TEST
https://doi.org/10.1007/s11749-024-00919-9

ORIGINAL PAPER

A copula formulation for multivariate latent Markov models

Alfonso Russo[1] · **Alessio Farcomeni**[1]

Received: 15 June 2023 / Accepted: 3 January 2024

Abstract
We specify a general formulation for multivariate latent Markov models for panel data, where outcomes are possibly of mixed-type (categorical, discrete, continuous). Conditionally on a time-varying discrete latent variable and covariates, the joint distribution of outcomes simultaneously observed is expressed through a parametric copula. We therefore do not make any conditional independence assumption. The observed likelihood is maximized by means of an expectation–maximization algorithm. In a simulation study, we argue how modeling the residual contemporary dependence might be crucial in order to avoid bias in the parameter estimates. We illustrate through an original application to assessment of poverty through direct and indirect indicators in a cohort of Italian households.

Keywords Frank copula · Mixed responses · Panel data

Mathematics Subject Classification 62H99 · 62J12 · 62P20

1 Introduction

Latent Markov (LM) models for panel data provide a flexible framework to analyze univariate and multivariate responses (Zucchini and MacDonald 2009; Bartolucci et al. 2013, 2014). They can be seen as (multivariate) mixed models, based on a latent discrete random variable that captures the dynamic unobserved heterogeneity and is assumed to follow a first-order Markov chain with $k \in \mathcal{N}$ latent masses. Response variables are modeled with assumptions of *local independence*, i.e., each outcome is independent of its history and the history of other outcomes conditionally on covariates and the latent process. While many works deal with multivariate outcomes, a

✉ Alfonso Russo
alfonso.russo@uniroma2.it

Alessio Farcomeni
alessio.farcomeni@uniroma2.it

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

Published online: 07 February 2024

conditional independence assumption is additionally often used, i.e., each outcome is independent of the other outcomes at the same time point for the same unit, conditionally on covariates and the latent process (e.g., Bartolucci and Farcomeni 2015; DeRuiter et al. 2017; Dotto et al. 2018; Anderson et al. 2019a; Russo et al. 2022). This assumption can very often be found to be restrictive: In our experience, tests for conditional independence are regularly rejected. Models using this assumption can therefore easily lead to biased results. In our application we will investigate three ways of measuring poverty (based on income, work intensity, and material deprivation). The latent variable can be argued to summarize poverty levels. Conditional independence would imply for instance that poverty completely explains the relationship between income below a poverty threshold and work intensity, which is not tenable since one can expect complex phenomena to be in place for some households (e.g., undeclared work, inherited assets, etc.).

The conditional independence assumption can be relaxed using a multivariate normal formulation when all outcomes are Gaussian (e.g., Anderson et al. 2019; Punzo et al. 2021). Alternatives are explored for instance in Punzo et al. (2021), and Merlo et al. (2022) use a multivariate asymmetric Laplace to jointly model multivariate quantiles. For categorical outcomes, Bartolucci and Farcomeni (2009) formulate a class of LM models where the joint distribution is marginally parameterized through logits, log-odds-ratios, and higher-order loglinear interactions (which are anyway often assumed to be null). Orfanogiannaki and Karlis (2018) model multivariate count data using multivariate formulations of the Poisson distribution that also have Poisson marginals.

To the best of our knowledge in the LM framework, there are no available alternatives to the conditional independence assumption when modeling multivariate outcomes that can be a mix of count, binary, categorical, discrete, and continuous variables. In this paper, we develop such an extension. Specifically, the dependency structure is captured by a copula formulation. This allows us to flexibly specify a generalized linear model (Farcomeni 2015) for each of the outcomes and express the joint distribution as a function of the marginal cumulative density functions (CDF)s through a parametric copula, whose parameter captures the residual contemporary dependence among outcomes. Copula-based formulations for latent Markov models have been previously considered in Hardle et al. (2015), Martino et al. (2020), Orfanogiannaki and Karlis (2018), Otting and Karlis (2023), Otting et al. (2023). Our contribution with respect to previous works mainly resides in the fact that we allow for a mix of measurement levels in the response variables (which was mentioned in some works but never explicited), we include the effects of covariates, and do not restrict our formulation to two or three dimensions.

We illustrate using an original real data application on a panel of Italian households, where our outcome is three-dimensional (a binary indicator of equivalised disposable income below a poverty line, a count of items materially lacking, and a measure of work intensity). The main aim is to shed light into the surprising but common mismatch among these indicators. We show that part of the mismatch is clearly due to unobserved heterogeneity, that is, there actually exist a non-negligible share of households that might be poor only according to one or two indicators. We also show that conditionally on the latent variable, the dependence among contemporary outcomes clearly exists,

but it is not strong, indicating possible problems in the very definition of the indicators themselves.

The rest of the paper is as follows: In the next section, we give some background on copulas. Our class of copula-based latent Markov models is defined in Sect. 3, and inference described in Sect. 4. A simulation study is reported in Sect. 5. In Sect. 6, we describe more in detail our real data application and show results of the data analysis. Some concluding remarks are given in Sect. 7.

R functions with an implementation of our approach are available at https://github.com/aruxxo/LMcopula.

2 Background on copula modeling

We start by giving some background on copula functions. The main advantage of copulas for our purposes arises from the possibility of obtaining proper joint distributions by separating the specification of the margins from the modeling of the dependency structure. For a detailed review, see for instance Joe (2014), and Nikoloulopoulos (2013) for the case of discrete outcomes. A nice review of copula models for a mix of discrete and continuous outcomes is given in Wu et al. (2013).

A d-dimensional copula, $\boldsymbol{\mathcal{C}} : [0,1]^d \rightarrow [0,1]$, is a joint cumulative distribution function. We will specify parametric copulas $\boldsymbol{\mathcal{C}}(\boldsymbol{m} \mid \boldsymbol{\xi}) = \boldsymbol{\mathcal{C}}(\mathrm{m}_1, \ldots, \mathrm{m}_d \mid \boldsymbol{\xi})$, where an association parameter $\boldsymbol{\xi}$ captures association among marginals $\boldsymbol{m}$.

Consider a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ with continuous and increasing margins $F_{Y_j}(Y_j \mid \boldsymbol{\psi}_j)$ for some parameter $\boldsymbol{\psi}_j$, $j = 1, \ldots, d$. We assume

$$\begin{aligned} \boldsymbol{\mathcal{C}}(\mathrm{m}_1, \ldots, \mathrm{m}_d \mid \boldsymbol{\xi}) &= \mathbb{P}\left[F_{Y_1}(Y_1 \mid \boldsymbol{\psi}_1) \leq \mathrm{m}_1, \ldots, F_{Y_d}(Y_d \mid \boldsymbol{\psi}_d) \leq \mathrm{m}_d\right] \\ &= F_{\boldsymbol{Y}}\left[F_{Y_1}^{-1}(\mathrm{m}_1 \mid \boldsymbol{\psi}_1), \ldots, F_{Y_d}^{-1}(\mathrm{m}_d \mid \boldsymbol{\psi}_d)\right] \end{aligned} \tag{1}$$

and therefore

$$F(y_1, \ldots, y_d) = \boldsymbol{\mathcal{C}}\left[F_1(y_1 \mid \boldsymbol{\psi}_1), \ldots, F_d(y_d \mid \boldsymbol{\psi}_d) \mid \boldsymbol{\xi}\right] \tag{2}$$

for all $y_i \in [-\infty, +\infty]$, $i = 1, \ldots, d$. The density takes the usual form

$$\boldsymbol{c}(\mathrm{m}_1, \ldots, \mathrm{m}_d \mid \xi) = \frac{\partial^d \boldsymbol{\mathcal{C}}(\mathrm{m}_1, \ldots, \mathrm{m}_d \mid \boldsymbol{\xi})}{\partial \mathrm{m}_1, \ldots, \mathrm{m}_d} \prod_{r=1}^{d} f_{Y_r}(y_r \mid \boldsymbol{\psi}_r) \tag{3}$$

where $\mathrm{m}_i = F_{Y_i}(y_i \mid \boldsymbol{\psi}_i) \overset{d}{=} U(0,1)$, $\forall\, i = 1, \ldots, d$.

There are several possible choices. Gaussian copulas, and Student's T counterparts, work well when modeling multivariate data with symmetric dependency patterns. Archimedean copulas are also very popular, being both flexible and convenient due to the existence of a closed form expression. Special cases that will be considered in this work are Frank, Clayton, Joe, and Gumbel copulas. Joe and Gumbel models capture asymmetric dependency structures, exhibiting stronger forms of positive dependence.

Table 1 Families of Archimedean copula models for the d-variate case and the support of the relative association parameter ξ

Family	$\boldsymbol{C}(\mathrm{m}_1, \ldots, \mathrm{m}_d \mid \xi)$	Support of ξ
Frank	$-\xi^{-1} \log \left\{ 1 + \left[\exp(-\xi \mathrm{m}_1) - 1\right] \cdot \ldots \cdot \left[\exp(-\xi \mathrm{m}_d) - 1\right] / \left[\exp(-\xi) - 1\right]^{d-1} \right\}$	$\mathbb{R}/\{0\}$
Clayton	$\left[\mathrm{m}_1^{-\xi} + \ldots + \mathrm{m}_d^{-\xi} - (d-1) \right]^{-1/\xi}$	$[-1, \infty)/\{0\}$
Joe	$1 - \left\{ 1 - [1 - (1 - \mathrm{m}_1)^{\xi}] \cdot \ldots \cdot [1 - (1 - \mathrm{m}_d)^{\xi}] \right\}^{1/\xi}$	$(1, \infty)$
Gumbel	$\exp \left\{ - \left[(-\log \mathrm{m}_1)^{\xi} + (-\log \mathrm{m}_2)^{\xi} + \ldots + (-\log \mathrm{m}_d)^{\xi} \right]^{1/\xi} \right\}$	$(1, \infty)$

The Clayton copula is also asymmetric, and better captures negative association. The Frank copula is, finally, comprehensive. Functional forms and domains for the parameter ξ of the copula models that will be used in this work are reported in Table 1; a comprehensive guide can be found in Nelsen (2006).

The brief introduction we have given has been restricted to the classical case of continuous margins. Cases in which one, some, or even all margins are not continuous require some more care. Some additional challenges arise also from the computational perspective. First of all, whenever all marginals are discrete, the copula might lack uniqueness, and some of the properties derived for the continuous case may not hold. Discrete random variables violate in general the assumption that margins are uniformly distributed. The discussion in Genest and Nevslehová (2007) clarifies that copula models are still valid for discrete outcomes, since Eq. (2) is still a correct representation of the joint distribution function. Moreover, inference is still possible with fully parametric likelihood-based methods; see, for example, Henn (2022). Nikoloulopoulos (2013) provides not only a review of parametric copulas with discrete margins, but also an extensive discussion on properties. Trivedi and Zimmer (2017) show that the identification/uniqueness issue attenuates when the model has a regression structure and responses are count data. They note though that asymmetric copulas (e.g., Clayton and Gumbel) might require large data sets to mitigate the issue of uniqueness.

In the following, we will give a specification of a copula regression model that is valid under any general combination of continuous and discrete margins and discuss the details where appropriate.

3 The model

Let Y_{itr} denote the rth endpoint of interest, measured for the ith subject at time t; with $r = 1, \ldots, d$, $t = 1, \ldots, T_i$, and $i = 1, \ldots, n$. We also assume there exists a discrete unidimensional latent variable U_{it} with support in $\{1, \ldots, k\}$, where k is known. The latent variable is assumed to evolve over time according to a homogeneous first-order Markov chain, with initial probabilities $\Pr(U_{i1} = u) = \pi_u$, collected in a vector $\boldsymbol{\pi}$, and transition probabilities $\Pr(U_{it} = v | U_{i,t-1} = u) = \pi_{uv}$, collected in a transition matrix $\boldsymbol{\Pi}$.

For the marginal distributions, conditionally on U_{it}, we assume that Y_{itr} follows a natural exponential family

$$p(Y_{itr}|U_{it} = u, \boldsymbol{\eta}, \lambda_r) = \exp\left\{(Y_{itr}\eta_{utr} - q(\eta_{utr}))/(a(\lambda_r) - b(Y_{itr}, \lambda_r))\right\}, \quad (4)$$

where $p(\cdot)$ can either be a probability density function (PDF) or probability mass function (PMF); and functions $a(\cdot)$, $b(\cdot)$, and $q(\cdot)$ are known. The parameter of interest is η_{utr}, while λ_r is a nuisance parameter. The current specification relies on homogeneity assumptions on the nuisance parameters λ_r, that is, that the nuisance parameters do not vary with the latent variable. This provides usually sufficiently good fit, but it can easily be relaxed, at the price of an increase in the number of free parameters. For instance, for conditional Gaussian outcomes, it might happen that latent states associated with large means are also associated with a larger variance, in which case one should assume a state-specific variance σ_u^2.

A regression model will usually be specified after reparameterization through a known link function $h(\cdot)$, as

$$h(\eta_{utr}) = \alpha_{ru} + \boldsymbol{\beta}_r' \boldsymbol{X}_{itr}, \quad (5)$$

where $\boldsymbol{X}_{itr}$ is a vector of time-subject-outcome specific covariates. In (5), only the intercept depends on U_{it}, as customary with latent Markov models. Also this assumption can be easily relaxed.

Usually, a conditional independence assumption is put forward so that Y_{itr} is independent of Y_{its} for all $s \neq r$, conditionally on U_{it}. We relax this assumption through a copula model. Specifically, we assume that $\boldsymbol{Y}_{it}$, conditionally on $U_{it} = u$, has joint CDF

$$\boldsymbol{C}(\mathrm{m}_{it1u}, \ldots, \mathrm{m}_{itdu} \mid \xi) \quad (6)$$

with $\mathrm{m}_{itru} = F_{Y_r}(Y_{itr} \mid U_{it} = u, \boldsymbol{\eta}, \lambda_r)$. Details on the joint PMF/PDF are given in Sect. 4. Local independence and part of the conditional independence assumption still holds as we assume $\boldsymbol{Y}_{it}$ to be independent of $\boldsymbol{Y}_{is}$, for $s < t$, conditionally on U_{it}. Clearly, Y_{itr} is also assumed to be independent of Y_{jsh} for $i \neq j$ as usual.

4 Inference

In this section, we show how to estimate the parameters of the model proposed in the previous section.

4.1 Observed likelihood

We first present the functional form of the joint PMF/PDF. Without loss of generality we assume that among the d margins, the first d_1 correspond to continuous variables,

and the remaining are discrete. Define

$$\tilde{\boldsymbol{C}}(\boldsymbol{m}_{itu} \mid \xi) = \frac{\partial^{d_1}}{\partial m_{it1u}, \ldots, m_{itd_1u}} \boldsymbol{\mathcal{C}}(m_{it1u}, \ldots, m_{itdu} \mid \xi) \tag{7}$$

which is based on partial derivatives only with respect to the first d_1 margins. The joint density function is obtained by taking the first differences of (7) with respect to the remaining $d - d_1$ margins:

$$\begin{aligned} \tilde{\boldsymbol{c}}(\boldsymbol{Y}_{it} \mid \xi, U_{it} = u) = & \prod_{r=1}^{d_1} f_r(y_{itr} \mid U_{it} = u) \\ & \times \sum_{j_{d_1+1}=0}^{1} \cdots \sum_{j_d=0}^{1} (-1)^{j_{d_1+1}+\cdots+j_d} \tilde{\boldsymbol{C}} \left(\left\{ m_{it1u}, \ldots, m_{itd_1u} \right\}, \right. \\ & \left. \times \quad \left\{ m_{itd_1+1u}^{j_{d_1+1}}, \ldots, m_{itdu}^{j_d} \right\} \mid \xi \right) \end{aligned} \tag{8}$$

where for the discrete outcomes $s = d_1 + 1, \ldots, d$ $m_{itsu}^1 = F_{Y_s}(Y_{its} \mid \boldsymbol{\psi}_s, U_{it} = u)$, and $m_{itsu}^0 = F_{Y_s}(Y_{its}^- \mid \boldsymbol{\psi}_s, U_{it} = u)$ is the left-hand limit at Y_{its}. See also Zilko and Kurowicka (2016) on (8).

The observed likelihood cannot be computed directly, as it would involve a telescopic sum over all possible configurations of the latent variable. A simple forward recursion can anyway be used. The computational complexity is linear in $\sum_i T_i$, and the recursion allows to exactly evaluate the observed likelihood. Define $a_{it}(u) = f(Y_{i1}, \ldots, Y_{it}, U_{it} = u)$. Let, by definition,

$$a_{i1}(u) = \pi_u \, \tilde{\boldsymbol{c}}(\boldsymbol{Y}_{i1} \mid \xi, U_{it} = u).$$

It is possible to show with some algebra that if $T_i > 1$, for $t = 2, \ldots, T_i$,

$$a_{it}(u) = \tilde{\boldsymbol{c}}(\boldsymbol{Y}_{it} \mid \xi, U_{it} = u) \sum_{h=1}^{k} a_{i,t-1}(j) \pi_{hu}.$$

The recursion shall be repeated for each $i = 1, \ldots, n$. By definition of $a_{it}(u)$, the observed log-likelihood at the parameter $\boldsymbol{\theta}$ (which is a short hand notation for the vector of free parameters involved in the model) is then

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left(\sum_{u=1}^{k} a_{iT_i}(u) \right).$$

A backward recursion is also implemented, as it is required within the E-step of the EM algorithm described below. Define $s_{it}(u) = f(Y_{i,t+1}, \ldots, Y_{i,T_i} | U_{it} = u)$. Clearly,

$s_{iT_i}(u) = 1$. The recursion then proceeds for $t = T_i - 1, \ldots, 1$ by setting

$$s_{it}(u) = \sum_{v=1}^{k} \pi_{uv}\, s_{i,t+1}(v)\, \tilde{\boldsymbol{c}}(\boldsymbol{Y}_{it} \mid \xi, U_{it} = u).$$

4.2 Expectation–maximization algorithm

For the maximization of $\ell(\boldsymbol{\theta})$, an EM algorithm can be then implemented. In order to limit the possibility of converging to local optima, a multi-start strategy is recommended.

The forward and backward recursions are repeated before each expectation step, using the current parameter values. In order to derive the EM algorithm, we introduce the log-likelihood of the complete data

$$\begin{aligned}\ell^*(\boldsymbol{\theta}) = &\sum_{i=1}^{n}\sum_{u=1}^{k} w_{i1u}\log(\pi_u) + \sum_{i=1}^{n}\sum_{t=2}^{T_i}\sum_{u=1}^{k}\sum_{v=1}^{k} z_{ituv}\log(\pi_{uv}) \\ &+ \sum_{i=1}^{n}\sum_{t=1}^{T_i}\sum_{u=1}^{k} w_{itu}\log\left[\tilde{\boldsymbol{c}}(\boldsymbol{Y}_{it} \mid \xi, U_{it} = u)\right]\end{aligned}$$

where $w_{itu} = 1$ if subject i is in state u at time t, and it is zero otherwise; while $z_{ituv} = 1$ if subject i at time t moves to state v from state u, and it is zero otherwise.

The EM algorithm iterates an expectation and a maximization step, until convergence. The *E-step* amounts to computing the conditional expectation of $\ell^*(\boldsymbol{\theta})$ given the data and the current value of the parameters. This is actually equivalent to plug-in of the conditional expected value of the variables w_{itu} and z_{ituv}, which are available in closed form as:

$$\begin{aligned}\mathbb{E}\left[w_{itu} \mid \boldsymbol{Y}\right] &= \frac{a_{it}(u)s_{it}(u)}{\sum_{u=1}^{k} a_{iT_i}(u)} \\ \mathbb{E}\left[z_{ituv} \mid \boldsymbol{Y}\right] &= \pi_{uv}\frac{a_{it}(u)s_{i,t+1}(v)}{\sum_{u=1}^{k} a_{it}(u)s_{it}(u)} \times \tilde{\boldsymbol{c}}(\boldsymbol{Y}_{i,t+1} \mid \xi, U_{it} = v).\end{aligned}$$

At the *M-step*, the parameters are updated by maximization of the expected value of $\ell^*(\boldsymbol{\theta})$ calculated in the previous step. Closed form expressions are available for the parameters of the latent distribution:

$$\begin{aligned}\hat{\pi}_u &= \frac{\sum_{i=1}^{n}\mathbb{E}\left[w_{i1u} \mid \boldsymbol{Y}\right]}{\sum_{i=1}^{n}\sum_{h=1}^{k}\mathbb{E}\left[w_{i1h} \mid \boldsymbol{Y}\right]} \\ \hat{\pi}_{uv} &= \frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i-1}\mathbb{E}\left[z_{ituv} \mid \boldsymbol{Y}\right]}{\sum_{i=1}^{n}\sum_{t=1}^{T_i-1}\sum_{v=1}^{k}\mathbb{E}\left[z_{ituv} \mid \boldsymbol{Y}\right]}\end{aligned}$$

For the regression parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of the marginal regression models, we implement d separate optimization steps, each involving one Newton–Raphson iteration. The objective function corresponds to

$$\sum_{i=1}^{n}\sum_{t=1}^{T_i}\sum_{u=1}^{k}\mathbb{E}\left[w_{itu} \mid \boldsymbol{Y}\right]\log\left\{\tilde{\boldsymbol{c}}(\boldsymbol{Y}_{it} \mid \xi, U_{it}=u)\right\}. \tag{9}$$

Following Anderson et al. (2019), we tackle label switching by reordering parameters at the end of each M-step so that $\alpha_{11} \leq \cdots \leq \alpha_{1k}$. The first dimension of $\boldsymbol{\alpha}$ is chosen for reference without loss of generality. The copula parameter ξ is finally updated by optimizing (9) through a univariate numerical optimization procedure.

5 Simulations

The modeling approach and inference have been so far introduced in complete generality. In order to assess the performance of the proposed approach, we conduct an extensive simulation study, where we consider scenarios involving either mixed margins, and fully discrete margins as in the empirical application.

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. A second outcome Y_{it2} is a binary outcome distributed as a Bernoulli, for which we specify a logistic regression model. The last dimension Y_{it3} is a count variable with Poisson distribution and log link function. For each dimension, we generate two zero-centered Gaussian covariates and have two sets of β parameters, to obtain a total of sixteen simulation settings involving mixed margins. Full details are given in the accompanying Web "Appendix."

We consider also the case of fully discrete response variables in order to mimick 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 Poisson-distributed count variable. A third dimension Y_{it3} is an ordered categorical variable taking values $\{0, 1, 2, 3, 4\}$, with global logit parameterization 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 $k = \{2, 3\}$, and the remaining details can be found in the accompanying Web Appendix.

For each scenario, we generate data, estimate a LM model based on a copula formulation as proposed, and a classical LM model under the full conditional independence assumption. We repeat the operation $B = 250$ times and evaluate model performance

through computation of the average mean squared error (MSE) for each group of parameters.

Tabel 2 reports results for the settings involving mixed responses. For the cases concerning fully discrete margins, we report in Table 3 results corresponding to a latent Markov model with $k = 2$ latent states, while Table 4 summarizes settings in which $k = 3$. We denote with LMC the results for a model using a copula formulation, and LMI the results corresponding to a full conditional independence assumption.

It is clear that whenever the classical LMI model is well-specified (that is, data are generated with $\xi = 0$), there is little loss of information in the use of an additional parameter for the LMC model. On the other hand, when the true association parameter is as low as $\xi = 2$, corresponding to $\tau = 0.21$ in a continuous-marginals model, the LMI approach clearly is sub-optimal and brings about a bias in estimation of both manifest and latent parameters.

Additionally, we study simulation scenarios in which we test the performance of the Bayesian Information Criterion (BIC) in recovering the correct copula structure. We generate data through a Frank copula with $\xi = 2$, with three mixed outcomes (Gaussian, Bernoulli, Poisson) as described above. We then estimate our model with four different assumptions on the copula structure (Frank, Clayton, Gumbel, and Joe) and select by minimizing BIC. It can be seen from Table 5 that BIC consistently leads to select the data generating Frank copula. We finally note that the computational cost associated with a single replicate varies with the complexity of the data generating process, the sample size, and especially the number of latent states; average time was around two hours on a 3.49 GHz Apple M2 Pro with 32 GB RAM. Our `R` code is not optimized and does not make use of `C` or `Fortran` code.

6 An analysis of multidimensional aspects of poverty in Europe

In this section, we describe our motivating example. We explore a multivariate measure of poverty in the manifest model, mostly to disentangle unobserved heterogeneity from measurement error. Use of copula models in this area of research is not new, see, e.g., Hohberg et al. (2021).

We here measure poverty through the following two indicators: income levels, and material deprivation. The first is the most natural and frequently used measure, but it can be argued that it is an indirect one. Material deprivation indicators aim at obtaining a direct measurement of poverty by assessing living conditions. A detailed motivation and definition can be found in Townsend (1987). It is generally recommended to measure both income and deprivation to identify the poor (Townsend and Gordon 1991); but a well-known issue in the literature is the surprisingly common occurrence of mismatch (e.g., Whelan et al. 2004, and references therein and thereof).

In this work, we give some insights into the nature of this mismatch and report about a further surprising mismatch: the one between work intensity and poverty. It will be seen indeed that non-negligible fractions of households at very low work intensity are *not* poor and *vice versa*.

We have data from the most recent (2014–2017, $T_i = T = 4$) panel component of the European Union Survey on Income and Living Conditions (EU SILC). We adopt

Table 2 Simulation results for data generated from a Latent Markov model with $k = 2$ latent masses and associated mixed (Gaussian and discrete) response variables having joint distribution modeled via a Frank copula with parameter ξ.

MSE				$\hat{\boldsymbol{\alpha}}$		$\hat{\boldsymbol{\beta}}$		$\hat{\sigma}$		$\hat{\boldsymbol{\xi}}$		$\hat{\boldsymbol{\pi}}$		$\hat{\boldsymbol{\Pi}}$	
n	T	$\boldsymbol{\beta}$	ξ	LMC	LMI	LMC	LMI	LMC	LMI	LMC	LMI	LMC	LMI	LMC	LMI
800	4	$(\mathbf{1}, \mathbf{1})$	2	0.045	0.233	0.271	0.311	0.013	0.213	0.111	2.000	0.019	0.020	0.012	0.014
800	4	$(-\mathbf{1}, \mathbf{0})$	2	0.047	0.231	0.274	0.324	0.014	0.213	0.118	2.000	0.019	0.020	0.012	0.014
1200	4	$(\mathbf{1}, \mathbf{1})$	2	0.039	0.228	0.241	0.255	0.012	0.213	0.089	2.000	0.015	0.014	0.010	0.011
1200	4	$(-\mathbf{1}, \mathbf{0})$	2	0.038	0.230	0.254	0.262	0.012	0.215	0.090	2.000	0.015	0.015	0.010	0.011
800	6	$(\mathbf{1}, \mathbf{1})$	2	0.036	0.227	0.247	0.259	0.011	0.214	0.075	2.000	0.018	0.018	0.009	0.011
800	6	$(-\mathbf{1}, \mathbf{0})$	2	0.037	0.228	0.240	0.263	0.011	0.214	0.080	2.000	0.017	0.018	0.009	0.010
1200	6	$(\mathbf{1}, \mathbf{1})$	2	0.030	0.229	0.198	0.202	0.010	0.215	0.069	2.000	0.016	0.016	0.008	0.009
1200	6	$(-\mathbf{1}, \mathbf{0})$	2	0.030	0.230	0.189	0.192	0.009	0.215	0.075	2.000	0.016	0.017	0.008	0.009
800	4	$(\mathbf{1}, \mathbf{1})$	0	0.047	0.231	0.295	0.307	0.012	0.219	0.019	0.000	0.095	0.019	0.011	0.011
800	4	$(-\mathbf{1}, \mathbf{0})$	0	0.046	0.231	0.298	0.302	0.012	0.218	0.034	0.000	0.019	0.018	0.012	0.012
1200	4	$(\mathbf{1}, \mathbf{1})$	0	0.036	0.228	0.253	0.254	0.009	0.219	0.001	0.000	0.015	0.015	0.010	0.010
1200	4	$(-\mathbf{1}, \mathbf{0})$	0	0.040	0.228	0.260	0.253	0.010	0.219	0.011	0.000	0.015	0.015	0.010	0.010
800	6	$(\mathbf{1}, \mathbf{1})$	0	0.038	0.229	0.255	0.255	0.011	0.219	0.001	0.000	0.017	0.017	0.008	0.008
800	6	$(-\mathbf{1}, \mathbf{0})$	0	0.039	0.230	0.245	0.242	0.010	0.219	0.001	0.000	0.017	0.017	0.009	0.009
1200	6	$(\mathbf{1}, \mathbf{1})$	0	0.030	0.228	0.196	0.189	0.008	0.217	0.011	0.000	0.016	0.016	0.008	0.008
1200	6	$(-\mathbf{1}, \mathbf{0})$	0	0.030	0.229	0.191	0.185	0.009	0.217	0.016	0.000	0.016	0.016	0.008	0.008

We report MSE for each group of parameters. LMC refers to our proposed framework, LMI to an LM model assuming full conditional independence. Results are based on $B = 250$ replicates

Table 3 Simulation results for data generated from a Latent Markov model with $k = 2$ latent masses and associated response variables having joint distribution modeled via a Frank copula with parameter ξ

MSE				$\hat{\boldsymbol{\alpha}}$		$\hat{\boldsymbol{\beta}}$		$\hat{\xi}$		$\hat{\boldsymbol{\pi}}$		$\hat{\boldsymbol{\Pi}}$	
n	T	$\boldsymbol{\beta}$	ξ	LMC	LMI	LMC	LMI	LMC	LMI	LMC	LMI	LMC	LMI
800	4	$(\mathbf{1}, \mathbf{1})$	2	0.091	0.103	0.315	0.337	0.161	2.000	0.022	0.024	0.016	0.054
800	4	$(-\mathbf{1}, \mathbf{0})$	2	0.091	0.106	0.309	0.321	0.165	2.000	0.024	0.026	0.015	0.056
1200	4	$(\mathbf{1}, \mathbf{1})$	2	0.070	0.089	0.268	0.298	0.123	2.000	0.017	0.017	0.013	0.054
1200	4	$(-\mathbf{1}, \mathbf{0})$	2	0.055	0.069	0.197	0.224	0.083	2.000	0.018	0.020	0.010	0.050
800	6	$(\mathbf{1}, \mathbf{1})$	2	0.068	0.082	0.259	0.295	0.123	2.000	0.020	0.021	0.012	0.050
800	6	$(-\mathbf{1}, \mathbf{0})$	2	0.067	0.085	0.256	0.295	0.114	2.000	0.018	0.021	0.011	0.050
1200	6	$(\mathbf{1}, \mathbf{1})$	2	0.052	0.069	0.201	0.227	0.081	2.000	0.019	0.022	0.010	0.050
1200	6	$(-\mathbf{1}, \mathbf{0})$	2	0.055	0.069	0.197	0.224	0.083	2.000	0.018	0.020	0.010	0.050
800	4	$(\mathbf{1}, \mathbf{1})$	0	0.075	0.075	0.329	0.328	0.001	0.000	0.020	0.020	0.014	0.014
800	4	$(-\mathbf{1}, \mathbf{0})$	0	0.075	0.075	0.324	0.333	0.001	0.000	0.020	0.020	0.013	0.014
1200	4	$(\mathbf{1}, \mathbf{1})$	0	0.061	0.061	0.295	0.293	0.001	0.000	0.017	0.017	0.011	0.011
1200	4	$(-\mathbf{1}, \mathbf{0})$	0	0.062	0.061	0.272	0.276	0.000	0.000	0.017	0.017	0.011	0.011
800	6	$(\mathbf{1}, \mathbf{1})$	0	0.062	0.062	0.276	0.275	0.010	0.000	0.019	0.019	0.010	0.010
800	6	$(-\mathbf{1}, \mathbf{0})$	0	0.063	0.064	0.275	0.279	0.010	0.000	0.019	0.019	0.010	0.010
1200	6	$(\mathbf{1}, \mathbf{1})$	0	0.049	0.050	0.213	0.216	0.001	0.000	0.019	0.019	0.008	0.008
1200	6	$(-\mathbf{1}, \mathbf{0})$	0	0.062	0.061	0.272	0.276	0.000	0.000	0.017	0.017	0.011	0.011

We report MSE for each group of parameters. LMC refers to our proposed framework, LMI to an LM model assuming full conditional independence. Results are based on $B = 250$ replicates

Table 4 Simulation results for data generated from a Latent Markov model with $k = 3$ latent masses and associated response variables having joint distribution modeled via a Frank copula with parameter ξ

MSE				$\hat{\boldsymbol{\alpha}}$		$\hat{\boldsymbol{\beta}}$		$\hat{\xi}$		$\hat{\boldsymbol{\pi}}$		$\hat{\boldsymbol{\Pi}}$	
n	T	$\boldsymbol{\beta}$	ξ	LMC	LMI	LMC	LMI	LMC	LMI	LMC	LMI	LMC	LMI
800	4	$(\mathbf{1}, \mathbf{1})$	2	0.128	0.152	0.311	0.389	0.174	2.000	0.046	0.031	0.034	0.129
800	4	$(-\mathbf{1}, \mathbf{0})$	2	0.122	0.148	0.316	0.367	0.161	2.000	0.041	0.032	0.034	0.129
1200	4	$(\mathbf{1}, \mathbf{1})$	2	0.107	0.142	0.254	0.303	0.126	2.000	0.036	0.030	0.030	0.133
1200	4	$(-\mathbf{1}, \mathbf{0})$	2	0.107	0.140	0.251	0.304	0.140	2.000	0.036	0.030	0.029	0.133
800	6	$(\mathbf{1}, \mathbf{1})$	2	0.094	0.127	0.251	0.322	0.126	2.000	0.037	0.033	0.025	0.125
800	6	$(-\mathbf{1}, \mathbf{0})$	2	0.095	0.129	0.260	0.323	0.124	2.000	0.037	0.033	0.026	0.125
1200	6	$(\mathbf{1}, \mathbf{1})$	2	0.077	0.110	0.220	0.263	0.101	2.000	0.025	0.026	0.018	0.126
1200	6	$(-\mathbf{1}, \mathbf{0})$	2	0.079	0.111	0.217	0.253	0.099	2.000	0.028	0.027	0.018	0.126
800	4	$(\mathbf{1}, \mathbf{1})$	0	0.101	0.102	0.352	0.353	0.016	0.000	0.031	0.030	0.026	0.025
800	4	$(-\mathbf{1}, \mathbf{0})$	0	0.102	0.102	0.358	0.360	0.017	0.000	0.029	0.029	0.027	0.024
1200	4	$(\mathbf{1}, \mathbf{1})$	0	0.085	0.085	0.272	0.270	0.031	0.000	0.026	0.025	0.023	0.021
1200	4	$(-\mathbf{1}, \mathbf{0})$	0	0.084	0.084	0.267	0.272	0.026	0.000	0.027	0.026	0.022	0.021
800	6	$(\mathbf{1}, \mathbf{1})$	0	0.080	0.081	0.283	0.284	0.012	0.000	0.027	0.026	0.019	0.018
800	6	$(-\mathbf{1}, \mathbf{0})$	0	0.083	0.083	0.276	0.274	0.005	0.000	0.025	0.026	0.018	0.018
1200	6	$(\mathbf{1}, \mathbf{1})$	0	0.066	0.066	0.223	0.226	0.020	0.000	0.021	0.021	0.014	0.014
1200	6	$(-\mathbf{1}, \mathbf{0})$	0	0.066	0.066	0.219	0.221	0.002	0.000	0.021	0.022	0.014	0.014

We report MSE for each group of parameters. LMC refers to our proposed framework, LMI to an LM model assuming full conditional independence. Results are based on $B = 250$ replicates

Table 5 Simulation results for data generated from Latent Markov models with $k = \{2, 3\}$ latent masses, three response variables (Gaussian, Bernoulli, Poisson), and Frank copula with parameter $\xi = 2$

n	T	$\boldsymbol{\beta}$	k	$\mathbb{P}(\widehat{\mathcal{C}^* = \mathcal{C}_x})$				Median(BIC)			
				$\mathcal{C}$ Frank	$\mathcal{C}$ Clayton	$\mathcal{C}$ Gumbel	$\mathcal{C}_{\text{Joe}}$	$\mathcal{C}_{\text{Frank}}$	$\mathcal{C}$ Clayton	$\mathcal{C}_{\text{Gumbel}}$	$\mathcal{C}_{\text{Joe}}$
800	4	$(\mathbf{1}, \mathbf{1})$	2	1.00	0.00	0.00	0.00	23,722.96	23,817.11	23,794.89	23,880.43
800	4	$(-\mathbf{1}, \mathbf{0})$	2	1.00	0.00	0.00	0.00	23,720.20	23,820.55	23,801.79	23,892.50
1200	4	$(\mathbf{1}, \mathbf{1})$	2	1.00	0.00	0.00	0.00	35,452.78	35,603.85	35,579.91	35,723.19
1200	4	$(-\mathbf{1}, \mathbf{0})$	2	1.00	0.00	0.00	0.00	35,438.32	35,594.87	35,563.49	35,717.81
800	6	$(\mathbf{1}, \mathbf{1})$	2	1.00	0.00	0.00	0.00	35,577.56	35,734.63	35,703.31	35,844.51
800	6	$(-\mathbf{1}, \mathbf{0})$	2	1.00	0.00	0.00	0.00	35,579.66	35,743.60	35,698.39	35,837.89
1200	6	$(\mathbf{1}, \mathbf{1})$	2	1.00	0.00	0.00	0.00	53,094.97	53,326.82	53,269.49	53,483.50
1200	6	$(-\mathbf{1}, \mathbf{0})$	2	1.00	0.00	0.00	0.00	53,083.78	53,330.06	53,263.33	53,488.72
800	4	$(\mathbf{1}, \mathbf{1})$	3	1.00	0.00	0.00	0.00	23,537.12	23,590.47	23,582.64	23,634.30
800	4	$(-\mathbf{1}, \mathbf{0})$	3	1.00	0.00	0.00	0.00	23,546.99	23,597.27	23,597.01	23,648.26
1200	4	$(\mathbf{1}, \mathbf{1})$	3	1.00	0.00	0.00	0.00	35281.72	35,364.22	35,349.38	35,432.09
1200	4	$(-\mathbf{1}, \mathbf{0})$	3	1.00	0.00	0.00	0.00	35,270.00	35,356.02	35,334.29	35,412.88
800	6	$(\mathbf{1}, \mathbf{1})$	3	1.00	0.00	0.00	0.00	35,192.26	35,286.73	35,269.12	35,357.88
800	6	$(-\mathbf{1}, \mathbf{0})$	3	1.00	0.00	0.00	0.00	35,177.79	35,275.21	35,255.63	35,345.85
1200	6	$(\mathbf{1}, \mathbf{1})$	3	1.00	0.00	0.00	0.00	52,719.42	52,860.02	52,822.75	52,957.72
1200	6	$(-\mathbf{1}, \mathbf{0})$	3	1.00	0.00	0.00	0.00	52,693.43	52,837.51	52,815.18	52,940.47

For each setting, we report the probability of selecting a Clayton, Gumbel, Joe, or Frank copula via Bayesian information criterion (BIC), and median BIC. Results are based on $B = 250$ replicates

households as unit of the analysis. We restrict the sample to Italian households having at least two members, and whose housing status (ownership or rental) was constant over the considered period, finally having a panel of $n = 1311$ households.

First, we consider income, as an obvious and very well-established indicator. We focus on equivalised disposable income, defined as "the total income of a household, after tax and other deductions, that is available for spending or saving, divided by the number of household members converted into equalised adults". Traditional measures of poverty require checking whether an individual's income is below a certain threshold considered to be the minimum required for a reasonable standard of living. This poses however some challenges, being the choice of poverty lines arbitrary and debatable. For the purpose of international comparability, a standard poverty line is often considered to be US$1.90 per person per day in purchase parity power (PPP). Since we restrict to Italian data, we here rely on the European convention, according to which the poverty threshold is set at 60% of the equivalised median national income. Our first outcome Y_{it1}, therefore, is a binary indicator of the household equivalised disposable income being below 60% of the national median.

Secondly, we consider material deprivation. EUROSTAT traditionally measures material deprivation through a counting approach of the number of items that an household is lacking (in a given year) within the following list:

1. Ability to keep the house adequately warm
2. Ability to afford 1-week annual holiday
3. Ability to afford a meal (meat, chicken, fish or eqv.) every other day
4. Ability to pay for unexpected expenses
5. Ability to afford a telephone
6. Ability to afford a color television
7. Ability to afford a washing machine
8. Ability to afford a car for private use
9. Avoid arrears on mortgage, rent, utility bills or loans.

Our second outcome Y_{it2}, therefore, is a count. Finally we consider work intensity, whose measurement has become complex due to very heterogeneous work spells and conditions in the population. According to the standards set by the International Labour Office the employment rate is simply the proportion of working-age people who have been working for at least an hour in a reference week. However, if we considered only this extensive margin, we would have people working exactly one hour per week being counted as those working forty hours. We rely here on Brandolini and Viviano (2016), who propose a generalized employment rate (GER) measure. We define the GER measure as $\text{GER}_{it} = \sum_{j=1}^{N_i} \left(w_{jt}\right) e_{jt}/N_i$; where N_i is the number of working-age members of the ith household, e_{jt} takes value 1 if individual j has worked in year t, and w_{jt} is a measure of work intensity for the jth individual at time t. In our implementation, we define $w_{jt} = (m_{jt}/12)(h_{jt}/40)$, where m_{jt} are the number of months worked in year t and h_{jt} is the individual's average weekly working time in year t. The resulting GER indicator is actually discrete, based on a few classes with little interpretability, and with a spike at $w_{it} = 1$. Even after transformation, it cannot be well-approximated by any continuous parametric distribution. In order to obtain a meaningful and interpretable indicator, we divide this outcome in four

ordered classes, where $Y_{it3} = 0$ if $\text{GER}_{it} \geq 1$ (the individuals in the household work an average of forty hours per week, or more), $Y_{it3} = 1$ if $0.75 \leq \text{GER}_{it} < 1$, $Y_{it3} = 2$ if $0.5 \leq \text{GER}_{it} < 0.75$, and $Y_{it3} = 3$ otherwise. A data analysis based on a different transformation of the variables, leading one Gaussian, one Bernoulli, and one Poisson outcome, is reported in the Web Appendix. We finally select two covariates for each poverty dimension: we include three dummy variables to discriminate between territorial regions (North–East, South and islands, North–West; where Central Italy is used as reference category), and time. About 14% of the households have an income below the poverty threshold in each year. The quantiles for the counting measure of material deprivation are constant over time, with a median of zero, third quartile of two, and maximum value of seven. The average number of active items is slightly above one in each year. Also work intensity is marginally approximately constant over time, with 40% of the households having the highest level of work intensity, approximately one third with a value of one, and about 10% for each of the other two categories (corresponding to lower and lower work intensity). Predictably, a positive association can be seen among each couple of indicators. The point biserial correlation between Y_{it1} and Y_{it2} is slightly less than 0.4 each year, similarly the one between Y_{it1} and Y_{it3} is about 0.2 each year. Kendall's τ between Y_{it2} and Y_{it3} is 0.25 or slightly less in each year. These association measures are actually slightly low when one realizes that the first two indicators should actually measure the same latent trait. Similarly, a mismatch is easily observed between poverty and work intensity. We now give some examples to fix the ideas, restricting to 2017 for simplicity. In 2017, 46% of the families with an income above the poverty line have high work intensity. This proportion decreases only to 42% for households whose income is below the poverty line: a surprisingly large proportion of households is made of "working poors" who have very low income despite constantly being involved in full time jobs. Similarly, a surprising share of 22% of households with low income do not have any active material deprivation items, and 19% only one. The mismatch between work intensity and material deprivation is even more surprising, as families with highest work intensity are actually more likely to have two or more active items in the deprivation list than families with $0.75 \leq \text{GER}_{it} < 1$ (24.5% vs. 18.2%). This might be explained by the occurrence of more qualified jobs with higher wages and less working hours. On the other hand, even when $0.5 \leq \text{GER}_{it} < 0.75$ households with two or more active items are not common (only 26.8%). Additionally, more than half of the households with the lowest level of work intensity have at most one active item in the material deprivation list: they are able to make ends meet despite their members working only occasionally. In Fig. 1, we visualize the complex marginal interrelationships among our three endpoints. On the horizontal and vertical axes of the figure, we condition on the different levels of material deprivation (Y_{it2}) and work intensity (Y_{it3}). In each block, as obtained by a combination of the levels of these two outcomes, we show the proportion of households with equivalised disposable income below the poverty threshold, where darker color indicates larger proportions.

We now fit Latent Markov models with dependency structures modulated by the four parametric copulas described in Table 1, and $k = 1, \ldots, 5$ latent masses. In order to perform model selection, we use the Bayesian Information Criterion (BIC). Results are reported in Table 6. It can be seen that the model corresponding to the lowest BIC

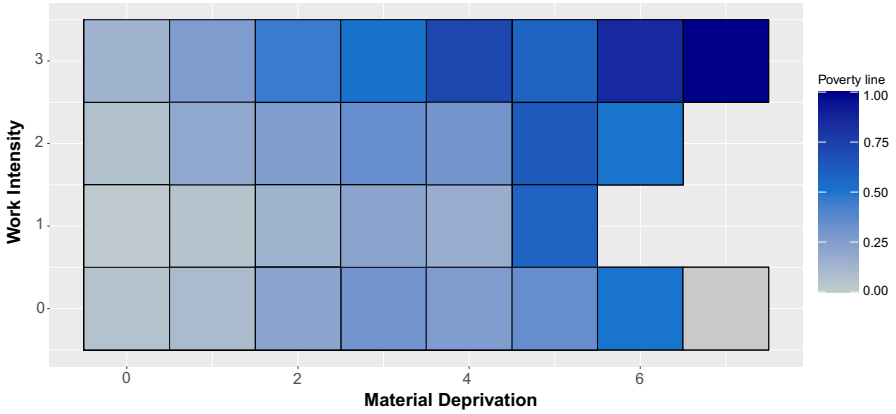

Fig. 1 EU SILC data. Fraction of households with equivalised disposable income below the poverty threshold, by material deprivation and work intensity levels

Table 6 EU SILC data

Frank				Clayton			
k	$\ell^*(\boldsymbol{\theta}^*)$	$\#par$	BIC	k	$\ell^*(\boldsymbol{\theta}^*)$	$\#par$	BIC
1	−15,755.79	18	31,619.26	1	−15,857.04	18	31,821.76
2	−13,964.45	26	28,115.55	2	−13,976.39	26	28,139.42
3	−13,653.87	36	27,566.18	3	−13655.35	36	27,569.12
4	**−13,300.74**	**48**	**26,946.05**	4	−13,318.14	48	26,980.85
5	−13,300.56	62	27,046.20	5	−13,318.09	62	27081.25
Joe				Gumbel			
k	$\ell^*(\boldsymbol{\theta}^*)$	$\#par$	BIC	k	$\ell^*(\boldsymbol{\theta}^*)$	$\#par$	BIC
1	−15,679.96	18	31,467.59	1	−15,676.34	18	31,460.36
2	−13,977.36	26	28,141.35	2	−13,974.31	26	28,135.26
3	−13,648.69	36	27,555.81	3	−13,647.20	36	27,552.83
4	−13,322.28	48	26,989.12	4	−13,317.56	48	26,979.70
5	−13,322.21	62	27,089.50	5	−13,317.56	62	27,080.19

Log-likelihood at convergence, number of parameters and Bayesian Information Criterion (BIC) for latent Markov models with different values of k and copula functions
Bold is used to highlight the model specification associated to the lowest BIC value (selected model)

is based on $k = 4$ latent masses, with a Frank copula modulating the dependency structure among the outcomes.

For the case of a Frank copula and $k = 4$ we report estimates for $\boldsymbol{\alpha}$ in Table 7 and for $\boldsymbol{\beta}$ in Table 8. Parameters' estimates for the latent structure are summarized in Table 9. Standard errors in parentheses are obtained through a parametric bootstrap procedure. The estimated association parameter $\hat{\xi}$ is 1.021, with 95% confidence interval equal to (0.617, 1.424). When marginals are continuous the Kendall's τ for a Frank copula with association parameter ξ corresponds to $\tau = 1 + 4\left(\mathcal{D}_1(\xi) - 1\right)/\xi$, where $\mathcal{D}_1(\xi) =$

Table 7 EU SILC data

$\hat{\alpha}$	Y_{it1}	Y_{it2}	$Y_{it3}=0$	$Y_{it3}=1$	$Y_{it3}=2$	$Y_{it3}=3$
k_1	−6.332 (0.174)	−1.788 (0.085)	–	−0.648 (0.135)	−1.930 (0.135)	−3.348 (0.155)
k_2	−5.737 (0.095)	0.418 (0.075)	–	−0.758 (0.198)	−3.508 (0.094)	−4.216 (0.084)
k_3	−1.225 (0.191)	0.689 (0.061)	–	4.662 (0.047)	2.094 (0.134)	0.355 (0.124)
k_4	0.403 (0.264)	0.500 (0.146)	–	−1.253 (0.146)	−2.975 (0.079)	−3.124 (0.097)

Estimated latent intercepts for a LM model with $k = 4$ latent masses and Frank copula. Standard errors in parentheses

Table 8 EU SILC data

$\hat{\beta}$	NW	SI	NE	Time
Y_{it1}	0.233 (0.239)	1.343 (0.207)	0.786 (0.260)	0.047 (0.054)
Y_{it2}	0.137 (0.068)	0.364 (0.057)	0.055 (0.079)	-0.048 (0.015)
Y_{it3}	-0.367 (0.139)	0.122 (0.118)	-0.169 (0.181)	-0.051 (0.028)

Estimated marginal-specific regression coefficients for a LM model with $k = 4$ latent masses and Frank copula. Standard errors in parentheses. NW: Northwest, SI: South and Islands, NE: Northeast. Central Italy is reference category

Table 9 EU SILC data

$\hat{\pi}$	k_1	k_2	k_3	k_4
	0.525 (0.024)	0.206 (0.023)	0.186 (0.015)	0.083 (0.012)
$\hat{\Pi}$	k_1	k_2	k_3	k_4
k_1	0.939 (0.011)	0.050 (0.011)	0.009 (0.004)	0.002 (0.002)
k_2	0.081 (0.022)	0.874 (0.025)	0.024 (0.010)	0.020 (0.010)
k_3	0.044 (0.017)	0.063 (0.018)	0.864 (0.019)	0.029 (0.009)
k_4	0.006 (0.010)	0.089 (0.035)	0.025 (0.018)	0.879 (0.032)

Estimated initial and transition probabilities for a LM model with $k = 4$ latent masses and Frank copula. Standard errors in parentheses

$\xi^{-1}\int_0^{\xi} t\cdot(e^t-1)^{-1}\mathrm{d}t$ is the Debye function of the first type. This mapping is not entirely accurate with discrete marginals, still it gives $\hat{\tau} = 0.112$ (95% CI 0.068–0.155) in our example.

The results of the data analysis lead to the following considerations. First of all, part of the mismatch/low association among outcomes can be explained by unobserved heterogeneity. The latent states indicate different levels of overall poverty, where the first indicates wealthy households that are unlikely to have a disposable income below the poverty line, have a low number of active material deprivation items on average, and have the highest work intensities. The other latent states can be characterized by

different (increasing) levels of monetary poverty, but material deprivation and work intensity are not similarly monotone. For instance, propensity to material deprivation in latent states two to four is approximately constant (with a peak in latent state three) despite a growing propensity to low income. Similarly, in latent state four we can see a slightly large propensity to high work intensity despite high likelihood of being low income and materially deprived. Consequently, it can be concluded that, for idiosyncratic reasons, there actually exist a non-negligible share of peculiar households with mismatch among two or even three outcome levels.

Secondly, even after taking into account this fact, mismatch among outcomes persists. This is testified by the low estimate for the association parameter, which indicates a low residual association among the outcomes (that is, conditionally on the latent state). It can be concluded then that one or more outcomes are possibly not well measured. A deep assessment about issues with measuring material deprivation can be found in Dotto et al. (2018) and Farcomeni et al. (2022) and can be summarized in the possible lack of unidimensionality and measurement invariance of the official indicators. It might also be supposed that equivalised disposable income by itself is not a very good indirect measure of living conditions, as households with similar income levels might be better or worse at making ends meet depending on several internal and contextual factors (e.g., purchase power in the local area). Additionally, there might be unemployed people who have a very low income but are not materially deprived due to support from relatives, which is a well-known safety net in Italy, or inherited wealth assets. A further source of unobserved heterogeneity might be in work intensity, due to the occurrence of undeclared or informal work spells. Additional comments involve the regression coefficients and latent distribution. It can be seen that as common with many indicators, southern Italy and its islands are associated with a significantly higher risk of low income and high material deprivation. Households living in the northwestern regions are more likely to have higher work intensity than those living in central Italy, but they are also more likely to be materially deprived (possibly due to lower purchase power in those regions). On the other hand, households living in the northeast of Italy are more likely to have low income, but we do not have evidence that they experience more material deprivation than households located in central Italy. It can also be seen that over time the propensity to material deprivation and low work intensity have generally decreased in the period considered. Finally, the estimated transition matrix indicates that households are generally trapped in their latent state, with 6% to 13% switching to another latent state at each time occasion.

7 Conclusions

In this work, we show a natural way of relaxing the conditional independence assumption in multivariate latent Markov models. The copula framework is flexible and includes conditionally independent marginals as a special case. We have proposed a rather general framework for moderately dimensional models with arbitrary marginals, in a regressive framework. Parameterization of copula parameters in terms of covariates, as for instance done in Donat and Marra (2018), is rather straightforward in our framework; as is a similar parameterization for the latent distribution. The full likeli-

hood approach is very effective and useful, and it provably reduces bias in real data examples, where residual contemporary dependence is the norm rather than the exception. Our interest resided mainly in modeling discrete outcomes, possibly together with continuous ones. In the development of our theory, we have evaluated several approximation strategies to the likelihood, all of which had strong limitations. The need to compute PDFs in the presence of discrete marginals somehow restricts the approach to Archimedean copulas for moderately sized data sets (as the motivating example), as Gaussian and T copulas do not scale well. The problem with Gaussian and T copulas resides in the fact that, in order to compute (8), the CDF is necessary; and this might be computationally demanding even for small dimensional problems. These computational issues do not apply to the adopted one-parameter copula functions, for which the CDF is available in closed form. On the other hand, the estimate of the copula parameter $\hat{\xi}$ might be slightly more difficult to interpret, and post-analysis mapping to measures of association (like we did in the real data example) might be necessary. Furthermore, the dependency structure might be somehow restricted within the family. Flexible and simple copula structures are available, but they might be restricted to two or anyway low dimensions, e.g., Zachariah et al. (2024). In further work, we plan to tackle these issues by exploring the use of Vine copulas. Another interesting route for further research involves reparameterizing the copula parameter as a function of covariates as well. This would be straightforward from a modeling perspective, but it might complicate interpretability and identifiability of the model.

We have also described a novel real data example about an open problem in economics, related to measurement of poverty. We have underlined how commonly used indicators do not seem to be able to catch the complexity of the poverty phenomenon, leading to surprising mismatch between income, material deprivation, and work intensity. This seems to be due both to the presence of a non-negligible share of peculiar households, and to how the constructs are measured, which leaves many open questions for further work.

Supplementary information A Web Appendix reports details on the simulation scenarios, and real example with one continuous and two discrete outcomes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11749-024-00919-9.

Acknowledgements The authors are grateful to Prof. Roberto Zelli for advice on a first draft, and to two referees for constructive suggestions. This study was funded by the European Union—NextGenerationEU, in the framework of the GRINS—Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 - CUP E83C22004690001). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

Funding Open access funding provided by Università degli Studi di Roma Tor Vergata within the CRUI-CARE Agreement.

References

Anderson G, Farcomeni A, Pittau MG, Zelli R (2019) Multidimensional nation wellbeing, more equal yet more polarized: an analysis of the progress of human development since 1990. J Econ Dev 44(1):1–22

Anderson G, Farcomeni A, Pittau MG, Zelli R (2019) Rectangular latent Markov models for time-specific clustering, with an analysis of the well being of nations. J R Stat Soc (Ser C) 68:603–621

Bartolucci F, Farcomeni A (2009) A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. J Am Stat Assoc 104:816–831

Bartolucci F, Farcomeni A (2015) A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates. Biometrics 71:80–89

Bartolucci F, Farcomeni A, Pennoni F (2013) Latent Markov models for longitudinal data. Chapman & Hall/CRC Press, Boca Raton

Bartolucci F, Farcomeni A, Pennoni F (2014) Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates (with discussion). TEST 23:433–486

Brandolini A, Viviano E (2016) Behind and beyond the (head count) employment rate. J R Stat Soc Ser A (Stat Soc) 179(3):657–681

DeRuiter SL, Langrock R, Skirbutas T, Goldbogen JA, Calambokidis J, Fiedlaender AS, Southall BL (2017) A multivariate mixed hidden Markov model for Blue Whale behaviour and responses to sound exposure. Ann Appl Stat 11:362–392

Donat F, Marra G (2018) Simultaneous equation penalized likelihood estimation of vehicle accident injury severity. J R Stat Soc (Ser C) 87:979–1001

Dotto F, Farcomeni A, Pittau MG, Zelli R (2018) A dynamic inhomogeneous latent state model for measuring material deprivation. J R Stat Soc (Ser C) 182:495–516

Farcomeni A (2015) Generalized linear mixed models based on latent Markov heterogeneity structures. Scand J Stat 42:1127–1135

Farcomeni A, Pittau MG, Viviani S, Zelli R (2022) A European measurement scale for material deprivation. https://doi.org/10.21203/rs.3.rs-2250804/v1

Genest C, Nevslehová J (2007) A primer on copulas for count data. ASTIN Bull J IAA 37(2):475–515

Hardle WK, Okhrin O, Wang W (2015) Hidden Markov structures for dynamic copulae. Econom Theory 31:981–1015

Henn LL (2022) Limitations and performance of three approaches to Bayesian inference for Gaussian copula regression models of discrete data. Comput Stat 37:909–946

Hohberg M, Donat F, Marra G, Kneib T (2021) Beyond unidimensional poverty analysis using distributional copula models for mixed ordered-continuous outcomes. J R Stat Soc (Ser C) 70:1365–1390

Joe H (2014) Dependence modeling with copulas. CRC Press, Boca Raton

Martino A, Guatteri G, Paganoni AM (2020) Multivariate hidden Markov models for disease progression. Stat Anal Data Min ASA Data Sci J 13:499–507

Merlo L, Maruotti A, Petrella L, Punzo A (2022) Quantile hidden semi-Markov models for multivariate time series. Stat Comput 32:61

Nelsen R (2006) An introduction to copulas. Springer, New York

Nikoloulopoulos AK (2013) Copula-based models for multivariate discrete response data. In: Jaworski P, Durante F, Härdle WK (eds) Copulae in mathematical and quantitative finance. Springer, Berlin, pp 231–249

Orfanogiannaki K, Karlis D (2018) Multivariate Poisson hidden Markov models with a case study of modelling seismicity. Aust N Z J Stat 60:301–322

Otting M, Karlis D (2023) Football tracking data: a copula-based hidden Markov model for classification of tactics in football. Ann Oper Res 325:167–183

Otting M, Langrock R, Maruotti A (2023) A copula-based multivariate hidden Markov model for modelling momentum in football. AStA Adv Stat Anal 107:9–27

Punzo A, Ingrassia S, Maruotti A (2021) Multivariate hidden Markov regression models: random covariates and heavy-tailed distributions. Stat Pap 62:1519–1555

Russo A, Farcomeni A, Pittau MG, Zelli, R (2022) Covariate-modulated rectangular latent Markov models with an unknown number of regime profiles. Stati Model (**in press**)

Townsend P (1987) Deprivation. J Soc Policy 16:125–146

Townsend P, Gordon D (1991) What is enough? New evidence on poverty allowing the definition of a minimum benefit. In: Alder M, Bell C, Clasen J, Sinfield A (eds) The sociology of social security. Edinburgh University Press, Edinburgh, pp 35–69

Trivedi P, Zimmer D (2017) A note on identification of bivariate copulas for discrete count data. Econometrics 5:10

Whelan CT, Layte R, Maître B (2004) Understanding the mismatch between income poverty and deprivation: a dynamic comparative analysis. Eur Sociol Rev 20(4):287–302

Wu B, de Leon AR, Withanage N (2013) Joint analysis of mixed discrete and continuous outcomes via copula models. In: de Leon AR, Chough KC (eds) Analysis of mixed data. Chapman and Hall/CRC, New York

Zachariah SR, Arshad M, Pathak AK (2024) A new class of copulas having dependence range larger than FGM-type copulas. Statistics & Probability Letters 206:109988

Zilko AA, Kurowicka D (2016) Copula in a multivariate mixed discrete-continuous model. Comput Stat Data Anal 103:28–55

Zucchini W, MacDonald IL (2009) Hidden Markov models for time series: an introduction using R. Springer, New York