Probabilistic Modeling

Christopher Lee

Department of Chemistry and Biochemistry, UCLA

Announcements

- New reading assignment for today's material will be up on the CCLE site by tonight, and overview of 3 common likelihood models.
- Today we start to apply the basic probability concepts we've been reviewing to learning how to build probabilistic models (of bioinformatics problems).
- Mission training will be posted today, and also this week's programming project mission.

Discovering Single Nucleotide Polymorphisms from Short-Read Sequencing

- SNPs are the most common form of genetic variation, and thus one of the best sources of genetic markers for high-resolution genetic mapping.
- short-read sequencing is the most powerful, inexpensive way to sequence a new genome or population.
- SNPs can be discovered directly from short-read sequencing of a population, as sites where different individuals have single letter differences.
- But short-read sequencing is error-prone!

Your Mission: Accurate SNP Discovery Scoring

- you have been given a large set of short-read data identifying candidate sequence variations. However the vast majority of them are probably sequencing errors.

Your mission is to call which are real SNPs, by accurately estimating the odds for each candidate. It must robustly integrate many aspects of the evidence:

- the amount of evidence (more reads is better...)
- the strength of evidence (sequencing error variations etc.)
- allele frequency effects
- multiple reads per individual (advanced)

7 days to complete your Mission

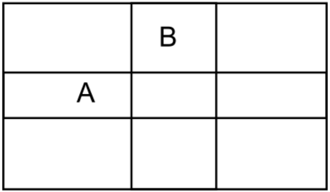
By:

- **Wednesday**: complete crash course on what you need to know to solve this
- **Friday**: complete your mathematical solution of this problem.
- **Monday**: complete your analysis of the short-read data, by programming your solution and running it on the full dataset.

Venn Diagram of Independence

Draw a Venn diagram representing two *independent* variables. To keep the diagram clear and simple, restrict each variable to just three states, and draw the diagram so *area* represents (i.e. is proportional to) probability.

Venn Diagram of Independence Answer

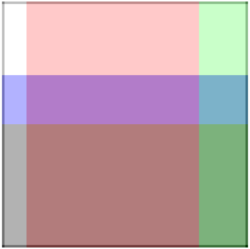


The X variable is shown as being split into three discrete values on the x-axis. The Y variable is shown as being split into three discrete values on the y-axis. The probability of a given X value or (X,Y) pair is shown by the area of that region in the figure. For example, the cell labeled A represents the joint probability of the intersection $(X = x_1, Y = y_2)$, and B represents $(X = x_2, Y = y_3)$. Since the probability (area) of any given (X,Y) pair is simply proportional to the product of the probabilities (area fractions) of that value of X and of that value of Y, they are independent.

Venn Diagram of Independence Errors

- Some people said the space should be split in two disjoint regions, one representing X (split into its three values), and the other representing Y (also split into three values). NO, each random variable covers the entire space (i.e. its three states sum to the entire space).
- "One variable should be a subset of the space covered by the other variable". NO, each random variable covers the entire space.
- Some people simply weren't sure how to translate $p(X, Y) = p(X)p(Y)$ into a graphical representation, implying that they weren't thinking of this product relation in terms of *orthogonality*.

Color Venn Diagram of Independence



- X={white, red, green}; Y={gray, blue, white2}
- key: any given Y slice (e.g. blue) has the same proportion of *every* X slice (e.g. red). This depicts $p(Y|X) = p(Y)$ graphically.
- Note: the actual widths / heights of the slices don't matter; in particular the widths need not be equal.
- The only thing that matters it that the two sets of slicings be *orthogonal* to each other.

Independence is about the entire joint distribution

Independence means "zero information" in the sense that X gives *absolutely no information* about Y.

- Even the slightest bias in just a part of the joint distribution would violate that definition, i.e. *not* strictly independent.
- Think of this like *correlation*, which you measure by plotting the entire dataset of X, Y values. If they show some correlation, they each give information about each other. Absolutely no correlation equals independence.

Learning How to Model Problems

- surprisingly, this is rarely taught as a subject in itself, yet it's one of the most important parts of real-world problem solving.
- "make everything as simple as possible, but no simpler" (Einstein) At least as a starting point, it is better to define a model that solves a core problem, and that you can actually understand, rather than piling on extra complications you haven't yet proved are needed. An 80-20 rule often applies here: 80% (or more) of cases can be modeled well by a core model that only deals with 20% of the full complexity.
- The flip side of this is: immediately test your model on real data to see if it is missing something actually essential for typical cases.

Modeling Steps

- define a *core problem* that captures the basic essence of what you are trying to solve.
- that means: choose assumptions that simplify the problem but still describe it relatively well.
- these assumptions correspond to choosing specific hidden parameters and associated probability models (e.g. uniform density assumption).
- while it may seem our *observables* are strictly pre-specified by the original dataset, often they can be considerably simplified: a *sufficient statistic* is a parameter that fully captures the information in a dataset that is relevant to a specific model. (e.g. for Poisson, just the observed count).

Putting the Model Together

- A model commonly has multiple pieces (variables and probability distributions) that link together in a specific structure.
- This structure is the essence of the model. It determines the algorithms we use to compute it; its computational complexity; and shows us the differences between alternative models.
- For probabilistic models this means the structure of how we compute the joint probability for our specific model of the problem.
- We can draw this very intuitively as *information graphs*.

The Conditional Probability Chain Rule

- The joint probability of a set of random variables can always be factored into the conditional probability of any (one or more) of those variables, given the remaining variables, times the unconditional probability of the latter.
- For random variables $X_1, X_2, ..., X_n$,

$$p(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} p(X_i | \vec{X}^{i-1})$$

Note that the conditioning can be performed in any order, due to the symmetry of set theory intersection, i.e.

$$p(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} p(X_{J_i} | X_{J_1}, X_{J_2}, ..., X_{J_{i-1}})$$

where J is any permutation of the index sequence $1, 2, ..., n$, and J_i is the i -th index in that permuted sequence.

Rules of the Road

- you can always write a joint probability in terms of the chain rule, in any order you like...

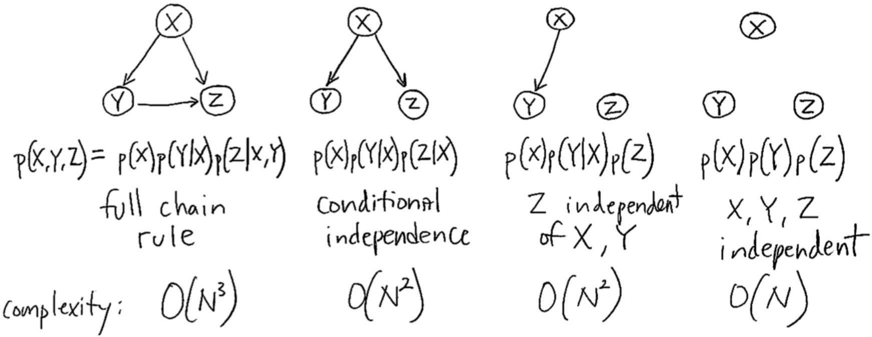
$$p(X_1, X_2, ...X_n) = p(X_1)p(X_2|X_1)...p(X_n|X_1, X_2, ...X_{n-1})$$

- you can always derive any marginal (unconditional, e.g $p(X_1, X_2)$) or conditional probability (e.g. $p(X_n|X_1)$) you like...
- but you cannot assume that some particular product of marginals and conditionals will correctly give you the joint probability, e.g. $p(X_1)p(X_2)...p(X_n)$. Only true if that *specific independence assertion* is actually true, across the whole joint probability function!

Graphical Models: "Information Graphs"

- gives a picture of a chain rule factoring of a joint probability distribution.
- nodes are the random variables in that joint distribution.
- edges are the conditional probability relations that appear in your chosen chain rule factoring.
- edges represent non-zero information links, i.e. where X is directly informative about Y i.e. $p(Y|X,\cdot) \neq p(Y|\cdot)$.
- They point from *condition* $\rightarrow$ *subject*.
- if the joint probability factoring can be simplified (due to independence) relative to the general chain rule, that should be reflected in the information graph as *missing edges* (some nodes are not directly connected).

Information Graphs for 3 Variables



Key principle: we generally seek models that *reduce* the complexity of the problem both to a *smaller number* of variables and *fewer connections* between them than the general chain rule.

Models Usually Propose a Causal Process

- One way to think about a model is as a *scientific theory*, i.e. a hypothesis about how the world works -- what causes what.
- arrows point from *causes* to *consequences*.
- Usually, from *hidden variable(s)* to *observable variable(s)*, i.e. "model" = *likelihood model* $p(O|H)$.
- Because O is observable, $p(O|H)$ is accessible to measurement, whereas $p(H|O)$ is not.
- Bayesian view: we *must* treat models as hypotheses (i.e. consider *every* possible model) and let Bayes' Law identify the "winner" (and tell us how confident we can be).

Writing Information Graphs As Text

Let's practice constructing a series of simple models.
Rules for entering your answers (standard `dot` format):

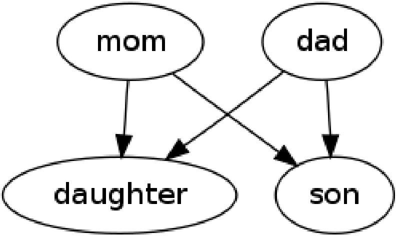
- write an undirected edge as `A -- B`
- write a directed edge as `A -> B`
- write variables as written in the problem, e.g. `X, f`
- spell Greek letters in Roman characters, e.g. `kappa`.
- write multiple edges separated by semi-colons, e.g. `A -> B; B -> C`

Modeling a Genetic Marker Over a Family

Consider a family consisting of a mom, dad, daughter and son, and a specific recessive mutation that causes a genetic disease. Let’s construct a model of the "genetic state" (genotype) of each person for this mutation. Specifically, draw an information graph for these four people, and briefly state the definition of your variables.

Modeling a Genetic Marker Over a Family Answer

mom -> daughter;
mom -> son;
dad -> daughter;
dad -> son;

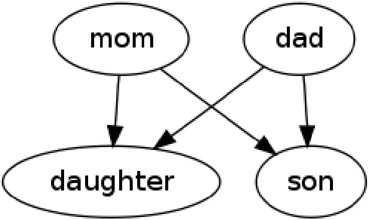


Each variable is the *copy number* of the mutation in that individual (i.e. 0, 1, or 2).

Info Graph vs. Equation

Now consider the relationship between the information graph for this family, versus the equation for their joint probability distribution.

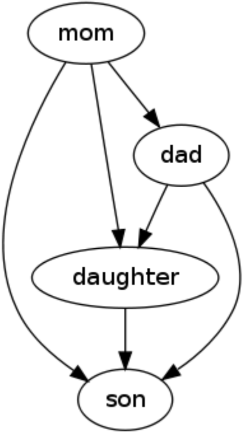
- write an equation for $p(mom, dad, daughter, son)$ that represents our pedigree information graph:



- draw an information graph representing the *general chain rule* for *mom, dad, daughter, son*.

Info Graph vs. Equation Answer

$$p(mom, dad, daughter, son) = p(mom)p(dad)$$
$$p(daughter|mom, dad)p(son|mom, dad)$$



NB: any order (of the four nodes) is equally valid

Info Graph vs. Equation Errors

- Some people wrote the same graph for the general chainrule as for the pedigree model, implying they didn't understand the general chainrule does not include *any* conditional independence.
- Some people expanded the joint probability as a product of terms with the same subject variable e.g. *p(mom,dad,daughter,son) = p(son)p(son|mom)p(son|mom,dad)p(son|daughter,mom,dad)*. No, each factor is for a different subject variable.
- Some people had edges going in both directions, e.g. mom -> dad; dad -> mom. No, an info graph contains no cycles (directed path from a node back to the same node).
- Some people wrote the chainrule as a linear graph e.g. mom -> dad -> daughter -> son. No, that implies many conditional independence assertions, contrary to the general chain rule.
- Some people wrote the genetics model as *p(daughter | mom)p(son | mom)p(daughter | dad)p(son | dad)*. No, two errors: 1. you cannot have a variable appear as the subject of two probabilities multiplied together (this is like counting the same

Info Graph vs. Equation Errors

- Some people wrote the genetics model as *p(daughter | mom)p(son | mom)p(daughter | dad)p(son | dad)*. No, two errors: 1. you cannot have a variable appear as the subject of two probabilities multiplied together (this is like counting the same person twice; it gives an invalid probability). What you really meant: $p(son|mom, dad) = f(son, mom)g(son, dad)$ a product of two separate factors for mom and dad. 2. But actually even that is not valid for the genetics model, e.g. no way to factor

$$p(son = 1|mom, dad) = \frac{mom}{2}\left(1 - \frac{dad}{2}\right) + \left(1 - \frac{mom}{2}\right)\frac{dad}{2}$$

$$= \frac{mom - mom(dad) + dad}{2}$$

Info Graph vs. Equation Errors

- Some people left out p(mom)p(dad) from the genetics model since they had no incoming edges. No, (of course) we still have to include their probabilities; the lack of incoming edges just means they have no *conditional dependencies*.
- Some people wrote their genetics model equation in terms like $p(son) = p(son|mom, dad)$. I think you meant that the factor for *son* in the joint probability should be written $p(son|mom, dad)$. But in the form you wrote, it means the *opposite* (i.e. it asserts that *son* is independent of both *mom* and *dad*, which is wrong).

Choosing Simplifying Assumptions

- Choosing a model really means choosing what simplifying assumptions are appropriate for the problem. "As simple as possible but no simpler".
- To first approximation, what variables *matter* for this problem both in terms of hidden causes and observable consequences. We try to reduce this down to the bare minimum that allows us to predict accurately what will happen.
- default: "assumed (conditionally) independent until proven coupled".
- Info graph states *explicitly* what's assumed independent vs. coupled.

"As Simple as Possible" = Seeking Conditional Independence

- by the simplicity criterion, the general chain rule is the *worst* possible model (maximal number of dependency edges, computational complexity).
- you must *think* about how the data are related, and especially about the underlying process that "emitted" them, to propose plausible ways to "cut" edges from the information graph representing the model.
- your proposals need to be reasonable (relative to what you know about the data and underlying process), but not *proved* -- you can only test the model later, against real data.
- Example: in genetics, each "inheritance event" is an independent random draw (i.e. the child randomly gets one of the parent’s two copies).

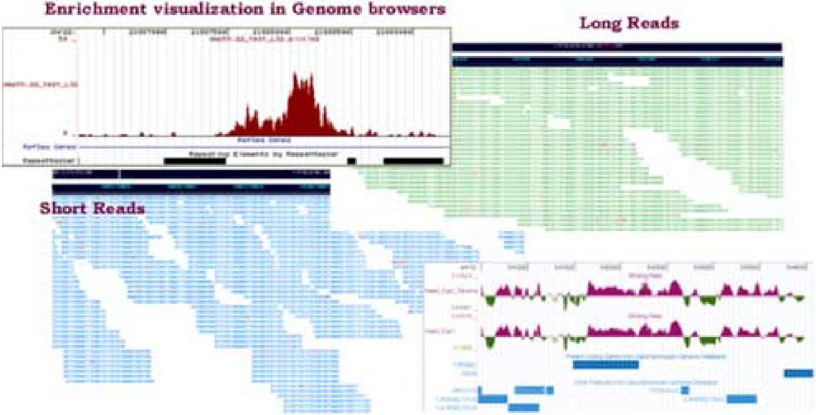
Sequence Variation

- *polymorphism*: Mutations in a population that have not yet been lost or "fixed" are called "polymorphisms", which simply means that some individuals have the mutation while others do not.
- *SNP*: A single nucleotide polymorphism (SNP) is a single letter in the DNA sequence that differs between individuals.
- *allele frequency*: the probability of finding a specific polymorphism on any given copy of the genome in a population.

The Key to Genomics Today: Short Read Sequencing

- *short read sequencing*: a recent technology for rapid sequencing of entire genomes. Current systems can read up to 500 million short sequence fragments each approximately 150 nucleotides in length, in one machine run. Due to the intrinsic error rates in the sequencing process, many fragment reads must be compared to arrive at a reliable "base call" at each letter position in the genome. Experimentally, DNA is randomly fragmented, and the resulting "short read" sequence strings must be aligned to each other (computationally, based on matching strings from overlapping reads) as the starting point for all analysis.

Short Read Sequencing



- *sequencing coverage*: the average number of independent reads covering any position in the genome (or region) being sequenced. Typically, sequencing projects aim for coverage of 70x or higher.

Basic SNP Scoring Model?

You are seeking to identify SNPs from short read sequencing of multiple people, pooled together (i.e. we have no information which person each read is from). Assume that you are scoring a single genomic position from the set of base-calls at that position from different reads, and that each base-call has a known error rate ε (i.e. occasionally the base-call is wrong).
Propose a simple model of the probability of these basecall observations, in terms of what hidden and observed variables, and their information graph.

Basic SNP Scoring Model? Answer

The simplest possible model for the probability of the observations would be to simply assume that the polymorphism b' occurs (or not) randomly and independently in each read DNA fragment. We can also adopt the simplest assumption that a sequencing error (incorrect basecall) occurs (or not) randomly and independently in each basecall. This implies the following variables:

- **population frequency of polymorphism b'** (θ, hidden)
- **presence vs. absence of b' in this DNA fragment** (κ, hidden)
- **observed base-call in this read** (B, observable)

The information graph is just $\theta \rightarrow \kappa \rightarrow B$.

Announcements

- CCLE Week 2 links to access the **online programming project** on Stepik.
- grad students have additional Mission Training to understand a more advanced scoring model, but do the same online programming project as undergrads.
- The **SNP project mission training** is due Friday at noon.

Basic SNP Scoring

You are seeking to identify novel SNPs from short read sequencing of multiple people, pooled together (i.e. we have no information which person each read is from). Assume that you are scoring a single genomic position from the set of base-calls at that position from different reads, and that each base-call has a known error rate ε. Based on the information graph $\theta \rightarrow \kappa \rightarrow B$, propose a simple likelihood equation for the probability of observing the SNP in one read i.e. $p(B = b' | \theta)$.

Basic SNP Scoring Answer

According to our information graph

$$p(B,\kappa|\theta) = p(\kappa|\theta)p(B|\kappa)$$

And we can eliminate κ by summing over its two possible values:

$$p(B|\theta) = p(\kappa = 1|\theta)p(B|\kappa = 1) + p(\kappa = 0|\theta)p(B|\kappa = 0)$$

Simplistically, the likelihood for κ is just binomial (SNP *b'* present with probability θ), and the likelihood of observing *B=b'* is just κ with a slight adjustment for the sequencing error rate e.g.

$$p(B = b'|\theta,\varepsilon) = \theta(1-\varepsilon) + (1-\theta)\varepsilon$$

Note that in this problem ε is a constant.

Basic SNP Scoring Errors

- Some people left out the SNP allele frequency θ; both ε, θ are critical for the likelihood of our observation.
- Some people did not explicitly state the likelihood model should be binomial, although this may have been implicit in their minds.
- Some people did not explicitly state that the basecalls in different reads are conditionally independent (or equivalently, that they follow a binomial likelihood model). Actually, most of the point of this question was to see whether you would explicitly state your conditional independence assumptions.
- Some people included a copy number variable for each person in the analysis. But since all the samples were pooled, and we have no way of knowing which read comes from which person, adding such variables doesn't help; only the population allele frequency matters.
- Some people diagrammed the *reverse* modeling process (i.e. the posterior probability of whether a SNP is present or not) rather than proposing a *likelihood model* (forward model) as the

SNP Scoring Models

As in the forensics test, we would like to measure the evidence for whether a SNP is present (or not) by comparing the likelihoods of the observations under two different models:

- *SNP*: a real SNP is present at this position.
- *no-SNP*: no SNP is present, i.e. any letter mismatch must be due purely to sequencing error.

Propose hidden variable(s) for these two models, and state precisely how the two models *differ*.

SNP Scoring Models Answer

- The *SNP* model has the population allele frequency θ as a hidden variable; e.g. we might assume an uninformative prior $p(\theta) = 1$ for all possible values in [0,1].
- By contrast, the *no-SNP* model asserts that $\theta = 0$, i.e. there is no actual polymorphism, so all letter mismatches must be due to the sequencing error ε.

SNP Scoring Models Errors

- Some people left out θ, a key hidden variable. The question defined the *SNP* simply as "a real SNP is present"; we certainly do not know the value of θ *a priori*.
- Some people thought of this as $\theta = 0$ (*no-SNP*) vs. $\theta = 1$ (*SNP*). That would be valid for a *single* observation (read), but here we have multiple observations sampled from the population.
- Some people talked about the observation of a "SNP" vs. a "miscall" as if they were *observably* different. No, in both cases we simply see a basecall that mismatches the reference genome. The only *difference* between the *SNP* vs. *no-SNP* cases is the expected *frequency* (approximately θ vs. ϵ).
- Some people asserted the *SNP* model can ignore sequencing error completely. No, even if there is a real SNP in the population, sequencing error can still occur in our sequencing machine. Why would the machine suddenly become perfect, just because some DNA out in the real world has a SNP? (how could the machine even know that?!)

Comparing Models using Odds Ratios

- Recall that in Bayes Law we compare different model states $\Gamma \in \{SNP, no - SNP\}$ via their *posterior probabilities*
- For comparing a pair of model states it's convenient to compute their *posterior odds ratio*:

$$\frac{p(\Gamma = SNP|X, S, F)}{p(\Gamma = no - SNP|X, S, F)} = \frac{p(X, S|\Gamma = SNP, F)p(\Gamma = SNP)}{p(X, S|\Gamma = no - SNP, F)p(\Gamma = no - SNP)}$$

- If we assume the priors are equal $p(\Gamma = SNP) = p(\Gamma = no - SNP)$, then the posterior odds ratio is just equal to the *likelihood odds ratio*:

$$\frac{p(\Gamma = SNP|X, S, F)}{p(\Gamma = no - SNP|X, S, F)} = \frac{p(X, S|\Gamma = SNP, F)}{p(X, S|\Gamma = no - SNP, F)}$$

A SNP Detection Problem

You are using a microarray to detect single nucleotide polymorphisms (SNPs) in samples from multiple people. For a given site you wish to compare two hypotheses: h_1, a SNP is present in the population (i.e. there is genetic variation at this site); h_0, no SNP is present. The microarray gives a fluorescence observation X for a given sample, and you are given likelihoods $p(X|\kappa)$ for the possible number of copies of the variant in that sample $\kappa = 0, 1, ...2N$, where N is the number of people pooled in that sample. You are also given the two probability models $p(\kappa|h_1, N), p(\kappa|h_0, N)$, and the prior ratio $p(h_1)/p(h_0)$. Given X measurements for multiple independent samples as your *obs*, can you compute the posterior odds ratio $p(h_1|obs)/p(h_0|obs)$? Briefly justify your answer.

1. Yes, the data are sufficient.
2. No, the κ values are unknown.
3. More information is needed in order to decide this.

A SNP Detection Problem Answer

Yes, just sum over all possible values of κ.

$$\frac{p(h_1|obs)}{p(h_0|obs)} = \frac{p(obs|h_1)p(h_1)}{p(obs|h_0)p(h_0)}$$

$$= \frac{\sum_\kappa p(obs,\kappa|h_1)p(h_1)}{\sum_\kappa p(obs,\kappa|h_0)p(h_0)}$$

$$= \frac{\sum_\kappa p(obs|\kappa)p(\kappa|h_1)p(h_1)}{\sum_\kappa p(obs|\kappa)p(\kappa|h_0)p(h_0)}$$

A SNP Detection Problem Errors

- Some people didn't consider that we can always eliminate a variable by summing over all its possible values (projection).

Modeling Multiple Observations

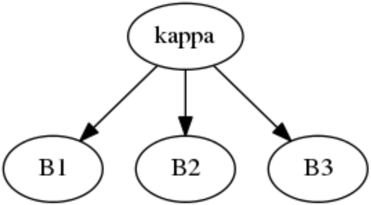
Say we have multiple short reads, giving us multiple basecalls B for a polymorphic position in a genome (you can write them $B_1, B_2, B_3, \ldots$). Say also that we represent the genetic state of a given individual person for that polymorphism (SNP) by a copy number variable κ. Now consider two different scenarios:

- All of the basecalls B are from a single person, i.e. single value of κ.
- Each basecall B is from a different person, i.e. different values of κ.

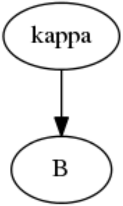
Do these two scenarios have the same information graph structure? If so, briefly explain why; if not, draw how they should differ. (For simplicity, let's exclude population allele frequency θ from our infograph).

Modeling Multiple Observations Answer

- in the first scenario we are drawing a *single* point $(\kappa, B_1, B_2, B_3, \ldots)$ from a joint probability $p(\kappa, B_1, B_2, B_3, \ldots)$



- in the second scenario we are drawing multiple points $(\kappa_1, B_1), (\kappa_2, B_2), (\kappa_3, B_3), \ldots$ from a joint probability $p(\kappa, B)$



Info Graph as a Random Draw From a Joint Probability

- $p(X, Y, Z, W)$ means *each* draw (datapoint) from the distribution is a vector of values (x, y, z, w), one value for *each* of our variables in the info graph.
- So if you want to show multiple variables C_i being drawn from the same value θ, you must explicitly draw that (as in the previous example).
- You can also represent this with the "one-to-many" arrow symbol.
- By contrast, the info graph $\theta \rightarrow C$ means each draw of C is drawn from a *different* value of θ!

Advanced SNP Scoring

You are seeking to identify novel SNPs from short read sequencing of multiple people. Each read has a unique tag that identifies the individual person that it comes from. Assume that you are scoring a single genomic position from the set of base-calls at that position from different reads, and that the probability that a given base-call is wrong is a known value ϵ.

Propose a set of hidden variables and associated likelihood models for modeling this problem.

Advanced SNP Scoring Answer

We are trying to find new SNPs so clearly we can't assume we know its frequency in the population beforehand: we must treat this as a hidden variable θ. We don't know whether the putative SNP is actually present in a given person, so we treat its copy number in that person as a hidden variable κ. Assuming each copy is chosen independently, this implies a binomial likelihood.

Finally, we observe some number of reads N, each of which has a basecall B for this position. Simplistically, the likelihood for B is just binomial (unmutated vs. mutated with probability $\kappa/2$), with a slight adjustment for the sequencing error rate e.g. $(1-\varepsilon)\frac{\kappa}{2}+\varepsilon(1-\frac{\kappa}{2})$. Note that in this problem ε is a constant.

Advanced SNP Scoring Errors

- Some people didn't see how knowing which person each read came from can help. If each person had only *one* read, it doesn't help. But if each person has multiple reads, it can help a lot, because the minimum frequency of a real SNP in one person is *high* (50%). Another way of saying this: real SNP observations should be strongly clustered (among reads from the same person), whereas sequencing errors should be randomly scattered over all reads, with no correlation vs. who they came from.
- Some people did not explicitly propose how the reads (obs) would be connected to an individual person (i.e. via a hidden variable representing that person's copy number for the SNP).
- Some people did not explicitly propose a likelihood model, as the question asked. The main thing when you build a model is to decide what conditional independence(s) you can assume.
- Some people claimed all reads from same person should report same basecall B_j. Implies still forgetting that human genome has

Variable vs. Constant

- A model cannot represent all the complexity in the universe; instead it confines itself to a small "subcase" consisting of just a few variables, while all other variables are treated as fixed (by the model's assumptions).
- e.g. in some cases we treat the allele frequency θ as a hidden variable to be estimated by inference; in other cases we assume we *know* its value.
- Constants only show up as *conditions* in your model's joint probability equation.
- Sometimes it's useful to explicitly represent a variable that our model treats as a constant. Mark it accordingly, e.g. $\theta = 0.2$, or write it in parentheses (θ).

Independent and Identically Distributed?

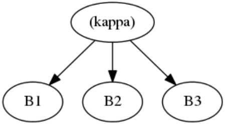
Say we have multiple short reads, giving us multiple basecalls B for a polymorphic position in a genome (you can write them $B_1, B_2, B_3, \ldots$). Say also that we represent the genetic state of a given individual person for that polymorphism (SNP) by a copy number variable κ. Now consider two different scenarios:

- All of the basecalls B are from a single person, i.e. single value of κ.
- Each basecall B is from a different person, i.e. different values of κ.

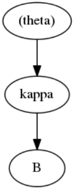
Do you consider the basecall observations B to be "independent and identically distributed" (I.I.D.)? If so, briefly explain why, in each of the two scenarios; if not, indicate the minimal assumption in each scenario that would render the basecalls I.I.D.

Independent and Identically Distributed? Answer

- in the first scenario $B_1, B_2, B_3, \ldots$ are conditionally independent given κ, so we can make them I.I.D. by explicitly marking κ as a constant (but otherwise not!):

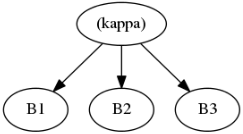


- in the second scenario $B_1, B_2, B_3, \ldots$ are I.I.D. as drawn in our original infograph (since the κ are independent). Optionally, we could make that assumption more explicit, by emphasizing that the κ are conditionally independent given θ:

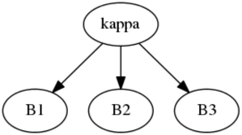


Conditionally Independent and Identically Distributed

Any two variables B_1, B_2 that people call "I.I.D." are drawn from some distribution (symbolized here by κ), and are independent only if κ is held **constant**.



If κ is not held constant, then B_1, B_2 can only be said to be **conditionally independent** given κ. Concretely, observing B_1 tells us information about κ, which in turn changes our probability distribution for B_2, i.e. "posterior likelihood" $p(B_2|B_1) \neq p(B_2)$.



Announcements

- This week: more in-depth exercises on building probabilistic models
- Next Monday: half hour quiz on building a small probabilistic model and assessing it vs. the evidence as you have already done.

Your Mission: Make Your First Solo Flight

- We have now trained on a variety of real-world model building problems, including taking your model all the way to a real computational implementation (SNP scoring).
- Now it's time for you to apply your skills to solving new problems by building your own models, on your own -- just like what you'll have to do in the real world.
- Friday: complete one more round of **mission training**, building a forensics model and getting immediate feedback.
- Monday: **quiz** where you'll be asked to build a model (using the same principles) and scored on your ability to use the concepts soundly.

Monday’s Quiz

- designed to be 30 minutes (but we’ll give you 45 minutes).
- Will give you quick feedback on how you’re doing relative to previous students.
- Your first solo flight building a probabilistic model "for real money". If you can do it here, you’ll be able to do it in the real world too.
- Bio glossary and equation sheet provided in the exam -- memorization is not our focus, but rather whether you can use the ideas to figure out problems.
- We will provide additional practice problems on Courselets (see CCLE Week 3).

# end of third lecture

Information Graphs: Why Bother?

- an alternate (and much more intuitive) representation of the *equation* for the joint probability. Use both!
- it can be translated directly to the equation and vice versa.
- if you can't draw the information graph, you don't understand your problem / your model.
- it shows you how the summation over the joint probability can be factored (separate branches factor).
- it shows you the computation complexity of your model at a glance.

Half Your DNAs Is Belong to Us!

A Great Business Opportunity

After watching Maury Povitch you decide paternity testing would be more profitable.

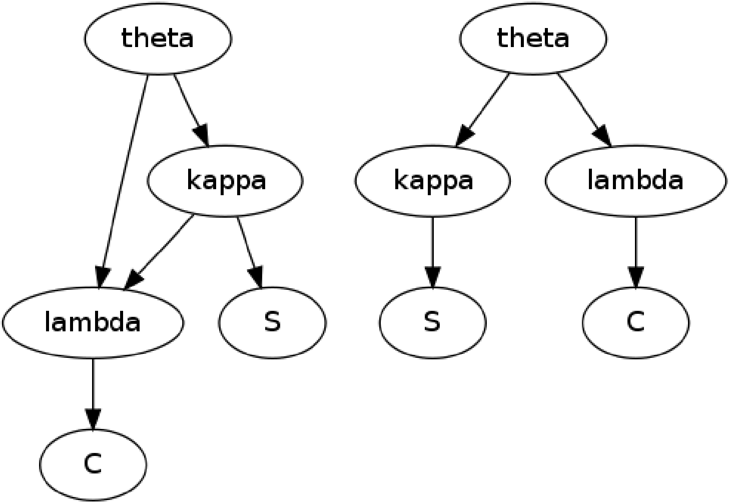
Paternity Test Info Graph

Say we have two microarray measurements for a given SNP marker: C, from a DNA sample from a child; S, from the "suspect" (man who may be the father). Propose two probabilistic models:

- *dad*, which asserts that the suspect is the father of the child;
- *not-dad*: they are unrelated.

In each case propose the simplest model that you think captures what's essential for these two hypotheses. Draw each model as an information graph structure, and briefly define any variable(s) you add, if any. Note that we assume no sample from the *mother* is included in this analysis.

Paternity Test Info Graph Answer



Paternity Test Info Graph Errors

- Some people wrote *dad* model as: $\theta \rightarrow S; \theta \rightarrow C$. (or inserted copy number variables between θ and each of the observables). No, that treats them as *unrelated*.
- Some people wrote *dad* model just like the forensic *match* model. No, *S* and *C* samples should not be from the *same person*, but instead as father and child, i.e. they need *separate* copy-number variables. It sounds like you simply *remembered* the *match* model, rather than thinking about what this question was asking.
- Some people left out κ, λ entirely. No, these variables are the crucial connections between *S*,*C* in the *dad* model.
- Some people proposed *dad* simply as "S and C are related in some way" rather than making a specific model of one being the father of the other. No, the question asked you to propose a *dad* model.
- Some people left out λ from the *dad* model, instead writing $S \rightarrow C$. This is wrong on several levels:
  - implies C, κ conditionally independent given S. That leads to

Why Do We Need Individual Copy Number Variables?

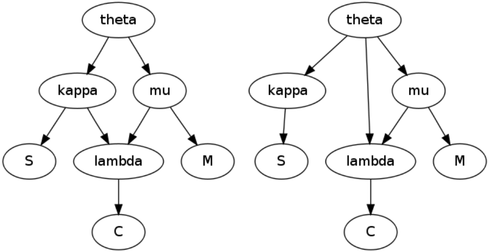
- remember: "as simple as possible, **but no simpler**"
- if omitting a hidden variable would **change the observation likelihoods** (versus including it), then getting rid of it is an **oversimplification**. That would change a correct model (i.e. correct likelihoods) into a wrong model (a different set of likelihoods).
- In particular, pay attention to **connectivity**. If a hidden variable adds *connections* between the observable variables, then getting rid of it is equivalent to forgetting about those connections. That will definitely change the observation likelihoods, i.e. make them **wrong**.
- Another way of saying this: including the copy number variables (κ_i) improves our ability to distinguish a real SNP from sequencing error, specifically if it adds such connections (i.e. we have *multiple reads from the same person*).

What About Mom?

Say we are given the opportunity to include a DNA sample from the child's mother, from which we obtain a microarray measurement *M*.

- draw information graphs for *dad* vs. *not-dad* models including *M*.
- based on these models, do you think *M* can improve our ability to reliably distinguish *dad* vs. *not-dad* cases? Try to give a concrete example supporting your argument.

What About Mom? Answer



Yes, including *M* can greatly improve discrimination (e.g. consider $S \approx M \approx 0, C \approx 0.5$). Concretely, about half the child's SNPs will come from mom (and most of these will not be found in dad). In the absence of the *M* information, these "genetic differences between C and S" will favor the *not-dad* model (by about two-fold, weakening our total evidence for the *dad* model). By contrast, with the *M* measurement, the model can see that these SNPs are just from the mom, and do not constitute evidence of "genetic differences versus S".

What About Mom? Error

Some people added μ in a way inconsistent with the rest of their model, e.g. $\mu \rightarrow C$, or $\theta \rightarrow C$ in the *dad* model. Genetics tells us mom and dad contribute to child's genotype in the same way, and are the only sources of the child's DNA. I.e. once you know κ and μ, θ cannot tell you anything more about λ.

What About Mom? Error

Some people left out μ entirely, plugging in directly to *M* instead. To my mind that is like *not* saying what the model is, by not specifying *how* (i.e. via *what* nodes) *M* connects to everything else. It's again inconsistent with both how we modeled dad, and basic genetics.

What About Mom? Error

Some people included mom in the *dad* model but not the *not-dad* model. No, we want to calculate the *same likelihood* under the two competing models, e.g. $p(S, C, M|\theta, dad)$ vs. $p(S, C, M|\theta, not - dad)$.

What About Mom? Error

Some people added edges directly between observables, e.g. $M \rightarrow C$. To my mind this implies not thinking about the causal process, but simply saying "M and C are somehow correlated". That is not a model! A model should specify *how* you think the observations are produced causally.

- In general for scientific problems our observables (measurements) are based on some hidden variable, so almost always edges to an observable should come from a hidden variable.
- A good model should stand up to challenging cases: e.g. say child's sample was first read by a reliable lab, and later mom's sample measurement was completely messed up by a sloppy lab. Can the fact that the M measurement is screwed up somehow travel back in time and space to affect the C measurement?
- Of course not. The only thing that should affect C (apart from the measurement process itself) is the child's DNA (here represented by λ).

What About Mom? Error

Some people asserted that the *M* sample cannot help "because mom and dad are conditionally independent" (given θ). Yes, they are... but what's that got to do with distinguishing *dad* vs. *not-dad*? I get the feeling you were just remembering the definition of "independence" as "X gives zero information about Y". Applying that to this problem, yes, *M* gives zero information about *S* (if we already know θ). But that's not what the question asked!

- *S* (observable) is not the *dad* vs. *not-dad* variable (hidden)!! Some people seemed to confuse these.
- **warning**: one of the most common and avoidable errors I see on exams is *not thinking about what the question asked*, e.g. instead *remembering* answer from a different question.

What About Mom? Error

Some people got the information graphs right, but asserted that *M* would not help distinguish them. Implies not thinking much about the connectivity of the variables. Since knowing *M* would tell us information about μ and hence λ, and λ is the key connection between *S* and *C* in the *dad* model, this strongly suggests *M* should help us distinguish *dad* vs. *not-dad*.

- Example: having unambiguous measures of *S,C,M* would actually create cases that are "impossible" under the *dad* model, e.g. $S \approx M \approx 0, C \approx 0.5$, which therefore become *powerful* evidence for *not-dad* -- but only if you have the *M* measurement!
- how often would this happen? If *S* is not the father, then about half of the child's SNPs will come from "mystery dad" (and are unlikely to be found in *S* or *M*). So about half the child's SNPs would fall into this "impossible" case.

Why Not Include Mystery Dad?

- say we propose to include a copy number variable ν for someone we don't have any sample observation for.
- at first glance, this still seems relevant to the variables we care about, e.g. the unknown father in the *not-dad* model will affect the child C.

Question: what is the probability ϕ of inheriting the SNP from "Mystery Dad"?

$$\phi = \sum_{\nu=0}^{2} \frac{\nu}{2} p(\nu|\theta) = (0)(1-\theta)^2 + \frac{1}{2} 2\theta(1-\theta) + (1)\theta^2$$

$$= \theta - \theta^2 + \theta^2 = \theta$$

Exactly the same as if we had drawn an edge directly from θ to the child, completely ignoring Mystery Dad!

Announcements

- This week's Mission Training (Forensics Testing) posted on CCLE Week 3, due Friday at noon. Come to discussion section with all your questions about anything that you're wondering about.
- All lecture slides (with solutions) will be updated on CCLE Week 3 today.
- Additional practice quiz problems will also be posted.
- Equation sheet and bio terms glossary (provided to you in the quiz) will be posted.
- Next Monday: half hour quiz on building a small probabilistic model and assessing it vs. the evidence as you have already done.

Principle: Ignore Irrelevant Variables

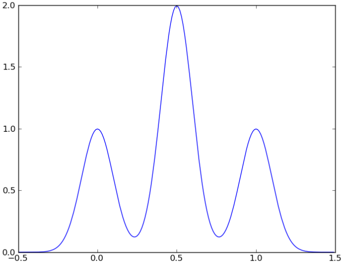
- I told you that the information graph for a model should be "fully reduced". But formally, what does that mean?

Principle: *if the inclusion of a specific hidden variable does not change the joint probability of the observations under your model, it is irrelevant to your model.*

- Example: $p(S, C, M|\theta)$ is the same regardless of whether we include Mystery Dad ν in the *not-dad* model or not.
- Example: what about the hidden variables for mother and child (μ, λ)? No, they are the key variables that connect the observations M, C. Removing them (and connecting M, C straight to θ) completely changes $p(S, C, M|\theta)$.

Principle: the "Single Peak" Likelihood Rule

Here's what $p(X|\theta)$ actually looks like for $\theta = 0.5, \sigma = 0.1$:



- It is not "one model", but *three models mixed together*.
- The three peaks reflect three distinct states of a hidden variable that this model is ignoring.

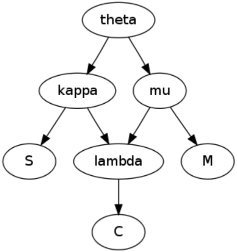
Principle: *one edge = one likelihood model = one peak*

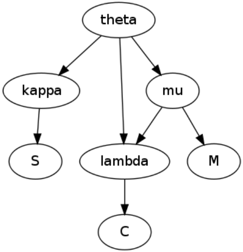
Another Example

- Do we need a hidden state λ for the child in the paternity test models?
- Yes, due to the single-peak rule: the child's observable C has multiple likelihood peaks (reflecting our uncertainty about the child's copy number).
- Trying to work out the likelihood for $p(C|\kappa,\mu)$ straight from the dad and mom variables κ,μ makes my head explode (too complicated!).
- But I can handle $p(\lambda|\kappa,\mu)$ and $p(C|\lambda)$ one at a time just fine.
- Breaking things down into pieces makes your life easier!

Paternity Test Factoring

Based on the information graphs for the *dad* model vs. the *not-dad* model, indicate how you could factor the joint probability $p(M, C, S|\theta)$, by writing the equation for this with appropriate parentheses to indicate which terms if any can be factored. Based on this, state the computational complexity for computing $p(M, C, S|\theta)$ in each of the models in big-O notation (assuming N possible states for each variable).





Paternity Test Factoring Answer

dad: $O(N^3)$

$$\sum_{\kappa} p(\kappa|\theta)p(S|\kappa)\sum_{\mu} p(\mu|\theta)p(M|\mu)\sum_{\lambda} p(\lambda|\kappa,\mu)p(C|\lambda)$$

not-dad: $O(N^2)$

$$\left(\sum_{\kappa} p(\kappa|\theta)p(S|\kappa)\right)\left(\sum_{\mu} p(\mu|\theta)p(M|\mu)\sum_{\lambda} p(\lambda|\mu,\theta)p(C|\lambda)\right)$$

Paternity Test Factoring Errors

- Some people only summed over κ, μ because they depended on θ. No, the value of λ is unknown too.
- Some people did not see that $p(\lambda|\kappa, \mu)$ couples together 3 summations, so they cannot be factored apart.
- Some people did not see that decoupling κ from λ reduces the *not-dad* complexity.
- Some people did not know how to translate the info graph to an equation.
- Some people did not include some of the hidden variables in the equation. But then how can it represent this info graph, or capture the complexity of the summation? E.g. someone wrote *dad* as $p(M|\theta)p(S|\theta)p(C|S, M, \theta)$ (OK, that assumes you've already eliminated all 3 hidden vars by summation), and then concluded complexity is *O(N)*. Conclusion: forgetting the hidden vars just leads to confusion, so why not just stick to the info graph?

This Week’s Mission: a DNA Forensics Test

- You are a forensics scientist who has been asked to score whether a crime scene DNA sample matches a suspect who has been arrested for the crime.
- You are provided two DNA samples, one from a crime scene (X), and another from a suspect (S), which are compared using a microarray that scores a large panel of SNPs by measuring the fraction of the sample that had each SNP.
- Your mission is to develop a scoring function that tells you the strength of evidence for whether the two DNA samples are a match or not!
- (There is no programming project for this week’s mission).

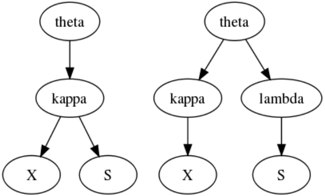
Forensic Test Causality

You are modeling a forensic test in which two samples, one from a crime scene (X), and another from a suspect (S), are compared using a microarray that scores a large panel of SNPs by measuring the fraction of the sample that had each SNP. Assume that you are given the frequency θ of each SNP in the general population.
Propose causal descriptions (i.e. what causes what) of two models for a single SNP: a *match* model that assumes both samples came from the same person; vs. a *mismatch* model that assumes they came from unrelated people. Define your variables and draw an information graph for each model.

Forensic Test Causality Answer

Clearly the microarray measurements approximate the copy number in the person(s) sampled. So we need to include that factor and the likelihoods of getting the same copy number in two unrelated people by random chance. This evidently depends on how frequent the SNP is in the general population. So we choose the following variables:

- θ: frequency of the SNP
- κ, λ: copy number of the SNP in a given person



Note that the model follows how we would actually calculate the probability: given θ we can calculate a likelihood for κ; given κ we can calculate a likelihood for X etc.

Forensic Test Causality Errors

- Some people drew *undirected* edges in their info graphs. But the edges represent *conditional probability* relations, which are inherently directional. Although we could conceivably expand the chain rule in any order we want, a model is a *specific* directed traversal representing how we think the data are produced.
- A common mistake was thinking in terms of *correlation* rather than a *causal model*. E.g. some people drew something like MATCH: $\theta \to X; \theta \to S; X \to S$. Two basic problems:
    1. *what is the causal model*? Do we believe that the microarray measurement X somehow travels through space and time to affect the S sample we analyze from the suspect? Such a model would violate not only basic genetics but also basic physics. I suspect you simply meant there should be *some connection* between them. But the model should *show* what we think connects them. If we're vague about what the connection is, our model is going to *suck*.
    2. *edges should represent likelihood models*: since the information graph shows the variables that matter for our problem, it should tell us everything we need to write down the joint probability of the

Forensic Test Causality Errors

- A common mistake was thinking in terms of *correlation* rather than a *causal model*. E.g. some people drew something like MATCH: $\theta \rightarrow X; \theta \rightarrow S; X \rightarrow S$. Two basic problems:
  1. *what is the causal model*? Do we believe that the microarray measurement X somehow travels through space and time to affect the S sample we analyze from the suspect? Such a model would violate not only basic genetics but also basic physics. I suspect you simply meant there should be *some connection* between them. But the model should *show* what we think connects them. If we're vague about what the connection is, our model is going to *suck*.
  2. *edges should represent likelihood models*: since the information graph shows the variables that matter for our problem, it should tell us everything we need to write down the joint probability of the observations. Specifically (each) edge should translate to a simple likelihood model representing how we think the data are produced. That's not true for *any* of these proposed edges, which tells us we've left out some key variables.

Confusion Over Correlation vs. Causation

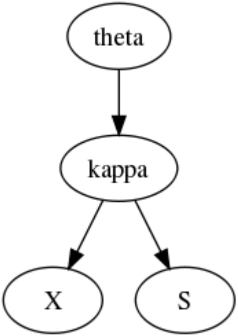
- the general chain rule and the general idea of conditional probability should be thought of in terms of *correlation*. Remember: the chain rule is inherently symmetric (non-directional); it lets you look at a set of variables in *any* direction you like and does not distinguish any one direction as better than any other direction.
- Since causality is by definition directional ("A causes B"), that is an *acausal* mindset, i.e. it ignores causality.
- Causality is a hidden variable, so we (have to) infer it from observable data, just like any other hidden variable!
- I.e. we propose different causal models and test them to see which best predicts the observable data.

The general chain rule is non-directional (acausal) but modeling is directional (causal), flowing from hidden to observable.

Forensic Test Causality Errors

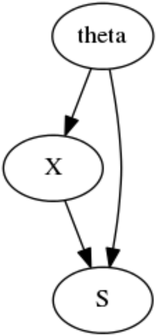
- Some people left out the hidden variable that connects X, S in the *match* model. When we say "the same person" we mean that both X, S are sampled from the same genetic state, which in this case is most easily described as the copy number of the SNP in that person (κ). Leaving this out implies that you are not thinking about the problem *causally*; you are not proposing a *specific model* for how X, S are connected in the *match* model. κ is crucial for the linkage of the observables, and thus is an essential part of the model.

What Happens to the information graph if we eliminate kappa?



Say we decide to eliminate κ (by summing over all its possible values). How would we draw the information graph for $p(X, S, \theta)$?

Info Graph Without Kappa



$$p(X, S|\theta) \neq p(X|\theta)p(S|\theta)$$

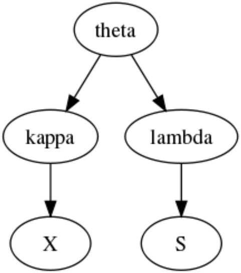
Not conditionally independent given θ! So we can't just draw $\theta \rightarrow X; \theta \rightarrow S$, implying they are conditionally independent. Strictly speaking we have to fall back to the general chain rule structure, to represent their lack of conditional independence.

What If a Hidden Variable Is Not Crucial for Linkage?

- Do we really need κ, λ in the *mismatch* model? e.g. we could eliminate κ:

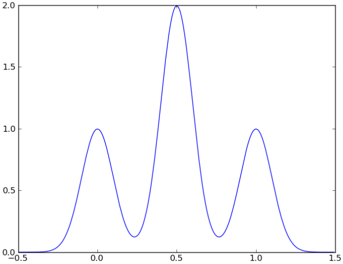
$$p(X|\theta) = \sum_{\kappa} p(X|\kappa)p(\kappa|\theta)$$

- Why not just write an edge $\theta \rightarrow X$ representing $p(X|\theta)$?
- Note that these variables *do not* connect together multiple observations. So are they irrelevant?



Principle: the "Single Peak" Likelihood Rule

Here's what $p(X|\theta)$ actually looks like for $\theta = 0.5, \sigma = 0.1$:



- It is not "one model", but *three models mixed together*.
- The three peaks reflect three distinct states of a hidden variable that this model is ignoring.

Principle: *one edge = one likelihood model = one peak*

Why Follow the Single Peak Rule?

- Keeps us honest by making us work with standard models we understand (normal, binomial etc.). A model built exclusively of such standard parts is completely unambiguous and "meaningful" in the sense that there is no question about "what the model means".
- By contrast, if we can force any arbitrary pdf we like into an edge, the "meaning of the model" is no longer visible in the information graph structure. Instead it has been "swept under the rug", hidden inside that edge.
- The person that this helps is *you* -- it helps you come up with the right model. For example, how likely is it that you're going to come up with the correct, properly mixed three-peak likelihood model $p(X|\theta)$, if you haven't even realized that there's a hidden variable κ that plays a role in this problem?

Forensic Test Causality Errors

- Some people proposed something like $X, S \rightarrow$ MATCH/MISMATCH, presumably meaning that we're going to use observations X,S to infer whether MATCH vs. MISMATCH is right. Of course, but how?? Using a model... but you didn't propose any model.

  Another way of saying this: drawing the information graph of the model is the *forward* direction of inference. The reverse direction (infering posterior probabilities of hidden variables) just means applying Bayes' Law to the (forward) models.

Forensic Test Factoring

Based on the information graphs for a. the *match* model; b. the *mismatch* model, indicate how you could factor the joint probability $p(X, S|\theta)$ (i.e. summing over all possible values of κ, λ to eliminate them from the info graph joint probability), by writing the equation for this with appropriate parentheses to indicate which terms if any can be factored outside of the required summations.

Forensic Test Factoring Answer

For the match model

$$p(X,S|\theta) = \sum_{\kappa} p(\kappa|\theta)p(X|\kappa)p(S|\kappa)$$

For the mismatch model

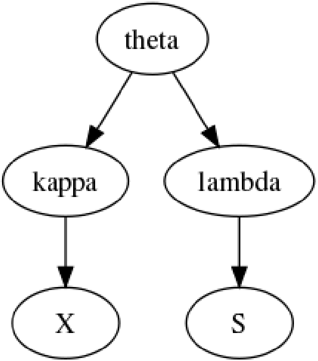
$$p(X,S|\theta) = \left(\sum_{\kappa} p(\kappa|\theta)p(X|\kappa)\right)\left(\sum_{\lambda} p(\lambda|\theta)p(S|\lambda)\right)$$

Forensic Test Factoring Errors

- Some people completely ignored κ, λ. Since the info graphs included κ, λ you needed to include them in your equation.
- Some people wrote the *mismatch* model as $p(X|\theta)p(S|\theta)$, but didn't show the required summations.
- Some people's equations left out edges from the info graph.
- Some people included $p(\kappa|\theta)$ *twice* in the *match* model ("I was thinking it had to look symmetric to the mismatch model"). No, each variable must appear as subject of exactly *one* probability factor, in any chain rule factoring.
- Some people forgot to sum over κ, λ, which is required to reduce these info graphs to yield $p(X, S|\theta)$.
- Some people wrote *match* as $p(X, S|\kappa)$.... This does not capture the conditional independence of *X,S* specified by the info graph.
- Some people left out θ entirely. No, you were asked about $p(X, S|\theta)$.
- Some people thought they could apply the summation to just $p(\kappa|\theta)$, eliminating it via normalization. No, other terms depend

Forensic Test Computation

Based on the information graph (factoring) of the *mismatch* model, write pseudocode for computing $p(X, S|\theta, mismatch)$ by summing over κ, λ, assuming you are given functions for computing each individual edge likelihood, e.g. for $p(X|\kappa)$, `calc_p_x_k(x, k)`.



Forensic Test Computation Answer

Pseudocode:

```
def calc_p_x_s_theta(x, s, theta):
    p = 0.
    for k in (0, 1, 2):
        p += calc_p_x_k(x, k) \
           * calc_p_k_theta(k, theta)
    p2 = 0.
    for l in (0, 1, 2):
        p2 += calc_p_x_k(s, l) \
            * calc_p_k_theta(l, theta)
    return p * p2
```

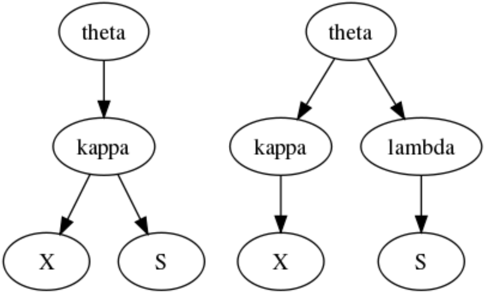
The two summations are calculated separately, exactly as shown in the factoring of the summation equation.

Forensic Test Computation Errors

- Some people nested the two loops, which makes the calculation slower (but will still give the right answer). No, they can be summed separately and then multiplied.

Forensic Test Complexity

What is the computational (time) complexity for computing $p(X = x, S = s|\Theta = \theta)$ in the *match* vs. *mismatch* models? For the sake of simplicity, write your answer in big-O notation, assuming that all variables have N possible states, and that each individual likelihood calculation (e.g. $p(X|\kappa)$) can be done in constant-time.



Forensic Test Complexity Answer

For the match model, we are summing over a single variable

$$p(X,S|\theta)=\sum_{\kappa}p(\kappa|\theta)p(X|\kappa)p(S|\kappa)$$

so the computational complexity is $O(N)$.
For the mismatch model we are summing over two variables, but as *separate factors*:

$$p(X,S|\theta)=\left(\sum_{\kappa}p(\kappa|\theta)p(X|\kappa)\right)\left(\sum_{\lambda}p(\lambda|\theta)p(S|\lambda)\right)$$

so the computational complexity is still $O(N)$.

Forensic Test Complexity Errors

- Some people called the *mismatch* model complexity $O(N^2)$ because they counted the number of "product terms". No, the two summations can be calculated as separate factors.
- Some people called the *mismatch* model complexity $O(N^2)$ because they thought the two separate $O(N)$ multiply to yield $O(N^2)$. No, they add to yield $O(N+N) = O(N)$.
- Some people called the *match* model complexity $O(N^3)$, implying they simply counted "maximum connection" (e.g. κ has three edges)? No, the problem defines X, S, θ as constant (i.e. no summations). Just think about the pseudocode for how to compute this -- how many loops and whether or not they are nested.

Study Suggestions

- Study together -- your grades are not in competition with each other.
- Talk about the things that are confusing you, until you get those concepts nailed down.
- Use the concept tests that confused you to show you where you need to clear up your understanding.
- Ask questions: in class; Friday's review session; office hours.

Check Your Assumptions

- I often see answers that show students didn't read the question fully, and didn't answer what the question was asking.
- Example: Basic SNP Scoring, "propose **likelihood models**"; but some people assumed this meant "draw an information graph", presumably because the previous question had asked about information graphs.
- we've defined "likelihood model" and "information graph" pretty carefully in this class, as core concepts that are essential "vocabulary".
- I always ask whether there are any words you're not sure about in what the question is asking you. Please ask -- please check your assumptions!

What to do when you don’t know where to start

- You’re looking at a question and don’t see a path to a solution.
- Your problem may be that you are trying to *remember* a solution, i.e. to see straight from the question to the answer, but it’s not ringing a bell.
- Is it really the case that you know *nothing* relevant to the question? Probably not. Start using your three parallel approaches (our language; pictures; equations) to see what helps you get clear about what exactly you’re being asked.
- This will typically lead somewhere -- either relevant steps that get you partial credit, or a path to a solution.

Example: multiple hidden variables

- Say X_1, X_2, X_3 are observed rolls of a dice, and $\Theta_1, \Theta_2, \Theta_3$ are hidden variables associated with each roll, e.g. Fair vs. Loaded dice.
- Show how you can compute $p(\Theta_2 = L|\vec{X})$ directly from the values you just computed. (given some info graph).

Clear Language can save your @$!

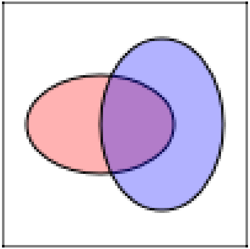
- Some people give an answer that summed $\sum_{\Theta_1,\Theta_3} p(\vec{X} = 266, \vec{\Theta})$ over the cases where $\Theta_2 = L$.
- OK, that's the right start, but we're not done yet:

$$p(\Theta_2 = L, \vec{X}) \neq p(\Theta_2 = L|\vec{X})$$

- What are the **words** for these things?
- What **picture** can we draw of these things?

Venn Diagrams of Conditional Probability

If you find yourself getting confused by distinctions like $p(\Theta_2 = L, \vec{X})$ vs. $p(\Theta_2 = L|\vec{X})$, please go back to Venn diagram picture defining these distinct probabilities ($\vec{X}$ = red, $\Theta_2 = L$ = blue):



$$p(\Theta_2 = L, \vec{X}) = \frac{purple}{white}$$

$$p(\Theta_2 = L|\vec{X}) = \frac{purple}{red} = \frac{purple}{white}\frac{white}{red} = \frac{p(\Theta_2 = L, \vec{X})}{p(\vec{X})}$$