MUDIM – A software system for MUltidimensional DIstribution Model handling

Vladislav Bína
Fac. of Management, Jindřichův Hradec
University of Economics, Prague
bina@fm.vse.cz

Radim Jiroušek
Inst. of Information Th. and Automation
Academy of Sciences of the Czech Rep.
radim@utia.cas.cz

Václav Kratochvíl
Inst. of Information Th. and Automation
Academy of Sciences of the Czech Rep.
vaclav@kratochvil.biz

Data – Algorithms – Decision Making, December 10 – 12, 2006, Třešť

Compositional models

- an effective representation of multidimensional distributions
- an alternative to graphical models (e.g. Bayesian networks)
- perfect sequence models and Bayesian networks are equivalent – there exist transformation algorithms
- multidimensional distribution is directly composed from a system of low-dimensional distributions

Basic notions

Distribution $\pi(x_K)$ – $|K|$-dimensional table of numbers from $[0;1]$.

Marginal distributions ($L \subset K$, $M = K \setminus L$): $\pi(x_L)$, $\pi^{\downarrow\{L\}}$, π^{-M}

Definition 1 For any two distributions $\pi(x_K)$, $\kappa(x_L)$ their composition is

$$\pi(x_K) \triangleright \kappa(x_L) = \begin{cases} \frac{\pi(x_K)\kappa(x_L)}{\kappa(x_{K\cap L})} & \text{when} \quad \pi(x_{K\cap L}) \ll \kappa(x_{K\cap L}), \\ \text{undefined} & \text{otherwise,} \end{cases}$$

where symbol $\pi(x_M) \ll \kappa(x_M)$ denotes that $\pi(x_M)$ is dominated by

$$\kappa(x_M): \quad \forall x_M \in \mathbf{X}_M \quad (\kappa(x_M) = 0 \implies \pi(x_M) = 0).$$

Iteration of compositions:

$$\pi_1 \triangleright \pi_2 \triangleright \pi_3 = (\pi_1 \triangleright \pi_2) \triangleright \pi_3 \neq \pi_1 \triangleright (\pi_2 \triangleright \pi_3)$$

Compositional model:

$$\pi_1 \triangleright \pi_2 \triangleright \pi_3 \triangleright \ldots \triangleright \pi_{n-1} \triangleright \pi_n = (\ldots((\pi_1 \triangleright \pi_2) \triangleright \pi_3) \triangleright \ldots \triangleright \pi_{n-1}) \triangleright \pi_n.$$

Perfect sequence model:

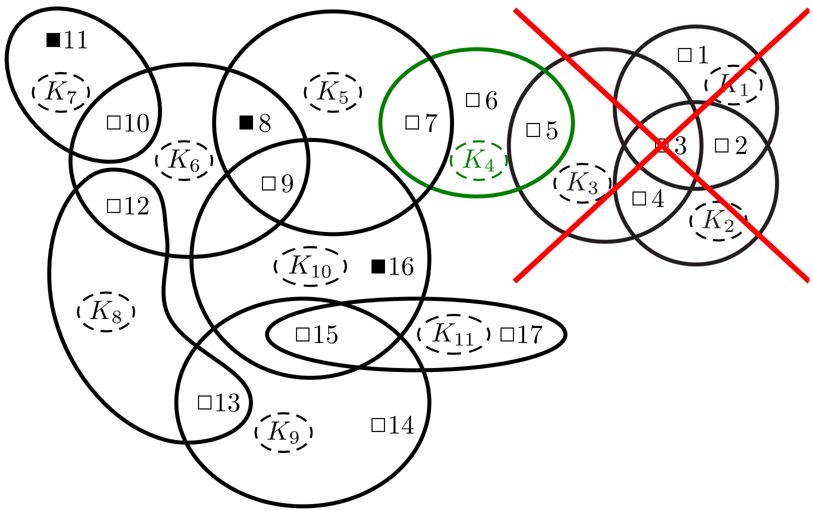
Sequence $\pi_1, \pi_2, \ldots, \pi_n$ is perfect iff $\pi_1, \ldots \pi_n$ are marginals of $\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n$.

Marginalization

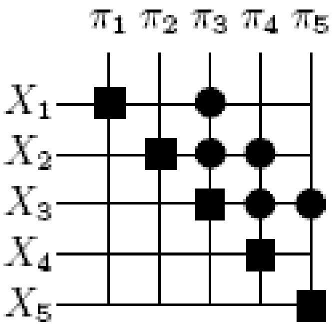
- Crucial algorithm – need of effective computation.
- Classical approach – equivalent to Bayesian networks – deletion of variable appearing only once in the sequence (deletion of childless node) and procedure for marginalization of one general variable (edge reversal and node deletion).
- Marginalizing of multiple variables in one step – developed in framework of compositional models.

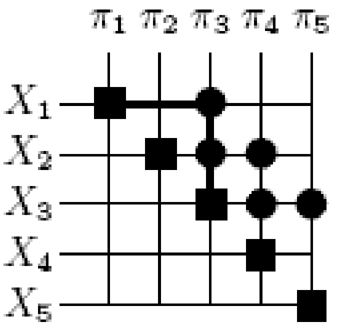
Visualization

$$(\pi_1 \triangleright \pi_2 \triangleright \pi_3 \triangleright \pi_4 \triangleright \pi_5 \triangleright \pi_6 \triangleright \pi_7 \triangleright \pi_8 \triangleright \pi_9 \triangleright \pi_{10} \triangleright \pi_{11})^{\downarrow\{8,11,16\}}$$



Persegrams





- Simple visualization of distributions and their variables.
- Independence and conditional independence relations (not implemented yet).

Piece of information theory

Shannon entropy (for $\pi \in \Pi^{(N)}$):

$$H(\pi) = -\sum_{x \in \mathbf{X}_N : \pi(x) > 0} \pi(x) \log \pi(x)$$

Informational content (for $\pi \in \Pi^{(N)}$):

$$IC(\pi) = \sum_{x \in \mathbf{X}_N : \pi(x) > 0} \pi(x) \log \frac{\pi(x)}{\prod_{j \in N} \pi(x_j)}$$

Kullback–Liebler divergence (for $\kappa, \pi \in \Pi^{(N)}$):

$$Div(\pi \parallel \kappa) = \begin{cases} \sum_{x \in \mathbf{X}_N : \pi(x) > 0} \pi(x) \log \frac{\pi(x)}{\kappa(x)} & \text{if } \pi \ll \kappa, \\ +\infty & \text{otherwise.} \end{cases}$$

Thank you for your attention.