Model Fusion: A New Approach To Processing Heterogenous Data

Omar Ochoa
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

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.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

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 spatial resolution;
  - a seismic data point comes from a narrow trajectory of a seismic signal – so spatial resolution is higher.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

3. Joint Inversion: An Ideal Future Approach

- *At present:* each of the datasets is often processed separately.
- *It is desirable:* to 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.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

4. 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 estimation error $\Delta x^{(i)}$: $|\Delta x^{(i)}| \le \Delta_i$.
- In this case, we can conclude that $|x - \widetilde{x}^{(i)}| \le \Delta^{(i)}$, i.e., that $x \in \mathbf{x}^{(i)} \stackrel{\text{def}}{=} [\widetilde{x}^{(i)} - \Delta^{(i)}, \widetilde{x}^{(i)} + \Delta^{(i)}]$.
- Based on each estimate $\widetilde{x}^{(i)}$, we know that the actual value x belongs to the interval $\mathbf{x}^{(i)}$.
- Thus, we know that the (unknown) actual value x belongs to the intersection of these intervals:

$$\mathbf{x} \stackrel{\text{def}}{=} \bigcap_{i=1}^{n} \mathbf{x}^{(i)} = [\max(\widetilde{x}^{(i)} - \Delta^{(i)}), \min(\widetilde{x}^{(i)} + \Delta^{(i)})].$$

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

5. 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 estimates $\widetilde{x}^{(1)}, \ldots, \widetilde{x}^{(n)}$ of the same quantity x.
  - each estimation error $\Delta x^{(i)} \stackrel{\text{def}}{=} \widetilde{x}^{(i)} - x$ is normally distributed with 0 mean and known st. dev. $\sigma^{(i)}$;
  - Least Squares: find x that minimizes $\sum\limits_{i=1}^{n} \dfrac{(\widetilde{x}^{(i)} - x)^2}{2 \cdot (\sigma^{(i)})^2}$;
  - solution: $x = \dfrac{\sum\limits_{i=1}^{n} \widetilde{x}^{(i)} \cdot (\sigma^{(i)})^{-2}}{\sum\limits_{i=1}^{n} (\sigma^{(i)})^{-2}}$.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

6. Towards Formulation of a Problem

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

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

- *Objective:* based on the estimates $\widetilde{x}_i$ and $\widetilde{X}_j$, we must provide more accurate estimates for x_i.
- *Geophysical example:* we are interested in the densities x_i.

Need to Combine Data ...
Proposed Solution – ...
Numerical Example: ...
Additional Problem: ...
Auxiliary Problem: ...
Conclusions
Acknowledgments

7. Model Fusion: Case of Probabilistic Uncertainty

We take into account several different types of approximate equalities:

- Each high spatial resolution value $\widetilde{x}_i$ is approximately equal to the actual value x_i, w/known accuracy $\sigma_{h,i}$:

$$\widetilde{x}_i \approx x_i.$$

- Each lower spatial resolution value $\widetilde{X}_j$ is approximately equal to the weighted average, w/known accuracy $\sigma_{l,j}$:

$$\widetilde{X}_j \approx \sum_i w_{j,i} \cdot x_i.$$

- 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 spatial resolution value $\widetilde{X}_j$ is $\approx$ the value within each of the smaller cells:

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

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

8. Case of Probabilistic Uncertainty: Details

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

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

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

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

where

$$E_j \stackrel{\text{def}}{=} \frac{1}{k_j} \cdot \sum_{l=1}^{k_j} \widetilde{x}_{i(l,j)}.$$

Need to Combine Data . . .
Proposed Solution − . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

9. 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:

$$\sum_{i=1}^{n} \frac{(x_i - \widetilde{x}_i)^2}{\sigma_{h,i}^2} + \sum_{j=1}^{m} \frac{1}{\sigma_{l,j}^2} \cdot \left(\widetilde{X}_j - \sum_{i=1}^{n} w_{j,i} \cdot x_i \right)^2 +$$

$$\sum_{i=1}^{n} \frac{(x_i - x_{pr,i})^2}{\sigma_{pr,i}^2} + \sum_{j=1}^{m} \sum_{l=1}^{k_j} \frac{(\widetilde{X}_j - x_{i(l,j)})^2}{\sigma_{e,j}^2}.$$

Need to Combine Data ...
Proposed Solution – ...
Numerical Example: ...
Additional Problem: ...
Auxiliary Problem: ...
Conclusions
Acknowledgments

10. 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:

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

$$\frac{1}{\sigma^2_{pr,i}} \cdot (x_i - x_{pr,i}) + \sum_{j:j\ni i} \frac{1}{\sigma^2_{e,j}} \cdot (x_i - \widetilde{X}_j) = 0,$$

where $j \ni i$ means that the j-th low spatial resolution estimate covers i-th cell.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

11. Simplification: Fusing High Spatial Resolution Estimates and Prior Estimates

- *Idea:* fuse each high spatial resolution estimate $\widetilde{x}_i$ with a prior estimate $x_{pr,i}$.
- *Detail:* instead of $\dfrac{1}{\sigma^2_{h,i}} \cdot (x_i - \widetilde{x}_i) + \dfrac{1}{\sigma^2_{pr,i}} \cdot (x_i - x_{pr,i})$, we have a single term $\sigma^{-2}_{f,i} \cdot (x_i - x_{f,i})$, where

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

- *Resulting simplified equations:*

$$\sigma^{-2}_{f,i} \cdot (x_i - x_{f,i}) + \sum_{j:j\ni i} \frac{1}{\sigma^2_{l,j}} \cdot w_{j,i} \cdot \left(\sum_{i'=1}^{n} w_{j,i'} \cdot x_{i'} - \widetilde{X}_j \right) +$$

$$\sum_{j:j\ni i} \frac{1}{\sigma^2_{e,j}} \cdot (x_i - \widetilde{X}_j) = 0.$$

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

12. Case of a Single Low Spatial Resolution Estimate

- *Simplest case:* we have exactly one low spatial resolution estimate $\widetilde{X}_1$.
- *In general:* we only have high spatial resolution estimates for *some* of the cells.
- *In geosciences:* such a situation is typical: e.g.,
  - we have a low spatial resolution gravity estimates which cover a huge area in depth, and
  - we have high spatial resolution seismic estimates which only cover depths above the Moho.
- *For convenience:* let us number the cells for which we have high spatial resolution estimates first.
- Let h denote the total number of such cells.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

13. Case of a Single Low Spatial Resolution Estimate: Simplified Algorithm

First, we compute the auxiliary value

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

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

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

and

$$D = \sigma_{l,1}^2 + \sum_{i=1}^{h} \frac{w_{1,i}^2 \cdot \sigma_{f,i}^2}{1 + \dfrac{\sigma_{f,i}^2}{\sigma_{e,1}^2}} + \left(\sum_{i=h+1}^{n} w_{1,i}^2 \right) \cdot \sigma_{e,1}^2.$$

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

14. Case of a Single Low Spatial Resolution Estimate: Simplified Algorithm (cont-d)

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

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

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

$$x_i = \widetilde{X}_1 - w_{1,i} \cdot \sigma_{e,1}^2 \cdot \mu.$$

Need to Combine Data ...
Proposed Solution – ...
Numerical Example: ...
Additional Problem: ...
Auxiliary Problem: ...
Conclusions
Acknowledgments

15. Numerical Example: Description

- *Objective:* to illustrate the above formulas.
- *Idea:* consider the simplest possible case, when we have
  - exactly one low spatial resolution estimate $\widetilde{X}_1$
  - that covers all n cells,

  and when:
  - all the weights are all equal $w_{1,i} = 1/n$;
  - there is a high spatial resolution estimate corresponding to each cell $(h = n)$;
  - all high spatial resolution estimates 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, $x_{f,i} = \widetilde{x}_i$ and $\sigma_{f,i} = \sigma_h$.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

16. Additional Simplification

- *In general:* there are cells for which there are no high spatial resolution estimates.
- *How to deal with these cells:* we added a heuristic rule that
  - each lower spatial 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 spatial resolution estimate in each cell.
- So, there is no need for this heuristic rule.
- The corresponding heuristic terms in the least squares approach are proportional to $\dfrac{1}{\sigma_{e,1}^2}$, so we take $\sigma_{e,1}^2 = \infty$.

Need to Combine Data ...
Proposed Solution – ...
Numerical Example: ...
Additional Problem: ...
Auxiliary Problem: ...
Conclusions
Acknowledgments

17. Formulas for the Simplified Case and Numerical Example

- *Resulting formulas:* $x_i = \widetilde{x}_i - \lambda$, where

$$\lambda \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} \widetilde{x}_i - \widetilde{X}_1.$$

- *Case study:* $n = 4$ cells,
  - with the high spatial resolution accuracy $\sigma_h = 0.5$
  - and the high spatial resolution estimates (in each of these cells)

$$\widetilde{x}_1 = 2.0, \quad \widetilde{x}_2 = 3.0, \quad \widetilde{x}_3 = 5.0, \quad \widetilde{x}_4 = 6.0;$$

  - the corresponding low spatial resolution estimate is $\widetilde{X}_1 = 3.7$.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

18. Estimates of High and Low Spatial Resolution: Illustration

$\widetilde{x}_1 = 2.0$	$\widetilde{x}_2 = 3.0$
$\widetilde{x}_3 = 5.0$	$\widetilde{x}_4 = 6.0$

$\widetilde{X}_1 = 3.7$

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

19. Numerical Example: Discussion

- We assume that the low spatial resolution estimate is accurate ($\sigma_l \approx 0$).
- So, the average of the four cell values is equal to the result $\widetilde{X}_1 = 3.7$ of this estimate:

$$\frac{x_1 + x_2 + x_3 + x_4}{4} \approx 3.7.$$

- For the high spatial resolution estimates $\widetilde{x}_i$, the average is slightly different:

$$\frac{\widetilde{x}_1 + \widetilde{x}_2 + \widetilde{x}_3 + \widetilde{x}_4}{4} = \frac{2.0 + 3.0 + 5.0 + 6.0}{4} = 4.0 \neq 3.7.$$

- *Reason:* high spatial resolution estimates are much less accurate: $\sigma_h = 0.5$.
- We use the low spatial resolution estimate to "correct" the high spatial resolution estimate.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

20. Numerical Example: Results

- Here, the correcting term takes the form

$$\lambda = \frac{\widetilde{x}_1 + \ldots + \widetilde{x}_n}{n} - \widetilde{X}_1 =$$

$$\frac{2.0 + 3.0 + 5.0 + 6.0}{4} - 3.7 = 4.0 - 3.7 = 0.3.$$

- So, the corrected ("fused") values x_i take the form:

$$x_1 = \widetilde{x}_1 - \lambda = 2.0 - 0.3 = 1.7; \;\; x_2 = \widetilde{x}_2 - \lambda = 3.0 - 0.3 = 2.7;$$

$$x_3 = \widetilde{x}_3 - \lambda = 5.0 - 0.3 = 4.7; \;\; x_4 = \widetilde{x}_4 - \lambda = 6.0 - 0.3 = 5.7.$$

- For these corrected values, the arithmetic average is equal to the low spatial resolution estimate:

$$\frac{x_1 + x_2 + x_3 + x_4}{4} = \frac{1.7 + 2.7 + 4.7 + 5.7}{4} = 3.7.$$

Need to Combine Data ...
Proposed Solution – ...
Numerical Example: ...
Additional Problem: ...
Auxiliary Problem: ...
Conclusions
Acknowledgments

21. The Result of Model Fusion: Simplified Setting

$\widetilde{x}_1 = 1.7$	$\widetilde{x}_2 = 2.7$
$\widetilde{x}_3 = 4.7$	$\widetilde{x}_4 = 5.7$

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

22. Taking $\sigma_{e,j}$ Into Account

- *Idea:* take into account the requirement that
  - the actual values in each cell are approximately equal to $\widetilde{X}_1$,
  - with the accuracy $\sigma_{e,1}$ equal to the empirical standard deviation.
- *Resulting formulas:* $\mu = \dfrac{\lambda}{\dfrac{1}{n}\cdot\sigma_h^2} = \dfrac{\dfrac{1}{n}\cdot\sum\limits_{i=1}^{n}\widetilde{x}_i - \widetilde{X}_1}{\dfrac{1}{n}\cdot\sigma_h^2}$, and

$$x_i = \frac{\widetilde{x}_i - \lambda}{1 + \dfrac{\sigma_h^2}{\sigma_{e,1}^2}} + \widetilde{X}_1 \cdot \frac{\dfrac{\sigma_h^2}{\sigma_{e,1}^2}}{1 + \dfrac{\sigma_h^2}{\sigma_{e,1}^2}}.$$

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

23. Taking $\sigma_{e,j}$ Into Account: Numerical Example

- *General idea:* the actual values in each cell are approximately equal to $\widetilde{X}_1$.
- *In our example:* $x_i \approx \widetilde{X}_1$, with the accuracy

$$\sigma^2_{e,1} = \frac{1}{4} \cdot \sum_{i=1}^{4} (\widetilde{x}_i - E_1)^2, \text{ where } E_1 = \frac{1}{4} \cdot \sum_{i=1}^{4} \widetilde{x}_i.$$

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

$$\sigma^2_{e,1} = \frac{(2.0 - 4.0)^2 + (3.0 - 4.0)^2 + (5.0 - 4.0)^2 + (6.0 - 4.0)^2}{4} =$$

$$\frac{4 + 1 + 1 + 4}{4} = \frac{10}{4} = 2.5.$$

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

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

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

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

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

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

$$x_1 \approx 0.91 \cdot (2.0 - 0.3) + 0.34 \approx 1.89;$$
$$x_2 \approx 0.91 \cdot (3.0 - 0.3) + 0.34 \approx 2.79;$$
$$x_3 \approx 0.91 \cdot (5.0 - 0.3) + 0.34 \approx 4.62;$$
$$x_4 \approx 0.91 \cdot (6.0 - 0.3) + 0.34 \approx 5.53.$$

Need to Combine Data ...
Proposed Solution – ...
Numerical Example: ...
Additional Problem: ...
Auxiliary Problem: ...
Conclusions
Acknowledgments

25. The Result of Model Fusion: General Setting

$\widetilde{x}_1 \approx 1.89$	$\widetilde{x}_2 \approx 2.79$
$\widetilde{x}_3 \approx 4.62$	$\widetilde{x}_4 \approx 5.53$

- 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 low spatial resolution estimate $\widetilde{X}_1 = 3.7$.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

26. Model Fusion: Case of Interval Uncertainty

- We take into account three different types of approximate equalities:
  - Each high spatial resolution estimate $\widetilde{x}_i$ is approximately equal to the actual value x_i:
  $$\widetilde{x}_i - \Delta_{h,i} \le x_i \le \widetilde{x}_i + \Delta_{h,i}.$$
  - Each lower spatial resolution value $\widetilde{X}_j$ is $\approx$ to the average of values of all the cells $x_{i(1,j)}, \ldots, x_{i(k_j,j)}$:
  $$\widetilde{X}_j - \Delta_{l,j} \le \sum_i w_{j,i} \cdot x_i \le \widetilde{X}_j + \Delta_{l,j}.$$
  - Finally, we have prior bounds $\underline{x}_{pr,i}$ and $\overline{x}_{pr,i}$ on the values x_i, i.e., bounds for which
  $$\underline{x}_{pr,i} \le x_i \le \overline{x}_{pr,i}.$$
- Our objective is to find, for each $k = 1, \ldots, n$, the range $[\underline{x}_k, \overline{x}_k]$ of possible values of x_k.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

27. Additional Results

- Additional problem: need to fuse discrete and continuous data
- Auxiliary problem: estimating accuracy of fused models

Need to Combine Data . . .
Proposed Solution − . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

28. Additional Problem: Need to Fuse Discrete and Continuous Models

- Traditionally, seismic models are *continuous*: the velocity smoothly changes with depth.
- In contrast, the gravity models are *discrete*: we have layers, in each of which the velocity is constant.
- The abrupt transition corresponds to a steep change in the continuous model.
- Both models locate the transition only approximately.
- So, if we simply combine the corresponding values value-by-value, we will have *two* transitions instead of one:
  - one transition where the continuous model has it, and
  - another transition nearby where the discrete model has it.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

29. What We Plan to Do

- *We want* to avoid the misleading double-transition models.
- *Idea:* first fuse the corresponding transition locations.
- *In this paper,* we provide an algorithm for such location fusion.
- *Specifically*, first, we formulate the problem in the probabilistic terms.
- *Then*, we provide an algorithm that produces the most probable transition location.
- *We show* that the result of the probabilistic location algorithm is in good accordance with common sense.
- *We also show* how the commonsense intuition can be reformulated in fuzzy terms.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

30. Available Data: What is Known and What Needs to Be Determined

- For each location, in the discrete model, we have the exact depth z_d of the transition.
- In contrast, for the continuous model, we do not have the abrupt transition.
- Instead, we have velocity values $v(z)$ at different depths.
- We must therefore extract the corresponding transition value z_c from the velocity values.
- To be more precise, we have values $v_1, v_2, \ldots, v_i, \ldots, v_n$ corresponding to different depths.
- We need to find i for which the transition occurs between the depths i and $i + 1$.

31. Probabilistic Approach

- The difference $\Delta v_j \stackrel{\text{def}}{=} v_j - v_{j+1}$ $(j \neq i)$ is caused by many independent factors.
- Due to the Central Limit Theorem, we thus assume that it is normally distributed, with probability density

$$p_j \stackrel{\text{def}}{=} \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot \exp\left(-\frac{1}{2 \cdot \sigma^2} \cdot (\Delta v_j)^2\right).$$

- The value Δv_i at the transition depth i is *not* described by the normal distribution.
- We assume that differences corresponding to different depths j are independent, so:

$$L_i = \prod_{j \neq i} p_j = \prod_{j \neq i} \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot \exp\left(-\frac{1}{2 \cdot \sigma^2} \cdot (\Delta v_j)^2\right).$$

Need to Combine Data . . .
Proposed Solution − . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

32. How to Find the Location: The General Idea of the Maximum Likelihood Approach

- *Reminder:* the likelihood of each model is:

$$L_i = \prod_{j\ne i} p_j = \prod_{j\ne i} \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma}\cdot\exp\left(-\frac{1}{2\cdot\sigma^2}\cdot(\Delta v_j)^2\right).$$

- *Natural idea:* select the parameters for which the likelihood of the observed data is the largest.
- The value L_i is the largest if and only if $-\ln(L_i)$ is the smallest: $-\ln(L_i) = \text{const} + \dfrac{1}{2\cdot\sigma^2}\cdot\sum\limits_{j\ne i}(\Delta v_j)^2 \to \min\limits_i$.
- This sum is equal to $\sum\limits_{j\ne i}(\Delta v_j)^2 = \sum\limits_{j=1}^{n-1}(\Delta v_j)^2 - (\Delta v_i)^2$.
- The first term in this expression does not depend on i.
- Thus, the difference is the smallest $\Leftrightarrow$ the value $(\Delta v_i)^2$ is the largest $\Leftrightarrow$ $|\Delta v_i|$ is the largest.

Need to Combine Data ...
Proposed Solution – ...
Numerical Example: ...
Additional Problem: ...
Auxiliary Problem: ...
Conclusions
Acknowledgments

33. Resulting Location

- *We want:* to select the most probable location of the transition point.
- *We select:* the depth i_0 for which the absolute value $|\Delta v_i|$ of the difference $\Delta v_i = v_{i+1} - v_i$ is the largest.
- This conclusion seems to be very reasonable:
  - the most probable location of the actual abrupt transition between the layers
  - is the depth at which the measured difference is the largest.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

34. The Results of the Probabilistic Approach are in Good Accordance with Common Sense

- Intuitively, for each depth i, our confidence that i a transition point depends on the difference $|\Delta v_i|$:
  - the smaller the difference, the less confident we are that this is the actual transition depth, and
  - the larger the difference, the more confident we are that this is the actual transition depth.
- In our probabilistic model, we select a location with the largest possible value $|\Delta v_i|$.
- This shows that the probabilistic model is in good accordance with common sense.
- This coincidence increases our confidence in this result.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

35. It May Be Useful to Formulate the Common Sense Description in Fuzzy Terms

- Fuzzy logic is known to be a useful way to formalize imprecise commonsense reasoning.
- Common sense: the degree of confidence d_i that i is a transition point is $f(|\Delta v_i|)$, for some monotonic $f(z)$.
- It is reasonable to select a value i for which our degree of confidence is the largest $d_i = f(|\Delta v_i|) \to \max$.
- Since $f(z)$ is increasing, this is equivalent to

$$|\Delta v_i| \to \max .$$

- Of course, to come up with this conclusion, we do not need to use the fuzzy logic techniques.
- However, this description may be useful if we also have other expert information.

36. How Accurate Is This Location Estimate?

- *Reminder:* the likelihood has the form
$$L_i = \prod_{j\ne i} p_j = \prod_{j\ne i} \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma}\cdot\exp\left(-\frac{1}{2\cdot\sigma^2}\cdot(\Delta v_j)^2\right).$$
- *We have found* the most probable transition i_0 as the value for which L_i is the largest.
- *Similarly:* we can find σ for which L_i is the largest:
$$\sigma^2 = \frac{1}{n-2}\cdot\sum_{j\ne i_0}(\Delta v_j)^2.$$
- The probability P_i that the transition is at location i is proportional to L_i: $P_i = c\cdot L_i$.
- The coefficient c can be determined from the condition that the total probability is 1: $1 = \sum\limits_i P_i = c\cdot\sum\limits_{i=1}^{n} L_i$.
- So, $c = \left(\sum L_i\right)^{-1}$.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

37. Accuracy of the Location Estimate (cont-d)

- The mean square deviation σ_0^2 of the actual transition depth from our estimate i_0 is defined as

$$\sigma_0^2 = \sum_{i=1}^{n-1} (i - i_0)^2 \cdot P_i.$$

- We know that $P_i = c \cdot L_i$, and we have formulas for computing L_i and c, so we can compute σ_0.
- We applied this algorithm to the seismic model of El Paso area, and got $\sigma_0 \approx 1.5$ km.
- This value is of the same order (1-2 km) as the difference between:
  - the border depth estimates coming from the seismic data and
  - the border depth coming from the gravity data.

Need to Combine Data ...
Proposed Solution – ...
Numerical Example: ...
Additional Problem: ...
Auxiliary Problem: ...
Conclusions
Acknowledgments

38. How to Fuse the Depth Estimates

- Now, we have two estimates for the transition depth:
  - the estimate i_d from the discrete (gravity) model;
  - the estimate i_0 from the continuous (seismic) model.
- The estimate i_d comes from a standard statistical analysis, so we know standard deviation σ_d.
- For i_0, we already know the standard deviation σ_0.
- It is reasonable to assume that both differences $i_d - i$ and $i_0 - i$ are normally distributed and independent:
$$p_i = \exp\left(-\frac{(i_d - i_f)^2}{2 \cdot \sigma_d^2}\right) \cdot \exp\left(-\frac{(i_0 - i_f)^2}{2 \cdot \sigma_0^2}\right).$$
- The most probable location i is when $p_i \to \max$, i.e.:
$$i_f = \frac{i_d \cdot \sigma_d^{-2} + i_0 \cdot \sigma_0^{-2}}{\sigma_d^{-2} + \sigma_0^{-2}}.$$

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

39. Towards Fusing Actual Maps

- In the discrete model:
  - values $i < i_d$ correspond to the upper zone;
  - values $i > i_d$ correspond to the lower zone.
- Similarly, in the continuous model:
  - values $i < i_0$ correspond to the upper zone;
  - values $i > i_0$ correspond to the lower zone.
- So, for depths $i \le \min(i_0, i_d)$ and $i \ge \max(i_0, i_d)$, both models correctly describe the zone.
- For these depths, we can simply fuse the values from both models.
- We can fuse them similarly to how we fused the depths.
- For intermediate depths, we need to adjust the models: e.g., by taking the nearest value from the correct zone.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

40. How to Fuse the Actual Maps: First Stage

- *First:* we adjust both models so that they both have a transition at depth i_f.
- *Adjusting the discrete model* is easy: we replace
  - the original depth i_d
  - with the new (more accurate) fused value i_f.
- *Adjusting the continuous model:*
  - when $i_f < i_0$, the values at depths i between i_f and i_0 are erroneously assigned to the the upper zone;
  - these values v_i must be replaced by the the value of the nearest point at the lower zone v_{i_0+1};
  - when $i_f > i_0$, the values at depths i between i_0 and i_f are erroneously assigned to the the lower zone;
  - these values v_i must be replaced by the the value of the nearest point at the upper zone v_{i_0}.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

41. How to Merge the Adjusted Models

- For each depth i, we now have two adjusted values v_i' and v_i'' corresponding to two adjusted models.
- Let σ' and σ'' be the corresponding standard deviations.
- It is reasonable to assume that both differences $v_i' - v_i$ and $v_i'' - v_i$ are normally distributed and independent:

$$p(v_i) = \exp\left(-\frac{(v_i' - v_i)^2}{2 \cdot (\sigma')^2}\right) \cdot \exp\left(-\frac{(v_i'' - v_i)^2}{2 \cdot (\sigma'')^2}\right).$$

- The most probable value $\widetilde{v}_i$ is when $p(v_i) \to \max$, i.e.:

$$\widetilde{v}_i = \frac{v_i' \cdot (\sigma')^{-2} + v_i'' \cdot (\sigma'')^{-2}}{(\sigma')^{-2} + (\sigma'')^{-2}}.$$

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

42. Auxiliary Problem: How to Estimate Accuracy of Fused Models

- *Calibration* is possible when we have a "standard" (several times more accurate) measuring instrument (MI).
- In geophysics, seismic (and other) methods are state-of-the-art.
- No method leads to more accurate determination of the densities.
- In some practical situations, we can use two similar MIs to measure the same quantities x_i.
- In geophysics, we want to estimate the accuracy of a model, e.g., a seismic model, a gravity-based model.
- In this situation, we do not have two similar applications of the same model.

Need to Combine Data ...
Proposed Solution – ...
Numerical Example: ...
Additional Problem: ...
Auxiliary Problem: ...
Conclusions
Acknowledgments

43. Maximum Likelihood (ML) Approach Cannot Be Applied to Estimate Model Accuracy

- We have several quantities with (unknown) actual values $x_1, \ldots, x_i, \ldots, x_n$.
- We have several measuring instruments (or geophysical methods) with (unknown) accuracies $\sigma_1, \ldots, \sigma_m$.
- We know the results x_{ij} of measuring the i-th quantity x_i by using the j-th measuring instrument.
- At first glance, a reasonable idea is to find all the unknown quantities x_i and σ_j from ML:

$$L = \prod_{i=1}^{n} \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi} \cdot \sigma_j} \cdot \exp\left(-\frac{(x_{ij} - x_i)^2}{2\sigma_j^2}\right) \to \max.$$

- *Fact:* the largest value $L = \infty$ is attained when, for some j_0, we have $\sigma_{j_0} = 0$ and $x_i = x_{ij_0}$ for all i.
- *Problem:* this is not physically reasonable.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

44. How to Estimate Model Accuracy: Idea

- For every two models, the difference $x_{ij} - x_{ik} = \Delta x_{ij} - \Delta x_{ik}$ is normally distributed, w/variance $\sigma_j^2 + \sigma_k^2$.
- We can thus estimate $\sigma_j^2 + \sigma_k^2$ as

$$\sigma_j^2 + \sigma_k^2 \approx A_{jk} \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (x_{ij} - x_{ik})^2.$$

- So, $\sigma_1^2 + \sigma_2^2 \approx A_{12}$, $\sigma_1^2 + \sigma_3^2 \approx A_{13}$, and $\sigma_2^2 + \sigma_3^2 \approx A_{23}$.
- By adding all three equalities and dividing the result by two, we get $\sigma_1^2 + \sigma_2^2 + \sigma_3^2 = \dfrac{A_{12} + A_{13} + A_{23}}{2}$.
- Subtracting, from this formula, the expression for $\sigma_2^2 + \sigma_3^2$, we get $\sigma_1^2 \approx \dfrac{A_{12} + A_{13} - A_{23}}{2}$.
- Similarly, $\sigma_2^2 \approx \dfrac{A_{12} + A_{23} - A_{13}}{2}$ and $\sigma_3^2 \approx \dfrac{A_{13} + A_{23} - A_{12}}{2}$.

Need to Combine Data ...
Proposed Solution – ...
Numerical Example: ...
Additional Problem: ...
Auxiliary Problem: ...
Conclusions
Acknowledgments

45. How to Estimate Model Accuracy: General Case and Challenge

- *General case:* we may have $M \geq 3$ different models.
- Then, we have $\dfrac{M \cdot (M-1)}{2}$ different equations $\sigma_j^2 + \sigma_k^2 \approx A_{jk}$ to determine M unknowns σ_j^2.
- When $M > 3$, we have more equations than unknowns,
- So, we can use the Least Squares method to estimate the desired values σ_j^2.
- *Challenge:* the formulas $\sigma_1^2 \approx \widetilde{V}_1 \stackrel{\text{def}}{=} \dfrac{A_{12} + A_{13} - A_{23}}{2}$ are approximate.
- Sometimes, these formulas lead to physically meaningless negative values $\widetilde{V}_1$.
- It is therefore necessary to modify the above formulas, to avoid negative values.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

46. An Idea of How to Deal With This Challenge

- The negativity challenge is caused by the fact that the estimates $\widetilde{V}_j$ for σ_j^2 are approximate.
- For large n, the difference $\Delta V_j \stackrel{\text{def}}{=} \widetilde{V}_j - \sigma_j^2$ is asymptotically normally distributed, with asympt. 0 mean.
- We can estimate the standard deviation Δ_j for this difference.
- Thus, $\sigma_j^2 = \widetilde{V}_j - \Delta V_j$ is normally distributed with mean $\widetilde{V}_j$ and standard deviation Δ_j.
- We also know that $\sigma_j^2 \ge 0$.
- As an estimate for σ_j^2, it is therefore reasonable to use a conditional expected value $E\left(\widetilde{V}_j - \Delta V_j \,\middle|\, \widetilde{V}_j - \Delta V_j \ge 0\right)$.
- This new estimate is an expected value of a non-negative number and thus, cannot be negative.

47. Resulting Algorithm

- *Input:* for each value x_i ($i = 1, \ldots, n$), we have three estimates x_{i1}, x_{i2}, and x_{i3} corr. to three diff. models.
- *Objective:* to estimate the accuracies σ_j^2 of these three models.
- First, for each $j \neq k$, we compute

$$A_{jk} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_{ij} - x_{ik})^2.$$

- Then, we compute

$$\widetilde{V}_1 = \frac{A_{12} + A_{13} - A_{23}}{2}; \quad \widetilde{V}_2 = \frac{A_{12} + A_{23} - A_{13}}{2};$$

$$\widetilde{V}_3 = \frac{A_{13} + A_{23} - A_{12}}{2}.$$

- After that, for each j, we compute

$$\Delta_j^2 = \frac{1}{n} \cdot \left(\left(\widetilde{V}_j \right)^2 + \widetilde{V}_j \cdot \widetilde{V}_k + \widetilde{V}_j \cdot \widetilde{V}_\ell + \widetilde{V}_k \cdot \widetilde{V}_\ell \right).$$

Need to Combine Data ...
Proposed Solution – ...
Numerical Example: ...
Additional Problem: ...
Auxiliary Problem: ...
Conclusions
Acknowledgments

48. Resulting Algorithm (cont-d)

- *Reminder:* we compute $\widetilde{V}_j = \dfrac{A_{jk} + A_{j\ell} - A_{kl}}{2}$ and
$$\Delta_j^2 = \frac{1}{n} \cdot \left(\left(\widetilde{V}_j \right)^2 + \widetilde{V}_j \cdot \widetilde{V}_k + \widetilde{V}_j \cdot \widetilde{V}_\ell + \widetilde{V}_k \cdot \widetilde{V}_\ell \right).$$
- Then, we compute the auxiliary ratios $\delta_j = \dfrac{\widetilde{V}_j}{\Delta_j}$.
- Finally, we return as an estimate $\widetilde{\sigma_j^2}$ for σ_j^2, the value
$$\widetilde{\sigma_j^2} = \widetilde{V}_j + \frac{\Delta_j}{\sqrt{2\pi}} \cdot \frac{\exp\left(-\dfrac{\delta_j^2}{2} \right)}{\Phi(\delta_j)}.$$
- These non-negative estimates $\widetilde{\sigma_j^2}$ can now be used to fuse the models: for each i, we take $x_i = \dfrac{\sum \widetilde{\sigma}_j^{-2} \cdot x_{ij}}{\sum \widetilde{\sigma}_j^{-2}}$.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments

49. Conclusions

- In many practical situations, there is a need to combine (fuse) data from different datasets.
- Ideal approach of *joint inversion* – which uses all the data from all the datasets – is often not yet practical.
- Main idea of *model fusion:* process each dataset separately and fuse the resulting models.
- In this thesis, algorithms are proposed for fusing models with different accuracy and spatial resolution.
- This thesis also addresses additional challenge:
  - fusing discrete and continuous models;
  - estimating the accuracy of fused models.
- This work can help geophysicists combine complementary models.

Need to Combine Data ...
Proposed Solution – ...
Numerical Example: ...
Additional Problem: ...
Auxiliary Problem: ...
Conclusions
Acknowledgments

50. Acknowledgments

- This work was supported by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence).
- The author is greatly thankful:
  - to Drs. Ann Gates, Vladik Kreinovich, and Aaron Velasco for their help and support, and
  - to family and friends for being there with me.

Need to Combine Data . . .
Proposed Solution – . . .
Numerical Example: . . .
Additional Problem: . . .
Auxiliary Problem: . . .
Conclusions
Acknowledgments