Towards Model Fusion in Geophysics: How to Estimate Accuracy of Different Models

**Omar Ochoa, Aaron Velasco,
Christian Servin, and Vladik Kreinovich**

Cyber-ShARE Center
University of Texas at El Paso
El Paso, TX 79968, USA

omar@miners.utep.edu, aavelasco@utep.edu
christians@utep.edu, vladik@utep.edu

1. Need to Fuse Models: Geophysics

- One of the main objectives of geophysics: find the density $\rho(x, y, z)$ at different depths z and locations (x, y).
- There exist several methods for estimating the density:
  - we can use seismic data,
  - we can use gravity measurements.
- Each of the techniques for estimating ρ has its own advantages and limitations.
- Example: seismic measurements often lead to a more accurate value of ρ than gravity measurements.
- However, seismic measurements mostly provide information about the areas above the Moho surface.
- It is desirable to combine ("fuse") the models obtained from different types of measurements into a single model.

2. Fusion: General Problem

- Similar situations are frequent in practice:
  - we are interested in the value of a quantity, and
  - we have reached the limit of the accuracy achievable by using a single measuring instrument.
- *Objective:* to further increase the estimation accuracy.
- *Idea:* perform several measurements of the desired quantity x_i.
- *Comment:* we may use the same measuring instrument or different measuring instruments.
- Then, we combine the results $x_{i1}, x_{i2}, \ldots, x_{im}$ of these measurement into a single more accurate estimate $\widehat{x}_i$.

3. Motivation for Using Normal Distributions

- The need for fusion comes when we have extracted all possible accuracy from each measurements.
- This means, in particular, that we have found and eliminated the systematic errors.
- Thus, the resulting measurement error has 0 mean.
- It also means that that we have found and eliminated the major sources of the random error.
- Since all big error components are eliminated, what is left is the large number of small error components.
- According to the Central Limit Theorem, the distribution is approximately normal.
- Thus, it is natural to assume that each measurement error $\Delta x_{ij} \stackrel{\text{def}}{=} x_{ij} - x_i$ is normally distributed.

4. Statistical Fusion: Formulas

- Each measurement error is normally distributed:

$$\rho_{ij}(x_{ij}) = \frac{1}{\sqrt{2\pi}\cdot\sigma_j}\cdot\exp\left(-\frac{(x_{ij}-x_i)^2}{2\sigma_j^2}\right).$$

- It is reasonable to assume that measurement errors corr. to different measurements are independent, so

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

- According to the Maximum Likelihood Principle, we select most probable value x_i s.t. $L \to \max$:

$$x_i = \frac{\sum\limits_{j=1}^{m}\sigma_j^{-2}\cdot x_{ij}}{\sum\limits_{j=1}^{m}\sigma_j^{-2}};$$ so, we must know the accuracies σ_j.

5. Traditional Methods of Estimating Accuracy Cannot Be Directly Used in Geophysics

- *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.

6. 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.

7. 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 $\sigma_3^2 \approx \dfrac{A_{13} + A_{23} - A_{12}}{2}$.

8. 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.

9. 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 $\sigma_j^2 \geq 0$.
- As an estimate for σ_j^2, it is therefore reasonable to use a conditional expected value $E\left(\widetilde{V}_j - \Delta V_j \,\middle|\, \widetilde{V}_j - \Delta V_j \geq 0\right)$.
- This new estimate is an expected value of a non-negative number and thus, cannot be negative.

10. Derivation of the Corresponding Formulas

- Based on the values $x_{ij} = x_i + \Delta x_{ij}$, where the st. dev. of Δx_{ij} is σ_j^2, we compute $A_{jk} = \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} (x_{ij} - x_{ik})^2$.
- Then, we compute $\widetilde{V}_j = \dfrac{A_{jk} + A_{j\ell} - A_{k\ell}}{2}$.
- For $\Delta_j^2 = E\left[\left(\widetilde{V}_j - \sigma_j^2\right)^2\right]$, we get the value

$$\Delta_j^2 = \frac{1}{n} \cdot (2\sigma_j^4 + \sigma_j^2 \cdot \sigma_k^2 + \sigma_j^2 \cdot \sigma_\ell^2 + \sigma_k^2 \cdot \sigma_\ell^2).$$

- We do not know the exact values σ_j^2, but we do no know the estimates $\widetilde{V}_j$ for these values.
- Thus, we can estimate Δ_j as follows:

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

11. Derivation of the Corr. Formulas (cont-d)

- We want to estimate $E\left(\widetilde{V}_j - \Delta V_j \,\middle|\, \widetilde{V}_j - \Delta V_j \geq 0\right)$.
- The Gaussian variable ΔV_j has 0 mean and standard deviation Δ_j.
- Thus, ΔV_j can be represented as $t \cdot \Delta_j$, where t is a Gaussian random variable with 0 mean and st. dev. 1.
- In terms of the new variable t, the non-negativity condition $\widetilde{V}_j - \Delta V_j \geq 0$ takes the form $t \leq \delta_j \stackrel{\text{def}}{=} \dfrac{\widetilde{V}_j}{\Delta_j}$.
- So, the desired conditional mean is equal to

$$E\left(\widetilde{V}_j - \Delta_j \cdot t \,\middle|\, t \leq \delta_j\right) = \widetilde{V}_j + \frac{\Delta_j}{\sqrt{2\pi}} \cdot \frac{\exp\left(-\dfrac{\delta_j^2}{2}\right)}{\Phi(\delta_j)}.$$

12. 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

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

13. 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 $x_i = \dfrac{\sum \widetilde{\sigma}_j^{-2} \cdot x_{ij}}{\sum \widetilde{\sigma}_j^{-2}}$.

14. Acknowledgments

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence).