Z-Numbers: How They Describe Student Confidence and How They Can Explain (and Improve) Laplacian and Schroedinger Eigenmap Dimension Reduction in Data Analysis

Vladik Kreinovich, Olga Kosheleva, and Michael Zakharevich

Abstract Experts have different degrees of confidence in their statements. To describe these different degrees of confidence, Lotfi A. Zadeh proposed the notion of a Z-number: a fuzzy set (or other type of uncertainty) supplemented by a degree of confidence in the statement corresponding to fuzzy sets. In this chapter, we show that Z-numbers provide a natural formalization of the competence-vs-confidence dichotomy, which is especially important for educating low-income students. We also show that Z-numbers provide a natural theoretical explanation for several empirically heuristic techniques of dimension reduction in data analysis, such as Laplacian and Schroedinger eigenmaps, and, moreover, show how these methods can be further improved.

1 Need to Take Into Account Accuracy and Reliability When Processing Data

Need for data processing. In many practical situations, we are interested in the values of the quantities $x_1,\ldots,x_n$ which are difficult (or even impossible) to measure directly. For example, in GPS-based localization, we want to find where different objects (and we) are, i.e., we want to find the coordinates of different objects. However, it is not possible to directly measure coordinates.

What we can measure in such situations is some auxiliary quantities $y_1,\ldots,y_m$ that depend on the desired quantities x_i in a known way, i.e., for which $y_j = f_j(x_1,\ldots,x_n)$ for known algorithms f_i. For example, to find the location of an ob-

Vladik Kreinovich and Olga Kosheleva
University of Texas at El Paso, El Paso, TX 79968, USA
e-mail: {vladik,olga}@utep.edu

Michael Zakharevich
SeeCure Systems, Inc., 1040 Continentals Way #12, Belmont, CA 94002, USA
e-mail: michael@seecure360.com

ject, we can measure distances between objects and/or angles between directions towards different objects.

Once we know the results $\widetilde{y}_j$ of measuring the quantities y_j, we need to reconstruct the desired quantities x_i from the corresponding system of equations:

$$\widetilde{y}_1 \approx f_1(x_1,\ldots,x_n), \quad \ldots, \quad \widetilde{y}_m \approx f_m(x_1,\ldots,x_n); \tag{1}$$

we write approximately equal, because measurements are never absolutely accurate, there is usually a difference $\Delta y_j \stackrel{\text{def}}{=} \widetilde{y}_j - y_j$ (known as *measurement error*) between the measurement result $\widetilde{y}_j$ and the actual value y_j of the measured quantity.

The process of reconstructing x_i from $\widetilde{y}_j$ is an important case of *data processing*.

In some applications – e.g., in many medical situations – it is difficult to find related easier-to-measure quantities y_j, but we *can* find related quantities that can be well estimated by an expert: e.g., by the patient's appearance or reaction to different tests. In such situations, to reconstruct the desired quantities x_i, instead of the measurement results, we can use the expert estimates $\widetilde{y}_j$.

How to take accuracy into account when processing data: probabilistic case. In many practical situations, based on the previous experience of using the corresponding measuring instruments, we know the probabilities of different values of measurement errors. In precise terms, we know the corresponding probability density functions $\rho_j(\Delta y_j) = \rho_j(\widetilde{y}_j - f_j(x_1,\ldots,x_n))$.

Measurement errors corresponding to different measurements are usually independent. As a result, the overall probability of given observations is equal to the product of the corresponding probabilities:

$$\prod_{j=1}^{m} \rho_j(\widetilde{y}_j - f_j(x_1,\ldots,x_n)). \tag{1}$$

For different values of x_i, this probability is different. It is therefore reasonable to select the most probable combination $(x_1,\ldots,x_n)$, i.e., the combination for which the product (1) attains the largest possible value. This quantity (1) is known as *likelihood*, and the above idea is known as the *Maximum Likelihood method*; see, e.g., [17].

The probabilities $\rho_j(\Delta y_j)$ can be reasonably small, and the number of measurements is often large. The product of a large number of small values is often too small, sometimes smaller than the smallest positive real number in a usual computer representation. To avoid this problem, practitioners use the fact that maximizing a function is equivalent to minimizing its negative logarithm

$$\sum_{j=1}^{m} \psi_j(\widetilde{y}_j - f_j(x_1,\ldots,x_n)), \tag{2}$$

where we denoted $\psi_j(z) \stackrel{\text{def}}{=} -\ln(\rho_j(z))$.

Often, the measurement error is the result of a joint effect of a large number of independent factors. In such situations, due to the Central Limit Theorem (see, e.g.,

[17]), the distributions are close to Gaussian – and, be re-calibrating the measuring instruments, we can usually safely assume that the mean of the measurement error is 0. In this case, $\rho_j(\Delta y_j) \sim \exp\left(-\frac{(\Delta y_j)^2}{\sigma_j^2}\right)$, where σ_j is the standard deviation of the j-th distribution. Thus, minimizing the expression (2) is equivalent to minimizing the sum

$$\sum_{j=1}^{m} \frac{(\widetilde{y}_j - f_j(x_1,\ldots,x_n))^2}{\sigma_j^2}. \tag{3}$$

This is known as the *Least Squares method*.

In particular, if we do not have any reason to believe that different measurements have different accuracy, it makes sense to assume that they all have the same accuracy $\sigma_1 = \sigma_2 = \ldots$ In this case, (3) becomes equivalent to minimizing the sum

$$\sum_{j=1}^{m} (\widetilde{y}_j - f_j(x_1,\ldots,x_n))^2. \tag{3a}$$

How to take accuracy into account when processing data: fuzzy case. Often, instead of the probabilities of different values of the approximation error, we only have expert opinions about the possibility of different values. Describing these opinions in computer-understandable terms was one of the main motivations for fuzzy logic; see, e.g., [6, 10, 12, 14, 18]. It is therefore reasonable to describe these opinions in terms of the membership functions $\mu_j(\Delta y_j) = \mu_j(\widetilde{y}_j - f_j(x_1,\ldots,x_n))$.

In line with the general ideas of fuzzy logic, to describe the expert's degree of confidence that:

- the first approximation error is Δy_1, *and*
- the second approximation error is Δy_2,
- etc.,

we can apply the corresponding "and"-operation (t-norm) $f_\&(a,b)$, and get the value

$$f_\&(\mu_1(\widetilde{y}_1 - f_1(x_1,\ldots,x_n)),\ldots,\mu_m(\widetilde{y}_m - f_m(x_1,\ldots,x_n))). \tag{4}$$

It is thus reasonable to select the values x_i for which the degree (4) is the largest possible.

It is known (see, e.g., [13]) that for every $\varepsilon > 0$, each t-norm can be approximated by an Archimedean one, i.e., by a t-norm of the type $f_\&(a,b) = g^{-1}(g(a)\cdot g(b))$ for some increasing function $g(a)$. Thus, without losing generality, we can assume that our t-norm has this form. For such t-norms, the expression (4) takes the form

$$g^{-1}(g(\mu_1(\widetilde{y} - f_1(x_1,\ldots,x_n))\cdot\ldots\cdot g(\mu_m(\widetilde{y} - f_m(x_1,\ldots,x_n))).$$

So, maximizing the expression (4) is equivalent to maximizing the product of type (1), where we denoted $\rho_j(z) \overset{\text{def}}{=} g(\mu_j(z))$, and is, hence, equivalent to minimizing the corresponding sum (2).

Need to take reliability into account. In the above text, we implicitly assumed that every measuring instrument functions absolutely reliably and thus, every number $\widetilde{y}_j$ that we get comes from the actual measurement. In practice, measuring instruments are imperfect, sometimes they malfunction, and thus, once in a while, we get a value that has nothing to do with the measured quantity – i.e., an *outlier*.

Some outliers are easy to detect and filter out: e.g., if we measure body temperature and get 0 degrees, clearly the device is not working. In many other cases, however, it is not so easy to detect outliers. Similarly, some expert estimates can be way off.

In such cases, when processing data, we need to take into account that the values $\widetilde{y}_j$ are un-reliable: some of these values may be un-related to measurements.

2 What Do We Know About Reliability: Enter Z-Numbers

What do we know about the possible outliers: analysis of the problem. Information about accuracy of measurements (or expert estimates) comes from our past experience:

- we know how frequent were different deviations between the measured and actual values, and
- we can thus estimate the probabilities of different deviations Δy_j.

Similarly, based on our past experience:

- we can determine how frequently the values produced by the measuring instrument (or by an expert) turned out to be outliers, and
- thus, estimate the probability p_j that a given value $\widetilde{y}_j$ is an outlier.

In both cases, we arrive at the following description.

What do we know about the possible outliers: probabilistic case. In the probabilistic case, for each j:

- we know the probability distribution function $\rho_j(\Delta y_j)$, and
- we know the corresponding probability p_j.

What do we know about the possible outliers: fuzzy case. In the fuzzy case, for each j:

- we know the corresponding membership function $\mu_j(\Delta y_j)$ – or, equivalently, the corresponding function $\rho_j(\Delta y_j)$ – and
- we also know the corresponding probability p_j.

General case. L. Zadeh called such a pair (ρ_j, p_j) or (μ_j, p_j) – that describes both the accuracy and the reliability – a *Z-number*; see, e.g., [1, 19].

3 Z-Numbers and Teaching

Up to know, we considered the case when Z-numbers describe measurements or expert estimates, but there is another important area where Z-numbers are useful: teaching.

Namely, usually, the success of teaching is gauged by how accurate are the students' answers. However, it is important to also take into account how confident the students are in their answers:

- if a student gives the right answer, but he or she is not confident, this means there is still room for improvement,
- on the other hand, if a student gives the wrong answer, but he or she is not sure, the situation is not so bad: it means that in a similar future real-life case, the student will probably doublecheck or consult someone else and thus, avoid making a wrong decision.

In [11], we showed how to take both accuracy and reliability into account when gauging the result of teaching.

The need to take both accuracy and confidence into account is especially important for female students, low-income students, and students from under-represented minority groups, since these students typically show decreased confidence – even when their accurate answers show that they have reached a high level of competence; see, e.g., [8].

4 How to Take Into Account Accuracy and Reliability When Processing Data: Idea and Resulting Algorithm

Problem. How can we extend the formulas from [11] – designed specifically for the teaching case – to the general data processing situation?

If we knew which values $\widetilde{y}_j$ are outliers, we could simply ignore these values and process all others. In practice, however, we do not know which measurement results are outliers, we only know the probabilities of each of them being an outlier. In principle, we could consider all possible outlier subsets – but since there are exponentially many such possible subsets, this would require an un-feasible exponential time. So what can we do?

Idea. We do not know which values $\widetilde{y}_j$ are outliers, but knowing the probability p_j means that we know that if we repeat the measurements N times, than in approximately $p_j \cdot N$ cases we will have accurate estimates – and in the remaining $N - p_j \cdot N$ cases, we will have outliers.

To utilize this information, let us consider an imaginary situation in which each value $\widetilde{y}_j$ is repeated N times.

Good news is that if all values $\widetilde{y}_j$ were absolutely reliable, and simply repeat each value $\widetilde{y}_j$ the same number of times N, the result of data processing will not change. Indeed, e.g., in the minimization formulation (2), repeating each value N times simply increases the minimized expression by a factor of N – and, of course, both the original expression (2) and the same expression multiplied by N attains their minimum on the exact same tuple x_i.

So, it makes sense to consider repetitions. But once we have many (N) repetitions, we kind of know which values are outliers – namely, we know that only $N \cdot p_1$ of copies of $\widetilde{y}_1$ are accurate estimates, etc. So, in processing data, we take into account:

- only $N \cdot p_1$ copies of the value $\widetilde{y}_1$,
- only $N \cdot p_2$ copies of the values $\widetilde{y}_2$,
- etc.

When we apply the expression (2) to these values, we end up with selecting the tuple $(x_1, \ldots, x_n)$ that minimizes the sum

$$\sum_{j=1}^{m} (N \cdot p_j) \cdot \psi_j(\widetilde{y}_j - f_j(x_1, \ldots, x_n)).$$

Strictly speaking, this expression depends on the unknown number of repetitions N, but good news is that if we divide the above expression by N, we get a new expression that no longer depends on N – but which attains its minimum at exactly the same tuple $(x_1, \ldots, x_n)$. Thus, we arrive at the following recommendation.

Resulting algorithm. When we know the reliability p_j of each value $\widetilde{y}_j$, then we should select the tuple $(x_1, \ldots, x_n)$ that minimizes the expression

$$\sum_{j=1}^{m} p_j \cdot \psi_j(\widetilde{y}_j - f_j(x_1, \ldots, x_n)). \tag{5}$$

In particular, in the case of normal distribution, applying the same idea to formula (3) leads to the need to minimize the expression

$$\sum_{j=1}^{m} p_j \cdot \frac{(\widetilde{y}_j - f_j(x_1, \ldots, x_n))^2}{\sigma_j^2} = \sum_{j=1}^{m} \frac{(\widetilde{y}_j - f_j(x_1, \ldots, x_n))^2}{(\sigma'_j)^2}, \tag{6}$$

where we denoted $\sigma'_j \stackrel{\text{def}}{=} \dfrac{\sigma_j}{\sqrt{p_j}}$.

How good is this algorithm? To check whether this algorithm is good, we will show, on the case study of dimension reduction, that the ideas behind this algorithm provide a natural explanation for an empirically successful heuristic approach.

5 Case Study: Dimension Reduction

Dimension reduction: formulation of the problem. In many practical situations, we analyze a large number of objects of a certain type. For example, in medical research, we study all the patients that suffer from a given disease.

In many such situations, we do not know which quantities will turn out to be relevant. Thus, not to miss any relevant quantity, we measure as many quantities as possible. As a result, for each object, we have a large number of measurement results and/or expert estimates. In other words, each object is represented by a point in a very high-dimensional space.

Processing such high-dimensional data is often very time-consuming. It is therefore desirable to reduce the amount of data. Good news – coming from our experience – is that in most practical situations, most of the collected data is irrelevant, that there are usually a few important combinations of the original parameters that are relevant for our specific problem. In other words, with respect to the corresponding problem, we can as well use a low-dimensional representation of the data.

To use this possibility, we need to be able to reduce the data dimension.

Reformulating the dimension reduction problem in terms of Z-numbers. We want to assign, to each point s_i in the multi-D space, a point q_i in the lower-dimensional space. The main criterion that we want to satisfy is that if s_i and s_j are close, then the corresponding points q_i and q_j should also be close.

If we had a clear (crisp) idea of which pairs (s_i, s_j) are close and which pairs are not close, we would simply require that the values q_i and q_j corresponding to these pairs are close, i.e., that

$$q_i \approx q_j$$

for all such pairs. By applying the Least Squares approach to this situation, we would then arrive at the problem of minimizing the sum $\sum \|q_i - q_j\|^2$, where the sum is taken over all such pairs. Of course, to avoid the trivial and useless solution $q_1 = q_2 = \ldots$, we need to "normalize" these solutions: e.g. by requiring that $\sum\limits_i \|q_i\|^2 = 1$.

In practice, we usually do not have an absolutely clear idea of which points are close to each other and which are not. A reasonable idea is to describe closeness in probabilistic terms. Since there can be many different reasons why objects are somewhat different, it makes sense to apply the same Central Limit theorem argument that we used before and conclude that closeness corresponds to a normal distribution.

Since we do not have a priori knowledge of which components of the original vectors s_i are more relevant and which are less relevant, it is therefore reasonable to assume that the corresponding Gaussian distribution is invariant with respect to all permutations of these components (and changing their signs), and thus, that the normal distribution has the form $\text{const} \cdot \exp\left(-\frac{\|s_i - s_j\|^2}{2\sigma^2}\right)$ for some $\sigma > 0$. Thus, we arrive at the following Z-number-type problem:

$$q_i \approx q_j \text{ with probability } p_{ij} = \text{const} \cdot \exp\left(-\frac{\|s_i - s_j\|^2}{2\sigma^2}\right). \tag{7}$$

Applying our algorithm to the resulting Z-number problem leads to a known successful heuristic. If we apply the above algorithm to this problem, we arrive at the need to minimize the expression

$$\sum_{i,j} p_{ij} \cdot \|q_i - q_j\|^2, \tag{8}$$

where the values p_{ij} are defined by the formula (7). (Of course, some normalization like $\sum_i \|q_i\|^2 = 1$ is needed.) This is equivalent to minimizing the sum

$$\sum_{i,j} w_{ij} \cdot \|q_i - q_j\|^2, \tag{8a}$$

where we denoted

$$w_{ij} = \exp\left(-\frac{\|s_i - s_j\|^2}{2\sigma^2}\right). \tag{7a}$$

This is indeed one of the most successful heuristic methods for dimension reduction – it is known as the *Laplacian eigenmap*, since its solution can be described in terms of eigenvectors of the corresponding Laplacian operator $\nabla^2 \varphi = \sum_{i=1}^{d} \frac{\partial^2 \varphi}{\partial^2 x_i}$; see, e.g., [2, 3, 4, 5, 9, 15, 16].

So, *Z-numbers provide a theoretical explanation for the empirical success of Laplace eigenmaps – a heuristic approach to dimension reduction.*

Taking into account that some objects may be not relevant as well. In the above analysis, we assumed that for each object, we are 100% sure that this object belongs to the desired class. In practice, we are often not fully confident about this. For example, when we study a certain disease, we are not always sure that a patient suffers from this very disease – and not from some similar one.

In general, the further away the object from the "typical" (average) situation – which, by shifting, we can always assume to be 0 – the less probable it is that this object actually belongs to the desired class. In making this conclusion, we should not take into account irrelevant components of the points s_i. Thus, this conclusion should be based only on the values q_i - which contain only relevant combinations.

Similar to the above argument, we can safely assume that the corresponding distribution is Gaussian, with probability P_i proportional to $\exp\left(-\frac{\|q_i\|^2}{2\sigma_i^2}\right)$. Here, different values σ_i correspond to different degrees of confidence that this object belongs to the class:

- when $\sigma_i = 0$, this means that the probability does not depend on q_i at all: in other words, we are so confident, that no matter how big the deviation from the typical object, our degree of confidence does not change;

- on the other hand, if σ_i is large, then even a small deviation from the typical value will make us conclude that this object does not belong to the desired class.

In this case, to get a more adequate description of the situation, to the product (1), we need to add the factors corresponding to different objects. After taking negative logarithm, these terms are equivalent to adding terms proportional to $V_i \cdot \|q_i\|^2$ to the sum (2), where we denoted $V_i \overset{\text{def}}{=} \sigma_i^{-2}$.

In particular, for the dimension reduction problem, this means that instead of minimizing the expression (8a), we minimize a more complex expression

$$\sum_{i,j} w_{ij} \cdot \|q_i - q_j\|^2 + \alpha \cdot \sum_i V_i \cdot \|q_i\|^2. \tag{9}$$

This expression has indeed been proposed and successfully applied – on a heuristic basis – in [7]. This approach is known as *Schroedinger eigenmap*, since it corresponds to using eigenvectors of the operator $\nabla^2 \varphi + \text{const} \cdot V \cdot \varphi$ from Schroedinger's equations in quantum physics.

Thus, *Z-numbers provide a theoretical explanation for the empirical success of this a heuristic approach as well.*

Can we go beyond justification of existing approaches? A theoretical justification of known heuristic approaches is nice, but can we learn something new from this approach? Yes, we can.

While, as we have shown, the Schroedinger approach is well-justified for the case when we are not sure whether objects belongs to the class, this approach is also used in a completely different situation: when:

- we have an additional discrete value V_i characterizing each object, and
- we want to require $q_i \approx q_j$ only for objects that have close values of V_i and V_j.

For this situation, the Schroedinger approach is not perfect: indeed, even in the simplest case when V_i takes two possible values – which we can describe as 0 and 1 – the result of minimizing the expression (9) depends on which of the two possible value we associate with 0 and which with 1.

In view of our analysis, it is more adequate to add the similarity between the value V_i and V_j to the description of closeness, i.e., to use an expression

$$w_{ij} = \exp\left(-\frac{\|s_i - s_j\|^2}{2\sigma^2} - \frac{(V_i - V_j)^2}{2\sigma_0^2}\right), \tag{10}$$

for some $\sigma_0 > 0$. The resulting probabilities does not change if we swap 0 and 1 value of V_i – thus, the resulting minimized expression (8) will not change after this swap, and hence, the produced optimizing arrangement q_i will not change – which is exactly what we wanted.

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

References

1. R. A. Alief, O. H. Huseynov, R. R. Aliyev, and A. A. Alizadeh, *The Arithmetic of Z-Numbers: Theory and Applications*, World Scientific, Singapore, 2015.
2. M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation", *Neural Computation*, 2003, Vol. 15, pp 1373–1396.
3. M. Belkin and P. Niyogi, "Convergence of Laplacian eigenmaps", In: D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds) *Advances in Neural Information Processing Systems 21 (NIPS'2008)*, Springer Verlag, Berlin, Heidelberg, New York, 2008.
4. M. Belkin and P. Niyogi, "Towards a theoretical foundation for Laplacian-based manifold methods", *Journal of Computer and System Sciences*, 2008. Vol. 74, No. 8, pp. 1289–1308.
5. M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: a geometric framework for learning from labeled and unlabeled examples", *Journal of Machine Learning Research*, 2006, Vol. 7, pp. 2399–2434.
6. R. Belohlavek, J. W. Dauben, and G. J. Klir, *Fuzzy Logic and Mathematics: A Historical Perspective*, Oxford University Press, New York, 2017.
7. W. Czaja and M. Ehler, "Svchoedinger eigenmaps for the analysis of bio-medical data", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, Vol. 35, No. 5, pp. 1274–1280.
8. C. L. Frisby, *Meeting the Psychoeducational Needs of Minority Students: Evidence-Based Guidelines for School Psychologists and Other School Personnel*, Wiley, Hoboken, New Jersey, 2013.
9. A. J. Izenman, "Spectral embedding methods for manifold learning", In: Y. Ma and Y. Fu, *Maniforld Learning: Theory and Applications*, CRC Press, Boca Raton, Florida, 2012, pp. 1–36.
10. G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, New Jersey, 1995.
11. O. Kosheleva, J. Lorkowski, V. Felix, and V. Kreinovich, "How to take into account student's degree of confidence when grading exams", *Proceedings of the 5th International Conference "Mathematics Education: Theory and Practice" MATHEDU'2015*, Kazan, Russia, November 27–28, 2015, pp. 29–30.
12. J. M. Mendel, *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions*, Springer, Cham, Switzerland, 2017.
13. H. T. Nguyen, V. Kreinovich, and P. Wojciechowski, "Strict Archimedean t-norms and t-conorms are universal approximators, *International Journal of Approximate Reasoning*, 1998, Vol. 18, Nos. 3–4, pp. 239-249.
14. H. T. Nguyen and E. A. Walker, *A First Course in Fuzzy Logic*, Chapman and Hall/CRC, Boca Raton, Florida, 2006.
15. L. K. Saul, K. Q. Weinberger, F. Sha, J. Ham. and D. D. Lee, "Spectral methods for dimensionality reduction", Chapter 16 in: O. Chapelle, B. Scholkopf, and A. Zien (eds.), *Semi-Supervised Learning*, MIT Press, 2013.
16. B. Shaw, *Graph Embedding and Nonlinear Dimensionality Reduction*, PhD Dissertation, Columbia University, 2011.
17. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.
18. L. A. Zadeh, "Fuzzy sets", *Information and Control*, 1965, Vol. 8, pp. 338–353.
19. L. A. Zadeh, "A Note on Z-Numbers", *Information Sciences*, 2011, Vol. 181, pp. 2923–2932.