Using Bayesian Networks to Analyze Expression Data

Nir Friedman • Michal Linial
Iftach Nachman • Dana Peér

Hebrew University
Jerusalem, Israel

Presented By
Ruchira Datta

April 4, 2001

Ways of Looking At Gene Expression Data

- *Discriminant analysis* seeks to identify genes which sort the cellular snapshots into previously defined classes.
- *Cluster analysis* seeks to identify genes which vary together, thus identifying new classes.
- *Network modeling* seeks to identify the causal relationships among gene expression levels.

Why Causal Networks?

Explanation and Prescription

- *Explanation* is practically synonymous with an understanding of causation. Theoretical biologists have long speculated about biological networks (e.g., [Ros58]). But until recently few were empirically known. Theories need grounding in fact to grow.
- *Prescription* of specific interventions in living systems requires detailed understanding of causal relationships. To predict the effect of an intervention requires knowledge of causation, not just covariation.

Why Bayesian Networks?

Sound Semantics . . .

- Has well-understood algorithms
- Can analyze networks *locally*
- Outputs confidence measures
- Infers causality within probabilistic framework
- Allows integration of prior (causal) knowledge with data
- Subsumes and generalizes logical circuit models
- Can infer features of network even with sparse data

A philosophical question

What does probability mean?

- *Frequentists* consider the probability of an event as the expected frequency of the event as the number of trials grows asymptotically large.
- *Bayesians* consider the probability of an event to reflect our degree of belief about whether the event will occur.

Bayes's Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

"We are interested in A, and we begin with a *prior* probability $P(A)$ for our belief about A, and then we observe B. Then Bayes's Theorem . . . tells us that our revised belief for A, the *posterior* probability $P(A|B)$, is obtained by multiplying the prior $P(A)$ by the ratio $P(B|A)/P(B)$. The quantity $P(B|A)$, as a function of varying A for fixed B, is called the *likelihood* of A. . . . Often, we will think of A as a possible 'cause' of the 'effect' B . . . " [Cow98]

The Three Prisoners Paradox [Pea88]

- Three prisoners, A, B, and C, have been tried for murder.
- Exactly one will be hanged tomorrow morning, but only the guard knows who.
- A asks the guard to give a letter to another prisoner—one who will be released.
- Later A asks the guard to whom he gave the letter. The guard answers “B”.
- A thinks, “B will be released. Only C and I remain. My chances of dying have risen from 1/3 to 1/2.”

Wrong!

Three Prisoners (Continued)

More of *A*'s Thoughts

- When I made my request, I knew at least one of the other prisoners would be released.
- Regardless of my own status, each of the others had an equal chance of receiving my letter.
- Therefore what the guard told me should have given me no clue as to my own status.
- Yet now I see that my chance of dying is 1/2.
- If the guard had told me "C", my chance of dying would also be 1/2.
- So my chance of dying must have been 1/2 to begin with!

Huh?

Three Prisoners (Resolved)

Let's formalize ...

$$\begin{aligned} P(G_A|I_B) &= \frac{P(I_B|G_A)P(G_A)}{P(I_B)} \\ &= \frac{P(G_A)}{P(I_B)} = \frac{1/3}{2/3} = 1/2. \end{aligned}$$

What went wrong?

- We failed to take into account the context of the query: what other answers were possible.
- We should condition our analysis on the observed event, not on its implications.

$$\begin{aligned} P(G_A|I'_B) &= \frac{P(I'_B|G_A)P(G_A)}{P(I'_B)} \\ &= \frac{1/2 \cdot 1/3}{1/2} = 1/3. \end{aligned}$$

Dependencies come first!

- Numerical distributions may lead us astray.
- Make the qualitative analysis of dependencies and conditional independencies first.
- Thoroughly analyze semantic considerations to avoid pitfalls.

We *don't* calculate the conditional probability by first finding the joint distribution and then dividing:

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

We *don't* determine independence by checking whether equality holds:

$$P(A)P(B) = P(A, B)$$

What's A Bayesian Network?

Graphical Model & Conditional Distributions

- The *graphical model* is a DAG (directed acyclic graph).
- Each vertex represents a random variable.
- Each edge represents a dependence.
- We make the *Markov assumption*:

Each variable is independent of its non-descendants, given its parents.

- We have a conditional distribution $P(X|Y_1, \ldots, Y_k)$ for each vertex X with parents $Y_1, \ldots, Y_k$.
- Together, these completely determine the joint distribution:

  $P(X_1, \ldots, X_n) = \Pi_{i=1}^{n} P(X_i|\text{parents of } X_i)$.

Conditional Distributions

- Discrete, discrete parents (multinomial): table
  - Completely general representation
  - Exponential in number of parents
- Continuous, continuous parents: linear Gaussian

  $$P(X|Y_i\text{'s}) \propto N(\mu_0 + \sum_i a_i \cdot \mu_i,\ \sigma^2)$$

  - Mean varies linearly with means of parents
  - Variance is independent of parents
- Continuous, discrete parents (hybrid): conditional Gaussian
  - Table with linear Gaussian entries

Equivalent Networks

Same Dependencies, Different Graphs

- Set of conditional independence statements does not completely determine graph
- Directions of some directed edges may be undetermined
- But relation of having a common child is always the same (e.g., $X \to Z \leftarrow Y$)
- Unique PDAG (partially directed acyclic graph) for each class

Inductive Causation

[PV91]

- For each pair X, Y:
  - Find set S_{XY} s.t. X and Y are independent given S_{XY}
  - If no such set, draw undirected edge X, Y
- For each (X, Y, Z) such that
  - X, Y are not neighbors
  - Z is a neighbor of both X and Y
  - $Z \notin S_{XY}$

  add arrows: $X \rightarrow Z \leftarrow Y$

Inductive Causation (Continued)

- Recursively apply:
  - For each undirected edge $\{X, Y\}$, if there is a strictly directed path from X to Y, direct the edge from X to Y
  - For each directed edge (X, Y) and undirected edge $\{Y, Z\}$ s.t. X is not adjacent to Z, direct the edge from Y to Z
- Mark as *causal* any directed edge (X, Y) s.t. there is some edge directed at X

Causation vs. Covariation [Pea88]

- Covariation does not imply causation
- How to infer causation?
  - chronologically: cause precedes effect
  - control: changing cause changes effect
  - negatively: changing something else changes the effect, not the cause
    - ∗ turning sprinkler on wets the grass but does not cause rain to fall
    - ∗ this is used in Inductive Causation algorithm
- Undirected edge represents covariation of two observed variables due to a third *hidden* or *latent* variable

Causal Networks

- Causal network is also a DAG
- *Causal Markov Assumption:* Given X's *immediate causes* (its parents), it is independent of earlier causes
- PDAG representation of Bayesian network may represent multiple latent structures (causal networks including hidden causes)
- Can also use interventions to help infer causation (see [CY99])
  - If we experimentally set X to x, we remove all arcs *into* X and set $P(X = x|\text{what we did}) = 1$, before inferring conditional distributions

Learning Bayesian Networks

- Search for Bayesian network with best score
- Bayesian scoring function: posterior probability of graph given data

$$\begin{aligned} S(G:D) &= \log P(G|D) \\ &= \log P(D|G) + \log P(G) + C \end{aligned}$$

- $P(D|G)$ is the *marginal likelihood*, given by

$$P(D|G) = \int P(D|G, \Theta) P(\Theta|G)\, d\Theta$$

- Θ are parameters (meaning depends on assumptions)
  - parameters of a Gaussian distribution are mean and variance
- choose priors $P(G)$ and $P(\Theta|G)$ as explained in [Hec98] and [HG95] (Dirichlet, normal-Wishart)
- graph structures with right dependencies maximize score

Scoring Function Properties

With these priors:

- if assume *complete data* (all variables always observed):
  - equivalent graphs have same score
  - score is decomposable as sum of local contributions (depending on a variable and its parents)
  - have closed form formulas for local contributions (see [HG95])

Partial Models

Gene Expression Data:
Few Samples, Many Variables

- too few samples to completely determine network
- find partial model: family of possible networks
- look for features preserved among many possible networks
  - *Markov relations:* the *Markov blanket* of X is the minimal set of X_i's such that given those, X is independent of the rest of the X_i's
  - *order relations:* X is an ancestor of Y

Confidence Measures

- Lotfi Zadeh complains: conditional distributions of each variable are too crisp
  - (He might prefer fuzzy cluster analysis: see [HKKR99])
- assign *confidence measures* to each feature f by bootstrap method

$$p_N^*(f) = \frac{1}{m} \sum_{i=1}^{m} f(\hat{G}_i)$$

where G_i is graph induced by dataset D_i obtained from original dataset D

Bootstrap Method

- *nonparametric bootstrap:* re-sample with replacement N instances from D to get D_i
- *parametric bootstrap:* sample N instances from network B induced by D to get D_i
  - "We are using simulation to answer the question: If the true network was indeed B, could we induce it from datasets of this size?" [FGW99]

Sparse Candidate Algorithm [FNP99]

- Searching space of all Bayesian networks is NP-hard
- *Repeat*
  - *Restrict* candidate parents of each X to those most relevant to X, excluding ancestors of X in the current network
  - *Maximize* score of network among all possible networks with these candidate parents
- *Until*
  - *score* no longer changes; *or*
  - set of *candidates* no longer changes, or a fixed iteration limit is reached

Sparse Candidates

Relevance: Mutual Information

- standard definition:

  $$I(X;Y) = \sum_{X,Y} (\hat{P})(x,y) \log \frac{\hat{P}(x,y)}{\hat{P}(x)\hat{P}(y)}$$

  problem: only pairwise

- distance between $\hat{P}(X,Y)$ and $\hat{P}(X)\hat{P}(Y)$

  $$I(X;Y) = D_{KL}(\hat{P}(X,Y)\|\hat{P}(X)\hat{P}(Y))$$

  where $D_{KL}(P\|Q)$ is the *Kullback-Leibler divergence*:

  $$D_{KL}(P(X)\|Q(X)) = \sum_{X} P(X) \log \frac{P(X)}{Q(X)};$$

  this measures how far X and Y are from being independent

Sparse Candidates

Relevance: Mutual Information

- once we already have a network B, measure the *discrepancy*

  $$M_{\text{Disc}}(X_i, X_j|B) = D_{KL}(\hat{P}(X_i, X_j)|P_B(X_i, X_j));$$

  this measures how poorly our network already models the relationship between X and Y

- Bayesian definition: defining *conditional mutual information* $I(X; Y|Z)$ to be

  $$\sum_Z \hat{P}(Z) D_{KL}(\hat{P}(X, Y|Z) \| \hat{P}(X|Z)\hat{P}(Y|Z)),$$

  define

  $$M_{\text{Shield}}(X_i, X_j|B) = I(X_i; X_j|\text{parents of } X_i);$$

  this measures how far the Markov assumption is from holding

Sparse Candidates

Optimizing

- greedy hill-climbing
- divide-and-conquer
  - could choose maximal weight candidate parents at each vertex, except need acyclicity
  - decompose into strongly connected components (SCC's)
  - within an SCC, find separator (bottleneck), break cycle at separator using complete order of vertices in separator
  - to this end, first find cluster tree
  - then use dynamic programming to find optimum for all separators, all orders

Local Probability Models

Cost-Benefit

- multinomial loses information about expression levels
- linear Gaussian only detects near-linear dependencies

Robustness Analysis

- analyzed dataset: 76 gene expression levels of *S. cerevisiae*, measuring six time series along cell cycle ([SSZ$^+$98])
- perturbed datasets:
  - randomized data: permuted experiments
  - added genes
  - changed discretization thresholds
  - normalized expression levels
  - used multinomial or linear-Gaussian distributions
- robust persistence of findings
- Markov relations more easily disrupted than order relations

Biological Features Found

- order relations found dominating genes: “indicative of causal sources of the cell-cycle process”
- Markov relations reveal biologically sensible pairs
- some Markov relations revealed biologically sensible pairs not found by clustering methods (e.g., contrary to correlation)

References

[Cow98] Robert Cowell. Introduction to inference for bayesian networks. In Michael Jordan, editor, *Learning in Graphical Models*, pages 9–26. Kluwer Academic, 1998.

[CY99] Gregory F. Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In Kathryn B. Laskey and Henri Prade, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Fifteenth Conference*, pages 116–125. Morgan Kaufmann, 1999.

[FGW99] Nir Friedman, Moises Goldszmidt, and Abraham Wyner. Data analysis with bayesian networks: A bootstrap approach. In Kathryn B. Laskey and Henri Prade, editors, *Uncertainty in Artificial Intelligence: Proceedings of the*

Fifteenth Conference, pages 196–205. Morgan Kaufmann, 1999.

[FNP99] Nir Friedman, Iftach Nachman, and Dana Peér. Learning bayesian network structure from massive datasets: The 'sparse candidate' algorithm. In Kathryn B. Laskey and Henri Prade, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Fifteenth Conference*. Morgan Kaufmann, 1999.

[Hec98] David Heckerman. A tutorial on learning with bayesian networks. In Michael Jordan, editor, *Learning in Graphical Models*, pages 301–354. Kluwer Academic, 1998.

[HG95] David Heckerman and Dan Geiger. Learning bayesian networks: A unification for discrete and gaussian domains. In Philippe Besnard and Steve Hanks, editors, *Uncertainty in Artificial Intelligence: Proceedings of the*

Eleventh Conference, pages 274–284. Morgan Kaufmann, 1995.

[HKKR99] Frank Höppner, Frank Klawonn, Rudolf Kruse, and Thomas Runkler. *Fuzzy Cluster Analysis*. John Wiley & Sons, 1999.

[Pea88] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[PV91] Judea Pearl and Thomas S. Verma. A theory of inferred causation. In James Allen, Richard Fikes, and Erik Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference (KR '91)*, pages 441–452. Morgan Kaufmann, 1991.

[Ros58] Robert Rosen. The representation of biological systems from the standpoint of the theory of categories. *Bulletin of Mathematical Biophysics*, 20:317–341, 1958.

[SSZ^{+}98] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.