BTRY 7210: Topics in Quantitative Genomics and Genetics

Jason Mezey
Biological Statistics and Computational Biology (BSCB)
Department of Genetic Medicine
jgm45@cornell.edu

April 9, 2015
Lecture 6: Intro to Probabilistic Graphical Models: Markov Random Fields (Gaussian Graphical Models)
Review: using eQTL to discover regulatory relationships between genes

• What we would like to discover:

• The core of this modeling / discover technique are Probabilistic Graphical Models (PGMs)

• Today, we will not discuss this class of PGMs with “directed” edges but rather a class with “undirected” edges that is one of the most important classes of PGMs used to analyze genomic data (and often used when analyzed eQTL networks)
Review: foundation of PGMs

- A PGM is a “picture” that represents the structure of a multivariate probability distribution for several random variables under a set of assumptions that restrict the possible model:

\[ Pr(Y_1, Y_2, Y_3, Y_4) \]

- A complete characterization would require information on this joint distribution and each of the possible marginal distributions and conditional distributions:

\[ Pr(Y_1) \quad Pr(Y_1, Y_2, Y_3) \quad Pr(Y_1|Y_2) \quad Pr(Y_1, Y_2|Y_3, Y_4) \]

- We are not going to worry about the specific “parameterized” probability distribution(s), i.e. it could be anything admissible (!!), we are just going to think about the marginals and conditionals

- More specifically, we will be concerned about which of these relationships are independent and non-independent:

\[ Pr(Y_1|Y_2) \neq Pr(Y_1) \quad Pr(Y_1, Y_2|Y_3) \neq Pr(Y_1, Y_2) \]

\[ Pr(Y_1|Y_3) = Pr(Y_1) \quad Pr(Y_1, Y_2|Y_3, Y_4) = Pr(Y_1, Y_2) \]
What the “graph” represents

- We assume a graphical model $G=(V, E)$ describes the joint distribution where each random variable is represented as a vertex and each edge represents a conditional relationship:

  \[ Pr(Y) = Pr(Y_1, Y_2, \ldots, Y_k) \]
  \[ V = (Y) = Y_1, Y_2, \ldots, Y_k \]
  \[ E(Y_i, Y_j) \implies Pr(Y_i|Y_j) \neq Pr(Y_i) \]

- Different “classes” of PGMs are represented by graphs with different types of edges, which reflect the “types” of conditional relationships allowed.
Today, we are going to discuss the basics of a class of PGMs that have “undirected” edges.

These are (usually) Markov Random Fields (MRF) and, if each of the variables is normally distributed (!!) we refer to the MRF as a Gaussian Graphical Model (GGM), i.e. a GGM is a MRF where all the variables are normally distributed.

The literature on MRFs is extensive (and scattered across fields...) involving many topics.

Our goal today is to give you a sense of what the underlying MRF model represents and how we might “learn” the structure of an MRF from data, when assuming that the true model that generated the data can be described by an MRF.
Markov Random Fields I

- As with any PGM, an MRF assumes that among a set of variables, some of the variables can be made *conditionally independent* by conditioning on other variables and some of the variables cannot be made independent no matter what variables they are conditioned on:

  \[ \Pr(Y_1, Y_2, Y_3) \]

  \[ \Pr(Y_1, Y_2) \neq \Pr(Y_1)\Pr(Y_2) \quad \Pr(Y_1, Y_2, | Y_3) = \Pr(Y_1|Y_3)\Pr(Y_2|Y_3) \]

  \[ \Pr(Y_1, Y_3) \neq \Pr(Y_1)\Pr(Y_3) \quad \Pr(Y_1, Y_3, | Y_2) \neq \Pr(Y_1|Y_2)\Pr(Y_3|Y_2) \]

  \[ \Pr(Y_2, Y_3) \neq \Pr(Y_2)\Pr(Y_3) \quad \Pr(Y_2, Y_3, | Y_1) \neq \Pr(Y_2|Y_1)\Pr(Y_3|Y_1) \]

- We represent the relationship between variables that cannot be made independent with an undirected edge between the variables and put no edge between variables that can be made conditionally independent:
Markov Random Fields II

- This suggests that, assuming an MRF generated our data, we could correctly infer the MRF by considering whether each pair of variables can be made conditionally independent when conditioning on a subset of other observed variables.

- For example, if we are considering four variables, 1. start with a “complete” graph (i.e. each pair of variables is connected by an undirected edge), 2. consider every possible pair-wise conditioning relationship, 3. remove any edge that can be made conditionally independent by any subset:

\[
\Pr(Y_1, Y_2, Y_3, Y_4)
\]

If any of the following are true:

\[
\begin{align*}
\Pr(Y_1, Y_2, | \emptyset) &= \Pr(Y_1 | \emptyset) \Pr(Y_2 | \emptyset) = \Pr(Y_1) \Pr(Y_2) \\
\Pr(Y_1, Y_2, | Y_3) &= \Pr(Y_1 | Y_3) \Pr(Y_2 | Y_3) \\
\Pr(Y_1, Y_2, | Y_4) &= \Pr(Y_1 | Y_4) \Pr(Y_2 | Y_4) \\
\Pr(Y_1, Y_2, | Y_3, Y_4) &= \Pr(Y_1 | Y_3, Y_4) \Pr(Y_2 | Y_3, Y_4)
\end{align*}
\]

Remove the edge between \(Y_1\) and \(Y_2\)

Repeat for all pairs...
Markov Random Fields III

- How might we apply this approach in practice? As an example of one approach, assume that our variables are normally distributed:
  - Data: consider the $m$ variables each measured for $n$ individuals
  - When consider a pair of variables $(Y_1, Y_2)$ regress each of these variables on a subset of the other variables (where these subsets do not include $Y_1$ and $Y_2$)
  - Take the “residual” of the $Y_1$ and $Y_2$ for each regression
  - Assess is residual variables are independent, if so, remove the edge and stop (!!)
  - Repeat for this approach for all subsets (not including $Y_1$ and $Y_2$) and if none make $(Y_1, Y_2)$ conditionally independent, leave the edge
- Repeat for all $(Y_i, Y_j)$ pairs

\[
Y_1 = \beta_0 + \sum_{i \not\in 1,2} Y_i \beta_i + \epsilon \quad Y_2 = \beta_0 + \sum_{i \not\in 1,2} Y_i \beta_i + \epsilon \quad Pr(Y_{1,r}, Y_{2,r}) = Pr(Y_{1,r})Pr(Y_{2,r}) \\
\hat{Y}_1 = \hat{\beta}_0 + \sum_{i \not\in 1,2} Y_i \hat{\beta}_i \quad \hat{Y}_2 = \hat{\beta}_0 + \sum_{i \not\in 1,2} Y_i \hat{\beta}_i \\
Y_{1,r} = Y_1 - \hat{Y}_1 \quad Y_{2,r} = Y_2 - \hat{Y}_2
\]
The reason we can use this type of approach is an MRF assumes specific restrictions on the “types” of conditional relationships that can exist, the most important of which is “global Markov property”.

We won’t go into more detail here but keep in mind not all PGMs have this property so they allow different types of conditional relationships, requiring different approaches to recover the model (= graph)!

Note that the “algorithm” on the previous slide is not optimal (but the simple structure provides intuition about how these algorithms work in general)

In practice, there are many algorithms / approaches used to recover MRFs but no matter what the approach remember the goal is to find the conditional independence relationships we have discussed (!!)

As one example, if all the variables are normally distributed (i.e. the MRF is a GGM) one approach is to take the inverse of the covariance matrix calculated for our variables and, for any value in the inverse matrix that we determine to not be different from zero, we remove an edge.

However even in this case, while the recovery algorithm looks different, the goal is to determine if (and which) pairs of variables are conditionally independent
• Say we have eQTL data (i.e. genotypes and gene expression variables measured for \( n \) individuals) and we recover an MRF that connects the genotypes to the expression variables AND the variables among each other, what is the value of this analysis?

• This is a tricky question, where the problem is two-fold: 1. Is an MRF a good model describing how the data was generated? 2. What conclusions can we draw about biology if two variables are connected by an edge in the MRF?

• The underlying problem is there are MANY reasons why two variables may be connected by an edge in an MRF (e.g. one variable regulates another, regulation by an unobserved variable, the relationship between them is “induced”, etc.)

• This makes it difficult to draw any strong conclusions from an MRF beyond there might be a relationship (although this doesn’t stop people from doing so!)
That’s it for today!