Supporting Materials for "A Statistical Method for Empirical Testing of Competing Theories"

Kosuke Imai* Dustin Tingley†

November 20, 2011

1 Comparison with Standard Approaches

Within each simulation study conducted in Section 4 of the manuscript, we also examine the performance of the standard model selection procedures commonly used in the literature; the Bayesian Information Criterion (BIC) (Schwarz, 1978) and the Vuong test (Vuong, 1989). Specifically, we record, according to these model selection methods, which model is selected or if no model is selected at all. For the Vuong test the p-value of 0.05 was used as the threshold, whereas for the BIC the absolute difference of 8 was used.

Figure 1 presents the results. Here, we plot the percentage of times Model 1 is chosen given that one of the two models is selected. Both methods tend to support one model disproportionately when the mixture probability, π_1 is close to 0 or 1. For example, if 80% of observations were generated by Model 1 and 20% by Model 2, both methods would almost never choose Model 2. Unlike the mixture model which recovers π_1 fairly well in this case, the standard model selection procedures ignore the fact that some observations are consistent with Model 2. More concerning is the performance of these tests

*Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609–258–6610, Email: kimai@princeton.edu, URL: http://imai.princeton.edu

†Ph.D. student, Department of Politics, Princeton University. Email: dtingley@princeton.edu, URL: http://www.princeton.edu/~dtingley

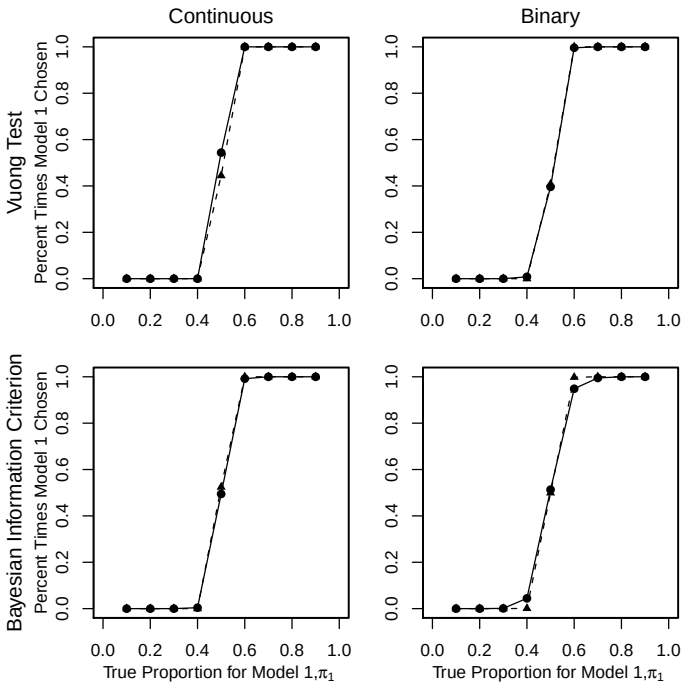

Figure 1: Percentage of Times Model 1 Chosen by the Vuong Test (top row) and the Bayesian Information Criterion (bottom row) When A Model is Selected. Simulations are conducted for continuous (left column) and binary (right column) outcome variables with the sample size of $1,000$ (dashed lines) and $5,000$ (solid lines). The vertical axis represents the percentage of times Model 1 is chosen given that either model is chosen.

at intermediate values of π_1. For example, when an exactly half of observations comes from each of the two theories, i.e., $\pi_1 = 0.5$, these tests may incorrectly select one of the two models (though the Vuong test often fails to reject the null hypothesis of no difference).

These results hold irrespective of the outcome variable type and the sample size. While standard model selection procedures have some desirable characteristics when all of the data is generated by one model versus the other (Clarke, 2007), when the data generating process follows a mixture of different models these tests can be misleading. In contrast, the mixture models can handle the situation where almost all observations come from one theory, by yielding a large estimated value for the proportion of observations consistent with that theory.

2 Semi-parametric Mixture Modeling without Clustering

Next, we assess the robustness of the above results by relaxing the clustering assumption to allow for the possibility that different individual votes for the same bill may be consistent with different theories. We

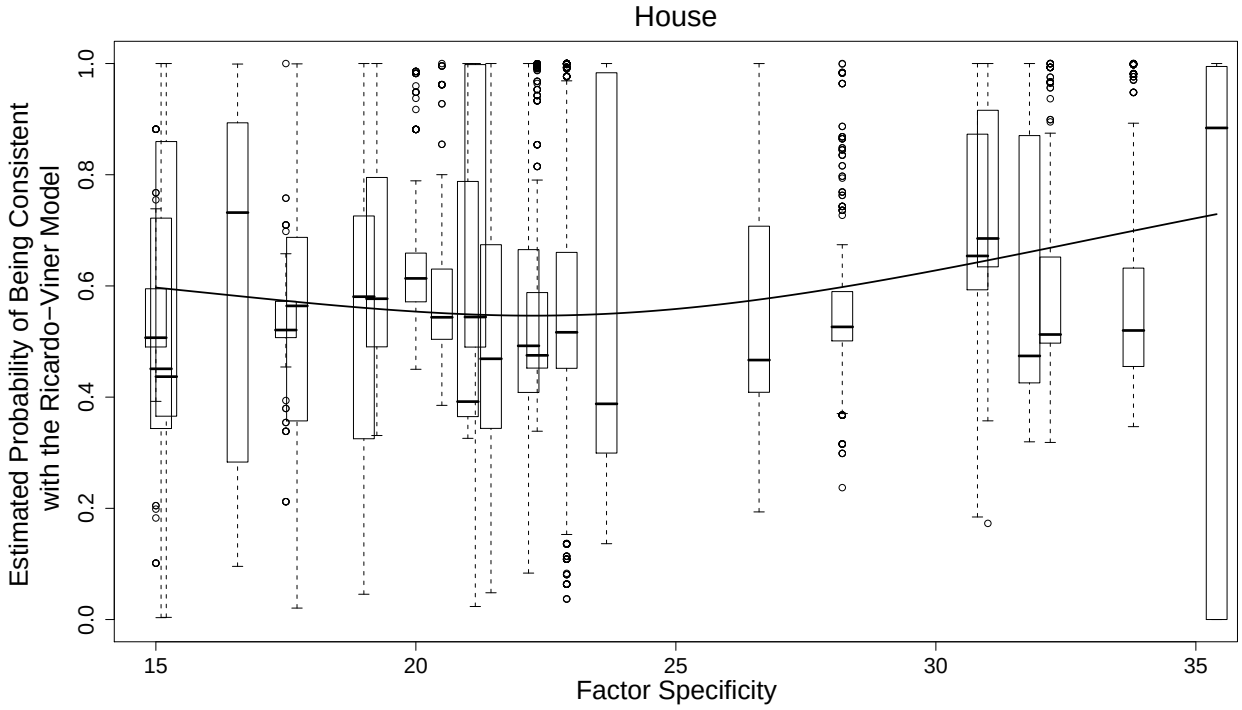

Figure 2: π of Ricardo-Viner model as a function of factor specificity. Model estimated EM with a generalized additive model (GAM) allowing for a smooth influence of the factor specificity variable. Model includes bill level fixed effects.

also relax the parametric assumption made about the relationship between the level of factor specificity and the appropriateness of each theory so that we do not a priori assume the relationship is monotonic. Finally, following Hiscox, we include bill fixed effects. In sum, we use the following model,

$$f_{\text{SS}}(Y_{ij} \mid X_{ij}, \theta_{\text{SS}}) = \text{logit}^{-1}(\beta_{0j} + \beta_1 \texttt{profit}_{ij} + \beta_2 \texttt{manufacture}_{ij} + \beta_3 \texttt{farm}_{ij}) \quad (1)$$

$$f_{\text{RV}}(Y_{ij} \mid X_{ij}, \theta_{\text{RV}}) = \text{logit}^{-1}(\gamma_{0j} + \gamma_1 \texttt{export}_{ij} + \gamma_2 \texttt{import}_{ij}) \quad (2)$$

$$\pi_{\text{RV}}(W_j, \phi_{\text{RV}}) = \text{logit}^{-1}(\delta_0 + s(\texttt{factor}_j)) \quad (3)$$

where $s(\cdot)$ represents the unknown smooth function. We use the Generalized Additive Models (Hastie and Tibshirani, 1990) within the EM algorithm to estimate the smooth function as well as other model parameters. This illustrates the flexibility of finite mixture models where sophisticated models such as semi-parametric models can be fitted in a relatively straightforward manner.

Figure 2 summarizes the results for the House where the solid line represents the estimated mixing probability π and box plots represent the distribution of vote-specific posterior probabilities ζ for each trade bill. The overall result is similar to what we found above. Across much of the range, the estimated

mixing probability monotonically increases as the level of factor specificity goes up although a slight nonlinear relationship appears to exist when the level of factor specificity is low. Unlike the previous analysis, however, the posterior probabilities vary considerably across legislators' votes for any given bill. This is expected given that we have relaxed the clustering assumption. The finding suggests that while the results are consistent with the theoretical expectation, the amount of unexplained variation is rather large.

3 Classified Trade Bills

Table 1 provides a list of trade bills that are classified to either the Stolper-Samuelson or Ricardo-Viner model according to our proposed method and assuming each vote from one rival theory.

4 Illustration of Pitfalls of Mixture Modeling via Simulation

We conduct another set of two theory mixture simulation studies to empirically illustrate some of the pitfalls discussed above. To make the simulation setup realistic, we use the trade policy preference data analyzed in Section 4.1 of the main text. Specifically, the true values of model parameters are set to their estimates from the mixture model, presented in Table 1 of the main text. The exception is that the coefficient for the factor specificity variable is varied in order to produce different proportions of observations consistent with each theory. We then sample covariates from the multivariate normal distribution with the mean and variance equal to their sample counterparts. Finally, the two types of outcome variables are sampled by adding an independent standard normal error term for the continuous outcome variable and using the bernoulli distribution with the logit link function for the binary outcome variable.

Figure 3 shows the results where the format of all plots is identical to that of Figure 1 of the main text (see the caption of Figure 1 of the main text for details). The results show that the mixture model performs well for the linear model with the continuous outcome while the performance is not ideal when the proportion becomes more extreme. This makes sense because binary variables are in general less informative than continuous variables. As expected, a larger sample size improves the performance of the mixture model. With $5,000$ observations, the mixture model recovers π_1 almost perfectly in the

House		Senate	
Stolper-Samuelson	Ricardo-Viner	Stolper-Samuelson	Ricardo-Viner
Adams Compromise (1832)	Tariff Act (1824)	Tariff Act (1824)	Clay Compromise (1833)
Clay Compromise (1833)	Tariff Act (1828)	Tariff Act (1828)	Tariff Act (1857)
Tariff Act (1842)	Gorman Tariff (1894)	Adams Compromise (1832)	RTA Extension (1937)
Walker Act (1846)	Underwood Tariff (1913)	Tariff Act (1842)	RTA Extension (1945)
Tariff Act (1857)	RTAA (1934)	Walker Act (1846)	RTA Extension (1955)
Morrill Act (1861)	RTA Extension (1937)	Morrill Act (1861)	Trade Expansion Act (1962)
Tariff Act (1875)	RTA Extension (1945)	Tariff Act (1875)	Fast-Track (1991)
Morrison Bill (1884)	RTA Extension (1955)	Mills Bill (1988)	NAFTA (1993)
Mills Bill (1888)	Trade Expansion Act (1962)	McKinley Tariff (1890)	GATT (1994)
McKinley Tariff (1890)	Mills Bill (1970)	Gorman Tariff (1894)	
Dingley Tariff (1894)	Trade Reform Act (1974)	Dingley Tariff (1894)	
Payne-Aldrich Tariff (1909)	Fast-Track (1991)	Payne-Aldrich Tariff (1909)	
Fordney-McCumber Tariff (1922)	NAFTA (1993)	Underwood Tariff (1913)	
Smoot-Hawley Tariff (1930)	GATT (1994)	Fordney-McCumber Tariff (1922)	
Trade Remedies Reform (1984)		Smoot-Hawley Tariff (1930)	
		RTAA (1934)	

Table 1: Bill Classifications to the Stolper-Samuelson or Ricardo-Viner Model. Classifications based on posterior probability ζ of a given bill being consistent with each of the two competing theories. The (posterior) expected number of incorrect classifications on each list is less than 0.01. Vote dates are in parentheses.

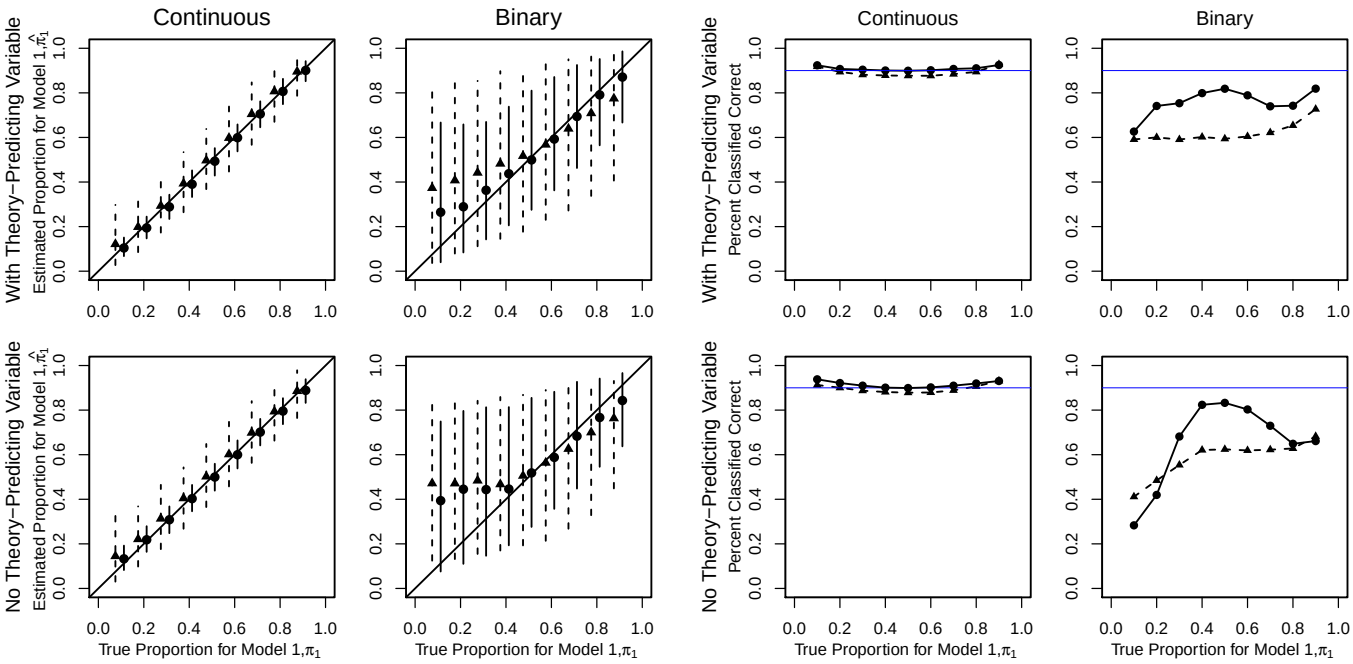

Figure 3: Results of the Two Theory Mixture Model Simulation Study based on the Trade Policy Preference Data. The format of the figure is identical to that of Figure 1 of the main text. See the caption of Figure 1 of the main text for details. The eight plots of the figure together show that the proposed method performs better when the outcome variable is continuous, the sample size is larger, and a theory-predicting variable is present.

linear case and even in the binary case the performance is significantly improved. Finally, in the case of binary model, the availability of the theory-predicting variable helps especially in the cases where one theory receives much greater support than the other. Indeed, when the sample size is large and the theory-predicting variable is available, the binary mixture model performs reasonably well.[1]

A similar observation can be made for the performance of our proposed method for identifying observations that are statistically significantly consistent with each theory. The right four plots in Figure 3 show the classification success rates of the proposed method. Each of 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%.

As it is evident in the figure, the proposed classification method works well in all the linear cases

[1]We note that with a low sample size, binary dependent variable, no theory predictive variable, a low proportion of observations from Model 1, the estimate of π_1 is higher than it should be. In this case model 1 included more covariates than model 2, possibly suggesting a bias towards less parsimonious theories in these extreme conditions.

with the continuous outcome variable (third column), regardless of the sample size and the availability of theory-predicting variable. In contrast, in the binary outcome variable case (fourth column), the method has a difficult time, correctly classifying the observations to theories. In addition, the proportion of classified observations is quite small when the outcome variable is binary and the theory-predicting variable is not available. For example, with a sample size of $N = 5,000$ in the binary case around only 10% of observations are classified without the theory-predicting variable but this roughly triples when the theory-predicting variable is included. As expected, we find that a larger sample size (solid line) and the availability of theory-predicting variable (upper-right corner) help dramatically improve the performance of the method.

Together with the evidence given in Section 4 of the main text, the set of simulation studies presented here illustrate the fact that mixture models demand far more from the data than standard regression models. When the data contain less information (e.g., smaller sample, binary outcome, lack of theory-predicting variable), the performance of the proposed mixture modeling approach may not be desirable.

References

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

Hastie, T. J. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman Hall, London.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 2, 307–333.