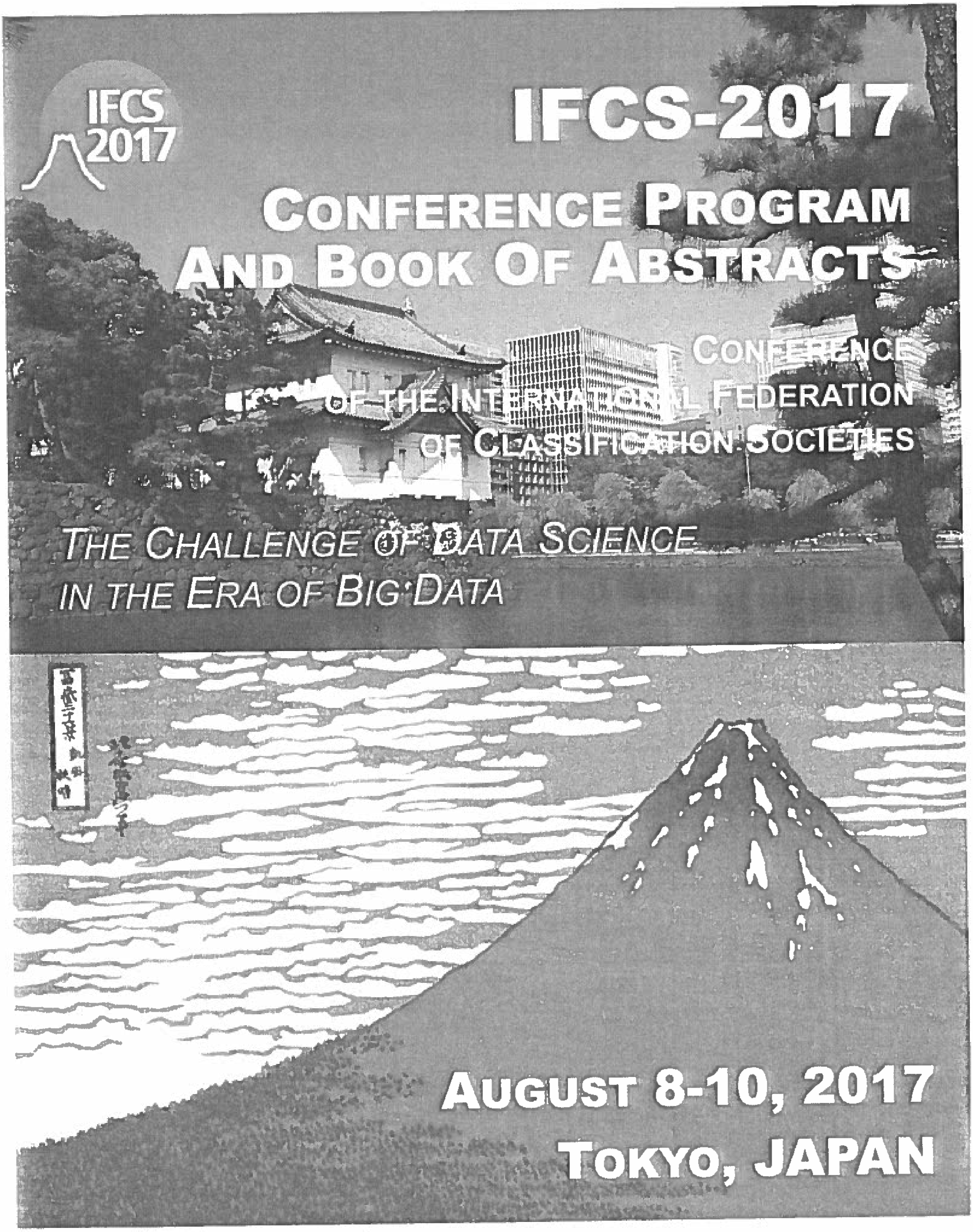

IFCS-2017

Conference Program And Book Of Abstracts

Conference of the International Federation of Classification Societies

The Challenge of Data Science in the Era of Big Data

August 8-10, 2017

Tokyo, Japan

JCS

Japanese Classification Society

ifcs

TOKAI UNIVERSITY

In cooperation with Japan National Tourism Organization and The Institute of Statistical Mathematics

IFCS-2017
Conference Program and Book of Abstracts

Conference
of the International Federation
of Classification Societies

The Challenge of Data Science in the Era of Big Data

August 8-10, 2017
Tokyo, JAPAN

IFCS-2017 Organizing Team
Japanese Classification Society
Tokai University
International Federation of Classification Societies

Organizing Team

Local Organizer

Tadashi Imaizumi

Scientific Program Committee

Tadashi Imaizumi, SPC Chair
Sadaaki Miyamoto, SPC Vice Chair, JCS
Yoshiro Yamamoto, LOC Chair
Akinori Okada, IFCS President
Maurizio Vichi, IFCS Past-President
Berthold Lausen, IFCS President-Elect
Christian Hennig, IFCS Secretary
Paul McNicholas, IFCS Publication Officer
Nema Dean, IFCS Treasurer
Vladimir Batagelj, SSS
Theodoros Chadjipantelis, GDSA
Sonya Coleman, IPRCS
José Gonclaves Dias, CLAD
Carlos Cuevas Covarrubias, SoCCCAD
Brian C. Franczak, CS
Salvatore Ingrassia, CLADAG
Hans Kestler, GfKl
Éva Laczka, HSA-CMSG
Sugnet Lubbe, SASA-MDAG
Fionn Murtagh, BCS
Mohamed Nadif, SFC
Heon Jin Park, KCS
Józef Pociecha, SKAD
Abderrahmane Sbihi, MCS
José Fernando Vera, SEIO-AMyC
Jeroen K. Vermunt, VOC
Patrick Groenen, IASC Delegate

Local Organizing Committee

Yoshiro Yamamoto, LOC Chair
Takafumi Kubota
Koji Kurihara
Masahiro Mizuta
Miki Nakai
Junji Nakano
Atsuho Nakayama
Makiko Oda
Takuya Ohmori
Kosuke Okusa
Fumitake Sakaori
Kumiko Shiina
Akinobu Takeuchi
Makoto Tomita
Yuki Toyoda
Hiroshi Yadohisa
Satoru Yokoyama
Christian Hennig (IFCS Secretary)

Secretariat

Fumitake Sakaori, Secretary-General
Satoru Yokoyama, Assistant Secretary-General
Makiko Oda, General Affairs

Exhaustive relabeling experiments for biomarker selection, Ludwig Lausser, Alexander Groß, and Hans A. Kestler.

8/10 12:35 - 14:15 **SP47: Methods of data analysis and statistical measures in the social sciences** - Chair: *Theodore Chadjipantelis* (Room F)

Optimal model-based clustering with multilevel data, Fulvia Pennoni, Francesco Bartolucci, and Silvia Bacci.

A comparison of different applications of functional linear discriminant analysis, Sugnet Lubbe.

Changes in the gendered division of labor and women's economic contributions within Japanese couples, Miki Nakai.

Determining the similarity index in electoral behavior analysis: An issue voting behavioral mapping, Theodore Chadjipantelis and Georgia Panagiotidou.

Title: Optimal model-based clustering with multilevel data
Authors: Silvia Bacci, Francesco Bartolucci, Fulvia Pennoni
Track: Methods of Data Analysis and Statistical Measures in the social sciences (SP08)

In many contexts, sample units are clustered in groups according to a certain criterion, for instance employees in firms, students in classes, or patients in hospitals. These data are analyzed by multilevel models (Goldstein, 2011) and have important applications in the evaluation of public services, particularly in education and health. For instance, it may be of interest to make comparisons between schools or classes at national and international level on the basis of the students'acquired knowledge. Accountability systems in education have been promoted in the statistical literature mainly since the 90's by Goldstein and Spiegelhalter (1996), who supported the idea that the performance monitoring approach may improve efficiency.

In this work, we focus on models in which the multilevel structure is accounted for by a hierarchical set of discrete latent variables, even in the presence of multivariate responses; these latent variables are used to represent the unobserved heterogeneity between clusters (i.e., groups) of units and between units in each cluster, extending the Latent Class (LC) approach (Lazarsfeld and Henry, 1968) to the multilevel setting. In particular, two cases are of interest. The first is when the observed outcomes are polytomous, as they correspond to item responses, and data are collected at the same time occasion. This approach has been applied by many authors in the educational context, see among others Vermunt (2008) and Gnaldi *et al.* (2016). The second case of interest is when the data have a longitudinal dimension and heterogeneity between units is represented in a dynamic fashion by a Latent Markov (LM) chain, as proposed in Bartolucci *et al.* (2011); see also Bartolucci *et al.* (2013).

While maximum likelihood estimation through the Expectation-Maximization algorithm (Dempster *et al.* 1977) of the models mentioned above is already well established, an issue that still deserves attention is that of predicting the latent variables at cluster and individual level on the basis of the observed data. In the LC literature, the Maximum A-Posteriori (MAP) approach is commonly used for this aim; for each latent variable, it consists in selecting the value having the highest posterior probability, which corresponds to the conditional distribution of this variable given the observed data. For the models at issue, the MAP approach may be applied in two different ways: (*i*) the latent variables at cluster and unit levels are separately dealt with for each cluster and unit; (*ii*) we first predict the latent variable for each cluster and then we predict each individual-specific latent variable (or variables in longitudinal case) conditional on the value predicted for the corresponding cluster-level latent variable. Both approaches may lead to suboptimal predictions, in the sense that the predictions may not correspond to the MAP probability of all latent variables. A similar problem exists in the LM model literature, where the sequence of latent states predicted by the local decoding method may not correspond to the MAP sequence of latent states that may be found by the global decoding method (Viterbi, 1967, Juang and Rabiner, 1991).

We propose an alternative rule for the posterior classification that *jointly* considers individuals and groups. More in detail, the proposed rule is built by formulating the multilevel LC model in terms of an LM model (Bartolucci *et al.* 2013) and, then, considering a suitable adaptation of the Viterbi algorithm. The Viterbi algorithm applied in the hidden Markov literature has the advantage to have a linear complexity since it consists in finding the most likely sequence of latent classes on the basis of a forward and a backward recursion. The involved quantities may be interpreted as posterior probabilities by which we allocate each individual and cluster of individuals to a latent class.

To illustrate the proposed approach, we show the results of some applications related to two educational effectiveness studies by considering data collected with the purpose to assess differences in the education level. The first dataset is a collection of measures related to the entire Italian population of schools and classes at the end of the compulsory education period (having at least 10 years of education). These Italian data have been collected by the National Institute of Evaluation of the Educational System of Instruction and Training (INVALSI). They refer to the competences assessed in 2009 by a set of multiple choice items which are dichotomously scored and concern Italian reading and grammar and mathematics; the student gender is available as well as the geographical location of the school.

Another type of measurement on reading, mathematics, and science competences has been collected on the large-scale assessment surveys TIMSS (Trends in International Mathematics and Science Study) and PIRLS (Progress in International Reading Literacy Study). The surveys have been conducted in 2011 according to a sampling design that also accounts for the geographical area. We consider the achievement scores at the fourth grade when the Italian pupils are 9 to 10 years old. They have been related to a set of covariates collected by the background parents' questionnaires and by the principals' questionnaire of the schools (see also Grilli *et al.* 2016). The data are released according to five achievement scores for each subject and their variability should be due to the estimation process. These scores known as plausible values (Von Davier and Sinharay, 2013) result from the expected quantities calculated by the E step of the EM algorithm and they are an approximation of the conditional distribution of proficiency when the generalized partial credit model (Muraki, 1992) is used to estimate the performance of examinee subgroups.

Main references

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Chapman and Hall/CRC press, Boca Raton.

Bartolucci, F., Pennoni, F., and Vittadini, G. (2011). Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioral Statistics*, **36**, 491–522.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Gnaldi, M., Bacci, S., and Bartolucci, F. (2016). A multilevel finite mixture item response model to cluster examinees and schools. *Advances in Data Analysis and Classification*, **10**, 53-70.

Grilli, L., Pennoni, F., Rampichini, C., and Romeo, I. (2016). Exploiting TIMSS and PIRLS combined data: Multivariate multilevel modelling of student achievement. *The Annals of Applied Statistics*, **10**, 2405-2426.

Goldstein, H. (2011). *Multilevel Statistical Models*, John Wiley & Sons, Chichester, UK.

Goldstein, H. and Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, Series A*, **3**, 385-443.

Juang, B. H. and Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, **33**, 251–272.

Lazarsfeld, P. F and Henry, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.

Muraki, E. (1992). A generalized partial credit model: Application of the EM algorithm. *Applied Psychological Measurement*, **16**, 159-177.