7-1-2024

For 2 x n Cases, Proportional Fitting Problem Reduces to a Single Equation

Olga Kosheleva
The University of Texas at El Paso, olgak@utep.edu

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Comments:
Technical Report: UTEP-CS-24-36

For $2 \times n$ Cases, Proportional Fitting Problem Reduces to a Single Equation

Olga Kosheleva and Vladik Kreinovich

Abstract In many practical situations, for each of two classifications, we know the probabilities that a randomly selected object belong to different categories. For example, we know what proportion of people are below 20 years old, what proportion is between 20 and 30, etc., and we also know what proportion of people earns less than 10K, between 10K and 20K, etc. In such situations, we are often interested in proportion of people who are classified by two classifications into two given categories. For example, we are interested in the proportion of people whose age is between 20 and 30 and whose income is between 10K and 20K. If we do not have detailed records of all the objects, we select a small sample and count how many objects from this sample belong to each pair of categories. The resulting proportions are a good first-approximation estimate for the desired proportion. However, for a random sample proportions of each category are, in general, somewhat different from the proportions in the overall population. Thus, the first-approximation estimates need to be adjusted, so that they fit with the overall-population values. The problem of finding proper adjustments is known as the proportional fitting problem. There exist many efficient iterative algorithms for solving this problem, but it is still desirable to find classes for which even faster algorithms are possible. In this paper, we show that for the case when one of the classifications has only two categories, the proportional fitting problem can be reduced to solving a polynomial equation of order equal to number n of categories of the second classification. So, for $n = 2,3,4$, explicit formulas for solving quadratic, cubic, and quartic equations lead to explicit solutions for the proportional fitness problem. For $n > 4$, fast algorithms for solving polynomial equations lead to fast algorithms for solving the proportional fitness problem.

Olga Kosheleva
Department of Teacher Education, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: olgak@utep.edu

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: vladik@utep.edu

1 Formulation of the Problem

Practical problem. In general, objects can be classified differently. For example:

- people can be classified into several categories based on their age, and
- people can also be classified into different categories based on their income.

For each of these classifications, we usually have a good understanding of how many objects belong to each of its categories:

- let us denote the estimated proportion of objects that are classified into category i in the first classification by f_i, and
- let us denote the estimated proportion of objects that are classified into category j in the second classification by s_j.

We are often also interested in the relation between the two categories. For example, we may be interested in the proportion of people from the given age category has income from the given category. In general, for each category i from the first classification and from each category j from the second classification, we want to know the proportion p_{ij} of objects that are classified into these two categories.

In the ideal situation, when we have a lot of information about each object – e.g., after a census – we can simply count the number of objects n_{ij} that are classified into the two given categories. However, in many practical situations, we do not have that information about each individual object. In such cases, a natural idea is:

- to have a poll, i.e., to randomly pick a small – but statistically significant – number of objects (usually about 1000), and
- for all the objects from the selected group, find their category in each classification and thus, count the proportion $\widetilde{p}_{ij}$ of objects – within this sample – that belong to both categories.

The problem is that since the sample is randomly selected, in this sample, the proportion of people who are classified, e.g., into category i, is somewhat different from the overall proportion e_i:

$$\sum_j \widetilde{p}_{ij} \neq f_i \tag{1}$$

and, similarly,

$$\sum_i \widetilde{p}_{ij} \neq s_j. \tag{2}$$

It is desirable to adjust the estimates $\widetilde{p}_{ij}$ into more accurate estimates p_{ij}, estimates that are consistent with the known proportions f_i and s_j.

Proportional Fitting problem. The problem is caused by the fact that for each classification, the proportion of each category in the sample is somewhat different from the proportion in the whole population. So, a natural idea is to adjust for this difference, by multiplying by an appropriate coefficient. For example, in the sample, the proportion classified into category i is 1.2 times larger than in the population as a whole, we should divide all corresponding numbers by 1.2 – i.e., equivalently, multiply them by $1/1.2$.

A similar correction needs to be made for the second classification. So, there should be some coefficients a_i and b_j so that instead of the original values $\widetilde{p}_{ij}$, we should consider the adjusted values $a_i \cdot b_j \cdot \widetilde{p}_{ij}$. The coefficients a_i and b_j should be selected in such a way that the corrected proportions should be in perfect agreement with the known values f_i and s_j, i.e., that we should have

$$\sum_j a_i \cdot b_j \cdot \widetilde{p}_{ij} = f_i \text{ for all } i, \text{ and} \tag{3}$$

$$\sum_i a_i \cdot b_j \cdot \widetilde{p}_{ij} = s_j \text{ for all } j, \tag{4}$$

The problem of finding the values a_i and b_j based on the known values $\overline{p}_{ij}$, s_i, and f_j is known as the *Proportional Fitting problem*; see, e.g., [1, 2].

How we can solve this problem: what is known. There are several efficient iterative algorithms for solving this problem [1, 2].

Remaining problem. While iterative methods are reasonably fast, it is desirable to find class of problems for which even faster methods are possible.

What we do in this paper. In this paper, we show that for the $2 \times n$ case, when one of the classifications has only two categories, the proportional fitting problem can be reduced to a single n-th order polynomial equation. This means that:

- for $n = 2$, $n = 3$, and $n = 4$, we can use known explicit formulas for solving the corresponding equations to come up with explicit formulas for the proprtional fitting problem; and
- for $n > 4$, we can use known fast algorithms for solving polynomial equations to come up with fast algorithms for solving the proportional fitting problem

2 Reduction

Let us derive the desired reduction. In the $2 \times n$ case, we need to find the values $a_1, a_2, b_1, \ldots, b_n$.

First, let us notice that the solution is not unique: if we divide all the values a_i by some positive value λ and multiply all the values b_j by the same number λ, then the products $a_i \cdot b_j$ remain the same for all i and j. Indeed, if we take

$$a'_i = \frac{a_i}{\lambda}$$

and $b'_j = \lambda \cdot b_j$, then we have

$$a'_i \cdot b'_j = \frac{a_i}{\lambda} \cdot \lambda \cdot b_j = a_i \cdot b_j.$$

We can use this non-uniqueness to simplify the problem. Namely, let us perform this division-and-multiplication for $\lambda = a_1$. Then, the new value a'_1 of a_1 will be equal to 1. So, without losing generality, we can assume that $a_1 = 1$.

In this case, each equation (4) takes the form

$$b_j \cdot \widetilde{p}_{1j} + b_j \cdot a_2 \cdot \widetilde{p}_{2j} = s_j,$$

i.e., equivalently,

$$b_j \cdot (\widetilde{p}_{1j} + a_2 \cdot \widetilde{p}_{2j}) = s_j.$$

If we divide both sides of this equality by the coefficient at b_j, we conclude that

$$b_j = \frac{s_j}{\widetilde{p}_{1j} + a_2 \cdot \widetilde{p}_{2j}}. \tag{5}$$

The equation (3) for $i = 1$ takes the form

$$b_1 \cdot \widetilde{p}_{11} + \ldots + b_n \cdot \widetilde{p}_{1n} = f_1. \tag{6}$$

Substituting the expressions (5) for b_j into the formila (6), we conclude that

$$\frac{s_1 \cdot \widetilde{p}_{11}}{\widetilde{p}_{11} + a_2 \cdot \widetilde{p}_{21}} + \ldots + \frac{s_n \cdot \widetilde{p}_{1n}}{\widetilde{p}_{1n} + a_2 \cdot \widetilde{p}_{2n}} = f_1. \tag{7}$$

If we multiply both sides by the product of all n denominators, we get the following polynomial equation of order n:

$$s_1 \cdot \widetilde{p}_{11} \cdot \prod_{j \ne 1} (\widetilde{p}_{1j} + a_2 \cdot \widetilde{p}_{2j}) + \ldots + s_n \cdot \widetilde{p}_{1n} \cdot \prod_{j \ne n} (\widetilde{p}_{1j} + a_2 \cdot \widetilde{p}_{2j}) =$$

$$f_1 \cdot \prod_{j=1}^{n} (\widetilde{p}_{1j} + a_2 \cdot \widetilde{p}_{2j}). \tag{8}$$

Once we solve this equation and find the value a_2, we can then find the values b_j by using the formula (5). So, we arrive at the following algorithm.

Resulting algorithm. Suppose that we are given the values $\widetilde{p}_{ij}$, f_i, and s_j. Then, we take $a_1 = 1$, and as a_2, we take the solution of the following polynomial equation of n-th order:

$$s_1 \cdot \widetilde{p}_{11} \cdot \prod_{j \ne 1} (\widetilde{p}_{1j} + a_2 \cdot \widetilde{p}_{2j}) + \ldots + s_n \cdot \widetilde{p}_{1n} \cdot \prod_{j \ne n} (\widetilde{p}_{1j} + a_2 \cdot \widetilde{p}_{2j}) =$$

$$f_1 \cdot \prod_{j=1}^{n} (\widetilde{p}_{1j} + a_2 \cdot \widetilde{p}_{2j}). \tag{8}$$

Once we compute a_2, we can then compute all the values b_j by using the following formula:

$$b_j = \frac{s_j}{\widetilde{p}_{1j} + a_2 \cdot \widetilde{p}_{2j}}. \tag{5}$$

Comment. We copied the formulas into the algorithm-describing subsection, to make it easier for readers who are only interested in the resulting algorithm – and not in its derivation.

Examples. For $n = 2$, the formula (8) leads to the following quadratic equation:

$$s_1 \cdot \widetilde{p}_{11} \cdot (\widetilde{p}_{12} + a_2 \cdot \widetilde{p}_{22}) + s_1 \cdot \widetilde{p}_{11} \cdot (\widetilde{p}_{11} + a_2 \cdot \widetilde{p}_{21}) =$$

$$(\widetilde{p}_{11} + a_2 \cdot \widetilde{p}_{21}) \cdot (\widetilde{p}_{12} + a_2 \cdot \widetilde{p}_{22}). \tag{9}$$

For $n = 3$, we get the following cubic equation:

$$s_1 \cdot \widetilde{p}_{11} \cdot (\widetilde{p}_{12} + a_2 \cdot \widetilde{p}_{22}) \cdot (\widetilde{p}_{13} + a_2 \cdot \widetilde{p}_{23}) +$$

$$s_2 \cdot \widetilde{p}_{12} \cdot (\widetilde{p}_{11} + a_2 \cdot \widetilde{p}_{21}) \cdot (\widetilde{p}_{13} + a_2 \cdot \widetilde{p}_{23}) +$$

$$s_3 \cdot \widetilde{p}_{13} \cdot (\widetilde{p}_{11} + a_2 \cdot \widetilde{p}_{21}) \cdot (\widetilde{p}_{12} + a_2 \cdot \widetilde{p}_{22}) =$$

$$(\widetilde{p}_{11} + a_2 \cdot \widetilde{p}_{21}) \cdot (\widetilde{p}_{12} + a_2 \cdot \widetilde{p}_{22}) \cdot (\widetilde{p}_{13} + a_2 \cdot \widetilde{p}_{23}). \tag{10}$$

For $n = 4$, we get the following quartic equation:

$$s_1 \cdot \widetilde{p}_{11} \cdot (\widetilde{p}_{12} + a_2 \cdot \widetilde{p}_{22}) \cdot (\widetilde{p}_{13} + a_2 \cdot \widetilde{p}_{23}) \cdot (\widetilde{p}_{14} + a_2 \cdot \widetilde{p}_{24}) +$$

$$s_2 \cdot \widetilde{p}_{12} \cdot (\widetilde{p}_{11} + a_2 \cdot \widetilde{p}_{21}) \cdot (\widetilde{p}_{13} + a_2 \cdot \widetilde{p}_{23}) \cdot (\widetilde{p}_{14} + a_2 \cdot \widetilde{p}_{24}) +$$

$$s_3 \cdot \widetilde{p}_{13} \cdot (\widetilde{p}_{11} + a_2 \cdot \widetilde{p}_{21}) \cdot (\widetilde{p}_{12} + a_2 \cdot \widetilde{p}_{22}) \cdot (\widetilde{p}_{14} + a_2 \cdot \widetilde{p}_{24}) +$$

$$s_4 \cdot \widetilde{p}_{14} \cdot (\widetilde{p}_{11} + a_2 \cdot \widetilde{p}_{21}) \cdot (\widetilde{p}_{12} + a_2 \cdot \widetilde{p}_{22}) \cdot (\widetilde{p}_{13} + a_2 \cdot \widetilde{p}_{23}) =$$

$$(\widetilde{p}_{11} + a_2 \cdot \widetilde{p}_{21}) \cdot (\widetilde{p}_{12} + a_2 \cdot \widetilde{p}_{22}) \cdot (\widetilde{p}_{13} + a_2 \cdot \widetilde{p}_{23}) \cdot (\widetilde{p}_{14} + a_2 \cdot \widetilde{p}_{24}). \tag{11}$$

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), HRD-1834620 and HRD-2034030 (CAHSI Includes), EAR-2225395 (Center for Collective Impact in Earthquake Science C-CIES), and by the AT&T Fellowship in Information Technology. It was also supported by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

The authors are greatly thankful to all the participants of the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) Workshop on Synthetic Data and Generative AI (El Paso, Texas, June 17, 2024), especially to Amy Wagler, for valueable discus sions.

References

1. Y. M. M. Bishop, S. E. Feinberg, and P. W. Holland, *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, Massachusetts, USA, 1975.
2. M. Idel, *A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps*, arXiv:1609.06349v1, 2016, https://arxiv.org/pdf/1609.06349.pdf