Supplementary material to: A Copula formulation for multivariate latent Markov models

Alfonso Russo and Alessio Farcomeni

[*]Department of Economics and Finance, Tor Vergata University of Rome, Via Columbia,2, Rome, 00133, Italy.

*Corresponding author(s). E-mail(s): alfonso.russo@uniroma2.it; Contributing authors: alessio.farcomeni@uniroma2.it;

A. Details on simulation settings

We generate panels of dimensions $n = \{800, 1200\}$, $T_i = \{4, 6\}$, where outcomes are three-dimensional responses whose dependency structure is captured by a Frank copula with association parameter $\xi = \{0, 2\}$.

For the case of mixed margins, we simulate from a latent Markov model with $k = 2$ latent components. The first dimension Y_{it1} is a Gaussian response variable with identity link function; the corresponding modeling assumption is therefore

$$\mu_u := \mathbb{E}\left[Y_{it1} \mid U_{it} = u\right] = \alpha_{1u} + \boldsymbol{\beta}_1 \boldsymbol{X}_{it1}$$

$$Y_{it1} \mid U_{it} = u \sim \mathcal{N}(\mu_u, \sigma)$$

where σ is a common nuisance parameter.

A second outcome Y_{it2} is a binary outcome distributed as a Bernoulli, for which we specify a logistic regression model of the form:

$$\log\left(\frac{Pr(Y_{it1} = 1|U_{it} = u)}{Pr(Y_{it1} = 0|U_{it} = u)}\right) = \alpha_{1u} + \boldsymbol{\beta}_1 X_{it1}.$$

The last dimension Y_{it3} is a count variable with Poisson distribution and log link function, implying:

$$\log\left(E(Y_{it2}|U_{it} = u)\right) = \alpha_{2u} + \boldsymbol{\beta}_2 X_{it2}.$$

For each dimension, we generate two zero-centered Gaussian covariates with standard deviation equal to 0.1, and fix marginal parameters:

$$\boldsymbol{\alpha} = \begin{pmatrix} \text{-1.50} & 1.50 \\ \text{-1.50} & 1.50 \\ \text{-1.50} & 1.50 \end{pmatrix},$$

$$\boldsymbol{\beta} = \left\{ \begin{pmatrix} 1.00 & 1.00 \\ 1.00 & 1.00 \\ 1.00 & 1.00 \end{pmatrix} ; \begin{pmatrix} \text{-1.00} & 1.00 \\ \text{-1.00} & 1.00 \\ \text{-1.00} & 1.00 \end{pmatrix} \right\};$$

to obtain a total of sixteen simulation settings involving mixed margins. We fix

$$\boldsymbol{\Pi} = \begin{pmatrix} 0.80 & 0.20 \\ 0.20 & 0.80 \end{pmatrix},$$

and $\boldsymbol{\pi} = (0.50\ ,\ 0.50)$.

We consider also the case of fully discrete response variables in order to mimic the model that will be specified for the real data analysis. In these settings, two dimensions Y_{it1} and Y_{it2} are, respectively, a binary outcome and a Poission-distributed count variable, with link functions and resulting regression models that are identical to the case of mixed margins described above. We replace the Gaussian outcome with a third dimension Y_{it3} which is an ordered categorical variable taking values $\{0, 1, 2, 3, 4\}$, and specify a global logit parameterisation of the form:

$$\log \left(\frac{Pr(Y_{it3} \geq z | U_{it} = u)}{Pr(Y_{it3} < z | U_{it} = u)} \right) = \alpha_{3uz} + \boldsymbol{\beta}_3 X_{it3},$$

for $z = 0, 1, 2, 3$. The latter can be simply inverted to obtain the conditional distribution of Y_{it3}. We let the number of latent states vary as $k = \{2, 3\}$. We set

$$\boldsymbol{\beta} = \left\{ \begin{pmatrix} 1.00 & 1.00 \\ 1.00 & 1.00 \\ 1.00 & 1.00 \end{pmatrix} ; \begin{pmatrix} \text{-1.00} & 0.00 \\ \text{-1.00} & 0.00 \\ \text{-1.00} & 0.00 \end{pmatrix} \right\};$$

as coefficients modulating the effect of two dimension-specific Gaussian covariates with zero mean and standard deviation equal to 0.1. When $k = 2$ we fix

$$\boldsymbol{\alpha} = \begin{pmatrix} \text{-1.00} & 1.00 \\ \text{-1.00} & 1.00 \\ \text{-0.50} & 2.50 \\ \text{-1.50} & 1.50 \\ \text{-2.50} & 0.50 \end{pmatrix}.$$

When $k = 3$ we fix $\boldsymbol{\alpha}$ as

$$\begin{pmatrix} \text{-1.50} & 0.00 & 1.50 \\ \text{-1.50} & 0.00 & 1.50 \\ \text{-0.50} & 1.00 & 2.50 \\ \text{-1.50} & \text{-0.50} & 1.50 \\ \text{-2.50} & \text{-1.50} & 0.50 \end{pmatrix}.$$

When $k = 2$ we specify the same initial distribution and transition matrix as before, while when $k = 3$ we fix $\boldsymbol{\Pi}$ as

$$\begin{pmatrix} 0.80 & 0.10 & 0.10 \\ 0.10 & 0.80 & 0.10 \\ 0.10 & 0.10 & 0.80 \end{pmatrix},$$

and $\boldsymbol{\pi} = (0.40\ ,\ 0.30\ ,\ 0.30)\,.$

B. A multidimensional poverty analysis with continuous and discrete outcomes.

We here report results for a case study in which we analyse a different multivariate measure of poverty. We will now have a mix of continuous and discrete endpoints.

In order to allow for a direct comparison with the setup described in Section 6 of the main paper, we let $k = 4$ and specify a Frank copula to capture the dependency structure. We also, clearly, use the same households considered in Section 6 of the main paper, analysing a panel of $n = 1311$ units observed over four waves from 2014 to 2017.

Our new endpoint is a trivariate indicator of living conditions consisting of a continuous, a binary, and a count response; as follows. First, we still rely on equivalised disposable income as one indicator poverty, yet this time we do not threshold it. We standardize households' incomes by dividing each observed disposable income by the MAD of the wave, and the compute the difference with respect to the maximum income in order to obtain an increasing measure of poverty. We assume that conditionally on the latent variable, and covariates, this endpoint is Gaussian; and use an identity link function. We then consider work intensity as our second endpoint, where generalised employment rates are computed as described in the main paper. We then let $Y_{it2} = I(GER_{it} \leq 1)$, where $I(\cdot)$ is the indicator function. Finally, Y_{it3} is a count variable measuring active material deprivation items, as described in the main paper.

Estimates for the latent intercepts $\hat{\boldsymbol{\alpha}}$ are reported in Table 1, while Table 2 summarizes parameters' estimates for the $\hat{\boldsymbol{\beta}}$ coefficients capturing the effects of covariates. Table 3, finally, reports on the parameters modulating the latent structure. Estimated standard errors are reported in parentheses for all parameters. The association parameter is estimated as $\hat{\xi} = 0.301$, with 95% confidence interval $(0.091, 0.529)$.

The conclusions that can be made from additional analysis are closely related to the ones reported in the main paper, where we analysed three discrete endpoints. The endpoints in the main paper are in our opinion more appropriate, most importantly

$\hat{\boldsymbol{\alpha}}$	Y_{it1}	Y_{it2}	Y_{it3}
$\boldsymbol{k}_1$	9.782 (0.608)	0.280 (0.365)	-1.515 (0.167)
$\boldsymbol{k}_2$	13.553 (0.179)	0.949 (0.185)	-1.483 (0.208)
$\boldsymbol{k}_3$	15.180 (0.050)	1.578 (0.139)	-1.050 (0.115)
$\boldsymbol{k}_4$	16.128 (0.051)	1.460 (0.148)	0.484 (0.070)

Table 1: *EU SILC data. Estimated latent intercepts for a LM model with $k = 4$ latent masses and Frank copula. Standard errors in parentheses. The responses are assumed to be Gaussian, Bernoulli, and Poisson.*

$\hat{\boldsymbol{\beta}}$	NW	SI	NE	Time
Y_{it1}	0.049 (0.082)	0.617 (0.062)	0.222 (0.100)	-0.011 (0.008)
Y_{it2}	-0.009 (0.139)	0.061 (0.136)	0.034 (0.143)	0.077 (0.025)
Y_{it3}	0.039 (0.080)	0.597 (0.083)	0.118 (0.178)	-0.039 (0.014)

Table 2: *EU SILC data. Estimated marginal-specific regression coefficients for a LM model with $k = 4$ latent masses and Frank copula. Standard errors in parentheses. NW: North West, SI: South and Islands, NE: North East. Central Italy is reference category. The responses are assumed to be Gaussian, Bernoulli, and Poisson.*

the thresholding of the disposable income leads to a measure of monetary poverty. In this analysis we are on the other hand capturing heterogeneity among non-poor households in terms of income. This is testified for instance by the fact that we find one small mass clustering households with solid working and living standards, and richer in monetary terms. This is likely to result from the different nature of the considered monetary indicator included in the analysis, with the latent structure adapting to a right-tailed continuous variable.

The main conclusion is still that, since latent intercepts lack overall monotonicity and $\hat{\xi}$ is moderately low, even conditionally on the covariates there is some missmatch among the three living conditions indicators. For instance we observe households that are not deeply lacking in monetary terms, do not display strong material deprivation, yet have low levels of work intensity (state 2). Main interest for institutional authorities and policy makers might reside in households in latent state 4, who exhibit the overall worst living and working conditions: intercepts detect the strongest propensity to material deprivation, low work intensity, and largest monetary margins with respect to the richest households.

Finally, also in this analysis we observe high persistence of units in their latent states; and estimated regression coefficients describe the same picture of the main paper, with southern Italy and its islands having households at the highest risk of

$\hat{\boldsymbol{\pi}}$	$\boldsymbol{k}_1$	$\boldsymbol{k}_2$	$\boldsymbol{k}_3$	$\boldsymbol{k}_4$
	0.018 (0.005)	0.128 (0.016)	0463 (0.019)	0.391 (0.023)
$\hat{\boldsymbol{\Pi}}$	$\boldsymbol{k}_1$	$\boldsymbol{k}_2$	$\boldsymbol{k}_3$	$\boldsymbol{k}_4$
$\boldsymbol{k}_1$	0.725 (0.062)	0.230 (0.056	0.045 (0.032)	0.000 (0.001)
$\boldsymbol{k}_2$	0.038 (0.011)	0.943 (0.018)	0.011 (0.014)	0.009 (0.006)
$\boldsymbol{k}_3$	0.001 (0.001)	0.011 (0.005)	0.988 (0.006)	0.000 (0.001)
$\boldsymbol{k}_4$	0.000 (0.001)	0.001 (0.001)	0.011 (0.0182)	0.987 (0.012)

Table 3: *EU SILC data. Estimated initial and transition probabilities for a LM model with $k = 4$ latent masses and Frank copula. Standard errors in parentheses. The responses are assumed to be Gaussian, Bernoulli, and Poisson.*

material and monetary deprivation, compared to the North-East and the industrialised North-West where there might be more job opportunities than in other areas.