VARIABLE SELECTION FOR LATENT CLASS ANALYSIS WITH APPLICATION TO LOW BACK PAIN DIAGNOSIS

BY MICHAEL FOP[*,1], KEITH M. SMART[†] AND
THOMAS BRENDAN MURPHY[*,1]

University College Dublin[*] and St. Vincent's University Hospital[†]

The identification of most relevant clinical criteria related to low back pain disorders may aid the evaluation of the nature of pain suffered in a way that usefully informs patient assessment and treatment. Data concerning low back pain can be of categorical nature, in the form of a check-list in which each item denotes presence or absence of a clinical condition. Latent class analysis is a model-based clustering method for multivariate categorical responses, which can be applied to such data for a preliminary diagnosis of the type of pain. In this work, we propose a variable selection method for latent class analysis applied to the selection of the most useful variables in detecting the group structure in the data. The method is based on the comparison of two different models and allows the discarding of those variables with no group information and those variables carrying the same information as the already selected ones. We consider a swap-stepwise algorithm where at each step the models are compared through an approximation to their Bayes factor. The method is applied to the selection of the clinical criteria most useful for the clustering of patients in different classes. It is shown to perform a parsimonious variable selection and to give a clustering performance comparable to the expert-based classification of patients into three classes of pain.

1. Introduction. Musculoskeletal pain is the pain concerning muscles, bones and joints that arises in different conditions. Low back pain (LBP) is the muscoloskeletal pain related to disorders in the lumbar spine, low back muscles and nerves and it may radiate to the legs. Although there is a lack of homogeneity in the studies, a considerable proportion of the population experiences LBP during their lifetime [Hoy et al. (2012), Walker (2000)], with effects on social and psychic traits and working behaviour [Froud et al. (2014)]. Several LBP classification systems have been developed in order to group patients into classes with similar characteristics, with the purpose of effective pain management; see Stynes, Konstantinou and Dunn (2016) and references therein. Among the different systems, mechanism-based classification of pain is based on the underlying neurophysiological mechanisms responsible for its generation and maintenance. The system has been advocated in clinical practice on the ground of better pain treatment and

improved patient outcomes [Smart, O'Connell and Doody (2008), Woolf et al. (1998)]. In the absence of any diagnostic gold standards for mechanisms-based pain diagnoses, such a categorization may be identifiable on the basis of sets of symptoms and signs characteristic to each category by means of a standard clinical examination process and experienced clinical judgement [Graven-Nielsen and Arendt-Nielsen (2010), Katz et al. (2000), Nijs et al. (2015), Smart et al. (2010)]. Furthermore, with the aim of a diagnosis of the nature of the LBP suffered by a subject, identifying a smaller collection of signs or symptoms which best relates the manifestation of pain to its neurophysiological mechanism is a critical task. Focusing the attention only on few manifest pain characteristics can guide a preliminary patient evaluation and can constitute a valid basis for additional investigations and immediate pain treatment.

Model-based clustering [Fraley and Raftery (2002), McNicholas (2016)] is a well-established framework for clustering multivariate data. In this approach, the data generating process is modelled through a finite mixture of probability distributions, where each component distribution corresponds to a group. When the observations are measured on categorical variables (such as data arising from questionnaires), the most common model-based clustering method is the latent class analysis model (LCA) [Lazarsfeld and Henry (1968)]. Typically, all the variables are considered in fitting the model, but often only a subset of the variables at hand contains the useful information about the group structure of the data. When performing variable selection for clustering, the goal is to remove *irrelevant* variables, which do not carry group information, and *redundant* variables, which convey similar group information, retaining only with the set of *relevant* variables, which contains the useful information [Dy and Brodley (2003/04)]. Therefore, considering all the variables unnecessarily increases the model complexity and can produce model identifiability problems. Moreover, using variables that do not contain group information or that contain unneeded information frequently leads to a poor classification performance.

In recent years, wide attention has been given to the problem of variable selection in clustering multivariate data. The problem has been generally tackled through two approaches: the *wrapper* approach, which combines clustering and variable selection at the same time, and the *filter* approach, where the variables are selected after or before the clustering is performed [Dy and Brodley (2003/04)]. Model-based clustering for continuous data has seen the prevalence of the wrapper approach; we cite the works from Law, Figueiredo and Jain (2004), Tadesse, Sha and Vannucci (2005), Kim, Tadesse and Vannucci (2006), Raftery and Dean (2006), Maugis, Celeux and Martin-Magniette (2009a, 2009b), Murphy, Dean and Raftery (2010), Scrucca and Raftery (2017), Malsiner-Walli, Frühwirth-Schnatter and Grün (2016), Marbac and Sedki (2017). Moreover, further works in the wrapper approach considered the introduction of a penalty term in the log-likelihood in order to induce sparsity in the features; for example, Pan and Shen (2007), Wang and Zhu (2008), Xie, Pan and Shen (2008), Meynet and Maugis-Rabusseau (2012)

In LCA, the variable selection problem has been assessed only recently. Under a filter approach, Zhang and Ip (2014) propose two measures for quantifying the discriminative power of a variable for mixed mode data, but the method is limited only to binary variables. Under the wrapper approach, Dean and Raftery (2010) recast the variable selection problem as a model selection problem, Bontemps and Toussile (2013) suggest an approach designed on the minimization of a risk function, Silvestre, Cardoso and Figueiredo (2015) propose a method adapted from Law, Figueiredo and Jain (2004) and based on feature saliency, White, Wyse and Murphy (2016) present a full Bayesian framework with a collapsed Gibbs sampler and Bartolucci, Montanari and Pandolfi (2016) present a method based on the work of Dean and Raftery (2010) for item selection in questionnaires.

All of the above mentioned wrapper methods for LCA have a drawback: they consider a variable to be added or removed to the already selected set of clustering ones assuming that the former is independent of the latter. By this assumption, two (or more) informative correlated variables are selected, even if they contain similar group information. However, retaining only one (or a subset) of them can lead to a clustering of comparable quality with a more parsimonious variable selection. Thus the result is the methods are capable of discarding non-informative variables, but not the redundant variables.

In this work, we develop a variable selection method for LCA based on the model selection framework of Dean and Raftery (2010), which overcomes the limitation of the above independence assumption. By adapting the variable role modelling of Maugis, Celeux and Martin-Magniette (2009b) in the variable selection procedure, we propose a method capable of discarding variables that do not contain group information and variables that are redundant. This variable selection method assesses a variable usefulness for clustering by comparing models via an approximation to their Bayes factor.

We apply the proposed method to cluster a set of patients suffering of low back pain. Each patient was diagnosed as having a different type of pain by a group of experienced physiotherapists using a list of several clinical indicators. The aim is to recover in an unsupervised setting a classification of the patients comparable to the expert-based one and at the same time selecting a reduced collection of clinical indicators that can be used for a preliminary assessment of the characteristics of pain.

Section 2 presents the low back pain data, which gave the motivation for the improvement in the variable selection approach for LCA. In Section 3, we give a brief description of model-based clustering and latent class analysis. The general variable selection methodology for LCA is presented in Section 4. First, we review the Dean and Raftery (2010) procedure and subsequently we present our proposed variable selection method characterized by the relaxation of the independence assumption between the clustering variables and the proposed one. Section 5 is dedicated to the results of the variable selection method applied to the LBP data. Section 6 presents a simulation study on two different scenarios. The paper ends with a brief discussion in Section 7.

2. Low back pain data. A mechanisms-based classification of pain relates the generation and maintenance of pain to its underlying neurophysiological mechanisms. To this purpose, the following categories have been suggested for a clinically meaningful classification of pain [Merskey and Bogduk (2002), Woolf (2004)]:

- *Nociceptive*: Pain that arises from actual or threatened damage to non-neural tissue, occurring with a normally functioning somatosensory nervous system.
- *Peripheral neuropathic*: Pain initiated or caused by a primary lesion or dysfunction in the peripheral nervous system.
- *Central sensitization*: Pain initiated or caused by a primary lesion or dysfunction in the central nervous system.

It is thought that classifying patients low back pain based on a clinical judgement regarding the likely dominant category of neurophysiological mechanisms responsible for its generation and/or persistence may usefully inform treatment by inviting clinicians to select treatments either known or hypothesized to target those mechanisms in an attempt to optimize clinical outcomes [Smart, O'Connell and Doody (2008)]. In this regard, a list of 38 clinical criteria (signs and symptoms) whose presence or absence can best discriminate the three types of pain has been generated on an expert-consensus basis. See Smart et al. (2010) and the Supplementary Material, Section 4 [Fop, Smart and Murphy (2017)] for the complete clinical criteria checklist.

Smart et al. (2011) conducted a preliminary discriminative validity study of such mechanisms-based classification of musculoskeletal pain in clinical practice. The aim of the study was to assess the discriminative validity of the above classification system for low back disorders. The data are a sample of 464 patients, each one assigned to one of the three categories of pain by a group of experienced physiotherapists. For each patient, information regarding the presence/absence of the 38 binary clinical indicators is recorded.

In the present work, in analysing these data the aim is twofold:

1. Implement an unsupervised partition of the patients to form groups of patients with similar characteristics. Thus we can establish if the clusters found using the unsupervised method agree with the expert-based classification or not. This allows for the discovery of a potentially novel partition of the patients into homogeneous groups or a further validation of the expert-based classification;

2. Select a subset of most relevant clinical criteria for partitioning the patients. Most of the indicators (if not all) have good discriminative power and a large part of them carry the same information about the pain categories. The interest here is to discard redundant and non-informative indicators in order to reduce the list of signs and symptoms to check for a preliminary assessment of a patient condition.

In collecting the data, the presence/absence of some criteria was indicated as "don't know" for some patients as the corresponding information was unavailable.

In particular, Criteria 20 records if a subject condition was responsive or not to non-steroidal anti-inflammatory drugs (NSAIDs) and for a set of patients it was not known if they actually took or not any NSAIDs. In Smart et al. (2011), these entries were discarded. To be consistent with the authors approach and consider the same set of data, we discard them as well in the following analysis: Section 1 of the Supplementary Material [Fop, Smart and Murphy (2017)] contains a discussion and a brief analysis with these entries included as extra category. Furthermore, Criteria 17 and 21 are not available in the data and are not considered. The final data set is then composed of 425 patients examined on 36 binary variables.

3. Latent class analysis. Let $\mathbf{X}$ the $N \times M$ data matrix, where each row $\mathbf{X}_n$ is the realization of a M-dimensional vector of random variables $\mathbf{X}_n = (X_{n1}, \ldots, X_{nm}, \ldots, X_{nM})$. Model-based clustering assumes that each $\mathbf{X}_n$ arises from a finite mixture of G probability distributions, each representing a different cluster or group. The general form of a finite mixture distribution is specified as follows:

$$(1) \qquad p(\mathbf{X}_n) = \sum_{g=1}^{G} \tau_g \, p(\mathbf{X}_n | \boldsymbol{\theta}_g),$$

where the τ_g are the mixing probabilities and $\boldsymbol{\theta}_g$ is the parameter set corresponding to component g. The component densities fully characterize the group structure of the data and each observation belongs to the corresponding cluster according to a set of unobserved cluster membership indicators $\mathbf{z}_n = (z_{n1}, z_{n2}, \ldots, z_{nG})$, such that $z_{ng} = 1$ if $\mathbf{X}_n$ arises from the gth subpopulation [Fraley and Raftery (2002), McLachlan and Peel (2000)].

When clustering multivariate categorical data a common model-based approach is the latent class analysis model (LCA). In this framework, within each class each variable X_m is modelled using a multinomial distribution, therefore,

$$p(X_m | \theta_g) = \prod_{c=1}^{C_m} \theta_{\text{gmc}}^{\mathbb{1}\{X_m = c\}},$$

where $c = 1, \ldots, C_m$ are the possible categories values for variable m, θ_{gmc} is the probability of the variable taking value c given class g, and $\mathbb{1}\{x_m = c\}$ is the indicator function equal to 1 if the variable takes value c, 0 otherwise. In LCA it is assumed that the variables are statistically independent given the class value of an observation. This is a basic assumption known as the *local independence assumption* [Clogg (1988)] and it allows the following factorization of the joint component density:

$$p(\mathbf{X}_n | \boldsymbol{\theta}_g) = \prod_{m=1}^{M} \prod_{c=1}^{C_m} \theta_{\text{gmc}}^{\mathbb{1}\{X_{nm} = c\}};$$

consequently the overall density in (1) becomes

$$p(\mathbf{X}_n) = \sum_{g=1}^{G} \tau_g \prod_{m=1}^{M} \prod_{c=1}^{C_m} \theta_{\mathrm{gmc}}^{\mathbb{1}\{X_{nm}=c\}}.$$

For a fixed value G, the set of parameters $\{\tau_g, \theta_{\mathrm{gmc}}\colon m = 1, \ldots, M; c = 1, \ldots, C_m; g = 1, \ldots, G\}$ is usually estimated by the EM algorithm, but also a Newton–Raphson algorithm or a hybrid form of the two can be considered [McLachlan and Krishnan (2008)]. In any case, the algorithm is initialized through a set of randomly generated starting values and there is no guarantee of reaching the global maximum. For this reason, it is usually a good practice to run the procedure a number of times and select the best solution [Bartholomew, Knott and Moustaki (2011)].

More details about the model and the parameter estimation are provided in Lazarsfeld and Henry (1968), Goodman (1974), Haberman (1979), Clogg (1995), Agresti (2002) and Bartholomew, Knott and Moustaki (2011).

Regarding parameters interpretation, in the LCA model the parameter θ_{gmc} represents the probability of occurrence of attribute c for variable X_m in class g. Thus for the binary variables of the LBP data, θ_{gmc} will represent the probability of having a certain symptom or clinical criteria for each patient belonging to class g.

Model selection. Different LCA models are specified by assigning different values to G. Here, the selection of the best model and of the related number of latent classes is carried out using an approximation to their Bayes factor. When comparing two competing models specified to describe the data $\mathbf{X}$, say $\mathcal{M}_A$ against $\mathcal{M}_B$, the extent to which the data support model $\mathcal{M}_A$ over $\mathcal{M}_B$ is measured by their posterior odds. In absence of prior preference for one of the two models, this quantity is given by

$$\frac{p(\mathcal{M}_A|\mathbf{X})}{p(\mathcal{M}_B|\mathbf{X})} = \frac{p(\mathbf{X}|\mathcal{M}_A)}{p(\mathbf{X}|\mathcal{M}_B)},$$

where $p(\mathbf{X}|\mathcal{M}_A) = \int p(\mathbf{X}|\boldsymbol{\theta}, \mathcal{M}_A) p(\boldsymbol{\theta}|\mathcal{M}_A)\, d\boldsymbol{\theta}$ is the integrated likelihood. The ratio of the integrated likelihoods of the two models is the Bayes factor, $\mathcal{B}_{A,B}$. The quantity $p(\mathbf{X}|\mathcal{M}_A)$ is conveniently approximated using the Bayesian Information Criterion (BIC), defined by

$$\mathrm{BIC}(\mathbf{X}|\mathcal{M}_A) = 2\log(L_A^*) - \nu_A \log(N),$$

where L_A^* is the maximized likelihood and ν_A is the number of model parameters [Schwarz (1978)]. Then the following approximation to twice the logarithm of the Bayes factor holds [Kass and Raftery (1995)]:

$$(2) \qquad 2\log(\mathcal{B}_{A,B}) \approx \mathrm{BIC}(\mathbf{X}|\mathcal{M}_A) - \mathrm{BIC}(\mathbf{X}|\mathcal{M}_B),$$
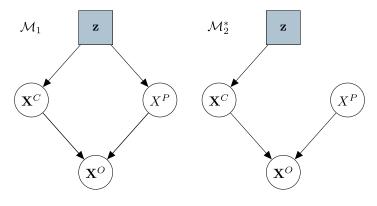
FIG. 1. *The two competing models in Dean and Raftery (2010).*

and if this difference is greater than zero the evidence is in favour of model $\mathcal{M}_A$, otherwise in favour of $\mathcal{M}_B$. Several arguments in favour of BIC for model selection in mixture models have been given in the literature; see McLachlan and Rathnayake (2014) for a recent review.

For a given number of variables, not all the models specified by assigning different values to G are identifiable. In fact, a necessary (though not sufficient) condition to the identifiability of a model with G latent classes is

$$(3) \qquad \prod_{m=1}^{M} C_m > \left(\sum_{m=1}^{M} C_m - M + 1 \right) G,$$

with C_m the number of categories taken by variable X_m [Goodman (1974)]. Thus when selecting the number of classes, hereafter we will consider values of G for which this identifiability condition holds.

4. Variable selection for latent class analysis. To select the variables relevant for clustering in LCA, Dean and Raftery (2010) suggested a stepwise model comparison approach. At each step of their method, the authors specify a partition of the variables into:

- $\mathbf{X}^C$, the current set of relevant clustering variables, dependent on the cluster membership variable $\mathbf{z}$,
- X^P, the variable proposed to be added or removed from the clustering variables,
- $\mathbf{X}^O$, the set of the other variables which are not relevant for clustering.

Then the decision of adding or removing the considered variable is made by comparing two models: model $\mathcal{M}_1$, in which the variable is useful for clustering, and model $\mathcal{M}_2^*$ in which it does not. Figure 1 gives a graphical sketch of the two competing models.

Both models make the realistic assumption that the relevant variables are not independent from the irrelevant ones (the edge between $\mathbf{X}^O$ and $\mathbf{X}^C$), but they

differ in the specification of the relationship with X^P. In $\mathcal{M}_1$, there is no edge between $\mathbf{X}^C$ and X^P because the model states that the proposed variable is useful for clustering and we have that the joint distribution $p(\mathbf{X}^C, X^P|\mathbf{z})$ factorizes into $p(\mathbf{X}^C|\mathbf{z})p(X^P|\mathbf{z})$ by the local independence assumption of LCA. In $\mathcal{M}_2^*$, there is no edge between $\mathbf{z}$ and X^P because under this model the proposed variable is not useful for clustering. Also, the edge between $\mathbf{X}^C$ and X^P is missing, since Dean and Raftery (2010) assume the *independence* of the proposed variable from $\mathbf{X}^C$, even when it is not relevant for clustering. However, this assumption seems to be misleading for two reasons. On one hand because if model $\mathcal{M}_2$ holds, actually X^P *belongs* to $\mathbf{X}^O$, contradicting the fact that the latter is not assumed independent from $\mathbf{X}^C$. On the other hand, because to this assumption the model does not take into account that the proposed variable could be redundant for clustering given the set of already selected relevant variables. In fact, as it has already pointed out in previous works [Law, Figueiredo and Jain (2004), Raftery and Dean (2006), White, Wyse and Murphy (2016)], assuming the independence between the proposed variable and the current set of clustering variables can wrongly lead to declare as relevant a variable that could be explained by (some or all) the variables in $\mathbf{X}^C$, even if actually it contains redundant group information that is no needed or it does not contain further information at all.

4.1. *Relaxing the independence assumption.* Now let us consider the models depicted in Figure 2. Model $\mathcal{M}_1$ is exactly the same model as before, where the proposed variable is useful for clustering. On the other hand, $\mathcal{M}_2$ is the model in which the proposed variable is not relevant for clustering, but there is an edge between X^P and $\mathbf{X}^C$, which defines the conditional distribution $p(X^P|\mathbf{X}^C)$. Therefore, $\mathcal{M}_2$ is specified by relaxing the independence assumption between the proposed variable and the set $\mathbf{X}^C$ and taking into account the potential redundancy of X^P. Hence, if the evidence is in favour of model $\mathcal{M}_2$, the proposed variable is discarded from $\mathbf{X}^C$ for two reasons: because it does not contain information about the
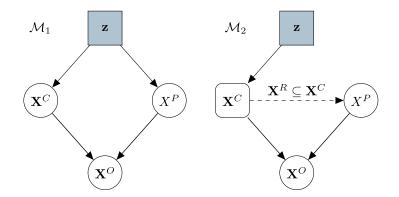

FIG. 2. *The two competing models specified relaxing the independence assumption between X^P and $\mathbf{X}^C$.*

latent classes at all, or because it does not add further useful information about the groups given the information already contained in the current clustering variables.

Moreover, another assumption is considered in model $\mathcal{M}_2$: we let the proposed variable to be related only to a subset $\mathbf{X}^R$ contained in the current set of clustering variables, since could be the case that not all the variables in $\mathbf{X}^C$ are associated to X^P [Maugis, Celeux and Martin-Magniette (2009a, 2009b)]. In this way, we do not induce spurious dependencies, avoiding the inclusion in the model of additional parameters without effectively increasing its likelihood. In addition, a more realistic modelling framework for the relationship between X^P and $\mathbf{X}^C$ is outlined, letting it to be as much flexible as possible. Clearly, it ranges between two extrema: if $\mathbf{X}^R = \mathbf{X}^C$, all the current clustering variables explain X^P, which could likely be redundant for clustering; if $\mathbf{X}^R = \varnothing$, the proposed variable is not related to the current clustering set, recalling the assumption of Dean and Raftery (2010).

Hence the two models are specified as follows:

$$\mathcal{M}_1\colon p(\mathbf{X}|\mathbf{z}) = p(\mathbf{X}^C, X^P, \mathbf{X}^O|\mathbf{z})$$
$$= p(\mathbf{X}^O|\mathbf{X}^C, X^P)p(\mathbf{X}^C, X^P|\mathbf{z});$$
$$\mathcal{M}_2\colon p(\mathbf{X}|\mathbf{z}) = p(\mathbf{X}^C, X^P, \mathbf{X}^O|\mathbf{z})$$
$$= p(\mathbf{X}^O|\mathbf{X}^C, X^P)p(\mathbf{X}^C|\mathbf{z})p(X^P|\mathbf{X}^R \subseteq \mathbf{X}^C).$$

Following Dean and Raftery (2010), models $\mathcal{M}_1$ and $\mathcal{M}_2$ are then compared via the Bayes factor:

$$\mathcal{B}_{1,2} = \frac{p(\mathbf{X}|\mathcal{M}_1)}{p(\mathbf{X}|\mathcal{M}_2)}.$$

Model $\mathcal{M}_1$ is specified by the probability distribution of the latent class model $p(\mathbf{X}^C, X^P|\boldsymbol{\theta}_1^C, \boldsymbol{\theta}_1^P, \mathcal{M}_1)$ and the distribution $p(\mathbf{X}^O|\mathbf{X}^C, X^P, \boldsymbol{\theta}_1^O, \mathcal{M}_1)$. We denoted $\boldsymbol{\theta}_1^C$, $\boldsymbol{\theta}_1^P$ and $\boldsymbol{\theta}_1^O$ the parameters vectors that identify these distributions and we assume that their prior probability distributions are independent. Hence the integrated likelihood factors as follows:

$$p(\mathbf{X}|\mathcal{M}_1) = p(\mathbf{X}^O|\mathbf{X}^C, X^P, \mathcal{M}_1)p(\mathbf{X}^C, X^P|\mathcal{M}_1),$$

with

$$p(\mathbf{X}^O|\mathbf{X}^C, X^P, \mathcal{M}_1) = \int p(\mathbf{X}^O|\mathbf{X}^C, X^P, \boldsymbol{\theta}_1^O, \mathcal{M}_1)p(\boldsymbol{\theta}_1^O|\mathcal{M}_1)\, d\boldsymbol{\theta}_1^O;$$

$$p(\mathbf{X}^C, X^P|\mathcal{M}_1) = \iint p(\mathbf{X}^C, X^P|\boldsymbol{\theta}_1^C, \boldsymbol{\theta}_1^P, \mathcal{M}_1)p(\boldsymbol{\theta}_1^C, \boldsymbol{\theta}_1^P|\mathcal{M}_1)\, d\boldsymbol{\theta}_1^C\, d\boldsymbol{\theta}_1^P.$$

Similarly the integrated likelihood of model $\mathcal{M}_2$ factors in

$$p(\mathbf{X}|\mathcal{M}_2) = p(\mathbf{X}^O|\mathbf{X}^C, X^P, \mathcal{M}_2)p(\mathbf{X}^C|\mathcal{M}_2)p(X^P|\mathbf{X}^R \subseteq \mathbf{X}^C, \mathcal{M}_2),$$

with

$$p(\mathbf{X}^O|\mathbf{X}^C, X^P, \mathcal{M}_2) = \int p(\mathbf{X}^O|\mathbf{X}^C, X^P, \boldsymbol{\theta}_2^O, \mathcal{M}_2) p(\boldsymbol{\theta}_2^O|\mathcal{M}_2) \, d\boldsymbol{\theta}_2^O;$$

$$p(\mathbf{X}^C|\mathcal{M}_2) = \int p(\mathbf{X}^C|\boldsymbol{\theta}_2^C, \mathcal{M}_2) p(\boldsymbol{\theta}_2^C|\mathcal{M}_2) \, d\boldsymbol{\theta}_2^C;$$

$$p(X^P|\mathbf{X}^R \subseteq \mathbf{X}^C, \mathcal{M}_2) = \int p(X^P|\mathbf{X}^R \subseteq \mathbf{X}^C \boldsymbol{\theta}_2^P, \mathcal{M}_2) p(\boldsymbol{\theta}_2^P|\mathcal{M}_2) \, d\boldsymbol{\theta}_2^P.$$

Assuming that the prior distributions for $\boldsymbol{\theta}_1^O$ and $\boldsymbol{\theta}_2^O$ are the same under both models, we obtain that $p(\mathbf{X}^O|\mathbf{X}^C, X^P, \mathcal{M}_1) = p(\mathbf{X}^O|\mathbf{X}^C, X^P, \mathcal{M}_2)$. Therefore,

$$\mathcal{B}_{1,2} = \frac{p(\mathbf{X}^C, X^P|\mathcal{M}_1)}{p(\mathbf{X}^C|\mathcal{M}_2) p(X^P|\mathbf{X}^R \subseteq \mathbf{X}^C, \mathcal{M}_2)}.$$

Note that in the Bayes factor the distribution of the non-clustering variables given the rest cancels out; this represents an advantage in terms of computations because there is no need to specify the joint distribution of all the non-clustering variables, unlike in White, Wyse and Murphy (2016), for example. Then this Bayes factor is estimated by the BIC approximation outlined in (2), leading to the following criterion:

$$\begin{aligned} \text{BIC}_{\text{diff}} &= \text{BIC}(\mathbf{X}^C, X^P|\mathcal{M}_1) - \text{BIC}(\mathbf{X}^C, X^P|\mathcal{M}_2) \\ &= \text{BIC}(\mathbf{X}^C, X^P|\mathbf{z}, \mathcal{M}_1) \\ &\quad - [\text{BIC}(\mathbf{X}^C|\mathbf{z}, \mathcal{M}_2) + \text{BIC}(X^P|\mathbf{X}^R \subseteq \mathbf{X}^C, \mathcal{M}_2)], \end{aligned}$$

where $\text{BIC}(\mathbf{X}^C, X^P|\mathbf{z}, \mathcal{M}_1)$ and $\text{BIC}(\mathbf{X}^C|\mathbf{z}, \mathcal{M}_2)$ are the BIC of the LCA model on the sets $\mathbf{X}^C \cup X^P$ and $\mathbf{X}^C$, respectively, while $\text{BIC}(X^P|\mathbf{X}^R \subseteq \mathbf{X}^C, \mathcal{M}_2)$ is the BIC of the model for the conditional distribution of the proposed variable (note that we made explicit the dependence on the latent variable $\mathbf{z}$). If this difference is greater than zero, there is evidence in favour of X^P adding further information about the clusters to the information already contained in the current set $\mathbf{X}^C$. On the other hand, if the difference is less than zero there is evidence that no useful information is added by the proposed variable.

4.2. *Proposed variable conditional distribution.* The conditional distribution of the proposed variable given $\mathbf{X}^C$ is modeled by a multinomial logistic regression using the *softmax* link function:

$$(4) \qquad p(X^P = c|\mathbf{X}^R \subseteq \mathbf{X}^C) = \frac{e^{\mathbf{X}^R \boldsymbol{\beta}_c}}{\sum_{c=1}^{C_P} e^{\mathbf{X}^R \boldsymbol{\beta}_c}},$$

where $\boldsymbol{\beta}_c$ is the vector of regression parameters for category c and $c = 1, \ldots, C_P$ are the categories for the proposed variable; the model reduces to a standard logistic regression with logit link if the proposed variable is binary. We refer to Ripley

(1996) and Agresti (2002) for a detailed description of the model and its estimation.

In the regression model (4), the subset $\mathbf{X}^R$ contains the relevant predictors of the proposed variable. Their selection is carried out using a standard stepwise algorithm described in the Supplementary Material, Section 2 [Fop, Smart and Murphy (2017)]. When selecting the variables that compose $\mathbf{X}^R$, we allow it to be the empty set, thus taking into account the general variable role modelling described in Maugis, Celeux and Martin-Magniette (2009b).

If the proposed variable is highly correlated with the predictors, the problem of *separation* may occur. Separation arises when a linear combination of the predictors perfectly or quasi-perfectly separates the classes of the response variable, leading to infinite estimates of the regression coefficients and large standard errors [Albert and Anderson (1984), Lesaffre and Albert (1989)]. Different remedies have been proposed in literature in order to perform inference on the parameters, for example, Heinze and Schemper (2002), Zorn (2005) and Gelman et al. (2008). In the present framework, separation does not represent a problem, as the regression coefficients are only accessory to the computation of the maximum of the log-likelihood of the logistic regression. In fact, even in case of separation the log-likelihood surface is concave, bounded above and has a finite maximum [Albert and Anderson (1984)]. In practice, if separation occurs the log-likelihood surface becomes flat, approaching a limiting value as some (or all) regression coefficients are going to infinity. So convergence criteria are satisfied, and the log-likelihood is numerically maximized and computation of quantities based on that maximum, such as the BIC, are still valid [Agresti (2015), Albert and Anderson (1984)].

4.3. *Swap-stepwise selection algorithm.* The clustering variables are selected using a stepwise algorithm which alternates between exclusion, inclusion and swapping steps. In the removal step, all the variables in $\mathbf{X}^C$ are examined in turn to be removed from the set. In the inclusion step, all the variables in $\mathbf{X}^O$ are examined in turn to be added to the clustering set. In the swapping step, a non-clustering variable is swapped with a clustering variable.

In the removal and inclusion step, we compare model $\mathcal{M}_1$ against model $\mathcal{M}_2$. Instead, in the swapping steps we actually compare two different configurations of model $\mathcal{M}_2$ that differ in the fact that one clustering variable is replaced by one of the non-clustering variables. The rationale for the swap step lies in the assumptions of model $\mathcal{M}_2$. In model $\mathcal{M}_2$, the proposed variable is assumed independent from $\mathbf{z}$ conditionally on the set of already selected variables and not marginally (which would be a special case). Therefore, X^P is actually allowed to contain some information about the clusters, which in some situations may be the best information available if one of the variables of the optimal set for $\mathbf{X}^C$ has been discarded during the search. Hence the algorithm could converge to a sub-optimum. To avoid it, we compare two different sets of clustering variables in the swapping step. Then if

a "true" clustering variable has been removed during the search, when compared to a less informative one is likely to be added back to the clustering set.

The algorithm also performs the selection of the number G of latent classes, finding at each stage the optimal combination of clustering variables and number of classes. The procedure stops when no change has been made to the set $\mathbf{X}^C$ after consecutive exclusion, swapping, inclusion and swapping steps.

A detailed description of the algorithm is in the Appendix.

4.4. *Comparing selected and discarded variables.* By means of the outlined variable selection procedure, we aim to remove variables that do not contain any information about the clustering *and* variables that contain *additional* information, which are redundant given the already selected relevant variables. Since it is likely that related variables carry similar information about the groups, it is of interest to analyze the association between each discarded variable and each selected one after the selection is performed. We accomplish this task as a result of simple considerations.

Let $X_o \in \mathbf{X}^O$ be one of the discarded variables, and $X_c \in \mathbf{X}^C$ be one of the selected ones; let also $\hat{\mathbf{z}}$ be the estimated cluster membership allocation vector. In the light of the described general modelling framework, we analyze the association between X_c and X_o by comparing the following two models for the joint conditional distribution $p(X_c, X_o | \hat{\mathbf{z}})$:

$$\mathcal{M}_{\text{as}}: \quad p(X_c, X_o | \hat{\mathbf{z}}) = p(X_c | \hat{\mathbf{z}}) p(X_o | X_c);$$
$$\mathcal{M}_{\text{no as}}: \quad p(X_c, X_o | \hat{\mathbf{z}}) = p(X_c | \hat{\mathbf{z}}) p(X_o).$$

In a similar fashion to the models involved in the variable selection procedure, this two models are compared via the Bayes factor $\mathcal{B}_{\text{as, no as}} = p(X_c, X_o | \hat{\mathbf{z}}, \mathcal{M}_{\text{as}}) / p(X_c, X_o | \hat{\mathbf{z}}, \mathcal{M}_{\text{no as}})$. Applying the same arguments of Section 4.1 and noting that $p(X_c | \hat{\mathbf{z}}, \mathcal{M}_{\text{as}}) = p(X_c | \hat{\mathbf{z}}, \mathcal{M}_{\text{no as}})$, we obtain that the above Bayes factor reduces to

$$\mathcal{B}_{\text{as, no as}} = \frac{p(X_o | X_c, \mathcal{M}_{\text{as}})}{p(X_o | \mathcal{M}_{\text{no as}})}.$$

Then using the BIC approximation of (2) leads to

(5) $\qquad \mathcal{B}_{\text{as, no as}} \approx \text{BIC}_{\text{diff as}} = \text{BIC}(X_o | X_c, \mathcal{M}_{\text{as}}) - \text{BIC}(X_o | \mathcal{M}_{\text{no as}}).$

The quantity $\text{BIC}_{\text{diff as}}$ corresponds to the difference between the BIC of a multinomial logistic regression where X_o depends on X_c and the BIC of the regression with only the constant terms. Then if this difference is greater than zero, there is evidence of the association between the considered selected variable and the discarded one.

Clustering summary of the LCA model for different sets of variables and different number of classes for the LBP data (*note that the BIC are not comparable for differing sets of variables*)

Selection method	Variables	N. latent classes	BIC	ARI
–	All	5	−12,586.48	0.50
–	All	3*	−12,763.81	0.82
Dean and Raftery	35 Criteria	5	−12,116.32	0.50
Stepwise	10 Criteria	3	−3462.82	0.66
Swap-stepwise	11 Criteria	3	−3946.31	0.75

* We fixed the number of classes to this value in advance.

5. Latent class model and clinical criteria selection. The proposed model is applied to the low back pain data. We measure the agreement between the model-based partition of the data and the expert-based classification using the adjusted Rand index (ARI), which is equal to 1 when two partitions are exactly the same; otherwise, it is close to 0 when they do not agree [Hubert and Arabie (1985)]; compared to other indices, Milligan and Cooper (1986) recommended the ARI as the index of choice for clustering validation.

We consider LCA models with the number of latent classes G ranging from 1 to 7. The clustering results for the different models are summarized in Table 1. When fitting a LCA model on all of the clinical criteria, the BIC selects a model with 5 classes, providing an ARI of 0.50. By fixing the number of classes equal to 3 in advance, we obtain a model with an ARI of 0.82. By performing the variable selection with the independence assumption of Dean and Raftery (2010), only one variable is discarded, Criterion 36, and the BIC selects again a model with 5 classes, identifying the same clusters of the model on all the variables. Note that also in White, Wyse and Murphy (2016) only one variable is discarded. Using the variable selection method proposed here with swap-stepwise search, we retain 11 variables and the BIC selects a 3-class model on these. The ARI for the model on the 11 selected clinical criteria is 0.75, thus the number of variables is reduced by about two thirds, identifying a partition of the patients that agrees well with the physiotherapists' classification. For comparison, we also performed the same variable selection with a standard stepwise search, selecting a model on 10 criteria, but with a smaller ARI. Therefore, the use of the swap move in the search avoided selection of sub-optimal informative clustering variables.

A cross-tabulation of the estimated partition on the 11 selected variables versus the expert-based classification is reported in Table 2. It seems reasonable to match the 3 detected classes to the nociceptive, peripheral neuropathic and central sensitization group, respectively.

Table 3 lists the 11 selected clinical criteria and the estimated probability of occurrence given the class which a patient is assigned to; also the observed proportion of occurrence is reported in brackets. Figure 3 is a heatmap of the estimated

TABLE 2
Cross-tabulation between the estimated partition on the 11 *clustering variables and the expert-based classification of the LBP data*

		Estimated		
		Class 1	Class 2	Class 3
Expert-based	Nociceptive	210	21	4
	Peripheral Neuropathic	5	88	2
	Central Sensitization	3	3	89

class conditional probabilities: the selected variables present good degree of separation between the three classes which are generally characterized by the almost full presence or almost complete absence of the selected criteria.

Smart et al. (2011) fit a logistic regression of each type of pain versus the others, ending with the selection of a set of 14 features whose presence or absence

TABLE 3
Estimated class conditional probability of occurrence and actual frequency (in brackets) for the selected clinical criteria in the low back pain data

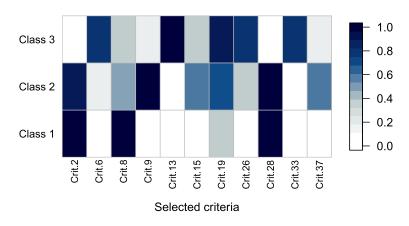| Crit.	Description	Class 1	Class 2	Class 3
2	Pain associated to trauma, pathologic process or dysfunction	0.94 (0.94)	0.90 (0.92)	0.04 (0.04)
6	More constant/unremitting pain	0.05 (0.04)	0.13 (0.17)	0.79 (0.79)
8	Pain localized to the area of injury/dysfunction	0.97 (0.97)	0.50 (0.42)	0.31 (0.33)
9	Pain referred in a dermatomal or cutaneous distribution	0.06 (0.12)	1.00 (0.97)	0.11 (0.13)
13	Disproportionate, non-mechanical, unpredictable pattern of pain	0.01 (0.01)	0.00 (0.01)	0.91 (0.87)
15	Pain in association with other dysesthesias	0.03 (0.06)	0.51 (0.51)	0.34 (0.34)
19	Night pain/disturbed sleep	0.34 (0.37)	0.70 (0.68)	0.86 (0.85)
26	Pain in association with high levels of functional disability	0.07 (0.09)	0.36 (0.36)	0.79 (0.78)
28	Clear, consistent and proportionate pattern of pain	0.97 (0.95)	0.94 (0.94)	0.07 (0.12)
33	Diffuse/non-anatomic areas of pain/tenderness on palpation	0.03 (0.03)	0.01 (0.01)	0.73 (0.73)
37	Pain/symptom provocation on palpation of relevant neural tissues	0.07 (0.09)	0.57 (0.58)	0.19 (0.21)

FIG. 3.   *Heatmap plot of the estimated probability of occurrence of the* 11 *selected clinical criteria.*

best describes each class of pain. They selected Criteria 3, 4, 5, 7, 8, 9, 11, 13, 15, 19, 25, 27, 29, 33. Six out of eleven of our selected criteria match those selected in a supervised setting. Furthermore, the estimated parameters reported in Table 3 agree with the description of the factors related to each class of pain given by the authors: Nociceptive pain (Class 1) is well described by the presence of a pain localized to the area of injury or dysfunction, and by the absence of dysaesthesias (unpleasant sensations, e.g., crawling) and pain at night; peripheral neuropathic (Class 2) is characterized by the presence of a dermatomal distribution of pain and pain on palpation of nerve tissue, and central neuropathic (Class 3) is linked to the presence of pain that is more constant and has a disproportionate and un-predictable pattern of provocation and is associated with diffuse areas of pain on palpation as well as the absence of pain in proportion to trauma or pathology in addition to consistent and proportionate pain on clinical provocation tests. Also, fitting a LCA model on the criteria selected by Smart et al. (2011), a model with 3 latent classes is chosen, with an ARI of 0.77. By comparing the latter partition with the classification of the LCA model on the 11 criteria of Table 3, an ARI of 0.79 is obtained. Thus the classification attained by the variable selection method in an unsupervised setting has a satisfying rate of agreement with the classifica-tion of patients based on the variables selected in a supervised setting, and with a smaller set of relevant clinical criteria. These findings provide some confirmatory discriminative validity evidence for a three-category mechanisms-based classifica-tion system for musculoskeletal pain. Furthermore, it is shown that the proposed method is able to reduce the number of useful clinical criteria to be checked for elaborating a preliminary assessment of the pain characteristics.

Discarded clinical criteria.   The clinical criteria in the data are specified in advance on a expert-consensus basis [Smart et al. (2010)]. Indeed they were chosen such that most are good in discriminating between the three types of pain. Here, we

TABLE 4
Cross-tabulation between the estimated 3-class partition on the discarded variables and the expert-based classification of the LBP data

		Estimated		
		Class 1	**Class 2**	**Class 3**
Expert-based	Nociceptive	208	25	2
	Peripheral Neuropathic	2	91	2
	Central Sensitization	8	1	86

want to point again the fact that the discarded criteria are removed from the set of clustering ones not only because they may not contain discriminative information about the pain classes, but also because they may carry information that is not needed, as it is already included in the set of selected ones.

We fit a LCA model on the 25 removed clinical criteria, selecting a model with 4 latent classes with a BIC of -9355.407 and an agreement to the experts' classification of 0.51. By setting in advance the number of classes equal to 3, we obtain a model with a BIC of -9470.552 and an ARI of 0.73. The cross tabulation of the fitted classification and the expert-based one for the 3-class model is presented in Table 4.

The partition thus obtained is comparable to the partition estimated on the selected clinical criteria. Therefore, by taking into consideration the discarded clinical criteria, it is still possible to get an acceptable classification of patients into clusters that sufficiently agrees with the expert-based classification. Thus it seems reasonable to consider the fact that the removed criteria are discarded mostly because they are redundant given the set of 11 selected clustering clinical criteria.

We check the association between each discarded clinical criterion and each selected one by calculating the BIC difference in (5). The computed differences range from -6.05 to 366.55 and the results are reported in Figure 4. Apart from Criterion 1 and Criterion 36, all the discarded criteria present evidence of association with some of the selected criteria. It is also worth to notice that Criterion 36 is the only discarded criterion in the Dean and Raftery (2010) modelling framework with the independence assumption between the proposed variable and the clustering ones.

6. Simulation study. In this section, we evaluate the proposed variable selection method through two different simulated data scenarios, also discussing the robustness of our method and comparing the results with the Dean and Raftery (2010) modelling framework. In both scenarios, we simulate 100 datasets for different sample sizes. The scenarios are sketched in Figures 5 and 6. The details of the simulation methodology are exposed in the Supplementary Material, Section 3 [Fop, Smart and Murphy (2017)].
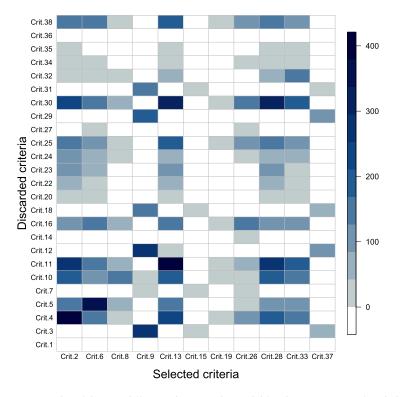
FIG. 4. *Heatmap plot of the BIC difference between the model for the association of each discarded clinical criteria and each selected one and the model for the independence. A white coloured cell indicates no evidence of association between the two variables.*
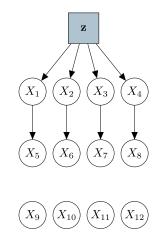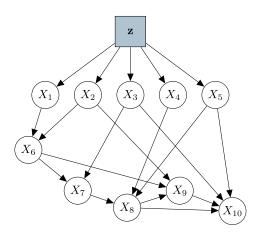

FIG. 5. *First simulated data scenario.*

FIG. 6. *Second simulated data scenario.*

6.1. *First scenario.* In the first simulation setting, we consider 12 categorical random variables. Figure 5 presents the scenario. Variables X_1, X_2, X_3, X_4 are the clustering variables, distributed according to a mixture of $G = 3$ multinomial distributions with mixing proportions 0.3, 0.5 and 0.2. Variables X_5, X_6, X_7, X_8 are redundant variables, each one generated dependent on one of the clustering variables. The last four variables, X_9, X_{10}, X_{11}, X_{12} are irrelevant variables not related to the previous ones. We consider three sample sizes: $N = 500, 750, 1000$. Figure 7 shows the proportion of times each variable was declared a clustering variable by our variable selection method and the variable selection with the independence assumption of Dean and Raftery (2010). Both methods are able to discard the noisy variables. Only the proposed method never selects almost any of the redundant variables, while the Dean and Raftery (2010) method includes also the redundant variables in the clustering set, especially as the sample size increases. Figure 8 displays the boxplots of the ARI between the actual classification of the data and the estimated classification from the LCA model fitted on: (i) all the variables (all), (ii) the "true" clustering variables (clus), (iii) the variables selected by the method with the Dean and Raftery (2010) assumption (selInd), (iv) the proposed method (selSwap). As expected, the inclusion of the redundant variables in the clustering set leads to a poor performance in terms of classification. Figure 9 presents the three most frequent sets of variables declared as clustering variables by our variable selection procedure. The most selected set is the one composed by the "true" clustering variables, and is the only one chosen for a sample size of 750. It is also worth noting that the other selected subsets contain mainly clustering variables.

6.2. *Second scenario.* In the second simulation setting, we consider 10 binary random variables. Figure 6 shows the scenario. Variables X_1, X_2, X_3, X_4, X_5
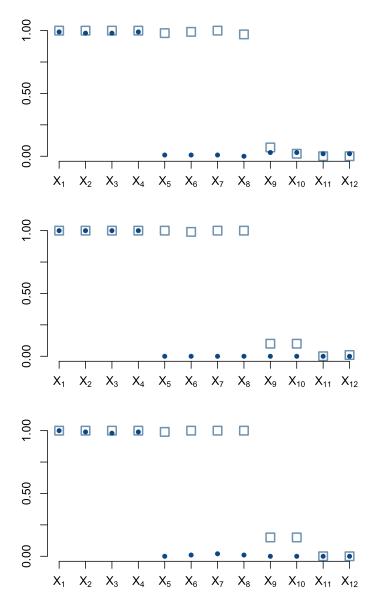
FIG. 7. *First simulation scenario*: *Proportions of times each variable has been declared a clustering variable by the proposed variable selection method* (*circle*) *and the variable selection method with the independence assumption of* Dean and Raftery (2010) (*square*). *From top*: *sample sizes corresponding to* 500, 750, 1000.

are the clustering variables, distributed according to a mixture of $G = 2$ binomial distributions with mixing proportions equal to 0.3 and 0.7. Variables $X_6, X_7, X_8, X_9, X_{10}$ are redundant variables; each one of these is generated in
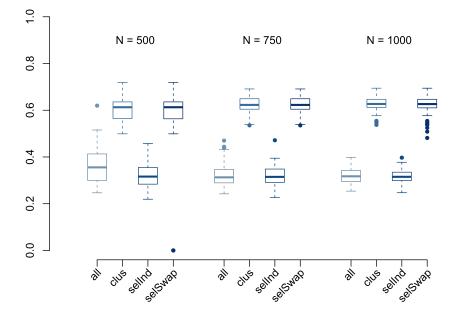
FIG. 8. *First simulation scenario*: *Boxplots of the ARI between the actual classification of the data and the estimated classification from the LCA model fitted on*: (i) *all the variables* (all), (ii) *the "true" clustering variables* (clus), (iii) *the variables selected by the method with the* Dean and Raftery (2010) *assumption* (selInd), (iv) *the proposed method* (selSwap).

order to be dependent on more than one of the clustering variables and the other redundant variables. We consider three sample sizes: $N = 750, 1000, 1500$. Figure 10 displays the proportion of times each variable was declared a clustering variable. The figure shows that the selection with the Dean and Raftery (2010) as-
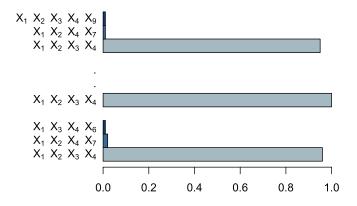

FIG. 9. *First simulation scenario*: *Proportions of the three most frequent sets of variables declared as relevant for clustering by the presented variable selection method. From top*: *sample sizes corresponding to* 500, 750, 1000.
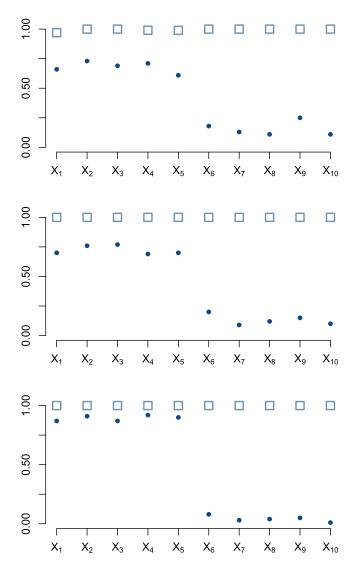
FIG. 10. *Second simulation scenario*: *Proportions of times each variable has been declared a clustering variable by the proposed variable selection method. From top*: *sample sizes corresponding to* 750, 1000, 1500.

sumption almost never discard any of the redundant variables. Furthermore, with the proposed method the probability of selecting a "true" clustering variable increases as N becomes larger. Figure 11 presents the boxplots of the ARI between the actual classification of the data and the estimated classifications. The classification of the observations based on the selected variables gives on average a better performance in terms of ARI. However, there are some situations in which the
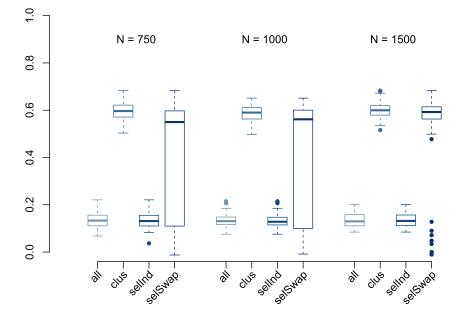
FIG. 11. *Second simulation scenario*: *Boxplots of the ARI between the actual classification of the data and the estimated classification from the LCA model fitted on*: (i) *all the variables* (all), (ii) *the "true" clustering variables* (clus), (iii) *the variables selected by the method with the Dean and Raftery* (2010) *assumption* (selInd), (iv) *the proposed method* (selSwap).

proposed method does not converge to the selection of the correct set of relevant variables. In Figure 12, the three most frequent sets declared as clustering variables are shown. Again, the set of "true" clustering variables is the one selected more often.
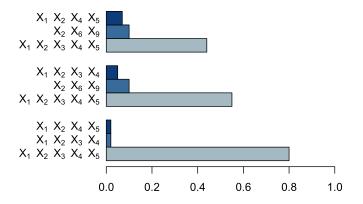

FIG. 12. *Second simulation scenario*: *Proportions of the three most frequent sets of variables declared as relevant for clustering. From top*: *sample sizes corresponding to* 750, 1000, 1500.

7. Discussion and further work.   In this paper, we have presented an improved variable selection method for LCA that overcomes the limitations of the Dean and Raftery (2010) and White, Wyse and Murphy (2016) methods, which lies in the independence assumption between the selected clustering variables and the variable proposed for removal or inclusion. The proposed method performs the selection of the most informative clustering variables, discarding those that are not informative and those that are redundant. The ability of the method of discriminating among relevant variables and redundant or non-informative variables has been shown in two simulated data settings.

The work was motivated by the nature of the LBP data examined. In the data, all the variables possess good discriminative power, since the clinical criteria list was built by experts in order to best identify the traits of the three classes of pain. The aim was to remove those criteria that are not needed because they contain similar group information to that already included in the selected clinical criteria. This resulted in a smaller set of criteria to be considered in order to derive a mechanisms-based classification of pain. Further, the modelling of the data in an unsupervised manner allowed for the validation of the mechanisms-based classification of pain because the patients clustered into groups that closely correspond to this classification.

We built the variable selection method on the model comparison framework pioneered by Law, Figueiredo and Jain (2004) and completely defined in Raftery and Dean (2006). However, another framework for performing variable selection is the regularization approach, although to the authors knowledge it has not been explored yet in categorical data clustering. Furthermore, for continuous data, Celeux et al. (2014) showed that the model comparison approach is a better methodology in terms of classification and variable selection accuracy than the recent regularization method of Witten and Tibshirani (2010).

We considered a greedy swap-stepwise searching algorithm to perform the variable selection. The idea of replacing a selected variable with one of the discarded variables has already been considered. For example, Miller (2002) in subset selection for regression presents a sequential replacement heuristic where in sequence each of the selected predictors is replaced by one of the non-selected variables. In a model-based clustering context, Tadesse, Sha and Vannucci (2005) and Kim, Tadesse and Vannucci (2006) use a stochastic search for Bayesian variable selection where the values of a latent variable selection indicator are randomly swapped. Many other searching strategies and metaheuristics could be used in order to conduct a robust search through the solution space and avoid local optima. For example, genetic algorithms [Goldberg (1989)] have already been applied for variable selection in cluster analysis for market segmentation [Liu and Ong (2008)] and subset selection for model-based clustering of continuous data [Scrucca (2016)]. In a high-dimensional problem with many variables, a forward algorithm and a headlong search [Badsberg (1992)] can be considered, as has been done in Dean

and Raftery (2010); although in this case, the problem of a good initialization of the clustering variables arises.

The variable selection method is developed in application to clinical criteria selection. However, it can be applied to any kind of multivariate categorical data, although its use is limited to only unordered categorical variables. A further extension in that direction is the incorporation of the capability of dealing with ordinal data, which often arise from likert scale questionnaires. In this context it is worth mentioning the work of Arima (2015), where a Bayesian approach is developed to reduce the items of a questionnaire used to evaluate patients' quality of life, with the goal that the reduced questionnaire will provide the same information of the complete questionnaire. Another limitation of our methodology lies in the local independence assumption of the LCA model. Much work has been done towards relaxing this assumption and allowing class-conditional dependencies between the variables. Among the most recent, Gollini and Murphy (2014) present a setting where it is assumed that the class distribution of the categorical variables depends on a number of continuous latent variables, which allow to model the dependences among the observed categorical variables. Another approach is the one by Marbac, Biernacki and Vandewalle (2015), where conditional on a class, the variables are grouped into independent blocks, each one following a specific distribution that takes into account the dependency between variables. Including these frameworks in a variable selection method for clustering categorical data could be promising and may be of interest for further developments.

R [R Core Team (2017)] functions implementing the variable selection method for LCA are available in the Supplementary Material [Fop, Smart and Murphy (2017)]. In the code, the LCA model is fitted using the poLCA package [Linzer and Lewis (2011)].

APPENDIX: SWAP-STEPWISE SELECTION ALGORITHM

Here, we give a more detailed description of the swap-stepwise variable selection algorithm for the LCA model. At each stage of the algorithm, a greedy search over the model space is conducted and all the variables are examined for being removed, added or swapped.

Note that in fitting the LCA model we perform multiple runs with random starting values. Also in this case the aim is to allow the search for the global maximum of the log-likelihood rather than a local one; then the model with the greatest log-likelihood is retained. In the following, in the notation we drop the conditioning on the model $\mathcal{M}$ for ease of reading.

Initialization. Set $G_{\max}$, the maximum number of clusters to be considered for the data. Then when fitting the LCA models, a maximum number $G^* \leq G_{\max}$ of latent classes will be considered at each stage. Here, G^* is the maximum number of latent classes that satisfies the identifiability condition in (3) for the set of variables currently taken into account in fitting the LCA model.

Initialize the set of clustering variables and the set of non-clustering variables by assigning $\mathbf{X}^C = \mathbf{X}$ and $\mathbf{X}^O = \varnothing$, respectively.

Removal step. Fit a LCA model on all the elements contained in the current set of clustering variables $\mathbf{X}^C$, for $1 \leq G \leq G^*$ and set

$$\text{BIC}_{\text{clus}} = \max_G \{\text{BIC}(\mathbf{X}^C | \mathbf{z})\}.$$

Then for each variable $X_j^C \in \mathbf{X}^C$ compute

$$\text{BIC}_{\text{no clus}}(X_j^C) = \max_G \{\text{BIC}(\mathbf{X}_{-j}^C | \mathbf{z})\} + \text{BIC}(X_j^C | \mathbf{X}_j^R \subseteq \mathbf{X}_{-j}^C),$$

where $\mathbf{X}_{-j}^C = \mathbf{X}^C \setminus X_j^C$; $\text{BIC}(\mathbf{X}_{-j}^C | \mathbf{z})$ is the BIC of the latent class model on the current clustering variables after removing variable X_j^C, maximized over $1 \leq G \leq G^*$; $\text{BIC}(X_j^C | \mathbf{X}_j^R \subseteq \mathbf{X}_{-j}^C)$ is the BIC of the multinomial logistic regression model of variable X_j^C given the set $\mathbf{X}_j^R$ of selected predictors obtained using the algorithm outlined in the Supplementary Material, Section 2 [Fop, Smart and Murphy (2017)].

Subsequently, for each variable in $\mathbf{X}^C$ estimate the evidence of being a relevant clustering one versus the evidence of not being useful for clustering by computing the difference:

$$\text{BIC}_{\text{diff}}(X_j^C) = \text{BIC}_{\text{clus}} - \text{BIC}_{\text{no clus}}(X_j^C).$$

According to the values of $\text{BIC}_{\text{diff}}(X_j^C)$ rank the current clustering variables in increasing order, generating the ordered set $\{X_{(1)}^C, X_{(2)}^C, \ldots, X_{(M_C)}^C\}$, with M_C the number of variables in the current set $\mathbf{X}^C$. Then $X_{(1)}^C$ is such that

$$X_{(1)}^C = \arg \min_{X_j^C \in \mathbf{X}^C} \text{BIC}_{\text{diff}}(X_j^C).$$

Set $X^P = X_{(1)}^C$ and propose it for removal. Next, if $\text{BIC}_{\text{diff}}(X^P) < 0$, remove the proposed variable from $\mathbf{X}^C$ and set $\mathbf{X}^C = \mathbf{X}^C \setminus X^P$ and $\mathbf{X}^O = \mathbf{X}^O \cup X^P$; otherwise leave the set of clustering variables unchanged. Go to the swapping step 1.

Swapping step 1. If a variable has been removed in the removal step, set $X_{\text{swap}} = X_{(2)}^C$, otherwise set $X_{\text{swap}} = X_{(1)}^C$.

Swap each variable $X_k^O \in \mathbf{X}^O$ with X_{swap} generating the sets $\mathbf{X}_k^C = \mathbf{X}^C \setminus X_{\text{swap}} \cup X_k^O$. Fit a LCA model on the set of variables $\mathbf{X}_k^C$ for $1 \leq G \leq G^*$ and compute

$$\text{BIC}_{\text{clus}}(X_k^O) = \max_G \{\text{BIC}(\mathbf{X}_k^C | \mathbf{z})\} + \text{BIC}(X_{\text{swap}} | \mathbf{X}_{\text{swap}}^R \subseteq \mathbf{X}_k^C),$$

where $\text{BIC}(\mathbf{X}_k^C | \mathbf{z})$ is the BIC of the latent class model on the current clustering variables after swapping the variable X_{swap} with the variable X_k^O, and

$\text{BIC}(X_{\text{swap}}|\mathbf{X}^R_{\text{swap}} \subseteq \mathbf{X}^C_k)$ is the BIC of the multinomial logistic regression of variable X_{swap} given the selected predictors $\mathbf{X}^R_{\text{swap}}$.

Then calculate

$$\text{BIC}_{\text{no clus}}(X^O_k) = \max_G\{\text{BIC}(\mathbf{X}^C|\mathbf{z})\} + \text{BIC}(X^O_k|\mathbf{X}^R_k \subseteq \mathbf{X}^C),$$

where $\text{BIC}(X^O_k|\mathbf{X}^R_k \subseteq \mathbf{X}^C)$ is the BIC of the multinomial logistic regression model of variable X^O_k given the set $\mathbf{X}^R_k$ of relevant predictors in $\mathbf{X}^C$.

Subsequently, for each variable in $\mathbf{X}^O$ estimate the evidence of carrying more clustering information than X_{swap} versus the evidence of containing less clustering information by computing the difference

$$\text{BIC}_{\text{diff}}(X^O_k) = \text{BIC}_{\text{clus}}(X^O_k) - \text{BIC}_{\text{no clus}}(X^O_k),$$

and propose for swapping with X_{swap} the variable X^P such that

$$X^P = \arg \max_{X^O_k \in \mathbf{X}^O} \text{BIC}_{\text{diff}}(X^O_k).$$

Then if $\text{BIC}_{\text{diff}}(X^P) > 0$, replace X_{swap} by the proposed variable and set $\mathbf{X}^C = \mathbf{X}^C \setminus X_{\text{swap}} \cup X^P$ and $\mathbf{X}^O = \mathbf{X}^O \setminus X^P \cup X_{\text{swap}}$; otherwise, leave the set of clustering variables unchanged. Go to the inclusion step.

Inclusion step. For each variable $X^O_k \in \mathbf{X}^O$, compute

$$\text{BIC}_{\text{clus}}(X^O_k) = \max_G\{\text{BIC}(\mathbf{X}^C_{+k}|\mathbf{z})\},$$

where $\mathbf{X}^C_{+k} = \mathbf{X}^C \cup X^O_k$; $\text{BIC}(\mathbf{X}^C_{+k}|\mathbf{z})$ is the BIC of the latent class model on the current clustering variables after adding variable X^O_k.

Then compute

$$\text{BIC}_{\text{no clus}}(X^O_k) = \max_G\{\text{BIC}(\mathbf{X}^C|\mathbf{z})\} + \text{BIC}(X^O_k|\mathbf{X}^R_k \subseteq \mathbf{X}^C).$$

Subsequently, for each variable X^O_k estimate the evidence of being a clustering variable versus the evidence of not being useful for clustering by computing the difference:

$$\text{BIC}_{\text{diff}}(X^O_k) = \text{BIC}_{\text{clus}}(X^O_k) - \text{BIC}_{\text{no clus}}(X^O_k).$$

According to the values of $\text{BIC}_{\text{diff}}(X^O_k)$ rank the current non-clustering variables in decreasing order, generating the ordered set $\{X^O_{(1)}, X^O_{(2)}, \ldots, X^O_{(M_O)}\}$, with M_O the number of variables in the current set $\mathbf{X}^O$. Then $X^O_{(1)}$ is such that

$$X^O_{(1)} = \arg \max_{X^O_j \in \mathbf{X}^O} \text{BIC}_{\text{diff}}(X^O_j).$$

Set $X^P = X^O_{(1)}$ and propose it for inclusion in the clustering set. Next, if $\text{BIC}_{\text{diff}}(X^P) > 0$, add the proposed variable to $\mathbf{X}^C$ and set $\mathbf{X}^C = \mathbf{X}^C \cup X^P$ and

$\mathbf{X}^O = \mathbf{X}^O \setminus X^P$; otherwise, leave the set of clustering variables unchanged. Go to the swapping step 2.

Swapping step 2. If a variable has been added in the inclusion step, set $X_{\text{swap}} = X^O_{(2)}$; otherwise, set $X_{\text{swap}} = X^O_{(1)}$.

Swap each variable $X^C_j \in \mathbf{X}^C$ with X_{swap} generating the sets $\mathbf{X}^C_j = \mathbf{X}^C \cup X_{\text{swap}} \setminus X^C_j$. Compute

$$\text{BIC}_{\text{clus}}(X^C_j) = \max_G \{\text{BIC}(\mathbf{X}^C|\mathbf{z})\} + \text{BIC}(X_{\text{swap}}|\mathbf{X}^R_{\text{swap}} \subseteq \mathbf{X}^C).$$

Then fit a LCA model on the set of variables $\mathbf{X}^C_j$ for $1 \leq G \leq G^*$ and calculate

$$\text{BIC}_{\text{no clus}}(X^C_j) = \max_G \{\text{BIC}(\mathbf{X}^C_j|\mathbf{z})\} + \text{BIC}(X^C_j|\mathbf{X}^R_j \subseteq \mathbf{X}^C_j),$$

where $\text{BIC}(\mathbf{X}^C_j|\mathbf{z})$ is the BIC of the latent class model on the current clustering variables after swapping the variable X_{swap} with the variable X^C_j, and $\text{BIC}(X^C_j|\mathbf{X}^R_j \subseteq \mathbf{X}^C_j)$ is the BIC of the multinomial logistic regression model of variable X^C_j given the set $\mathbf{X}^R_j$ of relevant predictors in $\mathbf{X}^C_j$.

Subsequently, for each variable in $\mathbf{X}^C$ estimate the evidence of carrying more clustering information than X_{swap} versus the evidence of containing less clustering information by computing the difference

$$\text{BIC}_{\text{diff}}(X^C_j) = \text{BIC}_{\text{clus}}(X^C_j) - \text{BIC}_{\text{no clus}}(X^C_j),$$

and propose for swapping with X_{swap} the variable X^P such that

$$X^P = \arg \min_{X^C_j \in \mathbf{X}^C} \text{BIC}_{\text{diff}}(X^C_j).$$

Then if $\text{BIC}_{\text{diff}}(X^P) < 0$, replace X_{swap} by the proposed variable and set $\mathbf{X}^C = \mathbf{X}^C \setminus X_{\text{swap}} \cup X^P$ and $\mathbf{X}^O = \mathbf{X}^O \setminus X^P \cup X_{\text{swap}}$; otherwise, leave the set of clustering variables unchanged.

The algorithm starts with two successive removal steps, then it iterates alternating between removal, swapping, inclusion, swapping steps. It stops when all the moves are rejected since no further change can be produced on the set of clustering variables. In the swapping steps, we do not look at all possible pairs of variables because it could be too computational demanding. Instead we consider the variable with the largest evidence of being removed or added, because it is the one most likely to be swapped.

SUPPLEMENTARY MATERIAL

Supplementary information, data and R code [Fop, Smart and Murphy (2017)]. (DOI: 10.1214/17-AOAS1061SUPP; .zip). The .zip folder contains a document with: further considerations regarding the "don't know" entries, a description of the backward-stepwise selection algorithm for the multinomial logistic regression, a detailed description of the simulated data experiments, a complete list of clinical criteria and a notation page for reference. The folder also contains the data used in this paper and R code implementing the variable selection method.

REFERENCES

AGRESTI, A. (2002). *Categorical Data Analysis*, 2nd ed. Wiley, New York. MR1914507

AGRESTI, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley, Hoboken, NJ. MR3308143

ALBERT, A. and ANDERSON, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71** 1–10. MR0738319

ARIMA, S. (2015). Item selection via Bayesian IRT models. *Stat. Med.* **34** 487–503. MR3301590

BADSBERG, J. H. (1992). Model search in contingency tables by CoCo. In *Computational Statistics*, *Vol.* 1 (Y. Dodge and J. Whittaker, eds.) 251–256. Physica-Verlag, Heidelberg.

BARTHOLOMEW, D., KNOTT, M. and MOUSTAKI, I. (2011). *Latent Variable Models and Factor Analysis*: *A Unified Approach*, 3rd ed. Wiley, Chichester. MR2849614

BARTOLUCCI, F., MONTANARI, G. E. and PANDOLFI, S. (2016). Item selection by latent class-based methods: An application to nursing home evaluation. *Adv. Data Anal. Classif.* **10** 245–262. MR3505059

BONTEMPS, D. and TOUSSILE, W. (2013). Clustering and variable selection for categorical multivariate data. *Electron. J. Stat.* **7** 2344–2371. MR3108816

CELEUX, G., MARTIN-MAGNIETTE, M.-L., MAUGIS-RABUSSEAU, C. and RAFTERY, A. E. (2014). Comparing model selection and regularization approaches to variable selection in model-based clustering. *J. SFdS* **155** 57–71. MR3211754

CLOGG, C. C. (1988). Latent class models for measuring. In *Latent Trait and Latent Class Models* (R. Langeheine and J. Rost, eds.) 173–205. Plenum, New York.

CLOGG, C. C. (1995). Latent class models. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (G. Arminger, C. C. Clogg and M. E. Sobel, eds.) 311–360. Plenum, New York.

DEAN, N. and RAFTERY, A. E. (2010). Latent class analysis variable selection. *Ann. Inst. Statist. Math.* **62** 11–35. MR2577437

DY, J. G. and BRODLEY, C. E. (2003/04). Feature selection for unsupervised learning. *J. Mach. Learn. Res.* **5** 845–889. MR2248002

FOP, M., SMART, K. M. and MURPHY, T. B. (2017). Supplement to "Variable selection for latent class analysis with application to low back pain diagnosis." DOI:10.1214/17-AOAS1061SUPP.

FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635

FROUD, R., PATTERSON, S., ELDRIDGE, S., SEALE, C., PINCUS, T., RAJENDRAN, D., FOSSUM, C. and UNDERWOOD, M. (2014). A systematic review and meta-synthesis of the impact of low back pain on people's lives. *BMC Musculoskeletal Disorders* **15** 1–14.

GELMAN, A., JAKULIN, A., PITTAU, M. G. and SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2** 1360–1383. MR2655663

GOLDBERG, D. (1989). *Genetic Algorithms in Search*, *Optimization*, *and Machine Learning*. Addison-Wesley, Reading.

GOLLINI, I. and MURPHY, T. B. (2014). Mixture of latent trait analyzers for model-based clustering of categorical data. *Stat. Comput.* **24** 569–588. MR3223542

GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215–231. MR0370936

GRAVEN-NIELSEN, T. and ARENDT-NIELSEN, L. (2010). Assessment of mechanisms in localized and widespread musculoskeletal pain. *Nat Rev Rheumatol* **6** 599–606.

HABERMAN, S. J. (1979). *Analysis of Qualitative Data*, *Vol.* 2: *New Developments*. Academic Press, New York.

HEINZE, G. and SCHEMPER, M. (2002). A solution to the problem of separation in logistic regression. *Stat. Med.* **21** 2409–2419.

HOY, D., BAIN, C., WILLIAMS, G., MARCH, L., BROOKS, P., BLYTH, F., WOOLF, A., VOS, T. and BUCHBINDER, R. (2012). A systematic review of the global prevalence of low back pain. *Arthritis & Rheumatism* **64** 2028–2037.

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. MR3363402

KATZ, J. N., STOCK, S. R., EVANOFF, B. A., REMPEL, D., MOORE, J. S., FRANZBLAU, A. and GRAY, R. H. (2000). Classification criteria and severity assessment in work-associated upper extremity disorders: Methods matter. *Am. J. Ind. Med.* **38** 369–372.

KIM, S., TADESSE, M. G. and VANNUCCI, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika* **93** 877–893. MR2285077

LAW, M. H. C., FIGUEIREDO, M. A. T. and JAIN, A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **26** 1154–1166.

LAZARSFELD, P. F. and HENRY, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston, MA.

LESAFFRE, E. and ALBERT, A. (1989). Partial separation in logistic discrimination. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **51** 109–116. MR0984997

LINZER, D. A. and LEWIS, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *J. Stat. Softw.* **42** 1–29.

LIU, H. H. and ONG, C. S. (2008). Variable selection in clustering for marketing segmentation using genetic algorithms. *Expert Systems with Applications* **34** 502–510.

MALSINER-WALLI, G., FRÜHWIRTH-SCHNATTER, S. and GRÜN, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Stat. Comput.* **26** 303–324. MR3439375

MARBAC, M., BIERNACKI, C. and VANDEWALLE, V. (2015). Model-based clustering for conditionally correlated categorical data. *J. Classification* **32** 145–175. MR3369412

MARBAC, M. and SEDKI, M. (2017). Variable selection for model-based clustering using the integrated complete-data likelihood. *Stat. Comput.* **27** 1049–1063. MR3627562

MAUGIS, C., CELEUX, G. and MARTIN-MAGNIETTE, M.-L. (2009a). Variable selection for clustering with Gaussian mixture models. *Biometrics* **65** 701–709. MR2649842

MAUGIS, C., CELEUX, G. and MARTIN-MAGNIETTE, M.-L. (2009b). Variable selection in model-based clustering: A general variable role modeling. *Comput. Statist. Data Anal.* **53** 3872–3882. MR2749931

MCLACHLAN, G. J. and KRISHNAN, T. (2008). *The EM Algorithm and Extensions*, 2nd ed. Wiley, Hoboken, NJ. MR2392878

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York. MR1789474

MCLACHLAN, G. J. and RATHNAYAKE, S. (2014). On the number of components in a Gaussian mixture model. *Wiley Interdisciplinary Reviews*: *Data Mining and Knowledge Discovery* **4** 341–355.

MCNICHOLAS, P. D. (2016). Model-based clustering. *J. Classification* **33** 331–373. MR3575621

MERSKEY, H. and BOGDUK, N., eds. (2002). *Classification of Chronic Pain*. IASP Press, Washington, DC.

MEYNET, C. and MAUGIS-RABUSSEAU, C. (2012). A sparse variable selection procedure in model-based clustering. *INRIA Report*.

MILLER, A. (2002). *Subset Selection in Regression*, 2nd ed. *Monographs on Statistics and Applied Probability* **95**. Chapman & Hall/CRC, Boca Raton, FL. MR2001193

MILLIGAN, G. W. and COOPER, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research* **21** 441–458.

MURPHY, T. B., DEAN, N. and RAFTERY, A. E. (2010). Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *Ann. Appl. Stat.* **4** 396–421. MR2758177

NIJS, J., APELDOORN, A., HALLEGRAEFF, H., CLARK, J., SMEETS, R., MALFLIET, A., GIRBES, E. L., KOONING, M. D. and ICKMANS, K. (2015) Low back pain: Guidelines for the clinical classification of predominant neuropathic, nociceptive, or central sensitization pain. *Pain Physician* **18** E333–E346.

PAN, W. and SHEN, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.* **8** 1145–1164.

R CORE TEAM (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

RAFTERY, A. E. and DEAN, N. (2006). Variable selection for model-based clustering. *J. Amer. Statist. Assoc.* **101** 168–178.

RIPLEY, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge Univ. Press, Cambridge. MR1438788

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SCRUCCA, L. (2016). Genetic algorithms for subset selection in model-based clustering. In *Unsupervised Learning Algorithms* (M. E. Celebi and K. Aydin, eds.) 55–70. Springer, Berlin.

SCRUCCA, L. and RAFTERY, A. E. (2017). Clustvarsel: A package implementing variable selection for model-based clustering in R. *J. Stat. Softw.* To appear.

SILVESTRE, C., CARDOSO, M. G. M. S. and FIGUEIREDO, M. (2015). Feature selection for clustering categorical data with an embedded modelling approach. *Expert Systems* **32** 444–453.

SMART, K. M., O'CONNELL, N. E. and DOODY, C. (2008). Towards a mechanisms-based classification of pain in musculoskeletal physiotherapy? *Physical Therapy Reviews* **13** 1–10.

SMART, K. M., BLAKE, C., STAINES, A. and DOODY, C. (2010). Clinical indicators of "nociceptive", "peripheral neuropathic" and "central" mechanisms of musculoskeletal pain. A Delphi survey of expert clinicians. *Manual Therapy* **15** 80–87.

SMART, K. M., BLAKE, C., STAINES, A. and DOODY, C. (2011). The discriminative validity of "nociceptive","peripheral neuropathic", and "central sensitization" as mechanisms-based classifications of musculoskeletal pain. *The Clinical Journal of Pain* **27** 655–663.

STYNES, S., KONSTANTINOU, K. and DUNN, K. M. (2016). Classification of patients with low back-related leg pain: A systematic review. *BMC Musculoskelet Disord* **17** 226.

TADESSE, M. G., SHA, N. and VANNUCCI, M. (2005). Bayesian variable selection in clustering high-dimensional data. *J. Amer. Statist. Assoc.* **100** 602–617. MR2160563

WALKER, B. F. (2000). The prevalence of low back pain: A systematic review of the literature from 1966 to 1998. *Journal of Spinal Disorders & Techniques* **13** 205–217.

WANG, S. and ZHU, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics* **64** 440–448, 666. MR2432414

WHITE, A., WYSE, J. and MURPHY, T. B. (2016). Bayesian variable selection for latent class analysis using a collapsed Gibbs sampler. *Stat. Comput.* **26** 511–527. MR3439388

WITTEN, D. M. and TIBSHIRANI, R. (2010). A framework for feature selection in clustering. *J. Amer. Statist. Assoc.* **105** 713–726. MR2724855

WOOLF, C. J. (2004). Pain: Moving from symptom control toward mechanism-specific pharmaco-
    logic management. *Ann. Intern. Med.* **140** 441–451.
WOOLF, C. J., BENNETT, G. J., DOHERTY, M., DUBNER, R., KIDD, B., KOLTZENBURG, M.,
    LIPTON, R., LOESER, J. D., PAYNE, R. and TOREBJORK, E. (1998). Towards a mechanism-
    based classification of pain? *Pain* **77** 227–229.
XIE, B., PAN, W. and SHEN, X. (2008). Penalized model-based clustering with cluster-specific
    diagonal covariance matrices and grouped variables. *Electron. J. Stat.* **2** 168–212. MR2386092
ZHANG, Q. and IP, E. H. (2014). Variable assessment in latent class models. *Comput. Statist. Data
    Anal.* **77** 146–156. MR3210054
ZORN, C. (2005). A solution to separation in binary response models. *Polit. Anal.* **13** 157–170.

M. FOP
T. B. MURPHY
SCHOOL OF MATHEMATICS & STATISTICS
    AND INSIGHT RESEARCH CENTRE
UNIVERSITY COLLEGE DUBLIN
BELFIELD, DUBLIN 4
IRELAND
E-MAIL: michael.fop@ucdconnect.ie
        brendan.murphy@ucd.ie

K. M. SMART
ST. VINCENT'S UNIVERSITY HOSPITAL
DUBLIN 4
IRELAND
E-MAIL: k.smart@svuh.ie