Final meeting of the FIRB research group on *"Mixture and Latent Variable Models for Causal Inference and Analysis of Socio- Economic Data"*

Section: *Unit of Milano-Bicocca*

Principal Investigator and Coordinator

Fulvia Pennoni

Department of Statistics and Quantitative Methods

University of Milano-Bicocca

fulvia.pennoni@unimib.it

Research's team and main topics

- Four researchers from *Milano-Bicocca University* (Lombardy), 2 young researchers (Italian and Spanish); 3 researchers from *University of Catania* (Sicily);
- Research Topics mainly related to:
  - The development of the Latent Markov (LM) model for causal effects in the observational studies with respect to latent and observed confounders;
  - Comparisons of the LM models with existing proposals and development of a simple R environment for practitioners.

Research's team and main topics

- Developments of extended versions of Component Analysis (CA) to estimate a latent variable of interest by considering exogenous, endogenous, external and concomitant indicators;
- Implementation of SAS macros (www.sas.com) to specify and fit CA models with and without external covariates or concomitant indicators;
- Developments of the Cluster-Weighted Models (CWM) with random covariates for mixed-type variables (i.e. both categorical and numerical) and high-dimensional data;
- Development of the R packages for the CWM.

Main research results

- Proposal of a new estimation method for the causal effect of multiple treatments in observational studies with the maximum likelihood inferential approach;
- Development of a new secondary data analysis method concerning multivariate multilevel data to allow for a more proper effectiveness evaluation;
- Proposal of a nonparametric estimation algorithm and a data analysis method called Generalized Redundancy Analysis (GRA);
- Development of CWM for the data analysis when there is a hierarchical structure and mainly for the hospital's evaluation and monitoring.

Main applications with real data

- Self-related *health* assessment;
- Self-related first job *satisfaction*;
- Italian student *achievement* on Reading, Maths and Science on large scale assessment surveys;
- Evaluation of *cognitive and behavioural development* of children born in 2000 according to the family situation;
- *Human capital* development according to earnings, skills and type of contracts of the graduates from some Italian universities;
- Hospital's *effectiveness* to improve healthcare outcomes.

First Research's Work

- First presentation: Garriga, A., Pennoni, F., Romeo. I.
  - *Conditional average treatment effect: an application related to the partner union quality and divorce on the child's psychological wellbeing*

Research's work

- Main interest: Partner union quality and divorce
    - A challenge since we had to handle complex observational data to detect the effect of divorce on young children;
    - The data are collected within the Millennium Cohort Study (University of London) and are representative for UK;
    - Main aspects: Interesting to work with a sociologist, a challenge to account for the survey structure and weights into the models.

The survey data

- The survey is important and complex due to the rotating sample scheme;
- Children's cognitive development and their internalizing and externalizing behaviours are collected by various items at 9 months, age 3 and 5;
- Sum of the raw scores has been considered and they have been normalized (Bracken Basic Concept Scale) and grouped into 5 categories;
- A measure of the child's behaviour and temperament at age 9 months derived from the Carey Infant Temperament Scale (Carey and McDevitt, 1977) has been considered;
- It has been included in order to account for possible reciprocal effect between parenting practices and children's behaviour at age 3.

From the literature

- The following three research hypotheses have been considered:
  - *i*) parent relationship quality and family disruption are unrelated processes that have independent effects on children (*independent hypothesis*);
  - *ii*) the apparent effect of family disruption is mediated by the parent relationship quality (*selection hypothesis*);
  - *iii*) the effect of family disruption on children depends on the quality of their parents' relationship (*heterogeneity hypothesis*).

The model and data analysis

- In the following for the heterogeneity hypothesis we illustrate the model, the missing data pattern and some results:

- Main notation:
  - N number of units;
  - T binary treatment;
  - Y_j Potential Outcome (PO) for $j = 0, 1$;
  - *ps*: scalar propensity score equal to $P(T = 1|\boldsymbol{X})$ where $\boldsymbol{X}$ is a set of pre-treatment covariates;
  - $m_j(\boldsymbol{X}) = E[Y|\boldsymbol{X}, T = j]$ conditional expected value of the response;
  - The unconfoundness assumption is still required as for other statistical models to identify the causal effect.

The model

- The proposed PO model accounts jointly for the outcome and for the treatment probability by considering an *augmented inverse-propensity weighted* (AIPW) estimator;
- Like other IPW estimators it weights the observed outcome by the inverse probability of receiving the treatment but it combines regression with ps;
- It has the double robust property (Robins et al. 1994) since it reduces the sensitivity to the parametric model mispecification and improves precision;
- It is sufficient to correctly specify only one model to consistently estimate the treatment effects;
- The estimator has the property to be $\sqrt{n}$ consistent and asymptotically Gaussian distributed (Tsiatis, 2006).

Main notation

- The AIPW estimator can be broadly defined as in the following:

$$
\begin{aligned}
\hat{\theta} \quad = \quad & \frac{1}{N}\sum_{i}^{n}\left(\frac{T_i Y_i}{\hat{ps}_i} - \frac{(T_i - \hat{ps}_i)\hat{m}_1(\boldsymbol{X}_i)}{\hat{ps}_i}\right) \\
& -\frac{1}{N}\sum_{i}^{n}\left(\frac{(1 - T_i)Y_i}{1 - \hat{ps}_i} - \frac{(T_i - \hat{ps}_i)\hat{m}_1(\boldsymbol{X}_i)}{\hat{ps}_i}\right)
\end{aligned}
$$

- It is a consistent estimator of the average causal effect even when the covariates are not balanced between the treated and the control units (Waernbaum, 2011).

Main features of the data

- The first sweep (MCS1) involved 18,819 babies in 18,533 families, the babies were 9-11 months old;
- During MCS2 and MCS3 the children were aged 3 and 5 years;
- Survey weights are used to account ofr the initial sampling design, and adjustments were made for non-response;
- The response rates achieved for the second (2004/05) and third (2006) waves were around 78%.

Main features of the data

- For the MCS3:
  - The child cognitive development (at five years old) is assessed by: *Naming Vocabulary test* to assess expressive language; *Picture Similarities test*, to measure pictorial reasoning; and *Pattern Construction test*, to establish spatial ability;
  - The child behaviour is assessed by the family main respondent on: *emotional symptoms, conduct problems, hyperactivity or inattention problems, peer problems, and pro-social behaviour*.

Main features of the data

- Partnership quality is considered according to the Golombok Rust Inventory of Marital State (GRIMS, Rust *et al.*, 1990) to assess couple discord (mother, at MSC1 with 7 items);
- We obtained a continuous variable measuring the parents's relationship (PRQ) which is considered as an ordinal variable according to its quartiles;
- We consider three type of situations for the couple: stable, temporary separation and divorce.

Main features of the data

- We use many pre-treatment covariates collected at MSC1 (control variables) related to the socio-demographic characteristics, the mother's health and some life course variables such as mother's attitudes to single-parent upbringing;

- 9,222 children resulted to have a recorded family structure at all the three MSCs whose parents were stably married or cohabiting from the birth of the child until the age of the first test.

Main features of the data

- To handle missing covariates within the missing at random assumption we used multiple imputation by chained equations (MICE);
- We accounted jointly for each type of available variable (continuous, categorical, ordinal or nominal);
- The imputed dataset (the sample weights have been included in the imputation model) have been up to five so as to dispose of multiple predictions for each missing value;
- This data analysis step was performed to have a look to all the imputed datasets and to evaluate the sensitivity of the estimated parameters to the imputed values.

Some descriptive statistics

Parents' relationship quality by types of family instability

	Parental divorce	Total
Highest Relationship Quality (Q1)	208 16.09%	3390 30.87%
High Quality (Q2)	288 22.27%	3082 28.07%
Low Quality (Q3)	274 21.19%	2174 19.80%
Lowest Quality (Q4)	523 40.45%	2335 21.26%
Total	1293	10981

- Relationship quality according to quantity quantiles (Radaelli and Zenga, 2008).

Some descriptive statistics

Socio-demographic characteristics by family type:

Stable family (SF), Parental temporary separation (PTS), Parental Divorce (PD).

		Types of family instability*			
		SF (%)	**PTS (%)**	**PD (%)**	**Total (%)**
Household income	Q1	26.89	11.34	13.13	24.85
	Q2	22.37	12.71	14.98	21.24
	Q3	25.35	16.49	27.10	25.32
	Q4	17.94	33.33	28.11	19.55
	Q5	7.46	26.12	16.68	9.04
Number of children in the household	1 child	39.64	38.49	40.26	39.68
	2 children	38.59	35.05	35.94	38.19
	3 children	15.35	17.18	15.22	15.39
	4+ children	6.42	9.28	8.58	6.75

- Psychological variables with a score ranging from 0 to 10 and cognitive variables from 0 to 60.

Some results

Estimated effect of PD on the quartiles of parents' relationship quality

		Q1 highest		Q2 high		Q3 low		Q4 lowest	
	Effect	**ATE**	**PO**	**ATE**	**PO**	**ATE**	**PO**	**ATE**	**PO**
Conduct problems	Coef.	0.349	0.960	-0.013	1.232	0.053	1.399	0.210	1.639
	P>|t|	0.002	0.000	0.916	0.000	0.513	0.000	0.013	0.000
Hyperactivity problems	Coef.	0.288	2.507	0.025	2.970	0.207	3.137	0.403	3.538
	P>|t|	0.156	0.000	0.898	0.000	0.116	0.000	0.003	0.000
Internalizing problems	Coef.	0.200	1.013	0.087	1.254	0.156	1.308	0.174	1.487
	P>|t|	0.099	0.000	0.535	0.000	0.235	0.000	0.068	0.000

- According to the estimated Average Treatment Effect and PO values: *PD causes conduct problems to be increased on average of 0.35 from the averages of 0.96 for children whose parents are in stable situation when at the time of the first survey the partnership quality was excellent.*

Results: brief summary

- We analyze multiple domains of children's school readiness: cognitive, social and emotional well-being;
- We focus on very young children who are at a key point of their development (transition to school) while most research focuses on children in middle childhood or older;
- We account for family instability types scarcely covered in prior literature;
- We find that a non-negligible proportion of children from divorced families did not experience parent relationship problems;
- From our findings the dissolution of high-quality parental unions has the most harmful effects on children's psychological well-being. The magnitude of this effect is lower for children with the lowest level of parent relationship quality than for children with the highest level.

Main References

- Carey, W.B. and McDevitt, S.C. (1977). Infant Temperament Questionnaire, Philadelphia: Temple University.
- Garriga, A., Pennoni, F., Romeo. I. (2016). Parent relationship quality, family instability and children's school readiness. *Submitted*.
- Rust, J., Bennun, I., Crowe, M., Golombok, S., (1990). The GRIMS. A psychometric instrument for the assessment of marital discord. *J. Fam. Ther.*, **12**, 45–57.
- Robins, J.M., Rotnitzky, A., Zhao, L.P., 1994. Estimation of regression coefficients when some regressors are not always observed, *J. Am. Stat. Assoc.*, **89**, 846-866.

Main References

- Radaelli, P. and Zenga, M. (2008). Quantity quantiles linear regression, *Statistical methods and applications*, **17**, 455-469.
- Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.
- University of London, Institute of Education, Centre for Longitudinal Studies, *Millennium Cohort Study: First Survey, 2001-2003, Second Survey 2003-2005, Third 2006*, UK Data Archive, SN 4683, SN 5350, SN 5795.
- Waernbaum I. (2011). Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation, *Statistics in Medicine*, **31**, 1572-1581.

Second Research's Work

- Second presentation: Pennoni F., Grilli L., Rampichini C. and Romeo I.
  - *A multivariate multilevel model to analyze educational achievement in Reading, Mathematics and Science in Italy.*

Research's work

- For the Italian sample of the TIMSS&PIRLS 2011 Combined International Database for fourth grade students we consider a joint analysis of achievement in Reading, Mathematics and Science;
- We aimed to point out the association of covariates collected at the hierarchical levels on the outcomes;
- We aimed to explore the correlations among the outcomes at student and class level to account also for the residual class level correlations and to evaluate effectiveness.

TIMSS and PIRLS surveys

- TIMSS and PIRLS are **large scale assessment surveys** held by the International Association for the Evaluation of Educational Achievement (IEA); (Rutkowski, *et. al.*, 2010):
  - **TIMSS** (Trends in International Mathematics and Science Study): at *fourth and eighth* grades every four years since 1995;
  - **PIRLS** (Progress in International Reading Literacy Study): at *fourth* grade every five years since 2001;
- In 2011 - for the first time - TIMSS and PIRLS cycles coincided (Martin and Mullis, 2012, 2013);
- The **TIMSS&PIRLS 2011 Combined International Database** concerns *fourth grade* students and collects data from questionnaires administrated to students, parents, teachers, and school principals.

TIMSS & PIRLS 2011 data

- The data are collected according to two stage sample design (by accounting the hierarchical structure):
  - schools (primary units) to their size (number of students)
  - classes (secondary units, 1 or 2) are randomly sampled and all the students are assessed;

  - We consider the survey weights referred to the students (given by the sample weight and the adjustment weight) and the unconditional class weights and we assess the variability of the weights by computing the design effect in the sample by verifying a negligible bias when this weights are omitted in the analysis.

Plausible Values (PVs)

- The items are administered by a rotating scheme;
- The student provides responses only with respect to subsets of items
- This allow to:
  - minimize the testing burden;
  - ensure accurate population estimates;
- For each student the total score is then replaced by **five PVs**;
- Main challenges of the study: to deal with PVs and to make comparisons among countries; to consider the huge amount of variables collected at any level, to account for the survey design.

Plausible Values (PVs)

- PVs are random draws (imputed values) from the distribution of the total score derived by a suitable IRT model (Mislevy, 1991, Wu, 2005);
- PVs are handled by running separate analyses with each PV and combining the results through a multiple imputation procedures Rubin (1987);
- *To the best of our knowledge, all reports and papers exploit multilevel models for a single outcome.*

Model equation

- We specify the following *multivariate two-level* model:

$$Y_{mij} = [\alpha_m + \boldsymbol{\beta}_m \mathbf{x}_{mij} + \boldsymbol{\gamma}_m \mathbf{w}_{mj}] + u_{mj} + e_{mij}$$

  - outcome m (1: Reading, 2: Math, 3: Science)
  - student i
  - class j
  - $\mathbf{x}_{mij}$ vector of student-level covariates
  - $\mathbf{w}_{mj}$ vector of class-level covariates (also including covariates at higher level, e.g. school or province);
  - u_{mj} class-level errors
  - e_{mij} student-level errors

- The model is also suitable to include outcome-specific covariates such as teacher's experience (Snijders and Bosker, 2011).

Covariance matrices

Student-level errors: $\mathbf{e}'_{ij} = (e_{1ij}, e_{2ij}, e_{3ij})$ Class-level errors: $\mathbf{u}'_j = (u_{1j}, u_{2j}, u_{3j})$

- $\mathbf{e}_{mij}$ indep. across students, $\mathbf{u}_{mj}$ indep. across classes
- $\mathbf{e}_{mij}$ independent from $\mathbf{u}_{mj}$
- $\mathbf{e}_{mij}$ and $\mathbf{u}_{mj}$ are assumed to have a multivariate Gaussian distribution cantered in zero and with variance-covariance matrices

$$Var(\mathbf{e}_{ij}) = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ & \sigma_2^2 & \sigma_{23} \\ & & \sigma_3^2 \end{pmatrix}$$

$$Var(\mathbf{u}_j) = \boldsymbol{T} = \begin{pmatrix} \tau_1^2 & \tau_{12} & \tau_{13} \\ & \tau_2^2 & \tau_{23} \\ & & \tau_3^2 \end{pmatrix}$$

$\mathbf{Y}_{ij} = (Y_{1ij}, Y_{2ij}, Y_{3ij})'$ has residual var-cov matrix given by $\boldsymbol{\Sigma} + \boldsymbol{T}$.

Data description according the geographical area

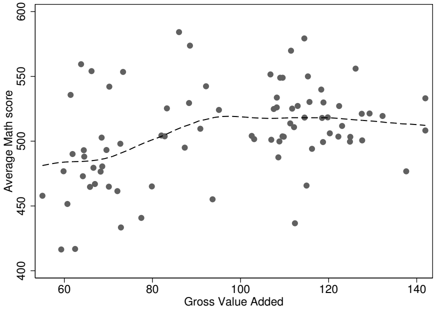
TABLE 1

TIMSS&PIRLS 2011 *sample sizes and average Gross Value Added* 2010 *for Italy by geographical area, alongside with MI combined average scores and their MI combined between-class standard deviations* (*in parenthesis*)

	Sample sizes				**Average score (between-class SD)**		
Area	**Classes**	**Teachers**	**Students**	**GVA**	**Reading**	**Math**	**Science**
North-West	48	103	849	122	540 (23)	516 (28)	538 (28)
North-East	49	103	920	120	534 (38)	508 (42)	529 (43)
Centre	48	97	852	113	530 (31)	506 (35)	524 (36)
South	49	97	832	66	517 (40)	499 (54)	508 (53)
South–Islands	45	83	672	69	502 (48)	475 (57)	489 (58)
Italy	239	483	4125	100	525 (39)	502 (46)	518 (47)

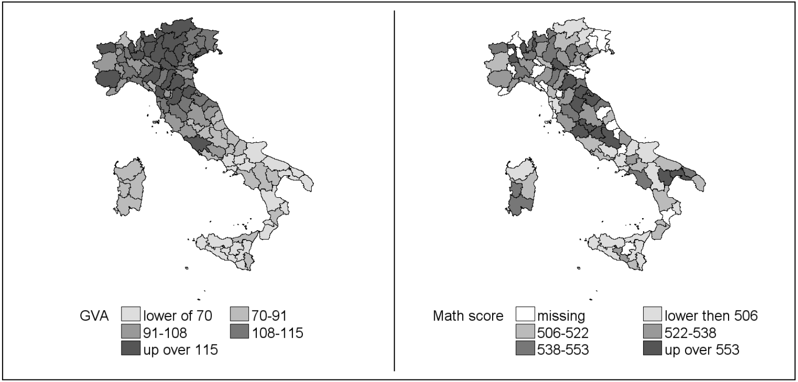
Additional covariate

- We could not find among the variables of the surveys a proper indicator of the territorial differences in wealth;
- We considered the estimated pre capita Gross Value Added at province level (Istituto Tagliacarne, 2010);
- We found that the student achievement is positively related to wealth only for province below the national average;



- The line for GVA$<$ 100 (national average) has a significant positive slope,
- the line for GVA$>$ 100 is nearly flat and the slope is not significantly different from zero. $\Rightarrow$ We constrained to zero the slope of the second line of the spline (i.e. GVA$>$ 100).

Cartograms by provinces



- Cartograms by province of the Gross Value Added 2010 (left panel) and the Math average score from TIMSS&PIRLS 2011 (right panel).

Selected covariates

The following *hierarchical order* is considered to include the covariates:

1 Student

- **Gender**
- **Language spoken at home**
- **Pre-school**
- **Home resources for learning**[1]
- **Early literacy/numeracy tasks**[2]

[1] Derived variable from number of books and study supports available at home, parents' levels of education and occupation.
[2] Derived from parents' responses on the perceived child's ability on early literacy/numeracy activities at the beginning of the primary school.

2 Teacher

- Gender
- Years been teaching

3 Class, School and Province

- % Students attended pre-school
- % Language spoken at home is not Italian
- Average of home resources for learning
- Average of Early literacy/numeracy tasks
- School is safe and orderly
- School with Italian students >90%[1]
- $<$ 10% of students has a low SES[1]
- School is located in a big area[1]
- **Adequate environment and resources**[1]
- **GVA**[2]

[1] Declared by the school principal.
[2] *per capita* Gross Value Added (GVA) at market prices in 2010 (proxy of the school socio-economic context).

Some results

TABLE 3

Multivariate multilevel model: main steps of the model selection process, first plausible value (TIMSS&PIRLS 2011, Italy)

Model	**n. par.**	**log *L***	**Significant covariates (on at least one subject)**
M0: null	15	−59,625.44	–
M1: student covariates	30	−59,119.09	Female, Preschool, Language spoken at home is not Italian, Home resources for learning, Early literacy/numeracy tasks
M2: student and class/school covariates	36	−59,109.75	M1 covariates + Class average Early literacy/numeracy tasks, School adequate environment and resources
M3: student, class/school and province covariates	36	−59,106.25	M1 covariates + School adequate environment and resources, Gross Value Added by province

Some results

TABLE 4

Correlation matrix decomposition and between-class percentage of (co)variances. Estimates from null model (M0) and final model (M3), MI combined estimates (TIMSS&PIRLS 2011, Italy)

	Correlations									**%Between-class (co)variances**		
	Within-class			**Between-class**			**Total**					
	Read	**Math**	**Scie**	**Read**	**Math**	**Scie**	**Read**	**Math**	**Scie**	**Read**	**Math**	**Scie**
M0 null												
Reading	1.00			1.00			1.00			19.8		
Math	0.71	1.00		0.93	1.00		0.76	1.00		29.5	28.8	
Science	0.81	0.74	1.00	0.97	0.98	1.00	0.85	0.81	1.00	28.2	35.0	29.4
M3 final												
Reading	1.00			1.00			1.00			16.3		
Math	0.67	1.00		0.93	1.00		0.72	1.00		27.6	27.6	
Science	0.77	0.70	1.00	0.97	0.97	1.00	0.80	0.78	1.00	25.3	34.1	26.9

Some results

TABLE 5

Multivariate multilevel model: parameter estimates and robust standard errors from the final model, MI combined results (TIMSS&PIRLS 2011, Italy)

	Read		**Math**		**Science**		***F* test**†
	Coef.	**s.e.**	**Coef.**	**s.e.**	**Coef.**	**s.e.**	***p*-value**
Intercept	531.73	3.57	514.99	4.25	531.47	3.92	0.0006
Student covariates							
Female	2.92	2.41	−11.96	3.05	−10.64	2.28	0.0000
Lang. home not Italian	−22.57	3.12	−14.94	3.27	−23.74	3.53	0.0161
Preschool	8.85	3.01	8.46	2.51	10.91	3.15	0.6386
Home res. for learning	14.04	0.84	10.64	0.84	13.23	0.93	0.0009
Early lit./num. tasks	7.24	0.77	10.07	0.76	6.53	0.83	0.0051
School covariates							
Adequate envir. & res.	5.28	1.92	8.61	3.19	7.00	2.96	0.1950

Main results

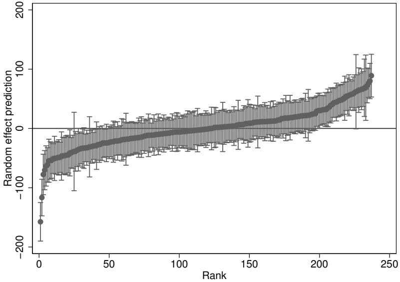
- The intercepts are related to the average scores for the baseline student: a male, whose language spoken at home is Italian, who did not attend at least three years of preschool, and with all the other covariates set at their mean;

- The performance of the baseline student is beyond the international mean of 500 in all the three outcomes, though the average score in Math is substantially lower than the average scores in Reading and Science;

- The regression coefficients are significant for all the three outcomes, except for being female, which is not significantly associated with Reading.

Empirical Bayes residuals

- The *class random effect* u_{mj} may be seen as the contribution of class j to the student's achievement in m;
- Empirical Bayes residuals for Math with their 95% CI:



- *good* classes (CI above 0): students on average achieve substantially more than expected on the basis of the covariates;
- *poor* classes (CI below 0): students on average achieve substantially less than expected

- Closer inspection of residuals reveals further territorial differences:
  - for example in the South there are high percentages of both *good* and *poor* classes ⇒ greater variability in achievement (we also tried to specify heteroscedastic random effects).

Final remarks

- Achievement in Reading, Math and Science from large-scale assessment surveys is studied jointly;
- By means of a *multivariate multilevel* model we could:
  - *estimate the correlations among outcomes*: they are higher at class level than at student level;
  - background covariates are equally associated with Reading and Science, but not with Math;
- We use the Gross Value Added (GVA) at province level instead of dummy variables for geographical areas and we gain a more refined interpretation of the territorial differences;
- The class-level Empirical Bayes residuals provide a measure to distinguish among *good* and *poor* classes and to inspect for further territorial patterns.

Main References

- Foy, P. (2013): *TIMSS and PIRLS 2011 user guide for the fourth grade Combined International Database* , Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Foy, P., O'Dwyer, L.M.: *Technical Appendix B. School Effectiveness Models and Analyses*, in: TIMSS and PIRLS 2011 Relationships report. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Grilli, L., Pennoni, F., Rampichini, C., Romeo, I. (2016). *Exploiting TIMSS and PIRLS combined data: multivariate multilevel modelling of student achievement, Annals of Applied Statistics*, **4**, 2405-2426.
- Martin, M.O., Mullis, I.V.S. (2013): *TIMSS and PIRLS 2011: Relationships Among Reading, Mathematics, and Science Achievement at the Fourth Graded Implications for Early Learning*, Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Main References

- Rabe-Hesketh, S., Skrondal, A. (2012).*Multilevel and Longitudinal Modeling using Stata (3th ed)*. Stata Press, College Station.

- Raudenbush, S.W., Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage, Thousand Oaks.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Sage, Wiley, New York.
- Martin, M. O., Mullis, I. V. S. (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Snijders, T.A.B., Bosker, R.J. (2011). *Multilevel analysis: an introduction to basic and advanced multilevel modelling (2nd ed)*. SAGE Publications, Inc.
- Rutkowski, L., Gonzalez, E., Joncas, M. von Davier, M. (2010). International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting. *Educational Researcher*.
- Wu, M. (2005). The role of plausible values in large-scale surveys. Studies in Educational Evaluation **31**, 114–128.

Third Research's Work

- Third presentation Pennoni F. and Romeo I.
  - *A comparison between two statistical models to analyse and predict individual changes over time*

LM and GMM comparison

- We propose a comparison between the latent Markov (LM) models and the Growth Mixture Models (GMMs) when the interest lies in modelling longitudinal ordinal responses and time-fixed and time-varying individual covariates;

- The interest on this topic is relevant since in many different contexts the ordinal data are a way to account for the importance given by an item or to measure something which is not directly observable;

LM and GMM comparison

- The LM models are observation-driven models tailored for many types of longitudinal categorical data (Bartolucci, Farcomeni, Pennoni, 2013);
- The evolution of the individual characteristics of interest over time is represented by a latent process with state occupation probabilities which are time-varying;
- The conventinal growth model or growth curve model (GCM) are viewed either as hierarchical linear models or as structural equation models;
- Their use in analyzing continuous response variables has been widely discussed in the literature (see, among others Duncan et. al, 1999).
- Their use in modeling and analyzing categorical data has received more attention (Muthén, 2002).

LM and GMM comparison

- We illustrate two recent extensions of the LM model and GMM when the ordinal response variable is made by thresholds imposed on an underlying continuous latent response variable;
- The models are compared on how they allow covariates, how they make inference, on their computational features required to achieve the estimates, and on their ability to classify units and their predictive power;
- Our proposal is an attempt to joint the recent literature on these models for panel data.

LM and GMM comparison

- The observed variables are obtained by categorizing the latent continuous response which may be related for example to the amount of understanding, attitude, required to respond in a certain category;
- Let Y_{it} be the observed ordinal variable for individual i , for $i = 1, \ldots, n$ at time t, $t = 1, \ldots, T$;
- We assume an underlying continuous latent variable Y^*_{it}, via a threshold model given by

$$Y_{it} = s \quad \text{iff} \quad \tau_{s-1} < Y^*_{it} \leq \tau_s$$

where $s = 1, 2, \ldots, S$ and $-\infty = \tau_0 < \tau_1 < \tau_2 < \ldots < \tau_{s-1} < \tau_s = +\infty$ are the cut-points by which it is possible to achieve a unique correspondence. With S response categories, there are $S - 1$ threshold parameters, τ_s, $s = 1, 2, \ldots, S - 1$.

LM and GMM comparison

- Under the basic model we assume the existence of a discrete latent process such that

$$Y^*_{it} = \alpha_{it} + \epsilon_{it},$$

with

- $\alpha_{i1}, \ldots, \alpha_{iT}$ following an hidden Markov chain; with state space denoted by $\xi_1, \ldots, \xi_k$,
- initial $\pi_u = p(\alpha_{i1} = \xi_u)$ and transition $\pi_{u|\bar{u}} = p(\alpha_{it} = \xi_u | \alpha_{i,t-1} = \xi_{\bar{u}})$, $\bar{u}, u = 1, \ldots, k$ probabilities,
- ϵ_{it} is a random error with normal or logistic distribution, see also Pennoni, F., Vittadini, G. (2013).

LM and GMM comparison

- In the case of time-varying or time-fixed covariates collected in the column vectors $\boldsymbol{x}_{it}$, the model is extended as

$$Y_{it}^{*} = \alpha_{it} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \epsilon_{it},$$

  as to include these covariates in the measurement model concerning the conditional distribution of the response variables given the latent process (McCullagh, P., 1980);

- The latent process is assumed to be a first-order homogeneous Markov process and we assume that it can generate independences among the responses $i = 1, \ldots, n$, $t = 1, \ldots, T$.

LM and GMM comparison

- A generalized linear model parameterization allow us to include properly the covariates in the measurement model;
- We carry out the estimation of the model parameters by using the maximum likelihood method through the EM algorithm;
- We select the number of latent states according to the information criteria like the Akaike or the Bayesian Information Criterion;
- The global decoding employing the Viterbi algorithm (Viterbi, 1967) allows us to obtain the most a posteriori likely predicted sequence of states for each individual.

LM and GMM comparison

- The LGCM without covariates is defined by the following equations

$$
\begin{aligned}
Y^*_{it} &= \alpha_i + \lambda_t \beta_i + \lambda_t^2 q_i + \epsilon_{it}, \\
\alpha_i &= \mu_\alpha + \zeta_{\alpha i}, \\
\beta_i &= \mu_\beta + \zeta_{\beta i}, \\
q_i &= \mu_q + \zeta_{q_i},
\end{aligned}
$$

- where α_i and β_i are the intercept and slope growth factor respectively, and q_i is the quadratic growth factor, for $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

LM and GMM comparison

- Time-varying covariates can only be included as predictors in the measurement model (*in the following the quadratic term is deleted to simplify the notation)*

$$
\begin{aligned}
Y_{it}^* &= \alpha_i + \lambda_t \beta_i + \boldsymbol{\omega}_{it} \boldsymbol{\gamma}_t + \epsilon_{it}, \\
\alpha_i &= \mu_\alpha + \boldsymbol{x}_i' \boldsymbol{\gamma}_\alpha + \zeta_{\alpha i}, \\
\beta_i &= \mu_\beta + \boldsymbol{x}_i' \boldsymbol{\gamma}_\beta + \zeta_{\beta i},
\end{aligned}
$$

for $i = 1, \ldots, T$ and $t = 1, \ldots, T$, where $\boldsymbol{\gamma}_\alpha$ and $\boldsymbol{\gamma}_\beta$ are vectors of parameters for the time-fixed covariates $\boldsymbol{x}_i$ on α_i and β_i, respectively, and $\boldsymbol{\gamma}_t$ is the vector of parameters for the time-varying covariates $\boldsymbol{\omega}_{it}$ on the measurement model;

- Different constrains can be imposed on the variance-covariance matrix.

LM and GMM comparison

- Maximum likelihood estimation of the model parameters requires numerical methods with categorical response variables and continuous latent variables;

- Model selection concerns the choice of the number of the latent classes and the order of the polynomial of the group's trajectories. The most common applied empirical procedure starts by choosing the order of the polynomial for the trajectories without accounting for the covariates within a model with just 1 latent group;

- Then, the number of latent classes is determined according to the unconditional model in order to avoid an over-extraction of the latent classes (Nylund *et al*. 2008);

- Finally, the covariates are added into the model as predictors of the latent classes.

LM and GMM comparison

- This relative entropy measure is commonly employed to state the goodness of classification

$$E_k = 1 - \frac{\sum_{i=1}^{n} \sum_{u=1}^{k} -\hat{p}_{iu} \log(\hat{p}_{iu})}{n \log(k)},$$

- where $\hat{p}_{iu}$ is the estimated posterior probability for unit i of belonging to the u-th latent class, k is the number of latent classes and n is the sample size. The values approach one when the latent classes are well separated;

- However, we notice that it differs from the normalized entropy criterion (NEC) proposed by Celeux and Soromenho (1996) within the literature on finite mixture models (McLachlan and Peel, 2000).

LM and GMM comparison

- We applied the model by considering a longitudinal study aimed at describing the self-perceived health status;
- The data is taken from version I of the RAND HRS data;
- $T = 8$ approximately equally spaced occasions, from 1992 to 2006; where we censored on the complete cases ended up with a sample of $n = 7,074$ individuals;
- The response variable SRHS is measured on a scale based on five ordered categories: "poor", "fair", "good", "very good" and "excellent";
- We used the following individual time-varying covariates gender, race, education and age (at each time occasion).

LM and GMM comparison

- i) The model estimation and selection procedure leading to the choice of the number of the latent states or classes:
  - The model choice is more complex for the GMM, and a stage procedure is suggested by the main statistical literature;
  - By following the suggested steps we found that the Monte Carlo integration for the GMM with a number of latent classes up to three leads to improper solutions;
  - The selection of the best model is more straight for the LM model, however it requires a search strategy to properly initialize the EM algorithm and therefore it is time and computational demanding for higher levels of latent states.

LM and GMM comparison

- ii) The way they relate the conditional probabilities of the responses to the available individual covariates:
  - The covariates are better handled by the LM model since they influence is allowed according to a suitable parametrization for categorical data such as global logits;
  - While in the LM model the covariates may affect the measurement part of the model or may influence the latent process, in the GMM they can affect both but in the measurement model only time-fixed covariates are allowed;
  - The LM model is more appropriate when the interest is on detecting subpopulations in which individuals may be arranged according to their perceived health status;
  - The GMM can be useful when just a mean trend is of interest and the expected subpopulations are not too many.

LM and GMM comparison

- iii) The model capability to use the posterior probabilities in order to get the predicted profiles for each latent class membership;
- We showed that the LM model outperforms the GMM mainly because it is more rigorous on each of the above points;
  - The predictions of the LM model are based on local and global decoding. The first is based on the maximization of the estimated posterior probability of the latent process and the second on a well known algorithm developed in the hidden Markov model literature to get the most aposterior likely predictive sequence;
  - In the GMM the prediction is based on maximum posterior probability and as showed in the example it may not be precise when the internal reliability of the model is poor.

Main References

- Bartolucci, F., Pandolfi S. and Pennoni F. LMest: An R package for latent Markov models for longitudinal categorical data. *Journal of Statistical Software*, pages 1- 38, *forthcoming*.
- Bartolucci, F. and Farcomeni, A. and Pennoni, F. (2014). Latent Markov Models: A Review of a General Framework for the Analysis of Longitudinal Data with Covariates (with Discussion), *Test*, **23**, 433–465.
- Bartolucci, F. and Farcomeni, A. and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*, Chapman and Hall/CRC press.
- Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification*, **13**, 195–212.
- McLachlan, G. J. Peel, D. (2000). Finite Mixture Models. New York: John Wiley Sons.
- McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society*, Series B, **42**, 109–142.

Main References

- Pennoni F., Romeo I. (2017). Latent Markov and growth mixture models for ordinal individual responses with covariates: A comparison, *Statistical Analysis and Data Mining: The ASA Data Science*, DOI: 10.1002/sam.11335.

- Muthén, B. O. (2002). Beyond SEM: general latent variable modeling, *Behaviormetrika*, **29**, 81–11.

- Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Structural Equation Modeling*, **14**, 535-569.

- Pennoni, F., Vittadini, G. (2013). Two competing models for ordinal longitudinal data with time-varying latent effects: an application to evaluate hospital efficiency. *Quaderni di Statistica*, **15**, 53-68.

- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.