Computational Systems Biology: Biology X

Bud Mishra

Room 1002, 715 Broadway, Courant Institute, NYU, New York, USA

L#3:(Feb-14-2011)
Genome Wide Association Studies

Outline

The law of causality ... is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm ...

–Bertrand Russell, *On the Notion of Cause*. Proceedings of the Aristotelian Society 13: 1-26, 1913.

Outline

Terminology

- Data used in population-based genetic studies have three components:
  1. the *genotype* of the organism under investigation;
  2. a single trait or multiple *traits* (also, called *phenotypes*) that are associated with disease progression or disease status; and
  3. patient specific *covariates*, including treatment history and additional clinical and demographic information.

- **Primary Aim**: Characterize the relationship between the first two of these components: the genotype and a trait
- *Pharmacogenomics*: analyzes how genotypes modify the effects of drug exposure (a *covariate*) on a trait.
- *Focus on certain epidemiological principles*: Confounding and effect mediation, effect modification and conditional association.
- *Haplotypes and Phase Ambiguities*

Genetic Information

- **Genotype**: Observed genetic sequence information — a categorical variable.
- **Genes**: Regions of DNA that are eventually translated into proteins, or involved in the regulation of transcription.
- **Candidate Gene Studies**: A set of genes under investigation is chosen based on known biological function.
- In whole or partial *genome-wide association studies (GWAS)*, segments of DNA across large regions of the genome are considered and may not be accompanied by an a priori hypothesis abut the specific pathways to disease.

Polymorphisms

- Polymorphism refers to multiple alleles of a gene (more generally, genomic region) within a population, usually (but not always) expressing different phenotypes.
- **SNP: Single Nucleotide Polymorphisms**: describe a single base-pair change that is variable across the general population at a frequency of at least 1%. Common SNPs (frequency $>$ 5%); Rare SNPs (5% $>$ frequency $>$ 1%).
- Regions of DNA are said to have *genetic variability* if the alleles within the region vary across a population. *Conserved regions* exhibit no variability in a population.

Multilocus Genotype

- **Multilocus Genotype** describes the observed genotype across multiple SNPs or genes... *Genotype and multilocus genotype are often used interchangeably*.
- A *locus* or *site* refer to the portion of the genome, which encodes a single gene or the location of a single nucleotide on the genome. Multilocus genotype data consist of a set/sequence of categorical variables with elements corresponding to the genotype at each of multiple sites on the genome.

Haplotype

- **Haplotype** refers to the specific combination of alleles that are in *alignment* on a single *homolog* (one of the two homologous chromosomes in humans).
- The corresponding pair of haplotypes is referred to as the individual's *diplotype*.
- **Missingness**: Unobserved haplotypes and haplotype ambiguities. (More to be said on this topic.)

Zygousity

- **Zygousity** is the comparative genetic makeup of two homologous chromosomes.
- An individual is said to be *homozygous* at a given locus (e.g., SNP) if two locus pairs are the same.
- An individual is said to be *heterozygous* at a given locus (e.g., SNP) if it has more than one allele at that site.
- The *loss of heterozygousity* (LOH) refers to the loss of function of an allele when the second allele is already inactive (through inheritance of the heterozygous genotype.)

Allele Frequency

- The *minor allele* frequency (also, referred to as the *variant allele* frequency) refers to the frequency of the less common allele at a variable site.
- Note: in genetics, **frequency** refers to a proportion in the population.
- The terms *homozygous rare* and *homozygous variant* refer to homozygous with two copies of the minor allele.

Traits

- **Population Based Genetic Association Studies**: aim to relate genetic information to a *clinical outcome* or *phenotype* — also, called a *trait*.
- **Quantitative Trait**: A trait described by a continuous variable (taking values in a range), e.g., cholesterol level.
- **Binary Trait**: A trait described by a binary variable (more generally discrete values), e.g., diseased or non-diseased; HIV-positive or HIV-negative.

Phenotype

- **Phenotype**: is defined as a physical attribute or the manifestation of a trait (e.g., measure of disease progression). For instance, IC_{50}, an *in vitro* measure called "50% inhibitory concentration" — Amount of drug required to reduce the replication rate of a virus by 50%.
- **Outcome**: Presence of a disease, but more generally, values taken by a random variable.

Measuring Traits

- Traits can be measured cross-sectionally or over multiple time points (spanning several weeks to years). Data measured over time are referred to as *longitudinal* or *multivariate* data.
- Data and sample size are important; they relate to the questions of *multiple-hypotheses testing* and *overfitting*.

Covariates

- Collection of other information on patient specific characteristics.
- Example: Trait = Cholesterol level among patients at risk for cardiovascular diseases; Genotype = SNP polymorphisms in coding regions of certain genes; Covariates: Gender, age, smoking status and BMI (body mass index).
- A covariate may be *confounding* (associated with the outcome, either because it is a secondary trait or because a genotype is a common cause of the trait and the covariate).
- A covariate may be *causal* (associated with the outcome only in conjunction with a genotype).

Population Based Genetic Association Studies

- **Aim**: Relate genetic sequence information derived from unrelated individuals to a (i) measure of disease progression or (ii) disease status.
- **Uncovering Disease Etiology**: *Determining mechanisms for causation*. Inferred from complex disease association studies in humans. Usually association between single nucleotide polymorphisms (SNPs) and a trait.
- GWAS (Genome-Wide Association Studies): tend to be less hypothesis driven and involve very large number of SNPs in genome-wide scans.
- Partial scans: 100 Kb – 500 Kb segments of DNA; Whole-genome Scans: 500 Kb – 1 Mb.

Types of Investigations

- Population-based genetic association studies can be divided into four categories:
  1. **Candidate Polymorphism**
  2. **Candidate Gene**
  3. **Fine Mapping**
  4. **Whole or Partial Genome-Wide Scans**

Candidate Polymorphism Studies

- Based on an *a priori* hypothesis about functionality and its association with a particular polymorphism.
- Primary hypothesis: The variable site (under investigation) is *functional*.
- **Missense & Nonsense Mutations** suggest possible hypotheses.
- With a missense mutation, the new nucleotide alters the codon so as to produce an altered amino acid in the protein product.
- With a nonsense mutation, the new nucleotide changes a codon that specified an amino acid to one of the STOP codons (TAA, TAG, or TGA). Therefore, translation of the messenger RNA transcribed from this mutant gene will stop prematurely.

Candidate Gene Studies

- Based on an *a priori* hypothesis about functionality of a gene and a multiple SNPs within a single gene.
- The choice of SNPs depends on defined linkage disequilibrium (LD) blocks.
- **Assumption**: *The SNPs under investigation capture information about the underlying genetic variability of the gene — though the SNPs may not be serve as the true disease-causing variants.*
- While precise causal-variant SNP may not be known, lot more information may be available from the multiple SNPs “close” to the causal site on the genome.

- The “proximate” SNPs are known as “markers,” (or “Biomarkers”) — Observed genotype at the marker locations tends to be associated with the genotype at the true disease-causing locus.
- Over evolutionary time, the disease allele was inherited alongside variants at the marker loci — the correlation is high if the probability of a recombination event in the DNA region between the disease locus and the marker locus is small.

Fine Mapping Studies

- *Fine mapping* studies set out to identify (with a high level of precision) the location of a disease causing variant.
- Within the context of mapping studies, the term *quantitative trait loci (QTL)* refers to a chromosomal position that underlies a trait.
- Usually these methods are used for mapping and controlled experiments on inbred model animal lines...

Genome-Wide Association Studies (GWAS)

- It refers to an examination of genetic variation across a given genome, designed to identify genetic associations with observable traits.
- These studies normally require two groups of participants: people with the disease (cases) and similar people without (controls). After genotyping each participant, the set of markers, such as SNPs, are scanned. Then statistical algorithms are applied to survey participants' genomes for markers of genetic variation.

- If genetic variations are more frequent in people with the disease, the variations are said to be “associated” with the disease. The associated genetic variations are then considered as pointers to the region of the human genome where the disease-causing problem is likely to reside.
- Since the entire genome is analysed for the genetic associations of a particular disease, this technique allows the genetics of a disease to be investigated in a non-hypothesis-driven manner.

Genotypes vs. Gene Expressions

- Associations Studies: Now refer to studies relating *sequence* information to a phenotype. Gene Expression Studies (based on microarray data) characterize association among gene *products* – e.g., RNA or protein abundance – and disease outcomes.
- Note: Sequence based data are more stable — not easily affected by local conditions. Time-course gene-expression data capture more temporal information – missing from the sequence data.

- Genotype data explain the variability in terms of the predictor (independent) variables based on polymorphisms. Gene-expression studies give insights into the structure of the biochemical pathways and the genetic circuits controlling the traits – transcription, translation, protein-protein interactions, degradation, transportation, etc.
- Genotype is treated as a *categorical variable*; Transcriptome profiles are described in terms of *continuous variables*
- These two studies are, however, intimately related by the basic molecular biology of regulation, metabolism and signaling.

Genotype precedes the trait in a Causal-chain.

- Our studies will be anchored on genotype (sequence) data and physiological traits.
- Confirmation/explanation in terms of mechanisms will come later. Primarily, through a Bayesian inference process.

Epigenetics.

- The term *epigenetics* is used to describe heritable features that control the functioning of genes within an individual cell (but do not constitute a physical change in the corresponding DNA sequence.
- EPIGENOME: Above the genome.
- *Epigenetic Code*: Information on methylation and histone patterns (*epigenetic tags*. Have an essential role in the control of gene expressions. Through gene inhibition or silencing, the epigenetics environment (Hypo- or hyper-methylated) plays an important role in many cancers.

Complex Diseases.

- Many diseases such as cardiovascular disease and cancer are thought to have a complex origin — require multiple genetic and environmental factors.
- Complex disorders often cluster in families — but, they do not have a clear-cut pattern of inheritance. This makes it difficult to determine a person's risk of inheriting or passing on these disorders.
- Complex disorders are also difficult to study and treat because the specific factors that cause most of these disorders are not easy to identify exhaustively.

Cell Division

- New generation of maternal and paternal gametes combine to form a *zygote*
- **Cross-over or Recombination Event**: In the process of meiosis, **cross-over** between the maternal and paternal chromatids can occur.
- An important aspect of meiosis is that whole portions or *segments* of DNA within a chromosome tend to be passed from one generation to another. — However portions of DNA within a chromosome that are far from one another are less likely to be inherited together. SNPs in each such segments may play different roles.

SNPs and LD Blocks

- **Functional SNPs**: They affect a trait directly — a component within the causal pathway of a disease.
- **Haplotype Tagging SNPs**: They capture over-all variability within the gene under consideration.
- **Linkage Disequilibrium Blocks**: Its structure depends on probability of recombination within a region.

Outline

Random Variables

- A (discrete) random variable is a numerical quantity that in some experiment (involving randomness) takes a value from some (discrete) set of possible values.
- More formally, these are measurable maps

$$X(\omega), \omega \in \Omega,$$

from a basic probability space (Ω, F, P) ($\equiv$ outcomes, a sigma field of subsets of Ω and probability measure P on F).

- *Events*

$$...\{\omega \in \Omega | X(\omega) = x_i\}...$$

same as $\{X = x_i\}$ [X assumes the value x_i].

Few Examples

- Example 1: Rolling of two six-sided dice. Random Variable might be the sum of the two numbers showing on the dice. The possible values of the random variable are 2, 3, . . ., 12.
- Example 2: Occurrence of a specific word *GAATTC* in a genome. Random Variable might be the number of occurrence of this word in a random genome of length 3×10^9. The possible values of the random variable are 0, 1, 2, . . ., 3×10^9.

The Probability Distribution

- The *probability distribution* of a discrete random variable Y is the set of values that this random variable can take, together with the set of associated probabilities.
- Probabilities are numbers in the range between zero and one (inclusive) that always add up to one when summed over all possible values of the random variable.

Bernoulli Trial

- A *Bernoulli trial* is a single trial with two possible outcomes: "success" & "failure."

$$P(\text{success}) = p \text{ and } P(\text{failure}) = 1 - p \equiv q.$$

- Random variable S takes the value -1 if the trial results in failure and $+1$ if it results in success.

$$P_S(s) = p^{(1+s)/2} q^{(1-s)/2}, \quad s = -1, +1.$$

The Binomial Distribution

- A *Binomial random variable* is the number of successes in a fixed number n of independent Bernoulli trials (with success probability = p).
- Random variable Y denotes the total number of successes in the n trials.

$$P_Y(y) = \binom{n}{y} p^y q^{n-y}, \quad y = 0, 1, \ldots, n.$$

The Uniform Distribution

- A random variable Y has the *uniform distribution* if the possible values of Y are a, $a+1$, ..., $a+b-1$ for two integer constants a and b, and the probability that Y takes any specified one of these b possible values is b^{-1}.

$$P_Y(y) = b^{-1}, \quad y = a, a+1, \ldots, a+b-1.$$

The Geometric Distribution

- Suppose that a sequence of independent Bernoulli trials is conducted, each trial having probability p of success. The random variable of interest is the number Y of trials before but not including the first failure. The possible values of Y are 0, 1, 2, ...

$$P_Y(y) = p^y q, \quad y = 0, 1, \ldots.$$

The Poisson Distribution

- A random variable Y has a Poisson distribution (with parameter $\lambda > 0$) if

$$P_Y(y) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y = 0, 1, \ldots.$$

- The Poisson distribution often arises as a limiting form of the binomial distribution.

Continuous Random Variables

- We denote a continuous random variable by X and observed value of the random variable by x.
- Each random variable X with range I has an associated density function $f_X(x)$ which is defined, positive for all x and integrates to one over the range I.

$$\text{Prob}(a < X < b) = \int_a^b f_X(x)dx.$$

The Normal Distribution

- A random variable X has a normal or Gaussian distribution if it has range $(-\infty, \infty)$ and density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where μ and $\sigma > 0$ are parameters of the distribution.

Expectation

- For a random variable Y, and any function $g(Y)$ of Y, the expected value of $g(Y)$ is

$$E(g(Y)) = \sum_y g(y) P_Y(y),$$

when Y is discrete; and

$$E(g(Y)) = \int_y g(y) f_Y(y)\, dy,$$

when Y is continuous.

- Thus,

$$\text{mean}(Y) = E(Y) = \mu(Y),$$

$$\text{variance}(Y) = E(Y^2) - E(Y)^2 = \sigma^2(Y).$$

Conditional Probabilities

- Suppose that A_1 and A_2 are two events such that $P(A_2) \neq 0$. Then the conditional probability that the event A_1 occurs, given that event A_2 occurs, denoted by $P(A_1|A_2)$ is given by the formula

$$P(A_1|A_2) = \frac{P(A_1 \& A_2)}{P(A_2)}.$$

Bayes Rule

- Suppose that A_1 and A_2 are two events such that $P(A_1) \neq 0$ and $P(A_2) \neq 0$. Then

$$P(A_2|A_1) = \frac{P(A_2)P(A_1|A_2)}{P(A_1)}.$$

Bayes Nets

- **Bayes Nets** or **Bayesian networks** are graphical representation for probabilistic relationships among a set of random variables.
- Given a finite set $X = \{X_1, \ldots, X_n\}$ of discrete random variables where each variable X_i may take values from a finite set, denoted by $Val(X_i)$.
- A Bayesian network is an annotated directed acyclic graph (DAG) G that encodes a joint probability distribution over X.

- The graph G (Bayesian Network) is defined as follows:
- The nodes of the graph correspond to the random variables

$$X_1, X_2, \ldots, X_n.$$

- The links of the graph correspond to the direct influence from one variable to the other.
  1. If there is a directed link from variableX_i to variable X_j, variable X_i will be a **parent** of variable X_j.
  2. Each node is annotated with a conditional probability distribution (CPD) that represents

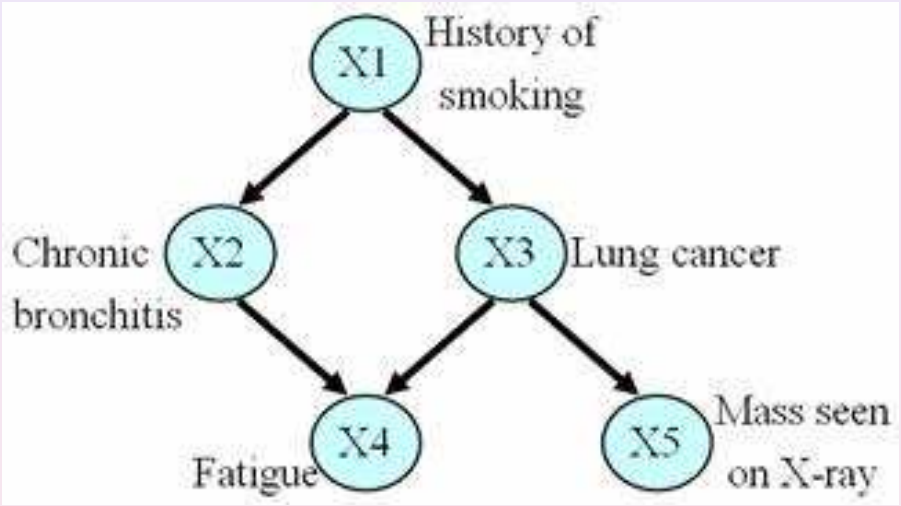
$$Pr(X_i|Pa(X_i)),$$

where $Pa(X_i)$ denotes the parents of X_i in G.

- The pair (G, CPD) encodes the joint distribution $Pr(X_1, \ldots, X_n)$.

A unique joint probability distribution over X from G is factorized as:

$$Pr(X_1, \ldots, X_n) = \prod_i (Pr(X_i | Pa(X_i))).$$


X1
History of smoking
Chronic bronchitis
X2
X3
Lung cancer
Fatigue
X4
X5
Mass seen on X-ray

$Pr(X_1 = no)$	=	0.8	$Pr(X_1 = yes)$	=	0.2
$Pr(X_2 = absent \mid X_1 = no)$	=	0.95	$Pr(X_2 = absent \mid X_1 = yes)$	=	0.75
$Pr(X_2 = present \mid X_1 = no)$	=	0.05	$Pr(X_2 = present \mid X_1 = yes)$	=	0.25
$Pr(X_3 = absent \mid X_1 = no)$	=	0.99995	$Pr(X_3 = absent \mid X_1 = yes)$	=	0.997
$Pr(X_3 = absent \mid X_1 = no)$	=	0.00005	$Pr(X_3 = absent \mid X_1 = yes)$	=	0.003
$Pr(X_4 = absent \mid X_2 = absent, X_3 = absent)$	=	0.95	$Pr(X_4 = absent \mid X_2 = absent, X_3 = present)$	=	0.5
$Pr(X_4 = absent \mid X_2 = present, X_3 = absent)$	=	0.9	$Pr(X_4 = absent \mid X_2 = present, X_3 = present)$	=	0.25
$Pr(X_4 = present \mid X_2 = absent, X_3 = absent)$	=	0.05	$Pr(X_4 = present \mid X_2 = absent, X_3 = present)$	=	0.5
$Pr(X_4 = present \mid X_2 = present, X_3 = absent)$	=	0.1	$Pr(X_4 = present \mid X_2 = present, X_3 = present)$	=	0.75
$Pr(X_5 = absent \mid X_3 = absent)$	=	0.98	$Pr(X_5 = absent \mid X_3 = present)$	=	0.4
$Pr(X_5 = present \mid X_3 = absent)$	=	0.02	$Pr(X_5 = present \mid X_3 = present)$	=	0.6

Causal Bayesian networks

- A **causal Bayesian network** of a domain is similar as the normal Bayesian network — the difference is in the explanation of the links in the Bayesian networks.
  - In the normal Bayesian networks, the links between variables can be explained as correlation or association.
  - In a causal Bayesian network, the links mean that the parent variables will causally influence the values of the child variables.

- The causal influence in this thesis is defined based on "*manipulation criteria*:"

Manipulation Criteria

Suppose there are two variables A and B in the domain; If we can manipulate the variables in the domain, set the value of variable A as a_1 or a_2 and measure its effect on variable B, then the probability distribution of variable B will change under the conditions of the different values of variable A

$$Pr(B|do(A = a_1)) \neq Pr(B|do(A = a_2)).$$

Bayesian network structure learning

- The main task in Bayesian network structure learning:

Find a structure of Bayesian network that describes the observed data the best.

- The problem is NP-complete.
- Many heuristics have been proposed to learn Bayesian network structure. There are two categories of approaches for Bayesian network structure learning: **The score-and-search-based approach** and **The constraint-based approach**.

The score-and-search-based approach

The score-and-search-based approach:

- The methods in this category start from an initial structure (generated randomly or from domain knowledge) and move to the neighbors with the best score in the structure space determinately or stochastically until a local maximum of the selected criteria is reached.
- The greedy learning process can re-start several times with different initial structures to improve the result.

The constraint-based approach

The constraint-based approach:

- The methods under this category start to test the statistical significance of the pairs of variables conditioning on other variables to induce conditional independence.
- The pairs of variables which pass some threshold are deemed as directly connected in the Bayesian networks.
- The complete Bayesian network structure is constructed from the induced conditional independence and dependence information.

Structure space

- The search space in Bayesian network structure learning is all the possible structures of directed acyclic graphs (DAGs) given the number of variables in the domain.
- Note that the number of possible DAGs containing n nodes is super exponential in n:

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i).$$

It is impossible to enumerate all possible structures and score them, even for a reasonably-sized number of nodes n in the Bayesian network. Then heuristic-based methods have been proposed to find a local maximum in the structure space. The representative methods are K2 algorithm, greedy search, etc.

Size of the Structure space

Number of variables in DAG	Number of the possible DAGs
1	1
2	3
3	25
4	543
5	29,281
6	3,781,503
7	1,138,779,265
8	78,370,2329,343
9	1,213,442,454,842,881
10	4,175,098,976,430,598,100

K2 algorithm

- Since one cannot enumerate all the possible DAGs for Bayesian network structure learning, various heuristic methods are used.
- Since DAGs are acyclic and the parents of the variables come before children in causal ordering, knowing the **ordering of the variables** can reduce the structure space. If we know a total ordering on the nodes, finding the best structure amounts to picking the best set of parents for each node independently.
- This is what the K2 algorithm does.

K2 algorithm

- K2 algorithm is a greedy search algorithm that works as follows:
  1. Suppose one knows the total ordering of the nodes.
  2. Initially each node has no parents. The algorithm then incrementally adds the parent whose addition most increases the score of the resulting structure.
  3. When no addition of a single parent can increase the score, it stops adding parents to the node.
- Since an ordering of the nodes is known beforehand, the search space under this constraint is much smaller than the entire space.

K2 algorithm

- Note: K2 does not need to check for cycles, since the total ordering guarantees that there is no cycle in the deduced structures.
- Furthermore, based on some reasonable assumptions, one can choose the parents for each node independently.
- If the ordering is unknown, K2 can search over various "plausible" orderings. The space of orderings is much smaller and more regular than the space of the structures, and has a smoother posterior "landscape." As a result, the search over ordering is more efficient than searching over DAGs.

Greedy search

- If we know nothing about the structure, we can treat the structure learning problem as a general optimization problem in a discrete space.
- The intuitive way to solve such a problem is “greedy search.”
- Greedy search starts at a specific point (an initial structure) in the structure space, considers all nearest neighbors of the current point, and moves to the neighbor that has the highest score; if no neighbors have higher score than the current point (i.e., we have reached a local maximum), the algorithm stops.

Greedy search

The greedy search process for score-and-search-based approach is as follows:

Algorithm 1: GREEDY - pseudo code

Input: Observational data set;
Output: A Bayesian network.

1. Generate the initial Bayesian network, evaluate it and set it as the current Bayesian network;
2. Evaluate the neighbors of the current Bayesian network;
3. Set the neighbor with the best score as the current Bayesian network;
4. **repeat**
5. | Steps 2. and 3.
6. **until** *the best score of the neighbors is NOT better than the score of the current Bayesian network* ;
7. **return** *The Best Bayesian Network found so far*;

Scoring Function

- In Bayesian network structure learning, a scoring function evaluates how well a given network G matches the data D. Given a scoring function, the best Bayesian network is the one that maximizes this scoring function.
- An ad-hoc scoring function is based on the maximum likelihood (ML) principle:

$$\hat{\theta_G} = \arg\max_{\theta_G}\{Pr(D|\theta_G, G^k)\}$$

where G^k denotes the hypothesis of the Bayesian structure, θ_G is the vector of parameters $(\theta_1, \ldots, \theta_n)$, θ_i is the vector of parameters for the distribution $Pr(x_i|Pa_i, \theta_i, G^k)$ of variable and $\hat{\theta_G}$ is the maximum likelihood configuration of θ_G.

Scoring Function

- Since the number of parameters (degrees-of-freedom or model complexity) can be very large depending on G, this approach may lead to an overfitting to Bayesian networks that match the given data well, but have low generalization quality for new data.
- Therefore, a constrained optimization of the scoring function is often used.
- Bayesian Information Criterion (BIC) and Bayesian score combine the likelihood with some penalty relating to the complexity of the model.

$$BIC = \log Pr(D|\hat{\theta_G}, G^k) - \frac{d}{2}\log N.$$

N is the number of data points and d is the degree of freedom.

Constraint-based approach

Assumptions

- **Causal sufficiency assumption:** *There exist no common unobserved (also known as hidden or latent) variables in the domain that are parent of one or more observed variables of the domain.*
- **Causal Markov assumption:** *Given a Bayesian network model G, any variable is independent of all its non-descendants in G given its parents.*
- **Faithfulness assumption:** *A Bayesian network structure G and a probability distribution P generated by G are faithful to one another if and only if every conditional independence relationship valid in P is entailed by the Causal Markov assumption in G.*

Constraint-based approach

- Constraint-based methods assume: Since a Bayesian network structure encodes many dependencies and (conditional) independencies of the underlying model, try to discover the dependencies and conditional independencies from the data, and then use these to infer the Bayesian network structure.
- The dependency and conditional independency relationships are measured by using some kind of Conditional Independence (CI) test. (Based on the earlier assumptions.)

Constraint-based approach

- Based on these assumptions, try to ascertain the existence of an edge between two variables, or the direction of that link, (only possible in certain cases). The output of constraint-based methods will be a partial DAG (PDAG) to represent the whole Markov equivalence class.
- The representative algorithms in this category are SGS algorithm, CI algorithm, and PC algorithm.

Hybrid methods

- We may combine the score-and-search-based approach with constraint-based approach for Bayesian network structure learning.
- Use the learned network from constraint-based methods as the start point for the search-and-score-based methods.
- For instance, generate the ordering of the variables with constraints-based approach and learn the Bayesian network structure with the score-and-search-based approach when the ordering is generated.

Causation and Correlation

- A fallacy, known as *cum hoc ergo propter hoc* (Latin for "with this, therefore because of this"): Correlations do not imply causation.
- Statements associated with *necessity* and *sufficiency*
- **The INUS condition**: An Insufficient but Non-redundant part of an Unnecessary but Sufficient condition.
- **The Probability Raising condition**
- **Temporal Priority**

How can Bayes Nets be Causal?

- The causal Markov condition is a relative of Reichenbach's thesis that "conditioning on common causes will render joint effects independent of other another."
- One can then add the assumption of faithfulness or stability as well as to assume that all underlying systems of causal laws are deterministic.
- Similarly, using causal minimality (or sufficiency) assumption, one may try to justify the claim that "Bayesian Nets Are All There Is to Causality..."
- "Bayes nets encode information about probabilistic independencies. Causality, if it has any connection with probability, would seem to be related to probabilistic dependence." (Cartwright, N.)

Probability Raising

"Causes produce their effects; they make them happen. So, in the right kind of population we can expect that there will be a higher frequency of the effect (E) when the cause (C) is present than when it is absent; and conversely for preventatives. What kind of populations are "the right kinds"? Populations in which the requisite causal process sometimes operates unimpeded and its doing so is not correlated with other processes that mask the increase in probability, such as the presence of a process preventing the effect or the absence of another positive process."

Cartwright

[End of Lecture #3]