as the most exemplary researchers of their respective countries, placed *qualitative data analysis* at the center of their work. Converging thoughts led to mutual interest, fed by international meetings and exchange programs.

The first Japanese-French Scientific Seminar in Tokyo, held during 1987, finally made this common viewpoint official, thus allowing for an increase number of researches plus an enhanced reciprocal flow of ideas. Should the parties concerned gather again to assess the latest developments? All who took part in the Tokyo conference favored another meeting. Accordingly it convened during 1992 in Montpellier and the outcome of that highly successful seminar -this book- clearly showed that the decision to assemble was most worthy.

Since 1987 computer science has steadily provided scientists with increasingly powerful means to automatically collect and store data. Organizational concepts have been developed that allow for prompt retrieval of stored data either partly or in mass. Static or dynamic graphics, with or without color, are efficient and esthetic. Meanwhile, computing power has been enhanced tenfold. New computer potential multiplied the number of possible tools for *data analysis*. Unless clearly based on a consistent set of concepts, newly developed tools cannot experience their full potential in data processing and interpretation. All of this by no means waives overall reflection on the course these general concepts will take, but rather demands it. Bearing this in mind is how the present volume should be used. The authors propose ways to formalize steps in *data analysis*. They have studied their properties, drawing from them methods of analysis. They then propose software with which to implement the methods.

Such an approach gives birth to a new science with data at its core. Its nature, numerical, qualitative or symbolic, determines the type of operations possible with them. Their origin, whether exhaustive collection or sample, conditions the objective expected in their analysis. It seems justified to coin the term *data science* for this particular activity. Of course this construction must rest on solid ground. Several texts presented herewith would find a place in publications on multidimensional statistical analysis, exploratory data analysis, or clustering. Bringing these texts together in this book expresses the common conviction of its authors: To take data as a starting point provides a complementary vision of theory and practice, and avoids creating an unfortunate gap between these two steps both of which essential in any scientific process.

The Editors

Data Science at the Unité de Biométrie - Montpellier -

Yves Escoufier

ENSAM-INRA-UM II
Unité de Biométrie
9, Place P. Viala
34060 Montpellier Cedex 01, France

1 Introduction

In this section we present a view of the peculiarities of the organization and localization of data science specialists in Montpellier. In the following section we show how this situation brings forth original research in "data science": analysis of data collected under constraints, inference in data analysis, and non-linear data analysis.

To start with, I am truly pleased to welcome you to Montpellier for this second Japanese-French Scientific meeting. The relations between Japanese and French specialists in data science are well established. French scientists, present in this room like Maurice Roux and Ludovic Lebart, have had the opportunity to spend several months in Japan, while Japanese researchers like Noboru Ohsumi have been in France for quite some time. The length and quality of these individual relationships have been enhanced through various general meetings held in France or in Japan, like those of the International Biometric Society, or those regularly organized by Edwin Diday within the framework of INRIA. The Compstat conferences present other occasions when Japanese and French have opportunities to meet and I have had the pleasure of seeing that their scientific objectives are close at hand and deserve a mutually deepening sense of achievement.

The Japanese-French Scientific meeting held in Tokyo in 1987 was the first to materialize through this convergence. The memories retained by the participants and the record of proceedings clearly show that this first meeting

In opening this second meeting, we have the opportunity and the responsibility to develop our relationship, exchange knowledge of our works and plans, and cooperate in developing "data science." Let us hope that the preparations made by our local organizing committee will contribute to the success of this meeting.

2. The Unité de Biométrie, Montpellier

Next, I would like to say a few words about the organization in which we work on data science in Montpellier. First of all, our organization is a good example of complicated French construction. More important, though, the agronomic and biological environment in which we live has had powerful effects on the trend of our research.

Early in the 1980s here, representatives and leaders of the numerous institutions engaged in education and research on agronomy thought it necessary to coordinate their efforts in order to promote Montpellier as an important place for agronomy. With this, the idea arose that statistics and, more precisely, biometry would be reinforced if the few specialists spread out in the University, the Agricultural School, and the Agricultural National Research Institute worked together.

The Unité de Biométrie evolved from this simple idea and from a scientific point of view that life is easy: We are interested in computer intensive methods in statistics and their agronomic applications. The difficulties arise from our French administrative organizations.

We have three administrative supervisors. Each has his specific rules. Our Japanese colleagues should know that this situation is very common in France. A research leader must be a tightrope walker and a juggler to reach his goals. As a representative of a French university, I would be glad to learn from our Japanese participants a few details on how administrative problems are seen and resolved in Japan.

But we are here to talk about "data science." So let me set aside our administrative woes and proceed to the scientific aspect of our situation. At the moment, we are a group of fifteen permanent teachers or researchers working on statistics. The majority of us are teachers at University Montpellier II, the

scientific and technological university of Montpellier. We teach students of mathematics, computer science or biology. Because of these different orientations, some teachers concentrate more on classical and traditional statistical problems, while others specialize in new domains where using the computer and handling large amounts of real data presents one of the difficulties. My philosophy is that our postgraduate students must have a wide-ranging view of what is meant by the word "statistics." For this reason we appreciate opportunities to meet with statisticians working in various fields of endeavor.

The teachers who instruct classes in biology at the University or the Agronomical School, and researchers of the National Agricultural Research Institute meet agronomists and biologists continuously. The two specific aspects of the work done by these specialists is that they use experimental designs and want to test the differences between controlled responses.

Data analysis in France originated as a strong reaction against what seemed to be an exaggeration of the mathematical developments which masked the fact that one had to deal with real life problems. So, inspired by the classical French data analysis approach, we were torn between the pressure of our environment and the reluctance of the French data analysis community for inference. We have taken up the challenge of at the same time being faithful to data analysis and honest with the information derived from controlled experiments. I think we can now say that this new point of view has been productive.

3. Recent Developments in Data Analysis at the Unité de Biométrie, Montpellier

3.1. When data come from an experimental design, we have for each case i ($i = 1, 2, \ldots, n$) two types of information: a vector y_i of observed data and a vector x_i describing the conditions of observation of the case i. We will note $\mathbf{Y}$ as the matrix of observed data, and $\mathbf{X}$ as that of design.

The standard practice in data analysis is to run a principal component analysis, or a correspondence analysis, or a cluster analysis of the $\mathbf{Y}$ matrix and, after obtaining the results to use the $\mathbf{X}$ matrix to construct relevant supplementary points which are projected on the $\mathbf{Y}$ results. Thus, we obtain qualitative information on how the design is responsible for the responses. Scientifically speaking, this approach is not completely satisfactory. We know $\mathbf{X}$ from the beginning and use it only at the end. In a certain sense, we hide from our study on $\mathbf{Y}$ what we know about the cases. Based on previous work by C. R. Rao [9] and L. Lebart [7], R. Sabatier [10] developed a systematic approach to these situations, enlarging the problem in various ways: $\mathbf{X}$ can be a matrix of concomitant observed variables: we can have a $\mathbf{Z}$ matrix giving information on the variables: we can consider a three-way dataset of data and not only a matrix. Papers presented at this meeting by F. Kazi-Aoual et al [6] are related to this problem.

3.2. Let $\mathbf{P}_x$, be the projector on the subspace of $\mathbf{R}^n$ spanned by the column of $\mathbf{X}$. To summarize, let us consider $\hat{\mathbf{Y}} = \mathbf{P}_x(\mathbf{Y})$ and $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}}$. Roughly speaking, the solution to the problem stated in *3.1* comes from the simultaneous analysis of $\hat{\mathbf{Y}}$ and $\mathbf{R}$.

If C is a quantity such that $C(\mathbf{Y}) = C(\hat{\mathbf{Y}}) + C(\mathbf{R})$, then a criterion like $C(\hat{\mathbf{Y}})/C(\mathbf{Y})$ or $\dfrac{C(\mathbf{Y}) - C(\hat{\mathbf{Y}})}{C(\mathbf{Y})}$ can be used for quantifying the part taken by $\mathbf{X}$ in the results obtained for $\mathbf{Y}$. It is easy to see that a classical quantity used in data analysis, like the sum of the eigenvalues, could play the role of C.

The problem now in a specific analysis would be to decide when the value obtained for the selected criteria is large enough for us to say that $\mathbf{X}$ determines the response $\mathbf{Y}$. To start with, we used simulation tests (El Faouzi et al [3]), permuted the rows of $\mathbf{X}$ and computed the criteria for each permutation. We obtained a distribution of values for the criteria associated with random allocations of the cases in the design. If the value of the criteria obtained for the initial design exceeds the values obtained for the random allocations, we can say that the design is responsible for the responses.

Work by Mardia [8] led to analytic results on distributions of the values of some criteria. Papers presented at this meeting by F. Kazi-Aoual et al [6], or S. Holmes [5] give such results.

3.3. When $\mathbf{X}$ is the matrix of observed concomitant variables, it could be unsatisfactory to approximate $\mathbf{Y}$ with $\hat{\mathbf{Y}} = \mathbf{P}_x(\mathbf{Y})$. In other words, the explanation of $\mathbf{Y}$ by $\mathbf{X}$ could be non-linear, and we must try to perform a substitution of the subspace $\mathbf{R}^n$ spanned by the columns of $\mathbf{X}$ by a larger subspace, embodying the preceding one. Different approaches are available.

J. F. Durand [1] presents a non-linear transformation of the $\mathbf{X}$ data obtained from a set of spline functions to solve this problem. The same set of spline functions has been used by N. El Faouzi [2] to obtain a non-linear principal components analysis. Beyond the numerical and computational difficulties linked to this approach, are very difficult theoretical developments for establishing the existence or unicity of the solutions. The proximity of theoretical statisticians has been invaluable.

4. Conclusion

If I must conclude, I would say that looking at the papers submitted by our Japanese participants, several titles struck me as related to the three problems I have stated. I hope this meeting will give us an opportunity to progress together on these topics.

References

[1] Durand, J. F. (1990). Principal components analysis with respect to instrumental variables via univariate spline transformations; *Compstat 90*; K. Momirovic and V. Mildner (eds.); Physica-Verlag, 109-114.

[2] El Faouzi N. (1992). *Extensions non linéaires de l'analyse en composantes principales*; Thése de Doctorat; Université Montpellier II.

[3] El Faouzi, N. and Escoufier, Y. (1991). Modélisation I-spline et comparison de courbes de croissancees; *Rev. Stat. Appl.*, Vol. XXXIX, 1, 51-64.

[4] Escoufier, Y., Fraile, L., and Raibaut, A. (1991). Analyse des correspondances de données planifiées: étude de la chémotaxie de la larve infestante d'un parasite; accepté pour publication dans "Biometrics".

[5] Holmes, S. (1992). Randomization and bootstrap tests in the multivariate context; *Data Science and its Applications.*

[6] Kazi-Aoul, F., Sabatier, R. and Lebreton, J. D. (1992). Approximation of permutation tests for multivariate inference: application to species-environment relationship; *Data Science and its Applications.*

[7] Lebart, L., Morineau, A. and Fenelon, J. P. (1979). *Traitement des données statistiques: méthodes et programmes*; Dunod, Paris.

[8] Mardia, K. V. (1971). The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model; *Biometrika*, **58**, 1, 105-121.

[9] Rao, C. R. (1987). The use and the interpretation of principal component analysis in applied research; *Sankya*, **A 26**, 329-359.

[10]Sabatier, R. (1987). *Méthodes factorielles en analyse des données, Approximations et prise en compte de variables concomitantes;* Thése d'Etat; Université Montpellier II.

Quantification, Multiway Data Matrices, and Generalized Latent Equations

Chikio Hayashi
The Institute of Statistical Mathematics
Sakuragaoka Birijian #304
15-8, Sakuragaoka, Shibuya-ku
Tokyo 150, Japan

1. Introduction

Some discussions on data analysis of multiway matrices have been already given in [7], where it was mentioned that the future was not promising without any development of new mathematics of multiway matrix and determinant. The innovation of mathematics of multiway matrix and determinant will be promoted by best-fitting applied such a theory, for example, by necessary demands of data analysis on multiway matrices. In the present situation, there are few severe demands of applications. The development of useful methods is hopeless without any good application on exploratory ideas for the analysis of actual multiway data. Some useful methods of data analysis on multiway data matrices are found. However there are a very few widely-used methods as there are a very few problems which cannot be essentially solved without such a multiway-method. We meet frequently the situation that the formal or trivial or apparent (not essential) generalization of analysis of two-way data matrices to more than three way matrices is found too. Suppose that there are any usual two-way data matrices. The construction of three-way data matrices by the addition of time, space and other such kinds of factors as the third dimension is, generally speaking, a wrong course of generalization, because separate data