Bayesian Model Selection, Model Comparison, and Model Averaging*

Florian M. Hollenbach† & Jacob M. Montgomery‡

First Draft: December, 2018
This Draft: November 12, 2019

Forthcoming in
The SAGE Handbook of Research Methods in Political Science and International Relations
(Pre-Proof)

*Author order in alphabetical order. We thank Bruce A. Desmarais, Andrew Gelman, and Richard Nielsen for helpful comments on earlier drafts. All remaining errors are our own.

†Assistant Professor, Department of Political Science, Texas A&M University, 2010 Allen Building, 4348 TAMU, College Station, TX, USA, 77843-4348. Email: fhollenbach@tamu.edu. Phone: 979-845-5021. URL: fhollenbach.org

‡Associate Professor, Department of Political Science, Washington University in St. Louis. Email: jacob.montgomery@wustl.edu. URL:jacobmontgomery.com

1 Introduction

Applied researchers are often interested in testing competing theories against each other. Most often, the goal is to determine whether and how a limited number of variables are related to a single outcome. The question in these cases is, "which of these theoretical accounts is most consistent with the data?" Or, more ambitiously, "which of these theories is most consonant with the true data generating process (DGP)?" Despite the ubiquity and importance of this research task, many scholars are still uncertain as to how to proceed in these situations. The purpose of this chapter is to explain how this analytical objective can be accomplished effectively using Bayesian model comparison, selection, and averaging, while also highlighting the key assumptions and limitations of these methods. This chapter's overall purpose is to provide readers with a larger set of tools for tackling this task and to discourage the kinds of haphazard (and often incorrect) practices for comparing theories often seen in the literature.

There are two interlocking problems in comparing and contrasting alternative theories via standard statistical methods. First, in many cases the alternative theories are not "nested" in a way that allows them to be tested simultaneously in a single regression model. When this is true – and it often is – the common practice of placing all of the variables from all of the theories into a single regression is inappropriate and can lead researchers to incorrect conclusions. Yet, there appears to be no widely accepted framework in the methods literature that allows scholars to compare non-nested models for the purposes of theory testing.[1]

[1]See Clarke (2001) for a review of some of the many methods in this area that have been proposed.

At the same time, researchers testing any theory need to "control for" additional covariates in order to rule out potential confounding factors, shrink the standard errors for key coefficients, or improve model fit. They must also choose from among many potential modeling options, including choosing functional forms, link functions, and more. Yet, in many cases theory offers limited guidance as to which or how many potential confounders should be included or what exact modeling strategy is most appropriate. This leaves scholars facing the challenge of having to choose among many different models and yet having little guidance as to how to arbitrate between them. In response, researchers often engage in a haphazard search through a large implied model space and report only a handful of results to readers for evaluation. Even worse, scholars may either intentionally or unintentionally try out alternative model specifications only until they find a result that confirms their research hypotheses – a practice that can lead to higher false positive rates for published research (see Montgomery and Nyhan, 2010, for additional discussion).

What criteria should scholars use when choosing between competing models or when considering alternative modeling strategies or model configurations? In this chapter, we present a number of tools from Bayesian statistics that allow scholars to approach these challenges in a more principled manner. The Bayesian framework significantly facilitates this task since the model configuration itself can be viewed as an unknown quantity to which Bayesian reasoning can be applied. We can use the tools of Bayesian statistics

Clarke (2007) provides further improvements to the Vuong test that allows for comparison of two non-nested models. Desmarais and Harden (2013) note that the Clarke test can be based on biased estimates and inconsistent but suggest an alternative implementation using cross-validation. Our claim is not that there are no methods available, only that many applied scholars appear to be very uncertain about which, if any, method to use.

to compare the relative evidence in favor of various models to select the "best" model, an approach that can be loosely labeled *model selection*. We can also examine posterior estimates to assess the degree to which a candidate model has adequately captured the true data generating process, which we label *model evaluation*. Finally, we can take a more agnostic approach and incorporate the uncertainty about the appropriate model configurations directly into final estimates, or *model averaging*.

Below, we provide a broad overview of the tools available for model selection, evaluation, and averaging with a special emphasis on theory testing. First, we briefly discuss "traditional" approaches to model selection via Bayes factors and model fit comparisons. This latter category of tools includes approximations of Bayes factors as well as criteria based on out-of-sample prediction. We then discuss the idea that no "correct" model exists, and therefore researchers should incorporate the uncertainty about model configuration directly into their statistical approach. Specifically, we cover Bayesian mixture models, Bayesian model averaging, and the recently developed Bayesian stacking. Throughout the chapter, although we do provide some details of the mathematics, our focus is on providing a general intuition about these various methods along with their relative strengths and weaknesses. Readers interested in more thorough treatments of this subject are directed to the works cited below. All code to reproduce the models and model selection examples we describe in this chapter will be available online.[2]

[2]Code and dat can be found on GitHub under https://github.com/fhollenbach/BayesModelSelection.

2 Motivating example: Testing theories of congress

As our working example throughout this chapter, we rely on Richman (2011), an article that appeared in the *American Political Science Review* that seeks to test competing models of policymaking in the US Congress. Specifically, Richman (2011) tests competing models that make predictions about which status quo policies are likely to be enacted as the composition of the House, Senate, and Presidency shift (Brady and Volden, 1998; Krehbiel, 1998; Cox and McCubbins, 2005). Using novel estimates of status quo locations in different policy areas, policy changes in those areas, and the ideal points of pivotal actors, Richman (2011) empirically tests four competing theories.[3]

Richman (2011) estimates predictions for where the status quo in 42 policy areas *should be* according to each theory for the 103rd through 110th Congress. Richman (2011) then estimates a simple regression for the status quo policy in issue area i at time period t using the following formula:

$$y_{it} = \beta_1 \text{prediction}_{it} + \beta_2 \text{inflation}_{it} + \epsilon_{it}$$

The main difference across models is, therefore, how the *prediction* variable is calculated. In all cases, the theory is that the β_1 coefficient should be equal to 1, although the main criteria is determining whether the coefficient is positively related to the outcome as expected. Richman (2011) generates predictions for the status quo to be located at the

[3]Our specific focus here is on Table 3 in the original paper. We deviate from Richman (2011) in not estimating panel corrected standard errors, as this sort of "correction" does not easily translate into a Bayesian framework, and we want to present a very simple example. The *brms* package we use does allow users to estimate models with autoregressive terms or other solutions to serial correlation in time and space.

position of the median voter of the house (*Model 1*), as predicted by the pivotal politics theory (*Model 2*), as predicted by the party cartel model (Cox and McCubbins, 2005) with only negative agenda control (*Model 3*), and as predicted by a hybrid cartel theory that assumes some degree of positive agenda control by party leaders (*Model 4*). The only control variable considered in Richman (2011) is a measure of inflation to reflect the natural change in status quo positions in some policy areas that result from inflation rates. Richman (2011) calculates two versions of the inflation measure, one for models one and two, and one for the third and fourth models. The two inflation measures differ based on the relevant policy area according the relevant theory. Richman (2011) then evaluates the different models based on their ability to predict the status quo based on a linear model. Specifically, the models are ranked based on their individual R^2 values.

We replicate the four models presented in Table 3 in Richman (2011) using the *brms* package in R (Bürkner, 2017, In Press).[4] The *brms* package provides users with a large number of pre-specified Bayesian models that are then estimated in Stan using C++ (Team, 2017; Carpenter et al., 2017). Stan is a relatively young probabilistic programming language, similar in spirit to WinBugs. In fact, writing model code in Stan is quite similar to doing so in WinBug. At this point, the vast majority of Bayesian models can be fit in Stan, and a fast growing number of R packages provide users with pre-programmed routines for an extensive number of Bayesian models. Stan allows users to estimate models using fully Bayesian sampling via the Hamiltonian Monte Carlo (HMC) methods or approximating posterior means and uncertainty using variational inference. The HMC

[4]To have the same sample in all our models, we delete three observations that have missing data on the lagged dependent variable.

approach to Markov chain Monte Carlo methods is particularly attractive because of its high scalability and ability to succeed in highly dimensional spaces.[5]

For each of the four models described above, we estimate a standard Gaussian linear model, where $y = \mathbf{X}\beta + \epsilon$, and $\epsilon \sim N(0, \sigma)$. We specify Gaussian priors with mean zero and a standard deviation of five for the regression coefficients. For the residual standard deviation (σ), we keep the default half student t prior with three degrees of freedom and scale parameter ten. In addition to the four models presented in Richman (2011), we add two additional models. First, we estimate a model that includes all six possible covariates and a lagged dependent variable. Second, we estimate the model with all covariates and the lag DV but also add random intercepts for each congress and issue area. For the standard deviation of the random effects, we use the same half student t prior as above.

The median estimates and 95% credible intervals for the estimates in all six models are shown in Table 1. There are two aspects of these results that are notable. First, Richman (2011) correctly identifies that these competing theories cannot be tested within a single model and does not attempt to do so. The result, however, is that we end up with four non-nested models that must be compared against each other. To arbitrate between them, Richman (2011) makes interpretive claims based on the overall model fit (as assessed by R^2 values). For instance, Richman (2011, p 161) states that Model 2 "dramatically improves upon the predictions that can be made" relative to Model 1. Likewise, in comparing Model 3 and 4, he concludes that "the differences in fit between

[5]The intricacies behind Hamiltonian Monte Carlo go beyond the scope of this chapter; for a more thorough introduction to Hamiltonian Monte Carlo, see Neal (2011) and Betancourt (2017).

Table 1: Evaluating Theories of Congress: Median Estimates and 95% Credible Intervals for Table 3 in Richman (2011) [Models 1-4] and two garbage can models

Variable	Model 1	Model 2	Model 3	Model 4	Full Model	Full Model & RE
lag DV					0.59	0.39
					(0.38,0.8)	(0.09,0.71)
Median only	0.5				0.02	-0.13
	(0.07,0.93)				(-0.25,0.29)	(-0.61,0.24)
Pivotal politics		0.92			-0.72	-0.61
		(0.66,1.17)			(-1.22,-0.23)	(-1.19,-0.08)
Party cartel open rule			0.89		1.1	0.81
			(0.72,1.06)		(0.23,1.94)	(-0.16,1.84)
Party cartel closed rule				0.82	-0.4	-0.11
				(0.65,0.99)	(-1.1,0.34)	(-0.96,0.71)
Inflation (median/pivot)	0.07	0.05			-0.01	0
	(-0.1,0.25)	(-0.1,0.19)			(-0.11,0.09)	(-0.1,0.1)
Inflation (party)			0.06	0.07	0.07	0.1
			(0.02,0.1)	(0.03,0.11)	(0.04,0.11)	(0.05,0.15)
Sigma	4.52	3.83	3.09	3.15	2.49	2.21
	(3.99,5.17)	(3.38,4.41)	(2.74,3.54)	(2.77,3.62)	(2.19,2.86)	(1.85,2.69)
N	117	117	117	117	117	117
Random Effects	No	No	No	No	No	Yes

the models is modest enough that no definitive conclusion can be drawn" (Richman, 2011, p 161). While the model fits superficially suggest that these conclusions are true, without a formalized approach to non-nested model comparison these competing claims cannot be formally tested. Fit statistics such as R^2 are simply not designed to allow us to say clearly that one model is better than another in a statistical sense. That is, there is no threshold we can establish for when R^2 values are "different enough" to show that one is statistically superior to another.

Second, as in nearly always the case, the models reported in Richman (2011) are not the only ones that were considered.

> I have also analyzed the data using a wide range of assumptions, including ordinary least squares, fixed effects by issue, random effects with and without AR(1) errors, panel heteroskedastic errors with an AR(1) process, and

> dynamic GMM models without analytic weights. All analytic approaches produced statistically significant effects in the expected direction (except for the median model), and all produced the same relative ranking ... (Richman, 2011, p 160).

From this description, it seems likely that these alternative specifications were tried in response to or in anticipation of reviewer questions and serve as robustness checks for the main model. Nonetheless, they do indicate that there are other potential model configurations that were considered and could have been evaluated relative to the reported results if appropriate criteria were available.

3 How *not* to test competing theories

How can we arbitrate among these competing theories? Before answering, we discuss one approach *not* to take: tossing everything into a single model and examining which coefficients are significant. As Achen (2005) shows, even in trivial models with only two covariates, this approach will not only fail to appropriately test between theories but may actively mislead researchers by, for example, switching the sign of key coefficients once we have conditioned on competing (and likely correlated) concepts. In short, simply combining variables from competing models in the hope that the results will somehow point to the "right" theory is a deeply flawed approach to science. While under strict conditions it *can* be correct,[6] in a more general setting it is ill advised and conclusions

[6]Imagine, for instance, that we had conducted a large experiment with 12 different treatment arms. In this case, simply tossing all of the variables into a single regression would indeed be an appropriate way to test the effectiveness of each treatment. However, social scientists are rarely, if ever, blessed with explanatory variables where the effects are strictly linear and co-linearity is sufficiently low to allow for this approach to work.

based on this approach should not be trusted.

As an example, recall that Columns 5 and 6 in Table 1 show what happens if we simply drop all of the various predictions into a single model. Column 5 is the same standard Gaussian model with all six predictors and a lag DV included, and Column 6 also adds congress and issue random effects. The results in this case are dramatic and revealing about the nonsensical nature of the approach. For instance, all four of the critical variables are hypothesized to be positive and significant. However, in the full model several have credible intervals that now include zero. The *pivotal politics* variable actually flips to become negative.[7] The party cartel (with open rule) has a 95% credible interval that excludes zero on Column 5 but includes zero once we include random effects (Column 6). In general, we get a mishmash of results that are hard to interpret in some cases and in others do not correspond with the theory at all.

Of course the problem here is that these coefficients have no interpretable meaning since the key explanatory variables are deeply inter-dependent. The coefficient for the pivotal politics variable does not reflect the *independent* effect of this prediction since a change in one variable almost necessarily implies changed values in the other variables. After all, each of the variables are functions of shared precursors (e.g., the median ideological position of the senate or the position of president). In a causal setting we might consider them to be post-treatment (Acharya, Blackwell and Sen, 2016; Montgomery, Nyhan and Torres, 2018), but even with observational data there is no justification for including them all in the same model and no way to interpret the coefficients directly

[7] This is not a function of the lagged dependent variable being included. Without the lagged dependent variable but all other covariates included, the estimate of the pivotal politics coefficient is -1.37 and the 95% credible interval ranges from -1.87 to -0.86.

when we do. In general, the simultaneous inclusion of all covariates or their *step-wise* selection based on p-values is ill-advised in the vast majority of settings.

4 Model selection and Bayes factors

This simplest approach to Bayesian model selection is via the construction of Bayes factors. This approach is attractive due to its simplicity, and perhaps for this same reason was among the earliest proposed Bayesian methods for model selection. For these same reasons we present this approach first. However, we note up front that this approach has been extensively criticized by some authors, leading to the alternative methods discussed below.

A (seemingly) simple way of evaluating models from a Bayesian perspective is nothing more than applying Bayes rule to model probabilities (Gill, 2009). Bayes rule is simply a formalization of basic human intuition as to how we can take evidence to inform our beliefs about the "true" state of the world:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{1}$$

Here, $P(B)$ is our prior beliefs about B before any data is collected. $P(A|B)$ is the conditional distribution of observing A given that we have observed B, which is simply another way of expressing the likelihood of a statistical model. Finally, $P(A)$ is the marginal probability of observing A.

Bayes factors attempt to apply this same logic to the problem of using observed data to inform our beliefs about the probability in favor of a specific model. Let $\pi(\mathcal{M}_k)$ be the

prior probability that model k is "true," and let $p(\mathbf{y}|M_k)$ be the probability of observing the data $\mathbf{y}$ under the assumption that k is true.[8] Further, assume that we are considering $k \in 1, 2, \ldots, K$ alternative model configurations. With a finite set of potential models, we can then calculate the marginal distribution of the data as,

$$p(\mathbf{y}) = \sum_{k=1}^{K} \pi(\mathcal{M}_k) p(\mathbf{y}|M_k). \tag{2}$$

Simply applying Bayes rule, we can express the posterior probability of any particular model k as:

$$p(\mathcal{M}_k|\mathbf{y}) = \frac{\pi(\mathcal{M}_k) p(\mathbf{y}|M_k)}{\sum_{k=1}^{K} \pi(\mathcal{M}_k) p(\mathbf{y}|M_k)} = \frac{\pi(\mathcal{M}_k) p(\mathbf{y}|\mathcal{M}_k)}{p(\mathbf{y})}. \tag{3}$$

If we then want to compare two models (a vs. b), we can construct a ratio between the two models' posteriors. This has the advantage that the denominators will simply cancel out:

$$\frac{p(\mathcal{M}_a|\mathbf{y})}{p(\mathcal{M}_b|\mathbf{y})} = \frac{\pi(\mathcal{M}_a) p(\mathbf{y}|M_a)}{\pi(\mathcal{M}_b) p(\mathbf{y}|M_b)} = \text{Prior odds}(\mathcal{M}_a; \mathcal{M}_b) \times \text{Bayes factor}(\mathcal{M}_a; \mathcal{M}_b) \tag{4}$$

Since the Bayes factor contains all of the "objective" information about the models (i.e., information that is separate from the model priors), this has been the traditional quanitity of interest. Higher values for this calculation represent evidence in favor of M_a and smaller values represent evidence in favor of M_b. Jeffreys (1961) provided an early attempt to set thresholds for Bayes factors where this evidence could be considered "con-

[8] The assumption that the *true* model is in the model space is not necessary for methods of model selection discussed below.

clusive" or merely "suggestive." A widely accepted threshold is that a Bayes Factor of 3 is "substantial" evidence in favor of M_a and values above 10 are considered "strong."[9]

The advantage of the Bayesian approach to model evaluation is that estimating a posterior probability for each model allows us to talk about model selection in an intuitive way. Unlike alternatives such as likelihood ratio tests that rely on confusing p-values and null hypothesis testing, we can talk directly about model probabilities. Statements like, "there is a 90% chance that this is the best model" have some possibility of making sense. That is, we can directly assess various models and determine which one is most supported by the data and with what degree of certainty. Moreover, the models do not have to be nested to be comparable.

Despite these superficial advantages, however, the simplified description above obscures several complexities that make the problem of model comparison and selection difficult. To begin with, in almost all cases we are not just interested in choosing the right model, but rather in estimating some set of model parameters $\boldsymbol{\theta}$ conditioned on our model choice and data. That is, our actual learning target is often the posterior distribution, $p(\boldsymbol{\theta}_k|\mathbf{y}, \mathcal{M}_k)$. Further, the presentation above makes implicit assumptions about prior structures for $\boldsymbol{\theta}$ that will not always hold. In more realistic settings, evaluating model fit based on Bayes factors comes with several additional difficulties that have, in combination, worked against their widespread adoption: computational intractability, prior sensitivity, incomplete accounting for uncertainty, and open model spaces.

[9] If these two models have equal prior distributions, the models are nested, and we estimate them via maximum likelihood, then Equation 4 simply reduces to a likelihood ratio statistic.

4.1 Marginal likelihoods

In order to construct Bayes factors, we need to be able to calculate the probability of the data given the model after marginalizing out the model parameters. More concretely, let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ be some set of model parameters of interest, let $L(\boldsymbol{\theta}_k) = p(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{M}_k)$ represent the standard likelihood function for the data from model k, and $\pi(\boldsymbol{\theta}_k|\mathcal{M}_k)$ be the prior distribution for $\boldsymbol{\theta}_k$.[10] In order to calculate model probabilities or Bayes factors, we need to marginalize out $\boldsymbol{\theta}_k$:

$$p(\mathbf{y}|\mathcal{M}_k) = \int_{\boldsymbol{\Theta}} p(y|\boldsymbol{\theta}_k, \mathcal{M}_k)\pi(\boldsymbol{\theta}_k|\mathcal{M}_k)d\boldsymbol{\theta}. \tag{5}$$

Unfortunatley, the integral in Equation 5 is often not analytically tractable. Moreover, even where it can be approximated with fidelity for one model, the set of models being considered may be sufficiently large such that Equation 2 cannot be calculated in a reasonable amount of time. Note, that in order to find the marginal distribution in Equation 2, we would need to complete these calculations K times, and without $p(\mathbf{y})$ individual model probabilities cannot be directly calculated.

4.2 Approximations and BIC

Statisticians have developed several methods designed to approximate Equation 3 quickly. For example, one can use Laplace's method to approximate $p(\mathbf{y}|\mathcal{M}_k)$ using the formula

[10] The k subscript on $\boldsymbol{\theta}$ allows that each model under consideration may have a different set of parameters.

(Ando, 2010, 114):

$$p(\mathbf{y}|\mathcal{M}_k) \approx p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \mathcal{M}_k)\pi(\hat{\boldsymbol{\theta}}_k|\mathcal{M}_k) \times \frac{(2\pi)^{\frac{p}{2}}}{n^{\frac{p}{2}}|J(\hat{\boldsymbol{\theta}})|^{\frac{1}{2}}}, \tag{6}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator (MLE), p is the number of parameters in the model, and $J(\boldsymbol{\theta}) = -\frac{1}{n}\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$ is a function of the Hessian of the log-likelihood such that the second term is related to the asymptotic covariance of the MLE.

This can be simplified further in instances where a large n and appropriate prior structure allows us to ignore the prior. In such cases, Schwarz (1978) proposes a simpler approximation of $p(\mathbf{y}|\mathcal{M}_k)$, although subsequent work has raised questions about whether BIC can actually be considered an approximation of any valid quantity (Gelman and Rubin, 1995).[11] Taking the log of Equation 6 and disregarding portions that are constant in large-n settings, we get the Bayesian information criterion (BIC)[12]

$$\text{BIC} = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k) + p \log n,$$

where we let $p(y|\hat{\boldsymbol{\theta}}_k, \mathcal{M}_k) = p(y|\hat{\boldsymbol{\theta}}_k)$ to simplify the notation. Likewise, we get

$$\log BF[M_a; M_b] \approx (BIC_b - BIC_a)/2. \tag{7}$$

The advantage of using BIC is its simplicity. We can evaluate and compare models

[11] Despite the name, notice that BIC is primarily useful for evaluating the model fit from the MLE in an asymptotic setting where priors are irrelevant. Further, the motivation and derivation for BIC differs significantly from the other "information criteria" covered below.

[12] Note that some texts and software packages may define BIC as the negative of how we have defined it.

using straightforward calculations from the likelihood based on the MLE. The obvious drawback is that BIC will give inaccurate and even misleading approximations in small sample settings or when prior structures cannot be ignored (e.g., improper priors). For example, as discussed more below, if the compared models are quite similar, prior choices can be decisive when evaluating models using BIC or Bayes factor. Indeed, Berger, Ghosh and Mukhopadhyay (2003) shows that even in fairly simple models, BIC can lead to incorrect conclusions even as $n \to \infty$.[13] For these reasons, we advise that researchers be cautious in interpreting BIC as a proper approximation of a Bayes factor or, even better, avoid the use of BIC altogether.

4.3 Approximation via simulation and bridge sampling

Several other approaches focus on approximating marginal probabilities using methods that take advantage of the simulation methods (e.g., Markov chain Monte Carlo) typically used to estimate posterior distributions of $\boldsymbol{\theta}_k$. What these methods have in common is that they attempt to avoid the potentially high-dimensional integration problem in Equation 5 using Monte Carlo-like approximations. These estimation methods themselves allow for estimation of a wide array of models in a Bayesian framework including generalized linear models, hierarchical linear models, item response models, ARIMA modes, network models, and more.

One of the simplest simulation approaches again rests on the the Laplace method. Assume that we have $s \in [1, 2, \ldots, S]$ samples of $\boldsymbol{\theta}_k$ from model k, denoted $\boldsymbol{\theta}_k^{(s)}$. We can then approximate the posterior mode as (Ando, 2010, p. 170):

[13] More advanced approaches for estimating marginal model probabilities include the generalized Bayesian information criterion (Konishi, Ando and Imoto, 2004), which allows for more informative priors.

$$\hat{\boldsymbol{\theta}}_k \approx \max_s \left\{ p(\boldsymbol{\theta}_k^{(s)}|\mathbf{y}) \right\} = \max_s \left\{ p(\mathbf{y}|\boldsymbol{\theta}_k^{(s)}, \mathcal{M}_k) \pi(\theta_k^{(s)}|\mathcal{M}_k) \right\}.$$

The posterior covariance can be approximated as:

$$\hat{\mathbf{V}} \approx \frac{1}{S} \sum_{s=1}^{S} \left\{ \left(\boldsymbol{\theta}_k^{(s)} - \bar{\boldsymbol{\theta}}_k \right)^T \left(\boldsymbol{\theta}_k^{(s)} - \bar{\boldsymbol{\theta}}_k \right) \right\}$$

where $\bar{\boldsymbol{\theta}}_k$ is the posterior mean. We can then get:

$$p(\mathbf{y}|\mathcal{M}_k) \approx p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k) \pi(\hat{\boldsymbol{\theta}}) \times (2\pi)^{\frac{p}{2}} |\hat{\mathbf{V}}|.$$

More advanced examples of numerical approximations in the literature include reversible jump MCMC (Green, 1995), Chib's method (Chib, 1995), path sampling (Gelman and Meng, 1998), the harmonic mean estimator (Gelfand and D.K., 1994), and more. Each of these approaches has its relative advantages and disadvantages, but all can be technically difficult to implement and in some cases require completing difficult analytical calculations or setting up new samplers. Further, estimators that rely on evaluations of the likelihood (e.g., the harmonic mean estimator) can be numerically unstable since these are technically unbounded.

Perhaps the most generally useful approach within this family is bridge sampling (Meng and Wong, 1996; Meng and Schilling, 2002), which has a fairly "black box" implementation available for applied researchers (Gronau et al., 2017; Gronau, Singmann and Wagenmakers, 2017). The basic idea is to introduce a "bridge function," $h(\theta)$, and a proposal distribution, $\psi(\theta)$. We can then set up the identity

$$1 = \frac{\int_{\Theta} h(\boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k|\mathbf{y}, \mathcal{M}_k) \psi(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k}{\int_{\Theta} h(\boldsymbol{\theta}_k) \psi(\boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k|\mathbf{y}, \mathcal{M}_k) d\boldsymbol{\theta}_k} = \frac{\frac{1}{p(\mathbf{y}|\mathcal{M}_k)} \int_{\Theta} h(\boldsymbol{\theta}_k) p(y|\boldsymbol{\theta}_k, \mathcal{M}_k) \pi(\boldsymbol{\theta}_k|\mathcal{M}_k) \psi(\boldsymbol{\theta}_k) d\boldsymbol{\theta}}{\int_{\Theta} h(\boldsymbol{\theta}_k) \psi(\boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k|\mathbf{y}, \mathcal{M}_k) d\boldsymbol{\theta}_k}$$

$$p(\mathbf{y}|\mathcal{M}_k) = \frac{\int_{\Theta} h(\boldsymbol{\theta}_k) p(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{M}_k) \pi(\boldsymbol{\theta}_k|\mathcal{M}_k) \psi(\boldsymbol{\theta}) d\boldsymbol{\theta}_k}{\int_{\Theta} h(\boldsymbol{\theta}_k) \psi(\boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k|\mathbf{y}, \mathcal{M}_k) d\boldsymbol{\theta}_k} \tag{8}$$

For the denominator in Equation 8, we can then use draws of $\boldsymbol{\theta}_k$ to approximate the integral numerically. Likewise, we can take draws from the proposal density $\psi(\boldsymbol{\theta}_k)$ to accomplish the same task in the numerator. Assuming we take L draws from $\psi(\cdot)$, we can estimate Equation 5 as:

$$p(\mathbf{y}|\mathcal{M}_k) = \frac{\frac{1}{L} \sum_{l=1}^{L} h(\boldsymbol{\theta}_k^{(l)}) p(\mathbf{y}|\boldsymbol{\theta}_k^{(l)}, \mathcal{M}_k) \pi(\boldsymbol{\theta}_k^{(l)}|\mathcal{M}_k)}{\frac{1}{S} \sum_{s=1}^{S} h(\boldsymbol{\theta}_k^{(s)}) \psi(\boldsymbol{\theta}_k^{(s)})}.$$

Obviously, in order to implement this model we must choose $h(\cdot)$ and $\psi(\cdot)$. Meng and Wong (1996) provide an optimal choice for $h(\cdot)$ in terms of the the mean-squared error of the estimator. For efficiency, the proposal density will ideally be as close as possible to the posterior distribution. The `bridgesampling` package relies on either a multivariate normal distribution where the mean vector and covariance matrix are calculated from the full posterior for θ and a "warped" posterior (Meng and Schilling, 2002). For practical reasons, however, it is often useful to calculate $\log\{p(\mathbf{y}|\mathcal{M}_k)\}$ as we do below.

Once we have have estimated $p(\mathbf{y}|\mathcal{M}_k)$ for various models, we can then compare specific models by constructing Bayes factors. Once again, when comparing models a and b we can get,

$$BF[\mathcal{M}_a, \mathcal{M}_b] = \frac{p(\mathbf{y}|\mathcal{M}_a)}{p(\mathbf{y}|\mathcal{M}_b)}.$$

Alternatively, to maintain comparability with the BIC approach, we can calculate,

$$\log BF[\mathcal{M}_a, \mathcal{M}_b] = \log\left\{p(\mathbf{y}|\mathcal{M}_a)\right\} - \log\left\{p(\mathbf{y}|\mathcal{M}_b)\right\}.$$

4.4 Prior sensitivity, $\mathcal{M}$-closed assumption, and uncertainty

Perhaps the greatest limitation of Bayes factors is that the results can be very sensitive to priors. The approach to model selection discussed above rests on the $\mathcal{M}$-closed assumption, meaning that we are assuming that one of the $\mathcal{M}_k \in \mathcal{M}$ is the *true* model, even if the researcher does not know which it is. Especially when this is untrue, Bayes factors will be driven by the choice of prior distributions chosen for $\boldsymbol{\theta}$. Yao et al. (2018) provide the following powerful example of this problem:

> [C]onsider a problem where a parameter has been assigned a normal prior distribution with center 0 and scale 10, and where its estimate is likely to be in the range $(-1, 1)$. The chosen prior is then essentially flat, as would also be the case if the scale were increased to 100 or 1000. But such a change would divide the posterior probability of the model by roughly a factor of 10 or 100. (Yao et al., 2018, p 919)

In essence, if we were to compare two models that are exactly the same except one has a prior over β with standard deviation 10 and the other with standard deviation 100, the Bayes factor would strongly favor the first, despite the fact that the posterior

estimates of θ and even predictions for each observation would be essentially the same across models. Further, placing (most) improper vague priors on θ – arguably the most agnostic approach to model building – will lead to the Bayes factor not existing at all. This leads to the awkward result that the ultimate decision about model quality depends on choices we make about the prior structures on $\boldsymbol{\theta}$ parameters – choices that may be only incidental to the scientific question at hand.

A final concern is that some of these approaches to model comparisons are not "truly Bayesian" in the sense that they rely on a single estimate of $\hat{\boldsymbol{\theta}}$ rather than reflecting the entire posterior. Even the approaches that leverage the full posterior over $\boldsymbol{\theta}$ do so only to marginalize the quantities away rather than truly incorporating our posterior uncertainty into our estimates.

4.5 Example Application

Keeping these limitations in mind, we turn back to our running example. Based on the non-sensical results when including all covariates, we from now on only compare the four original models presented by Richman (2011). First, we calculate BIC for each of the four models, presented in Table 2. As we can see, Model 3 has the smallest BIC value, which would indicate the highest model fit. This is in line with the evaluation by Richman (2011) based on the R^2 values. Similarly, as with Richman's evaluation, the difference between the BIC values of Model 3 and 4 is very small. Nevertheless, the BIC for Model 2 is only 50 points larger than that of Model 3. Thus, again one would have to conclude that Model 3 seems to be slightly better than the other two models. In the bottom part of Table 2 we show the approximate Bayes factor (on the log scale),

calculated as in equation 7 above, for Model 3 compared to the three other models. The results lend additional support to the idea that Model 3 is clearly better than Model 1 and 2 while only being a slight improvement compared to Model 4. On the original scale, the Bayes factor between Model 3 and Model 4 is 9; i.e., in the language discussed above, this could be considered "substantial" but not quite "strong" evidence in favor of Model 3 over Model 4.

Table 2: BIC and Approximation to Bayes Factor

BIC scores			
Model 1	Model 2	Model 3	Model 4
696.3	657.2	607.2	611.5
Log Bayes Factor Approximation for Model 3			
$M_3 : M_1$	$M_3 : M_2$	–	$M_3 : M_4$
44.4	25.0		2.2

Next we use bridge sampling, with a *bridge function* as in equation 8 above, to first estimate the log marginal likelihood for the four models. We also generate estimations of the log Bayes factor comparing Model 3 to the others. The *brms* package again makes this very easy for applied users, as the estimations are integrated into the package. The top half of Table 3 below shows the log marginal likelihood for each of the four models as estimated via bridge sampling. Similar to previous results, the differences are largest with respect to Models 1 and 2 compared to Models 3 and 4. The lower half of Table 3 presents the Bayes factor of $M_3 : M_k$ on the log scale. For this exercise, the approximate Bayes factor using BIC and the estimation via bridge sampling are very similar. We note, however, that these different approximations are likely to diverge in more complicated settings.

Table 3: Log Marginal Likelihood and Bayes Factor using Bridge sampling

Log Marginal Likelihood			
Model 1	Model 2	Model 3	Model 4
-351.0	-332.5	-309.4	-311.5
Log Bayes Factor via Bridgesampling for Model 3			
$M_3 : M_1$	$M_3 : M_2$	–	$M_3 : M_4$
41.7	23.2		2.2

5 Predictive model evaluation

While model selection via Bayes factors and marginal model probabilities seems intuitive, as the above discussion indicates it is not always so straightforward in practice. This is particularly true in instances where the model parameters $\boldsymbol{\theta}$ take on continuous values requiring both informative priors and marginalization to complete the calculations. In addition to the difficulty of calculating the marginal likelihoods discussed above, Bayes factors in general are sensitive to priors on elements of $\boldsymbol{\theta}$ in ways that can be undesirable.

As an alternative to these approaches, the literature contains a number of approaches intended to help scholars evaluate the quality of any given model based on their predictive capacity. Here we follow the presentation in Gelman et al. (2013) and Gelman, Hwang and Vehtari (2014). The most common approaches are based on information theory, with the goal of minimizing the Kullback-Leibler (KL) divergence between the *true* (but unknown) data generating function and the predictive distribution implied by model M_k .

Let $\hat{\boldsymbol{\theta}}$ be our current estimate for the model parameters, $p(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{\theta}}_k, \mathcal{M}_k)$ be predictions

from model k for some new observation $\tilde{\boldsymbol{y}}$ that were not used to fit the model, and $g(\cdot)$ be the *true* data generating function. Rather than evaluating model fit based on *in-sample* performance, the idea is that we would like to choose the model that best fits *all* of the data, not just that which we have collected.

One way to summarize the predictive fit of a model is the log predictive density (lpd), $\log\{p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta})\}$. This quantity has the nice feature that, in the limit, the model with the lowest KL information also has the highest lpd (Gelman, Hwang and Vehtari, 2014). For a single point, we can define the lpd as the point prediction from that point after marginalizing out $\boldsymbol{\theta}$,

$$\log\{p(\tilde{y}_i|\boldsymbol{\theta}_k)\} = \log E\left[p(\tilde{y}_i|\theta)\right] = \log \int_{\Theta} p(\tilde{y}_i|\boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k|\mathbf{y}, \mathcal{M}_k) d\boldsymbol{\theta}_k.$$

Note that the expectation is taken over the posterior of $\boldsymbol{\theta}$, where the posterior is estimated based on the in-sample data $\mathbf{y}$.

We have to go further here, however, because the future data ($\tilde{y}$) is itself unknown. However, we can again use Bayesian reasoning to calculate the expected log predictive density (elpd) as:

$$E\left[\log\{p(\tilde{y}_i|\boldsymbol{\theta}_k, \mathcal{M}_k)\}\right] = \int g(\tilde{y}_i) \log\{p(\tilde{y}_i|\boldsymbol{\theta}_k, \mathcal{M}_k)\}\, d\tilde{y} = \text{elpd}.$$

In this case, the expectation is taken in terms of the unknown function $g(\cdot)$.

For more than one data point, we can simply sum this value to create the expected log pointwise predictive density (elppd),

$$\text{elppd} = \sum_{i=1}^{n} E\left[\log\left\{p(\tilde{y}_i|\boldsymbol{\theta}_k, \mathcal{M}_k)\right\}\right]. \tag{9}$$

The model that scores highest on this value can be considered the "best" model in terms of its predictive accuracy and, with large samples, optimal in terms of KL information. However, since this quantity cannot be calculated directly, we must again turn to approximations below.

Before moving onto specific approximations, however, it is helpful to define two additional quantities. First, if we assume some specific point estimate $\hat{\boldsymbol{\theta}}$, we can calculate the elpd as $E\left[\log\left\{p(\tilde{y}_i|\hat{\boldsymbol{\theta}})\right\}\right]$. Further, in this case and given standard *iid* assumptions, we can simplify the notation to get

$$p(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{\theta}}) = \prod_{i=1}^{n} p(\tilde{y}_i|\hat{\boldsymbol{\theta}})$$

.

5.1 Information criteria

In the particular case where we use the MLE, the elpd can be approximated accurately using the Akaike information criterion (AIC) (Ando, 2010).

$$AIC = -2\log L(\hat{\boldsymbol{\theta}}_{MLE}) + 2p,$$

where elpd $\approx -\frac{1}{2}AIC$. However, it is important to note that this assumes (i) we have not included informative priors on $\boldsymbol{\theta}_k$, (ii) the posterior distribution for $\boldsymbol{\theta}_k$ can be repre-

sented as a multivariate normal, and (iii) the model is correct (the true data generating process corresponds to some unknown member of the specified parametric family of distributions) (Ando, 2010, p 199). Thus, while it is simple to calculate, it is probably not applicable in many situations.

Note that AIC has two additive components, which is a feature of all of these similar criteria. The first represents the degree to which the model fits well given the data already collected; i.e., the in-sample fit. The problem, of course, is that models that better explain the data we have will *not* always be a superior representation of the underlying DGP. Instead, more complex models – models that include more variables, interactions, non-linearieties and the like – may simply capture random noise in the dataset, mistaking it for true information. More formally, improved in-sample model fit may actually decrease the ability of the model to explain (or predict) new observations generated by the same process.[14]

Following this logic, the second term in the AIC formula is a penalty term that punishes for complexity to work against selecting models that overfit the data. Under the assumptions stated above, the penalty term in the AIC is exact. However, when we move beyond a world of flat priors and linear models, this penalty term will no longer be correct (Gelman et al., 2013).

The deviance information criterion (DIC) overcomes these issues by approaching the problem from a more strict Bayesian perspective. Spiegelhalter et al. (2002) proposed the following criteria:

[14]See Hastie, Tibshirani and Friedman (2016) for a fuller discussion of the problems of in-sample and out-of-sample model fit. See Kung (2014) and Bagozzi and Marchetti. (2017) for recent applictions of AIC for model selection.

$$DIC = -2\log\left\{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{EAP})\right\} + P_D$$

where P_D is a Bayesian measure of model complexity that is based on the posterior mean $\hat{\boldsymbol{\theta}}_{EAP}$ and posterior covariance $\text{Var}(\boldsymbol{\theta})$. The exact penalty is

$$P_D = 2\left(\log\left\{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{EAP})\right\} - \int_{\boldsymbol{\Theta}} \log\ \left\{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})\right\} d\boldsymbol{\theta}\right).$$

Since the second term is simply the expected value for the log predictive distribution where the expectation is taken over the posterior of θ_k, we can use Monte Carlo integration using the $s = 1, \ldots, S$ draws from the posterior,

$$P_{D(1)} \approx 2\left(\log\left\{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{EAP})\right\} - \frac{1}{S}\sum_{s=1}^{S} \log\left\{p(\mathbf{y}|\boldsymbol{\theta}^{(s)}\right\}\right).$$

An alternative approximation,

$$P_{D(2)} \approx 2\text{Var}\left[\log\left\{p(\mathbf{y}|\boldsymbol{\theta})\right\}\right],$$

has the advantage of always providing a positive value (Gelman et al., 2013).

5.2 Model evaluation based on pointwise predictive densities

While both AIC and DIC aim to approximate (or proxy for) the out-of-sample fit, they are subject to two criticisms. First, the models use the same training data both to fit the model and evaluate the appropriate complexity penalty. This can bias estimates towards models that are too complex, leading researchers to the wrong decision. Further, both

criteria use only a single point estimate (either the MLE or the EAP) to evaluate model fit. This means that we ignore the full posterior of the model parameters, making the methods not "fully Bayesian" and can lead to problems, including negative estimates for the effective number of parameters (Vehtari, Gelman and Gabry, 2017, p 1414).

From a Bayesian perspective, several other criteria are generally preferred for model evaluation. These allow us to make use of the full posterior distribution but also avoid the problems of prior sensitivity that plague Bayes factors discussed above. These methods are also, to differing degrees, closely related to out-of-sample performance, which is more consonant with recent trends in model evaluations and further protects against over-fitting. For these criteria, we move towards evaluating fit based on how well predictive *densities* approximate the true data generating process for individual data points.

The goal is try to evaluate each model based on its predictive accuracy, where accuracy is evaluated based on the predictive distribution rather than the point estimate. Let $\tilde{\boldsymbol{y}}$ be some set of data points we are trying to predict (either new data or a data "held out" during fitting) and $\mathbf{y}^{obs}$ be the data we are currently using to fit the model. We can then write the posterior predictive distribution as:

$$p(\tilde{\boldsymbol{y}}|\mathcal{M}_k) = \int_{\boldsymbol{\Theta}} p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}_k, \mathcal{M}_k) p(\boldsymbol{\theta}|\mathbf{y}^{obs}) d\theta = E\big[p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}_k, \mathcal{M}_k)\big].$$

For similar reasons as noted above, we want to evaluate the model based on some function of the logged value, $\log\big\{p(\tilde{\boldsymbol{y}}|\mathcal{M}_k)\big\}$. Using $s = \{1, \ldots, S\}$ draws of $\boldsymbol{\theta}_k$ from a sampler, we can approximate using Monte Carlo integration. Summing over all observations in the held-out dataset, we then get the "computed log pointwise predictive density"

(clppd) (Gelman et al., 2013, p. 169):

$$\text{clppd} = \sum_{i=1}^{n} \log \left\{ \frac{1}{S} \sum_{s=1}^{S} p\left(\tilde{y}_i | \boldsymbol{\theta}_k^{(s)}\right) \right\}.$$

For a fully Bayesian treatment, we would then *like* to calculate the elppd shown in Equation 9. The general problem is that, since we cannot marginalize over the unknown function $g(\cdot)$, we must again settle for approximations based on clppd. When clppd is calculated within sample, we will overestimate the elppd which we then need to adjust (similar to the penalties for AIC and DIC discussed above). Alternatively, we might rely on true out-of-sample forecasts for calculating clppd. However, this comes with the problem of either introducing bias in the clppd as an estimate of elppd (because each $\boldsymbol{\theta}_k$ is estimated on a subset of the data and therefore our out-of-sample forecasts may be biased) or imposing considerable computational requirements.

5.2.1 WAIC

The widely available information criteria (alternatively Watanabe information criterion, or WAIC) is intended to provide a computationally friendly way of evaluating the performance of models based on predictive distributions (Watanabe, 2010, 2013). To generate the WAIC, we compute the clppd within the training sample and then apply a penalty term for complexity:

$$WAIC = \frac{1}{n} \sum_{i=1}^{n} \log \left\{ p(\tilde{y}_i | \mathcal{M}_k) \right\} - P_W.$$

As with DIC the trick is to find the correct penalty term P_W. One approach is to let

$P_w = V/n$, where V is the functional variance (Piironen and Vehtari, 2017*b*):

$$V = \sum_{i=1}^{n} \left\{ E[p(\tilde{y}_i|\boldsymbol{\theta})^2] - E[p(\tilde{y}_i|\boldsymbol{\theta})]^2 \right\}.$$

Gelman et al. (2013, Equation 7.12), however, recommend an alternative penalty equivalent to the summed variance of the log predictive density of each data point:

$$P_w = \sum_{i=1}^{n} \text{Var}_S \left[log \left\{ p\left(y_i|\boldsymbol{\theta}^{(s)}\right) \right\} \right],$$

where $\text{Var}_S[\cdot]$ is the sample variance function, and we calculate the variance across the S draws from the posterior.[15]

5.2.2 Cross validation

Rather than trying to arrive at a correct penalty for complexity, another approach is cross-validation to reduce the bias from over-fitting.[16] First, the data is split into A subsets (e.g, $A = 10$) of approximately equal size ($\frac{A-1}{A} \times n$). The model is then estimated on each of the subsets and predictions are made for the observations that were held out. Once completed A times, we now have out-of-sample predictions for each observation in the dataset. Based on this procedure, we can define a "cross validated predictive density" (Ando, 2010),

$$\text{cvpd}_k = \prod_{\alpha=1}^{A} \int_{\Theta} f(\mathbf{y}_\alpha|\boldsymbol{\theta}_k^{(\alpha)}, \mathcal{M}_k) p(\boldsymbol{\theta}_k^{(\alpha)}|\mathbf{y}_{\neg\alpha}, \mathcal{M}_k).$$

[15] See Kim, Londregan and Ratkovic (2018) for a recent application of WAIC in political science.

[16] In most literatures this approach is referred to as "k-fold" cross validation. However, to avoid notational confusion with the model space, we do not adopt this language here.

In this case, we "hold out" observations in partition α and calculate the posterior on θ_k. Based on this distribution, we can then calculate the clppd for model evaluation.[17]

While relying on out-of-sample predictions does reduce the bias in our estimate of elppd, reducing the sample size during model fitting can decrease the accuracy of the overall model itself. When A is too small, the inherent bias in the model utility estimate increases substantially. Generally, A between 8 and 16 has been recommended as reasonable to trade off bias and computational cost (Vehtari and Lampinen, 2002). Of course, estimating even a single Bayesian model, especially with large N or many parameters, can be computationally intensive and time consuming. Cross-validation requires to repeatedly (A-times) estimate each of the models one is interested in comparing. When we have many models to compare, cross-validating each can involve fitting thousands of total models. Since this can be parallelized easily, in simple cases cross-validation may be a good option. But even with modern computing, full cross-validation is only practical when comparing relatively few models and when each of the models is relatively computationally inexpensive. Further, depending on the type of dataset and estimated models (e.g., consider hierarchical data or spatial models), scholars have to give serious consideration as to how to partition the data for analysis (and for some there may be no clean approach to doing so).

5.2.3 LOO-CV

The extreme case of cross-validation methods is leave-one-out cross-validation (loo-cv). Here the model is estimated n times, each time leaving out one observation in the estima-

[17] Note that each competing model should be fit based on the same partitioning of the data.

tion and predicting that particular left-out observation. As in cross-validation, the predictive densities for all separately left out and predicted observations are then evaluated using some criteria. This approach has been shown to have a number of good properties, and, when feasible, is perhaps the best way to evaluate alternative models. Watanabe (2010) shows that WAIC is asymptotically equivalent to the Bayesian leave-one-out cross validation (loo-cv). Some argue that WAIC and loo-cv "give a nearly unbiased estimate of the predictive ability of a given model"(Piironen and Vehtari, 2017*a*, 712). Of course, since all models need to be estimated n times, loo-cv is likely to be too computationally intensive for many applied researchers.

To make loo-cv computationally tractable, Gelfand, Dey and Chang (1992) and Gelfand (1996) suggested importance sampling leave-one-out cross-validation. Effectively, we want to avoid estimating each model n times and sampling a new $\boldsymbol{\theta}_i$ to create a predictive distribution for each left-out observation i. To do so, we approximate $\boldsymbol{\theta}_i$, using the posterior draws for $\boldsymbol{\theta}$ taking into account the degree to which data point i affects the estimate. In the original approach, the importance ratio to approximate the model with the left out observation i would be: $r_i^s = \frac{1}{p(y_i|\theta^s)}$ (Vehtari and Lampinen, 2002; Vehtari, Gelman and Gabry, 2017). Thus, instead of having to estimate each model n-times, we only generate one estimate of $\boldsymbol{\theta}$ based on the whole data and then approximate the posterior for each of the data subsets by re-weighting with the importance ratios.

This strategy, however, can be problematic under several circumstances, such as for data with highly influential cases or high-dimensional models.[18] More recently, Vehtari,

[18]In particular, the variance of the importance weights may be too large or infinite since the denominator is not bounded away from zero.

Gelman and Gabry (2017) have suggested an improvement to importance sampling for loo-cv by smoothing the importance ratios, such that extreme values are not too influential or problematic. This is done using the Pareto distribution with its heavy tails, i.e., "Pareto smoothed importance sampling (PSIS)" (Vehtari, Gelman and Gabry, 2017, 1413).

Specifically, psis-loo-cv smoothes the 20% largest, and therefore potentially problematic, importance ratios. To do so, first a generalized Pareto distribution is fit to the largest importance ratios. These potentially problematic ratios are then replaced with "the expected values of the order statistics of the fitted generalized Pareto distribution" (Vehtari, Gelman and Gabry, 2017, 1415). Not only should the resulting smoothed weights perform better, the shape parameter of the fitted Pareto distribution can be used to check the reliability of the new importance weights. A large estimated shape parameter of the fitted Pareto distribution can indicate problems with the underlying distribution of the original importance samples. In that case, the estimates from the Pareto smoothed importance sampling may also be problematic. One immediate advantage is that one can then identify those problematic observations. This allows scholars to then estimate the full leave-on-out posterior for those observations that were identified as problematic. We can then directly sample from this actual leave-one-out posterior $p(\theta|y_{-k})$ for those observations. The full model evaluation would then be based on both psis-loo and full loo estimates. Claassen (2018) has recently used loo-cv for model evaluation in political science.

As Vehtari, Gelman and Gabry (2017) argue, psis-loo-cv has considerable *computational* advantages over exact leave-one-out cross-validation. It also performs better than

WAIC, traditional importance sampling, or truncated importance sampling loo-cv on a variety of models. For hierarchical models with few data points per group and high variation in the parameters between groups, however, the performance of WAIC and psis-loo decreases and exact loo-cv becomes more valuable.[19]

5.3 Application

We now return to our example application from above and compare the different models in terms of WAIC, 10-fold-cross-validation, and psis-loo. The *loo* package in *R* allows scholars to easily generate these model evaluation scores for models estimated in *Stan* (Vehtari et al., 2019). Moreover, the *loo* package is again integrated into the *brms* package and the information criteria scores are readily available for our estimated models.

The *loo* package produces the expected log predictive density (elpd) as well as the information criteria on the deviance scale (i.e., $-2 \times$ elpd) for all three information criteria. In Table 4 below we present the different information criteria scores on the deviance scale for the four models. First, in our simple example with 117 observations, two parameters per model, and an estimated Gaussian linear model, the values of the different information criteria for each model are quite similar to each other. For example, for Model 3 the WAIC score is 600.65, the psis-loo-ic score is 600.74, and the k-fold-ic score is 604. The same is true for the other three models, where the WAIC, psis-loo-ic, and k-fold-ic closely correspond with each other. In line with the results of the previous information criteria presented, Model 3 performs best, i.e., has the lowest information criteria scores.

[19]Similarly, for models with spatial or temporal dependence, psis-loo-cv is likely to be problematic. Ongoing work is considering possible approaches for these cases.

Table 4: Information Criteria for Evaluating Theories of Congress

	Model 1	Model 2	Model 3	Model 4
psis-loo-ic	687.6	650.1	600.7	605.3
WAIC	687.6	650.0	600.7	605.2
k-fold-ic	688.7	650.4	600.1	603.8

Readers might wonder how to decide whether the evidence in favor of a particular model is strong enough to make a claim about it being the *best* model. As mentioned above, for the Bayes factor a value of 10 or larger would, according to some, be considered *strong* evidence in favor of the better model. As a first step when comparing two models in terms of the information criteria, we can calculate the difference in their scores. If we want to evaluate whether Model 3 should be strictly preferred to Model 1, we calculate $\textit{psis-loo-ic}_1 - \textit{psis-loo-ic}_3 = 86.9$. Since smaller values on the scores are preferable, a negative value on the difference indicates support for the first model, whereas a positive value indicates support for the second model in the comparison. In Table 5 below we present the psis-loo-ic scores of the individual models in the top four rows, but in the bottom part of the table we present the calculated difference in psis-loo-ic scores between the four different models. Recall that positive values indicate a better score for the second model in the difference. As we can see, the second model is preferred for all comparisons except when comparing Model 3 and 4. The difference in psis-loo-ic scores again indicates that Model 3 performs better than Model 4. But is this difference large enough to strictly prefer Model 3?

Fortunately, the model evaluation criteria discussed here allow us to calculate uncertainty estimates, which can be used to judge whether differences between models are large enough to draw conclusions. Each of these methods are estimated using functions

applied to the individual observations in the data to create the information criteria; i.e., the total information criteria are combinations of n scores. Using the standard deviation of those n components one can estimate an approximate standard error for the information criteria. For example, in the case of psis-loo-cv, we can estimate a standard error based on the standard deviation of the n individual components of the expected log pointwise predictive density $\widehat{elpd_{i,loo}}$ (Vehtari, Gelman and Gabry, 2017, p 1426). Similarly, the individual components can be used to calculate an approximate standard error for the difference in information criteria scores.

The top half of Table 5 shows the estimated $\widehat{elpd_{loo}}$ score on the deviance scale for each of the four models as well as the estimated standard error. In the bottom half of the table we present the difference in psis-loo-ic scores and the standard error for each difference. There is no clear and hard rule about how large the difference compared to its standard error would have to be to conclude that a model is strictly superior. One might take the general rule of thumb that we would like to see a difference that is *at least* the size of twice its standard error. It has been suggested, however, that these standard error estimates are an optimistic approximation and, especially for smaller sample sizes, might not be appropriate (Vehtari, Gelman and Gabry, 2017). When model differences are sufficiently small, scholars may want to prefer the less complex that provides comparable fit. Again, both differences in information criteria and their standard errors are easily available in the *loo* package and integrated into the *brms* package.

Based on the results presented in Table 5, we would conclude that Model 3 is significantly better than Models 1 and 2. On the other hand, comparing Model 3 and Model 4 suggests that both do a similar job at explaining the variation in the data. That is, there

is no strong reason to prefer one to the other.

Table 5: Loo Information Criteria for Evaluating Theories of Congress with Standard Errors

Model	psis-loo-ic	SE
Model 1	687.6	17.5
Model 2	650.1	24.0
Model 3	600.7	24.1
Model 4	605.3	24.7
Differences and SEs		
$M_1 - M_2$	37.6	12.2
$M_1 - M_3$	86.9	16.4
$M_1 - M_4$	82.3	16.8
$M_2 - M_3$	49.3	9.5
$M_2 - M_4$	44.7	11.0
$M_3 - M_4$	-4.6	4.4

6 Finite mixture models, Bayesian model averaging, and stacking

In this section we turn to a somewhat different approach to handling multiple potential models. In particular, we consider statistical approaches where each of the competing models is considered a component of an overarching model. That is, we eschew the task of selecting or even comparing models and instead consider how much each *component* model contributes to the model combination. In the end, we may heuristically prefer the component that contributes the most, making this distinction seem somewhat arbitrary. However, from a statistical standpoint, the approaches we cover next can be quite distinct from those discussed above.

Specifically, we return to the finite model space approach discussed in Section 4.[20]

[20]For reasons of space, we confine ourselves here to finite mixture models with a known number of

From this perspective, the true data generating process is not "best" represented by any particular model, but rather by a combination of data generating processes. We can then engage in model selection either by choosing the model that receives the most weight in this mixture, or we can skip the task of model selection entirely and attempt to make inferences that actually reflect our uncertainty about the true DGP.

6.1 Mixture models

One family of models is to attempt to assign each observation to one of the potential candidate models and estimate $\boldsymbol{\theta}_k$ based on this assignment (Imai and Tingley, 2012; McLachlan and Peel, 2000). Let $p(y_i|\boldsymbol{\theta}_k, \mathcal{M}_k)$ represent the predictive distribution of observations i from model k and let $\boldsymbol{\tau} = [\tau_1, \tau_2, \ldots, \tau_k]$ index which model actually generated each observation such that $\tau_i \in [1, 2, \ldots, K] \forall i \in [1, 2, \ldots, n]$. We can construct our mixture model as:

$$p(y_i|\boldsymbol{\tau}, \boldsymbol{\theta}_k) \sim \sum_{k=1}^{K} p(y|\boldsymbol{\theta}_k)\mathcal{I}(\tau_i = k),$$

where $\mathcal{I}(\cdot)$ is the standard indicator function. We can then complete the model by placing appropriate priors over $\boldsymbol{\theta}_k$ and a hierarchical prior structure on $\boldsymbol{\tau}$,

$$\pi(\boldsymbol{\tau}) \sim \text{Multinomial}(\boldsymbol{\omega})$$

$$\pi(\boldsymbol{\omega}) \sim \text{Dirichlet}(\boldsymbol{\alpha}).$$

potential components. However, we note that there is also a large literature on Bayesian models that can relax or eliminate this assumption.

In this case the ω_i parameter represents the probability a generic observation is assigned to each model, and we can interpret the posterior estimate in a fashion similar to (but distinct from) the model weight shown in Equation 3. Likewise, we can look at the posterior distributions on the τ vector to get an estimate of how many observations are assigned to each model.

One important factor to note is that the model parameters for each component θ_k are estimated for observations "assigned" to that component using standard Bayesian methods. That is, we are imagining that all of the models are operating at the same time but that different units belong to each (Imai and Tingley, 2012).

In many cases this can be desirable, although it can increase uncertainty for components assigned few observations. Scholars must also be careful in how they interpret individual parameters as the parameters do not correspond to estimates for the entire population. Further, simultaneous estimation of model weights and model parameters can also lead to model degeneracy and identification problems during estimation. Standard regression models can be estimated in a fully Bayesian fashion using the `BayesMix` package in R or with alternative estimation routines (viz. the EM algorithm) using the `FlexMix` package. However, with attention to issues of identification, they can also be fit using `brms`.

6.2 EBMA

Ensemble Bayesian model averaging (EBMA), largely deriving from the literature on forecasting, is different in that we fit component models separately using the entire dataset and then combine them into a weighted ensemble (Raftery et al., 2005; Mont-

gomery, Hollenbach and Ward, 2012). In this case, we can divide our dataset into three partitions: a training set used to fit each individual model ($\mathbf{y}^{train}$), a calibration used to determine the model weights ($\mathbf{y}^{cal}$), and a true test set that we are hoping to accurately predict ($\mathbf{y}^{test}$). We assume that each of the component models is fit to the training data (although the component models need not be statistical models at all). The calibration set represents observations that were predicted out-of-sample by each component model and allows us to appropriately weight them without having to develop penalties for complexity. The goal is then to combine the forecasts in order to make accurate predictions of the test observations.

Let $w_k = p(\mathcal{M}_k|\mathbf{y}^{cal})$, and $p(\mathbf{y}^{test}|\mathcal{M}_k)$ represent the predictive pdf for the test set from model k. Our goal is then to generate an ensemble prediction,

$$p(\mathbf{y}^{test}) = \sum_{k=1}^{K} w_k p(\mathbf{y}^{test}|M_k).$$

To complete the model, therefore, we need to estimate the model weights for each component. Formally, we need to find the values of $\mathbf{w}$ that will maximize the log-likelihood function,

$$\mathcal{L}(\mathbf{w}, \mathbf{\Theta}) = \sum_{i=1}^{n^{cal}} \log \left\{ \sum_{k=1}^{K} w_k p(y^{cal}|\boldsymbol{\theta}_k) \right\},$$

subject to the constraint that $\sum \omega_k = 1$, which can be calculated efficiently using an EM in the `EBMAforecast` package in R.

6.3 BMA

As the name suggests, EBMA is closely related to Bayesian model averaging (Madigan and Raftery, 1994; Raftery, 1995; Bartels, 1997; Gill, 2004; Montgomery and Nyhan, 2010; Cranmer et al., 2017; Plümper and Traunmüller, 2018). In the traditional approach to BMA,[21] each model is again fit to the entire dataset. Model weights are calculated according to Equation 3.

Of course, this again leaves us with the problem of needing to estimate marginal likelihoods. One option is to use one of the several proxies discussed above, such as AIC or BIC. Another option is to select priors that allow for closed form solutions. For instance, Zellner's g-prior (Zellner, 1986) is:

$$\pi(\beta|\sigma^2) \sim \text{Normal}(\mathbf{0}, g\sigma^2(\mathbf{X}'\mathbf{X})^{-1}),$$

where $\mathbf{X}$ is some set of covariates and σ^2 is the variance for the residuals in a standard regression model.[22] This yields a marginal model likelihood of

$$p(\mathbf{y}|\mathbf{X}, \mathcal{M}_k) = \frac{\Gamma(n/2)}{\pi^{n/2}}(1+g)^{-p}S^{-1},$$

where S is a function only of the data and the prior mean for β.[23]

The result is that for any quantity of interest, we can simply construct a weighted ensemble from the full posterior. For instance, to estimate the posterior distribution of a

[21] For large model spaces, several scholars have developed stochastic samplers that simultaneously estimate the model probabilities and model parameters (George and McCulloch, 1993).

[22] We also place an improper prior on σ^2 of $\frac{1}{\sigma^2}$.

[23] See Chapter 5 in Ando (2010) for a complete proof.

specific regression coefficient, we need only calculate:

$$p(\beta|\mathbf{y}) = \sum_{k=1}^{K} p(\mathcal{M}_k|\mathbf{y})p(\beta_k|\mathbf{y}, \mathcal{M}_k) = \sum_{k=1}^{K} w_k p(\beta_k|\mathbf{y}, \mathcal{M}_k).$$

For models where this coefficient is excluded, we will then have a point mass at zero. The rest will create an ensemble posterior that reflects our uncertainty based on the set of covariates. It also allows us to focus on two separate quantities of interest that are commonly confused in interpreting regression analysis. First, we might be interested in the posterior probability that some particular variable should be in the model. This will be the sum of the model weights where that variable is included. Second, we might be interested in the distribution of β conditioned on the fact that it is included in the model, $p(\beta|\mathbf{y}, \beta \neq 0)$.

6.4 Stacking

BMA is based on the marginal likelihood of each model under an $\mathcal{M}$-closed assumption. Thus, there are several problems with the BMA approach that largely correspond to the issues with Bayes factors discussed above. First, model weights can be sensitive to prior specifications. Second, it is not always clear how to place priors on the probability of specific models, since seemingly innocuous assumptions can also affect weights in unintended ways. For instance, placing an agnostic prior that each coefficient has a 50% prior probability for inclusion biases the ensemble towards models that include 50% of the covariates. Finally, an underlying assumption of BMA is that the true model is included within the model space (i.e., the $\mathcal{M}$-closed case). When true, BMA will place

all of its posterior weight on this one model asymptotically.

Estimating model weights in stacking is done in a very different two-step process. First, the candidate models are estimated based on the data available. Second, model weights are calculated for each of the estimated candidate models. In the original stacking, model weights are generated by minimizing the leave-one-out mean-squared-error for each observation based on each model's point prediction. Ergo, stacking based only on the point estimate and not the full predictive distribution, was not considered as a true alternative to Bayesian model averaging (Yao et al., 2018). However, one of the advantage of stacking is its appropriateness in the $\mathcal{M}$-open setting; i.e., a true model does not have to exist, let alone be in the estimated model space.

Yao et al. (2018) built on several recent developments to further develop stacking to use the full leave-one-out predictive distribution instead of the point estimates. Adjusting the notation in Yao et al. (2018), let

$$\hat{p}_{ki}(\tilde{y}_i) = \int_{\Theta} p(\tilde{y}_i|\theta_k, \mathcal{M}_k) p(\boldsymbol{\theta}_k|\mathbf{y}, \mathcal{M}_k) d\theta.$$

Applying a logarithmic scoring rule,[24] we can then find the stacking weights for each model by solving the optimization problem:

$$\max_{w} \frac{1}{n} \sum_{i=1}^{n} \log \sum_{k=1}^{K} w_k \hat{p}_{ki}(\tilde{y}_i).$$

If we adopt the psis-loo-ic approximations above, we can simplify this further to give us weights based on estimates of the elpd, giving us stacking weights that can be con-

[24] We surpress discussion of alternaive scoring rules for simplicity.

structed from a single posterior but which reflect each models' out-of-sample properties.[25]

Stacking has the advantage in that it does not assume $\mathcal{M}$-closed, but rather allows that the true DGP may not be well represented by any of the candidate models. Additionally, because weights are calculated based on the unit-specific elpd for each model, stacking takes into account when models have different strengths in predicting certain observations. One of the advantages of stacking is, therefore, that it is able to combine weights from models that are very similar to the better model, instead of splitting the weight between very similar models as often occurs in BMA.[26]

6.5 Application

As a last example based on our application, we generate stacking model weights based on the psis-loo-ic scores presented above. Stacking can be easily done using the *loo* package (Vehtari et al., 2019) and integrated into *brms* (Bürkner, 2017). Once the psis-loo-ic scores are calculated, stacking weights are readily available. Additionally, a warning would be given if the Pareto smoothed importance sampling is questionable for some observations and full leave-one-out resampling is then suggested for those observations.

As we can see in Table 6 below, stacking weights are highly concentrated on Model 3, even though based on the information criteria Model 3 and 4 performed quite similar (see Table 5 above). This suggests that while BMA may have split the weight between Models 3 and 4 due to their similar performance, stacking combines the weights to the

[25] This last variant is referred to as pseudo-BMA.

[26] Interested readers may want to consult the the original article introducing stacking via psis-loo (Yao et al., 2018) and the full discussion appended to the article.

slightly better model. While Model 3 may be only slightly better than Model 4, the two models seem to be quite similar in the cases they predicted well, thus adding weight to Model 4 beyond Model 3 must not improve predictive accuracy. The remaining weight (0.04) is assigned to Model 1. In this case, this model is weighted poorly because it is not as accurate as the others. However, it is still weighted more highly than Model 2 and 4 because it gives sufficiently different predictions.

Table 6: Model Weights Based on Stacking

Model 1	Model 2	Model 3	Model 4
0.04	0.00	0.96	0.00

7 Conclusion

The current state of the literature when it comes to testing competing models is deeply unsatisfactory. In many cases the conclusions researchers draw from their analysis are sensitive to modeling choices. Traditional null hypothesis testing strategies in particular lead analysts to focus excessively or even exclusively on whether or not key variables are "significant" in the models they consider. However, significance can change easily depending on parametric assumptions, the set of covariates included, and more. In total, in many cases competing theories are not meaningfully tested against each other, and the conclusions we draw from our data are driven by modeling choices irrelevant or orthogonal to the scientific question at hand. Even worse, the most common practice we observe in the literature – simply tossing all variables from all theories into a single regression – is known to produce incorrect and misleading conclusions. In all, we are left with a picture that for a fundamental task for science – arbitrating between competing

theories – standard practices are pushing researchers towards either building incorrect garbage-can models or haphazardly exploring the potential space of models with little guidance other than important p-values.

Throughout this chapter, we have provided researchers with a set of alternative strategies for arbitrating between theories and models. The advantage of the Bayesian framework in tackling this problem is that we can use the laws of probability to think clearly about model probabilities directly. There are no null hypotheses required and no misleading p-values. Instead, we can talk meaningfully about whether a given model is better either in terms of Bayes factors or KL divergence (an optimal criteria implied by Bayesian decision theory). While imperfect, therefore, we believe that this set of tools offers a superior approach to arbitrating among theories than standard practices in the discipline.

With that said, it is important to keep the limitations of these methods in mind. In general, the appropriateness of these model evaluation techniques depends on the specific model specification and settings in which they are used. Further, the use of information criteria or predictive accuracy for model selection, for example, should not be a substitute for theoretical considerations of which covariates are important to include given the question asked. Similarly, these model selection techniques are not able to distinguish between pre- and post-treatment variables. In fact, full garbage-can models may perform better on these scores, even though the associated results are not theoretically meaningful. Thus, as with all methodological tools, they should be used with the necessary understanding of the actual problem being considered and theories of how the data was generated.

Finally, given recent advances in this area in the field of statistics, we hope that this overview will renew attention in the field of political science to these questions. Surely, none of the methods above represent the last word on this topic. Nor does our presentation take into account many practical difficulties applied scholars face in practice such as clustering, spatial correlations, time-series, confounding requiring an identification strategy, and measurement error. Further research is needed to evaluate the usefulness of these various methods in these circumstances.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining causal findings without bias: Detecting and assessing direct effects." *American Political Science Review* 110(3):512–529.

Achen, Christopher H. 2005. "Let's put garbage-can regressions and garbage-can probits where they belong." *Conflict Management and Peace Science* 22(4):327–339.

Ando, Tomohiro. 2010. *Bayesian Model Selection and Statistical Modeling*. New York: CRC Press.

Bagozzi, Benjamin E. and Kathleen Marchetti. 2017. "Distinguishing Occasional Abstention from Routine Indifference in Models of Vote Choice." *Political Science Research and Methods* 5(2):277–294.

Bartels, Larry M. 1997. "Specification, Uncertainty, and Model Averaging." *American Journal of Political Science* 41(2):641–74.

Berger, J.O., J.K. Ghosh and N. Mukhopadhyay. 2003. "Approximations to the Bayes factor in model seletion problems and consistency issues." *Journal of Statistical Planning and Inference* 112:241–258.

Betancourt, Michael. 2017. "A Conceptual Introduction to Hamiltonian Monte Carlo." *CoRR* .

URL: *https://arxiv.org/pdf/1701.02434.pdf*

Brady, David W and Craig Volden. 1998. *Revolving gridlock: Politics and policy from Carter to Clinton*. Westview Pr.

Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models using Stan." *Journal of Statistical Software* 80(1):1–28.

Bürkner, Paul-Christian. In Press. "Advanced Bayesian Multilevel Modeling With The R Package BRMS." *The R Journal* .

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li and Allen Riddel. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software, Articles* 76(1):1–32.

Chib, Siddhartha. 1995. "Marginal likelihood from the Gibbs output." *Journal of the american statistical association* 90(432):1313–1321.

Claassen, Christopher. 2018. "Estimating Smooth Country–Year Panels of Public Opinion." *Political Analysis* In Press.

Clarke, Kevin A. 2001. "Testing Nonnested Models of International Relations: Reevaluating Realism." *American Journal of Political Science* 45(3):724–744.

Clarke, Kevin A. 2007. "A Simple Distribution-Free Test for Nonnested Model Selection." *Political Analysis* 15(3):347–363.

Cox, Gary W. and Matthew D. McCubbins. 2005. *Setting the Agenda: Responsible Party Government in the US House of Representatives*. New York: Cambridge University Press.

Cranmer, Skyler J, Philip Leifeld, Scott D McClurg and Meredith Rolfe. 2017. "Navigating the range of statistical tools for inferential network analysis." *American Journal of Political Science* 61(1):237–251.

Desmarais, Bruce A. and Jeffrey J. Harden. 2013. "An Unbiased Model Comparison Test Using Cross-Validation." *Quality & Quantity* 48(4):2155–2173.

Gelfand, A.E. and Dey D.K. 1994. "Bayesian model choice: Asymptotics and exact calculations." *Journal of the Royal Statistical Society. Series B (Methodological)* 56:510–514.

Gelfand, Alan E. 1996. Model Determination Using Sampling-Based Methods. In *Markov Chain Monte Carlo In Practice* , ed. W.R. Gilks, Sylvia Richardson and David J. Spiegelhalter. Boca Raton, FL: Chapman & Hall chapter 9, pp. 145–162.

Gelfand, Alan E., Dipak K. Dey and Hong Chang. 1992. Model Determination Using Predictive Distributions With Implementation Via Sampling-Based Methods. In *Bayesian Statistics*, ed. José M. Bernardo, James O. Berger, Dawid A. Philip and Adrian F.M. Smith. Vol. 4 Oxford University Press pp. 147–167.

Gelman, Andrew and Donald B. Rubin. 1995. "Avoiding Model Selection in Bayesian Social Research." *Sociological Methodology* 25(nil):165.
URL: *https://doi.org/10.2307/271064*

Gelman, Andrew, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari and Donald B Rubin. 2013. *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, Andrew, Jessica Hwang and Aki Vehtari. 2014. "Understanding predictive information criteria for Bayesian models." *Statistics and computing* 24(6):997–1016.

Gelman, Andrew and Xiao-Li Meng. 1998. "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling." *Statistical science* pp. 163–185.

George, Edward I and Robert E McCulloch. 1993. "Variable selection via Gibbs sampling." *Journal of the American Statistical Association* 88(423):881–889.

Gill, Jeff. 2004. "Introduction to the Special Issue." *Politi* 12(4):647–674.

Gill, Jeff. 2009. *Bayesian methods: A Social and Behavioral Sciences Approach*. Boca Raton, FL: CRC Press.

Green, Peter J. 1995. "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika* 82(4):711–732.

Gronau, Quentin F, Alexandra Sarafoglou, Dora Matzke, Alexander Ly, Udo Boehm, Maarten Marsman, David S Leslie, Jonathan J Forster, Eric-Jan Wagenmakers and Helen Steingroever. 2017. "A tutorial on bridge sampling." *Journal of mathematical psychology* 81:80–97.

Gronau, Quentin F, Henrik Singmann and Eric-Jan Wagenmakers. 2017. "Bridgesampling: an r package for estimating normalizing constants." *arXiv preprint arXiv:1710.08162* .

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2016. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second ed. New York, NY: Springer.

Imai, Kosuke and Dustin H. Tingley. 2012. "A Statistical Method for Empirical Testing of Competing Theories." *American Journal of Political Science* 56(1):218–236.

Jeffreys, H. 1961. *Theory of Probability*. Oxford: Oxford University Press.

Kim, In Song, John Londregan and Marc Ratkovic. 2018. "Estimating Spatial Preferences From Votes and Text." *Political Analysis* 26(2):210–229.

Konishi, S., T. Ando and S. Imoto. 2004. "Bayesian information criteria and smoothing parameter seletion in radial asis function networks." *Biometrika* 91:27–43.

Krehbiel, Keith. 1998. *Pivotal Politics: A Theory of U.S. Lawmaking*. Chicago: University of Chicago Press.

Kung, James Kai-sing. 2014. "The Emperor Strikes Back: Political Status, Career Incentives and Grain Procurement during China's Great Leap Famine." *Political Science Research and Methods* 2(2):179–211.

Madigan, David and Adrian E. Raftery. 1994. "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window." *Journal of the American Statistical Association* 89(428):1535–1546.

McLachlan, Geoffrey and David Peel. 2000. *Finite Mixture Models*. New York, NY: John Wiley & Sons, Ltd.

Meng, Xiao-Li and Stephen Schilling. 2002. "Warp bridge sampling." *Journal of Computational and Graphical Statistics* 11(3):552–586.

Meng, Xiao-Li and Wing Hung Wong. 1996. "Simulating ratios of normalizing constants via a simple identity: a theoretical exploration." *Statistica Sinica* pp. 831–860.

Montgomery, Jacob M. and Brendan Nyhan. 2010. "Bayesian Model Averaging: Theoretical Developments and Practical Applications." *Political Analysis* 18(2):245–270.

Montgomery, Jacob M, Brendan Nyhan and Michelle Torres. 2018. "How conditioning on posttreatment variables can ruin your experiment and what to do about it." *American Journal of Political Science* 62(3):760–775.

Montgomery, Jacob M., Florian M. Hollenbach and Michael D. Ward. 2012. "Improving Predictions Using Ensemble Bayesian Model Averaging." *Political Analysis* 20(3):271–291.

URL: *http://pan.oxfordjournals.org/content/20/3/271.abstract*

Neal, Rashford M. 2011. MCMC Using Hamiltonian Dynamics. In *Handbook of Markov chain Monte Carlo*, ed. Steve Brooks, Andrew Gelman, Galin L. Jones and Xio-Li Meng. Boca Raton, FL: CRC Press chapter 5, pp. 113–162.

Piironen, Juho and Aki Vehtari. 2017*a*. "Comparison Of Bayesian Predictive Methods For Model Selection." *Statistics and Computing* 27(3):711–735.

Piironen, Juho and Aki Vehtari. 2017*b*. On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior. In *Artificial Intelligence and Statistics*. pp. 905–913.

Plümper, Thomas and Richard Traunmüller. 2018. "The Sensitivity Of Sensitivity Analysis." *Poltical Science Research and Methods* In Press.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25(HUH):111–163.

Raftery, Adrian E., Tillman Gneiting, Fadoua Balabdaoui and M. Polakowski. 2005. "Using Bayesian Model Averaging to Calibrate Forecast Ensembles." *Monthly Weather Review* 133:1155–1174.

Richman, Jesse. 2011. "Parties, pivots, and policy: the status quo test." *American Political Science Review* 105(1):151–65.

Schwarz, G. 1978. "Estimating the dimension of a model." *Annals of Statistics* 6:461–464.

Spiegelhalter, David J, Nicola G Best, Bradley P Carlin and Angelika Van Der Linde. 2002. "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(4):583–639.

Team, Stan Development. 2017. "Stan Modeling Language User's Guide and Reference Manual: Stan Version 2.17.0." https://github.com/stan-dev/stan/releases/download/v2.17.0/stan-reference-2.17.0.pdf (Accessed September 2017).

Vehtari, Aki, Andrew Gelman and Jonah Gabry. 2017. "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC." *Statistics and Computing* 27(5):1413–1432.

Vehtari, Aki, Jonah Gabry, Yuling Yao and Andrew Gelman. 2019. *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. R package version 2.1.0.
URL: *https://CRAN.R-project.org/package=loo*

Vehtari, Aki and Jouko' Lampinen. 2002. "Bayesian Model Assessment And Comparison Using Cross-Validation Predictive Densities." *Neural Computation* 14(10):2439–2468.

Watanabe, Sumio. 2010. "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." *Journal of Machine Learning Research* 11(Dec):3571–3594.

Watanabe, Sumio. 2013. "A widely applicable Bayesian information criterion." *Journal of Machine Learning Research* 14(Mar):867–897.

Yao, Yuling, Aki Vehtari, Daniel Simpson, Andrew Gelman et al. 2018. "Using stacking to average Bayesian predictive distributions." *Bayesian Analysis* .

Zellner, Arnold. 1986. "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." *Bayesian inference and decision techniques* .