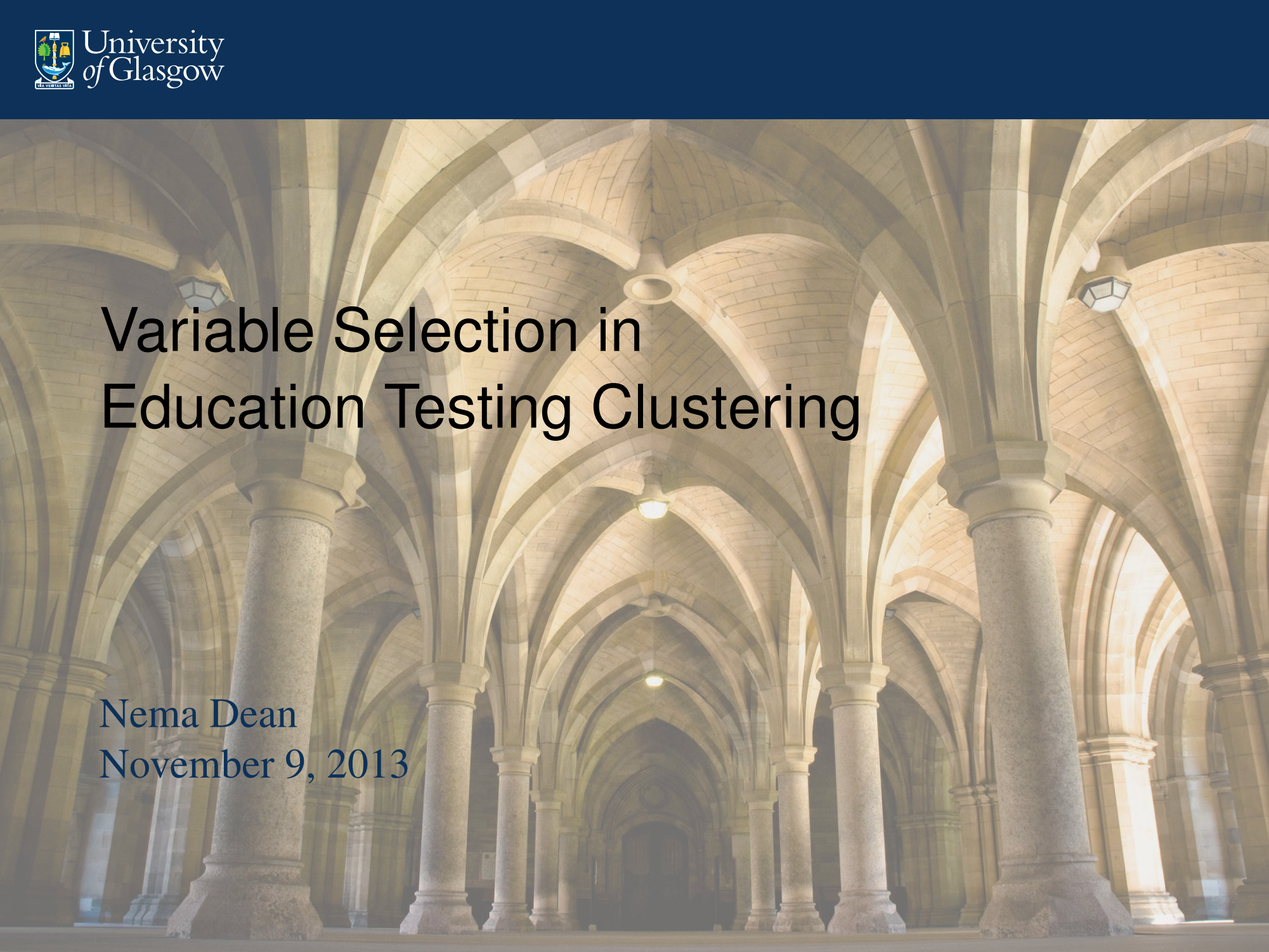


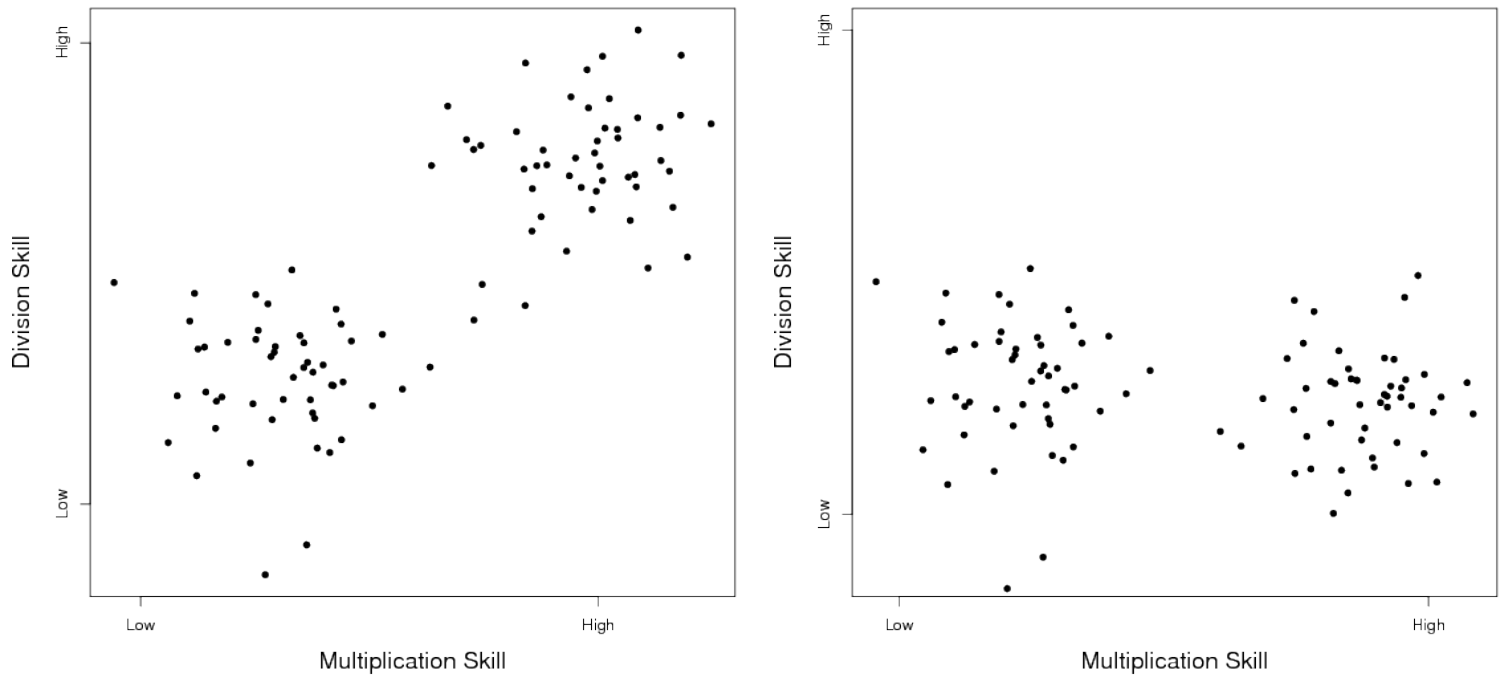
Variable Selection in Education Testing Clustering

Nema Dean

November 9, 2013

Describing/Characterizing Students' Abilities

Common educational research objective:
To identify students' *current* stage of skill mastery



- Help identify weaknesses and strengths
- How many different subgroups of students do we have?
- Why are they different?

How can we even measure skill mastery?

Cognitive Diagnosis Modeling Goals:

- For each student, estimate mastery for each skill
- Assign each student a *latent* skill set profile

- Student response matrix (Y)

$$Y = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,J} \\ \vdots & \ddots & & \vdots \\ y_{I,1} & y_{I,2} & \cdots & y_{I,J} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 1 \\ \vdots & \ddots & & \vdots \\ NA & 1 & \cdots & 0 \end{bmatrix}$$

I students, J items/questions

$Y_{ij} = 1$ if student i answered item j correctly; 0 if incorrectly; NA if not answered

- Assignment matrix of skills needed for each item (Q)

$$Q = \begin{bmatrix} q_{1,1} & q_{1,2} & \dots & q_{1,K} \\ \vdots & \ddots & & \vdots \\ q_{J,1} & q_{J,2} & \dots & q_{J,K} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & 1 & \dots & 1 \end{bmatrix}$$

J items, K skills

$Q_{jk} = 1$ if item j requires skill k; 0 if not.

Q-matrix is usually expert-elicited (i.e. the classroom teacher)

Q-matrix can be balanced or unbalanced.

- If balanced, same number of items for each skill (or combination of skills).
- If unbalanced, can have uneven numbers or not all combinations.

Cognitive Diagnosis Modelling

- Models for estimating latent skill set profiles
  - Cognitive Diagnosis Models
- Do not scale well for large number of students/items/skills
- Often assume skill mastery to be binary rather than interval

Introduction of cognitive or intelligent tutoring systems in the classroom has resulted in an explosion of data.

Now able to easily collect data on student performance for as many students and as many skills as wanted.

Data storage is relatively minimal/cheap.
Our bottleneck is the methodology.

Simpler Estimate Using Clustering

Current methods can't estimate large numbers of skill masteries, need another method.

One option is the Capability Matrix *(Nugent, Ayers, Dean)*

$$B_{ik} = \frac{\sum_{j=1}^{J} I_{Y_{ij} \neq NA} \cdot Y_{ij} \cdot q_{jk}}{\sum_{j=1}^{J} I_{Y_{ij} \neq NA} \cdot q_{jk}}$$

B_{ik}: % of items student i answered correctly for skill k.

B_{ik} scales for the number of items seen; reduces influence of over-represented skills; incorporates missingness

$$B_{ik} \in [0, 1]$$

Maps students into a unit hyper-cube (like CDM estimates).

Could also use sum-scores (Chiu, Douglas): $W_{ik} = \sum_{j=1}^{J} Y_{ij} q_{jk}$

Clustering in Cognitive Diagnosis

Once we have the estimates, how do we assign students to profiles?

Ideally:

- the estimates are close to a profile/corner of the hypercube.
- often the case when CDM estimation methods are used

B_{ik} usually not as close to corners.

Instead we cluster the B_{ik} to find groups of similar students; *(these clusters can be assigned to a corner if desired.)*

Clustering can more easily handle high numbers of skills

Can answer:

- How do the students group together?
- How could we aggregate the groups if need be?
- Which skills or combination of skills separate students?

Clustering Approaches

Two common approaches in CDM:

- Hierarchical Linkage: dendogram cut to form 2^K clusters
- K-means: set the number of clusters to be 2^K

Past work *(Nugent, Ayers, Dean)* showed that the quick estimates combined with clustering perform very similarly to more traditional estimation procedures *in a fraction of the time* which allows for real-time feedback in the classroom.

Model-Based Clustering

Suggested standard clustering approaches problematic:

- Need to decide on G (number of profiles/clusters)
- For K-means, results are dependent on starting values
- Spherical shapes appropriate for unit hypercube?

Model-Based Clustering: more flexible choice

- Model our overall population by a finite mixture model with mixing weights π_g $(i.e. \sum_g \pi_g = 1)$:

$$p(x) = \sum_{g=1}^{G} \pi_g p_g(x; \theta_g)$$

- Often $p_g(x; \theta_g)$ is assumed to be Gaussian: $\theta_g = (\mu_g, \Sigma_g)$
- Number of groups, G, chosen by BIC

Issues with standard MBC on B matrix

- B_{ik} ignores information about certainty of estimation of skill mastery
- MBC ignores restricted space
- With coarsely-gridded skills, MBC solution will overfit

Possible Solution

- Use Sum-scores instead: $W_{ik} = \sum_{j=1}^{J} I_{Y_{ij} \neq NA} \cdot Y_{ij} \cdot q_{jk}$
- Along with $N_{ik} = \sum_{j=1}^{J} I_{Y_{ij} \neq NA} \cdot q_{jk}$
- Model with finite mixture of conditionally independent binomials

A Better Mixture Model

Possible solution: Use a mixture of Binomial distributions

- Each p_g = product of different binomials (one per skill variable)

  i.e. for $W_i = (W_{i1}, \ldots, W_{iK}), N_i = (N_{i1}, \ldots, N_{iK})$

$$p(W_i, N_i) = \sum_{g=1}^{G} \pi_g \left(\prod_{k=1}^{K} p_{kg}(W_{ik}, N_{ik}; \theta_{kg}) \right)$$

where $p_{kg}(W_{ik}, N_{ik}; \theta_{kg})) = \binom{N_{ik}}{W_{ik}} \theta_{kg}^{W_{ik}} (1 - \theta_{kg})^{N_{ik} - W_{ik}}$

- Still has the advantages of a mixture model approach
- Estimated parameters fall in the unit hypercube

Finite Mixture of Binomials (FMB) Estimation

Standard EM estimation

- Define $z_i = \{z_{i1}, \ldots, z_{iG}\}$ as missing class membership labels
- Alternate M-step with E-step until convergence
- M-step:
  - $\hat{\pi}_g = \frac{\sum_{i=1}^{I} z_{ig}}{I}$
  - $\hat{\theta}_{kg} = \frac{\sum_{i=1}^{I} z_{ig} \cdot W_{ik}}{\sum_{i=1}^{I} z_{ig} \cdot N_{ik}}$
- E-step: $\hat{z}_{ig} = \frac{\pi_g p_g(W_i, N_i; \hat{\theta}_g)}{p(W_i, N_i; \hat{\theta})}$

Start algorithm with class estimates from k-means applied to (transformed) B matrix

Use BIC to select G

Possible issues with FMB

- Assuming conditional independence of variables given cluster membership (similar to latent class analysis), not currently accounting for skill correlation
- Ignoring correlation from questions with multiple skills
- What to do about skills some students do not see?

All skills the same?

- All skills treated as if they were clustering variables
- Need to perform variable selection to identify skills that separate groups of students

Variable Selection for FMB

Take the approach of Raftery and Dean (2006), Dean and Raftery (2010)

Variable Selection Procedure

Stepwise procedure, iterate inclusion exclusion steps:

- Inclusion step: check each variable not currently selected for evidence of its inclusion improving the current set of clustering variables
- Exclusion step: check each variable currently selected for evidence of its exclusion improving the current set of clustering variables

Inclusion step

Current set of clustering variables indexed by $S = \{s_1, \ldots\}$
Current set of not selected variables indexed by $NS = \{ns_1, \ldots\}$

For each index ns_i in NS:

- Calculate the BIC for the best FMB model ($G \geq 2$) with variables S and ns_i - $\text{BIC}_{clust}(S, ns_i)$
- Calculate the BIC for the best FMB model ($G \geq 2$) just with variables S - $\text{BIC}_{clust}(S)$
- Calculate the BIC for a single group FMB with variable ns_i - $\text{BIC}_{no.clust}(ns_i)$

Inclusion step

Evidence for include in clustering variable ns_i:

$$\text{BIC}_{clust}(S, ns_i) - (\text{BIC}_{clust}(S) + \text{BIC}_{no.clust}(ns_i))$$

- Propose variable in NS with largest evidence for inclusion
- If evidence for variable is greater than required level, update S to include variable (and remove from NS)
- If evidence for variable is not greater than required level, sets S and NS remain unchanged

BIC = 2·max. log likelihood - $\log(n) \cdot \nu$
$\Rightarrow$ larger BIC indicates better fit

Exclusion step

Current set of clustering variables indexed by $S = \{s_1, \ldots\}$
Current set of not selected variables indexed by $NS = \{ns_1, \ldots\}$

For each index s_i in S:

- Calculate the BIC for the best FMB model ($G \geq 2$) with variables S not s_i - $\text{BIC}_{clust}(S, -s_i)$
- Calculate the BIC for the best FMB model ($G \geq 2$) just with variables S - $\text{BIC}_{clust}(S)$
- Calculate the BIC for a single group FMB with variable s_i - $\text{BIC}_{no.clust}(s_i)$

Exclusion step

Evidence for include in clustering variable ns_i:

$$\text{BIC}_{clust}(S) - (\text{BIC}_{clust}(S, -s_i) + \text{BIC}_{no.clust}(s_i))$$

- Propose variable in S with smallest evidence, for exclusion
- If evidence for variable is less than required level, update S to exclude variable (and include in NS)
- If evidence for variable is not less than required level, sets S and NS remain unchanged

Simulated Data: Example

- 6 variables: 2 clustering, 4 noise
- 2 groups of equal size
- Randomly generated number of trials for each student and skill
- Success probabilities:

	Var.1	Var.2	Var.3	Var.4	Var.5	Var.6
Group 1	0.9	0.1	0.5	0.2	0.9	0.4
Group 2	0.3	0.8	0.5	0.2	0.9	0.4

Simulated Data: Example Results

Step Type	Variable Proposed	Evidence for Clustering	Step Accepted?
Inclusion	2	3137	Y
Inclusion	1	1388	Y
Inclusion	6	-3	N
Exclusion	1	1388	N

BIC selects 2 clusters and ARI (with truth) is 1

Simulations

2 groups, 2 clustering variables, 4 noise variables
Randomly generated differences between clustering variables
success probabilities (0.1 - 0.5)

	ARI for Sel. Variables	ARI for All Variables	No. of clusters chosen with Sel. Variables	No. of clusters ch with All Variabl
Equal Mix. Prop. (0.5, 0.2)	0.942($\pm$0.084)	0.942 ($\pm$ 0.084)	2 ($\pm$ 0)	2 ($\pm$ 0)
Unequal Mix Prop. (0.8, 0.2)	0.927 ($\pm$ 0.115)	0.923 ($\pm$ 0.122)	2 ($\pm$ 0)	2.05 ($\pm$ 0.22)

The Assistment Project:

- Web-based tutoring program developed by Carnegie Mellon University, Carnegie Learning, and Worcester Polytechnic Institute
- Blends tutoring "assistance" with "assessment" reporting
- Over 4000 students in Massachusetts and Pennsylvania utilized the system in 2007-2008
- System currently tracks/reports on about 120 skills per grade level

Goals:

- Help prepare students for end-of-year exams, e.g. MCAS
- Help teachers identify weaknesses/strengths in their students and in their curriculum
- Allow teachers to use their time more effectively
- Help researchers discover how students learn

Assistment Data

- 13 skills: Addition, Discount, Equation Solving, Interpreting Linear Equations, Interpreting Numberline, Multiplication, Multiplying Decimals, Order of Operations, Percent of, Probability, Reading graph, Substitution, Symbolization Articulation
- 128 students
- Variable selection procedure chooses all variables
- BIC selects 3 clusters (top, middling, poor students)

Issues/Future Work

- Independence of non-clustering variables
  - strong assumption
  Could mimic Maugis, Celeux and Martin-Magniette
  - more general variable role modelling
- Conditional independence of binomials
  - restrictive assumption (but allows for easy of modelling with missing data)
- Clustering on Subsets of Attributes approach

References

- A. E. Raftery and N. Dean (2006), “Variable Selection for Model-Based Clustering”, Journal of the American Statistical Association, 101 (473), 168-178.
- E. Ayers, R. Nugent and N. Dean (2008), ”Skill Set Profile Clustering Based on Student Capability Vectors Computed from Online Tutoring Data”, Educational Data Mining 2008: 1st International Conference on Educational Data Mining, Proceedings
- C. Maugis, G. Celeux and M.-L. Martin-Magniette, (2009), “Variable selection in model-based clustering: A general variable role modeling”, Computational Statistics and Data Analysis, 53, 3872-3882
- N. Dean and A. E. Raftery (2010), “Latent Class Analysis Variable Selection”, Annals of the Institute of Statistical Mathematics, 62 (1), 11-35