Spring 5-15-2022

Development and Deference to Legal Doctrine at the US Supreme Court

Jbrandon Duck-Mayr
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Political Science

Dissertation Examination Comittee:
James F. Spriggs II, Chair
Scott Baker
Randall Calvert
Lee Epstein
Keith Schnakenberg

Development and Deference to Legal Doctrine at the US Supreme Court

by
JBrandon Duck-Mayr

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2022
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

There are many people whose aid and assistance I would like to acknowledge. First and foremost, I want to say a special thanks to my loving partner Sarah Duck-Mayr, without whom I would surely have lost the will to complete this work long ago. I also want to thank my advisor, Jim Spriggs, for his endless patience and exceptional mentorship, and the other members of my dissertation committee, Randall Calvert, Lee Epstein, Keith Schnakenberg, and Scott Baker, for their incredibly helpful comments and advice. A number of other faculty at Washington University in St. Louis deserve special mention as well, including Jacob Montgomery, who is well-known in the department for his generosity with his time and willingness to collaborate with students, and Matt Gabel, whose support and uncanny insight was invaluable. Additionally, I want to thank the amazing staff for the Department of Political Science, particularly Colleen Skaggs and Heather Sloan-Randick, without whom I could not have navigated my time at Washington University in St. Louis. Finally, I want to acknowledge the assistance I received during my last year from the COVID Research Disruption Extension (CORDE) program. The novel coronavirus disrupted all our lives, and made it impossible to complete my research on my original time table; the university's aid through this program was crucial to the completion of my research.

JBrandon Duck-Mayr

Washington University in St. Louis

May 2022

ABSTRACT OF THE DISSERTATION

Development and Deference to Legal Doctrine at the US Supreme Court

by

JBrandon Duck-Mayr

Doctor of Philosophy in Political Science

Washington University in St. Louis, 2022

James F. Spriggs II, Chair

How do judicial institutions and the choices judges make affect how the law develops? And how does existing law in turn affect judges' decisions? In this dissertation, I address important aspects of both of these fundamental questions of judicial politics. First I explore why courts create inconsistent legal doctrine. Because judges cannot describe how the rules they craft will apply to every conceivable factual variation in cases, they must describe them more abstractly. I use a social choice theoretic model to show that absent unrealistic restrictions on judges' preferences, decision making on collegial courts in this context can result in inconsistent doctrine. I then examine the constraining effect of law at the US Supreme Court. I generate a measure of the legal status quo, or the outcome implied in cases heard by the Court from its past precedents. I show how to control for the justices' own contributions to the legal status quo to identify the law's constraining effect, and find it exerts a statistically reliable constraining effect on the decisions of a supermajority of Supreme Court justices. The methodology for this study required extending a class of models from the machine learning literature, Gaussian process models, which I also devote a chapter to.

Chapter 1

Introduction

What can we say about the legal policy judges will make as they craft law? How does existing law affect judges' decisions? These questions are fundamental to the study of judicial politics. In this dissertation, I tackle important aspects of both of these questions, examining why judges generate inconsistent legal doctrine, and whether and how much existing precedent constrains the choices of US Supreme Court justices. This research advances not only our understanding of legal inconsistency and legal constraint at the US Supreme Court, but I also generate a new measure of the "legal status quo"—an important concept in the judicial politics literature more broadly—and advance the state of the art for a class of flexible machine learning models (Gaussian process models) that are useful for the social sciences.

In Chapter 2, "Explaining Legal Inconsistency", I use a social choice theoretic model to show one reason why inconsistent legal doctrine develops. Kornhauser (1992) highlights a difficulty in judgement aggregation on collegial (i.e. multi-member) courts when the legal rule dictating how intermediate legal conclusions (such as whether parties to a case have formed a contract) are related to ultimate case outcomes is fixed, and Landa and Lax (2009) highlights a similar difficulty when intermediate legal conclusions are fixed but judges' preferences over legal rules are aggregated. However, judges often engage in a sort of multi-step reasoning, where they determine

both the intermediate legal conclusions using the specific facts of cases *and* how those intermediate conclusions determine the ultimate outcome of the case. I prove that only very strict (and in the context of judicial decision making, typically unreasonable) assumptions on preferences can ensure multi-step judgment aggregation on collegial bodies such as multi-member courts does not result in inconsistency.

In Chapter 4, "The Force of the Law: The Constraining Effect of Precedent at the US Supreme Court", I assess whether existing law constrains Supreme Court justices' decisions in the cases before them. Prior literature lacks a convincing answer to this question, instead offering mixed findings, with studies such as Bartels (2009) and Bailey and Maltzman (2011) finding some evidence of a constraining effect of law while others such as Segal and Spaeth (1996) and Segal and Spaeth (2002) find little evidence of such an effect. I argue the literature's mixed findings are due to conceptual shortcomings and methodological issues: While a number of clever approaches have been taken to operationalize the law in statistical models of judges' decisions, so far we lack a measure of the legal status quo, or the outcome implied in the current case by the Court's past cases. Moreover, a difficult obstacle to inference in this context is the problem of state dependence; since a justice's vote on a case today impacts what the law is tomorrow, correlation between the legal status quo and a justice's votes could simply be measuring their tendency to agree with their own past selves. I demonstrate how to address both of these issues using Gaussian process classification, a model from the machine learning literature. Using this approach, I find the legal status quo has a statistically reliable and substantively meaningful effect on the decisions of a supermajority of Supreme Court justices.

In addition to providing the most reliable evidence to date on the law's constraining effect at the Court, the measure of the legal status quo developed in Chapter 4 is an important contribution to the study of judicial politics. Explicitly or implicitly, the legal status quo plays a key role in much of the judicial politics literature (e.g. Martin and Quinn 2002; Black and Owens 2009). Some studies try to capture the legal status quo of a case at the Supreme Court by coding the lower court ruling as

liberal or conservative, which makes sense if we want to focus on the Supreme Court as a manager that wants to correct specific errors of its subordinate courts. However, if we think Supreme Court justices care more about the overall state of the law, my new measure better captures the legal status quo. Further, since I show in Chapter 4 how to control for the information about individual justices' preferences that have leaked into this status quo measure, I help solve a long-standing and serious danger to inference in studies of legal constraint.

To conduct the analysis in Chapter 4, I had to advance the state of the art in Gaussian process (GP) models, a method which is also quite new to the political science literature, so in Chapter 3, "Inference in Gaussian Process Models for Political Science", I introduce the method and derive inferential quantities of interest to social scientists for Gaussian process models. As this family of models hails from the machine learning literature, almost all inference for these models previously developed related to prediction, whereas social scientists are typically more interested in inference regarding predictors' effect on outcomes, so I derive the distribution of sample average treatment effects in these models and procedures for inference in a variety of common contexts social scientists encounter. Machine learning methods such as GP models can help social scientists avoid misspecification bias (see Hainmueller and Hazlett 2014), and GP models in particular represent a superior solution to spatial and temporal error correlation than existing approaches (Carlson 2021; see also Gill 2021) and are more flexible and extensible than methods such as kernel-regulated least squares advocated by Hainmueller and Hazlett (2014).

In sum, this dissertation presents substantive research on fundamental questions in the study judicial politics—how the law develops and how it constrains judges' behavior—as well as useful advances in political methodology. In the conclusion I further explore the contributions of this dissertation and avenues it presents for potential future research.

Chapter 2

Explaining Legal Inconsistency

A wide range of observers have noted particularly inconsistent rules being produced by courts across several areas of the law.[1] For example, legal scholars complain the U.S. Supreme "Court's numerous [federal] preemption cases follow no predictable jurisprudential or analytical pattern" (Dinh 2000).[2] Political commentators criticize the Court's "Establishment Clause decisions that have been, in the words of Alice in Wonderland, curiouser and curiouser," and hope the Court will "leaven with clarity the confusion it has sown" (Will 2019). Supreme Court Justice Clarence Thomas bemoans "an Establishment Clause jurisprudence in shambles," claiming the Court's "jurisprudence has confounded the lower courts and rendered the constitutionality of displays of religious imagery on government property anyone's guess. . . " (*Utah Highway Patrol Assoc. v. American Atheists Inc.*, 565 U.S. 994 (2011) at 994, Thomas, J., dissenting).

Courts' policies are implemented by others, from lower courts applying appellate court rules, to outside actors enforcing judicially created policies (Maltzman, Spriggs, and Wahlbeck 2000, 5). When courts' rulings are unpredictable, and their rules are confusing, it impedes these actors'

1. This chapter has been published as "Explaining Legal Inconsistency" in the *Journal of Theoretical Politics* (DOI: 10.1177/09516298211061159). I would like to thank Randall Calvert, Keith Schnakenberg, Jim Spriggs, Lee Epstein, Morgan Hazelton, Jordan Carr Peterson, and anonymous reviewers for their helpful comments. A previous version of this paper was presented at the 2019 Annual Meeting of the Midwest Political Science Association.

2. Drahozal (2004) gives a book-length review of the inconsistency rampant in federalism cases.

ability to implement judicial policies. Moreover, inconsistency in legal doctrine reduces judicial legitimacy (Landa and Lax 2009, 959). Why would courts create confusing policies that endanger judicial legitimacy and their efficacy as policymakers? Perhaps judges are free to act relatively unconstrained (e.g. Segal and Spaeth 2002), and current court members simply prefer outcomes inconsistent with prior cases. Or perhaps courts' decisions are well explained by pronounced rules, even when scholars and commentators believe an area of the law is in disarray (Segal 1984). Maltzman, Spriggs, and Wahlbeck (2000) explain that bargaining over opinion content among justices may produce results inconsistent with what we might otherwise expect. However, none of these accounts explain why courts' *descriptions* of their decision rules do not provide clear guidance for lower court judges and other policy enforcers.

I use a social choice theoretic model to show preference aggregation on collegial courts can result in inconsistent rules when judges communicate policy in terms of subjective criteria that depend on objective facts.[3] That is, judges often explain rules using a low number of abstract determinations that in turn are derived from specific facts of cases. I show this kind of multi-step reasoning in appellate review can result in inconsistent collegial rules.

For example, in Fourth Amendment search and seizure cases, the constitutionality of police conduct can depend on (1) the intrusiveness of the search or severity of the seizure, and (2) whether the police had the requisite level of suspicion (e.g. probable cause) required to support such conduct. The court must determine how intrusiveness and police suspicion translate into outcomes, and further use the specific facts of cases to determine the level of police suspicion: "As the Court recognizes, determinations of probable cause and reasonable suspicion involve a two-step process. First, a court must identify all of the relevant historical facts . . . and second, it must decide whether . . . those facts would give rise to a reasonable suspicion justifying a stop or probable cause to search" (*Ornelas v. United States*, 517 U.S. 690 (1996) at 700–701, Scalia, J., dissenting).

3. For formal statements of results and proofs of propositions, see Appendix A.

To make this even more concrete, consider the case *Terry v. Ohio*. In *Terry*, a police officer observed Terry and two compatriots suspiciously "casing" a store. Although he had no other information about the men, he believed a robbery was imminent, and "feared 'they may have a gun"', so he approached them, stopped them, and frisked them for weapons. He found weapons on Terry and one of the other men, and they were convicted of weapons charges. These concrete events that happened, and the evidence collected, are the specific facts of the case, or the "historical facts" as Justice Scalia puts it. While the Court did not find these facts amounted to probable cause, they said the evidence of criminal conduct amounted to "reasonable suspicion". Again, though the Court did not find these facts constituted an arrest, the seizure of Terry did constitute an investigatory stop. These findings are the abstract determinations I mentioned above, which I will call *doctrinal facts* throughout the article. The Court announced investigatory stops may be justified by reasonable suspicion; in other words, the Court updated *doctrine*.

When courts engage in such multi-step reasoning, opportunity for inconsistency in the resulting collegial rules arises, even when all the judges possess well-behaved preferences. The problem arises because with multiple levels of judgment or preference aggregation, judges can agree on outcomes while disagreeing on the proper justification for that outcome, so that applying the reasoning relied on by a majority coalition in any one case can be inconsistent with collegial outcomes in other cases. This source of inconsistency in the law is understudied despite related results in the literature (e.g. Kornhauser 1992; Landa and Lax 2009) because models have left unexplored the interaction between disagreements over doctrine and disagreements over intermediate legal determinations, or doctrinal facts.[4]

This mechanism leading to doctrinal inconsistency raises implications for some areas of research in judicial politics. For example, there is a large literature that uses case facts as explanatory variables in empirical models of judicial behavior (e.g. Segal 1984; Richards and Kritzer 2002;

4. That is, while Kornhauser (1992) allows disagreement over intermediate determinations, and Lax (2007) and Landa and Lax (2009) allow disagreement over doctrine, neither allow disagreement over both.

Bartels and O'Geen 2015; Epstein, Parker, and Segal 2018). Studies utilizing doctrinal facts may ignore that individual judges can have different determinations of their own on such doctrinal facts, while if only historical facts are used, important inconsistencies in the reasoning presented by courts can be obscured. There is also a large literature on principal-agent relationships in judicial hierarchies (e.g. Cameron, Segal, and Songer 1994; Westerland et al. 2010; Baker and Kornhauser 2015). This paper raises an important question for future research of these relationships: the decision to engage in the multi-step reasoning studied here is itself a strategic decision. If the appellate court defers to trial court findings of doctrinal facts, this multi-step reasoning does not occur. (See also the appendix titled "Deference to Trial Court Findings" for discussion of situations in which appellate courts may even revisit findings on historical facts, another setting in which such multi-step reasoning can occur). For example, in the *Ornelas* decision quoted above, the Supreme Court resolved a circuit split over whether findings of probable cause should be reviewed *de novo* or with deference (in favor of *de novo* review), resulting in multi-step reasoning in Fourth Amendment cases. When will collegial appellate courts choose increased control over trial court agents, even with the risk of the type of doctrinal inconsistency studied here, rather than defer to agents' findings?[5]

After a short survey of the substantive literature, I provide a brief overview of related models before detailing the setup of a model that allows for courts' multi-step reasoning. I then show why inconsistency in the law can result when appellate courts communicate policy this way, as well as when they can safely do so while maintaining clear policy; I illustrate these results with a simple Fourth Amendment example.

5. See Lax (2012) for exploration of a similar tradeoff in judicial hierarchies.

2.1 Causes and Consequences of Inconsistency

If an appellate court's "jurisprudence [confounds] the lower courts" and makes the proper decision in future cases "anyone's guess" (*Utah Highway Patrol Assoc.*, 565 U.S. at 994, Thomas, J., dissenting), the court will be less effective as a policy maker. Such inconsistency also raises normative concerns—crafting an inconsistent doctrine leaves citizens potentially less empowered to assert their rights (since they can't tell when they apply). Nevertheless, legal scholars highlight time and time again various doctrines that have grown inconsistent, from death penalty jurisprudence (Robinson and Simon 2006) to First Amendment jurisprudence (Post 1995) to federalism jurisprudence (Drahozal 2004).

Empirical work has well documented the effects of unclear doctrine on courts' policy-making efficacy. Spriggs (1996) argues administrative agencies will be more likely to follow Supreme Court opinions that offer clearer guidance, and finds evidence that agencies more closely follow opinions that were more specific and explicit. Westerland et al. (2010) hypothesize that unclear signals from the U.S. Supreme Court will lead to lower compliance by the appellate courts, finding an increased number of concurrences indeed reliably correlated with lower compliance.

Empirical work has also uncovered some causes of inconsistency or complexity in judicial behavior. Collins (2008) finds individual justices' choices are more variable in complex cases. Owens and Wedeking (2011) use text analysis methods to measure the cognitive complexity of court decisions,[6] finding, for example, that some justices provide clearer guidance in their opinion than others on average, and that majority opinions are less clear than dissents, perhaps due to the bargaining entailed in crafting a binding precedent (1032–1033; Maltzman, Spriggs, and Wahlbeck 2000).

Related theoretical work includes the discovery of the "doctrinal paradox" (Kornhauser 1992) and its extension (Landa and Lax 2009),[7] as well as work on rules vs. standards (e.g. Clark 2016; Lax

6. Though they acknowledge that *doctrinal complexity*—the topic of this article—is another aspect of clarity of Supreme Court opinions (Owens and Wedeking 2011, 1038).

7. Study of the phenomenon identified by Kornhauser (1992) as the doctrinal paradox spread throughout legal

2012). The doctrinal paradox shows that outcomes depend on whether judges on collegial courts decide cases by majority vote over outcomes or by majority vote over intermediate determinations, such as whether police had probable cause. Interestingly, Kornhauser (1992, 447) explicitly envisions the cases as coming from a fact space that the judges must then map to these intermediate conclusions, but does not model *how* the judges make these intermediate determinations; accounting for this step in judicial reasoning is one of the principal technical contributions of this article.

However, Kornhauser (1992) assumes legal rules are fixed, while appellate courts themselves create legal rules. So Landa and Lax (2009) instead assume the intermediate conclusions are fixed, but allow each judge on a collegial court to have their own preferred legal rule. With this setup, the paradox that arises is that the rule implied for the court is different if the judges directly vote over rules or vote over outcomes in cases. Additionally, "it might not be possible to form the same type of rule for a court as a whole as any individual judge might have. That is, to the extent that individual rules are each representative of coherent legal philosophies, it may not be possible to construct a similarly principled collegial doctrine" (949). This captures a type of legal incoherence, and I build on these two models to additionally capture uncertainty, or the type of incoherent policy that renders the proper decision in a case "anyone's guess" as Justice Thomas complained.

The rules vs. standards literature tackles a separate but related issue to the doctrinal inconsistency I study. These studies seek to explain when judges will issue specific policies and when they will use vague policy. For example, Staton and Vanberg (2008) shows courts may use vague rules to prevent observed noncompliance with rulings by ideologically divergent governments or to allow leeway to governments that are ideologically aligned with the court.

Most on point for the present article in this vein are Clark (2016) and Lax (2012). Clark studies the trade-off between an opinion that clearly disposes of cases closely related to the present case and an opinion that is less precise but has more impact on dissimilar cases. Clark finds judges will

theory and social choice theory and became known also as the "discursive dilemma" (e.g., List 2012; List and Pettit 2002; Nehring and Puppe 2006, 2010).

be more precise when the instant case is most representative of potential disputes and when they anticipate being able to issue additional clarifying rulings in the future. This analysis starts from the important point that judges generally cannot specify a complete mapping from cases to outcomes in a single opinion. The import of Proposition 2.2 below, detailing the general susceptibility of doctrine to inconsistency, involves this issue; inconsistency has real bite precisely when judges cannot perfectly map every potential future dispute to an outcome.

Lax (2012) considers the ability of an appellate court to promulgate a bright-line rule that depends only on an objective fact, or a standard based also on a subjective dimension such as severity of the weather. In this context, we may say bright-line rules are specific or precise, whereas standards based on a subjective dimension are less precise, either because the Court cannot perfectly observe the subjective dimension or because it is difficult to specify doctrinal requirements on that dimension. In the first case, standards are preferred despite their vagueness when the ability to observe the subjective dimension is relatively higher, or there is lower risk of ideologically opposed lower courts. In the second, standards can be attractive despite imprecision if the weight placed on the subjective dimension in the Court's preferences is high enough, or if the cost of writing more precise opinions is low enough. This provides a nuanced account of incentives to rely on potentially vague doctrine, but again, does not wrestle with inconsistency in doctrine.

Evidence exists that courts' policy-making efficacy depends on legal clarity, and normatively we may expect courts to consistently interpret legal rights. Empirical work has uncovered some correlates of lack of clarity in the law, and theoretical work has shown conditions under which judges may choose vagueness over precision and clarity. I extend models of case-based adjudication (Kornhauser 1992) and rulemaking (Lax 2007) to show an explanation for inconsistent doctrine embedded in legal reasoning: Judges generally engage in multiple steps of judgment aggregation, and this multi-step reasoning provides more opportunity for inconsistency in aggregation than previous models have accounted for.

2.2 Rule Making on Collegial Courts

I use a case space model to study rule making on collegial courts (Lax 2011). A case space model considers the set of all possible cases, or factual scenarios, a court could be presented with, and represents judicial policy as dividing that space into outcomes. That is, the set of possible cases is divided into two sets: the set of cases where plaintiffs win and the set of cases where defendants win; or, the set of cases where government activity is permissible, and the set where it is unconstitutional.

In a traditional case space model, the court is presented with a case $x \in X \subseteq \mathbb{R}^n$, the set of all possible cases the court could hear.[8] Each judge j then has a preferred *rule* mapping cases to outcomes $\rho_j : X \to \{-1, 1\}$.[9] The dimensions of X are interpreted as "whatever facts might be considered relevant to the judges" (Landa and Lax 2009, 593). Often models consider these facts to be high-level doctrinal concerns, such as the intrusiveness of a police search (Clark and Carrubba 2012), or sometimes specific "historical" facts, such as the speed at which a car is travelling (Lax 2012).

I will use as a running example the constitutionality of a seizure of a person—an investigatory stop or an arrest—under the Fourth Amendment.[10] The Fourth Amendment to the U.S. Constitution provides the "right of the people to be secure . . . against unreasonable searches and seizures, shall not be violated. . . " (U.S. Const. Amend. IV). However, courts "must evaluate the reasonableness of a particular search or seizure in light of the particular circumstances" (*Terry v. Ohio*, 392 U.S. 1 (1968) at 21). For example, while arrests require probable cause, investigatory stops are less intrusive seizures that require only "reasonable suspicion" (*Terry*).

8. The dimensions of the case space could be the set of real numbers (e.g. Clark and Carrubba 2012), real intervals such as $[0, 1]$ (e.g. Lax 2007), or discrete sets such as $\{0, 1\}$ (e.g. Landa and Lax 2009); the space may be unidimensional (e.g. Hübert 2019) or multidimensional (e.g. Badawi and Baker 2015).

9. The dichotomous outcomes are sometimes presented with other labels, such as Y and N (e.g. Landa and Lax 2009).

10. Fourth Amendment doctrine is a familiar example both to empirical (e.g. Segal 1984) and theoretical (e.g. Clark and Carrubba 2012) studies of case-based judicial decision making.

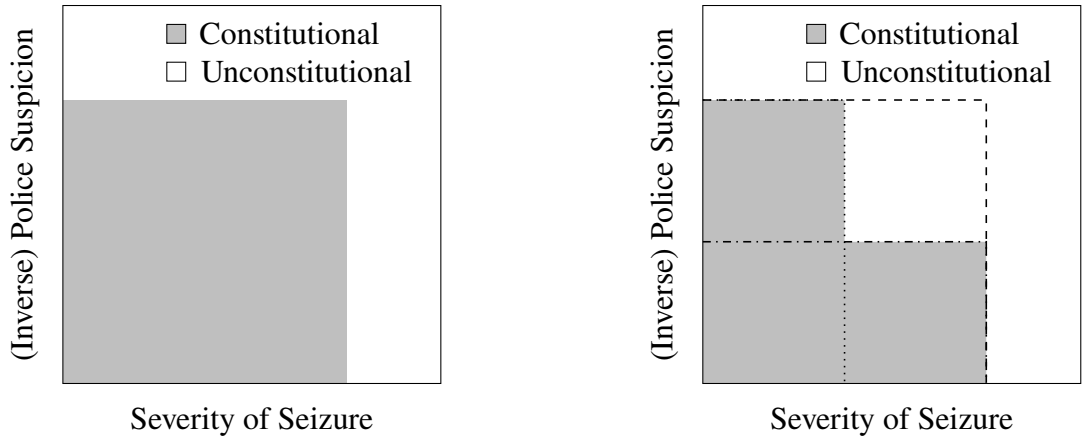
So, we might think of the case space dimensions as the doctrinal concerns of the level of police suspicion and severity of the seizure; an example of a rule in such a space is depicted in Figure 2.1a. In this example, there are some seizures so severe they could never be found constitutional, some circumstances under which there is so little evidence of criminality that no seizure could be constitutional, but as long as the seizure is sufficiently not severe and the police have sufficient certainty that criminal conduct has occurred, the judge will find the seizure was constitutional.

Judges on collegial courts decide cases by majority rule over dispositions. The *implicit collegial rule*, or ICR, is the mapping between cases and outcomes that results from these majority votes over outcomes (Lax 2007, 595). In other words, the ICR represents "the law."[11] An example of a three judge panel's individual preferences and the resulting ICR is depicted in Figure 2.1b. In this case the judges' preferences aggregate to an ICR in which for the lowest range of police suspicion, no seizure is warranted, for a moderate range of police suspicion low levels of seizure are permissible, and at the highest range of police suspicion a much broader range of seizures are found constitutional.

2.3 Model

Case space dimensions that capture high level doctrinal concerns are generated from historical facts, as Justice Scalia discusses in the *Ornelas* exerpt quoted in the introduction. As Lax (2007) explains, "in equal protection cases . . . the dimensions might include (1) how 'suspect' the class invoked is . . . (2) how compelling the state interest is . . . and (3) how necessary the classification is . . . . Or, these dimensions could be broken down further" (594). While the technology of traditional case space models can be used to model decisions based on historical facts, doctrinal concerns, or both, it lacks the ability to model the relationship *between* doctrinal concerns and the dimension of historical facts they are derived from. Abstracting away from this relationship is useful for

11. I join Justice Oliver Wendell Holmes, Jr in claiming, "The prophecies of what the courts will do in fact, and nothing more pretentious, are what I mean by the law" (Holmes 1897, 461).



(a) Judge 1's rule. The shaded region is the set of cases the judge finds the police seizure to be constitutional and in the unshaded region the judge finds the police conduct unconstitutional.

(b) The ICR. Dashed and dotted lines mark the set of cases in which each judge rules police conduct constitutional. In the shaded region, the court as a whole rules it constitutional.

Figure 2.1: An example individual rule and ICR for Fourth Amendment police seizure cases.

Note: The case space is comprised of two dimensions: severity of the police seizure, where larger values indicate a more intrusive seizure, and inverse police suspicion, where larger values indicate less certainty that criminal conduct has occurred.

analyzing other aspects of judicial decision making. However, to understand why outside observers are confused by judicial doctrine, it will be useful to separately represent the high dimensional space of all possible historical facts and the lower dimensional doctrinal space, and the relationship between these spaces.

A legal case presented to a court can be uniquely identified by its historical facts, such as whether a person seized by the police was placed in handcuffs or not, or how long a person was detained. We will say there are N potentially relevant dimensions of historical facts, so that $H \subseteq \mathbb{R}^N$ is the set of all possible combinations of historical facts.

A set of judges J (with $|J|$ odd) must decide cases presented to it from H, and assign them one of two outcomes $\{-1, 1\}$. So, as in other case space models, we will discuss policy as a partition of cases into outcomes. However, judges (and the public they communicate policies to) do not think about policy by considering every possible combination of historical facts, even if they

could. They think about and communicate policy in more abstract terms informed by the historical facts, such as the severity of a police seizure or the degree of police certainty of criminality that supports the seizure. So we also need to define a lower dimensional doctrinal space, $D \subseteq \mathbb{R}^n$, with $1 < n < N$.[12] Then each judge j has a preferred doctrine δ_j mapping D to $\{-1, 1\}$. A doctrine is *monotonic* if for any two points $d, d' \in D$, $d_i \geq d_i{}' \; \forall \; i$ implies $\delta(d) \geq \delta(d')$. We will assume the judges (and other relevant actors such as the public or lower court judges attempting to comply with the collegial appellate court's rulings) can "consistently label" the dimensions of H and D such that higher values of any h_k or d_i should lead to a weakly higher outcome, all else equal.

Unfortunately, as we will see, judges can disagree not only over doctrine, but how historical facts map onto doctrinal facts.[13] Not only could judges disagree whether a particular type of police seizure needs to be supported by probable cause or only by reasonable suspicion, but they could disagree about whether the historical facts support a finding of probable cause or not. So, we add the last moving part to the model: each judge j maps historical facts on to D; I will call this mapping a "fact finding function" $f_j : H \rightarrow D$.[14] For convenience, for a case h and a fact finding function f_j, we will write d_{ij} to mean the ith element of $f_j(h)$. A fact finding function is *monotonic* if for any two points $h, h' \in H$, $h_k \geq h'_k \; \forall \; k$ implies $d_{ij} \geq d'_{ij} \; \forall \; i$. For the remainder of the article, I assume all f_j and δ_j are monotonic.

12. A one dimensional doctrinal space could be possible but would be rare. Some case space models present a simplified unidimensional case even when the lower dimensional abstract doctrine space is still multidimensional. For example, Clark and Carrubba (2012) discuss a unidimensional case space for police search cases, where the dimension is the intrusiveness of the search, when in fact such a formulation must (at least) be some combination of the search's intrusiveness and police certainty of criminality. (Consider that a search at intrusiveness level x that would be constitutional given probable cause may still be unconstitutional in the absence of probable cause). Such an abstraction is useful for studying some questions but not for examining why judicial policy can appear inconsistent to the citizens and other judges charged with following them.

13. Note that Landa and Lax (2008) also discuss the various types of disagreements which judges on a collegial court may face. They make connections from related models to a couple of aspects of disagreement over the mapping between historical facts to doctrinal facts, though none directly address the difficulty discussed in this article.

14. In practice, the judges would be presented with not only the lower court's determination of the case's placement in H given the evidence presented at trial, but also D, but we will ignore for now the lower court's determination of a case's placement in doctrinal space. I discuss appellate courts' deferenece to trial courts' findings as to a case's placement in H and D in Appendix B. In short, appellate judges generally defer to trial courts regarding historical facts, though exceptions can apply in constitutional cases (Hoffman 2001; Redish and Gohl 2017), but are less deferential regarding doctrinal concerns. Readers interested in this procedural issue should consult Appendix B.

In sum, each judge's preferred disposition is thus determined by $\delta_j(f_j(h))$; the judge is presented with the historical facts, they determine how those facts relate to the doctrinal dimensions they find relevant, and thus how the case should be decided according to their preferred doctrine. Thus, a judge's preferred *rule*, or mapping from unique cases to outcomes, is a pair $\rho_j = (f_j, \delta_j)$. This process is depicted in Figure 2.2.

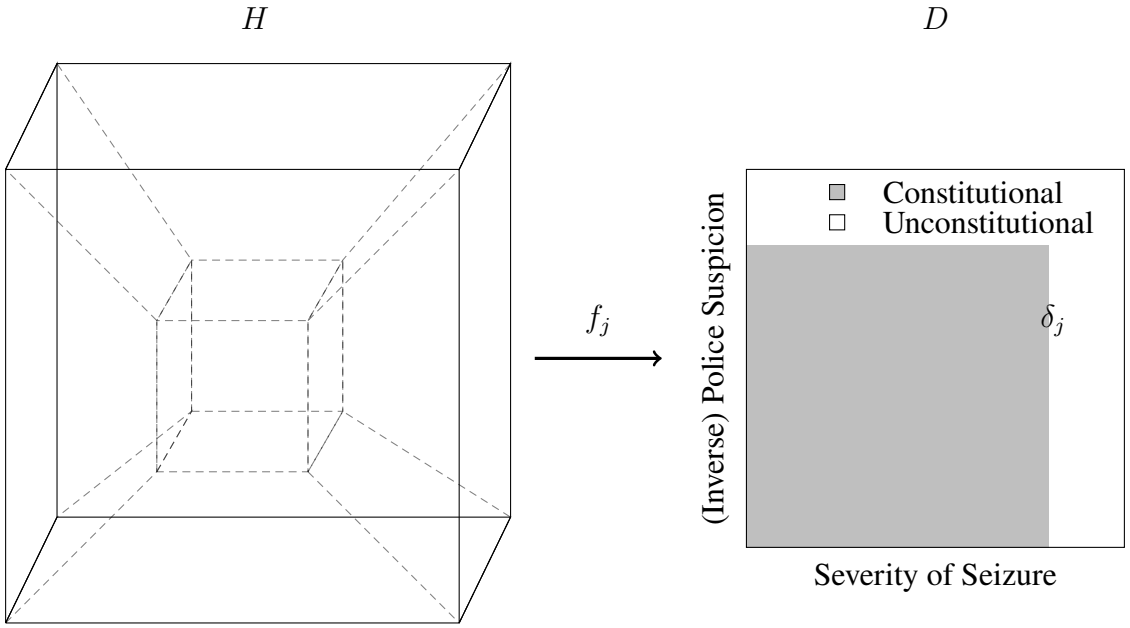


Figure 2.2: Assigning outcomes by translating a fact space to a doctrine space.

Note: A judge j is presented with a set of historical facts, a point in a pontentially high dimensional space H. Cases in this issue area are discussed using broader doctrinal terms—the lower dimensional space D. So, the judge uses the function f_j to translate the case from a point in H to a point in D, the space in which she describes her preferred partition (δ_j) of cases into -1 outcomes and 1 outcomes.

Judges decide cases by majority vote over outcomes. Similarly to Landa and Lax (2009), define an *outcome set* as specifying the outcome (-1 or 1) with each case $h \in H$, and the *collegial outcome set* as the outcome set formed by majority voting among J over the outcome in each case h. A *consistent rule* is a rule $\rho = (f, \delta)$ such that f is monotonic and δ is monotonic in f. The *implicit collegial rule* (ICR) is the rule $\rho_m = (f_m, \delta_m)$ constructed as follows: f_m takes the (dimension by dimension) median value of the f_j for every j in the majority coalition for every

$h \in H$; and δ_m maps D_m to $\{-1, 1\}$ using the collegial outcome set. A summary of notation used is presented in Table 2.1.

Table 2.1: Notation Used

j	A judge on the collegial appellate court.
J	The set of judges on the collegial appellate court.
H	The set of all possible combinations of historical facts.
h_k	One of the N dimensions of H.
D	The set of all possible combinations of doctrinal determinations.
d_i	One of the n dimensions of D.
f_j	The mapping from historical facts to doctrinal dimensions as seen by judge j.
δ_j	The mapping from doctrinal determinations to outcomes preferred by judge j.
ρ	A pair (f, δ) mapping H to outcomes through D such that the outcome in case h is $\delta(f(h))$.
ρ_m	The implicit collegial rule (f_m, δ_m), where f_m takes the (dimension by dimension) median value of the f_j for every j in the majority coalition for every $h \in H$ and δ_m maps D_m to $\{-1, 1\}$ using the collegial outcome set.

2.4 Inconsistency from Multi-step Reasoning

Let us start with the simplest case, where the judges happen to agree on doctrine; that is, δ_j is the same for all j.[15] For example, suppose the judges agree that some seizures of a person are never justified, probable cause is needed to justify others, and that some seizures can be justified merely by reasonable suspicion, but that the judges disagree on the set of historical facts that support a finding of probable cause or reasonable suspicion.

Three types of doctrines in particular will be of interest, both because they are common types

15. This is the case examined by, for example, Kornhauser (1992).

of legal doctrines and because of their aggregation properties. Call a doctrine δ such that

$$\delta(d) = \begin{cases} 1 & \text{if } d \cdot w \geq \tau \\ -1 & \text{otherwise,} \end{cases}$$

where τ is a scalar threshold and w is a vector of weights on the dimensions of D, a *balancing test*.[16] A doctrine δ such that

$$\delta(d) = \begin{cases} 1 & \text{if } d_i \geq \tau_i \; \forall \; i \\ -1 & \text{otherwise,} \end{cases}$$

where τ is a vector of thresholds of length n, shall be called a *conjunctive test*.[17] Finally, define a *disjunctive test* as a doctrine δ such that

$$\delta(d) = \begin{cases} 1 & \text{if } \exists \; i : d_i \geq \tau_i \\ -1 & \text{otherwise,} \end{cases}$$

where τ is a vector of thresholds of length n.

Then we can state the following:

Proposition 2.1. *If all $\delta_j = \delta^*$, and δ^* is a balancing test, then ρ_m is a consistent rule. If δ^* is a conjunctive or disjunctive test, ρ_m need not be a consistent rule.*

Call the situation in the first sentence of Proposition 2.1 a "shared balancing test." Then let $\boldsymbol{\delta} = \{\delta_j\}$ (and similarly for $\mathbf{f}$) and let $\mathcal{F}(\boldsymbol{\delta})$ be the set of combinations of monotonic fact finding functions for the judges such that ρ_m is not consistent given $\boldsymbol{\delta}$ and $\mathbf{f}$. Now we will deal with the more general case where judges may disagree on doctrine and state a more ominous result, which is a more general form of the second sentence in Proposition 2.1:

16. One may note the similarity between such a doctrine and what Landa and Lax (2009) call "base rules."
17. Note that balancing, conjunctive, and disjunctive tests are all monotonic doctrines.

Proposition 2.2. *If $\boldsymbol{\delta}$ is not a shared balancing test, $\mathcal{F}(\boldsymbol{\delta})$ is nonempty.*

The implications of Proposition 2.2 explain a structural reason embedded in our common law system for inconsistent doctrine. Because the judges are engaging in multi-step reasoning to determine case outcomes, in general the court's opinions taken as a whole can be inconsistent in the sense that doctrine is not monotonic in the findings of legal facts. To understand why such monotonicity is crucial, consider a situation in which we have not observed the court's rulings in all of (the infinite number of) the potential cases, nor has the court completely revealed in its opinions ρ_m. (Of course, this is in fact the situation we find ourselves in at all times).[18] Then what we can say about the law, or "the prophecies of what the courts will do in fact" (Holmes 1897), becomes very limited. If δ_m is guaranteed to be monotonic in f_m, we could deduce outcomes in some regions of D_m, and we will have some information about the set of fact finding functions that could be f_m. However, if δ_m is *not* guaranteed to be monotonic in f_m, much less could be said about the outcomes we should expect in cases not observed. Whereas Clark (2016) models the Court's strategy for reducing the uncertainty lower courts (and perhaps other actors) have about outcomes in cases so far unobserved, this result reveals a source of uncertainty courts have no choice over.

Moreover, when the revealed outcomes show δ_m to be non-monotonic in f_m, the collegial doctrine is revealed to be "perverse" (Lax 2007, 594; Landa and Lax 2009, 952).[19] In other words, a person observing two different cases may believe in case one, the police had probable cause and conducted a seizure of a person amounting to an arrest, and in the second case, the police arrested a suspect with *more* evidence of criminality than in the first case, but find the court rules the police

18. See, for example, the discussion of the imperfect ability of judges to communicate their preferences in Clark (2016).

19. Callander and Clark (2017) consider not expecting monotonicity of legal rules, and discuss expecting only a reliance on the dictate that like cases be treated alike (187). This makes sense in the context explored there, where the outcome is not binary, but a latent legal outcome in $\mathbb{R}$. Then similarity across a unidimensional case space that is non-monotonic is easy to understand. However, when we allow ourselves to describe doctrine in multiple dimensions, "wrinkles" or "cut outs" that Callander and Clark (2017) account for with non-monotonicity can often be accounted for by a monotonic rule in a richer historical fact and/or doctrine space. It's also harder to say what similarity in points "close" to each other in the case space is when we are dealing only with the dichotomous outcomes and not a latent legal outcome, absent monotonicity. Additionally, in at least some areas of the law, like the Fourth Amendment example discussed here, monotonicity *is* a normative expectation.

conduct constitutional in the first case but unconstitutional in the second. The ICR may even assign different outcomes to cases at the same location in the doctrine space. For example, a person may view two different set of historical facts and determine that in both cases police had probable cause and conducted a seizure of a person amounting to an arrest, and therefore acted in accordance with the Fourth Amendment, but observe the court rule the actions as constitutional in one case and unconstitutional in the other.

Let us make this example concrete, with $H = [0,1]^4, D = [0,1]^2$, and the judges' fact-finding functions and doctrines as given in Table 2.2.[20] Each of the judges has monotonic doctrines (disjunctive tests) and monotonic fact-finding functions; these are depicted in panels (a)–(c) of Figure 2.3, which show how each judge would place every case in D and which outcome they would choose for those cases if they were deciding cases unilaterally. However, the collegial rule is decidedly inconsistent, as depicted in panel (d), which shows the implicit collegial rule, or how the collegial fact-finding function f_m would place every case in D and which outcome is assigned under the collegial outcome set. Although difficult to depict, in the darkly shaded region where both outcomes occur, the density of cases receiving each outcome varies, and importantly sometimes in an alternating fashion. We see both types of problems mentioned in the previous paragraph: opposing outcomes occurring at the same point in D, and violations of strict monotonicity as well.

Table 2.2: Doctrines and fact-finding functions for the judges on the collegial court.

j	f	δ
1	$(0.5h_1 + 0.5h_2, 0.5h_3 + 0.5h_4)$	$1 \Leftrightarrow d_1 > 0.750 \vee d_2 > 0.750$
2	$(0.6h_1 + 0.4h_2, 0.4h_3 + 0.6h_4)$	$1 \Leftrightarrow d_1 > 0.375 \vee d_2 > 0.750$
3	$(0.4h_1 + 0.6h_2, 0.6h_3 + 0.4h_4)$	$1 \Leftrightarrow d_1 > 0.750 \vee d_2 > 0.375$

We can highlight a few specific cases to make this easier to see. Consider the cases listed in Table 2.3; these cases are labeled with their number in Figure 2.3. In case 1, both judges 2 and 3 find the case satisfies one element of their disjunctive test (though different ones), so both vote

20. This configuration of preferred doctrines is also used in Figure 2.1.

Table 2.3: Example cases showing inconsistency.

Case	h	j	d	Outcome
1	$(0.70, 0.05, 0.75, 0.00)$	1	$(0.375, 0.375)$	1
		2	$(0.440, 0.300)$	
		3	$(0.310, 0.450)$	
		m	$(0.375, 0.375)$	
2	$(0.15, 0.60, 0.15, 0.60)$	1	$(0.375, 0.375)$	-1
		2	$(0.330, 0.420)$	
		3	$(0.420, 0.330)$	
		m	$(0.375, 0.375)$	
3	$(0.15, 0.65, 0.40, 0.40)$	1	$(0.400, 0.400)$	-1
		2	$(0.350, 0.400)$	
		3	$(0.450, 0.400)$	
		m	$(0.375, 0.400)$	

for outcome 1, while in case 2, both judges 1 and 3 find the case satisfies neither element of their disjunctive test, and so both vote for outcome -1. So, while f_m places both cases at $(0.375, 0375)$, they receive opposing outcomes! This is so because while the judges' individual preferences at both levels of aggregation are assumed to be very well behaved, the different levels of aggregation do not always agree with each other. Then, in case 3, judges 1 and 2 find the case satisfies neither element of their disjunctive test, and so both vote for outcome -1, resulting in a case at $f_m = (0.375, 0.4)$ having an outcome of -1 even though a case at $f_m = (0.375, 0.375)$ has an outcome of 1.

Not only is the implicit collegial rule inconsistent, the collegial outcome set is also not monotonic in any of the judges' projection of historical facts into doctrine space, as depicted in Figure 2.4. Table 2.4 singles out three more cases for consideration, all of which are labeled in the panels of Figure 2.4. Cases 4 and 5 cause inconsistency under both f_1 and f_3. For judge 1, cases 4 and 5 occupy the same point in D_1 but receive opposite collegial outcomes. For judge 3, case 4 is more extreme on both doctrinal dimensions than case 5, but receives a -1 outcome where case 5 receives a 1 outcome; that is, this is a situation where in case 4, judge 3 considers that there is both less evidence of criminality and a more severe seizure than in case 5, but the court rules that the seizure

Table 2.4: Further example cases showing inconsistency.

Case	h	j	d	Outcome
4	$(0, 0.875, 0.875, 0)$	1	$(0.438, 0.438)$	-1
		2	$(0.350, 0.350)$	
		3	$(0.525, 0.525)$	
		m	$(0.394, 0.394)$	
5	$(0.125, 0.75, 0.75, 0.125)$	1	$(0.438, 0.438)$	1
		2	$(0.375, 0.375)$	
		3	$(0.500, 0.500)$	
		m	$(0.438, 0.438)$	
6	$(0.875, 0, 0, 0.875)$	1	$(0.438, 0.438)$	-1
		2	$(0.525, 0.525)$	
		3	$(0.350, 0.350)$	
		m	$(0.394, 0.394)$	

in case 4 is constitutional whereas the seizure in case 5 is not. For judge 2, cases 5 and 6 reveal inconsistency in a similar manner to cases 4 and 5 for judge 3.

Of course, in this example, the structure of H is relatively simple, and the f_j appear easy enough to communicate. Even if lower court judges and members of the public have not had a chance to observe the full mapping from H to D_m to outcomes, assuming the judges had full knowledge of their preferences they could simply announce when deciding any case the association between H and the collegial outcome set. However, it is important to note the differences between an easily understood toy example like this and the even worse situation we generally find ourselves in. Generally H will be of a much higher dimensionality than 4; consider our Fourth Amendment example, where it is relevant whether the police restrained the suspect, the duration of the seizure, the credibility of information the police are acting on, what the supect was doing at the time of the seizure, etc. (many of which could be further broken down into multiple historical fact dimensions, but were not for simplicity here). Moreover, *the judges are unlikely to know even their own full mapping* f_j. For example, in the model of Callander and Clark (2017), the High Court does not know with certainty their preferred legal outcome in a particular set of factual circumstances until

they observe such a case, a reasonable assumption in many contexts. Additionally, judges are even commonly presented with new historical factual dimensions that they have never considered before, and they often do not know how such facts affect where the judge will place the case in D until they have occasion to consider it.

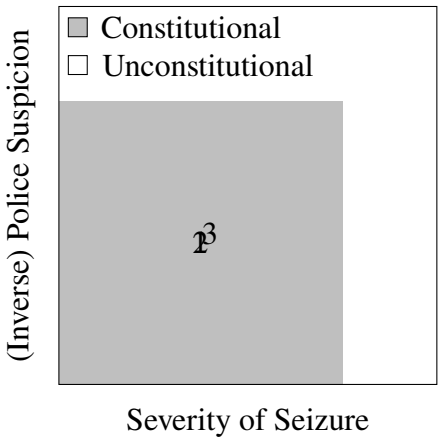
In other words, judges often *cannot* create general doctrinal statements in terms of H; they must communicate their general decision principles in terms of D, and relate cases $h \in H$ to D as they come. In this setting, I have shown a very troublesome result; policies generated and communicated using such multi-step reasoning are generally subject to doctrinal inconsistency.

2.5 Discussion

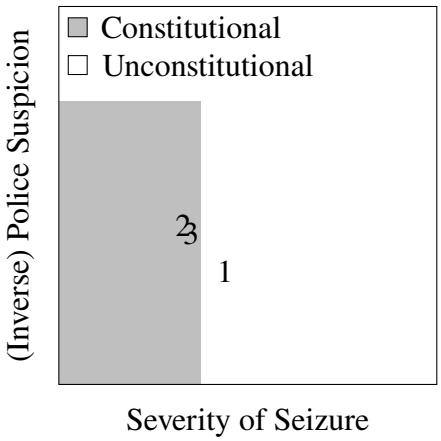
Legal inconsistency is a problem, both for judges as policy makers, since agents and outside actors cannot follow or implement rules they do not understand, and for the public, who might normatively expect consistent application of legal rights. The prior literature offers explanations for inconsistency in individual judges' choices and their preferences (e.g. Collins 2008; Maltzman, Spriggs, and Wahlbeck 2000) or for lack of precision in doctrine (e.g. Clark 2016; Fox and Vanberg 2014; Lax 2012; Staton and Vanberg 2008). Some sources of inconsistency in doctrine have been highlighted by Lax (2007) and Landa and Lax (2008, 2009); by expanding on such case-space models to account for judges' multi-step reasoning, I highlight a new source of legal inconsistency. When we allow for disagreement over both how historical facts should be aggregated to doctrinal dimensions and how the doctrine space should be partitioned into outcomes, the resulting judgment and preference aggregation among judges displays inconsistency even under strict assumptions about how well-behaved the individual judgments and preferences are. The general presence of the danger of this inconsistency explains why so often, courts' doctrines become inconsistent (Drahozal 2004; Post 1995; Robinson and Simon 2006; Will 2019).

This model also raises implications for other areas of judicial politics research. Related to

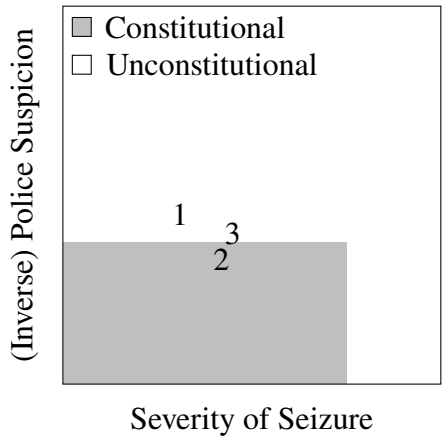
research on the principal-agent relationship between appellate courts and trial courts, when will collegial appellate courts defer to lower court agents' placement of cases in doctrine space to avoid this source of inconsistency on doctrine, and when will they exert more control despite the danger of doctrinal inconsistency shown in this article? For empirical research that uses case facts as explanatory variables, care should be used to recognize that individual judges can have different determinations of their own on such doctrinal facts, and if only historical facts are used, important inconsistencies in the reasoning presented by courts may be obscured.



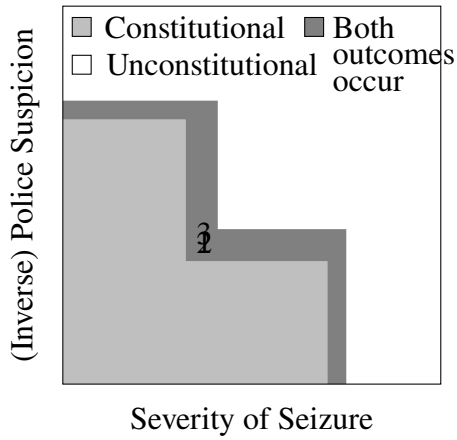
(a) δ_1 with cases placed according to f_1



(b) δ_2 with cases placed according to f_2



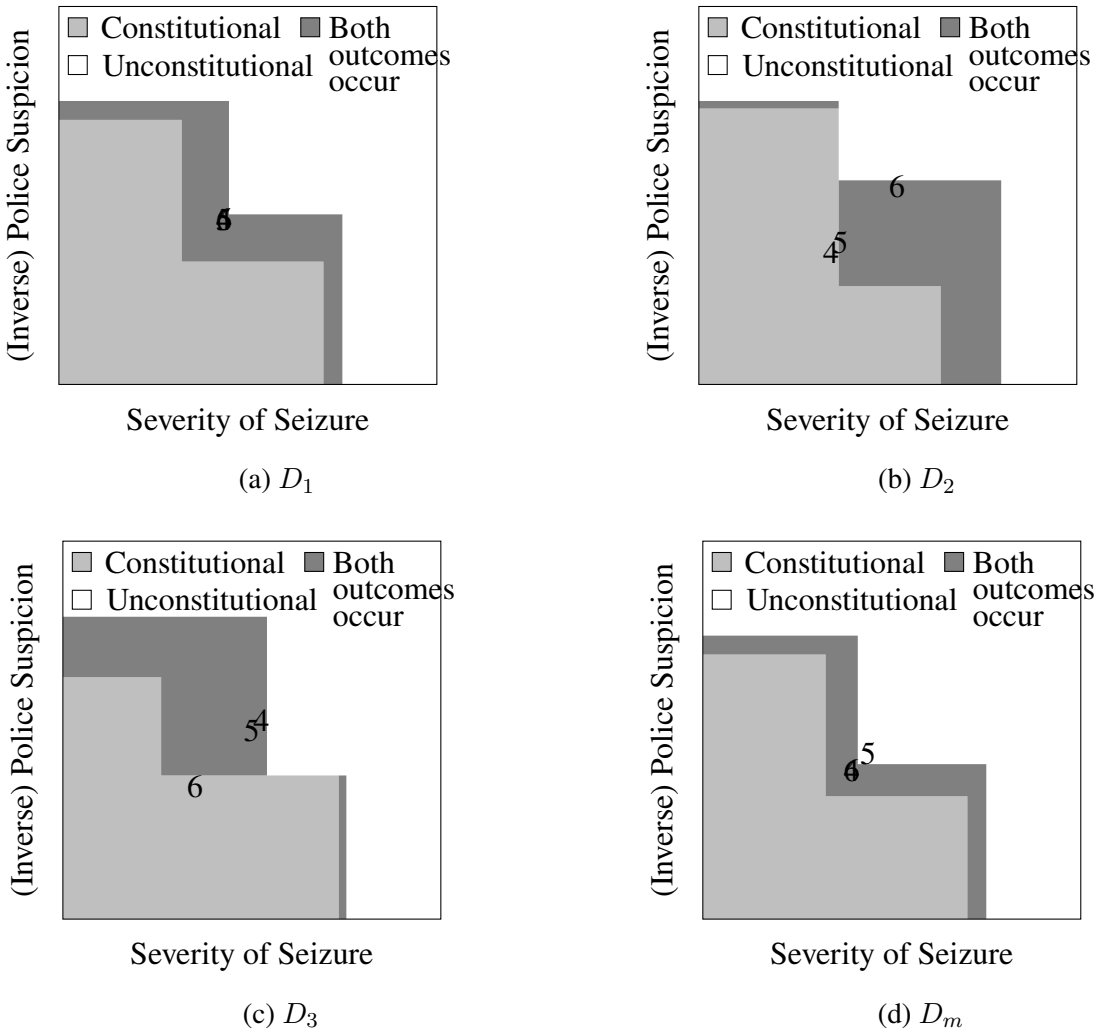
(c) δ_3 with cases placed according to f_3



(d) δ_m in D_m

Figure 2.3: An example of an inconsistent doctrine.

Note: The doctrine space is comprised of two dimensions: severity of the police seizure, where larger values indicate a more intrusive seizure, and inverse police suspicion, where larger values indicate less certainty that criminal conduct has occurred. The judges all have preferred monotonic doctrines and monotonic fact-finding functions; the judges' preferred rules are depicted in panels (a)–(c). However, the implicit collegial rule is inconsistent as depicted in panel (d). In each panel, the three cases from Table 2.3 are labeled with their identifying number.



(a) D_1 (b) D_2

(c) D_3 (d) D_m

Figure 2.4: An example of an inconsistent doctrine.

Note: The doctrine space is comprised of two dimensions: severity of the police seizure, where larger values indicate a more intrusive seizure, and inverse police suspicion, where larger values indicate less certainty that criminal conduct has occurred. The judges all have preferred monotonic doctrines and monotonic fact-finding functions, but the collegial outcome set is not monotonic in any of the judges' projection of historical facts into doctrine space, or even the projection taking the dimension-by-dimension median placement of the majority coalition in every case.

Chapter 3

Inference in Gaussian Process Models for Political Science

Gaussian process (GP) models, a class of machine learning techniques, are increasingly being employed to study politics, from measuring ideology (Gill 2021; Duck-Mayr, Garnett, and Montgomery 2020)[1] to dealing with violations of conditional independence in time-series cross-sectional data (Carlson 2021).[2] GP models are powerful tools to model the relationship between predictors and outcomes when the functional form mapping predictors to response is imperfectly known or observations may not be conditionally independent, common settings in political science. However, inference in the machine learning setting often focuses (sometimes exclusively) on prediction, while other inferential quantities are often of interest to political scientists. I show how these models can be used to obtain quantities of interest to political scientists, including a novel derivation of average marginal effects for GP models.[3] I first provide a primer on GP models, explaining their particular

1. Gill (2021) uses "spatial kriging," which is a type of GP model, to extrapolate ideology measures spatially across the U.S., while Duck-Mayr, Garnett, and Montgomery (2020) develop a novel GP item response theoretic model (GPIRT).

2. I would like to thank Jacob Montgomery and participants at Washington University in St. Louis' Political Data Science Lab for their helpful comments.

3. The full derivations can be found in the appendix; in the main body of the paper I present results and practical guidance.

import for political science specifically. I highlight their nascent use in political science, briefly explaining existing approaches to inference with GP models in the political science literature. I then outline how to obtain a number of other quantities of interest and provide practical guidance in the use of these models to enable the discipline to better harness these powerful tools for social scientific inference.

3.1 A Primer on GP Models

Generally we want to reason about the relationship between predictors X and outcomes y. So, we generally say

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \tag{3.1}$$

where y_i is our observed outcome for observation i, $\mathbf{x}_i$ is our observed vector of predictors for observation i, and ε_i is the error term—some added random noise. Then, we want to learn about $f(X)$. A stereotypical approach in political science is to assume a functional form for f, and that the noise elements ε_i are independently distributed. In that case our task is to perform inference on the parameters of f. A number of non-parametric approaches are available when the form of f is unknown, and a variety of statistical fixes have been developed for various correlation structures of $\boldsymbol{\varepsilon}$.

A non-parametric approach common in the machine learning literature and now starting to gain traction in political science is to model f as a *Gaussian process* (GP) (Rasmussen and Williams 2006).[4] While there are many methods that have been developed to accomplish non-parameteric inference or handle error correlation, GP models have risen to prominence because

4. Rasmussen and Williams (2006) is a comprehensive textbook for GP classification and regression. A reader seeking a treatment that is much more in-depth should consult Rasmussen and Williams (2006).

in addition to their flexibility, they represent a principled, probabilistic approach that presents a very general framework applicable in a variety of settings (Cheng et al. 2019). Moreover, they can outperform even tailored models for many inferential problems; for example, Carlson (2021) finds GP regression to be more effective at handling error correlation in time-series cross-sectional data than other existing approaches.

A GP is an infinite dimensional generalization of the normal distribution, where any finite subcollection of the process' variables are normally distributed. This is accomplished by specifying the mean and covariance of the distribution as a function of the predictors:

$$f \sim GP\left(\mu\left(X\right), K\left(X, X\right)\right), \tag{3.2}$$

where μ is a function (such as, for example, $X\beta$ for a vector of coefficients β) that gives the mean of the distribution of f at X and $K\left(X, X'\right)$ is a matrix-valued function that gives the covariance between values of f at X and X', such that for any finite set of observations $\mathbf{X}$, $\mathbf{f} = f\left(\mathbf{X}\right)$ has a prior distribution

$$\mathbf{f} \sim N\left(\mu\left(\mathbf{X}\right), K\left(\mathbf{X}, \mathbf{X}\right)\right). \tag{3.3}$$

In the regression case, where y is continuous and we use a normal likelihood with variance σ_y^2, we can then learn about f after observing $\mathbf{X}$ and $\mathbf{y}$ by appyling Bayes' theorem along with Gaussian identities to derive the exact posterior over $\mathbf{f}$,

$$\mathbf{f} \mid \mathbf{X}, \mathbf{y} \sim N\left(\mathbf{m}^*, \mathbf{C}^*\right), \tag{3.4}$$

$$\mathbf{m}^* = \mu\left(\mathbf{X}\right) + KK_y^{-1}\left(\mathbf{y} - \mu\left(\mathbf{X}\right)\right), \tag{3.5}$$

$$\mathbf{C}^* = K - KK_y^{-1}K, \tag{3.6}$$

$$K_y = K + \sigma_y^2 I, \tag{3.7}$$

where we write $K = K(X, X)$ for more compact notation.[5] Notice, then, that (for example) Bayesian linear regression is simply a special case of GP regression; GP regression with $\mu(X) = 0$ and $K(X) = XX^T$ returns the same solution as linear regression with standard normal priors on the coefficients.

This allows us to learn about *potentially* nonlinear functions of X with very mild assumptions. The assumptions we are making about f are largely through our choice of the mean function μ and the covariance function, or *kernel*, K. The common choice in the machine learning literature is to choose $\mu(X) = 0$, giving a vague prior over f where all learning about f goes through the kernel. We may also choose a linear mean, $\mu(X) = X\beta$, encoding an assumption that f should have a linear trend, though perhaps with nonlinear deviations or correlated errors.

An overwhelmingly popular choice for the kernel is the squared exponential covariance function, in which the covariance between $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ is given by

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \left(-0.5 \sum_d \frac{(x_{id} - x_{jd})^2}{\ell_d^2} \right), \tag{3.8}$$

where σ_f^2 is a *scale factor* scaling the entire prior covariance matrix and $\boldsymbol{\ell}$ is a vector of *length scales*. This kernel corresponds to assuming that (1) f is smooth,[6] and (2) the correlation between values of f should decrease with distance in the covariate space. Then $\boldsymbol{\ell}$ determines how we define "closeness" along each dimension of X. Often an isotropic version of this kernel is used where ℓ is instead a scalar, treating distance in every dimension the same. This kernel should similarly be

5. Although this derivation is available in multiple sources in varying levels of detail, including Rasmussen and Williams (2006), I provide a derivation in the appendix as well.

6. For scholars interested in modeling functions that are not smooth, other kernel options are available with a similar proximity assumption. Please consult Chapter 4 in Rasmussen and Williams (2006) for details.

most useful in political science; these assumptions match up to problems where we must account for correlated errors across space or time (Carlson 2021; Gill 2021), and notably is also equivalent to a linear model with infinite basis expansion.

To make this more concrete, consider the following example, where

$$f(x) = 2\sin(2x) + x, \tag{3.9}$$

$$y = f(x) + \varepsilon, \tag{3.10}$$

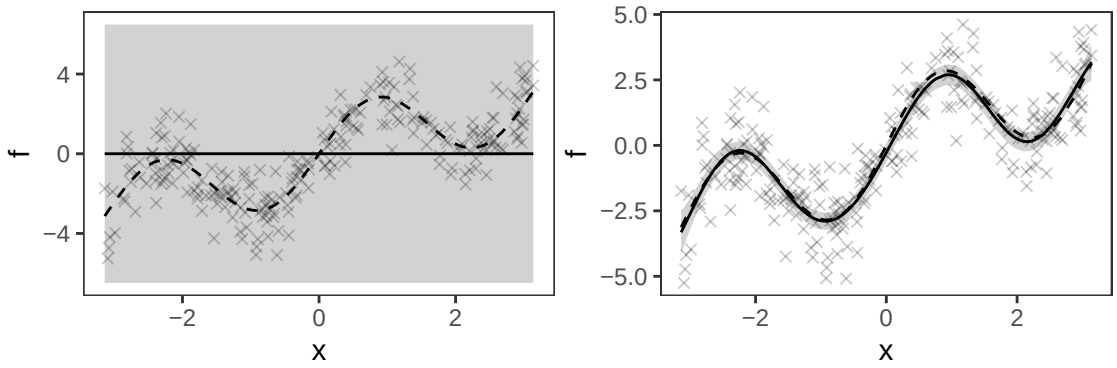
$$\varepsilon \sim N(0, 1). \tag{3.11}$$

So, we have a function mapping the single predictor variable x to outcomes y with a linear trend and we have independent noise, but the function also has systematic nonlinear deviations. Figure 3.1a shows a vague, zero-mean GP prior over f, using the squared exponential covariance function. I simulated 250 x values, drawn from a uniform distribution with bounds $-\pi$ and π (which allows for two full oscillations of f), then simulated corresponding y values with standard normal noise. We can see the posterior from Equation (3.4) depicted in Figure 3.1b; essentially we have taken the somewhat mild assumptions encoded by the covariance function that f is smooth and that covariance between function outputs decreases with distance in x to derive a reasonable estimate of f (depicted with the solid line) with a measure of uncertainty (depicted with the shaded region).

For modeling discrete outcomes, the posterior becomes intractable; however, good approximations of the posterior in important cases have been derived. For example, for dichotomous outcomes, we simply say

$$\Pr(y_i = 1) = \sigma(f(\mathbf{x}_i)), \tag{3.12}$$

$$f \sim GP(\mu(X), K(X)), \tag{3.13}$$



(a) The GP prior (b) The GP regression posterior

Figure 3.1: An example GP prior and posterior for the function $f(x) = 2\sin(2x) + x$. Simulated data points are depicted with crosses, the prior (posterior) mean with a solid line and the 95% CI with a shaded region, and the true $f(x)$ with a dashed line.

where f is then a latent function fed though σ, which is some sigmoid "squashing" function mapping the reals to $[0, 1]$, such as using a logistic or probit likelihood, to obtain the probability of a postitive response. This gives us a very similar setup to a generalized linear model such as the probit or logit regression, with the difference being that we do not assume as much about the structure of the latent function f. While the posterior does not have a closed form as in the regression case, we can solve for a Laplace approximation to the posterior centered at the posterior mode $\hat{\mathbf{f}}$,

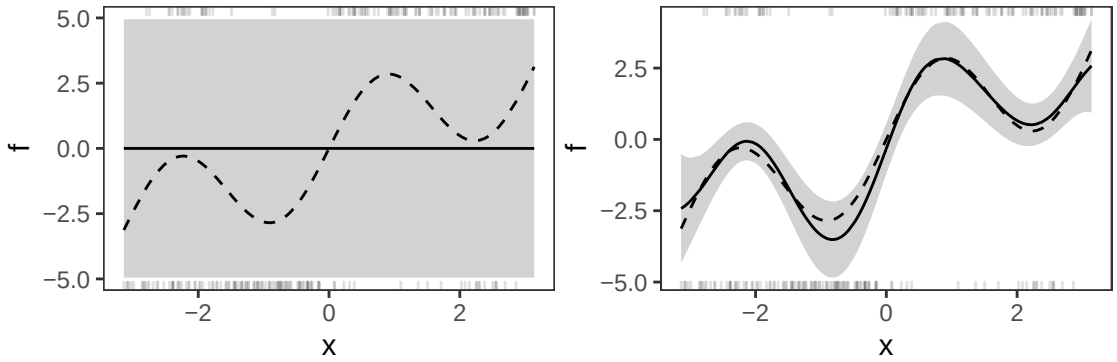
$$\mathbf{f} \mid \mathbf{X}, \mathbf{y} \sim N\left(\hat{\mathbf{f}}, \left(K^{-1} + W\right)^{-1}\right), \tag{3.14}$$

$$\hat{\mathbf{f}} = \mu(\mathbf{X}) + K\left(\nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right)\right), \tag{3.15}$$

$$W = -\nabla\nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right), \tag{3.16}$$

and other approximations based on minimizing Kullbeck-Liebler divergence are available as well, in addition to being able to simulate the posterior via MCMC sampling, commonly utilizing an elliptical slice sampler (Murray, Adams, and MacKay 2010).

Taking the same example function and simulated x values we used to illustrate GP regression,



(a) The GP prior (b) The GP regression posterior

Figure 3.2: An example GP prior and posterior for the function $f(x) = 2\sin(2x) + x$, where the data were simulated as $x \sim U(-\pi, \pi), \Pr(y = 1) = \sigma(f(x))$, where σ is the logistic function. Observations receiving positive labels are depicted in the rug on the top margin, while observations receiving negative labels are depicted in the rug on the bottom margin; the prior (posterior) mean with a solid line and the 95% CI with a shaded region, and the true $f(x)$ with a dashed line.

I draw corresponding dichotomous y values where $\Pr(y = 1) = \sigma(f(x))$, where σ is the logistic function. A zero mean GP prior with squared exponential covariance function is depicted in Figure 3.2a, and the Laplace approximation of the posterior from Equation (3.14) is depicted in Figure 3.2b.

Modelling the relationship between predictors and outcomes as a GP offers a flexible but principled approach with a number of advantages over other approaches. Unlike parametric approaches, we can be agnostic *a priori* as to the shape of the relationship between predictors and outcomes, accounting for our typical uncertainty over functional form as social scientists. When compared to many non-parametric approaches that are similarly agnostic, the GP approach provides a more principled probabilistic approach that is more readily extended to varied settings. Finally, as I show in Section 3.3, the GP approach still allows us to recover and make probabilistic statements about inferential quantities of interest to political scientists.

For many political science applications, Gaussian process regression or classification represents a useful approach to guard against misspecification bias and handle error correlation, though there are some tradeoffs to consider. The main drawback of the method is scalability; as you can

see in Equations (3.4) and (3.14), obtaining the posterior distribution requires matrix inversion, which scales cubically with the number of observations, though there are methods to ameliorate the scaling problem at the cost of some precision (Liu et al. 2020). Additionally, if there are very strong theoretical reasons to believe (1) the exact model specification is known, so that we need not guard against misspecification bias, and (2) we can actually credibly assume conditional independence of observations, slightly more precision in estimation may be gained from parametric models over the GP models. Finally, we may worry about overfitting, particularly when discussing relationships found in the data that were not hypothesized. However, in model selection steps where hyperparameters are chosen to maximize the log marginal likelihood (see Section 3.3), the log marginal likelihood includes a complexity penalty term to help guard against this, resulting in a sort of "automatic trade-off" between data fit and model complexity (Rasmussen and Williams 2006, 110–113). Moreover, this concern can be easily ameliorated via k-fold cross-validation to ensure overfitting has not occurred (Rasmussen and Williams 2006, 109, 111–112).

3.2 GP Models in Political Science

While GP models have a longer history in statistics and machine learning, they are just beginning to take hold in the study of politics. The GP approach is particularly suited to the study of politics as political scientists often confront situations in which we should acknowledge some uncertainty regarding functional form, or (as may be the modal case in political science) the common assumption of independent errors is violated.

Carlson (2021) considers time series cross-sectional (TSCS) data, common in many areas of political science, and recommends GP regression for those settings. Carlson (2021) shows GP regression outperforms a variety of previous approaches such as lagged dependent variable, fixed effects, and random effects specifications, as well as panel-corrected standard errors in the TSCS setting. Gill (2021) similarly utilizes a GP model to handle spatial autocorrelation.

Duck-Mayr, Garnett, and Montgomery (2020) develop an IRT model where, rather than assuming the functional form of the response functions, a GP prior is placed over latent response functions fed into a logistic likelihood. Among their applications demonstrating the method, the authors estimate ideology of members of the House of Representatives in the 116th U.S. Congress. They show this more flexible approach that acknowledges uncertainty over the functional form of the roll call votes' response functions allows for more plausible estimates of extremist members' ideology; while parametric methods that impose monotonicity of responses in ideology force extreme members such as Rep. Alexandria Ocasio-Cortez who often vote against the moderate proposals of her own party to be placed closer to the opposing party, a flexible GP approach can recognize that members such as she should be placed at the end of the ideological spectrum and instead allow the item response function to bend downwards in such situations.

As political scientists are increasingly taking advantage of the flexible GP approach to handle data that pose difficulties for inference using traditional approaches, I next provide practical guidance for applied researchers and derive distributions of inferential quantities of interest to political scientists that are not covered in the machine learning literature where the lion's share of study of GP models has occurred.

3.3 Tools for Inference in GP Models

A common goal for those employing GP models is out of sample prediction, so the typical inferential quantity of interest is the distribution of the unknown function at test points. The machine learning literature on GP models has largely focused on this sort of inference in various settings. Political scientists are more often interested in parsing out the effects of the predictors. After explaining the usual process for inference in GP models, I will also show how to derive two types of effects of predictors: the distribution of coefficients if a linear mean function is employed, and the distribution of average marginal effects of predictors whether or not a linear mean function is employed. The

former is useful for determining the contribution of a predictor to a linear trend within f, while the latter is necessary for uncovering the full average effect of a predictor on outcomes, since we allow for a potentially nonlinear relationship between predictors and outcomes. A convenient interface to obtain the distribution of all of these quantities is provided in an `R` package.[7]

When the goal is prediction, inference for GP models often follows the following sequence: first, make choices about the GP prior, including the structure of the mean and covariance functions; next, set the parameters of the mean and covariance functions (such as the coefficients of a linear mean function, or the scale factor of a squared exponential covariance function) as a model selection step by choosing them to use the GP prior that maximizes the log marginal likelihood of the model given the training data; finally, use the training data and the selected hyperparameters to predict out of sample for test data. This workflow is very effective for generating probabilistic predictions for unknown data generating processes. The posterior predictive distribution for GP regression at test cases $\mathbf{X}^*$ is

$$\mathbf{f}^* \mid \mathbf{y}, \mathbf{X} \sim N\left(\mu\left(\mathbf{X}^*\right) + K_* K_y^{-1}\left(\mathbf{y} - \mu\left(\mathbf{X}\right)\right), K_{**} - K_* K_y^{-1} K_*^T\right), \tag{3.17}$$

with a common shorthand of $K_* = K\left(\mathbf{X}^*, \mathbf{X}\right)$ and $K_{**} = K\left(\mathbf{X}^*, \mathbf{X}^*\right)$, and the posterior predictive distribution for classification (using the Laplace approximation) is

$$\mathbf{f}^* \mid \mathbf{y}, \mathbf{X} \sim N\left(\mu\left(\mathbf{X}^*\right) + K_* \nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right), K_{**} - K_*\left(K + W^{-1}\right)^{-1} K_*^T\right). \tag{3.18}$$

However, often other inferential quantities are of more interest than out of sample predictions.

7. A number of packages are available for fitting and predicting out of sample for GP models, both in R and in a number of other programming languages and software environments including python and MATLAB. However, no currently available software implements average marginal effects or provides posterior inference over linear mean function coefficients.

For example, suppose you believe the important underlying relationship between your predictors and outcomes is in fact linear, but you also want to explicitly model and account for correlated errors, as in Carlson (2021). Then you may use a model of the following form:

$$y \sim N\left(f\left(X\right), \sigma_y^2 I\right), \tag{3.19}$$

$$f \sim GP\left(X\beta, K\left(X, X\right)\right), \tag{3.20}$$

where your quantity of interest is β. Then it would not be useful to treat the mean function parameters as a model selection problem; rather we want to find the distribution of those parameters themselves instead of the distribution of f. One approach would be to place priors on the mean function *and* covariance function parameters to get the posterior distribution of all the GP prior's hyperparameters; this approach, taken in Carlson (2021), however, generally requires MCMC sampling, as priors on the covariance function hyperparameters generally result in the posterior being analytically intractible.[8]

If the covariance function parameters are not directly of interest, however, we can set those using Bayesian model selection as in the general prediction workflow; then, with covariance function parameters in hand, the posterior distribution of β is

$$\beta \mid \mathbf{y}, \mathbf{X} \sim N(\bar{\beta}, \Sigma_\beta), \tag{3.21}$$

$$\bar{\beta} = (B^{-1} + \mathbf{X}^T K_y^{-1} \mathbf{X})^{-1} (B^{-1}\mathbf{b} + \mathbf{X}^T K_y^{-1} \mathbf{y}), \tag{3.22}$$

$$\Sigma_\beta = (B^{-1} + \mathbf{X}^T K_y^{-1} \mathbf{X})^{-1}, \tag{3.23}$$

8. A bespoke MCMC sampler for GP regression with priors on all hyperparameters is also offered in the R package, which provides samples *much* faster than the general-purpose Stan implementation used for Carlson (2021).

where $\mathbf{b}$ is the prior mean of β and B is the prior covariance of β.[9]

However, if our motivation for using a GP approach is flexibility in the form of f rather than being interested only in posterior inference over a linear trend contained in f, we should instead perform inference on the *average marginal effect* of our predictors. As this is not a task common in the use of GP models in computer science, the machine learning literature on GP models provides no explicit derivation for this quantity, although for other reasons, building blocks we need for it have previously been derived.

To formalize, we want to reason about the relationship between predictors X and outcomes y, and specifically wish to know the marginal effect of a particular predictor d, i.e., the dth feature of X. In a parametric regression model where we assume $f(X) = X\beta$ and simply estimate β, the marginal effect of X_d is easy to see: $\frac{\partial f(\mathbf{x}_i)}{\partial x_{id}} = \hat{\beta}_d, \forall i$. In GP regression and classification, we gain a much more flexible model that allows for non-independence of observations and non-linear mappings from X to y, but then feature d does not have a constant effect on y. We can summarize the effect of feature d on y with the sample average marginal effect,[10] which in the regression context is defined as

$$\gamma_d \triangleq \frac{1}{N} \sum_{i=1}^{N} \frac{\partial f(\mathbf{x}_i)}{\partial x_{id}}. \tag{3.24}$$

For classification, γ_d gives us the sample average partial effect, or the sample average effect on the latent function f, which does not directly translate to the average marginal effect on our dichotomous outcomes y. In this case, often of more interest than γ_d is

$$\pi_d \triangleq \frac{1}{N} \sum_{i}^{N} \frac{\partial \sigma(f(\mathbf{x}_i))}{\partial x_{id}}, \tag{3.25}$$

which gives the average marginal effect of predictor d on the function $\sigma(f(X))$ that gives the probability of a positive response.

9. See Rasmussen and Williams (2006) Section 2.7. A full derivation is also offered in the appendix.
10. See Hainmueller and Hazlett (2014) for a similar approach with kernel-regularized least squares.

Moreover, when d is discrete, instanteous change in f at our observed points is not meaningful; then we want the average discrete change in f at levels of predictor d. For binary variables, let $\mathbf{X}_1^*$ be a set of test points where all feature observactions are identical to $\mathbf{X}$ except that all observations of feature d have been set to 1, and analogously for $\mathbf{X}_0^*$.[11] Then a more appropriate quantity of interest rather than γ_d is

$$\delta_d \triangleq \frac{1}{N}\sum_{i=1}^{N} f\left(\mathbf{x}_{1i}^*\right) - f\left(\mathbf{x}_{0i}^*\right), \tag{3.26}$$

which gives the average marginal effect on y of taking a 1 vs a 0 value in the regression case, or the average partial effect on f of taking a 1 vs a 0 value in the classification case. For classification, the effect on the probability scale is

$$\psi_d \triangleq \frac{1}{N}\sum_{i=1}^{N} \sigma\left(f\left(\mathbf{x}_{1i}^*\right)\right) - \sigma\left(f\left(\mathbf{x}_{0i}^*\right)\right). \tag{3.27}$$

For categorical variables, we simply find δ_d or ψ_d for all substantively interesting pairwise comparisons of levels of the categorical variable. (Often this is comparing the various categorical labels to one "baseline" label).

Our starting point for deriving these quantities is noting that "[s]ince differentiation is a linear operator, the derivative of a Gaussian process is another Gaussian process" (Rasmussen and Williams 2006, 191). Let

$$\mathbf{f}_d = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_{1d}} \\ \vdots \\ \dfrac{\partial f_n}{\partial x_{nd}} \end{bmatrix}. \tag{3.28}$$

Using Equation 9.1 in Rasmussen and Williams (2006),

11. You can replace 1 and 0 with other binary value labels as needed.

$$K_d \triangleq \mathbb{C}\left[\mathbf{f}_d, \mathbf{f}\right] = \begin{bmatrix} \frac{\partial\, k\left(\mathbf{x}_1, \mathbf{x}_1\right)}{\partial\, x_{1d}} & \cdots & \frac{\partial\, k\left(\mathbf{x}_1, \mathbf{x}_n\right)}{\partial\, x_{1d}} \\ \vdots & \ddots & \vdots \\ \frac{\partial\, k\left(\mathbf{x}_n, \mathbf{x}_1\right)}{\partial\, x_{nd}} & \cdots & \frac{\partial\, k\left(\mathbf{x}_n, \mathbf{x}_n\right)}{\partial\, x_{nd}} \end{bmatrix}, \tag{3.29}$$

$$K_{dd} \triangleq \mathbb{C}\left[\mathbf{f}_d, \mathbf{f}_d\right] = \begin{bmatrix} \frac{\partial^2 k\left(\mathbf{x}_1, \mathbf{x}_1\right)}{\partial\, x_{1d}\, \partial\, x_{1d}} & \cdots & \frac{\partial^2 k\left(\mathbf{x}_1, \mathbf{x}_n\right)}{\partial\, x_{1d}\, \partial\, x_{nd}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 k\left(\mathbf{x}_n, \mathbf{x}_1\right)}{\partial\, x_{nd}\, \partial\, x_{1d}} & \cdots & \frac{\partial^2 k\left(\mathbf{x}_n, \mathbf{x}_n\right)}{\partial\, x_{nd}\, \partial\, x_{nd}} \end{bmatrix}, \tag{3.30}$$

To make the notation more compact, as is usual we set $K = K\left(\mathbf{X}, \mathbf{X}\right)$, and additionally set $\mu = \mu\left(\mathbf{X}\right)$ and

$$\mu_d = \begin{bmatrix} \frac{\partial\, \mu\left(\mathbf{x}_1\right)}{\partial\, x_{1d}} \\ \vdots \\ \frac{\partial\, \mu\left(\mathbf{x}_n\right)}{\partial\, x_{nd}} \end{bmatrix}. \tag{3.31}$$

Then we can describe the joint prior on $\mathbf{f}$ and $\mathbf{f}_d$:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_d \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu_d \end{bmatrix}, \begin{bmatrix} K & K_d^T \\ K_d & K_{dd} \end{bmatrix} \right), \tag{3.32}$$

and for regression and under normal approximations of the posterior for classification,[12] the posterior distribution of $\mathbf{f}_d$ given $\mathbf{X}$ and $\mathbf{y}$ is normal with mean

12. When simulating the posterior for classification rather than using an analytical approximation, $\mathbf{f}_d$ is normally distributed given each $\mathbf{f}$ draw (with mean $\mu_d + K_d K^{-1}\left(\mathbf{f} - \mu\right)$ and variance $K_{dd} - K_d K^{-1} K_d^T$), so $\mathbf{f}_d$ samples can simply be taken conditioned on the $\mathbf{f}$ samples.

$$\mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = \int \mathbb{E}\left[\mathbf{f}_d \mid \mathbf{f}, \mathbf{X}\right] p\left(\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right) d\mathbf{f}$$
$$= \mu_d + K_d\, K^{-1} \left(\mathbb{E}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] - \mu\right), \tag{3.33}$$

and variance

$$\mathbb{V}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = K_{dd} - K_d\, K^{-1}\, K_d^T + \mathbb{E}\left[\left(\mathbb{E}\left[\mathbf{f}_d \mid \mathbf{f}\right] - \mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right]\right)^2\right]$$
$$= K_{dd} - K_d \left(K^{-1} - K^{-1}\, \mathbb{V}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] K^{-1}\right) K_d^T. \tag{3.34}$$

(The full derivations for all results in this section are provided in the appendix for the interested reader). In the regression case,

$$\mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = \mu_d + K_d\, K_y^{-1} \left(\mathbf{y} - \mu\right), \tag{3.35}$$
$$\mathbb{V}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = K_{dd} - K_d\, K_y^{-1}\, K_d^T, \tag{3.36}$$

For classification, under the Laplace approximation to the posterior,

$$\mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = \mu_d + K_d \left(\nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right)\right), \tag{3.37}$$
$$\mathbb{V}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = K_{dd} - K_d \left(K + W^{-1}\right)^{-1} K_d^T. \tag{3.38}$$

Since then γ_d is a constant ($1/N$) times the sum of correlated normal random variables,

$$\gamma_d \sim N\left(\frac{1}{N}\sum_{i=1}^{N} m_{\gamma_d i}\,,\ \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} c_{\gamma_d ij}\right), \tag{3.39}$$

$$\mathbf{m}_{\gamma_d} = \mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] \tag{3.40}$$

$$\mathbf{C}_{\gamma_d} = \mathbb{V}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] \tag{3.41}$$

Importantly, we may also get the average marginal effect of feature d within subgroups of $\mathbf{X}$ rather than the full sample average by simply altering the indices of summation in Equation (3.39). The distribution of δ_d is analogous:

$$\delta_d \sim N\left(\frac{1}{N}\sum_{i=1}^{N} m_{\delta_d i}\,,\ \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} c_{\delta_d ij}\right), \tag{3.42}$$

$$\mathbf{m}_{\delta_d} = \mathbb{E}\left[\mathbf{f}_1^* \mid \mathbf{X}, \mathbf{y}\right] - \mathbb{E}\left[\mathbf{f}_0^* \mid \mathbf{X}, \mathbf{y}\right], \tag{3.43}$$

$$\mathbf{C}_{\delta_d} = \mathbb{V}\left[\mathbf{f}_1^* \mid \mathbf{X}, \mathbf{y}\right] + \mathbb{V}\left[\mathbf{f}_0^* \mid \mathbf{X}, \mathbf{y}\right] + \mathbb{C}\left[\mathbf{f}_1^*, \mathbf{f}_0^* \mid \mathbf{X}, \mathbf{y}\right] + \mathbb{C}\left[\mathbf{f}_0^*, \mathbf{f}_1^* \mid \mathbf{X}, \mathbf{y}\right]. \tag{3.44}$$

Unfortunately, the distribution π_d cannot be analytically expressed, though we can readily simulate from it. First note that

$$\frac{\partial\, \sigma\left(f\left(\mathbf{x}_i\right)\right)}{\partial\, x_{id}} = \frac{\partial\, \sigma\left(f\left(\mathbf{x}_i\right)\right)}{\partial\, f}\frac{\partial\, f\left(\mathbf{x}_i\right)}{\partial\, x_{id}}. \tag{3.45}$$

Generally the sigmoid function σ has a known derivative; for example, in the logistic case,

$$\frac{\partial\, \sigma\left(f\left(\mathbf{x}_i\right)\right)}{\partial\, f} = \sigma\left(f\left(\mathbf{x}_i\right)\right)\left(1 - \sigma\left(f\left(\mathbf{x}_i\right)\right)\right). \tag{3.46}$$

Since we have an approximation to the posterior on f, and given f, the posterior over $\mathbf{f}_d$ is

$$\mathbf{f}_d \mid \mathbf{f} \sim N\left(\mu_d + K_d\, K^{-1}\left(\mathbf{f} - \mu\right), K_{dd} - K_d\, K^{-1}\, K_d^T\right), \tag{3.47}$$

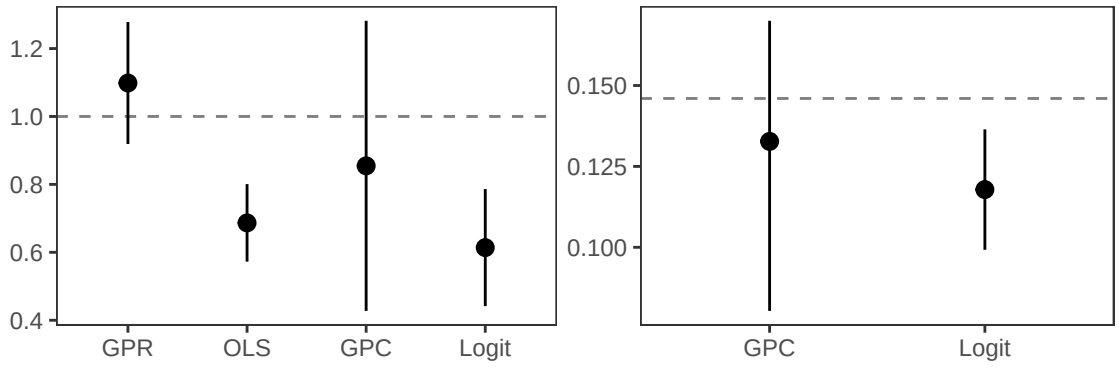
we can obtain M samples of π_d by

- drawing $\mathbf{f}^t$ from the chosen posterior approximation, such as the Laplace approximation $N\left(\mu + K\left(\nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right)\right), \left(K^{-1} + W\right)^{-1}\right)$,[13]
- drawing $\mathbf{f}_d^t$ from $N\left(\mu_d + K_d\, K^{-1}\left(\mathbf{f}^t - \mu\right), K_{dd} - K_d\, K^{-1}\, K_d^T\right)$,
- and calculating $\pi_d^t = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \frac{\partial\, \sigma\left(f_i^t\right)}{\partial\, f} f_{id}^t$,

so that we can summarize the distribution of π_d using the M draws, similar to the CLARIFY procedure (King, Tomz, and Wittenberg 2000). We can also similarly simulate the distribution of ψ_d by simply generating values of $f\left(\mathbf{X}_1^*\right)$ and $f\left(\mathbf{X}_0^*\right)$ from the posterior approximation, pushing those samples through the chosen sigmoid σ, and calculate the average of the differences to get the average marginal effect for each sample so that we can summarize the distribution of average marginal effects.

We can illustrate the use of average marginal effects in GP models by returning to our previous example. For the function $f(x) = 2\sin(x) + x$ where $x \sim U(-\pi, \pi)$, the true average marginal effect of x on $f(x)$ is 1—the oscillations cancel out and we are left with the effect of the linear term. In other words, on average, $f(x)$ increases with x at a rate of 1, which is often the type of information we want to have about functions of interest as political scientists. For the dichotomous simulated data we may also be interested in the average slope of $\sigma\left(f\left(x\right)\right)$; here, the true average marginal effect of x on $\sigma\left(f\left(x\right)\right)$ is 0.146. Figure 3.3 shows the posterior means and 95% CIs for the average marginal effects in the regression and classification cases, with linear model baselines

13. At this point, since we have resorted to simulation, it may be tempting to use ESS to draw from the posterior. This can be done, but then we lose the ability to perform simple model selection for GP prior hyperparameters, resulting in markedly increased computation time.



(a) AME of x on $f(x)$ (b) AME of x on $\sigma(f(x))$

Figure 3.3: Average marginal effect (AME) of x. The left panel shows the AME of x on $f(x)$ for both regression and classification; the right panel shows the AME of x on the probability of a positive outcome in classification. The true theoretical AME is given by the dashed line in both panels.

depicted for comparison; we can see the average marginal effect on both $f(x)$ and $\sigma(f(x))$ is captured well by the GP models. Even though there is a true linear trend in f and the nonlinear deviations are designed to cancel out over the range of the simulated x values, the linear models by contrast poorly estimate the average marginal effect of x on both the link and probability scales.

3.4 Application: Economic Voting in Presidential Elections

I now turn to an application to real-world data to demonstrate strengths of the method. Campello and Zucco (2016) present an important finding in the literature on economic voting: Presidential reelection prospects (in the Latin American context) are affected by economic conditions that are unambiguously unrelated to presidents' policies. Relying on evidence from economics "that Latin American countries generally do well when international interest rates are low and commodity prices are high, and are hurt when the opposite happens", Campello and Zucco (2016) generate a "good economic times" (GET) index that captures both international interest rates and commodity prices (592). This should affect commodity exporting countries and countries with low savings (since they "depend on foreign inflows of capital to foster investment and economic growth") even

more, so they also code for which countries are low-savings and commodity exporting (LSCE) countries (592). Latin American presidents' actions do not affect international interest rates or commodity prices, so if voters' economic voting is tied to these factors, it illustrates a failure of economic voting to lead to electoral accountability. In presidential elections in South America from 1980 to 2012, they code the outcome of the election as "Reelected" if the incumbent ran and was reelected, or if the incumbent's chosen successor wins election if the incumbent doesn't run, and run logit models to assess the effect of the GET index. They show that an increase in the GET index increases the probability of reelection in LSCE countries. However, analyzing their data using GP classification yields a more nuanced story.[14]

Campello and Zucco (2016) report results for a few different model specifications; first they show a basic model that includes only an interaction between the GET index and the LSCE indicator (and the constituent terms). For this basic model, their central result holds: An increase in the GET index improves presidential reelection prospects in LSCE countries, though the difference between LSCE countries and non-LSCE countries is somewhat muted in the GP classification model relative to the logit model, and the average effect size for an increase in the GET index is somewhat smaller. Campello and Zucco (2016) report the "first difference" for the GET index for the LSCE group of countries and for the non-LSCE countries; this is the difference in the predicted probability of reelection when the GET index is set to its mean plus its standard deviation compared to when the GET index is set to its mean minus its standard deviation. In Figure 3.4, I depict the GET first differences for both the GP classification model and the logit model; they are very similar, but the effects are slightly smaller in the GP model, and the difference between LSCE and non-LSCE countries is not quite as pronounced.

The reason for this becomes clear if we look at how the GET index's marginal effect varies across its range, depicted in Figure 3.5, which illustrates one of the strengths of GP models: They do not

14. The replication data and code for Campello and Zucco (2016) are available on the *Journal of Politics* Dataverse (Campello and Zucco 2015).

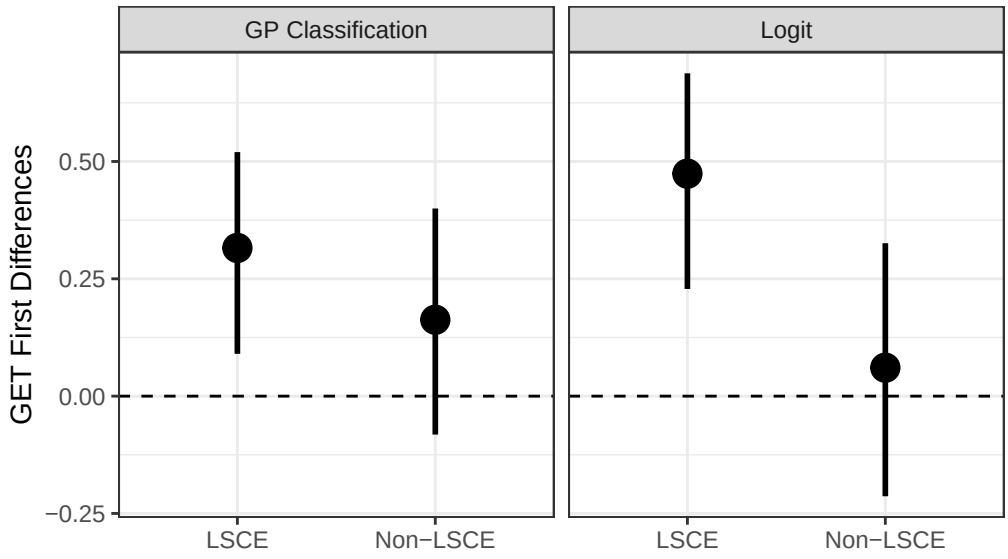


Figure 3.4: Comparing GET first differences between GP classification and logit; the first difference depicted is the sample average difference in predicted probability of reelection between setting the GET index equal to its mean plus its standard deviation and setting it equal to its mean minus its standard deviation, for both LSCE and non-LSCE countries, with 95% credible/confidence intervals

force a linear relationship between predictors and outcomes, but allow more flexible relationships where they should exist. While the logit model forces a constant marginal effect, the more flexible GP model allows the effect to vary across GET values. Keep in mind precisely what the plot is depicting: At every GET value, it shows the instantaneous rate of change for an increase to the GET index. So, what we see is interesting, but quite sensible once the model is able to illuminate the relationship. For very negative values of the GET index, an increase does not change much, but the effect grows as we reach more moderate values, and eventually wanes at extreme positive values, which makes perfect sense: When times are very good, it can be hard to distinguish between small increases.

We can further visualize this result by looking at the predicted probability of reelection as the GET index varies, depicted in Figure 3.6; as the index increases, initially there is a sharp increase in probability of reelection which then flattens off at more extreme values for the GP model. So, for their basic model, the result is well in keeping with the argument and conclusions of Campello and Zucco (2016), but provides a nice nuance to the story.

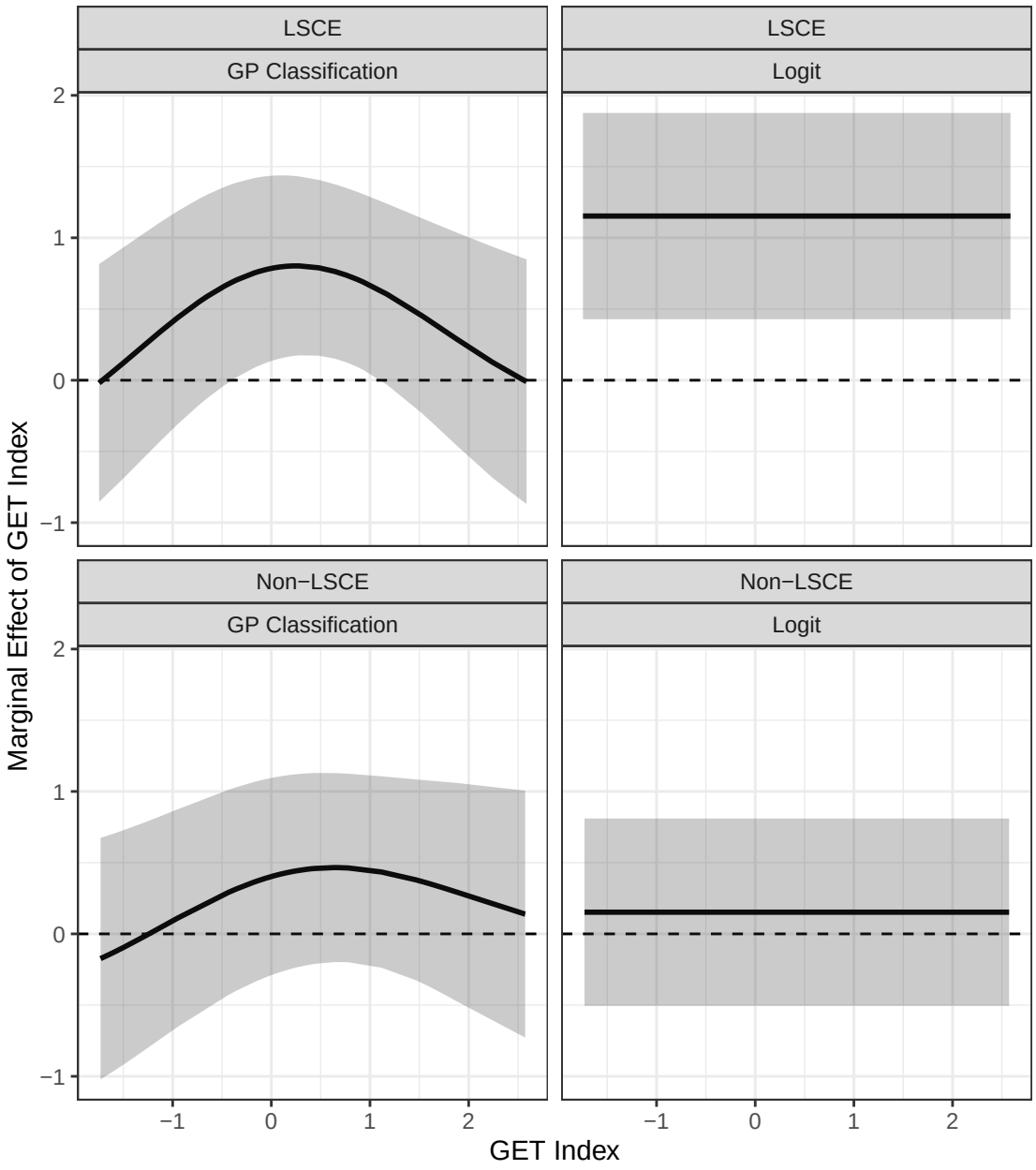


Figure 3.5: Marginal effect of an increase to the GET index across its observed range. Effects are on the link scale; 95% credible bands depicted with shading.

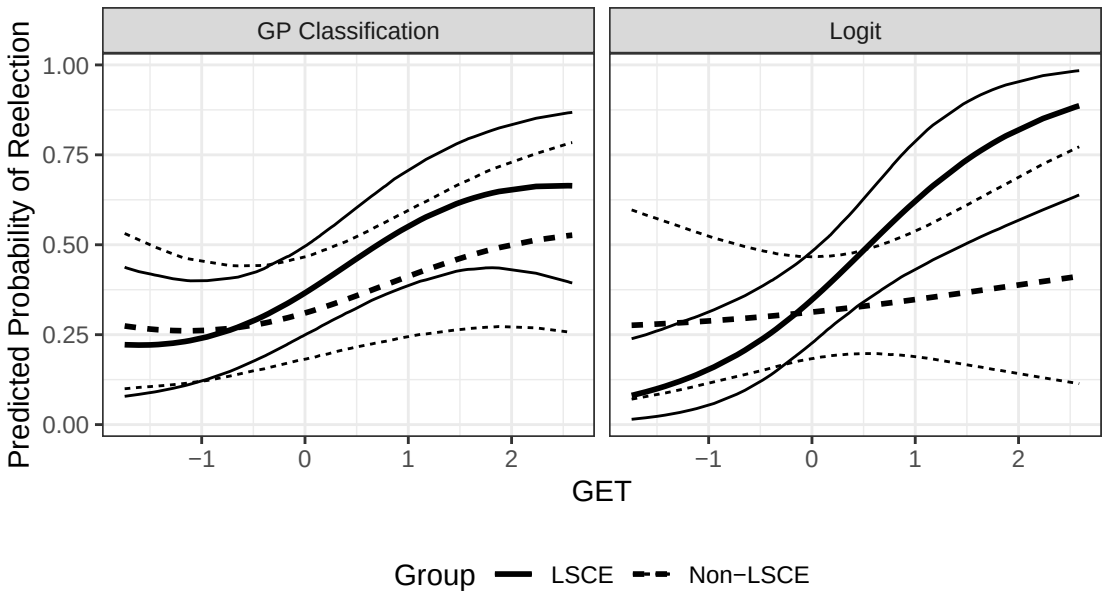


Figure 3.6: Comparing predicted probabilities between GP classification and logit, computed across the range of the GET index setting the LSCE indicator to 1 and to 0, with 95% credible/confidence bands

However, they also estimate models with controls. Specifically, they estimate a model that controls for whether the incumbent seeks reelection. In the South American context, this is important because "Incumbent presidents running for reelection in Latin America rarely lose an election" (Campello and Zucco 2016, 595), so we should see an overall higher probability of reelection for the incumbent group. When we add this control to the GP classification model, the sample average marginal effect of the GET index is no longer reliably positive (mean: 0.27; 95% credible interval: [-0.08, 0.61]), even when focusing on the LSCE group alone (mean: 0.26; 95% credible interval: [-0.08, 0.60]). The reason for this becomes clear when we look at the marginal effect of the GET index separately for the incumbent and non-incumbent groups, depicted in Figure 3.7: The GP model is able to uncover an interactive effect between incumbency and the GET index!

From the standpoint of democratic accountability, this makes sense; even if good economic times exerts an irrational influence on voters' evaluations of the parties/candidates, when incumbents themselves are up for reelection, the voters have much more specific information on what the incumbent will do and what effects that will have, which could mute the more irrational effect of

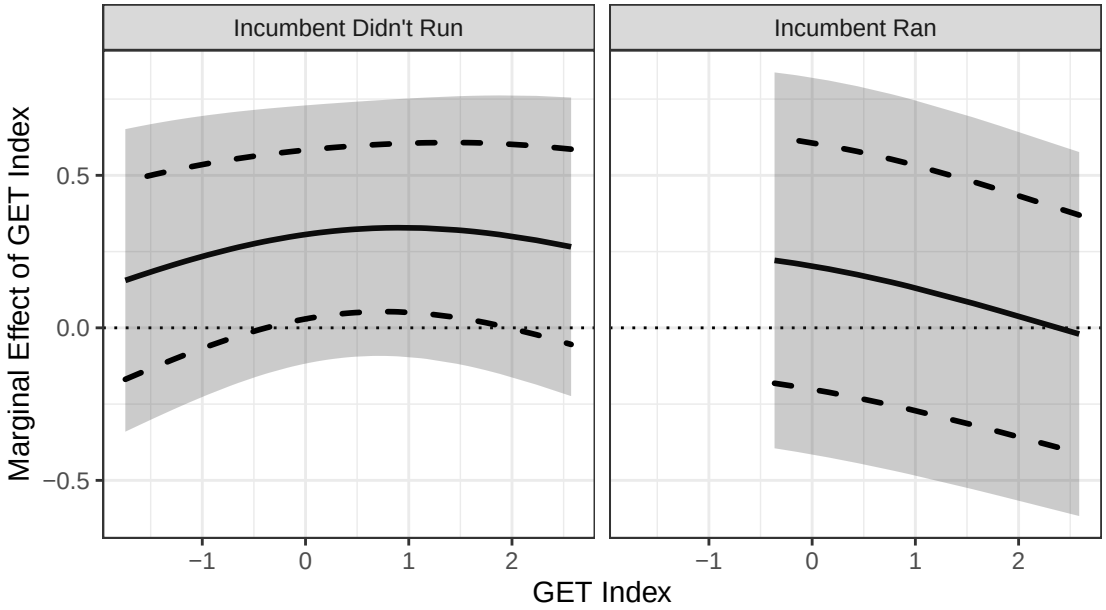


Figure 3.7: Marginal effect of an increase to the GET index across its observed range when incumbency is controlled for. Effects are on the link scale; 95% credible bands are depicted with shading and 80% credible bands are depicted with dashed lines.

generally good economic times. In the context of Latin American elections specifically, another factor that may support an interactive effect is just how heavily favored incumbents are—they are almost universally reelected; it may be that the effect of incumbency so drowns out all other factors in the voters' decision making process that they do not consider the things the GET index represents. Thus the GP classification model was once again able to guard against misspecification bias, this time through uncovering an interactive effect the original researchers did not account for. Again, this moderates the story originally told in Campello and Zucco (2016) and provides more nuance. Notice in the left panel of Figure 3.7 that in the non-incumbent group, the marginal effect of the GET index is estimated to be positive across its range, and though it is not statistically reliable at the 95% level, for moderate values of the index, the 80% credible interval does not contain zero, providing (admittedly weak) evidence that the mechanism Campello and Zucco (2016) propose may be at play when incumbents do not run. At the least it suggests more work should be done in this area to determine if we can uncover a statistically reliable effect with broader data; these data

contain only 106 observations, so a higher powered study may be needed to detect a reliable effect when accounting for this interaction.[15]

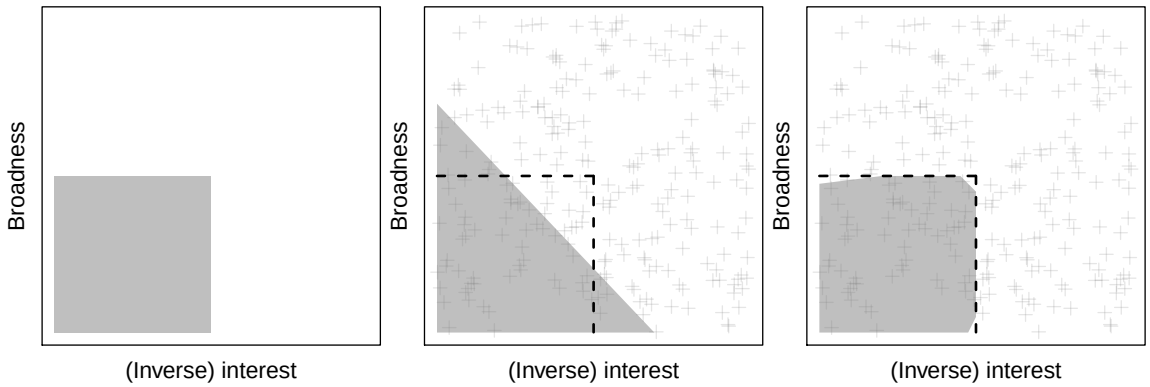
3.5 Benefits of GP Models for the Study of Judicial Politics

One area where using GP models can be particularly useful is judicial politics. An important concept in judicial politics is legal doctrine, or the policy rules pronounced by judges. Judicial scholars often incorporate some aspect of doctrine, either as an outcome (for example, studying legal change as in Wahlbeck 1997) or as (a) predictor(s) of interest or control(s) (for example, studying legal constraint as in Bartels 2009). One way to conceptualize legal rules or doctrine is a mapping from real-world circumstances to outcomes that judges would dictate in a case presented to them with those circumstances. However, as noted by Kastellec (2010), using models with strong functional form assumptions, particularly generalized linear models, inappropriately restricts our estimation of that mapping.

To make this clear, let's consider a concrete example using a common type of legal rule, called a "conjunctive rule", where all elements of the rule must be met for a positive outcome, such as a law facing strict scrutiny analysis. For a law to pass constitutional muster under a strict scrutiny test, it must both serve a compelling governmental interest and be narrowly tailored to achieve that interest. Such a rule is depicted in Figure 3.8. The x-axis captures whether the governmental interest is "compelling"; the axis is reverse coded, such that low values indicate a compelling interest and high values indicate an interest that is less than compelling. The y-axis is for the broadness of the regulation at issue; low values indicate a narrowly tailored regulation and high values a regulation that is not narrowly tailored.

Taking each dimensions as being a continuum from 0 to 1, with 0.5 being the threshold on

15. At a conventional power level of 0.8, we would need at least 144 observations (a sample size increase of 35.8%) to detect an effect of the size seen in the basic model at the 0.05 significance level once accounting for the interactive effect of incumbency and the GET index.



(a) True strict scrutiny test (b) Logit estimate (c) GP classification estimate

Figure 3.8: Comparing estimation of legal rules between linear models and GP classification. In panel (a), the true strict scrutiny test is depicted with the set of cases receiving a "constitutional" outcome indicated with gray shading. In panels (b) and (c), estimates of the rule from a logit model and GP classifier respectively are depicted with gray shading. In those panels the true rule is indicated with a dashed line, and the simulated cases the models were trained on are indicated with light gray crosses.

each dimension for meeting that element of the strict scrutiny test, I simulated 250 cases, drawing the case factors independently from a standard uniform distribution for each case. I assigned them outcomes according to the rule that a "constitutional" outcome should be given if and only if both broadness and inverse interest were less than 0.5. I then trained both a logit model (the standard approach in prior literature) and a GP classifier on these cases. I then determined the set of cases each model would assign a "constitutional" outcome to, and compared these estimated rules to the true strict scrutiny test; the results are depicted in Figure 3.8.

While the linear model struggled to find the relationship legal doctrine dictates between states of the world and outcomes, the GP classifier was able to very closely approximate the true legal rule. When studying legal rules, non-parametric models such as GP classification should be strongly encouraged.

3.6 Conclusion

Political scientists commonly face uncertainty over functional form in their models or issues of error correlation. Sometimes assuming a linear functional form masks systematic deviations from linearity in the true relationship between our predictor of interest and our outcomes. We may also fail to specify all necessary interactive effects in our model specifications. GP models solve both of these misspecification bias pitfalls, and are more flexible and extensible than other approaches to alleviating misspecification bias such as Hainmueller and Hazlett (2014) (which does not cover, for example, binary outcomes). Carlson (2021) shows GP regression is even more adept at handling error correlation in time series cross-sectional data than bespoke solutions such as panel-corrected standard errors or lagged dependent variable or fixed effect approaches.

I offer a thorough introduction to GP models for political scientists, and expand the state of the art for these models to provide the distribution of inferential quantities of interest to political scientists that the machine learning literature does not cover. I provide both simulation and replication evidence of the usefulness of GP models as well as practical guidance on performing inference with these models that are increasingly gaining popularity with political scientists.

Chapter 4

The Force of the Law: The Constraining Effect of Precedent at the US Supreme Court

A fundamental question at the heart of judicial politics is whether, how, and to what extent "law" impacts judges' decisions.[1] Judges often profess to simply apply existing law to the cases before them, untainted by their own preferences, exemplified by Chief Justice Roberts' (in)famous statement, "Judges are like umpires. Umpires don't make the rules, they apply them... I will decide every case based on the record, according to the rule of law... and I will remember that it's my job to call balls and strikes, and not to pitch or bat." Nevertheless, judges are typically portrayed as conservative or as liberal just as other political actors are, and with good reason: A long line of political science research finds ideology to be a strong predictor of US Supreme Court justices' decisions (e.g. Epstein and Knight 1998; Segal and Spaeth 2002) and media coverage may highlight political or negative aspects of the Court's decisions (Denison, Wedeking, and Zilis 2020; Johnson and Socker 2012).

1. A previous version of this paper was presented at the 2021 annual conference of the American Political Science Association. I would like to thank Jim Spriggs, Jacob Montgomery, and Jongyoon Baik for their helpful comments.

Although the public need not view judges as "legal automatons" to view the courts as legitimate (Gibson and Caldeira 2011), to the extent judges are viewed as mere "politicians in robes", the courts' legitimacy may suffer. More fundamentally, understanding the way legal constraint interacts with political ideology in judicial decision making is crucial for generating useful theories of judicial behavior and accurate inferences in quantitative studies of judicial politics. Given the stakes for understanding the extent to which the law constrains judicial decision making, it is no surprise the literature offers a variety of clever ways to test for the law's influence, such as examining the behavior of dissenting justices in future related cases (Segal and Spaeth 1996), studying the use of precedent (Hansford and Spriggs 2006; Hinkle 2015), considering legal forces particular to agenda-setting decisions (Black and Owens 2009), interacting legal and ideological variables to find conditional effects (Bartels 2009, 2011), and isolating ideological effects by considering positions taken by political actors outside the Court (Bailey and Maltzman 2008).

While the political science literature is unambiguous that ideology plays a large role at the Court, the evidence on what constraining effect, if any, the law exerts on the justices' decisions is mixed. Segal and Spaeth (2002) posited Supreme Court justices as more or less unconstrained by law, deciding cases as they like, but other scholars advocate a more nuanced view, with effects of ideology as well as constraint of ideological behavior by legal factors (e.g. Bartels 2009). Even some work that concludes that "law matters" finds that legal *constraint* is not among the chief legal effects; for example, Hansford and Spriggs (2006) conclude that law (that is, precedent) can act as both an opportunity (to interpret existing precedent positively or negatively in line with their own ideological preferences) in addition to a constraint—but the weight of their presented evidence is on the side of law as opportunity rather than constraint.

Perhaps because constraining effects of law are difficult to uncover in Supreme Court justices' ultimate decisions on the merits, work on this topic often focuses on slightly different effects or settings. For example, Bailey and Maltzman (2008) estimate justices' willingness to explicitly overrule precedent, controlling for justices' preferences by using positions taken on Supreme Court

cases by members of Congress and the president.[2] Black and Owens (2009) show factors such as circuit conflicts predict agenda-setting decisions at the Supreme Court, controlling for whether the justices would ideologically prefer to grant or deny cert given the status quo from the lower court decision. Hinkle (2015) finds Circuit Court judges are more likely to cite precedent that is binding in their jurisdiction, utilizing random panel assignment at the U.S. Courts of Appeals to identify the effect.

Moreover, even when empirical studies uncover constraining effects of law, methodological critiques cloud the findings. For example, Richards and Kritzer (2002) find evidence of "jurisprudential regimes": When a (typically landmark) decision dictates an important change to how certain facts are treated under the law, their work suggests justices seem to vote differently afterward, suggesting these precedents exert some binding force on the justices' choices. However, Lax and Rader (2010a) show that work in this area failed to account for crucial forms of error correlation, such as by term, resulting in overconfident statistical tests. Some work, such as Bartels (2009), has used mutli-level modeling to account for this issue, though we may still worry about "parametric assumptions of error distributions" since we "must assume what sort of clustering can exist [and] assume that other forms... do not exist" (Lax and Rader 2010b, 289). A particularly difficult problem for inference in this area is a type of time-varying confounding: as a justice serves on the Court, their preferences influence the law, so finding an impact of "law" on the justice's decisions may just be picking up on agreement with their own past selves. Some work, such as Segal and Spaeth (1996), focuses on the actions of dissenters in future related cases for this reason, finding little constraining effect of the law.

I provide a novel examination of this issue by using Gaussian process (GP) classification, a machine learning technique, to measure the law and its effect on justices' decisions. I conceptualize the law as the implication in a particular case from all the cases that came before it; given the outcome

2. Bailey and Maltzman (2008) also look at two other legal effects: Whether the justices exhibit "judicial restraint"—i.e., defers to Congress—and whether they adhere to a strict interpretation of the First Amendment.

the Court has assigned to similar cases in the past, what should we think about this case if we did not let anything else influence our thinking? I operationalize this as the predicted outcome in the present case given its facts from a GP classification model trained only on the cases that came before it. This approach aligns more closely than past empirical approaches with the process typically identified as legal reasoning: determining how the present case, as a whole, relates to the Court's past decisions (Levi 1949). It provides a single variable whose effect in justice-level models will give the impact of the law on their decisions. This contrasts with approaches in traditional case fact studies, such as Richards and Kritzer (2002) and Bartels (2009), who look at the effect of a particular (set of) legally relevant case fact(s), where I instead look at the effect of the implied legal outcome given all the legally relevant case facts on justices' individual decisions. Importantly, I show how to control for justices' own contribution to the current state of the law to properly identify the impact of the current state of precedent on justices' individual decisions, independent of their own preferences, addressing the time-varying confounding issue. I find ample effects of ideological and policy preferences, but also that law provides a significant and substantial effect on Supreme Court decision making, for some justices more than others.

This study provides several main contributions. First, I provide a fresh theoretical perspective on legal constraint by conceptualizing the law as the implied outcome in each case given the cases that came before it. Second, I overcome methodological issues with past studies to provide the best available evidence that the law exerts a constraining force on the actions of (some) justices. Third, I provide a new measure of the law, or of the legal status quo. This new measure can help reinvigorate studies of legal constraint and modeling law at the Court, as the volume of work on this important issue has diminished in part because some think the justices are largely unconstrained, but also due to difficult methodological issues (see, e.g., Klein 2017). A measure of the legal status quo is also an important quantity in a number of contexts judicial politics researchers face (see, e.g., Black and Owens 2009) just as a status quo is important in many areas of political science more broadly (e.g. Krehbiel 1998). I also provide suggestive evidence differentiating between reasons

why the law matters, indicating some justices have a preference for following the law rather than seeing it as a constraint due to (for example) legitimacy needs, which few studies have attempted. Finally, I highlight a widespread mismatch between theory and methods in judicial politics and show how to apply a method better suited to studying decision making in the setting of adjudication. By restricting attention to linear models, past work was susceptible to misspecification bias (see Kastellec 2010), while the modeling strategy taken here avoids that issue while more closely matching the concept of legal reasoning. In so doing, I offer the most compelling evidence to date of how law constraints Supreme Court Justices' decisions.

4.1 Measuring the Law

Part of the difficulty in studying this issue is one of measurement; how do you measure the law in a way to facilitate analyzing its constraint on justices' decisions? A number of clever approaches to measuring the law have been utilized. For example, the law can be conceptualized as a collection of precedents; Hansford and Spriggs (2006) use this idea, and analyze citations to those precedents to assess legal change. Bailey and Maltzman (2008) measure justices' willingness to overrule past precedent as a measure of legal constraint. Bartels (2009) interacts Martin-Quinn scores (Martin and Quinn 2002)[3] with legally relevant case facts to determine if facts constrain the effect of ideology, and in particular to determine whether some types of cases provide more constraint on ideology than others.

However, lack of a measure of what outcome the law implies in each case has led some to feel "we have reached a point of rapidly diminishing returns in our study of [the] issue" of legal constraint, though with perhaps some hope that advances in text analysis methods may provide better measurement tools (Klein 2017). I provide a different approach: Joining Justice Oliver Wendell Holmes in claiming, "the prophecies of what the courts will do . . . are what I mean by the

3. Martin-Quinn scores provide ideological estimates for Supreme Court justices somewhat analogous to (e.g.) NOMINATE scores (Poole and Rosenthal 1985) for members of Congress.

law" (Holmes 1897), I propose estimating a *predictive* mapping between case characteristics and legal outcomes as a measure of "the law." That is, for each case, we place ourselves in the shoes of a decision maker at the time the case was decided, and say "the law" in that case is what our best predicted outcome in the case would be if we examined *only* the past decisions of the Court at the time.

This conception matches a number of theoretical perspectives on the law, judicially-created policy, and judicial decision making. Consider the attitudinal model, which posits that judges make decisions based on the characteristics of the cases presented to them and their own sincere attitudes and values. Holmes' legal realist approach comports well with this approach. "Case space" models (Lax 2011) can be considered a formalization of the attitudinal model; judges' policy preferences in these models are a partition of a case space, where each dimension corresponds to a case fact, into outcomes. Lax (2007) shows how to represent judicial policy-making using such models. My predictive approach can be viewed as a Bayesian update after each case as to the mapping from the case space to outcomes, or an online update about the justices' attitudes.

This approach can also comport with a legal model of judicial decision making; Levi (1949) explains that in legal analysis, factually similar cases should receive similar outcomes. So this is exactly how law students learn and how lawyers typically argue cases, and how judges would make decisions under the legal model: We take all past cases of the Court, compare the present case to those cases, and assign it the outcome implied by the cases most similar to the present case. Callander and Clark (2017) show how to extend case space models so that the case space maps onto a continuous latent legal outcome rather than onto discrete dispositions; under an assumption that the high court has perfect knowledge of the "correct" mapping from the case space to legal outcomes, lower courts update their knowledge of that mapping after every observed decision of the high court by comparing similarity of the cases they are presented with the past observed decisions of the high court.

Thus, I use the following approach to measure the law: Let $X \subset \mathbb{R}^n$ be the n case factors

relevant in the area of law at issue and $\mathbf{X}$ be a finite number of observed cases from the space X; for each case i, we estimate the mapping between the characteristics $\overline{\mathbf{X}}_{i-1}$ of all cases observed *prior* to i and outcomes $\overline{\mathbf{y}}_{i-1}$ in those cases, so that "the law" is the outcome predicted from that model given the characteristics of the instant case, $\mathbf{x}_i$.

There is a long history in the study of judicial politics of estimating the relationship between case characteristics and Court decisions. For example, in the context of Fourth Amendment challenges to police searches and seizures, Segal (1984) shows the relationship between case facts such as the existence of probable cause and the Court's decisions. However, scholars have stopped short of using such models to develop measures of the law. In addition to this innovation, I also use GP classification for estimating this mapping rather than a generalized linear model, as in, e.g., Segal (1984). I use this model because it flows naturally from my conception of law here: The GP classifier is a nonparametric model that compares the test case to all training cases on the basis of all combinations of predictors. That is, a judge hearing a case is not so concerned with any particular fact in isolation. Instead, they care about how the facts as a collective bundle affect the legal outcome; the effect of one fact almost always depends on the presence or absence of other facts. Therefore a linear model unreasonably constrains the class of legal rules we can recover from observed cases (see Kastellec 2010), resulting in misspecification bias. GP classification allows us to uncover any smooth function mapping case characteristics to legal outcomes, as explained in Section 4.3.2, and provides a convenient way to separate out the effects of this predictor and judges' own preferences as explained in Section 4.3.3.

4.2 Modeling Judges' Decisions

I start from the assumptions of an attitudinal model or the case space model (Kornhauser 1992)[4]. For any issue area, or type of case (e.g. cases about police searches and seizures), a judge has

4. See Lax (2011) for an overview of case space models in political science.

a number of factors relevant to their preferred outcome. For example, in a police search case, a judge's preferred outcome (i.e., whether they prefer to rule that the search was constitutional or unconstitutional) may be influenced by whether the search was of a home or not, or if the search was incident to arrest; these case factors may interact in potentially complex ways to determine the judge's sincerely preferred outcome. However, we also must take into account that judges may not follow their own preferred dispositions precisely; they may feel constrained by "the law". For example, suppose a judge would always prefer to rule that searches of a home require a warrant to be lawful. However, they have also observed the Supreme Court's ruling in *Maryland v. Buie* (1990)[5] that officers may sweep the area during an in-home arrest to uncover hidden persons who could pose a danger to those at the scene. This judge may—or may not—then feel constrained to rule differently from their preferred disposition in a closely related case.

To formalize this idea, let ϕ_j represent judge j's sincere policy preferences; in other words, ϕ_j is a function that maps X to outcomes. Now, Supreme Court justices are likely less concerned with the outcome in any particular case they hear than broader issues of how the law is interpreted, so it may seem awkward to define judicial preferences in this way. However, think about the implications of a case outcome in light of the discussion in the previous section; every time the Supreme Court puts out a new decision, it clarifies their interpretation of the mapping between X and y. That is, the justices' utility from deciding a case one way or another may come in part from preferring a particular party wins the case, but it also provides a vehicle for updating observers on judicial policy. Let λ represent the Court's mapping from X to y, or the association between cases in X and the outcomes that would result from majority voting over outcomes among the justices on the Court. λ_i will indicate the outcome implied in case i from the prior cases $\overline{D}_{i-1} = \left(\overline{\mathbf{X}}_{i-1}, \overline{\mathbf{y}}_{i-1}\right)$. Then each justice j's overall preferences f_j are a function both of λ and of ϕ_j.

For example, let's revisit the *Buie* case mentioned above. An armed robber was described as wearing a red track suit. The police obtained a warrant for Buie, and executed it at his home.

5. 494 U.S. 325.

An officer coaxed Buie out of the basement and arrested him. Another officer then "entered the basement 'in case there was someone else' down there", at which point he found a red track suit in the basement. Buie argued the red track suit should be excluded as evidence. There are a number of prior decisions of the Court in $\overline{D}_{i-1}$ that influence our estimate of λ_i, or the legal outcome. In *Terry v. Ohio* (1968)[6], the Court ruled officers can "stop and frisk" an individual without violating the Fourth Amendment if they have reasonable suspicion the person is armed and thus presents a threat to the officers. In *Michigan v. Tyler* (1978)[7], the Court held government actors may seize evidence of a crime if it is in "plain view" from a place they were allowed to be in. Thus our estimate of λ_i may well be that the law implies a ruling that the track suit was legally seized, since the officer suspected "someone else" (who could pose a danger to the officers at the scene) was "down there" in the basement where he found the track suit in "plain view". The sincere preferences ϕ_j of judge j may be to rule it was a constitutional seizure or an unconstitutional one.

It will be useful to consider λ to be a mapping from cases to continuous outcomes rather than dichotomous ones. Judges must assign a dichotomous outcome to each case (government action is constitutional or unconstitutional; or the plaintiff wins or defendant wins, etc.), but some cases are a "closer call" than others. Then with some function σ mapping $\mathbb{R}$ to $[0, 1]$, which I will take to be the logistic function,[8] $\sigma(\lambda_i)$ gives the probability case i should receive a positive outcome. Cases with large positive λ_i are cases where we are quite certain the case should receive a positive outcome, while cases with large negative λ_i are cases where we have high confidence the case should receive a negative outcome; cases with λ_i close to zero are "close calls".

Approaches to model or account for judicial preferences in models of case adjudication differ.

6. 392 U.S. 1.

7. 436 U.S. 499. The plain view doctrine relied on in *Tyler* was first articulated by the Court in *Coolidge v. New Hampshire*, 403 U.S. 443 (1971), though in that case the Court described it only to state the facts of *Coolidge* did not meet the requirements of the plain view doctrine. In *Horton v. California*, 496 U.S. 128 (1990), decided the same term as *Buie*, the Court held that does not even matter if the discovery of the evidence "in plain view" was inadvertent or not.

8. I will assume the use of the logistic function throughout the paper, but a number of other functions such as the normal cumulative distribution function could be used.

Case space models typically concern themselves only with ϕ_j, and how judges act strategically to enact their preferences within particular institutional arrangements or strategic situations. Bartels (2009) instead focused on controlling for legal effects via λ and capturing judicial preferences via Martin-Quinn scores. Here I focus on accounting for the fuller picture of judicial preferences ϕ_j and utilizing λ as the predictor of interest to capture the effect of the law on judicial preferences. That is, the judges' sincere preferences in these cases are not wholly captured by a unidimensional measure such as Martin-Quinn scores. The Martin-Quinn measure is useful as an overall measure of ideology, but issue-specific preferences are best described with reference to the case space defined by the relevant facts in such cases. However, the law, which we must measure or operationalize in some way to assess legal constraint, is also best described as a mapping between this space and outcomes.

Separating out these effects—the pull toward a particular vote from the judge's individually truly preferred outcome given the characteristics of the present case and the pull from the outcome implied by the Court's past decisions given the characteristics of the present case—is non-trivial. We cannot observe the justices' sincere policy preferences ϕ_j, and both ϕ_j and λ are themselves functions of the same case factors X.

However, there are some things we can say even with such a general theoretical setup. First, consider our quantity of interest is the average marginal effect of the law, or how much, on average, λ influences a justice's decision; considering all the cases this justice has heard, how much on average would the value of f_j change with an increase in λ (i.e. with an increase in the probability that *the law* says the outcome should go a particular way)? This average marginal effect is by definition

$$\gamma_j = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \lambda_i} f_j \left(\mathbf{x}_i \right). \tag{4.1}$$

Next, if the law exerts no constraining effect on judges' decisions, then $f_j = \phi_j$. This implies

that typically we should see an average effect of 0 if judges' behavior is unconstrained, or an average positive effect if the law is constraining their behavior.

Could λ have an average *negative* effect on f_j? Let's consider a basic sequential game where the justices are voting on cases one by one (as we see in reality). Absent legal constraint, under a wide variety of assumptions about preferences, it is a weakly dominant strategy for justices to always vote precisely in line with their preferences.[9] That is, no matter how the other justices are voting, if justice j votes how they would sincerely prefer, they can only make the outcome they prefer more likely rather than less, and further, their preferred outcome occurring in that case can only make the law get closer to their preferred policy rather than farther away.

In other words, in no circumstances will such a vote cause λ to deviate further from ϕ_j than it would given a contrary vote, and in some circumstances—when j would be a pivotal vote in a majority coalition—it will cause λ to come more into line with ϕ_j. So when j might be pivotal, if j is not constrained by law, they are strictly better off by voting their preferences. However, when there is no chance for j to be pivotal, they are indifferent between votes except perhaps for expressive reasons. If we assume some noise or imperfection in a judge's actions[10], this can result in a negative relationship between λ and f_j since then her expressive motivations break her indifference more often in more extreme cases (in terms of the difference between λ and ϕ_j).

It is also important to note that in the applied (empirical) setting, where we cannot directly observe ϕ_j, and may not even be certain about all its inputs, omitted variable bias could bias our *estimate* of γ—if some factor Z in truth is an input to both λ and ϕ_j, but is not included in the statistical model, we may see a spurious (positive or negative) effect. However, if this relationship is only seen in non-pivotal votes, it provides at least suggestive evidence that an expressive mechanism is at play rather than omitted variable bias.

We may also consider different reasons *why* the law would enter into judges' decisions. From

9. See Proposition 5 in Lax (2007).

10. Or perhaps we may assume some imperfect knowledge of their own preferences; sometimes judges might be confronted with a "hard case".

one perspective, judges may allow the law to influence their decisions contrary to their personal policy preferences because they also intrinsically care about the law itself. That is, they have both individual policy preferences, but also some preference for deciding cases consistent with the Court's past decisions (and these things may sometimes conflict). However, we may also think that judges only allow the law to constrain their decisions because they do not want to be *seen* as going against the law, perhaps for institutional legitimacy reasons. Bartels (2009) and Bartels (2011) suggest certain legal factors constrain ideological voting by judges more than others. In particular, Bartels (2009) finds that ideology influences Supreme Court justices' votes to a greater degree in cases involving "intermediate scrutiny" than in cases involving either a "rational basis" or "strict scrutiny" test. In the latter two cases, legal doctrine strongly suggests the correct legal outcome should likely be a ruling that government action is constitutional and unconstitutional respectively, while there is much more variation in the implied legal outcome for intermediate scrutiny cases. In the context of the model of judicial decision making considered here, when the absolute value of λ is higher, the implied legal outcome is more certain; large positive values of λ indicate a high probability the present case should, legally speaking, receive a positive outcome while large negative values indicate a high probability the present case should receive a negative outcome. When the absolute value of λ is close to zero, there is a roughly 0.5 probability that either outcome is legally correct. So if judges have some intrinsic concern for following the law, the effect of λ should be more or less constant over the range of λ values, whereas if the mechanism argued for in Bartels (2009) drives law's constraining effect, then the effect of λ should be high when the absolute value of λ is high, but lower when λ is closer to zero.

In sum,

1. If $\gamma_j > 0$, it implies the law exerts some positive constraining effect on j's decision making.

2. If $\gamma_j = 0$, then j's decisions are not constrained by the law and they act only according to their own preferences ϕ_j.

3. If $\gamma_j < 0$, we should expect the negative relationship between λ and f_j to exist only in cases where j is unlikely to be a pivotal justice.

4. If γ_j is increasing in $|\lambda|$, it implies j follows the law only when the legal outcome is more certain (offering less ideological discretion), whereas if γ_j is constant across $|\lambda|$, j may have some intrinsic concern for following the law.

4.3 Data and Methods

To assess legal constraint at the Supreme Court, I will utilize both court- and justice-level data on case dispositions. A court-level model provides predicted values to serve the role of λ in Section 4.2; at both levels I use Gaussian process (GP) classification, a method described further in Section 4.3.2.

4.3.1 Data

I use data on First Amendment Free Expression cases at the U.S. Supreme Court. This setting has been frequently studied, both in the context of studying legal change (Richards and Kritzer 2002; Bartels and O'Geen 2015) as well as studying legal constraint (Bartels 2009) as here. Richards and Kritzer (2002) original coded the facts of all Free Expression cases heard by the Supreme Court in the 1953 to 1998 terms, and Bartels and O'Geen (2015) backdated and updated this data to include cases from the 1946 to 2004 terms.[11] I updated this data to include cases from the 1946 to 2019 terms (adding the cases from the 15 most recent terms). These free expression cases are coded for: whether the court-level outcome was liberal (i.e., pro-expression) or conservative (i.e., anti-expression), measures of the facts of each case (discussed further below), and the median Martin-Quinn score (Martin and Quinn 2002) for the Court at the time of that decision.[12] I use

11. The data for Bartels and O'Geen (2015) are published as Bartels and O'Geen (2014).

12. As Martin and Quinn (2005) explain, despite the tautological issue of "votes explaining votes", use of Martin-Quinn scores as explanatory variables is appropriate when the votes in question are only from one issue area (see also

the justice-level Supreme Court Database (Spaeth et al. 2020) and the justice-level Martin-Quinn score data (Martin and Quinn 2020) to expand the data to the justice level. The relevant case facts in Free Expression cases are: the *Category* of restriction, or whether the restriction on speech is content-based, content-neutral, or of a less protected category of speech; the *Actor*, or who is restricting speech, such as the federal government, a state government, or a private actor; what the *Restriction* is, such as a criminal sanction or a loss of employment; and the *Identity* of the speaker, such as whether they are a politician, an alleged communist, or a racial minority (Bartels and O'Geen 2015). After merging records, there are 677 court-level outcomes to analyze and 5,922 justice-level votes to analyze.[13]

4.3.2 Gaussian process classification

Gaussian process (GP) models are a class of flexible machine learning models (Rasmussen and Williams 2006) that political scientists have recently begun to utilize (Carlson 2021; Duck-Mayr, Garnett, and Montgomery 2020; Gill 2021).[14] Chapter 3 provides a fuller introduction to the method, but I briefly refresh that introduction here.

GP models are used to learn the mapping from predictors to outcomes when its functional form is unknown. Think back to my discussion in Section 4.2 of what the law might imply in the *Buie* case. I did not have to specify some linear relationship between individual relevant facts and outcomes, or even tell you anything about the shape of the relationship between X and y; I told you about the cases that were close to *Buie* in X and what their outcomes were, which gave you some idea of what the legal outcome would be in *Buie*. That is essentially what GP models do: They learn about what the outcome should be in our test cases by assuming it will be similar to the

Bartels 2009).

13. There are 6,076 justice-decision combinations, but sometimes justices do not participate in a particular case (such as for ethical reasons, or because arguments were given before they joined the Court), leaving 5,922 observations remaining after accounting for missingness in the outcome.

14. They share some similarities to the kernel-regulated least squares model introduced by Hainmueller and Hazlett (2014), but are more versatile, accommodating categorical outcomes, as well as a variety of kernels, and providing a Bayesian approach (see Cheng et al. 2019).

outcome in the cases that are closest in the covariate space. That GP models can accommodate a wide variety of "shapes" is of particular interest in the context of judicial politics; Kastellec (2010) rightly points out that linear models often used to relate case factors to case outcomes are too inflexible to accommodate many common types of legal rules.

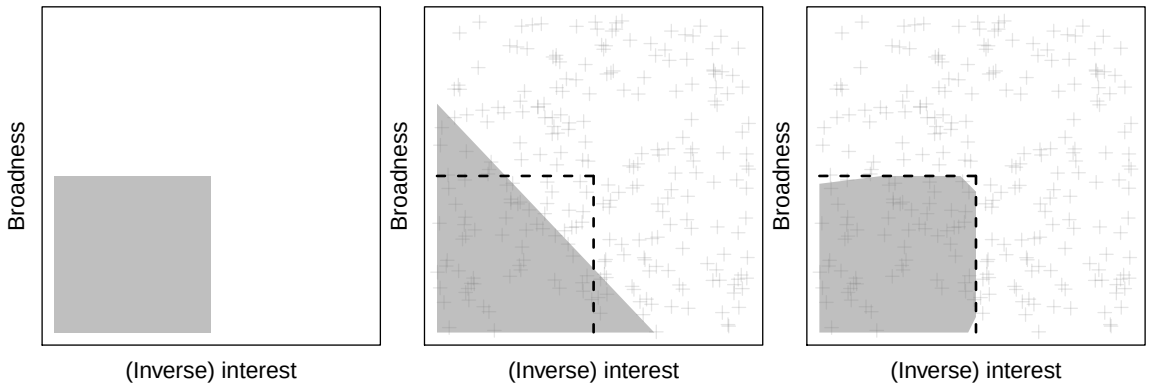
To see this, recall the example from Chapter 3, in Section 3.5. This example uses a common type of legal rule, called a "conjunctive rule", to illustrate the issue. For a conjunctive rule, all elements of the rule must be met for a positive outcome. One example is the "strict scrutiny" test; for a law to pass constitutional muster under a strict scrutiny test, it must both serve a compelling governmental interest and be narrowly tailored to achieve that interest. A strict scrutiny rule is depicted in Figure 4.1.[15]. The x-axis captures whether the governmental interest is "compelling"; the axis is reverse coded, such that low values indicate a compelling interest and high values indicate an interest that is less than compelling. The y-axis is for the broadness of the regulation at issue; low values indicate a narrowly tailored regulation and high values a regulation that is not narrowly tailored.

Taking each dimensions as being a continuum from 0 to 1, with 0.5 being the threshold on each dimension for meeting that element of the strict scrutiny test, I simulated 250 cases, drawing the case factors independently from a standard uniform distribution for each case. I assigned them outcomes according to the rule that a "constitutional" outcome should be given if and only if both broadness and inverse interest were less than 0.5. I then trained both a logit model (the standard approach in prior literature) and a GP classifier on these cases. I then determined the set of cases each model would assign a "constitutional" outcome to, and compared these estimated rules to the true strict scrutiny test; the results are depicted in Figure 4.1.

The GP classifier is able to develop a much more accurate estimate of rules such as the conjunctive rule in Figure 4.1.[16] This is accomplished by making less restrictive assumptions.

15. This figure is a replication of Figure 3.8

16. To be more specific, using a linear model is equivalent to the very restrictive assumption that judges' preferred rules must be a hyperplane in the case space.



(a) True strict scrutiny test (b) Logit estimate (c) GP classification estimate

Figure 4.1: Comparing estimation of legal rules between linear models and GP classification. In panel (a), the true strict scrutiny test is depicted with the set of cases receiving a "constitutional" outcome indicated with gray shading. In panels (b) and (c), estimates of the rule from a logit model and GP classifier respectively are depicted with gray shading. In those panels the true rule is indicated with a dashed line, and the simulated cases the models were trained on are indicated with light gray crosses.

GP classifiers typically only make some (usually fairly minimal) assumptions about the covariance between outcomes conditional on the predictors. Then Bayes rule combined with some linear algebra allows us to derive a posterior distribution. More specifically, the model uses a logistic likelihood as in a typical logit model,

$$\Pr\left(\mathbf{y} \mid \mathbf{X}\right) = \prod_{i} \sigma\left(y_i f\left(\mathbf{x}_i\right)\right), \tag{4.2}$$

(where σ is the logistic function). However, rather than assuming a linear (or any particular) form for f, we simply put a prior distribution on the latent outcomes,

$$f\left(\mathbf{X}\right) \sim \mathcal{N}\left(\mathbf{0}, K\left(\mathbf{X}, \mathbf{X}\right)\right), \tag{4.3}$$

where K is a matrix-valued function whose i, j entry gives the prior covariance between latent outcomes for observations i and j.[17] An overwhelmingly popular choice for K, which I use in this

17. Technically this prior distribution need not have a zero mean either; see Chapter 3 for more details.

article, is the squared exponential covariance function,

$$K\left(\mathbf{x}_i, \mathbf{x}_j, \sigma_f, \boldsymbol{\ell}\right) = \sigma_f^2 \exp\left(-0.5 \sum_d \frac{\left(x_{id} - x_{jd}\right)^2}{\ell_d^2}\right), \tag{4.4}$$

where σ_f is a scaling factor for the covariance matrix and $\boldsymbol{\ell}$ is a vector of length scales. Essentially, the assumption of the model here is simply that $\mathbf{X}$ observations that are closer together in the covariate space will be more likely to have latent outcomes that are close together; $\boldsymbol{\ell}$ basically tells us what "close" means on each dimension. From this simple assumption, the posterior over f—in other words, an update of our belief about what sort of relationship there is between X and y after observing some data—can actually be quite easily approximated with a Taylor expansion (for details, see Chapter 3).

This posterior, to return back to the notation of our specific context, is referred to above as λ from the court-level model. In Chapter 3 I also derive average marginal effects for GP models, which I use to calculate γ, the ultimate quantity of interest.

4.3.3 Identification

I am interested in determining the extent to which the decisions of the justices on the United States Supreme Court are influence by the law, or the latent legal outcome implied by the Court's past decisions. As discussed above, this conception of the law clearly implies a measure of the law: the prediction in case i given its characteristics $\mathbf{x}_i$ from a model trained only on the Court's cases $1, \ldots, i-1$, denoted by $\boldsymbol{\lambda}$, or λ_i in case i. However, I am more specifically interested in the effect of the law on justices' decisions, *independent of* the justices' own preferences ϕ over outcomes. Case characteristics can influence justice j's decisions not only through the law, but also through their own preferred mapping from case characteristics to outcomes. However, we cannot directly observe ϕ_j, so we must include in our model of justice j's decisions not only $\boldsymbol{\lambda}$, but also $\mathbf{X}$ itself. But, moreover, we must account for the fact that justice j's decisions also have the potential to

dynamically impact our variable of interest, λ. This issue becomes clearer by considering the directed acyclic graph for the influences on justices' decisions.

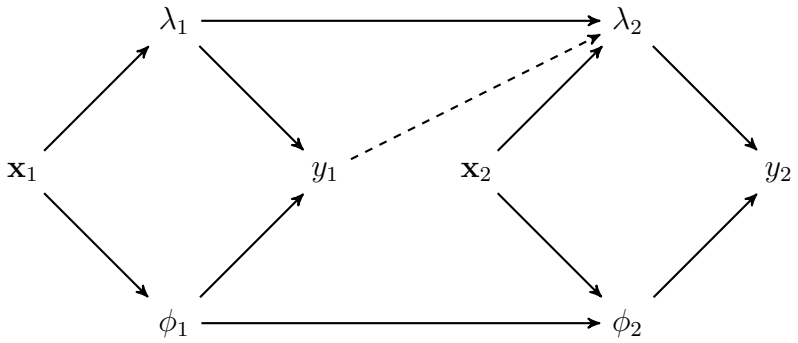


Figure 4.2: Directed acyclic graph for the influence on a justice's decisions. $\mathbf{x}_i$ denotes the facts of case i, y_i denotes the justice's decision in case i, λ_i denotes the predicted legal outcome in case i, and ϕ_i represents the justice's policy preferences in case i.

We want to estimate the influence of the law on justices decisions; in Figure 4.2, this is represented by the arrows from λ_i, the predicted legal outcome in case i to y_i, the justice's decision in case i. However, this relationship can be confounded by the justice's own policy preferences that get "baked in" to the law when they are a part of the majority coalition in a case; in Figure 4.2, this is represented by the dashed arrow from y_i to λ_{i+1}. This line is dashed to illustrate that it is a *potential* pathway;[18] the justice's preferences only become part of the law if they are in the majority coalition. If we do not block these dashed pathways, *when they are active*, our estimate of the effect of λ may be biased[19] by the justice's mere agreement with their own past self rather than acquiescence to the law. So, consider the sequence of events. In case 1, controlling for $\mathbf{x}_1$ is sufficient to identify the effect of λ_1; then λ_1 has no path to y_1 through $\mathbf{x}_1$ and ϕ_1. However, in case 2, *if* the justice was part of case 1's majority coalition, controlling for $\mathbf{x}_2$ is no longer sufficient; a

18. Technically in a DAG all pathways are considered potential pathways, but here we observe whether the pathway is active or not, which is what we emphasize.

19. In the Bayesian setting, we are not strictly speaking concerned with the frequentist conception of bias. However, we are concerned with the correct interpretation of the parameters we estimate. Without conditioning on the proper set of variables, the average marginal effect of λ_i we estimate will not represent solely the constraining effect of the law, but may also include an effect of agreement with your own past self, whose views have been partially enshrined in the law.

path then runs from λ_2 to y_2 through $y_1, \phi_1,$ and ϕ_2.

This is a form of time-varying confounding, a particularly difficult issue for inference. Political science has recently begun taking cues from biostatistics to deal with various forms of time-varying confounding, such as using structural nested mean models (Acharya, Blackwell, and Sen 2016), or marginal structural models (MSM) (Robins, Hernán, and Brumback 2000; Torres 2020) to estimate controlled direct effects of past treatments when there are intervening treatments, or to estimate cumulative effects of dynamic treatments with MSM (Blackwell 2013). However, while these frameworks account for dynamic treatments, they do not account for dynamics where past outcomes can affect current treatments.[20]

Sequential conditional mean models (Liang and Zeger 1986) are sometimes used with longitudinal data. These models are closer to our setting, where past outcomes, treatments, and confounders can all have an effect on present treatment. However, the approach does not directly extend to our setting, as the past outcomes' effects on current treatments are actually a function of present covariates, rather than an unconditional effect.

In our case, we can take advantage of the nature of $\boldsymbol{\lambda}$ as the prediction from a GP model to give us a direct way to control for the effect of past values of $\mathbf{y}$ that have impacted $\boldsymbol{\lambda}$. For case i,

$$\lambda_i = \sum_{t=1}^{i-1} k\left(\mathbf{x}_i, \mathbf{x}_t\right) \alpha_t, \tag{4.5}$$

$$\boldsymbol{\alpha} = \nabla \log p\left(\overline{D}_{i-1} \mid \overline{\boldsymbol{\lambda}}_{i-1}\right), \tag{4.6}$$

where $\overline{D}_{i-1}$ and $\overline{\boldsymbol{\lambda}}_{i-1}$ indicate the decisions of the Court and the predicted legal outcomes in cases before case i. Letting $\mathcal{Y}_i$ be the set of indices of the justice's $\mathbf{y}$ decisions that have been in the majority up to case i, then

20. Additionally, the MSM framework is used for discrete treatment values rather than our continuous λ.

$$\rho_i \triangleq \sum_{t \in \mathcal{Y}_i} k\left(\mathbf{x}_i, \mathbf{x}_t\right) \alpha_t \tag{4.7}$$

gives the total impact justice j's preferences have had on λ_i. In other words, λ_i is a weighted sum, and the elements of that sum attributable to justice j's chosen outcomes are directly identifiable, so we can control for them using that portion of the weighted sum. Thus, controlling for ρ_i blocks the backdoor path from λ_i to y_i through $\overline{\mathbf{y}}_{i-1}$ and $\overline{\boldsymbol{\phi}}_{i-1}$, allowing us to obtain a correct estimate of the effect of λ on y, *assuming* all other confounders on present decisions are accounted for.

4.3.4 Model specification

For the Court-level model, I use as predictors each of the case factors identified by prior literature as relevant: the *Category* of restriction, the *Actor* imposing the restriction, the *Restriction* itself, and the *Identity* of the speaker. I also use the *Term* of the Court as a predictor, to allow doctrine to fluctuate over time. Finally, I include the *Median Martin-Quinn Score* on the Court to capture any remaining ideological nature of these cases outside the dimensions of preference given by the case factors. To ensure sufficient training data in the Court-level model, and sufficient observations in the justice-level models, I use data before the appointment of Rehnquist as purely training data and focus on the final natural court in the Rehnquist court—consisting of Chief Justice Rehnquist and Justices Breyer, Ginsburg, Kennedy, O'Connor, Scalia, Souter, Stevens, and Thomas—for analysis.[21]

I cannot optimize the prior in the Court-level model as a model-selection step or ρ would be infected by other information in the prior optimization step, so I specify at the outset a somewhat agnostic prior for the Court-level model. I use a scale factor of approximately 2.2, or the 90% quantile of the logistic distribution. This corresponds to an assumption that on average, cases that make it to the Supreme Court will have a probability of between 0.2 and 0.8 of having a liberal

21. For completeness I also conduct analysis on the other justices as well; the results in these other models are substantively similar to those presented in the main paper and are reported in Appendix C.2.

outcome; that is, we keep the model flexible enough to predict extreme outcomes but assume that most cases that make it to the Supreme Court are not "easy cases". For the length scales, I use a length scale of one for each category of every case factor so that we are not *a priori* imposing a relative importance of case factors but letting the model learn which case factors are more important over time as we add data. Finally, for the *Term* and *Median Martin-Quinn Score* variables, I use the inter quartile range for the length scale, which is about 24 and 0.11 respectively. This allows the model to understand these variables are on a much larger and smaller scale respectively than the others.

In the justice-level models, I use each of the case factors as predictors to capture the justice's own preferences, as well as the justice's individual *Martin-Quinn Score* to capture any residual ideological influence on their votes. Again, I include the *Term* to allow justices' preferences to vary over time. Finally, I include λ, or the predicted outcomes from the Court-level model, to represent the law, and ρ to control for the justice's own contribution to λ. As is standard for GP classification, I first engage in a model-selection step to set the prior's hyperparameters, then fit the justice-level models to their individual voting records and calculate the average marginal effect γ as described in Chapter 3. All analyses were conducted in `R` (R Core Team 2021) utilizing the R extension package `gpmss` (Duck-Mayr 2021) for GP classification-specific functionality.

4.4 Results

The effect of λ is listed for each justice in Table 4.1 with 95% credible intervals.[22] For interpretability, I calculated the effect on the probability scale, and as a discrete difference (also called "first differences") rather than marginal effect of instantaneous change. As choosing any particular points at which to calculate the discrete difference is somewhat arbitrary, I report two different choices of difference points: A change between the mean of λ minus one standard deviation and

22. Average marginal effects for all predictors are given in Appendix C.1

the mean plus one standard deviation (-1.05 and 1.27, respectively), and a change between the minimum and maximum values of λ (-2.42 and 3.09 respectively); both are common choices for reporting first differences. The mean plus and minus a standard deviation also offers a nice substantive interpretation: We are comparing a point at which we are fairly certain the law implies a liberal ruling ($\lambda = 1.27$ corresponds to a 78% probability the ruling should be liberal according to the Court's past decisions) and a point at which we are fairly sure the law implies a conservative ruling ($\lambda = -1.05$ corresponds to a 26% probability the ruling should be liberal according to the law). So the effect tells us how much more likely a justice is to vote liberally in a case if the Court's past decisions imply a 78% probability the "correct legal outcome" in the case is liberal versus a 26% probability it is liberal. This effect is averaged over all observations in the sample. For example, the average marginal effect of λ for Justice Kennedy is 0.13 with a 95% credible interval of [0.06, 0.22]. This means that for every set of case facts Kennedy was actually presented, on average Kennedy's probability of voting liberally would increase by 13% if past cases implied a probability of 78% the correct legal outcome is liberal versus if past cases implied a probability of 26% the correct outcome is liberal. For the mean plus and minus a standard deviation, I also depict the average marginal effect estimates in Figure 4.3 with 90% and 95% credible intervals.

We see that for most justices, the law exerts a reliable influence on their decision making; the 95% credible interval for Justices Breyer, Kennedy, O'Connor, Scalia, Stevens and Chief Justice Rehnquist contains only positive values at both difference levels. However, for other justices, the credible intervals bound zero, indicating the law does not have a reliable effect on their decisions. Justices Souter and Thomas fall into this group. In some ways, this affirms past results in the literature; some justices act relatively unconstrained as argued in Segal and Spaeth (2002). However, contrast the closest analogous attitudinalist finding: Segal and Spaeth (1996) studied dissenters to landmark cases and their subsequent votes in related cases. They found the justices *most* deferential to the precedent they disagreed with still voted in line with their preferences (and against the legal outcome) two thirds of the time, and for a supermajority of justices, that occurred

Table 4.1: Average marginal effect of λ on judges' decisions.

	Mean of $\lambda \pm$ sd of λ	Range of λ
Breyer	0.07	0.18
	[0.01, 0.13]	[0.12, 0.25]
Ginsburg	-0.14	-0.29
	$[-0.23, -0.05]$	$[-0.39, -0.18]$
Kennedy	0.13	0.31
	[0.06, 0.22]	[0.22, 0.39]
O'Connor	0.08	0.21
	[0.04, 0.13]	[0.16, 0.25]
Rehnquist	0.05	0.12
	[0.00, 0.09]	[0.07, 0.16]
Scalia	0.10	0.22
	[0.02, 0.17]	[0.13, 0.31]
Souter	0.04	0.11
	$[-0.12, 0.21]$	$[-0.07, 0.27]$
Stevens	0.06	0.16
	[0.02, 0.11]	[0.11, 0.20]
Thomas	0.02	0.06
	$[-0.06, 0.10]$	$[-0.04, 0.15]$

Note: For each justice, I report the estimate and 95% confidence interval for the difference in the probability the justice will vote liberally between two different values of λ, averaged over all observations in the sample.

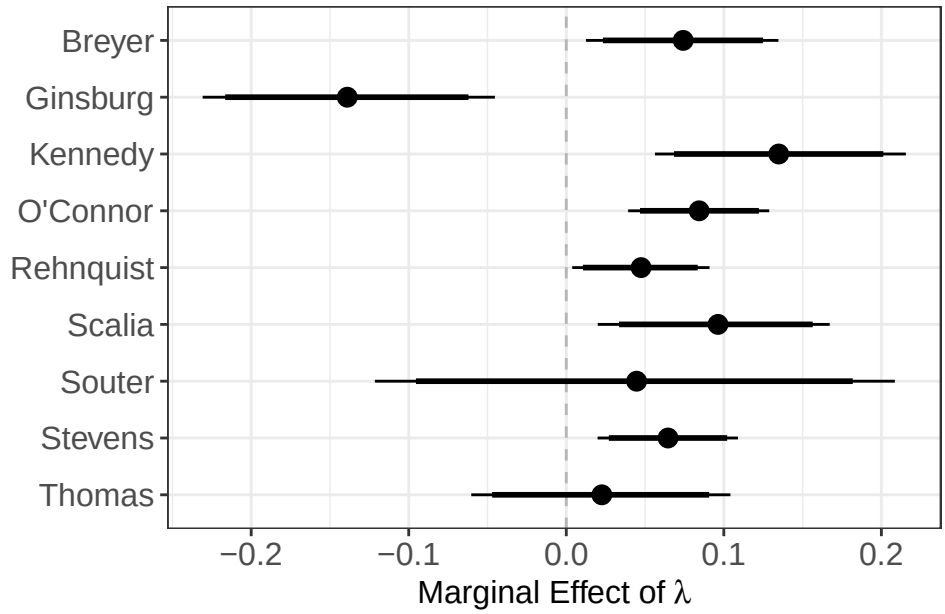


Figure 4.3: Average marginal effect of λ on judges' decisions. Estimates are depicted with circles, 90% credible intervals are depicted with thick line segments, and 95% credible intervals are depicted with thinner line segments. Effects are on the probability scale and calculated for a concrete difference; they reflect the difference in the probability the justice will vote liberally when the Court's past decisions give a 0.78 probability the outcome should be liberal vs. a 0.26 probability the outcome should be liberal (corresponding to latent legal outcomes of 1.27 and -1.05 respectively—the mean of λ plus and minus one standard deviation).

over 90% of the time. By contrast, here we see reliable evidence that the law on average affects the decisions of a supermajority of the Court![23]

Similarly, when we consider studies that do find some constraining effect of law, there are important differences to the results here. First, and perhaps most importantly, past work in this area has often been clouded by methodological issues, casting doubt on even the mixed evidence offered in favor of legal constraint (see Lax and Rader 2010a). This study uses an approach that not only handles error correlation over (e.g.) term,[24] but crucially takes into account the "state dependence" problem, or the problem that justices' own preferences from the past get incorporated into the law for the future. Without devising a control for the justices' own influence on the current legal status quo, any constraining effect of law found could be a spurious effect. An important limitation in this approach is it requires us to identify the case factors relevant to judges' preferences and court outcomes. However, it allows us to control for this issue when we believe we have identified these factors, and there is strong evidence in several contexts of the case factors we must include (see Segal 1984; Richards and Kritzer 2002). By addressing the state dependence issue in contexts where we can identify these factors, I offer the best available evidence to date of the constraining effect of law.[25]

Next, there are contextual differences. For example, Bailey and Maltzman (2008) focus on justices' willingness to explicitly overrule precedent; while important to assess, this is decidedly a narrower focus than the present study, which looks for the average influence the law has on justices' decisions. Similar contextual differences separate the present study from past work (e.g. Black and Owens 2009).

23. That is, a majority of the natural court studied.

24. As Carlson (2021) explains, GP models are a natural answer to the common methodological problem in political science of violating an assumption of conditional independence.

25. A potential additional consideration in the Free Expression context is the ideological nature of the expression in question. Epstein, Parker, and Segal (2018) show justices may provide more protection to speech they agree with (for example, a conservative justice ruling to protect religious expression) than speech they disagree with (for example, a liberal justice refusing protection to commercial speech). I include the variables from Epstein, Parker, and Segal (2018) in the justice-level models in Appendix C.3, which gives substantively similar results.

There are two crucial differences to Bartels (2009). First, the approach here allows for law to influence justices differently; I allow Justice Breyer to consider the current state of the law more important in his decision making than Justice Scalia does. In contrast, Bartels (2009) focuses on an effect that is homogeneous across justices by construction. Moreover, Bartels (2009) treats the case factors as relevant *legal* factors, but uses Martin-Quinn scores *only* to capture the justices' preferences, while we would expect from theoretical approaches like the attitudinal model or the case space model that we should in fact also treat justices' issue-specific preferences as defined with reference to the case factors. The modeling approach here allows us to capture both justice-level preferences and legal implications with reference to the case factors.

The average marginal effect of law for Justice Ginsburg is reliably negative. This means that as the Court's past decisions indicate a higher probability that the correct legal outcome is liberal, Justice Ginsburg becomes more likely to vote conservative. While it is possible there is some unmeasured aspect of these cases that cause Justice Ginsburg's preferred outcomes to differ from the law, which could result in the negative marginal effect we see, it is also possible the mechanism discussed in Section 4.2 is at play: Ginsburg may be affected by the law in cases where her vote is not pivotal. As shown in Figure 4.4, that is indeed the case. The correlation between Ginsburg's latent outcomes and the latent legal outcome is 0.19 [-0.12, 0.47] when Ginsburg might be pivotal (i.e., the correlation is not reliably negative), but is -0.32 [-0.50, -0.12] when Ginsburg is likely not pivotal.[26] This provides suggestive though not conclusive evidence that the mechanism from Section 4.2 is driving the result for Ginsburg rather than omitted variable bias.

Another way to contextualize the effect of the law on justices' decisions—and case outcomes—is to determine how often a change in the legal status quo would cause a justice to change their

26. Notice that I use a pivotality threshold of six here, whereas strictly speaking a justice is only pivotal if the majority coalition is five justices. Using the five justice threshold gives similar results: 0.44 [0.06, 0.71] and -0.32 [-0.48, -0.14] respectively. However, I use the higher threshold to add observations to the pivotal group, to avoid concern that the pattern was driven by too few observations in the pivotal group. This also serves to cover cases where Ginsburg is in truth pivotal but her presence in the coalition causes another justice to join, or where Ginsburg has some uncertainty about whether she would be pivotal.

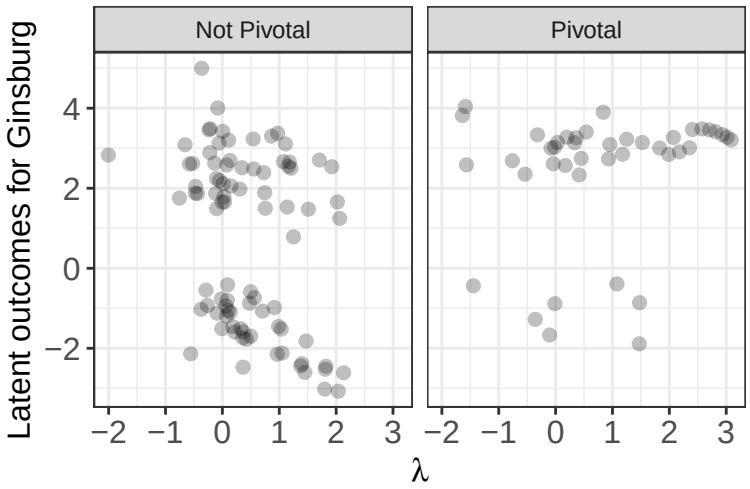


Figure 4.4: Comparing predicted outcomes for Justice Ginsburg against λ. Values are depicted with gray circles. The right panel labeled "Pivotal" contains observations from all cases where Ginsburg was part of a majority coalition of six or fewer justices—the situations where Ginsburg may have been pivotal. The left panel labeled "Not Pivotal" contains observations from all remaining cases.

vote, and in turn how often the outcome in the case would change. For each vote cast by a justice, I simulate whether their vote would switch from an anti-expression vote to a pro-expression vote when λ moves from its mean minus its standard deviation (i.e., a situation where the legal status quo says there is a 0.26 probability the legal outcome should be pro-expression) to its mean plus its standard deviation (i.e., a situation where the legal status quo says there is a 0.78 probability the legal outcome should be pro-expression). I generate 10,000 draws of whether a switch occurs for each justice's vote in each case to generate a credible interval for the proportion of their votes that change. For each case, I also determine how often the outcome the Court would assign to the case changed. The proportion of cases whose justice- and case-level outcomes would change are presented in Figure 4.5. The effect of the law is substantial enough that over 11% of the cases the Court heard would move from an anti-expression outcome to a pro-expression outcome with such a change in the status quo.

Is the constraining effect of law here driven by situations in which doctrine allows little ideological discretion, while judges can follow their policy preferences more in less certain cases? Bartels (2009) presents evidence that ideological constraint is higher in cases where more "certain"

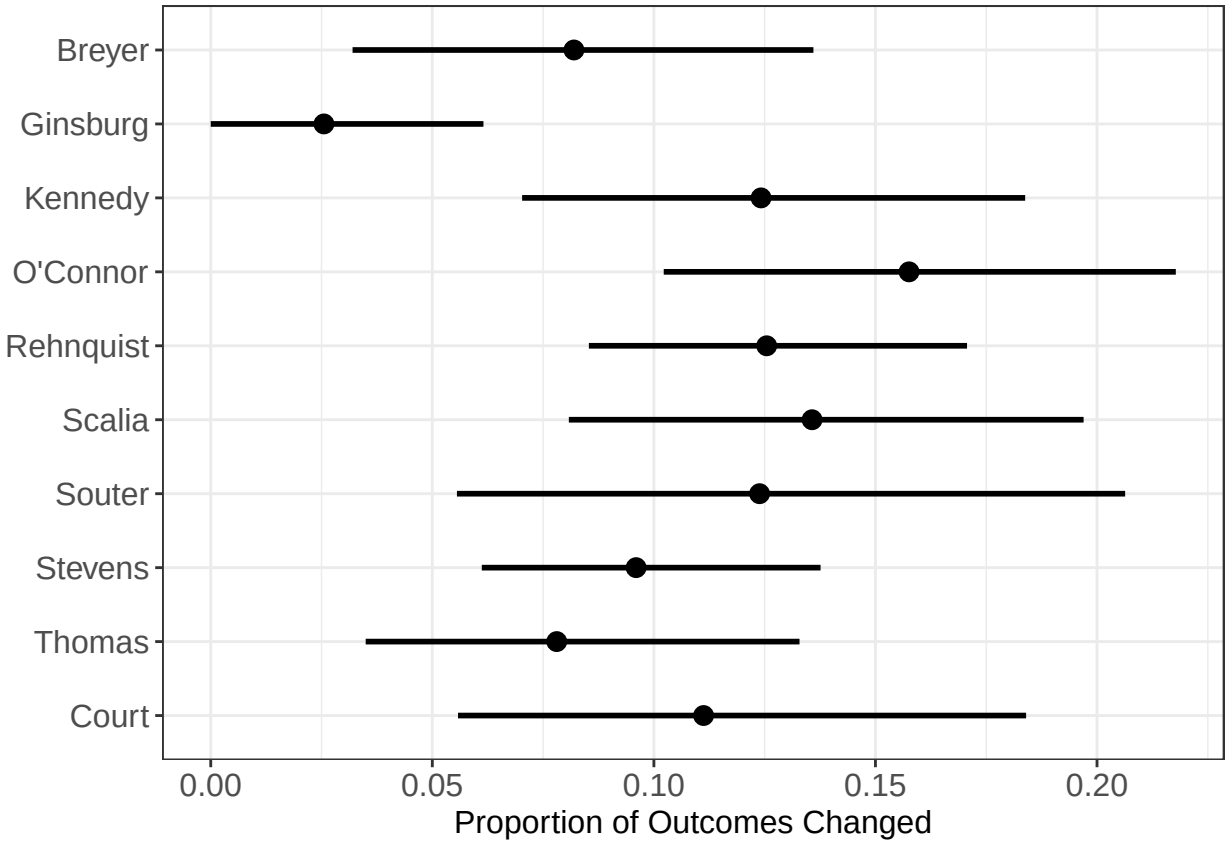


Figure 4.5: Proportion of justice- and case-level outcomes that would change from anti-expression to pro-expression given a change in λ from -1.05 to 1.27 with 95% credible interval.

legal tests such as the rational basis or strict scrutiny tests should apply than in cases where a more fluid test such as intermediate scrutiny should apply. In the Free Expression context, this means we should see a higher marginal effect of law in "Content based" restriction and "Less protected" forms of expression cases than in "Content neutral" restriction cases. Figure 4.6 shows the marginal effect of λ averaged over each case that each justice heard from each of those three categories.

While we do not see a difference in magnitude of the effect of law by category as Bartels (2009) found, we see an effect in certainty; the credible intervals for the "Content based" and "Less protected" categories are somewhat tighter than that for the "Content neutral" category. However, this is likely simply caused by the fact that the "Content neutral" category has fewer cases in it, as shown in Table 4.2.

Table 4.2: Number of observations in each category for each justice.

Justice	Less protected	Content neutral	Content based
Breyer	35	19	71
Ginsburg	36	19	75
Kennedy	53	25	107
O'Connor	73	27	125
Rehnquist	145	42	188
Scalia	58	26	114
Souter	37	19	70
Stevens	120	36	171
Thomas	42	21	80

So we can consider the more general formulation from Section 4.2: If justices only feel the constraining effect of law when legal outcomes are more certain (thus giving them less ideological discretion), the pointwise partial derivatives, giving the effect of law in each case, should be increasing in $|\lambda|$. However, as shown in Figure 4.7, this is not the case, providing suggestive evidence that for justices constrained by the law, they feel some *internal* constraint of the law, or in other words, a preference for following the law that can conflict with—and perhaps in some cases override—their sincere policy preferences.

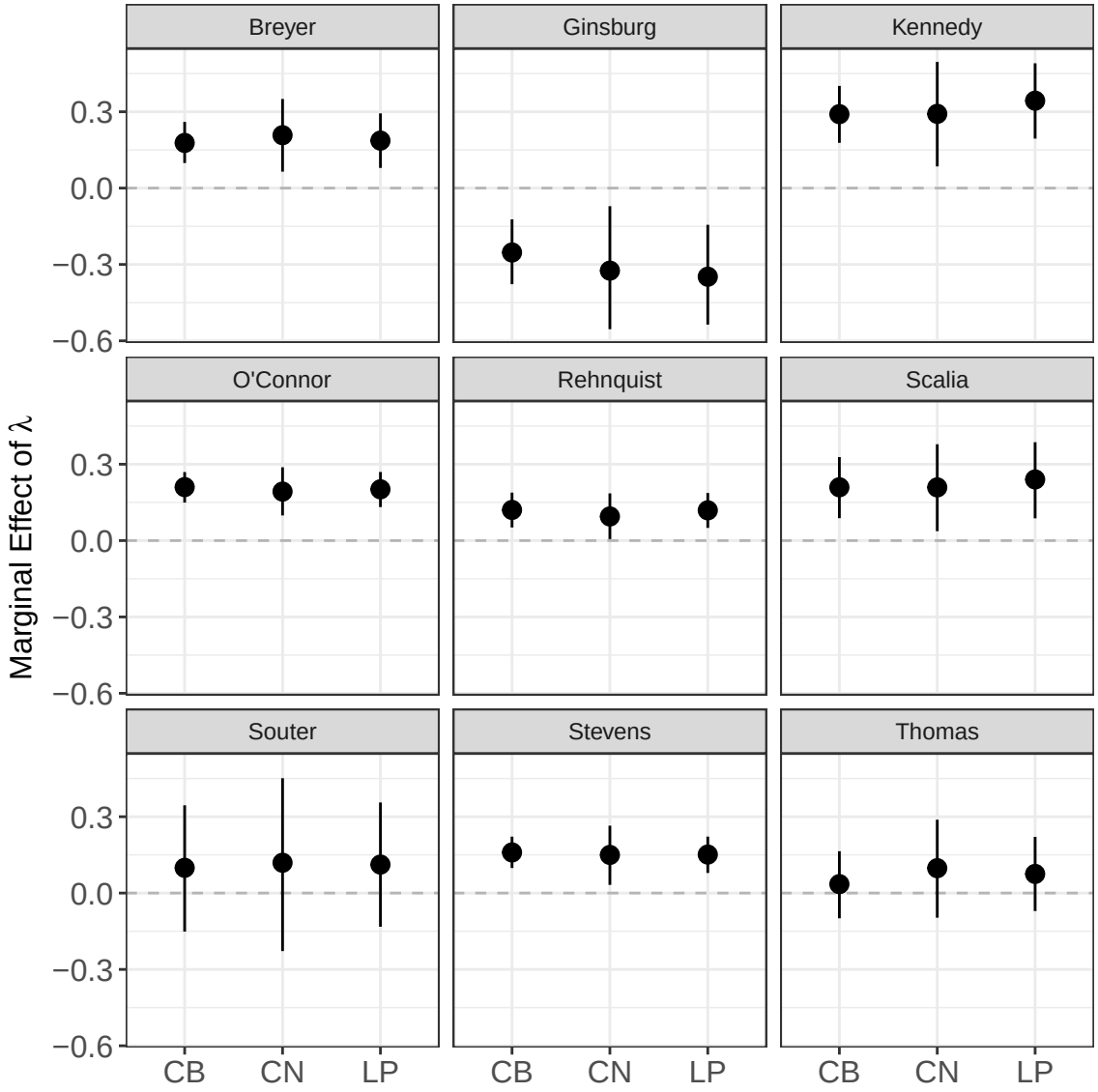


Figure 4.6: Average marginal effect of λ by category. The effect displayed in this figure uses the difference between minimum and maximum values of λ. Content based is abbreviated as CB, Content neutral as CN, and Less protected as LP.

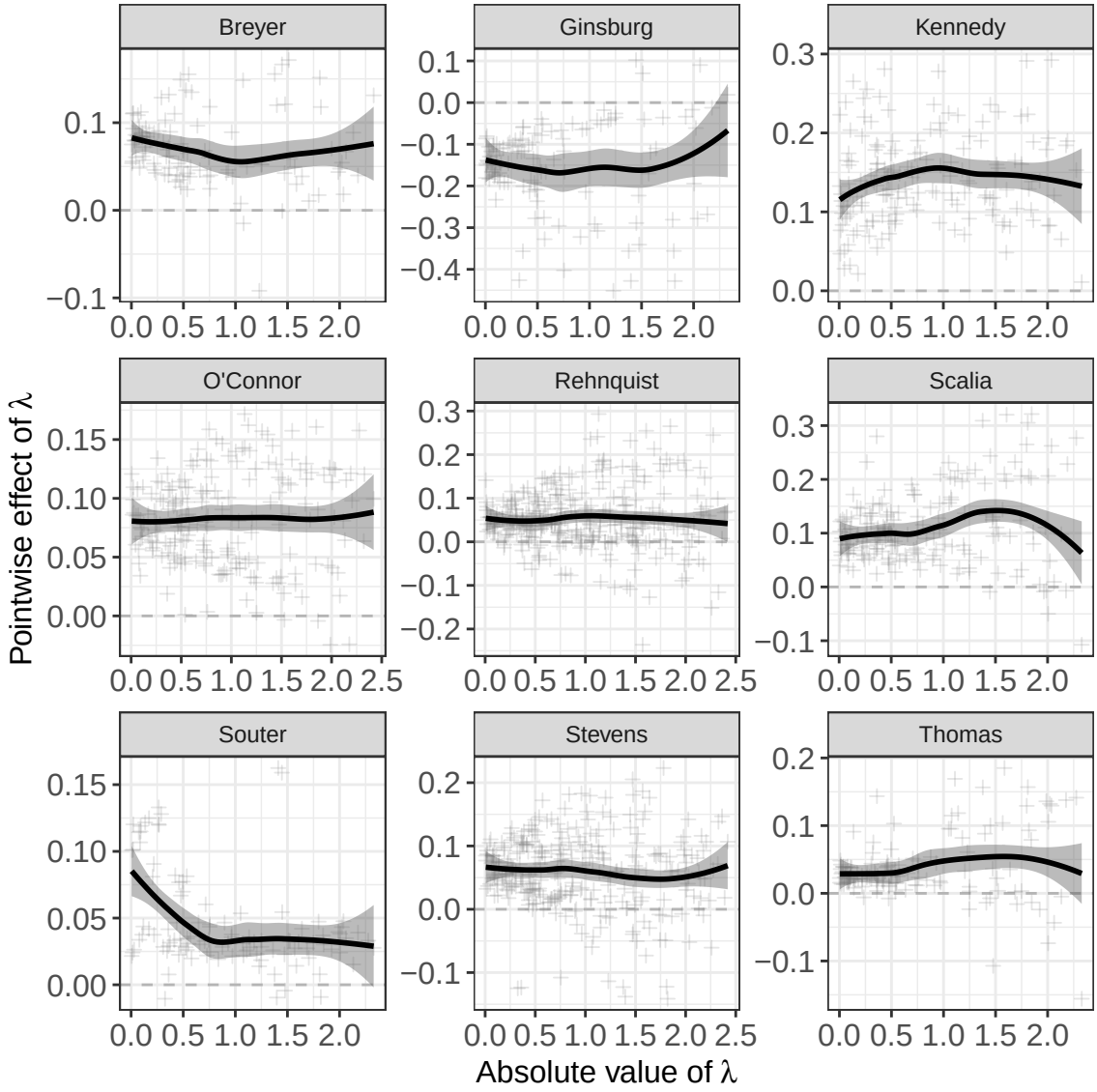


Figure 4.7: Effect of λ by absolute value of λ for each justice. Pointwise effects depicted with light gray crosses, LOESS smoother depicted with a black line, 95% CI for the smoother depicted with gray shaded region.

4.5 Conclusion

I asked at the outset, "(How much) does the law affect judges decisions?" The answer is the law does affect judges decisions, substantially... at least some of them. Several studies approach the question of legal influence on judicial decision making using contexts other than votes on the merits and find the legal effects in decisions such as agenda-setting decisions (Black and Owens 2009), citation choices (Hansford and Spriggs 2006; Hinkle 2015), or willingness to explicitly overrule precedent (Bailey and Maltzman 2008). Bartels (2009) studies justices' votes on the merits and finds a constraining effect of law, but with an analysis that pools the justices together rather than an analysis that allows some justices to be unfazed by the force of law.

I start from a general theoretical approach based on case space models (Lax 2011) and extended it to consider the law as a potential explicit influence on judges' preferred outcomes. Decision makers in this model can consult the set of cases previously decided by the Court, comporting with the typical model of legal reasoning (Levi 1949): we consider that like cases ought to be treated alike, and determine the outcome reached in the most closely analogous cases to the one we are deciding. This framework suggests a measure of the law: The predicted outcome in each case given a model trained only on the cases that came before it. I use a model to generate these predictions, GP classification (Rasmussen and Williams 2006), that accommodates realistic forms of Court doctrine better than some more restrictive approaches to modeling doctrine taken in the past (see Kastellec 2010). Importantly, this approach readily allows for a way to not only directly test the average influence the law exerts on justices' decisions, but crucially while controlling for the justices' own impact on the law—an issue of time-varying confounding that would otherwise pose a danger to inference. With the effect of law thus carefully identified, I apply the method to the natural court beginning with the appointment of Stephen Breyer and ending with the end of the Rehnquist Court. I show several justices exhibit reliable constraint from the law in their decision making, though others do not.

This study makes several major contributions. First, I provide a fresh theoretical perspective on legal constraint by conceptualizing the law as the implied outcome in each case given the cases that came before it. Second, I present credible evidence of a substantial constraining effect of law for some justices by addressing the danger to inference presented by the justices' own votes affecting the legal status quo whose effect we are trying to measure. Third, developing this measure of the legal status quo in itself is a contribution; status quo policy is an important consideration in a wide variety of political contexts. While, for example, Black and Owens (2009) simply uses the median Martin-Quinn score of the lower court panel to measure the legal status quo, the measure I develop provides the legal outcome implied by the Court's past decisions, a quantity that is a closer theoretical match to the concept of the legal status quo. I also provide suggestive evidence differentiating between reasons why the law matters, indicating some justices have a preference for following the law rather than seeing it as a constraint due to (for example) legitimacy needs, which few studies have attempted. Finally, I highlight common mismatches between methods and theory in studies of judicial politics: judges' preferences within an issue area are best conceptualized as multidimensional, and with respect to case facts; and empirical models seeking to capture either the law or judges' individual preferences with respect to case facts should be flexible enough to accommodate any shape rather than imposing the strict assumption of linearity as in past studies. Thus, I provide the most convincing evidence to date of a substantively important effect on judicial decision making as well as provide methodological tools and measurement strategies useful for future research in judicial politics.

Chapter 5

Conclusion

At base, the study of judicial politics is about determining how judicial behavior and/or institutions affect the decisions and policy output of courts, and how existing law affects the choices judges make. In this dissertation I tackle important aspects of both questions, and make advancements in political methodology in the process.

I first examine how preference and judgment aggregation on collegial courts affects the consistency of legal doctrine. Using a social choice theoretical model, I show that because judges must rely on abstraction to describe their preferred legal rules, only a very restrictive class of preferences for the judges can ensure consistency of legal doctrine from collegial courts. In addition to its intrinsic substantive importance, this result also has important implications for other areas of judicial politics, highlighting another area in which appellate court principals must balance control over lower court agents with their other aims (such as consistency of doctrine) and new concerns and avenues of research regarding predicting judges' choices using facts of cases.

I also study the constraining effect of law at the US Supreme Court, which first requires some methodological advancements. To identify the effect of law on the justices' choices, I need both (1) a measure of the legal status quo (i.e. what the outcome implied by the law is in the case before the justices) and (2) a way to control for the information about the justices' own preferences over

outcomes that has leaked into the current state of the law because of their votes in past cases. As explained by Kastellec (2010), models with strict functional form assumptions are inappropriate for measuring legal rules, so I turn to a model from the machine learning literature: Gaussian process (GP) classification.

Political scientists are beginning to make use of GP models (Duck-Mayr, Garnett, and Montgomery 2020; Carlson 2021; Gill 2021), but because these models hail from the machine learning literature where the focus is on prediction rather than inference regarding predictors, I had to advance the state of the art for these models and derive the distribution of marginal effects of predictors on outcomes. I thus present a chapter where a detailed primer on GP models is given for political scientists, including practical guidance in their use for applied researchers and deriving inferential quantities of interest for political scientists as well as providing a free, open-source, user-friendly `R` package for these models, `gpmss` (Duck-Mayr 2021).

I use these tools in my investigation of the effect of legal constraint at the US Supreme Court. I use GP classification to develop a measure of the legal status quo, which trains a model on court-level outcomes from all previously decided cases and predicts out of sample for the case before the justices in each case. I show how to exploit mathematical properties of this measure to control for the information about justices' preferences that has leaked into the measure of the legal status quo, allowing me to more plausibly identify the causal effect of past precedent. I find the law exerts a statistically reliable constraining effect on a supermajority of Supreme Court justices, both a normatively important finding and one that informs our understanding and future study of judicial behavior.

The work in this dissertation presents multiple avenues for future research. From a methodological standpoint, the technological innovations in Chapters 2 and 3 provide scaffolds for further advancement. As mentioned in Chapter 2, study of the doctrinal dilemma from Kornhauser (1992) I build on spread throughout the social choice literature (where it was known as the "discursive dilemma" in more general contexts). The extension to multi-step reasoning could be explored

further in future theoretical papers. Additionally, Gaussian process models are ripe for further extension for tasks useful for political scientists and other social scientists, such as estimating causal effects via regression discontinuity designs (Ornstein and Duck-Mayr 2022).

Perhaps most importantly, the legal status quo measure developed in Chapter 4 will be useful for a wide array of research questions in judicial politics. First, generating this measure in additional issue areas to cover all the cases heard by the Supreme Court will result in a dataset useful for the discipline at large. Additionally, there are specific substantive questions that dovetail with the studies presented in this dissertation. For example, this approach can help bring a new perspective and evidence to the study of legal change. Unlike past studies such as Richards and Kritzer (2002), Bartels and O'Geen (2015) that look at when decision making at the Court changes, using this measure of the legal status quo will allow me to explore when "the prophecies of what the courts will do" (Holmes 1897) change. That is, the law shapes outside actors' behavior by shaping their expectations, so an important but less explored question regarding legal change is not when the judges' behavior in cases changes, but when our expectations about what they will do in cases changes, which this measure will help illuminate.

This dissertation presents several implications for judicial politics. First, when collegial appellate courts craft legal rules, unless the appellate judges' preferences over legal rules meet very strict assumptions or they defer to lower courts' findings on intermediate legal conclusions, inconsistent legal doctrine may result. This has important normative implications, can affect courts' efficacy as policy makers, and exposes another aspect of the principal-agent problem in the judicial hierarchy. I also show that the most common approach in prior literature to empirically modelling the relationship between case facts and outcomes is often inappropriate, and offer Gaussian process classification as a solution. I also show how to use GP classification to generate a measure of the legal status quo—what outcome would be implied in cases using only the information we can derive from past cases alone—a concept that features prominently in many studies of judicial politics. Finally, I provide the most reliable evidence to date on the extent to which past precedent

constrains the choices of US Supreme Court justices, a fundamental and long-running debate in judicial politics.

Bibliography

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110 (3): 512–529.

Badawi, Adam B., and Scott Baker. 2015. "Appellate Lawmaking in a Judicial Hierarchy." *The Journal of Law & Economics* 58 (1): 139–172.

Bailey, Michael A., and Forrest Maltzman. 2008. "Does Legal Doctrine Matter? Unpacking Law and Policy Preferences on the U.S. Supreme Court." *American Political Science Review* 102 (3): 369–384.

———. 2011. *The Constrained Court: Law, Politics, and the Decisions Justices Make.* Princeton, NJ: Princeton University Press.

Baker, Scott, and Lewis A. Kornhauser. 2015. "A Theory of Judicial Deference." Unpublished manuscript. https://perma.cc/C4QG-LJ46.

Bartels, Brandon L. 2009. "The Constraining Capacity of Legal Doctrine on the U.S. Supreme Court." *American Political Science Review* 103 (3): 474–495.

———. 2011. "Choices in Context: How Case-Level Factors Influence the Magnitude of Ideological Voting on the U.S. Supreme Court." *American Politics Research* 39 (1): 142–175.

Bartels, Brandon L., and Andrew J. O'Geen. 2014. *Replication data for: The Nature of Legal Change on the U.S. Supreme Court: Jurisprudential Regimes Theory and Its Alternatives.* Version V3. Harvard Dataverse. https://doi.org/10.7910/DVN/26522.

———. 2015. "The Nature of Legal Change on the U.S. Supreme Court: Jurisprudential Regimes Theory and Its Alternatives." *American Journal of Political Science* 59 (4): 880–895.

Black, Ryan C., and Ryan J. Owens. 2009. "Agenda Setting in the Supreme Court: The Collision of Policy and Jurisprudence." *Journal of Politics* 71 (3): 1062–1075.

Blackwell, Matthew. 2013. "A Framework for Dynamic Causal Inference in Political Science." *American Journal of Political Science* 57 (2): 504–520.

Callander, Steven, and Tom S. Clark. 2017. "Precedent and Doctrine in a Complicated World." *American Political Science Review* 111 (1): 184–203.

Cameron, Charles M., Jeffrey A. Segal, and Donald Songer. 1994. "Strategic auditing in a political hierarchy: An informational model of the Supreme Court's certiorari decisions." *American Political Science Review* 94 (1): 101–116.

Campello, Daniela, and Cesar Zucco Jr. 2015. *Presidential Success and the World Economy.* V. 2. Harvard Dataverse, November 4, 2015. https://doi.org/10.7910/DVN/XG6QQX.

———. 2016. "Presidential Success and the World Economy." *Journal of Politics* 78 (2): 589–602.

Carlson, David. 2021. "Modeling Without Conditional Independence: Gaussian Process Regression for Time-Series Cross-Sectional Analyses." Under review, https://mysite.ku.edu.tr/dcarlson/research/.

Cheng, Lu, Siddharth Ramchandran, Tommi Vatanen, Niina Lietzén, Riitta Lahesmaa, Aki Vehtari, and Harri Lähdesmäki. 2019. "An Additive Gaussian Process Regression Model for Interpretable Non-Parametric Analysis of Longitudinal Data." *Nature Communications* 10 (1): 1–11.

Clark, Tom S. 2016. "Scope and precedent: judicial rule-making under uncertainty." *Journal of Theoretical Politics* 28 (3): 353–384.

Clark, Tom S., and Clifford J. Carrubba. 2012. "A Theory of Opinion Writing in a Political Hierarchy." *Journal of Politics* 74 (2): 584–603.

Collins, Paul M., Jr. 2008. "The Consistency of Judicial Choice." *Journal of Politics* 70 (3): 861–873.

Denison, Alexander, Justin Wedeking, and Michael A. Zilis. 2020. "Negative Media Coverage of the Supreme Court: The Interactive Role of Opinion Language, Coalition Size, and Ideological Signals." *Social Science Quarterly* 101 (1): 121–143.

Dinh, Viet D. 2000. "Appellate Lawmaking in a Judicial Hierarchy." *Georgetown Law Journal* 88:2085–2118.

Drahozal, Christopher R. 2004. *The Supremacy Clause: A Reference Guide to the United States Constitution.* Westport, Connecticut: Praeger.

Duck-Mayr, JBrandon. 2021. *gpmss: Gaussian Process Models for Social Science.* R package version 0.1.1. https://github.com/duckmayr/gpmss.

Duck-Mayr, JBrandon, Roman Garnett, and Jacob Montgomery. 2020. "GPIRT: A Gaussian Process Model for Item Response Theory." In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI),* edited by Jonas Peters and David Sontag, 124:520–529. Proceedings of Machine Learning Research. PMLR. http://proceedings.mlr.press/v124/duck-mayr20a.html.

Epstein, Lee, and Jack Knight. 1998. *The Choices Justices Make.* CQ Press.

Epstein, Lee, Christopher M. Parker, and Jeffrey A. Segal. 2018. "Do Justices Defend the Speech They Hate? An Analysis of In-Group Bias on the U.S. Supreme Court." *Journal of Law and Courts* 6 (2): 237–262.

Fox, Justin, and Georg Vanberg. 2014. "Narrow versus borad judicial decisions." *Journal of Theoretical Politics* 26 (3): 355–383.

Gibson, James L., and Gregory A. Caldeira. 2011. "Has Legal Realism Damaged the Legitimacy of the U.S. Supreme Court?" *Law & Society Review* 45 (1): 195–219.

Gill, Jeff. 2021. "Measuring Constituency Ideology Using Bayesian Universal Kriging." *State Politics & Policy Quarterly* 21 (1): 80–107.

Hainmueller, Jens, and Chad Hazlett. 2014. "Kernel Regularized Least Squares: Reducing Misspecification Bias With a Flexible and Interpretable Machine Learning Approach." *Political Analysis* 22 (2): 143–168.

Hansford, Thomas G., and James F. Spriggs II. 2006. *The Politics of Precedent.* Princteon University Press.

Hinkle, Rachael K. 2015. "Legal Constraint in the US Courts of Appeals." *Journal of Politics* 77 (3): 721–735.

Hoffman, Adam. 2001. "Corralling Constitutional Fact: De Novo Fact Review in the Federal Appellate Courts." *Duke Law Journal* 50:1427–1466.

Holmes, Oliver Wendell, Jr. 1897. "The Path of the Law." *Harvard Law Review* 10:457–478.

Johnson, Tyler, and Erica Socker. 2012. "Actions, Factions, and Interactions: Newsworthy Influences on Supreme Court Coverage." *Social Science Quarterly* 93, no. 2 (434–463).

Kastellec, Jonathan P. 2010. "The Statistical Analysis of Judicial Decisions and Legal Rules With Classification Trees." *Journal of Empirical Legal Studies* 7 (2): 202–230.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the most of statistical analyses: Improving interpretation and presentation." *American Journal of Political Science* 44 (2): 341–355.

Klein, David. 2017. "Law in Judicial Decision-Making." Chap. 12 in *The Oxford Handbook of U.S. Judicial Behavior,* edited by Lee Epstein and Stefanie A. Lindquist. Oxford University Press.

Kornhauser, Lewis A. 1992. "Modeling Collegial Courts II. Legal Doctrine." *Journal of Law, Economics, & Organization* 8 (3): 441–470.

Krehbiel, Keith. 1998. *Pivotal Politics.* University of Chicago Press.

Landa, Dimitri, and Jeffrey R. Lax. 2008. "Disagreements on Collegial Courts: A Case-Space Approach." *Journal of Constitutional Law* 10 (2): 305–329.

———. 2009. "Legal Doctrine on Collegial Courts." *Journal of Politics* 71 (3): 946–963.

Lax, Jeffrey R. 2007. "Constructing Legal Rules on Appellate Courts." *American Political Science Review* 101 (3): 591–604.

———. 2011. "The New Judicial Politics of Legal Doctrine." *Annual Review of Political Science* 14:131–157.

———. 2012. "Political Constraints on Legal Doctrine: How Hierarchy Shapes the Law." *Journal of Politicse* 74 (3): 765–781.

Lax, Jeffrey R., and Kelly T. Rader. 2010a. "Legal Constraints on Supreme Court Decision Making: Do Jurisprudential Regimes Exist?" *Journal of Politics* 72 (2): 273–284.

———. 2010b. "The Three Prongs of a Jurisprudential Regimes Test: A Response to Kritzer and Richards." *Journal of Politics* 72 (2): 289–291.

Levi, Edward. 1949. *An Introduction to Legal Reasoning.* University of Chicago Press.

Liang, Kung-Yee, and Scott L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73 (1): 13–22.

List, Christian. 2012. "The Theory of Judgment Aggregation: An Introductory Review." *Synthese* 187 (1): 179–207.

List, Christian, and Phillip Pettit. 2002. "Aggregating Sets of Judgments: An Impossibility Result." *Economics and Philosophy* 18:89–110.

Liu, Haitao, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. 2020. "When Gaussian Process Meets Big Data: A Review of Scalable GPs." *IEEE Transactions on Neural Networks and Learning Systems* 31 (11): 4405–4423. https://doi.org/10.1109/TNNLS.2019.2957109.

Maltzman, Forrest, James F. Spriggs II, and Paul J. Wahlbeck. 2000. *Crafting Law on the Supreme Court: The Collegial Game*. Cambridge University Press.

Martin, Andrew D., and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court." *Political Analysis* 10 (2): 134–153.

———. 2005. "Can Ideal Point Estimates Be Used as Explanatory Variables?" Working paper, October. https://mqscores.lsa.umich.edu.

———. 2020. *Martin-Quinn Scores*. Version 2019 Term Release, July. https://mqscores.lsa.umich.edu/measures.php.

Murray, Iain, Ryan Prescott Adams, and David J. C. MacKay. 2010. "Elliptical slice sampling." *The Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* 9:541–548.

Nehring, Klaus, and Clemens Puppe. 2006. "Consistent Judgement Aggregation: The Truth-Functional Case." *Social Choice and Welfare* 31 (1): 41–57.

———. 2010. "Justifiable Group Choice." *Journal of Economic Theory* 145 (2): 583–602.

Ornstein, Joseph, and JBrandon Duck-Mayr. 2022. "Gaussian Process Regression Discontinuity." Working Paper, https://joeornstein.github.io/publications/gprd.pdf.

Owens, Ryan J., and Justin P. Wedeking. 2011. "Justices and Legal Clarity: Analyzing the Complexity of U.S. Supreme Court Opinions." *Law & Society Review* 45 (4): 1027–1061.

Poole, Keith T., and Howard Rosenthal. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29 (2): 357–384.

Post, Robert. 1995. "Recuperating First Amendment Doctrine." *Stanford Law Review* 47 (6): 1249–1281.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Version 4.1.0. R Foundation for Statistical Computing. https://www.R-project.org/.

Rasmussen, Carl Edward, and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

Redish, Martin H., and William D. Gohl. 2017. "The Wandering Doctrine of Constitutional Fact." *Arizona Law Review* 59:289–338.

Richards, Mark J., and Herbert M. Kritzer. 2002. "Jurisprudential Regimes in Supreme Court Decision Making." *American Political Science Review* 96 (2): 305–320.

Robins, James M., Miguel Ángel Hernán, and Babette Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11 (5): 550–560.

Robinson, Matthew B., and Kathleen M. Simon. 2006. "Logical and Consistent? An Analysis of Supreme Court Opinions Regarding the Death Penalty." *Justice Policy Journal* 3 (1): 1–59.

Segal, Jeffrey A. 1984. "Predicting Supreme Court Cases Probabilistically: The Search and Seizure Cases, 1962–1981." *American Political Science Review* 78 (4): 891–900.

Segal, Jeffrey A., and Harold J. Spaeth. 1996. "The Influence of Stare Decisis on the Votes of United States Supreme Court Justices." *American Journal of Political Science* 40 (4): 971–1003.

———. 2002. *The Supreme Court and the Attitudinal Model Revisited.* Cambridge University Press.

Spaeth, Harold J., Lee Epstein, Andrew D. Martin, Jeffrey A. Segal, Theodore J. Ruger, and Sara C. Benesh. 2020. *2020 Supreme Court Database.* Version 2020 Release 01. http://supremecourtdatabase.org.

Spriggs, James F., II. 1996. "The Supreme Court and Federal Administrative Agencies: A Resource-Based Theory and Analysis of Judicial Impact." *American Journal of Political Science* 40 (4): 1122–1151.

Staton, Jeffrey K., and Georg Vanberg. 2008. "The Value of Vagueness: Delegation, Defiance, and Judicial Opinions." *American Journal of Political Science* 52 (3): 504–519.

Torres, Michelle. 2020. "Estimating Controlled Direct Effects Through Marginal Structural Models." *Political Science Research and Methods* 8 (3): 391–408.

Wahlbeck, Paul J. 1997. "The Life of the Law: Judicial Politics and Legal Change." *Journal of Politics* 59 (3): 778–802.

Westerland, Chad, Jeffrey A. Segal, Lee Epstein, Charles M. Cameron, and Scott Comparato. 2010. "Strategic Defiance and Compliance in the U.S. Courts of Appeals." *American Journal of Political Science* 54 (4): 891–905.

Will, George F. 2019. *The Supreme Court can undo past confusion with its ruling on this WWI memorial.* https://www.sltrib.com/opinion/commentary/2019/02/24/george-f-will-supreme/, February.

Appendix A

Appendix to Chapter 2

A.1 Formal Results and Proofs

Proposition A.1. *If all $\delta_j = \delta^*$, and δ^* is a balancing test, then ρ_m is a consistent rule. If δ^* is a conjunctive or disjunctive test, ρ_m need not be a consistent rule.*

Proof.

Balancing tests:

Suppose all $\delta_j = \delta^*$, a balancing test. Then when the collegial outcome is 1, d_j must lie on or beyond the hyperplane in D described by δ^* for every j in the majority coalition for any case h. Then the point in D that is a dimension-by-dimension median for the majority coalition at h must also lie on or beyond the hyperplane described by δ^*. Similarly, when the collegial outcome is -1, d_j must not lie as far as the hyperplane described by δ^* for any j in the majority coalition for any case h, and therefore the point in D that is a dimension-by-dimension median for the majority coalition at h must also not lie as far as the hyperplane described by δ^*. Therefore $\delta_m = \delta^*$ and is monotonic in f_m.

Conjunctive and disjunctive tests:

For conjunctive tests, when the collegial outcome is 1, $d_{ij} \geq \tau_i \; \forall \; i$ for every j in the majority coalition for any case h, and analogously for disjunctive tests and collegial outcomes of -1. Then the point in D that is a dimension-by-dimension median for the majority coalition at h must also satisfy that condition. However, when the collegial outcome is -1 for a conjunctive test, we can only say $\exists \; i : d_{ij} \leq \tau_i$ rather than $d_{ij} \leq \tau_i \; \forall \; i$. Then for each dimension i, there may be a member of the majority coalition with a high enough placement of d_{ij} such that $d_{im} \geq \tau_i \; \forall \; i$ even though no member of the majority coalition would assign the outcome 1. (Again, an analogous argument applies for disjunctive tests). Therefore δ_m may not be monotonic in f_m. □

Lemma A.1. *If δ_j is monotonic in f_j for every $j \in J$, then for every two cases $h, h' \in H$, $h_k \geq h'_k \; \forall \; k$ implies that the collegial outcome at h is weakly greater than the collegial outcome at h'.*

Proof. If f_j is monotonic and δ_j is monotonic in f_j, then $\delta_j(f_j)$ is monotonic in H. Then the collegial outcome set is monotonic in H by Proposition 3 from Lax (2007). □

Proposition A.2. *If $\boldsymbol{\delta}$ is not a shared balancing test, $\mathcal{F}(\boldsymbol{\delta})$ is nonempty.*

Proof. If $\boldsymbol{\delta}$ is not a shared balancing test, but every δ_j is monotonic, then by Lemma 1, the collegial outcome set is monotonic in H. Then define two parallel hyperplanes in H, partitioning it into three regions, such that cases in the least extreme region contain only cases with -1 outcomes, cases in the most extreme region contain only cases with 1 outcomes, and cases in the "middle" region contain cases with both outcomes (this third region is guaranteed to exist since $\boldsymbol{\delta}$ is not a shared balancing test). Then since $N > n$, a monotonic fact finding function $\hat{f}$ can then be constructed such that corresponding hyperplanes exist in D, where δ_m is monotonic in the most and least extreme regions in $\hat{D}$, but non-monotonicity is induced in the middle region. □

Remark. The contrast between Lemma 1 and Proposition 2 may explain why legal scholars so roundly critiqued Fourth Amendment doctrine although Segal (1984) found decisions well explained by historical facts. Segal (1984) was right that if you have enough observations (possibly

a lot depending on the structure of H), a clear relationship can be found between H and outcomes. However, because humans cannot generally think in such high dimensional spaces, a clear relationship between D and outcomes is what's needed for clear doctrine; unfortunately Proposition 2 shows that even in general settings with well-behaved individual preferences, such a relationship may not obtain.

A.2 Deference to Trial Court Findings

I refer to two types of facts, *historical facts* and *doctrinal facts*, the latter of which are sometimes what legal texts may refer to as mixed questions of law and fact. A historical fact is what a lay person may typically think of as a fact: whether or not an accused murderer's victim is deceased, whether or not a traffic light was green, etc. A doctrinal fact is a fact that requires some level of legal analysis to determine: whether or not there was probable cause for a search, whether or not a contract was formed, etc.

The model assumes the collegial appellate court determines for itself all the doctrinal facts. When the doctrinal facts are questions of law, this is appropriate. When the doctrinal facts are mixed questions, sometimes appellate courts give greater deference to trial courts' findings, and sometimes they review them *de novo*. What kinds of intermediate factual determinations do they subject to greater scrutiny? Extant legal reasoning suggests the answer to this question perhaps should be based on which "judicial actor is better positioned . . . to decide the issue in question.' ' *Miller v. Fenton*, 474 U.S. 104 (1985), at 114. For example, appellate and trial judges are equally capable of examining the text of an unambiguous contract, while those present at trial are in a better position to determine the credibility of witness testimony, having been present to observe their demeanor. In *Ornelas v. United States*, 517 U.S. 690 (1996), quoted in the main text, the U.S. Supreme Court settled a circuit split on whether findings of probable cause should be reviewed *de novo* or with deference (in favor of *de novo* review).

In addition to standard of review choices regarding mixed questions, appellate courts sometimes apply a "constitutional fact doctrine' ' that could allow for independent review even of historical facts (see, e.g., *Ohio Valley Water Co. v. Borough of Ben Avon*, 253 U.S. 287 (1920), allowing *de novo* review of property valuation on appeal from state agency determinations). Some legal scholars have criticized the evolution of this doctrine, worried it will be applied in inappropriate situations (see, e.g., Hoffman 2001; Redish and Gohl 2017). The results here provide a formal theoretic result to support such a concern.

Perhaps most troubling, heightened standards of review tend to be applied in cases regarding civil liberties such as First and Fourth Amendment rights (Hoffman 2001; Redish and Gohl 2017). For example, consider the determination in *Ornelas* to subject probable cause determinations to *de novo* review. Similarly, some circuits hold that whether speech constitutes a "true threat' ' unprotected by the First Amendment is a fact reviewable *de novo*, though other appellate courts disagree (Redish and Gohl 2017, 292). Heightened review of facts is applied in many First Amendment contexts, from cases involving freedom of speech to those implicating religious freedoms (Hoffman 2001, 1453–1455). The legal rules governing arguably our most important freedoms, for which there is often disagreement such that the first sentence of Proposition 1 may not apply, may be the areas in which courts most often apply increase scrutiny on findings of doctrinal facts.

Appendix B

Appendix to Chapter 3

B.1 Derivation of posterior over β in GP regression

In Gaussian process regression, we have the following model specification:

$$y \sim \mathcal{N}(f(X) + X\beta, \sigma_y^2 I), \tag{B.1}$$

$$\beta \sim \mathcal{N}(b, B), \tag{B.2}$$

$$f \sim \mathcal{GP}(0, K(X)). \tag{B.3}$$

Suppose we want the posterior distribution of β. We find

$$p(\beta \mid y, X) \propto p(y \mid \beta)p(\beta) \tag{B.4}$$

$$= N(y; X\beta, K_y) \times N(\beta; b, B), \tag{B.5}$$

where

$$K_y = K(X) + \sigma_y^2 I. \tag{B.6}$$

We can show

$$\beta \mid y \sim \mathcal{N}(\bar{\beta}, \Sigma_\beta), \tag{B.7}$$

$$\bar{\beta} = (B^{-1} + X^T K_y^{-1} X)^{-1}(B^{-1} b + X^T K_y^{-1} y), \tag{B.8}$$

$$\Sigma_\beta = (B^{-1} + X^T K_y^{-1} X)^{-1} \tag{B.9}$$

with a bit of tedious algebra:

$$p(\beta \mid y, X) \propto \exp\left(-\frac{1}{2}\left[(y - X\beta)^T K_y^{-1}(y - X\beta) + (\beta - b)^t B^{-1}(\beta - b)\right]\right) \tag{B.10}$$

$$= \exp\Bigg(-\frac{1}{2}\Bigg[y^T K_y^{-1} y - \beta^T X^T K_y^{-1} y - y^T K_y^{-1} X\beta + \beta^T X^T K_y^{-1} X\beta \tag{B.11}$$

$$+ \beta^T B^{-1}\beta - b^T B^{-1}\beta - \beta^T B^{-1} b + b^T B^{-1} b\Bigg]\Bigg)$$

$$= \exp\Bigg(-\frac{1}{2}\Bigg[\beta^T B^{-1}\beta + \beta^T X^T K_y^{-1} X\beta - \beta^T B^{-1} b - \beta^T X^T K_y^{-1} y \tag{B.12}$$

$$- b^T B^{-1}\beta - y^T K_y^{-1} X\beta\Bigg]$$

$$- \frac{1}{2}\Bigg[y^T K_y^{-1} y + b^T B^{-1} b\Bigg]\Bigg)$$

$$= \exp\Bigg(-\frac{1}{2}\Bigg[\beta^T (B^{-1} + X^T K_y^{-1} X)\beta - \beta^T (B^{-1} b + X^T K_y^{-1} y) \tag{B.13}$$

$$- (b^T B^{-1} + y^T K_y^{-1} X)\beta\Bigg]$$

$$- \frac{1}{2}\Bigg[y^T K_y^{-1} y + b^T B^{-1} b\Bigg]\Bigg)$$

$$= \exp\Bigg(-\frac{1}{2}\Bigg[\left(\beta - (B^{-1} + X^T K_y^{-1} X)^{-1}(B^{-1} b + X^T K_y^{-1} y)\right)^T \tag{B.14}$$

$$
\begin{aligned}
&\times (B^{-1} + X^T K_y^{-1} X) \\
&\times \left(\beta - (B^{-1} + X^T K_y^{-1} X)^{-1}(B^{-1} b + X^T K_y^{-1} y)\right)\Bigg] \\
&- \frac{1}{2}\left[y^T K_y^{-1} y + b^T B^{-1} b\right]\Bigg) \\
\propto \exp&\left(-\frac{1}{2}\left[\left(\beta - \bar{\beta}\right)^T \Sigma_\beta^{-1} (\beta - \bar{\beta})\right]\right), \qquad \text{(B.15)}
\end{aligned}
$$

which is clearly the core of a normal distribution.

B.2 Distribution of derivatives of Gaussian processes

Luckily, "[s]ince differentiation is a linear operator, the derivative of a Gaussian process is another Gaussian process' ' (Rasmussen and Williams 2006, 191).

Let

$$\mathbf{f}_d = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_{1d}} \\ \vdots \\ \dfrac{\partial f_n}{\partial x_{nd}} \end{bmatrix}. \tag{B.16}$$

Using Equation 9.1 in Rasmussen and Williams (2006),

$$K_d \triangleq \mathbb{C}\left[\mathbf{f}_d, \mathbf{f}\right] = \begin{bmatrix} \dfrac{\partial k(\mathbf{x}_1, \mathbf{x}_1)}{\partial x_{1d}} & \cdots & \dfrac{\partial k(\mathbf{x}_1, \mathbf{x}_n)}{\partial x_{1d}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial k(\mathbf{x}_n, \mathbf{x}_1)}{\partial x_{nd}} & \cdots & \dfrac{\partial k(\mathbf{x}_n, \mathbf{x}_n)}{\partial x_{nd}} \end{bmatrix}, \tag{B.17}$$

$$K_{dd} \triangleq \mathbb{C}\left[\mathbf{f}_d, \mathbf{f}_d\right] = \begin{bmatrix} \dfrac{\partial^2 k(\mathbf{x}_1, \mathbf{x}_1)}{\partial x_{1d} \partial x_{1d}} & \cdots & \dfrac{\partial^2 k(\mathbf{x}_1, \mathbf{x}_n)}{\partial x_{1d} \partial x_{nd}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 k(\mathbf{x}_n, \mathbf{x}_1)}{\partial x_{nd} \partial x_{1d}} & \cdots & \dfrac{\partial^2 k(\mathbf{x}_n, \mathbf{x}_n)}{\partial x_{nd} \partial x_{nd}} \end{bmatrix}, \tag{B.18}$$

To make the notation more compact, as is usual we set $K = K(\mathbf{X}, \mathbf{X})$, and additionally set $\mu = \mu$ and

$$\mu_d = \begin{bmatrix} \dfrac{\partial \mu(\mathbf{x}_1)}{\partial x_{1d}} \\ \vdots \\ \dfrac{\partial \mu(\mathbf{x}_n)}{\partial x_{nd}} \end{bmatrix}. \tag{B.19}$$

Then we can describe the joint prior on $\mathbf{f}$ and $\mathbf{f}_d$:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_d \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu_d \end{bmatrix}, \begin{bmatrix} K & K_d^T \\ K_d & K_{dd} \end{bmatrix} \right). \tag{B.20}$$

Then the posterior distribution of $\mathbf{f}_d$ given $\mathbf{X}$ and $\mathbf{y}$ is normal with mean

$$\begin{aligned} \mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] &= \int \mathbb{E}\left[\mathbf{f}_d \mid \mathbf{f}, \mathbf{X}\right] p\left(\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right) d\mathbf{f} && \text{(B.21)} \\ &= \int \mu_d + K_d K^{-1} \left(\mathbf{f} - \mu\right) p\left(\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right) d\mathbf{f} && \text{(B.22)} \\ &= \mu_d + K_d K^{-1} \left(\mathbb{E}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] - \mu\right), && \text{(B.23)} \end{aligned}$$

and variance (using the law of total variance)

$$\begin{aligned} \mathbb{V}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] &= K_{dd} - K_d K^{-1} K_d^T + \mathbb{E}\left[\left(\mathbb{E}\left[\mathbf{f}_d \mid \mathbf{f}\right] - \mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right]\right)^2\right] && \text{(B.24)} \\ &= K_{dd} - K_d K^{-1} K_d^T \\ &\quad + \mathbb{E}\Bigg[\Bigg(\mu_d + K_d K^{-1} \left(\mathbf{f} - \mu\right) \\ &\qquad - \left(\mu_d + K_d K^{-1} \left(\mathbb{E}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] - \mu\right)\right)\Bigg)^2\Bigg] && \text{(B.25)} \\ &= K_{dd} - K_d K^{-1} K_d^T \\ &\quad + \mathbb{E}\left[\left(K_d K^{-1} \left(\mathbf{f} - \mu\right) - K_d K^{-1} \left(\mathbb{E}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] - \mu\right)\right)^2\right] && \text{(B.26)} \\ &= K_{dd} - K_d K^{-1} K_d^T \\ &\quad + \mathbb{E}\Bigg[\Bigg(K_d K^{-1}\mathbf{f} - K_d K^{-1}\mu \end{aligned}$$

$$- K_d K^{-1} \mathbb{E}[\mathbf{f} \mid \mathbf{X}, \mathbf{y}] + K_d K^{-1} \mu \Big)^2 \Big] \tag{B.27}$$

$$= K_{dd} - K_d K^{-1} K_d^T + \mathbb{E}\left[\left(K_d K^{-1} \mathbf{f} - K_d K^{-1} \mathbb{E}[\mathbf{f} \mid \mathbf{X}, \mathbf{y}] \right)^2 \right] \tag{B.28}$$

$$= K_{dd} - K_d K^{-1} K_d^T + \mathbb{E}\left[K_d K^{-1} \Big(\mathbf{f} - \mathbb{E}[\mathbf{f} \mid \mathbf{X}, \mathbf{y}] \Big)^2 K^{-1} K_d^T \right] \tag{B.29}$$

$$= K_{dd} - K_d K^{-1} K_d^T + K_d K^{-1} \mathbb{E}\left[\Big(\mathbf{f} - \mathbb{E}[\mathbf{f} \mid \mathbf{X}, \mathbf{y}] \Big)^2 \right] K^{-1} K_d^T \tag{B.30}$$

$$= K_{dd} - K_d K^{-1} K_d^T + K_d K^{-1} \mathbb{V}[\mathbf{f} \mid \mathbf{X}, \mathbf{y}] K^{-1} K_d^T \tag{B.31}$$

$$= K_{dd} - K_d \left(K^{-1} - K^{-1} \mathbb{V}[\mathbf{f} \mid \mathbf{X}, \mathbf{y}] K^{-1} \right) K_d^T. \tag{B.32}$$

B.2.1 The regression case

In the regression case,

$$\mathbb{E}[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}] = \mu_d + K_d K^{-1} (\mathbb{E}[\mathbf{f} \mid \mathbf{X}, \mathbf{y}] - \mu) \tag{B.33}$$

$$= \mu_d + K_d K^{-1} \left(\mu + K K_y^{-1} (\mathbf{y} - \mu) - \mu \right) \tag{B.34}$$

$$= \mu_d + K_d K^{-1} \left(K K_y^{-1} (\mathbf{y} - \mu) \right) \tag{B.35}$$

$$= \mu_d + K_d K_y^{-1} (\mathbf{y} - \mu), \tag{B.36}$$

where, as usual, $K_y = K + \sigma_y^2 I$. Then

$$\mathbb{V}[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}] = K_{dd} - K_d \left(K^{-1} - K^{-1} \mathbb{V}[\mathbf{f} \mid \mathbf{X}, \mathbf{y}] K^{-1} \right) K_d^T \tag{B.37}$$

$$= K_{dd} - K_d \left(K^{-1} - K^{-1} \left(K - K K_y^{-1} K \right) K^{-1} \right) K_d^T \tag{B.38}$$

$$= K_{dd} - K_d \left(K^{-1} - \left(I - K_y^{-1} K \right) K^{-1} \right) K_d^T \tag{B.39}$$

$$= K_{dd} - K_d \left(K^{-1} - \left(K^{-1} - K_y^{-1}\right)\right) K_d^T \tag{B.40}$$

$$= K_{dd} - K_d \, K_y^{-1} \, K_d^T. \tag{B.41}$$

Note the similarity to the predictive distribution of $\mathbf{f}_*$ for new cases $\mathbf{X}_*$.

B.2.2 The classification case

In the classification case, under the Laplace approximation to the posterior,

$$\mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = \mu_d + K_d \, K^{-1} \left(\mathbb{E}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] - \mu\right) \tag{B.42}$$

$$= \mu_d + K_d \, K^{-1} \left(\mu + K \left(\nabla \log p \left(\mathbf{y} \mid \hat{\mathbf{f}}\right)\right) - \mu\right) \tag{B.43}$$

$$= \mu_d + K_d K^{-1} \left(K \left(\nabla \log p \left(\mathbf{y} \mid \hat{\mathbf{f}}\right)\right)\right) \tag{B.44}$$

$$= \mu_d + K_d \left(\nabla \log p \left(\mathbf{y} \mid \hat{\mathbf{f}}\right)\right), \tag{B.45}$$

$$\mathbb{V}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = K_{dd} - K_d \left(K^{-1} - K^{-1} \, \mathbb{V}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] K^{-1}\right) K_d^T \tag{B.46}$$

$$= K_{dd} - K_d \left(K^{-1} - K^{-1} \left(K^{-1} + W\right)^{-1} K^{-1}\right) K_d^T \tag{B.47}$$

$$= K_{dd} - K_d \left(K + W^{-1}\right)^{-1} K_d^T, \quad \text{(by the matrix inversion lemma)} \tag{B.48}$$

$$W = -\nabla\nabla \log p \left(\mathbf{y} \mid \hat{\mathbf{f}}\right). \tag{B.49}$$

B.3 Distribution of average marginal effects

Since then γ_d is a constant ($1/N$) times the sum of correlated normal random variables,

$$\gamma_d \sim \mathcal{N}\left(\frac{1}{N}\sum_{i=1}^{N} m_{\gamma_d i}\,,\ \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} c_{\gamma_d ij}\right), \tag{B.50}$$

$$\mathbf{m}_{\gamma_d} = \mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] \tag{B.51}$$

$$\mathbf{C}_{\gamma_d} = \mathbb{V}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] \tag{B.52}$$

Importantly, we may also get the average marginal effect of feature d within subgroups of $\mathbf{X}$ rather than the full sample average by simply altering the indices of summation in Equation~B.50.

However, for binary classification, we may be more interested in the distribution of

$$\pi_d \triangleq \frac{1}{N}\sum_{i}^{N} \frac{\partial\, \sigma\left(f\left(\mathbf{x}_i\right)\right)}{\partial\, x_{id}} \tag{B.53}$$

than γ_d. Unfortunately, that distribution cannot be analytically expressed, though we can readily simulate from it. First note that

$$\frac{\partial\, \sigma\left(f\left(\mathbf{x}_i\right)\right)}{\partial\, x_{id}} = \frac{\partial\, \sigma\left(f\left(\mathbf{x}_i\right)\right)}{\partial\, f}\frac{\partial\, f\left(\mathbf{x}_i\right)}{\partial\, x_{id}}. \tag{B.54}$$

Generally the sigmoid function σ has a known derivative; for example, in the logistic case,

$$\frac{\partial\, \sigma\left(f\left(\mathbf{x}_i\right)\right)}{\partial\, f} = \sigma\left(f\left(\mathbf{x}_i\right)\right)\left(1 - \sigma\left(f\left(\mathbf{x}_i\right)\right)\right). \tag{B.55}$$

Since we have an approximation to the posterior on f, and given f, the posterior over $\mathbf{f}_d$ is

$$\mathbf{f}_d \mid \mathbf{f} \sim \mathcal{N}\left(\mu_d + K_d\, K^{-1}\left(\mathbf{f} - \mu\right), K_{dd} - K_d\, K^{-1}\, K_d^T\right), \tag{B.56}$$

we can obtain M samples of π_d by

- drawing $\mathbf{f}^t$ from $\mathcal{N}\left(\mu + K\left(\nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right)\right), \left(K^{-1} + W\right)^{-1}\right)$,
- drawing $\mathbf{f}_d^t$ from $\mathcal{N}\left(\mu_d + K_d\, K^{-1}\left(\mathbf{f}^t - \mu\right), K_{dd} - K_d\, K^{-1}\, K_d^T\right)$,
- and calculating $\pi_d^t = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \frac{\partial\, \sigma\left(f_i^t\right)}{\partial\, f} f_{id}^t$,

so that we can summarize the distribution of π_d using the M draws, similar to the CLARIFY procedure (King, Tomz, and Wittenberg 2000).

Moreover, another issue to consider is discrete variables in $\mathbf{X}$. For binary variables, let $\mathbf{X}_1^*$ be a set of test points where all feature observactions are identical to $\mathbf{X}$ except that all observations of feature d have been set to 1, and analogously for $\mathbf{X}_0^*$.% [1] Then a more appropriate quantity of interest rather than γ_d is

$$\delta_d = \frac{1}{N}\sum_{i=1}^{N} f\left(\mathbf{x}_{1i}^*\right) - f\left(\mathbf{x}_{0i}^*\right), \tag{B.57}$$

which gives the average marginal effect on y of taking a 1 vs a 0 value in the regression case, or the average partial effect on f of taking a 1 vs a 0 value in the classification case. Letting $\mathbf{f}_1 = f\left(\mathbf{X}_1^*\right)$ and analogously for $\mathbf{f}_0$, δ_d is distributed

$$\delta_d \sim \mathcal{N}\left(\frac{1}{N}\sum_{i=1}^{N} m_{\delta_d i}\,,\ \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} c_{\delta_d ij}\right), \tag{B.58}$$

$$\mathbf{m}_{\delta_d} = \mathbb{E}\left[\mathbf{f}_1^* \mid \mathbf{X}, \mathbf{y}\right] - \mathbb{E}\left[\mathbf{f}_0^* \mid \mathbf{X}, \mathbf{y}\right], \tag{B.59}$$

$$\mathbf{C}_{\delta_d} = \mathbb{V}\left[\mathbf{f}_1^* \mid \mathbf{X}, \mathbf{y}\right] + \mathbb{V}\left[\mathbf{f}_0^* \mid \mathbf{X}, \mathbf{y}\right] + \mathbb{C}\left[\mathbf{f}_1^*, \mathbf{f}_0^* \mid \mathbf{X}, \mathbf{y}\right] + \mathbb{C}\left[\mathbf{f}_0^*, \mathbf{f}_1^* \mid \mathbf{X}, \mathbf{y}\right]. \tag{B.60}$$

For classification we can also simulate the distribution of

1. You can replace 1 and 0 with other binary value labels as needed.

$$\psi_d = \frac{1}{N}\sum_{i=1}^{N} \sigma\left(f\left(\mathbf{x}_{1i}^*\right)\right) - \sigma\left(f\left(\mathbf{x}_{0i}^*\right)\right) \tag{B.61}$$

by simply generating values of $f\left(\mathbf{X}_1^*\right)$ and $f\left(\mathbf{X}_0^*\right)$ from the posterior approximation, pushing those samples through the chosen sigmoid σ, and calculate the average of the differences to get the average marginal effect for each sample so that we can summarize the distribution of average marginal effects, similar to the continuous case. In the case of a categorical variable that has been one-hot encoded into $\mathbf{X}$, we can simply follow the above procedures for all the substantively interesting pairwise comparisons between category labels. Moreover, in some cases using this procedure to find the distribution of difference in MAP predictions at two discrete values of a continuous variable may be more readily interpretable than π_d.

B.4 Derivatives of mean and covariance functions

Regarding the derivatives of the mean function,

$$\mu(X) = 0 \Rightarrow \frac{\partial \mu(\mathbf{x}_i)}{\partial \mathbf{x}_{id}} = 0, \tag{B.62}$$

$$\mu(X) = X\beta \Rightarrow \frac{\partial \mu(\mathbf{x}_i)}{\partial \mathbf{x}_{id}} = \beta_d, \tag{B.63}$$

to cover a couple of popular choices.

Note that calculating the mean and variance of the distribution of γ_d requires a twice-differentiable covariance function. For the squared exponential covariance function with automatic relevance determination,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_d \left[\frac{(x_{id} - x_{jd})^2}{\ell_d^2}\right]\right). \tag{B.64}$$

Then,

$$\frac{\partial k(\mathbf{x}_i, \mathbf{x}_j)}{\partial \mathbf{x}_{id}} = k(\mathbf{x}_i, \mathbf{x}_j)\frac{x_{jd} - x_{id}}{\ell_d^2} \tag{B.65}$$

$$\frac{\partial^2 k(\mathbf{x}_i, \mathbf{x}_j)}{\partial \mathbf{x}_{id}\partial \mathbf{x}_{jd}} = k(\mathbf{x}_i, \mathbf{x}_j)\left(\frac{1}{\ell_d^2} + \frac{(x_{jd} - x_{id})(x_{id} - x_{jd})}{\ell_d^4}\right), \tag{B.66}$$

Note that when $i = j$, the cross partial simplifies to σ_f^2/ℓ_d^2.

Appendix C

Appendix to Chapter 4

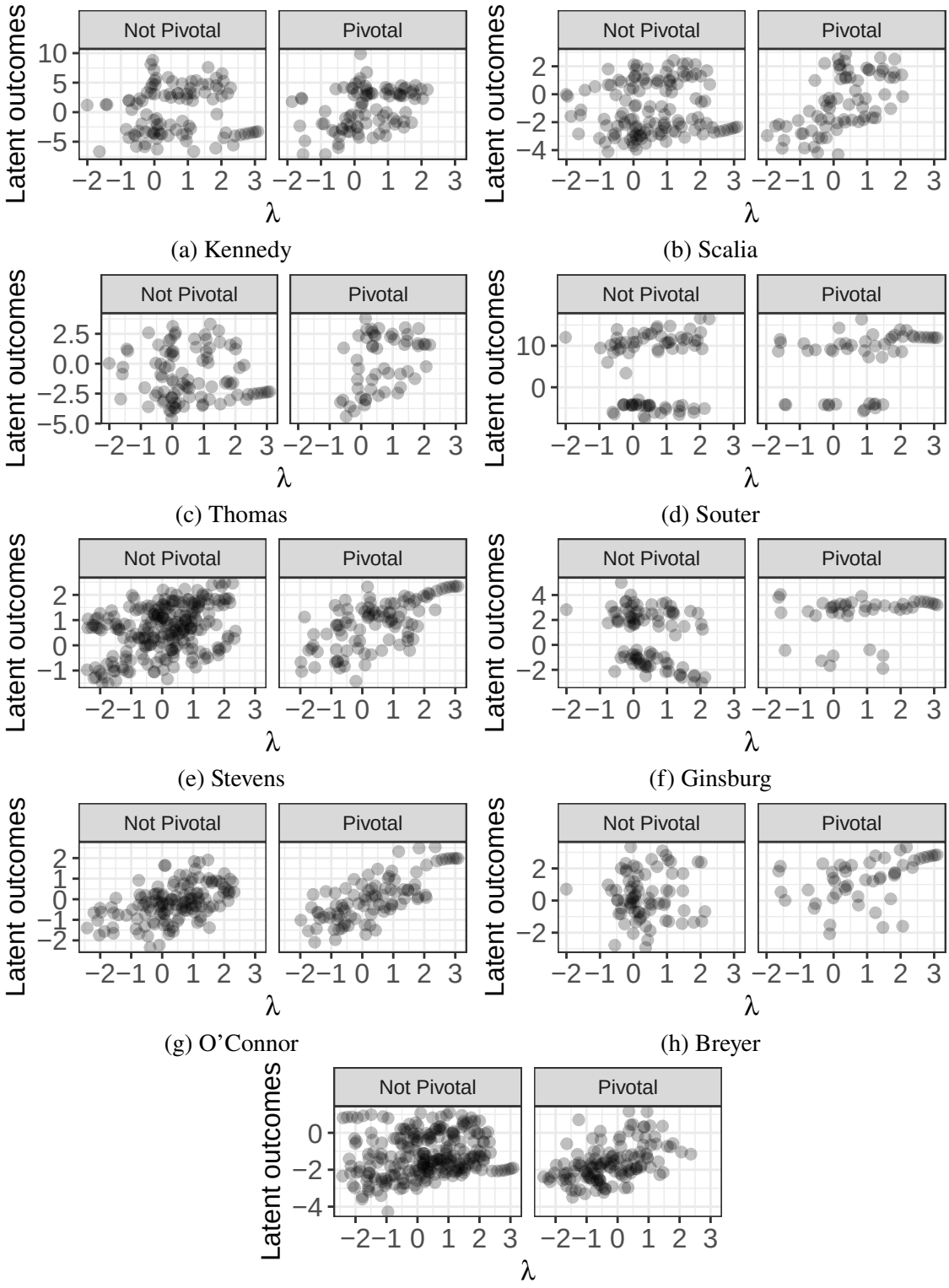
C.1 Full results of main model

The main text of the paper presents results of the average marginal effect of λ, or the law. For those interested in the average marginal effect of other predictors, I provide the average marginal effect of every variable in the models in Table C.1. The average marginal effect here is on the link, or latent, scale rather than on the probability scale; that is, it provides $\partial x/\partial f$ rather than $\partial x/\partial \sigma(f)$. Of note is that the results for λ are consonant with the probability scale results in the main paper, with the exception that Justice Souter has a reliable positive average marginal effect of λ on the latent scale even though the result is not reliable at a 95% level on the probability scale. Looking at the marginal effects of other predictors, we see that most justices are more likely to vote liberally in cases with content-based restrictions, which we may interpret as in line with the prior literature (see Bartels 2009; Bartels and O'Geen 2015; Richards and Kritzer 2002). Other case factors appear to have more nuanced effects, and may be highly conditional on the values of other variables. A reader may also want to see breakdowns like Figure 4.4 between cases where a justice is pivotal or not pivotal for justices other than Ginsburg; I provide such figures for each justice analyzed in the main paper in Figure C.1.

Table C.1: Average marginal effects for all predictors in the justice-level models.

	Breyer	Ginsburg	Kennedy	O'Connor	Rehnquist	Scalia	Souter	Stevens	Thomas
λ	0.2	−0.5	0.8	0.2	0.2	0.3	1.1	0.2	0.1
	[0.0, 0.4]	[−0.9, −0.2]	[0.6, 0.9]	[0.1, 0.3]	[0.0, 0.3]	[0.1, 0.5]	[0.1, 2.0]	[0.0, 0.3]	[−0.1, 0.3]
ρ	0.4	0.2	−0.6	0.0	−0.1	−0.4	−1.1	0.1	−0.2
	[0.2, 0.6]	[−0.3, 0.7]	[−0.8, −0.4]	[−0.1, 0.1]	[−0.2, 0.0]	[−0.5, −0.2]	[−2.5, 0.2]	[0.0, 0.2]	[−0.4, 0.0]
Term	0.0	−0.1	0.0	0.0	0.0	0.0	−0.1	0.0	0.0
	[−0.3, 0.3]	[−0.7, 0.5]	[−2.4, 2.3]	[−0.2, 0.2]	[−0.2, 0.2]	[−0.5, 0.5]	[−22.8, 22.7]	[−0.2, 0.2]	[−0.9, 0.9]
MQ Score	0.1	0.2	−2.8	−0.5	0.0	−0.1	0.4	0.1	0.1
	[−0.1, 0.4]	[−0.6, 1.0]	[−4.2, −1.3]	[−0.6, −0.3]	[−0.2, 0.2]	[−0.7, 0.4]	[−8.3, 9.2]	[−0.1, 0.3]	[−0.6, 0.8]
Category; reference 'Less Protected';'Content-based' restriction abbreviated 'CB''Content-neutral' abbreviated 'CN'									
CB	1.5	1.7	4.1	0.8	0.9	1.3	−1.1	−0.1	1.9
	[1.1, 1.9]	[0.7, 2.7]	[2.6, 5.6]	[0.5, 1.0]	[0.5, 1.2]	[0.6, 2.0]	[−14.2, 12.1]	[−0.4, 0.2]	[1.1, 2.7]
CN	1.1	1.1	−0.8	0.1	−0.2	−0.6	−1.0	−0.2	0.1
	[0.7, 1.6]	[0.0, 2.2]	[−2.6, 0.9]	[−0.2, 0.3]	[−0.5, 0.2]	[−1.4, 0.3]	[−14.4, 12.3]	[−0.5, 0.1]	[−0.8, 1.0]
Government; reference 'Other'									
Education	−0.1	−0.1	−2.2	0.0	0.5	−0.3	1.2	0.0	−0.3
	[−0.6, 0.3]	[−1.3, 1.2]	[−4.1, −0.3]	[−0.3, 0.2]	[0.1, 0.9]	[−1.1, 0.6]	[−14.3, 16.6]	[−0.3, 0.3]	[−1.3, 0.8]
Federal	−0.1	−0.5	−1.2	0.8	0.8	1.1	−2.8	−0.5	2.2
	[−0.5, 0.4]	[−1.6, 0.7]	[−2.9, 0.4]	[0.5, 1.0]	[0.4, 1.1]	[0.3, 1.9]	[−17.4, 11.8]	[−0.8, −0.2]	[1.2, 3.1]
Local	0.0	−0.2	−0.6	0.3	−0.4	−0.4	−0.6	0.1	−0.6
	[−0.5, 0.5]	[−1.4, 1.1]	[−2.3, 1.2]	[0.0, 0.6]	[−0.8, 0.0]	[−1.3, 0.5]	[−16.8, 15.6]	[−0.2, 0.4]	[−1.6, 0.4]
Private	0.1	0.0	−1.8	−0.1	−0.2	0.3	−1.6	−0.3	0.0
	[−0.4, 0.6]	[−1.2, 1.3]	[−3.6, 0.1]	[−0.4, 0.2]	[−0.6, 0.2]	[−0.6, 1.2]	[−17.6, 14.5]	[−0.7, 0.0]	[−1.0, 1.1]
State	−0.6	−0.8	1.4	0.3	0.3	0.7	−4.8	−0.1	1.3
	[−1.1, −0.1]	[−2.0, 0.4]	[−0.3, 3.1]	[0.0, 0.6]	[−0.1, 0.6]	[−0.2, 1.5]	[−19.3, 9.6]	[−0.4, 0.2]	[0.3, 2.2]
Action; reference 'Civil suit'									
Criminal	0.1	−0.5	1.2	0.2	0.0	0.5	−0.4	1.1	0.9
	[−0.3, 0.6]	[−1.9, 0.8]	[−0.7, 3.1]	[−0.1, 0.4]	[−0.4, 0.4]	[−0.4, 1.3]	[−14.5, 13.6]	[0.7, 1.4]	[−0.1, 1.9]
Deny benefit	−0.3	−0.2	−1.1	−0.5	−0.7	−1.8	−4.7	0.1	0.0
	[−0.8, 0.1]	[−1.6, 1.2]	[−3.1, 1.0]	[−0.7, −0.2]	[−1.1, −0.4]	[−2.8, −0.9]	[−19.7, 10.3]	[−0.2, 0.4]	[−1.1, 1.0]
Censor	−1.2	−1.6	1.7	0.2	0.2	0.5	−5.6	0.5	2.3
	[−1.6, −0.7]	[−2.8, −0.4]	[−0.2, 3.5]	[0.0, 0.5]	[−0.1, 0.6]	[−0.3, 1.3]	[−19.7, 8.6]	[0.2, 0.8]	[1.4, 3.2]
Disciplinary	−0.7	−1.0	0.8	−0.1	−0.3	−0.5	−4.0	0.9	1.3
	[−1.1, −0.2]	[−2.4, 0.4]	[−1.3, 2.8]	[−0.4, 0.1]	[−0.7, 0.1]	[−1.5, 0.4]	[−19.0, 11.1]	[0.6, 1.2]	[0.3, 2.3]
Lose job	−0.5	−1.1	−1.2	−0.6	−0.4	−1.3	−2.0	0.8	−0.4
	[−0.9, 0.0]	[−2.6, 0.3]	[−3.3, 0.9]	[−0.9, −0.3]	[−0.8, −0.1]	[−2.2, −0.3]	[−16.5, 12.6]	[0.4, 1.1]	[−1.4, 0.6]
Regulation	−0.6	−1.1	−0.7	−1.0	0.0	−0.9	−1.6	0.3	0.3
	[−1.1, −0.2]	[−2.4, 0.3]	[−2.7, 1.3]	[−1.2, −0.7]	[−0.4, 0.3]	[−1.9, 0.0]	[−15.6, 12.4]	[0.0, 0.6]	[−0.7, 1.3]
Speaker identity; reference 'Other'									
Communist				0.6	0.7			−0.1	
				[0.3, 0.9]	[0.3, 1.0]			[−0.4, 0.2]	
Broadcast	−0.3	0.4	−3.2	1.0	1.1	0.7	−1.3	−0.1	0.7
	[−0.8, 0.1]	[−0.7, 1.5]	[−4.9, −1.4]	[0.7, 1.3]	[0.7, 1.4]	[−0.1, 1.5]	[−17.3, 14.7]	[−0.4, 0.2]	[−0.2, 1.6]
Business	−1.2	−0.4	2.9	0.8	0.9	2.0	0.9	−0.2	2.3
	[−1.7, −0.8]	[−1.5, 0.6]	[1.4, 4.5]	[0.5, 1.0]	[0.6, 1.3]	[1.2, 2.7]	[−12.0, 13.8]	[−0.5, 0.1]	[1.5, 3.2]
Protester		0.6	−0.8	0.7	0.5	0.9	2.7	0.1	1.1
		[−0.5, 1.7]	[−2.5, 1.0]	[0.4, 1.0]	[0.1, 0.8]	[0.1, 1.7]	[−12.7, 18.0]	[−0.2, 0.4]	[0.2, 2.0]
Politician	−0.3	0.7	−3.7	1.0	0.8	−0.1	4.0	0.0	−1.0
	[−0.7, 0.2]	[−0.3, 1.8]	[−5.3, −2.0]	[0.8, 1.3]	[0.4, 1.1]	[−0.9, 0.7]	[−10.9, 18.8]	[−0.3, 0.3]	[−1.9, −0.1]
Print			−2.3	0.7	0.3	0.5	1.0	−0.1	
			[−4.1, −0.6]	[0.5, 1.0]	[−0.1, 0.7]	[−0.3, 1.3]	[−14.8, 16.8]	[−0.4, 0.2]	
Religious	−0.3	1.2	−1.5	1.0	0.7	0.9	6.2	0.4	0.5
	[−0.8, 0.1]	[0.0, 2.3]	[−3.2, 0.2]	[0.7, 1.2]	[0.3, 1.1]	[0.1, 1.7]	[−7.9, 20.2]	[0.1, 0.7]	[−0.4, 1.4]

Note: For each justice, I report the estimate and 95% confidence interval for the average marginal effect on the link scale. For the continuous variables 'MQ Score', 'Term', 'λ', and 'ρ', this is the instantaneous rate of change. For the categorical variables 'Category', 'Government', 'Action', and 'Speaker Identity', this is a discrete difference against a reference category (listed in the table).



(a) Kennedy

(b) Scalia

(c) Thomas

(d) Souter

(e) Stevens

(f) Ginsburg

(g) O'Connor

(h) Breyer

(i) Rehnquist

Figure C.1: Comparing predicted outcomes for each justice against λ.

C.2 Marginal effect of law for additional justices

In the main paper I focus on results for the last natural court of the Rehnquist court to ensure that

1. a sufficient number of observations are used as purely training data for the court-level model to ensure the quality of the λ measure used in the justice-level models;
2. we are using every case decided by each justice we're analyzing; and
3. we have a sufficient number of observations for each justice we're analyzing to ensure the quality of estimates.

In this appendix I provide supplemental results for additional justices. However, I do not analyze the Trump appointees, who each have observations in the single digits in this data. Additionally, even the Bush and Obama appointees all have less than 50 observations, so we may want to see more data from them as well (all justices whose results are presented in the main paper have over 100 observations), since GP classification can be a somewhat data-hungry procedure (see Duck-Mayr, Garnett, and Montgomery 2020). The number of observations in the data for each justice is given in Table C.2.

I present the average marginal effect of λ for the justices who decided any cases outside of the training data that were not presented in the main paper (except for the Trump appointees) in Table C.3. These results are substantively similar to the results for the natural court studied in the main paper, with a majority of justices exhibiting a reliably positive average marginal effect of law, while others are more or less unconstrained.

Table C.2: Number of complete observations for each justice.

Justice	N
Kavanaugh	6
Gorsuch	8
Kagan	25
Sotomayor	31
Alito	46
Roberts	49
Douglas	67
Breyer	125
Souter	126
Ginsburg	130
Thomas	143
Stewart	156
Kennedy	185
Scalia	198
Powell	218
Burger	219
O'Connor	225
Brennan	262
Marshall	279
Blackmun	302
White	302
Stevens	327
Rehnquist	375

Table C.3: Average marginal effect of λ on judges' decisions.

	Mean of $\lambda \pm$ sd of λ	Range of λ
Alito	-0.03	-0.08
	$[-0.15, 0.07]$	$[-0.21, 0.04]$
Blackmun	0.12	0.29
	$[0.10, 0.15]$	$[0.26, 0.31]$
Brennan	0.02	0.04
	$[-0.08, 0.12]$	$[-0.06, 0.15]$
Burger	0.11	0.27
	$[0.09, 0.14]$	$[0.25, 0.29]$
Douglas	0.11	0.25
	$[-0.05, 0.29]$	$[0.08, 0.44]$
Kagan	-0.50	-0.94
	$[-0.55, -0.45]$	$[-0.95, -0.92]$
Marshall	0.02	0.05
	$[-0.10, 0.13]$	$[-0.07, 0.16]$
Powell	0.26	0.58
	$[0.23, 0.30]$	$[0.55, 0.61]$
Roberts	-0.28	-0.57
	$[-0.29, -0.26]$	$[-0.59, -0.56]$
Sotomayor	0.27	0.57
	$[0.25, 0.29]$	$[0.55, 0.58]$
Stewart	0.11	0.26
	$[0.07, 0.16]$	$[0.21, 0.30]$
White	0.13	0.30
	$[0.08, 0.18]$	$[0.25, 0.35]$

Note: For each justice, I report the estimate and 95% confidence interval for the difference in the probability the justice will vote liberally between two different values of λ, averaged over all observations in the sample.

C.3 Robustness check: In-group bias

Epstein, Parker, and Segal (2018) find that justices may offer greater protection to speech they agree with. We may worry this feeds into ϕ and thus may bias the γ estimate unless we control for whether the speech at issue is liberal or conservative. I re-run the analysis with a variable coding whether the speech at issue in a case is liberal (for example, obscenity), conservative (for example, commercial or religious speech), or neutral (such as where, for example, campaign spending of both Republicans and Democrats is implicated), as well as other variables from Epstein, Parker, and Segal (2018) such as whether the law at issue is conservative (e.g. anti-obscenity laws), liberal (e.g. a law criminalizing depiction of animal cruelty), or neutral, whether the challenge is an "as applied" or "facial" challenge, and the type of expression (spoken, written, other expression, or association) at issue. As the Epstein, Parker, and Segal (2018) data covers the 1953-2014 terms, I update and backdate that data, as well as filling in some missingness from cases that were included in the Richards and Kritzer (2002) data but not in the Epstein, Parker, and Segal (2018) data.

The results are largely substantively similar. The main differences to note are that our estimate of the average marginal effect of law for Kennedy is now essentially zero and not reliably positive or negative, and the result for Rehnquist is weakened. However, a majority of the Court still exhibits a reliable effect of the law on their decision making.

Table C.4: Average marginal effect of λ on judges' decisions.

	Mean of $\lambda \pm$ sd of λ	Range of λ
Breyer	0.04	0.10
	$[-0.01, 0.10]$	$[0.05, 0.16]$
Ginsburg	-0.15	-0.34
	$[-0.22, -0.09]$	$[-0.41, -0.27]$
Kennedy	0.00	0.00
	$[-0.10, 0.11]$	$[-0.11, 0.12]$
O'Connor	0.05	0.12
	$[0.01, 0.08]$	$[0.08, 0.16]$
Rehnquist	0.02	0.02
	$[-0.03, 0.06]$	$[-0.02, 0.07]$
Scalia	0.04	0.10
	$[-0.03, 0.11]$	$[0.01, 0.18]$
Souter	0.12	0.28
	$[0.04, 0.20]$	$[0.19, 0.37]$
Stevens	0.05	0.12
	$[0.00, 0.10]$	$[0.07, 0.18]$
Thomas	-0.08	-0.18
	$[-0.16, 0.01]$	$[-0.27, -0.09]$

Note: For each justice, I report the estimate and 95% confidence interval for the difference in the probability the justice will vote liberally between two different values of λ, averaged over all observations in the sample.