Submitted to the Annals of Applied Statistics

COMPOSITE MIXTURE OF LOG-LINEAR MODELS WITH APPLICATION TO PSYCHIATRIC STUDIES

BY EMANUELE ALIVERTI [1], AND DAVID B. DUNSON[2]

[1]*University Ca' Foscari Venezia, emanuele.aliverti@unive.it*

[2]*Duke University, dunson@duke.edu*

Psychiatric studies of suicide provide fundamental insights on the evolution of severe psychopathologies, and contribute to the development of early treatment interventions. Our focus is on modelling different traits of psychosis and their interconnections, focusing on a case study on suicide attempt survivors. Such aspects are recorded via multivariate categorical data, involving a large numbers of items for multiple subjects. Current methods for multivariate categorical data – such as penalized log-linear models and latent structure analysis – are either limited to low-dimensional settings or include parameters with difficult interpretation. Motivated by this application, this article proposes a new class of approaches, which we refer to as Mixture of Log Linear models (MILLS). Combining latent class analysis and log-linear models, MILLS defines a novel Bayesian approach to model complex multivariate categorical data with flexibility and interpretability, providing interesting insights on the relationship between psychotic diseases and psychological aspects in suicide attempt survivors.

1. Introduction. We are motivated by a psychiatric study of suicide attempts, focused on investigating the psychological profiles of survivors of a suicidal act (e.g. Scocco et al., 2020; Nock et al., 2008; De Leo et al., 2004). Studies on suicide attempts are crucial for the development of novel interventions, based on early identification of key psychological symptoms, such as depression or hallucination (e.g. Hawton and Fagg, 1988; Kelleher et al., 2011). Detailed characterisation of the psychological profiles in suicide attempts provide important insights on the dynamics of suicidal acts, and the relationships between psychotic symptoms and other psychological traits, such as empathy (De Beurs et al., 2019). We are interested in analysing traits of suicide attempt patients, including psychoses and empathic profiles, while also characterizing interactions across these classes of traits.

In the psychological literature, the investigation of the relationship between psychoses and empathy has received considerable attention, remaining a challenging research objective which is routinely explored (e.g. McCormick et al., 2012; Ladisich and Feil, 1988). In general, specific empathic profiles are also associated with depression (Cusi et al., 2011; Schreiter, Pijnenborg and Aan Het Rot, 2013), obsessive compulsive disorders (Fontenelle et al., 2009), anxiety (Perrone-McGovern et al., 2014) and hostility (Guttman and Laporte, 2002). For example, a frequent symptom of depression is the inability to perceive our own feelings, which is also realistically associated with the inability to comprehend other individuals' emotions (e.g. Cusi et al., 2011). Similar examples involve different empathic conditions, such as personal distress and severe hostility, which are likely to be associated with acute anxiety (Guttman and Laporte, 2002).

Although there are many studies focusing on the interconnections among these psychological aspects, their mutual influence in patients attempting suicide is not completely understood. Indeed, preliminary evidence suggests that individuals who attempted suicide can exhibit unexpected association patterns across psychotic symptoms and specific empathic profiles, and such interactions could be relevant for characterising underlying psychological

Keywords and phrases: Psychiatric profiles, Bayesian modelling, Categorical data, Contingency table, Log-linear models, Mixture model.

mechanisms (Scocco et al., 2020; Wang et al., 2020; Zhang et al., 2019). For instance, depressed individuals with a high level of empathic concern may suffer inconsistent thoughts and feelings, exacerbating their clinical condition and potentially increasing the risk of re-attempting suicide.

Subjects analysed in the study correspond to a sample of 56 inpatients hospitalized after an attempted suicide at the psychiatric ward of Padova Hospital (Italy) between January 2017 and December 2018 (Scocco et al., 2020). Suicide attempts can be intentional or not, depending on whether the individual consciously realizes that his actions are intended to kill him. This distinction can be blurred for many episodes; for example, with poisoning or drug overdoses (Britton et al., 2012). In this study, we rely on clinicians' evaluations about intentionality. Individuals were labelled as "attempted suicide" if they harmed their body and consciously realized that such an act could kill them (e.g. Goodfellow, Kõlves and De Leo, 2019). During hospitalisation, clinicians submit self-reported questionnaires to each patient to supervise their psychological evolution over time. Such tools are developed to investigate different aspects of individuals' psychology, with the main focus being on the evaluation of the psychotic profiles and the empathic status (Scocco and De Leo, 2002). Specifically, these facets are evaluated through the Symptom Check List (SCL-90; Derogatis, Lipman and Covi, 1973) and the Interpersonal Reactivity Index (IRI; Davis, 1980) questionnaires.

The SCL-90 is commonly used to describe psychiatric symptoms, using 90 items scored on a five-point Likert scale; additionally, scores can be grouped into nine subscales (somatization, obsessive-compulsive, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation, psychoticism) corresponding to well-defined psychiatric profiles (Derogatis, Lipman and Covi, 1973). As suggested by our clinician collaborators, it is of particular interest to focus on 4 subscales of the questionnaire: obsessive-compulsive (OC), depression (DEP), anxiety (ANX) and hostility (HOS), encompassing a total of 39 items measuring the psychotic aspects which are more relevant in suicide attempts evaluation. We have further removed from analysis 4 items with a large fraction of missing observations, resulting in a total of 35 items for SCL-90. Although we could have used imputation methods, the high proportion of missingness and our small sample size led us to instead remove these items. See Table 6 in Appendix A for a detailed illustration of the items under investigation.

The IRI is a 28-item instrument scored on a five-point Likert scale that measures the emotional and cognitive components of a person's empathy, with four subscales. The IRI measures the cognitive capacity to see things from the point of view of others (Perspective Taking, PT), the tendency to experience reactions of sympathy, concern and compassion for other people undergoing negative experiences (Empathic Concern, EC), the tendency to experience distress and discomfort in witnessing other people's negative experiences (Personal Distress, PD) and the capacity to strongly identify oneself with fictitious characters in movies, books, and plays (Fantasy, FS). We will focus only on the 22 items that were uniquely associated with a specific empathic subscale, and without missing observations. For a detailed illustration, see Table 7 in Appendix A.

Following the notation convention of Lauritzen (1996), we will indicate with $V = \{1, \ldots, k\}$ the set of $k = 57$ categorical items collected from the two psychological questionnaires combined. We also denote with $(Y_j, j \in V)$ the variables taking values in the finite set $\mathcal{I}_j$, with dimension $|\mathcal{I}_j| = d_j$ corresponding to the number of categories of the j-th item. In the psychological study under investigation, $d_j = 5$ and $\mathcal{I}_j = \{0, \ldots, 4\}$, for each $j = 1, \ldots, 57$. Data collected from patients consist of an $n \times k$ matrix with elements $y_{ij} \in \{0, \ldots, 4\}$, where $i = 1, \ldots, 56$, $j = 1, \ldots, 57$. Table 1 illustrates the univariate frequencies for the items under investigation, sorted according to the subscale they refer to.

Preliminary findings suggest that most subjects generally report high scores of hostility (HOS). Such a subscale focuses on measuring different dimensions of hostility, including

TABLE 1

Univariate descriptive statistics. SCL-90 *questionnaire (left) and* IRI-*28 (right). Second column refer to the specific subscale the items refer to. Subjects answer with their level of agreement with numbers ranging from* 0 *("Not at all") to* 4 *("Extremely").*

ITEM	SUB	0	1	2	3	4
SCL-2	ANX	12	15	17	8	4
SCL-17	ANX	4	4	9	11	28
SCL-23	ANX	3	4	10	10	29
SCL-33	ANX	6	9	7	12	22
SCL-39	ANX	5	6	7	11	27
SCL-72	ANX	4	7	9	6	30
SCL-78	ANX	6	6	7	8	29
SCL-80	ANX	5	5	8	7	31
SCL-86	ANX	5	6	17	13	15
SCL-5	DEP	16	5	6	5	24
SCL-14	DEP	10	15	12	10	9
SCL-15	DEP	12	3	10	14	17
SCL-20	DEP	4	11	6	14	21
SCL-22	DEP	9	5	6	9	27
SCL-26	DEP	6	8	13	14	15
SCL-29	DEP	18	12	7	10	9
SCL-30	DEP	16	14	14	9	3
SCL-31	DEP	9	13	9	12	13
SCL-32	DEP	13	14	5	11	13
SCL-71	DEP	8	12	8	12	16
SCL-79	DEP	10	13	5	15	13
SCL-11	HOS	6	8	8	22	12
SCL-63	HOS	2	2	6	6	40
SCL-67	HOS	2	4	7	2	41
SCL-74	HOS	3	2	9	9	33
SCL-3	OC	14	13	11	8	10
SCL-9	OC	7	6	8	22	13
SCL-10	OC	2	8	13	18	15
SCL-28	OC	9	6	11	20	10
SCL-38	OC	7	8	9	19	13
SCL-45	OC	3	9	7	14	23
SCL-46	OC	9	5	8	19	15
SCL-51	OC	6	5	8	13	24
SCL-55	OC	7	10	11	16	12
SCL-65	OC	1	2	6	11	36

ITEM	SUB	0	1	2	3	4
IRI-2	EC	4	7	9	17	19
IRI-4	EC	19	10	13	8	6
IRI-9	EC	3	6	7	14	26
IRI-14	EC	21	15	8	6	6
IRI-18	EC	27	7	7	7	8
IRI-1	FS	10	10	22	9	5
IRI-5	FS	8	12	12	14	10
IRI-7	FS	10	11	18	12	5
IRI-12	FS	19	13	9	7	8
IRI-16	FS	15	8	14	9	10
IRI-23	FS	8	12	15	4	17
IRI-26	FS	12	11	8	14	11
IRI-10	PD	4	9	14	12	17
IRI-13	PD	13	12	14	9	8
IRI-17	PD	11	10	12	11	12
IRI-19	PD	11	12	7	10	16
IRI-3	PT	9	19	12	14	2
IRI-11	PT	5	8	17	12	14
IRI-15	PT	10	9	13	14	10
IRI-21	PT	5	9	14	16	12
IRI-25	PT	12	13	15	10	6
IRI-28	PT	4	11	12	15	14

thoughts, feelings, and actions that are characteristic of the negative affect state of anger (Derogatis, Lipman and Covi, 1973). High scores demonstrate that resentment, irritability and rage are common in the patients under investigation. Similarly, subjects respond with high scores to items belonging to the Anxiety (ANX) and Obsessive-Compulsive (OC) subclasses. These items are devoted to measuring nervousness, tension and impulses that are experienced as irresistible (Derogatis, Lipman and Covi, 1973). The prevalence of high scores in these questions indicate that patients who attempted suicide demonstrate feelings of apprehension and panic, and that they often feel the need to obsessively check what they do.

Interestingly, we observe heterogeneous responses to items measuring depressive profiles (DEP). For example, subjects respond to the item SCL-15 (*"Thoughts of ending your life"*) both with low and high scores. Similarly, responses to most questions referring to empathic traits are heterogeneous, and indicate that the sample is characterized by different profiles in terms of empathic feelings. As an exception, it is of interest to focus on the Empathic-Concern subscale (EC), which is characterised by more polarized answers; see for example, item IRI-18 (*"When I see someone being treated unfairly, I sometimes don't feel very much pity for them"*) and IRI-14 (*"Other people's misfortunes do not usually disturb me a great deal"*), where most patients respond with low scores (disagreement) indicating feelings of sympathy and concern for unfortunate others.

These preliminary descriptions indicate that patients under investigation have non-trivial psychopathological traits, characterised by different psychotic symptoms and interesting empathic profiles. To provide deeper insights into the psychopathology of attempted suicide, it is important to characterize the association structure across the items, in order to evaluate which profiles are mostly associated with specific symptoms. Therefore, the focus of further analysis will be on making inference on the dependence structure across the different pairs of categorical variables $(Y_j, Y_{j'})$, $j = 2, \dots, k$, $j' = 1, \dots, j$, providing a measure of the intensity of the pairwise dependence and an assessment of uncertainty in estimation. Several studies have described the design and the empirical dependence structure across the SCL-90 and IRI items, focusing on random samples (e.g., Prunas et al., 2012; Gilet et al., 2013) or subjects with moderate psychotic symptoms (Prinz et al., 2013). However, related information is not available for suicide attempt survivors, who might show unexpected association patterns that differ from other psychotic profiles (Scocco et al., 2020).

Associations and interactions across categorical variables are generally investigated through multi-way contingency tables, where individuals are cross classified according to their values for the different items. These tools are routinely used to investigate the association across the items and to test for the presence of specific dependence structures; see for example Agresti (2003) for an introduction. Under the adopted notation, the contingency table is denoted as $\mathcal{I}_V = \times_{j \in V} \mathcal{I}_j$, while its generic elements $\boldsymbol{i} = (i_1, \dots, i_p) \in \mathcal{I}_V$ are referred to as the *cells*. Given a sample of size n, the number of observations falling in the generic cell $\boldsymbol{i}$ is denoted as $y(\boldsymbol{i})$, with $\sum_{\boldsymbol{i} \in \mathcal{I}_V} y(\boldsymbol{i}) = n$. The joint table has a number of elements equal to $|\mathcal{I}_V| = \prod_{j=1}^{k} d_j = 5^{57}$ in our motivating application, which is exponential in the number of categorical variables and tremendously large. Indeed, computation of the joint cell counts is unfeasible even for moderate values of k, and is basically limited to settings with at most 15 binary variables (e.g. Johndrow et al., 2018). In addition, most cells will contain zero observation, leading to issues during estimation; for example, non existence of maximum-likelihoods estimates (e.g. Fienberg and Rinaldo, 2007). The huge dimensionality and severe sparsity motivate novel methods to adequately characterise the interactions among categorical variables in multivariate categorical data, with sparse log-linear models and latent structure modelling being popular options.

1.1. *Relevant literature.* The development of methods to analyse categorical data began well back in the 19th century, and remains a very active area of research (e.g. Fienberg and Rinaldo, 2007). Log-linear models are particularly popular. Logarithms of cell probabilities are represented as linear terms of parameters related to each cell index, and with coefficients that can be interpreted as interactions among the categorical variables (Agresti, 2003). The relationship between multinomial and Poisson log-likelihoods allows one to obtain maximum likelihood (ML) estimates for log-linear models leveraging standard generalized linear model (GLM) algorithms (e.g. Fisher-Scoring), with the vectorized table of cell counts used as a response variable. As outlined in Section 1, when the number of variables increases the number of cells of the contingency table grows exponentially. Therefore, many cells will be empty and there will be infinite ML estimates (Fienberg and Rinaldo, 2007). To overcome this issue and obtain unique estimates, it is often assumed that many coefficients are zero, and estimation is performed via penalised likelihood (Nardi et al., 2012; Tibshirani, Wainwright and Hastie, 2015; Ravikumar et al., 2010). However, these methods require computation of the joint cell counts, which is unfeasible in our setting.

Bayesian approaches for inference in log-linear models often restrict consideration to specific nested model subclasses; for example, hierarchical, graphical or decomposable log-linear models (Lauritzen, 1996). Conjugate priors on the model coefficients are available (Massam et al., 2009), but exact Bayesian inference is still complicated since the resulting posterior distribution is not particularly useful, lacking closed form expressions for important functionals – such as credible intervals – and sampling algorithms to perform inference via Monte Carlo integration. As an alternative, the posterior distribution can be analytically approximated with a Gaussian distribution if the number of cells is not excessive (Johndrow et al., 2018). When the focus is on selecting log-linear models with high posterior evidence, stochastic search algorithms evaluating the exact or approximate marginal likelihood are available (Dobra and Massam, 2010; Dobra and Mohammadi, 2018).

A different perspective on analyzing multivariate categorical data relies on latent structures (Lazarsfeld, 1950). This family of models is specified in terms of one or more latent features, with observed variables modelled as conditionally independent given the latent features. Marginalising over the latent structures, complex dependence patterns across the categorical variables are induced (e.g. Andersen, 1982). Representative examples include latent class analysis (Lazarsfeld, 1950) and the normal ogive model (Lawley, 1943), where a univariate latent variable with discrete or continuous support, respectively, captures the dependence structure among the observed categorical variables; see also Fruhwirth-Schnatter, Celeux and Robert (2019, Chapters 9 and 11) and references therein. More flexible multivariate latent structures have also been introduced; for example, grade of membership models (Erosheva, 2005) and the more general class of mixed membership models (Airoldi et al., 2014). Specific latent variable models can also be interpreted as tensor decompositions of the contingency tables (Dunson and Xing, 2009; Bhattacharya and Dunson, 2012); see also Kolda and Bader (2009) for a discussion.

To conduct meaningful and interpretable inferences, it is important for marginal or conditional distributions and measures of association to have a low-dimensional structure. For example, it is often of substantial interest to characterise bivariate distributions and test for marginal or conditional independence (Agresti, 2003). Leveraging data-augmentation schemes, estimation of latent variable models is feasible in high-dimensional applications (e.g. Dunson and Xing, 2009); however, these approaches might require many components to adequately characterize complex data, and can lack simple interpretability of the model parameters and the induced dependence structure. On the other hand, log-linear model directly parameterize the interactions among the categorical variables (Agresti, 2003) or the lower-dimensional marginal distributions (Bergsma et al., 2002), but estimation is generally unfeasible when the number of variables is moderate to high, due to the huge computational bottlenecks and the massively large model space. Sparse log-linear models and latent class structures are deeply related in the way in which sparsity is induced in the resulting contingency table (Johndrow, Bhattacharya and Dunson, 2017), but a formal methodology mixing the benefits of the two model families is still lacking.

Motivated by the application to studies of suicide attempt, in this article we introduce a novel class of Bayesian models for categorical data, which we refer to as MILLS. We propose to model the multivariate categorical data as a composite mixture of log-linear models with first order interactions, characterising the bivariate distributions with simple and robust models while accounting for dependencies beyond first order via mixing different local models. Such a specification models categorical data with a simple, yet flexible, specification which can take into account complex dependencies with a relatively small number of components. The idea of mixing simple low-dimensional models to reduce the number of parameters needed to characterize complex data has a long history. One example is mixing first order Markov models to account for higher order structure (Raftery, 1985). See also Fruhwirth-Schnatter, Celeux and Robert (2019) for related ideas.

2. Log linear models. Following Lauritzen (1996), we fix an arbitrary reference cell $\boldsymbol{i}^\star$ of the contingency table, which can be assumed as $\boldsymbol{i}^\star = (0, \ldots, 0)$ without loss of generality. For each cell $\boldsymbol{i} \in \mathcal{I}_V$ of the table, we denote as $p(\boldsymbol{i}) = \mathrm{pr}(Y_1 = \boldsymbol{i}_1, \ldots, Y_k = \boldsymbol{i}_k)$ the probability of falling in cell $\boldsymbol{i}$. According to the notation of Section 1, we denote as $\mathbf{p} = (p(\boldsymbol{i})/p(\boldsymbol{i}^\star), \boldsymbol{i} \in \mathcal{I}_v)$ the vectorised ratio between cell probabilities and the reference cell $\boldsymbol{i}^\star$; see also Johndrow et al. (2018). A log-linear model is a generalised linear model for the resulting multinomial likelihood, which represents the logarithms of cell probabilities additively as a function of a set of log-linear parameters $\boldsymbol{\vartheta}$. Following Propostion 2.1 of Letac et al. (2012), it is possible to relate cell probabilities and log-linear coefficients as follows:

$$\log \mathbf{p} = \mathbf{X}\boldsymbol{\vartheta}, \tag{1}$$

where $\mathbf{X}$ is a full rank $|\mathcal{I}_V| \times |\mathcal{I}_V|$ matrix if the transformation is invertible; for example, when $\mathbf{X}$ is the identity matrix, the so-called identity parametrisation is obtained. Identifiability is imposed through careful specification of the matrix $\mathbf{X}$, which determines the model parametrisation and, consequently, constraints on the parameters, and fixing the first element of $\boldsymbol{\vartheta}$ to zero (Agresti, 2003); see also Letac et al. (2012, Proposition 2.1) for related arguments. Equation (1) can be extended to embrace a larger class of invertible and non-invertible log-linear parametrisations; for example, marginal parametrisations (e.g. Bergsma et al., 2002; Roverato, Lupparelli and La Rocca, 2013; Lupparelli, Marchetti and Bergsma, 2009).

In general, it is desirable to specify a sparse set of m coefficients with $m \ll |\mathcal{I}_v|$, corresponding to some notion of interactions among the categorical variables; for example, representing conditional or marginal independence (Agresti, 2003). When a sparse parameterisation is employed, it is common to remove in Equation (1) the columns of $\mathbf{X}$ associated with excluded coefficients, thereby obtaining a more parsimonious design matrix with dimension $|\mathcal{I}_V| \times m$. In this article we focus on the corner parameterisation, which is particularly popular in the literature for categorical data (Agresti, 2003; Massam et al., 2009; Letac et al., 2012), and is generally the default choice in statistical software. The columns of $\mathbf{X}$ under the corner parameterisation can be formally expressed in terms of Möbius inversion (e.g. Letac et al., 2012, Preposition 2.1); see also Massam et al. (2009, Lemma 2.2). For simplicity in exposition, we prefer to use matrix notation.

Let $\mathbf{y} = (y(\boldsymbol{i}), i \in \mathcal{I}_v)$ denote the vectorised cell counts. The likelihood function associated with the multinomial sampling and log-linear parameters can be expressed, in matrix form, as follows:

$$\prod_{\boldsymbol{i} \in \mathcal{I}_V} p(\boldsymbol{i})^{y(\boldsymbol{i})} = \exp\left\{\mathbf{y}^\intercal \mathbf{X}\boldsymbol{\vartheta} - n\kappa(\boldsymbol{\vartheta})\right\} = \exp\left\{\tilde{\mathbf{y}}^\intercal \boldsymbol{\vartheta} - n\kappa(\boldsymbol{\vartheta})\right\}, \tag{2}$$

with $\kappa(\boldsymbol{\vartheta}) = \log\left[\mathbf{1}^\intercal \exp(\mathbf{X}\boldsymbol{\vartheta})\right]$. Such a parametrisation yields a very compact data reduction, since the canonical statistics $\mathbf{y}^\intercal \mathbf{X} = \tilde{\mathbf{y}}^\intercal$ correspond to the marginal cell counts relative to the highest interaction term included in the model (Massam et al., 2009; Agresti, 2003). In particular, we will consider hierarchical log-linear models which include all the main effects and all the first-order interactions; under such a specification, the canonical statistics $\tilde{\mathbf{y}}$ correspond to the marginal bivariate and univariate tables (e.g., Agresti, 2003).

3. Composite likelihood. The log-partition function in Equation (2) involves a sum of $|\mathcal{I}_V|$ terms, the total number of cells. Due to the immense number of cells, the likelihood cannot be evaluated unless the number of variables k is very small. Approximations of intractable likelihoods have been proposed in the literature, with Monte Carlo maximum likelihood (Snijders, 2002; Geyer and Thompson, 1992) being one option. Composite likelihoods provide a computationally tractable alternative to the joint likelihood, relying on a product of marginal or conditional distributions; see Varin, Reid and Firth (2011) for an overview. Extending the work of Meng et al. (2013), Massam and Wang (2018) focused on composite maximum likelihood estimation for log-linear models, with a careful choice of the conditional and marginal distributions based on the conditional dependence graph. However, the dependence graph is typically unknown and its estimation can be very demanding and affected by large uncertainty (Dobra and Massam, 2010).

We propose to replace the joint likelihood with a simple and robust alternative. Denote as $\mathcal{P}_2$ the set of subsets of V with cardinality 2. For each $E_2 \in \mathcal{P}_2$, let $\boldsymbol{y}_{E_2}$ denote the vectorised E_2-marginal bivariate table of counts. We define, for each $\boldsymbol{y}_{E_2}$, a saturated log-linear model with corner parametrisation:

$$\mathbf{p}(\boldsymbol{y}_{E_2}; \boldsymbol{\vartheta}_{E_2}) = \exp\left\{\boldsymbol{y}_{E_2}^{\intercal}\mathbf{X}_2\boldsymbol{\vartheta}_{E_2} - n\kappa_2(\boldsymbol{\vartheta}_{E_2})\right\} = \exp\left\{\tilde{\boldsymbol{y}}_{E_2}^{\intercal}\boldsymbol{\vartheta}_{E_2} - n\kappa_2(\boldsymbol{\vartheta}_{E_2})\right\}, \tag{3}$$

where $\kappa_2(\boldsymbol{\vartheta}_{E_2}) = \log[\mathbf{1}^{\intercal}\exp(\mathbf{X}_2\boldsymbol{\vartheta}_{E_2})]$, $\dim\boldsymbol{\vartheta}_{E_2} = \dim\tilde{\boldsymbol{y}}_{E_2} = |\mathcal{I}_{E_2}| = \prod_{j\in E_2} d_j$ and $\boldsymbol{\vartheta}_{E_2} \in \mathbb{R}^{|\mathcal{I}_{E_2}|}$. In our motivating application, this choice implies $\boldsymbol{\vartheta}_{E_2} \in \mathbb{R}^{25}$, with the first element of $\boldsymbol{\vartheta}_{E_2}$ equal to 0 for identifiability. There is an important difference between $\boldsymbol{y}_{E_2}$ and $\tilde{\boldsymbol{y}}_{E_2}$. The former refers to the E_2-marginal bivariate table, while the latter refers to the sufficient statistics of the log-linear model with corner parametrisation, which are elements of the bivariate and univariate E_2-marginal table; see, for example, Agresti (2003).

We define a surrogate likelihood function combining the distributions defined in (3) as

$$\begin{aligned}&\prod_{E_2\in\mathcal{P}_2}\mathbf{p}(\boldsymbol{y}_{E_2};\boldsymbol{\vartheta}_{E_2})^{w_{E_2}}\\ &= \exp\left\{\sum_{E_2\in\mathcal{P}_2} w_{E_2}\log\mathbf{p}(\boldsymbol{y}_{E_2};\boldsymbol{\vartheta}_{E_2})\right\} = \exp\left\{\sum_{E_2\in\mathcal{P}_2} w_{E_2}\left[\tilde{\boldsymbol{y}}_{E_2}^{\intercal}\boldsymbol{\vartheta}_{E_2} - n\kappa_2(\boldsymbol{\vartheta}_{E_2})\right]\right\}.\end{aligned} \tag{4}$$

Equation (4) is constructed with the same motivation of composing simplified likelihoods from marginal densities in composite likelihood estimation; see, for example, Cox and Reid (2004); Varin, Reid and Firth (2011). Differently from Massam and Wang (2018), we include contributions for all the bivariate distributions in Equation (4), since the underlying graphical structure is not known a priori, and it is not possible to decide which marginal densities should be included accordingly. Instead, we include all bivariate terms and assign to each component a non-negative weight $w_{E_2} \in \mathbb{R}^+$, controlling the contribution of the E_2 component to the joint likelihood function.

Although it is common to choose unity weights $w_{E_2} = 1$ for each $E_2 \in \mathcal{P}_2$ (e.g. Cox and Reid, 2004), careful choice of composite weights can improve efficiency (Varin, Reid and Firth, 2011). Popular choices focus on selecting weights according to some optimality criteria; for example, to correct the magnitude (Pauli, Racugno and Ventura, 2011) or curvature (Ribatet, Cooley and Davison, 2012) of the likelihood-ratio test or, more generally, to improve statistical efficiency of the resulting estimating equation (e.g. Lindsay, Yi and Sun, 2011; Fraser and Reid, 2019; Pace, Salvan and Sartori, 2019). Beside asymptotic arguments, such procedures are also practically well justified since Equation (4) might include redundant terms, accounting for the same contribution (e.g. marginal univariate) multiple times. This has motivated the development of more efficient likelihood composition, with the focus on producing sparse estimating equations with few informative components by setting some weights to zero via constrained optimisation (Ferrari, Qian and Hunter, 2016; Huang and Ferrari, 2017). In this article, we build on a similar strategy and aggregate the different components under a Bayesian approach, imposing a sparsity-inducing prior on the weights which favours deletion of redundant terms.

Equation (4) can also be motivated from an inferential point of view. When interest focuses on inferences for low-dimensional marginal distributions, such as univariates and bivariates, estimates based on the pseudo likelihood in Equation (4) and the original likelihood in (2) are equivalent, since the joint model is a closed exponential family which includes only first order interactions in the sufficient statistics (Mardia et al., 2009, Theorem 2). With respect to this consideration, it is also worth highlighting that the sufficient statistics $\tilde{\boldsymbol{y}}_{E_2}$ of the simplified model in Equation (3) are actually a subset of the sufficient statistics of the joint model for $\tilde{\mathbf{y}}$ in (2) and that $\bigcup_{E_2 \in \mathcal{P}_2} \tilde{\boldsymbol{y}}_{E_2} = \tilde{\mathbf{y}}$.

Although in a variety of applications the focus of statistical inference is on low-dimensional margins and related measures of association, Equation (4) may be oversimplified and hence lead to a poor characterisation of multivariate categorical data. For example, there may be significant dependence in the data beyond first order. To improve flexibility, we propose to use Equation (4) to characterize variability within subpopulations using a mixture modeling approach. To formalize this, denote with $\boldsymbol{i}_{E_2}$ the elements of $\mathcal{I}_{E_2}$, cells of the E_2-marginal bivariate table. The contribution for a single observation $y_i = (y_{i1}, \dots, y_{ik})$ in Equation (4) can be expressed as

$$\tilde{\mathbf{p}}(y_i; \boldsymbol{\vartheta}, \boldsymbol{w}) = \exp\left\{ \sum_{E_2 \in \mathcal{P}_2} w_{E_2} \left[\mathbb{1}\left(y_i, \boldsymbol{i}_{E_2}\right) \mathbf{X}_2 \boldsymbol{\vartheta}_{E_2} - \kappa_2(\boldsymbol{\vartheta}_{E_2}) \right] \right\}, \tag{5}$$

with $\boldsymbol{\vartheta} = \{\boldsymbol{\vartheta}_{E_2}\}_{E_2 \in \mathcal{P}_2}$, $\boldsymbol{w} = \{w_{E_2}\}_{E_2 \in \mathcal{P}_2}$ and $\mathbb{1}(y_i, \boldsymbol{i}_{E_2})$ corresponding to a vector of length $|\mathcal{I}_{E_2}|$ with a 1 in the position for the cell in which the E_2 component of y_i falls and all other elements 0. We introduce a latent group indicator $z_i \in \{1, \dots, H\}$ with $\mathrm{pr}[z_i = h] = \nu_h$, indexing the subpopulation for the ith subject. We use Equation (4) as a local model for characterizing the dependence structure of subjects in the same latent group. By allowing the weights w_{E_2} to vary across subpopulations, we allow the complexity of the local model to vary substantially and adapt to the subpopulation-specific structure.

Considering only observations belonging to group h and denoting with $n_h = \sum_{i=1}^{n} \mathbb{1}[z_i = h]$ the number of units in group h, we interpret Equation (4) as a model for the contingency table conditional on group membership, as

$$\tilde{\mathbf{p}}(\boldsymbol{y}^h; \boldsymbol{\vartheta}^h, \boldsymbol{w}^h \mid \mathbf{z}) = \exp\left\{ \sum_{E_2 \in \mathcal{P}_2} w^h_{E_2} \left[\tilde{\boldsymbol{y}}^{h\intercal}_{E_2} \boldsymbol{\vartheta}^h_{E_2} - n_h \kappa_2(\boldsymbol{\vartheta}^h_{E_2}) \right] \right\}, \tag{6}$$

where the composite likelihood weights $\boldsymbol{w}^h = \{w^h_{E_2}\}_{E_2 \in \mathcal{P}_2}$ and the log-linear parameters $\boldsymbol{\vartheta}^h = \{\boldsymbol{\vartheta}^h_{E_2}\}_{E_2 \in \mathcal{P}_2}$ are allowed to vary across mixture components $h = 1, \dots, H$ to characterise different dependence patterns in different subpopulations. Marginalising over the latent feature $\mathbf{z}$ and considering the contribution for all the data points, we obtain a joint model with likelihood function equal to

$$\tilde{\mathbf{p}}(\boldsymbol{y}; \boldsymbol{\vartheta}, \boldsymbol{w}, \boldsymbol{\nu}) = \prod_{i=1}^{n} \sum_{h=1}^{H} \nu_h \, \tilde{\mathbf{p}}(y_i; \boldsymbol{\vartheta}^h, \boldsymbol{w}^h), \tag{7}$$

with $\boldsymbol{\vartheta} = \{\boldsymbol{\vartheta}^h\}_{h=1}^H$, $\boldsymbol{w} = \{\boldsymbol{w}^h\}_{h=1}^H$ and $\boldsymbol{\nu} = \{\nu_h\}_{h=1}^H$.

The adaptive log-linear structure imposed within each component of Equation (6) allows one to characterize complex dependence patterns with few components. Increasing the number of components H, any structure can be effectively characterised under MILLS. The following Lemma formalizes the ability of MILLS to represent any $\mathbf{p} \in \mathcal{S}_{|\mathcal{I}_V|}$, with $\mathcal{S}_{|\mathcal{I}_V|}$ denoting the $(|\mathcal{I}_V| - 1)$-dimensional simplex. See Appendix B for a proof.

LEMMA 3.1. *Any* $\mathbf{p} \in \mathcal{S}_{|\mathcal{I}_V|}$ *admits representation (7) for some* H, *with* $\nu_h \in (0, 1)$ *such that* $\sum_{h=1}^{H} \nu_h = 1$.

Equation (7) provides a compact model for efficiently making inference on low-dimensional marginals. For example, a natural estimate for the E_2 bivariate distribution is given by

$$\text{pr}(\hat{\boldsymbol{i}}_{E_2}) = \sum_{h=1}^{H} \nu_h \exp\left\{\mathbf{X}_2 \boldsymbol{\vartheta}_{E_2} - \kappa_2(\boldsymbol{\vartheta}_{E_2})\right\},$$

which corresponds to a weighted average of local estimates, with weights given by the mixture weights.

4. Bayesian inference. We proceed with a Bayesian approach to inference, and specify prior distributions for the parameters $\boldsymbol{\nu}$, $\boldsymbol{\vartheta}^h_{E_2}$ and $\mathbf{w}$. We rely on Dirichlet and Gaussian distributions, letting

(8)
$$(\boldsymbol{\nu} \mid H) \sim \text{DIR}\left(\frac{1}{H}, \dots, \frac{1}{H}\right), \quad (\boldsymbol{\vartheta}^h_{E_2} \mid \sigma^2) \stackrel{\text{iid}}{\sim} \text{N}_{|\mathcal{I}_{E_2}|}(\mu_{E_2}, \sigma^2_{E_2} I), \quad E_2 \in \mathcal{P}_2, \quad h = 1, \dots, H.$$

Estimation for the number of active components is performed by choosing a conservative upper bound H_0 for H, and specifying a sparse Dirichlet distribution on the mixture weights to automatically favour deletion of redundant components (Rousseau and Mengersen, 2011). In practical application, we found that values $H_0 \in [5, 10]$ often provide sufficiently large bounds for the number of mixture components. However, we recommend checking posterior estimates for the number of non-empty groups $\hat{H}$, specifying a larger value H_0 if $\hat{H}$ is close to the upper bound H_0, in order to guarantee that such value is sufficiently large to capture the correct number of components. The Gaussian priors on the log-linear parameters allow simple inclusion of prior information, for example reflecting knowledge on the expected direction and strength of the association between pairs of variables. Moreover, computations are particularly easy adapting the Pòlya-Gamma data-augmentation strategy for the multinomial likelihood and Gaussian prior (Polson, Scott and Windle, 2013). Under an exponential family representation, other conjugate priors are available for the natural parameters (e.g. Massam et al., 2009; Bradley, Holan and Wikle, 2019). However, Gaussian priors have simpler interpretation and facilitate computation.

As motivated in Section 3, the prior distribution for the composite weights $w^h_{E_2} \in \mathbb{R}^+$ should induce sparse configurations, deleting redundant components. To address this with computational tractability, we rely on a continuous spike and slab prior. Such a strategy focuses on introducing latent binary indicators $\delta^h_{E_2} \in \{0, 1\}$ encoding exclusion or inclusion of the E_2 component in (4), with $\text{pr}[\delta^h_{E_2} = 1] = \gamma^h_0$. Conditionally on $\delta^h_{E_2}$, each $w^h_{E_2}$ is drawn independently either from a distribution concentrated around zero, P_0, or from a diffuse distribution over the real positive line, which we denote as P_1. For computational convenience, we rely on the following hierarchical specification for $w^h_{E_2}$.

(9)
$$\begin{aligned} (\delta^h_{E_2} \mid \gamma^h_0) &\stackrel{\text{iid}}{\sim} \text{BERNOULLI}(\gamma^h_0) \\ (w^h_{E_2} \mid \delta^h_{E_2}) &\stackrel{\text{iid}}{\sim} \text{GAMMA}(1 + a^h_0 \delta^h_{E_2}, a^h_1), \quad E_2 \in \mathcal{P}_2, \quad h = 1, \dots, H \end{aligned}$$

Although it is possible to replace the spike with a Dirac mass at 0, we follow Ishwaran et al. (2005), and introduce a continuous shrinkage prior, which is shown to generally improve computation and mixing; see also Legramanti, Durante and Dunson (2020) for related arguments.

Marginalising out $\delta^h_{E_2}$ from (9), we obtain a discrete mixture between a Gamma distribution with shape 1 and rate a^h_1 (Exponential), and a Gamma distribution with shape $(1 + a_0)$

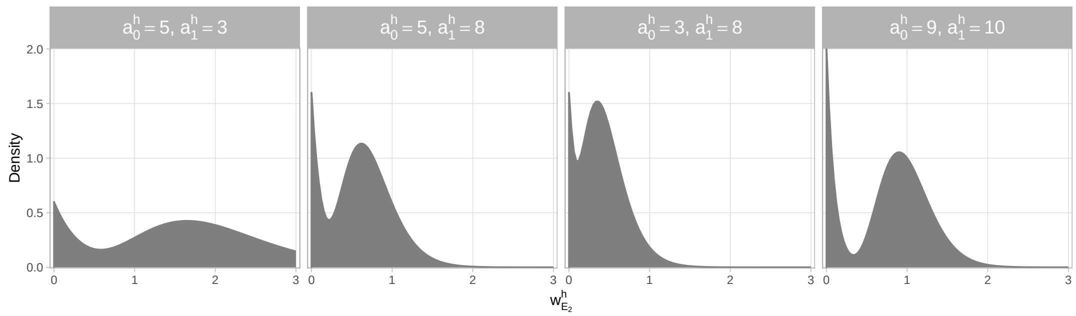

FIG 1. *Graphical illustration of the prior distribution of Equation 9 for different hyper-parameter values. In each panel,* $\gamma_0^h = 0.2$.

and rate a_1^h. The parameter γ_0 controls the prior proportion of active terms, and is assigned a symmetric BETA$(0.5, 0.5)$ prior (Ishwaran et al., 2005). Specifying large values for a_1^h, substantial mass around 0 is induced, while a_0^h controls the mean and variance for the Gamma distribution associated with the slab. See Figure 1 for a graphical illustration of the prior density over illustrative combinations of hyper-parameters. In the absence of explicit prior information on the composite likelihood weights, we recommend to elicit the prior distribution to include values around 1 with high probability in the slab component. Such choice guarantees that, when a component is included, default units weights are selected with high probability a priori, centering the model around a standard specification.

4.1. *Posterior computation.* There is a rich literature on the use of alternative likelihoods for Bayesian inference; for example, approximate likelihood (Efron, 1993), partial likelihood (Raftery, Madigan and Volinsky, 1995), empirical likelihood (Lazar, 2003) and adjusted profile likelihood (Chang and Mukerjee, 2006), among many others. See also Greco, Racugno and Ventura (2008) for related arguments. Although the use of composite likelihoods in Bayesian inference is more recent (e.g. Ribatet, Cooley and Davison, 2012; Pauli, Racugno and Ventura, 2011), it has received substantial attention (Miller, 2019). Related to these approaches, we conduct inference using the composite posterior distribution

$$\tilde{\pi}(\boldsymbol{\vartheta}, \boldsymbol{\nu} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\vartheta})\pi(\boldsymbol{\nu})\pi(\boldsymbol{w})\tilde{\mathbf{p}}(\boldsymbol{y}; \boldsymbol{\vartheta}, \boldsymbol{w}, \boldsymbol{\nu}). \tag{10}$$

Since the composite likelihood function $\mathbf{p}(\boldsymbol{y}; \boldsymbol{\vartheta}, \boldsymbol{w}, \boldsymbol{\nu})$ is not a proper distribution function, it is important to guarantee that the pseudo-posterior (10) is proper (Ribatet, Cooley and Davison, 2012). The following Lemma shows that our composite posterior does have this property. See Appendix B for a proof.

LEMMA 4.1. $\tilde{\pi}(\boldsymbol{\vartheta}, \boldsymbol{\nu} \mid \boldsymbol{y})$ *is a proper probability distribution.*

To make inference from (10), we rely on an MCMC algorithm whose main steps are described in Appendix C. We leverage the Pòlya-Gamma data augmentation strategy of Polson, Scott and Windle (2013) to obtain conditionally conjugacy between the Gaussian prior and the multinomial likelihood, while the mixture weights $\boldsymbol{\nu}$ and composite weights $\boldsymbol{w}$ are updated sampling from Dirichlet and Gamma full conditional distributions, respectively. Similarly, the mixture indicator z_i is sampled from its full conditional categorical distribution, for each $i = 1, \dots, n$. The main bottleneck is storage of the conditional bivariate terms, which have size $\mathcal{O}(Hk^2d^2)$. Although the introduction of the spike and slab strategy drastically improves estimation — since many components are effectively assigned to zero weight at each iteration and Equation (4) involves only few informative components — the storage of redundant terms is required during estimation and can be burdensome. However, the proposed algorithm easily scales up in our motivating application, relying on a mixed R and C++ implementation on a standard laptop; see Section 6. Scaling to much larger cases can potentially be accomplished by replacing the continuous spike with a mass at zero or thresholding redundant components as an approximation.

5. Simulation Study. In order to evaluate the model performance, we considered a simulation study over four different settings. In each scenario, we focus on an artificial sample of size $n = 400$, with $k = 15$ categorical variables and $d_1 = \dots\ d_{15} = 4$ categories. In the first scenario, multivariate categorical data are generated from a latent class model with $H = 5$ components and probabilities generated from a uniform prior on the simplex. The second scenario samples categorical variables $j \in \mathcal{J} = (1, 2, 3, 4, 5)$ from a dense log-linear model with first order interactions and coefficients randomly sampled from a Gaussian distribution with standard deviation 0.1, while the remaining categorical variables $j \notin \mathcal{J}$ are generated from independent Dirichlet-Multinomial distributions with hyper-parameter $(3, 3, 3, 3)$. In the third scenario, we focus on the same groups of variables, imposing more structure on the variables in the group $\mathcal{J}$, which are sampled from the joint probability mass function assigning probability 0.1 to the cells $\boldsymbol{i}_{\mathcal{J}} \in \{(1, \dots, 1), \dots, (4, \dots, 4)\}$ and probability 0.6 to the remaining cells in equal proportion; see also Russo, Durante and Scarpa (2018). The remaining variables $j \notin \mathcal{J}$ are generated from independent Dirichlet-Multinomial distributions with hyper-parameter $(3, 3, 3, 3)$. The fourth and last scenario further complicates the second one by introducing an additional group of variables $\mathcal{J}' = (5, 6, 7, 8, 9, 10)$, generated from a dense hierarchical log-linear model with first and second order interactions, and coefficients randomly sampled from a Gaussian distribution with standard deviation 0.1.

The focus of these settings is on inducing challenging data generating processes, characterised by heterogeneous dependence across subsets of categorical variables. Posterior inference for MILLS relies on 1000 iterations collected after a burn-in period of 1000, setting a conservative upper bound $H = 5$ and specifying $\mu^h_{E_2} = 0$, $\sigma^2_{E_2} = 3$ and $a^h_0 = 10, a^h_1 = 10$, with $h = 1, \dots, H$ and $E_2 \in \mathcal{P}_2$. Trace plots and MCMC diagnostics indicate good mixing in all the settings considered. As competitor approaches, we considered two flexible latent variable models, whose estimation is feasible in the settings under investigation. The first is a Bayesian specification of a latent class model with $H = 10$ classes, sparse Dirichlet priors over the mixture weights and unit Dirichlet priors on the class-specific probabilities. Such an approach corresponds to a finite mixture of product multinomial distributions; see, for example, Fruhwirth-Schnatter, Celeux and Robert (2019, Chapter 9) for an introduction. The second competitor is a simplex factor model (Bhattacharya and Dunson, 2012) with $H = 10$ latent factors. This approach can be interpreted as a mixed membership model (e.g. Airoldi et al., 2014) for multivariate categorical data. Specifically, the observed categorical variables are modeled as conditionally independent given a vector of subject-specific latent attributes lying on the simplex. Such latent features can be interpreted as the subject-specific partial membership to H extreme profiles, with each individual partially belonging to each extreme profile, to a different degree; see also Manrique-Vallier (2014) for a similar specification with longitudinal survey data. Again, we rely on a Bayesian specification relying on independent Dirichlet priors over the model parameters. As outlined in Section 1.1, both approaches induce a parsimonious low-rank decomposition of the probability mass function, and the connection between such decompositions and a log-linear model specification has been explored in Johndrow, Bhattacharya and Dunson (2017).

The focus of the simulations is on evaluating the ability of the approaches in estimating low-dimensional functionals of the data. We focus on the set $\mathcal{P}_2$ of bivariate distributions,

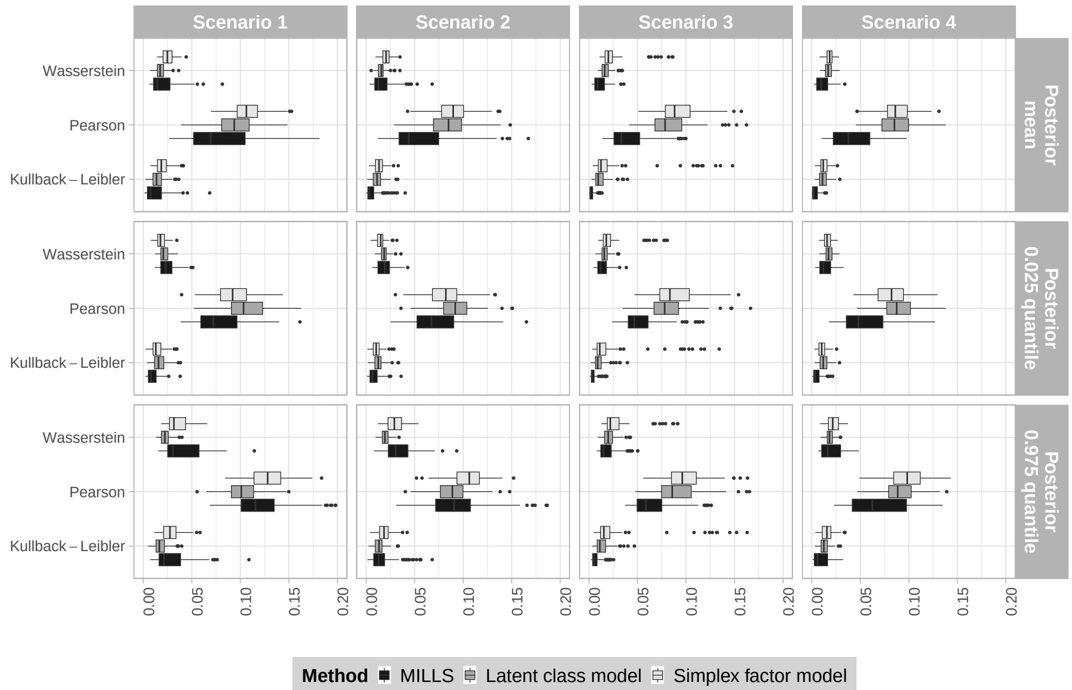

FIG 2. *Simulation studies. Wasserstein distance, normalised Pearson's residuals and absolute Kullback-Leibler divergence between estimates and observed quantities. First row refers to posterior means; second and third to posterior* 0.025 *and* 0.975 *quantiles, respectively. Black boxplots refer to* MILLS*. Gray and light-gray to latent class model and simplex factor model, respectively.*

whose precise estimation is crucial for computing measures of bivariate associations and making inference on the dependence structure. Figure 2 illustrates the variability across $\mathcal{P}_2$ under the four simulations settings and for the three approaches considered. The first row of Figure 2 shows estimated posterior mean for the three methods, compared with their empirical counterparts in terms of Kullback-Leibler divergence, Wasserstein distance and normalised Pearson's residuals.

The first column of Figure 2 illustrates results for the first scenario, and suggests that when data are generated from a latent class model, the three approaches are comparable in terms of goodness of fit, with MILLS resulting in predictions which are more accurate on average, but also more variable. The good performance of the latent class model was expected, since such an approach is correctly specified in the first scenario. As outlined in Section 3, MILLS can induce a latent class specification as a special case, and therefore its performance is on average similar with the competitors, but also characterized by a higher variability which might be due to the estimation of the richer dependence structure imposed within each mixture component. In the second and third scenario, results indicate the superiority of MILLS with respect to the latent class model and the simplex factor model. Such a result highlights the ability of the proposed approach to adapt to settings with heterogeneous dependence patterns across subsets of variables; the third column of Figure 2, in addition, confirms how MILLS achieves better performance than the competitors also when such dependence patterns go beyond first order interactions. Lastly, the fourth scenario illustrates the ability of MILLS to adapt better than the competitors to highly complex settings, dependence patterns beyond first order interactions and involving multiple sub-groups of variables. The superiority of MILLS in such settings might be due to the parsimonious composite likelihood specification of Equation (4), with adaptive estimation of the degree of dependence required by each component.

TABLE 2
Additional simulation study. Root mean squared error (RMSE) of the posterior mean estimator and coverage of 90% credible intervals across three simulation scenarios. Values are averaged across 100 replications.

	RMSE			COVERAGE, 90%		
	Cramer-V	H	$\boldsymbol{\vartheta}$	Cramer-V	H	$\boldsymbol{\vartheta}$
2 classes	0.043	0.021	0.000	0.871	0.860	0.910
5 classes	0.038	1.002	0.001	0.890	0.880	0.903
10 classes	0.069	2.013	0.012	0.903	0.890	0.880

Variability in the simulations is assessed considering the posterior 0.025 and 0.975 quantiles of the estimated bivariate distributions, graphically reported for each method in the second and third row of Figure 2 respectively. The main empirical findings are consistent with the discussion outlined above, indicating an overall better performance of MILLS under complex data generating processes.

5.1. *Additional simulations studies.* As suggested by an anonymous Referee, we conducted an additional simulation study to evaluate the performance of MILLS in estimating functionals of primary interest in our application. These quantities correspond to the dependence structure among the items, to the number of subpopulations and their specific structure, and they can be estimated using the posterior distribution for the Cramer-V, the number $\hat{H}$ of non-empty groups, and the group-specific parameters $\boldsymbol{\vartheta}^h$, respectively. We estimate these functionals via Monte Carlo integration, post-processing the MCMC sample to obtain point and interval estimates via posterior means and quantile-based credible intervals, respectively.

The simulation focuses on three additional settings characterized by the same sample size $(n = 56)$ and number of categorical variables $(k = 57)$ as in our motivating application, sampling categorical variable with $d = 5$ from a latent class model with $H = 2$, $H = 5$ and $H = 10$ groups, respectively, and probabilities generated from a uniform prior on the simplex. Each setting is replicated 100 times using different random seeds, and in each replication posterior inference for MILLS relies the same settings as in Section 5, increasing the upper bounds on the mixture components to $H_0 = 10$.

In Table 2, we evaluate the Root Mean Squared Error (RMSE) of the posterior mean and assess coverage of 90% quantile-based credible intervals. The first part of Table 2 reports the RMSE between the posterior mean and the functionals of interest, and results indicate that MILLS accurately estimates these quantities in simulations. Estimation for the number of components might be biased due to the more intricate structure introduced by MILLS, which requires fewer component than a latent class model to characterize the data. The second part of Table 2 focuses on the coverage of 90% credible intervals, and results indicate that intervals have a coverage close to the nominal level for all functionals of interest. As outlined in Section 3 and in Ribatet, Cooley and Davison (2012) and Pauli, Racugno and Ventura (2011), it is important to carefully weight each likelihood component to reduce the under coverage of credible intervals constructed from unadjusted composite-likelihood specifications. Since MILLS adjusts each component with a positive weight $w^h_{E_2}$, we do not observe signs of significant under coverage. Coverage can be potentially improved introducing a further level of adjustment to explicitly control for the curvature of the asymptotic distribution of the posterior; see Ribatet, Cooley and Davison (2012) for further arguments.

6. MILLS for psychopathological associations. We applied MILLS on the data described in Section 1. Posterior inference for MILLS uses the same specification as in the simulations, relying on 3000 iterations collected after a burn-in of 1000. Posterior computation requires approximately 7 minutes per 100 iterations and 4GB of RAM on a laptop with an

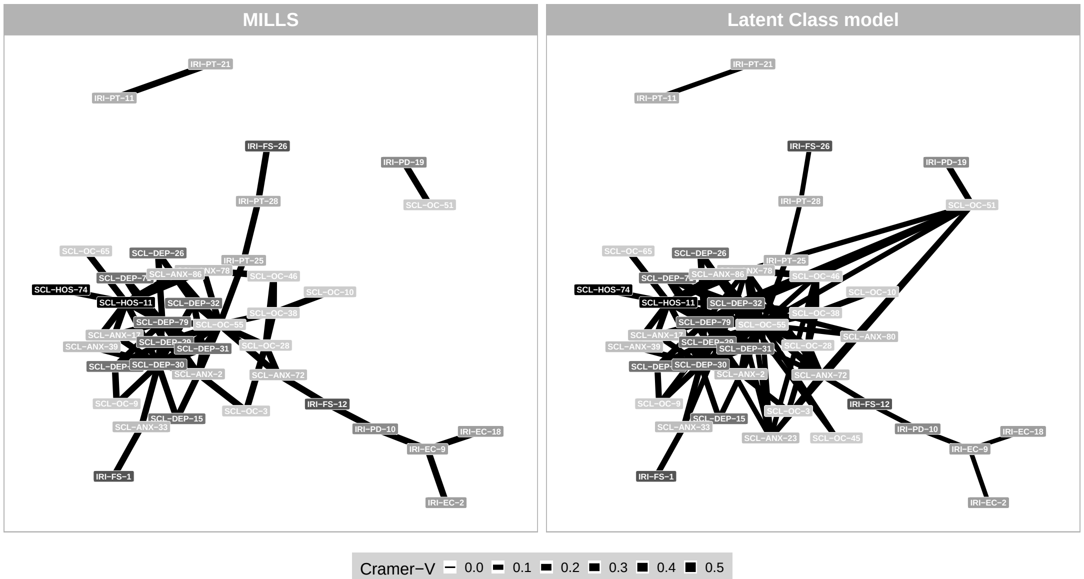

FIG 3. *Association structure of the items. Color of the labels varies with subscales, while edge widths vary with the value of the posterior mean of the pairwise Cramer-V.*

INTEL(R) CORE(TM) I7-7700HQ @ 2.8 GHZ processor running Linux. We conducted sensitivity analysis for different hyper-parameter specifications, replicating posterior computation with values $H_0 \in \{10, 15\}$, $a_0^h \in \{10, 100, 1000\}$, $a_1^h \in \{10, 100, 1000\}$ and $\sigma^2_{E_2} \in \{3, 10\}$. The overall empirical findings were robust across changes in hyper parameters.

Posterior inference focuses on bivariate associations measured via the Cramer-V, which can be easily computed via Monte Carlo integration leveraging the MCMC output. Figure 3 illustrates the dependence structure as a graph, with nodes corresponding to the categorical variables and edges to their associations, with thicker edges corresponding to stronger associations and higher Cramer-V. The left panel of Figure 3 refers to MILLS, and the right panel to a latent class model with $H = 10$ components and the same specification as in the simulations. In order to improve graphical visualisation, we have removed from the graph the items whose largest associations is below 0.1.

Our empirical findings highlight the presence of strong associations across several subscales, in particular within items associated with similar profiles. For example, the bulk of central nodes in Figure 3 denote items associated with depressive (SCL-DEP) and obsessive compulsive profiles (SCL-OC), suggesting significant interconnections within these two subscales. Similarly, items corresponding to the Empathic Concern (EC) subscale have different associations among them, and with other empathic subscales. To some extent, this result confirms the validity of the tools to measure psychopathological symptoms, which characterize consistent psychological profiles and highlights that such traits are strongly associated in suicide attempt survivors. In addition, some items corresponding to different profiles measured within the same questionnaire are characterized by strong interactions. For example, the empirical findings indicate an association between an anxious subject SCL-ANX-2 (*"Nervousness or shakiness inside"*) and SCL-DEP-15 (*"Thoughts of ending your life"*) in suicide attempt survivors. Similarly, we observe an association between IRI-EC-9 (*"When I see someone being taken advantage of, I feel kind of protective towards them."*) and IRI-PD-10 (*"I sometimes feel helpless when I am in the middle of a very emotional situation."*), which indicate how patients under investigation feel empathic to others, in particular in stressful situations.

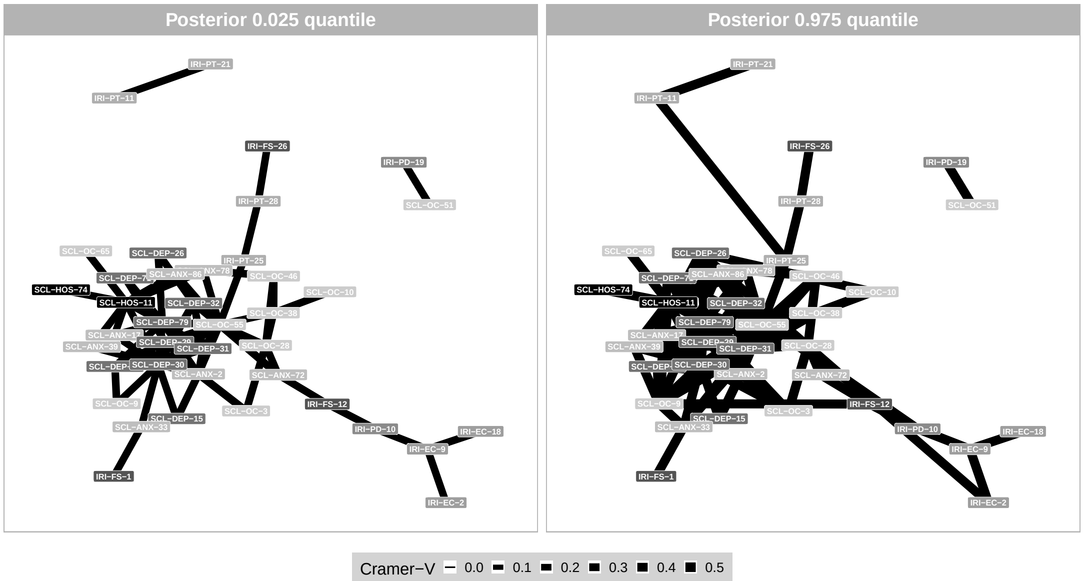

FIG 4. *Posterior quantiles of the pairwise Cramer-V under* MILLS

Other interesting associations involve items in different subscales. For example, there is an association between an item from the IRI questionnaire IRI-FS-1 (*“I daydream and fantasize, with some regularity, about things that might happen to me”*) with the item SCL-ANX-33 (*“Feeling fearful”*), and also the SCL-DEP-30 item (*“Feeling blue”*). This dependence structure is coherent with a paranoid profile, with fantasies about things that might happen and with such thoughts inducing substantial fear and sadness. Another interesting association involves the items SCL-OC-51 (*“Your mind going blank”*) and IRI-PD-19 (*“I am usually not effective in dealing with emergencies.”*), which are consistent with a profile with low-capacity to handle complex situations with calm. Panels of Figure 4 assess uncertainty in MILLS estimation considering the 0.025 and 0.975 posterior quantiles of the Cramer-V, and suggesting that the estimated structure is maintained considering such posterior summaries. These interconnections are further explored in Table 3, which reports the posterior means and 95% credible intervals for the Cramer-V referring to different bivariate associations of interest. Current empirical findings confirm the presence of strong associations within the depressive symptoms subscale (SCL-DEP) and between SCL-DEP and obsessive-compulsive subscale (SCL-OP). Worth mentioning are also the associations between the perspective-taking (IRI-PT) and other empathic components, as well as the already mentioned association between obsessive compulsive symptoms and personal distress. These results provide an overview of the dependence structure characterizing the psychopathology of suicide attempt survivors, highlighting the interdependence among psychological symptoms and empathic profiles.

Results from a latent class model on the overall association structure – reported in the right panel of Figure 3 – are roughly consistent with inference based on MILLS, suggesting dense associations among items related to the same pychopathologies. However, this approach required a larger number of mixture components to adequately characterise the data under investigation; see Table 4, where the posterior medians of the mixture weights under both approaches are reported, suggesting evidence of 2 non-empty components for MILLS and 5 for the latent class model. As discussed in Section 1.1, this result might be due to the richer structure imposed by MILLS within each subpopulation, which is expected to reduce the number of components required to characterize higher order dependencies.

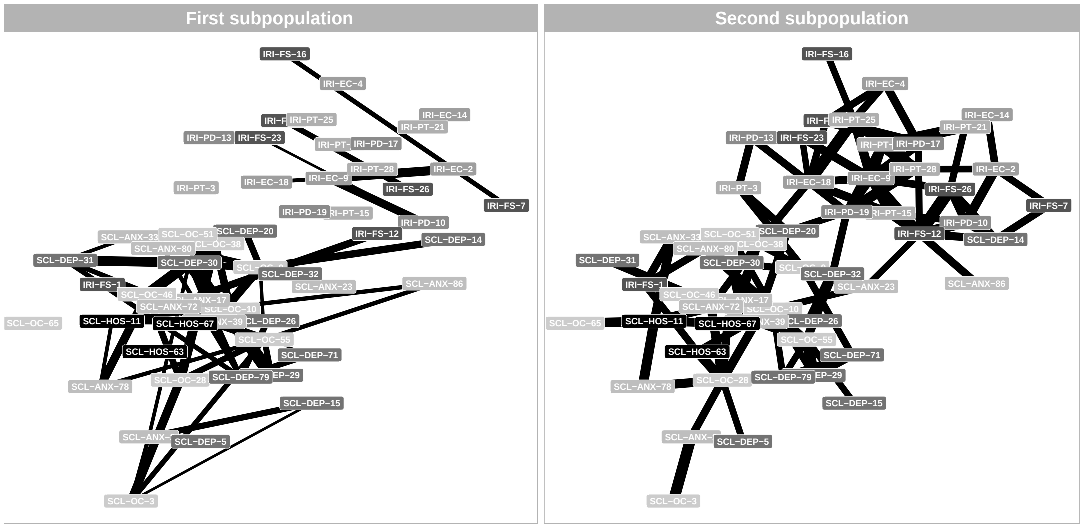

FIG 5. *Posterior means of the pairwise Cramer-V under* MILLS *for the two estimated subpopulations*

This property leads to relevant practical implications for the analysis of our motivating application. For example, when interest is on characterizing profiles specific to each subpopulation, inference for latent class models would focus on evaluating the parameters within each non-empty component, describing the univariate response patterns of the individuals belonging to that specific latent group (e.g., McHugh, 1956). Inference on other relevant quantities, such as the association structure within each component, is not possible under a standard latent class model, due to the independence assumption of the items conditionally on the group membership. Instead, under the proposed MILLS, we can easily conduct inference on such association structures, effectively characterising the interactions between psychopathological symptoms and empathic traits in each subpopulation.

Figure 5 compares the posterior means of the Cramer-V across items, within each of the two non-empty subpopulations – according to results summarized in Table 4. Associations reported in the left panel of Figure 5 refer to the first latent group, and highlight several connected psychopathological symptoms, in particular within depression and anxiety traits.

TABLE 3
Bivariate Cramer-V. Posterior means and 95% credibile intervals.

		Cramer-V	95% CI
SCL-DEP-29	SCL-DEP-30	0.471	[0.375, 0.570]
SCL-DEP-29	SCL-OC-55	0.428	[0.335, 0.519]
SCL-DEP-30	SCL-DEP-32	0.424	[0.318, 0.527]
SCL-DEP-30	SCL-DEP-31	0.422	[0.324, 0.522]
SCL-OC-38	SCL-OC-46	0.410	[0.304, 0.516]
SCL-ANX-2	SCL-OC-3	0.402	[0.306, 0.498]
SCL-OC-55	SCL-ANX-72	0.398	[0.294, 0.508]
SCL-ANX-2	SCL-DEP-15	0.391	[0.298, 0.494]
SCL-DEP-30	SCL-ANX-33	0.389	[0.297, 0.486]
SCL-OC-51	IRI-PD-19	0.381	[0.278, 0.476]
SCL-OC-9	SCL-DEP-14	0.378	[0.281, 0.472]
IRI-EC-9	IRI-PD-10	0.367	[0.271, 0.465]
IRI-PT-11	IRI-PT-21	0.362	[0.270, 0.465]
IRI-EC-2	IRI-EC-9	0.358	[0.262, 0.450]
IRI-FS-1	SCL-ANX-33	0.357	[0.275, 0.453]

TABLE 4
Posterior medians (and standard deviations) for the mixture weight parameters. Values are sorted in decreasing order. Results for the latent class approach are reported until the first empty group.

	$\hat{\nu}_1$	$\hat{\nu}_2$	$\hat{\nu}_3$	$\hat{\nu}_4$	$\hat{\nu}_5$	$\hat{\nu}_6$
Latent Class	0.530 (0.065)	0.208 (0.055)	0.157 (0.048)	0.053 (0.031)	0.030 (0.023)	0.000 (0.004)
MILLS	0.671 (0.089)	0.318 (0.088)	0.000 (0.008)	0.000 (0.008)	0.000 (0.007)	–

Items measuring empathic profiles, instead, show a more sparse structure in the first subpopulation, indicating strong associations only across few items. The second subpopulation (right panel of Figure 5), is instead characterized by more interconnected associations, both in terms of empathic profiles and psychopathological symptoms. Although many items are similarly associated across the subpopulations, it is interesting to observe that some association patterns deviate across groups. For example, SCL-DEP-14 (*"Feeling low in energy or slowed down"*) is associated with obsessive compulsive symptoms in the first subpopulation (SCL-OC-9, *"Trouble remembering things"*), while in the second group it is linked with empathic profiles (e.g., IRI-PD-10, *"I sometimes feel helpless when I am in the middle of a very emotional situation"*). Similarly, different anxiety symptoms (SCL-ANX-86, *"Feeling pushed to get things done"* and SCL-ANX-23, *"Suddenly scared for no reason"*) are associated with some psychopathological items in the first subpopulation (SCL-OC-10, *"Worried about sloppiness or carelessness"*) and with empathic items in the second (IRI-FS-12, *"Becoming extremely involved in a good book or movie is somewhat rare for me"*) .

These aspects are further detailed in Table 5, which reports the posterior means and credible intervals for the Cramer-V for a subset of bivariate distributions, separately across the two subpopulations. Subjects in the first group are characterized by several associations across different SCL-90 items, in particular with respect to depressive and obsessive compulsive symptoms, reporting posterior means for the bivariate Cramer-V above 0.4. The structure across empathic items indicates instead interesting interconnections across the fantasy scale and between fantasy and obsessive-compulsive symptoms. The second group characterizes latent profiles more driven by empathic aspects, in particular referring to the IRI-PT subscale, and items measuring depressive, obsessive-compulsive and anxiety symptoms.

These information, combined with the results in Table 4, provide a richer interpretation of the psychology underling suicide attempt survivors. Subjects in the first profile show indications of high mental distress, characterized by important associations across severe psychopathological symptoms. The estimated proportion of the population in this class is 0.67 (first column of Table 4), so that the majority of the suicide attempt survivors belong to this group. The second profile is associated with roughly a third of the population (0.32, second column of Table 4) and differs from the first one reporting more dense associations across empathic aspects. Therefore, patients in this subpopulation are characterized by a psychopathology more driven by the emotional and cognitive components of empathy.

Such results indicate that the patients under investigation are characterized by different latent profiles that vary in terms of the association structure between psychopathological symptoms and empathic traits. Also, investigation of the subpopulation specific structure indicates that the proposed approach has concrete advantages over a latent class specification, since it allows investigation of the association structure characterising different subpopulations, providing additional insights on the psychology of suicide attempt survivors. These findings suggest that empathy and psychotic symptoms are deeply related in the characterisation of the psychosis of suicide attempt survivors, and deserve further attention.

6.1. *Model checking.* In order to check if MILLS provides a reasonable representation of the observed psychological data, we follow the approach illustrated in Section 5 and rely

TABLE 5
Bivariate Cramer-V. Posterior means and 95% credibile intervals for the two estimated subpopulations

			CRAMER-V	95% CI
FIRST GROUP	SCL-DEP-29	SCL-DEP-30	0.491	$[0.377, 0.603]$
	SCL-OC-38	SCL-OC-46	0.463	$[0.343, 0.570]$
	SCL-OC-28	SCL-OC-38	0.447	$[0.340, 0.555]$
	SCL-DEP-30	SCL-DEP-31	0.428	$[0.335, 0.549]$
	SCL-OC-55	SCL-ANX-72	0.403	$[0.320, 0.520]$
	IRI-EC-9	IRI-PD-10	0.379	$[0.280, 0.506]$
	IRI-FS-16	SCL-FS-7	0.363	$[0.284, 0.487]$
	IRI-FS-12	SCL-OC-9	0.363	$[0.284, 0.487]$
	SCL-OC-9	SCL-DEP-14	0.352	$[0.264, 0.418]$
	SCL-OC-9	SCL-DEP-31	0.337	$[0.201, 0.425]$
SECOND GROUP	IRI-PT-11	IRI-PT-28	0.486	$[0.345, 0.611]$
	IRI-FS-12	IRI-FS-26	0.482	$[0.339, 0.625]$
	IRI-EC-18	IRI-PT-25	0.472	$[0.332, 0.606]$
	IRI-PT-25	IRI-PT-28	0.466	$[0.346, 0.572]$
	SCL-ANX-2	SCL-OC-3	0.455	$[0.320, 0.545]$
	SCL-DEP-29	SCL-OC-55	0.446	$[0.326, 0.509]$
	SCL-DEP-26	SCL-DEP-32	0.419	$[0.284, 0.512]$
	SCL-OC-38	SCL-OC-51	0.381	$[0.302, 0.498]$
	IRI-PD-10	IRI-FS-12	0.370	$[0.271, 0.415]$
	IRI-EC-9	IRI-FS-12	0.353	$[0.262, 0.421]$

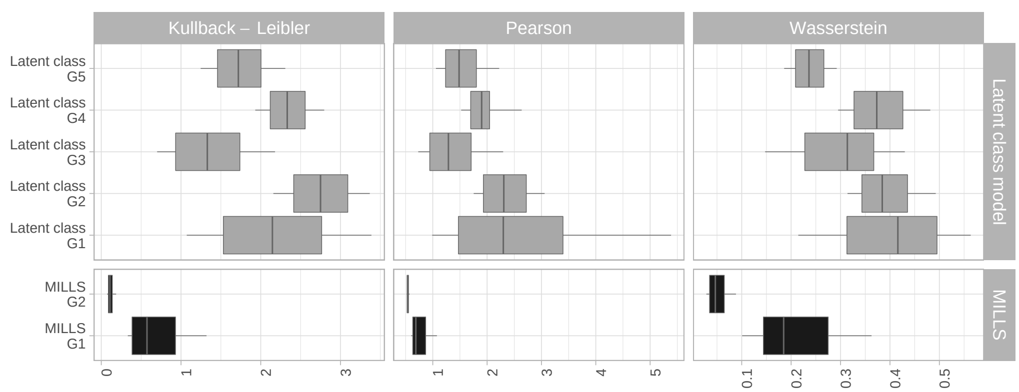

FIG 6. *Absolute Kullback-Leibler, normalised Pearson's residuals and Wasserstein distance between estimated and observed bivariate distributions. Black and grey boxplots refer to* MILLS *and a latent class model, respectively.*

on posterior checks to validate our model (e.g. Gelman et al., 2013). Specifically, MILLS assumes that conditionally on the group membership, the specification in Equation 6 provides a flexible characterization of the psychopathological patterns characterizing the subpopulation. We are therefore interested to measure if such group-specific structures are adequately accounted for, comparing the posterior predictive distribution for a functional of interest with its empirical value. We will focus on the posterior predictives for the bivariate distributions, conditionally on the subpopulation membership, for MILLS and the latent class model.

According to Table 4, posterior inference provides evidence for two subpopulations for MILLS and five for the latent class models. Figure 6 illustrates the Kullback-Leibler divergence, normalized Pearson's residuals and Wasserstein distance between the observed and estimated population-specific bivariate distributions, focusing on the subpopulations estimated by MILLS and the latent class model. Current empirical findings suggest that MILLS provides a good fit for both the subpopulation specific structures, providing estimates for the bivariate distributions that are close to their empirical counterparts, and with more accurate results for the second subpopulation. In addition to estimating a larger number of subpopulations, the latent class model is also characterized by an overall worse fit within each group, likely due to the conditional independence assumption across items which is not met in practice.

7. Discussion. Motivated by a case study on suicide attempt survivors, this article has proposed a new approach for the analysis of categorical data relying on a mixture of log linear models, with a computationally convenient composite likelihood-type specification facilitating implementation. Although multivariate categorical data are very commonly collected in many different areas, we still lack methods for doing inferences on associations among variables in a flexible manner that can accommodate more than a small number of variables. Current log-linear models do not scale up to large contingency tables and latent structure methods sacrifice some of the key advantages of log-linear models in terms of providing a direct and interpretable model on the association structure. Hence, latent structure models are in some sense too black box and unstructured, potentially leading to a non-parsimonious characterization of the data, and necessitating a moderately large number of latent components.

The goal of the proposed framework is to borrow the best of both worlds between latent structure and log linear models. The proposed methods have shown practical advantages in our motivating application, highlighting the presence of clinically interesting associations between psychopathological symptoms and empathy in suicide attempt survivors. There are many interesting next steps in terms of including further computational simplifications to facilitate scaling up, and to include more complex data structure which are routinely collected in psychological studies; for example, having missing data or mixed measurement scales. Also, it is of substantial interest to develop a formal testing procedure based on MILLS to assess whether psychiatric patients that did not attempt suicide differ in terms of psychopathologies from patients under investigation.

APPENDIX A: ITEMS DETAILS

Table 6 and 7 report, respectively, the description of the items included in the analysis. Subject respond to the questions with their level of agreement, with 0 indicating “Not at all“ and 4 indicating “Extemely”. Items were selected according to the subscale they belong to – reported in the second column of Table 6 and 7 – as suggested by our clinician collaborators.

APPENDIX B: PROOFS

PROOF OF LEMMA 3.1. The proof for the full generality of MILLS relies on illustrating how such a specification induces a finite mixture of independent multinomial distributions as a special case. Without loss of generality, consider equal number of categories $d_j = d$ for $j = 1, \dots, k$ and equal weights $\bar{w}^h_{E_2} = 1/(k-1)$ for $E_2 \in \mathcal{P}_2$ and $h = 1, \dots, H$. Introduce a set of constrained log-linear coefficients $\bar{\boldsymbol{\vartheta}}^h_{E_2}$ as $\bar{\boldsymbol{\vartheta}}^h_{E_2} = \mathbf{L} \otimes \boldsymbol{\vartheta}^h_{E_2}$, where $\mathbf{L}$ denotes a vector of length d^2 with the first $1 + k(d-1)$ elements equal to 1 and the remaining 0, and with $\otimes$ denoting element-wise product. Therefore, each $\bar{\boldsymbol{\vartheta}}^h_{E_2}$ induces a log-linear independence model, which includes only main effects. Under the above constraints,

$$\sum_{h=1}^{H} \nu_h \exp\left\{ \sum_{E_2 \in \mathcal{P}_2} \bar{w}^h_{E_2} \left[\mathbf{X}_2 \bar{\boldsymbol{\vartheta}}^h_{E_2} - \kappa_2(\bar{\boldsymbol{\vartheta}}^h_{E_2}) \right] \right\}, \tag{11}$$

TABLE 6
SCL-90 *subscales.*

ID		SUBSCALE
2.	Nervousness or shakiness inside	(ANX)
3.	Unwanted thoughts, words, or ideas that won't leave your mind	(OC)
5.	Loss of sexual interest or pleasure	(DEP)
9.	Trouble remembering things	(OC)
10.	Worried about sloppiness or carelessness	(OC)
11.	Feeling easily annoyed or irritated	(HOS)
14.	Feeling low in energy or slowed down	(DEP)
15.	Thoughts of ending your life	(DEP)
17.	Trembling	(ANX)
20.	Crying easily	(DEP)
22.	Feeling of being trapped or caught	(DEP)
23.	Suddenly scared for no reason	(ANX)
26.	Blaming yourself for things	(DEP)
28.	Feeling blocked in getting things done	(OC)
29.	Feeling lonely	(DEP)
30.	Feeling blue	(DEP)
31.	Worrying too much about things	(DEP)
32.	Feeling no interest in things	(DEP)
33.	Feeling fearful	(ANX)
38.	Having to do things very slowly to insure correctness	(OC)
39.	Heart pounding or racing	(ANX)
45.	Having to check and double-check what you do	(OC)
46.	Difficulty making decisions	(OC)
51.	Your mind going blank	(OC)
55.	Trouble concentrating	(OC)
63.	Having urges to beat, injure, or harm someone	(HOS)
65.	Having to repeat the same actions such as – touching, counting, washing	(OC) ·
67.	Having urges to break or smash things	(HOS)
71.	Feeling everything is an effort	(DEP)
72.	Spells of terror or panic	(ANX)
74.	Getting into frequent arguments	(HOS)
78.	Feeling so restless you couldn't sit still	(ANX)
79.	Feelings of worthlessness	(DEP)
80.	Feeling that familiar things are strange or unreal	(ANX)
86.	Feeling pushed to get things done	(ANX)

corresponds to a discrete mixture of product multinomial distribution, for which Theorem 1 of Dunson and Xing (2009) follows directly, after noticing that

$$\psi_h^{(j)} = \mathbf{M} \prod_{E_2 \in \mathcal{P}_2 : j \in E_2} \left[\exp\left(\mathbf{X}_2 \bar{\boldsymbol{\vartheta}}^h_{E_2} - \kappa_2(\bar{\boldsymbol{\vartheta}}^h_{E_2}) \right) \right]^{\bar{w}^h_{E_2}}, \tag{12}$$

where $\mathbf{M}$ denotes a $d \times d^2$ marginalisation matrix, comprising zeros and ones in appropriate positions (e.g. Lupparelli, Marchetti and Bergsma, 2009). □

PROOF OF 4.1. In order to show that (10) is a proper probability distribution, it is necessary to show that the normalising constant is finite, which correspond to showing that

$$\int\int \pi(\boldsymbol{\vartheta})\pi(\boldsymbol{\nu})\pi(\boldsymbol{w})\tilde{\mathbf{p}}(\boldsymbol{y}; \boldsymbol{\vartheta}, \boldsymbol{w}, \boldsymbol{\nu}) \mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\nu}\mathrm{d}\boldsymbol{w} = \tag{13}$$

$$\int\int \pi(\boldsymbol{\vartheta})\pi(\boldsymbol{\nu})\pi(\boldsymbol{w}) \prod_{i=1}^{n} \sum_{h=1}^{H} \nu_h \, \tilde{\mathbf{p}}(y_i \mid \boldsymbol{\vartheta}^h, \boldsymbol{w}^h) \mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\nu}\mathrm{d}\boldsymbol{w} < \infty \tag{14}$$

TABLE 7
IRI-28 *questionnaire. Subjects answer with their level of agreement with numbers ranging from 0 ("Does not describe me") to 4 ("Describes me very well").*

ID		SUB
1.	I daydream and fantasize, with some regularity, about things that might happen to me.	(FS)
2.	I often have tender, concerned feelings for people less fortunate than me.	(EC)
3.	I sometimes find it difficult to see things from the "other guy's" point of view.	(PT)
4.	Sometimes I don't feel very sorry for other people when they are having problems.	(EC)
5.	I really get involved with the feelings of the characters in a novel.	(FS)
7.	I am usually objective when I watch a movie or play, and I don't often get completely caught up in it.	(FS)
8.	I try to look at everybody's side of a disagreement before I make a decision.	(PT)
9.	When I see someone being taken advantage of, I feel kind of protective towards them.	(EC)
10.	I sometimes feel helpless when I am in the middle of a very emotional situation.	(PD)
11.	I sometimes try to understand my friends better by imagining how things look from their perspective.	(PT)
12.	Becoming extremely involved in a good book or movie is somewhat rare for me.	(FS)
13.	When I see someone get hurt, I tend to remain calm.	(PD)
14.	Other people's misfortunes do not usually disturb me a great deal.	(EC)
15.	If I'm sure I'm right about something, I don't waste much time listening to other people's arguments.	(PT)
16.	After seeing a play or movie, I have felt as though I were one of the characters.	(FS)
17.	Being in a tense emotional situation scares me.	(PD)
18.	When I see someone being treated unfairly, I sometimes don't feel very much pity for them.	(EC)
19.	I am usually pretty effective in dealing with emergencies.	(PD)
21.	I believe that there are two sides to every question and try to look at them both.	(PT)
23.	When I watch a good movie, I can very easily put myself in the place of a leading character.	(FS)
25.	When I'm upset at someone, I usually try to "put myself in his shoes" for a while.	(PT)
26.	When I am reading an interesting story or novel, I imagine how I would feel if the events in the story were happening to me.	(FS)
28.	Before criticizing somebody, I try to imagine how I would feel if I were in their place.	(PT)

Since the priors specified in (8) are proper, it is sufficient to show that

$$\sup_{\boldsymbol{\vartheta},\boldsymbol{\nu}} \prod_{i=1}^{n} \sum_{h=1}^{H} \nu_h \, \tilde{\mathbf{p}}(y_i \mid \boldsymbol{\vartheta}^h, \boldsymbol{w}^h) < \infty \tag{15}$$

which is always bounded being a product of probabilities. □

APPENDIX C: ALGORITHMS FOR POSTERIOR INFERENCE

Acknowledgements. The case study illustrated in this work has been motivated by a collaboration with doctor Paolo Scocco from Padova Hospital, which is kindly acknowledged for providing the data and the stimulating discussions. Emanuele Aliverti would also like to acknowledge Prof. Giovanna Capizzi, Massimiliano Russo and Daniele Durante for the stimulating discussions on the first draft of this work. A preliminary version of this work was included in the Ph.D. dissertation of Emanuele Aliverti. This work was partially funded by MIUR–PRIN 2017 project 20177BR-JXS, as well as grant R01ES027498 of the National Institute of Environmental Health Sciences of the United States Institutes of Health, and grant N00014-16-1-2147 from the United States Office of Naval Research.

REFERENCES

AGRESTI, A. (2003). *Categorical data analysis* **482**. John Wiley & Sons.

AIROLDI, E. M., BLEI, D., EROSHEVA, E. A. and FIENBERG, S. E. (2014). *Handbook of mixed membership models and their applications*. CRC press.

ANDERSEN, E. B. (1982). Latent structure analysis: A survey. *Scandinavian Journal of Statistics* 1–12.

BERGSMA, W. P., RUDAS, T. et al. (2002). Marginal models for categorical data. *The Annals of Statistics* **30** 140–159.

Algorithm 1: One cycle of Gibbs sampler for MILLS.

for $h = 1, \ldots, H$ **do**

  for $E_2 = 1, \ldots, |\mathcal{P}_2|$ **do**

    It is convenient to reparametrize the MILLS likelihood as $\tilde{\boldsymbol{\vartheta}}^h_{E_2} = \mathbf{X}_2 \boldsymbol{\vartheta}^h_{E_2}$, corresponding to the cell-specific multinomial log-odds. The Gaussian prior on $\boldsymbol{\vartheta}^h_{E_2}$ induces a Gaussian prior on $\tilde{\boldsymbol{\vartheta}}^h_{E_2}$ with covariance matrix $\mathbf{X}_2^\intercal \mathbf{X}_2$. Therefore, the prior precision of each element of $\tilde{\boldsymbol{\vartheta}}^h_{E_2}$ given the others is given by the diagonal elements of $(\mathbf{X}_2^\intercal \mathbf{X}_2)^{-1}$.

    Sample each $\tilde{\boldsymbol{\vartheta}}^h_{E_2}$ from a conditionally-conjugate Gaussian distribution, adapting the Pòlya-Gamma strategy to the multinomial likelihood (Polson, Scott and Windle, 2013).

  end

end

for $h = 1, \ldots, H$ **do**

  for $E_2 = 1, \ldots, |\mathcal{P}_2|$ **do**

    Sample each $\delta^h_{E_2}$ from a Bernoulli distribution with probability of success equal to

$$\frac{\gamma^h_0 \text{GAMMA}(w^h_{E_2}; 1 + a^h_0, a^h_1 - \ell^h_{E_2})}{\gamma^h_0 \text{GAMMA}(w^h_{E_2}; 1 + a^h_0, a^h_1 - \ell^h_{E_2}) + (1 - \gamma^h_0)\text{GAMMA}(w^h_{E_2}; 1, a^h_1 - \ell^h_{E_2})},$$

    with $\ell^h_{E_2} = \log\big[\tilde{\boldsymbol{y}}^{h\intercal}_{E_2} \boldsymbol{\vartheta}^h_{E_2} - n_h \kappa_2(\boldsymbol{\vartheta}^h_{E_2})\big]$ and with $\text{GAMMA}(x; a, b)$ denoting the density of a Gamma distribution with shape a, rate b evaluated in x. Note that $\ell^h_{E_2}$ is always negative, and therefore there is no ambiguity in the evaluation of the Gamma density.

  end

  for $E_2 = 1, \ldots, |\mathcal{P}_2|$ **do**

    Sample the composite weight $w^h_{E_2}$ from

$$\text{GAMMA}\left(1 + a^h_0 \delta^h_{E_2}, a^h_1 - \ell^h_{E_2}\right)$$

  end

  Sample the slab probability γ^h_0 from

$$\text{BETA}\left(\frac{1}{2} + \sum_{E_2 \in \mathcal{P}_2} \delta^h_{E_2}, \frac{1}{2} + |\mathcal{P}_2| - \sum_{E_2 \in \mathcal{P}_2} \delta^h_{E_2}\right)$$

end

for $i = 1, \ldots, n$ **do**

  Sample z_i from

$$\text{CATEGORICAL}\left(\frac{\nu_1 \tilde{\mathbf{p}}(y_i; \boldsymbol{\vartheta}^1, \boldsymbol{w}^1)}{\sum_{h=1}^H \nu_h \tilde{\mathbf{p}}(y_i; \boldsymbol{\vartheta}^h, \boldsymbol{w}^h)}, \cdots, \frac{\nu_H \tilde{\mathbf{p}}(y_i; \boldsymbol{\vartheta}^H, \boldsymbol{w}^H)}{\sum_{h=1}^H \nu_h \tilde{\mathbf{p}}(y_i; \boldsymbol{\vartheta}^h, \boldsymbol{w}^h)}\right)$$

  with $\mathbf{p}(y_i; \boldsymbol{\vartheta}^h, \boldsymbol{w}^h)$ defined in (5).

end

Sample $\boldsymbol{\nu}$ from

$$\text{DIRICHLET}\left(n_1 + \frac{1}{H}, \cdots, n_H + \frac{1}{H}\right),$$

with $n_h = \sum_{i=1}^n \mathbb{1}[z_i = h]$.

BHATTACHARYA, A. and DUNSON, D. B. (2012). Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association* **107** 362–377.

BRADLEY, J. R., HOLAN, S. H. and WIKLE, C. K. (2019). Bayesian hierarchical models with conjugate full-conditional distributions for dependent data from the natural exponential family. *Journal of the American Statistical Association* 1–16.

BRITTON, P. C., BOHNERT, A. S., WINES JR, J. D. and CONNER, K. R. (2012). A procedure that differentiates unintentional from intentional overdose in opioid abusers. *Addictive Behaviors* **37** 127–130.

CHANG, I. H. and MUKERJEE, R. (2006). Probability matching property of adjusted likelihoods. *Statistics & Probability Letters* **76** 838 - 842.

COX, D. R. and REID, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91** 729–737.

CUSI, A. M., MACQUEEN, G. M., SPRENG, R. N. and MCKINNON, M. C. (2011). Altered empathic responding in major depressive disorder: relation to symptom severity, illness burden, and psychosocial outcome. *Psychiatry Research* **188** 231-236.

DAVIS, M. H. (1980). A muntidimensional approach to individual differences in empathy. *JSAS Catalogue of Selected Documents in Psychology* **10**.

DE BEURS, D., FRIED, E. I., WETHERALL, K., CLEARE, S., O'CONNOR, D. B., FERGUSON, E., O'CARROLL, R. E. and O'CONNOR, R. C. (2019). Exploring the psychology of suicidal ideation: A theory driven network analysis. *Behaviour Research and Therapy* **120** 103419.

DE LEO, D., BURGIS, S., BERTOLOTE, J. M., KERKHOF, A. and BILLE-BRAHE, U. (2004). Definitions of suicidal behaviour. *Suicidal Behaviour: Theories and Research Findings* 17-39.

DEROGATIS, L., LIPMAN, R. and COVI, L. (1973). Scl-90: an outpatient psychiatric rating scale–preliminary report. *Psychopharmacology Bulletin* **9** 13.

DOBRA, A. and MASSAM, H. (2010). The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. *Statistical Methodology* **7** 240–253.

DOBRA, A. and MOHAMMADI, R. (2018). Loglinear model selection and human mobility. *The Annals of Applied Statistics* **12** 815–845.

DUNSON, D. B. and XING, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104** 1042–1051.

EFRON, B. (1993). Bayes and Likelihood Calculations from Confidence Intervals. *Biometrika* **80** 3–26.

EROSHEVA, E. A. (2005). Comparing latent structures of the grade of membership, Rasch, and latent class models. *Psychometrika* **70** 619–628.

FERRARI, D., QIAN, G. and HUNTER, T. (2016). Parsimonious and Efficient Likelihood Composition by Gibbs Sampling. *Journal of Computational and Graphical Statistics* **25** 935–953.

FIENBERG, S. E. and RINALDO, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *Journal of Statistical Planning and Inference* **137** 3430–3445.

FONTENELLE, L. F., SOARES, I. D., MIELE, F., BORGES, M. C., PRAZERES, A. M., RANGÉ, B. P. and MOLL, J. (2009). Empathy and symptoms dimensions of patients with obsessive– compulsive disorder. *Journal of Psychiatric Research* **43** 455-463.

FRASER, D. and REID, N. (2019). Combining likelihood and significance functions. *Statistica Sinica* **29** 1–15.

FRUHWIRTH-SCHNATTER, S., CELEUX, G. and ROBERT, C. P. (2019). *Handbook of mixture analysis*. Chapman and Hall/CRC.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian data analysis*. CRC press.

GEYER, C. J. and THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society: Series B (Methodological)* **54** 657–683.

GILET, A.-L., MELLA, N., STUDER, J., GRÜHN, D. and LABOUVIE-VIEF, G. (2013). Assessing dispositional empathy in adults: A French validation of the Interpersonal Reactivity Index (IRI). *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* **45** 42.

GOODFELLOW, B., KÕLVES, K. and DE LEO, D. (2019). Contemporary definitions of suicidal behavior: a systematic literature review. *Suicide and Life-Threatening Behavior* **49** 488-504.

GRECO, L., RACUGNO, W. and VENTURA, L. (2008). Robust likelihood functions in Bayesian inference. *Journal of Statistical Planning and Inference* **138** 1258 - 1270.

GUTTMAN, H. and LAPORTE, L. (2002). Alexithymia, empathy, and psychological symptoms in a family context. *Comprehensive Psychiatry* **43** 448-455.

HAWTON, K. and FAGG, J. (1988). Suicide, and other causes of death, following attempted suicide. *The British Journal of Psychiatry* **152** 359-366.

HUANG, Z. and FERRARI, D. (2017). Fast construction of efficient composite likelihood equations. *arXiv preprint arXiv:1709.03234*.

ISHWARAN, H., RAO, J. S. et al. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics* **33** 730–773.

JOHNDROW, J. E., BHATTACHARYA, A. and DUNSON, D. B. (2017). Tensor decompositions and sparse log-linear models. *Annals of Statistics* **45** 1–38.

JOHNDROW, J., BHATTACHARYA, A. et al. (2018). Optimal Gaussian approximations to the posterior for log-linear models with Diaconis–Ylvisaker priors. *Bayesian Analysis* **13** 201–223.

KELLEHER, I., HARLEY, M., MURTAGH, A. and CANNON, M. (2011). Are screening instruments valid for psychotic-like experiences? A validation study of screening questions for psychotic-like experiences using in-depth clinical interview. *Schizophrenia Bulletin* **37** 362–369.
KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM review* **51** 455–500.
LADISICH, W. and FEIL, W. (1988). Empathy in psychiatric patients. *British Journal of Medical Psychology* **61** 155–162.
LAURITZEN, S. L. (1996). *Graphical models* **17**. Clarendon Press.
LAWLEY, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics* **61** 273–287.
LAZAR, N. A. (2003). Bayesian empirical likelihood. *Biometrika* **90** 319-326.
LAZARSFELD, P. F. (1950). The logical and mathematical foundation of latent structure analysis. *Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction* 362–412.
LEGRAMANTI, S., DURANTE, D. and DUNSON, D. B. (2020). Bayesian cumulative shrinkage for infinite factorizations. *Biometrika (in press)*.
LETAC, G., MASSAM, H. et al. (2012). Bayes factors and the geometry of discrete hierarchical loglinear models. *The Annals of Statistics* **40** 861–890.
LINDSAY, B. G., YI, G. Y. and SUN, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica* 71–105.
LUPPARELLI, M., MARCHETTI, G. M. and BERGSMA, W. (2009). Parameterization and fitting of discrete bi-directed graph models. *Scandinavian Journal of Statistics* **36** 559–576.
MANRIQUE-VALLIER, D. (2014). Longitudinal mixed membership trajectory models for disability survey data. *The Annals of Applied Statistics* **8** 2268.
MARDIA, K. V., KENT, J. T., HUGHES, G. and TAYLOR, C. C. (2009). Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika* **96** 975–982.
MASSAM, H. and WANG, N. (2018). Local conditional and marginal approach to parameter estimation in discrete graphical models. *Journal of Multivariate Analysis* **164** 1–21.
MASSAM, H., LIU, J., DOBRA, A. et al. (2009). A conjugate prior for discrete hierarchical log-linear models. *The Annals of Statistics* **37** 3431–3467.
MCCORMICK, L. M., BRUMM, M. C., BEADLE, J. N., PARADISO, S., YAMADA, T. and ANDREASEN, N. (2012). Mirror neuron function, psychosis, and empathy in schizophrenia. *Psychiatry Research: Neuroimaging* **201** 233–239.
MCHUGH, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika* **21** 331–347.
MENG, Z., WEI, D., WIESEL, A. and HERO III, A. (2013). Distributed learning of Gaussian graphical models via marginal likelihoods. In *Artificial Intelligence and Statistics* 39–47.
MILLER, J. W. (2019). Asymptotic normality, concentration, and coverage of generalized posteriors. *arXiv preprint arXiv:1907.09611*.
NARDI, Y., RINALDO, A. et al. (2012). The log-linear group-lasso estimator and its asymptotic properties. *Bernoulli* **18** 945–974.
NOCK, M. K., BORGES, G., BROMET, E. J., ALONSO, J., ANGERMEYER, M., BEAUTRAIS, A., BRUFFAERTS, R., CHIU, W. T., DE GIROLAMO, G., GLUZMAN, S. et al. (2008). Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *The British Journal of Psychiatry* **192** 98-105.
PACE, L., SALVAN, A. and SARTORI, N. (2019). Efficient composite likelihood for a scalar parameter of interest. *Stat* **8** e222.
PAULI, F., RACUGNO, W. and VENTURA, L. (2011). Bayesian composite marginal likelihoods. *Statistica Sinica* **21** 149–164.
PERRONE-MCGOVERN, K. M., OLIVEIRA-SILVA, P., SIMON-DACK, S., LEFDAHL-DAVIS, E., ADAMS, D., MCCONNELL, J., HOWELL, D., HESS, R., DAVIS, A. and GONÇALVES, Ó. F. (2014). Effects of empathy and conflict resolution strategies on psychophysiological arousal and satisfaction in romantic relationships. *Applied Psychophysiology and Biofeedback* **39** 19-25.
POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association* **108** 1339–1349.
PRINZ, U., NUTZINGER, D. O., SCHULZ, H., PETERMANN, F., BRAUKHAUS, C. and ANDREAS, S. (2013). Comparative psychometric analyses of the SCL-90-R and its short versions in patients with affective disorders. *BMC psychiatry* **13** 104.
PRUNAS, A., SARNO, I., PRETI, E., MADEDDU, F. and PERUGINI, M. (2012). Psychometric properties of the italian version of the scl-90-r: a study on a large community sample. *European Psychiatry* **27** 591-597.
RAFTERY, A. E. (1985). A model for high-order Markov chains. *Journal of the Royal Statistical Society: Series B (Methodological)* **47** 528–539.

RAFTERY, A., MADIGAN, D. and VOLINSKY, C. T. (1995). Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance. *Bayesian Statistics* **5** 323–349.

RAVIKUMAR, P., WAINWRIGHT, M. J., LAFFERTY, J. D. et al. (2010). High-dimensional Ising model selection using $\ell 1$-regularized logistic regression. *The Annals of Statistics* **38** 1287–1319.

RIBATET, M., COOLEY, D. and DAVISON, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica* 813–845.

ROUSSEAU, J. and MENGERSEN, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 689–710.

ROVERATO, A., LUPPARELLI, M. and LA ROCCA, L. (2013). Log-mean linear models for binary data. *Biometrika* **100** 485–494.

RUSSO, M., DURANTE, D. and SCARPA, B. (2018). Bayesian inference on group differences in multivariate categorical data. *Computational Statistics & Data Analysis* **126** 136–149.

SCHREITER, S., PIJNENBORG, G. and AAN HET ROT, M. (2013). Empathy in adults with clinical or subclinical depressive symptoms. *Journal of Affective Disorders* **150** 1-16.

SCOCCO, P. and DE LEO, D. (2002). One-year prevalence of death thoughts, suicide ideation and behaviours in an elderly population. *International Journal of Geriatric Psychiatry* **17** 842–846.

SCOCCO, P., ALIVERTI, E., TOFFOL, E., ANDRETTA, G. and CAPIZZI, G. (2020). Empathy profiles differ by gender in people who have and have not attempted suicide. *Journal of Affective Disorders Reports* **2** 100024.

SNIJDERS, T. A. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure* **3** 1–40.

TIBSHIRANI, R., WAINWRIGHT, M. and HASTIE, T. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

VARIN, C., REID, N. and FIRTH, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* 5–42.

WANG, W., ZHOU, Y., WANG, J., XU, H., WEI, S., WANG, D., WANG, L. and ZHANG, X. (2020). Prevalence, clinical correlates of suicide attempt and its relationship with empathy in patients with schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 109863.

ZHANG, K., SZANTO, K., CLARK, L. and DOMBROVSKI, A. Y. (2019). Behavioral empathy failures and suicidal behavior. *Behaviour Research and Therapy* **120** 103329.