A Statistical Method for Empirical Testing of Competing Theories

Kosuke Imai Princeton University
Dustin Tingley Harvard University

Empirical testing of competing theories lies at the heart of social science research. We demonstrate that a well-known class of statistical models, called finite mixture models, provides an effective way of rival theory testing. In the proposed framework, each observation is assumed to be generated either from a statistical model implied by one of the competing theories or more generally from a weighted combination of multiple statistical models under consideration. Researchers can then estimate the probability that a specific observation is consistent with each rival theory. By modeling this probability with covariates, one can also explore the conditions under which a particular theory applies.We discuss a principled way to identify a list of observations that are statistically significantly consistent with each theory and propose measures of the overall performance of each competing theory. We illustrate the relative advantages of our method over existing methods through empirical and simulation studies.

Empirical testing of competing theories lies at the heart of social science research. Since there typically exist alternative theories explaining the same phenomena, researchers can often increase the plausibility of their theory by empirically demonstrating its superior explanatory power over rival theories. In political science, Clarke set forth this argument most forcefully by claiming that "theory confirmation is not possible when a theory is tested in isolation, regardless of the statistical approach" (2007b, 886). In order to quantitatively test competing theories, however, most political scientists fit a regression model with many explanatory variables that are derived from multiple theories. Achen (2005) strongly condemns such practice as atheoretical and calls it a "garbage-can regression." To address this critique, some researchers have relied upon various model comparison procedures to assess the relative performance of statistical models implied by different theories under investigation.[1]

In this article, we demonstrate that a general and well-known class of statistical models, called *finite mixture models*, provides a more effective method for comparative theory testing. The basic idea of mixture modeling is intuitive. Each observation is assumed to be generated either from a statistical model implied by one of the rival theories or more generally from a weighted combination of multiple statistical models under consideration. In addition to the parameters of each model, researchers can estimate the probability that a specific observation is consistent with either of the competing theories. These observation-specific probabilities can be averaged to serve as an overall performance measure for each model, thereby also achieving most of what standard model selection methods are designed to do. Finite mixture models are typically used to make a parametric model flexible by allowing model parameters to vary across groups of observations. However, we show that a mixture of non-nested statistical models can be used

Kosuke Imai is Assistant Professor, Department of Politics, Princeton University, Corwin Hall 036, Princeton, NJ 08544 (kimai@princeton.edu, http://imai.princeton.edu). Dustin Tingley is Assistant Professor, Department of Government, Harvard University, 1737 Cambridge St., Cambridge, MA 02138 (dtingley@gov.harvard.edu, http://scholar.harvard.edu/dtingley). The replication code and data archive for this article are available at http://hdl.handle.net/1902.1/16378. We thank Mike Hiscox and Todd Allee for kindly sharing their data. Thanks to Will Bullock, Christina Davis, Marty Gilens, Michael Hiscox, Simon Jackman, Evan Lieberman, Helen Milner, Grigo Pop-Eleches, Brandon Stewart, Teppei Yamamoto, Carlos Velasco Rivera, Jaquilyn Waddell Boie, Robert Walker, and seminar participants at Harvard University, Princeton University, the University of California, Berkeley, and the University of Chicago, Harris School for helpful suggestions. We also thank the editor and the four anonymous reviewers for extensive comments that have significantly improved this article. Imai acknowledges the financial support from the National Science Foundation (SES–0918968).

[1]Popular methods include Bayesian information criteria, the Vuong test, the J test, and the Clarke test (see Clarke 2000).

American Journal of Political Science, Vol. 56, No. 1, January 2012, Pp. 218–236

 DOI: 10.1111/j.1540-5907.2011.00555.x

to empirically test competing theories in social science research.

The fundamental difference between the mixture modeling approach we advocate and standard model selection procedures such as the ones mentioned in footnote 1 is that the former allows for the possibility that multiple theories can coexist; some observations are better explained by one theory and others are more consistent with another theory.[2] In contrast, standard model selection procedures are based on the hypothesis that one theory explains all observations. Indeed, finite mixture models make it possible to determine the conditions under which each of the competing theories applies. This is an important advantage given that theoretical refinement is often achieved by considering the conditions under which existing theories apply. Moreover, in some cases, these conditions can be directly derived from the differences in the underlying assumptions of competing theories. Thus, the mixture modeling approach enables one to test a set of competing theories as a whole including their assumptions.[3]

The mixture modeling approach also provides a significantly more flexible framework for theory testing than standard model selection methods. For example, it can handle both nested and non-nested models in the same framework. The components of mixture models can be either frequentist or Bayesian statistical models. Moreover, mixture modeling can simultaneously test more than two theories. This contrasts with some of the popular non-nested model comparison procedures such as the Vuong, *J,* and Clarke tests, which can test only two theories at a time (see, e.g., Clarke 2008, for an exception). By testing all theories at once, finite mixture models can overcome the potential indeterminacy problem of these non-nested tests (i.e., paper beats rock, rock beats scissors, and scissors beat paper).

Another advantage of the proposed mixture modeling approach is its ability to account for uncertainty about which theory explains each observation, thereby avoiding the potential multiple testing problem resulting from preliminary testing. In particular, if one estimates the model chosen by a model selection procedure and computes standard errors for estimated model parameters, these standard errors will not incorporate the uncertainty about model selection and hence will be underestimated. In contrast, mixture modeling avoids the problem by fitting all competing models in one step while accounting for estimation uncertainty about the explanatory power of alternative theories.

We emphasize that finite mixture models are a well-known class of statistical models developed in a large body of statistical literature (see, e.g., Frühwirth-Schnatter 2007).[4] Although these models have begun to attract the attention of other social scientists (e.g., Harrison and Rutström 2009), relatively few political scientists have used them.[5] Using examples from international relations, we show that the same modeling technology can be used to test competing (and possibly non-nested) statistical models implied by alternative theories.

Furthermore, we propose a principled way to identify a list of observations that are *statistically significantly consistent* (as formally defined below) with each rival theory. The identification of such observations opens up the possibility that researchers directly connect the results of quantitative analysis with their qualitative knowledge (Lieberman, 2005). To construct such a list, our strategy is to conduct statistical hypothesis tests on whether a particular observation is consistent with each of the competing theories. We do so by formally addressing the well-known problem of false positives associated with multiple testing. Thus, researchers can control the expected number of falsely classified observations on the resulting list such that it does not exceed a predetermined threshold. Finally, the resulting lists provide alternative measures of the overall performance of each theory. For example, the proportion of observations that are statistically significantly consistent with one theory can serve as such a measure.

While our proposed method is widely applicable, it is motivated by an influential study from international relations, concerning alternative accounts of trade policy preferences. In what follows, we first briefly describe this motivating empirical example. We then demonstrate how to use finite mixture models for empirical testing of competing theories and develop methods to classify observations for each theory. Simulation studies are then conducted, and the proposed method is applied to the trade policy preference example as well as another example from international relations with three competing theories. We discuss the pitfalls of the mixture modeling

[2] See Gordon and Smith (2004), who share our motivation.

[3] We emphasize that the proposed approach does not require researchers to specify such conditions. In fact, researchers may wish to use the proposed mixture modeling approach in order to explore what factors determine the relative applicability of each rival theory. Of course, the findings obtained from such exploratory analyses must be interpreted with caution, and deductive and systematic theory testing is required to draw more definitive conclusions.

[4] In this article, we consider a mixture of regressions, which is also known as switching regressions (e.g., Quandt 1972). In political science, Brandt and Freeman (2006) use Markov switching model (particular mixture models) for time-series data.

[5] Notable exceptions include the work by Hill and Kriesi (2001), Kedar (2005), and Iaryczower and Shum (2009).

approach and ways to avoid them before giving concluding remarks.

A Motivating Empirical Example

In this section, we briefly describe the background of the motivating empirical example regarding the competing theories of trade policy preferences. An enduring theme in the international political economy literature is the explanation of preferences for free trade. In a seminal contribution, Hiscox (2002) analyzes legislative voting on trade bills in the United States by drawing on political economy interpretations of two canonical theories from the trade literature: the Stolper-Samuelson (SS) and Ricardo-Viner (RV) models of international trade. The two competing theories differ critically in the extent to which they emphasize *factoral* versus *sectoral* cleavages. The SS model suggests that cleavages on trade policy will be along factoral lines and predicts that the owners of factors which the United States is relatively abundant in (compared to the rest of the world) will favor trade liberalization.[6] In contrast, the RV model suggests an alternative cleavage between supporters and opponents of free trade that runs along sectoral lines.[7] These two models of support for trade policy figure centrally in this long tradition of international political economy research (e.g., Ladewig 2006; Rogowski 1989; Scheve and Slaughter 2001).

A key observation made by Hiscox (2002) is that the applicability of these competing models depends on how *specific* factors of production are to particular industries. If capital is highly mobile in the national economy, meaning it can easily move across industries, then the SS model is likely to be supported because the winners and losers of trade will be found among owners of abundant and scarce forms of factors, respectively. On the other hand, if capital is more specific (i.e., less mobile), then cleavages should fall along sectoral lines since capital is unable to easily adjust across industries. Hence, Hiscox hypothesized that whether congressional voting on trade bills can be explained by the SS or RV model will depend on the degree of factor specificity in the U.S. economy.

To empirically test this hypothesis, Hiscox collected the data on factor specificity in the U.S. economy over nearly two centuries. His measures varied considerably over time, suggesting that during some eras voting should be along factor lines (capital/land versus labor) and in other eras along sectoral lines (exporters versus importers). To leverage these changes over time, Hiscox estimated separate regressions for different eras in time. Using a conventional model selection procedure called the J test, Hiscox provides evidence that support for liberalization is best accounted for by the SS model during eras where specificity was low. In contrast, he finds that in periods where specificity was high, the RV is the preferred model.

Although breaking up the votes into different eras constitutes one informal way to test the factor specificity argument, the continuous measure of the factor specificity variable created by Hiscox does not provide natural breakpoints which can then be used to group votes. Thus, any grouping might be criticized as arbitrary. As we demonstrate, finite mixture models offer a relatively straightforward and yet formal way to directly incorporate the factor specificity measure. In particular, mixture models use the level of factor specificity to predict whether the SS or RV model is appropriate for each trade bill or even each vote. Thus, in addition to the overall assessment of the two models, we are also able to identify the list of trade bills in which the voting pattern is consistent with each theory.

The Proposed Methodology

In this section, we first briefly review the specification, estimation, and inference for finite mixture models in the context of empirical testing of competing theories. We then discuss a method to identify the observations that are statistically significantly consistent with each theory. We also propose several ways to measure the overall performance of each competing theory. Finally, we compare the proposed approach with the standard model selection procedures.

Before describing the proposed methodology, we emphasize an important distinction between *causal* and *predictive* inferences. For causal inferences, ignoring relevant confounders may result in omitted variable bias.[8] In contrast, the existence of omitted variables alone does not invalidate predictive inferences.[9] Indeed, it is well known

[6] For the United States, this means that capital and land owners will support free trade, whereas those specializing in labor should oppose liberalization.

[7] Those in export industries should favor liberalization, whereas those in import competing industries should oppose it.

[8] Although controlling for irrelevant variables can sometimes result in bias, in typical observational studies analyzed by social scientists it is difficult to argue that researchers should not adjust for the observed differences between the treatment and control groups.

[9] For example, suppose that we have a simple random sample from a population and find that in this sample, the turnout of overweight voters is significantly lower than that of other voters. Clearly, even

that for the purpose of *predictive* inferences, parsimonious models tend to outperform unnecessarily large models (see, e.g., Hastie, Tibshirani, and Friedman 2001). Thus, if the goal of researchers is to construct a theory with strong predictive power (as opposed to testing causal mechanisms; see Imai, Tingley, and Yamamoto 2011, for relevant methodological issues), parsimonious models that can capture systematic patterns in the data are preferred. While our method can be used for both purposes, the causal inference approach would require strong research designs that enable the identification of causal effects. Our empirical examples should be thought of as the instances of predictive inference. Nevertheless, whenever using mixture models, well-specified theories play an essential role in model specification.

Finite Mixture Models: A Review

Model Specification. Consider a finite number of M different statistical models, each of which is implied by one of the competing theories explaining the same phenomena. Beyond the fact that it can handle more than two theories at the same time, the proposed method is applicable without modification regardless of whether these statistical models are nested or not.

Finite mixture models are based on the assumption that each observation is generated either from one of the M statistical models or more generally from a weighted combination of multiple statistical models. This does not necessarily imply that researchers must identify all relevant theories. It is also possible that any observation, which is consistent with one of M theories under consideration, is also consistent with other theories that may or may not be included in the analysis. Rather, the goal of finite mixture models is to measure the *relative* explanatory power of the competing theories under consideration by examining how well a statistical model implied by any of the rival theories predicts each observation in the sample. For example, it is perhaps the case that the Stolper-Samuelson and Ricardo-Viner theories do not exhaust all possible theories for trade policies. And yet, it is of interest to investigate the relative performance of each theory explaining the variation in the voting behavior of legislators.

though body weight is not randomly assigned, the turnout difference between these two subsamples of voters is an unbiased and hence valid estimate for the difference in turnout between their corresponding subpopulations. However, this does not necessarily imply a causal relationship; making people diet may not increase turnout.

Formally, let $f_m(y \mid x, \theta_m)$ denote a statistical model implied by theory m where y is the value of the outcome variable Y, x is the value the vector of covariates X takes, and θ_m is the vector of model parameters. In statistics, typical applications of finite mixture models involve the same distributional and functional-form assumptions with identical covariates. Similar to random coefficient models, such an approach makes parametric models flexible by allowing different groups of observations to have different parameter values. However, these alternative statistical methods can neither provide a measure of overall support for each theory nor classify each observation to one of the competing theories. Furthermore, different theories usually require different sets of predictors. In fact, Hiscox (2002) employed logistic regression models with different sets of covariates for the Stolper-Samuelson and Ricardo-Viner theories. One may also wish to specify different statistical models for rival theories. For example, when analyzing the duration of cabinet dissolution, the underlying risk of cabinet dissolution may be constant (as in the exponential model) or increases over time (as in the Weibull model) (see King et al. 1990; Warwick and Easton 1992). Unlike random coefficient models and regressions with interaction terms, mixture models can handle these situations.

Given this setup, we formalize the idea that each observation is generated from one of M statistical models, but we do not know a priori which model generates a specific observation. Specifically, we use the latent (unobserved) variable Z_i to represent the theory with which observation i is consistent. Thus, Z_i can take one of M values, i.e., $Z_i \in \{1, 2, \ldots, M\}$, depending on which statistical model generates the ith observation. The data-generating process is given by,

$$Y_i \mid X_i, Z_i \sim f_{Z_i}(Y_i \mid X_i, \theta_{Z_i}) \tag{1}$$

for each $i = 1, \ldots, N$.

Next, assuming the conditional independence across observations given the covariates and the latent variable, the model specified in equation (1) yields the following observed-data likelihood function where the latent variable Z_i has been integrated out,

$$L_{obs}\left(\Theta, \Pi \mid \{X_i, Y_i\}_{i=1}^N\right) = \prod_{i=1}^{N} \left\{ \sum_{m=1}^{M} \pi_m f_m(Y_i \mid X_i, \theta_m) \right\}. \tag{2}$$

In this equation, $\pi_m = \Pr(Z_i = m)$ represents the population proportion of observations generated by theory m with $\sum_{m=1}^{M} \pi_m = 1$ and $\pi_m > 0$ for each m, $\Theta = \{\theta_m\}_{m=1}^{M}$ is the set of all model parameters, and $\Pi = \{\pi_m\}_{m=1}^{M}$ is the set of all model probabilities. The

parameter π_m can be interpreted as one measure of the overall performance of theory m.

The mixture model does not necessarily "assume" each observation is implied by one and only one of the rival theories under consideration. An alternative and more general interpretation of the above mixture model, as seen clearly from equation (2), is that each observation is implied by a weighted combination of the rival theories where the relative weights are represented by π_m. The distinction between these two interpretations is not important when fitting the mixture model because the data cannot distinguish them, but as discussed later, it becomes crucial when using the statistical test we propose.

As mentioned earlier, finite mixture models can be extended to determine the conditions under which a particular theory applies. In the example given earlier, the level of factor specificity in the national economy determined the relative applicability of the Stolper-Samuelson and Ricardo-Viner models. Such variables can be easily incorporated in the finite mixture modeling framework. This is done by directly modeling the probability that an observation is consistent with theory m, i.e., π_m, as a function of the observed *theory-predicting variables*, W_i (note that W_i may overlap with X_i), in the following manner,

$$\Pr(Z_i = m \mid W_i) = \pi_m(W_i, \psi_m), \tag{3}$$

where ψ_m is a vector of unknown model parameters. If W_i turns out to be a powerful predictor, then we may conclude that the applicability of rival theories depends on this variable. In practice, the multinomial logistic regression is often used for modeling this probability. However, more flexible models such as the multinomial probit model (Imai and van Dyk 2005) and the semiparametric multinomial logit model can also be used to model the relationship between Z_i and W_i (see the first section of the supporting materials).

Estimation and Inference. Estimation and inference can proceed by either a frequentist approach of maximizing the observed-data likelihood function or a Bayesian approach of sampling from the posterior distribution after specifying prior distributions. To obtain the maximum likelihood estimates, the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1997), an iterative numerical optimization algorithm consisting of the expectation (or E) step and the maximization (or M) step, can be applied to the following complete-data log-likelihood function, which is derived by assuming that Z_i is observed,

$$\begin{aligned} &l_{com}\big(\Theta, \Pi \mid \{X_i, Y_i, Z_i\}_{i=1}^N\big) \\ &\quad = \sum_{i=1}^{N} \sum_{m=1}^{M} \mathbf{1}\{Z_i = m\}\{\log \pi_m + \log f_m(Y_i \mid X_i, \theta_m)\}. \end{aligned} \tag{4}$$

Then, the E-step will compute the conditional expectation of the latent variable Z_i given the observed data and the values of parameters at the previous iteration. This is given by,

$$\begin{aligned} &Q\big(\Theta, \Pi \mid \Theta^{(t-1)}, \Pi^{(t-1)}, \{X_i, Y_i, Z_i\}_{i=1}^N\big) \\ &\quad = \sum_{i=1}^{N} \sum_{m=1}^{M} \zeta_{i,m}^{(t-1)}\{\log \pi_m + \log f_m(Y_i \mid X_i, \theta_m)\}, \end{aligned} \tag{5}$$

where the conditional expectation has the following expression,

$$\begin{aligned} \zeta_{i,m}^{(t-1)} &= \Pr\big(Z_i = m \mid \Theta^{(t-1)}, \Pi^{(t-1)}, \{X_i, Y_i\}_{i=1}^N\big) \\ &= \frac{\pi_m^{(t-1)} f_m(Y_i \mid X_i, \theta_m)}{\sum_{m'=1}^{M} \pi_{m'}^{(t-1)} f_{m'}(Y_i \mid X_i, \theta_{m'})}. \end{aligned} \tag{6}$$

This parameter has an intuitive interpretation because it represents the posterior probability (evaluated at the current values of parameters) that observation i arises from the statistical model implied by theory m. In other words, $\zeta_{i,m}$ represents the degree to which a specific observation is consistent with each theory. Later, we use the estimate of this probability to construct a list of observations that are statistically significantly consistent with each theory.

After the E-step, the M-step maximizes the function defined in equation (5). This step can be achieved by separately maximizing the weighted log-likelihood function for each model, i.e., $f_m(y \mid x, \theta_m)$, where the weight is given by $\zeta_{i,m}^{(t-1)}$. This is again intuitive because the weight for an observation is greater when fitting the statistical model with which this observation is consistent. The updated estimate of π_m can then be obtained by averaging $\zeta_{i,m}^{(t-1)}$ across all observations,

$$\pi_m^{(t)} = \frac{1}{N} \sum_{i=1}^{N} \zeta_{i,m}^{(t-1)}. \tag{7}$$

This step confirms the notion that π_m represents a measure of the overall performance of theory m while $\zeta_{i,m}$ measures the consistency between a specific observation and a particular theory.

When π_m is modeled as a function of covariates as in equation (3), then maximizing the weighted log-likelihood function (based on the multinomial logit regression, for example) will give the updated estimate of model parameters ψ_m. While the use of a parametric model means that the relationship given in equation (7)

no longer holds, the "averaging" of $\zeta_{i,m}$ is still used to estimate $\pi_m(W_i, \psi_m)$ because $\zeta_{i,m}$ is used as a weight when fitting the model.[10]

The E-step and M-step are repeated until convergence. The advantage of the EM algorithm is its numerical stability (each iteration increases the observed-data likelihood function in equation (2)) and its relatively straightforward implementation (the M-step can be implemented through successive fitting of standard statistical models with the weighted likelihood function). The disadvantage is that convergence can be slow and standard errors must be computed separately (bootstrap or the numerical estimation of the Hessian matrix can provide approximate standard errors for all parameters including π_m).

Alternatively, Bayesian inference can be applied to finite mixture models. Here, the standard approach is to use the Markov chain Monte Carlo (MCMC) algorithm with data augmentation where the latent variable Z_i is sampled along with model parameters. Bayesian inference requires the specification of prior distributions. For example, Dirichlet distribution is often used as the prior distribution for π_m. Given the prior distribution, the MCMC algorithm takes the following general form,

1. Sample Z_i given the current values of all parameters with the following probability,

$$\begin{aligned}\Pr\left(Z_i^{(t)} = m \mid \Theta^{(t-1)}, \Pi^{(t-1)}, \{Y_i, X_i\}_{i=1}^N\right) \\ = \zeta_{i,m}^{(t-1)} \propto \pi_m^{(t-1)} f_m\left(Y_i \mid X_i, \theta_m^{(t-1)}\right),\end{aligned} \tag{8}$$

   for $i = 1, \ldots, N$ and $m = 1, \ldots, M$.
2. Given $Z_i^{(t)}$, sample all parameters.
   (a) Given the subset of the data with $Z_i^{(t)} = m$, update θ_m using the standard MCMC algorithm for this particular model.
   (b) Update π_m using the standard MCMC algorithm. For example, if the Dirichlet distribution is used as the prior distribution, then we have,

$$\left(\pi_1^{(t)}, \ldots, \pi_M^{(t)}\right) \sim \text{Dirichlet}\left(s_1 + \sum_{i=1}^N \mathbf{1}\{Z_i = 1\}, \ \ldots, \ s_M + \sum_{i=1}^N \mathbf{1}\{Z_i = M\}\right), \tag{9}$$

   where $(s_1, \ldots, s_M)$ is the vector of Dirichlet prior parameters.

Again, the advantage of this MCMC algorithm is its simple implementation and ability to produce uncertainty estimates of all parameters including π_m. In particular, standard MCMC algorithms for each of the submodels can be used to sample parameters from their joint posterior distribution.

Finally, note that fitting mixture models can be computationally difficult given that the likelihood often contains multiple modes, which may pose problems for the EM algorithm. The mixing of the standard MCMC algorithm can also be poor. Thus, it is important to check the convergence carefully and run multiple independent chains with overdispersed starting values (Gelman et al., 2004).

Grouped Observations. In some situations, multiple observations are grouped and researchers may wish to assume that all observations of one group arise from the same statistical model implied by either a particular theory or more generally from the same weighted combination of statistical models under investigation. For example, multiple observations may be collected over time for each individual in a study, and all observations from one individual are assumed to be consistent with one of the competing theories or a particular weighted combination of these theories. In the trade policy example discussed earlier, it may be reasonable to assume that all votes on a particular bill are generated by a single theory because they all share the same level of factor mobility.

In such a situation, finite mixture models can be formulated as,

$$Y_{ij} \mid X_{ij}, Z_i \sim f_{Z_i}(Y_{ij} \mid X_{ij}, \theta_{Z_i}), \tag{10}$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, J_i$ where the latent variable Z_i is indexed by group i alone while both the outcome and the covariates are observed J_i times for a given group i. As specified in equation (3), one may model the conditions under which a particular theory applies using a vector of covariates observed at group level, W_i, which may be a function of X_{ij}.

We note that when the number of groups is small, the parameter ψ_m in $\pi_m(W_i, \psi_m)$ may not be precisely estimated. On the other hand, this grouping approach may result in a greater posterior probability that each group is consistent with a particular theory because multiple observations for each group provide more information about the relative appropriateness of each theory. To see this formally, note that the E-step of the EM algorithm is given by calculating the following conditional

[10] The introduction of the covariates W_i does not alter the basic relationship between π_m and $\zeta_{i,m}$. To see this clearly, consider the following setup. Suppose that W_i is discrete and we employ a nonparametric model for $\pi_m(W_i, \psi_m)$. This can be done by fitting the saturated model, which includes indicator variables for all possible values of W_i. Then, for any w, the estimate of $\pi_m(w, \psi_m)$ can be obtained by computing the mean value of $\zeta_{i,m}$ among the observations which have this specific covariate value $W_i = w$. That is, $\pi_m(w, \psi_m)$ still represents the population proportion of observations generated by theory m within the subpopulation of observations with $W_i = w$ and equation (7) holds within each of these strata.

expectation,

$$\zeta_{i,m}^{(t-1)} = \frac{\pi_m^{(t-1)} \prod_{j=1}^{J_i} f_m(Y_{ij} \mid X_{ij}, \theta_m)}{\sum_{m'=1}^{M} \pi_{m'}^{(t-1)} \prod_{j=1}^{J_i} f_{m'}(Y_{ij} \mid X_{ij}, \theta_{m'})} \quad (11)$$

instead of equation (6) where the multiplication of densities can yield a greater discrimination power across different theories.[11]

Calculating Usual Quantities of Interest. Along with π_m and $\zeta_{i,m}$, one may also be interested in calculating usual quantities of interest such as predicted and expected values under finite mixture models. To do this, given the specified values of the observed covariates, i.e., X_i and W_i if the mixing probability π_m is modeled as a function of W_i, estimate the quantities of interest under each of the competing models as well as π_m for each m. Then, the weighted average of these estimates where the weights are given by the estimated values of π_m represents the quantity of interest under the mixture model. To account for the estimation uncertainty, standard methods such as bootstrap, Monte Carlo approximation (King et al., 2000), and Bayesian simulation can be applied to calculate confidence intervals and standard errors. Note that it is important to account for the estimation uncertainty associated with the mixing probability π_m as well as other model parameters.

Identification of Observations Consistent with Each Theory

One advantage of the proposed mixture modeling approach is its ability to yield a list of observations for which researchers have sufficiently strong evidence that they are consistent with one of the competing theories. To do this, we focus on the posterior probability, $\zeta_{i,m}$, that observation i is consistent with theory m. This parameter can be estimated as part of either the EM algorithm or the MCMC algorithm (given in equations (6) and (8), respectively, or equation (11) in the case of grouped observations). Our proposal is to apply a prespecified threshold λ_m and call observation i *statistically significantly consistent* with theory m if its corresponding probability is greater than this threshold, i.e., $\hat{\zeta}_{i,m} > \lambda_m$.

How shall we choose an optimal value of λ_m? A naive selection of the value of λ_m will result in a list with many falsely classified observations due to the well-known multiple testing problem of false positives. For example, suppose that we conduct m independent hypothesis tests with the conventional 5% significance level. Even when the null hypothesis is true in all cases, the probability that at least one null hypothesis is falsely rejected increases rapidly with m; it equals $0.4 \approx 1 - 0.95^{10}$ when $m = 10$.

Thus, we propose to construct the longest list possible while at the same time controlling for the expected proportion of incorrect classifications on the resulting list. This can be done by applying the key insight from the fast-growing statistical literature on multiple testing. Specifically, for each theory m, we choose the smallest value of λ_m (so that we can include as many observations on the list as possible) while ensuring that the posterior expected value of false discovery rate on the resulting list does not exceed a certain threshold α_m.[12] The value of α_m needs to be selected by a researcher a priori. For example, we may choose $\alpha_m = 0.05$. This strategy yields the following expression for the optimal value of λ_m under the proposed criterion (see Genovese and Wasserman 2003; Newton et al., 2004; Storey 2003),

$$\lambda_m^* = \inf \left\{ \lambda_m : \frac{\sum_{i=1}^{N} (1 - \hat{\zeta}_{i,m}) \mathbf{1}\{\hat{\zeta}_{i,m} \geq \lambda_m\}}{\sum_{i=1}^{N} \mathbf{1}\{\hat{\zeta}_{i,m} \geq \lambda_m\} + \prod_{i=1}^{N} \mathbf{1}\{\hat{\zeta}_{i,m} < \lambda_m\}} \leq \alpha_m \right\}, \quad (12)$$

where the numerator represents the posterior expected number of falsely classified observations on the resulting list and the denominator represents the total number of observations on the list (the second term denominator avoids the division by zero).

Alternatively, we can obtain the optimal threshold applicable to all theories by controlling the expected false discovery rate across all lists, which yields the following thresholding formula for a single value of λ that is applicable to all rival theories,

$$\lambda^* = \inf \left\{ \lambda : \frac{\sum_{i=1}^{N} \sum_{m=1}^{M} (1 - \hat{\zeta}_{i,m}) \mathbf{1}\{\hat{\zeta}_{i,m} \geq \lambda\}}{\sum_{i=1}^{N} \sum_{m=1}^{M} \mathbf{1}\{\hat{\zeta}_{i,m} \geq \lambda\} + \prod_{i=1}^{N} \prod_{m=1}^{M} \mathbf{1}\{\hat{\zeta}_{i,m} < \lambda\}} \leq \alpha \right\}. \quad (13)$$

Together, this provides a principled way of identifying observations consistent with each of the competing theories under investigation.

[11] Similarly, in the MCMC algorithm, the conditional posterior for $\zeta_{i,m}$ is proportional to $\pi_m^{(t-1)} \prod_{j=1}^{J_i} f_m(Y_{ij} | X_{ij}, \theta_m)$ rather than what is given in equation (8).

[12] False discovery rate is known as FDR due to Benjamini (1995). See Ho and Imai (2008) for the first application of the FDR method in political science.

We emphasize that this test makes sense only if researchers interpret the mixture model as generating each observation from one rival theory. Thus, researchers who believe that each observation is implied by a weighted combination of all rival theories may not employ this test. Even in this case, however, $\zeta_{i,m}$ still represents the degree to which observation i is consistent with theory m. Regardless of which interpretation they adopt, researchers can also measure the overall performance of rival theories, the topic to which we now turn.

Measuring the Overall Performance of Rival Theories

Finally, we propose two ways to formally assess the overall performance of each theory. First, we can estimate the population proportion of observations consistent with each theory. For each theory m, the estimate of π_m represents this proportion and either its maximum likelihood or Bayesian estimate is obtained as a result of the EM or MCMC algorithm. Alternatively, π_m can be interpreted as the average degree to which observations are consistent with one of the competing theories. When π_m is modeled as a function of observed covariates, one can use the expected sample proportion of observations consistent with each theory. This measure can be calculated as the average of $\hat{\zeta}_{i,m}$ across all observations in the data, i.e., $\sum_{i=1}^{N} \hat{\zeta}_{i,m}/N$.

In addition, if we assume that each observation is consistent with only one theory, we may use the number of observations that are identified as statistically significantly consistent with one of the competing theories as a measure of overall performance. The idea is to focus on the observations for which we have strong evidence rather than to construct a measure by including ambiguous cases. In particular, the overall performance of a competing theory can be measured with the sample proportion of observations statistically significantly consistent with the theory. This measure is attractive because the observations for which the value of $\hat{\zeta}_{i,m}$ is neither close to zero or one may correspond to the cases explained by a theory other than those included in the mixture model.

Implementation in Statistical Software

A straightforward way to estimate finite mixture models for comparative theory testing is to use `flexmix` (Grün and Leisch, 2008a), which is an add-on package freely available for the statistical software **R**.[13] The `flexmix` package uses the EM algorithm to obtain the maximum likelihood estimates for a wide range of mixtures of regression models. Along with the replication code and data for this article, we provide an example syntax below so that others can use it as a template.

If researchers are capable of simple statistical programming, it is also possible to estimate finite mixture models that are not available in the existing software. We provide such examples in a later section (bivariate probit regression model) and the supporting materials (semiparametric logistic regression model). Such an extension is straightforward because we can rely upon the existing functionalities within the general framework of finite mixture models. Similarly, if researchers wish to implement Bayesian finite mixture models, they can take advantage of the existing MCMC algorithm implementation, including the ones available in the `MCMCpack` package (Martin, Qunn, and Park 2009), for fitting various models.

Comparison with the Other Common Approaches

Next, we briefly compare the proposed mixture modeling approach with some of the alternative methods often used by applied researchers. Perhaps the most common approach to empirical testing of competing theories is to construct a regression model that encompasses all relevant theories and then examine the magnitude and statistical significance of coefficients corresponding to each theory. Achen (2005) criticizes this widespread approach as atheoretical and calls it a "garbage-can regression" because no single theory can justify the model specification of such regressions with many explanatory variables. A number of other scholars share this concern (e.g., Braumoeller 2003; Clarke 2000, 2007b; Gordon and Smith 2004; Granato and Scioli 2004).

Parsimony is also regarded by many social scientists as an important criterion for theory development. For example, Friedman states, "A hypothesis is important if it 'explains' much by little, that is, if it abstracts the common and crucial elements from the mass of complex and detailed circumstances surrounding the phenomena to be explained and permits valid predictions on the basis of them alone" (1966, 14). The mixture modeling approach is consistent with this view. Rather than fitting a regression model with many covariates which encompass

[13] There also exists a STATA module, called `FMM`, to fit finite mixture models, but its capability is currently quite limited for the purpose of comparative theory testing.

all theories under consideration, it allows for empirical testing of several parsimonious statistical models, each of which is justified by a particular theory.

Political scientists who have abandoned the "garbage-can" approach have used various model selection methods to test competing theories. Popular methods include Bayesian information criteria, the Vuong test (Vuong 1989), the J test (Davidson and MacKinnon 1981), and the Clarke test (Clarke 2007a). These methods are useful because they enable the comparison of two non-nested models (Clarke 2000). Indeed, some of the methods share the same motivation as the mixture modeling approach.

In particular, the J test, which is used by Hiscox (2002) in the application described in an earlier section, is based on the following mixture setup with the mixing probability π,

$$Y_i = (1 - \pi) f(X_i, \beta) + \pi g(X_i, \gamma) + \epsilon_i, \quad (14)$$

where the null hypothesis, $H_0 : Y_i = f(X_i, \beta) + \epsilon_i$, is tested against the alternative hypothesis, $H_1 : Y_i = g(X_i, \gamma) + \epsilon_i$. Under this setup, the statistical test is conducted to see whether π is equal to zero or not (if it is, the null hypothesis is retained). Unlike the mixture approach, therefore, it is difficult to use the covariates W_i to model the mixing probability, i.e., $\pi(W_i)$. In addition, the J test does not allow researchers to make inferences about the applicability of rival theories to each observation in the sample. Finally, in order for the J test to be applicable, one needs to be able to write a model as in the following form, $Y_i = h(X_i, \beta) + \epsilon_i$, for a possibly nonlinear function $h(\cdot\,, \cdot)$ (Davidson and MacKinnon 1981).[14] In contrast, finite mixture models can incorporate virtually all of the likelihood-based models.

More generally, a fundamental difference between the mixture modeling approach and the standard model selection methods is that the latter hypothesize one theory applies to all observations, whereas the former allows for competing theories to coexist. In the presence of theoretical heterogeneity, standard model selection procedures may yield an ambiguous conclusion (appropriately so!). In contrast, the mixture modeling approach can quantify the degree of such heterogeneity and identify the conditions under which each theory applies, which facilitates further theoretical development. This difference is evident in the mixture setup of the J test given in equation (14) where the test is conducted with the null hypothesis of $\pi = 0$ against the alternative hypothesis $\pi = 1$, ignoring the possibility that π may take a value other than 0 and 1. Thus, unless one theory applies to the entire population, for the purpose of testing alternative theories, mixture modeling is more appropriate than standard model selection procedures.[15]

Another general problem of standard model selection procedures is the potential bias arising from the fact that the usual standard errors do not incorporate the uncertainty concerning model selection because they are calculated assuming that a particular model is correct. This means that since any model selection procedure yields false positives, the standard errors associated with the estimated parameters of the selected model are inaccurate and often too small (see, e.g., Freedman, 1983; Freedman, Navidi, and Peters, 1988). In contrast, mixture models take into account all the estimation uncertainty including the one concerning the applicability of each model to specific observations.

Finite mixture models are similar to random coefficient models (also known as multilevel models), which are essentially a generalization of models with interaction terms (e.g., Beck and Katz 2007; Gelman and Hill 2007). For example, both methods can easily incorporate grouping of observations that naturally arise in substantive problems. This is difficult to do within the framework of standard model selection procedures. However, several notable differences exist. While random coefficient models account for theoretical heterogeneity within a single-regression framework by varying coefficients across groups of observations, finite mixture models explicitly use a different regression model for each theory and yield both overall and observation-specific measures of different theories' explanatory power. Another difference is that whereas standard random coefficient models require, a priori, the specification of groups across which coefficients are allowed to vary, finite mixture models use the data to decide which group (or theory) each observation belongs to. Thus, we argue that for the purpose of empirical testing of competing theories, finite mixture models are more appropriate than random coefficient models.

Finally, Bayesian model averaging offers an approach that is conceptually quite similar to finite mixture modeling (see Hoeting et al. 1999; Imai and King 2004). The idea is to build a final model by computing the weighted average of multiple models according to the Bayes factor of each model. Like mixture models, this method therefore accounts for model uncertainty and avoids the preliminary testing problem of standard model selection

[14] Nevertheless, applied researchers often use the J test for logistic regression models and others that cannot be written in this form (see Collier and Hoeffler 2004; Hiscox 2002; Ladewig 2006, for recent examples).

[15] Even in this case, the mixture model is applicable because it will estimate π_m to be close to zero for the theories that completely lack explanatory power.

methods discussed above. Nevertheless, there are important differences. Aside from the fact that it is applicable only within the framework of Bayesian inference, Bayesian model averaging focuses on the overall assessment of competing theories and the improvement of prediction capability by combining multiple models. In contrast, finite mixture models allow researchers to explore the conditions under which each theory is applicable and identify a set of observations that are consistent with a specific theory. It is also much easier to group observations for each theory using the clustering formulation discussed earlier.

Simulation Studies

In this section, we conduct simulation studies to explore the conditions under which the proposed method works (or does not work) well. We investigate cases with two and three competing theories and also compare the results with other common procedures. In general, we find, as expected, that more information in the data (e.g., larger sample size, continuous outcome instead of binary outcome) improves the performance of mixture models.

Two-Theory Mixture Model Simulation

We begin with a simple data-generation process with two competing regression models, each of which consists of a different covariate and an intercept. These two covariates are sampled independently from a bivariate normal distribution with zero mean, unit variances, and correlation equal to 0.5. We use a binary logistic regression with one theory-predicting variable, which is independently sampled from normal distribution with mean 10 and variance 2.[16] Given this setup, we vary the logit coefficients so that the population proportion of observations consistent with Model 1 ranges from 0.1 to 0.9. Finally, two sets of outcome variables are generated. The continuous outcome variable is sampled from a linear regression with the standard normal variate error while the binary outcome is generated according to the logistic regression. The results are based on 1,000 Monte Carlo experiments.

The four left plots in Figure 1 show the estimated proportion of observations consistent with Model 1 $\hat{\pi}_1$ (the vertical axis) against their true values π_1 (horizontal axis) for eight different simulation settings; sample size is set to either 1,000 (solid triangles with dashed lines) or 5,000 (solid circles with solid lines), and the model is either the linear regression for continuous outcome variables (first column) or binary logistic regression for dichotomous outcomes (second column), with (top row) or without (bottom row) theory-predicting variables. The results show that for both continuous and binary outcomes, the mixture model approach recovers the true proportion of observations consistent with each theory. The model works somewhat better when the outcome is continuous and when the sample size is larger.

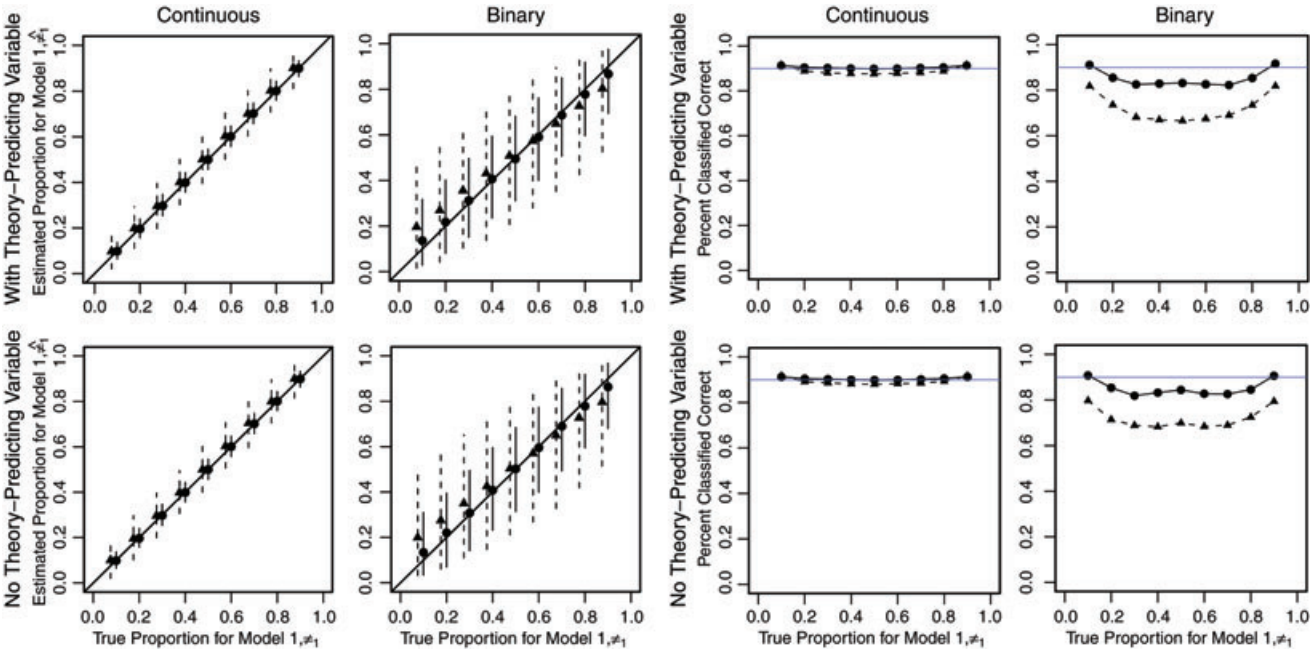
We next examine the performance of the proposed classification method for identifying observations that are statistically significantly consistent with each theory. The right four plots in Figure 1 show the classification success rates of the proposed method. Each plot of the two right columns uses the same simulation setup as the corresponding plot in the two left columns. We set the false discovery rate to $\alpha = 0.1$, which means that if the method is working appropriately, we would expect the classification success rates to be approximately 90%.

Again, the results show that the proposed method works best when the data are most informative. The best performance is obtained when the outcome variable is continuous with the large sample size. In the binary outcome case, the proposed method has larger classification error than its nominal rate, but the performance significantly improves when the sample size is larger. In addition, although not shown in the plots, the number of classified observations increases along with the amount of information. For example, simulations with a binary outcome and no theory-predicting variable often had less than 20% of observations classified to either theory, whereas simulations with a continuous outcome and theory-predicting variable regularly had greater than 50% of observations classified.

Finally, we compare the proposed method of classification with the two alternative methods—the Bayesian information criteria and the Vuong test. Figure 1 of the supporting materials presents the proportion of times when Model 1 is viewed as a better model according to these methods. As expected, this proportion becomes larger as the number of observations consistent with Model 1 increases. However, this result is not comforting when all observations do not come from a single theory, in which case it is misleading to conclude from these methods one model completely dominates the other.

[16] We have also examined the situation where the theory-predicting variable is correlated with other covariates. This changes the results relatively little except for slightly decreasing the number of observations that get classified to one of the models.

FIGURE 1 Estimated Population Proportion of Observations Consistent with Model 1 (four left plots) and Classification Success Rates (four right plots) in the Two-Theory Mixture Model Simulation Study



Note: The results of eight simulations are reported in the figure; sample size is set to either 1,000 (solid triangles with dashed lines) or 5,000 (solid circles with solid lines), the model is either the linear regression for continuous outcome variables (first and third columns) or binary logistic regression for dichotomous outcomes (second and fourth columns), and with (top row) or without (bottom row) the theory-predicting variable. In the four left plots, the horizontal axis represents the true proportion of observations consistent with Model 1, π_1, while the vertical axis represents the estimated proportion, $\hat{\pi}_1$. Solid symbols indicate the average of estimates and vertical lines represent the range from 5 percentile to 95 percentile of the sampling distribution of $\hat{\pi}_1$. The expected false discovery rate is set to $\alpha = 0.1$. The vertical axis represents the proportion of successful classification among the observations that are classified to either theory. Since this classification success rate equals 1 minus false discovery rate, we should expect the proposed procedure to give a classification success rate approximately 0.9 (indicated by blue solid horizontal line) when it is working appropriately. The eight plots together show that the proposed method performs better when the outcome variable is continuous and the sample size is larger.

Three-Theory Mixture Model Simulation

Next, we examine the performance of the mixture model with three competing theories. The simulation setup is nearly identical to the above case with two competing theories. First, we sample three covariates (one for each theory) from a multivariate standard normal distribution with all pair-wise correlations set to 0.5. Next, each observation is assigned to one of the models according to the predetermined proportions, which range from 0.2 to 0.8 for Model 1. This step is achieved by fixing coefficients to certain values in the multinomial logit model. As before, we consider two sample sizes (1,000 and 5,000), two outcome variable types (continuous and binary), and without a theory-predicting variable.[17]

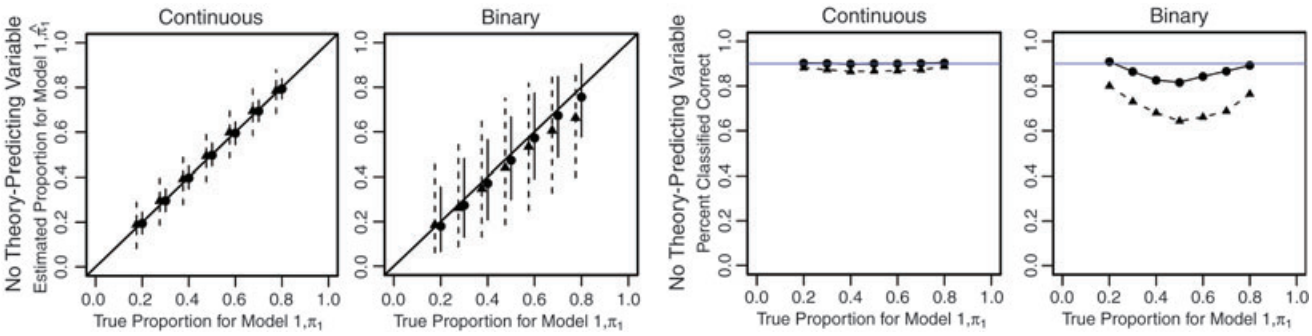
[17]We also conducted simulations with a theory-predicting variable and found a pattern similar to the one in the two-theory simulation. The results are omitted due to the space constraint.

The results are based on 1,000 Monte Carlo simulations and are presented in Figure 2, whose plots are formatted in the manner identical to those in the upper row of Figure 1. The plots reveal a pattern similar to the results for the two-theory simulation studies presented above. The proposed method performs best when the outcome variable is continuous and the sample size is large. The direct comparison between two-theory and three-theory simulations is difficult because the models are different, but the simulation results suggest that the observed pattern is similar between the two scenarios.

Empirical Results

In this section, we apply the proposed mixture modeling approach to the motivating empirical example concerning competing theories of trade policy preferences

FIGURE 2 Estimated Population Proportion of Observations Consistent with Model 1 (two left plots) and Classification Success Rates (two right plots) without a Theory-Predicting Variable in the Three-Theory Mixture Model Simulation Study



Note: The format of these plots is identical to those given in the upper row of Figure 1. See its caption for details.

introduced in an earlier section. We also test three competing theories of democratic peace by revisiting the work of Huth and Allee (2002).

Competing Theories of Trade Policy Preferences

Background and Data. The data set Hiscox (2002) collected spans over 150 years and contains the information about roll-call voting regarding 26 and 29 trade bills in the U.S. House and Senate, respectively. Each bill is coded as either protectionist or protrade. The outcome variable, a vote, is coded 1 if a legislator votes on a particular bill against liberalization and 0 if the legislator votes for. The data set also contains covariates regarding the factoral and industrial makeup of each state, which operationalize the two competing theories. For the Stolper-Samuelson (SS) theory, Hiscox codes the variable `profit` as state-level measures of profits, the variable `manufacture` as employment in manufacturing, and the variable `farm` as agricultural production.[18] For the Ricardo-Viner (RV) model, Hiscox uses the measures of the export and import orientation of a state, `export` and `import`, respectively.[19]

Finally, the data contain the national-level measure of factor specificity, which is the key theory-predicting variable. Hiscox was unable to collect a single measure of specificity over the entire period. Instead, he uses various measures and shows that all trend closely together over time in terms of the coefficient of variation across industries (see Figures 1 and 2 of Hiscox 2002). To create a single measure of factor specificity for each year, we use the coefficient of variation based on one of the two following measures given their availability: the annual earnings in 20 industries and the annual earnings of productive workers measures.[20] As can be seen from Figure 1 of Hiscox (2002), the resulting measure `factor` spans the entire period and tracks other measures well. Thus, this variable takes a greater value when factors are relatively specific (i.e., immobile) and a smaller value when factors are relatively nonspecific.

Statistical Analysis. We begin our analysis by estimating a mixture model of two logistic regression models, one with the three covariates corresponding to the SS model as main effects and the other with the three covariates corresponding to the RV model as main effects. Furthermore, instead of using fixed effects for each logistic regression as done in Hiscox (2002), we use a mixture model with clustering where all votes for a particular trade bill are assumed to be consistent with the same theory. This is a reasonable approach given that factor specificity is measured and operates at the level of national economy. Finally, we model the mixing probability (or the population proportion of observations consistent with the RV model), π,

[18]Specifically, these three measures refer to profits earned by capital in manufacturing (value-added minus wage payments) as a fraction of the state income, total employment in manufacturing as a proportion of each state's population, and total value of agricultural production as a fraction of state income, respectively.

[19]They are measured as total production in the 10 top export and import competing industries as a proportion of the state income, respectively.

[20]Not surprisingly, alternative combinations of measures to cover the entire time period produce similar results. Hiscox was also not able to collect the factor specificity data for every year in which there was a vote. For these years, we linearly impute the missing data as a rough approximation.

using a logistic regression with factor specificity variable as the only covariate. Under the maintained hypothesis, we expect the coefficient for this variable to be positive since a greater level of factor specificity is more likely to yield support for the RV model. In sum, using our previous notation, we have the mixture model with the following components,

$$f_{SS}(Y_{ij} \mid X_{ij}, \theta_{SS}) = \text{logit}^{-1}(\beta_0 + \beta_1 \texttt{profit}_{ij} + \beta_2 \texttt{manufacture}_{ij} + \beta_3 \texttt{farm}_{ij})$$

$$f_{RV}(Y_{ij} \mid X_{ij}, \theta_{RV}) = \text{logit}^{-1}(\gamma_0 + \gamma_1 \texttt{export}_{ij} + \gamma_2 \texttt{import}_{ij})$$

$$\pi_{RV}(W_j, \phi_{RV}) = \text{logit}^{-1}(\delta_0 + \delta_1 \texttt{factor}_j)$$

where i and j index votes and bills, respectively.

We estimate this model using the R package, `flexmix`, and here we provide a syntax to illustrate the simplicity of implementation in the hope that it may serve as a template for other researchers. First, the explanatory variables for each model and nesting structure must be specified. In this case, the models are completely non-nested except that both models include the intercept. Thus, we specify two separate formulas in the following manner,

```
model <- FLXMRglmfix(family = "binomial",
          nested = list(k = c(1, 1),
          formula = c(~ profit + manufacture
          +farm, ~ export + import)))
```

where `family` specifies the logistic (default link) regression for binary outcome, `k` within the `nested` argument specifies two models being fitted, each of which has one component, and `formula` tells which variables belong to each model.

Next, we specify the outcome variable `vote`, whether all votes for the same bill should be clustered, and the model for how specificity influences the mixture probabilities. In this step, we pass the `model` object produced in the first step to the function `stepFlexmix()`, which estimates the model using the EM algorithm with different random starting values to avoid local maxima. The syntax is as follows,

```
result <- stepFlexmix(cbind(vote, 1 - vote) ~
           1|bill, k = 2, model = model,
           concomitant = FLXPmultinom(factor),
           data = Hiscox, nrep = 20)
```

where `|bill` represents a standard R syntax for clustering for each bill, `k` is the number of competing models, `concomitant` specifies the logistic regression model with `factor` as the sole covariate to model the mixing probabilities, `data` is the data frame, and `nrep` specifies the total number of EM algorithm runs with different starting values.

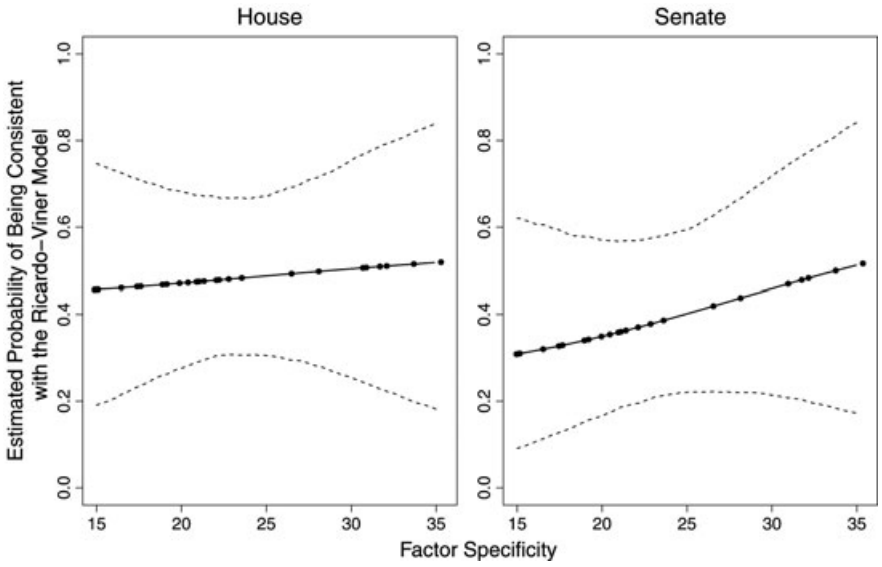
Figure 3 plots the estimated population proportion of observations consistent with the RV model, $\hat{\pi}$, across the range of factor specificity variable. This serves as an overall measure of applicability of each theory. For both House and Senate, the point estimates are consistent with the hypothesis that a greater level of factor specificity makes it more likely for legislators' votes to be explained by the RV model. The fact that the estimated probability ranges from 0.3 to 0.5 implies that factor specificity only partially explains the theoretical heterogeneity, and there may exist other important determinants of the applicability of each theory.

While the point estimates are consistent, the statistical insignificance of the factor specificity variable and the resulting wide confidence intervals in the figure suggest that the evidence for Hiscox's hypothesis is rather weak. This does not necessarily refute his hypothesis because the statistical power is low (the data contain only 26 and 29 bills for the House and Senate, respectively). On the other hand, the fact that the model was able to classify many bills with high probabilities means that legislative voting on many of these bills can be explained well by either the RV or SS variables. In sum, our reanalysis suggests that these two trade models have high explanatory power, but a more precise test of Hiscox's argument requires a larger data set with more bills.

Next, we illustrate the method described earlier and identify the list of trade bills, which are statistically significantly consistent with each of the rival theories. Here, we assume that all votes for any given bill are consistent with the same theory. While this assumption is rather strong, it is similar to that of the original analysis where all votes in one period are hypothesized to be consistent with either the SS or RV model. In fact, we are able to classify all bills even when we set the (posterior) expected number of incorrect classification to be very small, e.g., $\alpha = 0.01$. Table 1 of the supporting materials provides the resulting classification lists of trade bills for each model. Scholars may use these lists to examine whether quantitative evidence is in agreement with qualitative knowledge about each trade bill.[21]

[21] While the majority of the bills passed within the House and Senate in a given year are classified as belonging to the same theory (18/23), there nevertheless exist several instances where this is not the case (5/23). Such differences might arise from differences in voting dynamics across the institutions, the particulars of the bills across the institutions if prior to reconcilement in conference committee, or limitations in the application of our approach to this particular example.

FIGURE 3 Estimated Probability of Votes for a Bill Being Consistent with the Ricardo-Viner Model as a Function of Factor Specificity



Note: Solid line is the estimated probability with actual observations indicated by sold circles, and dashed lines represent 95% confidence intervals based on the Monte Carlo approximation. Although there is a considerable degree of uncertainty due to the small number of bills, the positive slopes in the House (the left panel) and Senate (the right panel) are consistent with the hypothesis that the Ricardo-Viner model rather than the Stolper-Samuelson model is supported when the level of factor specificity is high.

Comparison with Other Methods. Finally, the mixture modeling approach also yields estimated model parameters for each of the competing theories, i.e., θ_{SS}, θ_{RV}, as well as the estimated coefficients on variables that are used to estimate mixing probabilities, i.e., ψ_{RV}. In Table 1, we report these estimates and compare them to the "garbage-can" regression (the last four columns), which Achen (2005) and others (e.g., Clarke 2000; Gordon and Smith 2004) argue should be avoided. Here, the "garbage-can" regression refers to the single logistic regression, which contains all five variables taken from both SS and RV models. Following Hiscox's original analysis, we also include bill fixed effects in this model.

The table shows that for the mixture modeling approach, *all* estimated coefficients of the two models have expected signs and are statistically significant. For example, the estimated coefficient for the `farm` variable is negative, implying that states with high levels of agricultural production are more likely to oppose protectionism as expected under the Stolper-Samuelson model. In contrast, in the "garbage-can" regression the coefficients are considerably smaller and their standard errors are larger (relative to the size of the coefficients). For example, the `farm` variable is not statistically significantly different from zero both in the House and Senate.[22] This suggests the superior discrimination power of each variable in the mixture model despite the fact that the "garbage-can" regression was fit to the entire data.

The results based on the mixture model also improve upon those reported in the original article. For example, the application of the J test indicates that the SS model is selected for the period between 1945 and 1962. However, Hiscox found that the `farm` and `manufacture` variables in this model have opposite signs than what is predicted (2002, 603). In contrast, the results of the mixture model show no such inconsistency. Furthermore, when the SS model (with bill fixed effects) is fitted to the subset of votes classified to the RV model given in the second and fourth columns of Table 1 of the supporting materials, the estimated coefficient for the `farm` variable has a positive sign (statistically insignificant in the House and

[22]We also ran the same "garbage-can" regression model with the interaction terms between the `factor` variable and each of the covariates. The results are somewhat puzzling. The coefficients of some main effects do not have expected signs, and others are no longer statistically significantly different from zero. In addition, some signs of the coefficients for these interaction terms are not in the expected direction.

TABLE 1 Parameter Estimates and Their Standard Errors from the Mixture Model for the House and Senate

		Mixture Model				"Garbage-can" Model			
		House		Senate		House		Senate	
Models	Variables	coef.	s.e.	coef.	s.e.	coef.	s.e.	coef.	s.e.
Stolper-Samuelson	intercept	−0.23	0.14	0.02	0.21	0.47	0.12	0.78	0.25
	profit	−1.60	0.53	−5.69	1.19	−0.93	0.56	−3.58	1.23
	manufacture	17.60	1.54	19.79	2.59	10.01	1.11	7.82	2.27
	farm	−1.33	0.29	−1.27	0.43	−0.14	0.24	−0.03	0.42
Ricardo-Viner	intercept	−0.61	0.05	−0.83	0.13				
	import	3.09	0.33	2.53	0.80	1.03	0.34	2.22	0.76
	export	−0.85	0.16	−2.80	0.77	−1.45	0.14	−2.58	0.36
Mixture Probability	intercept	−0.39	1.48	−1.60	1.62				
	factor	0.01	0.06	0.05	0.07				

Each model is the logistic regression with model intercepts omitted in order to ease presentation. The first set of models uses the proposed mixture model approach with bill clustering. The second set is based on a "garbage-can" regression that uses all variables from both Stolper-Samuelson and Ricardo-Viner models as well as bill fixed effects.

significant in the Senate), which is opposite to what the SS model predicts. On the other hand, when the model is fitted separately to the "correct" subset of the trade bills, then *all* estimated coefficients are statistically significant and have the expected sign in both the House and Senate. This provides evidence supporting the appropriateness of bill classifications as a whole.

Competing Theories of Democratic Peace

Next, we apply the proposed mixture modeling approach and test three competing theories of democratic peace. Specifically, we revisit the work of Huth and Allee (2002), who empirically test three competing models—the accountability model, the norms model, and the affinity model. Here, we focus on the military escalation stage where a defending state has already refused negotiations and each state in a conflict dyad must choose whether or not to escalate the dispute.

Background and Data. The original data set consists of 374 military confrontations between 1919 and 1995 in which a challenging state had initiated a conflict against a defender. Huth and Allee construct separate dichotomous dependent variables for a challenger and a defender, which equal 1 if a state chose high levels of military escalation and 0 for low or limited escalation. For each of the competing models, the authors estimate a bivariate probit model to allow for correlation between the challenger's decision to escalate crisis and the corresponding decision of the defender.

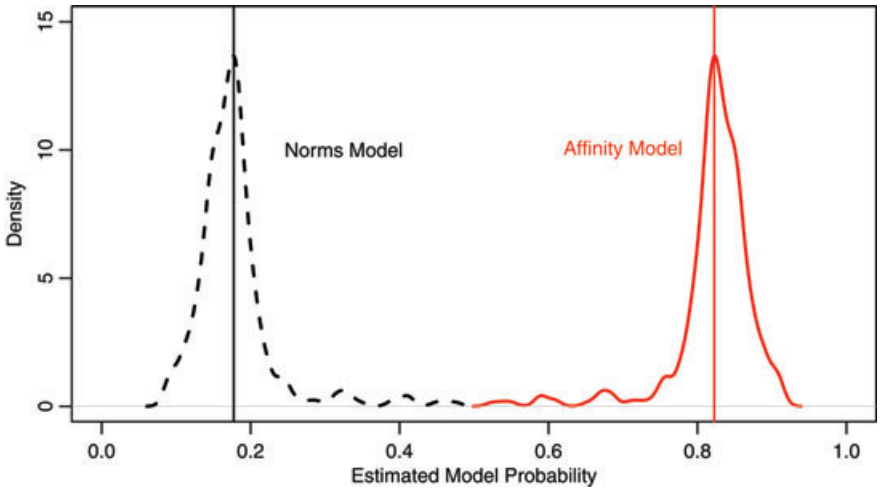
The accountability model argues that leaders are constrained in their ability to use force by domestic political institutions and threats from rivals. In particular, competitive elections can constrain leaders in a crisis. To operationalize this idea, Huth and Allee use measures of democracy levels for challenger and defender and interact them with various characteristics of disputes.[23] The norms model emphasizes the role of beliefs held by political leaders about how to negotiate and deal with political conflict. Specifically, the argument is that norms about domestic bargaining transfer to the international level. They use a measure of how strong nonviolent norms are in the state and interact it with several characteristics of the dispute.[24] Finally, the affinity model argues that conflict decision making is driven by shared interests and ideologies. As a measure of similarity, Huth and Allee use an indicator variable representing whether countries have the same regime type and another variable indicating if this similarity measure has changed in the last five years. In addition, the authors include a set of "realist" control variables in each of the three models (see Tables 9.4, 9.13, and 9.19).

Statistical Analysis. Based upon the lack of statistical significance of the key estimated coefficient and its wrong sign, Huth and Allee conclude that out of the three models the affinity model "produced the weakest results" (2002, 283) and that between the accountability

[23]They include indicator variables for whether the dispute is a stalemate, part of an enduring rivalry, if ethnic conationals are involved in the dispute, whether there is high military risk, and a measure of the target's resolve.

[24]They include a dichotomous variable indicating whether the dispute is a stalemate and another dichotomous variable indicating whether the state possesses military advantage.

FIGURE 4 Smoothed Histograms of Estimated Probabilities That Each Observation Is Consistent with Each Competing Theory of Democratic Peace, $\hat{\zeta}_{i,m}$



Note: Solid vertical lines represent the estimated overall probability that observations are consistent with each model, $\hat{\pi}_m$. The affinity model receives the greatest support. The estimated probability for the accountability model is essentially zero for all observations.

and norms model the accountability model was superior (2002, 286). Here, we formally test the three competing theories using the proposed mixture modeling approach. We begin our analysis by fitting the mixture model consisting of Huth and Allee's three bivariate probit models.[25] Figure 4 presents the smoothed histogram of estimated posterior probabilities that each observation is consistent with each competing theory, $\hat{\zeta}_{i,m}$. We find that the affinity model receives the greatest support where slightly more than 80% of observations are estimated to be consistent with this model.[26] However, the mixture model shows essentially no support for the accountability model, which contradicts the original finding. Interestingly, as shown in Table 2, the estimated coefficients for the other models—affinity and norms models—from the mixture model are quite similar, though the coefficients are estimated less precisely for mixture models. Note that the standard errors on nearly all of the variables are very large, making inferences about the signs of coefficients inappropriate. Here, the mixture model is choosing the most parsimonious model. This suggests that the realist control variables are explaining most of the variation in the outcome variable.

[25]The bivariate probit model is not available in the `flexmix` package, and hence we have programmed the EM algorithm.

[26]This conclusion is similar to the one given by Clarke (2008) using the Friedman test and Bayesian information criteria.

What Can Go Wrong?

What are potential pitfalls of finite mixture models? In this section, we list several limitations of mixture modeling and discuss practical recommendations that help applied researchers avoid them (see also Section 4 of the supporting materials for empirical illustration via simulation). First, the proposed mixture modeling approach provides one way to assess the relative *predictive* performance of rival theories, but like any statistical method, the method in itself does not solve endogeneity and other fundamental problems of *causal* inference in observational studies. For example, one may estimate causal effects using a mixture model which consists of causal submodels. In such cases, the inclusion of relevant confounders in each of the submodels will be required in order to identify causal effects.

Second, one should not test too many competing theories at once. Fitting a mixture model demands much more from the data than fitting each of the submodels separately. The fact that each submodel is identified does not necessarily imply a mixture of all submodels is identified. Even if a mixture model is identified, like any statistical modeling, overfitting can be a problem too. For example, including too many statistical models, especially the ones that are similar to each other and/or have many parameters, can lead to inefficient and sensitive inference in a small sample. The data may simply

TABLE 2 Parameter Estimates and Their Standard Errors for the Affinity and Norms Models from the Mixture Model and Standard Bivariate Probit Model Used in the Original Analysis by Huth and Allee (2002)

	Mixture Model		Huth and Allee	
	coef.	**s.e.**	**coef.**	**s.e.**
Affinity Model				
Challenger				
Political Similarity	0.005	0.005	0.005	0.005
Change in Political Similarity	0.003	0.009	0.003	0.007
Defender				
Political Similarity	−0.204	0.265	−0.233	0.260
Change in Political Similarity	0.784	1.419	0.929	1.330
Norms Model				
Challenger				
Nonviolent Norms	0.004	0.002	0.004	0.002
Stalemate	0.015	0.027	0.014	0.029
Nonviolent Norms × Stalemate	−0.003	0.003	−0.002	0.003
Nonviolent Norms × Military Advantage	−0.003	0.002	−0.003	0.002
Defender				
Nonviolent Norms	0.047	0.028	0.073	0.023
Stalemate	0.216	0.656	0.283	0.531
Nonviolent Norms × Stalemate	−0.004	0.098	−0.001	0.051
Nonviolent Norms × Military Advantage	−0.016	0.026	−0.025	0.021

Note: Each model also contains a set of control variables, which are omitted from this table. Standard errors are based upon the nonparametric bootstrap. The two methods give similar results for both models.

lack enough information to distinguish all models. For this reason, we recommend that researchers test only two or three competing theories with typical political science data sets. Overfitting can also be avoided by making sure that out-of-sample predictions of mixture models are as good as their in-sample predictions.

Third, while identification is still possible (see Grün and Leisch 2008b; Hennig 2000), high correlations across predictors may reduce the statistical power of mixture models.[27] There may also be a bias toward the selection of a submodel with a greater number of predictors, especially when a more parsimonious model generates relatively few observations and/or correlations across predictors are high (see Section 4 of the supporting materials).[28] We emphasize that definitive theoretical results about this question do not exist and indeed in one of our empirical applications, the most parsimonious model is selected. Unlike other model selection procedures such as the Bayesian information criteria, however, the mixture model does not explicitly penalize models with a large number of parameters. Therefore, when using the proposed approach, substantive theory (rather than statistical methods) must guide model specification.

Finally, while mixture modeling allows one to model the conditions under which different theories are applicable, these conditions must be directly derived from the underlying assumptions of each rival theory. This is exactly the contribution made by Hiscox (2002), who realized that the relative applicability of Stolper-Samuelson and Ricardo-Viner depends on their assumption about factor mobility. Although the inclusion of theory-predicting variables is appealing for both theoretical and statistical reasons, this does not mean that one can use any variables to predict the applicability of rival theories.[29] Even if it

[27]Note that most applications of mixture models in the statistical literature use the same predictors in all submodels. Thus, the suggested use of mixture modeling should have fewer problems than typical applications.

[28]Such issues have been reported in the case of the *J* test (e.g., Godfrey and Pesaran 1983).

[29]In addition, we emphasize that as observed in our analysis of the trade policy preference example, the statistical power may be low for detecting the factors that determine the applicability of each rival theory, thereby requiring a large sample size.

avoids the "garbage-can" regression, such an approach may be condemned as a "garbage-truck" model![30]

Concluding Remarks

We have shown that finite mixture models can be used to effectively conduct empirical testing of competing theories. Given that the mixture modeling strategy outlined in this article can accommodate a wide range of statistical models, we believe that the applicability of the proposed methodology is potentially high. Although finite mixture models have a long history in statistics, their main use has been to make parametric models flexible so that they fit the data better. We have shown that this methodology can be used for the empirical testing of competing theories, which is a central goal of social science research.

One important advantage of the proposed mixture modeling strategy is its ability to model the conditions under which different theories are applicable. Any theory rests upon certain assumptions, without which the theory is not applicable. However, much of empirical research takes for granted these assumptions when conducting theory testing. Evaluating the underlying assumptions is especially critical when testing rival theories because the applicability of each theory depends upon the appropriateness of different assumptions. With finite mixture models, researchers can now test a theory as a whole, including its assumptions, and explore the factors that determine when each rival theory is applicable. Given the ease of using finite mixture models, we believe that more scholars should collect variables like Hiscox's factor mobility measure and then use them directly in their statistical analysis.

References

Achen, C. H. 2005. "Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong." *Conflict Management and Peace Science* 22(4): 327–39.

Beck, N., and J. N. Katz. 2007. "Random Coefficient Models for Time-Series-Cross-Section Data: Monte Carlo Experiments." *Political Analysis* 15(2): 182–95.

Benjamini, Y., and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B* 57(1): 289–300.

Brandt, P. T., and J. R. Freeman. 2006. "Advances in Bayesian Time Series Modeling and the Study of Politics: Theory Testing, Forecasting, and Policy Analysis." *Political Analysis* 14(1): 1–36.

Braumoeller, B. 2003. "Causal Complexity and the Study of Politics." *Political Analysis* 11(3): 209–33.

Clarke, K. A. 2000. "Testing Non-nested Models of International Relations: Reevaluating Realism." *American Journal of Political Science* 45(3): 724–44.

Clarke, K. A. 2007a. "A Simple Distribution-Free Test for Non-nested Model Selection." *Political Analysis* 15(3): 347–63.

Clarke, K. A. 2007b. "The Necessity of Being Comparative: Theory Confirmation in Quantitative Political Science." *Comparative Political Studies* 40(7): 886–908.

Clarke, K. A. 2008. *A Nonparametric Approach to Testing Multiple Competing Models.* Unpublished manuscript, University of Rochester.

Collier, P., and A. Hoeffler. 2004. "Greed and Grievance in Civil War." *Oxford Economic Papers* 56(4): 563–95.

Davidson, R., and J. G. MacKinnon. 1981. "Several Tests for Model Specification in the Presence of Alternative Hypotheses." *Econometrica* 49(3): 781–93.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion)." *Journal of the Royal Statistical Society, Series B, Methodological* 39: 1–37.

Freedman, D. A. 1983. "A Note on Screening Regression Equations." *The American Statistician* 37(2): 152–55.

Freedman, D. A., W. C. Navidi, and S. C. Peters. 1988. "On the Impact of Variable Selection in Fitting Regression Equations." *In On Model Uncertainty and Its Statistical Implications,* ed. T. K. Dijkstra. Berlin: Springer-Verlag, 1–16.

Friedman, M. 1966. "The Methodology of Positive Economics." *In Essays in Positive Economics*. Chicago: University of Chicago Press, 3–160.

Frühwirth-Schnatter, S. 2007. *Finite Mixture and Markov Switching Models*. New York: Springer.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. *Bayesian Data Analysis*. 2nd ed. London: Chapman & Hall.

Gelman, A., and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Genovese, C., and L. Wasserman. 2003. "Bayesian and Frequentist Multiple Testing." In *Bayesian Statistics 7*, ed. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West. Oxford: Oxford University Press, 145–61.

Godfrey, L. G., and M. H. Pesaran. 1983. "Tests of Non-nested Regression Models: Small Sample Adjustments and Monte Carlo Evidence." *Journal of Econometrics* 21(1): 133–54.

Gordon, S., and A. Smith. 2004. "Quantitative Leverage Through Qualitative Knowledge: Augmenting the Statistical Analysis of Complex Cases." *Political Analysis* 12(3): 233–55.

Granato, J., and F. Scioli. 2004. "Puzzles, Proverbs, and Omega Matrices: The Scientific and Social Significance of Empirical Implications of Theoretical Models (eitm)." *Perspectives on Politics* 2(2): 313–23.

Grün, B., and F. Leisch. 2008a. "FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters." *Journal of Statistical Software* 28(4): 1–35.

Grün, B., and F. Leisch. 2008b. "Finite Mixtures of Generalized Linear Regression Models."*In Recent Advances in*

[30]We thank John Kastellec for coining this term.

Linear Models and Related Areas, ed. C. Heumann. Heidelberg: Physica-Verlag, 205–30.

Harrison, G. W., and E. E. Rutström. 2009. "Expected Utility Theory and Prospect Theory: One Wedding and a Decent Funeral." *Experimental Economics* 12(2): 133–58.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.

Hennig, C. 2000. "Identifiability of Models for Clusterwise Linear Regression." *Journal of Classification* 17(2): 273–96.

Hill, J. L., and H. Kriesi. 2001. "Classification by Opinion-Changing Behavior: A Mixture Model Approach." *Political Analysis* 9(4): 301–24.

Hiscox, M. J. 2002. "Commerce, Coalitions, and Factor Mobility: Evidence from Congressional Votes on Trade Legislation." *American Political Science Review* 96(3): 593–608.

Ho, D. E., and K. Imai. 2008. "Estimating Causal Effects of Ballot Order from a Randomized Natural Experiment: California Alphabet Lottery, 1978–2002." *Public Opinion Quarterly* 72(2): 216–40.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. "Bayesian Model Averaging: A Tutorial." *Statistical Science* 14(4): 382–417.

Huth, P., and T. Allee. 2002. *The Democratic Peace and Territorial Conflict in the Twentieth Century*. Cambridge: Cambridge University Press.

Iaryczower, M., and M. Shum. 2009. "The Value of Information in the Court: Get It Right, Get It Tight." *American Economic Review*. Forthcoming.

Imai, K., L. Keele, D. Tingley, and T. Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(4): 765–89.

Imai, K., and G. King. 2004. "Did Illegal Overseas Absentee Ballots Decide the 2000 U.S. Presidential Election?" *Perspectives on Politics* 2(3): 537–49.

Imai, K., and D. A. van Dyk. 2005. "A Bayesian Analysis of the Multinomial Probit Model Using Marginal Data Augmentation." *Journal of Econometrics* 124(2): 311–34.

Kedar, O. 2005. "When Moderate Voters Prefer Extreme Parties: Policy Balancing in Parliamentary Elections." *American Political Science Review* 99(2): 185–99.

King, G., J. E. Alt, N. E. Burns, and M. Laver. 1990. "A Unified Model of Cabinet Dissolution in Parliamentary Democracies." *American Journal of Political Science* 34(3): 846–71.

King, G., M. Tomz, and J. Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2): 341–55.

Ladewig, J. W. 2006. "Domestic Influences on International Trade Policy: Factor Mobility in the United States, 1963 to 1992." *International Organization* 60(1): 69–103.

Lieberman, E. S. 2005. "Nested Analysis as a Mixed-Method Strategy for Comparative Research." *American Political Science Review* 99(3): 435–52.

Martin, A. D., K. M. Quinn, and J. H. Park. 2009. *MCMCpack: Markov Chain Monte Carlo MCMC Package*.

Martin, A. D., K. M. Quinn, and J. H. Park. 2011. "*MCMCpack: Markov chain Monte Carlo in R.*" *Journal of Statistical Software* 42(9): 1–21.

Newton, M. A., A. Noueiry, D. Sarkar, and P. Ahlquist. 2004. "Detecting Differential Gene Expression with a Semiparametric Hierarchical Mixture Method." *Biostatistics* 5(2): 155–76.

Quandt, R. E. 1972. "A New Approach to Estimating Switching Regressions." *Journal of the American Statistical Association* 67(338): 306–10.

Rogowski, R. 1989. *Commerce and Coalitions: How Trade Affects Domestic Political Alignments*. Princeton NJ: Princeton University Press.

Scheve, K. F. and M. J. Slaughter. 2001. "What Determines Individual Trade Policy Preferences?" *Journal of International Economics* 54(2): 267–92.

Storey, J. D. 2003. "The Positive False Discovery Rate: A Bayesian Interpretation and the *q*-value." *Annals of Statistics* 31(6): 2013–35.

Vuong, Q. H. 1989. "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses." *Econometrica* 57(2): 307–33.

Warwick, P., and S. Easton. 1992. "The Cabinet Stability Controversy: New Perspectives on a Classic Problem." *American Journal of Political Science* 36(1): 122–46.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

- **S1:** Comparison with Standard Approaches
- **S2:** Semi-parametric mixture modeling without clustering
- **S3:** Classified trade bills
- **S4:** Illustration of pitfalls of mixture modeling via simulation

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.