Flexible Causal Inference for Political Science

Bear F. Braumoeller,[*] Giampiero Marra,[†] Rosalba Radice,[‡] and Aisha Bradshaw.[*]

May 2, 2017

ABSTRACT

Measuring the causal impact of state behavior on outcomes is one of the biggest methodological challenges in the field of political science, for two reasons: behavior is generally endogenous, and the threat of unobserved variables that confound the relationship between behavior and outcomes is pervasive. Matching methods, widely considered to be the state of the art in causal inference in political science, are generally ill-suited to inference in the presence of unobserved confounders. Heckman-style multiple-equation models offer a solution to this problem; however, they rely on functional form assumptions that can produce substantial bias in estimates of average treatment effects. We describe a category of models, *flexible simultaneous likelihood models,* that account for both features of the data while avoiding reliance on rigid functional form assumptions. We then assess these models' performance in a series of neutral simulations, in which they produce substantial (55% to >90%) reduction in bias relative to competing models. Finally, we demonstrate their utility in a reanalysis of Simmons' (2000) classic study of the impact of Article VIII commitment on compliance with the IMF's currency-restriction regime.[*]

[*]Department of Political Science, The Ohio State University, Columbus.

[†]Department of Statistical Science, University College, London.

[‡]Department of Economics, Mathematics and Statistics, Birkbeck University of London.

[*]The authors are grateful to Justin Esarey, William Minozzi, and the participants in the Ohio State University Research in International Politics workshop and the Rice University Symposium on Agreements, Law, and International Politics for helpful comments.

1. INTRODUCTION

The study of political science has undergone a methodological renaissance in the past 20 years. A field that was once content to base its conclusions on garden-variety logit and probit results has created or imported methods for modeling strategic interaction (Signorino 1999; Smith 1999; Lewis 2003; c.f. Carrubba, Yuen and Zorn 2007), selection bias (Sartori, 2003; von Stein, 2005; Boehmke, Morey and Shannon, 2006; Chiba, Martin and Stevenson, 2014), split-population or partial-observability models (Xiang, 2010; Braumoeller and Carson, 2011), zero-inflated or rare events data (King and Zeng, 2001; Bagozzi, Hill, Moore and Mukherjee, 2015), network analysis (Dorussen and Ward, 2008; Hafner-Burton and Kahler, 2009; Maoz, 2009; Cranmer, Desmarais and Menninga, 2012), and more.

Unfortunately, observational studies in political science continue to be plagued by the problem of endogeneity. Endogeneity occurs when an omitted variable or variables confounds the relationship between cause and effect, thereby introducing bias into the estimate of the causal effect. Contrary to what most practitioners seem to believe, simply adding omitted confounders to the right-hand side of a linear model in an observational study does not ensure that endogeneity bias has been eliminated. And while the number of experimental studies in political science is on the rise, the field is still dominated by analyses of observational data, in which unmeasured and mis-modeled confounders are a threat to inference in even the most well-designed studies.

Fortunately, political methodologists have increasingly focused on the development of methods designed to account for it. While scholars during the so-called "age of regression" (Morgan and Winship, 2007, ch. 1) were often content to base their conclusions on simple correlational methods such as regression, hazard models, logit, and the like, the 21st century has brought with it a renewed

appreciation of the hazards of basing causal claims on observational data. Except in the case of randomized experiments, one can rarely assume that variables of interest are truly exogenous—or, put differently, that assignment to a "treatment condition" of interest, like commitment to an institution, is truly randomized. Greater appreciation of this point has brought with it an increased interest in experiments as well as a renewed focus on statistical techniques designed to account for the confounding variables that represent a major threat to causal inference in observational studies.

In this paper we elaborate one such methodology, flexible simultaneous likelihood models, that we believe conveys a host of advantages to students of political science. By virtue of their structure, these models can account for endogeneity. Unlike matching methods, the current state of the art for causal inference in political science, they can also account for the unmeasured confounders that plague observational studies. Finally, relative to traditional simultaneous likelihood models, they are far less reliant on rigid distributional assumptions that can bias results.

To establish the validity of these models, we use a series of simulations to explore their ability to reduce endogeneity bias. To illustrate their utility, we then take up the debate among Simmons (2000), von Stein (2005), and Simmons and Hopkins (2005) on the impact of formal commitment to Article VIII of the International Monetary Fund's Articles of Agreement on states' willingness to refrain from currency restrictions. By loosening the parametric assumptions of the recursive bivariate probit model and allowing for unobserved confounders we show both that von Stein was correct to argue that Article VIII produces a screening effect and that Simmons was correct to argue that it produces a constraining effect. Moreover, we demonstrate that the constraining effect, while smaller than the original study suggested, is much more persistent.

2. THE PROBLEM OF ENDOGENEITY

Endogeneity occurs when one or more omitted variables confounds the relationship between cause and effect, rendering estimates of the causal effect problematic (Kish, 1959). Because confounding is a potential threat to inference whenever the causal variable is itself caused by something else, and because virtually everything in social science is caused by something else, endogeneity is ubiquitous. For instance, democracy is said to cause peace (Maoz and Russett, 1993; Bueno de Mesquita, Morrow, Siverson and Smith, 1999; Bausch, 2015), but democracy itself is thought to be endogenous to such variables as GDP, trade, and (most troublingly) peace (Gates, Knutsen and Moses, 1996; Reuveny and Li, 2003). Similarly, military alliances are widely accepted as a tool that can reduce the risk of conflict (Leeds, 2003; Johnson and Leeds, 2011; Benson, 2011; Fang, Johnson and Leeds, 2014), but both alliance formation and conflicts are driven by the (typically unobserved) interests and security environment of the states involved. Failing to account for these interests leads to bias in the estimation of the impact of alliances on conflict (Levy, 1981; Bearce, Flanagan and Floros, 2006). In the same vein, poor economic conditions are thought to increase the risk of terrorism (Blomberg, Hess and Weerapana, 2004; Freytag, Krüger, Meierrieks and Schneider, 2011; Meierrieks and Gries, 2012), but important variables such as political freedom affect both the state of the economy and the incidence of terrorism (Grier and Tullock, 1989; Krieger and Meierrieks, 2011). Even weather, often used as an instrumental variable because of its independence from human influence, has become endogenous to anthropogenic climate change in studies of civil conflict (see, e.g., Tir and Stinnett 2012 and Theisen 2012).

In statistical terms, the endogeneity of a given variable manifests itself as an association between that variable and the error term. For example, in the linear regression $y = \mathbf{X}\boldsymbol{\beta} + \epsilon$ each of the

variables in design matrix $\mathbf{X}$ is assumed to be independent of ϵ. If some omitted variable, w, influences both y and one of the right-hand-side variables, x_1, w is said to *confound* the relationship between y and x_1 because x_1 is no longer uncorrelated with ϵ. In this case the estimator for $\boldsymbol{\beta}$ will be biased and inconsistent, with β_1 (the impact of x_1 on y) being the most affected (Wooldridge, 2010). The same logic applies to techniques like logit and probit that are designed to handle binary dependent variables.

Unfortunately, there is no perfect answer to the problem of endogeneity in observational studies. As a theoretical solution, Manski (1990) did derive a nonparametric technique that calculates the bounds within which the average treatment effect of an endogenous causal variable must lie, but those bounds are typically so broad as to be of little use to practitioners. It is worth noting, however, that any solution that offers a more precise answer than Manski's must leverage *some* assumptions in order to do so. In general, the technique that is best able to recover the causal effect of a treatment with both accuracy and precision should be the technique whose assumptions are best-suited to the circumstances. In the following sections we argue that flexible simultaneous likelihood methods are well-suited to the challenging circumstances that are typically found in the study of political science.

2.1. *Existing Solutions*

While the majority of political scientists are aware of the problem of endogeneity, at least in principle, most seem to believe that it can be taken care of simply by adding observed confounding variables to the right-hand side of a linear, logit, or probit regression (see, e.g., Simmons 2000, 829, fn 25 and Doyle and Sambanis 2000). Unfortunately, doing so generally does not resolve

the problem. While it is technically not impossible to address endogeneity bias in this manner, these technique assume both that all possible confounders have been measured and included in the equation *and* that the functional form of their relationship to y has been correctly specified—assumptions that will almost certainly not be met in practice.

A simple substantive example from the voting literature illustrates this point. In a classic study, Bartels (2000) models voting behavior in American national elections as a function of partisan identification, which is meant to measure a voter's long-term political affiliation. He admits, however, that the characteristics of a particular candidate might drive both reported partisanship and vote in a given election. It is not difficult to imagine other variables (age, socioeconomic status, the state of the economy, and so on) that might have an impact on both partisanship and vote. Unless *all* of those variables can be accounted for, the causal impact of partisanship on vote choice will be biased due to unobserved confounding. And with observational data, it is typically impossible to account for those variables with any degree of certainty.

The desire to account for confounding variables' effects on estimates of causal impact has led researchers to explore instrumental variables approaches (e.g., Simmons 2009). These typically take the linear form

$$\begin{aligned} y &= \beta_1 x_1 + \{\mathbf{X}\boldsymbol{\beta}\}_{[-\beta_1 x_1]} + \epsilon, \\ x_1 &= \mathbf{Z}\boldsymbol{\gamma} + \upsilon, \end{aligned}$$

where $\{\mathbf{X}\boldsymbol{\beta}\}_{[-\beta_1 x_1]}$ excludes the component $\beta_1 x_1$ from $\mathbf{X}\boldsymbol{\beta}$ and $\mathbf{Z}$ is a design matrix. This approach is generally preferable to ordinary single-equation models in that it explicitly addresses the endogeneity of x_1 (Angrist and Krueger, 2001). It can create some practical difficulties in that

it relies on the existence of an *instrument*—a variable in $\mathbf{Z}$ that is independent of the unobserved confounders (i.e., independent of the error terms), independent of y conditional on the unobserved and unobserved confounders, and associated with x_1. Instruments can be problematic if they are invalid, in that they are correlated with the error term, ϵ, in the outcome equation, or if they are only weakly correlated with the endogenous variable x_1 (Murray, 2006). Another, less commonly noted problem with instrumental variable estimators is that their ability to identify causal effects typically hinges on their functional form, and functional form assumptions are often made arbitrarily:

> *Given the model,* least squares and its variants can be used to estimate parameters and to decide whether or not these are zero. However, the model cannot in general be regarded as given, because current social science theory does not provide the requisite level of technical detail for deriving specifications. (Freedman and Sekhon, 2010, 46)

As we demonstrate in our simulations below, the bias introduced by even minor differences in functional form assumptions can be substantial.

It is perhaps not surprising, then, that practitioners have turned to matching methods to estimate causal effects (Simmons and Hopkins, 2005; Gilligan and Sergenti, 2008; Hill, 2010; Lupu, 2013). Matching is a simple and powerful methodology for dealing with confounding variables. The basic idea behind it is to match each "treated" unit ($x_1 = 1$) with one "control" unit ($x_1 = 0$) (or vice versa) based on observed confounders. In this way the matched data resemble, as much as possible, observations from a natural experiment in which one comparable set of units was given a treatment while the other constituted a control. Once matching has been used to achieve balance, the impact of x_1 on y can be measured via a simple difference of means, as long as the identifying assumption, selection on observables, is met. Because the technique is nonparametric, concerns about functional

form assumptions are eliminated—a feature that is often emphasized by the method's proponents.

2.2. *The Threat of Unobserved Confounders*

Matching methods are a compelling way to cut through the thicket of problems surrounding instrumental variable approaches. They are not without problems of their own, however. Foremost among them is the assumption that all confounders have been measured and incorporated into the analysis—an assumption known as "selection on observables." To the extent that this assumption has been recognized as being potentially problematic, the solution is typically to measure unmeasured confounders and include them in the analysis (see, e.g., Simmons and Hopkins 2005; Lupu 2013). As Keele (2015, 322) notes in a recent review, however, "selection on observables is a very strong assumption. It is often difficult to imagine that selection on observables is plausible in many contexts."

The selection-on-observables assumption is especially problematic in the context of observational studies in political science, for two reasons. First, measures of quantities of interest are often either approximated using existing but tenuously related data ("proxies") or simply omitted due to the cost of additional data collection. Accordingly, accounting for all unobserved confounders, even in principle, is challenging at best. Second, using matching alone, there is no way to know whether unobserved confounders remain.

Indeed, ensuring that the selection-on-observables assumption has been met can prove difficult even in actual experiments. Consider, for example, the role that unobserved confounders played in a significant controversy from the American politics literature on the efficacy of get-out-the-vote efforts prior to the 1998 election. Gerber and Green (2000) carried out a field experiment involving

30,000 registered voters in New Haven, Connecticut and concluded that, while personal canvassing increased voter turnout, telephone calls did not. The experimental nature of the study should have ruled out the influence of unobserved confounders. However, the discovery of inadvertent deviations from the authors' experimental protocols led Imai (2005) to re-balance the data using propensity score matching and conclude that telephone calls do in fact increase voter turnout, by five percentage points.

Imai's strategy was explicitly designed to rectify the deviations from randomness in the experimental design and thereby eliminate the influence of unobserved confounders. Yet when Gerber and Green (2005) used Imai's matching technique to compare the turnout rate of people in the treatment group who were not called to that of people in the control group who were not called, they found something surprising. Because neither group was contacted, assignment to treatment should be irrelevant to the outcome—in other words, the causal impact must be zero. Yet Gerber and Green found a statistically significant, 5.6% difference in turnout rates between people in the two groups. In other words, as the authors put it, "placing phone numbers on a list and *not* calling them depresses turnout" (310).

Finding a causal effect where one could not possibly exist was an elegant way of demonstrating that, even under near-ideal circumstances, statistical adjustment that relied on selection on observables for identification had failed to produce a credible estimate of the causal effect. As Sekhon (2009, 502-503) concludes in a review of the controversy, "it is clear that the selection-on-observables assumption is not valid in this case." More generally, he argues that

> [s]election on observables and other identifying assumptions not guaranteed by the design should be considered incorrect unless compelling evidence to the contrary is

provided. (503)

Consider the import of this conclusion for students of politics. The overwhelming majority of our research designs are observational, and absent a natural experiment such designs cannot make the selection on observables assumption remotely plausible in and of themselves. That leaves users of matching techniques with the second-best strategy of measuring all conceivable confounders, including them in the analysis, and hoping that they have all been accounted for. The odds that they have been, however, would seem to be extraordinarily low in any realistic application. This problem gets even worse if we consider individual subfields: If students of American politics, conducting an experiment in which the subjects are individuals, the treatment is a telephone call, and the outcome is voting cannot meet the selection-on-observables assumption, how can students of comparative politics or international relations ever hope to do so?

All of these considerations point toward a straightforward conclusion: causal inference in political science would benefit greatly from a methodology that accounts for unobserved confounders.

2.3. *Accounting for Unobservables*

Fortunately for students of political science, it is possible to account for the impact of unobserved confounders, which manifests itself as an association between the error terms of the treatment and outcome equations. Multiple-equation instrumental variable approaches accomplish this, though they rely on generally arbitrary functional form assumptions. Moreover, in the case of binary dependent and endogenous right-hand-side variables, which are ubiquitous in political science, the two-stage estimation technique typical of such studies has been shown to be deeply problematic

(Freedman, Collier, Sekhon and Stark, 2010). A more promising approach is the use of simultaneous likelihood methods—in particular, multiple equation probit models with endogenous dummy regressors, also known as recursive models. These models address endogeneity directly by estimating the coefficients in two (or more) equations simultaneously. They allow one to capture the impact of unobserved confounders by modeling the correlation between the error terms of the equations.

Despite these advantages, simultaneous likelihood methods are not often used in studies of politics (for exceptions see von Stein 2005, McLaughlin Mitchell and Hensel 2007, and Gartner 2011). This fact may to some extent be due to their relative obscurity and the recent popularity of matching methods within the discipline. A more serious concern, however, has to do with their reliance on distributional and functional form assumptions and their sensitivity to violations of those assumptions (Winship and Mare, 1992; Sartori, 2003; Simmons and Hopkins, 2005; Freedman and Sekhon, 2010). Given the number of distributional assumptions in standard simultaneous likelihood methods, this sensitivity can be nontrivial. The classic recursive bivariate probit model assumes that the latent errors of the equations follow a standard bivariate normal distribution with correlation $\theta \in [-1, 1]$, does not allow for flexible functional dependence of the responses on covariates, and only makes use of symmetric (e.g., probit) link functions. Mismodeled dependencies that appear, for instance, in the tails of the distribution linking the two equations (that a linear measure of association can not fully capture), undetected nonlinear covariate-response relationships and mismodeled probabilities related to the outcomes of two equations can have severe consequences for parameter estimation (e.g., Chib and Greenberg, 2007; Little, 1985; Monfardini and Radice, 2008; Marra and Radice, 2011, 2015).

Fortunately, each of these assumptions can be relaxed using flexible likelihood-based methods.

As the next section will demonstrate, doing so holds great promise for causal inference under the challenging circumstances that typically characterize the study of politics: pervasive endogeneity, binary dependent and endogenous right-hand-side variables, and excluded confounders.

3. FLEXIBLE SIMULTANEOUS LIKELIHOOD METHODS

The key to relaxing the bivariate normal distributional assumption of simultaneous likelihood models lies in using a function, called a *copula,* which separates the univariate marginal distribution functions from the dependence structure between the dependent variables. In a simple bivariate probit model, for example, the two marginal distributions are Normal cumulative density functions—as in a standard probit—and the copula that describes the correlation of the error terms of the dependent variables is bivariate Normal.

Once the structure of the multivariate distribution has been broken down in this manner, it is possible to relax each of the distributional assumptions in turn. The assumption that the marginal distributions are cumulative Normal can be relaxed, for example, by utilizing logit or complementary log-log distributions. The assumption that the copula is multivariate Normal can be relaxed in any number of ways. While these modifications should ideally follow concrete theory regarding the likely impact of unobservable confounders, they can also be implemented in alternative models as a way of assessing the fragility of one's findings and reducing the bias of the resulting estimator.

For example, to model covariate-response relationships in a more flexible way Chib and Greenberg (2007) and Marra and Radice (2011) introduced theoretically founded Bayesian and likelihood estimation approaches based on penalized regression splines, thereby allowing for a number of different flexible covariate-response structures. Examples in political science include Chiba,

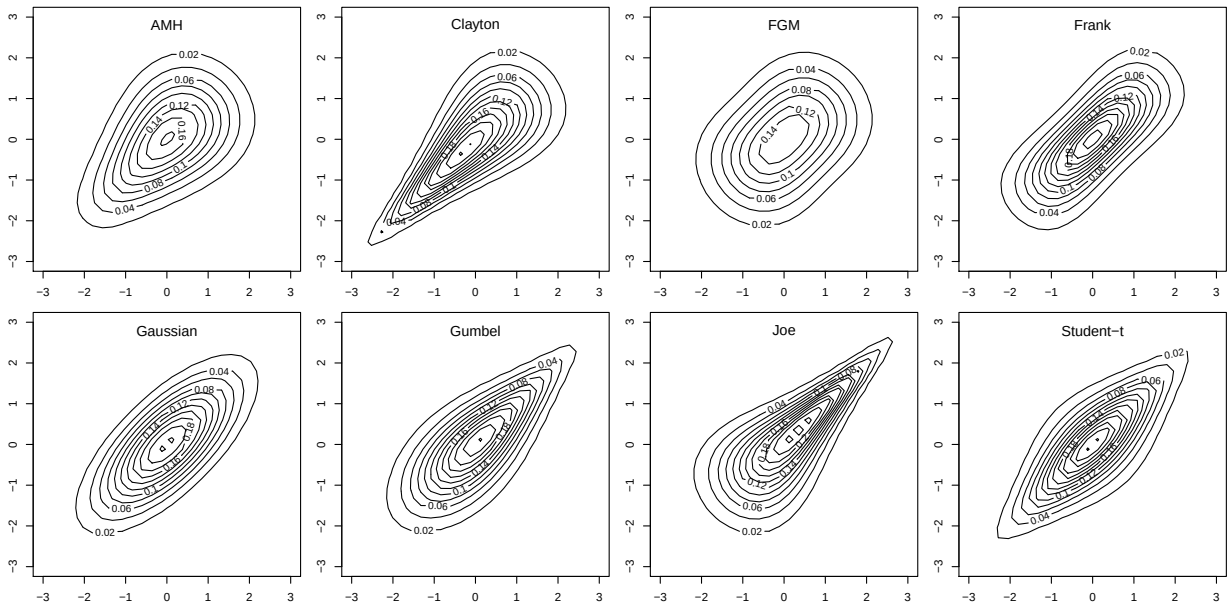

Figure 1: Examples of copulae implemented in SemiParBIVProbit that can be used to capture symmetric or asymmetric dependences among error terms of endogenous variables in a simultaneous likelihood model. A narrowing of the copula indicates stronger dependence; so, for example, the Clayton copula describes a strong dependence between negative shocks to the endogenous variables, while the Joe copula describes a strong dependence between positive shocks. The Gaussian, Frank, FGM, Student-t, and AMH copulae allow for both positive and negative dependence. AMH and Clayton are asymmetric, with a strong lower tail dependence for Clayton but a weaker upper tail dependence. The converse is true for the Gumbel and Joe copulae. The copulae that do not allow for both positive and negative dependence (Clayton, Gumbel, and Joe) can be rotated to change the direction of the tail dependence. Specifically, rotation by 180 degrees leads to the survival copula, while rotation by 90 and 270 degrees allows for negative dependence, which is not possible with the non-rotated and survival versions.

Metternich and Ward (2015) and Fukumoto (2015). To deal with the problem of non-linear dependence between outcomes, Winkelmann (2012) discussed a modification of the recursive bivariate probit which introduces non-normal dependence between the marginal distributions of the two equations using the Frank and Clayton copulae. Radice, Marra and Wojtys (2015) took a more general approach and extended the procedures discussed in Marra and Radice (2011) and Winkelmann (2012) to make it possible to deal simultaneously with unobserved confounding, flexible covariate effects and non-linear dependencies between two binary responses. In particular, they

generalized the approach based on the assumption of bivariate normality presented in Marra and Radice (2011) by allowing for non-normal dependencies between the two equations through Clayton, Frank, Student-t, Gumbel, Joe, Ali-Mikhail-Haq (AMH) and Farlie-Gumbel-Morgenstern copulae and the rotated versions of the asymmetric copulae (Clayton, Gumbel and Joe).[1] For this work we extended the scope of the modeling approach by also allowing for logit and complementary log-log link functions—an innovation in its own right. Radice, Marra and Wojtys (2015) provided a theoretical argumentation related to the asymptotic behavior of the proposed estimator and the ensuing formula to calculate the average treatment effect (ATE). Importantly, efficient and stable algorithms implementing the ideas above have been made available through the R package SemiParBIVProbit (Marra and Radice, 2016). The model can be formulated as follows:

$$\begin{aligned} y^* &= \beta_1 x_1 + f_1\left(\{\mathbf{X}\}_{[-x_1]}\right) + \epsilon, \\ x_1^* &= f_2\left(\mathbf{Z}\right) + \upsilon, \end{aligned}$$

where ϵ and υ are allowed to follow one of the Normal, logistic or Gumbel distributions with zero mean and variance equal to 1 (hence yielding probit, logit and cloglog link functions, respectively), the observed outcome variables are determined by the classic rules $y = \mathbf{1}(y^* > 0)$ and $x_1 = \mathbf{1}(x_1^* > 0)$; f_1 and f_2 represent flexible functions of the variables in $\{\mathbf{X}\}_{[-x_1]}$ and $\mathbf{Z}$. For instance, $f_1\left(\{\mathbf{X}\}_{[-x_1]}\right)$ could be equal to $\beta_2 x_2 + s(x_3) + s(x_3)x_2$, where x_2 is a binary predictor with impact β_2, $s(x_3)$ is a smooth function of the continuous covariate x_3, and $s(x_3)x_2$ is an interaction term. The probability that $y = 1$ and $x_1 = 1$ is defined as $\mathbb{P}(y = 1, x_1 = 1) = \mathcal{C}(\mathcal{F}_y(y|\pi_y), \mathcal{F}_x(x|\pi_x); \theta)$ where $\mathcal{C}$ is a two-place copula function, $\mathcal{F}_y(y|\pi_y)$ and $\mathcal{F}_x(x|\pi_x)$ are cumulative distribution func-

[1] Note that it is also possible to model positive and negative tail dependencies simultaneously by combining asymmetric copulae; this gives rise to a switching model. For instance, mixing the Clayton copula with its 90 degree (counter-clockwise) rotation allows one to model positive and negative tail dependence.The options available for the asymmetric copulae are: standard and rotated 90 degrees copulae, standard and rotated 270 degrees copulae, survival and rotated 90 degrees copulae and survival and rotated 270 degrees copulae (Marra and Radice, 2017).

tions (Normal, logistic or Gumbel) of y and x_1 taking values in (0,1) and θ an association parameter (see Figure 1 for some copula shapes[2]). The marginal distribution parameters, π_y and π_x, are related to covariates and regression coefficients via link functions g: $g_{\pi_y}(\pi_y) = \beta_1 x_1 + f_1\left(\{\mathbf{X}\}_{[-x_1]}\right)$, $g_{\pi_x}(\pi_x) = f_2(\mathbf{Z})$. θ can also be modeled through a flexible linear predictor like those used for modeling y^* and x_1^*: since the strength of the association between the treatment and outcome equations may vary across groups of observations (specifically, across years), $\theta = g_\theta^{-1}(f_3(\mathbf{W}))$, where g_θ^{-1} is a one-to-one transformation which ensures that the dependence parameter lies in its range and $\mathbf{W}$ is a design matrix (e.g., Radice, Marra and Wojtys, 2015).

It is important to note that, because of the model's complex covariate structure, the ATE cannot be inferred in a classical way. Inference is best achieved by utilizing a useful connection between Bayesian and maximum likelihood penalized regression spline estimators (Marra and Wood, 2012). This implies that intervals with close-to-nominal frequentist coverage probabilities for non-linear functions of the model coefficients (e.g., ATE) can be conveniently obtained by posterior simulation (Radice, Marra and Wojtys, 2015). This has the obvious advantage of avoiding computationally expensive bootstrap methods, which would hinder the model building process that is pivotal for practical modeling.

4. SIMULATIONS

The aim of this section is to use simulations to assess the empirical effectiveness of the copula recursive bivariate model for binary outcomes. Because simulated results are of marginal interest

[2]Contour plots of copulae with standard normal margins for data simulated using a Kendall's τ of 0.5 for all copulae except AMH, where a value of $1/3$ (the maximum allowed for this copula) was used.

when the data-generating process conforms to the assumptions of the authors' preferred model and the parameters can be set to degrade the performance of its competitors, we focus mostly on simulations in which *none* of the models captures the correct distributional assumptions. The goal is to compare model performance under what we take to be the most likely scenario for political scientists attempting causal inference: at least one confounder is omitted, the marginal distributions and dependence structure are unknown, and the goal is to recover the best (i.e., least biased and lowest RMSE) estimate of the ATE.

In the following simulations a binary instrument, two continuous observed confounders, one continuous unobserved confounder, a binary treatment and a binary outcome are denoted as z_3, z_1, z_4, z_2, x and y, respectively. We constructed the responses y and x using several distributions for the unobserved confounding variable z_2 (standard Normal, Student's t with four degrees of freedom, χ^2 with one degree of freedom and uniform[-3, 3]) and used logit link functions for the treatment and outcome variable, giving four simulation scenarios in total. Variable z_3 was simulated with 2 categories (0 and 1) with $\Pr(z_3 = 1) = 0.5$. Variables z_1 and z_4 were generated from uniform distributions over [0,1]. Non-linear covariate effects between y and z_1 and z_4, and between x and z_1 and z_4 were also introduced. The coefficient of the unobserved confounder z_2 was set to -0.85 in the treatment equation. The sample size was 1000. Each scenario was replicated 250 times. Following a reviewer's suggestion, we conducted further simulations with different values of the coefficient of z_2 in order to assess the size of bias and RMSE in relation to the importance of the unobserved confounder. A full description of the simulations is found in the Appendix.

Table 1 compares the bias and root mean squared error (RMSE) for the ATE from a univariate model (i.e., a recursive model that assumes uncorrelated errors and therefore adjusts only on observed covariates), genetic matching (an approach that matches on individual observed covariates

using an automated search algorithm to balance covariates), a recursive bivariate model with Gaussian copula (Gaussian C in the table), and a flexible recursive bivariate model in which the preferred model is selected by AIC (this is referred to as Flexible C).[3] Note that the recursive bivariate model with Gaussian copula and probit link functions corresponds to a bivariate probit.

Given the existence of an unobserved confounder, we should expect the first two models, both of which assume selection on observables, to perform poorly, and they do: bias ranges from 32%–121% and 41%–145%, respectively, and RMSE from 0.17–0.40 and 0.22–0.49.

Moving to a standard recursive bivariate model with a Gaussian copula, we can see that simply allowing for unobserved confounders dramatically reduces both bias and RMSE relative to the models that assume selection on observables: decreases in average bias of 90% or more are not uncommon. By contrast, allowing different marginal distributions across recursive bivariate models with a Gaussian copula does virtually nothing to decrease bias or improve RMSE.[4]

Use of the flexible copula model improves our estimates even further. On the whole, use of the flexible copula model results in a reduction by about 55% of the bias of the ATE relative to a standard recursive model with the same marginal distributions. This difference is greatest in the case of non-Gaussian errors, where the performance of the traditional recursive bivariate model worsens significantly whereas that of the flexible copula model remains reasonably consistent. In these cases, reduction of the bias of the ATE averages about 64%. Differences in ATE estimates between Gaussian and flexible copula models were generally statistically significant at $p = 0.05$.[5]

[3] An anonymous reviewer suggested that we include the results from a flexible recursive bivariate model in which matching had been used as a pre-processing step. While the model was not designed to be combined with matching, we were nevertheless curious enough to try it. The bivariate model with matching gave *more* biased estimates of the ATE than the bivariate model without matching, though it was a significant improvement over matching alone.

[4] Experiments with more flexible marginal distributions that required the estimation of additional parameters—skewed probit, for example—resulted in little improvement but produced substantial identification issues.

[5] The only exceptions are the logit-logit and probit-probit cases with $\mathcal{N}(0, 1)$ errors, where the assumptions of the conventional recursive bivariate model most closely reflected the data-generating process.

	Link	Logit-Logit				Probit-Probit				Cloglog-Cloglog			
	Distribution	$\mathcal{N}(0,1)$	t_4	χ^2_1	$\mathcal{U}(-3,3)$	$\mathcal{N}(0,1)$	t_4	χ^2_1	$\mathcal{U}(-3,3)$	$\mathcal{N}(0,1)$	t_4	χ^2_1	$\mathcal{U}(-3,3)$
% Bias													
Model	Univariate	33	53	42	119	32	53	41	119	33	43	40	121
	Matching	42	90	80	145	42	89	80	144	41	87	80	143
	Gaussian C	3	3	12	11	2	4	13	6	3	4	14	10
	Flexible C	3	0	5	4	2	1	6	1	6	0	7	4
RMSE													
Model	Univariate	0.17	0.25	0.24	0.39	0.17	0.25	0.24	0.39	0.17	0.25	0.23	0.40
	Matching	0.23	0.44	0.46	0.49	0.22	0.44	0.46	0.48	0.22	0.43	0.46	0.48
	Gaussian C	0.05	0.06	0.08	0.07	0.06	0.06	0.08	0.08	0.07	0.06	0.08	0.07
	Flexible C	0.06	0.07	0.05	0.08	0.06	0.06	0.05	0.07	0.08	0.07	0.07	0.08

Table 1: % Percentage biases and RMSEs for the average treatment effect of an endogenous right-hand-side variable, estimated on data simulated using logit links for the treatment and outcome variable, and normal, Student's t, χ^2 and uniform distributions for the unobserved confounder, when setting the coefficient of the unobserved confounder to -0.85 in the treatment equation. The percentage bias is defined as the average difference between the estimator and the true parameter divided by the true parameter. The RMSE is the square root of the mean of the squared deviations of the ATE estimates from their true values. Bias tells us how well or poorly the estimator does on average in estimating the ATE, while the RMSE reflects both the bias and the precision of the estimator. For both statistics lower numbers indicate a better-performing estimator; examining both helps to diagnose whether a poorly performing estimator suffers from imprecision or bias or both. The 250 simulated datasets were used to fit Gaussian, Frank, Clayton$_{90}$, Clayton$_{270}$, Joe$_{90}$, Joe$_{270}$, Gumbel$_{90}$ and Gumbel$_{270}$ copulae with logit, probit and complementary-log-log link functions for the treatment and outcome variable. The recursive bivariate model with Gaussian copula is denoted as Gaussian C and the flexible recursive bivariate model in which the copula is selected by AIC is refereed to as Flexible C.

Although these conclusions are valid for the simulation settings considered here, it cannot be determined *a priori* whether relaxing the distributional assumptions will lead to dramatically different estimated ATE as the true structure in the data is unknown. However, these results do suggest that there are a variety of scenarios in which incorrect distributional assumptions lead to biased results and in which a flexible recursive bivariate model can substantially mitigate this bias.

In order to illustrate the utility of the flexible recursive bivariate model in practice, we now turn to an example from the literature on international institutions: the debate over the impact of ratification of Article VIII of the International Monetary Fund's Articles of Agreement on compliance.

5. THE IMPACT OF INTERNATIONAL INSTITUTIONS

To highlight the utility of these models in political science, we have reexamined the debate over Article VIII ratification and compliance that was first explored by Simmons (2000). This is an ideal example to examine for two reasons: its structure—endogenous ratification as a determinant of compliance—is quite typical of studies of international institutions (as well as for regimes, international organizations, and state behavior more generally), and in the course of the debate the original study's conclusions have been reexamined using both simultaneous likelihood methods (von Stein, 2005) and matching methods (Simmons and Hopkins, 2005).

At issue is the impact of treaty commitment on compliance, using the example of Article VIII of the International Monetary Fund's Articles of Agreement. Article VIII stipulates that signatories must keep their current accounts free from restrictions. Individual governments may be tempted to restrict current accounts to realize short-term gains, such as developmental objectives or the easing of balance-of-payment difficulties, but doing so is inimical to the longer-term collective

goal of open foreign exchange. Simmons chose to explore Article VIII because its effects are uncontaminated by external considerations: commitment to Article VIII is voluntary, and there are neither positive incentives to commit nor sanctions for noncompliance (Simmons, 2000, 820). The interesting question for students of institutions is whether, and to what extent, formal commitment to Article VIII increases compliance by reducing the probability that a state will impose current-accounts restrictions.

Clearly, Article VIII ratification cannot be viewed as a random "treatment." Simmons (2000) recognizes the endogeneity of Article VIII status but, citing the absence of a good instrument (fn. 25), utilizes a logit model on cross-sectional time-series data. In doing so, she leverages the insight of Beck, Katz and Tucker (1998), who pointed out that annual cross-series time-section data are equivalent to grouped duration data. Following their recommendations, she accounts for temporal dependency using two time splines of a measure of the time elapsed since the state's last currency restriction. She finds that Article VIII commitments increase the probability of compliance by up to 27% in the first year after commitment, though the effect subsequently diminishes and fades to insignificance after about five years.

In a followup study, von Stein (2005) argues that screening effects, rather than constraining effects, could easily account for Simmons' results—that is, that states preferentially opt in to treaties with which they are already willing to comply. She also underscores the endogeneity of Article VIII status and argues that unmeasured confounders—that is, variables that have an impact both on commitment and compliance—are potentially problematic for Simmons' results. Von Stein points to Vreeland (2003, 5-8), who lists two examples of unobservables that could confound the impact of IMF programs on outcomes: "political will," or the resolve of countries that are determined both to make a commitment in the first place and, subsequently, to uphold it; and trust in government, which

provides the societal support necessary both to permit a government to make an IMF commitment and to weather its possible adverse repercussions.

To estimate the impact of Article VIII commitment net of these unobservables, von Stein derives an original dual-selection model based on a standard bivariate probit in which the probability of restriction is modeled separately for signatories and non-signatories—that is, states are "selected" into either signatory or nonsignatory status. Like Simmons, von Stein uses time splines to account for temporal dependence (fn. 12). In addition, she utilizes two dummy variables in the selection equation that are equal to 1 if the state in question restricted current accounts in the present or the previous year, respectively, and equal to 0 otherwise, to ensure that the coefficient estimates are "based on the variables' effects before and when states sign, but not after" (618).

Based on her results, von Stein concludes that Simmons' estimate of the impact of treaty commitment on compliance is overly optimistic—roughly double what it should be. Moreover, this reduced effect can no longer be distinguished from zero at standard levels of statistical significance. Von Stein also finds a statistically significant correlation between the error terms of the selection and outcome equations for signatories, implying that the unmeasured confounders that produce commitment also produce compliance. Intriguingly, the correlation of the error terms of the equations for non-signatories is not statistically significant, implying that the converse is not true.

Simmons and Hopkins (2005) reply by pointing to the well-known frailty of "Heckman-style" models in the face of violations of their distributional assumptions (624-627). Moreover, they point out that the dummy variables used by von Stein to restrict the selection equation to the pre-commitment period are problematic, both in that they induce quasi-perfect separation and in that they account for almost all of the difference between von Stein's estimate of the effect of treaty

commitment and Simmons' original estimate (626).

Simmons and Hopkins' solution to the larger selection-on-unobservables issue involves theorizing and measuring the unobserved confounders—specifically, using capital account openness and GATT/WTO membership as proxies for unobserved political will—and then using matching to calculate the ATE of commitment to Article VIII. While their matching-based estimate of the ATE corresponds well with the estimate from Simmons' original paper, they admit in the final paragraph that,

> [t]o be sure, von Stein's critique is about nonrandom assignment to treatment owing to both observable and unobservable selection factors, and matching assumes that there is no selection on unobserved covariates. Certainly, though, matching can play a role in narrowing the range of possible unobservables, just as we demonstrated earlier. (630)

While this statement is true, it does raise important questions. How can we be reasonably certain that all unmeasured confounders have been accounted for? If unobserved confounders remain, to what extent does their omission bias the matching-based estimate of the ATE? And how can we obtain separate estimates of the screening and constraining effects of Article VIII commitment, rather than the impact of commitment net of both?

Rather than attempting to address these questions within the context of matching, we instead address Simmons and Hopkins' concerns about the impact of rigid functional-form assumptions in simultaneous likelihood models by relaxing those assumptions.

5.1. *Analysis and Results*

To explore the screening and constraining impacts of Article VIII, we utilized a flexible recursive bivariate binary model on the commitment and compliance variables from Simmons and Hopkins' reanalysis. We use the Ali-Mikhail-Haq (AMH) copula with two logit marginals: this model produced the lowest AIC score of any marginal-copula combination,[6] and a Clarke test (Clarke, 2007) indicated that it was to be preferred over the closest contender, a model with logit marginals and a Gaussian copula.[7] We account for temporal dependence with smoothing splines to capture the non-linear relationships between both years of IMF membership and commitment and years since last restriction and compliance. The exclusion restriction is clearly satisfied: universality and regional norms, in particular, are unlikely to produce currency restriction except via Article VIII commitment. Utilizing alternative marginal distributions produced no noticeable improvement in fit and empirical findings.

The results of the analysis, displayed in Table 2, comport very well with theoretical expectations and findings from Simmons' original study. Of more than 20 coefficients, only one—*Change in GDP*, an economic control in the Commitment equation—both changes sign and becomes statistically significant. Two variables in the Restriction equation, *Change in GDP* and *Reserves/GDP*, retain their predicted sign but become statistically insignificant, while two others—*Reserve Volatility* in the Commitment equation and *Openness* in the Restriction equation—retain their predicted sign but become statistically significant. Leaving aside the impact of Article VIII commitment on restriction for the moment, then, the remaining theoretical conclusions regarding the determinants

[6] We explored all nine possible pairs of logit, probit, and cloglog marginals with 17 different copulas each: rotated and unrotated Clayton, Joe, and Gumbel copulas (a total of 12) as well as Frank, AMH, FGM, Student's t, and Gaussian.

[7] The AMH copula, like the Gaussian, can capture both positive and negative associations among error terms—a feature which, as we describe below, is important in the context of this model.

Variable	Article VIII Commitment β	s.e.	Restriction γ	s.e.
Article VIII Commitment			-1.109	0.227^{***}
Terms of Trade Volatility			0.461	0.142^{***}
Balance of Payments/GDP			-0.018	0.008^{**}
Use of Fund Credits	-0.372	0.157^{**}	1.059	0.180^{***}
Openness	0.025	0.002^{***}	-0.006	0.003^{**}
Change in GDP	-0.027	0.012^{**}	-0.012	0.017
Reserves/GDP	-0.909	0.964	0.589	0.928
Democracy	0.016	0.009^{*}	0.017	0.014
GATT/WTO Member	-0.698	0.164^{***}	-0.149	0.208
Universality	0.025	0.054		
Regional Norm	0.065	0.004^{***}		
Flexible Exchange Rate	-0.247	0.180		
Surveillance	-0.620	0.265^{**}		
GNP/Capita	0.000	0.000^{***}		
Reserve Volatility	-1.168	0.146^{***}		
Year	-0.028	0.041		
s(Years of IMF Membership)†	3.07	14.1^{***}		
Years Since Last Restriction			‡	
Intercept	-6.625	1.736^{***}	0.874	0.509^{*}
	θ	s.e.		
s(Years Since Commitment)†	1.000	4.299^{**}		
Intercept	0.136	0.357		

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

† Reported spline values are effective degrees of freedom and Chi-square, respectively.

‡ Ten coefficients (for dummy variables for one year, two years, …, 10 years) omitted to save space. Coefficients ranged from -3.964 to -5.436; all were significant at $p < 0.01$.

n = 2,288, $average\ \theta$ = -0.024 (-0.457, 0.635), total edf = 39.1.

Table 2: Determinants of Article VIII ratification and compliance.

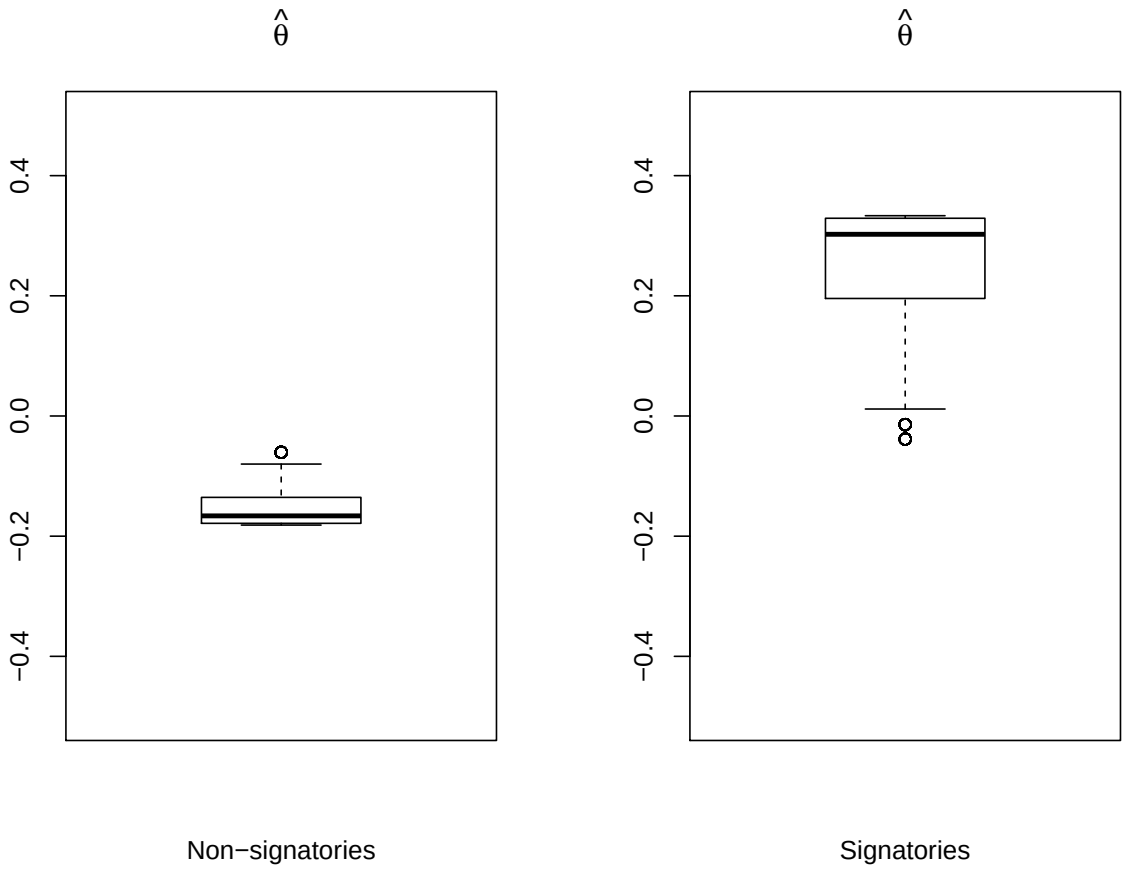

Figure 2: Estimated values of θ, the parameter that reflects the strength of the correlation between the error terms of the commitment and restriction equations, for non-signatories and signatories.

of commitment and restriction are largely similar to those in the original study.

The third equation, at bottom, models θ, the parameter that captures the strength of the correlation of the errors between the two equations, as a function of years since (or, if negative, prior to) commitment to Article VIII. The goal is to capture either transient shocks around the time of signing—the temporary rise of an unusually sympathetic government, say, or the presence of propitious short-term conditions—or longer-term internalization of (or disenchantment with) the norms of the treaty. The results show a weakly positive correlation, and a plot of the spline function (not shown) shows a linear relationship.

As Figure 2 demonstrates, the impact of unobserved confounders varies markedly between non-signatories and signatories. As we might expect, unobserved "shocks" were negatively correlated among non-signatories, indicating that unobserved confounders that decreased (increased) the propensity to commit also increased (decreased) the probability of currency restriction. The asymmetric tail of the AMH copula indicates that this association is especially strong. Among signatories, we see a positive correlation among shocks: unobserved confounders that increased the propensity to commit increased the probability of currency restriction. This result suggests that some unmodeled variable (changes in government, perhaps) decreases post-commitment enthusiasm for compliance. The fact that unobserved confounders pull in opposite directions across the two groups of states is one of the important conclusions that would have been missed by a less flexible standard recursive bivariate probit model. As we will soon see, it turns out to be an important one.

Using the chosen model, we calculate the ATE of Article VIII commitment, given by

$$\frac{1}{n}\sum_{i=1}^{n}\Lambda(\hat{\beta}_1 + \hat{f}_1\left(\{\mathbf{X}\}_{[-x_1],i}\right)) - \Lambda(\hat{f}_1\left(\{\mathbf{X}\}_{[-x_1],i}\right)),$$

where Λ is the cumulative distribution function of a standard logistic distribution. This formula captures the difference between the probability of restriction given Article VIII commitment (the first term) and the probability of restriction given no Article VIII commitment (the second term) for each observation i, summed across all n observations. We generate a 95% confidence interval for the ATE via posterior simulation.

The screening hypothesis, in which the states that sign are states for which compliance involves very little change in behavior, implies that there is a systematic difference in the average treatment

effect between signatories and non-signatories. To the extent that it holds, we should expect to see that the ATE among non-signatories is systematically greater than it is among signatories, indicating that they would have to change their behavior more radically than would signatories in order to comply. The compliance hypothesis, by contrast, suggests that the average treatment effect for signatories should be positive and statistically significant.

The ATE for signatories is found to be -0.059. That is, the average effect of signing Article VIII is to reduce the risk of current account restriction (that is, increase the probability of compliance) by 5.9%. These results suggest that signing on to Article VIII has a significant, if modest, impact on state behavior. For non-signatories, the model produces an ATE of -0.105—nearly twice that of signatories.[8] This difference, illustrated on the left of Figure 3, provides strong evidence of a screening effect. The states that sign are fundamentally different from the states that do not: in the counterfactual world in which these non-signatories were to sign, doing so would require a much more substantial change in behavior. In short, these results support von Stein's contention that Article VIII screens states that have a higher cost of compliance.

To explore the impact of commitment on compliance over time, we disaggregate the ATE by year since signing (Figure 3, right-hand graph). As one might expect, the effect of the treaty is not constant over time. When unobservables are properly accounted for in the bivariate model, the effect of Article VIII commitment shrinks to about a 10% increase in the probability of restriction immediately after signing.

What is much more striking is the fact that the ATE remains relatively constant, between 5% and 10%, for an entire decade after signing. While the results of Simmons' original model (2000,

[8]The difference between the two distributions, as gauged by a Kolmogorov-Smirnov test, is highly statistically significant (D = 0.7909, $p < 0.00$).

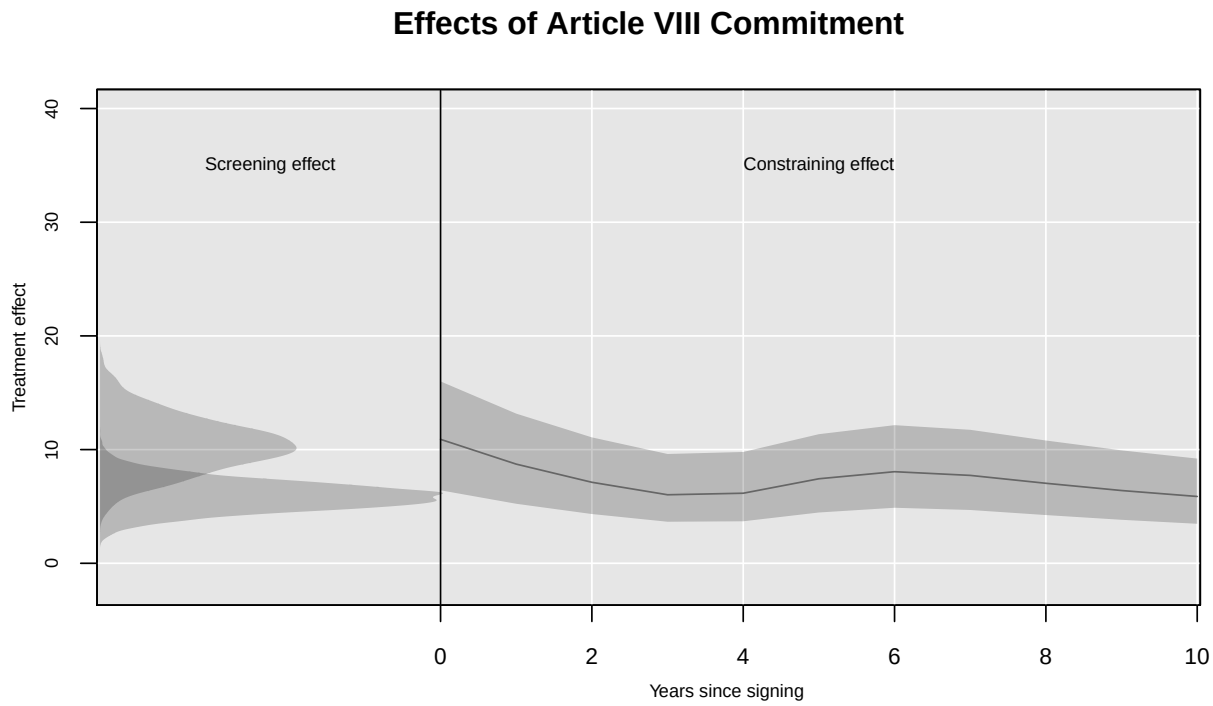

Figure 3: Average treatment effect of commitment to Article VIII, for signatories and non-signatories. The Y-axis represents the ATE measured in terms of percentage change in the probability of compliance. For the screening effect, the shaded areas represent the distributions of treatment effects for signatories (bottom) and non-signatories (top), while for the constraining effect the shaded area represents 95% confidence intervals.

831) suggested that the impact of signing drops below 5% after three years and fades to substantive and statistical insignificance after five years, our results indicate a much more robust and lasting effect—welcome news for students of international institutions.

The impact of unobserved confounders is worth a brief note as well. By comparing the ATE estimates from our final model with those of a *univariate* model—i.e., a model that assumes zero correlation between error terms—, we can see that unobserved confounders have little impact on the ATE for signatories: the increase in the probability of compliance associated with commitment was 5.9% in the full model and 5.7% in the univariate model. Unobserved confounders had a much

more substantial impact on the ATE for non-signatories, however: the ATE for non-signatories was 10.5% in the full model but only 5.7% in the univariate model. Given that Simmons and Hopkins went to a substantial amount of effort to include as many causes of compliance as possible but paid less theoretical attention to variables that might produce a screening effect, while von Stein relied on the error terms in her model to capture screening effects, this result makes quite a bit of sense.

In all, the results suggest that both Simmons and von Stein were correct. Simmons was correct about the main finding: commitment to Article VIII does increase compliance, and its impact is far more lasting than previously thought. Von Stein was correct about the magnitude of the average treatment effect (if not its statistical significance or duration) and about the existence both of significant screening effects and of important unobserved confounders.

5.2. *Discussion*

The malleability of the simultaneous flexible likelihood model and its ability to capture the impact of omitted confounders was essential in many ways to this analysis. As our simulations demonstrate, the ability to try out a wide range of functional form combinations and adjudicate among them based on fit significantly improves the accuracy and precision of our estimate of the ATE—an estimate that differs significantly from those of past studies both in magnitude and duration of impact. The ability to model the association between "shocks" to commitment and compliance, and the ability to allow that association to vary in magnitude and direction depending on signatory status, turned out to be essential for obtaining good estimates of the magnitude of screening effects in particular. Finally, although our estimates of constraining effects were similar to those that would have been obtained from a simple recursive bivariate logit model with no correlation among errors,

there is no way that we could have known that without estimating both models and comparing their results.

6. CONCLUSION

For many years, political scientists seeking to engage in causal inference have been forced to choose which unpalatable assumptions they wished to embrace in the face of endogeneity issues: they could utilize a potentially weak or invalid instrument, assume selection on observables, or embrace simultaneous likelihood methods with restrictive functional form assumptions. Nearly all statistical empirical work in political science suffers from these problems. Flexible simultaneous likelihood models represent a considerable improvement over this *status quo*. These models are capable of capturing the impact of unmeasured confounders, and their flexible functional form assumptions significantly reduce bias in the estimate of average treatment effects. As our results demonstrate, the increased flexibility of these models can greatly enhance our understanding of political science.

REFERENCES

Angrist, Joshua and Alan B. Krueger. 2001. Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. Technical report National Bureau of Economic Research.

Bagozzi, B. E., D. W. Hill, W. H. Moore and B. Mukherjee. 2015. "Modeling Two Types of Peace: The Zero-Inflated Ordered Probit (ZiOP) Model in Conflict Research." *Journal of Conflict Res-*

olution 59(4):728–752.

Bartels, Larry M. 2000. "Partisanship and Voting Behavior, 1952-1996." *American Journal of Political Science* 44(1):35–50.

Bausch, Andrew W. 2015. "Democracy, War Effort, and the Systemic Democratic Peace." *Journal of Peace Research* 52(4):435–447.

Bearce, David H., Kristen M. Flanagan and Katharine M. Floros. 2006. "Alliances, Internal Information, and Military Conflict Among Member-States." *International Organization* 60(3):595–625.

Beck, Nathaniel, Jonathan N. Katz and Richard Tucker. 1998. "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable." *American journal of Political Science* 42(4):1260–1288.

Benson, Brett V. 2011. "Unpacking Alliances: Deterrent and Compellent Alliances and Their Relationship with Conflict, 1816-2000." *Journal of Politics* 73(4):1111–1127.

Blomberg, S. Brock, Gregory D. Hess and Akila Weerapana. 2004. "Economic Conditions and Terrorism." *European Journal of Political Economy* 20(2):463–478.

Boehmke, Frederick J., Daniel S. Morey and Megan Shannon. 2006. "Selection Bias and Continuous-Time Duration Models: Consequences and a Proposed Solution." *American Journal of Political Science* 50(1):192–207.

Braumoeller, B. F. and A. Carson. 2011. "Political Irrelevance, Democracy, and the Limits of Militarized Conflict." *Journal of Conflict Resolution* 55(2):292–320.

Bueno de Mesquita, Bruce, James D. Morrow, Randolph M. Siverson and Alastair Smith. 1999. "An Institutional Explanation of the Democratic Peace." *American Political Science Review* 93(4):791 – 807.

Carrubba, C. J., A. Yuen and C. Zorn. 2007. "In Defense of Comparative Statics: Specifying Empirical Tests of Models of Strategic Interaction." *Political Analysis* 15(4):465–482.

Chib, S. and E. Greenberg. 2007. "Semiparametric modeling and estimation of instrumental variable models." *Journal of Computational and Graphical Statistics* 16:86–114.

Chiba, Daina, Lanny Martin and Randy Stevenson. 2014. "A Copula Approach to the Problem of Selection Bias in Models of Government Survival.".

Chiba, Daina, Nils W. Metternich and Michael D. Ward. 2015. "Every Story Has a Beginning, Middle, and an End (But Not Always in That Order): Predicting Duration Dynamics in a Unified Framework." *Political Science Research and Methods* 3(3):515–541.

Clarke, Kevin A. 2007. "A Simple Distribution-Free Test for Nonnested Hypotheses." *Political Analysis* 15(3).

Cranmer, Skyler J., Bruce A. Desmarais and Elizabeth J. Menninga. 2012. "Complex Dependencies in the Alliance Network." *Conflict Management and Peace Science* 29(3):279–313.

Dorussen, H. and H. Ward. 2008. "Intergovernmental Organizations and the Kantian Peace: A Network Perspective." *Journal of Conflict Resolution* 52(2):189–212.

Doyle, Michael W. and Nicholas Sambanis. 2000. "International Peacebuilding: A Theoretical and Quantitative Analysis." *The American Political Science Review* 94(4):779.

Fang, Songying, Jesse C. Johnson and Brett Ashley Leeds. 2014. "To Concede or to Resist? The Restraining Effect of Military Alliances." *International Organization* 68(4):775–809.

Freedman, D. A. and J. S. Sekhon. 2010. "Endogeneity in Probit Response Models." *Political Analysis* 18(2):138–150.

Freedman, David, David Collier, Jasjeet Singh Sekhon and Philip B. Stark. 2010. *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. Cambridge ; New York: Cambridge University Press.

Freytag, Andreas, Jens J. Krüger, Daniel Meierrieks and Friedrich Schneider. 2011. "The Origins of Terrorism: Cross-Country Estimates of Socio-Economic Determinants of Terrorism." *European Journal of Political Economy* 27(Supplement 1):S5–S16.

Fukumoto, Kentaro. 2015. "What Happens Depends on When It Happens: Copula-Based Ordered Event History Analysis of Civil War Duration and Outcome." *Journal of the American Statistical Association* 110(509):83–92.

Gartner, Scott Sigmund. 2011. "Signs of Trouble: Regional Organization Mediation and Civil War Agreement Durability." *The Journal of Politics* 73(02):380–390.

Gates, Scott, Torbjørn Knutsen and Jonathon W. Moses. 1996. "Democracy and Peace: A More Skeptical View." *Journal of Peace Research* 33(1):1–10.

Gerber, Alan S. and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94(3):653–663.

Gerber, Alan S. and Donald P. Green. 2005. "Correction to Gerber and Green (2000), Replication of Disputed Findings, and Reply to Imai (2005)." *American Political Science Review* 99(2):301–313.

Gilligan, Michael J. and Ernest J. Sergenti. 2008. "Do UN Interventions Cause Peace? Using Matching to Improve Causal Inference." *Quarterly Journal of Political Science* 3(2):89–122.

Grier, Kevin B. and Gordon Tullock. 1989. "An Empirical Analysis of Cross-National Economic Growth, 1951-80." *Journal of Monetary Economics* 24(2):259–276.

Hafner-Burton, Emilie M. and Miles Kahler. 2009. "Network Analysis for International Relations." *International Organization* 63:559–92.

Hill, Daniel W. 2010. "Estimating the Effects of Human Rights Treaties on State Behavior." *The Journal of Politics* 72(04):1161–1174.

Imai, Kosuke. 2005. "Do Get-out-the-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." *American Political Science Review* 99(2):283–300.

Johnson, Jesse C. and Brett Ashley Leeds. 2011. "Defense Pacts: A Prescription for Peace?" *Foreign Policy Analysis* 7(1):45–65.

Keele, Luke. 2015. "The Statistics of Causal Inference: A View from Political Methodology." *Political Analysis* 23(3):313–335.

King, Gary and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political analysis* 9(2):137–163.

Kish, Leslie. 1959. "Some Statistical Problems in Research Design." *American Sociological Review* 24(3):328–338.

Krieger, Tim and Daniel Meierrieks. 2011. "What Causes Terrorism?" *Public Choice* 147(1):3–27.

Leeds, Brett Ashley. 2003. "Do Alliances Deter Aggression? The Influence of Military Alliances on the Initiation of Militarized Interstate Disputes." *American Journal of Political Science* 47(3):427–439.

Levy, Jack S. 1981. "Alliance Formation and War Behavior: An Analysis of the Great Powers, 1495-1975." *Journal of Conflict Resolution* 25(4):581–613.

Lewis, J. B. 2003. "Revealing Preferences: Empirical Estimation of a Crisis Bargaining Game with Incomplete Information." *Political Analysis* 11(4):345–367.

Little, R. J. A. 1985. "A note about models for selectivity bias." *Econometrica* 53:1469–1474.

Lupu, Yonatan. 2013. "The Informative Power of Treaty Commitment: Using the Spatial Model to Address Selection Effects." *American Journal of Political Science* pp. 912–925.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *The American Economic Review* pp. 319–323.

Maoz, Zeev. 2009. "The Effects of Strategic and Economic Interdependence on International Conflict across Levels of Analysis." *American Journal of Political Science* 53(1):223–240.

Maoz, Zeev and Bruce Russett. 1993. "Normative and Structural Causes of Democratic Peace, 1946-1986." *American Political Science Review* 87(3):624–638.

Marra, G. and R. Radice. 2011. "Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity." *Canadian Journal of Statistics* 39:259–279.

Marra, G. and R. Radice. 2015. "Flexible Bivariate Binary Models for Estimating the Efficacy of Phototherapy for Newborns with Jaundice." *International Journal of Statistics and Probability* 4:46–58.

Marra, G. and S.N. Wood. 2012. "Coverage Properties of Confidence Intervals for Generalized Additive Model Components." *Scandinavian Journal of Statistics* 39:53–74.

Marra, Giampiero and Rosalba Radice. 2016. *SemiParBIVProbit: Semiparametric Copula Bivariate Regression Models*. R package version 3.7-1.
URL: *http://CRAN.R-project.org/package=SemiParBIVProbit*

Marra, Giampiero and Rosalba Radice. 2017. "Bivariate Copula Additive Models for Location, Scale and Shape." *Computational Statistics and Data Analysis* .

McLaughlin Mitchell, Sara and Paul R. Hensel. 2007. "International Institutions and Compliance with Agreements." *American Journal of Political Science* 51(4):721–737.

Meierrieks, Daniel and Thomas Gries. 2012. "Causality Between Terrorism and Economic Growth." *Journal of Peace Research* 50(1):91–104.

Monfardini, C. and R. Radice. 2008. "Testing Exogeneity in the Bivariate Probit Model: A Monte Carlo Study." *Oxford Bulletin of Economics and Statistics* 70:271–282.

Morgan, Stephen L. and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Meth-*

ods and Principles for Social Research. Analytical methods for social research New York: Cambridge University Press.

Murray, Michael P. 2006. "Avoiding Invalid Instruments and Coping with Weak Instruments." *Journal of Economic Perspectives* 20(4):111–132.

Radice, Rosalba, Giampiero Marra and Malgorzata Wojtys. 2015. "Copula Regression Spline Models for Binary Outcomes." *Statistics and Computing,* DOI 10.1007/s11222-015-9581-6.

Reuveny, Rafael and Quan Li. 2003. "The Joint Democracy-Dyadic Conflict Nexus: A Simultaneous Equations Model." *International Studies Quarterly* 47(3):325–346.

Sartori, Anne E. 2003. "An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions." *Political Analysis* 11(2):111–138.

Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12(1):487–508.

Signorino, Curtis S. 1999. "Strategic Interaction and the Statistical Analysis of International Conflict." *American Political Science Review* 93(2):279–297.

Simmons, Beth A. 2000. "International Law and State Behavior: Commitment and Compliance in International Monetary Affairs." *The American Political Science Review* 94(4):819.

Simmons, Beth A. 2009. *Mobilizing for Human Rights: International Law in Domestic Politics*. Cambridge: Cambridge University Press.

Simmons, Beth A. and Daniel J. Hopkins. 2005. "The Constraining Power of International Treaties: Theory and Methods." *American Political Science Review* 99(04).

Smith, Alastair. 1999. "Testing Theories of Strategic Choice: The Example of Crisis Escalation." *American Journal of Political Science* 43(4):1254–1283.

Theisen, Ole Magnus. 2012. "Climate Clashes? Weather Variability, Land Pressure, and Organized Violence in Kenya, 1989–2004." *Journal of Peace Research* 49(1):81–96.

Tir, Jaroslav and Douglas M. Stinnett. 2012. "Weathering Climate Change: Can Institutions Mitigate International Water Conflict?" *Journal of Peace Research* 49(1):211–225.

von Stein, Jana. 2005. "Do Treaties Constrain Or Screen? Selection Bias And Treaty Compliance." *American Political Science Review* 99(04):611–622.

Vreeland, James Raymond. 2003. *The IMF and Economic Development*. New York: Cambridge University Press.

Winkelmann, R. 2012. "Copula bivariate probit models: with an application to medical expenditures." *Health Economics* 21:1444–1455.

Winship, Christopher and Robert D. Mare. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 18:327–350.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

Xiang, Jun. 2010. "Relevance as a Latent Variable in Dyadic Analysis of Conflict." *The Journal of Politics* 72(02):484.

APPENDIX

1.1. *Simulation details*

The R code chunk below was used to generate the scenario with probit link for y and normal distribution for z_2, **with coefficient of z_2 set to -0.85**:

```
## set seed and sample size
set.seed(0)
n=1000

## observed confounders
z1 <- runif(n)
z4 <- runif(n)

# unobserved confounder
z2 <- rnorm(n)

## instrumental variable
z3 <- rbinom(n,1,0.5)

## non-linearities
f1   <- function(x) cos(pi*2*x) + sin(pi*x)
f2   <- function(x) x + exp(-30*(x-0.5)^2)

## treatment assignment
beta <- -0.85
prob.treated <-plogis(-0.5 + f1(z1) beta*z2 + 3*z3 + 1.3*z4)
x <- rbinom(n, 1, prob.treated)

## potential outcomes
p0 <- plogis(-3.5 + f2(z1) + 2*z2 -0.8*z4)
p1 <- plogis( 0.5 + f2(z1) + 2*z2 -0.8*z4)

y0 <- rbinom(n, 1, p0)
y1 <- rbinom(n, 1, p1)

## observed outcomes
y <- y0
y[x==1] <- y1[x==1]
```

To allow z_2 to be Student's t with four degrees of freedom, χ^2 with one degree of freedom and uniform[-3, 3], and the parameter of the unobserved confounder to have different impacts in the treatment equation, the above R code can be easily modified by replacing $\mathrm{z2 <- rnorm(n)}$ with $\mathrm{z2 <- rt(n, df=2)}$, $\mathrm{z2 <- rchisq(n, df=1)}$ or $\mathrm{z2 <- runif(n, -3, 3)}$ and replacing $\mathrm{beta <- -0.85}$ with $\mathrm{beta <- -1.5}$, $\mathrm{beta <- 0}$, $\mathrm{beta <- 0.85}$, or $\mathrm{beta <- 1.5}$.

1.2. *Further simulation results*

Results for additional simulation settings (where the value of coefficient of the unobserved confounder in the treatment equation has been set to 0 and -1.5) are reported in Tables 1 and 2. When the coefficient is set to 0 (i.e. no unobserved confounding problem), then, as expected, the univariate model and matching perform the best with the former being more efficient than the latter. Although the copula models (Gaussian and Frank copulae) show a poorer performance, the magnitudes of their bias and RMSE are comparable. When the problem of unobserved confounding becomes more severe (the coefficient is set to -1.5), then copula models still perform predictably well whereas the univariate and matching approaches deteriorate as compared to the case where the confounding issue is less severe. Though we do not report the results here, we observed similar patterns when setting the value of the coefficient of z_2 to 0.85 and 1.5.

	Link	Logit-Logit				Probit-Probit				Cloglog-Cloglog			
	Distribution	$\mathcal{N}(0,1)$	t_4	χ^2_1	$\mathcal{U}(-3,3)$	$\mathcal{N}(0,1)$	t_4	χ^2_1	$\mathcal{U}(-3,3)$	$\mathcal{N}(0,1)$	t_4	χ^2_1	$\mathcal{U}(-3,3)$
% Bias													
Model	Univariate	0	0	0	2	0	0	0	1	0	1	1	2
	Matching	0	0	0	4	0	0	1	4	0	0	0	4
	Flexible G C	2	3	0	9	1	3	1	8	3	4	3	10
	Flexible F C	2	3	1	9	1	3	1	8	2	4	4	10
RMSE													
Model	Univariate	0.03	0.03	0.04	0.04	0.03	0.03	0.04	0.04	0.03	0.03	0.04	0.04
	Matching	0.05	0.06	0.08	0.08	0.05	0.06	0.08	0.08	0.05	0.06	0.08	0.08
	Flexible G C	0.07	0.08	0.09	0.09	0.07	0.08	0.09	0.09	0.08	0.09	0.09	0.09
	Flexible F C	0.07	0.08	0.10	0.09	0.07	0.08	0.10	0.09	0.08	0.09	0.10	0.10

Table 1: % Percentage biases and RMSEs for the average treatment effect of an endogenous right-hand-side variable, estimated on data simulated using logit links for the treatment and outcome variable and normal, Student's t, χ^2 and uniform distributions for the unobserved confounder, when setting the coefficient of the unobserved confounder to 0 in the treatment equation. The 250 simulated datasets were used to fit Gaussian and Frank copulae, detonated as Flexible G C and Flexible F C, respectively, with logit, probit and complementary-log-log link functions for the treatment and outcome variable.

	Link	Logit-Logit				Probit-Probit				Cloglog-Cloglog			
	Distribution	$\mathcal{N}(0,1)$	t_4	χ^2_1	$\mathcal{U}(-3,3)$	$\mathcal{N}(0,1)$	t_4	χ^2_1	$\mathcal{U}(-3,3)$	$\mathcal{N}(0,1)$	t_4	χ^2_1	$\mathcal{U}(-3,3)$
% Bias													
Model	Univariate	37	58	45	135	37	58	41	135	37	58	42	135
	Matching	50	92	88	157	51	91	89	157	50	90	89	156
	Flexible G C	2	2	8	8	2	3	8	8	3	3	9	9
	Flexible F C	1	5	9	3	1	6	10	2	1	6	11	3
% RMSE													
Model	Univariate	0.19	0.27	0.30	0.44	0.19	0.27	0.25	0.44	0.19	0.27	0.25	0.44
	Matching	0.26	0.48	0.50	0.52	0.26	0.46	0.50	0.52	0.26	0.47	0.49	0.52
	Flexible G C	0.05	0.05	0.07	0.06	0.05	0.05	0.07	0.06	0.05	0.05	0.07	0.06
	Flexible F C	0.05	0.06	0.07	0.05	0.05	0.06	0.07	0.05	0.05	0.06	0.07	0.05

Table 2: % Percentage biases and RMSEs for the average treatment effect of an endogenous right-hand-side variable, estimated on data simulated using logit links for the treatment and outcome variable and normal, Student's t, χ^2 and uniform distributions for the unobserved confounder, when setting the coefficient of the unobserved confounder to -1.5 in the treatment equation. The 250 simulated datasets were used to fit Gaussian and Frank copulae, detonated as Flexible G C and Flexible F C, respectively, with logit, probit and complementary-log-log link functions for the treatment and outcome variable.