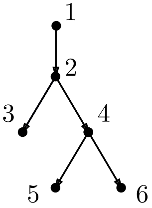
Probabilistic Modeling Using Tree Linear Cascades

Nick Landolfi & Sanjay Lall
Stanford University

Revision: June 2022

Motivation: modeling functional relations for high-dimensional observations

- example application: anomaly localization in cloud telemetry, a network of dynamical systems
- an approach: *structural equations*, model x_i as a function $x_i = f_i(x_{-i})$ of other metrics x_{-i}



  - e.g., $x_2 = f_2(x_1)$; $x_3 = f_3(x_2)$; $x_4 = f_4(x_2)$; $x_5 = f_5(x_4)$; $x_6 = f_6(x_4)$

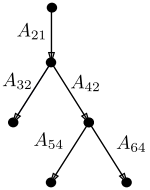
Cascade regression to simultaneously find tree and parameters

- **Problem 1**. Given random vector $x : \Omega \to \mathbf{R}^d$, find rooted tree (T, r) and $A \in \mathbf{R}^{d \times d}$ to

$$\begin{array}{rl} \text{minimize} & \mathbf{E}\, \|Ax - x\| \\ \text{subject to} & A \in \mathtt{sparse}(T, r) \end{array}$$

  - where $\mathtt{sparse}(T, r)$ has elements with sparsity pattern of directed adjacency matrix of (T, r); e.g.,

$$A = \begin{bmatrix} & & & & & \\ A_{21} & & & & & \\ & A_{32} & & & & \\ & A_{42} & & & & \\ & & & A_{54} & & \\ & & & A_{64} & & \end{bmatrix}$$



- **Solution.** find *maximum spanning tree* with edges weighted by $\mathbf{E}(x_i x_j)^2$ (Theorem 1)
  - and for selected edge $\{i, j\}$ with $j = \text{pa}_i$, choose $A^\star_{ij} = \mathbf{E}(x_i x_j)$

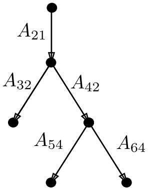
Tree linear cascades have identifiable structure

- $x : \Omega \to \mathbf{R}^d$ is a *tree linear cascade* on $e : \Omega \to \mathbf{R}^d$ with respect to $A \in \text{sparse}(T, r)$ if

$$x = Ax + e$$

where e is uncorrelated, $\mathbf{E}(e) = 0$ and $\text{sparse}(T, r)$ has sparsity like adjacency matrix of (T, r)

$$A = \begin{bmatrix} & & & & & \\ A_{21} & & & & & \\ & A_{32} & & & & \\ & A_{42} & & & & \\ & & & A_{54} & & \\ & & & A_{64} & & \end{bmatrix}$$



- **Result.** T is the *unique maximum spanning tree* with edges weighted by $\mathbf{E}(x_i x_j)^2$ (Theorem 2)
  - *cascade regression identifies the tree* of such a distribution (Corollary 1)
  - analogous to stochastic process variant studied in controls literature [Materassi and Innocenti, 2010]

Our formulation generalizes Gaussian Chow-Liu

- **Problem 2.** Given a density $g : \mathbf{R}^d \to \mathbf{R}$, find a tree T and a density $f : \mathbf{R}^d \to \mathbf{R}$ to

$$\begin{array}{ll} \text{minimize} & d_{kl}(g, f) \\ \text{subject to} & f \text{ factors according to } T \end{array}$$

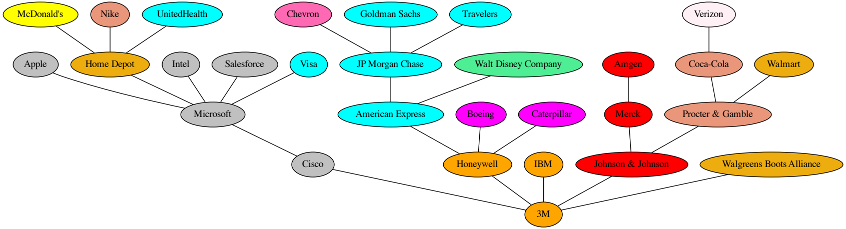
  - where d_{kl} is the Kullback-Leibler divergence
  - f factors according to T means $f = f_i \prod_{j \neq i} f_{j|\mathsf{pa}_j}$
- well-known prior solution: *maximum spanning tree* with edges weighted by *mutual informations*
  - if g is gaussian, then mutual information is $-1/2 \log(1 - \mathbf{E}(x_i x_j)^2)$
  - *monotonic transformation* of $\mathbf{E}(x_i x_j)^2$; so *trees coincide* (Corollary 2)
  - cascade regression did not require Gaussian assumption

Empirical cascade regression on real stock data

- in practice, we have data $x^{(1)}, \dots, x^{(n)} \in \mathbf{R}^d$ and we want to find (T, r) and $A \in \mathbf{R}^{d \times d}$

$$\text{minimize} \quad \sum_{k=1}^{n} \|Ax^{(k)} - x^{(k)}\| \quad \text{subject to} \quad A \in \mathrm{sparse}(T, r)$$

- ten years of daily stock price data from the Wall Street Journal for the Dow Jones 30, here's the tree



- nodes are stocks colored by industry; roughly speaking *stocks in similar industries are connected*

Conclusion: theoretical results build understanding, yield practical technique

- in summary, our contributions are
  1. posing and solving cascade regression
  2. analyzing tree linear cascades
  3. giving a non-Gaussian interpretation of Chow-Liu
- next steps include applications, other problem variants
  - e.g., non-linear featurized case, block case
- more details and full proofs available in paper and at poster session, *thanks!*