University of Zurich[UZH]

# Additive Bayesian Network approach applied to time series and longitudinal datasets

**Gilles Kratzer[1], Reinhard Furrer[1,2]**

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

**Contact:** gilles.kratzer@math.uzh.ch

## Motivation

- ABN[1] 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 a **correlated dataset**
- A suitable methodology to mastermind **complex and messy data** in an exploratory analysis
- Extending ABN to **correlated errors** and **mixed models**

## Summary

- **tsabn** is a time series **extension of abn**
- **tsabn** is distributed as an R package https://git.math.uzh.ch/reinhard.furrer/tsabn
- Several implemented scores: **AIC**, **BIC**, **MDL**
- Errors Autocorrelation: **iterative Cochrane-Orcutt** procedure with **Autoregressive modelling**

## Results

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

## Future Work

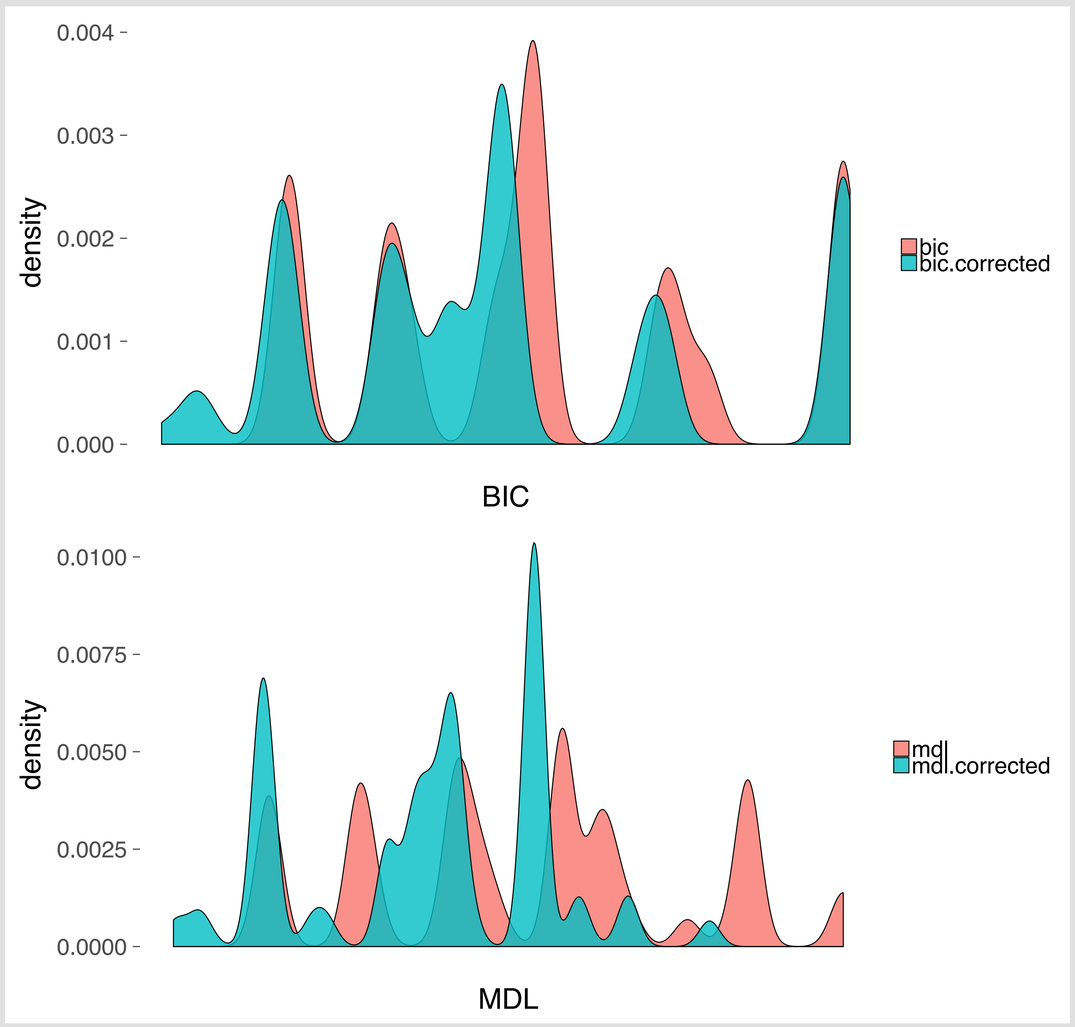
- Implementation of wider classes of **autocorrelation models**
- Implementation of **Granger causality** score for BN learning

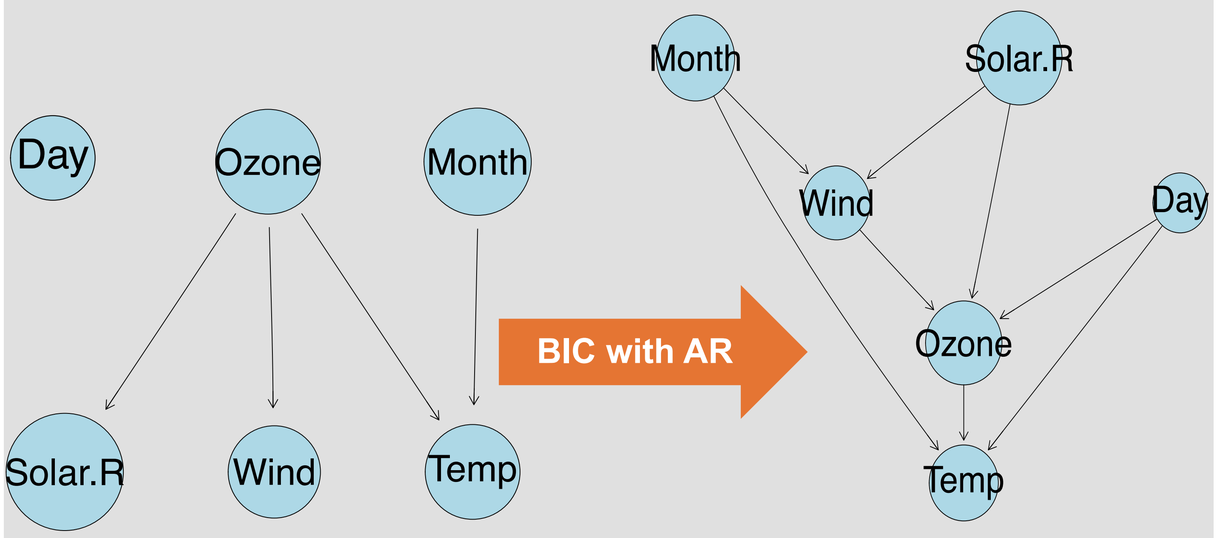
## New York air quality dataset[2]

**Daily readings of the air quality values from May to September 1973**

- 6 variables, n = 111, complete case analysis, unconstrained AR()
- **Ozone**: Mean ozone in parts per billion (Roosevelt Island)
- **Solar.R**: Solar radiation at Central Park
- **Wind**: Average wind speed at LaGuardia Airport
- **Temp**: Maximum daily temperature at LaGuardia Airport

### How the best BN changes with autocorrelation modelling?



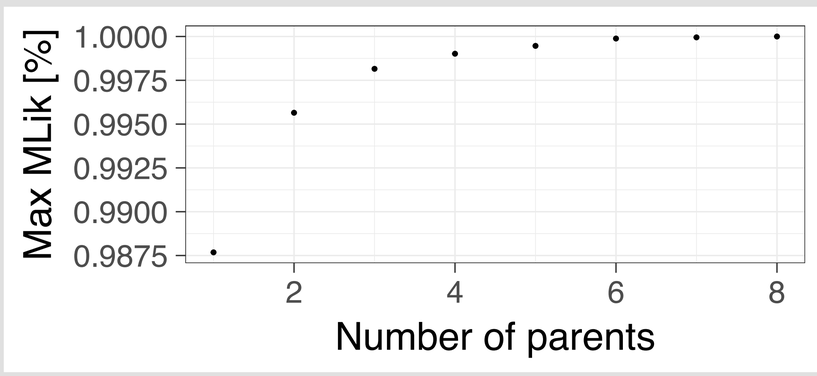



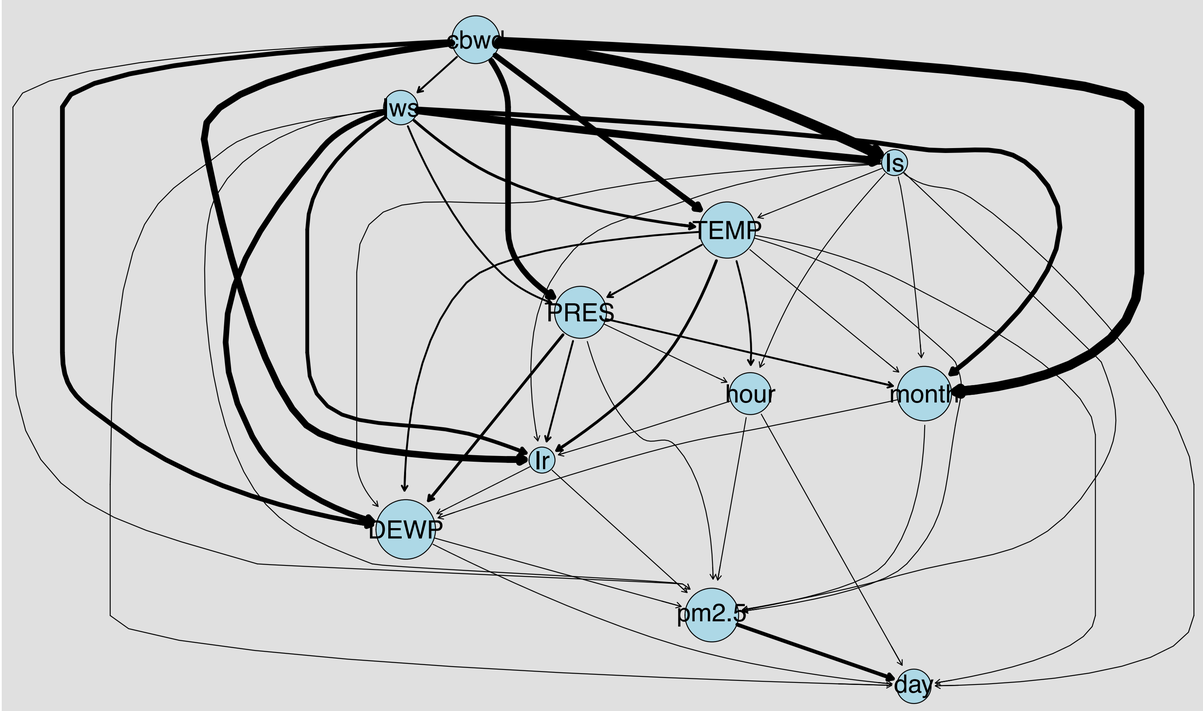

## Beijing air quality dataset[3]

**Hourly readings of the PM2.5 data of US embassy in Beijing with meteorological data from Beijing Capital International Airport from 2010 to 2014**

- 12 variables, n = 41'757, complete case analysis, unconstrained AR
- **pm2.5:** PM2.5 concentration
- **DEWP**: Dew Point
- **TEMP:** Temperature
- **PRES:** Pressure
- **cbwd:** Combined wind direction
- **Iws:** Cumulated wind speed
- **Is:** Cumulated hours of snow
- **Ir:** Cumulated hours of rain

### What is the relationship of the variables, taken into account that seasonal effect is random?




## References

1. Lewis, F. I., Brülisauer, F. and Gunn, G. J. "*Structure discovery in Bayesian networks: An analytical tool for analysing complex animal health data*", Preventive veterinary medicine 100.2 (2011): 109-115.
2. Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983) "*Graphical Methods for Data Analysis*", Belmont, CA: Wadsworth
3. Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H. and Chen, S. X. (2015). "*Assessing Beijing's PM2.5 pollution: severity, weather impact*", APEC and winter heating. Proceedings of the Royal Society A, 471, 20150257