International Frontier Science Letters *Online: 2015-12-29*
ISSN: 2349-4484, Vol. 6, pp 1-5
doi:10.18052/www.scipress.com/IFSL.6.1

Natural Transformations in Statistics.

Gennadii V. Kondratiev[1,a]

[1]24 Minina St., State Technical University of Nizhniy Novgorod n.a. R.E. Alexeev, Russia

[a]gennadii.kondratiev@gmail.com

Keywords: Data, internal property, uniform regularity, category theory, natural transformations, invariants, natural extensions.

Abstract. The old idea of internal uniform regularity of empirical data is discussed within the framework of category theory. A new concept and technique of statistical analysis is being introduced. It is independent on and fully compatible with the classical probabilistic approach. The absence of the model in the natural approach to statistics eliminates the model error and allows to use it in all areas with poor models. The existing error is fully determined by incompleteness of the data. It is always uniformly small by the construction of the data extension.

Introduction

In the past, when statistics was rather an empirical part of physics and was not thought as an applied probability area, the researchers spoke of a certain *internal regularity* of data. Recognition of this regularity gave them a pattern to predict behavior of the total population represented by the data. However, that time there was no adequate machinery capable to express the property of internal regularity. This technique appeared with the realization and introduction to mathematics by S. Eilenberg and S. Maclane, so-called natural transformations, that is, with the creation of category theory [1,2].

Even before this natural constructions were used in geometry by F. Klein [3] and later by E. Cartan [4] to study the *internal invariant* properties of geometric objects, regardless of their location or coordinate representation in the space under admissible transformations. The study of F. Klein's Erlangen program [3] has led to the understanding that any geometry is inseparable from the corresponding group of transformations of the space. The E. Cartan's method of moving frame [4] allowed to calculate the differential invariants of manifolds, foliations, differential equations for analytic transformation group and its various subgroups by a unified method based on the intrinsic properties of geometric objects.

Weaker invariants were introduced in algebraic topology. They captured the intrinsic properties of the space in the category of continuous transformations. It is, in algebraic topology the concept of the natural transformation was born.

The natural transformations provide an additional supporting information, which typically increases with the development of intuition in the area, but is rarely used explicitly by the researcher. Just as the natural laws require hard work in their discovery, natural transformations in mathematics are being recognized by the study. The word "natural" does not mean "obvious."

The Concept of Natural Transformation

As S. Maclane indicates [2], in order to introduce the central notion of *natural transformation* they [1] needed the notion of *functor*, which itself required the notion of *category*.

Definition 1. A *category* K is a class of arrows of the form $f : A \to B$ with domain A and codomain B. A and B are called objects of the category. Different arrows may have different domains and codomains. If the domain of one arrow coincides with the codomain of the other, as in $A \xrightarrow{f} B \xrightarrow{g} C$, these arrows can be multiplied to get a new arrow $g \circ f : A \to B$. The multiplication $\circ$ is associative, i.e. $(h \circ g) \circ f = h \circ (g \circ f)$ for the arrows $A \xrightarrow{f} B \xrightarrow{g} C \xrightarrow{h} D$.

For each object $A \in K$ there exists the identity arrow 1_A, such that $1_A \circ f = f$, $g \circ 1_A = g$ for any arrows f and g, when the multiplication is defined.

Typical examples of categories are the categories of sets or sets with additional structures with arrows, structure-preserving-maps.

Definition 2. A *functor* $F : K_1 \to K_2$ is a map of the objects of category K_1 to the objects of category K_2 and the arrows of category K_1 to the arrows of category K_2. The map F respects multiplication $F(g \circ f) = F(g) \circ F(f)$ and identities $F(1_A) = 1_{F(A)}$.

Examples of functors are the correspondence to each topological space the set of its connection components and to each continuous map the map of these components, or the assignment to each linear space its dual and to each linear map its transpose.

Definition 3. A *natural transformation* $\alpha : F \Rightarrow G$ of functor $F : K_1 \to K_2$ to functor $G : K_1 \to K_2$ is a class of arrows $\alpha_A : F(A) \to G(A)$ of category K_2, one for each object $A \in K_1$, such that $G(f) \circ \alpha_A = \alpha_B \circ F(f)$ for any arrow $f : A \to B$.

Examples of natural transformations are embedding of a linear space to its double dual, or taking the determinant of a quadratic matrix with an appropriate choice of the categories and functors, or the natural projection of a tangent bundle onto its base.

In the analysis of empirical data natural transformation appear as *natural extensions* of data $\alpha_A : A \to Ext(A)$, where Ext is an extension functor on the category of data with *infomorphisms* as arrows, that is, the maps not alternating information containing in the data. A and $Ext(A)$ are subsets of the base space, and their transformations are induced by the transformation of the total space. To describe it a convenient framework of fibered categories is used [5].

Typical examples of infomorphisms are the Euclidean motions, normalization of the coordinates $x_i \mapsto \dfrac{x_i}{x_{i\max}}$, repetition of some coordinates, as $(x_1,..., x_i,..., x_n) \mapsto (x_1,..., x_i, x_i,..., x_n)$.

Internal Uniform Regularity of Data

Empirical data, as a manifestation of some aspects of reality, necessarily carry its system properties. A general premise of the scientific description of reality is the axiom of local Cartesianess of the objects and processes of the real world. This means that the set of parameters, being somehow interconnected, approximately describes the process. Positivism asserts that this approximation can be arbitrarily fine.

It is assumed that the parameter set is complete, capturing the essential properties of the process. In this case, any set of empirical data will capture these properties. Extension of the data makes these properties more prominent. These properties *intrinsically* belong to the data and approximate the corresponding ones of the process.

Uniform regularity expresses local homogeneity of the process. It does not mean that the process should be little changeable locally, without any jumps. It means only local similarity of the laws, not excluding the singularity of the process.

Interior regularity is described in terms of the dependencies of the properties, which are attached to describe this kind of regularity. The author proposes to use the uniform regularity as *uniform continuity* with respect to the class of infomorphisms, i.e. transformations that leave the information unchanged. The *internality* inherent to the data property of the *best* uniform continuity is expressed in the fact that this attachment is *natural* with respect to the category of infomorphisms.

Approximately, modeling scheme of an object or process is as follows:

$$(Object, \mathrm{Re}lations_{Object}) \approx (X_1 \times X_2 \times \ldots \times X_n, \mathrm{Re}lations_{X_1 \times X_2 \times \ldots \times X_n}) \approx Data \approx Ext(Data),$$

where: *Object* is the investigated object, $\mathrm{Re}lations_{Object}$ are the relations between all, including hidden parameters of the object, $X_1 \times X_2 \times \ldots \times X_n$ is the Cartesian product of the observed

parameters, $Relations_{X_1 \times X_2 \times \ldots \times X_n}$ is an approximation of the actual relations, $Data$ are the observable data, $Ext(Data)$ is an internal extension of the data. All the approximations in the above equation are *natural*, that is *invariant* with respect to the category of admissible infomorphisms.

Lipschitz Uncertainty as an Expression of the Uniform Continuity

Geometrically, the data are specified with a map of a discrete set to the parameter space $Data : \{1,2,...,N\} \to X_1 \times X_2 \times \ldots \times X_n = X \times Y$, where: X_i, $i = 1,...,n$, are all the observed parameters, X and Y are the spaces of independent and dependent parameters, respectively.

In the case of a single metric ρ or a family of metrics $\{\rho_\alpha\}$, given on the space of parameters $X \times Y$, to each point $(x,y) \in X \times Y$ there is assigned a *Lipschitz subset* $U_{(x,y)} = \{(s,t) \in X \times Y \mid \forall \alpha \forall j \in \{1,2,...,N\}\ \rho_\alpha((s,t), Data(j)) \le \rho_\alpha((x,y), Data(j))\}$. Obviously, $U_{Data\,(j)} = \{Data(j)\}$.

Definition 4. *Lipschitz uncertainty* at point $(x,y) \in X \times Y$ is the diameter of the Lipschitz subset $U_{(x,y)}$, that is a non-negative real number $d_{(x,y)} = \max\limits_{\alpha,\,(s_1,t_1),\,(s_2,t_2)} (\rho_\alpha((s_1,t_1),(s_2,t_2)))$, where $(s_1,t_1),(s_2,t_2) \in U_{(x,y)}$.

A variant of Lipschitz uncertainty consistent with the transformations preserving the probability density is the probability measure of Lipschitz subset $U_{(x,y)}$. Although, in this case a full agreement of the proposed technique with the probabilistic approach to statistics is achieved, definition 4 for the prediction problem turns out to work better.

Definition 5. The *data extension* $Ext(Data)$ is a subset of the parameter space $\{(x,y) \in X \times Y \mid x \in X,\ d_{(x,y)} = \min\limits_{t \in Y}(d_{(x,t)})\}$.

Again, there are variants, what minimum should be considered, local or global one. As a rule, the presence of several local minima indicats different possible scenarios, due to the lack of parameters or data. Also, with the poor choice of a family of metrics on the parameter space, the set of minima attached to the point can be continual as an interval instead of a single value of the estimated parameter.

Proposition 1. The data extension $Ext(Data)$ is *natural* with respect to the isomorphisms of multimetric spaces.

Proposition 1 is quite obvious (for the proof, see [6]).

A Categorical Data Model

To describe the behavior of the data the author proposes fibered categories [5].

Definition 6. A functor $\pi : E \to B$ is called a *fibered category* with the total category E and the base category B, if for each arrow $f : X \to Y$ of category B there is a canonical lifting / Cartesian morphism $\bar{f} : \bar{X} \to \bar{Y}$ in category E, such that:

1) $\pi(\bar{f}) = f$,

2) for any arrow g of category E, for which there is an arrow k of category B, such that $\pi(g) = f \circ k$, there exists a unique arrow $\bar{k}$ in E, such that $\pi(\bar{k}) = k$ and $g = \bar{f} \circ \bar{k}$.

In modeling data, an object of the base category consists of a set M together with a family of metrics $O = \{\rho_\alpha\}$ and a distinguished foliation $S = \{s\}$, $\coprod s = M$. In Cartesian case the layers $s \in S$ of the parameter space $X \times Y$ are the subsets of the form $\{x\} \times Y$, $x \in X$. The data $Data$ appear in the objects of the total category as discrete subsets of the set M. Since the layers $s \in S$

are used to distinguish the data, in a good categorical model each layer contains at most one data point.

Definition 7. A *categorical data model* MD is the category consisting of quadruples $(M, \mathrm{O}, S, Data)$, where the values of the components are as described above. An arrow $f:(M_1,\mathrm{O}_1,S_1,Data_1)\to(M_2,\mathrm{O}_2,S_2,Data_2)$ of the category MD is a map $f:M_1\to M_2$, such that:

1) $\forall\rho_\alpha\in\mathrm{O}_2\ \ \rho_\alpha\circ(f\times f)\in\mathrm{O}_1$,

2) $\forall s\in S_1\ \ f(s)\subset t$ for some $t\in S_2$,

3) $f(Data_1)\subset Data_2$.

The functor $U_1:MD\to M$, forgetting the last component of the quadruple $Data$, is a fibered category with canonical liftings, determined by the Cartesian squares. Similarly, the functor $U_2:M\to MultiMet$ to the category of multimetric spaces, forgetting the last component of the triple, is a fibered category with canonical liftings, determined by the Cartesian squares.

Depending on the problem, in the category of data a subcategory of infomorphisms $MD_{\inf}\xrightarrow{\ i\ }MD$ is distinguished. It has the same objects as MD, and perhaps a more narrow class of arrows, that do not change the information contained in the data, depending on their representation.

Definition 8. The *data extension* is an endofunctor $Ext:MD_{\inf}\to MD_{\inf}$ of the subcategory $U_1\circ i:MD_{\inf}\to M$ of the fibered category $U_1:MD\to M$ together with a vertical natural transformation $\varepsilon:i\to i\circ Ext$, which components are not necessarily infomorphisms, that is, the following relations hold:

1) $U_1\circ i\circ Ext=U_1\circ i$,

2) $U_1\varepsilon=1_{U_1\circ i}$,

3) $\varepsilon_{(M,\mathrm{O},S,Data)}$ is an arrow of the category MD.

Proposition 2. All the possible data extensions form a *monoid* with respect to the operation $\varepsilon_2\bullet\varepsilon_1=(\varepsilon_2*1_{Ext_1})\circ\varepsilon_1:i\Rightarrow i\circ Ext_2\circ Ext_1$, where: $\varepsilon_2:i\Rightarrow i\circ Ext_2$, $\varepsilon_1:i\Rightarrow i\circ Ext_1$ are any two data extensions, $*$ and $\circ$ are the horizontal and vertical compositions of natural transformations, respectively. The unit of the monoid is the identity transformation $1_i:i\to i$.

The proof of proposition 2 is checking the associativity and unit laws (see also [6]).

In particular, if ε is not a trivial data extension, then ε^n is not a trivial data extension, too. The introduced above Lipschitz data extension is *idempotent*, $\varepsilon^2=\varepsilon$, i. e. its repeated application does not give any new information.

Remarks on Natural Statistics

The first gain as compared with the probability statistics is the *absence of the model* and therefore no modeling error is. Specifically, the model is completely determined with the only parameter space and data. In this kind of modeling the most important thing is which parameters should be selected. It may be that it is not so easy to choose the parameters of the model, such as the parameters determining the auction price of a picture, for example. At the same time even extra redundant parameters do not affect the accuracy of the prediction, only increasing the amount of computation.

The second gain is in *naturality*. In a sense, it is a proof of complete absence of an error. More specifically, the error is determined only by the incompleteness of the data and it cannot be reduced in the framework of the given information.

Also, the natural statistics is fully consistent and complementary to the probabilistic ones, since infomorphisms can always be taken as transformations preserving probability density, and the size of the Lipschitz set as its probability measure.

Some disadvantage of the method is in the large amount of computation in the case of global use of the data. When using only local data the volume of calculations becomes acceptable, but not all the data behavior is taken into account. In this sense, a great advantage of the probabilistic approach is the data compression by the equations of the model.

Conclusion

The article forces to introduce the idea of naturality into statistics as well as it was introduced into geometry, topology, and physics. In physics, too, the laws of nature are natural / invariant under the Lorentz transformations.

It is proposed to construct a natural extension of the data, being extracted only from the data themselves. Some arbitrariness of the method is associated with the choice of admissible transformations of the data representations. This is similar to a priori information in the probabilistic approach. Just how wrongly chosen a priori information does not ultimately spoil the prediction in the probabilistic statistics the wrongly selected category of infomorphisms does not ultimately spoil the prediction in the natural statistics and admits its improvement.

The method was tested for continuous and binary parameters in the real estate area and for the price prediction in the art. Everywhere the method showed much better results compared to the probabilistic one, with a *uniform error*, regardless of the case if it was typical or marginal.

The most appropriate and promising applications of the technique are those that do not have a good model of the process, such as the market, history, archeology, the field of experimental physics, art, betting forecasts, weather and others.

References

[1] S. Eilenberg, S. Maclane, General Theory of Natural Equivalences, Trans. Am. Math. Soc. 58 (1945) 231-294.

[2] S. Maclane, Categories for the Working Mathematician, second ed., Springer-Verlag, New York, 1998.

[3] F. Klein, Vergleichende Betrachtungen ber geometrische Forschungen, Erlanger Programm. 1872.

[4] E. Cartan, La méthode de repère mobile, la théorie des groupes continus, et les espaces généralisés, Actualités Scientifiques et Industrielles no. 194, 1935.

[5] B. Jacobs, Categorical Logic and Type Theory. Elsevier, North-Holland, 2001.

[6] G. V. Kondratiev, A Categorical-Informational Approach to the Value Prediction Problem, Atti della Accademia Peloritana dei Pericolanti: *Classe di Scienze Fisiche, Matematiche e Naturali*, Vol. **91**, No. 2, A3, 11 p.p., URL: http://dx.doi.org/10.1478/AAPP.912A3, 2013.