k$ can be obtained if we maximize $x_k$ under the same constraints.
29. Conclusions

- We propose a fast practical alternative to joint inversion of multiple datasets.

- Specifically, we consider measurements that have
  - not only different accuracy and coverage,
  - but also different spatial resolution.

- To fuse such models, we must account for three different types of approximate equalities:
  - each high-resolution value is approximately equal to the actual value in the corresponding cell;
  - each lower-resolution value is \( \approx \) to the weighted average of the values in the corresponding cells;
  - each lower-resolution value is approximately equal to the value within each of the constituent cells.
30. Conclusions (cont-d)

- *Possible situations:* probabilistic or interval uncertainty.
- *Solution:* use the least squares or interval technique to combine the approximate equalities.
- *Example:* the least squares approach
  – we find the desired combined values
  – by minimizing the resulting sum of weighted squared differences.
- *Case study:* simulated (synthetic) geophysical data.
- *We show:* that model fusion indeed improves the accuracy and resolution of individual models.
- *Future plans:* apply model fusion techniques to more realistic simulated data and to real geophysical data.
31. Acknowledgments

- This work was supported in part by NSF grants:
  - Cyber-ShARE Center of Excellence (HRD-0734825),
  - Computing Alliance of Hispanic-Serving Institutions CAHISI (CNS-0540592),

and by NIH Grant 1 T36 GM078000-01.

- The author is thankful to all the participants of
  - Int’l Conf. on Scientific Computing SCAN’2008, El Paso, Texas, September 29 – October 3, 2008,
  - AGU’08, San Francisco, California, December 15–19, 2008, and
  - CAHSI’09, Mountain View, California, January 15–18, 2009

for valuable suggestions.