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

Omar Ochoa1,2,* , Aaron Velasco2,3,4 , Christian Servin2,4,5

1Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA
2Cyber-SHARE Center, University of Texas at El Paso, El Paso, TX 79968, USA
3Department of Geological Sciences, University of Texas at El Paso, El Paso, TX 79968, USA
4Computational Science Program, University of Texas at El Paso, El Paso, TX 79968, USA
5Information Technology Department, El Paso Community College, 919 Hunter, El Paso TX 79915, USA

Received 1 April 2012; Revised 22 March 2013

Abstract

In geophysics, we usually have several Earth models based on different types of data: seismic, gravity, etc. Each of these models captures some aspects of the Earth structure. To get the more description of the Earth, it is desirable to “fuse” these models into a single one. To appropriately fuse the models, we need to know the accuracy of different models. In this paper, we show that the traditional methods cannot be directly used to estimate these accuracies, and we propose a new method for such estimation.

Keywords: model fusion, accuracy of a model, geophysics

1 To Properly Fuse Geophysical Models, It is Important to Estimate Accuracy of Different Models

Need to fuse models: geophysics. One of the main objectives of geophysics is to determine the density \( \rho(x, y, z) \) at different depths \( z \) and at different geographical locations \( (x, y) \). There exist several methods for estimating the density: e.g., we can use seismic data [1], or we can use gravity measurement. Each of the techniques for estimating \( \rho \) has its own advantages and limitations: e.g., seismic measurements often lead to a more accurate value of \( \rho \) than gravity measurements, but 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 that would combine the advantages of all of these models.

Fusion: statistical approach. Similar situations are frequent in practice: we are interested in the value of a quantity, and we have reached the limit of the accuracy that can be achieved by using a single available measuring instrument. In this case, to further increase the estimation accuracy, we perform several measurements of the desired quantity \( x_i \) — by using the same measuring instrument or different measuring instruments — and combine the results \( x_{i1}, x_{i2}, \ldots, x_{im} \) of these measurement into a single more accurate estimate \( \hat{x}_i \); see, e.g., [3, 7].

The need for fusion appears when we have already extracted as much accuracy from each type of measurements as possible. This means, in particular, that we have found and eliminated the systematic errors (thus, the resulting measurement error has 0 mean), and 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 of the sum of a large number of independent small random variables is approximately normal. Thus, it is natural to assume that each measurement error \( \Delta x_{ij} \) is normally distributed with 0 mean and some variance \( \sigma_j^2 \). Then, the

*Corresponding author. Email: omar@miners.utep.edu (O. Ochoa).
probability density corresponding to $x_{ij}$ is
\[
\frac{1}{\sqrt{2\pi} \cdot \sigma_j} \cdot \exp \left( - \frac{(x_{ij} - x_i)^2}{2\sigma_j^2} \right).
\]
It is also reasonable to assume that measurement errors corresponding to different measurements are independent. Under this assumption, the overall probability density is equal to the product of the corresponding probability distributions
\[
L = \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi} \cdot \sigma_j} \cdot \exp \left( - \frac{\sum_{j=1}^{m} (x_{ij} - x_i)^2}{2\sigma_j^2} \right). \tag{1}
\]
According to the Maximum Likelihood Principle, we select the value $x_i$ for which the above probability $L$ is the largest possible. Since $\exp(a) \cdot \exp(b) = \exp(a + b)$, we get
\[
L = \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi} \cdot \sigma_j} \cdot \exp \left( - \frac{\sum_{j=1}^{m} (x_{ij} - x_i)^2}{2\sigma_j^2} \right). \tag{2}
\]
Maximizing $L$ is equivalent to minimizing $-\ln(L)$, i.e., to minimizing the sum $\sum_{j=1}^{m} (x_{ij} - x_i)^2/\sigma_j^2$. Differentiating this sum w.r.t. $x_i$ and equating the derivative to 0, we conclude that $\sum_{j=1}^{m} \sigma_j^{-2} \cdot (x_{ij} - x_i) = 0$, so
\[
x_i = \frac{\sum_{j=1}^{m} \sigma_j^{-2} \cdot x_{ij}}{\sum_{j=1}^{m} \sigma_j^{-2}}. \tag{3}
\]
This idea has been successfully applied to geophysics; see, e.g., [2, 3, 4, 6].

Need to estimate accuracy of the corresponding models. To apply the above formula, we need to know the accuracies $\sigma_j$ of different models.

2 Traditional Methods of Estimating Accuracy Cannot be Directly Used in Geophysics

Let us describe the traditional methods of estimating accuracy (see, e.g., [5]) and let us show that these methods can be directly applied to the above geophysical problem.

First method: calibration. The first method is to calibrate the corresponding measuring instrument. Calibration is possible when we have a “standard” measuring instrument which is several times more accurate than the instrument which we are calibrating. We then repeatedly measure the same quantity by using both our measuring instrument and the standard one. Since the standard instrument is much more accurate than the one we testing, the result $x_{i,\text{st}}$ of using this instrument is practically equal to the actual value $x_i$, and thus, the measurement error $\Delta x_{ij} = x_{ij} - x_i$ is well approximated by the difference $\Delta x_{ij} \approx x_{ij} - x_{i,\text{st}}$ between the measurement results.

Since all the measurements $x_{ij}$, $i = 1, \ldots, n$, are performed by the same measuring instrument $j$, all these measurements have the same standard deviation $\sigma_j$. In this case, the likelihood (2) take the simplified form
\[
L = \frac{1}{(\sqrt{2\pi})^n \cdot \sigma_j^n} \cdot \exp \left( - \frac{\sum_{i=1}^{n} (x_{ij} - x_i)^2}{2\sigma_j^2} \right). \tag{4}
\]
We need to find the value $\sigma_j$ for which the likelihood $L$ attains the largest possible value. Maximizing $L$ is equivalent to minimizing
\[
-ln(L) = \text{const} + n \cdot \ln(\sigma_j) + \sum_{i=1}^{n} \frac{(x_{ij} - x_i)^2}{2\sigma_j^2}.
\]
Differentiating this sum w.r.t. \( \sigma_j \) and equating the derivative to 0, we get the usual estimation

\[
\sigma_j^2 = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - x_i)^2 .
\]  

(5)

Since we know approximate values of \( x_{ij} - x_i \), we can thus estimate \( \sigma_j \).

**It is not possible to directly use calibration.** For calibration to work, we need to have a measuring instrument which is several times more accurate than the one that we currently use. In geophysics, however, seismic (and other) methods are state-of-the-art, no method leads to more accurate determination of the densities. As a result, calibration techniques cannot be directly applied to estimating approximation errors in the geophysics problems.

**Second method: using several similar instruments.** In some practical situations, when we do have a standard measuring instrument, we can instead compare the results \( x_{i1} \) and \( x_{i2} \) of using two similar measuring instruments to measure the same quantities \( x_i \). The two instruments are independent and have the same accuracy \( \sigma \), so the likelihood function has the form

\[
L = \frac{1}{(\sqrt{2\pi} \cdot \sigma)^n} \cdot \exp\left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_{i1} - x_i)^2 \right) \cdot \frac{1}{(\sqrt{2\pi} \cdot \sigma)^m} \cdot \exp\left( -\frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_{i2} - x_i)^2 \right).
\]

In this case, we do not know \( \sigma \) and we do not know the actual values \( x_1, \ldots, x_m \); in the spirit of the Maximum Likelihood method, we will select the values of all these parameters for which the likelihood attains the largest possible value. Maximizing \( L \) is equivalent to minimizing

\[
-\ln(L) = \text{const} + 2n \cdot \ln(\sigma) + \sum_{i=1}^{n} \frac{(x_{i1} - x_i)^2}{2\sigma^2} + \sum_{i=1}^{m} \frac{(x_{i2} - x_i)^2}{2\sigma^2}.
\]

(6)

Minimizing with respect to \( x_i \) leads to \( x_i = (x_{i1} + x_{i2})/2 \). Substituting these values \( x_i \) into the formula (7) and minimizing the resulting expression with respect to \( \sigma \), we get

\[
\sigma^2 = \frac{1}{2n} \sum_{i=1}^{n} (x_{i1} - x_{i2})^2.
\]

(7)

**It is not possible to directly use this method either.** In usual measurements, when we estimate the accuracy of measurements performed by a measuring instrument, we can produce two similar measuring instruments and compare their results. In geophysics, we want to estimate the accuracy of a model, e.g., a seismic model, a gravity-based model, etc. In this situation, we do not have two similar applications of the same model, so the second method cannot be directly applied either.

Moreover, **Maximum Likelihood approach cannot be applied to estimate model accuracy.** Let us now consider the most general situation: we have several quantities with (unknown) actual values \( x_1, \ldots, x_n \), we have several measuring instruments (or geophysical methods) with (unknown) accuracies \( \sigma_1, \ldots, \sigma_m \), and we know the results \( x_{ij} \) of measuring the \( i \)-th quantity by using the \( j \)-th measuring instrument. At first glance, a reasonable idea is to find all the unknown quantities – i.e., the actual values \( x_i \) and the \( \sigma_j \) – from the Maximum Likelihood method. In this case, the likelihood takes the form

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

(8)

The problem with this approach is that, in contrast to the previous cases, this expression does not attain a finite maximum, it can reach values which are as large as possible. Namely, if we pick some \( j_0 \) and take \( x_i = x_{ij_0} + \varepsilon \) and \( \sigma_{j_0} = \varepsilon \), then we get \((x_{ij_0} - x_i)^2/2\sigma_{j_0}^2 = 1/2\), so the corresponding exponential factor is equal to \( \exp(-1/2) \); all other factors are also finite (and positive) in the limit \( \varepsilon \to 0 \) except for the terms \( 1/(\sqrt{2\pi} \cdot \sigma_{j_0}) \) which tends to infinity.
One can check that if all the values \( \sigma_j \) are positive, then the above likelihood expression attains finite values. Thus, the largest possible – infinite – value is attained when one of the standard deviations \( \sigma_{j_0} \) is equal to 0. In this case, in accordance with the formula (3), we get \( x_i = x_{i,j_0} \). In other words, for this problem, the Maximum Likelihood method leads to a counterintuitive conclusion that one of the measurements was absolutely accurate. This is not physically reasonable, so Maximum Likelihood method cannot be directly used to estimate random errors.

3 How to Estimate Model Accuracy: Proposed Idea

Analysis of the problem. We know that \( x_{ij} = x_i + \Delta x_{ij} \), where approximation errors \( \Delta x_{ij} = x_{ij} - x_i \) are independent normally distributed random variables with 0 mean and (unknown) standard deviations \( \sigma_j^2 \). For every two estimation methods (e.g., measuring instruments) \( j \) and \( k \), the difference \( x_{ij} - x_{ik} \) between the results of estimating the same quantity \( x_i \) by these two methods has the form

\[
x_{ij} - x_{ik} = (x_i + \Delta x_{ij}) - (x_i + \Delta x_{ik}) = \Delta x_{ij} - \Delta x_{ik}.
\]

Derivation of the resulting formula. The difference between two independent normally distributed random variables \( \Delta x_{ij} \) and \( \Delta x_{ik} \) is also normally distributed. The mean of the difference is equal to the difference of the means, i.e., to 0 - 0 = 0, and the variance of the difference is equal to the sum of the variances, i.e., to \( \sigma_j^2 + \sigma_k^2 \).

Thus, the difference \( x_{ij} - x_{ik} = \Delta x_{ij} - \Delta x_{ik} \) is normally distributed with 0 mean and variance \( \sigma_j^2 + \sigma_k^2 \). For each \( j \) and \( k \), we have \( n \) values \( x_{1j} - x_{1k}, \ldots, x_{nj} - x_{nk} \) from this distribution. Based on this sample, we can apply the usual formula (5) to estimate the standard deviation \( \sigma_j^2 + \sigma_k^2 \) as \( \sigma_j^2 + \sigma_k^2 \approx A_{jk} \), where

\[
A_{jk} \overset{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (x_{ij} - x_{ik})^2.
\] (9)

In particular, for every three different measuring instruments, with unknown accuracies \( \sigma_1^2, \sigma_2^2 \), and \( \sigma_3^2 \), we get the equations

\[
\sigma_1^2 + \sigma_2^2 \approx A_{12}, \quad \sigma_1^2 + \sigma_3^2 \approx A_{13}, \quad \sigma_2^2 + \sigma_3^2 \approx A_{23}.
\] (10)

By adding all three equalities (10) and dividing the result by two, we get

\[
\sigma_1^2 + \sigma_2^2 + \sigma_3^2 = \frac{A_{12} + A_{13} + A_{23}}{2}.
\] (11)

Resulting formulas. Subtracting, from (11), each of the equalities (10), we conclude that \( \sigma_j^2 \approx \tilde{V}_j \), where

\[
\tilde{V}_1 = \frac{A_{12} + A_{13} - A_{23}}{2}; \quad \tilde{V}_2 = \frac{A_{12} + A_{23} - A_{13}}{2}; \quad \tilde{V}_3 = \frac{A_{13} + A_{23} - A_{12}}{2}.
\] (12)

Comment. In general, when we have \( M \) different models, we have \( M \cdot (M - 1)/2 \) different equations \( \sigma_j^2 + \sigma_k^2 \approx A_{jk} \) to determine \( N \) unknowns \( \sigma_j^2 \). When \( M > 3 \), we have more equations than unknowns, so we can use the Least Squares method to estimate the desired values \( \sigma_j^2 \).

Challenge. The formulas \( \sigma_j^2 \approx \tilde{V}_j \) are approximate. If we use an estimate \( \tilde{V}_j \) for \( \sigma_j^2 \), we may get physically meaningless negative values for the corresponding variances.

It is therefore necessary to modify the formulas (12) so as to avoid negative values.

An idea of how to deal with this challenge. The negativity challenge is caused by the fact that the estimates in (12) are approximate. So, to come up with the desired modification, we will first estimate the accuracy of each of the formulas (12), i.e., the standard deviation \( \Delta_j \) for the difference \( \Delta V_j = \tilde{V}_j - \sigma_j^2 \).

For large \( n \), the difference \( \Delta V_j \) between the actual value of \( \sigma_j^2 \) and its statistical estimate is asymptotically normally distributed, with asymptotically 0 mean; see, e.g., [7]. In the next section, we will estimate the
standard deviation $\Delta_j$ for this difference. Thus, we can conclude that the actual value $\sigma_j^2 = \bar{V}_j - \Delta V_j$ is normally distributed with mean $V_j$ and standard deviation $\Delta_j$. We also know that $\sigma_j^2 \geq 0$. As an estimate for $\sigma_j^2$, it is therefore reasonable to use a conditional expected value $E \left( \bar{V}_j - \Delta V_j \bigg | \bar{V}_j - \Delta V_j \geq 0 \right)$. This new estimate is an expected value of a non-negative number and thus, cannot be negative. In the next section, we will show how to compute this new estimate.

4 Derivation of the Corresponding Formulas

Estimating accuracies $\Delta_j$ of the estimates $\bar{V}_j$ for $\sigma_j^2$. Let us estimate the accuracy $\Delta_j$ of $\bar{V}_j$, i.e., the expected value $\Delta_j^2 = E \left( (\bar{V}_j - \sigma_j^2)^2 \right)$. According to (12), $\bar{V}_j$ is computed based on the values

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

To simplify notations, let us denote $a_i = \Delta x_{ij}$, $b_i = \Delta x_{ik}$, and $c_i = \Delta x_{ik}$; then, we conclude that

$$\bar{V}_j = \frac{1}{2} \left[ \frac{1}{n} \sum_{i=1}^{n} (a_i - b_i)^2 + \frac{1}{n} \sum_{i=1}^{n} (a_i - c_i)^2 - \frac{1}{n} \sum_{i=1}^{n} (b_i - c_i)^2 \right],$$

e.i.,

$$\bar{V}_j = \frac{1}{2n} \sum_{i=1}^{n} [(a_i - b_i)^2 + (a_i - c_i)^2 - (b_i - c_i)^2].$$

Opening parentheses inside the sum, we get

$$(a_i - b_i)^2 + (a_i - c_i)^2 - (b_i - c_i)^2 = a_i^2 - 2a_i \cdot b_i + b_i^2 + a_i^2 - 2a_i \cdot c_i + c_i^2 - b_i^2 + 2b_i \cdot c_i - c_i^2.$$ 

Thus, the formula (13) takes the form

$$\bar{V}_j = \frac{1}{n} \sum_{i=1}^{n} (a_i^2 - a_i \cdot b_i - a_i \cdot c_i + b_i \cdot c_i).$$

Therefore,

$$\Delta_j^2 = E \left( (\bar{V}_j - \sigma_j^2)^2 \right) = E \left[ (\bar{V}_j)^2 - 2\bar{V}_j \cdot \sigma_j^2 + \sigma_j^4 \right] = E_1 - 2\sigma_1^2 \cdot E_2 + \sigma_4^2,$$

where

$$E_1 \overset{\text{def}}{=} E \left[ (\bar{V}_j)^2 \right] = E \left[ \left( \frac{1}{n} \sum_{i=1}^{n} (a_i^2 - a_i \cdot b_i - a_i \cdot c_i + b_i \cdot c_i) \right)^2 \right],$$

$$E_2 \overset{\text{def}}{=} E \left[ \bar{V}_j \right] = E \left[ \frac{1}{n} \sum_{i=1}^{n} (a_i^2 - a_i \cdot b_i + a_i \cdot c_i) \right].$$

The expected value $E_2$ is equal to linear combination of the expected values of the expressions $a_i^2$, $a_i \cdot b_i$, $a_i \cdot c_i$, and $b_i \cdot c_i$:

$$E_2 = \frac{1}{n} \sum_{i=1}^{n} (E[a_i^2] - E[a_i \cdot b_i] - E[a_i \cdot c_i] + E[b_i \cdot c_i]).$$

All variables $a_i$, $b_i$, and $c_i$ are independent and normally distributed with 0 mean and the corresponding variances $V_j = \sigma_j^2$. Due to independence, $E[a_i \cdot b_i] = E[a_i] \cdot E[b_i] = 0 \cdot 0 = 0$; similarly $E[a_i \cdot c_i] = E[b_i \cdot c_i] = 0$, and the only non-zero term is $E[a_i^2] = \sigma_j^2$. Thus, in the sum in $E_2$, only $n$ terms $a_1^2, \ldots, a_n^2$ lead to non-zero expected value $\sigma_j^2$, hence

$$E_2 = \frac{1}{n} \cdot n \cdot \sigma_j^2 = \sigma_j^2.$$
Let us now compute $E_1$. In general, the square of a sum can be represented as $(\sum z_i)^2 = \sum z_i^2 + \sum z_i \cdot z_i'$. In our case, $z_i = a_i^2 - a_i \cdot b_i - a_i \cdot c_i + b_i \cdot c_i$. Thus, the expected value $E_2$ can be presented as

$$E_1 = \frac{1}{n^2} \cdot \sum_{i=1}^{n} E[z_i^2] + \frac{1}{n^2} \cdot \sum_{i \neq i'} E[z_i \cdot z_i'] .$$  \hspace{1cm} (17)$$

Here, the expression $z_i^2 = (a_i^2 - a_i \cdot b_i - a_i \cdot c_i + b_i \cdot c_i)^2$ takes the form

$$z_i^2 = a_i^4 + a_i^2 \cdot b_i^2 + a_i^2 \cdot c_i^2 + b_i^2 \cdot c_i^2 + \text{terms which are odd in } a_i, b_i, \text{ or } c_i.$$  

Due to independence and the fact that all normally distributed variables $a_i$, $b_i$, and $c_i$ have 0 mean and thus, 0 odd moments, the expected values of odd terms like $a_i^3 \cdot b_i$ is zero: e.g., $E[a_i^3 \cdot b_i] = E[a_i^2] \cdot E[b_i] = 0$. Thus,

$$E[z_i^2] = E[a_i^4] + E[a_i^2 \cdot b_i^2] + E[a_i^2 \cdot c_i^2] + E[b_i^2 \cdot c_i^2] .$$

For the normal distribution, $E[a_i^4] = 3\sigma_j^4$; due to independence, $E[a_i^2 \cdot b_i^2] = E[a_i^2] \cdot E[b_i^2] = \sigma_j^2 \cdot \sigma_k^2$. Thus,

$$E[z_i^2] = 3\sigma_j^4 + \sigma_j^2 \cdot \sigma_k^2 + \sigma_j^2 \cdot \sigma_l^2 + \sigma_k^2 \cdot \sigma_l^2 ,$$

and

$$\frac{1}{n^2} \cdot \sum_{i=1}^{n} E[z_i^2] = \frac{1}{n} \cdot (3\sigma_j^4 + \sigma_j^2 \cdot \sigma_k^2 + \sigma_j^2 \cdot \sigma_l^2 + \sigma_k^2 \cdot \sigma_l^2) .$$  \hspace{1cm} (18)$$

For $z_i \cdot z_i'$ with $i \neq i'$, we similarly have

$$z_i \cdot z_i' = (a_i^2 - a_i \cdot b_i - a_i \cdot c_i + b_i \cdot c_i) \cdot (a_i'^2 - a_i' \cdot b_i' - a_i' \cdot c_i' + b_i' \cdot c_i') = a_i^2 \cdot a_i'^2 + \text{odd terms with 0 mean} .$$

Thus, $E[z_i \cdot z_i'] = E[a_i^2 \cdot a_i'^2] = E[a_i^2] \cdot E[a_i'^2] = \sigma_j^2 \cdot \sigma_j' = \sigma_j^2$ and so, after adding over all $n^2 - n$ pairs $(i, i')$ with $i \neq i'$, we get

$$\frac{1}{n^2} \cdot \sum_{i \neq i'} E[z_i \cdot z_i'] = \frac{n^2 - n}{n^2} \cdot \sigma_j^2 = \left(1 - \frac{1}{n}\right) \cdot \sigma_j^2 .$$  \hspace{1cm} (19)$$

Substituting the expressions (18) and (19) into the formula (17), we conclude that

$$E_1 = \frac{1}{n} \cdot (3\sigma_j^4 + \sigma_j^2 \cdot \sigma_k^2 + \sigma_j^2 \cdot \sigma_l^2 + \sigma_k^2 \cdot \sigma_l^2) + \left(1 - \frac{1}{n}\right) \cdot \sigma_j^4 .$$

Substituting this expression for $E_1$ and the formula $E_2 = \sigma_j^2$ into the formula (14), we get

$$\Delta_j^2 = \frac{1}{n} \cdot (3\sigma_j^4 + \sigma_j^2 \cdot \sigma_k^2 + \sigma_j^2 \cdot \sigma_l^2 + \sigma_k^2 \cdot \sigma_l^2) + \left(1 - \frac{1}{n}\right) \cdot \sigma_j^4 - 2\sigma_j^4 + \sigma_j^4 ,$$

i.e.,

$$\Delta_j^2 = \frac{1}{n} \cdot (2\sigma_j^4 + \sigma_j^2 \cdot \sigma_k^2 + \sigma_j^2 \cdot \sigma_l^2 + \sigma_k^2 \cdot \sigma_l^2) .$$  \hspace{1cm} (20)$$

We do not know the exact values $\sigma_j^2$, but we do no know the estimates $\tilde{V}_j$ for these values; thus, we can estimate $\Delta_j$ as follows:

$$\Delta_j^2 \approx \frac{1}{n} \cdot \left((\tilde{V}_j)^2 + \tilde{V}_j \cdot \tilde{V}_k + \tilde{V}_j \cdot \tilde{V}_l + \tilde{V}_k \cdot \tilde{V}_l\right) .$$  \hspace{1cm} (21)$$
From estimating $\Delta_j$ to a non-negative estimate for $\sigma_j^2$. So far, we have an estimate $\overline{V}_j$ for $\sigma_j^2$ (as defined by the formula (12)), we know that the difference $\Delta V_j = \overline{V}_j - \sigma_j^2$ is normally distributed with 0 mean, and we know the standard deviation $\Delta_j$ of this difference. Since, as we mentioned in the previous section, the original estimate $\overline{V}_j$ may be negative, it is desirable to use a new estimate $E\left(\overline{V}_j - \Delta V_j \mid \overline{V}_j - \Delta V_j \geq 0\right)$.

The Gaussian variable $\Delta V_j$ has 0 mean and standard deviation $\Delta_j$; thus, it can be represented as $t \cdot \Delta_j$, where $t$ is a Gaussian random variable with 0 and standard deviation 1. In terms of the new variable $t$, the non-negativity condition $\overline{V}_j - \Delta V_j \geq 0$ takes the form $\overline{V}_j - \Delta_j \cdot t \geq 0$, i.e., $t \leq \delta_j \equiv \overline{V}_j/\Delta_j$. Thus, the desired conditional mean is equal to

$$E\left(\overline{V}_j - \Delta_j \cdot t \mid t \leq \delta_j\right) = E\left(\overline{V}_j \mid t \leq \delta_j\right) - \Delta_j \cdot E\left(t \mid t \leq \delta_j\right) = \overline{V}_j - \Delta_j \cdot E\left(t \mid t \leq \delta_j\right).$$

(22)

So, to compute the desired estimate, it is sufficient to be able to compute the value $E\left(t \mid t \leq \delta_j\right)$ for the standard Gaussian variable $t$, with the probability density function

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

By definition, this conditional mean is equal to the ratio $E\left(t \mid t \leq \delta_j\right) = N_j/D_j$, where

$$N_j = \int_{-\infty}^{\delta_j} t \cdot \rho(t) \, dt; \quad D_j = \int_{-\infty}^{\delta_j} \rho(t) \, dt.$$

(23)

The denominator $D_j$ is equal to $\Phi(\delta_j) \equiv \text{Prob}(t \leq \delta_j)$. The numerator $N_j$ of this formula is equal to

$$N_j = \int_{-\infty}^{\delta_j} t \cdot \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{t^2}{2}\right) \, dt.$$

(24)

By introducing a new variable $s = t^2/2$ for which $ds = t \cdot dt$, we reduce (24) to

$$N_j = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{\delta_j^2/2} \exp(-s) \, ds.$$

This integral can be explicitly computed, so we get

$$N_j = -\frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{\delta_j^2}{2}\right),$$

and thus,

$$E\left(t \mid t \leq \delta_j\right) = -\frac{1}{\sqrt{2\pi}} \cdot \frac{\exp\left(-\frac{\delta_j^2}{2}\right)}{\Phi(\delta_j)}.$$

So,

$$E\left(\overline{V}_j - \Delta_j \cdot t \mid t \leq \delta_j\right) = \overline{V}_j - \Delta_j \cdot E\left(t \mid t \leq \delta_j\right) = \overline{V}_j + \frac{\Delta_j}{\sqrt{2\pi}} \cdot \frac{\exp\left(-\frac{\delta_j^2}{2}\right)}{\Phi(\delta_j)}.$$

5 Resulting Algorithm

Let us assume that for each value $x_i$ ($i = 1, \ldots, n$), we have three estimates $x_{i1}$, $x_{i2}$, and $x_{i3}$ corresponding to three different models. Our objective is to estimate the accuracies $\sigma_j^2$ 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
\[ \tilde{V}_1 = \frac{A_{12} + A_{13} - A_{23}}{2}, \quad \tilde{V}_2 = \frac{A_{12} + A_{23} - A_{13}}{2}, \quad \tilde{V}_3 = \frac{A_{13} + A_{23} - A_{12}}{2}. \]

After that, for each \( j \), we compute
\[ \Delta_j^2 = \frac{1}{n} \cdot \left( (\tilde{V}_j)^2 + \tilde{V}_j \cdot \tilde{V}_k + \tilde{V}_j \cdot \tilde{V}_\ell + \tilde{V}_k \cdot \tilde{V}_\ell \right). \]

Once we compute the preliminary estimates \( \tilde{V}_j \) and their accuracies \( \Delta_j \), we then compute the auxiliary ratios \( \delta_j = \tilde{V}_j / \Delta_j \) and return, as an estimate \( \tilde{\sigma}_j^2 \) for \( \sigma_j^2 \), the value
\[ \tilde{\sigma}_j^2 = \tilde{V}_j + \frac{\Delta_j}{\sqrt{2\pi}} \cdot \frac{\exp \left( -\frac{\delta_j^2}{2} \right)}{\Phi(\delta_j)}. \]

**Acknowledgments**

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

**References**


