Computational and Mathematical Modeling in the Social Sciences

SCOTT DE MARCHI
Duke University

CAMBRIDGE
UNIVERSITY PRESS

Contents

Acknowledgments

Without a larger research community, it would have been difficult for me to complete a project of this scope. I am particularly fortunate because an astonishing number of people have read versions of this manuscript and taken the trouble to try to correct the numerous blemishes and mistakes present in my work, some of which remain despite their efforts. I owe a great debt to Jennifer Harrod, Mike Munger, and Lyle Scruggs who (for different reasons) have been forced to talk to me about this project for the last several years. Bob Keohane and I taught a course on qualitative methods in the fall of 2003 at Duke University, and without our weekly conversations and his close reading of the manuscript, I would not have finished. John Aldrich and George Rabinowitz, at numerous coffee breaks and lunch meetings, have both been very influential in how I have approached the issues raised in this book. Ken Kollman, John Miller, and Scott Page are responsible for my interest in applying computational methods to social science, and they offered good advice at every stage of writing this manuscript. My editor Scott Parris at Cambridge has made the final steps of completing this manuscript much easier than I had thought possible.

Some really smart people read the manuscript and sent me comments, including Chris Bond, Jorge Bravo, John Brehm, Russ Denton, Charles Franklin, Chris Gelpi, Hein Goemans, Jeff Grynaviski, Jay Hamilton, Mel Hinich, Jerry Hough, Seth Jolly, Bill Keech, Dean Lacy, Karl Lietzan, Emerson Niou, Brendan Nyhan, Phillip Rehm, Jason Reifler, Tom Scotto, Curt Signorino, Terry Sullivan, Mike Tofias, Camber Warren, and Steven Wilkinson. There are also several groups

of people who have helped in indirect ways: the Emerging Solutions Group at PricewaterhouseCoopers, my current poker group (which is just as likely to do Monte Carlo work as not after a game), and the brave souls I have played Diplomacy with over the last two decades. Finally, I'd like to thank the Tooks, Daniel and Jennifer, for everything else.

Prelude

When Aeneas fled from burning Troy, he had some difficult decisions to make. His first priority was to rescue his country gods and relics, but he was covered in gore from combat and did not want to carry these sacred artifacts with his own hands. His solution was novel: Anchises, his father, could carry the artifacts and Aeneas would carry him upon his back. His second priority was to guard the safety of his wife Creusa and his son. With his heavy burden, he "satisficed" by holding the hand of his son and bidding his wife to follow him. Unfortunately, though he succeeded in rescuing the country gods and his son, he lost his wife during his flight from the doomed city.

Earning the appellation "pious" involved some cruel choices for Aeneas, but despite this offense to modern sensibilities (I daresay many of us would have tossed the country gods and told Anchises to walk on his own two feet), it is hard to blame him. Weary from battle, burdened with both his family and the country gods, it would be difficult to pay attention to everything of merit. It is not surprising that he did not even know when or how he had lost his wife.

Graduate school has some similarities. Granted, most students do not have to face a ravaging horde of Greek soldiers, nor are they surrounded by burning buildings. But the press of time is a constant weight, and one is forced to attend to some matters more than others. It is not a coincidence that if you ask students trained in the top research programs in the social sciences what their field is they may answer "mathematical methods" or even something more precise such as "game theory" or "econometrics." Most students spend a large fraction of their time learning these methods, and this comes at the expense

of other sorts of work such as history and case studies. Like pious Aeneas, we make choices, and even the most heroic of us are forced to ignore many worthwhile subjects.

The important thing to note is that many of the social sciences, most notably political science and economics, have made a wager. This wager involves both time and space. From graduate students to faculty, we spend our time learning and practicing mathematical methods, in particular game theory and statistical modeling. For the journals and presses, the lion's share of space is devoted to the results generated by mathematical methods. One does not find the best journals customarily publishing case studies of individual countries, firms, or political campaigns. Nor, in the case of the top journals in political science, is much advice (either prescriptive or predictive) given to real-world political actors. Based on the 2002–2003 Report of the Editor of the *American Political Science Review*, 69% of submissions were accounted for by the formal, quantitative, or formal *and* quantitative categories; 63% of accepted articles were in these categories – this during the tenure of an editor striving for diversity.

The presumption of this book will be to examine this epistemological gamble more closely and recommend a set of changes to current practice. It is not as if every scholar has embraced the increasing emphasis on mathematical methods. The last two decades have seen many critiques, most lamenting the sacrifices incurred in pursuit of mathematical rigor. If, as the critics argue, our ability to understand the world has not improved during the mathematization of the social sciences, we might want to adopt a different paradigm. Historiography (or qualitative research) is most often presented as the alternative to the abstractions of mathematical methods. It might, say the critics, be better for the discipline to turn out area-specialists who at least know the history of their cases than to engage in bad modeling that lacks any clear connection to the real world.

I have the good fortune of better than adequate training in history,[1] and I can argue with some fervor that a turn to historiography would not

[1] I took undergraduate degrees in computer science and history. Because of latent schizophrenia, I completed the coursework and thesis for a Master's degree in European history at the University of North Carolina–Chapel Hill before switching to the social sciences.

be good for the social sciences. Despite its problems, I remain devoted to mathematical modeling, and the goal of this book is to improve current practice rather than to supplant it. Area-specialization and case studies are necessary prerequisites for the inspiration and understanding implicit in all good models, but in my mind they do not of themselves constitute a coherent methodology for discovering causal relationships.[2]

Required reading for those who wish to supplant mathematical methods with qualitative research should include Peter Novick's *That Noble Dream: The "Objectivity Question" and the American Historical Profession* (1998). From the end of the 19th century to the beginning of the Cold War, history as a discipline was very similar in outlook to modern political science.[3] Novick's book lays out the history of the professionalization project in American history departments over this time period. Much like current social science disciplines, historians believed in their ability to understand causal relationships in the world and sought to give answers to pressing questions about how one prevents war between nation-states or the republican cycle of decay highlighted by political theorists such as Machiavelli.

The problem, after a century of consensus on method, was that historiography foundered upon the shoals of the objectivity question. For Novick, historians who believed in scientific objectivity never adequately answered the fundamental questions of how to tell good research from bad and neutral research from biased. Many historians, spurred on by the emergence of social history and other trends, simply did not believe that the empirical, objectivist tradition produced superior research.[4]

[2] There is an enormous literature on qualitative versus quantitative research. For an examination of some of the problems implicit in historical research from a political science perspective, good examples are Lustick (1996) and Goemans (2000).

[3] Although there was not great technical skill present in most historical research, there was a belief in empirical work and the use of history for understanding causality in human affairs. The letters of Henry Adams (at Harvard) to Herbert Baxter Adams (at Johns Hopkins), for example, demonstrate a high level of familiarity and respect for the hard sciences among practicing historians at the end of the 19th century and a belief that scientific objectivity was a worthwhile aspiration for the social sciences.

[4] For an example of an alternative approach to historiography, read Natalie Zemon Davis's *The Return of Martin Guerre* (1983). Davis's work concerns a French tale from the 16th century in which a woman discovers her husband is an imposter and takes him to court. Because the penalty was death by hanging, this was no laughing

One of the more sobering examples Novick uses to illustrate the death of objectivity in the historical profession is the case of David Abraham. The role of German industrialists in the rise of the Nazi Party was a contentious one, and Abraham, a junior faculty member at Princeton in the early 1980s, wrote a manuscript that emphasized the structural relationships in German society that precluded a more moderate outcome to the political turmoil of the Weimar state.

Unfortunately for Abraham, his abstract modeling, which was influenced by Marxist theory, did not endear him to senior researchers in the field. Despite many positive book reviews, Henry Turner at Yale University and Gerald Feldman at the University of California at Berkeley led an assault on Abraham's book. They believed that the footnotes to Abraham's monograph contained serious, willful errors. Misattributed citations, missing or incorrect quotations, and other errors were, in fact, plentiful in Abraham's work. For Turner and Feldman, these mistakes were proof of a malicious agenda that violated norms of historiography. In a book review in *Political Science Quarterly*, Turner wrote:

> Invoking the familiar primacy of economics, Abraham presents a highly reductionist version of the dissolution of the Republic and the rise of Nazism, which he explains in terms of his vastly simplified model of German society. . . . Unfortunately, Abraham's footnotes do not marshal evidence adequate to support his thesis. Informed readers will also balk at his disparagement or omission of institutions, ideologies, and personalities vital to comprehension of the German calamity. (Turner 1982, 740)

It is hard to convey how contentious this affair became. The journal *Central European History*, for example, featured an exchange between Feldman and Abraham that even included a complete list by Abraham of his errors and whether or not the corrections helped, hurt, or were neutral to his argument. The exchange appeared in press in 1985, but by then Abraham had been driven from the field. For Novick, who was Abraham's advisor, the lesson for historians was that optimism

matter. Davis had completed a screenplay on the story and found that her "appetite was whetted" for a more scholarly investigation, despite the lack of an expansive historical record on the story. Her approach to this problem is distinct from previous understandings of historiography: "Watching Gerard Depardieu [the actor] feel his way into the role of the false Martin Guerre gave me new ways to think about the accomplishments of the real imposter, Arnaud du Tilh. I felt I had my own historical laboratory, generating not proofs, but historical possibilities" (Davis 1983, viii).

about the ability to discern causality in history had been replaced by a naïve and defensive empiricism. Other than getting one's footnotes right, there was no other avenue for attacking or defending a model.

There are more modern examples of the continuing crisis in historiography. Michael Bellesiles's book *Arming America: The Origins of a National Gun Culture*, which presented the argument that gun culture in early American society was not as widespread as believed, won the Bancroft Prize when it was released in 2000. Much like Abraham, Bellesiles riled opponents of a different political stripe, and upon scrutiny, it was discovered that much of the data underlying the book's quantitative analysis was either misused (in the case of probate data) or entirely missing from the archives. Despite these glaring problems, the question remained about whether or not his core argument was valid. Ultimately, like Abraham, Bellesiles was forced from the discipline, resigning his post at Emory University under pressure from the trustees at the end of 2002. The Bancroft Prize for his book was rescinded shortly thereafter.[5]

Although I do not believe that Novick has much of a remedy for historiography, I do accept his diagnosis of the problem. If a particular methodological paradigm is to survive, a large majority of practicing scholars has to believe that the costs involved in training and research are merited. Simply put, the output of a methodology has to be superior results, at least compared to existing alternatives. The question economists and political scientists should ask is whether or not Novick's history of the erosion of the belief in objectivity among historians holds any lessons for us.

Despite the enormous successes made possible by the mathematical approach – the Arrow, McKelvey, and Schofield work on social choice is an excellent example – many critics, rightfully, want to know what the last decade has produced. The argument that I will present in this book is that the practice of mathematical modeling is due for a revision. In particular, existing methods are brittle when confronted with complex problems, and there is a genuine lack of correspondence

[5] A special issue of the *William and Mary Quarterly* (2002) featured essays by Bellesiles and several other historians that examine the controversy and its implications for historiography.

between deductive models, on the one hand, and empirical tests of these models, on the other.

There are additional problems unique to each of the two major subfields within mathematical methodology. Game theory, for example, has a troubling answer to the question "Is game theory meant to predict what people do, to give them advice, or what?" As Camerer (2003, 5) notes, many game theorists believe that "game theory is none of the above – it is simply 'analytical,' a body of answers to mathematical questions about what players with various degrees of rationality will do. If people don't play the way the theory says, their behavior has not proved the mathematics wrong, any more than finding that cashiers sometimes give the wrong change disproves arithmetic." Although there are examples of formal modelers tackling real-world problems, such as the interesting work of Groseclose, Milyo, and Primo on topics that include the dollar value of a House of Representatives seat, campaign finance, and empirical measures of media bias, many game theorists do not believe that their work needs an empirical referent.[6]

Statistical methodology in the social sciences has its own set of problems that mirrors the opening passage in Dickens's *A Tale of Two Cities*. We have increasingly sophisticated forays into Bayesian and nonparametric techniques. At the same time, replication continues to be problematic, especially as the complexity of statistical methods increases. Recently, the laudable goal of linking formal theory with statistical models has received renewed attention in the research of Signorino and others. Yet, most published research continues to ignore the most basic tenet of statistical work, which requires out-of-sample testing to validate a model.[7] Never before has training in statistical modeling been so widespread in graduate departments around the nation. So, too, has suspicion deepened, as many researchers have adopted Achen's (2003) admonition that a model with more than three independent variables is immediate cause for concern.

[6] On the value of a seat, see Groseclose and Milyo (2004a); on campaign finance, see Primo and Milyo (2004); and, on media bias, see Groseclose and Milyo (2004b). Behavioral game theory also tries to put game theory on a more empirical footing – Camerer's book provides a nice introduction to the field. One also might visit Roth's Web site at http://www.economics.harvard.edu/~aroth/alroth.html.

[7] For an excellent statement on statistical modeling that also happens to make this point on the neglect of out-of-sample work, see Good and Hardin (2003).

While some might question whether or not mathematical methodology is in need of revision, it is the case that a sense of unease permeates the social sciences. Those who do not practice these methods are deeply suspicious of the validity of results generated from mathematical models. And those that do practice one field of mathematical methodology are often just as suspicious about the other fields. I will argue that at least some of this suspicion is warranted, and the goal of this book is to provide a set of tools designed to increase transparency and improve modeling. Part of this enterprise involves a constructive critique of existing practice. Despite the widespread belief that the problems that beset mathematical methods are idiosyncratic to each subfield, I will demonstrate that there are a set of underlying problems that span subfields (including analytic, empirical, and qualitative).

Of the problems detailed in this book, the most severe is the curse of dimensionality. In the nonparametric statistics and artificial intelligence literatures, the "curse of dimensionality" is incredibly important, but it is not well known in the social sciences. In brief, the curse states that for any interesting problem, one should count the size of the parameter space needed to model the problem, paying special attention to how large this space becomes as the problem increases in size. If the parameter space implied by a naïve encoding of the problem is huge, one must resort to domain-specific information and a good dose of cleverness to surmount the curse of dimensionality. A brief example will clarify this informal definition.[8]

In the social sciences, preferences are almost always the subject of assumption rather than study. We simplify preferences by imposing *a priori* that for most human decisions, preferences are unidimensional, single-peaked, symmetric, and so on.[9] There is little justification for these assumptions, so why do we make them? Mathematical convenience is the typical answer, but this masks a more serious difficulty. Without simplifying assumptions, many of our models would produce different or unpredictable results.

[8] An excellent overview of this problem for statistical models is found in Chapters 4 and 8 of Harrell (2001).

[9] Note that assumptions of this type go well beyond more fundamental (and defensible) axioms such as well-ordered preferences and transitivity.

To be more concrete, imagine you were in an expensive ice cream parlor and had never before tasted ice cream. In addition to the flavors of ice cream, you have the option of adding sprinkles, nuts, syrups, and the like. All told, you have 10 flavors of ice cream and 10 different optional ingredients and want to test every possible flavor so that you could determine a preference ordering. This natural enough desire would probably bankrupt the store (and require you to do some shopping for larger clothes), as $10 \cdot 2^{10}$ possible combinations (recipes) exist. Unless one imposes limiting assumptions on the nature of preferences, there are no shortcuts possible – you would need to test every flavor if you wanted to be certain about your preference ordering. In many cases, you would feel justified in asking for this huge number of samples, because most everyone would agree that although sprinkles and marshmallows taste great singly, in combination they might be too sweet. Recipes are one example where the different dimensions of choice are nonseparable. We do not independently sample each ingredient, arrive at a set of ideal points, and then throw them all together in a pot.

This problem worsens if the ice cream parlor subsequently adds ingredients. Imagine you had just completed the extensive taste tests outlined above and then strawberries were provided as a new option. Would you be able to somehow "save" the results of your previous search, or would you have to begin an entirely new set of tests?[10] Few of us would think that adding strawberries to a hot fudge sundae, for example, would improve the sundae, whatever our preference for strawberries. It is easy to see that as the number of ingredients increases, the size of the resultant parameter space for ice cream recipes expands exponentially – and this is not a good thing![11] In the context of recipes, making the assumption that preferences are always separable would be quite odd, and would likely lead to equally odd results. One should instead depend upon domain-specific knowledge about cooking to simplify matters, but it may not be obvious how to go about this.[12]

[10] This exercise is left to readers, especially for those who like ice cream.

[11] I will argue throughout this book that trying to understand a problem like preference formation, without assuming away the complexities of the phenomenon (e.g., nonseparability), is a very important activity despite the ugly combinatorics involved.

[12] Domain-specific knowledge is information about the problem under consideration. Unidimensionality, for example, is appropriate to some contexts and not others – for recipes, it would be inappropriate. In all cases, one has to justify assumptions by the final performance of the model, not by appeals to abstract and untested notions about rationality or mathematical simplicity.

Ice cream recipes aside, how ubiquitous is the curse of dimensionality? Some readers will immediately point to statistical work, where the curse of dimensionality appears in a nearly equivalent form. Often, our data are insufficient for testing the huge parameter spaces implied by our independent variables and modeling choices. Like the preferences literature, empirical modelers often resort to limiting assumptions (e.g., linearity of the functional form) to derive results. We rightfully question these results due to their dependence upon atheoretic modeling choices and data mining.

The curse of dimensionality is not, however, limited to statistical work. Game theoretic work falls prey equally often. Assumptions are also parameters, and the structure of game theory comes at the price that results are conditioned upon the values chosen for these assumptions. Additionally, not just any assumptions will do, as formal modelers have to find a way to fit problems into the encoding of game theory (i.e., an extensive or normal form representation of strategies, explicit utility functions, and backwards induction as the solution algorithm). Many "games" do not fit comfortably within this encoding; as a consequence, technical assumptions end up doing a great deal of heavy lifting in many formal models. The intellectual process involved in finding a set of assumptions, choosing an equilibrium concept, and choosing an abstract game to produce an outcome desired *a priori* is not different in kind from the curve fitting of some empirical researchers.

It is important to go beyond criticism, however. The more important objective of this book is to provide both a framework for evaluating models and a set of tools designed to deal with the problems sketched in this prelude. The curse of dimensionality highlights the difficulty of using mathematical models to study complex phenomena. Contributing to this difficulty is the gap between analytic models and empirical tests; it is not a coincidence that as we extend our reach to investigate more complex phenomena, concerns have grown about the quality of our results. One consistent answer to these difficulties is to keep modeling simple, such that one can understand

and test all the moving parts in a given model. This position is elaborated quite well by Axelrod (1984), but it is not surprising that his advice is largely ignored by scholars attempting to "push the envelope." The main question is how to build more complex models of behavior without sacrificing the ability to subject the results to exacting scrutiny.

Thus, I do not believe that mere ignorance accounts for the existing problems in mathematical modeling in the social sciences. Rather, the complexity inherent in many problems of interest has hampered our ability to generate models with clear empirical referents. In this book, I will integrate computational modeling into existing methods and demonstrate how many classes of problems demand a shared approach that includes computational modeling.[13] Computational methods are poorly understood (and sometimes poorly utilized) in the social sciences, despite an increasing presence in both training and research. Yet, it is my contention that computational modeling offers several advantages over traditional modeling strategies when confronted with a variety of games and decision contexts.

THE BOOK IN A NUTSHELL

There are three components to this book. The first builds a framework for evaluating models. Whatever the methodological orientation of a model, one should ask the following questions:

1) What are the assumptions/parameters of the model? Do the values chosen for the parameters come from qualitative or empirical research, or are they chosen arbitrarily (i.e., for convenience)? More important still, do the assumptions spring from a consideration of the problem itself, or are they unrelated to the main logic of the model?
2) Is there any assurance that the results of the model are immune to small perturbations of the parameters; that is, is there an equivalence class where the model yields the same results for a neighborhood around the chosen parameters? Or, is the model brittle?
3) Do the results of the model map directly to a dependent variable, or is the author of the model making analogies from the model to the empirical referent? Although toy models[14] have their place in developing intuition, they are difficult to falsify, and even more difficult to build on in a cumulative fashion.
4) Are the results of the model verified by out-of-sample tests? In this book, it will be argued that the only appropriate out-of-sample tests for a model are either
   a. a large-N statistical approach that tests the model directly;
   b. a logical implication derived deductively from the model.
5) Is the parameter space of the model too large to span with the available data? This, as noted earlier, is the curse of dimensionality, and one should never neglect the importance of bean counting. To cope with large parameter spaces, did the author of the model derive a domain-specific encoding, provide a feature space,[15] or use theory in other ways to lessen the impact of the curse of dimensionality?

Topics 1–3 are covered in Chapter 1 of this book. In addition, Chapter 1 presents a comprehensive statement on epistemology that justifies the above framework. Topics 4 and 5 are covered in Chapter 2, which also introduces the concept of feature spaces and their role in surmounting large parameter spaces. Examples using currency adoption and the security studies literature on militarized interstate disputes illustrate the main concepts.

While the first two chapters focus on how to assess models, Chapters 3 and 4 focus on the second component of this book: computational

[13] At the broadest level, computational models are numerical experiments where one uses computers to simulate a problem rather than solve it deductively – Monte Carlo statistical methods are one familiar example.

[14] Toy models are defined here as a class of simple models without any unique empirical referent. For example, the iterated prisoner's dilemma (IPD) is a simple game that investigates cooperation. It seems unlikely that all of human cooperation is a two-player contest with the exact strategy set of the IPD, and there is enormous difficulty in analogizing from the IPD to actual human behavior with enough precision to do any sort of predictive work.

[15] Feature spaces will be covered in Chapter 2. Feature spaces use domain-specific information (i.e., theory) to reduce the dimensionality/complexity of a problem.

methods and their role in addressing more complex phenomena. The use of computational methods makes it easier to build models that directly map to empirical tests. The main topics are:

1) How do game theoretic and computational models differ? Illustrations will be drawn from the artificial intelligence and combinatorics game theory literatures.
2) How does one "break up" a problem into smaller pieces, thereby overcoming the curse of dimensionality? The concepts of component games and idiosyncratic utility functions are examined in detail.
3) How does one use statistical work or logical implications to verify the results of a computational model (to the degree this is possible)?

In addition to these questions, these chapters provide a gentle introduction to the skills needed for computational modeling. Topics include programming languages, good programming style, and testing computational results.

The final component of the book provides two lengthy illustrations of the main concepts of the previous chapters. Chapter 4 presents the first example, which builds a complete encoding for a complex alliance game. Unlike most game theoretic models, the alliance game presented here has infinite strategies, four or more players, and the possibility of cooperation between different, endogenously created coalitions. Chapter 5 returns to the problem of the ice cream store and nonseparable preferences. Unlike situations in which one has enough high-quality data to do out-of-sample statistical work, studying nonseparable preferences requires the creation of logical implications to leverage existing survey data.

1

Not All Fun and Games

Challenges in Mathematical Modeling

INTRODUCTION

In large part, the inspiration for this book came from three sources, which can be categorized neatly as a failure, a challenge, and an ideal. First, the failure. When I began teaching in the profession, I was immediately assigned to graduate methods coursework. This is the experience of many professors trained in the last decade with a mathematical bent, and I was lucky enough to teach at an institution with an excellent culture. Unlike many other political science departments that exist in a state in which "there is war of every one against every one," Duke's political science department is almost entirely free of disputes about the value of mathematical modeling in the social sciences. Divisions of opinion certainly exist but, more or less, everyone in the department recognizes the virtue of mathematical methods for at least some problems.

Better still, even those who do not practice mathematical modeling believe in good research design. As many prospective faculty members discover during their job talks, "methods questions" and questions about research design are just as likely to come from the theorists of the department as anyone else (though couched in different terminology). Between job talks, faculty brown bags, and informal interactions graduate students have with faculty, it would be hard to finish a Ph.D. at Duke and not try your hand at mathematical modeling.

Despite this positive culture, teaching graduate methods coursework has not been easy. As has been noted in numerous places, the

shock most politics students experience on entering graduate school is severe. They expect to talk shop, debate the issues, and deal with "big" questions about the state of the world; instead, their first experience of graduate training at Duke involves a mathematics camp in the dull heat of August. No weighty matters of politics are discussed in this camp, unless one thinks that urns and the different colors of balls one places in them are of great import. Some students take years to get over this shock, essentially repeating much of their methods coursework when they come to a point in their own research where they have a pressing need for it. Others acquire good technical skills but nonetheless have great difficulty finding interesting questions or arriving at "good" models. Clearly, my best efforts were not sufficient and it drove me to think about issues of modeling in the social sciences and how one should attempt to improve matters.

In particular, why were so many bright graduate students, many of whom had good technical skills, unable to make the leap to generating testable theories? Why did many graduate students identify themselves primarily by their choice of method (e.g., game theory) rather than their research question? And, finally, were there any features of mathematical methodology in political science that added to the difficulty of training graduate students? These questions form a thread that continues throughout this book, and, hopefully, the questions offered here will demonstrate that many of the problems in training are related to conceptual problems in our mathematical methodologies.

The second influence on this book concerns a challenge to the discipline raised by Beck, King, and Zheng (hereafter, BKZ). Their paper – "Improving Quantitative Studies of International Conflict: A Conjecture" – appeared in the *American Political Science Review* in 2000. The paper was a broad challenge to empirical work throughout the social sciences, not just in international relations, and turned on the idea of what the proper relationship was between deductive models (usually represented by game theory) and empirical work (applied statistics). Normally, the ideal paper for the mathematical modeling crowd is a well-specified game that reaches some equilibrium outcome, which is then instantiated and tested in an appropriate statistical model. If hiring is any signal of departmental preferences, empirical work or game theoretic work alone is not as desirable as a combination of the two.

The importance of the paper by BKZ is that they argue for an entirely different approach. Instead of modeling the data generating process (DGP), they assume it is complex and interactive, and that prior efforts to model the origin of conflicts using game theory have not amounted to much (at least not anything testable). They conclude that the only reasonable standard for evaluating a statistical model is out-of-sample performance, without regard to the assumptions or specification of the statistical model. Not surprisingly, they adopt a nonparametric approach and use a neural network to generate an empirical model of conflict without regard to any underlying theory. Their article thus challenges the current methodological orientation of the discipline, insofar as they eschew the ideal of mapping a strategic game to an empirical specification.

I was confident that BKZ were wrong on several particulars, most notably whether their model actually outperformed the standard logit model used by many scholars in quantitative international relations. Along with Christopher Gelpi and Jeffrey Grynaviski, I wrote a reply addressing this problem. Additionally, we presented a general framework for comparing models when the goal is maximizing out-of-sample performance.[1] The larger epistemological questions raised by BKZ remained, however, and their challenge cast into doubt the proper relationship between deductive and empirical models. This dispute and how it relates to the broader themes of this book are dealt with starting in the next chapter.

The final source of inspiration that led to this book concerned an ideal of the proper approach to mathematical modeling in the social sciences. This ideal was first advanced in a set of workshops dubbed "Empirical Implications of Theoretical Models" (hereafter, EITM) funded by the National Science Foundation (NSF) in 2002. After these initial meetings, EITM evolved into a joint effort of Harvard, Michigan, Duke, and Berkeley to train advanced graduate students during the summer. Unlike other methods workshops that focus on particular skills (e.g., the Interuniversity Consortium for Political and Social Research's summer courses), EITM has the larger, epistemological goal of helping young researchers to bridge the divide between

[1] Our article, plus a response from BKZ, is in the May 2004 issue of the *American Political Science Review*.

deductive and empirical methodology. The goals of EITM were summarized in a 2002 report presented at the NSF:

> Significant scientific progress can be made by a synthesis of formal and empirical modeling. The advancement of this synthesis requires the highest possible levels of communication between the two groups. Formal modelers must subject their theories to closely related tests while, at the same time, empirical modelers must formalize their models before they conduct various statistical tests. The point is not to sacrifice logically coherent and mathematical models. Rather, it is to apply that same rigor to include new developments in bounded rationality, learning, and evolutionary modeling. These breakthroughs in theory will be accomplished with the assistance of empirical models in experimental and non-experimental settings.
>
> How will progress be measured? There are several performance indicators, including the number of articles that use formal and empirical analysis in the major professional journals. Another measurable indicator is the number of NSF grant proposal submissions by faculty and graduate students (doctoral dissertations) that use both approaches. However, the one area that may be the most difficult to measure is improvement in the quality of knowledge. In this regard, the ramifications of merging formal and empirical analysis is a transformation of how researchers think about problems and whether they take intellectual risks in synthesizing the model and testing it. When they do, the primary achievement of EITM will be a better understanding of the political and social world, more accurate predictions, and ultimately the provision of solid information to policymakers whose choices can profoundly affect citizens' quality of life.

Although out-of-sample forecasting is specifically emphasized in the above passage, it is obvious that the EITM founders have in mind something quite different than the nonparametric work of BKZ. Their goal is to rework the discipline so that the chasm between formal modelers and empirical researchers is bridged, with the hopes that this synthesis will lead to better models that have clearly testable empirical hypotheses.

By and large, I was very sympathetic to the goals of EITM, and was lucky enough to be invited to participate as a faculty member in the 2003 session at Michigan. My job seemed easy: take two days and present a framework for accomplishing EITM-style research. In my mind, this meant making an argument for how one might bridge the gap between models (usually deductive) and empirical tests; currently, the clearest statement of the difficulties inherent in this problem is found in two articles by Signorino (1999) and Ramsay and Signorino (2003). After a bit of reflection, the issues involved were more difficult than I at first realized. Many of the arguments presented in this text are a direct result of the questions I faced in formulating my talks for EITM. Chapter 3 lays the groundwork for this investigation, and Chapters 4 and 5 provide a set of tentative answers to how one might implement the EITM statement on methodology.

WHAT THIS BOOK IS NOT

Before proceeding, it is important to say what this book is not. This book, despite appearances in some places, is not a critique of game theory (or formal theory more broadly). Although I am critical of some current practices, it should be obvious that I firmly believe in the aspirations of those who wish to make political science an actual science, complete with predictions and policy advice about events in the real world. My main concern is that game theory has become confused with definitions of human rationality. In this text, I will argue that game theory is a mathematical tool, not a proxy for human rationality where if one departs from game theoretic models one automatically sacrifices any notion of rational agents. As a tool, it is one way to "solve" problems and is better suited to some classes of problems than others. Most of the examples I focus upon concern classes of problems that for a number of reasons are ill-suited for a game theoretic approach, and I propose a set of methods "rational" agents might employ to deal with these complications. The reason for providing tools that expand the class of problems one can deal with analytically is in my mind simple: better models, with more verisimilitude, allow an easier transition to empirical tests. This is the primary goal advanced by EITM.

Game theory also has been confused with pure mathematics, insofar as many practitioners feel no need to connect their models to empirical tests. Much that masquerades under the classification of "theory building" is not worth the appendices, and one should question the usefulness of models that rely upon limiting assumptions to produce whatever narrow result is desired by the researcher.[2] Following Granger (1999),

[2] Arrow's impossibility theorem, in contrast, depends upon assumptions that are of substantive interest and produces a result that is extraordinarily broad.

the viewpoint adopted here is that the connection between theoretical models and their empirical referents needs to be direct enough such that we can be satisfied that the tests we conduct are actually dispositive. Dispositive tests distinguish the actual model (or data generating process) from the universe of possible models. This viewpoint is by no means new; rather, it has been the subject of debate in economics for decades.[3] What is perhaps new will be the particular modeling approach adopted here, which combines traditional game theoretic investigations with computational models. The reason for this union hopefully will become clear in subsequent chapters.

This book is also not a critique of empirical work in political science, though, again, one might be confused given that in places I am critical of existing efforts. Just as models without empirical tests are suspect, so, too, are data-driven statistical investigations that fail to make apparent what model is being tested. Good statistical work allows us to distinguish useful models from the universe of irrelevant models; further, it allows us to investigate the generality of a model and the places where assumptions are carrying too much of the load. I will, however, place rather more emphasis on predictive work than is currently the norm within the social sciences, as much of the statistical research that has been conducted in the social sciences aims solely at comparing the in-sample performance (or "explanatory power") of various models.

In-sample comparisons should be seen as innately suspect, as one can easily overfit a statistical model and claim "success" for a theory. More time will thus be spent in this text addressing the curse of dimensionality that has to this point been largely ignored by social scientists.

A Simple Example: Applause, Applause

As is fitting for a book on modeling, let us begin with a simple question. Hopefully, this will introduce most of my essential arguments before we wade into the deep end of the book. The history of this example is

[3] See, for example, the October 1993 *Special Issue Anniversary of the American Journal of Agricultural Economics*. Castle (1993) and Leontief (1993) are particularly useful in this issue, insofar as they outline a set of requirements that would help connect deductive models with empirical tests.

a rich one, given that it was used for many years by John Miller and Scott Page at their Computational Economics Workshop at the Santa Fe Institute.[4]

Imagine you are asked to explain or predict the occurrence of standing ovations. You have a performance of some type, where each member of the audience receives a signal from the performance about how good it is (based upon their own internal preferences). Each audience member can then choose to do nothing, applaud, or stand and applaud. They also can sit down again at any point should they decide to stand initially. This is a highly stylized problem but has relevance for social scientists. We often want to understand who stands, or votes, or participates in a riot, and how individual characteristics and social dynamics lead to this behavior.

There are different approaches one might take to this problem, and in social science one can roughly describe the three methodological traditions that could be utilized: empirical, deductive (i.e., game theoretic), and computational. Let us investigate what sorts of answers these traditions, in isolation, might provide to the standing ovation problem.

An empirical researcher would likely start out with questions concerning what measures would be collected for both the dependent and independent variables, and not all of the forms of these measures would be obvious. For example, the dependent variable might be coded as a binary variable measuring whether or not the ovation occurred. If this encoding is adopted, what would the right threshold be for distinguishing an ovation? Would 90% have to stand? More? Less? The choice of scale for the dependent variable is also not obvious; one could change both the temporal and spatial characteristics of the dependent variable. For example, one encoding would measure the length in time of the ovation, but any such measure of time would retain the problem of choosing an appropriate threshold. Alternately, one could measure the likelihood that any given audience member participates in the ovation, thus changing the unit of analysis spatially from the entire audience to each individual member.

[4] Past answers to this problem are archived at http://zia.hss.cmu.edu/econ/homework95.html. For the most recent investigation of this problem, see Miller and Page (forthcoming).

Another more insidious problem would involve the nonindependence of observations.[5] Cleary, if subsets of the data set involved repeat performances by the same artist, "buzz" might result in a lack of independent, identically distributed (IID) observations. This problem also would complicate the measure of independent variables. Measures of performance quality and the like could easily be contaminated by interactions either between guests of the same performance (e.g., social pressure) or for members that attend multiple performances across observations in the data set. And members of different audiences are obviously not drawn from identical distributions, as people sometimes choose which performances they attend.

Problems aside, what sorts of questions would the empirical researcher answer? Likely, it would involve establishing relationships between such concepts as "performance quality" (as perceived by the audience), the type of performance, the number of audience members, and so on, and the likelihood or length of an ovation.

A deductive (or formal) modeler would come at this problem from a different angle, where the most important decision would involve specifying the benefits and costs that are present for members of the audience when they decide to ovate or not. Clearly, you do not want to be the only fool in the audience standing and clapping madly; people would stare. Just as clear, you do not want to be the grinch, sitting alone in a sea of excited fans. At some level, though "quality" matters, you only want to reward "good" performances with an ovation, given the effort involved in standing and clapping.

The structure of the game would also involve a set of important considerations on the part of the deductive modeler. How many periods would be included in the game where agents could update their information and make choices? If an ovation occurred, how would people get back to their seats? The same sorts of utility considerations discussed in the preceding paragraph would apply with equal force to agents making choices to sit back down again.

Given these modeling choices, and the input of a few "state of nature variables" such as the quality of the performance, the deductive

[5] One also might point out that the observations are not independent spatially – that is, whether or not one member of the audience stands (or later, sits) is likely correlated with the actions of other audience members.

modeler might well reach a good understanding of the individual decisions that work together to produce an ovation. A model might also help worried performance-goers in reaching decisions about whether or not to stand for an ovation in future performances. Ever present, however, would be the worry that the limiting assumptions relied upon to formulate a sufficiently simple model might cut against the usefulness of any insight gained.

The final tradition that might generate a solution for this problem is less well known in the social sciences. A computational (or dynamic systems) researcher, in contrast to the two preceding approaches, would specify a set of rules that governed the behavior of individual audience members, along with a set of contextual variables that described such features as the seating arrangement, the shape of the performance hall, relationships between audience members, and so on. What would these rules look like? On one level, the rules would be functional expressions that would be similar to the utility functions used by a game theorist, though these functions might well be allowed to vary both in time and by the individual type of audience member. On another level, these rules could add substantial verisimilitude to the computational model by incorporating features of the problem that would be difficult to model in a deductive framework (e.g., learning models based upon research in cognitive psychology). One such rule might involve adding vision to the model – given the shape of the performance hall, not all audience members can physically see all other audience members. Any utility function that involved peer pressure should be more sensitive to people within an agent's field of vision than agents outside this field.[6]

Unlike a game theoretic model, it is unlikely that a computational model would produce a set of deductive results. What is far more likely is that the researcher, confronted with the large parameter space generated by the rules used in formulating the computational model, would have to rely upon statistical investigations to understand

[6] The outcome of such a rule is that not all audience members are created equal – that is, audience members in the middle rows nearest the stage would have a disproportionate share of influence. One also might consider the type of individual audience members. For example, if a group of Catholics got together to watch a play, it might matter if the Pope were sitting in the audience. I would hazard that if the Pope ovates, so, too, would everyone else.

any "results" of the computational model, much as in the empirical tradition. Statistical relationships between parameters, rules of interest, and the likelihood of an ovation would then be presented, albeit substituting artificial data for real data.

This is a brief sketch of an interesting problem, but it raises questions of importance to all modelers. To begin with, are the approaches complementary or distinct? On the face of it, our three stereotypical methodologists would not have much to say to each other. The empirical researcher is establishing correlations between different measures and the likelihood of ovation; the game theorist provides advice on how rational audience members should select strategies; and the computational modeler incorporates aspects of both of the forgoing approaches to produce a dynamic model that recreates a standing ovation.

All of these models ostensibly explain the same phenomenon, but can one compare or integrate the results? Or, are these simply different answers to different questions? I will argue in the succeeding chapters that it is undesirable to let each type of modeler work in a vacuum; models need to produce results that are directly comparable to competing explanations. Even within each methodological approach, models are often not unique. Different modelers will produce different answers, and the job of social science should be to sort among them by insisting on out-of-sample tests of some kind. If, for example, we are confronted with several different game theoretic models, all explaining standing ovations, how do we decide which one is closest to being right? Unless one of the game theorists makes a deductive mistake, the models will differ because the assumptions differ. Arguing about assumptions is a little like arguing about whether Wolverine is tougher than the Hulk; ultimately, it comes down to taste. This book will argue that a different, integrated approach is required to make sense of these questions.

STRIFE BETWEEN METHODOLOGICAL CAMPS

Currently, there is a sense of mutual distrust between different methodological camps. Let us start with the more forceful critiques of the empirical tradition. As part of the EITM meetings, Christopher Achen argued that one must be suspicious of empirical modeling in the social sciences. Because many models are quite complex, researchers have

an abundance of parameter choices that allow them to overfit models, generating any outcome they wish:

> Empirical work, the way too many political scientists do it, is relatively easy. Gather the data, run the regression/MLE with the usual list of control variables, report the significance tests, and announce that one's pet variable "passed." This dreary hypothesis-testing framework is sometimes even insisted upon by journal editors. Being purely mechanical, it saves a great deal of thinking and anxiety, and cannot help being popular. But obviously, it has to go. Our best empirical generalizations do not derive from that kind of work. How to stop it? The key point is that no one can know whether regressions and MLEs actually fit the data when there are more than two or three independent variables. These high-dimensional explanatory spaces will wrap themselves around any data set, but typically by distorting what is going on. They find the crudest correlations of course: education increases support for abortion, for example. In the behavioral tradition, that counts as a reliable finding. But no one knows why education is associated with that moral position (higher intellect discovering the truth? mindless adoption of elite tribal norms? correlation with something else entirely?), and that leaves open the possibility that abortion attitudes do not work the way the literature says they do. Getting rid of this cheap sense of "empirical findings" is probably the central task that empirical political research faces. . . .
>
> As an instance of the altered perspective I have in mind, I propose the following simple rule: Any statistical specification with more than three independent variables should be disregarded as meaningless. With more variables than that, no one can do the careful data analysis to be sure that the model specification is what s/he says it is. (Achen in the National Science Foundation EITM Report, 2002, Appendix B)

Or, one might look farther back to Keynes, and his critique of the hapless Professor Tinbergen:

> I infer that he considers independence of no importance. But my mind goes back to the days when Mr. Yule sprang a mine under the contraptions of optimistic statisticians by his discovery of spurious correlation. In plain terms, it is evident that if what is really the same factor is appearing in several places under various disguises, a free choice of regression coefficients can lead to strange results. It becomes like those puzzles for children where you write down your age, multiply, add this and that, subtract something else, and eventually end up with the number of the Beast in Revelation. . . .
>
> To the best of my understanding, Prof. Tinbergen is not presented with his time-lags, as he is with his qualitative analysis, by his economist friends, but invents them for himself. This he seems to do by some sort of trial-and-error

method. That is to say, he fidgets about until he finds a time-lag which does not fit in too badly with the theory he is testing and with the general presuppositions of his method. No example is given of the process of determining time-lags which appear, when the come, ready-made. But, there is another passage where Prof. Tinbergen seems to agree that the time-lags must be given a priori. . . .

These many doubts are superimposed on the frightful inadequacy of most of the statistics employed, a difficulty so obvious and so inevitable that it is scarcely worth the time to dwell on it. (Keynes 1939)

At root, Achen and Keynes are addressing the same problem in empirical methods. Unbeknown to anyone save the original researcher, choices are made in empirical work. *Lots* of choices. Given the obvious problem of false correlation, it does not seem too much of a stretch to imagine that any empirical modeler, given time, can produce almost any result that is desired. Journals and monographs, by their nature, only report "positive" results and only the "final" model. How much pain or guesswork or outright cheating at the margins that goes into an empirical paper is never seen in print. One way to think of this is to imagine every salient choice made by the empirical modeler as a parameter; results are thus conditional statements made upon the particular set of parameter values chosen. Given how large these implied parameter spaces are, one cannot place much faith in a final report of in-sample performance.

A deductive modeler (typically relying upon game theory) would certainly agree with the forgoing critique of statistical methods/econometrics. Moreover, most formal theorists believe that their methodological approach is immune to the flaws that plague other approaches. Niou and Ordeshook, for example, cite the transparency of formal theory as an enormous advantage over both qualitative and empirical methodologies:

But the rational choice paradigm and formalism are not mushrooms that sprung up in an unattended intellectual forest. They are reactions to a discipline mired in imprecision, vagueness, obscure logic, ill-defined constructs, nontestable hypotheses, and ad hoc argument. They are a reaction to a discipline that in the 1920s proclaimed the Weimar constitution the greatest political-intellectual achievement of its age; a discipline that in the 1960s substituted correlation for cause; a discipline submerged in such conveniently vague and ill-defined ideas as "power," "leadership," "authority," "group," "alliance," "function," "ideology," "culture," "regime," "stability," and "balance." They

are reactions to a discipline that substituted the well-turned phrase for concrete constructs, operational measures for theoretical primitives, and the gloss of methodological sophistication for true theory. They are, in short, a reaction to a discipline that did and does precisely what Walt critiques the formal analyst of doing – burying key assumptions in an indecipherable format, although generally that format was a language more to the liking of those who studied French and Plato in college rather than calculus. (Niou and Ordeshook 1999, 87)

In addition, it is obvious that Niou and Ordeshook draw a sharp distinction between the results of formal theory (which are uncontestable; that is, the results follow deductively from the premises) and empirical work that could be rife with spurious correlation. Bueno de Mesquita and Morrow go even further in a defense of formal theory by arguing that of all the virtues one might discover in a social science theory, logical consistency is foremost:

Walt gives three criteria for evaluating social science theories: logical consistency, degree of originality, and empirical validity. We believe that logical consistency takes precedence over the other two criteria; without logical consistency, neither the originality of a theory nor its empirical validity can be judged. Logical consistency is the first test of a theory because consistency is necessary, though not sufficient, for understanding how international politics works.

A basic point in logic drives our view. A theory, in terms of logic, consists of a system of assumptions and conclusions derived from those assumptions. A logical inconsistency exists when two mutually contradictory statements can be derived from the assumptions of a theory. When such a contradiction exists in a theory, then any statement follows logically from the theory. There is, then, no discipline for arguments in a logically inconsistent theory; those using the theory are free to draw any conclusion they wish from the premises of the theory. Logical inconsistencies deny the possibility of a theory having empirical content. Theories derive empirical content by producing falsifiable hypotheses, conclusions that could be contradicted by evidence. A theory gains credence as more of its falsifiable propositions are supported by evidence, although there are no hard and fast rules here. However, because any pattern of evidence can be matched with some conclusion of a logically inconsistent theory, such theories cannot be falsified and so cannot have empirical content. A theory is falsified when an alternative is shown to fit the range of predictions better than the initial theory. Falsification of a theory cannot happen if any evidence can be interpreted as an implication of the theory. . . .

Again, any conclusion can be derived when a logical inconsistency exists, and so the choice of which conclusion to use for policy purposes falls entirely on the tastes or prejudices of the party making the prescription. Indeed, the use

of a logically inconsistent theory to justify a policy recommendation is worse than recommendations not supported by any theory....

For these reasons, we believe that logical consistency has pride of place among the criteria for judging social science theories. (Bueno de Mesquita and Morrow 1999, 56–7)

There is, of course, a problem with arguments that attempt to draw a sharp distinction between empirical modeling and formal theory. Although it is the case, as Bueno de Mesquita and Morrow note, that "any pattern of evidence can be matched with some conclusion of a logically inconsistent theory," the exact same statement is true of logically *consistent* theory. As should be obvious (but for some reason is not), one may achieve any outcome one desires with consistent theory; all it takes is the right combination of assumptions, solution concepts, and the like. The chore for formal theory cannot rest solely upon consistency, as *the class of "consistent" games that provide any given result is infinite.*

One can think of this argument in a different way. Imagine a researcher perceives an empirical regularity – for example, that candidates tend to take positions near the middle of a left-right ideological dimension. How many consistent models could the researcher construct that would produce center-seeking candidates? Infinitely many. And only some of the models are "right" in the sense that they are analogous to the real, underlying process. All the other (infinite) models are correlated with the empirical regularity in much the same way that an empirical specification is spuriously correlated with a given sample. Without finding novel data or deriving secondary conclusions, one cannot place much certainty in any single model of center-seeking candidates.[7]

So while one must agree with Bueno de Mesquita and Morrow that consistency is a necessary condition, the more important goal is to choose the "best" theory from the class of consistent theories that produce a desired result. Friedman (1953) sums up the problem created by attributing virtue to consistency alone:

Logical completeness and consistency are relevant but play a subsidiary role; their function is to assure that the hypothesis says what it is intended to say

[7] For a similar perspective worth further study, see Lave and March (1975).

and does so alike for all users – they play the same role here as checks for arithmetical accuracy do in statistical computations. One effect of the difficulty of testing substantive economic hypotheses has been to foster a retreat into purely formal or tautological analysis. As already, noted, tautologies have an extremely important place in economics and other sciences as a specialized language or "analytical filing system." Beyond this, formal logic and mathematics, which are both tautologies, are essential aids in checking the correctness of reasoning, discovering the implications of hypotheses, and determining whether supposedly different hypotheses may not really be equivalent or wherein the differences lie.... But economic theory must be more than a structure of tautologies if it is to be able to predict and not merely describe the consequences of action; if it is to be something different from disguised mathematics. And the usefulness of the tautologies themselves ultimately depends, as noted above, on the acceptability of the substantive hypotheses that suggest the particular categories into which they organize the refractory empirical phenomena. (Friedman 1953, 11–12)

Choosing a game that provides a given result (that you want to achieve *a priori*) is thus not at all different than the problem of false correlation in the statistical literature. Not only is this always possible, it is also the case that the mapping of formal theories to results is not a one-to-one correspondence.[8] One might appeal to maxims such as parsimony, or generalizability (or whatever) to discriminate between competing formal theories, but this is very slippery epistemological ground, and places such discrimination firmly in the land of taste rather than science.

Moreover, all choices that go into a particular formal theory that are left to the modeler should be seen as traversing a very large parameter space; again, this problem mirrors the corresponding complaint levied against empirical modelers. As Peltzman (1991) notes, "Game theory has introduced a rigor in the analysis of rational behavior that was missing [but] skepticism about the marginal value of recent theory is warranted [because] conclusions drawn tend to be very sensitive to the way problems are defined and to the assumptions that follow." Game theoretic results are conditional upon these choices, and given

[8] Other than trivial examples, it is clear that the mapping of theories X to results Y is a bijection but not an injection. Rather, mapping theories to results is a many-to-one process, and the goal of formal theory should be to sort between the class of possible theories.

the size of these parameter spaces, results must be seen as exceptionally brittle things when the only test is whether or not the formal theory produces an expected outcome. The problems that plague empirical methodologies thus have almost perfect analogues in formal modeling.

One example of the forgoing pathology in formal theory can be found in the dispute between Banks (2000) and Groseclose and Snyder (2000) in the pages of the *American Political Science Review*. Banks, in short, shows that one of the results in Groseclose and Snyder's original paper on creating supermajorities in legislatures is wrong; that is, it fails the consistency condition raised above. The response, however, by Groseclose and Snyder is illuminating, insofar as they simply change an assumption such that the original result holds. As they note, the changed assumption "is crucial for our [Groseclose and Snyder's] results" and "the opposite assumption is crucial for Banks's results" (Groseclose and Snyder 2000, 683).[9] If one perceives that minimal winning coalitions are rare in actual legislatures, this dispute proves that one can certainly arrive at a model that yields that general result, even if one stumbles along the way. Further, it shows that although game theoretic results are in principle transparent, this is not necessarily the case in practice. The Groseclose and Snyder result was in print for four years before the error was found.[10]

Does computational modeling have similar defects? Of course it does. Like the formal theorists, computational modelers often claim that they also have transparent models. Instead of presenting a list of assumptions as a *fait accompli* as formal theorists do, the best computational models typically provide not only the assumptions but also an idea of what happens to the model's results when the assumptions are modified. But, despite this potential advantage, the fact remains that most social scientists cannot be expected to wade through thousands of lines of C++ code to understand the inner workings of a computational model, nor do journals and book editors publish such details. Just as

[9] As Groseclose notes (personal communication), empirical work could in principle distinguish between these competing models. Yet, the articles in this debate are entirely absent empirical work, which forces one to argue about assumptions rather than the question at hand (that is, the actual frequency of minimal winning coalitions).

[10] In addition, the normal review process would indicate that three referees also missed the error, as did the dozen or so citations of the article that occurred before Banks's reply (according to the ISS Web of Science).

with empirical and formal models, we are left with a situation in which one can write a computational model (actually, infinitely many) that will (with the right parameter settings, rules, etc.) produce any given result.

The only qualitative differences between computational models and formal theory is that computational models are rather more ecumenical in how they encode problems. Additionally, computational models often possess more verisimilitude at the cost of deductive tractability. One does not "solve" a computational model; one uses it to generate simulated data that one tests with the tools of applied statistics. Computational models are thus related to game theoretic models, except that they usually address more complex problems and lack deductive solutions (but, more on this in Chapter 3).

What sorts of additional problems plague computational models? Take for example the outputs of three models captured in Figure 1.1. Absent any additional information, it is difficult to discern what these slides are showing. All three look very much alike, though there are some differences in the level of clustering apparent in the slides. It may come as a surprise, then, that each of these slides purports to demonstrate a different computational "result," explaining such diverse phenomena as state formation (Cederman 1994), culture dissemination

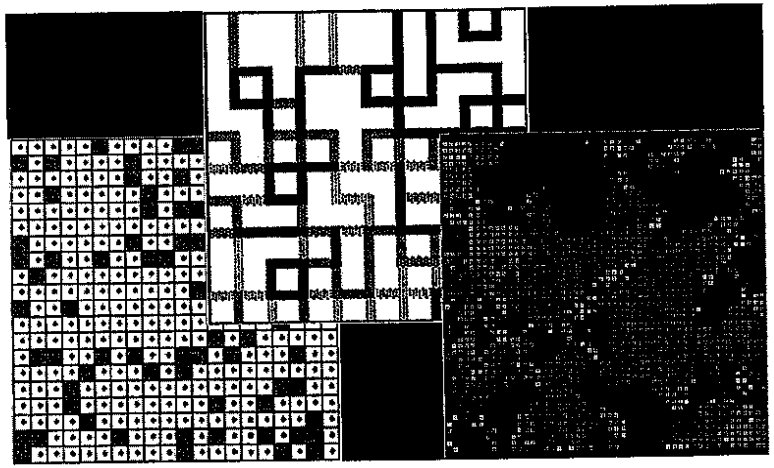

Figure 1.1. An Example from Computational Modeling

(Axelrod 1997), and collective identity (Lustick 1999).[11] While it is the case that the underpinnings of all of the models are related, this is not a cause for rejoicing – unless one has hopes that social science has stumbled on another iterated prisoner's dilemma. Each model is based upon an Ising model borrowed from physics, where it is used in statistical mechanics. By tweaking parameters (e.g., whether the neighborhood is Von Neumann or Moore), each author produced qualitative output that for whatever reason was suggestive to them. Given how slippery evaluations can be of such visual output, plus the huge parameter space underlying all such models, it is difficult to be any more optimistic about these results than those from any other methodological approach.

The main question is what one does about these difficulties. As we have seen, the problems that haunt the various methodological schools are more similar than they first appear, and the main goal of this book is to propose a solution to these problems. The short answer is that a structured combination of the methodological approaches I have listed is far superior to any approach taken separately. Much of the rest of this book will be spent examining what this "combination" looks like. While this position is at present only sketched, a brief tour of epistemology will help motivate the more detailed proposals that follow.

A SHORT STATEMENT ON EPISTEMOLOGY

From the preceding discussion, I have sketched a few of the problems that complicate the use of mathematical methods in the social sciences. If things are to improve, I would argue that a shift in our underlying epistemology is needed. The argument presented here is very close to the classic statement of Friedman (1953), and it is worth exploring how Friedman's view of epistemology has been critiqued

[11] It is important to note that Lustick's model is available for download in a format in which one can easily modify parameter values to test their impact on the results. This is extraordinarily helpful, but falls short of best practice, insofar as what computational modeling needs is a clear result that provides better predictive leverage on a question researchers care about. Like game theory, computational models most often serve as existence proofs.

(and subsequently ignored). As noted by Hausman, Friedman is not a standard instrumentalist:

> Friedman declares, "The ultimate goal of a positive science is the development of a 'theory' or 'hypothesis' that yields valid and meaningful (i.e., not truistic) predictions about phenomena not yet observed" (p. 7). This is the central thesis of instrumentalism. But from a standard instrumentalist perspective, in which all the observable consequences of a theory are significant, it is impossible to defend Friedman's central claim that the realism of assumptions is irrelevant to the assessment of a scientific theory. For the assumptions of economics are testable, and a standard instrumentalist would not dismiss apparent disconfirmations. (Hausman 1984, 217)

What troubles Hausman about Friedman's modification of instrumentalism? In Friedman's words,

> Viewed as a body of substantive hypotheses, theory is to be judged by its predictive power for the *class of phenomena which it is intended to "explain"*. Only factual evidence can show whether it is "right" or "wrong" or, better, tentatively "accepted" as valid or "rejected." As I shall argue at greater length below, the only relevant test of the validity of a hypothesis is comparison of its predictions with experience. The hypothesis is rejected if its predictions are contradicted ("frequently" or more often than predictions from an alternative hypothesis); it is accepted if its predictions are not contradicted; great confidence is attached to it if it has survived many opportunities for contradiction. Factual evidence can never "prove" a hypothesis; it can only fail to disprove it, which is what we generally mean when we say, somewhat inexactly, that the hypothesis has been "confirmed" by experience. (Friedman 1953, 8–9; emphasis added)

The distinction between proving a theory false and confirming a theory by experience is nothing new; most texts on modeling in the social sciences have adopted some version of Popper's work on falsification. What is new is that Freidman, as Hausman points out, limits the investigation or testing of a theory to the particular dependent variable the theory aims to explain. Under this limitation, attacking a rational choice model by "proving" the assumptions are not held by actual human actors is entirely beside the point. Experiments of the sort conducted by Kahneman and Tversky (1979) are useless in critiquing the results of a rational choice model designed to study a particular phenomenon; unless, of course, prospect theory has better

predictive ability for the phenomenon in question. Thus, in Friedman's terms, one can only compare theories based upon how well they predict out-of-sample, and the theorist is allowed to pick the dependent variable.

Hausman's problem with this statement is best revealed by his own example:

> I suggest that Friedman uses this view that science aims at narrow predictive success as a premise in the following implicit argument:
>
> 1. A good hypothesis provides valid and meaningful predictions concerning the class of phenomena it is intended to explain. (premise)
> 2. The only test of whether an hypothesis is a good hypothesis is whether it provides valid and meaningful predictions concerning the class of phenomena it is intended to explain. (invalidly from 1)
> 3. Any other facts about an hypothesis, including whether its assumptions are realistic, are irrelevant to its scientific assessment. (trivially from 2).
>
> If (1) the criterion of a good theory is narrow predictive success, then surely (2) the test of a good theory is narrow predictive success, and Friedman's claim that the realism of assumptions is irrelevant follows trivially. This is a tempting and persuasive argument.
>
> But it is fallacious. (2) is not true, and it does not follow from (1). To see why, consider the following analogous argument.
>
> 1′. A good used car drives safely, economically and comfortably. (oversimplified premise)
> 2′. The only test of whether a used car is a good used car is to check whether it drives safely, economically and comfortably. (invalidly from 1′)
> 3′. Anything one discovers by opening the hood and checking the separate components of a used car is irrelevant to its assessment. (trivially from 2′)
>
> Presumably nobody believes 3′. What is wrong with the argument? It assumes that a road test is a conclusive test of a car's future performance. (Hausman 1984, 218)

Hausman's example is quite nice: Assume one has a theory that predicts car performance (i.e., does the vehicle drive "safely, economically, and comfortably") based upon a test drive (in which "test drive" is the theory that produces an expectation about the car performance). Further assume that one can take the car to a mechanic, and that the mechanic can open the hood and evaluate (in his mind) the status of various components of the car.

But Hausman makes a logical mistake in the above analysis. The main problem with Hausman's argument concerns his use of the term "only" both in propositions (2) and (2′). Proposition (2) is (in Friedman's terms) correct, but proposition (2′) is a misuse of Friedman and contains a logical contradiction. 2′ states that the only test of a used car is to see if it drives well, but Friedman certainly does not mean this. Many theories could be proposed other than a test drive to determine the quality of a used car; the only qualification Friedman raises is that all theories have the same empirical referent. If the mechanic in 3′ points at a component and states that a component is flawed, one has two choices. Either this theory (i.e., flaws in components imply poor car quality) has an implication for the used car's quality, or else it has no bearing at all on overall quality. If the former is true, Friedman places the implicit theory of 3′ on equal footing with the test drive theory; the way one chooses between the two theories is to examine their out-of-sample performance. If the latter is true – a mechanic inspecting components is unwilling to make a statement about the used car's quality – one must agree with Friedman in saying that this statement has little bearing on evaluating our "test drive" theory.

Proposition (2′) is thus false, but only because Hausman failed to map his proposition to what Friedman is actually saying. Friedman believes that there is a universe of models, not just one. Thus, the use of the term "only" in reference to the "test drive" theory in proposition (2′) is unwarranted:

> Additional evidence with which the hypothesis is to be consistent may rule out some of these possibilities; it can never reduce them to a single possibility alone capable of being consistent with the finite evidence. (Friedman 1953, 9–10)

Proposition (2′) and (3′), properly restated to be in accord with Friedman, should be:

> 2′. Testing a used car's quality by seeing if it drives safely, economically and comfortably is one theory of many. One selects from the universe of possible theories by relying upon out-of-sample performance. This selection is also supplemented by consideration of the "fruitfulness" and "simplicity" of a theory.

3′. Anything one discovers by opening the hood and checking the separate components of a used car is irrelevant to its assessment unless one develops a mapping between the mechanic's assessment and the used car's quality.

Of course, what Hausman means to say is that if the mechanic looks at the engine and sees something wrong, we know for a fact that the mechanic is "right" and the test drive is "wrong." But this kind of classification makes no deductive sense, especially if one has ever visited a mechanic or listened to Car Talk on National Public Radio. Real-world mechanics get things wrong all the time, and there is no reason whatsoever to privilege the "mechanic's evaluation" over the "test drive" theory. Friedman's claim that one should treat both of these as competing theories, and adjudicate between them based upon out-of-sample performance, is thus not only logically consistent but far superior to Hausman's classification, which depends upon an unstated and unsupportable leap of faith in the mechanic.

In all of the above, one must distinguish between deductive logic and probabilistic knowledge.[12] If p is a model and q an implication or test of that model, a restatement of Hausman's critique of Friedman seems to be:

i. $p \Leftrightarrow q$
ii. $\sim p \rightarrow \sim q$
iii. Show $\sim p$ to prove that the model is wrong.

Step iii. contains Hausman's argument in a nutshell: showing that p is false – either because the model is inconsistent or because the assumptions are wrong – is all that is required to reject the model. In particular, many within the social sciences advocate scrutinizing the assumptions of a model and are reluctant to accept models that depend upon assumptions known to be false.

This line of attack misses something fundamental about research, however. Models are probabilistic in nature and one often chooses to model a phenomenon at a tractable level of granularity given the precise question asked or the data that are available. Thus, models are

[12] I am obliged to John Aldrich for this example.

rarely unique, as the use of "if and only if" in (i) implies. As Nagel notes:

> In any event, physicists show no noticeable compunction in using one theory for dealing with one class of problems and an apparently discordant theory for handling another class . . . They introduce considerations based on the theory of relativity in applying quantum mechanics to the analysis of the fine structure of spectral lines; they ignore such considerations when quantum theory is exploited for analyzing the nature of chemical bonds. (Nagel 1961, 133–4)

When testing a used car, it is possible that several models all seek to explain the performance of the car and no dispositive test exists to sort between them. Better to adopt Friedman's famous "as if" approach to theories and allow for multiple theories than to decide that some sorts of knowledge (e.g., the mechanic) are privileged.[13]

A better approach than a purely deductive formulation is illustrated by the following graph:

There exists, we would hope, a "true" data generating process that produces the sample q; to our sadness, we will never know what this process is with any accuracy for most phenomena of interest. Our model p, however, may reflect something systematic about the DGP, though it is unlikely it will capture everything systematic about a complex process.[14] A model p that captures something (though not everything) essential of the DGP trumps other models, however, that fail to mirror

[13] Compare Friedman's "as if" argument to Nagel's description of physics:

> Everything depends on the problem; there is no inconsistency in regarding the same firm *as if* it were a perfect competitor for one problem, and a monopolist for another, just as there is none in regarding the same chalk mark as a Euclidean line for one problem, a Euclidean surface for a second, and a Euclidean solid for a third. (Friedman 1953, 36; emphasis added)

This argument has caused a good deal of controversy – see footnote 14 in this chapter.

[14] As argued earlier, it is also likely that p is not unique – there are a multitude of models that all reflect different parts of the true DGP – the double arrow between p and the DGP thus reflects Friedman's controversial use of "as if" in his essay.

the DGP. Thus, it is not enough to say that if the assumptions of p are false then the model must be discarded.[15] In the case of statistical models, we accept arguments like this one without question and there is no reason the analogy does not hold for models more generally.

I would thus argue that Friedman's statement on epistemology is more compelling than critics have allowed. I would, however, make two amendments to Friedman's version of instrumentalism.

Amendment 1: Constraints upon Assumptions

On the one hand, I agree that assumptions are often proximate measures for more complicated phenomena, and to the extent they correlate with the real-world process, it is difficult to place much value in criticizing assumptions. As Friedman notes:

> Misunderstanding about this apparently straightforward process centers on the phrase "the class of phenomena the hypothesis is designed to explain." The difficulty in the social sciences of getting new evidence for this class of phenomena and of judging its conformity with the implications of the hypothesis makes it tempting to suppose that other, more readily available, evidence is equally relevant to the validity of the hypothesis-to suppose that hypotheses have not only "implications" but also "assumptions" and that the conformity of these "assumptions" to "reality" is a test of the validity of the hypothesis different from or additional to the test by implications. This widely held view is fundamentally wrong and productive of much mischief. . . . Truly important and significant hypotheses will be found to have "assumptions" that are wildly inaccurate descriptive representations of reality, and, in general, the more significant the theory, the more unrealistic the assumptions (in this sense). The reason is simple. 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. (Friedman 1953, 14)

On the other hand, there is the problem of false correlation. One may always discover a model that predicts a given empirical referent, and it is difficult at times to know whether this discovery constitutes

[15] It is always possible that the nonsystematic component of q is what the model actually exploits. One can thus have the appearance of a good model when in fact p only predicts a component of q that is accidental. This is the main reason why one must always compare models based upon out-of-sample tests.

an advance of knowledge. An oft-used example is using the winning conference in the Super Bowl to predict the outcome of the presidential race – whatever theory that is advanced in defense of this result would be met with a great deal of skepticism, no matter how accurately it predicted the presidential race out-of-sample.

Hinich and Munger (1997), in their text on analytic methods, by and large adopt Friedman's perspective on epistemology, with one exception that addresses the forgoing concern about assumptions. They add the criterion that assumptions must be plausible, because opaque assumptions make it difficult to understand how brittle a model's result is:

> We claimed above that a strength of mathematical models is the clarity of the statement of the assumptions. Yet clarity is only a strength if the assumptions themselves are plausible. One cannot tell if an argument works outside its own stylized context by looking only at the argument itself. Consequently, the external application, or "testing," of formal theory is by analogy: The theory is tested by measuring relationships among observable phenomena, in hopes that the observable phenomena are "like" the relationships the model focuses on. Without careful empirical tests, models would just be amusing mathematical exercises. (Hinich and Munger 1997, 5)

Like Friedman, Hinich and Munger also note that empirical failure or falsification is one of the key motive forces in improving models. Why then do Hinich and Munger add the consideration of "plausibility" to Friedman's statement?

While Hinich and Munger do not precisely define a plausibility standard for assumptions, this kind of concept is echoed throughout much of the public choice school, and there is an expectation that assumptions have something to do with the phenomena under investigation. Mueller (2003) and Aldrich (1995, 1997), for example, discuss the idea of adding a constant to models of turnout. As most everyone learns in their first formal theory class, rational voters, knowing that the odds of affecting the outcome of an election (either because their vote decides the election outright or because their vote causes a tie) are negligible, cannot justify turning out to vote because of whatever benefits might accrue due to their preferred candidate winning election. Following Aldrich, if p represents the odds of affecting turnout, B represents the benefit derived from one's preferred candidate winning, c is the cost

of voting, and d is the intrinsic benefit of voting (e.g., the expression of citizenship), then when

$$pB - c + d > 0$$

a citizen will vote. Given that p is in most elections arbitrarily close to 0, one has to believe that $d > c$ to avoid universal abstention.

Most models avoid the result of zero turnout by theorizing about the role of d; Mueller points out that although adding the assumption that $d > c$ to models of turnout may seem plausible, it is very difficult to say what the assumption represents. One may claim it stands for civic duty, but just as easily someone else could say the constant stands for the utility of voting as an expressive act. A plausible assumption, then, for Mueller boils down to one's ability to map the assumption to the phenomena in question (i.e., as with Hinich and Munger, the assumption should relate to turnout and decision making). Additionally, the assumption should allow the researcher to distinguish between rival hypotheses.

Here I will settle on a distinction that is broad and builds upon Hinich and Munger's argument that assumptions are related to the phenomena in question and Mueller's additional constraint that assumptions uniquely identify concepts. Assumptions are plausible if three conditions are met:

I. The assumption is related to the phenomena under investigation in a fashion that is not absurd. That is, assumptions may well be gross simplifications of reality. As Friedman notes, the best models are simple, yet nonetheless provide great predictive leverage. "Assume a frictionless surface" is certainly false in many contexts, but it is simple to relate this assumption to models that study motion. As such, it should be preserved without criticism and one should follow Friedman's advice that the way to compare models is to compare out-of-sample performance on the dependent variable selected by the researcher. Arguments that a given model's predictive power is void because actual surfaces are not frictionless would *not* be compelling.

II. The assumption is organically related to logic of the model. For example, one might create a rational investment model that predicts little or no turnout. If one adds an assumption of the sort detailed above that people have a consumption value for voting, this would likely improve the predictive ability of the model, but it fails this condition. The assumption that voting has consumption value and can be represented by the addition of a constant to the turnout calculus is indistinguishable from a universe of other explanations for this constant (e.g., civic virtue). Further, it violates the basic spirit of model, which explicitly focused on investment and not consumption.

III. The assignment of a particular value to an assumption is not, by itself, the dispositive factor in achieving a result. One way to think of this condition is to examine whether or not there is "result convergence" (a continuity condition) when the parameter embodying the assumption is subject to perturbations. As the assumption moves closer and closer to being satisfied, the outcome of the model should move closer and closer to the final result. To the extent there are discontinuities or knife-edge results based upon changing the value of an assumption, one would question the use of the assumption, as the assumption is doing all the work, not the core logic of the model. One example of an assumption that fails to satisfy this condition is the neighborhood metric relied upon in the social science version of the Ising model (see the earlier discussion and Figure 1.1) – slightly different metrics yield dramatically different results. Assumptions are always abstractions from reality, and as such, one would want to believe that results hold within a reasonable neighborhood for each parameter value. If results only hold for a particular value and no other, one would have to defend the choice of this value or else discard the model.[16]

[16] One might fault these criteria because they cut against the spirit of Friedman's instrumentalism – if a theory succeeds in out-of-sample tests, why bother with any consideration of the assumptions? The problem, as Nagel (1961, 1963) points out in his discussion of Craig's theorem, is that theoretical statements (whether true or false) have great value in organizing knowledge and in producing new theories. All theories are not equally useful in pursuing these ends, even if they have some empirical success. Moreover, Friedman does not explicitly deal with the problem that theories must include statements (which are themselves theories) on how to map theoretical objects to empirically observable objects. To the extent that this process allows for sleight of hand, insofar as arbitrary or shifting domain restrictions are often embedded in such practices, one should place less confidence in theories that violate the constraints raised here. As a concrete example, imagine N researchers provide N competing theories, all of which are consistent with a data set A, but all of which are wrong. If a new data set B is introduced, by chance alone some subset of the N

Readers will note that one class of assumptions would deserve particular scrutiny under the above conditions: technical assumptions. The name itself should raise a note of caution! To the extent that one finds technical assumptions in a model in which slight perturbations of these assumptions drive the behavior of the model, it is difficult to separate this problem from that of spurious correlation in the world of statistical methodology. At root, when technical assumptions drive results, one has to question how comfortable one is with the idea that an assumption that is unrelated to the phenomenon in question ends up accounting for a model's brittle results. To the extent it is difficult to justify one's choice of such assumptions (or values for them) endogenous to the problem under consideration, this seems exactly the same kind of practice that leads careless empirical researchers to include independent variables in a willy-nilly fashion until an arbitrarily high R^2 is reached.

At this point, many formal theorists may object to the above conditions as overly limiting. Providing a statement along with every deductive model of how changes in the values chosen for the assumptions would impact the results of a model would be quite difficult. Game theory, for example, has no ready-to-hand theory of equivalence classes of games, and it is typically the case that any change in an assumption or parameter value results in an incommensurable game.[17] To demonstrate that results are constant across perturbations in the assumption space would thus be impossible unless one adopts a different approach to modeling.

theories may still be seen as valid. There also will be new researchers who produce new and (let us assume) wrong models that comport with (A × B). If there are continual novel datasets, one would hope that at the end of the day, all the prior models are rejected and the difficulty of inventing a new model that comports with (A × B × ...) is progressively more difficult. Whenever data are sparse, however, it seems something else is needed to prevent random chance from usurping good judgment as the final arbiter between competing theories. Keep in mind that journals typically print only positive results, and thus condition III (equivalence classes in parameter space) is particularly useful – else, one cannot know when a model has failed over and again only to be resuscitated at the last moment by a fortuitous selection of parameter values or domain restrictions. In addition to Nagel, Simon (1963), Samuelson (1963), Boland (1979), and Hirsch and de Marchi (1984) represent high points in the lengthy debate over Friedman.

[17] Some of the inspiration for this concept comes from efforts to examine complex models for robustness. In particular, Miller (1998) was particularly influential, as well as the idea of parametric continuity from the optimization literature – see Sundaram (1996) for an overview.

But this is exactly the point. The parameter space generated by assumptions in deductive models should be viewed in exactly the same manner as the parameter spaces underlying statistical models. These "assumption spaces," as noted earlier, are usually quite large, even for simple games. To the extent we are critical of empirical modelers for loading the dice in the myriad of unreported choices they make in formulating a model, so, too, should we be critical of the choices made by formal theorists, especially as it is very difficult to know how crucial a given assumption or parameter value is in generating a result. In this sense, parsimony in parameter spaces is just as valuable in a formal model as it is in an empirical model.

One example of the importance of the concept of assumption spaces is demonstrated in Ramsay and Signorino (2003). Their goal is to derive a statistical model of the divide-the-dollar game directly from the extensive form of the game. The players each have a reservation value that is unobserved. To generate a unique maximum likelihood estimator (MLE), Ramsay and Signorino assign disturbance terms to these reservation values that are IID logistic variables. Their claim is that the MLE estimator they derive depends solely upon the form of the game; further, if one does statistical work using divide-the-dollar games as the data generating process (e.g., through experiments with human subjects), *only* their MLE estimator is appropriate. The problem they point to is important – using an Ordinary Least Squares (OLS) or some other estimator may not be appropriate to the game generating the data. But the fact that their estimator achieves different results than other statistical models should come as no surprise.

The main worry is that their assumption that the disturbance terms for each player's reservation value are IID and logistic violates condition III. For different distributions of the disturbance terms, Ramsay and Signorino would have to derive a unique MLE estimator for the divide-the-dollar game, and there is no logical implication from the structure of the game that one particular distribution is appropriate. Results generated with one assumption concerning the disturbance term would not be the same as results generated with other disturbance terms, even those with similar properties (e.g., a truncated normal bounded by 0 and 1). The fact that Ramsay and Signorino get different results with their method does not of itself cast doubt on prior work that makes different distributional assumptions. To demonstrate their

claim that their method is better, Ramsay and Signorino would have to show that the only logical choice of disturbance term was IID and logistic. Given that the subject of study is the divide-the-dollar game across different cultures, this would seem to be a difficult task that seems unrelated to the main problem under consideration (thereby violating condition II as well).[18]

Amendment 2: Logical Implications[19]

Reading this chapter might incline one to the belief that large N studies are the only appropriate way to test models. The main problem with mathematical modeling in the social sciences emphasized throughout is the disconnect between models and empirical tests that have the power to discriminate between competing models.

The reason models need to be clear about their empirical referent (i.e., the dependent variable that will test the model) is that, all too often, we resort to games such as the iterated prisoners' dilemma (IPD) and make broad claims about the results. The IPD purports to study cooperation, and surely it does detail cooperation of a kind. The most celebrated "result" of the IPD demonstrated by Axelrod's (1984) path-breaking tournaments is that tit-for-tat is the right strategy to employ when confronted by an IPD – many articles have taken this as a starting point and the literature on the IPD is vast. Unfortunately, this result is wrong on technical grounds (Binmore 1997), as the success of tit-for-tat depends upon the starting population. Another concern is more fundamental. Axelrod and other scholars use their results from the IPD to arrive at policy recommendations for phenomena ranging from Cold War deterrence strategies to regulatory compliance on the part of firms. Although results from the IPD might help one's intuition in

[18] One also should critique the assumption that the errors are IID. In many cultures, it is possible to imagine that the variance of the error term on the reservation value is correlated with the initial offer by player 1. For example, if player 1 makes a high offer (more than 50 cents, which some might label "irrational"), there would likely be very little variance in the error term. And if player 1 offers something close to 0 one would expect little variance. Across cultures, the variety of disturbance terms that might account for the data would be quite large. The logic of condition III would argue that how one's results change based on this choice should be the focus of study rather than a minor technical aside.

[19] This section was derived from conversations with Robert Keohane, during a seminar we jointly taught at Duke University on qualitative research methods in the fall of 2003.

facing the complexities of real-world problems, intuition is difficult to falsify. Once one decides that the IPD and tit-for-tat embody the essence of cooperation, everything looks like an IPD and it is difficult to know when one is in new territory.

Friedman's admonition to do out-of-sample testing curbs this kind of loose analogizing. Large N studies allow for the necessary and *repeated* confrontations with novel data that all modeling requires. There is, however, another approach to model testing that has a long history in the natural sciences as well as the social sciences (for an overview, see King, Keohane, and Verba 1994, section 1.1.3). One can derive logical implications of a model, and see if the implications in fact hold. The example provided by King, Keohane, and Verba concerns the study of dinosaur extinction:

> Nevertheless, dinosaur extinction can be studied scientifically: alternative hypotheses can be developed and tested with respect to their observable implications. One hypothesis to account for dinosaur extinction, developed by Luis Alvarez and collaborators at Berkeley in the late 1970s (W. Alvarez 1990), posits a cosmic collision: a meteorite crashed into the earth at about 72,000 kilometers an hour, creating a blast greater than that from a full-scale nuclear war. If this hypothesis is correct, it would have the observable implication that iridium (an element common in meteorites but rare on earth) should be found in the particular layer of the earth's crust that corresponds to sediment laid down sixty-five million years ago; indeed, the discovery of iridium at predicted layers in the earth has been taken as partial confirming evidence for the theory. Although this is an unambiguously unique event, there are many other observable implications. For one example, it should be possible to find the meteorite's crater somewhere on Earth. (King, Keohane, and Verba 1994, 11)

Note that unlike many modeling exercises, the researchers studying extinction did not attempt to fit a model to existing facts.[20] Failure to succeed at this activity is a sign of mathematical ineptitude, rather than a signal of a model's strength. For logical implications to be used as a test, they must be novel and uniquely connected to the logic of the model. This is a harder set of conditions then one might expect; as King, Keohane, and Verba note in a footnote on the same page:

> However, an alternative hypothesis, that extinction was caused by volcanic eruptions, is also consistent with the presence of iridium, and seems more

[20] For example, if we know that parties converge in a two-party system, producing a model that has this as an implication is trivial.

consistent than the meteorite hypothesis with the finding that all the species extinctions did not occur simultaneously.

We thus have a conundrum. It must be the case that logical implications are able to test theories – the examples of such tests are abundant and important in the history of science. It seems impossible to imagine that anyone would not have been jubilant when Sir Arthur Eddington's expedition verified Einstein's claim that mass curves space, by observing how light bends around stars. When there are other models that (as in the case of iridium deposits) might produce the same logical implication, one has to adopt a betting mentality. The test provided by deriving logical implications matters more when the implications are novel and untested – after one finds, for example, that light bends around stars, future models that "predict" this fact do not gain the same amount of credibility.

Novel implications are thus important in testing models, but one must be careful that the implication follows uniquely from the model in question. Unlike out-of-sample statistical work with large numbers of observations, logical implications, as a kind of test, are of a more qualitative nature. One has to ask how surprising the implication is, and how likely it is that a large class of other models might yield the same result. As noted in King, Keohane, and Verba, one aspect of testing a model via logical implications is shared with statistical work. To the degree possible, one should maximize variance in one's implications thereby enhancing the odds that one might be wrong. For example, instead of merely predicting the presence of iridium, one might derive a specific amount or pattern of sediment around the impact crater, decreasing the odds that other models produce the same implication.

Fortunately, this practice has a long history in natural science and we can stick with Einstein to provide a final example. In a close analogue to the Eddington experiment, the astronomer Sergei Kopeikin measured the displacement of light from a quasar moving around the mass of Jupiter during an eclipse. Because displacement depends upon gravity, he tested whether Einstein's theory that the speed of gravity is equal to the speed of light is true or not (it is – see Whitfield 2003). This test, obviously, is much easier to falsify, as any deviation from the constant for the speed of light would signal disconfirmation. An example of how to derive and test logical implications in the social sciences will be presented in Chapter 5 of this book.

LOOKING AHEAD

One of the core arguments of this text will be that deductive models are most useful in generating intuition about a problem, especially when one investigates limiting cases. But just as important is the process of developing an equivalence class where one has some idea about how changes in assumptions or parameter values change the results of a model.

Computational models supplement game theoretic models by allowing the researcher to investigate explicitly the properties of the assumption space. Computational modeling thus extends purely deductive work by generating equivalence classes of models, thereby increasing the confidence we have in our results. Of course, I agree wholeheartedly with Friedman's original statement of epistemology: Theory without an empirical referent is almost always navel-gazing. To the extent we can all agree on what the salient dependent variables are, thereby avoiding vague mappings from models to empirical referents, so much the better.

The goal of this book will be to provide the tools necessary to develop links between the three methodological traditions in the social sciences and avoid the problems detailed above. As there are already excellent texts on statistical methodology and game theory, much of the focus will be on computational models and out-of-sample forecasting. If the use of out-of-sample empirical tests are the best way to sort between competing theories, computational models naturally lend themselves to drawing out the implications of a purely deductive theory by allowing the researcher to build models with more verisimilitude, thereby decreasing the "gap" between the analytic model and the empirical test.

2

Looking for Car Keys Without Any Street Lights

INTRODUCTION

As we saw in the first chapter, there is a great deal of suspicion in the social sciences about purely empirical research. All too often, one finds models that fit a sample rather too well, demonstrating how modeling choices allow a researcher to discover relationships that are not genuine. Worse still, it is often unclear what is being tested in empirical work when there are ambiguities in the underlying deductive models. In this chapter, I will use the problems that complicate empirical work to highlight more general problems with both deductive and computational modeling. I will focus on empirical modeling initially, however, because these problems appear in a very clear form in empirical models. Moreover, empirical work is more common in the social sciences than either of the other traditions and thus deserves early attention. The goal, however, is not simply to criticize empirical work. Without it, no model stands on very firm ground, so this critique aims at the higher goal of generating a set of standards that would allow empirical work to be tied more closely to testing deductive and computational models. To demonstrate the main points I wish to make, I will draw more from the nonparametric and neural networks literatures than is common in social science. Additionally, research I have conducted with Christopher Gelpi and Jeffrey Grynaviski, and an ensuing debate with Nathaniel Beck, Gary King, and Langche Zheng will provide an example from an ongoing research question in security studies.[1]

[1] See de Marchi, Gelpi, and Grynaviski (2004) and the response by Beck, King, and Zeng (2004).

With this in mind, there are two problems I wish to consider in empirical work. The first problem concerns the trade-off between overfitting and underfitting a model. The second problem is named the "curse of dimensionality" in the nonparametric statistics literature, and concerns the size of parameter spaces in models. Both of these problems are underappreciated in social science and so deserve some attention here before progressing to the section on neural networks and security studies. After the section on neural networks, I will conclude with a brief section that outlines how the problems presented in this chapter apply with equal force to deductive and computational modeling – a subject that will be taken up in more detail in Chapter 3 of this book.

CHALLENGES IN BUILDING EMPIRICAL MODELS

Overfitting

We all learn in our first statistical methods class that the data generating process (DGP) is a big part of empirical modeling. It is, after all, the underlying process that we (hopefully) capture in our empirical specification, and to do statistical work we are forced to make assumptions about the nature of the DGP. The more precise we can be in these assumptions and the more accurate our assumptions about the DGP are, the better our statistical work will be – or so the story taught in most seminars goes. Spanos (1986) relates the accepted view of the role between theorizing about the DGP and building a statistical model:

> *Observed data* in econometric modelling are rarely the result of the experiments on some isolated system as projected by a theory. They constitute a sample taken from an on-going real DGP with all its variability and "irrelevant" features (as far as the theory in question is concerned). These, together with the sampling impurities and observational errors, suggest that published data are far from being objective facts against which theories are to be appraised, striking at the very foundation of logical positivism. Clearly the econometrician can do very little to improve the quality of the published data in the short-run apart from suggesting better ways of collecting and processing data. On the other hand, bridging the gap between the isolated system projected by a theory and the actual DGP giving rise to the observed data chosen is the econometrician's responsibility. Hence, in view of this and the multitude of observed data series which can be chosen to correspond to the concepts of theory, a distinction is

suggested between a theoretical and an estimable model. A *theoretical model* is simply a mathematical formulation of a theory.... This is to be contrasted with an *estimable model* whose form depends crucially on the nature of the observed data series chosen.... In order to determine the form of the estimable model the econometrician might be required to use auxiliary hypotheses in an attempt to bridge the gap between the theory and the actual DGP. It is, however, important to emphasise that an estimable model is defined in terms of the concepts of the theory and not the observed data chosen. (Spanos, 664; emphases in original)

As with most coins, there is a flip side: data mining. Unfortunately (again, according to seminar wisdom), many researchers engage in data mining, which is a brute force approach of searching through the space of possible models (an infinite set) until one finds a model that "works" for the existing sample. Contrasted with the above approach outlined by Spanos, data mining is empirical work absent any consideration of the underlying DGP. Kmenta (1997) sums up the conventional wisdom neatly:

In current research practice, the availability of well-defined competing models is not that frequent. Economic theory can often indicate which explanatory variables should be included but does not give much guidance with respect to functional form, lags in behavior, inclusion of control variables (e.g., social or demographic), or measurement of variables. Typically a researcher is faced with a list of regressors of which some are clearly to be included in the equation but most are uncertain candidates. The researchers then resort to some ad hoc criteria that enable to them to make a choice.... Probably the most common way of choosing a model in empirical research is by "data mining." A researcher confronted by a list of regressors tries various combinations of variables until satisfactory results (high R^2 "correct" signs of regression coefficients, a reasonable value of the Durbin-Watson test statistic, etc.) are obtained. This is known as "torturing the data until they confess." (Kmenta, 598–9)

Data mining is one example of the more general problem of overfitting. As noted in the first chapter, many researchers distrust empirical results due to the large parameter spaces involved. Often, it is all too easy to "discover" models by leveraging nonsystematic characteristics of a fixed sample. Overfitting is dangerous because it confuses the partially idiosyncratic nature of any fixed sample with genuine characteristics of the data generating process.

A standard example of overfitting can be provided by a very low-tech Monte Carlo experiment. First, generate a standard uniform variable named *normals* for some number of observations.[2] Treat this as your dependent variable. Next, generate a simple index variable x that counts the number of observations (i.e., $x \sim [1..N]$). Obviously, *normals* and x are completely unrelated, but imagine a modeler does not know anything about the DGP that created these variables and is convinced that the two *are* related. The question is, how far wrong could a modeler go in pursuing a relationship between *normals* and x?

If our creative modeler tries a linear regression, disappointment will result. The R^2 is close to 0 and a histogram of the residuals does not look normal for most small samples.[3] Visually, the predicted linear regression line will have a slope close to 0, indicating that no relationship exists. If she turns to a more complex statistical model what might happen? Figure 2.1 shows the results of fitting a lowess regression that depends upon local neighborhoods to fit a function – similar results can be obtained for any neighborhood regression technique[4] (e.g., median splines) or for other techniques that allow dramatic changes in slope (e.g., including several higher order polynomials of x in a linear regression).

The Stata function for lowess allows one to easily change the size of the neighborhood used to fit the estimated function. As the neighborhood gets larger, the results of lowess approach that of the linear

[2] In Stata, one has to set the number of observations and then generate the standard normal variable and the index. To accomplish this, use the following commands:
set obs 15
gen normals = invnorm(uniform())
gen x = _n
To graph both the sample and regressions on the sample, use the following command with different parameter values for bwidth:
scatter normals x || lowess normals x, bwidth(.5)

[3] Our creative modeler sees this as evidence that a more complicated relationship is latent in the data. When the sample size exceeds 50 observations, however, the residuals will look normally distributed.

[4] Neighborhood techniques partition the data into different adjacent subsets and fit each subset separately. Imagine a real valued independent variable that ranges from 1 to 100 – dividing this domain into 10 equal intervals and fitting a linear regression to each of them is a crude example of a neighborhood technique. The key parameter is the size of the neighborhoods.

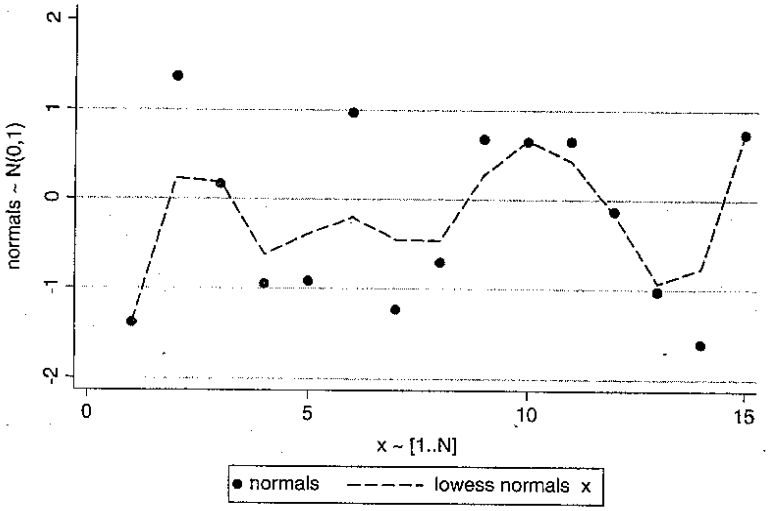

Figure 2.1. Creative Modeling

regression; as the neighborhood gets smaller, the lowess function becomes even more irregular, though better fitted to the sample. Figure 2.1 shows a middle value for the neighborhood size. Absent knowledge of the underlying DGP, a creative modeler will be satisfied with these overfit models (that depend upon small neighborhood values) given how well they conform to the data. In essence, a relationship has been created out of thin air, but visually it is easy to see that overfitting has occurred. The data in Figure 2.1 are evenly distributed around 0 with constant variance, and the lowess function is hopping from one point to another in an unpredictable fashion.

Repeating this experiment for a variety of samples demonstrates that things can go very badly for any modeling technique. Even if one avoids the mistake of using a complex model that overfits the sample, OLS models also can go astray if the sample has a trend (just as this sample has a slight positive slope). It is easy to forget that a sample is exactly that, and there is no direct way to verify that you have not overfit a model.[5]

[5] As I have argued in Chapter 1, out-of-sample testing is the only way to avoid the problems raised in this section.

Underfitting

It may not be as obvious that it is also possible to underfit a model, which would mean that one ignores systematic components latent in a sample. Sometimes, underfitting is justified by the researcher on the grounds that self-imposed handcuffs in the estimation process are a necessary safeguard, given the way results are presented in journals. As has been noted in many places, one only publishes the final model, and there is little room in journals for details of the journey that led to those results (or for negative results or replications). If one follows the admonition from Achen quoted in the first chapter, it makes sense to adopt a minimalist strategy, especially if one cares more about overall confirmation for a model than particular point estimates.

But for other research problems, this strategy will not do. In the quantitative literature on the causes of international conflicts, it is clear that better fit models are valuable. Underfitting, when one wishes to predict or assess the possibility of the outbreak of conflict between nations or any other important political event, is every bit as problematic as overfitting. One might imagine that decision makers desire as good a model as possible for use in the allocation of limited diplomatic and military resources for the prevention of war.

An Example: The Currency Game

For most modeling exercises, there is a trade-off between overfitting and underfitting that is difficult for a researcher to detect with any precision. Consider the following example borrowed from Young (2001). N actors in a society must decide on a currency, and there are two possibilities to choose from. Let the first currency be gold and the second silver. Initially, the N actors will be randomly assigned gold or silver with equal probability. Let p_t be the proportion of gold users in the population at time t and $(1 - p_t)$ be the silver users. At each subsequent time period, one actor will be chosen by a uniform draw from the population and will make a new decision according to the following rule:

1. With probability $(1 - \varepsilon)$, if $p_t > 0.5$ (i.e., gold is the dominant currency) the actor chooses gold or remains a gold user if one already; else, if $p_t < 0.5$ (i.e., silver is the dominant currency) the actor choose silver or remains a silver user if one already. If

$p_t = 0.5$ exactly, the actor continues with whatever currency they were using previously.

2. With probability $\varepsilon > 0$ (by assumption) the actor makes a new uniform draw between the two currencies (which means the actor changes currency with a chance of 50%).

Perl code that implements this model is available in the appendix to this chapter.

In this model, a natural dependent variable would be the average number of regime shifts one sees over any given time period; that is, if one allows the above game to go on for 100 iterations, how many times would the currency change from one standard to another? This variable is an integer with a range from [0..t]. Fortunately, there are not very many parameters that might explain this dependent variable. The two main candidates for independent variables are ε (the mutation rate) and N (the population size). For the sake of this initial example,[6] let us assume that the mutation rate is fixed at 0.5, reducing our investigation to a bivariate regression focusing on the role of N. One would imagine that this setup is straightforward (certainly simpler than a cross-sectional time series study involving multiple nation-states) and that finding a "good" model should be easy.

To look at the issue of overfitting versus underfitting, I created a training set of 500 observations varying N from 10 to 50 (in fact, there are 100 observations for each multiple of 10). Using this sample, I fit three candidate models: a linear, polynomial, and median spline regression. In Figure 2.2, the fit between the three models and a test data set of 500 new observations is presented; the linear, polynomial, and median spline model are graphed against the dependent variable for the number of regime shifts.[7] As is obvious, seemingly trivial

[6] We will, however, return to this simple currency game at several points (albeit with some amendment).

[7] Given that the parameter space is small and the model is deterministic, one does not strictly have to go to the trouble to produce a test data set. I have done so here for the sake of good pedagogy (many thanks to Ken Kollman for suggesting this). What may not be obvious to readers that have never done out-of-sample testing before is how easy this is. Simply fit your model(s) to the training set, record the equation, and then generate predicted values using this equation on a new test set. If there is not an actual test set available, one can do this artificially by taking a sample and removing some of the observations with uniform random draws to form the test set.

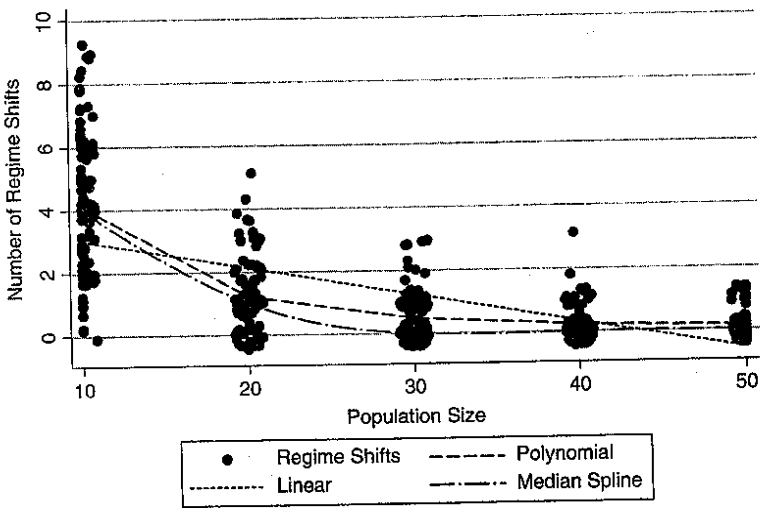

Figure 2.2. Models, Models Everywhere

modeling assumptions result in large differences in model fit, even though each model returns a qualitatively similar set of summary statistics for model performance. The good news in Figure 2.2 is that all three models seem relatively stable; although the linear model underfits the data, it would be difficult to argue that either of the more complex models is overfit or underfit. So as a baseline, one can see that if one is lucky enough to have data generated deterministically in a computational model, the problems of fit are not extreme.

Figure 2.3, however, demonstrates a different lesson. The DGP is the same, except a stochastic term (~N(0,16)) was added to the measurement of population size. One can imagine that in real-world datasets on currency use the measure for population size is a bit noisy, or there are recording errors, or any number of other problems, though in the data presented here the errors are both relatively small and centered on the true value. As is obvious from Figure 2.3, the addition of a stochastic term to the measurement of population size affects both the linear model and the spline model in unfortunate though different ways. On the one hand, the linear model is clearly underfit, and even within the sample it does fairly poorly at both the low and the high end of the scale for the independent variable. The spline, on the other

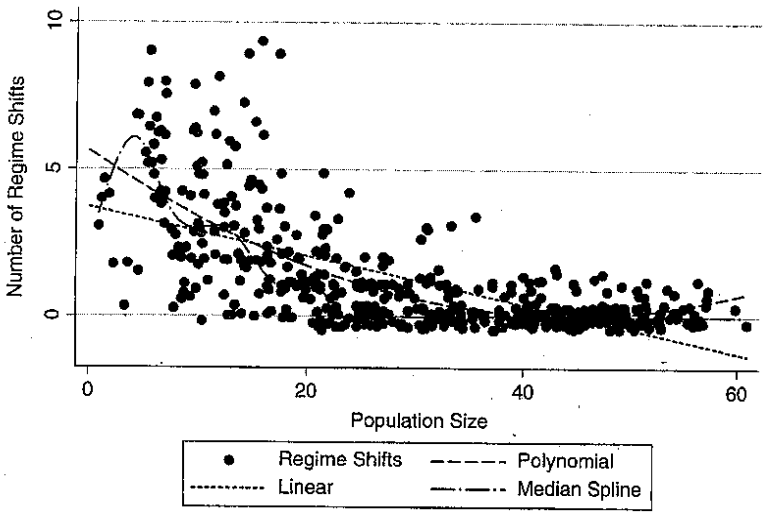

Figure 2.3. Once Again, With Noise

hand, is overfit to the data, with eight internal changes in the sign of the slope.

Only the polynomial model avoids these pathologies, and that is entirely because I limited the order of the highest polynomial term to 2 – allowing the polynomial to add additional terms would have resulted in an overfitted model. If these models were compared using a new test data set, the polynomial model would outperform the others as it is (by construction) less sensitive to idiosyncrasies in the training set. While careful graphical techniques and the availability of out-of-sample test sets can help researchers to select an appropriate model, as I have with the polynomial model, such techniques cannot be relied upon, especially when matters become more complex.[8] Keep in mind that the data in question for this example were generated according to a very simple DGP and that we are only estimating a bivariate

[8] Note that the polynomial model has only one parameter (i.e., the order of the highest term) and one independent variable. Thus, one only needs a couple of tries – each requiring a novel out-of-sample set – to arrive at good parameter values. If, however, one has many independent variables and a more complex statistical model, there is never enough data in the world to traverse the parameter space in any systematic fashion.

relationship (which adds simplicity to both the estimation process and any subsequent use of graphical techniques).

There are alternatives to graphical techniques. To deal with more complex modeling situations, a researcher might add a penalty term to summary measures of model fit to reflect the costs of increased model complexity; for example, for polynomial models, a penalty term might be based upon some function of the second derivative of the estimated function, thereby penalizing a model with large changes in slope (see Bishop 1995, sections 1.5, 1.6, and 5.4 for a review of the literature). Penalty terms of this kind essentially smooth the output of a model. "Smoothing" (or regularization) means nothing more than reducing the large slope changes and responsiveness of a model *to the sample*, in the hopes of eliminating the undue influence of outliers or nonsystematic variance in a sample. In Figure 2.3, one can see that flattening out the median spline model would have advantages.

Evaluating theories based upon statistical models is thus a very risky business. As we have seen, the consequences for overfitting or underfitting a model can be disastrous. Even simple decisions are problematic when one is aware of these difficulties. For example, imagine you are considering adding an X^2 term to a regression. Perhaps you have an excellent theoretical reason for doing so, but perhaps not. If the addition of X^2 improves your model, it may just be because you have overfit the sample. By contrast, omitting X^2 might be wrong, causing your model to underfit the sample and denying you vital information about the DGP. Without access to multiple out-of-sample datasets, how would you know which choice was correct?

Parameter Spaces and the Curse of Dimensionality

One way of looking at empirical modeling that may not be obvious is suggested by the above (brief) discussion of smoothing. If one has a model that is overfitted and to the eye appears to have large oscillations, it is natural to look for a way to smooth out these oscillations, thereby producing a better match to the underlying DGP.[9] Modeling

[9] And, it should be noted, a concomitantly worse match to the sample. As an example of this point, see Figure 2.1. Imagine a curve that simply connected the dots. This would fit the sample perfectly, but the true DGP in this case is a straight line at 0 on the y-axis. Deviation from this line is noise, and smoothing prevents one's

choices, then, can be thought of as rules for interpolating between different neighborhoods in a sample. Consider the example of a naïve researcher who reads the results of the bivariate regression detailed here on currency adoption. As one could easily imagine, our researcher is dissatisfied with the exclusion of other, possibly relevant, variables, such as culture, government type, urbanization, and the like.[10] He applies for a grant to supplement the above data set of 500 observations with more independent variables; in short, he wants to receive funds to add independent variables to each of the 500 cases in the data set so he can run "better" models that include more aspects of the real world phenomenon. On the face of it, this is a laudable goal, and anyone who has taken a semester or two of statistics in a political science department will have the rule of thumb that so long as the observations are greater than the number of independent variables, everything is fine. Given that 500 observations exist, this seems to be a grant worth funding.

Assume that the naïve researcher wants to collect 10 additional independent variables, and that each variable has a range of 10 possible values (e.g., there are 10 possible values for the culture variable). What size is the resultant parameter space? Sadly, it is 10^{10}, which is impossibly large.[11] One should immediately see that 500 observations will populate this parameter space very, very sparsely, and any results of models incorporating 10 independent variables are automatically suspect. Modeling choices, such as the reliance on OLS regression, are thus a way of forcing a particular method of interpolation on models that must span a given parameter space. In most parameter spaces found in empirical work in the social sciences, the data are quite sparse when one considers the parameter spaces they are asked to populate. OLS, as one type of model, simply imposes the constraint that whenever one finds blank spaces in a parameter space, one continues to draw a line through the voids. Other models, such as a polynomial model,

model from fitting the noise. It also should be noted that smoothing represents a bet – you could be wrong (the underlying DGP really could have multiple large changes in slope).

[10] Often, arguments of this sort are made by qualitative researchers who recommend using case studies as a way to investigate the intricacies of a problem.

[11] I am being generous here. One also has to add parameters that are required by the model to any calculation of the complete parameter space.

use more complex functions to interpolate against empty or minimally populated regions of a parameter space.

Since almost any model must interpolate, how does one choose a particular model form? By and large, empirical modelers in the social sciences reason from first principles or by analogy to preexisting problems. For example, when one allows both the population size and the mutation rate to vary in the currency adoption problem, a family of models that works quite well is a maximum likelihood regression using the Poisson distribution for the dependent variable. OLS models, in contrast, fit the data quite poorly (i.e., by underfitting). How would one arrive at the choice of a Poisson model? While some researchers might engage in the naughty practice of graphing a histogram of their dependent variable, one can instead consult a text on distributions and see if the properties of the Poisson distribution match the DGP of the currency adoption game. From Taylor and Karlin (1998) or DeGroot and Schervish (2002), one finds that a Poisson process has the following properties:

i. The number of events (in this case, the number of regime shifts in currency use) in any two disjoint time intervals are independent random variables;
ii. The probability of an event (regime shift) is proportional to the length of the time interval and not to the point in time at which the interval occurs (i.e., the process is stationary);
iii. For short intervals, the probability of two or more events occurring is of a smaller order of magnitude than just one event.

By and large, these properties seem to be a rough approximation of the currency game. Note, however, that (i) is violated, insofar as when a population tips into a different currency regime, it is easier to tip back into the former regime, given that the majority in favor of the new currency has only one extra vote (i.e., only one actor chooses in each time period t). Nevertheless, although the Poisson model is not quite right, it is a better choice than OLS, and probably good enough for most purposes.

There is, however, a quite different tradition for choosing model forms. Nonparametric statistics uses the sample itself to justify the choice of model form, and places much more emphasis on out-of-sample tests to curb tendencies toward overfitting. In order to

understand the basic issues involved in nonparametric estimation, it is useful to develop some notation.[12]

Let $Y \in \Re$ be a dependent variable, $X \in \Re^n$ be a vector of n independent variables, and $f(x, y)$ be the joint probability density function (p.d.f.) of X and Y. Our goal is simple: to develop a model $m(X)$ that matches the data generating process of Y as closely as possible. Our first choice thus involves the selection of a loss function.[13] As we have seen earlier in this chapter, the construction of a loss function can be complicated by the addition of penalty terms for model complexity and the like, but a natural first step would be to use mean squared error (MSE): $E[(Y - f(X))^2]$. Clearly, Y and X are given (i.e., they constitute the data), so how does one choose the model $m(\cdot)$?

In one adopts the assumptions of Ordinary Least Squares regression choosing $m(\cdot)$ is straightforward, and almost everyone at some point in their graduate training has derived the normal equations.[14] OLS regression replaces $m(\cdot)$ with X'b and uses calculus to determine the optimal values of the coefficients b. In the case of an arbitrary function $m(\cdot)$, things are not much more complicated:

1. $E[Y - m(X)^2] = \iint (y - m(x))^2 f(y, x)dydx$; by the definition of expected value.
2. $= \int (y - m(x))^2 f(y \mid x)dy/f(x)dx = E_x E_{Y|X}[(Y - m(X)^2 \mid X]$; by the definition of conditional probability (i.e., Pr(A|B) = Pr(AB)/Pr(B)).
3. $m(x) = \arg\min_c E_{Y|X}[(Y - c)^2 \mid X = x)]$; at each point X = x, find the optimal value for the model $m(x)$ by minimizing the relevant term in 2 (the other term can be discarded as it includes only X).
4. $\arg\min_c E_{Y|X}[(Y - c)^2 \mid X = x) = E[(Y^2 - 2Yc + c^2) \mid x] \rightarrow c = E(Y \mid x)$; using calculus, one can see that the optimal value for the model $m(x)$ is reached when c = Y in the above formula.

[12] There are many treatments of the curse of dimensionality and related issues; the notation and development used here is taken from Hastie, Tibshirani, and Friedman (Section 2.4, 2001, section 2.4). One should also look at Bishop (1995, section 6.1.3), DeGroot and Schervish (2002, section 9.6), and Gentle (2002, section 1.3).

[13] A loss function is the measure one uses to penalize models for deviating from the data.

[14] Griffiths, Hill, and Judge (1993) is an excellent introductory text on the subject. The normal equations, in matrix form, are derived by minimizing the residual sum of squares $(Y - Xb)^2 \rightarrow X'(Y - Xb) = 0 \rightarrow b = (X'X)^{-1}X'Y$.

We thus reach an answer that is hopefully not too surprising: the optimal model, at each point $x \in X$, selects the average of Y for all the observations at that particular point x. And, better still, this choice is optimal, in the sense that the mean squared error is minimized.

This is not, however, what most of us do in practice, though we may not realize how very different our normal practice is. Typically, we replace $m(x)$ with the linear function X'b.[15] What is remarkable about this choice is that we have made an incredible assumption. In our use of X'b as a model choice, you will note that we are *pooling* the data to derive our point estimate. Put another way, the entire data set X is used to determine the value of y at every point on the line. Even when there are gaps where no y's exist, we still draw the line to produce a predicted value. As noted earlier, this represents a leap of faith about how to pool data and interpolate across spaces where no (or little) data exist.

The result of equations 1–4, however, are quite different. The optimal choice is $E(Y \mid x)$. Although this looks familiar – it is simply the mean of Y conditional upon x – it presents a very different problem. By not making an assumption that allows us to pool our observations of Y, we must derive our predicted value for Y separately at each point $x \in X$. If we do not have any observations (x,y) at a given point x, then we are out of luck. More disturbing is the implication for the amount of data we need. To be certain of our estimates, we need several observations for each $x \in X$; for even simple modeling exercises, this demands a monstrous amount of data. As noted above, for a model with 10 independent variables, each one taking 10 possible values, we have a parameter space of 10^{10} – to employ the model where $m(x) = E(Y \mid x)$ and we want γ observations for each predicted value y, we would need $\gamma 10^{10}$ observations. This is in contrast to our linear model X'b, which is parsimonious in its demand for data because it assumes the relationship between X and Y is the same throughout the entire domain of X.

In the nonparametric modeling literature, the problem involved with estimating $E(Y \mid x)$ is known as the *curse of dimensionality*. As

[15] The point raised here is not unique to OLS regression. If one engages in maximum likelihood modeling, as an alternative example, one chooses a different distribution for the model but one is still pooling the data for the estimation of Y in exactly the same way as in OLS.

the parameter space of a model or problem increases, one's confidence in $E(Y \mid x)$ (which, remember, is the *optimal* estimate for minimizing MSE) drops precipitously. Fortunately, there are a number of avenues one can take to escape this problem, short of the dramatic assumptions involved in choosing *a priori* a well-behaved functional form and distribution for Y. One avenue relies upon the idea of a neighborhood. A neighborhood is a parameter δ that defines a local space around a point. When one has sparse (or nonexistent) data, for example, at a given point x_i, instead of relying upon the limited observations (or making no prediction at all) to predict y, one could instead estimate $E(Y \mid x)$ using the observations contained within $(x_i - \delta, x_i + \delta)$, thereby increasing the number of observations available and reducing the variance in the conditional mean of Y. The parameter δ can be adjusted to produce a more or less granular function, depending upon the modeling needs and the number of observations present in the sample.[16]

Sadly, one cannot escape the curse of dimensionality completely. The use of the neighborhood estimate $E(Y|(x_i - \delta, x_i + \delta)$ in place of $E(Y \mid x)$, for example, still becomes problematic as the parameter space increases. Imagine one has 10 independent variables each of which is selected from the Reals [0..1], and the desired granularity for each dimension requires that 10% of the domain of a single dimension make up the neighborhood. In 1 dimension, this means that $\delta = \pm 0.05$, which does not seem too difficult. But, in 10 dimensions, to achieve a hypercube that is equivalent to 10% of the total volume of the parameter space would require .794 of each individual dimension to make up such a neighborhood (i.e., δ would have to increase to $\pm .397$). As noted in Hastie, Tibshirani, and Friedman (2001, 22), "sampling density is proportional to $N^{1/p}$, where p is the dimension of the input [parameter] space and N is the [desired] sample size. Thus, if $N_1 = 100$ represents a dense sample for a single input problem, then $N_{10} = 100^{10}$ is the sample size required for the same sampling density with 10 inputs." Changing the size of the neighborhood can only help so much in grappling with combinatorics of this magnitude.

[16] Note that, for an OLS model, the neighborhood for each point x_i is the entire domain of x.

PARTIAL SOLUTIONS FROM THE NONPARAMETRICS LITERATURE

The preceding section has left us with significant problems. Ideally, we would like to be able to construct empirical models that avoid overfitting or underfitting the data. We also would like to sidestep the curse of dimensionality, without resort to extraordinarily limiting assumptions arrived at *a priori*. As argued in the first chapter, assumptions that allow us to "solve" a complex problem are always suspect, especially when they are not organic to the problem under consideration. One must admit that, in many contexts, choosing a particular distribution and functional form is inappropriate (or at the least, atheoretic). Fortunately, there are some things that one can do to arrive at a theoretically appropriate model that nonetheless avoids the problems noted here.

Feature Spaces

As one will quickly learn teaching a graduate research methods class, the modal graduate student sees enormous shortcomings in everything that has been published to date. This, obviously, is healthy, unless one subscribes to the notion that social science has discovered all that can be discovered – we will know we have achieved parity with the physical sciences when we produce our first Roger Penrose. Unfortunately, the most common fault students detect in existing research is that it is incomplete; that is, it is missing essential features or details of the "real-world" process under consideration.

This sort of idea is especially popular in "qualitative research" and the more historical schools within social science. Terms like "process tracing" along with citations to Clifford Geertz and thick description portray situations, which, from the perspective outlined above on the importance of the size of parameter spaces, seem completely impossible as objects of rigorous study. One example of this line of reasoning is exemplified in a thoughtful piece by Buthe (2002):

> The institutions within which actors interact are social constructs, as are the aggregate actors that populate so many of our models in political science. Due to factors such as uneven growth, increasing or diminishing marginal utility, and accumulation or ratcheting effects as well as the tendency of actors to attempt to manipulate or escape constraints, the passage of time makes it,

ceteris paribus, more likely that institutions, actors themselves, and their preferences may change. Recognizing this dynamic aspect of temporality does not mean that everything is constantly in flux. In fact, institutions and aggregate actors can be extremely stable for a long period of time. But the possibility of change implies that explanations of temporally large processes must allow for change in the constitution of actors as well as for change in their preferences. In the sense of the inherent dynamism of temporality, then, history is "the study of changes of things that change." Models of "history" must explain stability rather than assume it.... The dynamic quality of temporality suggests that models based on assumptions of stable institutional contexts, stable preferences, and constant units for which we record variable, independent attributes at any given point in time would be unsuited if we are concerned with explaining history, understood as a macroprocess. Endogenizing explanatory variables, however, comes at the expense of parsimony or worse: Scholars who seek causal explanations usually frown upon endogenization because when the dependent variable is not only explained by, but also (partly) explains the independent variables, we run the risk of circular reasoning. Can we avoid this problem? Sequence provides the answer.... Sequence allows us to endogenize the explanatory variables without having to abandon modeling and scientific aspirations because it enables us to avoid circular reasoning. Endogenization involves incorporating into the model some variation of causal feedback loops from the explanandum to the explanatory variables. (Buthe, 485; internal citations omitted)

The goal of this article is laudable – how to study historical processes scientifically – but the proposed solution is lamentable, given what we know about the effects of increasing the size of parameter spaces. If one actually believes that a process has actors, preferences, and institutions, none of which can be held constant, introducing time (here defined as sequence, with the possibility of feedback between different time periods) has the unsalutary effect of changing each observation into a statement that is conditional based upon time (i.e., $Pr(X)$ becomes $Pr(X \mid t)$). Simple bean counting of the magnitude of the parameter space, as we have engaged in throughout this chapter, would convince the prospective modeler that this is a bad idea.

Let us imagine, however, that we genuinely believe that time and the possibility of feedback are essential components of a given problem. Simply adding time will not do; the problem must first be simplified to a degree that an additional parameter will not result in a huge increase in the size of the parameter space. A better approach is to reduce the

size of the parameter space, such that the addition of a new variable (in this case, time), will not result in a sparsely populated space given the available data. In the artificial intelligence, nonparametric statistics, and related literatures this process is known as preprocessing or feature extraction. Feature extraction can be seen as a transformation of the raw data into a new set of inputs that dramatically reduces the size of the parameter or state space.

It is important to note that feature spaces are not unique; that is, any given input (parameter) space can be transformed in a large number of sensible ways. What is required to transform a given input space is either domain-specific knowledge[17] that allows the researcher to choose a theoretically justified feature space or a data-driven technique (e.g., principal component analysis) that produces an input space of reduced dimensionality and maximum variance on each of the remaining dimensions.[18] A ready example of the utility of feature spaces is provided by face recognition (a topic of much concern, given the world security situation).[19] One might think that better cameras would be useful in an automated system to recognize faces, but this is not obviously the case. Even a lousy camera with a pixel field that measures 100×100 with 8 bits for recording color would have a parameter space of $(2^8)^{10000}$ – allowing for an unimaginably large number of possible faces. Any number of observations would sparsely populate such a parameter space and increasing the resolution or color depth of the camera is the last thing one needs. Even if one argues, as Buthe does for the inclusion of time into historical models, that better cameras

[17] Readers may note that domain-specific knowledge is what qualitative researchers and historians are after. If qualitative researchers use their intimate knowledge of the details of a problem to derive better measures, features, and assumptions, so much the better. The argument here is against the idea that there is any free lunch in adding complexity to a causal argument – qualitative models are bound by the same laws as other models.

[18] For good treatments of the prevailing data-driven techniques, see Ballard (1997, Chapter 4) or Bishop (1995, Chapter 8). In additional to principal component analysis, one can rely upon a number of iterative techniques that sequentially discard inputs that fail to contribute to overall model performance.

[19] This example is taken from Ballard (1997, 85–6), who relies upon Turk and Pentland (1991). Turk and Pentland use eigenspaces to identify eigenvectors (features) that stratify different pixel-fields recording faces. The earlier discussion of theoretical feature selection owes much to Bishop (1995, section 1.3) and Russell and Norvig (1995, Chapter 4). In the social sciences, see Master (1988).

would provide details that inferior models miss (e.g., instead of a still shot, one could use a videoclip), without a feature space to reduce the complexity of the inputs, any proposed gain would be eclipsed by the curse of dimensionality.

To extract a useful feature space that would allow the available observations to span the reduced parameter space one might first adopt a data-driven method. The idea is simple: The population of faces follow rules that constrain the possible faces one could observe. The data never span the entire parameter space; rather, a lot of variance exists on a few dimensions but, otherwise, there are huge amounts of empty space. If one could find the reduced space that captures the systematic variance of human faces, one could safely do away with the larger parameter space that includes faces one would never observe in a human population (as opposed to a Martian or Plutonian one – on Earth, for example, we all have at most two eyes). A variety of techniques exist to do just this; a prominent example is principal component analysis. Fortunately, these techniques are familiar to social scientists, but it is important to keep in mind that these are linear transformations. It is possible that one can miss a great deal, even if "most" of the variance is retained in the reduced space.

Of more interest here is a second approach, which for lack of a better term I will call theory-driven feature extraction. Unlike game theoretic and statistical methods texts in the social sciences, artificial intelligence texts, which treat many of the same topics, always include a chapter on feature extraction. Russell and Norvig (1995) is one good example, and they make clear early on that most problems of interest are unsolvable without feature extraction/dimension reduction. In contrast to "uninformed search methods" (which is more or less what game theory equates with human rationality – but, see the next chapter), "informed search methods" are those "in which we see how information about the state [parameter] space can prevent algorithms from blundering about in the dark" (Russell and Norvig, 92). Solving complex problems without feature extraction, whether they are found in the context of a deductive or a statistical model is literally unthinkable within this tradition. Typically, researchers attempt to create feature spaces using domain-specific knowledge.

Using the example of face recognition, what would a theory-driven feature space look like? It would require that researchers deduce a set

of transforms to reduce the size of the input space into something much smaller using what we know about human features. Relevant measures might include the distance between the eyes, the size or shape of the nose, mouth, ears, and eyes, the shape of the hairline, and so on. The raw data would be used to form these summary measures and then discarded, resulting in an enormous reduction in the size of the parameter space. It is important to note that theory-driven feature extraction might not work perfectly. Given the size of the parameter space for problems such as face recognition, it is impossible to deductively show whether or not a given feature space will work. But, given the complexity of such problems, no other approach is possible.[20] Adding new variables or thickly describing complex phenomena is not a realistic alternative.

Out-of-sample Forecasting and Deriving Testable Implications

As the preceding section shows, one must strive to reduce the dimensionality of parameter spaces when one confronts complex problems; otherwise, one never has enough data to determine whether a model captures something essential about a problem or only some nonsystematic component of the sample. Determining whether a model "works" is difficult, however, and this difficulty is compounded by the fact that social scientists are loathe to compare models on out-of-sample performance. Rather, it is most often the case that the only results presented for a model are those for a fixed sample. A very persuasive monograph on this topic is provided by Granger (1999), where he notes that

> Cross-sectional and panel models are usually evaluated in-sample whereas time series models are also evaluated post-sample. To illustrate this difference, suppose one is interested in estimating the elasticity of demand for watermelons and has available some appropriate cross-sectional data set *I*. Two applied econometricians each build models, M_1 and M_2, using the same data and produce elasticity estimates e_1, e_2 ... If I ask a group of decision makers for advice, they are likely to expend a great deal of effort in comparing

[20] Explicit feature extraction also results in greater transparency. Rather than hiding such details through limiting assumptions – typically given short shrift in papers – one would have to present a case up front for the choice of a particular feature space. This of course assumes that researchers openly present their affirmative arguments for choosing a particular feature set.

the quality of the alternative models. Values of summary statistics such as R^2, likelihood values and model selection criteria, for instance BIC, can be compared and even tested for superiority of one model over the other. Doubtless the properties of the estimation techniques used in deriving the models will be discovered and compared, including (asymptotic) consistency and relative efficiency. Potential problems with the alternative estimating methods will be emphasized. It is possible to ask if M_1 encompasses M_2, or vice-versa, using alternative forms of encompassing. A single model can be viewed in terms of how well it performs under various "specification tests," against specific missing variables or linear trends or ARCH, general malaise such as non-linearity, the t-values of the included variables, can be discussed, and the coefficients of variables queried as to their economic meaning as well as asking if their signs are "correct" according to some particular naïve theory. However, all of this activity is aimed at discussing the (relative) quality of the models and ignores the quality of the outputs which I, as a decision maker, am most concerned about.

In contrast, consider a similar forecasting situation starting with a time series data set.... Although it is standard practice to pay some attention to the relative quality of the models, the majority of the evaluation effort is directed to comparing the quality of the forecasts, that is to the outputs. As a decision maker having to choose between two methods of forecasting, it is the quality of the output that is more important rather than the quality of the model.

Why this difference of approach? There seems to be two obvious distinctions. The first is the idea that a decision maker will be using the output of the model for some previously stated purpose and how well the models do in achieving this purpose provides a natural way of evaluating them.... A second idea, which I think is now widely accepted by time series econometricians, is that if M_1 produces better forecasts than M_2, it is unlikely that model M_2 will prove to be superior in other tasks such as testing theories or making control and policy statements. (Granger 1999, 63–5)

Note the close similarity between Granger's statement on the purpose of modeling and that of Friedman's discussed in Chapter 1. Models should be compared based upon their stated objective, and not on other grounds. Moreover, this stated objective should be as closely aligned as possible to the dependent variable used in the model, as loose analogies between a model and the empirical referent remain just that (loose).

Equally important is Granger's distinction between the quality of the *model* and the quality of a model's *output*. Discussions centering on model quality obviously consume the greatest share of methodological

debate within political science, much to our detriment. Instead of comparing ourselves based upon methodological allegiance (e.g., Bayesian vs. Frequentist), we might do better by comparing the results of our models using real or artificial out-of-sample performance. If novel data exist to decide between competing models, we should focus upon how models perform on these data. If we lack such data, we can divide our sample into a *training set* (i.e., data used to generate a model) and an artificial *test set* (i.e., a surrogate for novel out-of-sample data). One might even adopt a more complicated structure for model testing; that is, set aside 50% of a sample for training models, 25% for a validation set used to distinguish between different candidate models, and then a final 25% held independently to be used as a test set as well as to provide final results.

One might object that for some samples it is impossible to create an artificial test set to be used in generating out-of-sample results, because of a small sample size. In this case, one is pressing up against the curse of dimensionality, and there can be no great confidence in the results. Results, when predicated solely upon a fixed sample/training set, are woefully nonrobust, and usually present a rosy picture of model performance based upon overfitting the sample.[21] Thus, while out-of-sample forecasting, real or artificial, curbs many sins of modeling, it does present additional difficulties. As one increases the number of steps involved in modeling a given problem, it becomes more difficult for other researchers to replicate this process. Put another way, when one is left alone to train many candidate models on a training set, and then chooses among them by resorting to a validation set, and then presents results for a final, (hopefully) independent test set, other researchers require a great deal more information to replicate this process fully.

Out-of-sample forecasting, along with feature selection, offers a great deal of hope for addressing complex problems in the social sciences. But both avenues might inadvertently complicate the hoped for transparency of mathematical models, thereby requiring extra effort on the part of researchers for others to replicate their steps. The next section demonstrates how difficult this process can be, using a

[21] Though see Chapter 1, Section 4, and Chapter 5 on Logical Implications for an alternative route in testing models.

recent forum in the *American Political Science Review* as grist for the mill.

AN EXTENDED EXAMPLE: PREDICTING CONFLICT BETWEEN NATIONS[22]

During the 1990s, quantitative security studies became an increasingly prominent and sophisticated area of inquiry within our discipline. In particular, estimators based on the general linear model have been central to the development of extensive literatures on deterrence, the impact of democracy and trade on international conflict, and other issues. In 2000, BKZ offered a sweeping critique of these research programs and argued that nonparametric estimation was more appropriate for predicting the outbreak of war using the Correlates of War data on militarized interstate disputes. This viewpoint was challenged by de Marchi, Gelpi, and Grynaviski (2004), which elicited a response by BKZ (2004). In both their original article and the subsequent exchange, BKZ contend that standard parametric procedures underfit the data, missing systematic components best described as "highly non-linear, massively interactive, and heavily context dependent or contingent (22)." As noted earlier, underfitting the data is typically not a problem with most empirical models in the social sciences; in fact, given the overreliance on the sample (i.e., training set) and the qualitative features of the model rather than the output, it would take a leap of faith to imagine that underfitting had plagued a field for any appreciable amount of time. Too many researchers with too many free parameters (variables, model choices, etc.) have attacked the problem of predicting conflict.

BKZ, however, offer a partially convincing explanation for why underfitting might plague models of conflict. If, indeed, the outbreak of a militarized dispute is caused by highly nonlinear interactions between variables, termwise linear models or close analogs such as a logit specification might systematically miss these interactions, even if researchers designate a few modest nonlinear terms (e.g., by squaring the years a particular dyad had been at peace). Nonparametric techniques such

[22] This section borrows heavily from de Marchi, Gelpi, and Grynaviski (2004) and owes a huge debt to our continued conversations on these topics.

as the neural networks used by BKZ would thus pick up on these interactions in a way that more limited functional forms such as OLS could not.

As argued in the preceding section, nonparametric statistics offers a great deal of insight into how empirical research goes wrong and points to several interesting approaches to take in solving these problems. Moreover, the militarized interstate disputes data set offers what seems to be a great opportunity to utilize the approach detailed in section III of this chapter: a focus on out-of-sample forecasting. Given the number of observations in the disputes data set, one can easily divide the sample into several subsamples for training, validation, and testing. On the face of it, BKZ's approach seems reasonable, and their focus on out-of-sample results is noteworthy.

Unfortunately, a closer evaluation of the data and models relied upon by BKZ reveals several flaws in their approach, highlighting how difficult it can be to apply the principles outlined in this chapter without a great deal of care. As we will see, BKZ present results that are likely overfitted, and suffer from a lack of theorizing about what would be an appropriate feature space for predicting the outbreak of conflict. Additionally, BKZ forget one of the simple rules of research developed in Chapter 1: build models to test hypothesis, rather than engaging in atheoretic data mining.

Torturing Innocent Data Always Produces a Confession

The first place to begin is always the data. The data set used by BKZ records the initiation of militarized disputes within "politically relevant dyads" between 1947 and 1989 from the Correlates of War (COW) project.[23] The data include 23,529 dyad years; 976 of these years include a militarized dispute.

Dependent variable: **Dispute**, coded 1 for the presence of a conflict and 0 for peace. The threshold for the presence of a conflict is arbitrary–1,000 battle deaths, and is somewhat compounded by elaborate coding rules. One example is taken from the codebook for the COW 2 data: "One dispute, 3575 in MID 2.1, was removed from the MID 3.0 data

[23] Documentation on the project and the included variables may be found at http://cow2.la.psu.edu and http://www.umich.edu/~cowproj.

set. Papua New Guinea launched a raid against the Solomon Islands on March 12, 1992. Subsequently, Papua New Guinea apologized for the raid, said that it was not authorized and promised to pay compensation. That apology is sufficient for us to delete the dispute" (from the codebook at http://cow2.la.psu.edu). While there are enormous benefits for the scholarly community to agree on one particular set of coding rules, it does complicate matters for evaluating out-of-sample forecasts. The possibility that results are not robust under different codings of this variable is a distinct possibility, especially given the relatively rare nature of conflicts (they account for ~4% of the data). Applying the lessons of Chapter 1 on assumptions, one would not place much confidence in a model that provided varying levels of predictive capability based upon small changes in the threshold of battle deaths.[24]

What is more subtle, however, is to decide what it is that one is predicting when one uses this dependent variable. A naïve answer is that investigations using these data discover high-risk dyads; that is, those pairs of nation-states that are likely to go to war in a given year. This, however, is not quite sufficient for understanding the challenge presented by these data. Given the fact that the data are cross-sectionalized (i.e., time has been stripped away), what one is actually doing is discriminating between the observations in which a particular dyad chose to fight and the larger aggregation of observations in which the exact same dyad did not choose to fight – and all of this with sequence removed from the data!

A brief example will make the point. As part of a larger Arab-Israeli conflict, Egypt and Israel fought a war in 1967. Egypt and Israel, however, are in the data set for each year between 1947 and 1989, including some years when there was a conflict (e.g., 1973) and many more years when there was not. Sequence, obviously, is lost, given that the data are cross-sectionalized, so the real task is to distinguish between observations in which the dyad fought a conflict from observations in which the dyad did not. Under most investigations, Egypt and Israel are a

[24] Literally, current models including those of BKZ predict 1,000 or more battle deaths, and not conflict per se. It would be informative to see if one's results were consistent when noise is added to the above variable – for an example of checking for robustness by adding noise to a model, see Axelrod (1984).

high-risk dyad, but a model that "discovered" this fact would not be very informative, nor would it predict the outbreak of conflict with much accuracy in a particular year.

The task of predicting conflict out-of-sample turns out to be quite difficult and its importance is underappreciated in the literature. Simply determining that Egypt and Israel are a high-risk dyad results in a large number of false positives for the majority of observations that do not involve a conflict between the two nations. And, as we will see below, with one exception, none of the independent variables changes from observation to observation within a dyad, complicating this process significantly.[25]

Independent Variables

- **Contiguous landmass**, coded 1 if the nations are contiguous and 0 otherwise.
- **Distance**, coded as the actual distance between states. This has been shown to have a substantial impact on military conflict (Bremer 1992, Maoz and Russett 1993; Oneal and Russett 1999) and was derived from the EUGene program (Bennett and Stam 2000).
- **Similarity of alliance portfolios**, coded as a real from −1 to 1, where 1 indicates maximum similarity and −1 indicates dissimilarity. This variable measures whether or not each state in a dyad has similar relationships with other nation-states.

[25] Matters are worse in BKZ's chosen coding of the dependent variable. Their dependent variable includes multiple years of a conflict as observations – e.g., a dyad that participated in the Thirty Years' War would produce 30 observations coded as a war. The problem with this coding is that it violates the assumption of IID observations – and this is never trivial. The result is that BKZ's model (or any other using their dependent variable) only produces true positives (i.e., correct predictions of war at any reasonable threshold) when the Peace Years variable equals 0 (note that this is true even if one drops the classification threshold to very small thresholds – for example, .25, where one finds a huge number of false positives). Put another way, the presence of multiple years of a conflict as "independent" observations means that a flexible technique like a neural network can leverage the fact that if you fought in year t, you are likely to continue fighting in year t + 1. One should thus take BKZ's results in this chapter with an additional grain of salt. Their model does not predict war; rather, it says one should predict war whenever the dyad fought in the preceding year.

- **Alliance status**, coded 1 if the two nations share a treaty and 0 otherwise.
- **Asymmetry of Military Capabilities**, coded as a real from 0 to 1 where 1 represents imbalance and 0 represents parity in the dyad. Most work (e.g., Oneal and Russett 1999) has hypothesized that the relationship between military capabilities is curvilinear. The square of the asymmetry value is used in order to account for this relationship.
- **Major Power Status**, coded 1 if one of the dyad members is a major power and 0 otherwise. It is well established that major power states are much more likely to engage in military conflict. (Bremer 1992; Maoz and Russett 1993; Oneal and Russett 1999).
- **Democracy**, coded from the Polity III data set, and ranging from −10 (autocracy) to +10 (democracy); typically, one adds +11 to each variable (to eliminate negative numbers) and multiplies them to create a summary interaction of the joint level of democracy in the dyad (Bueno de Mesquita and Lalman 1992; Maoz and Russett 1993; Rousseau, Gelpi, Reiter, and Huth 1996). Additionally, a number of scholars suggest that the impact of democracy on conflict may be curvilinear (Snyder 1991; Mansfield and Snyder 1995; Goemans 2000), so the square of the joint variable is also included.
- **Peace Years**, coded as an integer from 0 to N, where 0 indicates the dyad fought in the previous year and N is the number of years in the data set. This variable is coded oddly, insofar as all states start out at either a 0 (indicating a conflict in 1946) or a 1 (indicating no conflict in 1946 and any number of years before that).

What should be noted about the above independent variables is that they *preclude* the data generating process stipulated by BKZ. Although one might imagine that war is a complex function of nonlinear interactions between variables (i.e., BKZ may be right), the data must allow one to test such a notion. Keep in mind that the real difficulty in this data set is to distinguish observations in which disputes occurred from the much more numerous observations in which peace occurred within each dyad. Merely identifying high-risk dyads is a recipe for large numbers of false positives and little predictive power. Thus, if one's hypothesis is that war is the complex accumulation of factors, the data must support a direct test of this (admittedly nebulous) hypothesis.

Unfortunately, it seems obvious that the above independent variables are poorly suited to the task. This is largely because all but one of these variables do not change significantly from observation to observation within a dyad. For Egypt and Israel, as one example, Contiguity, Distance, Similarity of Alliance Portfolios, Major Power Status, and so on do not change very much between observations. If these variables jointly indicate that conflict is likely, then they must do so for all observations in the data set featuring these two countries. If they fail to indicate that conflict is likely, then peace will be predicted instead for all observations for Egypt and Israel. This represents an enormous problem, and it is mitigated only partially by the inclusion of Peace Years as a variable. As stands to reason, Peace Years *does* change from observation to observation, and not surprisingly, it accounts for most of the performance of all models relying upon these data (alone, through functional transformations, and through modest interactions). Given the coding of this variable, it is necessarily a blunt instrument, but it is all one has in this data set.[26]

Thus, while BKZ might hope to model war with more complicated nonparametric models designed to capture massive nonlinearities and interactions between different independent variables, the data simply do not support such a venture. One should not be surprised to find that a simple logit model using only Peace Years along with splines of this variable yields a baseline for out-of-sample performance that is equivalent to BKZ's original model from 2000, and less than 2% worse than BKZ's subsequent effort in 2004 using a committee of neural networks and the entire complement of independent variables. Moreover, BKZ's 2004 effort represented their third effort at modeling the exact same out-of-sample test set (the data post-1985), which by any notion of forecasting is a violation of the spirit of the enterprise.

BKZ, despite repeated attempts to model conflict, have very little hope of improving upon previous efforts given the limitations of the data on militarized interstate disputes. Complicating their efforts

[26] A more sensible coding for Peace Years would be to count backward in time to arrive at real starting values for 1947. That is, instead of coding the United States and Britain as a 1 (indicating that they had fought a war two years earlier), one could code it as 125. If this is too arduous, use the average of Peace Years across the range of 1947–1989 as the starting value, and count from there. Otherwise, one confuses early years with years that genuinely saw violence.

are problems with the idiosyncratic coding of both the dependent variable (which depends upon an arbitrary number of battle deaths) and the most salient independent variable (peace years, which also is coded idiosyncratically given the arbitrary starting values – see the description of this variable earlier). Although there are many useful lessons in their approach to this long-standing problem in security studies (e.g., a focus on out-of-sample forecasting to compare models), this does not mean that one allows the data to speak without any intervention by theory. And to the extend their model does improve upon far simpler models, one has to question whether the improvement of ~2% represents genuine progress or overfitting.[27] Overfitting, even in what is supposed to be an out-of-sample experiment, is much exacerbated by repeated efforts against a fixed and known out-of-sample set. Arbitrary coding rules compounds the problem by calling into question whether one is explaining systematic components of the DGP or simply a small set of observations right at the threshold of 1,000 battle deaths (or the initial values of Peace Years, or any number of other idiosyncratic factors).

Is There a Theory in the House? Not Without a Feature Space . . .

In large part, BKZ's repeated attempts to model conflict with such paltry rewards is a byproduct of the fact that their theory is not only vague but unsupported by even rudimentary consideration of the data at hand. As noted in Chapter 1, empirical models serve to test theories. Even if BKZ had reaped greater rewards from their neural networks, it is not at all obvious what these solely empirical exercises would tell us.

[27] The difference between the best model of BKZ, arrived at after numerous attempts at predicting the same out-of-sample set, is within any reasonable confidence interval of our logit model for summary statistics such as the area under an ROC curve – see Figure 2.4. One should note that the results for BKZ in this table are taken from their published papers – we have subsequently discovered that they normalized their data incorrectly and these results are wrong. To help their optimization algorithm, they took the sensible step of normalizing each input variable to a mean of 0 and a variance of 1. Unfortunately, they normalized the training and predictive sets together – that is, they used a global mean and global standard deviation. The correct procedure is to normalize the training and test sets independently, else one is peeking into the future. BKZ's published results benefited substantially from this mistake.

One might defend the practice of using an atheoretically derived empirical model, especially given their reliance (however imperfect) upon out-of-sample forecasting. There is, however, a problem that is revealed by bean counting. Given the independent variables in the previous section, one can come up with a rough idea of the size of the parameter space by dividing the real-valued independent variables into bins. To accomplish a back-of-the-envelope calculation, I use Stata's formula for histograms:

$k = \min(\text{sqrt}(N), 10^* \ln(N)/\ln(10))$

where N is the (weighted) number of observations.

This results in

$$2 \cdot 43 \cdot 43 \cdot 2 \cdot 43 \cdot 43 \cdot 2 \cdot 43 \cdot 44 = 50{,}570{,}904{,}392$$

possible parameterizations for the models of conflict used by BKZ. If one adds the five splines for peace years and squared terms for joint democracy and asymmetry, one gets

$$= 50{,}570{,}904{,}392 \cdot 43^7$$

total parameterizations. Lest one forget, this already enormous space is further expanded by the choices made in the modeling process. In an OLS or MLE model, the researcher has comparatively few choices to make. But, in a neural network such as the one utilized by BKZ, researchers make an incredible number of choices related to the following features of the model:

i. random number generation and the distribution of seeds for the optimization procedure;
ii. the composition of training/validation/test datasets;
iii. smoothing or penalizing results to account for model complexity;
iv. the number of hidden neurons and the number of layers in the network;
v. the target evaluation function (e.g., maximizing area under the receiver-operating characteristic [ROC] curve);
vi. the type of committee system for producing predictive values.

Including these choices as parameters expands the parameter space even more, and allows researchers to do great damage to the data,

usually through overfitting.[28] Thus, 23,529 observations seems like a large number only if one ignores the vast parameter space BKZ ask these observations to span.

In more restricted models such as OLS or logit, one can do an end-run around such data problems if one is testing a theory arrived at before statistical work. In this case, a well-developed theory is buttressed by empirical findings that demonstrate the main tenets of the theory hold, even in an impossibly large parameter space. Think of this as a betting person would – even though the parameter space is large and the data populate a fraction of the total space, finding that the main tenets of the theory hold would increase one's priors about the usefulness/correctness of the overall theory. Simply stating a theory *prior* to looking at the data and then testing it increases the likelihood that one is "right" – inventing a theory *after* finding a statistical model cannot be counted as evidence supporting the theory.

BKZ thus have a problem because that they are not testing a theory, nor even presenting a specific statement about the DGP. Feed forward, multilayer neural networks such as the one relied upon by BKZ have the virtue that they are universal function estimators, but this is offset by the "black box" nature of interpreting these models.[29] Recovering the exact functional form for a particular neural network is impossible. With the huge size of the parameter space, tricks that work for simple models like logit – taking the first derivative for each independent variable and setting the other variables to their medians – fail miserably in a neural network model. The possibility of dramatically nonlinear and interactive functional forms means that one cannot look at a first derivative and recover anything meaningful, as the slope for a given independent variable at a particular point does not provide anything but an estimate for the variable in a narrow neighborhood. Literally anything can happen as one allows the values of the other variables to stray from their fixed, arbitrary values.

[28] Defining choices of this sort as parameters is a slight abuse of terminology, but is defensible given that all results are predicated upon these choices.

[29] For a general overview of the black box problem and an interesting mapping between feed forward, multilayer neural networks, and fuzzy rule-based systems (i.e., systems based upon fuzzy logic) see Benitez, Castro, and Requena (1997). Such approaches are obviously a bit far afield from mainline econometrics/statistical methods, and impose significant costs on researchers hoping to interpret the role of the independent variables in a model.

To simplify the coincident problems of an intractable parameter space and hard to interpret independent variables, one should follow the advice of section III of this chapter and derive a theoretically justified feature space.[30] Given the computational costs of estimating a neural network, it is difficult to simply try all combinations and transformations of the full set of possible explanatory variables and modeling choices. Without the "correct" set of explanatory variables, modeling choices, and a theoretically justified feature space, one cannot place much confidence in the output of a neural network or any other nonparametric technique. Studies of the performance of neural networks in the related literature of macroeconomics reveals a quite mixed set of outcomes (Gonzalez 2000), reinforcing the need for theoretical work prior to estimating a neural network. Even though neural networks should in theory encompass and outperform a simple logit model, the results of actual applications in the macroeconomics literature suggest that these models often overfit the data, due in large part to the enormous size of the parameter spaces being tested relative to the number of observations in the data.

Feature sets, as noted in section III, make everything easier, increasing one's confidence that the results of a nonparametric model are in response to systematic components of the DGP as well as allowing for easier interpretation of results. Without feature sets, one is searching in a very large space with a tiny flashlight, and the odds that you are fitting your model to nonsystematic aspects of the sample are heightened (see footnotes 25 and 27). Given the task at hand is predicting conflict, any reasonable feature set would focus upon factors that would discriminate the small proportion of years when a high-risk dyad would engage in a conflict. One would think that the dimensionality of many of the independent variables (e.g., distance,

[30] BKZ (2004) seem to confuse normalization with preprocessing/feature extraction. They cite Bishop (1995, Chapter 8) on feature extraction, but, in reality, they merely rescale their existing variables to a mean of 0 and unit standard deviation. If the independent variables have dramatically different scales, normalization can be useful insofar as it simplifies the task of choosing starting values for the optimization procedure used by the nonparametric estimation technique (usually some variant on hill-climbing). But in this data set, few variables are not on a similar scale, typically binary or [0..1]. Only distance has an extraordinary range (integers from 0 to ~12,000), but the log of this is used. The argument here is that BKZ should have engaged in the sort of feature extraction covered in Bishop necessary to reduce the pathologically large parameter space of their model; rescaling does not accomplish this task.

similarity of alliance portfolios, democracy) could be dramatically reduced.[31]

How Do You Know When You Are Right? Evaluating Models

Imagine that BKZ had avoided most of the mistakes listed earlier. How would we know whether their model was better than previous efforts? Almost every mathematical methods book (at least those in the empirical tradition) have a section on model selection. Typically, these sections offer vague advice, because in the most general case, this is a difficult question. Granger (2000), however, offers extraordinarily lucid advice. As noted previously, his focus is on comparing the quality of outputs, rather than features of the model itself. He notes that many models are "the result of considerable specification searching . . . data mining, or data snooping in which data are used several times" and suffer from the curse of dimensionality. "Unfortunate experience," Granger argues, has led time series econometricians to focus on out-of-sample testing.

Comparing models in this way is easier said than done. First, all parties must agree on a reasonable loss function; typically, mean squared error is used for regression problems, but there are other error functions that are more appropriate for dependent variables that are non-Gaussian.[32] Second, one has to agree on a standard for comparing different models. For mean squared error as the error function, it may seem natural to compare models based upon the ratio of the error

[31] Although a pessimist would conclude that most of the independent variables already contained within the COW data are not suitable for discriminating between conflicts and peaceful observations. Changes in economic conditions, demographics, and so on seem far more fruitful than the existing variables, and Peace Years is likely a blunt proxy for these processes. An inability to define a reasonable feature space that corresponds to a well-defined model of conflict is probably a sign that the data are not adequate to the task of testing a model. Yet, even with data of this quality, it is still better to define a feature space, rather than let your statistics package do it for you (atheoretically). Inevitably, you will find that you rarely produce stable models without a feature space.

[32] For example, one can use a Minkowski-R error function to minimize the influence of outliers in fat-tailed distributions (by setting the parameter $R < 2$). See Bishop (1995), Chapter 6, for a selection of error functions. Also note the arguments presented earlier in the chapter for smoothing, which complicates error functions by adding a term to account for model complexity.

functions using an F-test, but Granger notes that "there are good reasons for expecting that e_{1t}, e_{2t} [the errors] will be highly correlated" thereby decreasing the value of the F-test (Granger 2000, 69). Alternately, one could generate a linear regression comprised of two competing models M_1 and M_2:

$$x_{test} = \alpha + \beta_1 M_1 + \beta_2 M_2 + e_{test}, \quad \text{where } e \; \hbar \; N(0, \sigma^2)$$

If β_1 is significant and β_2 is not, one can conclude that M_1 dominates M_2; the converse also holds (Granger 2000, 69–70). Or, it may be the case that neither model dominates, but a committee of models (in this case, the number of members = 2) performs better than any individual model.

Fortunately, the special case in which one has a *binary* dependent variable is considerably simpler than the general case, though a bit of explanation is required to make this point. Once again, I will use the work by BKZ and de Marchi, Gelpi, and Grynaviski on predicting militarized interstate disputes to examine model comparison with a binary dependent variable.[33]

As BKZ (2000, 21) correctly note, out-of-sample results indicate whether a model reflects the "true" causal process driving the phenomena of interest and guards us against "taking advantage of some idiosyncratic feature of the data." Predictive success against a binary dependent variable, however, should never be judged on the basis against an arbitrary 0.5 probability threshold or a single classification table. BKZ, by relying upon such a poor standard for adjudicating out-of-sample performance, diminish their contribution to the literature.

In general, the use of any arbitrary cutoff point to discriminate between "peace" and "war" or "success" and "failure" in classification tasks is risky, and may simply be inappropriate (Greene 1997, 892–3; King and Zeng 2001, 11–13; Swets 1988, 1285–93). Theoretical work on international conflicts provides us with an additional worry in choosing such a threshold. Statistical models provide the predicted probability of a conflict, *but this probability may be low in all cases*. As noted in Greene, "0.5, although the usual choice may not be a very good value to use for the threshold. If the sample is unbalanced – that

[33] This subsection follows the "Evaluation of Dichotomous Forecasts" section of de Marchi, Gelpi, and Grynaviski (2004).

is, has many more 1's than 0's, or vice versa – then by this prediction rule, it might never predict a 1 (or 0) . . . The obvious adjustment is to reduce [the threshold]" (892).[34]

Even wars that ultimately *do* occur may have been generated by circumstances where the *ex ante* probability of war was less than 0.5 (Fearon 1995). For example, if we view war as "off-equilibrium" behavior (Gartzke 1999), then the precise timing of the outbreak of military conflict may result from some combination of idiosyncratic events. In this case, any attempt to build systematic statistical models that generate high *ex ante* probabilities of military disputes will inevitably become an exercise in overfitting a particular data set, especially given the nature of coding rules such as those used for the dependent variable (i.e., 1,000 battle deaths constitute a dispute) and the key independent variable Peace Years (i.e., the arbitrary starting values for all dyads).

A better alternative would be to use the criterion developed in de Marchi, Gelpi, and Grynavisky: examine the trade-offs between false-positives and false-negatives for a variety of predictive thresholds, and do not penalize a model predisposed to predictions biased too high or too low. One way to look at different thresholds would be to generate a huge number of classification tables. A better solution, however, is to use ROC curves. ROC curves are diagnostics that are able to cope with the trade-offs between false positives and false negatives in model assessment (Swets 1988). These curves plot the proportion of conflicts correctly predicted on the x-axis and the proportion of nonconflicts correctly predicted on the y-axis. The intuition behind the graph is that any threshold used as the cutoff between a conflict or peace prediction will correspond to a single point on this curve. The area below a single point on the curve corresponds to the proportion of true negatives for that cutoff, while the area above the point indicates the proportion of false positives. Similarly, the area to the left of a point corresponds to the proportion of true positives, while the area to the right of the point represents the proportion of false negatives. For example, if the cutoff is zero, then all disputes (but no cases of peace) are predicted correctly. Finally, as the cutoff varies over the range between zero and one, the

[34] See Morrow (1989) for an early attempt to address this problem with international conflict data.

curve will be negatively sloped, as fewer conflicts and greater numbers of peaceful dyads are forecast correctly.

The key point to glean from a pair of ROC curves used for model comparison is that the curve with more area underneath it corresponds to a greater proportion of successful predictions, regardless of what arbitrary threshold is settled upon for predicting the dependent variable. In the absence of a specified optimal threshold based upon decision-theoretic criteria (see Granger 2000, Chapter 3, or de Finetti 1974), the area under an ROC curve provides a useful summary statistic that can arbitrate between competing models.

As an example of how necessary ROC curves are, consider the model presented in BKZ (2000). They discover that neural networks predict wars with a probability greater than 0.5, whereas prior logit models do not. One might conclude that this demonstrates the superiority of the particular brand of nonparametric techniques used by BKZ at the expense of simpler logit models.[35] The question then is how does one compare models of conflict to make this determination? For this example, I present several models, ordered from the least complex to the most complex:

- a very simple linear discriminant in which each class is assumed to have an equivalent covariance matrix;
- a logit model derived from the existing security studies literature;
- (tie) a feed forward neural network presented in BKZ (2000) using an incorrect error function based upon the number of correct classifications at the 0.5 threshold;
- (tie) a neural network from de Marchi, Gelpi, and Grynaviski (2004) using the correct error function of area under the ROC curve;
- BKZ's 2004 committee of three neural networks estimated using the correct error function of area under the ROC curve.

Results are presented for two different test sets. One of these is the test set originally reported in BKZ (2000), consisting of all dyads in the years after 1985. A second test set was created by drawing a

[35] BKZ correct this mistake in their 2004 publication. As noted in this section, the choice of error function is crucial. BKZ in 2000 did not use MSE; rather, they used a weighted function that rewarded true positives and true negatives. Maximizing the area under the ROC curve is the correct error function for this problem, however.

Model \ Forecast Set	Uniform Draws (Pre-1985) & selected 95% confidence intervals	Post-1985 & selected 95% confidence intervals
Peace Years (+ splines)	0.805	0.904
Linear discriminant	0.815	0.872
Logit	0.837 [0.77 ... 0.9]	0.915 [0.89 ... 0.94]
Single neural net - classification	0.801	0.87
Single neural net - ROC area	0.856 [0.77 ... 0.91]	0.869 [0.79 ... 0.94]
Committee of neural nets	0.871 [0.82 ... 0.92]	0.927 [0.91 ... 0.95]

Figure 2.4. Area Under ROC Curves

5% uniform random sample of the dyad-years from 1947 to 1985. The latter test set was used to test for robustness, and serves as a useful tool to determine whether the particular cutoff of 1986 might be fortuitous, aiding or hurting different models. Second, I do not present results from the training set; to do so may artificially inflate the models' apparent performances and deflect attention from their out-of-sample performance.

Note that these models range from the simple linear discriminant to a very complex committee comprised of three different neural networks. Given the increase in the parameter space intrinsic to complex models, and the difficulty in understanding the moving parts in a complex model, it is always useful to see whether or not the complexity of a model is warranted by the data at hand. When in doubt, one should follow Friedman and Granger's advice and choose the simplest "best" model.

How do these very different models perform? Figure 2.4 reports the area under the ROC curves for all the models detailed above and Figure 2.5 plots the ROC curve for selected models. As Figure 2.4 indicates, the committee of neural networks that maximized the area under the ROC curve in the training set outperformed all of the other models in this forecast set for both the uniform draws test set and the post-1985 test set.

Given the rarity of militarized disputes, one should not dismiss even a modest increase in forecasting accuracy, but one cannot avoid the impression that the difference between all of these models is quite small. Figure 2.4 reports that the committee of neural networks had the greatest area under its ROC curve at 0.9271, while the logistic regression had the second greatest at 0.9152. Figure 2.5 demonstrates

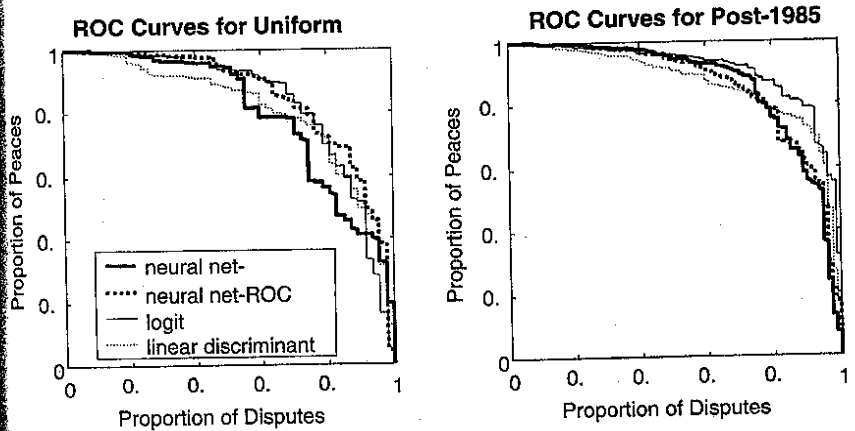

Figure 2.5. ROC Curves for Neural Net (classification), Neural Net (ROC) Logit, and Linear Discriminant Across Uniform Draw and Post-1985 Test Sets

that for practically any given tolerance of false positives, the logit model was nearly indistinguishable from its rivals. The overwhelming impression left by this set of results is that there is little difference between the various neural networks, logit and discriminant analysis in terms of their ability to distinguish between dispute and nondispute cases, once one controls for a model's inherent bias toward predictive probabilities that are either too high or too low.

But, as with many methodological ventures, such impressions may be wrong. How could we explicitly test whether or not one model outperforms another based upon the results of the ROC curves? Fortunately, there is an easy answer to this question. One can compare different ROC curves by using a chi-square test, or a number of related techniques such as the Kolmogorov-Smirnov test (see Scott and Fasli 2001) or the algorithm developed by DeLong, DeLong, and Clarke-Pearson (1988).[36] All of these approaches make no assumptions about the distributions in question, and are appropriate to comparing ROC curves. As one might guess from inspecting Figures 2.4 and 2.5, the differences between the competing models are in fact not significant (not even at the P = 0.20 level). Given the limitations in the data and

[36] See the Stata Reference Manual, Rel. 8 (2003). The latest version of Stata has implemented very robust code for computing ROC curves as well as comparing different models with either correlated or independent ROC curves.

the overall level of performance of all the models (which is quite high), this should not be seen as surprising.

As noted in Chapter 1, a better approach by far would be to use a genuinely new out-of-sample set, such as the recently released 1990–2001 data, in which the differences between the models would likely be far greater. Repeated efforts to achieve marginally better performance on the 1986–1989 data, by contrast, runs counter to the entire predictive enterprise.

A GENERAL STATEMENT ON MODELS

As noted in the first section, the problems discussed in this chapter are by no means limited to empirical modeling. Since the next chapter will concern itself with game theoretic and computational models, I will limit myself to three brief remarks.

First, just as one would argue that empirical models can overfit the data, so, too, can deductive models "overfit" a desired end result. With a combination of limiting assumptions, information restrictions, and equilibrium concepts, one can achieve any desired result with a formal or game theoretic model. The fact that one can prove something to be true is not of itself useful; there are an infinite number of models that prove any given result is true. As with empirical modeling, the task is to discriminate between competing models. As I argued in the first chapter of this book, one must avoid what Friedman defines as a "retreat into purely formal or tautological analysis."

Second, one must bean count for analytic models in much the same fashion as the preceding sections have bean counted the various parameters involved in formulating empirical models. Parameter spaces are parameter spaces, and to the extent that one is making choices for a deductive model, the stability of the result depends upon both how many choices are made as well as the nature of these choices.[37] A reasonable question to ask is how robust one's deductive results are given perturbations in any of these assumptions. And, if one believes in the epistemological arguments from Chapter 1, we should be particularly worried about technical assumptions or assumptions that seem unrelated to the empirical referent.

[37] See the section in Chapter 1 on conditions that should be placed upon assumptions.

There is, perhaps, a more subtle point to be made here. Unlike empirical models, deductive models do not depend upon stochastic components in producing results; that is, assumptions are not noisy in the way that data are. If you have a set of rotten assumptions in an empirical model, confronting that model with the data will immediately reveal this fact. For a deductive model, however, one might arrive unknowingly at a set of knife-edge values for what seem to be trivial or purely technical assumptions but account entirely for the result. Perturbing any of these assumptions may produce drastically different outcomes for the model.

As I will argue in Chapter 3, this is a real problem, especially if one adopts the normal science notion that models should build upon previous research (and previous models). Fragile models make this goal difficult. What one requires is a theory of *equivalence classes* for formal or game theoretic models. Much of the remainder of the book attempts to build such equivalence classes, so I will say no more about it here.

Third, one has to test deductive models. I believe most researchers would accept that the work of BKZ is incomplete given the fact that there is no real model being tested. So, too, must one be wary about deductive models that are "proven" with case studies or qualitative anecdotes. The exact same standards should apply to deductive work: When the parameter space of a deductive model is large, one needs a plentiful supply of out-of-sample data or a set of novel logical implications to test the deductive model. To the extent that the data do not span the deductive model's parameter space, we have less confidence in the utility of the deductive model. Granger's admonition that the consideration of empirical models should focus on the outputs and not characteristics of the modeling process also applies to deductive models. The attractiveness of one equilibrium concept over another, or whether the result was generated with one set of assumptions over another, is only so useful.[38] One should instead focus on how well models do when confronted with their empirical referents.

[38] In large part, it seems the conflation of "rationality" with game theory has diverted attention from the theories generated by game theoretic models. The rational choice debate, in Granger's terms, concerns features of the modeling process, not the quality of the results.

A final point concerns these outputs. It is often the case that the mapping from a deductive model's results to an empirical referent is attenuated. To the extent that a deductive model has only loose application to an empirical referent, we should ignore it, as no real testing can proceed. As noted in the first chapter, one obvious example is the iterated prisoner's dilemma. What, exactly, are the predictions of this class of models given the plethora of equilibria one can generate? That people in some settings cooperate?

APPENDIX: PERL CODE FOR THE CURRENCY GAME

```
1. use Math::BigFloat;
2
3 # parameters
4
5 $loops=100;          # number of times program is repeated
6 $N=10;               # population size
7 $ep=0.50;       # mutation rate
8 $time=100;           # length of horizon
9 $early=$N/2;         # number of early adopters/those changing
  each time period; must be <= $N
10
11 # do file I/O; in this case, output only
12
13 open DATA_DUMP, ">>dump.txt";
14 select DATA_DUMP;
15
16 # print header for data 1x
17 # print "NUMBER of regime shifts | OVERALL MEAN of
   populations | Delta |
   N | Mutatation Rate | Time \n";
18
```



```
19 # repeat program $loops times; output at the end of the
   program body
20
21 for ($z=0; $z<$loops; $z++)
22 {
23
24 # initialize population - uniform draws
25 $pop_mean=0;              # population mean
26 $current_standard=0;      # either 0 or 1 based upon pop mean
27 $regime_shifts=0;         # number of regime shifts
28 $pop_total=0;             # keeps total population values for
   use later in determining mean
29
30
31 for ($i=0; $i<$N; $i++)
32       {
33       if (rand()<0.5){
34                   $agents[$i]=0;
35                   }
36                   else {$agents[$i]=1; $pop_mean++;}
37       }
38
39 if ($pop_mean>$N/2) {$current_standard=1;} else
   {$current_standard=0;}
40
41 # start main loop; pick one agent at random and allow
   change
42
43 for ($i=0; $i<$time; $i++)
44       {
```

```
45      # set temp_regime to existing standard for later
        comparison at loop end
46      $temp_regime=$current_standard;
47
48      for ($e=0; $e<$early; $e++) # start loop of switchers
49          {
50          # draw the agent who considers change
51          $temp=int(rand($N));
52          $val=$agents[$temp];
53
54          # draw the epsilon
55          if (rand()<$ep) {$agents[$temp] = sprintf
            ("%.0f", rand ()); }
56          # otherwise, act like a sheep
57          else    {if ($agents[$temp]!=$current_standard)
58                  {$agents[$temp]=$current_standard;}
59          }
60
61          #check for change, update $pop_mean and $current_
            standard
62          if ($agents[$temp]>$val) {$pop_mean++;}else
63                  {if ($agents[$temp]<$val) {$pop_mean--;}}
64          if ($pop_mean>$N/2) {$current_standard=1;}else
    {$current_standard=0;}
65  }       }# end $early loop for all switchers
66
67          # save pop_total and print out values of interest
68          $pop_total=$pop_total+$pop_mean;
69
```



```
70          # print "Pop_mean: ", $pop_mean, "Standard: ",
    $current_standard, "\n";
71
72          # check to see if a regime shift occurs
73          if ($temp_regime!=$current_standard) {$regime_
            shifts++; }
74
75          }
76
77  # determine overall pop mean
78  $pop_total=$pop_total / $time;
79
80  # print data
81  print $regime_shifts, " ", Math::BigFloat->bceil
    ($pop_total), " ", abs(Math::BigFloat->bceil($pop_total)
    -$N/2), " ", $N, " ", $ep,," ", $time, " \n";
82
83      }       # end $loops
84
85  close DATA_DUMP;      # close output file
```

3

From Curses to Complexity

The Justification for Computational Modeling

INTRODUCTION

The study of international conflict, like many fields, has hosted a long-standing debate between rational choice theory and its critics.[1] Rational choice, in this debate, is in some sense a misnomer, as the critics almost solely reference game theoretic models in which the nation-state is the primary actor. The overall question, however, raised by critics of game theory is whether or not three decades of formal modeling has helped us to understand complicated phenomena such as the causes of war and alliance formation.

In this chapter, I will review the main positions taken in this debate, using some of the more prominent game theoretic models as illustrations. It will become obvious that I am critical of the efforts of game theory to date, despite my support for the idea that one must engage in modeling as an enterprise to understand complicated phenomena. My focus, however, is not solely on the limitations of existing models nor is it on any particular substantive area (although problems in security studies will be used as examples); rather, it is to investigate these models with the goal of determining which complications systematically hamper their effectiveness when applied to difficult problems. Obviously, game theory has been tremendously effective in solving other problems, so one must wonder where the stumbling blocks lie in

[1] See, in particular, Walt (1999a) which launched the thousand ships, and Walt (1999b), which is a response to various defenses of game theory in a special issue of *International Security* (Vol. 24:2, Fall 1999). The most salient defenses were written by Bueno de Mequita and Morrow (1999), Niou and Ordeshook (1999), and Powell (1999).

complex problem areas like security studies and what the implications are for game theory more generally.[2]

As we have seen in the previous chapter, there are two problems that complicate empirical research. First, the curse of dimensionality hampers one's ability to test the results of complex models. Second, one needs to avoid brittle models, in which seemingly trivial changes to a model's assumptions cause huge swings in the results. This chapter demonstrates that these problems affect game theoretic models in exactly the same fashion as empirical models. Fortunately, the solutions to these problems are much the same for game theory as they were for empirical models. We will see that one must develop feature spaces that encode problems in a parsimonious fashion. Further, without the development of equivalence classes, it is difficult to engage in cumulative research.

I will thus present an approach that borrows from both game theory and computational political economy, and, as such, will likely cause some discomfort to both methodological "camps." To illustrate the promise of a combined approach, I use two examples that serve as illustrations of how to apply the proposed methodology. The first example (beginning in this chapter and continuing to the next) concerns the problem of alliance formation. The second example (in Chapter 5) examines nonseparable preferences and survey data.

A DETOUR: A BRIEF CRITIQUE OF GAME THEORY (AND SUNDRY COMMENTS ON MACHINE CHESS)

In the debate between formal modelers and their naysayers, modeling complex processes such as war is most often framed as a dichotomous choice between game theoretic models, on the one hand, and qualitative/historical work, on the other.[3] It is worth repeating that

[2] One example involving a complicated auction would be Binmore and Klemperer (2002). In April 2000, they organized the license auction of Britain's third-generation mobile phones. The total revenue of the auction was equivalent to 2.5% of Britain's GNP. There are, of course, other examples of the utility of game theory.

[3] As argued in the first chapter of this book, there is a third approach, which is to adopt Friedman's (1953) maxim that models and their assumptions are not worth arguing over; rather, one should be ecumenical with regard to models and base comparisons solely upon their out-of-sample performance. Sadly, most social scientists

game theory has been widely accepted in large part due to analytic simplicity and broad applicability. Most would believe that game theory deserves the name – all games are potentially representable within the confines of game theory.

Thus, if game theory is to work as advertised, we have to have some way to encode most of the games we are interested in studying; or, more in the vein of mathematics proper, build upon previous efforts to derive increasingly complicated models that achieve more verisimilitude with the problem at hand. It is worth summarizing what is required by game theory to encode a given problem. To accomplish this task, game theory requires three moving parts. First, one needs an instantiation for the problems that humans confront, and ideally this encoding represents what players know at different points in the game. In game theory, this instantiation is most often an extensive form game, where the innovation of information sets provides a nice vehicle for understanding how knowledge impacts play. Second, one needs explicit utility functions that represent how players evaluate the outcomes of the game. Last, one needs a solution concept and an algorithm that "solves" a given problem; game theory typically utilizes a Nash equilibrium (or a refinement) as the solution concept and backwards induction as the algorithm. Given an extensive form and a utility function, one can readily apply backwards induction and test for the existence of a Nash equilibrium.

Although I am working toward a critique, I would like to distinguish the current work from the approach taken by most critics of game theory (and rational choice more generally). There are two main streams of criticism. The first has focused upon whether humans can actually frame problems correctly (i.e., can we satisfy information requirements?) or apply the appropriate solution concept (i.e., can we satisfy computational requirements?). The consensus of this line of research is that humans are quite stupid compared to the ideal rational choice player, even if the humans in question are heads of state.

The problem with the standard criticism is that it (bizarrely) concedes that the definition of human rationality coincides with a game theoretic player, and this is what Green and Shapiro (1994), and other critics of rational choice theory, have failed to recognize. Humans are in

have fetishized the different methodological approaches at the expense of focusing upon results and predictive performance.

many ways limited or less good than *homo economicus*, but it is also the case that humans typically make choices in difficult environments with limited information. Kahneman and Tversky (1979) style experiments point to our obvious flaws in relatively simple tasks; but up the ante a bit by presenting humans with more complicated games and the tables turn quite dramatically. We "outperform" rational choice players, and there is no reason to suspect that game theoretic models have much to say about certain (complex) classes of human games and decision contexts. So although we may sometimes make mistakes calculating simple expected value problems, it is also true that we thrash rational choice players at games like poker.[4]

The second, equally critical line of research, focuses upon the supposed ahistoric, simplifying assumptions relied upon in most game theoretic models.[5] The argument from proponents of a more historical approach is that the models are assumption driven, and that careful attention to case studies is a more appropriate avenue toward understanding complex problems.

History, regrettably, seems less than fruitful as a model of how to conduct research in the social sciences, unless one wishes to give up on any notion of causality.[6] If one adopts a King, Keohane, and Verba sensibility about the nature of epistemology underlying both mathematical methods and historiography, one has to admit that "case studies" of itself is not a methodology.[7] Much as some would wish, the use of history does not somehow avoid problems inherent in models of any kind (game theoretic or not). As I have argued in Chapter 2, a parameter space is a parameter space, whether the parameters involved are quantitative or qualitative.

Despite my belief that game theory, as currently employed, is inadequate to the challenge posed by many important problems in the

[4] The most ambitious poker research project is currently run by the University of Alberta Computer Poker Research Group. See http://www.cs.ualberta.ca/~games/poker/ for more details.

[5] For an overview in the security studies literature, see Walt (1999a); for an attempt at a general synthesis of game theory with case studies, see (Bates et al. 1998).

[6] For a considered treatment of the role of causality in historiography, see Novick (1998).

[7] Typically, proponents of qualitative/historical methods state that one should match assumptions with reality, incorporate dynamic/path dependent elements, and be sensitive to the possibility that preferences or institutions may change through time. Typically, they are not troubled by the complications that ensue due to a loss of IID observations and the explosion of the parameter space.

social sciences, I do believe that mathematical modeling is the best of all approaches, so long as we remember the arguments presented in Chapter 1. Models should be compared by examining how fruitful their predictions are, rather than the supposed accuracy of their assumptions. This implies that models should provide results that correspond to the real-world phenomenon that is being measured with a dependent variable; qualitative assessments and analogies that aim to "bridge the gap" between the model and the data are troubling.

Extensive Game Forms, Utility Functions, and Feature Spaces

In Chapter 2 of what is perhaps the most important book ever written in economics, Von Neumann and Morgenstern state what has now become a tenet of belief for most of us:

> It should be clear from the discussions of Chapter I that a theory of rational behavior – i.e., of the foundations of economics and of the main mechanisms of social organization – requires a thorough study of the "games of strategy" ... in the process of this analysis it will be technically advantageous to rely on pictures and examples which are rather remote from the field of economics proper, and belong strictly to the field of games of the conventional variety. Thus the discussions which follow will be dominated by illustrations from Chess, "Matching Pennies," Poker, Bridge, etc., and not from the structure of cartels, markets, oligopolies, etc. (1944, 46–7)

There is, however, one point of difference between the program specified by Von Neumann and Morgenstern and that advocated by more modern game theorists. The games Von Neumann and Morgenstern viewed as "essential" to any theory of rational behavior have been dramatically dumbed down or truncated from modern game theory – bridge and poker, for example, are not active research concerns. And it is not as if Von Neumann and Morgenstern were alone in their earlier belief that game theory's domain should involve complex human games. John Nash, famously, was interested in Go, and invented the game Hex while he was at graduate school at Princeton.[8] Ken Binmore, after finding one of Von Neumann's poker models counterintuitive (yet helpful for actual play), also decided to study game theory.[9] Many of

[8] See, for example, Kuhn and Nasar (2002, Chapter 3).

[9] Personal communication with Binmore and Binmore (1992, 573). As Binmore details in Chapter 12 of *Fun and Games*, many other game theorists were interested in poker, including Borel and Shapley.

the brightest game theorists had, at some point, a desire to use the tools of game theory to study real, human games.

Ironically, Go, Hex, chess, and poker are largely immune to game theoretic treatments. The main reason that the tools of game theory (and the early research program of Von Neumann and Morgenstern) have failed so completely is because of the encoding used to study games. What is wrong with using a normal form or extensive form to encode games? To understand the issues involved, a brief detour to machine chess is necessary.

Of all the games noted here, chess, on the face of it, is by far the most amenable to game theoretic treatments. Chess also has the beneficial feature that decades of work have been devoted toward building a machine player that is the equivalent of human masters. Periodically, one can measure the success of this program by taking stock of the frequent matches between machines and human opponents.

Of all these matches, the most famous is the second tournament between IBM's Deep Blue and Garry Kasparov that occurred in the spring of 1997. After watching Kasparov lose to Deep Blue, chess masters went on record with statements such as:

> "Nice style!" said Susan Polgar, the women's world champion. "Really impressive. The computer played a champion's style, like Karpov," she continued, referring to Anatoly Karpov, a former world champion who is widely regarded as second in strength only to Kasparov. "Deep Blue made many moves that were based on understanding chess, on feeling the position. We all thought computers couldn't do that." (New York Times, "Computer Defeats Kasparov, Stunning the Chess Experts," May 5, 1997)

The problem, of course, is that Deep Blue does not represent a triumph for artificial intelligence, a fact the IBM team is quite up-front about (though the media has not been). From the official IBM FAQ on Deep Blue:

> **Does Deep Blue use artificial intelligence?**
> The short answer is "no." Earlier computer designs that tried to mimic human thinking weren't very good at it. No formula exists for intuition. So Deep Blue's designers have gone "back to the future." Deep Blue relies more on computational power and a simpler search and evaluation function. (Deep Blue FAQ. *http://www.research.ibm.com/deepblue/meet/html/d.3.3.html*)

Even though Deep Blue is not what one might expect from artificial intelligence – a learning, strategic algorithm – Deep Blue *is* (rather

surprisingly) the ultimate game theoretic player. By this, I mean that Deep Blue approaches the problem of chess in much the same way a game theorist would recommend. It uses an extensive form for chess positions, and "solves" this tree with a variant of backwards induction (i.e., alpha-beta pruning, which is computationally more efficient than backwards induction).

So, Zermelo had it right all along; at some point, chess will be conquered by computers, given any reasonable increase in the power of hardware. And more important for the argument here is that chess has been conquered by an encoding and a solution algorithm that look very much like game theory. One could argue that success in such a complex game is an indication that game theory is an appropriate tool for studying all human games and decision contexts.

The problem is that looks are deceiving in this case. Two characteristics in particular of Deep Blue (and machine game players more generally) distinguish it from game theory proper; moreover, these distinctions are instrumental in highlighting what goes wrong when game theorists approach complex problems.

The first characteristic worth noting is the use of *idiosyncratic utility functions*. Chess, obviously, has a well-known utility function with three elements {win, lose, draw}. Unfortunately, encoding chess with game theory requires the ability to match strategies with outcomes; given the combinatorics of chess, this is not feasible. It is also obvious, to anyone who plays the game, that human players are able in most cases to evaluate a game before the terminal nodes are reached, based upon features such as material, position, pawn support, and the like. These features are mysterious, insofar as they bear no obvious connection to the ultimate utility function (or strategies) of the game as encoded by game theory.[10]

But perhaps after reading Chapter 2 they are not so mysterious to us! The terms of the utility function used in models of chess can be seen as analogous to the features described by Bishop and other researchers in non-parametric statistics. Just as in empirical models, not all features

[10] Concepts such as "material" do not bear any relation to the rules of chess either, thereby compounding the mystery. Whether one assigns a single point to a pawn and three to a bishop or entirely different values is not, in any way, deducible from the rules of the game.

in machine chess are created equal – a random feature of chess, such as how close to the center of a square one places one's pieces, would probably not be useful. Machine algorithms thus incorporate an insight that is common to both empirical and analytic work: By utilizing (linear) combinations of features that characterize utility derived from intermediate actions taken within the game, one can reduce the parameter space of an impossibly complex game. Without such a simplification, no analytic work would be possible.

Thus, one surmounts a logical problem. Even though Deep Blue could calculate hundreds of millions of moves a second, all of this processing power would be for naught unless it could span complete strategies. Game theory, as noted earlier, requires complete extensive forms, so that one can map terminal nodes to payoffs – an impossible feat in chess. Instead, Deep Blue used its formidable computational power to generate *partial* extensive forms (e.g., a tree that represents 10 moves out from the current game state) and assigned the terminal nodes of these partial extensive forms values taken from idiosyncratic utility functions. "Idiosyncratic" is taken here to mean a utility function comprised of features that bear no necessary relation to the utility function or rules of chess but are nonetheless helpful in evaluating intermediate positions in chess.

Where does such a utility function come from? The researchers involved used a combination of domain specific knowledge of chess to choose the terms in the utility function, plus empirical work drawn from past games of human chess masters:

> The evaluation hardware has four components. A piece placement evaluation scores pieces according to their central placement, their mobility and other considerations. A pawn structure evaluation scores pawns according to such parameters as their mutual support, their control of the center of the board and their protection of the king. A passed-pawn evaluation considers pawns that are unopposed by enemy pawns and can therefore be advanced to the eighth rank and promoted to queens. A file structure evaluation assigns values to more complicated configurations of pawns and rooks on a particular file.
>
> We also began to consider ways of tuning the evaluation function's 120 or so parameters, specified in software. Traditionally, programmers had hand-tuned the weights that programs assigned to material – pawns and pieces – and to positional considerations. We believe ours is the only major program to tune its own weights automatically.

> We acquired 900 sample master games and arbitrarily defined the optimum weights as those that produce the best match between the moves the machine judges to be best and those that the masters actually played. (Hsu et al. 1990, 18)

Essentially, the researchers relied upon a combination of hill-climbing and OLS estimation to provide a fit between feature weights and optimum play, where "optimum" is defined as a correspondence between good human play and the idiosyncratic utility function's evaluation of intermediate strategies.[11] A combination of deductive (particularly for the end game) and computational work thus chose the features, while clever empirical work using a data set comprised of expert human play determined the weights of these features in a linear function.

The second characteristic that distinguishes machine chess from game theory proper concerns the use of *components*, which may be thought of in similar terms as a subgame. The distinction between the two concepts is that subgames, starting with a node in the extensive form, include all possible histories generated from that node (i.e., all actions from the starting node to the corresponding terminal actions). Component games, in contrast, are here defined as any linked collection of actions, whether or not the partial strategy includes terminal actions.

In machine chess, components often are used to simplify play. In fact, independent algorithms are defined for different components that added together form a complete game of chess. One example would be opening move libraries; another would be a specialized algorithm for end games in chess. What should be obvious is that the use of components dramatically reduces the combinatorics of strategies by decomposing them into computationally independent parts. It should be equally clear that doing so represents an enormous leap of faith, insofar as the independence of components does not at all follow from the extensive form of a game.

[11] Hill-climbing refers to a computational technique akin to a gradient search. One defines a neighborhood as a number of perturbations originating from a given action, then chooses a perturbation that results in the greatest increase in utility. By following the "best" perturbation from each action's neighborhood, one can computationally determine the location of local optima. Note that the idiosyncratic utility function defined above is a real-valued function defined on incomplete extensive forms, while the actual utility function for chess is {win, lose, draw} defined on the complete extensive form. There will be more discussion of optimization in the next chapter.

Components, like subgames, have associated utility functions, but, as noted earlier, they do not necessarily include terminal nodes as in game theory. This complicates the issue of assigning utility to strategies taken in a component game, as one cannot simply use the overall utility function of the complete game for the partial strategies of a component. The lack of complete strategies demands that "solving" any given component rests upon the assignment of an idiosyncratic utility function tailored to that component. As we have seen, these utility functions have no necessary relationship to the overall utility function of the game itself.

One example of the relationship between components and idiosyncratic utility functions would be the idea (probably correct, but perhaps wrong) that a rook's-pawn opening is inferior to a king or queen's pawn opening. If our component is taken to be the first N moves of the game (spanning most opening move libraries), and our idiosyncratic utility function is taken as outlined earlier (some combination of position, attacks on central squares, etc.), rook's-pawn openings are inferior. In game theoretic terms, this kind of statement is impossible to make given the combinatorics involved – and would be quite suspect if made informally. After all, it is possible that an equilibrium strategy for white for the complete game of chess starts with a rook's-pawn move.

Machine chess thus provides a great deal of insight into how one might go about modeling complex games using feature spaces and domain specific encodings. One starts with deductive models to gain insight into a problem using transparent, tractable methods; one also may use deductive methods to solve "easy" parts of a problem (e.g., the end game in chess readily succumbs to purely deductive approaches). Instead of resting at this point, however, one should try to build cumulative models that add verisimilitude, which I have defined as a model that provides results directly connected to empirical tests (whether they be out-of-sample data or analytic implications of the model). This may well mean that early deductive models are expanded using computational models; in machine chess, this process was aided by empirical work that chose the key parameter values of the idiosyncratic utility functions. At the end of the day, the "closer" the final model is to the real-world referent, the better. After all, one would be far less impressed with machine chess algorithms if they did not actually

play the game but instead yielded vague aphorisms about cooperation, bluffing, and the like.[12]

"Brittle" Encodings and Equivalence Classes of Games: Empirical Implications

A quite different problem with the application of game theory to security studies concerns the "brittleness" of game theoretic models. But what does this mean? Simply put, it means that game theory should be able to model a broad range of games, and moreover, scale well with the different sorts of complexity one finds as problems in subfields like security studies. Else, we run into the difficulty that we model what we can and depend upon loose analogies to talk about the problems we are really interested in.[13]

This is potentially quite embarrassing, and much like the situation of teaching the IPD to undergraduates. In every class, at least one student, dissatisfied with the outcome of the single-stage game, attempts to change the payoffs.[14] We patiently explain that changing the payoffs, or the strategy sets, results in a *different* game; moreover, the "novel" game bears no obvious relationship to the original. This runs counter to most students' intuition and for good reason: Game theory demands a high level of precision because of the way it encodes games. The problem, of course, is that all this precision usually goes out the window when we make analogies and derive empirical implications of our models. One cannot assume that a two-player, complete information game of interactions between nations is useful for understanding the

[12] Note that there is an enormous "peace dividend" of writing models that play the actual game, rather than relying upon simplifying assumptions that change the game under consideration. By forcing modelers to focus on chess (rather than some simpler game that one believes, through unprovable analogies, is "like" chess), one can use data from real games to improve one's models.

[13] One example of this phenomenon is the short, unhappy literature on Colonel Blotto-type games. Despite enormous interest by military planners and think tanks, the literature petered out when results were not forthcoming. There is, however, relatively recent work (in the style of Axelrod's tournaments with the IPD) by the mathematician Jonathan Partington.

[14] You want to partially reward this type of behavior – at least it shows an understanding of the logic of the game, and there are examples where this type of skullduggery is rewarded. See, for example, Kreps and Wilson 1982 and their modification of the chain store paradox (Selten 1978).

nondyadic conflicts of the real world. Small changes (e.g., increasing the numbers of players) matter a lot; to date, there is no aspect of game theory that readily lends itself to theory that persists across "equivalence classes" of games.

The work of Signorino (1999) details exactly how disastrous this is for the empirical study of international conflict. On the one hand, Signorino notes that the measures typically employed in quantitative studies of conflicts (e.g., the balance of military forces) fail to "capture the structure of that strategic interdependence – that is, the set of states interacting, their sequence of decisions, options at decision points, the factors that influence their incentives, and the equilibrium effects of this interdependence on outcomes" (280). Further, if more than two states are involved in a conflict, using dyadic observations (as most everyone does), means that "each N-nation interaction becomes $N(N-1)/2$ independent observations, greatly expanding the size of the data set without adding any additional information to it" (280).[15] Given these problems, Signorino does the only sensible thing, and builds an empirical model that is more closely connected to the strategic game responsible for conflict.[16]

On the other hand, the obvious merit in including the structure of a game in the empirical specification leads to an unexpected difficulty in evaluating the predictive power of the model:

> We know that small changes to a theory (e.g., the number of players, the sequence of their moves, the choices and information available to them, and their incentives) can have large consequences in what the theory predicts. If a theory is vague, then it is unclear what statistical model would be consistent with that theory. Therefore, if we want to ensure consistency between a theory and a statistical model, we must be as precise as possible in the specification of the theory. Given the requirement for theoretical precision, how are we to specify and test strategic theories without doing so formally? . . . although the call for increased formalization of theories may be welcomed by many positivists, the importance of structures also seems to cut the other way. Consider the typical derivation and analysis of a positive theory. One major assumption

[15] I would quibble with the characterization of these expanded observations as independent – they are not IID, and this obviously sows confusion for empirical models.

[16] Essentially, he assigns probabilities to histories of the extensive game form. To avoid the problem of null probability histories, an error term is attached to these probabilities.

generally held – indeed, held throughout this article – is that the structure of the model remains constant across all observations in the data.... It does not seems unreasonable to suspect, however, that the true game structure changes over time and place. If even small changes in structure can make a large difference in likely outcomes, and if the true structure of the strategic interaction changes from observation to observation in our data, then what are we to make of any statistical results predicated on the assumption of a fixed game? (294–5)

One could easily expand the above point to include the possibility that the game in question is not quite right. Absent any theory of equivalence classes that would describe "similar" games, one has to suspect that empirical "verification" of a game has more to do with building a game to match a preconceived idea of the correct outcome than any genuine understanding of the causal processes involved. There are, after all, a universe of possible games. For any given empirical outcome, 10 different researchers might construct ten different games that all yield the "right" answer. Unless we have a source of novel data, it would be impossible to discriminate between them.

How brittle are game theoretic results? In the next section, I will review one of the more prominent models of international conflict, and see how well it holds up under the sorts of criticism advanced in this section.

AN EXISTING MODEL OF CONFLICT INITIATION

Perhaps the most famous paper that uses game theory to examine conflict is Fearon's (1995) "Rationalist Explanations for War." A game theoretic paper by Brito and Intriligator (1985) and work in military history by Blainey (1988) predate the work by Fearon in political science, and along with several decades of work in labor economics, all of these papers make essentially the same argument: wars, given what we know about bargaining theory, are not "rational" enterprises, insofar as one could always transfer resources rather than incur the costs of a conflict. Wars are thus the byproducts of incomplete or asymmetric information; or, as Gartzke (1999) puts it, "War is in the Error Term."

To make his argument, Fearon relies upon a standard bargaining game (see Osborne and Rubinstein 1990, for an overview) and

makes the following assumptions to justify the claim that completely informed, rational players do not initiate conflicts:

1. war is costly for both participants;
2. two nations constitute the set of players;
3. the game is a single-stage, not repeated;
4. there exists a single fungible, continuous resource (i.e., no non-separabilities or discontinuities exist over multiple issues).

Given the bargaining model and requisite assumptions employed by Fearon, is there any reason to place much stock in his conclusions?

A fair answer has to be "no." The problem, as noted by Signorino, is that the conclusions of Fearon's work are very much dependent upon the structure of the formal model; if any of these assumptions are incorrect (or simply different), one can make no prediction about the effect this perturbation would have in generating his conclusions. And as I argued in the first chapter, one has to be especially wary of models that do not have the property of result convergence; that is, if minor perturbations of assumptions produce dramatically different results, one cannot place much confidence in the model.

His first assumption, that war is costly, is far from obvious in the macroeconomics peace literature. In their review of the macroeconomics literature, Isard and Anderton (1999) detail the impact of defense spending and conflict on various measures of the domestic economy, including GDP, inflation, unemployment, and technological investment. This overview makes clear that it is difficult to discern what the overall impact of military spending or a conflict is, especially if one understands that "the good" is not a single feature (e.g., GDP might go up if the resources used for the military were optimally reallocated but unemployment might rise). If one adds more dramatic but longer term effects such as a nation's overall technology for war and the experience gained from fighting, the picture becomes murkier still.[17] Last, Fearon is reluctant to identify who the relevant agents are in his model – who,

[17] See Diamond (1997) or Kennedy (1987). While one can take issue with the analysis presented within these volumes, they do serve as warnings to how difficult it is to evaluate the costs and benefits of military resource allocations over long time spans. Of the more memorable examples of how decreased military spending can result in dramatic, unexpected costs is China's self-mandated destruction of its naval resources in the 15th century.

in particular, receives utility from conflict (or its avoidance)? As Goemans (2000) has pointed out, it may in fact be beneficial for state leaders to engage in conflict in some circumstances, and they are but one candidate for the agent under consideration in Fearon's work.[18]

Fearon's second assumption limits the set of players to two nation-states. This adds obvious analytic tractability but at the cost of compromising any general conclusion generated by such a model. It is worth repeating that war is often nondyadic, and models that assume dyadic conflicts cannot be expected to have anything general to say about nondyadic conflicts (i.e., structure matters). As noted earlier, N-player games are difficult, in most circumstances, for game theory to encode; thus, formal modelers often make the choice of making analogies to the multiplayer case from a two-player game.

Fearon's third assumption is equally injurious if one wishes to derive general conclusions from his work. Conflicts in the international system are not single-shot games; and again, we should have no expectation that a model with repeated plays would generate the same conclusion. In fact, working from similar assumptions but generalizing to the repeated case, Garfinkel and Skaperdas (2000) show that it is often "rational" to initiate conflicts. The reason for this is easy to understand – following Fearon's setup, if nation A has epsilon more power than nation B, it could extract some amount from B so long as this total is less than the cost to B of fighting. Unfortunately (for B), if the game is iterated, the extraction of resources by A would certainly continue, and the gap would grow while B's chances of winning a conflict would decrease monotonically. B's best chance, then, of winning a conflict would be on the first round. By considering the repeated game, one arrives at an equilibrium result that is exactly the opposite of that presented by Fearon in his section on "Preventive War."

The final assumption, which seems trivial, is that nations may exchange resources in a transferable, continuous commodity. Imposing the assumption of unidimensionality in this way avoids a problem that Fearon himself notes: If the issues at stake are discontinuous or nonseparable, war may be the inevitable result of an inability to reach a bargaining outcome all sides prefer to war. Fearon dismisses this as unlikely, but surely, this is an empirical question, and issues such as the

[18] One also could look to Machiavelli's *Discourses* for a distinct viewpoint on the costs and benefits of war for a republic.

status of Jerusalem in the Israeli-Palestinian conflict might lead one to reconsider this assumption.

At this point, one might object that attacking assumptions in this manner is counterproductive, and cite the treatment of Friedman (1953) in Chapter 1 of this book as proof of this position. Friedman's version of instrumentalism is consistent, however, only if one heeds his call to compare models based upon their predictive power, rather than the verisimilitude of their assumptions. Fearon's model has no hypotheses that one could test, so under Friedman's terms it would be impossible to support or reject his model by tying the results to existing (or even hypothetical) sources of data on conflicts.[19] Assumptions, in this case, are all we have, both for evaluating the model and for generating substantive conclusions. Slight changes in the assumptions result in huge changes in the conclusions, and that should be a cause for alarm, especially when the assumptions fail to comport with the phenomenon in question.

Put another way, *all* conflicts are characterized by asymmetric/incomplete information. But not all nations fight wars. For Fearon's model to be useful, one would have to provide a measurable dependent variable that could falsify the model.[20]

On a theoretical front, it seems that the state of affairs has not advanced much past the description offered in Niou and Ordeshook (1991):

> For those who are not a party to it, the debate between realists and neoliberals seems a curious circus. While realists struggle with the specification of state goals and with alternative conceptualizations of balance of power, neoliberals offer vague admonitions that goals depend on context. Realists see cooperation as secondary to the conflictual processes of politics even though stability

[19] As noted in Chapter 1, one might also test a model by deriving nonobvious implications of the model, especially for problems that are data-poor.

[20] This is not to say that all purely deductive models are without merit – contrast Fearon's model with Arrow's (1963) seminal work. For Fearon, one needs to be confident that one is observing a two-player noniterated game where war is costly. For Arrow, *any* preference aggregation mechanism fails to satisfy a set of conditions one would like to be true of democracy. Absent any way to directly test a deductive model, one must consider the breadth of the theory. In Fearon's case, it seems fair to say that the model does not obviously describe anything about the world. To the degree it is a simplification of reality, one needs to map the results of the model back to empirical work to see if the simplification is a useful one. In the case of Arrow's impossibility result, it applies to all conceivable voting rules, which clearly covers any real-world case. Empirical work is thus beside the point.

requires some minimal level of cooperation to maintain alliances, whereas neoliberals, aside from references to examples in game theory that do not necessarily model any specific international process, fail to define precisely the necessary and sufficient conditions for cooperation.... Neoliberalism argues that institutions matter because they somehow modify the actions of decision makers both directly by altering the costs and benefits of actions and indirectly by modifying goals, whereas realism has difficulty explaining the institutions and patters of cooperation that characterize human affairs. (481–2)

In large part, many of the debates about the nature of the international system stem from the underlying difficulty of modeling processes such as alliance formation and the initiation of conflicts. Fearon's model was used here because it is widely recognized as one of the better efforts to model conflict initiation; one should not be deceived, however, into believing that better models exist elsewhere.[21] What is needed is a different approach that avoids the shortfalls noted earlier.

A final note to the wary is in order. Critics of game theory often depend upon unstated assumptions about the nature of reason or other matters of taste. That is not the case with the criticism presented here. What is at issue is whether or not game theory, as one encoding of many that exists for investigating human games, is particularly effective in generating predictions or explaining current strategic problems. My answer is that it is not, and to make this case, I depend upon an examination of existing models.

Additional weight may be added to the argument presented here by borrowing from analytic results into the nature of the encoding offered by game theory. Conitzer and Sandholm (2002) have shown

[21] Powell's (1993) work on the trade-off between domestic spending and military allotments is less a well-stated game than it is a set of parameters, where Powell picks parameter values (by assumption) that allow him to make his argument. The Markov Perfect Equilibrium is also superfluous in this paper, as it does not serve to refine the equilibria past what a subgame perfect equilibrium would do. Niou and Ordeshook's (1991) paper is the best of the existing models, as it tries to incorporate N players with the possibility of coalitions. Unfortunately, the results are driven by the twin assumptions of a single, continuous resource and the fact that the game terminates when one coalition achieves exactly one half of the available resources. The main strategy, given this condition, is for the losing nations to achieve a coalition with exactly one half the resources, thereby artificially terminating the game before one nation dominates. Both of these papers violate the spirit of my arguments on the nature of assumptions in Chapter 1, insofar as technical assumptions (or parameter choices) unrelated to the phenomenon in question drive the results.

that for many classes of "simple" games, finding a Nash equilibrium (or a related refinement) is a nonpolynomial (or, NP) hard problem:[22]

Noncooperative game theory provides a normative framework for analyzing strategic interactions. However, for the toolbox to be operational, the solutions it defines will have to be *computed.* In this paper, we provided a single reduction that 1) demonstrates *NP*-hardness of determining whether Nash equilibria with certain natural properties exist, and 2) demonstrates the #*P*-hardness of counting Nash equilibria (or connected sets of Nash equilibria). We also showed that 3) determining whether a pure-strategy Bayes-Nash equilibrium exists is *NP*-hard, and that 4) determining whether a pure-strategy Nash equilibrium exists in a stochastic (Markov) game is *PSPACE*-hard even in invisible games (and remains *NP*-hard if the game is finite). All of our hardness results hold even if there are only two players and the game is symmetric. (10)

What does this result imply? It means that as the size of even the simplest games increase, the number of computations that must be performed to check for the existence of a Nash equilibrium increases in nonpolynomial time (i.e., greater than polynomial). And yes, this is a very, very bad thing. Again, simple bean counting of the sort performed in Chapter 2 comes in handy. Fearon's model, despite the apparent simplicity, has a quite large parameter space, where almost any deviation from his assigned parameter values causes substantially (and unpredictably) different results.[23]

[22] For an overview of computational theory, see Papadimitriou (1994).

[23] For his basic bargaining model, you have the following parameter space:

a. N players (set to 2 in Fearon's model)
b. A single issue with a real value from the set [0..1]
c. An ideal point for player A and B (or N in the most general case)
d. A choice of utility function for A and B
e. A number of iterations (set to 1)
f. A rule for conflict (in Fearon's model, a simple statement of probability p)
g. Costs C_a and C_b for the two players (assumed to be >0)
h. An equilibrium concept (perfect Bayesian, a refinement of the Nash)

One could quite easily generalize Fearon's model by resorting to a computational model, thereby increasing one's ability to track how changes in parameter values modify the results Fearon presents. Using a deductive model with a particular set of parameter values as Fearon does is a valid first step to gain intuition about a problem. The argument in this and subsequent chapters of this book is that building a more general model is an essential second step in this process, even if this means abandoning a purely game theoretic approach and moving toward computational models. The final step, of course, is empirical work to test one's results.

AN ALTERNATIVE METHODOLOGICAL APPROACH

Game theory is a powerful tool, but, as I have argued here, it has limitations that prevent it from aiding us in understanding some of the more complex human games and decision-making contexts. In this section, I will start to build a methodological approach that aims to:

a. allow researchers to explicitly model more complex games and build upon prior efforts using feature spaces and domain-specific encodings;
b. avoid brittle encodings that limit the generalizability of results (i.e., develop an equivalence class for your model);
c. introduce a methodology to model component games and their associated (idiosyncratic) utility functions.

So what should one do differently if one wishes to study more complex games that model conflict initiation, alliance formation, and the like? As noted earlier in this chapter, machine chess offers a great deal of guidance in solving this problem using the tools of computational political economy. Accordingly, I will argue that a combination of methodological approaches yields better answers for complex problems in the social sciences.

Is Computational Political Economy Different?

I have presented a number of examples of computational modeling in this book, focusing on problems ranging from standing ovations to machine chess. As I have argued, the best examples of computational models often model things as they are, without relying upon simplifications that distort the problem of interest. The merit in computational methods is that the space between the model and the empirical referent, as in machine chess, is entirely absent. This allows for a much easier transition to empirical tests, which curbs many of the problems of mathematical modeling highlighted by this book. Further, computational models can also serve as bridges between different methods. In the alliance game presented later in this chapter, game theory will provide most of the intuition about this problem, which will then be expanded on using computational methods.

It is, however, true that many computational models fall short of the research design standards presented here. Just as in game theory, researchers are often content to present a model that is more of an existence proof than anything dispositive. To take one example, consider the investigation of state formation by Cederman (1994). In his computational model, there are roughly two dozen parameters chosen for convenience; changing them produces qualitatively different results than those found in the paper.[24] As with Fearon's model of conflict, one needs out-of-sample work to increase confidence that the model has something genuine to say about the world.[25]

Combinatorial Game Theory

In addition to the computational political economy literature, a source of inspiration is the almost unknown (in the social sciences) combinatorial game theory literature.[26] Combinatorial game theory studies human games with the following properties:[27]

1. there are two players;
2. moves are sequential (rather than simultaneous);
3. there is complete information;
4. strategies are finite;
5. there is an ending condition, which specifies a constraint that determines the winner. Typically, a player loses if she cannot move; for example, checkmate in chess;

[24] Some of these parameters are reported in the paper and some are only found in the code. A more thorough treatment of the fragility of Cederman's model and the implications for computational political economy is the subject of ongoing research by a team of graduate students at Duke University (look for a paper by Jolly, Reifler, Tofias, and Warren in the near future). In Cederman's defense, his practice of using computational models essentially as existence proofs is no different from the standard practice in game theoretic articles. For a more modern computational model, see Lustick, Miodownik, and Eidelson (2004). Like Cederman, Lustick has constructed a brittle model without any empirical tests, but there is the advantage that Lustick has made the code available with an interface that allows one to modify the parameters easily.

[25] Predictive tests aside, one can still improve on current practice in choosing parameter values. Instead of relying upon a magician's hat or convenience, why not choose values based upon real-world data? One may still go wrong (e.g., any particular sample may be misrepresentative), but a bet placed on data trumps bets placed on fancy. Rabinowitz, MacDonald, and Listhaug (2004), for example, develop a computational model of issue voting using parameters derived from the 1989 Norwegian Election Study. Another compelling alternative is to use qualitative or historiographical methods to choose parameter values.

[26] For an introductory text, see Conway (1976).

[27] This partial list is taken from Berlekamp, Conway, and Guy (2001). The list included in their volume is somewhat more restrictive, but relatively recent work in the field has relaxed some of their assumptions (which I have dropped).

6. there is an effort to build equivalence classes for all results; that is, to see whether one's results apply to other games.

This strain of game theory has been the province of mathematics departments, and at first glance, seems to have much in common with game theory or even represents a quite drastic domain restriction of game theory.[28]

What is less obvious, however, is the way the above tenets are translated into practice, which *is* quite different from game theory in the social sciences:

I. encodings are specific to particular games;
II. rules are often expressed as constraints rather than a set of strategies and associated utility functions;
III. failing equilibrium play, the goal is for "better" play, typically measured against human performance;
IV. the focus is upon more complex games that humans actually play; for example, hex, Go, or poker.

A trade-off is taking place here. By giving up a ubiquitous encoding and an emphasis on backwards induction as the solution algorithm (with the goal of finding equilibrium play), the above approach has more latitude to attack harder problems. The cost for more verisimilitude, as we will see, is a change in the nature of the conclusions we may draw from our models and complications in encoding the model (see Chapter 4). Combinatorial game theory and computational political economy thus share common goals; the following examples will demonstrate how to apply these abstractions to research questions in the social sciences.

A Return to the Currency Game

Machine chess, as outlined earlier, is an example of a computational approach to modeling. What this entails may not be entirely clear, however, without an example from the social sciences. Let us return to the currency game of the preceding chapter to see what distinguishes computational modeling from game theory.

In the last chapter, I developed a model of the currency game using the programming language Perl. Building a computational model of

[28] It does, however, have a great deal to do with constraint satisfaction problems/logic programming, a field of artificial intelligence.

this kind was a straightforward process of translating the dynamics of the game into Perl – in many respects, this is much easier than constructing a game theoretic model of comparable complexity. What do the results of a computational model look like?[29]

In my initial presentation, epsilon (the parameter for mutation) was held constant while N (the parameter for population size) was allowed to vary in a limited range. Despite the fact that the game has only two parameters, it is complex enough such that one's intuition can be wrong. Young (1998), from whom I borrowed the setup of the currency game, is in part led astray because he considers a very small subset of the parameter space in his initial forays into modeling the game. His claim is that the currency game is useful in studying path dependent processes, such as competing technologies (see Arthur 1989):

> Qualitatively, this process evolves in the following manner. After an initial shakeout, the process converges quite rapidly to a situation in which most people are carrying the same currency – say, gold. This norm will very likely stay in place for a considerable period of time. Eventually, however, an accumulation of random shocks will "tip" the process into the silver norm. These tipping incidents are infrequent compared to the periods in which one or the other norm is in place (assuming ε is small). Moreover, once a tipping incident occurs, the process will tend to adjust quite rapidly to the new norm. This pattern – long periods of stasis punctuated by sudden changes of regime – is known in biology as the *punctuated equilibrium effect*. (Young, 11–12; emphases in original)

Along with the text, Young provides graphs of his experiments with the currency game, illustrating how one sees multiple regime shifts when the game is allowed to run through 30,000 iterations. Unfortunately, for the experiments he details in the text, he relied upon small population sizes ($N = 10$) and a very high mutation rate ($\varepsilon = .5$) to generate his intuition about the properties of the game.

Young's intuition that the currency game generates relatively long periods of convergence characterized by sudden shifts to the

[29] I should caveat this to say what the results of a "good" computational model do look like. As noted in Chapter 1, some computational modelers present qualitative results based upon a limited sample of synthetic data. This practice should be avoided. To have any hope of understanding a computational model of any complexity, one needs to develop a theoretically justified feature space and present statistical work for the parameters that make up that reduced space. Results are thus specific to a feature space.

alternative currency is supplemented by a deductive result he derives later in his book. The theorem is short enough to be stated here:

> Let G be a 2 × 2 symmetric coordination game with a strictly risk dominant equilibrium, and let Q_m, ε be adaptive learning in the playing the field model with population size m, complete sampling, and error [mutation] rate ε, $0 < \varepsilon < 1$. For every $\varepsilon' > \varepsilon$, the probability is arbitrarily high that at least $1 - \varepsilon'/2$ of the population is playing the risk dominant equilibrium when m is sufficiently large. (Young, 76)

What this theorem states is that if you have a suitably large population, a majority of the agents will be playing one currency. This majority is largest when ε is small. This adds quite a bit to his earlier chapter, but many readers might not be certain about what the properties of the currency game are for different values of n (m in Young's notation) and ε. Interpreting this theorem is elusive, even though it is deductively true.

Contrast Young's deductive approach with a computational model. In the appendix to Chapter 2, I list the Perl code that implements the currency game. The choice of language (Perl) is a matter of personal preference; what matters is that the goal of a computational model is twofold:

I. encode the rules of the game as accurately as possible;
II. iterate the computational model through a wide range of parameter values.

Once this process is accomplished, one has a tidy data set comprised of the following observations:

$$\text{mean regime shifts} \sim N, \varepsilon.$$

The final step is then to investigate the relationship of N and ε on the dependent variable (mean regime shifts) by estimating a statistical model. As noted in Chapter 2, a good model for the data generated by the currency game is a Poisson regression, which for a large number of observations spanning much of the parameter space described by N and ε is:[30]

[30] For the empirical work presented here, I generated 1,600 observations allowing N to range between [10 . . . 10,000] and ε to range between [.05 . . . .5]. This is not in actuality anywhere near enough observations for an accurate representation of the underlying DGP, but it is close enough for the purposes of this exposition.

Table 3.1. *Results for the Currency Game*

Poisson regression				**Number of obs**	=	**1600**
				LR chi2(2)	=	**19492.69**
				Prob > chi2	=	**0.0000**
Log likelihood = −2244.3969				**Pseudo R2**	=	**0.8128**
Num_reg	**Coef.**	**Std.Err.**	**z**	**P > |z|**	**[95% Conf.**	**Interval]**
N_pop	−.0351948	.0007884	−44.64	0.000	−.03674	−.0336496
Mutat_rate	15.4024	.9942932	15.49	0.000	13.45362	17.35118
_cons	−4.148296	.4970227	−8.35	0.000	−5.122443	−3.17415

The above model has a reasonable fit to the data and the residuals are normally distributed with very little variance. These results are different in character than Young's deductive work and easier to interpret. For a Poisson regression, the predicted values are given by e^{xb}. By examining representative values, it is easy to gain a more complete understanding of the currency game. For example, Young's preliminary analysis of the game is misleading, insofar as it only applies to very small populations. For N = 10 and $\varepsilon = .5$, $E(\cdot) = 0.015$ – in 1000 iterations of the game, this indicates that one would expect to see 15 regime shifts. If one considers a larger population and a smaller mutation rate – which more closely mirrors the nature of the real-world problem Young is using the currency game to investigate[31] – one quickly discovers that regime shifts never, ever happen. For N = 100,000 (still a tiny population) and $\varepsilon = .05$ (a reasonably high mutation rate), E(regime shifts) = 6.89E-1534 – and that is a lot of 0's.

As argued in the concluding remarks of Chapter 1, purely deductive investigations are most useful in starting an investigation and gaining some insight into how a problem works. Young's work is certainly helpful in understanding the dynamic process by which one regime, subject to small mutations, can move to another regime because of the nonzero probability that an arbitrarily large number of mutations will accumulate, pushing the system into a different regime. The possibility

[31] For example, the adoption of competing technologies such as a Wintel computer versus an Apple.

of regime changes, however, and the usefulness of this particular model is better understood when one develops a full-fledged computational model with the appropriate statistical analysis of the results. By doing so, one avoids intuition based upon a very small subset of the parameter space and brittle assumptions.

An Example Alliance Game

From the preceding example, we have seen how computational modeling can complement game theory; what we have not seen is what one does when presented with a more difficult problem. Often, computational and complex systems research in the social sciences has addressed different sorts of problems, such as path dependency, massively interactive social games, and the like. There are, however, domains of problems, such as those represented in security studies, where the focus is on complex interactions between highly motivated actors engaged in strategic games such as alliance formation and conflict. It is my contention that the lessons of computational modeling and combinatorial game theory offer insight into classes of difficult problems that elude purely game theoretic approaches.

To illustrate how one might develop models of complex phenomena in security studies, I now present an alliance game:

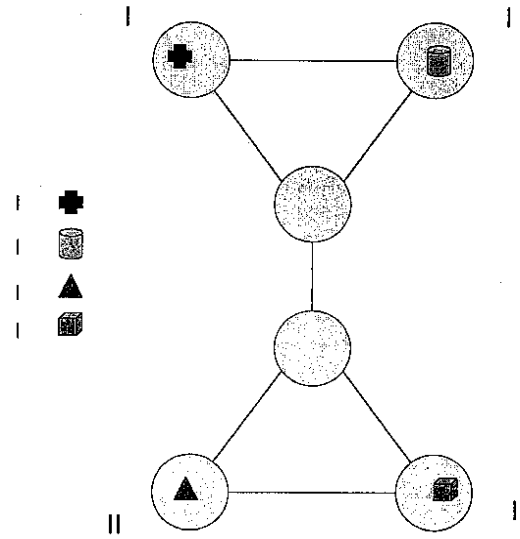

Figure 3.1. The Simple Diplomacy Game

Instead of representing this game as a set of strategies and their associated payoffs, one should note that Figure 3.1 uses a graph to represent the game. Additionally, the rules associated with this game are, in the spirit of combinatorial game theory, defined without explicit reference to the set of strategies:

The Alliance Game

1. The alliance game is represented by a graph G with a set of vertices X and edges U. There are six vertices in Figure 3.1; each vertex represents a location for a military unit. Each vertex also represents a supply (or resource) for building/maintaining a unit. Edges in G represent allowable moves for the units in the game.
2. Four players exist; each has a "home" territory where he or she starts with one unit on a vertex of the graph G (depicted with different shapes and colors on the graph). The home vertex is the only territory on the graph where a player may add units during Build/Remove (see below).
3. Each turn has three phases of play: Spring, Fall, and Build/Remove.
   a. Spring: units may move (change location between vertices), support (aid another unit in staying at a location or in moving to a new location), or do nothing (hold). *One action per unit is allowed per turn.* Territories captured (green circles on graph above) do *not* count for unit totals during the Spring turn.
   b. Fall: the allowable moves are identical to the Spring turn, except that territories captured DO count toward unit totals. If you capture a territory it remains yours unless another player occupies it for a Fall phase (n.b., moving into a territory in the Spring and leaving before Fall does *not* change ownership of a territory).
   c. Build/Remove: Count up the number of territories you own. Ownership, as noted above, is established by holding a territory for a Fall move (one also includes the starting territory). Ownership does not require a unit to stay in a territory; it is only disrupted by a different player occupying the territory for a Fall turn. If one owns more territory than one has units on the board, one adds units to the home territory (if it is owned by the player). If one owns less territory than one

has units on the board, one removes units from any territory on the board (player choice).

d. More than one unit of the same side may occupy a territory; but different players may not "share" a territory.
e. Units may support units of another player, as well as their own. Support may be given to defending units as well as to attacks.
f. When units come into conflict (i.e., different players move to the same territory, or a player attempts to dislodge a unit in a territory), numerical superiority – *adding all support orders* – wins. Losing units must retreat to an empty territory or be eliminated. For example, if two units attack a territory with one unit, the single unit must retreat. If another unit supported the defender, a tie would result and none of the four involved units would move.
g. *Ending condition*: the player who captures all six territories wins the game.[32]

It should be apparent that this game would be extraordinarily difficult to analyze within the confines of game theory. The strategies are (potentially) infinite, and despite the simple structure of the game, four players plus the ability to cooperate through support moves makes the combinatorics of this game quite ugly. One could, perhaps, argue that a few simplifying assumptions would somehow help one to gain leverage, but it is difficult to imagine what the complexion of these assumptions would be (unless they completely violated the spirit of the game).

So let us imagine we want to study this game, with the hope of improving our play against sophisticated human opponents or for developing insight into the dynamics of n-player alliances. Two definitions will help in examining the alliance game presented here:

Component game: A component of the game in Figure 3.1 is any proper, connected subgraph of G.

Component (idiosyncratic) utility function: Given that strategies are not necessarily finite (even in a component game), one needs to assign reasonable payoffs to actions taken in component games.

[32] This game is a simplification of the Avalon Hill game of Diplomacy, which is the study of some research in the artificial intelligence community.

One thus has to solve two problems before assigning payoffs to actions taken in a component game. First, one needs to choose an interval for analyzing payoffs; this can be difficult given the potential for nonterminating strategies. Second, one needs to choose a function that represents "progress" in the context of the component. For the purposes of the alliance game represented in Figure 3.1, one example of a component utility function has a horizon of one complete year, and the payoff is an integer that represents the delta of units in the last year (e.g., 0 = no change in the number of territories/units; +1 = one territory gained; −2 = two territories lost).

Some Example Components

So how does the forgoing discussion aid us in examining games like the alliance game? The (brief) inspection of several components will illustrate the main lessons of this section, and lay the groundwork for the more technical material in Chapter 4.

Example 1: Friends and Enemies, Together Forever. Figure 3.2 presents a typical component; in fact, this component eliminates the other two players of the game and focuses on one's own neighborhood. What, given this truncation of the game, would constitute good play? Given that we have already decided upon our component game, we also need to specify a component utility function. For the sake of the example, imagine we adopt the straightforward function that counts the delta in territories after each turn for each player (i.e., an integer from −1 to +1 for Figure 3.2). What type of play might transpire?

On the face of it, "safe" play might involve each player using their single unit to attack the other, as noted by the gray arrows in Figure 3.2.

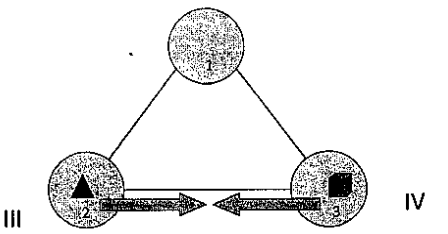

Figure 3.2. Friends and Enemies, Together Forever

What would the outcome of these actions be? Under the proposed utility function, the outcome would be 0 for both players. This infinite strategy {attack, attack, ...} would be supported in part by the fact that moving away in the Fall turn would give the other player your home territory, thereby awarding an additional unit to your opponent along with your home base. Mopping up the remaining territory would be easy at this point.

There is one caveat to the play proposed here. Suppose that the players on the other side of the board have not read a book on game theory, and for whatever reason, they fight and one of them wins their half of the graph. While you continue with a self-reinforcing strategy, the lucky (or foolish?) winner on the other half of the graph will then have superior forces with which to attack your half. This should cause some concern with the forgoing analysis and illustrates how difficult the selection of a component and corresponding utility function can be.

Example 2: For Better or For Worse? Does the strategy set forth for both players in Example 1 constitute an equilibrium of sorts? Imagine that one of the players "defected" from the strategy on a Spring turn, resulting in Figure 3.3. At this point in the game, the all-important Fall turn is about to occur, and one would like to know which player has the advantage. At first glance, many people evaluate Figure 3.3 by awarding the advantage to player 4; he does, after all, occupy player 3's home territory, and stands to gain a unit as outlined above.

It is important to remember that any such evaluation depends upon an (often unstated) component utility function. As always, the component we are studying bears no obvious relationship to the final outcome of the game, and so any associated component utility function is in some sense unsupportable within the context of the complete game. Given

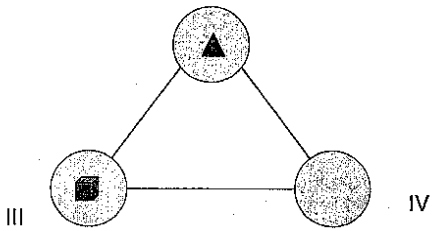

Figure 3.3. For Better or For Worse

the arbitrary nature of components and their utility functions, let us consider two candidates for a utility function:

Candidate 1: $u_1(\cdot) = \{-1, 0, +1\}$. This represents the delta in territories after the Fall move. It is plausible, insofar as gaining territory has to be seen as "good" and losing territory as "bad."

Candidate 2: $u_2(\cdot) = \{-1, 0, +1\}$, where -1 = your opponent has more territory than you do at turn end; 0 = equal territory; and $+1$ = you have more territory. Again, this is plausible, as it incorporates some notion of relative gains.

It cannot be overemphasized that there are a universe of utility functions that might be paired with any given component; these are simply two choices from that set. But, it is not the case that any component utility function will do – some are clearly better than others, though at this point it would be difficult to say why.

Nonetheless, what sort of action is implied by the two-candidate payoff structures? As in the preceding example, game theory will be used to build our intuition about the component game and these payoff functions.

Table 3.2. *Component Game u_1*

	Hold	**Attack Open Territory**	**Attack IV's Home**
Hold	(0,1)	(0,1)	(1,0)
Attack III's home	(0,1)	(0,1)	(0,0)
Attack IV's home	(0,0)	(1,0)	(0,1)

Nash Equilibria: [hold, hold]; [attack III's home, hold]; [attack IV's home, hold]; [attack IV's home, attack open territory]; [attack IV's home; ½ attack open territory and ½ attack IV's home]

Table 3.3. *Component Game with u_2*

	Hold	**Attack Open Territory**	**Attack IV's Home**
Hold	(−1,1)	(−1,1)	(1,−1)
Attack III's home	(−1,1)	(−1,1)	(0,0)
Attack IV's home	(0,0)	(1,−1)	(−1,1)

Nash Equilibrium: [1/3 hold and 2/3 attack IV's home, 2/3 hold and 1/3 attack IV's home]

With component utility function u_1 one finds multiple equilibria, and it would be difficult to guess what the outcome would be (or what strategy would result – see Binmore's (1992) discussion of chicken, 2825–6). One would imagine that player III would very much prefer the equilibrium [attack IV's home, attack open territory], but that player IV might well dissent given the comparative wealth of equilibria that advantage her.

With component utility function u_2 one finds a different outcome – a unique mixed equilibrium exists. Under this utility function, player III is at a disadvantage – his expected value is $-1/3$, while player IV can expect to receive $+1/3$.

Again one must raise the question of which utility function is "better." On the face of it, it would seem the second candidate does a better job capturing the reality of the game, but the only way to demonstrate the superiority of this selection would be to somehow show it results in better play (where "better" is defined as improving one's chances in the overall game).

Example 3: Tough Choices... The last example, displayed in Figure 3.4, represents a tough choice for both players III and IV. Ostensibly, they are potential enemies (i.e., there are no exogenously determined alliances), and player III has the opportunity to eliminate

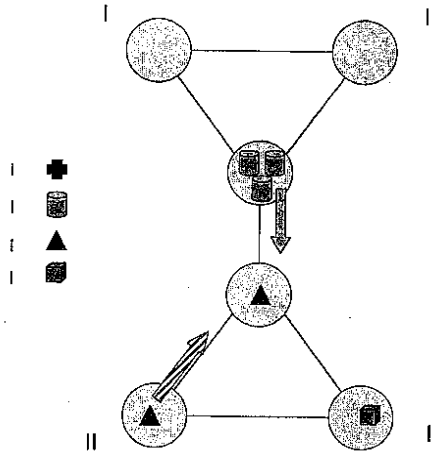

Figure 3.4. Tough Choices ...

player IV from the game. Player II is dominant, however, on the northern half of the graph, and will win the game (by taking the open territory on the southern half of the graph) before player III completes his conquest.

There is, however, one possible play that would salvage player III and IV – it would involve player IV supporting player III at the open territory and player III holding. One could pick a component utility function that would result in precisely this play, but it is hard to imagine this particular utility function performing well when player II is not threatening the southern half of the graph – in most cases, player III should eliminate player IV. If player IV supports player III and player IV refrains from attacking player IV, one would have to admit that this unmistakably represents an alliance, given that these actions are costly (in the short term) to both players.

What is needed (for players III and IV) is an indicator of what is happening on the other half of the graph. When no clear victor exists on the northern half of the graph, player III should crush player IV and player IV should seek allies elsewhere (perhaps in attacking the open territory held by player III). When there is a victor, as in Figure 3.4, players III and IV should instead be the best of friends.

CONCLUDING REMARKS ON A METHODOLOGY FOR CHOOSING COMPONENTS AND IDIOSYNCRATIC PAYOFFS

If one has been counting, I have so far defined three sets in the previous section. First, there is the set of all possible component games; even in the relatively simple alliance game, there are a great number of components; for example, the number of combinations with one to five vertices is 62. Second, there is the set of component utility functions, which in most cases is infinite, although usually this set is limited by and naturally formed by the rules of the game. For example, the rules of chess would naturally constrain the terms of the component utility functions to the squares of the board, the pieces, and relations between the pieces. Last, there is the infinite set of indicators that might aid a player's selection of components or utility functions in different states of the game. What is one to do?

Before proceeding, consider the alternative. One could define a simpler game that is "solvable" using the encoding and algorithms

of standard game theory and avoid the difficulties detailed earlier. Accordingly, one would then be forced to resort to analogies to discuss the behavior one is actually interested in. In the case of the alliance game, it would be difficult to imagine that the simplifying assumptions necessary for a game theoretic treatment would result in a game that had anything at all to do with the original version. And, as it stands, the unmodified alliance game presented here seems to illustrate alliance formation and conflict over resources in a much more realistic context than existing game theoretic treatments of these phenomena.

Seen this way, the preceding section has only made explicit what has been there all along – a huge search space (in terms of the combinatorics) that we can either account for in our models or assume away. We thus find ourselves in exactly the same situation as the last chapter when we considered empirical research. The curse of dimensionality and the possibility of brittle encodings are difficulties that models should grapple with. As we will see in the next chapter, developing a feature space that accounts for the complex strategic considerations highlighted in the preceding examples without running afoul of these difficulties is not easy. But, working through different encodings does develop a real understanding about the problem; this is to be contrasted with a game theoretic approach that would focus on deriving a set of limiting assumptions to produce an analogous game.

What methodology, then, is appropriate for constructing complex models with more verisimilitude? Given the difficulties in making analogies from existing game theoretic models to empirical tests, it is my belief that the only avenue available for studying more complex classes of games involves the use of computational models. As with the example of machine chess, a combination of deductive modeling to get one's feet wet and solve "easy" parts of the problem, computational modeling that builds upon this and adds verisimilitude, and statistical work to fix parameters and examine model performance, seems appropriate for most games.

This, in brief, would mean taking the following steps for the alliance game:

1. Use game theory to examine several candidate component games and associated utility functions, thereby developing intuition about the problem. For each possible encoding, examine the implied parameter space to see if it is reasonable.
2. Define a population of agents, their selection of components, associated utility functions, and features of the game's state space. In preliminary trials, components that result from most specifications are quite simple – the reason for this is that larger components (i.e., with a greater number of vertices) make learning more difficult and are selected against in an evolutionary framework. The same general rule also seems to be true of features. Within components and their associated utility functions, the preceding examples make clear that a good deal of "deductive" or game theoretic reasoning is employed by agents in the model. The main thrust of the computational work is to reduce an incredibly complex game into bite-size fragments (fit for chewing by limited game theoretic agents).
3. Optimization theory would be employed to traverse the space of possible components and idiosyncratic utility functions. For example, if one used a genetic algorithm, less "fit" members of the population would reproduce proportionately less than more fit members. Operators such as mutation and crossover would be applied at the level of component selection and the idiosyncratic utility functions. One might, of course, use an alternative stochastic optimization technique such as hill-climbing, simulated annealing, or genetic programming.
4. The results of the computational model would be probabilistic assessments of favorable components (i.e., those most often employed by agents in the population), idiosyncratic utility functions, and features.
5. To test for stability, one would test populations against "similar" games or against human play (which provides an infinite supply of novel data) – if a population was robust under perturbations in the model's form and diverse human strategies, one would have more confidence that the model was not overly brittle. E.g., one could expand the alliance game in Figure 3.1 to include N players (where N is even) by adding symmetric components that match the existing component games defined in Figures 3.2 and 3.3. In all the games of this class, one would expect that the components

discussed in the previous section and their utility functions would have similar properties.

As noted earlier, the knowledge that results from this approach is probabilistic in nature (i.e., regressions would be used to demonstrate the utility of different components, their associated utility functions, or indicators), much as it was for the computational model of the currency game above. Further, results from computational models, while not as brittle as a game theoretic treatment, would be specific to classes of similar games – there would not be a universal encoding suitable for all possible games.[33] This represents a trade-off, insofar as the qualitative character of our knowledge of games would change dramatically under this methodology. In compensation, however, one arrives naturally at a way to provide equivalence classes of games, which I have argued is necessary if one is to link deductive or computational models to their empirical referents.

In the next chapter, we will explore these issues further, and I will detail both the optimization theory and computational tools necessary to derive computational models of the sort hinted at above. As an extended illustration of this methodological approach, I will develop the specific algorithms that instantiate component games and the associated utility functions that provide an encoding for the alliance game. Examining the choices involved in developing such an encoding will demonstrate the advantages and trade-offs between computational modeling and purely deductive work.

[33] As we will see in the next chapter, components and their associated idiosyncratic utility functions will be useful (i.e., robust) under many modifications to the structure of the alliance game. For example, if one defines an equivalence class by adding additional players in a symmetric fashion (i.e., groups of three nodes akin to those found in the current game) thereby expanding the number of players from 4 to higher multiples of 2, most of the results of the simpler game carry over without modification.

4

Why Everything *Should* Look Like a Nail

Deriving Parsimonious Encodings for Complex Games

INTRODUCTION

In this text, there have been a number of models presented, ranging in complexity from the currency game to the alliance game. On the simpler end of things, the currency game (in Chapters 2 and 3) can be implemented with 85 lines of documented Perl code.[1] The alliance game, by contrast, represents a much more significant undertaking. If, in fact, one set out as I did for the currency game, fired up a text editor, and typed in the code without any prior planning, disaster would certainly result. While it may well be the case that Perl would serve for this project, one would need a different paradigm for building a computational model to accommodate the complexities of the alliance game.

The goal of this chapter is to cover some of the material that would allow a researcher who is not an expert in computer science to derive

[1] Perl is a programming language derived from C and originally used for system programming in the Unix environment. It is often referred to as the "Swiss Army Chainsaw" of programming languages (this last is attributed to the hacker Harry Spencer). For more on the origin of Perl, see the FAQ at http://www.perldoc.com/perl5.6/pod/perlfaq1.html. For our purposes, it is important to note that Perl does not impose a specific paradigm of programming (e.g., object oriented); rather, it is an easy-to-use language for many jobs. It is worth noting that after assigning this project to some smart undergraduates, I discovered that one can produce a smaller (though more opaque) program in about half the size of my version (see the appendix to Chapter 2). Instead of recording the preference of each member of the population in a vector, one can instead store only the proportion of gold adopters and modify this scalar value as the result of single agents changing their preferences. If one wishes to extend the model to allow social networks, for example, then one has to return to a vector representation.