



Information-Theoretic Scoring Rules to Learn Additive Bayesian Network Applied to Epidemiology

Gilles Kratzer¹, Reinhard Furrer^{1,2}

¹Department of Mathematics, ²Department of Computational Science; University of Zurich (Switzerland)

Contact: gilles.kratzer@math.uzh.ch

Motivation

- ABN¹ methodology extends the classical generalized linear model (GLM) framework to **multiple dependent variables**
- The key perspective of ABN is to extract the conditional independence information from an **observational dataset**
- ABN is a suitable methodology to mastermind **complex and messy data** in an exploratory analysis

Summary

- ABN is a mixture between **machine learning** and **statistical** approach
- **abn** is distributed as an R package
<https://CRAN.R-project.org/package=abn>
- Several implemented information theory scores: **AIC, BIC, MDL**
- **Bayesian** scoring function
- **Exact** and **Heuristic** search algorithm

Results

- Perform **structure discovery**
- ABN modelling empirically identifies associations in complex and high dimensional data as a **machine learning technique**

Future Work

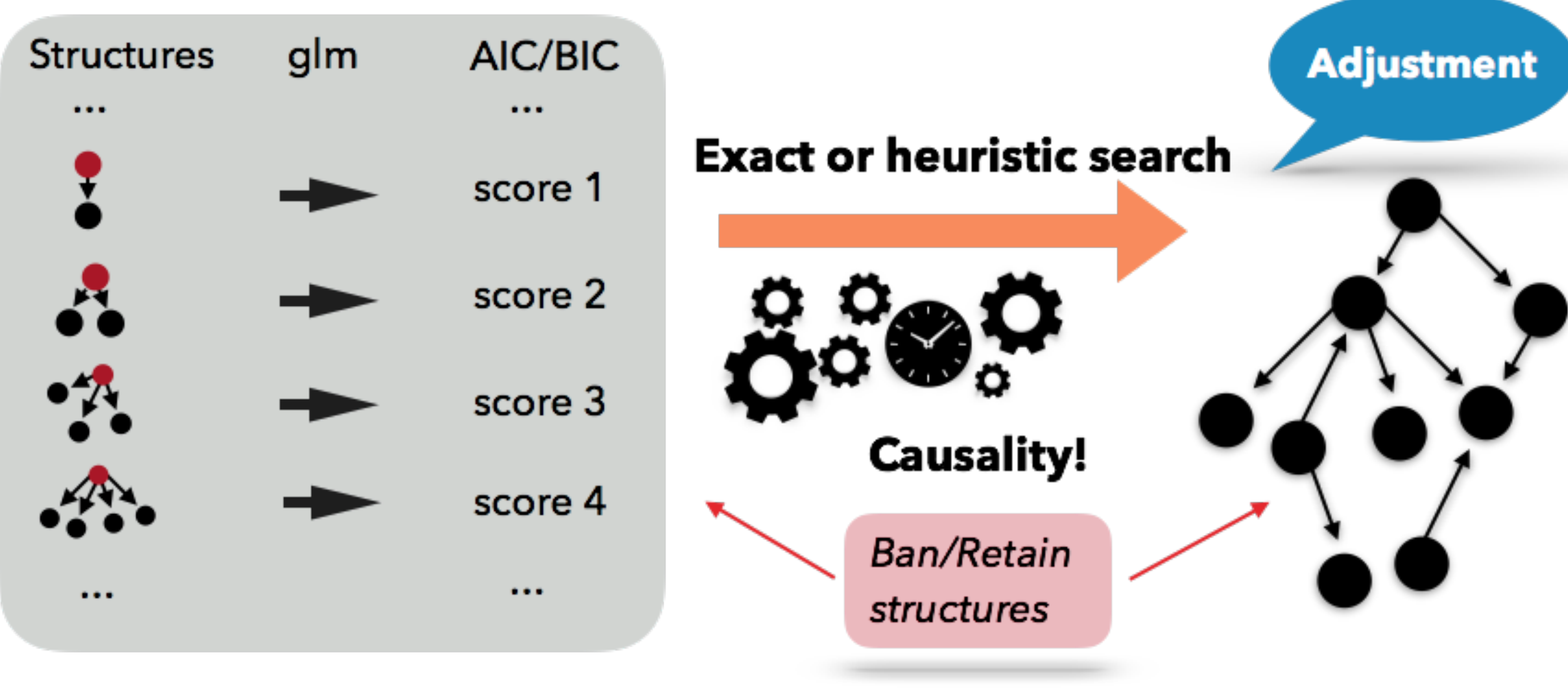
- Implementation of wider classes of **distributions**
- Implementation of **boosted information theoretic scores**

Why thinking in terms of system?



- System epidemiology = contributions at different levels
- Confounding factors
- Complex dependence structure
- Multicollinearity

How to learn Bayesian Networks from data using search and score method²?



Why Information-theoretic scores?

Data separation and sparsity
No optimal solution for Bayesian regression in the ABN context

Multinomial Distribution
Actually no implementation of multinomial Bayesian estimation with suitable prior

Effective computation (fast and reliable)
Iterative reweighting least square method is fast and highly reliable

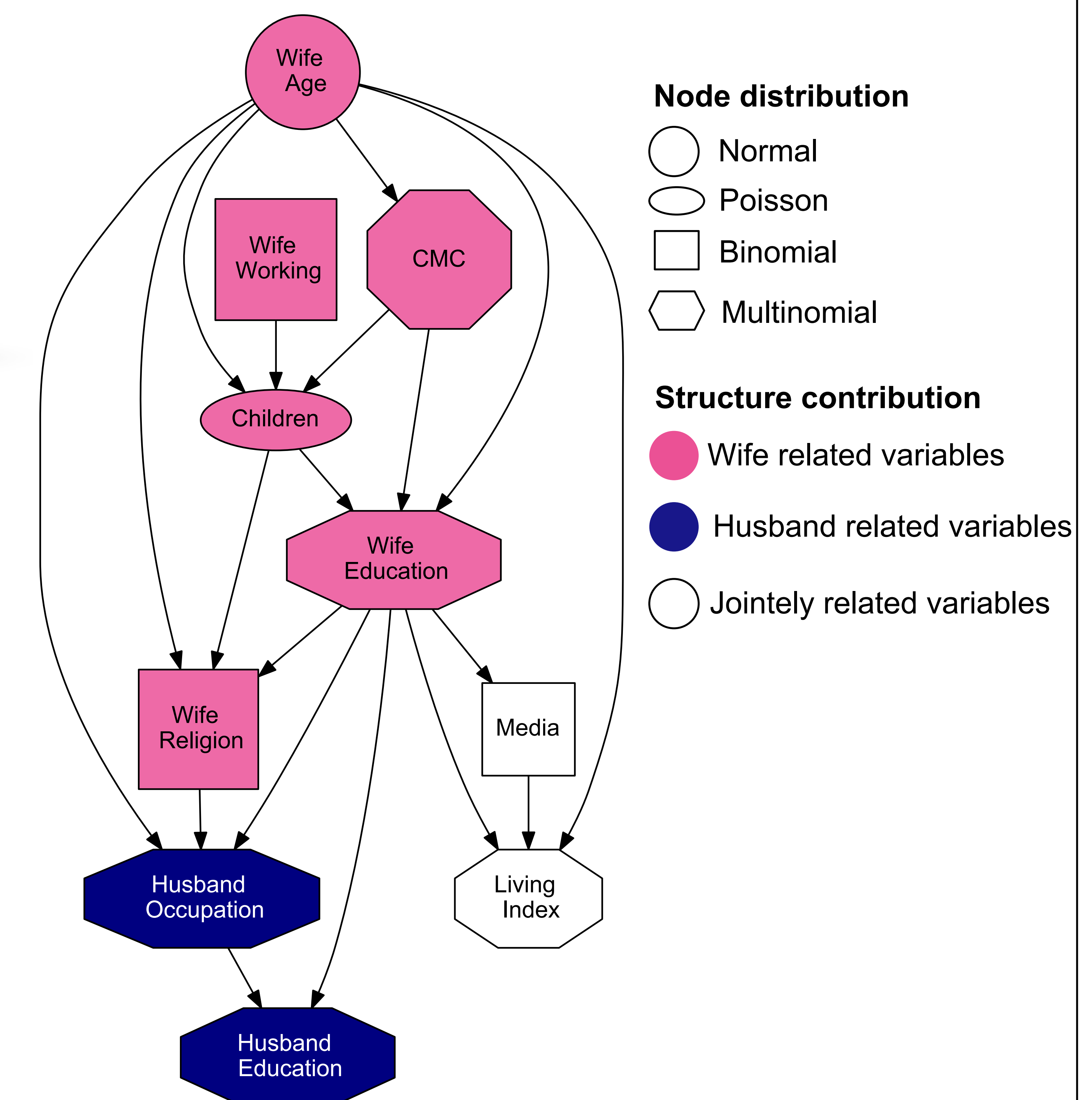
Adjustment for confounders
Possibility to compute an adjusted DAG with classical epidemiological confounders

Illustrative example³

Subset of the 1987 National Indonesia Contraceptive Prevalence Survey. Selection criteria: no pregnant and married womans.

10 variables, n = 1473 observations (no missing data)

- Wife's age (normal)
- Husband education (multinomial, 4 levels)
- Number of children (Poisson)
- Wife's religion (binomial)
- Wife is currently working (binomial)
- Husband's occupation (multinomial, 4 levels)
- Standard-of-living index (multinomial, 4 levels)
- Media exposure (binomial)
- Contraceptive method used (multinomial, 3 levels)



References

1. Lewis, F. I. et al. "Structure discovery in Bayesian networks: An analytical tool for analysing complex animal health data", Preventive Veterinary Medicine 100.2 (2011): 109-115.
2. Koivisto, M. et al. "Exact Bayesian structure discovery in Bayesian networks". Journal of Machine Learning Research, 549-573. (2004)
3. Lim, T.-S. et al. "A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms". Machine Learning (1999).

