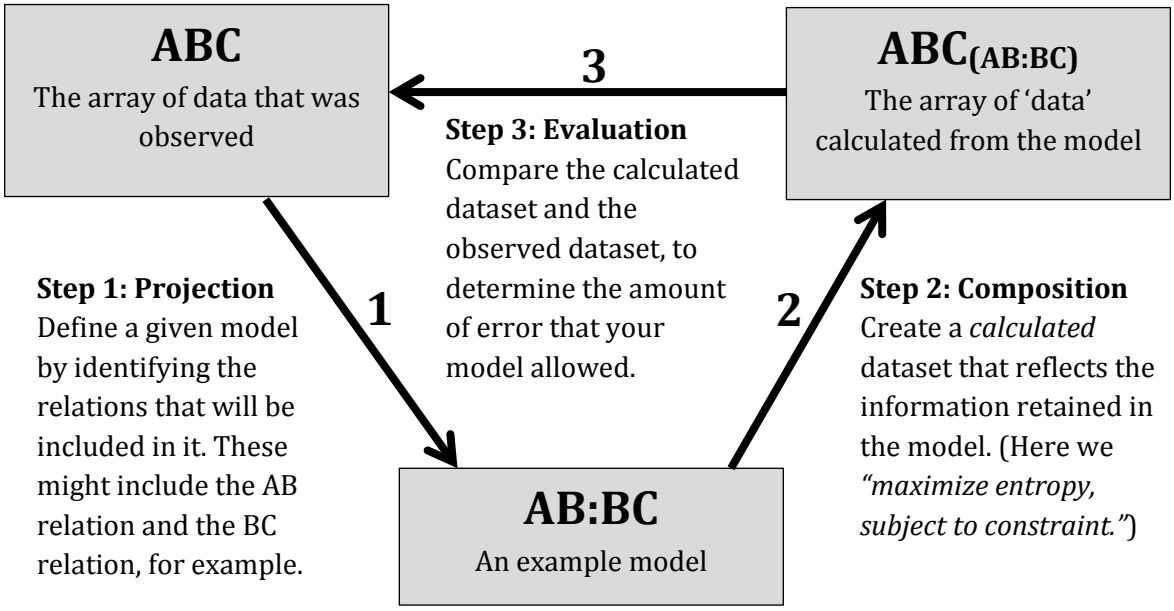
DISCRETE MULTIVARIATE MODELING: BASIC CONCEPTS

By Teresa Schmidt

In a Nutshell

At its core, Reconstructability Analysis (RA) is a data mining methodology that uses Information Theory and Graph Theory to detect deviations from mutual independence among a set of variables based on patterns in behavior.

The gist of RA is testing to see whether a relatively simple model can still capture the essence of a system. If, say, three variables are present in a system (A, B, and C), RA can help you to determine if it would be possible to "reconstruct" the whole dataset by paying attention to simple relations among the variables (like the AB relation and the BC relation). Since a model is always a simplified version of the data, a "good" model is one that doesn't lose too much information. The crux is a tradeoff: Create the simplest model possible, and also retain as much information as possible. To help you with this, RA has three steps you can conduct to test any model of a given dataset.



Why RA?

Reconstructability analysis is a member of the class of graphical models which also includes log-linear methods, Bayesian networks, and epsilon machines. It overlaps considerably with both, and where they overlap, they offer similar (if not identical) results.

However, RA has features that make it uniquely useful in many situations. For example, RA does not assume linear relations among variables, which makes it ideal for studying systems with complex nonlinearities. RA can also apply Set Theory (instead of Information Theory) for applications that are not amenable to statistical analysis. An example of this is a logical proposition where probabilities are not relevant.

1. Univariate Uncertainty (H), Diversity, & Information

Uncertainty, or entropy (abbreviated as H) is central in our evaluation of a "good" model. The error of a model is equal to the entropy of the model minus the entropy of the data.

$$Error_{model} = Entropy_{model} - Entropy_{data}$$

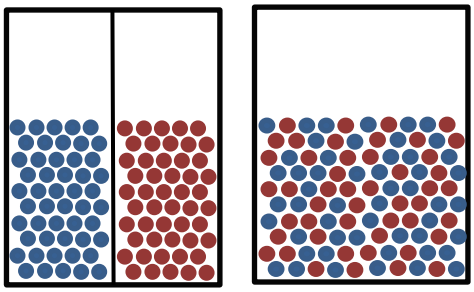
1.1. So What is Entropy?

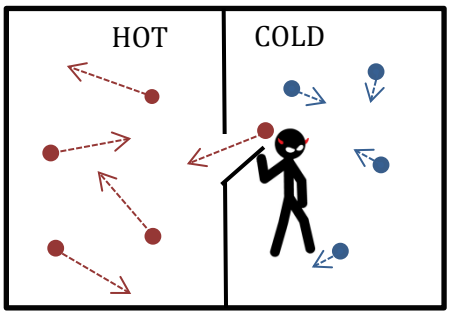
In the context of RA, Entropy (H) is a measure of univariate uncertainty. In other words, it's a way to quantify the average surprise or unpredictability of an outcome. Larger values of H always indicate greater uncertainty. The formalization of this measure started with Ralph Hartley (hence the *H*) in his work on Set Theory, and was augmented by Claude Shannon to involve Information Theory. When we use *H*, we mean Shannon entropy, because we're using Information Theory. Entropy (H) is a weighted measure of the probabilistic uncertainty of a set of outcomes.

1.1.1. An Aside: Entropy in Thermodynamics

The original notion of entropy comes from thermodynamics. The second law of thermodynamics suggests that entropy can only increase or stay constant in isolated systems; it can't decrease. This means energy will tend to go from heterogeneous to homogeneous. The temperature of your cup of coffee will consistently approach room temperature, no matter how many times you microwave it. This means the heterogeneity of temperature (heat of coffee vs. coolness of room) moves toward homogeneity (same temperature).

Why does this translate to uncertainty? Let's say you have a container with a divider in the middle, so half of it is filled with red marbles and half of it is filled with blue marbles. Your removal of the divider will allow the two to mix. Before, you might have had a lot of confidence about what color marble you would get if you picked one from the left side of the container. But now, since they are mixed, there is more entropy. You're not sure if you'll get red or blue.





Energy doesn't tend toward heterogeneity. Maxwell's Demon (pictured at left) is a thought experiment by James Maxwell where he imagined that a tiny door could possibly be used to increase entropy if someone (obviously with malicious intent) would let all the vectors with high velocity into one side, and all the vectors with low velocity into the other side. This would make one side get hotter, the other side cooler, and would decrease entropy. Without a demon, however, entropy is inclined to increase (or at least stay the same).

Note that Shannon and Hartley's entropy (H) are not restricted to energy transfer, and there is no rule in Information Theory that H will increase over time. For the purposes of RA, we will consider Entropy to be a measure of diversity: The fraction of something in one state vs. another state.

1.2. Entropy Equation

Below is the definition for the uncertainty of variable X:

$$H(x) = -\sum_{j=1}^{n} p(x_j) \log_2 p(x_j)$$

The entropy of a variable can be calculated from the probabilities of that variable having each possible state (e.g., probability that a coinflip results in heads or tails). You multiply the each probability by its log, and sum all the products together. This can be written more simply as

$$= -\sum p_j \log_2 p_j$$

We can break this equation down to understand what's inside of it. At the most basic level, we are interested in the probabilities of each state of a variable. If something has a 1 in 2 chance of happening (like heads v. tails), that is way less surprising than if something has a 1 in 100 chance of happening. We can express probabilities as fractions, such as $1/100$ or $1/2$.

Next, we need to take the inverse of these probabilities. Why? Let's compare the two fractions $1/100$ and $1/2$. When we just look at these, very surprising outcomes have very small numbers (.01), and less surprising outcomes have larger numbers (.50). For a measure of *uncertainty*, we want larger numbers to reflect more uncertainty. This will happen if we take 1 over the probability. Like $\frac{1}{\frac{1}{100}}$ or $\frac{1}{\frac{1}{2}}$.

OK, so what about the logs? Having $\log_2$ in the equation ensures that whenever we have a 50-50 chance of two states (say heads vs. tails), the uncertainty value will be 1. When H = 1, we can say that 1 bit of information is needed. Bits and Information will be discussed more later. But for now trust that we'll take the $\log_2$ of 1 over the probability of a given state: $\log_2 \frac{1}{p_j}$. And it turns out, by the miracle of algebra,[1] that this is equivalent to the $\log_2$ of that same probability. So we can rewrite $\log_2 \frac{1}{p_j}$ as: $\log_2 p_j$

Now, we are going to want to sum the uncertainty values of each state ($x_1, x_2, x_3, \ldots x_j$). But before we do that, we want to weigh each uncertainty factor by the probability of that state. Imagine that you have a loaded coin, so that heads appears 90% of the time. Even though the uncertainty of tails is really high, the overall uncertainty of the coin flip is low: it's usually heads. We want the probability of heads to be weighted more heavily than the probability of tails when we figure out the overall uncertainty in the coin flip. This means we'll take $p_j \log_2 p_j$ instead of using only $\log_2 p_j$.

Finally, we can sum all of the weighted uncertainty factors as a way to calculate the uncertainty (H) for our variable. We take the *negative* of this sum because logs of a fraction (like our probabilities) are always negative. Taking the negative of this sum gives us a positive value for H.

---

[1] the log of a fraction is equal to the log of the numerator minus the log of the denominator. For example, $\log_2 \frac{1}{1/2}$ can be solved as 1, or can be rewritten as $\log_2 1 - \log_2 \frac{1}{2}$, which equals $0 - (-1)$.

If you want to list the different probabilities to be logged, weighted, summed, and made negative, Zwick's shorthand is illustrated by:

$$= \Gamma(p1, p2, \dots)$$

With this shorthand, you can write $= \Gamma(.15, .32, .21 \ \dots)$, instead of writing $-\sum(.15 \log .15, .32 \log .32, .21 \ \log .21, \dots)$,

1.3. Nominal Variables

This is a good time to mention that standard RA uses only *nominal* variables, which means the order of states is assumed not to matter. A good example of a nominal variable is color, because there's no natural order in a set of colors. By contrast, an example of an ordinal variable would be height, where there is a natural order in the variables short, medium, and tall.

Here's why this is relevant now: If I have a bag filled with different colored marbles, I can calculate the uncertainty of my pulling a marble of any given color. If I were to calculate the uncertainty of my (randomly) picking a short, medium, or tall person, I do not have any regard for the natural order of those states. In RA, it is possible to use ordinal variables, but the calculations do not retain information about the order of states in those variables. It treats them as nominal.

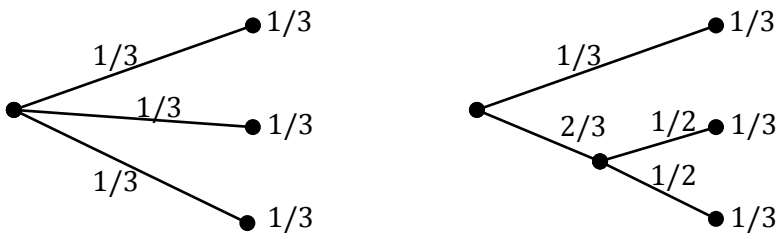
1.4. Factors influencing Uncertainty

Uncertainty, or entropy (H) is assumed to increase with

- A larger cardinality (i.e., the number of states or values in a set of *x*)
  In a bag filled with marbles, more colors means more uncertainty about what color you will randomly choose.

- With uniformity of probability
  There is more uncertainty in an even split between states than a lopsided split between states. There's less uncertainty in the result of a coin flip when using a loaded coin.

Entropy is also assumed to be decomposable. For example, uncertainty should be the same in both of the situations shown below:



The entropy on the left would be calculated as $\Gamma\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, which should be equal to the entropy of the situation on the right, calculated as $\Gamma\left(\frac{1}{3}, \frac{2}{3}\right) + \frac{2}{3}\Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$.

1.5. Uncertainty and Information

Information is the reduction of uncertainty. Our definition for uncertainty was:

$$H(x) = -\sum p(x_j) \log p(x_j)$$

So our definition for Information will be

$$I = -\Delta H$$

This can also be written as

$$I = -(H_{final} - H_{initial}), \text{ or}$$
$$I = H_{initial} - H_{final}$$

H_{final} will be 0 if we know the result with no residual uncertainty. This means that Information is equal to the total amount of initial uncertainty if it removes all uncertainty. Let's use coin flipping as an example. This is a binary variable, where each outcome has a probability of .5, or ½.

$$H_{initial} = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$

Now, we said that $Information = -\Delta H$

So if you tell me that the result was Heads, I have no residual uncertainty. I know the probability of it being Heads is 1, and the probability of it being Tails is 0. Hence, H_{final} will equal zero:

$$H_{final} = -1 \log 1 - 0 \log 0 = 0$$

Compare that equation with the $H_{initial}$ equation above. If we subtract H_{final} from $H_{initial}$, we know that the information provided was 1 unit.

$$I = H_{initial} - H_{final}$$
$$= 1 - 0$$

1.5.1. Bits as Units of Information

Units of Information are called Bits. A bit is the amount of information you could gain from the (truthful) answer of a yes-or-no question. For example, say you have a playing card, and I want to know what suit it is. I can first ask, "Is it red?" If you answer "Yes, it's red" I can ask whether it's a heart or a diamond. If you answer "No, it's not red" I can ask whether it's a club or a spade. Either way, I only need to ask two questions to know the card suit. So there are two bits of information necessary for me to remove all uncertainty.

2. Measures & Models

When we discuss measures (i.e., variables) and models, one notation is that lower case letters (x, y) represent variables and upper case letters XY represent models. Colons between upper case letters (e.g., X:Y) indicate independence between variables.

2.1. Two Variable Models

For two variables, there is only the saturated model XY (equivalent to the data), and an independence model X:Y. So in this situation you can only explore whether or not the two variables (x and y) are associated with each other. The independence model only includes the marginal probabilities of each variable. Say variable x is the season, and it's Winter (Sep-Feb) or Summer (Mar-Aug). The probability of each will be .5. Then say variable y is whether it's sunny or rainy, and the probability of that is also .5. Our independence model will be based on calculations from these marginal probabilities. We would guess rain and season are independent (model X:Y) and the arrangement of joint probabilities would look like this:

	y=0 (Sun)	y=1 (Rain)	
x=0 (Summer)	.25	.25	.5
x=1 (Winter)	.25	.25	.5
	.5	.5	

We could compare the above probabilities with data from the real world. If our independence model accurately matches real data, then we can conclude that the two variables are not associated. However, what if the real data might look more like this?

	y=0 (Sun)	y=1 (Rain)	
x=0 (Summer)	.33	.17	.5
x=1 (Winter)	.17	.33	.5
	.5	.5	

Here, we would say that there *is* a relationship between x and y. Our independence model did not match the observed probabilities, so we'd better go with the model XY. Model XY includes more information than X:Y. It means that the probabilities for each combination of variable states (e.g., summer + rain) cannot be accurately captured with only the marginal probabilities of each variable.

2.2. Three Variable Models

For three variables, more intermediate models are possible. The saturated model XYZ (the data), might contain a three-way association (XYZ), or the system might contain three two-way associations (XY:YZ:XZ), two two-way associations (XY:YZ), etc. The possible combinations are shown below, in what is known as a *lattice of structures*:

XYZ

XY:YZ:XZ

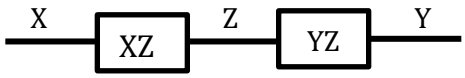
XY:YZ XY:XZ XZ:YZ

XY:Z XZ:Y YZ:X

X:Y:Z

Three-way and higher-way associations are difficult to depict with normal graph theory, because graphs typically use only dyadic relations. Another way to depict a model is to emphasize the relations as boxes, with lines representing the variables involved in them.

The model XZ:YZ can also be depicted as this:



We can calculate the entropy of the model XZ:YZ and compare it to the entropy of the saturated model XYZ (i.e., the data) to determine if this model accurately captures information or has error. This is just like the previous equation for H, only now we are doing it for a calculated distribution, q, which was generated from the relations we retained for our model, XY:YZ.

$$H(XY{:}\,YZ) = -\sum\sum\sum q_{XY:YZ}(x_j, y_k, z_l) \log q_{XY:YZ}(\ldots)$$

DMM notes by Teresa Schmidt

Choosing Models Statistically

1. Definitions

- **Type I error:** This is when I reject a null hypothesis and I shouldn't have.
    - Let's say two things are not different in reality (e.g., typing speed for men vs. women), but they *happen* to look different in my sample. If I reject the null hypothesis, claiming there *is* a difference in typing speed, I have done so incorrectly. This is a Type I error.

- **Type II error:** This is when I fail to reject a null hypothesis and should have.
    - Let's say two things really *are* different in reality (e.g., height for men vs. women), but it just so happens they don't look very different in my sample. If I don't reject the null, and say "we didn't find evidence of a height difference," this is a Type II error.

- **P-Value, or α (alpha):** This is the probability of making a Type I error.
    - Usually you want this to be small, because you don't want to go spouting off "I found a significant difference!" when it was just due to chance variations in your sample. You want to be confident that there's only a very small likelihood that this difference could have been caused by chance variations. When $p < .05$, it means there's less than a 5% chance that the difference you observed was due to chance alone.

- **Power, or β (beta):** This is your probability of being able to reject a false null hypothesis.
    - Alpha and beta are linearly related – the higher your statistical power, the better your chance at being able to reject a false null hypothesis (i.e., the more likely your *p* value will be less than .05).

- **"Good" Model:** Qualitatively speaking, a model is good if it captures a lot of the information in your data. Technically speaking, a model is good when its probability distribution (q) is really similar to the probability distribution (p) in your original dataset.
    - Let's say that you can exactly reproduce the values in the observed (p) probability distribution for AB by just knowing the marginal probabilities of A and B. That means that the calculated distribution (q) for the model A:B is a perfect fit for the probability distribution you observed. A:B captures all the information present in your data. Great!

- **"Good Enough" Model:** Usually the (q) distribution will not match your data perfectly, so when is it close enough? We'll need to test whether a model is significantly better than other models.

- **"Better" and "Worse" Models:** When we use statistical approaches to determine which models are better and worse, we need to know two things:
    - **Significance**, or is this model *statistically* different from the other model? (Is the difference unlikely to be due to chance alone?), and

- **Relative to what?** Our reference point and starting point might be either the independence model (bottom) or the data (top).

2. Basic Ideas

When the Independence Model is your Reference, Test if Models are Significantly Better

When the independence model is your reference, you test each model to see if it's significantly better (at replicating the p distribution) than the models below. You start at the bottom and work upward, and each time models are significantly different it means the higher model is significantly better.

Why go from the Bottom Up?

- This focuses on how complex of a model is justified by our data. It can tell me,
  - Which associations actually exist among these variables?
  - Are there simply 2-way associations among these variables?
  - Or more complex relations, such as 3-way and higher-way?
  - Can I be confident that this complex model is significantly better than a simpler model?

When the Data is your Reference, Test if Models are Significantly Worse

When the data is your reference, you test each model to see if it's significantly worse (at replicating the p distribution) than the models above it. You start at the top and work downward, and as soon as you find a significant difference it means that lower model is significantly worse.

Why go from the Top Down?

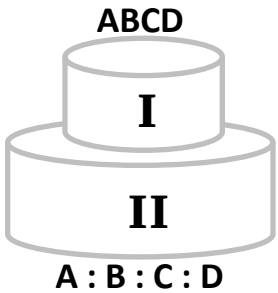
- This focuses on how simple a model can still decently capture the patterns in our data. It can tell me,
  - Do we really need a 4-way relation, ABCD, to capture the patterns we observed?
  - Would information about the nature of four 3-way relations, ABC:ABD:ACD:BCD, capture the patterns just as well?
  - Can I be confident that any simpler model would be a significantly worse representation of my data?

Using Cake to Understand Type I and Type II Errors

OK, look at the picture below. Imagine a 'Type I Error Zone' at the top of the lattice of structures, and a 'Type II Error Zone' at the bottom of the lattice of structures. (This is just symbolic, to help you remember.) The 'I' in "Type I" is smaller than the 'II' in "Type II," so you might imagine stacking the 'I' on top of the 'II' to keep it straight. Each time you cut the cake, pretend you only want to cut one layer. (Be polite.)

OK, so let's say the data is my reference. I'm starting from the top and working my way downward. I am going to see how far down I can go (how simple of a model I can get), but I want to make sure to stop before I get into the Type II Layer.

Alternatively, let's say the independence model is my reference. I'm starting from the bottom and working my way upward. I am going to see how far up I can go (seeing how complex of a model I can justify), but I want to make sure to stop before I get into the Type I Layer.

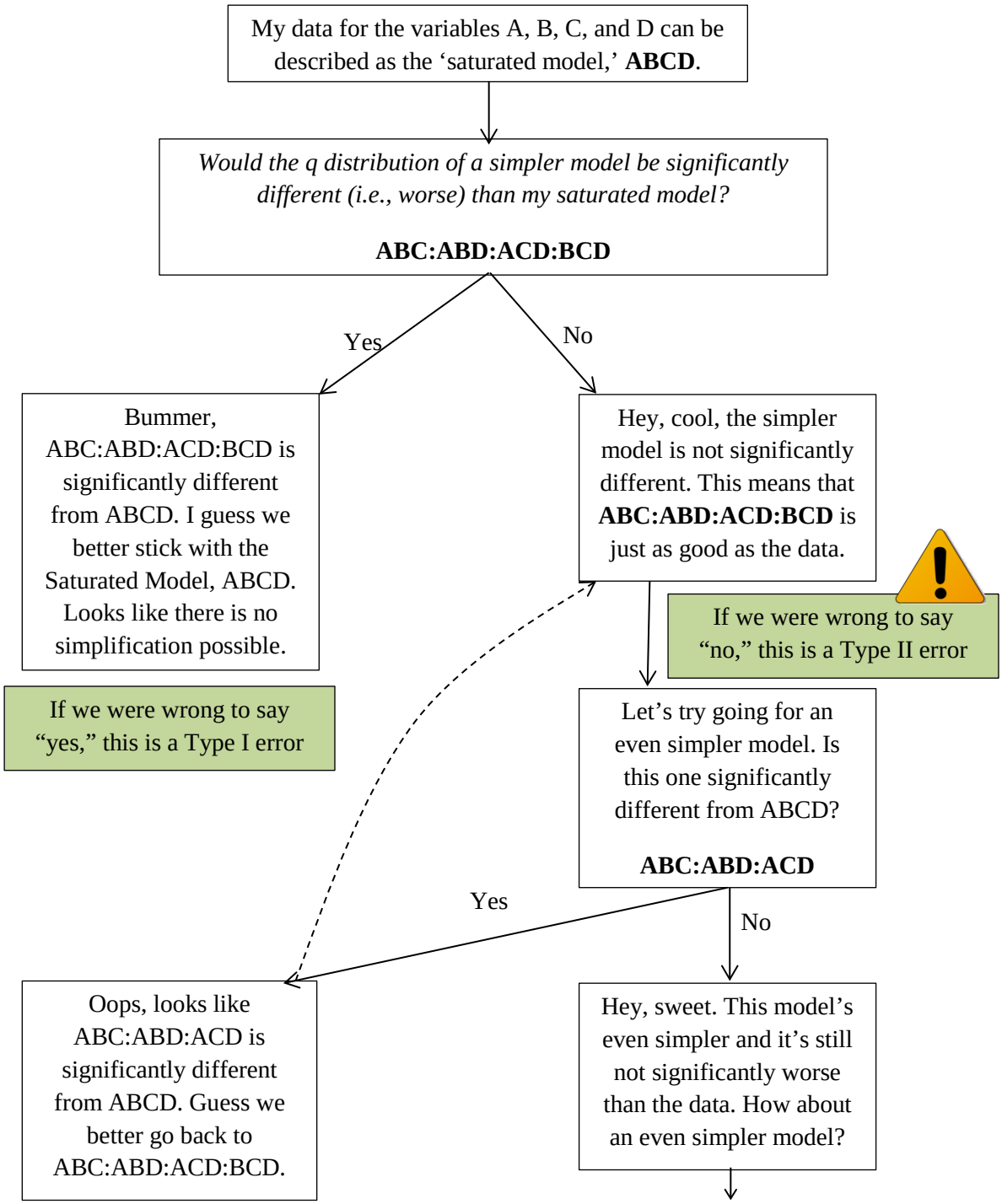


The principle is this: Whichever way you're going, you want to go as far as you can, but not too far. Going too far is like overstating findings that are not

warranted. And it's worse to overstate your findings than to understate. So when the bottom is your reference, don't go too far up (you'll get a Type I error, and be over fitting).
When the top is your reference, don't go too far down (you'll get a Type II error, and be over simplifying).

3. When the Reference is the Top

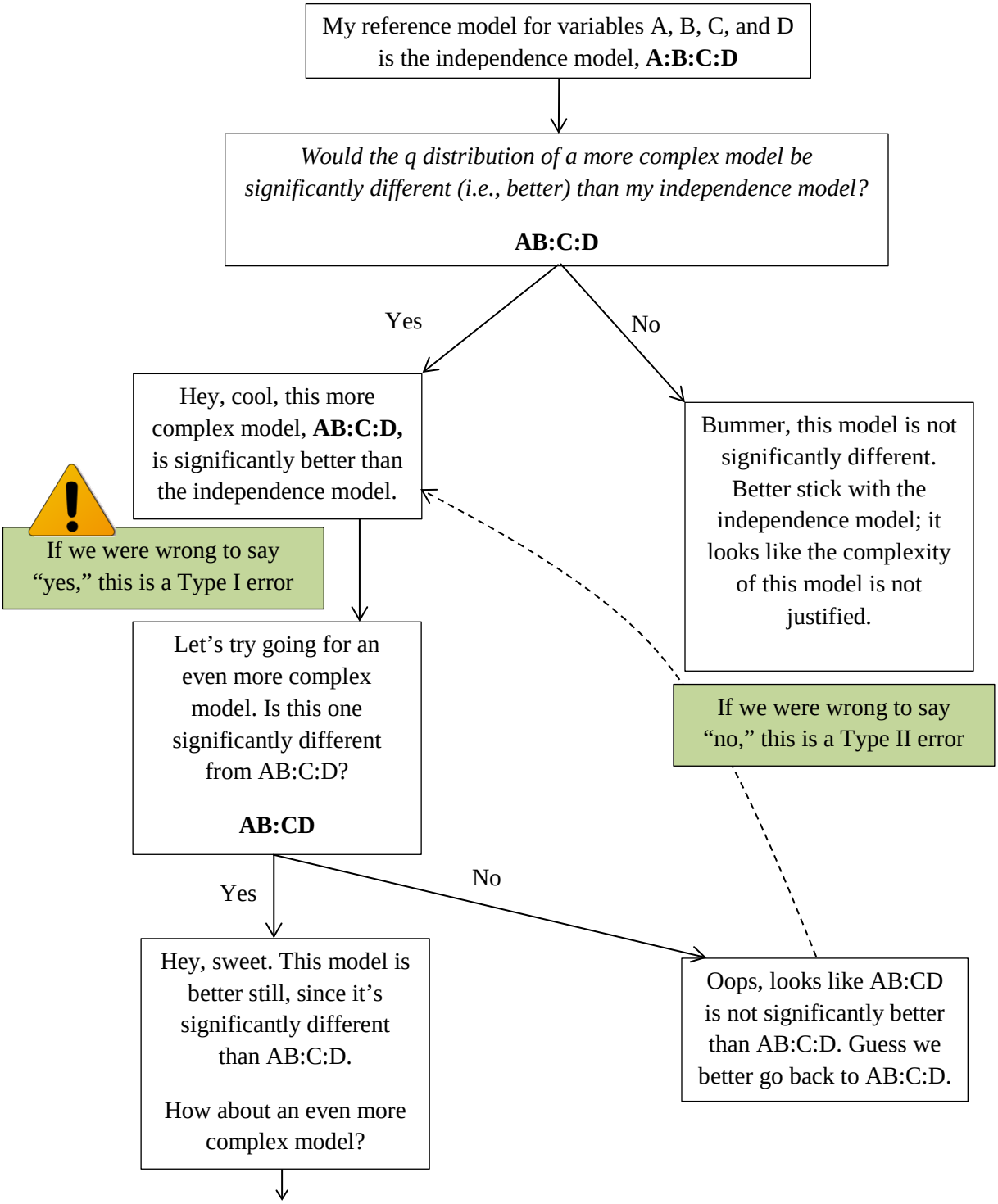
Here's a flow chart example for evaluating models when the reference is the top. Usually when the reference is the top, you work from the top down.



**Note that here, if our Type I error rate were really small ($p < .05$), we'd have to be really confident a model is significantly worse before we'd stop going down. We'll probably have Type II errors, which are very troublesome: We may be over confident that a simpler model is ''just as good.' To protect us from this, we should use a larger Type I error rate (like .3).*

4. When the Reference is the Bottom

Here's a flow chart example for evaluating models when the reference is the bottom. Usually when the reference is the bottom, you work from the bottom up.



Note that here, the Type I error rate is more intuitive, because we do want to be really confident that a model is significantly better before we keep going up. A small p value, such as $p < .05$, will keep us from being over confident that a complex model is justified.

Overall Patterns

Regardless of your reference model, rejection of the null always results in an "upward focus".

- If your reference is the top, rejecting the null means you will go back up to the previous level.
- If your reference is the bottom, rejecting the null means you will at least stay there, and maybe even try to move up another level.

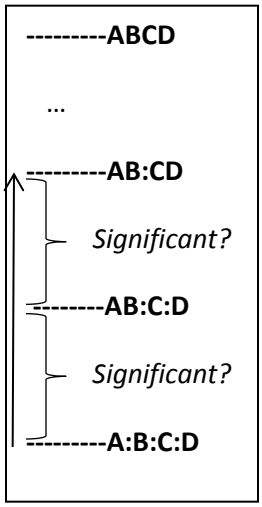
Also, regardless of your reference model, failure to reject the null results in a "downward focus".

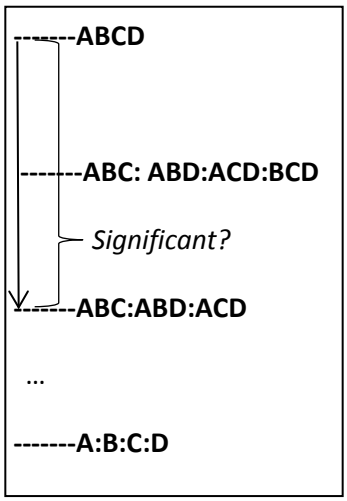
- If your reference is the top, failing to reject the null means that you will at least stay there, and maybe even try to move down another level.
- If your reference is the bottom, failing to reject the null means you will go back down to the previous level.

So remember: Rejection is upward (think of flipping the bird?), and non-rejection is downward.

Incremental Alpha, but not Beta

When you are going up the lattice, each additional model ought to be significantly different (i.e., significantly better) than the model below it. That is, if I go up from A:B:C:D to AB:C:D, and want to go up even further to AB:CD, I need to make sure that AB:CD is significantly better than AB:C:D (not only better than A:B:C:D). Why? Well, think of it this way: If the difference between A:B:C:D and AB:C:D is significant, then that significant difference will also be present in your test of whether A:B:C:D and AB:CD are significantly different. Finding a significant difference between A:B:C:D and AB:CD will be influenced (or "contaminated") by the significant difference between AB:C:D and A:B:C:D. Testing incrementally helps to "purify" your tests of significance, so you can be sure that each step up the lattice is incrementally significant (not just cumulatively significant). It helps protect you from committing a Type I error.





When going down the lattice, you actually want to compare each model with the data (rather than comparing it with the model directly above). The reason is this: We are more worried about Type II errors here, and they are less likely if we compare models that are further away from each other. Imagine you are climbing onto your roof. The step ladder is not significantly far from the ground, and your roof is not significantly higher from the stepladder. But falling off the roof onto the ground will be significant. In the same way, if I find that ABC:ABD:ACD:BCD is not significantly worse than ABCD, and that ABC:ABD:ACD is not significantly worse than ABC:ABD:ACD:BCD, it could still be the case that this lower model, ABC:ABD:ACD, *is* significantly worse than my data. I want to make sure I reject the null in this case, so that I won't commit a Type II error.