Complex Redundancy Analysis models with covariate effect: a simulation study

Modelli di Analisi di Ridondanza complessi con covariate: uno studio di simulazione

Pafundi Pia Clara and Vacca Gianmarco

Abstract The focus of the present work are Structural Equation Models in the Redundancy Analysis framework (SEM-RA) and, in particular, the extension of Redundancy Analysis to more than two sets of variables, with the recently developed Extended Redundancy Analysis as the major outline. Drawbacks of the model in presence of concomitant indicators will be highlighted, thus introducing a further extension, Generalized Redundancy Analysis, whose introduction will be motivated, along with a simulation study aimed to assess the performance of the model, in three path diagrams at increasing complexity.

Abstract *Il focus del presente lavoro sono i Modelli ad Equazioni Strutturali nell'ambito della Redundancy Analysis (SEM-RA) e, in particolare l'estensione della Redundancy Analysis a più di due insiemi di variabili, attraverso la Extended Redundancy Analysis, di cui si evidenzieranno i limiti in presenza di indicatori concomitanti. Si introdurrà, motivando assieme ad uno studio di simulazione che valuti le prestazioni del modello, in tre path diagrams a crescente complessità, l'introduzione come ulteriore estensione della Generalized Redundancy Analysis.*

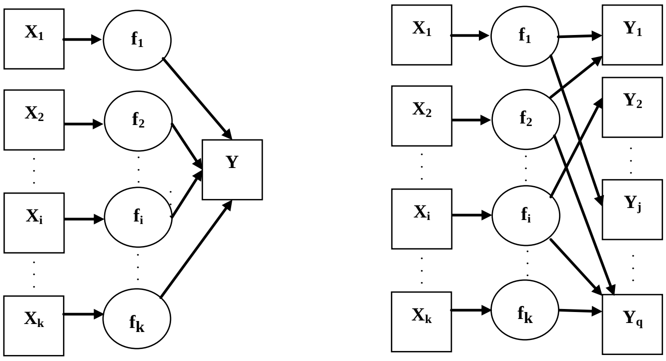
Key words: Redundancy Analysis, Structural Equation Modelling, Component Analysis

1 Introduction

In the Structural Equation Models - Redundancy Analysis framework (SEM-RA), our main focus will be on the extensions of the original Redundancy Analysis (RA; [10]) model, which analyses the causal relationship between two sets of multivariate data [4]. A few attempts have been made to extend RA to more than two sets of variables [8, 11], with the most relevant Multiblock Redundancy Analysis (MbRA; [1]) and Extended Redundancy Analysis (ERA; [3]) (see Figure 1).

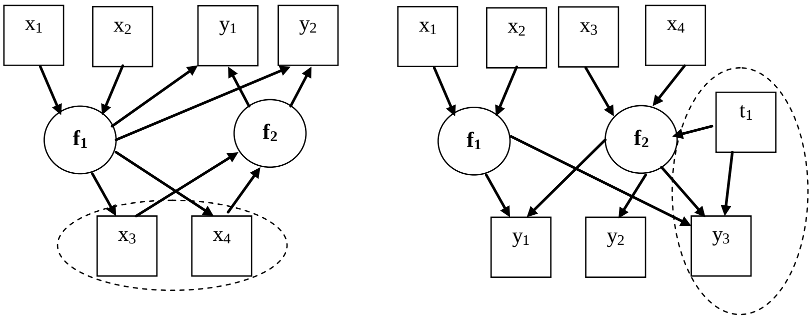
[1] Pia Clara Pafundi, e-mail: p.pafundi@campus.unimib.it; Gianmarco Vacca, g.vacca@campus.unimib.it;
Department of Economics, Quantitative Methods and Business Strategy, University of Milano Bicocca;

Figure 1: *MbRA (on the left) and basic ERA (on the right) path specifications.*



In MbRA the relationships between one block of dependent variables and several blocks of explanatory variables is modelled, maximizing the sum of the covariances between the latent constructs of each explanatory block and the latent construct of the dependent block, whereas in ERA a linear combination of the manifest variables is employed to obtain the latent composites (LCs), which are in turn fitted with the endogenous block. The estimates of the parameters are obtained minimizing a global LS criterion. ERA stands off as an incisive improvement in the SEM-RA framework, especially thanks to its versatility: it can accommodate more diverse and complex specifications than MbRA (in the ERA model in fig. 1 a LC does not necessarily have an impact on all the endogenous variables), and it can also include either (i) direct effects from covariates not strictly taking part in the formation of the latent composites or (ii) simultaneously exogenous and endogenous variables (see Fig. 2).

Figure 2: *Examples of two path diagrams not feasible with MbRA.*



2 Generalized Redundancy Analysis

Extending the ERA formal specification to evaluate direct effects without altering the model formulation or the estimation algorithm leads to inefficient solutions and misinterpretation of the coefficients, since (a) the estimation of the parameters forming the LCs is performed by ERA between the endogenous block (i.e. **Y**) and the exogenous and concomitant block altogether, in block matrix notation (i.e. **[X|T]**), ignoring direct effects; (b) **T** and **X** are typically correlated and present different causal effects on **Y**, making the separate contribution of each block to the determination of the LC scores not distinguishable. GRA [7] has been proposed as a further extension, to include concomitant indicators[1]:

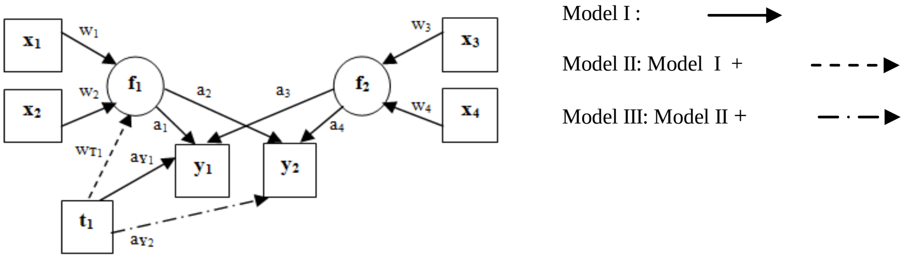
$$\mathbf{Y}= \mathbf{TA_Y}' + \mathbf{XWA}' + \mathbf{TW_T A}' + \mathbf{E}$$

where $\mathbf{W_T}$ is the corresponding weight matrix and $\mathbf{A_Y}$ is the corresponding direct coefficient matrix. Two further steps are then added to the algorithm: (1) the $\mathbf{TA_Y}'$ block is isolated and estimated separately; (2) **W** and $\mathbf{W_T}$ are estimated separately and correlation between **X** and **T** is accounted for, by matrix manipulation.

3 A simulation study

In this section the GRA model in three different path diagrams at increasing complexity (Fig. 3) will be evaluated through simulation (**SAS** software, [9]), focusing on GRA and how biases and accuracy of the estimates behave in presence of different concomitant links, rather than on comparing ERA and GRA in a single model specification (as in [7]). In fact, the ultimate aim of these simulations is to point out variations in bias patterns in a clear step-by-step increase in complexity: (i) a single covariate affecting a single endogenous variable (Model I); (ii) a single concomitant indicator affecting also LCs (Model II) and finally, (iii) a concomitant indicator affecting both the endogenous variables (Model III).

Figure 3: *Three GRA path specifications, with increasing complexity*.



The distributions underlying **X**, **T** and **E** simulated data is $\mathbf{X}\sim N_4(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{X})$, $\mathbf{T}\sim N(0, \sigma^2_T)$, and $\mathbf{E}\sim N_2(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{E})$, with covariance matrices and arbitrary parameters for **W** and **A'** defined as follows (as in [3]):

$$\boldsymbol{\Sigma}_\mathbf{X} = \begin{bmatrix} 1 & & & \\ 0.6 & 1 & & \\ 0.05 & 0.05 & 1 & \\ 0.05 & 0.05 & 0.6 & 1 \end{bmatrix} ; \quad \sigma^2_T = 1 \ ; \quad \boldsymbol{\Sigma}_\mathbf{E} = \begin{bmatrix} 1 & \\ 0.5 & 1 \end{bmatrix}; \quad \mathbf{W} = \begin{bmatrix} 0.6 & 0 \\ 0.6 & 0 \\ 0 & 0.6 \\ 0 & 0.6 \end{bmatrix} \ ; \quad \mathbf{A'} = \begin{bmatrix} 0.2 & 0.2 \\ 0.2 & 0.2 \end{bmatrix}$$

[1] A concomitant indicator is an exogenous covariate that does not strictly belong to the formative blocks of unobservable composites, but that may have a causal impact on observed endogenous variables and on composites too [8].

For Model I, with only one external link, $\mathbf{W_T}$ is null and $\mathbf{A_Y}$'= [0.3 0]; for Model II, with a concomitant link, $\mathbf{W_T}$ = [0.4 0] and $\mathbf{A_Y}$'= [0.3 0]; for Model III, with the second external link: $\mathbf{W_T}$ = [0.4 0] and $\mathbf{A_Y}$'= [0.3 0.3]. A population of N=1000 observations is then randomly generated and the endogenous matrix **Y** is calculated based on the above specifications. GRA is evaluated in each specification with increasing sample sizes, with two indices representing the accuracy of the estimates: (1) the congruence coefficient $\phi = \boldsymbol{\theta}^{*}\boldsymbol{\theta} / [(\boldsymbol{\theta}^{*}\boldsymbol{\theta}^{*})(\boldsymbol{\theta}\ \boldsymbol{\theta})]^{1/2}$ [5] between the estimates $\boldsymbol{\theta}^{*}$ and the real parameters $\underline{\boldsymbol{\theta}}$ and (2) the relative bias (in absolute value), $\gamma=100(|\theta - \theta^{*}|/\theta)$ of each estimate. Estimates have been obtained through Bootstrap resampling [2], for each sample size.

3.1 Simulation Results

In all three models convergence of the estimates was reached (see Table 1). Results show a general increase in Mean(ϕ) along with an increase in the sample size, indicating a satisfactory recovery of the population parameters, with n>200, for Models I and II ($\phi > 0.85$). The distribution of ϕ through the replications appears to be increasingly pointed towards high values, with decreasing standard deviations. Model III fails to recover efficiently the true values of the coefficients, with $\phi < 0.9$ until n = 600 and more variable results (StdDev(ϕ) = 0.110 at n = 400).

Table 1: *Descriptive statistics of* ϕ *for the GRA model* for n= {50; 100; 200; 400; 600} *and model complexity. For model III convergence criterion has been lowered from* 10^{-4} *to* 10^{-3}.

Model	I	II	III	I	II	III
n		Mean(ϕ)			StdDev(ϕ)	
50	0.653	0.597	0.494	0.226	0.228	0.226
100	0.770	0.734	0.642	0.158	0.172	0.219
200	0.855	0.845	0.784	0.113	0.107	0.173
400	0.912	0.903	0.880	0.074	0.065	0.110
600	0.929	0.925	0.917	0.058	0.053	0.067

Reasons are highlighted analyzing the estimates biases (Table 2), generally satisfactory in Models I and II, although the presence of the concomitant indicator affects some estimates peculiarly: with $\mathbf{t_1}$, **W** biases do not decrease, having w_1 and w_2 affected by its presence (→ $\mathbf{y_1}$ in I and also →$\mathbf{f_1}$ in II), for which $\gamma(w_1) > 50\%$ and $\gamma(w_2)$ has increasing bias for n = 50→300; without concomitant indicators, **A** biases do not decrease (a_3 and a_4 in I and II are affected by the absence of indicators related to $\mathbf{f_2}$, for which $\gamma(a_3)$ and $\gamma(a_4)$ are ~ 20%, regardless of n). Model III is more unstable for lower values of n, with biases mainly on loadings related to $\mathbf{f_1}$: having $\mathbf{t_1}$ and $\mathbf{f_1}$ sharing the same endogenous variables, should normally lower **A** biases, but the causal link between $\mathbf{t_1}$ and $\mathbf{y_2}$ is only external, influencing them similarly to

Model I. Moreover, the additional external link affects w_{T1} bias ($\gamma(w_{T1})$ > 20% at high sample size).

Table 2: *Estimates, S. Errors and biases for relevant parameters, by model and sample size.*

Model			I			II			III		
Par.	Pop.	n	Est.	S.E.	%Bias	Est.	S.E.	Bias%	Est.	S.E.	Bias%
w_1	0.6	50	0.18	0.701	70.1	0.17	0.614	70.1	0.16	0.558	73.3
		100	0.25	0.547	58.3	0.22	0.467	62.6	0.21	0.386	64.6
		200	0.27	0.384	55.1	0.28	0.352	53.7	0.24	0.320	60.5
		400	0.31	0.273	49.1	0.27	0.243	54.4	0.25	0.253	58.3
		600	0.32	0.216	45.9	0.3	0.202	51.8	0.27	0.208	55.2
w_2	0.6	50	0.61	0.631	2.3	0.52	0.537	14.7	0.49	0.504	19
		100	0.7	0.442	15.1	0.59	0.397	1.9	0.47	0.383	22.1
		200	0.74	0.301	23.7	0.63	0.301	5.4	0.61	0.310	1.3
		400	0.74	0.216	22.8	0.67	0.205	12	0.72	0.215	19.3
		600	0.73	0.175	21.6	0.68	0.17	13.1	0.69	0.167	14.8
a_1	0.3	50	0.23	0.140	16.9	0.3	0.297	48.4	0.37	0.399	85.2
		100	0.23	0.010	15	0.32	0.497	58	0.44	0.472	120
		200	0.22	0.068	10.1	0.25	0.152	26.7	0.29	0.260	45.9
		400	0.21	0.047	6.4	0.23	0.056	17	0.22	0.121	10.8
		600	0.21	0.039	4.1	0.23	0.230	15.1	0.23	0.040	15.4
a_2	0.3	50	0.23	0.138	15.3	0.28	0.280	40.2	0.27	0.153	76.9
		100	0.22	0.094	8.0	0.25	0.098	23	0.42	0.448	107.4
		200	0.2	0.072	1.2	0.23	0.069	14.8	0.28	0.281	40.5
		400	0.2	0.050	2.1	0.22	0.049	8.8	0.21	0.081	3.7
		600	0.2	0.041	1.5	0.21	0.042	7.2	0.21	0.041	6.5
a_3	0.3	50	0.27	0.152	33.9	0.26	0.151	31.5	0.27	0.153	34.5
		100	0.25	0.100	24.9	0.24	0.156	21.5	0.23	0.102	14.7
		200	0.24	0.071	22.1	0.25	0.071	22.9	0.22	0.065	10.9
		400	0.24	0.048	21.7	0.24	0.050	20.6	0.22	0.044	8.1
		600	0.24	0.039	21.7	0.24	0.042	19.7	0.24	0.043	13.7
a_4	0.3	50	0.27	0.136	37.2	0.26	0.144	30.9	0.26	0.151	30.7
		100	0.26	0.097	27.4	0.25	0.102	24.2	0.23	0.104	16.8
		200	0.24	0.070	21.9	0.24	0.072	20.2	0.22	0.069	12.2
		400	0.24	0.050	20.7	0.24	0.051	20.5	0.22	0.045	9.1
		600	0.24	0.039	21.7	0.24	0.04	19.9	0.24	0.041	19.2
w_{T1}	0.4	50	-	-	-	0.27	0.492	32.4	0.22	0.592	45.3
		100	-	-	-	0.31	0.407	23.7	0.15	0.662	62.9
		200	-	-	-	0.35	0.279	12.6	0.21	0.499	46.6
		400	-	-	-	0.38	0.208	6.3	0.25	0.315	36.9
		600	-	-	-	0.38	0.161	5.3	0.39	0.154	3.7

4 Conclusion

Several extensions of RA, among which MbRA [1] and ERA [3], have been proposed to investigate causal relationships between more than two datasets [8, 11]. However, both methodologies present limitations. ERA, in particular, studies more complex relationships among variables, also including possible direct effects of observed variables on endogenous variables [7], but its limits are clear when we want to analyze these effects without any change either in the model specification or in the estimation procedure. Thus, GRA [7] has been introduced, with innovative and theoretically valid features in the estimation of SEM, providing separate estimation for each different influential block and efficiently dealing with concomitant and external covariates simultaneously. GRA also provides good performances in recovering population parameters at sufficiently high sample size, either when compared to the ERA block-matrix counterpart [7], or in itself when challenged by increasingly complex path diagrams, as pointed out by the previous simulations.

Empirical validity of GRA has been proven in the analysis of Human Capital (HC), considered as a compound of latent traits derived from education (f_1) and working experience (f_2) variables, and of their impact on several economic indicators [6]: the related model has first been estimated without any concomitant indicator, adding it on second instance to test whether the economic background of the subjects (t_1) had a meaningful impact on the income variables. The provided results are fully interpretable and coherent, sustaining the hypothesis of a positive direct effect of the socioeconomic background on the income.

Some further complex specifications need to be tested, as the presence of correlations between different blocks, different underlying distribution, and concomitant indicators simultaneously linked to different LCs, for which correlations between LCs appear to be a crucial aspect to outline.

References

1. Bougeard S., Qannari E.M., Lupo C., Hanafi M. *From multiblock partial least squares to multiblock redundancy analysis. A continuum approach.* Informatica; **22**(1): 1–16, (2011).
2. Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia (1982).
3. Hwang, H., and Takane, Y. *Structural equation modeling by Extended Redundancy Analysis.* In Nishisato, S., Baba, Y., Bozdogan, H., and Kanefuji, K. (Eds.), *Measurement and multivariate analysis* (pp. 115-124). Tokyo: Springer Verlag, (2002).
4. Lambert Z. V., Wildt A. R., & Durand R. M. *Redundancy analysis: an alternative to canonical correlation and multivariate multiple regression in exploring interset associations.* Psychological Bulletin, 104, 282-289, (1988).
5. Lorenzo-Seva, U. & ten Berge, J.M.F. *Tucker's Congruence Coefficient as a Meaningful Index of Factor Similarity*. Methodology, *2*, 57-64, (2006).
6. Lovaglio P.G. Vacca G., Verzillo S. *Measuring Human Capital in Higher Education*. In “Advances in Latent Variables”, Eds. Brentari E., Carpita M., Vita e Pensiero, Milan (2013).
7. Lovaglio, P.G., Vittadini, G. *Component analysis for structural equation models with concomitant indicators.* In Giudici P, Ingrassia S, Vichi M (eds.) Statistical Models for Data Analysis, Springer: Milan (2013).
8. Reinsel, G. C., and R. P. Velu. *Multivariate Reduced-Rank Regression.* Springer, New York, (1998).
9. SAS® software. http://www.sas.com
10. Van Den Wollenberg, A.L.: *Redundancy Analysis: an Alternative for Canonical Analysis.* Psychometrika Vol.42(2), 207-219 (1977).
11. Velu, R.P. *Reduced rank models with two sets of regressors.* Applied Statistics 40, 159-170, (1991).