10-708: Probabilistic Graphical Models, Spring 2020

1: Introduction

Lecturer: Eric P. Xing *Scribe: Ini Oguntola, Vineet Jain, Samuel Levy, Zihao Chen, Sungjun Choi*

Basic Probability Concepts

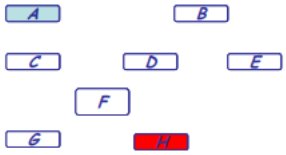

Figure 1: Multiple variables in a graph

- **Representation**: What is the joint probability distribution on multiple variables?

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \tag{1}$$

  There are 2^8 state configurations in total. But do they need to be all represented? Do we get any scientific/medical insight? One of the main benefits of graphical models is the cost savings in representing the joint distribution. Modeling the dependencies among the variables with a graph and conditionals can drastically reduce the number of parameters needed to describe the joint distribution, compared to what we would get with a full joint distribution table.

- **Learning**: Where do we get all these probabilities? Should we use maximum likelihood estimation? but how many data do we need for that? Could we use other estimation principles? Where do we incorporate domain knowledge in terms of plausible relationships between variables, and plausible values of the probabilities?

- **Inference**: If not all variables are observable, how do we compute the conditional distribution of latent variables given evidence? Computing $P(H \mid A)$ in Figure 1 would require summing over all 2^6 configurations of the unobserved variables: that requires a lot of compute power.

Multivariate Distribution in High-Dimensional Space

We start with an example from biology and represent cellular signal transduction as follows (Figure 2). Receptors A and B receive signal from cell surface, Kinases C, D and E read and decode the signal; TF F takes in the signal and triggers production of DNA with DNA template and Genes G and H are expressions of the DNA templates.

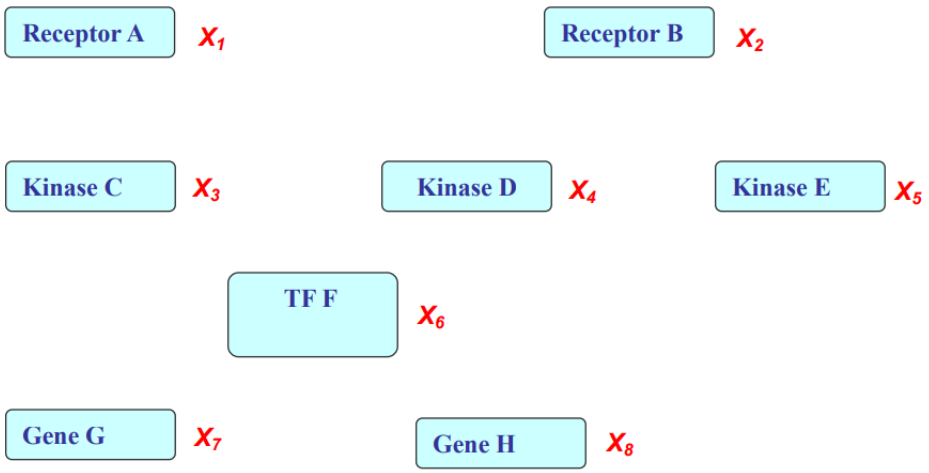

Figure 2: A possible world for cellular signal transduction

Although Figure 2 is a good start to model cellular signal transduction, the additional domain knowledge from biologist can be incorporated to impose a structure on the random variables A, B, C, D, E, F, G and H. Figure 3 partitions the random variab les into compartments they live in within a cell. The dependencies among the variables (nodes) are communication mechanisms and are modeled as edges. This representation allows us to derive the joint probabilities among the random variables using the factorization law.

A Structured View From Domain Experts

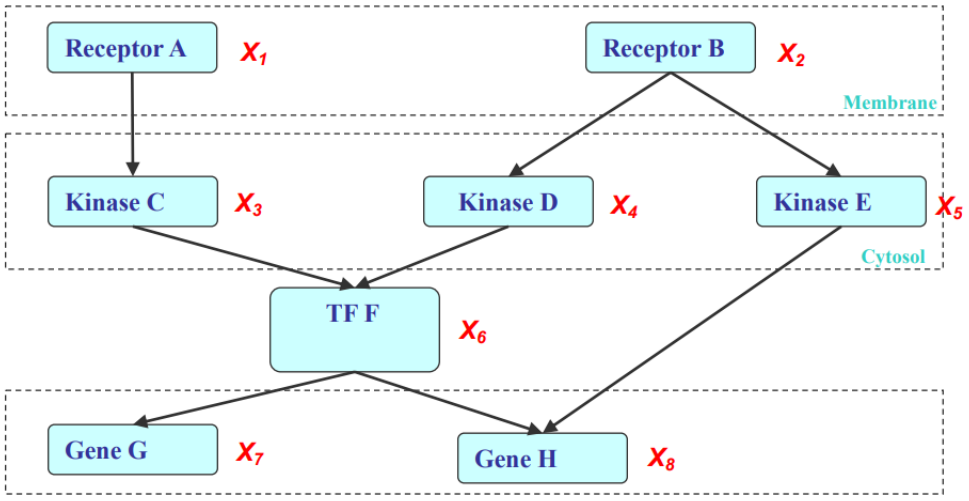

Figure 3: Dependencies among variables are represented in a directed acyclic graph

What are Graphical Models

Informally, a graphical model is a just a graph representing relationship among random variables. Nodes are random variables (features, not examples) and edges (or absence of edges) represent relationships or dependencies among random variables.

The notion of relationship varies depending on the graph. For example, in Figure 4, the graphical model is a representation of co-occurrences within a page between major Biblical figures.

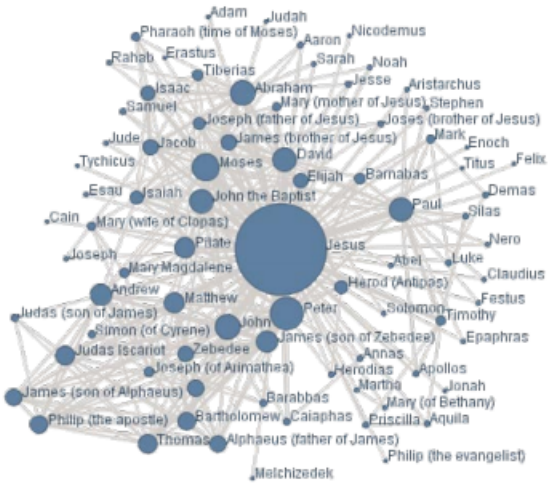

Figure 4: Graphical representation of co-occurrences of Biblical figures within a page

Relationship between Two Random Variables

Rigorously defining each component of a graphical model is crucial in avoiding multiple representations of the same phenomenon by different people, and as part of such effort we first delve into rigorously defining possible relationships between two random variables. The random variables may potentially have many types of relationships (some of which are listed in p.10), and to be rigorous we look for "one-number measures" to serve as summaries that quantitatively represent the presence, absence, or strength of such relationships.

Again, there are many such measures, some of which are listed and discussed below, and each has its adequacies and inadequacies. Choosing a measure is not a trivial task in the sense that, while one can arbitrarily choose one such measure, draw a graph out of data, and provide convincing "stories" out of the graph, unless the measure is chosen rigorously, the argument can be rather easily overthrown by counter-examples from the same data. It is thus vital to understand what each measure entails.

Pearson's Correlation

Pearson's correlation (denoted as ρ) is one of the most well-known and fundamental measures of association between random variables that is defined as follows:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Still, there are two important caveats to note:

1. Two independent variables are uncorrelated; however, the reverse does not hold. For example, let random variables X, Y be such that $X \sim U[-1, 1]$ and $Y = X^2$. Then while it is evident that Y (deterministically) depends on X, they are uncorrelated since:

$$\begin{aligned}\text{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \\ &= \mathbb{E}[XY] \quad (\because \mathbb{E}[X] = 0) \\ &= \mathbb{E}[X^3] = 0 \quad (\because X \text{ is not skewed})\end{aligned}$$

2. Pearson's correlation only captures linear dependence, as can be seen from the example above. This in turn means Pearson's correlation is very weak in terms of capturing independence.

Strong(er) Measures of Association

The limitations of Pearson's correlation introduced above calls for stronger measures, ones that can capture non-linear dependences and thus independences. In fact, for the following two measures we bring in the very definition of statistical independence between random variables to construct measures.

Exploiting the fact that the joint density P_{XY} of two jointly-distributed random variables X and Y can be factorized as $P_X P_Y$ if and only if they are independent of each other, we quantify the "distance" between the joint density P_{XY} and the product of marginals $P_X P_Y$. Indeed, this approach guarantees that the distance $= 0$ iff X and Y are independent.

Mutual Information

One of the most common measures of distance between two densities P and Q is **Kullback-Leibler divergence**, or KL-divergence in short:

$$KL(P, Q) = \int_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}\, dx$$

KL divergence returns 0 when P and Q are equal, i.e. $P(x) = Q(x)$, $\forall x \in \mathcal{X}$, and a larger positive value as P and Q deviate further from each other. Since we likewise want the distance to be 0 when $P_{XY}(x, y) = P_X(x)P_Y(y)$, $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$ and positive otherwise, we can utilize KL divergence to obtain our desired measure, known as **mutual information**:

$$I(X, Y) = KL(P_{XY}, P_X P_Y)$$

This measure indeed successfully captures non-linear dependences. However, it poses **computational issues** since integration over complex combination of non-Gaussian, multi-modal, and possibly even non-parametric densities is a significant challenge.

Hilbert-Schmidt Independence Criterion (HSIC)

A recent finding that also captures non-linear dependences is HSIC(Gretton et al. 2005). It's defined as the **maximum mean discrepancy (MMD)** between joint density P_{XY} and product of marginals $P_X P_Y$.

Definition of MMD:
Let P, Q be any two densities,

$$\begin{aligned} MMD(P,Q) &= ||\mu_k(P) - \mu_k Q||_{\mathcal{H}_k} \\ \mu_k(P) &= \mathbb{E}_{Z\sim P}[\phi(Z)] \qquad \text{(kernel embedding of } P\text{)} \\ \phi(Z) &= \text{feature map of kernel k} \end{aligned}$$

One important property of this measure is $HSIC(X,Y) = 0$ if and only if $X \perp Y$. More details will be covered in future lectures.

Towards Graphical Models: Partial Correlation

The measures of association between two random variables discussed in the previous sections can be used to define a **marginal correlation/dependency** graph. This is the most primitive form of graphical model in which we connect any pair of variables with a non-trivial pairwise correlation or mutual information or HSIC.

The drawback is that this type of graphical model is not very informative due to the reason that two random variables will have non-zero measure of association very rarely. We can almost always find some statistical association between a pair of variables, either due to some underlying process that affects both variables or sometimes due to random chance.

Consider the following example: define, X = height of kid, Y = vocabulary of kid, Z = age of kid. If we compute a pairwise measure of association between these variables, we expect to find all of them to be non-zero. However, we know from 'common sense' that the height of the kid and the vocabulary has no direct relation, rather the age of the kid is the underlying variable affecting both these values. In this case, a marginal dependency graph will have edges between all pairs of variables, but we can find a more informative structural relationship.

Partial Correlation

We can define a new measure of correlation between two variables **given another variable**. We can think of it as the correlation measured between two variables X and Y after conditioning on another variable Z, or after eliminating the linear effect of Z. This is known as the **partial/conditional correlation**.

$$\begin{aligned} \rho(X,Y|Z) = \rho(e_X, e_Y) &= \frac{\text{Cov}(e_X, e_Y)}{\sqrt{\text{Var}(e_X)}\sqrt{\text{Var}(e_Y)}} \\ e_X &= X - (\beta_X^T Z + \text{intercept}_X) \\ e_Y &= Y - (\beta_Y^T Z + \text{intercept}_Y) \end{aligned}$$

It is the correlation between the residuals from regressing Z to X and Z to Y linearly. In this sense, it is similar to Pearson's correlation.

$$\begin{aligned} X \perp\!\!\!\perp Y \mid Z &\implies \rho(X,Y|Z) = 0 \\ \rho(X,Y|Z) = 0 &\not\implies X \perp\!\!\!\perp Y \mid Z \end{aligned}$$

Partial Correlation Graph

We can now construct a more meaningful graphical model than the marginal dependency graph. We connect a pair of variables if they have non-trivial partial correlation given the rest of the variables.

One possible issue with this model is that it is computationally expensive to compute the partial correlation for every pair of variables conditioned on all the rest, since we need to first fit a (linear) regression for each of the conditioned variables. However, it turns out the partial correlation matrix R has a simple form related to the inverse covariance matrix Θ.

$$R_{ij} = \rho(X_i, X_j | X_{-ij})$$
$$R_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}}\sqrt{\Theta_{jj}}}$$

Conditional Independence

As revealed in previous sections, it's always helpful to reduce statistical and computational complexity if we can point out conditional independence. The classical notation for conditional independence is $X \perp Y|Z$, X, Y, Z are random variables. And we have the definition:

$$X \perp Y|Z \iff P(X, Y|Z) = P(X|Z)P(Y|Z)$$

It's a hard mission to extract conditional independence if we want to use strong dependency measures or partial correlation as a tool. One shortcut is simply impose Gaussian assumption to the random variables of interest. To be detailed, suppose (X, Y, Z) are jointly Gaussian, we have $\rho(X, Y|Z) = 0$ iff $X \perp Y|Z$. Many papers rely on this fact though may not state it explicitly.

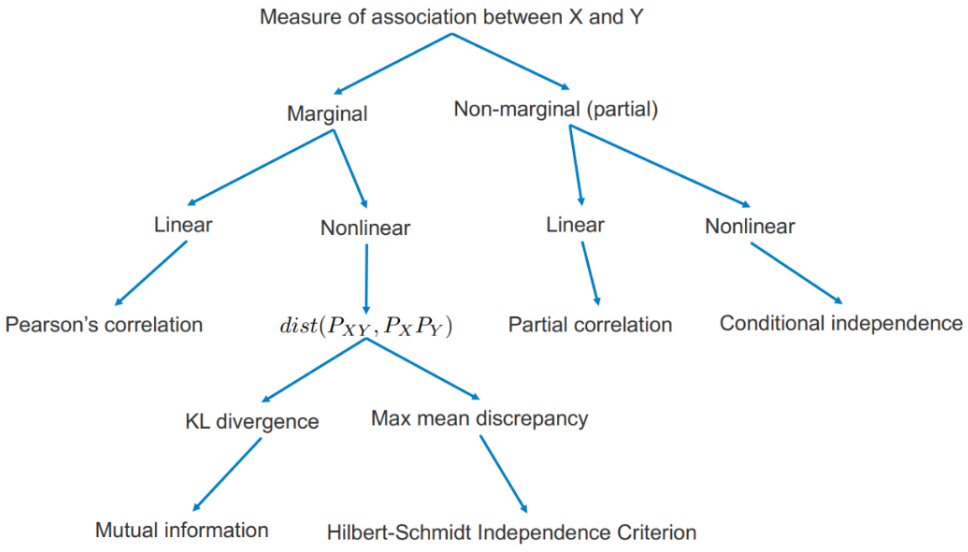

Figure 5: Summary of pairwise measures of association.

Why Graphical Models?

"Graphical models" really refer to a way of thinking, not necessarily to any particular individual model. They are a language for communication, computation, and development.

Probability theory provides the *glue* whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data.

The *graph theoretic* side of graphical models provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms.

Many of the classical multivariate probabilistic systems studied in fields such as statistics, systems engineering, information theory, pattern recognition and statistical mechanics are *special cases of the general graphical model formalism.*

The graphical model framework provides a way to view all of these systems as instances of a *common underlying formalism.*