Bayesian Model Averaging: Theoretical Developments and Practical Applications

Jacob M. Montgomery

Department of Political Science, Duke University, 326 Perkins Library, Box 90204, Durham, NC 27708
e-mail: jmm61@duke.edu

Brendan Nyhan

School of Public Health, University of Michigan, 1420 Washington Heights, M2208 SPH-II, Ann Arbor, MI 48109
e-mail: bnyhan@umich.edu (corresponding author)

Political science researchers typically conduct an idiosyncratic search of possible model configurations and then present a single specification to readers. This approach systematically understates the uncertainty of our results, generates fragile model specifications, and leads to the estimation of bloated models with too many control variables. Bayesian model averaging (BMA) offers a systematic method for analyzing specification uncertainty and checking the robustness of one's results to alternative model specifications, but it has not come into wide usage within the discipline. In this paper, we introduce important recent developments in BMA and show how they enable a different approach to using the technique in applied social science research. We illustrate the methodology by reanalyzing data from three recent studies using BMA software we have modified to respect statistical conventions within political science.

1 Introduction

Uncertainty about the ''correct'' model specification can be high in political science research. Classical methods offer researchers little guidance and few useful tools for dealing with this uncertainty. As a result, scholars often engage in haphazard searches of possible model configurations, a practice that can lead to incorrect inferences, fragile reported findings, and publication bias.

A better approach is Bayesian model averaging (BMA), which was introduced to political scientists by Bartels (1997) but has not come into wide use in the discipline. As

---

a result, ''a remarkable evolution'' in BMA methodology has been completely overlooked (Clyde and George 2004). In this paper, we explain how BMA can help applied researchers to ensure that their estimates of the effects of key independent variables are robust to a wide range of possible model specifications.

Our presentation follows in four stages. First, we review existing approaches to addressing model uncertainty and discuss their limitations. Next, we summarize recent developments in BMA methodology and advocate an approach to the technique that differs substantially from Bartels (1997). To facilitate usage of BMA, we have developed modified software code that addresses several obstacles to its use in applied social science research. We then illustrate our approach by reanalyzing data from three recent articles: the Adams, Bishin, and Dow (2004) study of proximity and directional voting in U.S. Senate elections, the Canes-Wrone, Brady, and Cogan (2002) study of the electoral consequences of extremist roll call voting in the U.S. House, and the Fearon and Laitin (2003b) analysis of civil war onset internationally. Finally, we conclude with words of caution about appropriate applications of the technique.

2 The Problem of Model Uncertainty

Political scientists who analyze observational data frequently encounter uncertainty about what variables to include in their statistical models. A typical researcher develops theory about a few key explanatory variables and then must choose from a set of possible control variables over which she has much weaker prior beliefs. In such cases, the appropriate set of control variables is often highly uncertain. As a result, researchers frequently estimate a variety of models before selecting one to include in the published version of their research.

This practice leads to a number of pathologies. First, it understates our uncertainty about the effects of the variables of interest. Basing inferences on a single model implicitly assumes that the probability that the reported model generated the data is 1, an assumption that is surely mistaken. Second, some researchers may search the model space until they find a specification in which a key variable is statistically significant, a practice that has led to indications of publication bias in top journals (Gerber and Malhotra 2008). As a result, reported results are often fragile to slight variations in model specification. Finally, the perceived necessity to control for large numbers of potential confounds has led to bloated specifications that decrease efficiency without necessarily decreasing omitted variable bias (Clarke 2005).

Addressing this problem is difficult because classical methods offer few tools for handling model uncertainty. Researchers who wish to test the robustness of their findings often estimate a handful of alternative models to see whether the sign and/or significance level of key coefficients change. However, these tests are conducted in a haphazard manner. In addition, frequentist hypothesis testing offers no method for resolving conflicting findings across alternative specifications. What is one to infer if a variable is significant in some specifications but fails to pass traditional thresholds in others?

Analysts may also try more formal methods to substantiate the models they report. Typically, researchers compare nonnested models using frequentist tests, such as the Cox and Vuong tests, select models based on a model fit statistic that penalizes complexity, such as the Bayesian information criterion (BIC), or compare nested models using likelihood-ratio tests.[1] Methodological objections can be raised concerning the limitations of each of these

---

[1]Previously, some researchers resorted to stepwise variable selection to find the ''best'' model when uncertainty is pervasive, but it is now commonly understood that this technique leads to upward bias in R^2 and estimated coefficients, downward bias in standard errors, and incorrect p values (Harrell 2001, 56–57).

techniques (Clarke 2001). But at a more philosophical level, we believe that the enterprise of searching for a ''best'' model is inappropriate to most political science data, which rarely yield clear proof that one specification is the ''true model.''

In addition, both approaches described above share a deeper underlying problem—the size of the potential model space. A model with p independent variables implies 2^p possible specifications. Uncertainty about even a few control variables thus makes it extremely difficult to ensure robustness to alternative specifications within a frequentist framework. Given the relatively large model space associated with even a modest number of variables, model uncertainty becomes a serious issue. At present, there is no way to combine the results of multiple hypothesis tests into more general measures of uncertainty over coefficients and/or models using frequentist techniques.[2]

3 BMA: An Overview

A more comprehensive approach to addressing model uncertainty is BMA, which allows us to assess the robustness of results to alternative specifications by calculating posterior distributions over coefficients and models. BMA came to prominence in statistics in the mid-1990s (Madigan and Raftery 1994; Draper 1995; Raftery 1995) and has expanded into fields such as economics (Fernandez, Ley, and Steel 2001), biology (Yeung, Bumgarner, and Raftery 2005), ecology (Wintle et al. 2003), and public health (Morales et al. 2006). (The state of research in the field is most recently summarized in Hoeting et al. 1999, Clyde 2003, and Clyde and George 2004.)

BMA is particularly useful in three specific contexts that we illustrate in our empirical examples below. First, BMA can be helpful when a researcher wishes to assess the evidence in favor of two or more competing measures of the same theoretical concept, particularly when there is also significant uncertainty over control variables. Second, when there is uncertainty over control variables, researchers can use BMA to test the robustness of their estimates more systematically than is possible under a frequentist approach. Finally, BMA may also be valuable for researchers who wish to estimate the effects of large numbers of possible predictors of a substantively important dependent variable (though there are important reasons to be cautious about the conclusions one can draw from such an approach). As we discuss below, recent methodological innovations have increased the usefulness of BMA in all these contexts.

3.1 *A Brief Review of BMA*

We first briefly review the basic theory of BMA in a linear regression context (following Clyde 2003), which provides the necessary vocabulary for the discussion of innovations in BMA methodology in the next section.[3] Let X denote the $n \times p$ matrix of all the

---

[2]The problems with idiosyncratic model specifications described above are related to the problem that King and Zeng (2006) call ''model dependence,'' which Ho et al. (2007) recommend addressing by estimating the treatment effect of a single-binary variable. Under this approach, researchers should drop observations for which appropriate counterfactuals are missing and use nonparametric matching to improve covariate balance. Although there are many good reasons to recommend this approach, it may not always be appropriate. For instance, some researchers are interested in continuous treatment variables or more than one independent variable. Others may have substantive reasons to prefer to estimate the most robust possible model for a full sample rather than dropping observations. Finally, some researchers will lack the sample size necessary to get good matches on relevant covariates. In all these cases, BMA is a potentially useful tool for improving the robustness of reported results.

[3]The approach described here extends naturally to generalized linear models. For the purposes of this article, we will assume that the functional form is known and that the standard linear regression assumptions are satisfied. Researchers who are concerned about serious violations of model assumptions should resolve these issues before employing BMA (or not using BMA at all).

independent variables theorized to be predictors of outcome Y.[4] Standard analyses would assume that $Y = X\beta + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2 I)$. However, we might have uncertainty about which of the $q = 2^p$ model configurations from the model space $\mathcal{M} = [\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_q]$ is the correct model.

The purpose of BMA is to explicitly incorporate this uncertainty into our model and therefore our inferences. The standard BMA approach represents the data as coming from a hierarchical mixture model. We begin by assigning a prior probability distribution to the model parameters β and σ^2 and the models $\mathcal{M}_k$. The model, $\mathcal{M}_k$, is assumed to come from the prior probability distribution $\mathcal{M}_k \sim \pi(\mathcal{M}_k)$, and the vector of model parameters is generated from the conditional distributions $\sigma^2|\mathcal{M}_k \sim \pi(\sigma^2|\mathcal{M}_k)$ and $\beta_\omega|\sigma^2,\mathcal{M}_k \sim \pi(\beta_\omega|\mathcal{M}_k,\sigma^2)$, where $\Omega = \omega_1, \ldots, \omega_p$ represents a vector of zeroes and ones indicating the inclusion (or exclusion) of variables in model $\mathcal{M}_k$.

Using this notation allows us parameterize the data-generating process using the following conditional model: $Y|\beta_\omega,\sigma^2,\mathcal{M}_k \sim N(X_\omega\beta_\omega,\sigma^2 I)$. The marginal distribution of the data under model M_k can therefore be written as follows:

$$p\left(Y\Big|\mathcal{M}_k\right) = \int \int p(Y|\beta_\omega, \sigma^2, \mathcal{M}_k)\pi(\beta_\omega|\sigma^2, \mathcal{M}_k)\pi(\sigma^2|\mathcal{M}_k)\mathrm{d}\beta_\omega\,\mathrm{d}\sigma^2. \tag{1}$$

The posterior probability of model M_k[5] is

$$p(\mathcal{M}_k|Y) = \frac{p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum\limits_{k=0}^{q} p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}. \tag{2}$$

Equation (2) provides a coherent way of summarizing model uncertainty after observing the data. For instance, we can easily derive the expected value for a given coefficient β after averaging across the model space $\mathcal{M}$:

$$E(\beta|Y) = \sum_{k=0}^{q} p(\mathcal{M}_k|Y)E(\beta_k|\mathcal{M}_k, Y). \tag{3}$$

$E(\beta|Y)$ represents the coefficient's weighted expected value across every possible model configuration (with the weights determined by our priors and the performance of the models).

---

[4]For the purposes of exposition, the constant is ignored in this discussion, which is equivalent to assuming that all variables in X have been centered at their means.
[5]In practice, the calculations of these quantities use Bayes's factors (Jeffreys 1935, 1961), a method for assessing the evidence in favor of two competing models, to compare each model with either the null model or the full specification (see, e.g., Kass and Raftery 1995). The reason for doing so is the appealing simplicity of calculating the Bayes's factor for each possible model against some base model, $\mathcal{M}_j$, rather than directly calculating the posterior probability for each model. Using Bayes's rule, we can show that the posterior odds of some model $\mathcal{M}_k$ to $\mathcal{M}_j$ can be calculated as $\frac{p(\mathcal{M}_k|Y)}{p(\mathcal{M}_j|Y)} = \frac{p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}{p(Y|\mathcal{M}_j)\pi(\mathcal{M}_j)}$, which is the Bayes's factor. As Clyde and George (2004, 82) point out, the posterior model probabilities in equation (2) above can then be expressed using only Bayes's factors $B[k{:}j]$ and prior odds $O[k{:}j]$: $p(\mathcal{M}_k|Y) = \frac{B[k:j]O[k:j]}{\sum_k B[k:j]O[k:j]}$.

3.2 *Publicly Available BMA Software*

The difficulties associated with implementing the BMA approach are primarily computational. Calculating any statistic of interest involves solving or approximating $p(\mathcal{M}_k|Y)$, which is often an intractable high-dimensional integral, for all $q = 2^p$ models under consideration. Given modest numbers of plausible covariates, even standard *Markov chain Monte Carlo (*MCMC) approaches become increasingly impractical as the model space expands. These computational difficulties led many early researchers to adopt simplifying assumptions and techniques that made BMA analyses more tractable but required significant trade-offs.

Since then, BMA computation has been radically improved. The combination of increased computing power, the development of more analytically tractable prior specifications, and the distribution of the BMA and Bayesian adaptive sampling (BAS) packages for R have made these techniques far more accessible. Nonetheless, both packages still have important limitations. We have therefore modified them for use in applied social science research (as discussed below) and will release our code for public use.[6]

4 A New Approach to BMA in Political Science

Since Bartels (1997) first introduced BMA to political science, applications of the technique within the discipline have been surprisingly rare. Gill (2004) provides a more recent overview of the approach, but the only published applications we have been able to locate are Bartels and Zaller (2001), Erikson, Bafumi, and Wilson (2001), Zaller (2004), Imai and King (2004), and Geer and Lau (2006). However, the BMA literature has developed substantially since 1997 and new software programs have become available. In this section, we discuss limitations of previous research (including the analysis in Bartels 1997, the most prominent presentation of the technique in the field) and propose a revised approach to using BMA in applied research.

4.1 *Interpreting Posterior Distributions Using Coefficient Plots*

Previous presentations of BMA in political science have placed a disproportionate emphasis on posterior summary statistics that can be reductive or misleading. For instance, Bartels relies almost exclusively on simple hypothesis tests using posterior means and standard deviations. But as Erikson, Wright, and McIver (1997) point out, Bartels computes t statistics for model-averaged coefficients that are invalid given the often irregular shapes of BMA posterior distributions. Bartels (1997), who initially described the t statistics as for "descriptive purposes only" (654), later conceded this point (Bartels 1998), noting that

---

[6]Both packages are freely modifiable under the GNU General Public License. The better known of the two is the BMA package (Raftery, Painter, and Volinsky 2005; Raftery et al. 2009), which covers linear regression, generalized linear models, and survival models. Although the package is very useful, it has several important limitations, including an ad hoc model selection criterion that may bias posterior estimates and an exclusive reliance on the BIC prior (Clyde 1999). Clyde's BAS package (Clyde 2009) improves on the BMA package in several important respects—it uses a stochastic model search algorithm that outperforms naïve sampling without replacement and MCMC model averaging algorithms in a variety of contexts (Clyde, Ghosh, and Littman 2009); it can search very large model spaces; and it offers a variety of prior specification options. However, BAS can only estimate ordinary least squares (OLS) models at the present time. We therefore recommend that applied analysts use the BAS package for OLS and the BMA package for generalized linear models and survival models.
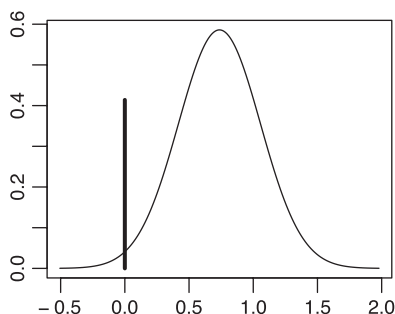
Fig. 1   Sample BMA posterior coefficient plot. The vertical line located at 0 on the *x* axis represents the posterior probability of models that exclude the variable ($p(\beta = 0 \mid Y)$). The density represents the conditional posterior probability for the parameter in those models in which it is included ($p(\beta \mid \beta \neq 0, Y)$).

"the posterior distribution of each parameter under the assumptions in my article is a mixture of normal distributions . . . and this mixture of normal distributions will not, in general, be a normal distribution" (18).

Rather than relying on summary statistics, the best way to understand the properties of posterior distributions is to plot them for each parameter, which is now trivial with publicly available software. Figure 1 illustrates what a coefficient posterior plot looks like. These plots allow us to answer two distinct questions:

1. Does the variable contribute to the model's explanatory power? (i.e., what is the posterior probability of all models that include this variable?)
2. Is it correlated with unexplained variance when it is included? (i.e., what is the conditional posterior distribution assuming that the variable is included?)

The vertical line located at 0 on the *x* axis represents the cumulative posterior probability of all models that *exclude* the relevant variable. One minus this value is the posterior probability of inclusion, $p(\beta \neq 0|Y)$, which can be used to answer question 1 above. The conditional posterior distribution, which is also included in the plot, represents the estimated value of the coefficient in the models in which it is included weighted by the likelihood of those models, $p(\beta|\beta \neq 0, Y)$. The location and density of this distribution allow us to answer question 2 above.

A related point is that BMA encourages researchers to be more clear about their statistical hypotheses. In practice, many scholars may wish to distinguish between the conditional posterior distribution and the posterior probability of inclusion depending on their goals and the nature of the data. For instance, some scholars are primarily interested in whether an independent variable is strongly correlated with a dependent variable across a range of potential model configurations. In such cases, BMA allows researchers to calculate $p(\beta > 0|\beta \neq 0, Y)$ or $p(\beta < 0|\beta \neq 0, Y)$ for the conditional posterior distribution, an option that we have added to the BMA and BAS packages. Alternatively, a scholar who is more interested in prediction (say, a scholar of interstate war) may want to know whether a predictor adds to the explanatory power of statistical models for a given dependent variable (or whether it offers more explanatory power than some alternate concept). In this context, it might be appropriate to focus on the posterior probability of inclusion. Finally,

other researchers may wish to consider both metrics and use the combined posterior distribution $p(\beta|Y)$.[7]

4.2 *Searching the Full Model Space*

A second major difference in our approach is that we advise researchers to consider the full set of 2^p possible models when conducting model averaging (excluding those that are theoretically or statistically inappropriate as described below). Some early presentations of BMA focused on averaging across very small subsets of the model space. For instance, in the two examples he presents, Bartels limits his model averaging to a handful of model specifications reported in published work, which implicitly places a zero prior on all other possible models. He concedes that his approach "can provide only a rough reflection of real specification uncertainty" but argues that it reflects the "substantive insight" of researchers (1997, 667–670).

However, putting a nonzero prior probability on only a handful of models when using BMA is almost always a mistake. Substantively, it typically will overstate our certainty that the included models are the only possible choices. In addition, such restrictions cripple the greatest strength of BMA—its ability to systematically search a model space and present posterior estimates that incorporate uncertainty in the model specification. Even Erikson, Wright, and McIver (1997)—the authors of one of the articles whose models were reanalyzed—dissent, noting that "the original model averaging literature is unambiguously clear in its rule that *all* models involving plausible variables must be considered."[8] Previously, researchers might have been forced to restrict the model space due to computational limitations, but the innovations in BMA software discussed above have made it possible to analyze large numbers of covariates.

4.3 *Alternative Prior Specifications to BIC*

In addition, most early BMA research, including Bartels (1997), approximated Bayes's factors using the BIC (Raftery 1995, 129–133).[9] Although this approach was computationally convenient, its consequences were not always desirable. For instance, BIC tends to place a relatively high posterior probability on sparse models (Kass and Raftery 1995; Erikson, Wright, and McIver 1997; Kuha 2004), a model prior that is not always substantively appropriate. In addition, though advocates of BIC argue that it is a reasonable approximation of the Bayes's factor under a unit information prior (Raftery 1995, 129–133), Gelman and Rubin (1995) note that BIC does not correspond to a proper Bayesian prior distribution (see also Weakliem 1999).

However, other prior specifications are now available to applied researchers. In conjunction with advances in techniques for sampling large model spaces, these new priors

---

[7]Current practices in the discipline rely heavily on p values, which awkwardly conflate these two concepts (Gill 1999). Separating them allows for useful distinctions in variable performance. For instance, it is possible to have variables that are "statistically significant" (i.e., their credible intervals do not overlap with zero) but have low posterior probabilities of inclusion. Likewise, it is possible for a variable with a high posterior probability of inclusion to have a model-averaged credible interval that overlaps with zero due to variation in sign and significance across models.

[8]Searching such a limited model space may also lead to an unwarranted emphasis on the selection of the "best" model, which is generally of limited substantive interest.

[9]The BIC for model $\mathcal{M}_k$ compared with the null model $\mathcal{M}_0$ is $\text{BIC}_k = -2\log(L_k - L_0) + p\log n$, where L_k is the maximized likelihood for $\mathcal{M}_k$ and p is the number of parameters in the model.

have allowed researchers to significantly improve the flexibility and power of BMA techniques while avoiding shortcuts such as BIC and Akaike information criterion (AIC).[10]

One option in the BAS package that has appealing properties is Zellner's g-prior (Zellner 1986), which is formulated as follows:

$$\pi\left(\beta_\omega \Big| \mathcal{M}_k, \sigma^2\right) \sim N_{p_\omega}\left(0, g\sigma^2\left(X'_\omega X_\omega\right)^{-1}\right) \qquad (4)$$

and

$$\pi\left(\beta_0, \sigma^2 \big| \mathcal{M}_k\right) \propto 1/\sigma^2 \qquad (5)$$

for some positive constant g where p_ω represents the number of predictor variables in the ωth model.[11] It yields closed form expressions for $p(Y|\mathcal{M}_k)$ that are rapidly calculable and requires the choice of only one hyperparameter, simplifying the prior specification process.[12]

Alternatively, one can place a hyper-prior on g to avoid possible misspecification. Here, we introduce two such hyperpriors for linear regression, which are analyzed in Liang et al. (2008) and available for use in the BAS package. The first, the so-called "hyper-g," puts the following hyperprior on g:

$$\pi(g) = \frac{a-2}{2}(1+g)^{\frac{a}{2}} \text{ for } = g > 0. \qquad (6)$$

Liang et al. (2008) use example values of 3 or 4 for a when specifying the hyper-g but state that values of $2 < a \leqslant 4$ are "reasonable" (the distribution is proper when $a > 2$). A related approach is the Zellner-Siow prior (Zellner and Siow 1980). To create this prior, we put a Gamma(1/2, n/2) prior on g, which induces a multivariate Cauchy prior on β_ω:

$$\pi\left(\beta_\omega \Big| \mathcal{M}_\omega, \sigma^2\right) \propto \int N\left(\beta_\omega \Big| 0, g\sigma^2\left(X'_\omega X_\omega\right)^{-1}\right)\pi\left(g\right)\mathrm{d}g. \qquad (7)$$

Both priors have desirable asymptotic properties and perform well in simulations (Liang et al. 2008).

How should one choose among the various prior options that are now available? As noted above, BIC tends to favor parsimonious models, whereas AIC tends to include more parameters (Kass and Raftery 1995; Kuha 2004). The hyper-g and Zellner-Siow priors will tend to fall somewhere in between. In practice, one's choice should depend on the goals of the research project, the nature of the data, and the type of model. However, the method we advocate—and which we use in our examples below—is to analyze data with respect to multiple priors to assess the sensitivity of one's results to prior choice.

---

[10]Although BIC and AIC are not proper Bayesian priors (Gelman and Rubin 1995), we will sometimes refer to them as "priors" for expositional clarity.

[11]All variables are assumed to be centered at zero in this notation.

[12]For instance, one can often choose a value for g that corresponds to the AIC and BIC approximations, although this value may not necessarily be known.

4.4 *Specifying Model Priors*

A related development in BMA methodology involves specifying more flexible priors over models. Per Clyde (2003) and Clyde and George (2004), we can think of placing a prior distribution on models $\mathcal{M}_1, \ldots, \mathcal{M}_k$ by treating the indicator variables ω as resulting from independent Bernoulli distributions.

$$\pi(\omega \mid \gamma) = \gamma^{p_\omega}(1-\gamma)^{p-p_\omega}. \tag{8}$$

This prior is fully specified by the selection of the hyperparameter $\gamma \in (0, 1)$, which can be thought of as the probability that each predictor variable is included in the model.

The vast majority of previous presentations have assumed a uniform distribution over models.[13] This assumption implies that $\gamma = .5$ and that the number of parameters is distributed binomial $(q, .5)$ over the $q = 2^p$ models, which means that the expected number of independent variables in a model is $p/2$ (Clyde 2003).

However, the assumption of a uniform distribution over models is not always appropriate (Erikson, Wright, and McIver 1997). The BAS package offers several options for specifying priors over models that reflect researchers' understanding of the data-generating process. First, analysts can select a value for γ that corresponds to their prior beliefs about the appropriate number of predictors in the model. Analysts with prior beliefs about the inclusion of specific variables can also represent γ as a vector $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p)$, where γ_i represents the prior probability that variable i should be included in the model. Finally, a third possible approach is to put a beta prior on the hyperparameter γ to reflect the range of complexity we expect in the posterior model space.

4.5 *Properly Handling Interaction Terms*

Finally, it is necessary to adjust BMA usage to account for the presence of interaction terms, which are frequently employed in social science data analysis. In his analysis, Bartels averages over models that vary in whether they include one or more interaction terms derived from variables of theoretical interest. However, the coefficient for a constitutive term of an interaction represents the marginal effect of that variable when the other constitutive term is equal to zero (Braumoeller 2004; Brambor, Clark, and Golder 2006). Combining coefficient estimates of constitutive terms with estimates of the same coefficients from models that omit the interaction creates an uninterpretable mixture of estimates of two different quantities.[14]

We recommend a different approach that is consistent with contemporary statistical practice. First, if an interaction term is one of the covariates under consideration, we should avoid averaging over models in which one or more of its constitutive terms are excluded (Braumoeller 2004; Brambor, Clark, and Golder 2006). To do otherwise assumes that the marginal effect of the excluded variable is zero. If this assumption is false, the interaction

---

[13]Bartels (1997) does so as well in his main analysis (669). (He also introduces "dummy-resistant" and "search-resistant" priors, but these have not come into wide usage and we therefore do not discuss them further.)

[14]For instance, the coefficients for state opinion and Democratic legislative strength in the Erikson, Wright, and McIver (1993) data that Bartels reanalyzes represent the marginal effect of those variables in "individualistic" states when interactions with state political culture indicators from Elazar (1972) are included (i.e., individualistic is the reference category and is therefore excluded). By contrast, when the interaction terms are omitted from the model, the coefficients for state opinion and Democratic legislative strength represent their unconditional marginal effects. A similar critique applies to Bartels's other example, which reanalyzes models of economic growth in Organisation for Economic Co-operation and Development countries by Lange and Garrett (1985) and Jackman (1987).

term will be incorrectly estimated. In addition, if an interaction term and its constitutive terms are quantities of theoretical interest (rather than control variables), it is desirable to average within the subset of models that include the constitutive terms and the interaction term. The resulting posterior distributions for the interaction and the constitutive terms will then have consistent conceptual definitions and can be interpreted properly.

Previously, it was impossible for the applied analyst to restrict the set of analyzed models in this way without writing new code. For instance, Erikson, Wright, and McIver (1997) express concern that BMA "does not seem adaptable to models containing mutually exclusive dummy variables or complicated interaction terms." To address this concern, we have modified the BMA and BAS packages to allow analysts to easily exclude theoretically inappropriate models from the averaging process. Using these software options, analysts can drop models that violate important theoretical or statistical assumptions. For instance, it is possible to drop all models that include an interaction term but omit one or more of its constitutive variables as described above.

5   Applying BMA: Three Illustrative Examples

In this section, we present three examples of how BMA can be applied in contemporary political science research using the methodological approach described above. Our first example examines the Adams, Bishin, and Dow (2004) study of voting in U.S. Senate elections, illustrating how BMA can be used to arbitrate between two possible measures of the same concept (voter utility from candidate positioning in one dimension). Second, we reanalyze the Canes-Wrone, Brady, and Cogan (2002) study of the effect of roll call extremity on incumbent support in the U.S. House of Representatives, which illustrates how BMA can be used to test the robustness of a single predictor against a wide array of alternative specifications including interactions. Our final example illustrates how BMA can help validate the robustness of one's statistical results in a vast model space using data from the analysis of Fearon and Laitin (2003b) of the onset of civil war.

5.1   *U.S. Senate Voting*

We begin with an example that demonstrates how the BMA approach can help arbitrate between competing predictors. Adams, Bishin, and Dow (henceforth ABD) use data from the 1988–1990–1992 Pooled Senate Election Study to "evaluate the discounting/directional hypothesis versus the alternative proximity hypothesis" (348). Using both an individual-level model of vote choice and an aggregate-level model of vote share, they "find a consistent role" for their directional variables, whereas results for their proximity variables are weaker and less consistent (368). We focus here only on their aggregate-level results (see Montgomery and Nyhan 2008 for a reanalysis of their individual-level results).

ABD follow the common approach of putting alternative measures into the same model and basing their inferences on the resulting coefficients—a practice that Achen (2005) refers to as a "pseudotheorem" of political science. Unfortunately, as Achen shows, this practice is likely to lead to incorrect inferences. A better approach is to use BMA, which allows us to test competing measures in a more coherent fashion.[15]

---

[15]Some studies have argued for combining directional and spatial approaches (e.g., Iversen 1994; Adams and Merrill 1999; Merrill and Grofman 1999). However, we interpret the ABD paper as an attempt to arbitrate between directional/discounting and proximity models.

ABD conduct an ordinary least squares analysis in which they predict the percentage of the two-party vote received by the Democrat in each election. They focus on two independent variables of interest, which they call Democratic directional advantage and Democratic proximity advantage, and estimate two types of models—one in which these variables are calculated using the average ideological placement of that candidate by all respondents in the relevant state and year (which we will refer to as "mean candidate placement") and one in which the variables are calculated using respondents' own placements of the two candidates (which we will refer to as "voter-specific placement").[16] These four different measures are then aggregated at the campaign level. In addition, ABD express some uncertainty about the correct set of control variables to include in the analysis, resulting in the reporting of two models for each variable of interest.

Columns 1–4 in Table 1 present our replication of ABD's table 2, which incorporates corrections of several errors in the published results (see the Appendix for a more extensive discussion of our replication). The corrected results, which serve as the basis for the BMA analysis below, show that the directional variable is consistently positive and statistically significant but that the proximity variable is consistently *negative* and significant.[17] This result contradicts spatial voting theory, which suggests that a party's ideological proximity to voters should be positively associated with its share of the vote. Moreover, the magnitude of the directional coefficients raises concerns about misspecification. For instance, the results in the first column of Table 1 indicate that a 1-unit increase in the Democratic directional advantage (a variable with a range that exceeds 4 in the data) results in a 11% increase in the Democratic share of the two-party vote.

As stated earlier, it is inappropriate to include two competing (and highly correlated) measures of a concept in the same model. Our BMA analysis therefore considers the entire model space implied by the five variables in the original model excluding those models containing *both* the directional and the proximity variables.[18] Because the dependent variable is continuous, we can use the BAS package. Columns 5–8 report our findings for the hyper-g prior ($a = 3$) with a uniform prior on the model space. Results were substantively identical under AIC, BIC, and Zellner-Siow, as illustrated by Fig. 2, which presents posterior plots for the variables of theoretical interest under all four priors.

We focus on the posterior probability of inclusion as the best metric for arbitrating between two possible measures of the same concept. In this case, our findings show considerably less support for ABD's conclusions than our replication of their original tables. The posterior probability of inclusion for the directional measures is consistently higher than the proximity variables. However, the Democratic directional advantage variable in the "mean candidate placement" model has a posterior probability of inclusion of only .30, which suggests that the variable is a relatively weak predictor of electoral outcomes.

---

[16]For the voter-specific evaluation, the proximity score is created by using the formula $[(x_R - x_i)^2 - (x_D - x_i)^2]$, where x_R and x_D are the respondent's placements of the Republican and Democratic candidates (respectively) on a 7-point Likert scale of ideology and x_i is the respondent's self-placement on that scale. For the voter-specific evaluation of the directional score, the relevant equation is $[(x_D - 4)(x_i - 4) - (x_D - 4)(x_i - 4)]$. The mean candidate variables are identical except that the average placements of each candidate from all respondents in that state and year are used for x_R and x_D.

[17]In the published version of the table, the proximity variable is insignificant and changes signs across specifications (see Appendix).

[18]In the text, ABD identify variables that were considered but not included in the final analysis. For expositional purposes, we do not consider them there. We demonstrate the utility of BMA for reanalyzing alternate control variables in our reanalysis of Fearon and Laitin (2003b) below.

Table 1 Ordinary least squares models of Democratic vote share in U.S. Senate campaigns 1988–92

| | ABD replication | | | | BMA reanalysis (hyper-g prior) | | | |
| | Reduced model | | Full model | | Mean candidate | | Voter-specific | |
| | Mean candidate | Voter-specific | Mean candidate | Voter-specific | Conditional mean (SD) | $p(\beta \neq 0 \mid Y)$ | Conditional mean (SD) | $p(\beta \neq 0 \mid Y)$ |
	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)				
Democratic proximity advantage	−3.137 (1.593)	−2.881 (1.270)	−3.053 (1.315)	−2.007 (1.056)	−0.222 (0.954)	.067	0.745 (0.677)	.033
Democratic directional advantage	11.146 (3.376)	5.729 (1.620)	7.953 (2.854)	4.177 (1.356)	3.576 (2.015)	.299	2.363 (0.840)	.787
Democratic incumbency advantage	6.376 (1.092)	6.609 (1.054)	1.060 (1.201)	1.139 (1.189)	1.607 (1.242)	.194	1.295 (1.237)	.159
Democratic quality advantage	5.972 (1.400)	5.035 (1.384)	3.117 (1.240)	2.378 (1.216)	2.964 (1.246)	.599	2.740 (1.223)	.541
Democratic spending advantage	—	—	0.270 (0.041)	0.265 (0.040)	0.3238 (0.041)	1.00	0.314 (0.037)	1.00
Democratic partisan advantage	—	—	0.055 (0.054)	0.060 (0.054)	0.0749 (0.057)	.201	0.0661 (0.055)	.181
Constant	54.759 (1.325)	52.786 (0.892)	53.309 (1.155)	52.028 (0.758)	51.381 (1.032)	1.00	51.556 (0.764)	1.00
N	95	95	94	94	94		94	
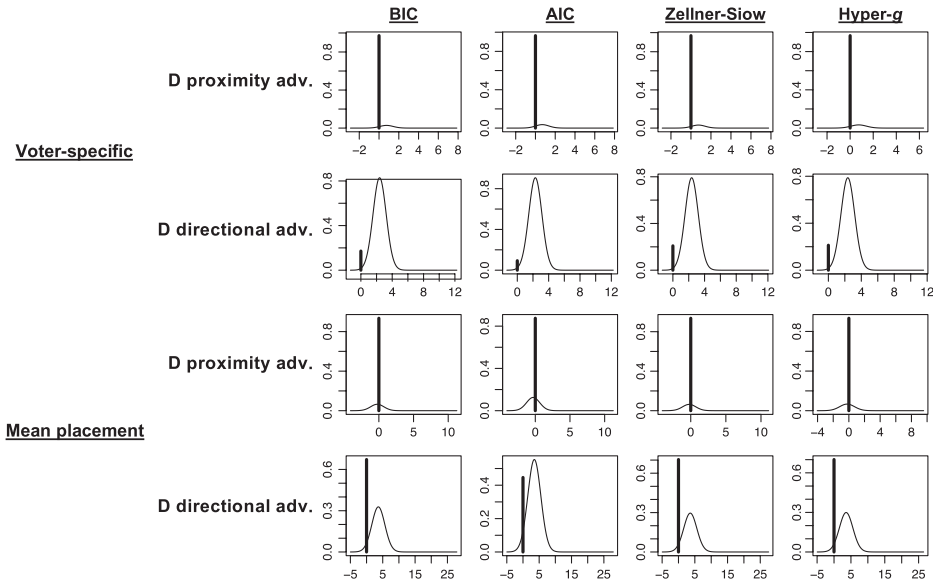
Fig. 2 Proximity/directional posteriors for Democratic vote share. These are the main posterior plots of interest under four different priors (additional posterior plots for each model available upon request). The vertical line located at 0 on the x axis represents the posterior probability of models that exclude the variable ($p(\beta = 0 \mid Y)$). The density represents the conditional posterior probability for the parameter in those models in which it is included ($p(\beta \mid \beta \neq 0, Y)$).

The results reported in columns 5–8 suggest that the negative coefficient on the proximity variables and the large coefficients associated with the directional measures were artifacts of including both measures in the same regressions. When we exclude models that include both variables and average across the remaining model space, we find that the proximity variable has a minimal rather than a negative coefficient. Second, the size of the coefficients for the directional variables is substantially reduced. These findings illustrate the inferential dangers of including competing measures of a single concept in a statistical model and demonstrate how BMA can help arbitrate between such measures in a systematic way.

5.2 *U.S. House Elections*

In their widely cited 2002 article, Canes-Wrone, Brady, and Cogan (henceforth CWBC) combine summary measures of roll call voting with electoral returns to show that legislative extremity reduces support for members of the U.S. House of Representatives in future elections. Their work demonstrates an important linkage between Congressional behavior and electoral outcomes. It also provides a classic example of a research design intended to demonstrate the robustness of the relationship between a single predictor and a dependent variable.

The focus of the CWBC analysis is their measure of roll call ideological extremity, which is based on ratings of House members provided by Americans for Democratic Action (ADA).[19] For expositional reasons, we focus here only on the full version of their

---

[19]It is calculated as the ADA score for Democratic members and 100 minus the ADA score for Republican members, so that higher values represent greater extremity by party (Canes-Wrone, Brady, and Cogan 2002, 131). The resulting score is then divided by 100.

Table 2 Ordinary least squares models of incumbent vote shares in U.S. House elections 1956–96

	CWBC replication	Unconditional models		Conditional models		All models	
	Coefficient (robust SE)	Conditional mean (SD)	$p(\beta \neq 0 \mid Y)$	Conditional mean (SD)	$p(\beta \neq 0 \mid Y)$	Conditional mean (SD)	$p(\beta \neq 0 \mid Y)$
Roll call ideological extremity (ADA)	−0.070 (0.005)	−0.0615 (0.0046)	.0000	−0.0550 (0.0048)	.0000	−0.0550 (0.0048)	.0000
Roll call ideological extremity (DW-NOM)	—	−0.2278 (0.0158)	1.0000	−0.2247 (0.0157)	1.0000	−0.2247 (0.0157)	1.0000
Presidential vote	0.454 (0.011)	0.3996 (0.0115)	1.0000	−0.0508 (0.0548)	1.0000	−0.0508 (0.0548)	1.0000
Presidential vote × extremity (ADA)	—	—		0.1804 (0.0354)	.0000	0.1804 (0.0354)	.0000
Presidential vote × extremity (DW-NOM)	—	—		0.8952 (0.1065)	1.0000	0.8952 (0.1065)	1.0000
Challenger quality	−0.053 (0.002)	−0.0495 (0.0023)	1.0000	−0.0492 (0.0023)	1.0000	−0.0492 (0.0023)	1.0000
Freshman	−0.020 (0.003)	−0.0203 (0.0028)	1.0000	−0.0198 (0.0027)	1.0000	−0.0198 (0.0027)	1.0000
Five or more terms	—	−0.0041 (0.0021)	.5723	−0.0033 (0.0020)	.3622	−0.0033 (0.0020)	.3622
In party	−0.040 (0.011)	−0.0594 (0.0131)	1.0000	−0.0593 (0.0131)	1.0000	−0.0593 (0.0131)	1.0000
Δ personal income (coded by in party)	0.042 (0.003)	0.0426 (0.0035)	1.0000	0.0426 (0.0035)	1.0000	0.0426 (0.0035)	1.0000
Presidential popularity (coded by in party)	−0.044 (0.010)	−0.0215 (0.0107)	.5687	−0.0218 (0.0107)	.5397	−0.0218 (0.0107)	.5397
Midterm loss (coded by in party)	−0.015 (0.003)	−0.0192 (0.0027)	1.0000	−0.0196 (0.0027)	1.0000	−0.0196 (0.0027)	1.0000
Years since 1956	—	0.0025 (0.0008)	1.0000	0.0022 (0.0008)	1.0000	0.0022 (0.0008)	1.0000
Years since 1956^2	—	0.000028 (0.000049)	1.0000	0.000047 (0.000045)	1.0000	0.000047 (0.000045)	1.0000
Years since 1956^3	—	−0.000002 (0.000001)	.9377	−0.000002 (0.000001)	.9737	−0.000002 (0.000001)	.9737
Proportion of African American (coded by party)	—	0.1320 (0.0077)	1.0000	0.1054 (0.0083)	1.0000	0.1054 (0.0083)	1.0000

Continued

258

Table 2 (*continued*)

	CWBC replication	Unconditional models		Conditional models		All models	
	Coefficient (robust SE)	*Conditional mean (SD)*	*p(β ≠ 0 | Y)*	*Conditional mean (SD)*	*p(β ≠ 0 | Y)*	*Conditional mean (SD)*	*p(β ≠ 0 | Y)*
Proportion of veterans (coded by party)	—	0.0048 (0.0006)	1.0000	0.0049 (0.0006)	1.0000	0.0049 (0.0006)	1.0000
Proportion of foreign born (coded by party)	—	0.1935 (0.0170)	1.0000	0.1915 (0.0168)	1.0000	0.1915 (0.0168)	1.0000
Proportion of employed in manufacturing (coded by party)		−0.1889 (0.0214)	1.0000	−0.1749 (0.0213)	1.0000	−0.1749 (0.0213)	1.0000
Proportion of rural (coded by party)	—	0.0151 (0.0115)	.3187	0.0154 (0.0114)	.2821	0.0154 (0.0114)	.2821
% Union workers in state (coded by party)	—	0.0487 (0.0086)	1.0000	0.0476 (0.0085)	1.0000	0.0476 (0.0085)	1.0000
Constant	0.739 (0.007)	0.7675 (0.0102)	1.0000	0.7651 (0.0101)	1.0000	0.7651 (0.0101)	1.0000
N	6521	6521		6521		6521	

259

pooled model (column 2 of table 2 in their article), which estimates the effect of extremity on a member's share of the total two-party vote for the 1956–96 period.

The pooled model that CWBC present, which is replicated in the first column of Table 2 below, includes a measure of the district presidential vote (which is intended to serve as a proxy for party strength in the district) and seven other control variables. These controls are presumably included to help ensure that any relationship they find is not spurious. However, we can use BMA to assess the robustness of their model across a wider range of plausible control variables. The literature on US elections suggests a number of other possible factors that might also be associated with electoral vote share. In this reanalysis, we consider variables measuring the demographic characteristics of the district (the proportion of district residents who live in rural settings, the proportion who work in the manufacturing sector, and the proportion who are African Americans, union members, foreign born, or veterans),[20] incumbency (an indicator for members who have served five or more terms), and a flexible function of years since 1956 (i.e., linear, squared, and cubed terms) to capture the changing magnitude of the incumbency advantage in this period.

We also use BMA to consider an alternate measure and a possible moderator. CWBC note (but do not show) that their results hold using an average of first- and second-dimension DW-NOMINATE scores (Poole and Rosenthal 1997, 2007) instead of ADA ratings. Since the average score across two dimensions is difficult to interpret, we instead transform DW-NOMINATE first-dimension scores by party (following the CWBC ADA measure) to assess how results compare between two possible measures of roll call extremity.[21] Finally, it is plausible that the electoral punishment for extremity may vary depending on the partisan composition of the district. As such, we separately interact both measures of roll call extremity with the CWBC measure of district presidential vote to assess whether the strength of the relationship is conditional on party strength in the district.

In each case, we also exclude all models that include competing measures of the same concept (i.e., those that include one or more ADA-based variables and one or more DW-NOMINATE–based variables), those that do not include a dummy variable for being in the incumbent president's party (it implicitly interacts with several other variables of interest), and all models that include the cubed or squared term for years since 1956 but exclude a lower order polynomial. Following our recommendations for analyzing interactions (described above), we analyze the unconditional effect of roll call extremity and the conditional effect in separate models before pooling terms to assess the posterior probability of inclusion for the interaction terms.

Table 2 provides model outputs from the BMA analysis using a Zellner-Siow prior on the coefficients and a Beta(3,2) prior on the model hyperparameter γ.[22] As noted above, the table contains three models. The first, which is reported in columns 2–3, considers the unconditional effect of extremity and therefore excludes all models with interaction terms. This analysis allows us to estimate the robustness of the CWBC finding across a large model space. The second model, which is reported in columns 4–5, examines the potential

---

[20]These variables are drawn from Adler (n.d.). In each case, the values are multiplied by −1 for Republicans to allow for differing effects by party.

[21]Specifically, we multiply Democrats' first dimension scores by −1 and then rescale the resulting variable to range from 0 to 1.

[22]The posterior probability plots for the main coefficients of interest are not shown for expositional reasons but are available upon request. In this case, they are regularly shaped and provide no additional information beyond the posterior summary statistics provided in Table 2.

moderating effects of district party strength and therefore excludes all models that do not include a properly specified interaction with both constitutive terms. In this case, we can interpret the posterior distributions of the extremity constitutive term and interaction as we would in a normal interaction model.[23] Finally, the third model, which is reported in columns 6–7, includes the interaction terms in the averaging process but does not *require* them to be included (though we again omit all models with an interaction that omit one or more constitutive terms). The resulting estimates for the constitutive terms are not necessarily interpretable, but this model allows us to use the posterior probability of inclusion to assess the importance of the interaction terms.

Comparing these results with those reported in the original study (column 1) leads us to three conclusions.[24] First, the unconditional effect of extremity on electoral support, as shown in columns 2–3, is robust to a large set of possible model configurations. The CWBC hypothesis is supported across a vast space of more than 98,000 models.

Second, we find that measures of roll call extremity constructed using DW-NOMINATE scores perform substantially better in all circumstances than those created using ADA scores. The DW-NOMINATE extremity variable dominates the posterior space in all the analyses with a posterior probability of inclusion approaching one. By contrast, the ADA extremity variable and its associated interaction term have extremely low posterior probabilities of inclusion. For instance, the posterior probability of inclusion for the ADA-based extremity measure in the unconditional model reported in columns 2–3 is 5.930×10^{-6}.

Third, the effect of roll call extremity on election results is moderated by party strength in the district (as measured by the CWBC presidential vote variable). The DW-NOMINATE interaction term is highly statistically significant in columns 4–5 ($p(\beta > 0 | \beta \neq 0, Y) > .999$) and its posterior probability of inclusion in the pooled model in columns 6–7 is approximately one. Substantively, these results indicate that members from very marginal districts suffer severe punishment for legislative extremity but the electoral cost of extremity declines rapidly as party strength in the district increases. In those districts in which the party is strongest, the marginal effect of roll call extremity is actually either negligible (i.e., the 95% confidence interval includes zero) or positive.[25] In other words, members are punished to the extent they are out of step *with their district*.[26]

5.3 *Civil War Onset*

In a groundbreaking study, Fearon and Laitin (2003b) seek to determine the most important predictors of civil war onset (a binary-dependent variable). Their reported logit models

---

[23]Note that we cannot interpret the constitutive term for district presidential vote as we would normally would (the marginal effect when the extremity variable equals zero) since it is interacted with two different measures of roll call extremity.

[24]The substantive inferences discussed below are consistent across multiple priors (results available upon request). The original CWBC analysis used robust standard errors, which are not available in BMA and thus not included in the analysis below.

[25]To fully understand this effect, it was necessary to estimate the marginal effect of extremity over the observed range of district presidential vote in a single model (Brambor, Clark, and Golder 2006). We selected the model containing the interaction and its constitutive terms with the highest posterior probability (.24). Since the sign and significance of the interaction and its constitutive terms were consistent with the conditional posterior distributions in the BMA analysis, the resulting marginal effect estimates should be representative of the set of models that include the interaction. All results of this analysis are available upon request.

[26]Griffin and Newman (2009) find a similar result using data from the 2000 and 2004 National Annenberg Election Study.

Table 3 Logit models of civil war onset 1945–99

	F&L M1 Coefficient (SE)	F&L M3 Coefficient (SE)	AIC		BIC	
			Conditional mean (SD)	$p(\beta \neq 0 \mid Y)$	Conditional mean (SD)	$p(\beta \neq 0 \mid Y)$
Constant	−6.731 (0.736)	−7.019 (0.751)	−7.007 (1.360)	1.00	−6.597 (0.717)	1.00
Prior war	−0.954 (0.314)	−0.916 (0.312)	−0.871 (0.315)	1.00	−0.724 (0.312)	1.00
Per-capita income	−0.344 (0.072)	−0.318 (0.071)	−0.304 (0.079)	1.00	−0.314 (0.068)	1.00
Log(population)	0.263 (0.073)	0.272 (0.074)	0.274 (0.080)	.998	0.323 (0.069)	.994
Log(% mountainous)	0.219 (0.085)	0.199 (0.085)	0.201 (0.088)	.880	0.201 (0.082)	.221
Noncontiguous state	0.443 (0.274)	0.426 (0.272)	0.479 (0.314)	.565	0.550 (0.285)	.073
Oil exporter	0.858 (0.279)	0.751 (0.278)	0.760 (0.298)	.904	0.820 (0.268)	.431
New state	1.709 (0.339)	1.658 (0.342)	1.697 (0.351)	1.00	1.770 (0.337)	.997
Instability	0.618 (0.235)	0.513 (0.242)	0.569 (0.250)	.835	0.690 (0.233)	.399
Democracy (Polity IV)	0.021 (0.017)		0.025 (0.017)	.200	0.028 (0.017)	.047
Ethnic fractionalization	0.166 (0.373)	0.164 (0.368)	0.159 (0.505)	.272	0.242 (0.372)	.016
Religious fractionalization	0.285 (0.509)	0.326 (0.506)	0.790 (1.564)	.300	0.076 (0.573)	.013
Anocracy		0.521 (0.237)	0.598 (0.245)	.618	0.748 (0.228)	.034
Democracy		0.127 (0.304)	0.166 (0.314)	.618	0.143 (0.306)	.034
Wars in neighboring countries			0.051 (0.093)	.290	0.068 (0.094)	.017
Proposition of largest ethnic group			−0.273 (0.587)	.285	−0.283 (0.430)	.016
Proposition of largest religious group			0.004 (0.018)	.283	−0.0002 (0.006)	.013
Percent Muslim			0.003 (0.003)	.358	0.004 (0.003)	.039
Log (number of languages)			−0.021 (0.156)	.263	0.009 (0.118)	.013
Former French colony			−0.202 (0.301)	.309	−0.284 (0.291)	.021

Continued

Table 3 (*continued*)

|	F&L M1 Coefficient (SE)	F&L M3 Coefficient (SE)	AIC		BIC	
			Conditional mean (SD)	*p(β ≠ 0 | Y)*	*Conditional mean (SD)*	*p(β ≠ 0 | Y)*
Eastern Europe			−0.269 (0.656)	.378	−0.515 (0.403)	.031
Asia			0.239 (0.685)	.303	0.037 (0.311)	.013
Sub-Saharan Africa			0.494 (0.661)	.381	−0.006 (0.268)	.013
Middle east and north Africa			0.419 (0.593)	.362	0.237 (0.273)	.019
Latin America			0.522 (0.663)	.392	0.447 (0.316)	.034
Second largest ethnic group			−0.386 (1.184)	.270	0.210 (1.006)	.013
N	6327	6327	6327		6327	

263
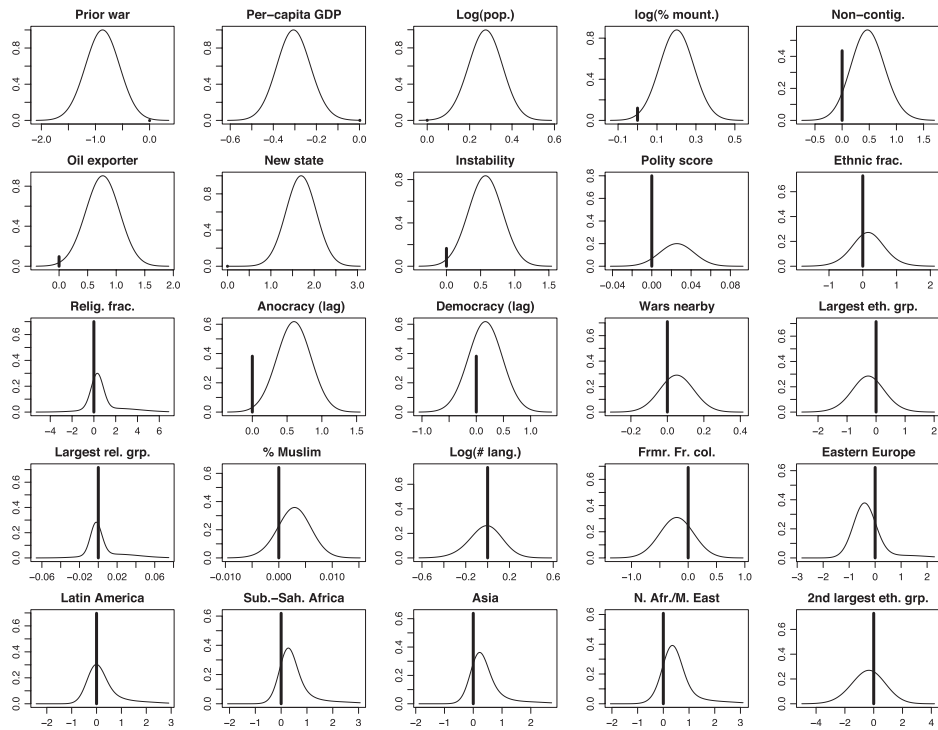
Fig. 3   Predictors of civil war onset 1945–99 (AIC). The vertical line located at 0 on the *x* axis represents the posterior probability of models that exclude the variable ($p(\beta = 0 \mid Y)$). The density represents the conditional posterior probability for the parameter in those models in which it is included ($p(\beta \mid \beta \neq 0, Y)$).

estimate the effects of 13 explanatory variables. However, throughout the text, footnotes, and additional results posted online (Fearon and Laitin 2003a), Fearon and Laitin (F&L) are unusually transparent in describing numerous other variables and interactions that were considered during the modeling process. In short, they acknowledge a great deal of uncertainty about the final model configuration that cannot be analyzed using traditional methods. Indeed, the length of their online supplement—which is 30 pages and contains 18 multicolumn tables—indicates the need for a more concise approach to specification uncertainty.

F&L's transparency allows us to identify a number of other variables that were considered to be plausible predictors of civil war onset. We estimate that F&L discuss approximately 74 possible independent variables (excluding various interpolation/missing data decisions), which implies a potential space of roughly 2×10^{22} potential models. As noted earlier, the traditional approach does not allow researchers to properly express uncertainty about their estimates when faced with such vast model spaces. For instance, consider the following quote (Fearon and Laitin 2003b, 84):

> When we add dummy variables for countries that have an ethnic or religious majority and a minority of at least 8% of the country's population, both are incorrectly signed and neither comes close to statistical significance. *This finding does not depend on which other variables are included in the model* (emphasis ours).

Obviously, F&L did not test these variables under all 20 sextillion possible specifications. One suspects that they tried adding relevant variables to their best models and found that they were insignificant (one such model is reported in table 3 of Fearon and Laitin 2003a).[27] BMA makes it possible to systematically justify such statements.

In this analysis, we chose a subset of F&L's variables to evaluate. One of the limits of BMA is that the model space $q = 2^p$ can quickly exceed the abilities of even the most advanced computers to fully explore the posterior model distribution. Clyde (2003) recommends that any models that use more than approximately 25 variables should be analyzed using stochastic sampling techniques rather than deterministic search algorithms. However, no publicly available BMA software performs stochastic sampling for GLM models (but see Pang and Gill 2009). As such, we chose 25 publicly available variables that had no missing data in the same universe of cases that F&L analyze, which allows us to explore the entire posterior distribution using the bic.glm function in the BMA package. To reduce the software limitations described in Clyde (1999), we effectively disable the model selection criterion, ensuring that the software returns the maximum number of relevant models, and create an option to use AIC instead of BIC (Akaike 1974).

Per our earlier discussion, we also place theoretically motivated limitations on the models we wish to explore. Specifically, we put a zero prior on all models that do not contain the key explanatory variable indicating the existence of a prior war. We also put a zero prior on models that contain both the Polity IV measure of democracy and dummy variables for democracy and anocracy derived from Polity IV or include only one of the anocracy and democracy dummy variables. In each case, we seek to adhere to standard procedures in the political science literature.

We replicate their primary models of civil war onset in columns 1 and 2 of Table 3.[28] Columns 3–6 of Table 3 provide conditional means, standard deviations, and posterior probabilities of inclusion under AIC and BIC.[29] Posterior plots under AIC are presented in Fig. 3.[30]

Although the table and figure contain a great deal of information, we highlight two key findings. First, conditional posterior distributions for the variables that F&L identify as statistically significant predictors of civil war onset—prior war, per-capita income, log(population), log(% mountainous), oil exporter, new state, instability, and anocracy—are consistent with their original results. BMA therefore provides a truly systematic demonstration of the robustness of F&L's results (and does not require 30 pages of tables to do so!). However, most of F&L's predictors (which are frequently measured imprecisely) have low posterior probabilities of inclusion under BIC (column 6). Besides the constant and the prior war variable (which we required to be included in each model), only per-capita income, logged population, and the indicator of a new state have a posterior probability of inclusion of more than .5—a result that underscores the need to examine the sensitivity of one's results to priors.

---

[27]This should not be interpreted as a criticism of F&L's important article. Many studies, including ones we have participated in, use this approach.

[28]These models correspond to models 1 and 3 in F&L. We do not address the three other models they report, which use different dependent variables.

[29]Before performing our analysis, we dropped a single observation with a miscoded value for the dependent variable from F&L's data. To assure that enough models were sampled, we set the leaps and bounds algorithm employed by bic.glm to return the 100,000 best models for each possible rank of X.

[30]It's worth noting that the BMA package assumes that the posterior distribution of each coefficient is normal, while BAS assumes they are distributed Student t. As a result, BMA plots tend to be more smooth than those generated by BAS.

6 Cautions and Conclusions

Political science researchers are often confronted with substantial uncertainty about the robustness of reported results. In many prominent literatures, researchers have proposed dozens (if not hundreds) of potential explanatory variables. Classical approaches to modeling techniques provide researchers with few tools for dealing with this uncertainty. As a result, readers are frequently concerned about alternative model configurations that were tried but not reported—and those that were never tried at all.

BMA offers researchers a comprehensive method for assessing model uncertainty that can easily be presented to readers. In this paper, we have reviewed recent developments in prior specifications and posterior computation techniques, presented a contemporary approach to the use of BMA, and applied this methodology to three prominent studies from the discipline. Our empirical analyses revealed substantive differences in the effects of the theoretical variables of interest from Adams, Bishin, and Dow (2004), demonstrated the conditional nature of the main effect reported in Canes-Wrone, Brady, and Cogan (2002), and gave a more rigorous foundation to the findings presented in Fearon and Laitin (2003b). In general, we strongly believe that BMA can strengthen the robustness of reported results in political science.

Despite the usefulness of the technique, we wish to conclude with words of caution about the appropriate use of BMA. First, we emphasize that it should not be used to conduct theory-free searches of the model space, particularly if such a step is not reported to readers. BMA also offers no solutions to the problems of endogeneity or causal inference. Statistical analysis should begin with the careful development of a model based on theory and previous research (Gelman and Rubin 1995). BMA is best used as a subsequent robustness check to show that our inferences are not overly sensitive to plausible variations in model specification.

On a related note, we also caution that BMA—like all statistical methods—cannot defeat unscrupulous researchers. Although it should be more difficult to manipulate BMA analysis than, say, a single-reported model specification, researchers *could* alter the set of variables that are averaged to try to support a desired finding. Similarly, one could use BMA to identify a model specification that maximizes fit to the data and then present that choice as the result of theory. As in all such cases, we must trust in the good intentions of the researcher and use theory to guide our judgments about the set of independent variables that should be considered.

With those caveats in mind, we hope that more analysts make use of BMA, which makes it possible to systematically test the robustness of our findings to a much wider array of model specifications than is otherwise possible.

Appendix: Adams, Bishin, and Dow (2004) Replication

Our reanalysis of Adams, Bishin, and Dow (2004) began with a replication of their published results. The authors generously provided us with code and data for their individual-level results, included an unpublished correction of their table 1. In collaboration with ABD, who provided us with further details on their data, we returned to the original 1988–1990–1992 Pooled Senate Election Study data and located several potential problems with their reported (and corrected) results.

Table A1 Aggregate Democratic vote share in Senate campaigns 1988–92

	ABD (2004) original				ABD (2004) replication			
	MC	VS	MC	VS	MC	VS	MC	VS
Democratic proximity advantage	−2.88 (1.78)	0.91 (1.80)	−2.67 (1.55)	0.78 (1.78)	−3.137 (1.593)	−2.881 (1.270)	−3.053 (1.315)	−2.007 (1.056)
Democratic directional advantage	9.45 (3.31)	8.68 (3.32)	7.08 (2.67)	7.29 (2.56)	11.146 (3.376)	5.729 (1.620)	7.953 (2.854)	4.177 (1.356)
Democratic incumbency advantage	6.83 (1.14)	5.96 (4.02)	7.50 (6.23)	7.41 (5.99)	6.376 (1.092)	6.609 (1.054)	1.060 (1.201)	1.139 (1.189)
Democratic quality advantage	5.70 (1.48)	4.38 (2.34)	1.52 (1.22)	1.23 (1.09)	5.972 (1.400)	5.035 (1.384)	3.117 (1.240)	2.378 (1.216)
Democratic spending advantage			3.16 (1.27)	3.00 (1.32)			0.270 (0.041)	0.265 (0.040)
Democratic partisan advantage			0.27 (0.04)	0.22 (0.04)			0.055 (0.054)	0.060 (0.054)
Constant	53.92 (1.36)	52.06 (1.44)	52.46 (1.14)	51.34 (1.37)	54.759 (1.325)	52.786 (0.892)	53.309 (1.155)	52.028 (0.758)
N	95	95	95	95	95	95	94	94

Note. MC, mean candidate placement; VS, voter-specific placement.

ABD's correction revised the coding of the dependent variable in the individual-level analysis from their published paper. Numerous nonvoting respondents were inadvertently coded as having voted for the Republican Senate candidate. In collaboration with ABD, we uncovered a few other discrepancies. Several of these appeared to be coding errors in the statistical analysis and in the data itself.[31] An additional concern is that the coding of the proximity advantage variable appears to differ from the one presented in the article (we used the coding $[(x_R - x_i)^2 - (x_D - x_i)^2]$, which conforms to equations (3) and (4) and footnote 17).

The model results reported in Table A1 of this paper incorporate all relevant corrections for the aggregate-level data. The table below reports the original published aggregate-level results and our replication of those results. Our replication of their individual-level results is reported in Montgomery and Nyhan (2008). Details and code are available upon request.

References

Achen, Christopher H. 2005. Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science* 22:327–39.

Adams, James, Benjamin G. Bishin, and Jay K. Dow. 2004. Representation in congressional campaigns: Evidence for discounting/directional voting in U.S. Senate elections. *Journal of Politics* 66:348–73.

Adams, James, and Samuel Merrill. 1999. Modeling party strategies and policy representation in multiparty elections: Why are strategies so extreme? *American Journal of Political Science* 43:765–91.

Adler, E.. Scott. n.d. *Congressional district data file*. Unpublished data-set, Boulder, CO: University of Colorado.

Akaike, Hirotugu. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19:716–23.

Bartels, Larry M. 1997. Specification uncertainty and model averaging. *American Journal of Political Science* 41:641–74.

———. 1998. Posterior distributions from model averaging: A clarification. *Political Methodologist* 8:17–9.

Bartels, Larry M., and John Zaller. 2001. Presidential vote models: A recount. *PS: Political Science and Politics* 34:8–20.

Brambor, Thomas, William Roberts Clark, and Matthew Golder. 2006. Understanding interaction models: Improving empirical analyses. *Political Analysis* 14:63–82.

Braumoeller, Bear. 2004. Hypothesis testing and multiplicative interaction terms. *International Organization* 58:807–20.

Canes-Wrone, Brandice, David W. Brady, and John F. Cogan. 2002. Out of step, out of office: Electoral accountability and House members' voting. *American Political Science Review* 96:127–40.

Clarke, Kevin A. 2001. Testing nonnested models of international relations: Reevaluating realism. *American Journal of Political Science* 45:724–44.

———. 2005. The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science* 22:341–52.

Clyde, Merlise. 1999. 'Bayesian model averaging: A tutorial': Comment. *Statistical Science* 14:401–4.

———. 2003. Model averaging. In *Subjective and objective Bayesian statistics*. 2nd ed., Chap. 13, ed. S. James Press, 320–35. Hoboken, NJ: Wiley-Interscience.

———. (with contributions from Michael Littman). 2009. *BAS: Bayesian model averaging using Bayesian adaptive sampling*. R Package Version 0.45. http://cran.r-project.org/package=BAS.

Clyde, Merlise, and Edward I. George. 2004. Model uncertainty. *Statistical Science* 19:81–94.

Clyde, Merlise, Joyee Ghosh, and Michael Littman. 2009. *Bayesian adaptive sampling for variable selection*. Unpublished manuscript, Department of Statistical Science Discussion Paper, Duke University.

Draper, David. 1995. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B (Methodological)* 57:45–97.

---

[31]These errors included some miscoded California Senate data, what we suspect was the inclusion of the wrong variable in one of the regressions in the corrected tables, and apparent data entry errors in the aggregate-level variables that are included in both the individual-level and the aggregate regressions.

Elazar, Daniel J. 1972. *American federalism: A view from the states*. 2nd ed. New York: Crowell.

Erikson, Robert S., Joseph Bafumi, and Bret Wilson. 2001. Was the 2000 presidential election predictable? *PS: Political Science and Politics* 34:815–19.

Erikson, Robert S., Gerald C. Wright, and John P. McIver. 1993. *Statehouse democracy: Public opinion and policy in the American states*. New York: Cambridge University Press.

———. 1997. *Too many variables? A comment on Bartels' model-averaging proposal.* Paper Presented at the 1997 Political Methodology Conference, Columbus, Ohio.

Fearon, James D., and David D. Laitin. 2003a. *Additional tables for 'ethnicity, insurgency, and civil war'*. Unpublished manuscript, Stanford University.

———. 2003b. Ethnicity, insurgency, and civil war. *American Political Science Review* 97:75–90.

Fernandez, Carmen, Eduardo Ley, and Mark F. J. Steel. 2001. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* 16:563–76.

Geer, John, and Richard R. Lau. 2006. Filling in the blanks: A new method for estimating campaign effects. *British Journal of Political Science* 36:269–90.

Gelman, Andrew, and Donald B. Rubin. 1995. Avoiding model selection in Bayesian social research. *Sociological Methodology* 25:165–73.

Gerber, Alan, and Neil Malhotra. 2008. Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Quarterly Journal of Political Science* 3:313–26.

Gill, Jeff. 1999. The insignificance of null hypothesis significance testing. *Political Research Quarterly* 52:647–74.

———. 2004. Introduction to the special issue. *Political Analysis* 12:323–37.

Griffin, John, and Brian Newman. 2009. *Assessing accountability*. Paper presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.

Harrell, Frank E. 2001. *Regression modeling strategies*. New York: Springer.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15:199–236.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Christopher T. Volinsky. 1999. Bayesian model averaging: A tutorial. *Statistical Science* 14:382–401.

Imai, Kosuke, and Gary King. 2004. Did illegal overseas absentee ballots decide the 2000 U.S. presidential election? *Perspectives on Politics* 2:537–49.

Iversen, Torben. 1994. Political leadership and representation in west European democracies: A test of three models of voting. *American Journal of Political Science* 38:45–74.

Jackman, Robert W. 1987. The politics of economic growth in the industrial democracies, 1974–80: Leftist strength or North Sea oil? *Journal of Politics* 49:242–56.

Jeffreys, Harold. 1935. Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophical Society* 31:203–22.

———. 1961. *Theory of probability*. 3rd ed. Oxford: Oxford University Press.

Kass, Robert E., and Adrian E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90:773–95.

King, Gary, and Langche Zeng. 2006. The dangers of extreme counterfactuals. *Political Analysis* 14:131–59.

Kuha, Jouni. 2004. AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods Research* 33:188–229.

Lange, Peter, and Geoffrey Garrett. 1985. The politics of growth: Strategic interaction and economic performance in the advanced industrial democracies, 1974–1980. *Journal of Politics* 47:792–827.

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* 103:410–23.

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* 89:1535–46.

Merrill, Samuel, and Bernard Grofman. 1999. *A unified theory of voting: Directional and proximity spatial models*. New York: Cambridge University Press.

Montgomery, Jacob, and Brendan Nyhan. 2008. *Bayesian model averaging: Theoretical developments and practical applications*. Working paper, Society for Political Methodology.

Morales, Knashawn H., Joseph G. Ibrahim, Chien-Jen Chen, and Louise M. Ryan. 2006. Bayesian model averaging with applications to benchmark dose estimation for arsenic in drinking water. *Journal of the American Statistical Association* 101:9–17.

Pang, Xun, and Jeff Gill. 2009. *Spike and slab prior distributions for simultaneous Bayesian hypothesis testing, model selection, and prediction, of nonlinear outcomes*. Unpublished manuscript, Washington University in St. Louis.

Poole, Keith T., and Howard Rosenthal. 1997. *Congress: A political-economic history of roll call voting*. New York: Oxford University Press.

————. 2007. *Ideology and congress*. New Brunswick, NJ: Transaction Publishers.

Raftery, Adrian E. 1995. Bayesian model selection in social research. *Sociological Methodology* 25:111–63.

Raftery, Adrian E., Jennifer Hoeting, Chris T. Volinsky, Ian S. Painter, and Ka Yee Yeung. 2009. *BMA: Bayesian model averaging*. R package version 3.12. http://cran.r-project.org/package=BMA.

Raftery, Adrian E., Ian S. Painter, and Christopher T. Volinsky. 2005. BMA: An R package for Bayesian model averaging. *R News* 5:2–8.

Weakliem, David L. 1999. A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research* 27:359–97.

Wintle, B. A., M. A. McCarthy, C. T. Volinsky, and R. P. Kavanagh. 2003. The use of Bayesian model averaging to better represent uncertainty in ecological models. *Conservation Biology* 17:1579–90.

Yeung, Ka Yee, Roger E. Bumgarner, and Adrian E. Raftery. 2005. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics* 21:2394–402.

Zaller, John R. 2004. Floating voters in U.S. presidential elections, 1948–2000. In *Studies in public opinion: Attitudes, nonattitudes, measurement error, and change*, eds. Willem Saris and Paul M. Sniderman, 166–214. Princeton, NJ: Princeton University Press.

Zellner, Arnold. 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In In *Bayesian inference and decision techniques: Essays in honor of Burno de Finetti*, eds. Prem K. Goel and Arnold Zellner, 233–43. North-Holland, The Netherlands: Elsevier.

Zellner, Arnold, and Aloysius Siow. 1980. Posterior odds ratios for selected hypotheses. In *Bayesian statistics*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith. Valencia, Spain: University Press.