



2-FUN

*Full-chain and **UN**certainty Approaches for Assessing Health Risks in
FUture ENvironmental Scenarios*

FP6 Project-2005-Global-4

Integrated Project - Contract n°: 036976

- A SET OF SCRIPTS FOR LEARNING THE GNOSTIC MULTI-DIMENSIONAL ADVANCED ANALYSIS -

Due date of delivery: 31/07/2010

Actual submission date: 20/08/2010

Start date of the project: 01/02/2007

Duration: *48 Months*

Lead contractor organisation name for this deliverable: *IPH*

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)	
Dissemination Level	
PU	Public

PROPRIETARY RIGHTS STATEMENT

This document contains information, which is proprietary to the 2-FUN Consortium. Neither this document nor the information contained herein shall be used, duplicated or communicated by any means to any third party, in whole or in parts, except with prior written consent of the 2-FUN consortium.



Document Information

Document Name *A set of scripts for learning the gnostic multidimensional advanced analysis*

ID D1.12.doc

Revision Version 2

Revision Date 20/08/2010

Author P. KOVANIC/IPH, T.OCELKA/IPH and L. PAVLISKA/ IPH

Approvals

	Name	Company	Date	Visa
Author	P. KOVANIC	IPH	20/08/2010	P. Kovanic
Co-Author	T. OCELKA	IPH	20/08/2010	T. Ocelka
Co-Author	L. PAVLISKA	IPH	20/08/2010	L. Pavliska
WP Leader	A. MARCOMINI	UNIVE	20/08/2010	po E. Giubilato
Coordinator	F. BOIS	INERIS		F. Bois

Documents history

Revision	Date	Modification	Author
Version 0		Template made available to the WP leaders	F. BOIS
Version 1	29/07/2010	First version	P. KOVANIC, T.OCELKA, L. PAVLISKA
Version 2	20/08/2010	Definitive version after format revision	P. KOVANIC



Contents

This deliverable consists of the following components:

1. LECTURE NOTES TO APPLICATIONS OF MATHEMATICAL GNOSTICS III. (ScriptsMDA: Multidimensional Analysis), file D1_12.docx
2. Presentation GSCHOOL1.pdf (Analysis of Small Data Samples, Introduction to Mathematical Gnostics, 1893 kB)
3. Presentation GSCHOOL2.pdf (Mathematical Gnostics: Advanced Data Analysis, 1605 kB)
4. Presentation GSCHOOL3.pdf (Gnostic Distribution Functions, 1542 kB)
5. GSCHOOL4.pdf (Gnostic Multidimensional Analysis, 1598 kB)
6. Presentation GSCHOOL5.pdf (Comparison of Methods, 1836 kB)

Introduction and references

This text is related to 2-FUN Summer School on Gnostic Approach to Small Sample Data Analysis organized by the Institute of Public Health (Prague, 22.-25. June 2010).

The following files were distributed among participants:

A) Installation

1. GUIGnostic:
 - i) Gui_6.7.zip (package of Gnostic functions for R-environment, 2901 kB)
 - ii) RGNGui_1.3.zip (graphical user interface to Gnostic package, 21 kB)
2. R-2.11.0-win32 (Environment R for Windows, 31.9 MB)
3. RAndFriendsSetup2110V3.1-4-1 (Extensions to R with RExcel and statconn, 254 MB)

B) Lessons:

1. Guide.pdf (Kovanic P.: Guide to Gnostic Analysis of Uncertain Data, 447 pp., 2906 kB)
2. GuideGui.pdf (Pavliska L.: Guide to Gnostic GUI Package RGNGui, 20 pp., 330 kB)
3. Four exercises for practicing gnostic analysis programmed by L. Pavliska in RExcel:
 - i) RExcelInstall (2.88 MB)
 - ii) RExcelIntroR (3.90 MB)
 - iii) RExcelGnostic1 (3.54 MB)
 - iv) RExcelGnostic2 (2.26 MB)

C) R scripts for the Tinn-R environment:

- i) Tools.R (1 kB)



- ii) IntroductionToR.R (2 kB)
- iii) GuiExamples.R (9 kB)
- iv) GuiExamplesMD.R (4 kB)
- v) IntroToR.xls (22 kB)

D) Optional:

Lectures of P. Kovanic in the form of Microsoft Power Point presentations:

1. GSCHOOL1.pdf (Analysis of Small Data Samples, Introduction to Mathematical Gnostics, 1893 kB)
2. GSCHOOL2.pdf (Mathematical Gnostics: Advanced Data Analysis, 1605 kB)
3. GSCHOOL3.pdf (Gnostic Distribution Functions, 1542 kB)
4. GSCHOOL4.pdf (Gnostic Multidimensional Analysis, 1598 kB)
5. GSCHOOL5.pdf (Comparison of Methods, 1836 kB)

The 3rd part of the Guide (item B.2 above) is called “Practice of Gnostic Analysis”. Chapter 17 (Easy Analysis) contains section Easy Marginal Analysis, which was submitted to the project 2-FUN in the form of the deliverable D1_11. It is followed by the sections “Easy Correlation Analysis” and “Easy Multidimensional Analysis” and chapter 18 – “Concise Glossary”, which are reproduced here as parts of the deliverable D1.12.

Both chapters 17 and 18 were to be used as scripts for the Summer School.

Following notation is used below referring to the Guide:

- Code e.g. G2.3.1 is to be read as chapter 2, section 3 and subsection 1 of the Guide.
- Symbol (x.y) refers to the formula of the Guide.
- Words written in italics are names of the Gnostic functions.

1. Easy Correlation Analysis

There are two classes of gnostic functions providing robust estimates of correlation coefficients (CC) and matrices (CM) based:

- on the linear dependence of the probability on irrelevance ((2.22), (2.47), G2.3.1),
- on the inherent dependence of the regression model on the correlations (G2.3).

Both they are advantageous under different conditions:

- ❖ **IF** (You need the CC of two samples, a part of which is censored) **THAN** (Run *Gcor2*.)
- ❖ **IF** (You need the CC of two samples, which are surely homogeneous) **THAN** (Run *Gcor2*.)
- ❖ **IF** (You need the CM of M samples, a part of which is censored) **THAN** (Run *Gcor2M*.)
- ❖ **IF** (You need the CM of M samples, which are surely homogeneous) **THAN** (Run *Gcor2M*.)
- ❖ **IF** (You need a fast estimate of the CC of two samples) **THAN** (Run *Gregcor2*.)



- ❖ **IF** (You need a fast estimate of the CM of M samples) **THAN** (Run *Gregcor2M*.)
- ❖ **IF** (You need to test the statistical significance of gnostic correlation coefficients) **THAN** (Use the p.values in results of *Gcor2*, *Gcor2M*, *Gregcor2* and *Gregcor2M*.)
- ❖ **IF** (You need a complete review of a gnostic CM) **THAN** (Run *AnCorMat*)
- ❖ **IF** (You need to review elements of a CM satisfying a given level of significance) **THAN** (Run *SignifCor*)

2. Easy Multidimensional Analysis

The Gnostic approach to multidimensional analysis is based on the fundamental publication¹, which has been generalized in dissertation² and in the book³ and which is described in G6.1. Its implementation is represented by the function *\$GWLS\$* and its modifications *\$Glogreg\$* and *\$Gprobreg\$* (denoted collectively by *GMs*). There are 13 arguments of these functions. It is therefore reasonable to give some hints to their selection.

2.1 Easy Doing the MD-models

- ❖ **IF** (You are sure that the regression equations have the standard (additive, linear) form) **THAN** (Run *GWLS*) **ELSE** (Run *Glogreg*.)
- ❖ **IF** (You are sure that the regression equations have the multiplicative (log-linear, additive in logarithms) form) **THAN** (Run *Glogreg*) **ELSE** (Run *GWLS*.)
- ❖ **IF** (You have distribution functions of all variables available) **THAN** (Run *Gprobreg*) **ELSE** (Run *GWLS* or *Glogreg*.)
- ❖ **IF** (x is a numerical matrix with linearly independent columns) **THAN** (Substitute x into *GMs*) **ELSE** (Stop.)
- ❖ **IF** (You wish to run a standard (explicit) model and y is the dependent vector) **THAN** (Use y in *GMs*) **ELSE** (Use y=0 in *GMs*.)
- ❖ **IF** (You have a first guess w for iterated weights) **THAN** (Use w in *GMs*) **ELSE** (Use the default.)
- ❖ **IF** (There is no constant column in x and you wish to include the intercept) **THAN** (Use int=T) **ELSE** (Use int=F.)
- ❖ **IF** (You wish to use the gnostic version of the *GMs*) **THAN** (Use Nmet=1) **ELSE** (Use Nmet ≠ 1.)

¹ Kovanic P., *A New Theoretical and Algorithmical Basis for Estimation, Identification and Control*, Automatica V22, (1986), 6, 657-674.

² Kovanic P., *Gnostická teorie neurčitých dat, (Gnostic Theory of Uncertain Data)*, In Czech, doctor (DrSc.) dissertation, The Institute of Information Theory and Automation, Czechoslovak Academy of Sciences, Prague (1990), 161 s., (<http://www.math-gnostics.com>)

³ Kovanic P., Humber B.M., *Economics of Information, Mathematical Gnostics for Data Analysis*, (2003), 714 s., (<http://www.math-gnostics.com>)



- ❖ **IF** (You wish to run one of statistical versions of the *GMs*) **THAN** (Nmet ≠ 1) **ELSE** (Use Nmet=1)⁴.
- ❖ **IF** (There are weights *w_x* available given a priori and remaining constant in iterations) **THAN** (Use *w_x* in *GMs*) **ELSE** (Use defaults.)
- ❖ **IF** (You wish to use the Minimum Penalty Estimate instead of the Least Square iteration) **THAN** (Use MPE=T) **ELSE** (MPE=F.)
- ❖ **IF** (You wish to apply the gnostic filter to the dependent variable) **THAN** (Use rob2=T) **ELSE** (rob2=F.)
- ❖ **IF** (You wish to limit the number of iterations by Nit) **THAN** (Use iter=Nit) **ELSE** (Use default.)
- ❖ **IF** (You wish to limit the precision of iterations by Kacc) **THAN** (Use acc=Kacc) **ELSE** (Use default.)
- ❖ **IF** (You wish to test the limit of iterations by the vector *w*) **THAN** (Use test.vec="w") **ELSE** (Use default.)
- ❖ **IF** (You wish to the limit of iterations by the vector of residuals) **THAN** (Use test.vec="resid") **ELSE** (Use default.)
- ❖ **IF** (You wish to test the orthogonality of residuals with *x*) **THAN** (Use test.vec=NULL) **ELSE** (Use default.)
- ❖ **IF** (You wish to get complete intermediate prints on the screen) **THAN** (Use Print=T) **ELSE** (Print=F.)
- ❖ **IF** (You wish to have prints with D valid digits) **THAN** (Use valdig=D) **ELSE** (Use default.)

2.2 Easy Using the MD-models

The shortcut *GMs* states again for any of functions *GWLS*, *Glogreg* or *Gprobreg*.

- ❖ **IF** (You wish to explore one-to-one dependencies between columns of a matrix) **THAN** (Run *plotgr*.)
- ❖ **IF** (You wish to explore one-to-one dependencies between logarithms of columns of a matrix) **THAN** (Run *plotgrlog*.)
- ❖ **IF** (You need to explore residuals of an MD-model) **THAN** (Run *GraphMDres*.)
- ❖ **IF** (You wish to test the significance of coefficients of an MD-model) **THAN** (Run *GMs* and use the p.values of the result.)
- ❖ **IF** (You wish to order MD-objects) **THAN** (Run *GMs* and use values Score of results.)
- ❖ **IF** (You wish to test homogeneity of an MD-sample) **THAN** (Run *GMs* and test the residuals for homogeneity.)

⁴ Classical non-robust Least Square version is available with Nmet=0, Robust statistical versions with Nmet=2, ..., Nmet=11 are available in S-PLUS implementation of *GMs*, while Nmet=2, 3 and 4 can be run in R-versions of *GMs*.



- ❖ **IF** (You need to extract the main cluster of an MD-sample) **THAN** (Run *MainClust.*)
- ❖ **IF** (You need predictions of a dependent variable of an MD-series of regression models) **THAN** (Run *MDpred.*)
- ❖ **IF** (You need to sequentially extract the homogeneous sub-clusters from an MD-series) **THAN** (Run *homogenizeM.*)
- ❖ **IF** (You wish to decompose an MD-sample into homogeneous MD-clusters) **THAN** (Run *homogenizeC.*)
- ❖ **IF** (You need to perform a complete analysis of a sequence of MD-objects) **THAN** (Run *ModelMDCS.*)
- ❖ **IF** (You need evaluation of impacts of explaining variables on the dependent ones) **THAN** (Run *ModelMDCS* and use the result *\$Impacts.*)
- ❖ **IF** (You need to continuously monitor an MD-process) **THAN** (Run *ModelMDser.*)

Concise Glossary

The most important terms used in this Guide are defined here. They are not ordered alphabetically, but logically.

For more details on these notions see Index of the Guide.

Symbol GX.Y.Z. refers again to the Guide to Gnostic Analysis of Uncertain Data, Chapter X, Section Y, Subsection Z. Symbol (x.y) refers to the formula of the same Guide.

Uncertainty: A lack of knowledge (G1.1, G1.3, G2.2, G2.3.1)

Data: Numerical images of real quantities (G2.2.1, G5).

Data sample: a set of observed data submitted to the analysis (G2.3, G5.13, G15.1).

Mathematical gnostics: Methodology of treatment of uncertain data consisting

- of the axiomatic theory of individual uncertain data and of small samples (Part I. of the Guide),
- of the methods for a robust treatment of data maximizing the information (G2, G3, G4),
- of algorithms for data treatment by gnostic methods (Part II. and III. of the Guide).

Geometry: A part of mathematics concerned with questions of size, shape, relative position of figures and the properties of space (G2.2.8).

Space of uncertain data: The set of uncertain data endowed with geometry determined by the data to be treated.

Euclidean geometry: The oldest ("elemental") geometry based on five postulates. Distance is measured by absolute values of difference between coordinates of points, lengths are square roots of sum of squared distances.

Minkowskian geometry: One of three geometries (Euclidean, Galilean and Minkowskian) which apply a constant weight to all elements of distances. Length is a square root of the difference between squared distances.

Riemannian geometry: Geometry applying inconstant weights to differentials of distances in dependence on the points of the space.



Quantification: Mapping of a real quantity (including its uncertainty) into a data item as its observed value (G2.1, G2.2).

Estimation: Mapping of an observed data item value into the estimate of the true quantity's value (G2.2).

Entropy: A measure of data uncertainty ((2.18), (2.19), G2.2.5).

Information: Alternative measure of data uncertainty evaluating the quality of a data item or of a data sample ((2.23, 2.26, G2.2.6).

Composition law: The way of accumulation of uncertain data and of their functions to summarize information from data (G2.3).

Robustness: Decreased sensitivity of estimates with respect to undesirable data:

- ❖ **external** robustness: suppressing the impact of the outlying data on the estimate,
- ❖ **internal** robustness: suppressing the impact of the inner disturbances of a data sample on the estimate.

Duality of estimates: Parallel existence of two versions of all gnostic formulae, quantification and estimation ones. The former ensure the robustness of internal and the latter of external type.

Finite data support: The numerical interval containing the observed data expressed by using their natural measuring scale (G5.2).

Infinite data support: A theoretical open interval of real dimension-less numbers (0, Inf), onto which the observed data and estimates are transformed for operations based on the gnostic theory (G5.2).

The lower (LB) and upper (UB) bounds of the data support: The bounds of expected data values objectively estimated from the data, or subjectively given by the user. They are expressed in their natural scale or in their dimensionless form transformed onto the infinite data support (G4.3, G5.2).

Additive data: Observed data, the values of which can be expected to reach an arbitrary finite real positive and negative value or zero. The natural numerical operation on additive data is the addition and subtraction (G5.2).

Multiplicative data: Observed data, the value of which can be expected to reach only positive and finite real values. The natural numerical operation on multiplicative data is multiplication and division (G5.2).

Irrelevance: Gnostic (Riemannian) measure of the data error. ((2.11), (2.16), (2.17), (2.30), (2.31), (4.5), (6.5) through (6.7)).

Gnostic data weight: Relevance of a data value estimated by the gnostic algorithm ((2.10), (2.16)).

A priori data weight: Relevance of a data value known before the analysis (G5.10, G8.2).

Censored data: Left-censored data having values less than LOD (Limit of Detection) , right-censored data reaching values exceeding the LOR (Limit of measuring Range) , or interval data, which have uncertain values from a closed interval (LID, UID) with some known bounds (G5.4).

Probability: Expectation of a data value (of a quantile) estimated by a gnostic formula expressed by a number between zero (unexpected value) and 1 (almost sure value) (G2.2.6, (2.22)).

WEDF (Weighted Empirical Distribution Function): Discrete probability distribution function computed directly from the observed data (G4.1, (4.1)-(4.4), G11.22).

GNDF: Gnostic probability distribution function. One of four functions *TKDF*, where *T* is the type (*E* ... Estimating, or *Q* ... Quantifying) and *K* is the kind (*G* ... Global or *L* ... Local) (G4.1-G4.3, G8.2).



Density of probability: Derivative $dP/d\log(q)$ or dP/dq of the probability distribution P (some of *EGDF*, *ELDF*, *QGDF* or *QLDF*) by the independent variable $\log(q)$ (G4.2, (4.9), (4.12), (4.16), (4.20), G11.10-G11.13).

Homogeneous sample: Data sample, the global distribution of which over the infinite data support has only one local maximum (G5.5).

Scale parameter: Multiplier of values of observed data unifying the data to make them comparable. It can be optimally and objectively estimated by the *GPDF* from the data or subjectively given by the user (G5.6, G11.17, G11.18, (2.1), (2.3)).

Membership bounds: The lower and upper (LSB and USB) bounds of the interval of data "properly" belonging to a certain homogeneous data sample. A data item leaving this interval would make the data sample non-homogeneous.

Homoscedastic data: Data sample, the scale parameter of which is constant (G5.7).

Heteroscedastic data: Data sample, the scale parameter of which is dependent on the data item's value (G5.7).

Cross-section data filtering: Using the smooth distribution function *GPDF* representing all the data to estimate the true data values by determination of quantiles to probabilities corresponding to the middle points of the steps of the *WEDF*.

Correlation coefficient: A numeric evaluation of similarity of two vectors of uncertain data. Unlike classical statistics, where correlation coefficient is the second unbiased mixed normalized statistical moment of two variables, more complex formulae are used in gnostics alternatively based

- on the linear dependence of the probability on irrelevance ((2.22), (2.47), G2.3.1),
- on the inherent dependence of the regression model on the correlations (G2.3).

Additive regression model: A system of (linear) equations defining the dependence of the variable to be explained on a linear combination of explaining variables with coefficients, which are to be estimated.

Multiplicative regression model: The additive regression model valid for logarithms of variables instead of the original ones.

Regression model in probabilities: The additive or multiplicative regression model describing dependencies between probabilities of variables instead of between the variables themselves.

Implicit regression model: Regression model with a constant (1) playing the role of the „dependent" variable.

Similarity of MD-samples: Multidimensional (MD-) samples are similar if they approximately satisfy the same regression model.

Ordering of MD-samples: Multidimensional samples are considered as ordered, if residuals of their robust regression models form a rising sequence.

Homogeneity of an MD-sample: A multidimensional sample is considered as homogeneous if the global distribution function of residuals of its robust regression model is homogeneous.

Cluster analysis of an MD-sample: Decomposition of an MD-sample into a set of homogeneous subsamples (clusters).