An Introduction to Model Uncertainty and Averaging for Categorical Data Analysis

Chris Franck

June 21, 2022

What this course is

- A gentle introduction to concepts used in Bayesian model selection and averaging. Approximate methods based on BIC.
- A good starting place to begin moving towards fully Bayesian methods.
- An overview of concepts, and an open forum for questions and discussion.
- I illustrate the approach with an R demo.

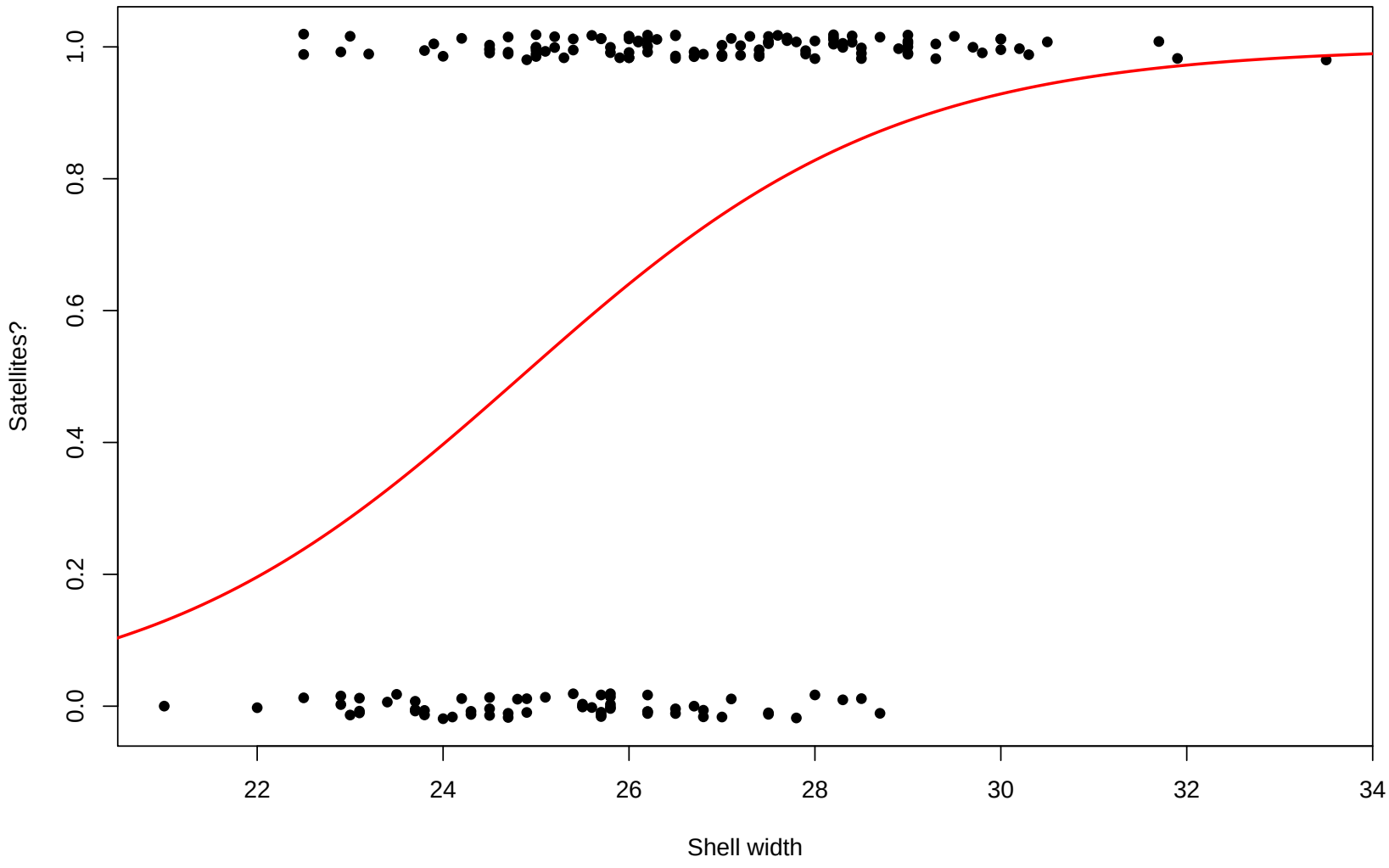
What this course is not

- A fully-described Bayesian approach to model averaging for categorical data. This would require specialized knowledge in Bayesian modeling, sampling algorithms and other computational techniques.
- I will present formulas and an R demo for the BIC approximation I describe here.
- I will also how to use the BAS package (M. Clyde 2022) for fully Bayesian analysis on the same data. Not enough time to fully describe the machinery behind this package. This is a technician's view.
- References for fully Bayesian approach: (Liang et al. 2008) (Li and Clyde 2018) (M. A. Clyde, Ghosh, and Littman 2011).

Today's lecture has three parts

1. Quick review of logistic regression
2. Bayesian model selection and averaging, approximate BIC approach + R demo.
3. Other approaches to weighing models, i.e., stacking (brief)

Part 1: Review of logistic regression.



Consider the crab satellite data

Figure 1: A horseshoe crab

Overview of crab data

- Data obtained from Alan Agresti's *Introduction to Categorical Data Analysis* book (Agresti 2018).
- The data measure whether or not a female horseshoe crab has additional male suitors beyond their current mate, essentially. The outcome is *binary*.
- There are four *candidate predictors* including: shell width, weight, shell color, spine condition.

The multiple logistic regression model

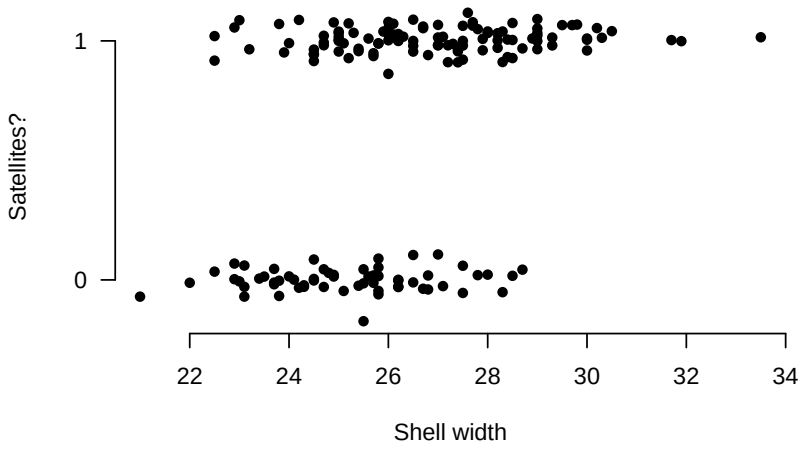
$$\text{logit}[P(Y=1)] = log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

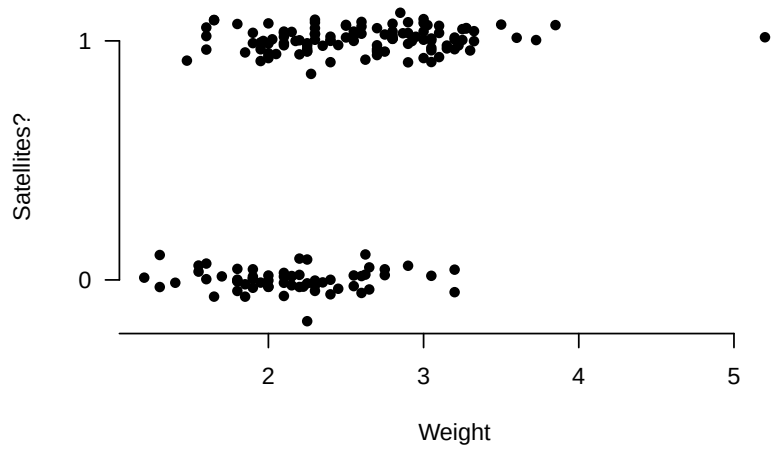
- Y is a binary outcome random variable.
- β_0 is the y-intercept, $\beta_1, \ldots, \beta_p$ are coefficients. The β terms are unknown parameters.
- The values $x_1, \ldots, x_p$ are candidate predictor variables that are assumed fixed.
- $\pi(x)$ is the probability of an event as a fucntion of predictors. For notational brevity, recognize that $\pi(x)$ depends on all of the x_j values, $j = 1, \ldots, p$.

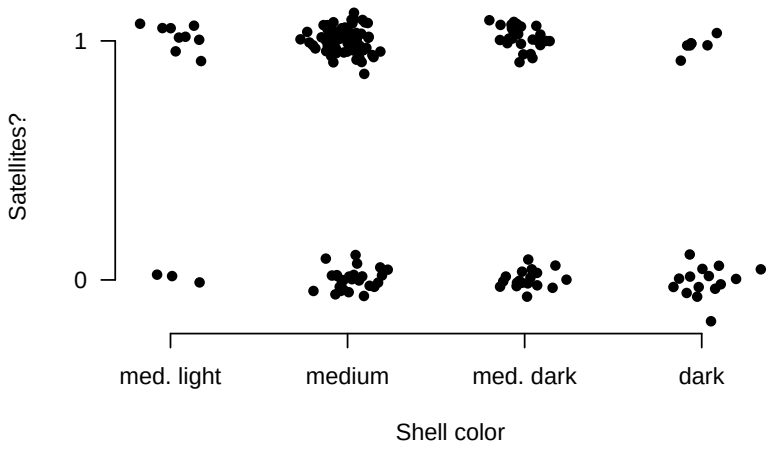
The data contain four potential predictors

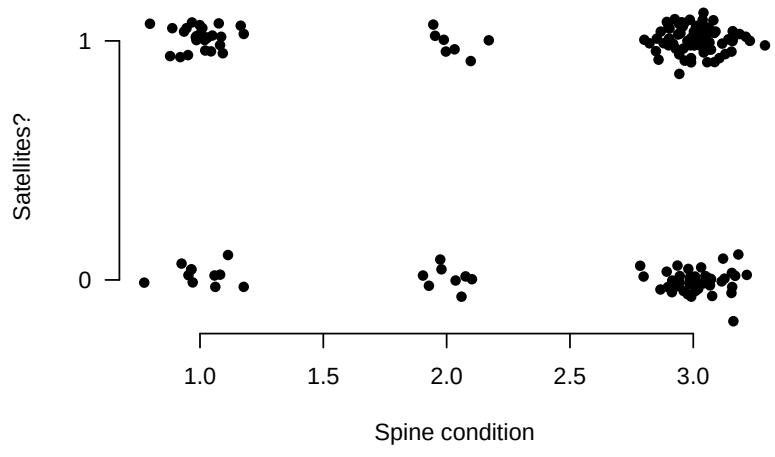
- Model selection involves choosing single "best" model, and all subsequent inference is based on that model. Many selection approaches available.
- There are 2^p possible models. Sixteen for the crab esxample. Model space grows *exponentially* with number of new predictors.
- The above approach *ignores* the uncertainty in the model selection process. Model averaging is one way to sensibly conduct inference in the context of all available models.
- In cases where the p is large, the model space becomes too large to computationally assess every model. Algorithms such as Markov Chain Monte Carlo Model Composition (MC^3) can be used to search the space effectively. See (Hoeting et al. 1999) for a good overview.

Plot the data

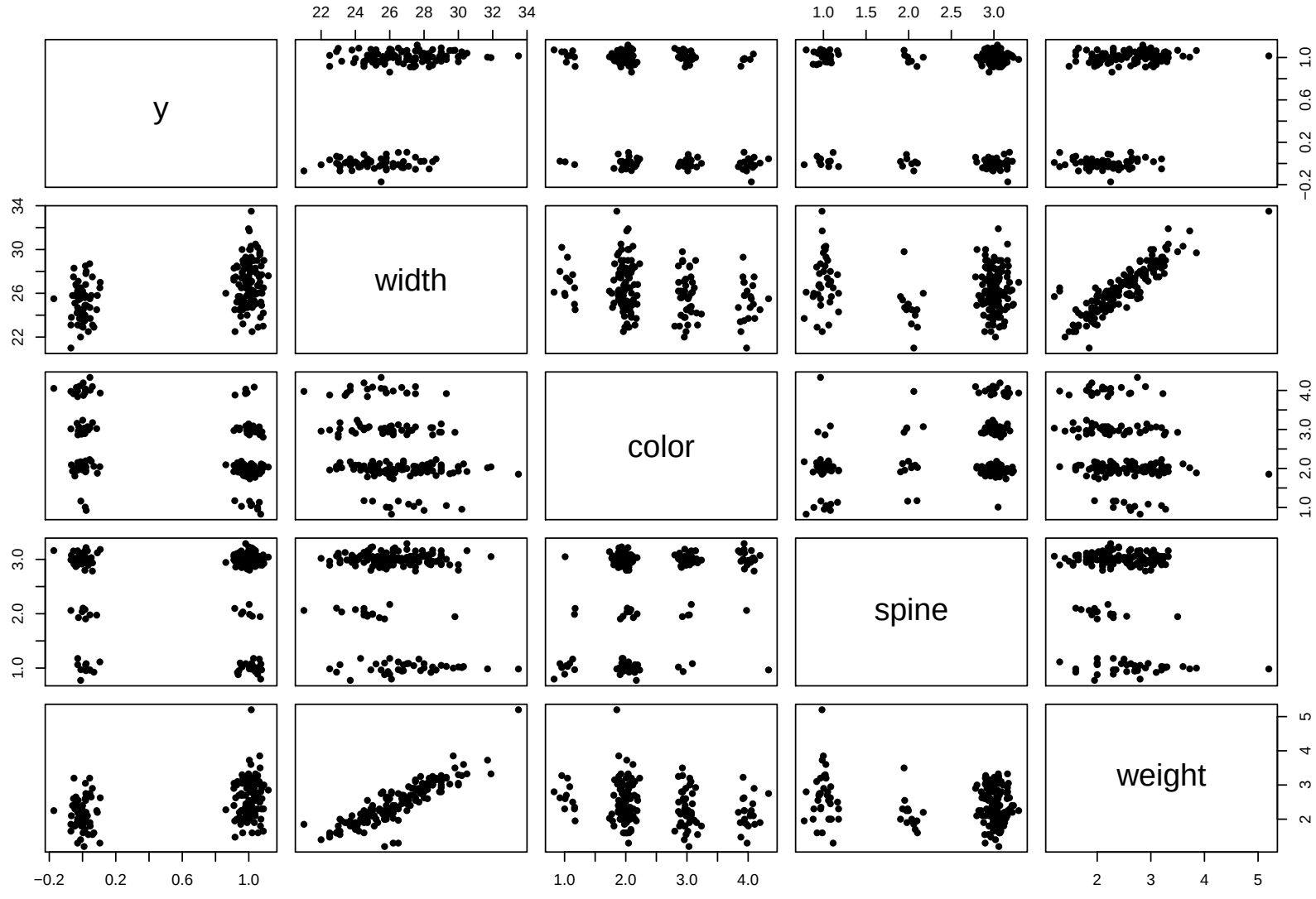








Multicollinearity is naturally a concern as well



History of GLMs

- GLMs generalized in 1970s. Logistic regression formalized in 1920s.
- Fisher scoring algorithm applies generally to GLMs which make them very popular and easy to implement in software.
- Here are some popular models

Table 3.3 Generalized linear models for statistical analysis.

Random Component	Link Function	Explanatory Variables	Model
Normal	Identity	Continuous	Regression
Normal	Identity	Categorical	Analysis of variance
Normal	Identity	Mixed	Analysis of covariance
Binomial	Logit	Mixed	Logistic regression
Multinomial	Logits	Mixed	Multinomial logit
Poisson	Log	Mixed	Loglinear

Logistic regression with categorical predictors

- Like all linear models, categorical variables can be included in logistic regression by forming an appropriate array of zeroes and ones in the design matrix.
- The two nomial variables (color and spine), can be handled in this manner. Analyses in the book justify treating these as linear trends, which I will do in the R demo subsequently

Part 2 : Bayesian model selection and averaging

Consider a set of candidate models

- For the variable subset selection problem, there are $K = 2^p$ models, with a model for every possible combination of variables. Typical to include intercept in every model.
- Denote the models M_k for $k = 1, \ldots, K$. For the Bayesian approach to model selection and averaging, each model requires a prior probability $P(M_k) \geq 0$. Note $\sum_{k=1}^{K} P(M_k) = 1$.
- We will consider a uniform prior on the model space in our R demo. A great paper on specification of model priors: (Scott and Berger 2010)

Review of Bayesian information Criterion

- A well known approach to model selection is to compute an information criterion (e.g., AIC, BIC, their extensions) for each of the $k = 1, \ldots, K$ models, then select the model with the optimal value of the information criterion
- AIC chooses models that are too big, but BIC is model selection consistent for regular problems.
- BIC has a term that rewards high likelihood and penalizes complexity (i.e., number of parameters.)
- $\text{BIC}_k = -2\ln(\hat{L}_k)$, where $\hat{L}_k$ is maximized log likelihood for model k.
- It is **very easy** to obtain BIC for models from software.

The integrated likelihood is an important quantity in model selection and averaging.

- Since the different candidate models have different numbers of parameters, the first step in Bayesian model selection is to integrate the parameters out of the joint density of the data and the parameters.

$$P(y|M_k) = \int_{\Theta} P(y|\theta_k, M_k)P(\theta_k|M_k)d\theta_k$$

- The $P(\theta_k|M_k)$ priors on parameters must be chosen with care in order for Bayesian model selection to work well. A great history of the problem is in the literature review here: (Ormerod et al. 2017).
- This step can sometimes be tedious - analytic solution infrequently available.

Bayes factor comparing model j to k

- The Bayes factor is the ratio of integrated likelihoods from two ccandidate models.

$$\mathrm{BF}_{jk} = \frac{P(y|M_j)}{P(y|M_k)}.$$

- The Bayes factor is the quantity by which the data update the prior odds to posterior odds. See (Lavine and Schervish 1999)
- For our approximate BIC approach, access to Bayes factors is a bit of an intermediate step.

Approximation to Bayes factors using BIC

$$BF_{kj} \approx e^{-\frac{1}{2}(BIC_k - BIC_j)},$$

- In the last decade have been surprised to learn of this approximation and especially how not-well-known it is compared to BIC. Doesn't it seem like everybody knows BIC?

Posterior model probabilities

- Once the user has specified prior model probabilities $P(M_k)$ for $k = 1, \ldots, K$, an application of Bayes rule provides posterior model probabilities.

$$P(M_k|y) = \frac{P(y|M_k)P(M_k)}{\sum_{j=1}^{K} P(y|M_j)P(M_j)}.$$

From Bayes factors to posterior model probabilities

$$P(M_k|y) = \frac{P(M_k)}{P(M_1)} BF_{k1} P(M_1|y)$$

- WLOG let model 1 be the baseline model

$$P(M_k|y) = \frac{P(M_k)}{P(M_1)} BF_{k1} P(M_1|y)$$

Obtaining the posterior model probability of the baseline model

$$P(M_1|y) = \Big(\sum_{k=1}^{K} \frac{P(M_k)}{P(M_1)} BF_{k1} \Big)^{-1}$$

Bayesian model averaging

- For a quantity of interest Δ (e.g. parameter value, inclusion in model), the law of total probability

$$P(\Delta|y) = \sum_{k=1}^{K} P(\Delta|M_k, y)P(M_k|y).$$

- KEY POINT: Whatever it is you care about, the way to get a Bayesian model averaged distribution for that quantity is to use posterior model probabilities $P(M_k|y)$ as weights, average across the models.
- This is a straightforward application of the law of total probability.

Posterior inclusion probabilities

- By setting Δ as an indicator in cases where an effect of interest is in a model, you can use BMA to calculate the probability that an effects is non-null.

$$\Delta = \begin{cases} 0, & \text{if candidate predictor is not in model} \\ 1, & \text{if candidate predictor is in the model,} \end{cases}$$

Then re-express BMA equation as

$$P(\Delta = 1|y) = \sum_{k=1}^{K} P(\Delta = 1, M_k|y)$$

Let's do an R demo!

- Scope: manually implement the BIC approximation to posterior model probabilities and inclusion probabilities.
- Further: I will demonstrate usage of the BAS package to implement fully Bayesian model averaging (including posterior distributions on coefficients)
- Live session: As time permits we can field questions with the code.

Downsides of BMA

- Tacitly assumes true model is in candidate set (m closed). More realistically, the true model is not in the candidate set (m open).
- BMA concentrates posterior model probability on a single model as sample size increases (Yao et al. 2018). The model that wins is closest to the true model in s KL divergence sense. So you end up giving basically 100% of posterior model probability to a model which is not technically the true one in m open setting.
- Recent work has been done on the stacking of Bayesian predictive distributions(Yao et al. 2018). The basic idea of stacking is to weigh the candidate models based on their ability to predict out-of-sample data rather than posterior model probability. This is useful for prediction and overcomes the tendency of BMA to put all posterior model probability on a single model as sample size increases.

References I

Agresti, A. 2018. *An Introduction to Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley. https://books.google.com/books?id=onZyDwAAQBAJ.

Clyde, Merlise. 2022. *BAS: Bayesian Variable Selection and Model Averaging Using Bayesian Adaptive Sampling*.

Clyde, Merlise A., Joyee Ghosh, and Michael L. Littman. 2011. "Bayesian Adaptive Sampling for Variable Selection and Model Averaging." *Journal of Computational and Graphical Statistics* 20 (1): 80–101. https://doi.org/10.1198/jcgs.2010.09049.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. 1999. "Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors." *Statistical Science* 14 (4): 382–417. https://doi.org/10.1214/ss/1009212519.

References II

Lavine, Michael, and Mark J. Schervish. 1999. "Bayes Factors: What They Are and What They Are Not." *The American Statistician* 53 (2): 119–22. https://doi.org/10.1080/00031305.1999.10474443.

Li, Yingbo, and Merlise A. Clyde. 2018. "Mixtures of g-Priors in Generalized Linear Models." *Journal of the American Statistical Association* 113 (524): 1828–45. https://doi.org/10.1080/01621459.2018.1469992.

Liang, Feng, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. 2008. "Mixtures of g Priors for Bayesian Variable Selection." *Journal of the American Statistical Association* 103 (481): 410–23. https://doi.org/10.1198/016214507000001337.

Ormerod, John T, Michael Stewart, Weichang Yu, and Sarah E Romanes. 2017. "Bayesian Hypothesis Tests with Diffuse Priors: Can We Have Our Cake and Eat It Too?" *arXiv Preprint arXiv:1710.09146*.

References III

Scott, James G., and James O. Berger. 2010. "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics* 38 (5): 2587–2619. https://doi.org/10.1214/10-AOS792.

Yao, Yuling, Aki Vehtari, Daniel Simpson, and Andrew Gelman. 2018. "Using Stacking to Average Bayesian Predictive Distributions." *Bayesian Anal.* https://doi.org/10.1214/17-BA1091.