

Two algorithms for Error Localization Problems

Scarnò Marco¹, Caramanna Laura²

Abstract A common method for solving error localization problems with linear constraints for numerical data is based on vertex generation technique, derived from the Chernikova's algorithm. This was initially designed for calculating the edges of a convex polyhedral cone; when the algorithm is applied to solve the error localization problem turns out to be rather slow. Many authors proposed various heuristics in order to accelerate the identification of the solutions. In this article we propose two algorithms that can be used to solve localization problems; the first one is a modification of the original, the second considers a new approach. This last could be more efficient for few variables.

1 Introduction

Errors in data can be detected by specifying several constraints that have to be satisfied by the values observed for each individual/respondent. These constraints are called edit rules (or edits for short); for continuous variables they are represented by a sets of linear equalities or inequalities.

When a record fails an edit, it is considered erroneous, otherwise it is considered corrected. The main issue of the error localization problem is that, if a record is erroneous, this not means that all its values are to be considered not valid; the real question is to find the minimum number of fields that could be changed in order to let that the record satisfies all the edits at the same time.

This problem is of great interest to statistical agencies and it is concerned to data editing problem, minimum (weighted) fields to impute, or Error Localization Problem (ELP). There are many possibilities to change an erroneous record.

The first description of a paradigm for automatically identifying errors in a record was made by Fellegi and Holt [5]. According to this paradigm the data of an erroneous

¹ Inter-university Consortium CASPUR, Rome, email: marco.scarno@caspur.it

² Inter-university Consortium CASPUR, Rome, email: laura.caramanna@caspur.it

record should be modified to satisfy all edits by considering the fewest possible changes in its fields. In due course the original Fellegi-Holt paradigm has been generalized to: the data of a record should be made to satisfy all edits by changing the values of the variables with the smallest possible sum of reliability weights.

A reliability weight of a variable is a non-negative number that expresses how reliable it is considered the value of this variable; a high reliability weight corresponds to a variable of which the values are considered trustworthy, a low reliability weight to a variable of which the values are considered not so trustworthy.

In order to solve ELPs there are, indeed, many tools; unfortunately their algorithms are not opened or replicable because they usually contain many heuristics or other not well described modifications. Our purpose is to add new procedures inside our Open Source statistical software, called ADaMSoft that can manage such kind of problems; in doing so we also tried to develop new approaches.

Obviously a paradigm for identifying the erroneous fields in a not corrected record is only a first step towards ELPs. The second step consists of actually solving the mathematical problem.

2 General problem formulation

It can be considered to have observed m quantitative variables on a sample of n individuals. We will address the variables with $\mathbf{X}_1, \dots, \mathbf{X}_m$ (they could be considered as m column vectors). It is possible to use the symbol $\mathbf{X}_j$ in order to represent a generic variable; its value, for the i -th individual, can be expressed with x_{ij} . It is, also, possible to introduce the symbol $\mathbf{x}_i$ to refers to the values that the i -individual assumes for each of the m variables.

An edit rule represent the constrain that a record must satisfy; it could be represented by an equality or an inequality that involves, linearly, the variables. For example:

$$b_1^h x_1 + \dots + b_m^h x_m \geq b_o^h \text{ (if the edit is an inequality), or}$$

$$b_1^h x_1 + \dots + b_m^h x_m = b_o^h \text{ (if the edit is an equality).}$$

In these expressions the index h refers to the h -th edit (we can assume to have H edits), while $b_1^h, \dots, b_m^h$ are the coefficients of the linear combination of the variables and b_o^h is a constant. All the b values can be real numbers (they can, also, assume a 0 value).

It is possible to indicate the coefficients of the edits and those of the constant terms by using, respectively, the matrix $\mathbf{B}$ and the column vector $\mathbf{K}$; so we will have:

$$\mathbf{B} = \begin{pmatrix} b_1^1 & \dots & b_m^1 \\ \dots & b_j^h & \dots \\ b_1^H & \dots & b_m^H \end{pmatrix} \text{ and } \mathbf{K} = \begin{pmatrix} b_0^1 \\ \dots \\ b_0^H \end{pmatrix}.$$

According to these notations, the product:

$$\mathbf{B}\mathbf{x}_i \leq \mathbf{K} \quad (1)$$

represents the system of inequalities (and equalities) that should be satisfied by the i -th individual.

It has, also, to be considered that in the ELPs all the values of the variables for each record must be greater or equal to 0; this could be represented by:

$$\mathbf{x}_i \geq \mathbf{0} \quad (2)$$

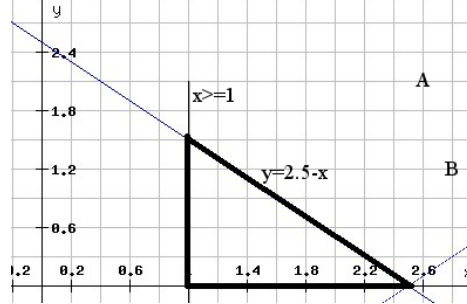
For what concerns the reliability weights, these are associated to the variables in such a way that it is possible to consider, respectively, $w_1, \dots, w_m$.

3 Localization in graphical terms

It is possible to suppose to have two variables (**X** and **Y**) observed on a series of individuals; each of these should satisfies the following edits: $x \geq 1$; $y \leq 2.5 - x$; $y \geq 0$.

Because the two dimensions of the problem, the edits can be graphed in an xy plane as reported in Fig. 1. It is possible to see that all the values that lie inside the triangular area, called the *feasible region*, satisfy simultaneously all the edits.

Figure 1: an example of the *feasible region*



The central problem of error localization is to determine the minimum number of changes that should be made to a failed record so that it can pass all the edits. In case of the point A, with coordinates [2.6: 2], it is possible to verify that it is out of the region for both the values of the two variables; it will require a correction to all of these in order to be moved inside the feasible region.

Instead, the point B, having the coordinates [2.6:1.2], will require only the modification of the variable x in order to be moved inside the region.

Note that, before to apply the edits to a real data set, it should be checked if the edits define a feasible region.

4 Mathematical solution of the ELP

The mathematical problem can be solved in several ways, see De Waal [3]. One of these ways is by identifying the vertices of a certain polyhedron.

There are several algorithms that efficiently generate a suitable subset of vertices of a polyhedron. An example of such a vertex generation algorithm is the one proposed by Chernikova [1, 2]. Probably most computer systems for automatic edit and imputation of numerical data are based on an adapted versions of this algorithm. Unfortunately the original algorithm of Chernikova is rather slow for solving the error localization problem. It has been accelerated by various modifications (see Rubin [7, 8]; Sande [9]; Schiopu-Kratina and Kovar [10]; Fillion and Schiopu-Kratina [6]).

Only the last three of these papers focus on the error localization problem. Sande discusses the error localization problems for numerical data, categorical data and mixed data.

It should, also, be noticed that Chernikova's algorithm and its modification made by Duffin [4] have been designed only to handle inequalities; but in common situations many edits involve equalities rather than inequalities. Of course, it is possible to represent each equality by two inequalities, and then use one of the algorithms proposed by the two authors to solve the error localisation problem. However this would result in an inefficient approach.

The standard method to solve the ELP based on the Chernikova approach introduces a new system of inequalities that has to be solved; in this are specified the corrections (positive and/or negative) for each variable of the erroneous records; using the expressions (1) and (2) this could be represented by:

$$\mathbf{B}(\hat{\mathbf{x}}_i + \mathbf{x}_i^+ - \mathbf{x}_i^-) \leq \mathbf{K} \quad (1a)$$

$$\hat{\mathbf{x}}_i + \mathbf{x}_i^+ - \mathbf{x}_i^- \geq \mathbf{0} \quad (2a)$$

$$\mathbf{x}_i^+, \mathbf{x}_i^- \geq \mathbf{0} \quad (3)$$

In this case $\hat{\mathbf{x}}_i$ represents the current observed values for the i -th record, while $\mathbf{x}_i^+$ and $\mathbf{x}_i^-$, respectively, are the positive and the negative corrections to apply to the fields of the record in order to let it satisfies the original edits. It should be noticed that the inequalities generated by (2a) represent the conditions that the final fields of the corrected record have to be greater or equal to 0.

In order to determine the minimum set of variables to change it is possible to consider the number of non-zero elements of the correction $\mathbf{x}_i^+ - \mathbf{x}_i^-$.

For this purpose it is defined a "cardinality function" for the i -th individual:

$$C_{\mathbf{x}_i^+ - \mathbf{x}_i^-} = \sum_{j=1}^m c_{ij},$$

where

$$c_{ij} = \begin{cases} 1, & \text{if } x_{ij}^+ \text{ or } x_{ij}^- > 0 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Obviously when reliability weights are defined, the expression (3) will consider:

$$c_{ij} = \begin{cases} w_{ij}, & \text{if } x_{ij}^+ \text{ or } x_{ij}^- > 0 \\ 0 & \text{otherwise} \end{cases}$$

Having the enumeration of all the possible vertices of the system, it will be so possible to identify the one (or the ones) to which the cardinality function assumes the smallest value. The non-zero coordinates of this vertex (of these vertices) will identify the variables that are assumed to be erroneous.

As many authors observed, this method is really “intensive”; it doubles the number of the variables and it introduces $2 \times \text{number of variables}$ new inequalities (and it requires that the equalities are transformed in two inequalities).

5 ELP in real cases

In real cases it happens that all the edits don’t involve all the variables; more specifically each edit tests only a limited set of the variables.

This situation can be associated to the possibility that there is a hierarchical “order” in the variables to consider in ELPs when one edit is not satisfied; obviously the corrections will involve at least one of the variables that are specified in the edit; it could happen that the required correction/s will not more satisfy other edits in which the same variable/s is/are present; it could happen that, in order to find valid solutions to these other edits, more variables will be founded not valid and their values will need to be changed. But, anyway, it could also happen that the space of the solutions is of a lower dimension of the original one; this just because the hierarchy of the variables that has to be checked.

Obviously an “extreme” situation happens when the edits can be divided in two or more disjoint sets (i.e. when the variables related to each set are different).

Furthermore it should be considered that more efficiency in the algorithms can be obtained by mean of the reliability weights; these could, in fact, be used to help the algorithm in considering in certain ways the variables associated to their smallest values.

6 How to solve ELP in practice

In order to solve error localization problems there are, indeed, many tools; most of them contain algorithms to solve a system of inequalities adapted for the specific problems (this means that they don’t find all the solutions but only those related to the

minimum cardinality); usually in these are defined some heuristics that can speed up the process.

Instead, in our approach, we will refer to the use of a general purpose library that can enumerate all the vertices and the extreme rays of a polyhedral cone. In particular we chose the *polco* library, distributed under the Simplified BSD License and written in Java. This library receives in input a system of linear equalities and of inequalities and produces, as output, a matrix with extreme rays as column vectors. For more details on how the algorithm works see Terzer [11].

Our purpose is to add new procedures inside our Open Source statistical software, called ADaMSoft, that can solve ELPs. It has to be noticed that the library is really efficient in managing equalities and in *not finding solutions* (i.e. the library takes less than one second to identify a system that has not solutions, due to incoherencies in its equalities and/or inequalities).

7 A simple transformation of the standard algorithm

According to the notes in the 5th paragraph, a simple transformation to the original algorithm can be made by considering the system of inequalities (1a), (2a) and (3) (referring, obviously, to the i -th record that has erroneous values).

We can consider an iterative process in which, at each step, a certain number of corrections for the given variables are equal to 0 (i.e. there will be $1, \dots, k$ variables for which we will have: $x_1^+, x_1^-, \dots, x_k^+, x_k^-, = 0$).

The process will be stopped when some solutions are found; inside these it will be possible to look for the one/s that satisfy/ies the condition of minimum cardinality (also by considering the weights).

At the first step we consider different from 0 only the corrections for the variables directly involved in the not satisfied edits (or those with a weight less than 1).

In the next step (if no solutions were found) to these we will add all the other variables that are involved in those edits in which they appear. As stated before, the process will continue till valid solutions will be found.

Obviously the algorithm will always converge; it will be much faster if the edits will contain disjoint variables.

8 An alternative approach to the standard algorithm

As observed, the standard algorithm doubles the number of variables in the system (and, also, adds new inequalities).

By doing this it defines a new $\Re^{2m}$ space in which the erroneous record will represent a “point” with all the coordinates equal to 0 (i.e. the record is the “centre” of the space). All the inequalities (defined by the possible corrections) will define a polyhedron, originated from that point, in which the vertices represent all the possible solutions.

Another way to solve the ELPs could be to consider only a generic correction (that can be positive or negative) for each variable; this can result in this system:

$$\mathbf{B}(\hat{\mathbf{x}}_i + \mathbf{x}_i^{+/-}) \leq \mathbf{K} \quad (1b)$$

$$\hat{\mathbf{x}}_i + \mathbf{x}_i^{+/-} \geq \mathbf{0} \quad (2b)$$

In this case with $\mathbf{x}_i^{+/-}$ it is specified such generic correction.

Unfortunately the solutions of such system (i.e. the vertices that will be found), are not able to directly specify valid corrections that can be applied to the original values (i.e. they are not able to move the record inside the feasible region).

In fact such system represent only a “translation” of the observed individual (i.e. of its point in the $\mathfrak{R}^m$ space) in such way that it becomes just the new centre of such space. So the solutions are the original vertices of the initial polyhedron, translated into the m -dimensional space characterized by the new centre.

In order to use such system to solve ELPs it is required a method that can evaluate the intersections of the polyhedron with one or more of the m axes; the coordinates of such intersections are the required solutions .

We solved such problem with an iterative approach; in particular we try to solve the system by considering that, in each step, some $\mathbf{x}_i^{+/-}$ could be equal to 0 (this means that we suppose the polyhedron intersects the corresponding axes). To limit the search process we introduced, also, a parameter that specifies the maximum cardinality value that the researcher wants to test; we will indicate this with p .

In the first step we consider all the variables involved in the not satisfied edits (to which we add all those to which correspond a weight less than 1); suppose these are the first k ; so we try to solve the system (1b), (2b) by considering all the possible *tuples* of the k variables taken from 1 to p (if p is greater than k we set $p=k$). So we will test:

$$\binom{k}{1}, \dots, \binom{k}{p}$$

combinations.

If no valid solutions are found, the k variables will be expanded by introducing those that are specified in the edits in which there are also the previous ones.

Also this algorithm will converge; it will be as faster less variables are in the system.

Acknowledgments

We would like to thanks Antonia Manzari, Giulio Barcaroli, Marco Di Zio and Ugo Guarnera (ISTAT) for their help in testing our methods and for the original idea of inserting the ELPs in our software.

We thanks, also, Marco Terzer (Institut of Computational Science, Zurich) for his help in using the *polco* library and for the modifications he made in order to let us use it inside our package.

References

1. Chernikova, N.V. Algorithm for finding a general formula for the non-negative solutions of a system of linear equations. *USSR Computational Mathematics and Mathematical Physics* (1964), 4, 151-158
2. Chernikova, N.V. Algorithm for finding a general formula for the non-negative solutions of a system of linear inequalities. *USSR Computational Mathematics and Mathematical Physics* (1965), 5, 228-233.
3. De Waal, T. Processing of erroneous and unsafe data. *Ph.D. Thesis, Erasmus University Rotterdam* (2003)
4. Duffin, R.J. On Fourier's analysis of linear inequality systems. *Mathematical Programming Study*. North-Holland Publishing Company (1974), 1, 71-97
5. Fellegi, I.P., and Holt, D. A systematic approach to automatic edit and imputation, *Journal of the American Statistical Association* (1976), 71, 17-35
6. Fillion, J.M., and Schiopu-Kratina, I. On the Use of Chernikova's Algorithm for Error Localization. *Report, Statistics Canada* (1993)
7. Rubin, D.S. Vertex generation and cardinality constrained linear programs. *Operations Research* (1975), 23, 555-565
8. Rubin, D.S. Vertex generation methods for problems with logical constraints. *Annals of Discrete Mathematics* (1977), 1, 457-466
9. Sande, G. An algorithm for the fields to impute problems of numerical and coded data. *Report* (1978), Statistics Canada
10. Schiopu-Kratina, I., and Kovar, J.G. Use of Chernikova's algorithm in the generalized edit and imputation system. *Report* (1989) Statistics Canada
11. Terzer, M. and Stelling, J., Large scale computation of elementary flux modes with bit pattern trees, *Bioinformatics* (2008) July 28