Maximum Entropy in Support of Semantically Annotated Datasets

Paulo Pinheiro da Silva, Vladik Kreinovich, and
Christian Servin

Department of Computer Science,
University of Texas at El Paso,
500 W. University,
El Paso, TX 79968, USA
paulo@utep.edu, vladik@utep.edu
christians@miners.utep.edu
http://www.cs.utep.edu/paulo
http://www.cs.utep.edu/vladik

Abstract. One of the important problems of semantic web is checking whether two datasets describe the same quantity. The existing solution to this problem is to use these datasets' ontologies to deduce that these datasets indeed represent the same quantity. However, even when ontologies seem to confirm the identify of the two corresponding quantities, it is still possible that in reality, we deal with somewhat different quantities. A natural way to check the identity is to compare the numerical values of the measurement results: if they are close (within measurement errors), then most probably we deal with the same quantity, else we most probably deal with different ones. In this paper, we show how to perform this checking, and how to use similar techniques to estimate the uncertainty of the results of data processing.

Key words: semantic web, ontology, uncertainty, probabilistic approach, Maximum Entropy approach, data processing, safety factors

1 Checking Whether Two Datasets Represent the Same Data: A Problem

Formulation of the problem. In the semantic web, data are often encoded in Resource Description Framework (RDF) [9]. In RDF, every piece of information is represented as a triple consisting of a *subject*, a *predicate*, and an *object*. For example, when we describe the result of measuring the gravitation field, the coordinates at which we perform the measurements for a subject, a predicate is a term indicating that the measured quantity is a gravitational field (e.g., a term *hasGravityReading*), an the actual measurement result is an object.

In general, an RDF-based scientific dataset can be viewed as a (large) graph of RDF triples. One of the hard-to-solve problems is that triples in two different datasets using the same predicate *hasGravityReading* may not mean the same

thing just because the predicates have the same name. One way to check this is to use semantics, i.e., to specify the meanings of the terms used in both datasets by an appropriate ontology, and then use reasoning to verify that the meaning of the terms is indeed the same. In the gravity example, we conclude that the predicate *hasGravityReading* has the same meaning in both datasets if in both datasets, this meaning coincides with *sweet:hasGravityReading*, the meaning of this term in one of the the Semantic Web for Earth and Environmental Terminology (SWEET) ontologies [10] that deals with gravity.

Need to take uncertainty into account. Even when ontologies seem to infer that we are dealing with the same concept, there is still a chance that the two datasets talk about slightly different concepts. To clarify the situation, we can use the fact that often, the two datasets contain the values measured at the same (or almost the same) locations. In such cases, to confirm that we are indeed dealing with the same concept, we can compare the corresponding measurement results $x'_1, \ldots, x'_n$ and $x''_1, \ldots, x''_n$. Due to measurement uncertainty, the measured values x'_i and x''_i are, in general, slightly different.

The question is: *Based on the semantically annotated measurement results and the known information about the measurement uncertainty, how can we use the uncertainty information to either reinforce or question whether two datasets namely representing the same data may not be the same data.*

2 Checking Whether Two Datasets Represent the Same Data: Towards a Solution

Probabilistic approach to measurement uncertainty. To answer the above question, we must start by analyzing how the measurement uncertainty is represented. In this paper, we consider the traditional probabilistic way of describing measurement uncertainty.

In the engineering and scientific practice, we usually assume that for each measuring instrument, we know the probability distribution of different values of measurement error $\Delta x'_i \stackrel{\text{def}}{=} x'_i - x_i$. This assumption is often reasonable, since we can *calibrate* each measuring instrument by comparing the results of this measuring instrument with the results of a "standard" (much more accurate) one. The differences between the corresponding measurement results form the sample from which we can extract the desired distribution.

Often, after the calibration, it turns out that the tested measuring instrument is somewhat *biased* in the sense that the mean value of the measurement error is different from 0. In such cases, the instrument is usually re-calibrated – by subtracting this bias (mean) from all the measurement results – to make sure that the mean is 0. Thus, without losing generality, we can also assume that the mean value of the measurement error is 0: $E[\Delta x'_i] = 0$.

The degree to which the measured value x'_i differs from the actual value x_i is usually measured by the *standard deviation* $\sigma'_i \stackrel{\text{def}}{=} \sqrt{E[(\Delta x'_i)^2]}$. In addition to

standard deviation, we can also estimate other characteristics of the corresponding probability distribution, e.g., its third central moment (skewness) describes the degree of a asymmetry), the fourth central moment (excess) describes the "heaviness" of the distribution's tails, etc.

Gaussian distribution: justification and consequences. The measurement error is usually caused by a large number of different independent factors. It is known that under certain reasonable conditions, the joint effect of a large number of small independent factors has a probability distribution which is close to Gaussian; the corresponding results – known as *Central Limit Theorems* [11] – are the main reason why Gaussian (normal) distribution is indeed widely spread in practice.

As a result, it is reasonable to assume that the distribution for $\Delta x_i'$ is Gaussian. It is known that a Gaussian distribution is uniquely determined by its mean and its standard deviation. Since the mean value of the measurement error is 0, to describe the measurement uncertainty, it is sufficient to describe the standard deviation σ_i'.

Towards a solution. We do not know the actual values x_i, we only know the measurement results x_i' and x_i'' from the two datasets. For each i, the difference between these measurement results can be described in terms of the measurement errors:

$$\Delta x_i \stackrel{\text{def}}{=} x_i' - x_i'' = (x_i' - x_i) - (x_i'' - x_i) = \Delta x_i' - \Delta x_i''.$$

It is reasonable to assume that this difference is also normally distributed. Since the mean values of $\Delta x_i'$ and $\Delta x_i''$ are zeros, the mean value of their difference Δx_i is also 0, so it is sufficient to find the standard deviation $\sigma_i = \sqrt{V_i}$ of Δx_i. In general, for the sum of two Gaussian variables, we have

$$\sigma_i^2 = (\sigma_i')^2 + (\sigma_i'')^2 + 2r_i \cdot \sigma_i' \cdot \sigma_i'',$$

where $r_i = \dfrac{E[\Delta x_i' \cdot \Delta x_i'']}{\sigma_i' \cdot \sigma_i''}$ is the correlation between the i-th measurement errors. It is known that the correlation r_i can take all possible values from the interval $[-1, 1]$:

- the value $r_i = 1$ corresponds to the maximal possible (perfect) positive correlation, when $\Delta x_i'' = a \cdot \Delta x_i' + b$ for some $a > 0$;
- the value $r_i = 0$ corresponds to the case when measurement errors are independent;
- the value $r_i = -1$ corresponds to the maximal possible (perfect) negative correlation, when $\Delta x_i'' = a \cdot \Delta x_i' + b$ for some $a < 0$.

Other values correspond to imperfect correlation. The problem is that usually, we have no information about the correlation between measurement errors from different datasets.

First idea: assume independence. A usual practical approach to situations in which we have no information about possible correlations is to assume that the measurement errors are independent.

A possible (somewhat informal) justification of this assumption is as follows. Each correlation r_i can take any value from the interval $[-1, 1]$. We would like to choose a single value r_{ij} from this interval.

We have no information why some values are more reasonable than others, whether non-negative correlation is more probable or non-positive correlation is more probable. Thus, our information is invariant with respect to the change $r_i \to -r_i$, and hence, the selected correlation value r_i must be invariant w.r.t. the same transformation. Thus, we must have $r_i = -r_i$, thence $r_i = 0$. A somewhat more formal justification of this selection can be obtained from the Maximum Entropy approach (see the following text). Under the independence assumption, we have $(\sigma_i)^2 = (\sigma_i')^2 + (\sigma_i'')^2$.

Once we know the values, we can use the χ^2 criterion (see, e.g., [11]) to check whether with given degree of confidence α, the observed differences are consistent with the assumption that these differences are normally distributed with standard deviations σ_i:

$$\sum_{i=1}^{n} \frac{(\Delta x_i)^2}{(\sigma_i)^2} \le \chi^2_{n,\alpha}.$$

If this inequality is satisfied, i.e., if

$$\sum_{i=1}^{n} \frac{(\Delta x_i)^2}{(\sigma_i')^2 + (\sigma_i'')^2} \le \chi^2_{n,\alpha},$$

then we conclude that the two datasets indeed describe the same quantity. If this inequality is not satisfied, then most probably, the datasets describe somewhat different quantities.

On the other hand, there is another possibility: that the two datasets do describe the same quantity, but the measurement errors are indeed correlated.

An alternative idea: worst-case estimations. If the above inequality holds for some values σ_i, then it holds for larger values σ_i as well. To take into account the possibility of correlations, we should only reject the similarity hypothesis when the above inequality does not hold even for the largest possible values σ_i.

Since $|r_i| \le 1$, we have $(\sigma_i)^2 \le V_i \stackrel{\text{def}}{=} (\sigma_i')^2 + (\sigma_i'')^2 + 2\sigma_i' \cdot \sigma_i''$. The value V_i is attained for $\Delta x_i'' = -\dfrac{\sigma_i''}{\sigma_i'} \cdot \Delta x_i'$. So, the largest possible value of σ_i^2 is equal to V_i. One can easily check that $V_i = (\sigma_i' + \sigma_i'')^2$. Thus, in this case, if

$$\sum_{i=1}^{n} \left(\frac{\Delta x_i}{\sigma_i' + \sigma_i''} \right)^2 \le \chi^2_{n,\alpha},$$

then we conclude that the two datasets indeed describe the same quantity. If this inequality is not satisfied, then most probably, the datasets describe somewhat different quantities.

Conclusion. In this section, we considered the following question: Based on the semantically annotated measurement results and the known information about the measurement uncertainty, how can we use the uncertainty information to either reinforce or question whether two datasets namely representing the same data may not be the same data.

Specifically, we assume the some values from the two datasets contain the results of measuring the same quantity at the same location and/or moment of time. Let n denote the total number of such measurements, let $x'_1, \ldots, x'_n$ denote the corresponding results from the first dataset, and let $x''_1, \ldots, x''_n$ denote the measurement results from the second dataset. We assume that we know the standard deviations σ'_i and σ''_i of these measurements, and that we have no information about possible correlation between the corresponding measurement errors. In this case, we apply the Maximum Entropy approach, and conclude that if $\sum\limits_{i=1}^{n} \dfrac{(\Delta x_i)^2}{(\sigma'_i)^2 + (\sigma''_i)^2} \le \chi^2_{n,\alpha}$, where $\chi^2_{n,\alpha} \approx n$ is the value of the χ^2-criterion for the desired certainty α, then this reinforces the original conclusion that the two datasets represent the same data. If the above inequality is not satisfied, then we conclude that either the two datasets represent different data (or, alternatively, that the measurement uncertainty values σ'_i and σ''_i are underestimated).

If we have reasons to suspect that the measurement errors corresponding to two databases may be correlated, then can be more cautious and reinforce the original conclusion even when a weaker inequality is satisfied: $\sum\limits_{i=1}^{n} \left(\dfrac{\Delta x_i}{\sigma'_i + \sigma''_i} \right)^2 \le \chi^2_{n,\alpha}$.

3 Estimating Uncertainty of the Results of Data Processing: Introduction to the Problem

So far, we have considered an important current problem of semantic web: how to check whether two datasets describe the same quantity. A similar approach turns out to be useful also in solving similar problems related to the future use of semantic web for data processing: namely, the problem of estimating uncertainty of the results of data processing.

Need for data processing. In many practical situations, we are interested in the value of a physical quantity y which is difficult (or even impossible) to measure directly. For example, in geophysics, we would like to know the density and the velocity of sound at different geographic locations and different depths.

Since it is difficult to measure y directly, a natural solution is to measure y *indirectly*:

- first, we measure easier-to-measure auxiliary quantities $x_1, \ldots, x_n$ which are related to y by a known algorithmic dependence $y = f(x_1, \ldots, x_n)$;
- then, we apply the algorithm f to the results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ of these auxiliary measurements and produce the desired estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ for y.

For example, to measure the velocity of sound at different locations and at different depths, we set up explosions and measure the time during which the sound waves generated by these explosions travel to different locations. Similarly, to measure densities at different locations, we can measure the local variations of the gravitational field generated by the different densities; see, e.g., [7].

Need for estimating uncertainty of the result of data processing. Measurements are never 100% accurate: the result $\widetilde{x}_i$ of measuring a physical quantity is, in general, different from the actual (unknown) value x_i of this quantity. Due to this uncertainty, the result $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of data processing is, in general, different from the ideal value $y = f(x_1, \ldots, x_n)$ of the desired quantity y.

Often, there is another reason why the estimate $\widetilde{y}$ is different from the actual value y: usually, the dependence $y = f(x_1, \ldots, x_n)$ is only approximately known and/or y may also depend on the values of some other quantities whose values we do not know. In this paper, we mainly concentrate on the situations in which the dependence f is known with high accuracy and for which, therefore, uncertainty in y is mainly due to the uncertainty of measuring x_i. For example, in the above geophysical example, we have a pretty good physical understanding of how the sound propagates through a solid body; however, since we only can only measure travel times with some uncertainty, we can only get the resulting velocities with uncertainty. To use our techniques in other situations, we must also take into account the uncertainty with which we know the dependence f.

In situations in which the uncertainty in f can be ignored, the resulting uncertainty in y is mainly caused by the uncertainties in x_i.

Linearization. Measurement errors Δx_i are usually relatively small. As a result, terms which are quadratic and of higher order in terms of Δx_i can be usually safely ignored. For example, if $\Delta x_i \approx 10\%$, then $(\Delta x_i)^2 \approx 1\% \ll 10\%$. The resulting *linearization* of the dependence of Δy on Δx_i can be used to simplify computations.

Specifically, by definition of the measurement error $\Delta x_i = \widetilde{x}_i - x_i$, hence $x_i = \widetilde{x}_i - \Delta x_i$. So, the expression for Δy takes the form

$$\Delta y = \widetilde{y} - y = f(\widetilde{x}_1, \ldots, \widetilde{x}_i, \ldots, \widetilde{x}_n) - f(x_1, \ldots, x_n) =$$

$$f(\widetilde{x}_1, \ldots, \widetilde{x}_i, \ldots, \widetilde{x}_n) - f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_i - \Delta x_i, \ldots, \widetilde{x}_n - \Delta x_n).$$

When we expand this expression in Taylor series in terms of Δx_i and only keep linear terms, we get

$$y = f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_i - \Delta x_i, \ldots, \widetilde{x}_n - \Delta x_n) \approx$$

$$f(\widetilde{x}_1, \ldots, \widetilde{x}_i, \ldots, \widetilde{x}_n) - \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \cdot \Delta x_i$$

and hence $\Delta y = \widetilde{y} - y = \sum\limits_{i=1}^{n} c_i \cdot \Delta x_i$, where we denoted $c_i \stackrel{\text{def}}{=} \dfrac{\partial f}{\partial x_i}$.

Gaussian distribution: justification and consequences. For individual measurements, we can have different types of probability distributions. However, for the results of data processing, the situation is often different. Indeed, in data processing, we usually combine a large number of measurement results with different measurement errors. We have already mentioned that under certain reasonable conditions, the joint effect of a large number of small independent factors has a probability distribution which is close to Gaussian. As a result, it is reasonable to assume that the distribution for Δy is Gaussian. Since a Gaussian distribution is uniquely determined by its mean and its standard deviation, it is sufficient to find the mean $E \stackrel{\text{def}}{=} E[\Delta y]$ and the standard deviation $\sigma = \sqrt{V}$ of Δy, where $V \stackrel{\text{def}}{=} V[\Delta y] = E[(\Delta y - E)^2]$.

The mean $E[\Delta y]$ is easy to estimate: since Δy is a linear combination of the measurement errors Δx_i, and the measurement errors are assumed to have 0 mean $E[\Delta x_i] = 0$, we conclude that $E = E[\Delta y] = \sum\limits_{i=1}^{n} c_i \cdot E[\Delta x_i] = 0$. For the variance V, we similarly get the formula

$$V = E[(\Delta y)^2] = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i \cdot c_j \cdot E[\Delta x_i \cdot \Delta x_j].$$

Here, for $i = j$, by definition of standard deviation, we have $E[(\Delta x_i)^2] = \sigma_i^2$. For $i \neq j$, we have $E[\Delta x_i \cdot \Delta x_j] = r_{ij} \cdot \sigma_i \cdot \sigma_j$, where $r_{ij} = \dfrac{E[\Delta x_i \cdot \Delta x_j]}{\sigma_i \cdot \sigma_j}$ is the correlation between the i-th and the j-th measurement errors. Thus, we have

$$\sigma^2 = V = \sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2 + \sum_{i \neq j} r_{ij} \cdot c_i \cdot c_j \cdot \sigma_i \cdot \sigma_j.$$

In the traditional data processing, practitioners usually thoroughly analyze the situation and come up with reasonable estimates for the correlations r_{ij}; thus, we get reasonable estimates for the the standard deviation σ of the approximation error Δy.

Based on the Gaussian character of the distribution, we can conclude, e.g., that with probability $\approx 90\%$, we have $|\Delta y| \leq 2\sigma$.

4 Estimating Uncertainty of the Results of Data Processing: Specific Problems of Web Services

Specific problem of uncertainty estimation for web services. Lately, more and more data processing is performed via web services, where different data at geographically different locations are automatically brought together and processed together. This new development leads to additional difficulties in estimating the uncertainty of the results of data processing.

Indeed, in the traditional data processing, practitioners usually thoroughly analyze the situation and come up with reasonable estimates for the correlations.

In the automatic data processing of web services, we often operate without an easy possibility of such an analysis, and thus, without any information about possible correlations.

First idea: assume independence. As we have mentioned, a usual practical approach to situations in which we have no information about possible correlations is to assume that the measurement errors are independent. Under the independence assumption, we have

$$\sigma^2 = V = \sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2.$$

Limitation of the independence assumption. The main limitation of the independence assumption is that in many practical cases, this assumption leads to a drastic underestimation of the corresponding uncertainty σ. In particular, this is the case with the above geophysics example of determining velocities, where the independence assumption leads to unrealistic estimates of $\sigma \approx 0.01$ km/sec at depths of 40 km – while a simple difference between two different method of estimating the velocity leads to differences of the size 1 km/sec or even more [7].

An alternative idea: worst-case estimations. The independence approach is reasonable if we try to find a *single* most representative set of correlation coefficients r_{ij}, and use this set to estimate uncertainty. Since this approach leads to an underestimation of σ, a natural next idea is to try *all* possible combinations of the correlations r_{ij}, and to use the worst-case (largest) value of σ as the desired estimate for the uncertainty of y.

From the fact that $|r_{ij}| \le 1$, we can easily conclude that $r_{ij} \cdot c_i \cdot c_j \cdot \sigma_i \cdot \sigma_j \le |c_i| \cdot |c_j| \cdot \sigma_i \cdot \sigma_j$, and thus, that $V \le V_w$, where $V_w \stackrel{\text{def}}{=} \sum_{i=1}^{n} \sum_{j=1}^{n} |c_i| \cdot |c_j| \cdot \sigma_i \cdot \sigma_j$. One can easily check that the expression V_w is a full square, i.e., that $V_w = \sigma_w^2$, where $\sigma_w \stackrel{\text{def}}{=} \sum_{i=1}^{n} |c_i| \cdot \sigma_i$. This value can indeed be attained for appropriate correlations. For example, if we take ξ as a normally distributed random variable with 0 mean and standard deviation 1, then for $\Delta x_i = \Delta_i \cdot \text{sign}(c_i) \cdot \xi$ (where $\text{sign}(a) \stackrel{\text{def}}{=} 1$ for $a > 0$ and $\text{sign}(a) \stackrel{\text{def}}{=} -1$ for $a < 0$), we get $r_{ij} = \text{sign}(c_i) \cdot \text{sign}(c_j)$, therefore, $r_{ij} \cdot c_i \cdot c_j \cdot \sigma_i \cdot \sigma_j = |c_i| \cdot |c_j| \cdot \sigma_i \cdot \sigma_j$ and thus, $V = V_w$.

So, in the worst-case approach, we return σ_w as the desired estimate for the uncertainty of y.

Comment: relation to interval computations. It is worth mentioning that from the computational viewpoint, a similar formula $\Delta = \sum_{i=1}^{n} |c_i| \cdot \Delta_i$ describes the largest possible value of $|\Delta y|$ in the *interval computations* case, when we do not have any information about the probabilities of different values Δx_i, we only know that $\Delta x_i \in [-\Delta_i, \Delta_i]$ for some known values Δ_i. Thus, to transform the

values σ_i into the worst-case estimate σ_w, we can use the techniques developed in interval computations to transform the values of Δ_i into the estimate Δ; see, e.g. [4].

Limitation of the worst-case approach. The main limitation of the independence assumption is that in many practical cases, this assumption leads to a drastic overestimation of the corresponding uncertainty σ. In particular, this is the case with the above geophysics example of determining velocities, where the worst-case approach leads to useless estimates of $\sigma_w \gg 10$ km/sec at depths of 40 km – while without any measurements, we know that $\sigma_w \leq 8$ km/sec [7].

In the following section, we describe a new approach that will lead us to more adequate estimates. This approach is based on the Maximum Entropy techniques.

5 A New Approach Based on Maximum Entropy Techniques

Maximum Entropy techniques: reminder. In the traditional error estimation for the result of data processing, we know the means (they are 0s), we know the standard deviations σ_i, we know the correlations r_{ij}, and thus (under the Gaussian assumption), we know the actual n-dimensional distribution on the set of all possible of measurement errors $(\Delta x_1, \ldots, \Delta x_n)$.

The main problem of uncertainty estimation for web services is that we often do not know the values of the correlations r_{ij}. As a result, instead of a single probability distribution in the n-dimensional space, we have many possible probability distributions corresponding to different possible combinations of the correlations r_{ij}.

This non-uniqueness situation is reasonably typical in mathematical statistics. A traditional statistical approach to the situation when several probability distributions are possible is to select the "most uncertain" distribution, i.e., the distribution which has the largest possible value of the entropy $S \stackrel{\text{def}}{=} -\int \rho(x) \cdot \ln(\rho(x))\,\mathrm{d}x$, where $\rho(x)$ denotes the probability density. For details on this Maximum Entropy approach and its relation to Laplace's principle of indifference, see, e.g., [1, 5, 6].

It is known that for a single variable x_1, among all distributions located on a given interval, the entropy is the largest when this distribution is *uniform* on this interval.

In general, if we know the marginal distributions $\rho_i(x_i)$ of several random variables ξ_i, and we do not have any information about their correlation, then the maximum entropy approach leads to the selection of a distribution in which all these variables are independent: $\rho(x_1, \ldots, x_n) = \rho_1(x_1) \cdot \ldots \cdot \rho_n(x_n)$. This is how the independence assumption is usually justified in the foundations of statistics.

In particular, if we only know that each of n random variables ξ_i is located on a certain interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$, then we can similarly conclude that the distribu-

tion with the largest value of the entropy is the one which is uniformly distributed in the corresponding box $\boldsymbol{x}_1 \times \ldots \times \boldsymbol{x}_i \times \ldots \times \boldsymbol{x}_n$, i.e., a distribution in which each variable ξ_i is uniformly distributed on the corresponding interval $[\Delta_i, \Delta_i]$, and variables corresponding to different inputs are statistically independent.

Comment: relation to interval uncertainty. This is indeed one of the main ways how interval uncertainty is treated in engineering practice: if we only know that the value of some variable is in the interval $[\underline{x}_i, \overline{x}_i]$, and we have no information about the probabilities, then we assume that the variable x_i is uniformly distributed on this interval, and that different variables x_i are independent.

A straightforward use of maximum entropy techniques leads to underestimation. As we have mentioned, a straightforward use of the Maximum Entropy technique leads to the independence assumption and thus, often, to underestimation of the uncertainty in y.

New idea. Instead of trying to find a single combination of values r_{ij}, let us find a single *probability distribution* on the set of all possible values of r_{ij} for $i \neq j$.

Derivations based on the new idea. Since $r_{ij} = r_{ji}$, it is sufficient to describe the values r_{ij} for $i < j$. The only information that we have about each value r_{ij} is that $r_{ij} \in [-1, 1]$. Thus, in line with the above consequences of the Maximum Entropy approach, we assume that each value r_{ij} is uniformly distributed on the interval $[-1, 1]$, and that different values r_{ij} are independent random variables.

For the following computations, we use the fact that for the uniform distribution on the interval $[-1, 1]$, the mean is 0, and the variance is $1/3$.

For large n, the resulting estimate for the variance

$$V = \sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2 + 2 \sum_{i<j} r_{ij} \cdot c_i \cdot c_j \cdot \sigma_i \cdot \sigma_j$$

is the sum of large number of independent small random variables. Therefore, due to the Central Limit theorem, the resulting distribution of V is close to Gaussian. Thus, we can safely assume that the variable V has a Gaussian distribution.

In general, the mean of a linear combination $\xi = \sum a_i \cdot \xi$ of n independent random variables with means E_i and variances V_i is equal to $\sum a_i \cdot E_i$, and the variance of ξ is equal to $\sum a_i^2 \cdot V_i$. Thus, based on the known mean and variance of each of the independent random variables r_{ij}, we can compute the mean E_V and the variance V_V of the corresponding random variable V. Specifically, the mean E_V is equal to $E_V = \sum\limits_{i=1}^{n} c_i^2 \cdot \sigma_i^2$, i.e., to the variance V_{ind} corresponding to the case of independent measurement errors. The variance V_V of V is equal to $V_V = \dfrac{4}{3} \cdot \sum\limits_{i<j} c_i^2 \cdot c_j^2 \cdot \sigma_i^2 \cdot \sigma_j^2$. The expression for V_V can be rewritten as

$$V_V = \frac{2}{3} \cdot \sum_{i \neq j} c_i^2 \cdot c_j^2 \cdot \sigma_i^2 \cdot \sigma_j^2 = \frac{2}{3} \cdot \left(\sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2 \right)^2 - \frac{2}{3} \cdot \sum_{i=1}^{n} c_i^4 \cdot \sigma_i^4.$$

We are interested in the case when we process a large number of data points, i.e., when n is large. For large n, the sum $\sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2$ is proportional to n and thus, its square is proportional to n^2. On the other hand, the second sum $\sum_{i=1}^{n} c_i^4 \cdot \sigma_i^4$ is proportional to n and is, thus, much smaller than the square term – which is of size $O(n^2)$. Thus, we can safely ignore the second sum, and conclude that $V_V \approx \frac{2}{3} \cdot \left(\sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2 \right)^2$, and thus, that the standard deviation σ_V of the value V is (approximately) equal to $\sigma_V = \sqrt{\frac{2}{3}} \cdot \sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2$, i.e., to $\sigma_V = \sqrt{\frac{2}{3}} \cdot V_{\rm ind}$.

Thus, in line with the general properties of the Gaussian distribution, we can conclude with the corresponding probability, the actual value V is below the value $\widetilde{V} = E_V + k_0 \cdot \sigma_V = \left(1 + k_0 \cdot \sqrt{\frac{2}{3}} \right) \cdot V_{\rm ind}$. Hence, we can use $\widetilde{V}$ as the desired "almost" worst-case estimates for the variance of the results of web services, and we can the square root $\widetilde{\sigma} = \sqrt{\widetilde{V}} = \sqrt{1 + k_0 \cdot \sqrt{\frac{2}{3}}} \cdot \sigma_{\rm ind}$ of the variance estimate as the desired estimate for the standard deviation for Δy, where $\sigma_{\rm ind}$ is the estimate corresponding to the assumption of independent measurement errors.

Conclusion. For independent measurement errors, the standard deviation σ_i of the result $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of data processing is equal to $\sigma_{\rm ind} = \sqrt{\sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2}$, where σ_i is the standard deviation of the i-th measurement result $\widetilde{x}_i$ and $c_i = \frac{\partial f}{\partial x_i}(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

For the case when we do not have any information about the dependence of different measurement errors, it is reasonable to conclude that the standard deviation σ of y is bounded by the value $\widetilde{\sigma} = \sqrt{1 + k_0 \cdot \sqrt{\frac{2}{3}}} \cdot \sigma_{\rm ind}$. Here, the value k_0 determines the probability with which this conclusion holds:

- for $k_0 = 2$, the inequality $\sigma \le 1.62 \cdot \sigma_{\rm ind}$ holds with probability 95%;
- for $k_0 = 3$, the inequality $\sigma \le 1.86 \cdot \sigma_{\rm ind}$ holds with probability 99.95%;
- for $k_0 = 6$, the inequality $\sigma \le 2.43 \cdot \sigma_{\rm ind}$ holds with probability $1 - 0.5 \cdot 10^{-6}$.

In order words, to take into account possible correlations between the measurement errors, we must multiply the estimate based on the independence assumption by a constant factor depending on the desired reliability of this estimate.

Cases when this estimate makes physical sense: relation to the engineering idea of safety factors. A similar idea has been used in engineering for several centuries: e.g., to take uncertainty into account, engineers have multiplied the desired strength of a building by an empirical constant factor called *safety factor*,

usually from 1.5-2 for regular building to 3-4 for more critical constructions; see, e.g., [2, 3]. So, our factors are in good accordance with the empirical values obtained from the engineering practice.

Cases when other estimates are needed. The above estimates do not always make physical sense. For example, as we have mentioned, for the geophysical problem, the estimates based on the independence assumption are too low, and even if we multiply these estimates by a factor of 1.65-2.43, the results will still be too low.

From the physical viewpoint, this underestimation is cased by the fact that the measurement errors are positively correlated. For such cases, an alternative approach have been proposed in [7].

Acknowledgments. This work was partly supported by NSF grant HRD-0734825 and by NIH Grant 1 T36 GM078000-01.

References

1. Chokr, B., and Kreinovich, V.: How far are we from the complete knowledge: complexity of knowledge acquisition in Dempster-Shafer approach. In: Yager, R.R., Kacprzyk, J., Pedrizzi, M. (eds.), Advances in the Dempster-Shafer Theory of Evidence, pp. 555–576, Wiley, N.Y. (1994)
2. Elishakoff, I.: Interrelation Between Safety Factors and Reliability, NASA Technical Report NASA/CR2001-211309 (2001), available at http://gltrs.grc.nasa.gov/reports/2001/CR-2001-211309.pdf
3. Elishakoff, I.: Safety Factors and Reliability: Friends or Foes? Kluwer Academic Publishers, Dordrecht, The Netherlands (2004)
4. Jaulin, L., Kieffer, M., Didrit, O., and Walter, E.: Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control and Robotics, Springer-Verlag, London (2001)
5. Jaynes, E. T.: Probability Theory: The Logic of Science, Cambridge University Press (2003)
6. Klir, G. J.: Uncertainty and Information: Foundations of Generalized Information Theory. J. Wiley, Hoboken, New Jersey (2005)
7. Pinheiro da Silva, P., Velasco, A., Ceberio, M., Servin, C., Averill, M. G., Del Rio, N., Longpré, L., and Kreinovich, V.: Propagation and provenance of probabilistic and interval uncertainty in cyberinfrastructure-related data processing and data fusion, In: Muhanna, R. L., and Mullen, R. L. (eds.), Proceedings of the International Workshop on Reliable Engineering Computing REC'08, Savannah, Georgia, February 20–22, 2008, pp. 199–234 (2008)
8. Rabinovich, S.: Measurement Errors and Uncertainties: Theory and Practice. Springer-Verlag, New York (2005)
9. Resource Description Framework (RDF) http://www.w3.org/RDF/
10. Semantic Web for Earth and Environmental Terminology SWEET ontologies http://sweet.jpl.nasa.gov/ontology/
11. Sheskin, D.: Handbook of Parametric and Nonparametric Statistical Procedures, Chapman & Hall/CRC, Boca Raton, Florida (2004)