Variable selection for model-based clustering of categorical data

Brendan Murphy

Wirtschaftsuniversität Wien Seminar, 2016

Alzheimer Dataset

- Data were collected on early onset Alzheimer patient symptoms in St. James' Hospital, Dublin.
- Two hundred and forty patients had six behavioural and psychological symptoms (Hallucination, Activity, Aggression, Agitation, Diurnal and Affective) recorded.
- Number of distinct groups of patients gives an idea of the number of subclasses or syndromes.
- Which symptoms distinguish the groups? Can some subset better distinguish syndromes?
- Previous studies: difficulty determining whether two or three groups are more suitable to describe data.

Back Pain Dataset

- A study to investigate the use of a mechanisms-based classification of muscoloskeletal pain in clinical practice.
- The aim of the study was to asses the discriminative power of the taxonomy of pain in *Nociceptive*, *Peripheral Neuropathic* and *Central Sensitization* for low-back disorders.
- There are $N = 464$ patients who were assessed according to a list of 36 binary clinical indicators ("Present" / "Absent").
- Some of the indicators carry the same information about the pain categories, thus the interest here is to select a subset of most relevant clinical criteria, performing a partition of the patients.
- Does the partition of the patients agree with the clinical taxonomy?

Clustering and Variable Selection

- The motivating examples show the need for:
  - **Clustering:** Can we establish the existance of subgroups? How can we characterize these subgroups?
  - **Variable Selection:** Can we use a subset of the variables to distinguish the subgroups?

Model-Based Clustering/Mixture Models

- Denote the $N \times M$ data matrix by $\mathbf{X}$
- The nth observation is denoted by $\mathbf{X}_n$.
- Model-based clustering assumes that $\mathbf{X}_n$ arises from a finite mixture model
- Assuming G classes (components)

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

- τ_g are mixture weights
- $p(\mathbf{X}_n|\boldsymbol{\theta}_g)$ is the component distribution.

Latent Class Analysis (LCA) model

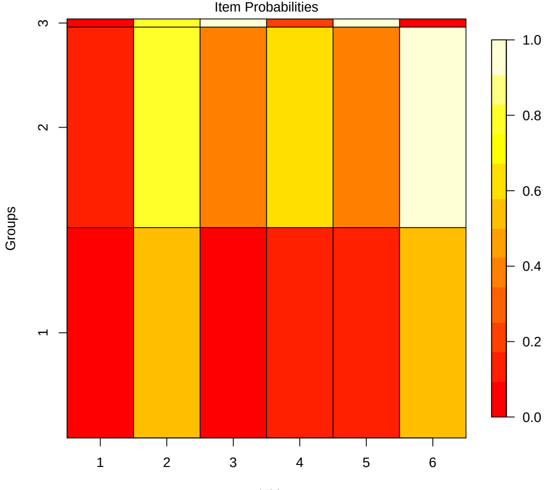
- Latent Class Analysis (LCA) is a model for clustering categorical data.
- Let $\mathbf{X}_n = (X_{n1}, X_{n2}, \ldots, X_{nM})$ where X_{nm} takes a value from $\{1, 2, \ldots, C_m\}$.
- In LCA we assume that there is local independence between variables, so that if we knew $\mathbf{X}_n$ was in class g we could write it's density as

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

  where $\{\theta_{gm1}, \ldots, \theta_{gmC_m}\}$ give the probabilities of observing the categories $\{1, \ldots, C_m\}$ in variable m
- $\boldsymbol{\theta}_g$ will characterize and embody the differences between groups

Example: Alzheimer Dataset

Result for three group model from BayesLCA package (White & Murphy, 2014)



LCA model (general)

- Model likelihood of the form,

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

- More convenient to work with completed data
- Augment data with class labels $\mathbf{Z}_n = (Z_{n1}, Z_{n2}, \ldots, Z_{nG})$ where

$$Z_{ng} = \begin{cases} 1 & \text{if observation } n \text{ belongs to group } g \\ 0 & \text{otherwise.} \end{cases}$$

- Then we can write down completed data likelihood for an observation

$$p(\mathbf{X}_n, \mathbf{Z}_n|\boldsymbol{\theta}, \boldsymbol{\tau}, G) = \prod_{g=1}^{G} \left\{ \tau_g \prod_{m=1}^{M} \prod_{c=1}^{C_m} \theta_{gmc}^{\mathrm{I}(X_{nm}=c)} \right\}^{Z_{ng}}.$$

LCA model (general)

- Estimation by EM algorithm or VB (see BayesLCA package)
- Note that G must be chosen in advance; possible to discriminate the best G for the data using information criteria (eg. BIC)
- Bayesian approaches: Pandolfi, Bartolucci and Friel (2014) use reversible jump to get posterior probability for G

Bayesian variable selection in LCA model

- Consider the variables that are useful for clustering in the model
- Let $\boldsymbol{\nu}_{\mathrm{cl}}$ be a vector containing the indexes of the set of variables used for clustering the data
- $\boldsymbol{\nu}_{\mathrm{n}}$ contain the remaining indexes
- This splits the observed categorical variables into those with discriminating power, and those without.

Bayesian variable selection in LCA model

- Then for the variables used in clustering

$$p_{\mathrm{cl}}(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{\tau}, G) = \prod_{n=1}^{N} \prod_{g=1}^{G} \left\{ \tau_g \prod_{m \in \nu_{\mathrm{cl}}} \prod_{c=1}^{C_m} \theta_{gmc}^{\mathrm{I}(X_{nm}=c)} \right\}^{Z_{ng}}$$

- ... and for those not used

$$p_{\mathrm{n}}(\mathbf{X}|\boldsymbol{\rho}, \boldsymbol{\nu}) = \prod_{n=1}^{N} \prod_{m \in \nu_{\mathrm{n}}} \prod_{c=1}^{C_m} \rho_{mc}^{\mathrm{I}(X_{nm}=c)},$$

- ρ_{mc} is the probability of variable m having category c and is the same for all items

Bayesian variable selection in LCA model

- Priors are Dirichlet on the item probabilities in each class for the discriminating variables, and also in the non-discriminating variables

$$p(\boldsymbol{\theta}_{gm}|\beta) = \frac{\Gamma\left(C_m\beta\right)}{\Gamma\left(\beta\right)^{C_m}} \prod_{c=1}^{C_m} \theta_{gmc}^{\beta-1}.$$

$$p(\boldsymbol{\rho}_{m}|\beta) = \frac{\Gamma\left(C_m\beta\right)}{\Gamma\left(\beta\right)^{C_m}} \prod_{c=1}^{C_m} \rho_{mc}^{\beta-1}.$$

- Prior on class probabilities also Dirichlet

$$p(\boldsymbol{\tau}|\alpha, G) = \frac{\Gamma\left(G\alpha\right)}{\Gamma\left(\alpha\right)^{G}} \prod_{g=1}^{G} \tau_{g}^{\alpha-1}$$

Bayesian variable selection in LCA model

- We aim to explore uncertainty in the number of groups G **and** the variables used for clustering

- Take a prior on G also. We employ the prior of Nobile and Fearnside (2007), that was justified for this problem in a similar context
$$p(G) \propto \frac{1}{G!}$$
normalized over $1, \ldots, G_{\max}$

- In fact the work we present here, brings that of Nobile and Fearnside (2007) (for Gaussian mixtures) into the categorical data domain

Bayesian variable selection in LCA model

- Variables are assumed to be included *a priori* following a Bernoulli with parameter π

$$p(\boldsymbol{\nu}|\pi) = \prod_{m \in \nu_{\mathrm{cl}}} \pi \prod_{m \in \nu_{\mathrm{n}}} (1 - \pi).$$

- Usually there will only be enough information to set something like $\pi = 0.5$ in a practical situation

- Ley and Steel (2009) investigate putting a Beta(a_0, b_0) hyperprior on π; we tried this but found no notable difference in results

Bayesian variable selection in LCA model

If we write down the model in it's full form, we get a joint posterior on item probabilities over classes, class probabilities, labels and the number of classes: the full completed likelihood is

$$p_{\mathrm{full}}(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\nu}, \boldsymbol{\tau}, G) = p_{\mathrm{cl}}(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{\tau}, G)p_{\mathrm{n}}(\mathbf{X}|\boldsymbol{\rho}, \boldsymbol{\nu})$$

and posterior is

$$\begin{aligned}
p(G, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\nu}, \boldsymbol{\tau}|\mathbf{X}, \alpha, \pi, \beta) \quad \propto \quad & p_{\mathrm{full}}(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\nu}, \boldsymbol{\tau}, G) \\
\times \quad & p(\boldsymbol{\tau}|\alpha, G)p(\boldsymbol{\nu}|\pi) \\
\times \quad & \prod_{m \in \nu_{\mathrm{n}}} p(\boldsymbol{\rho}_m|\beta) \\
\times \quad & \prod_{g=1}^{G} \prod_{m \in \nu_{\mathrm{cl}}} p(\boldsymbol{\theta}_{gm}|\beta) \\
\times \quad & p(G).
\end{aligned}$$

Marginalization approach

- Using normalizing constants for the Dirichlet distribution it turns out that

$$
\begin{aligned}
& p(G, \mathbf{Z}, \boldsymbol{\nu} | \mathbf{X}, \alpha, \pi, \beta) \\
\propto \quad & p(G) \int p(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\nu}, \boldsymbol{\tau} | \mathbf{X}, G, \alpha, \pi, \beta) \, d\boldsymbol{\theta} \, d\boldsymbol{\rho} \, d\boldsymbol{\tau}.
\end{aligned}
$$

  is actually available in closed form.

- Instead of doing a trans-dimensional search like reversible jump algorithm, why not search over the discrete space defined by $(G, \mathbf{Z}, \boldsymbol{\nu})$?

Marginalization approach

Doing the algebra gives

$$
\begin{aligned}
& p(G, \mathbf{Z}, \boldsymbol{\nu} | \mathbf{X}, \alpha, \pi, \beta) \\
\propto \quad & p(G) p(\boldsymbol{\nu} | \pi) \frac{\Gamma(G\alpha)}{\Gamma(\alpha)^G} \frac{\prod_{g=1}^{G} \Gamma(N_g + \alpha)}{\Gamma(N + G\alpha)} \\
\times \quad & \prod_{m \in \nu_{\mathrm{n}}} \frac{\Gamma(C_m \beta)}{\Gamma(\beta)^{C_m}} \frac{\prod_{c=1}^{C_m} \Gamma(N_{mc} + \beta)}{\Gamma(N + C_m \beta)} \\
\times \quad & \prod_{g=1}^{G} \prod_{m \in \nu_{\mathrm{cl}}} \frac{\Gamma(C_m \beta)}{\Gamma(\beta)^{C_m}} \frac{\prod_{c=1}^{C_m} \Gamma(N_{gmc} + \beta)}{\Gamma(N_g + C_m \beta)}
\end{aligned}
$$

N_g is the number of observations clustered to group g, N_{mc} is the number of times variable m takes category c, N_{gmc} is the number of items in group g that have category c for variable m

MCMC sampling algorithm

- Class memberships are sampled using a Gibbs sampling step which exploits the full conditional distribution of the class label for observation n, $n = 1, \ldots, N$

- A component is added or removed with probability 0.5

  - A component k is chosen at random to "eject" a new component from

  - A draw $u \sim \mathrm{Beta}(a, a)$ is made, and each element of the ejecting component is assigned to new component with prob u

- Components are removed by putting the elements of two randomly drawn clusters into a single cluster.

MCMC sampling algorithm

To sample the clustering variables

- A variable j is chosen randomly from $\{1, \ldots, M\}$
- If $j \in \boldsymbol{\nu}_{\mathrm{n}}$ it is proposed to move it to $\boldsymbol{\nu}_{\mathrm{cl}}$.
  Alternatively, if $j \in \boldsymbol{\nu}_{\mathrm{cl}}$ propose to move it to $\boldsymbol{\nu}_{\mathrm{n}}$

Acceptance prob for inclusion in $\boldsymbol{\nu}_{\mathrm{cl}}$ is $\min(1, R)$ with

$$
\begin{aligned}
R &= \frac{p(G, \mathbf{Z}, \tilde{\boldsymbol{\nu}} | \mathbf{X}, \alpha, \pi, \beta)}{p(G, \mathbf{Z}, \boldsymbol{\nu} | \mathbf{X}, \alpha, \pi, \beta)} \\
&= \left(\frac{\Gamma\left(C_j \beta\right)}{\Gamma\left(\beta\right)^{C_j}} \right)^{G-1} \prod_{g=1}^{G} \frac{\prod_{c=1}^{C_j} \Gamma(N_{gjc} + \beta)}{\Gamma(N_g + C_m \beta)} \\
&\times \left(\frac{\prod_{c=1}^{C_j} \Gamma(N_{jc} + \beta)}{\Gamma(N + C_j \beta)} \right)^{-1} \times \left(\frac{\pi}{1 - \pi} \right).
\end{aligned}
$$

Label switching

- Because the LCA likelihood is invariant to relabelling of the components, we need to deal with the label switching problem

- The reason is that

$$p(G, \mathbf{Z}, \boldsymbol{\nu} | \mathbf{X}, \alpha, \pi, \beta) = p(G, \mathbf{Z}_{\cdot\delta}, \boldsymbol{\nu} | \mathbf{X}, \alpha, \pi, \beta)$$

  where $\mathbf{Z}_{\cdot\delta}$ denotes the indicator matrix obtained by applying any permutation δ of $1, \ldots, G$ to the columns in $\mathbf{Z}$

- Need to post-process the samples of labels to undo any label switching that may have occurred; this has to be done to get the posterior probability of cluster membership

Post-hoc parameter estimation

- Use the conditional expection and variance formulae

$$\begin{aligned}\mathbb{E}[A] &= \mathbb{E}[\mathbb{E}[A|B]] \\ \mathbb{V}\mathrm{ar}[A] &= \mathbb{E}[\mathbb{V}\mathrm{ar}[A|B]] + \mathbb{V}\mathrm{ar}[\mathbb{E}[A|B]],\end{aligned}$$

- Let $N_g^{(t)} := \sum_{n=1}^N Z_{ng}^{(t)}, \qquad S_{gmc}^{(t)} := \sum_{n=1}^N Z_{ng}^{(t)} \mathrm{I}(X_{nm} = c).$

- Then we can estimate the expected values

$$\begin{aligned}\mathbb{E}[\theta_{gmc}|\mathbf{X}, \beta] &= \mathbb{E}[\mathbb{E}[\theta_{gmc}|\mathbf{X}, \mathbf{Z}, \beta]] \\ &\approx \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\theta_{gmc}|\mathbf{X}, \mathbf{Z}^{(t)}, \beta],\end{aligned}$$

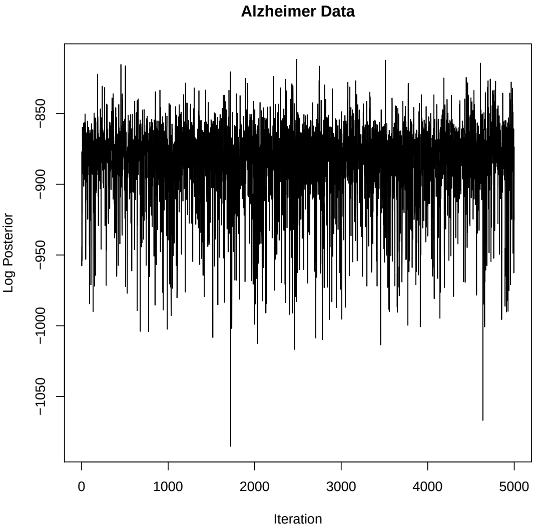
and similarly for the variance.

Post-hoc parameter estimation

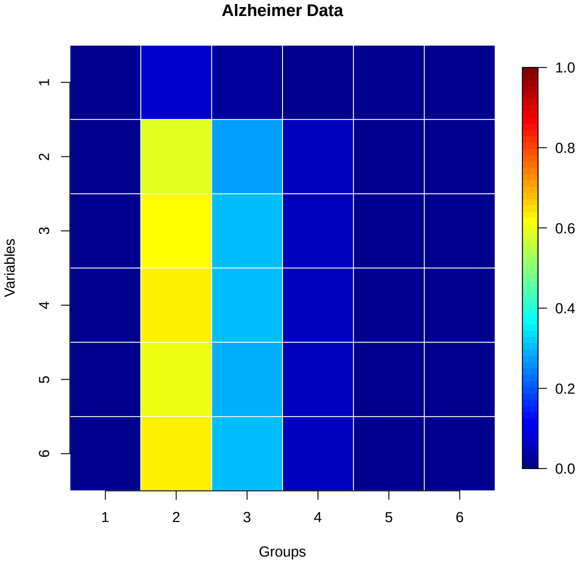
- This leads to nice formulae to estimate the posterior mean and variance of the item probabilities
- $$\mathbb{E}[\theta_{gmc}|\mathbf{X},\beta] \approx \frac{1}{T}\sum_{t=1}^{T}\frac{S_{gmc}^{(t)}+\beta}{N_g^{(t)}+C_m\beta}.$$
- $$\begin{aligned}&\mathbb{V}\mathrm{ar}[\theta_{gmc}|\mathbf{X},\beta]\\ \approx\quad & \frac{1}{T}\sum_{t=1}^{T}\frac{(S_{gmc}^{(t)}+\beta)(N_g^{(t)}+(C_m-1)\beta-S_{gmc}^{(t)})}{(N_g^{(t)}+C_m\beta)^2(N_g^{(t)}+C_m\beta+1)}\\ +\quad & \frac{1}{T}\sum_{t=1}^{T}\left(\frac{S_{gmc}^{(t)}+\beta}{N_g^{(t)}+C_m\beta}-\frac{1}{T}\sum_{t=1}^{T}\frac{S_{gmc}^{(t)}+\beta}{N_g^{(t)}+C_m\beta}\right)^2,\end{aligned}$$
- Similar formulae are available for τ_g.

Alzheimer data

The sampler was then run for 100,000 iterations, and thinned by subsampling every twentieth iterate.



Alzheimer data



Alzheimer data

The posterior probability for the number of syndromes in early onset Alzheimers where

$$p_j = \text{ Estimated posterior probability of } G \text{ classes}$$

Setting for π	p_2	p_3	p_4	p_5	p_6
$\pi = 0.5$	0.6284	0.2996	0.0622	0.0096	0.0002
$\pi \sim$Beta(1,1.5)	0.6600	0.2724	0.0584	0.0092	0

Alzheimer data

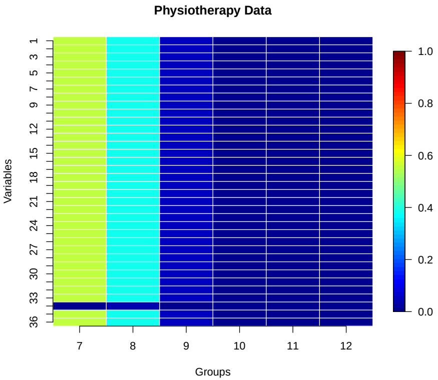
Collapsed Gibbs sampler post-hoc estimates

	Hallucination	Activity	Aggression	Agitation	Diurnal	Affective
Group 1	0.08 (0.03)	0.54 (0.06)	0.10 (0.04)	0.14 (0.06)	0.13 (0.05)	0.59 (0.08)
Group 2	0.10 (0.04)	0.80 (0.06)	0.40 (0.08)	0.64 (0.12)	0.39 (0.07)	0.94 (0.04)

Full model Gibbs sampler estimates

	Hallucination	Activity	Aggression	Agitation	Diurnal	Affective
Group 1	0.08 (0.03)	0.54 (0.06)	0.11 (0.05)	0.14 (0.06)	0.14 (0.05)	0.59 (0.08)
Group 2	0.10 (0.04)	0.79 (0.07)	0.39 (0.08)	0.64 (0.12)	0.38 (0.07)	0.93 (0.07)

Back Pain Data



p_7	p_8	p_9	p_{10}	p_{11}	p_{12}
0.5577	0.3879	0.0491	0.0046	0.0006	0.0001

Back Pain Data / Clinical Taxonomy

- The clustering closely follows the clinical taxonomy, but where the groups are subdivided into subtypes.

	CN	N	PN
Group 1	3	0	1
Group 5	52	0	0
Group 7	30	3	0
Group 3	6	96	1
Group 6	0	120	1
Group 2	1	16	79
Group 4	3	0	13

Local Independence (A Problem?)

- When analyzing the back pain data, we achieved very little data reduction.
- In fact, only one variable was labeled as non-clustering.
- An explanation for this is the *local independence* assumption in the model.
- Suppose we have two variables that are highly dependent and both exhibit clustering.
- The variable selection method will include both variables in the model, even if one variable contains no *extra* clustering information.

Dean & Raftery's Greedy Search

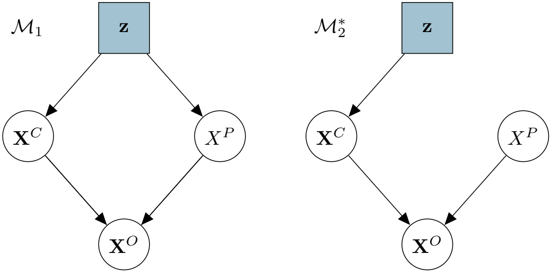
- Dean & Raftery (2010) proposed a greedy stepwise variable selection algorithm for LCA.
- The observation vector $\mathbf{X}_n$ is partitioned as

$$\mathbf{X}_n = (\mathbf{X}_n^C, \mathbf{X}_n^P, \mathbf{X}_n^O)$$

  where

  - $\mathbf{X}_n^C$ are the current clustering variables.
  - $\mathbf{X}_n^P$ is proposed to be added to the clustering variables.
  - $\mathbf{X}_n^O$ are the other variables.

- Two competing models are compared:

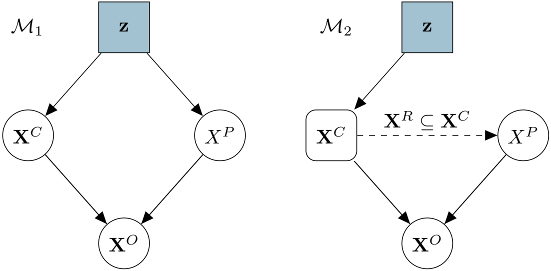


  - $\mathcal{M}_1$ assumes that the proposed variable has clustering structure.
  - $\mathcal{M}_2^*$ assumes that the proposed variable has no clustering structure.

- This framework reduces the independence assumption of the previously described approach.

Novel Extension: Relaxing Independence Further

- It is unrealistic to assume that $\mathbf{X}_n^C$ and $\mathbf{X}_n^P$ are conditionally independent.
- We propose replacing $\mathcal{M}_2^*$ with a different model.



  - $\mathcal{M}_1$ assumes that the proposed variable has clustering structure.
  - $\mathcal{M}_2$ assumes that the proposed variable has no clustering structure beyond that explained by the clustering variables.

Stepwise Search Algorithm

- We propose a stepwise search algorithm to find an *optimal* set of variables for clustering.
- The algorithm involves the following steps:
  - **Add:** Add a variable to the current clustering variables.
  - **Remove:** Remove a variable from the current clustering variables.
  - **Swap:** Swap a proposed variable with one already in the clustering variables.
- Model selection is implemented using BIC.

Back Pain Data

- The proposed model was applied to the back pain data:

Variables	*N. latent classes*	*BIC*	*ARI*
All	5	-12582.62	0.50
All	3^*	-12763.81	0.82
35 Criteria	5	-12116.32	0.50
35 Criteria	3^*	-12305.67	0.80
11 Criteria	3	-3965.24	0.75

- The new model achieves much greater data reduction.

Algorithm Run

Iter.	*Proposal*	*BIC diff.*	*Decision*	*Proposal*	*BIC diff.*	*Decision*
1	Remove Crit.5	-122.2	Accepted			
2	Remove Crit.23	-126.3	Accepted	Swap Crit.22 with Crit.5	-73.2	Rejected
3	Remove Crit.38	-109.0	Accepted	Swap Crit.25 with Crit.5	-81.5	Rejected
4	Remove Crit.4	-103.5	Accepted	Swap Crit.2 with Crit.38	-98.6	Rejected
5	Remove Crit.1	-78.3	Accepted	Swap Crit.29 with Crit.4	-23.1	Rejected
6	Remove Crit.29	-73.2	Accepted	Swap Crit.12 with Crit.1	2.7	Accepted
7	Remove Crit.1	-73.5	Accepted	Swap Crit.26 with Crit.29	3.2	Accepted
8	Remove Crit.29	-66.8	Accepted	Swap Crit.18 with Crit.12	-10.2	Rejected
9	Remove Crit.35	-63.0	Accepted	Swap Crit.7 with Crit.29	-9.0	Rejected
10	Remove Crit.7	-59.6	Accepted	Swap Crit.11 with Crit.35	-7.6	Rejected
11	Remove Crit.10	-62.9	Accepted	Swap Crit.8 with Crit.7	-76.1	Rejected
12	Remove Crit.11	-50.4	Accepted	Swap Crit.16 with Crit.10	6.8	Accepted
13	Remove Crit.8	-54.5	Accepted	Swap Crit.10 with Crit.16	-32.0	Rejected
14	Remove Crit.3	-44.2	Accepted	Swap Crit.31 with Crit.16	-9.5	Rejected
15	Remove Crit.31	-33.2	Accepted	Swap Crit.18 with Crit.16	-22.7	Rejected
16	Remove Crit.22	-30.9	Accepted	Swap Crit.24 with Crit.23	-1.7	Rejected
17	Remove Crit.14	-22.7	Accepted	Swap Crit.32 with Crit.31	-5.0	Rejected
18	Remove Crit.32	-19.2	Accepted	Swap Crit.37 with Crit.14	-8.0	Rejected
19	Remove Crit.10	-35.4	Accepted	Swap Crit.9 with Crit.3	-1.3	Rejected
20	Remove Crit.24	-17.6	Accepted	Swap Crit.30 with Crit.8	15.7	Accepted
21	Remove Crit.34	-15.7	Accepted	Swap Crit.37 with Crit.1	-0.7	Rejected
22	Remove Crit.25	-13.7	Accepted	Swap Crit.36 with Crit.1	3.3	Accepted
23	Remove Crit.18	-10.5	Accepted	Swap Crit.1 with Crit.31	8.5	Accepted
24	Remove Crit.27	-13.7	Accepted	Swap Crit.6 with Crit.26	6.1	Accepted
25	Remove Crit.31	-1.3	Accepted	Swap Crit.20 with Crit.6	5.6	Accepted
26	Remove Crit.37	1.4	Rejected	Swap Crit.6 with Crit.5	-3.1	Accepted
27	Remove Crit.5	0.4	Rejected	Swap Crit.37 with Crit.20	4.0	Rejected

Clustering / Clinical Taxonomy

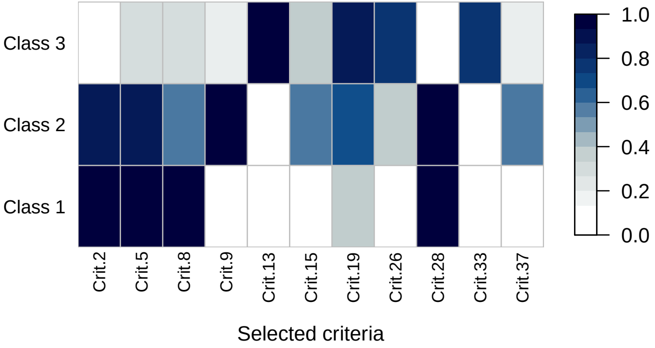
- The clustering closely follows the clinical taxonomy.

	Class 1	*Class 2*	*Class 3*
Nociceptive	210	21	4
Peripheral Neuropathic	5	88	2
Central Sensitiization	3	3	89

- It is not unusual for patients diagnosed as Nociceptive may have Peripheral Neuropathic aspects to their back pain.

Clustering Variables

- The selected variables exhibit strong clustering across the three groups.



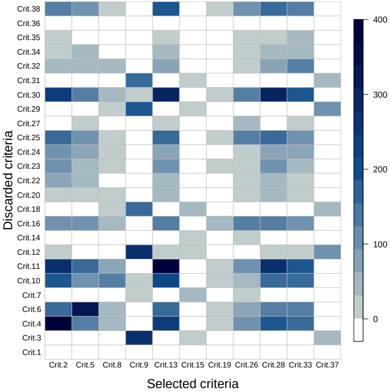
Chosen Variables with Descriptions

The chosen variables have the following descriptions.

Crit.	*Description*	*Class 1*	*Class 2*	*Class 3*
2	Pain associated to trauma, pathologic process or dysfunction	0.94	0.90	0.04
5	Usually intermittent and sharp with movement/mechanical provocation	0.94	0.84	0.24
8	Pain localized to the area of injury/dysfunction	0.97	0.50	0.31
9	Pain referred in a dermatomal or cutaneous distribution	0.06	1.00	0.11
13	Disproportionate, nonmechanical, unpredictable pattern of pain	0.01	0.00	0.91
15	Pain in association with other dysesthesias	0.03	0.51	0.34
19	Night pain/disturbed sleep	0.34	0.70	0.86
26	Pain in association with high levels of functional disability	0.07	0.36	0.79
28	Clear, consistent and proportionate pattern of pain	0.97	0.94	0.07
33	Diffuse/nonanatomic areas of pain/tenderness on palpation	0.03	0.01	0.73
37	Pain/symptom provocation on palpation of relevant neural tissues	0.07	0.57	0.19

Discarded Variables

- Many of the discarded variables are related with the clustering variables.

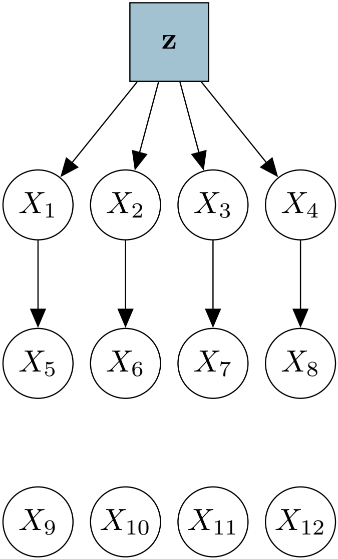


- These are not clustering variables because they don't exhibit clustering *beyond* what can be explained by the clustering variables.

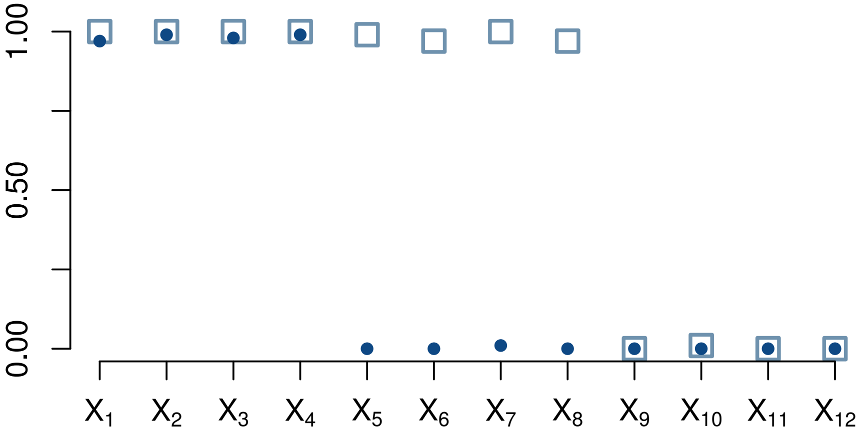
Summary

- Model-based approaches to clustering and variable selection achieve excellent performance.
- The collapsed MCMC scheme explores the model space effectively.
- Removing independence assumptions in the model achieves improved variable selection.
  Care needed interpreting the chosen/discarded variables.

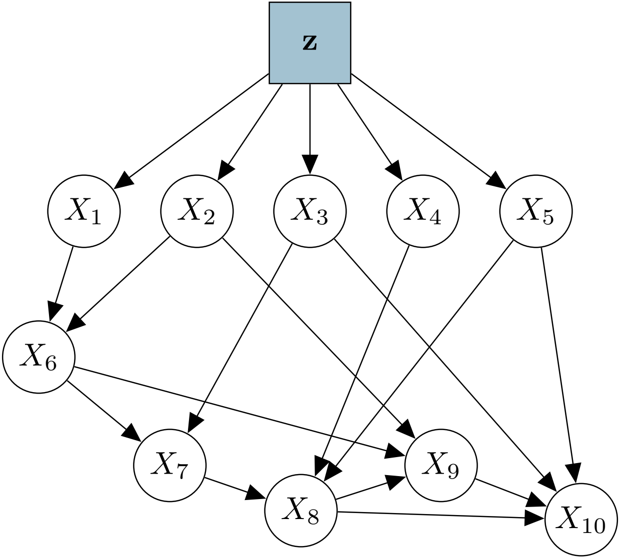
Simulation 1



Simulation 1 Results



Simulation 2



Simulation 2 Results