CEDAS-23-model-selection

Iain Johnston, University of Bergen

You are a statistical adventurer, keen to make your fortune discovering new science from existing datasets. You are armed with a device running R and access to this Github repo https://github.com/StochasticBiology/CEDAS-23-model-selection. A friendly, if rather scruffy, mage has advised you to download the `mombf` library before embarking on your quest. Talk to the friendly mage if this doesn't work.

How will you start your adventure?
Used `mombf` before? Turn to paragraph **4**.
Stats and R pro, but unfamiliar with `mombf`? Turn to paragraph **3**.
Got to here, and know what a linear model is? Turn to paragraph **2**.
Otherwise, let's begin our quest at paragraph **1**.

- 1 -

We'll be looking at linear models. A linear model looks like
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon$

y is the *response*, the dependent variable. The x_i are the *predictors*, the independent variables. ε is a random number drawn from a normal distribution with mean 0 and standard deviation σ – we usually write this distribution as $N(0, \sigma^2)$. β_0 is the intercept – the value y takes when all the x_i are zero. The β_i are the slopes associated with each predictor – how much a unit change x_i in affects y.

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$ is the deterministic "core" of the model – all the y values generated by this model would lie on a straight line. The addition of different (independent, identically distributed) values of ε for each y value capture random deviations from this deterministic behaviour. The *parameters* of a linear model are β_0, β_1, … and σ. The width of the noise distribution describes whether points are tightly clustered around a deterministic relationship or spread all over the place.

We are typically interested in the values of β_i. For model selection we are interested in which of them are nonzero (meaning that their predictor x_i has some relationship with y). In parallel, and in parameter inference, we are interested in the particular values they take.

NB – confusingly, a *linear model* does not mean that the *relationship* between x and y has to be linear. $y = \beta_0 + \beta_1 x^3 + \varepsilon$ is a *linear model* of a *cubic relationship*. The thing that makes the model linear is the dependence of y on the β_i parameters. $y = \exp(\beta_1 x) + \varepsilon$ is not a linear model (though we could easily transform it).

If your system of interest doesn't look like this, another modelling framework may be appropriate, and may also admit model selection approaches. Generalised linear models (GLMs), for example, generalize the form of y, so that we can deal with binary, ordinal, etc responses (logistic regression is a common example). The friendly mage would be delighted to unpack any of these ideas further; give him a shout (and consider coming to STAT200). Or, onwards to paragraph **2**!

We can fit a linear model in R using `lm`:

```
my.lm = lm(y ~ x)             # if y and x are vectors (x could be a matrix
                              where each column is a different predictor)
my.lm = lm(y ~ x, data=df)    # if df is a data frame with columns y and x
```

`summary(my.lm)` will then tell us information about parameter estimates, hypothesis tests about these parameters (usually, are they nonzero?), the residuals (what's left over after the model fit, useful for diagnostics), and goodness of fit (measured via R^2, the coefficient of determination).

Remember AIC? We can immediately get this:

```
AIC(my.lm)
```

But we'll also be looking at Bayesian pictures of linear models, for model selection. We'll use the `mombf` package, and the command

```
modelSelection(y, x, …)
```

We'll look at the later arguments later, but `modelSelection` wants y to be a vector and x to be a matrix, where rows are observations and columns are predictors. The examples below should help clarify this.

Somewhat awkwardly, some of the different approaches we will use will require their arguments in different formats. We can read in data from a CSV (comma-separated value) file into an R "dataframe" with

```
df = read.csv("file.csv")
```

A dataframe read in this way will have columns named by the first row in the file, which are referenced with $ to obtain vectors. Rows and columns can also be referenced numerically.

```
df$x
df[2,1]
df$x[2]
```

Matrices and vectors also exist in R; we'll meet them in the next step. Let's press on to **3**.

- 3 -

Bayesian model selection vignette
Look at code *mombf-vignette.R*.

a. Write down in words what synthetic data has been constructed.

b. The uncommented l.14-15 give some default choices of prior distribution for parameters and model index. Run the code with these. How can we interpret the output from l.31-32?

c. Explore the effect of uncommenting some of the alternative prior choices, including the output on l.36. Do the effects match your intuition from Bayes' theorem?

> The `momprior` prior is a somewhat complicated, but useful, concept. When choosing priors we need to decide a probability distribution and its parameters. The parameters of a prior in Bayesian analysis are referred to as *hyperparameters*. `momprior` takes an argument tau which determines the scale of our prior, then finds the hyperparameters of a prior (based on a normal distribution) that produce that scale. The priors implemented in the `mombf` package are so-called "non-local" priors, which avoid some issues when we have nested models (ie when one model structure is a subset of another, as is often the case). You can read more (quite technical!) information here https://cran.r-project.org/web/packages/mombf/vignettes/mombf.pdf ; but for us, `momprior` can serve as a useful out-of-the-box prior in model selection analysis (but we should always explore the effects of different prior structures and hyperparameters!).

d. What does a frequentist model selection approach tell us for this example? What can it not tell us that the Bayesian approach can?

Next, let's look at an example where identified model structures can inform us about what's scientifically going on. Turn to **4**!

- 4 -

Identifying biochemical dynamics from data
We have a chemical system where a metabolite X undergoes reactions that produce other metabolites. We suspect that reactions

X -> Y
X+X+X+X -> 2Y

might exist. Others might too.

The "law of mass action" says that reactants are produced at a rate that is

Rate = (a rate constant) × (the product of reactant concentrations).

So the first reaction would produce Y at a rate k_1 [X]; the second would produce Y at a rate 2 k_2 $[X]^4$. (The square brackets just mean "concentration of").

chem-data.csv contains samples from experiments where the rate of Y production is measured at different values of [X]. Different powers of [X]: $[X]^2$, $[X]^3$, $[X]^4$ are also given, in columns called c (concentration), c2, c3, c4.

a. What does a frequentist approach suggest about this system's reactions and rate constants?

b. What does Bayesian model selection suggest about this system's reactions and rate constants? You could consider using the code from *mombf-vignette.R*, for example.

Next, an example from research! Make sure your adventuring pack is adequately stocked and turn to **5**.

- 5 -

Discovering evolutionary mechanisms from data

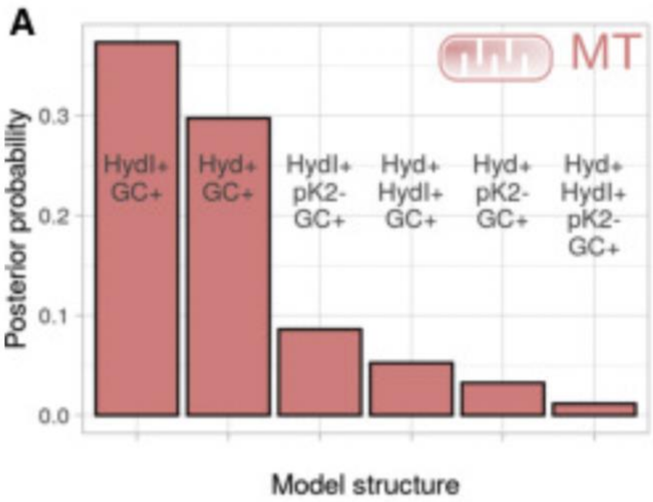
This example is taken directly from a paper published last year [1]. Mitochondria – the power plants of our cells – contain their own DNA, which encodes some vital genes. In different species, mitochondria contain different sets of genes, and it's an open question in biology why some genes are kept in mitochondrial DNA in more species than others.

mt-data.csv contains data on mitochondrial genes, labelled by "GeneLabel". Their retention across species is quantified in two ways: "Occurrence", a raw count of species possessing that gene (subject to substantial sampling bias) and "Index", a retention index controlling for sampling bias. The other, subsequent, columns store information corresponding to different hypotheses about why genes may be more or less retained.

> The hypothesis-linked stats are Length (gene length; are longer products harder to import to the organelle?); Hydro and Hydro_I (hydrophobicity measured in two different ways; are hydrophobic products harder to import?); MW (molecular weight; are bigger products harder to import?); pKa1 and pKa2 (amine and carboxyl pKa; does acidity of environment influence importability?); A_Glu and CW (two measures of the energy required to produce a product; is energy economics important?) and GC (GC content; does nucleic acid base balance influence retention?). We have looked at several more but keep this set for simplicity and generality.

a. What do you think? What features predict mtDNA gene retention?

b. Can you reproduce Fig. 2A in [1]? (Maybe without the top-right inset)



c. *pt-data.csv* contains similar data for chloroplast genes (pt stands for plastid, the more general – but less well-known – term for these organelles). What features predict chloroplast gene retention?

d. How might you go about testing the predictions of such models?

e. Can we predict mitochondrial gene retention from chloroplast data, and/or vice versa?

[1] Giannakis, K., Arrowsmith, S.J., Richards, L., Gasparini, S., Chustecki, J.M., Røyrvik, E.C. and Johnston, I.G., 2022. *Evolutionary inference across eukaryotes identifies universal features shaping organelle gene retention*. Cell Systems, 13(11), pp.874-884. https://www.cell.com/cell-systems/fulltext/S2405-4712(22)00351-9

We've learned about the world using data and model selection! We now have a choice. We can either see if we can use our understanding to help other, future adventurers (turn to **6**). Or we can celebrate our accomplishments with a flagon of ale (turn to **7**).

- 6 -

Teach us!

That's some Bayesian analysis with priors and posteriors, but we shouldn't neglect frequentist approaches (and their Bayesian analogues). Create an example, involving commented R code, demonstrating model selection via information criteria:
i. Obtaining an existing, or producing a new synthetic, dataset where some predictors are more tightly linked to the response than others;
ii. Evaluating an appropriate statistic for model quality for a single model;
iii. Comparing such a statistic across the range of possible models in your example;

- 7 -

Tired but happy, you return through the storm to the local tavern having discovered the arcane secrets of model selection. Now, faced with any dataset, you can make probabilistic statements about sets of hypothesized model structures – doing science with data! Remembering to be ever vigilant about the influence of subjectivity in your priors and violations of your modelling assumptions, you order a flagon of ale to celebrate and start musing about your next adventure.