Competing Ideologies: Fit, Simplicity, and Fear*

José Luis Montiel Olea (Cornell)
Andrea Prat (Columbia)

March 1, 2025

Abstract

We consider a political contest where data is common knowledge and candidates compete by announcing ideologies (statistical descriptions of the data-generating process). The election winner will make policy decisions according to the announced ideology. Voters form a prior on the data-generating process on the basis of the announced ideologies. They choose the candidate who maximizes their expected payoff. We prove that the ideology selected by candidates in equilibrium maximizes the product of three factors that can be interpreted as: fit (likelihood of the observed data within the ideology), simplicity (a penalty on the ability of the ideology to explain a lot of data realizations), and fear (the ability of the ideology to evoke bad outcomes of competing ideologies).

1 Introduction

Canonical models of electoral politics assume that parties compete over policy positions (Persson and Tabellini, 2002). The policy space is a set of alternatives over which voters have preferences. For instance, in Downsian competition, each of two parties select a point in the policy space and voters vote for the party whose policy choice would given them a higher expected utility.

In this account of electoral politics, voters come exogenously equipped with models of the world that help them make sense of the data they observe and the policy choices that are announced. However, a recent literature has emphasized the importance of understanding the origin of such

*We thank Chiara Aina, Nageeb Ali, Steve Callander, Ernesto Dal Bó, Laura Doval, Kfir Eliaz, Patrick Francois, Drew Fudenberg, John Glover, Takuma Habu, Navin Kartik, Elliot Lipnowski, Greg Martin, Pietro Ortoleva, Gerard Padró, Jacopo Perego, Rani Spiegler, Steve Yeh, and audiences at Columbia University, ESSET Gerzensee and the UBC Berkeley Political Economy Conference for useful comments.

models of the world. Shiller (2019) provides powerful examples of how "narratives" – sometimes incorrect narratives – diffuse and how they shape individual choices and policy decisions. Stantcheva (2021) shows that the partisan gap in US voters' views on tax policy is explained by differences in how voters reason about the issues. On the theory front, Eliaz, Galperti, and Spiegler (forthcoming) propose a formalization of political competition in narratives based on directed causal graphs. Izzo, Martin, and Callander (2023) consider electoral competition over *ideologies*, systems of beliefs about the underlying process that generates social outcomes.

Competition among 'model providers' tends to push toward models that fit the existing data as well as possible (Schwartzstein and Sunderam (2021)). However, in politics winning ideologies can also display other features. One example is fear ideology. Herman and Chomsky's (1988) propaganda model identifies fear as as a powerful ideological tool to build consent around political programs. Glassner (2018) documents how a culture of fear is built by evoking perceived dangers that are much larger than actual risks. Glassner describes Donald Trump's 2016 campaign as particularly effective in amplifying voters' concerns about crime and immigration and weaponizing them against the Democratic party.

Another reason for not selecting ideologies purely on the basis of fit is a desire to keep things simple. One of the tenets of political communication – sometimes referred to as KISS (Keep it simple, stupid) – is that complex narratives should be avoided (Lakoff (2004)), not just because they tax the receivers' cognitive capabilities but also because they generate mistrust in voters (Luntz (2007)). For instance, available evidence indicates that crime is a complex phenomenon with multiple causes, and there is limited consensus among experts on how to reduce it (Chalfin and McCrary (2017)). Only a rich multi-factor family of models has some hope of containing the true data-generating model. Yet, routinely politicians win elections on the basis of extremely simple one-factor causal models of crime, like Donald Trump's 'migrant crime' narrative.

The goal of the present paper is to offer a possible account of these three forces – fit, fear, and complexity – within a parsimonious model of competition between political ideologies.

In a nutshell, the paper considers a situation with two office-seeking candidates and a homogeneous electorate. Voters care about policy outcomes that depend on chosen policy and a potentially complex state of the world. They have access to data but do not have a model to make sense of it. Candidates announce statistical models of the world that explain how data and outcomes are related. In line with Izzo et al. (2023), we refer to such models as "ideologies". If elected, candidates will make policy decisions according to the announced ideology. However, ideology announcements have another effect: they provide voters with tools to make sense of the data. The paper provides a characterization of the features of ideologies that tend to emerge in equilibrium.

The game we consider is as follows. Before the election is held, voters and candidates observe past realizations of a potentially large vector of random variables. While voters have the ability to observe the data, they are unable to make sense of it, perhaps because they lack the analytical skills to imagine a statistical model for the observed data and to define a prior distribution over the model's parameters.

Candidates provide voters with ideologies to make sense of the data. There is a set of possible ideologies. An ideology is a statistical model (a collection of probability distributions) for the observed data combined with a prior distribution over the model's parameters. Each candidate selects one of the ideologies in the set and announces it.[1]

For instance, we can imagine a situation where policy must fit a state of the world (say fiscal policy must respond to aggregate demand), which in turn can potentially depend in a linear way on a large number of covariates. In this setting an ideology is determined by a subset of the set of potential covariates and a prior distribution over their magnitude (maintaining, for example, a Gaussian linear regression model). The ideology set is then defined over all possible subsets of

[1]Our theory is silent with respect to the set of possible ideologies that candidates can announce. In practice, the set is likely to depend on the cognitive skills of voters (the set of ideologies they can understand) and their cultural background (the set of ideologies that seem plausible to them). The ideology set need not contain the true data generating process, and in this sense, the politician is allowed to persuade voters using any ideology including 'alternative realities' as in Szeidl and Szűcs (2022). Our results hold, with some technical conditions, for any ideology set.

covariates. Each candidate announces the subset of covariates that, in her view, affect the outcome. The effect of each covariate on the outcome of interest is parameterized by a linear coefficient that will then be estimated on the basis of the data.

Let's continue the description of the set-up. After the candidates announce their ideologies, voters can then use the two ideologies to come up with their own statistical model for the available data. We assume that voters assign a priori equal probability to the two announced ideologies. From this point onward, voters have a prior distribution – the mixture of the two ideologies – and they are standard Bayesian decision-makers. They combine their prior with the data to compute posterior distributions over the probability that each ideology is correct and over the distribution of parameters within each ideology.

After the election, the winning candidate will have to select an action. The voter's utility depends on the action and the unknown distribution of the data.

In some examples, the data is divided into covariates, which are observed after the election but before the action is taken, and outcomes, which are observed only after the action is taken. For instance, in the linear regression example above, the winning candidates will observe new covariates, and use this information along with the model and the historical data to choose an action.

Voters select the candidate that offers the higher expected payoff given the announced ideologies and the observed data. Each candidate realizes how voters make electoral decisions and selects an ideology to maximize the probability of winning the election. Candidates do not care about outcomes, but – as mentioned above – they are bound to make decisions on the basis of the ideology they announced.

Given an observable dataset and a set of possible ideologies, what ideologies will be selected in equilibrium? In the linear regression example, which subset of covariates will candidates focus on?

Our first result (Theorem 1) is a necessary and sufficient condition for an ideology to be selected in equilibrium. Compared to all other possible ideologies, the selected one must maximize the product of three factors. (In the paper's title, we suggest three short 'nicknames' for the factors:

fit, simplicity, and fear. These terms should not be taken as accurate descriptions of the underlying mathematical concepts but rather as mnemonic references to those concepts.)

The first term can be interpreted as the ability of the ideology to fit the data. It is a maximized likelihood, namely the value of the probability of the observed data within the ideology maximized over the parameter set. The intuition for the first term resonates with the previous literature in this area: an ideology that fits the data better is, everything else equal, more attractive for the voter.

The second term may be interpreted as the simplicity of the ideology. It corresponds to the ratio between the ex ante probability of observing the data given the ideology and the ex post probability of observing it after maximizing over the parameter set. The maximal value of this ratio is one, and it is attained when the ideology has no free parameters. The minimal value is (asymptotically) zero, and it is attained with a 'theory of everything' where, given any data realization, we can always find parameter values that predicted the realization with certainty. An ideology with more parameters is better at explaining observed patterns but runs a risk of overfitting the data. Voters worry that policy made according to a complex ideology is suboptimal because it is affected by spurious patterns in the data. This term is related to the Bayesian Occam's Razor Principle (Murphy (2012)). A complex model that "can predict many things" engenders skepticism among voters.

Section 6 explores the implications of this simplicity bonus (or complexity discount) within the linear regression example discussed above. We will see how the winning ideology is not the one that fits the data best but one that offers an optimal compromise between fit and simplicity.

We will also show that the terms of the tradeoff can change depending on the number of observations. Small datasets tend to favor less complex ideologies. What happens with richer datasets crucially depends on whether the model is correctly or incorrectly specified. Larger datasets will tend to favor the model that is closest to the true data generating process. If at least one of the models in the ideology set is correctly specified, smaller or larger models (relative to the true data generating process) are unlikely to be chosen in equilibrium. If none of the models in the set of ideologies is correctly specified, and if larger models approximate better the true data generating

process, then more complex ideologies are more likely to be selected in equilibrium.

The third term measures the ability of the ideology to evoke potentially bad outcomes in competitive ideologies. It is captured by the difference between the expected payoff given the optimal action that would be chosen with the ideology under consideration and the expected payoff given the optimal action that would be chosen with any other ideology – both expectations being computed within the ideology under consideration. We will point out a connection between the fear factor and the argument behind Pascal's Wager (Pascal (1670)).

Section 5 illustrates how the fear term can lead to a situation where ideologies that provide 'reasonable' descriptions of the data-generating process are beaten by ideologies that fit the data worse but have the ability to exploit the fear factor. This point is made in a scenario with two possible ideologies: a *moderate* ideology that assigns a positive probability to events that have been observed in the past – and only to those – and an *alarmist* ideology, which assigns probability $1-p$ to the moderate ideology being correct and p to an extreme event that has never been observed. Proposition 4 shows that for any observed dataset, if voters are sufficiently risk averse and p is sufficiently low, the alarmist ideology wins. Namely, if one candidate announces the moderate ideology and the other announces the alarmist ideology, voters elect the latter. The intuition is that the alarmist ideology fits the data worse than the moderate ideology but it evokes a very costly event and then it provides some insurance against it. When p is small, the second effect dominates.

The paper also deals with existence and uniqueness of equilibrium. We begin by showing that an equilibrium – a pure strategy equilibrium where each candidate announces an ideology – may not exist. We do so in a simple example with three ideologies. The first ideology leads to choosing an action that performs well if the first ideology is correct, relatively well if the second ideology is correct, and poorly if the third ideology is correct; the second ideology leads to choosing an action that performs well if the second ideology is correct, relatively well if the third ideology is correct, and poorly if the first ideology is correct; and similarly for the third ideology. This creates a "cycle"

among ideologies that prevents pure-strategy equilibria. We then identify a class of environments where an equilibrium always exists. In a quadratic payoff setting, there always exists a winning ideology. This holds for any data and any ideology set.

The paper is organized as follows. Section 2 introduces the model. Section 3 reports the main result of the paper: the characterization of the features of ideologies that are selected in equilibrium (Theorem 1). Section 4 explores the role of evocativeness. Section 5 investigates existence and uniqueness. Section 6 studies the role of simplicity. Section 7 concludes.

Literature

Izzo et al. (2023) – henceforth IMC – is the main source of inspiration for the present work. They study an electoral competition where voters observe a public record of public policies and associated outcomes. Voters are initially "ideologically innocent". They look at parties as sources of theories of the world. Two policy-motivated candidates compete by providing voters with theories (*ideologies*) to explain outcomes as linear functions of policies. Voters prefer the ideology that minimizes the sum of squared residuals, with the addition of an idiosyncratic shock that is realized after ideologies are announced. In equilibrium parties propose different ideologies. The paper also considers the multi-dimensional case, where voters and parties also observe past realization of other random variables besides policies and outcomes. Equilibrium ideological conflict is shown to extend to those dimensions as well.

Our model extends IMC to a more abstract setting, allowing for ideologies to be defined as arbitrary classes of parameterized models and for voters' preferences to be expressed as arbitrary payoff functions. This allows us to show that fit is not the only criterion to choose between ideologies. Two other factors – which we nickname 'complexity' and 'fear' – also play a role and can ultimately override fit in determining the winning ideology.[2]

[2]There are a number of other differences with IMC that are highlighted in the model section. In particular, we abstract from one of IMC's main focuses: ideological polarization. We do so by assuming that candidates are pure office seekers.

Eliaz et al. (forthcoming) study the role of narratives in politics. The work is based on the theory of competing narratives developed in Eliaz and Spiegler (2020). A narrative is a causal model in the sense of Bayesian networks and can be represented as a directed acyclic graph. The narrative is a causal mapping between variables. Agents choose between competing narratives on the basis of anticipatory utility subject to the constraint that the narrative is consistent with available data. Eliaz et al. (forthcoming) extend the framework from one agent to a society composed of heterogeneous agents. A political platform is defined as a coalition of agents around a political narrative. The success of a platform is determined by the strength of the mobilization around it. The framework is used – among other things – to explain scapegoating, namely the rise of a false narrative that attributes a non-existent negative causal effect to a particular social group. Our paper focuses on different issues. Methodologically, our model differs from these two works for at least three reasons. First, our model is cast as an electoral contest as in IMC rather than a mobilization game. Second, our receivers have standard expected utility rather than anticipatory utility. Third, our narratives are represented as paramaterized classes of models rather than directed acyclic graphs, which allows us to discuss complexity in a direct way.[3]

Outside political economy, Schwartzstein and Sunderam (2021) develop a model where an agent observes some data and must choose between a default model and a model proposed by one or more persuaders.[4] The agent has a correct prior on the distribution of payoff-relevant states but prefers models the fit the data better. The persuader chooses among a set of available models and wishes to affect the decision of the receiver. This work obtains many results. The closest to our analysis is a characterization of the case with multiple persuaders. If they have different goals and there are no restrictions on the set of models they can announce, then in the only equilibrium persuaders will

[3]Levy, Razin, and Young (2022) study dynamic competition between two political groups holding different subjective models of the data generating process. One model is simpler than the other. They show that equilibrium features perpetual cycles and extreme policy under the simplistic model. Our definition of simple and complex models is related to Levy et al. (2022), especially in the class of examples developed in Section 6.

[4]See also Schwartzstein and Sunderam (2024) for a dynamic version of this framework where agents share models within a community.

all announce a model that perfectly fits the data and has no predictive power. Besides our focus on politics, our set-up differs from Schwarzstein and Sundaram's in a number of ways: we do not assume that the receiver has a correct prior on the distribution of states; as the statistical models we use are parametric, it is possible to discuss model complexity; finally, the receiver chooses between persuaders on the basis of expected utility rather than explanatory power (a side contribution of the paper is to understand when these two criteria lead to the same outcome).

Montiel Olea, Ortoleva, Pai, and Prat (2022) consider a competitive situation – like an auction for a productive asset – where multiple agents observe the same data and have different priors about how the data was generated. The paper shows that the predictive ability of a model can be decomposed into the ability to fit the data and a penalty for model complexity. If the dataset is small, the winner tends to be an agent with a simple model. As the number of observations grows, the winner is an agent with a complex model, possibly including irrelevant variables. The present paper also explores the tradeoff between fit and simplicity. However, it does so in a different context and with different results. In particular, in Montiel Olea et al. (2022), each agent is exogenously endowed with a single (linear regression) statistical model, while here the crux of the game is the selection of ideologies by parties and the selection of a party by a voter that is exposed to a pair of ideologies. The formal analysis is also entirely different, with the exception of an overlap in some steps of the linear regression example (which are highlighted in Section 6).[5]

More broadly, this paper creates a bridge between political economy and Bayesian statistics. The tradeoff between fit and simplicity is a theme of that literature (e.g. Van der Linde (2012), Murphy (2012)). While that theme is also present here, our paper differs because the tradeoff here occurs in a strategic multi-player environment. Our main result combines statistical concepts with game theoretic concepts in a way that is novel for both statistics and political economy. We explore this connection in Section 3, after stating Theorem 1.

[5]Aina (2024) also studies a set-up where the receiver selects a model on the basis of fit with observed data. However, the persuader announces her model(s) before a signal is realized, which the receiver observes before selecting a model. In our set-up all players observe the same data when they announce/select ideologies.

2 Set-Up

There are two candidates, A and B and one voter. The three agents observe the same data, D_n. Let $D_n \equiv (Z_1, \ldots, Z_n)$ be a random vector in which each component Z_i takes values in some subset $\mathcal{Z}$ of a Euclidean space $\mathbb{R}^d$.

We now present the definition of ideology used in this paper. Let $\mathcal{P}$ denote the set of all probability distributions over $\mathcal{Z}$. If we assume the data is i.i.d. we can define a *statistical model* for D_n simply as a collection of distributions over $\mathcal{Z}$ (Giné and Nickl, 2015, Chapter 1), or equivalently as a mapping $\mathcal{M} : \Theta_{\mathcal{M}} \to \mathcal{P}$ from some set $\Theta_{\mathcal{M}}$ to probability distributions over $\mathcal{Z}$. We refer to $\Theta_{\mathcal{M}}$ as the parameter space of model $\mathcal{M}$. Let π be a *prior* over the model's parameter space. An ideology h is a statistical model $\mathcal{M}$ together with a prior distribution π. Whenever convenient, we use $\mathcal{M}_h$ and π_h to denote, respectively, the statistical model and the prior parameter distribution under ideology h. Let H be the ideology set, which we assume to be finite.

The timeline is as follows. All three agents observe the data. Each candidate simultaneously announces an element of the ideology set H. Let h_A and h_B be their respective ideologies. The voter observes h_A and h_B and must select one of the two candidates. The voter has a payoff function $u : \mathcal{A} \times \mathcal{Z} \to \mathbb{R}$ that is known to both candidates.

The winning candidate is tasked with choosing the action a and she must do so according to her announced ideology and in order to maximize the voter's expected payoff. We allow for the possibility that the winning candidate observes a new public signal $X(Z_{n+1})$ after the election, where $X(Z_{n+1})$ is an $\mathcal{X}$-valued random variable that depends on Z_{n+1}. The presence of the additional signal X is not necessary for the analysis and, in many examples, one can assume X to be a constant. However, as we will explain later, the presence of the additional signal can sometimes be helpful to simplify the definition of actions and policies. In fact, we define the *policy* of candidate j as a mapping $\tau_j : \mathcal{X} \to \mathcal{A}$. The policy must maximize the expected value of $u(\tau_j(X(Z_{n+1})), Z_{n+1})$ given ideology h_j and data D_n.

At the beginning of the game, the voter does not hold any prior belief over the set of distributions that generated the data $\mathcal{D}_n$. He is also unaware of the ideology set H that the candidates have access to. We assume that the voter only becomes aware of the ideologies that are presented to him by the candidates. Moreover, we assume that the voter assigns a priori equal probability to each ideology. This means that the voter treats candidates symmetrically and assigns $1/2$ probability to h_A and $1/2$ probability to h_B.

The voter understands how candidates choose policies and therefore he knows the policies τ_A and τ_B. Using the two ideologies (and the equal prior probabilities over them), the voter can compute his expected utility under the two policies and chooses the candidate that offers the higher expected payoff.

The two candidates maximize the probability of being chosen and have no direct stake in the decision.

2.1 Discussion of Modeling Assumptions

1. *Connection with IMC.* As mentioned in the introduction, our model is inspired by Izzo et al. (2023). There are however important differences in the set-up, which we list below:

   (a) IMC's baseline model considers linear models of the form $y_t = \beta_0 x_t + \epsilon_t$, where ϵ_t is normally distributed with known variance σ^2. An ideology corresponds to selecting a particular value of β_0. This is a particular case of our set-up. It corresponds to assuming that $Z_t = (X_t, Y_t)$ and that an ideology h_{β_0} is the joint probability distribution generated by the linear model $y_t = \beta_0 x_t + \epsilon_t$. The ideology set is defined over all possible values of the coefficient β_0, namely $H = \{h_{\beta_0} : \beta_0 \in \mathbb{R}\}$.[6] Note that within each ideology h_{β_0} there is no parameter choice: all ideologies are simple models, and there is no role for complexity.

[6] One small technical difference is that we assume that H is finite, while IMC has a continuum of ideologies.

(b) A crucial point of departure is that we allow ideologies to be statistical models with a prior distribution over its parameters.

(c) Another crucial difference is that we allow the voter to have a generic payoff functions and that the voter chooses the winning candidate on the basis of expected payoff.

(d) We differ from IMC also because our candidates are office-seekers. Theirs instead have policy preferences, which – coupled with the assumption that an idiosyncratic shock ξ is realized after the ideology announcements – leads to the possibility of equilibrium policy divergence. See point below on candidate commitment.

(e) As mentioned, our voter uses model mixture rather than model selection. See point below.

However, note that, if we used IMC's ideology set, Theorem 1 would predict that voters would choose on the basis of fit only (because all available ideologies are equal in terms of simplicity and fear). Thus, our results are consistent with IMC's.

2. *Model Selection Process.* In IMC and Schwartzstein and Sunderam (2021), model selection is performed on the basis of explanatory power. The decision-maker selects the model that maximizes the likelihood of the available data. Instead, our set-up assumes that model selection is performed on the basis of expected payoff. The voter assigns equal prior probability to the two models (ideologies) and uses the data to compute a posterior probability of the outcome variable. In turn, the posterior is used to decide which model would yield a higher expected payoff. This approach aligns with one of the standard approaches in Bayesian statistics in the presence of multiple models (Hoeting, Madigan, Raftery, and Volinsky (1999)). The decision maker builds a prior distribution as a mixture of the individual models, usually with equal weights, and then uses the data to estimate a posterior (Titterington, Smith, and Makov (1985)).[7]

[7]While we make no claim that this is how people actually vote, there is some evidence that real decision making

A side contribution of our paper is to show when the two model selection approaches yield the same outcome. When the models under consideration do not have free parameters (as is the case in IMC and Schwartzstein and Sunderam (2021)) and the payoff function is quadratic (as in IMC), then model selection on the basis of likelihood yields the same choice as model selection on the basis of expected payoff (See Corollary 5 and related discussion on page 27). In this sense, our result confirms and extends the existing literature.

3. *Electoral Discourse.* Our model can also be interpreted within a broader theory of how candidates' ideological messages affect the electoral discourse. Vavreck (2009) examines the relationship between the economy, campaign messaging, and election outcomes over 60 years of US presidential elections. Candidates can choose to frame the economic situation in different ways. Voters in turn use the two candidates' messages to make sense of the current situation and decide who to vote for.

4. *Office-seeking Committed Candidates.* the assumption that candidates have no ex ante policy-preferences but are committed to the ideology they announce is in line with Downsian electoral competition, where elected officials stick to their announced electoral stance, even though they have no direct preferences over policy. At a theoretical level, such time consistency can be rationalized if voters care about 'character' (Kartik and McAfee (2007)). Empirically, there is evidence that voters dislike candidates who alter their ideological positions (e.g. Doherty, Dowling, and Miller (2016)).

   Our model could be extended to policy-motivated candidates and a probabilistic element in line with IMC. The three elements we identify – fit, fear, simplicity – would still affect which ideologies are chosen in equilibrium. For expositional reasons, we focus on the office-seeking case.

can be represented as a Bayesian inference process over alternative models. Kemp, Perfors, and Tenenbaum (2007) review a growing body of experimental psychological evidence documenting how human cognition can be represented as the acquisition and updating of multiple "overhypotheses" that help us learn causal models from sparse data.

5. *The Truth.* The set-up makes no mention of the true data-generating process. Obviously, the voter does not know it. There is also no assumption that the candidates know it. Theorem 1 holds as stated whether the candidates know the true model or not. The only point in the paper where we will assume a true data-generating process is in the numerical illustration of the results in Section 6.

6. *Prior Weights on Ideologies.* We assume that the voter assigns the same prior probability to the two ideologies announced by the candidate. The set-up can be extended to allow voters to find certain ideologies ex ante more convincing than others. We could assign a *plausibility factor* $Q_{h^*,h} \in [0,1]$ to each pair of ideologies $h^*, h \in H$. The factor would represent the ex-ante plausibility of ideology h^* relative to h. Theorem 1 would hold as stated with an additional plausibility ratio. See Remark 1 on page 26.

7. *Ideology set.* Our set-up is silent on how the ideology set is generated. Our key result (Theorem 1) holds for any ideology set as defined above. In practice, one may want to impose additional psychological or cultural constraints on the statistical models that voters can understand or accept, but this is outside the scope of the present work.[8]

2.2 Simple Example

We now present a simple example of our set-up. The goal is to develop intuition on the mechanics of our approach and on the main forces that determine the winning ideology. The rest of the paper will show that this intuition is general.

The data consists of past observations of two states of the world: an outcome variable $Y \in \{-1, 1, 10\}$ and a covariate $X \in \{\text{Green}, \text{Red}\}$. There are six past observations:

[8]For a discussion of theories of political ideology, see Freeden (2003).

i	1	2	3	4	5	6
Y_i	1	-1	1	-1	-1	1
X_i	Green	Red	Red	Green	Red	Green

The data has three features that will matter for the analysis. First, the "extreme outcome" $Y = 10$ has never been observed. Second, the other two outcomes have occurred the same number of times. Third, the outcome Y displays an association with the covariate X: Under Green, $Y = 1$ has appeared 2 times out of 3, while under Red it has appeared only 1 time out of 3.

The election is held at the beginning of time 7. After the election, the winner will observe the realization of the covariate X_7 and will select an action $a \in \Re$. Then Y_7 will be realized, and voter's payoff will be:

$$u(a, Y_7) = -|a - Y_7|^3.$$

The choice of a cubic loss function is meant to capture the voter's aversion to extreme deviations from his policy bliss point (the role of curvature will be investigated in Section 5).

There are three ideologies:

(H_S) *The Simple Ideology*: The two "moderate" outcome realizations have probability 50% each and they do not depend on the covariate. Namely, $\Pr(Y = 1) = \Pr(Y = -1) = \frac{1}{2}$ and $\Pr(Y = 10) = 0$.

(H_C) *The Complex Ideology*: The probabilities of the two "moderate" outcome realizations may depend on the covariates. Namely, $\Pr(Y = 1|\text{Green}) = \theta_G$, $\Pr(Y = 1|\text{Red}) = \theta_R$, and $\Pr(Y = 10) = 0$, with θ_G and θ_R independently distributed as uniforms on $[0, 1]$.

(H_A) *The Alarmist Ideology*: It is the same as the simple ideology excepts that it assigns a non-zero probability to the extreme outcome that has never been observed.[9] Namely, $\Pr(Y = 1) =$

[9]For the sake of completeness, assume that all three ideologies assign equal probability to the two realizations of the covariates. Namely, $\Pr(X_i = \text{Green}) = \Pr(X_i = \text{Red}) = 0.5$.

$\Pr(Y=-1)=\frac{1}{2}-0.01$ and $\Pr(Y=10)=0.02$

First, note that the three ideologies can be ranked by how well they can explain the data. The maximized likelihood of the six observations within the three ideologies are given, respectively, by:

$$\Lambda_C=\left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right)^2; \qquad \Lambda_S=\left(\frac{1}{2}\right)^6; \qquad \Lambda_A=\left(\frac{49}{50}\right)^6.$$

If the voter used Λ as the criterion to assess ideologies, the ranking would be: $H_C \succ H_S \succ H_A$. Office-seeking candidates would both announce the complex ideology.[10] However, we now show that in our setting the voter ranks the ideologies in the exact opposite order. We do that in two steps.[11]

Does H_S beat H_C?

Suppose candidate A announces the simple ideology while B announces the complex ideology. Who will the voter elect?

If A is in power, policy will be chosen according to H_S. The optimal policy will be the midpoint between $Y=-1$ and $Y=1$, namely $a_S^*=0$. If B is in power, the policy will be tailored to the realization of the covariate, slightly positive under Green and slightly negative under Red:

$$a_C^*(G)=5-2\sqrt{6}\approx 0.101 \text{ and } a_C^*(R)=-5+2\sqrt{6}\approx -0.101$$

This tailoring improves the voter's payoff if the complex ideology is correct, but it actually worsens it – by adding noise to the policy – if the simple ideology is correct.

What probability does the voter assign to the two ideologies? The voter starts with equal prior probabilities and uses the data to compute posteriors according to Bayes' theorem. Let $\Pr[D|H]$ be the likelihood of observing data D within ideology H. The posterior probability of the complex

[10] The same ordering would hold if instead of the likelihood we defined explanatory power as least squared residuals.
[11] Here, we present an intuitive summary of the results. The full calculations are available in Appendix ?.

ideology (given that the complex and simple ideologies are presented), denoted by P_{CS}, equals:

$$P_{CS} := \frac{\Pr[D|H_C]}{\Pr[D|H_S] + \Pr[D|H_C]} = \frac{64}{64+144} = \frac{4}{13} \approx 0.307.$$

The voter is skeptical of the complex ideology. That is a manifestation of the Bayesian Occam's Razor, that we will explore in more depth after the theorem. Given the two additional degrees of freedom (θ_G and θ_R), the complex technology has a large potential to explain observations. Here, H_C underperforms because the association between outcome and color is relatively weak.[12]

So, which candidate offers the higher expected payoff? The expected (and realized) payoff under the simple ideology is simply $U_S = -1$. One can show that the expected payoff under the complex ideology is lower: $U_C = -1.011$. The simple ideology beats the complex ideology.

2.2.1 Does H_S beat H_A?

Now suppose instead that candidate A announces the simple ideology and candidate B announces the alarmist ideology. Who will the voter select?

As in the previous subsection, the policy chosen according to H_S is still $a_S^* = 0$. Instead, the policy chosen under H_A is greater than zero because of the small probability that $Y = 10$. Standard computations show that $a_A^* = 59 - 7\sqrt{(}69) \approx 0.853$.

When H_A and H_S are announced, the posterior probabilty of the former, denoted P_{AS}, is

$$P_{AS} = \frac{49^6}{50^6 + 49^6} \approx 0.469,$$

which is below 50% because it predicts an outcome, $Y = 10$, that is never observed.

However, the alarmist ideology enjoys an advantage on payoffs. If H_A is correct and the voter chooses the candidate with H_S rather than the one with H_A, the expected payoff loss will be high

[12]For instance, if all three 1 outcomes were associated with Green and all three -1 outcomes were associated with Red, the posterior probability of the complex ideology would be much higher: $P_{CS} = \frac{4}{5}$.

(about 2.6). Instead, if H_S is correct and the voter chooses the candidate with H_A rather than the one with H_S, the expected payoff loss will be lower (about 2.1). Putting together the probability effect and the payoff effect, one can see that the expected payoff is higher under H_A. The alarmist ideology beats the simple ideology.

2.2.2 Summing Up

The analysis above has shown that H_A beats H_S and H_S beats H_C – the opposite ordering compared to the a ranking based on explanatory power. This is because there are two additional factors to be considered. In the comparison between H_C and H_S, the Bayesian Occam's Razor favored the simpler model. In the comparison between H_A and H_S, the alarmist ideology won by "making it expensive to trust" the other model.

To find the equilibrium of the game, we should also ask whether the alarmist ideology beats the complex one.[13] We omit that comparison, but one can show that if A announces H_A and B announces H_C, then A wins. The three comparisons taken together show that the only equilibrium of the game is when both candidates announce the alarmist ideology.

The example has introduced two effects that play alongside the explanatory power effect to determine the winning ideology. The next section will show that these three effects are universal and will derive a general formula to account for them and determine the winning ideology.

3 Equilibrium Analysis

An electoral game is defined by $(\mathcal{Z}, \mathcal{X}, \mathcal{A}, D, H, u)$. For a given electoral game, define a (pure-strategy) equilibrium as a situation where A announces ideology $h_A \in H$, B announces ideology $h_B \in H$ (possibly the same as h_A), and neither candidate has a strict incentive to deviate. To simplify analysis and avoid uninteresting equilibria, we assume that the voter must randomize

[13]This third comparison is not superfluous. Section 4 shows that, except under restrictive conditions, 'beats' is not a transitive relation in our set-up. Cycles are possible.

between the two candidates if he receives the same expected utility from both of them.

We introduce the following notation:

- Let $\Pr(Z|\theta;h)$ be the probability mass function of the random variable Z in ideology h at parameter value θ.[14]

- Let

$$\Pr(D_n|\theta;h) := \prod_{i=1}^{n} \Pr(Z_i|\theta;h)$$

be the likelihood of data $D_n := (Z_1, \ldots, Z_n)$ in ideology h at parameter value θ. Note that, given h and θ, $\Pr(D_n|\theta;h)$ is the likelihood of a dataset that consists of n independent and identically distributed draws of Z from $\Pr(Z|\theta;h)$. In what follows, we omit the subscript n from D_n.

- Given ideologies h_A and h_B, the voter uses a *mixture* of the two ideologies to model the distribution of the available data. We label the mixture h_v. The voter assumes that the distribution of the random variable Z is given by:

$$\Pr(Z|\theta_v;h_v) := \Pr(Z|\theta_A;h_A)\alpha + \Pr(Z|\theta_B;h_B)(1-\alpha) \text{ for all } Z \in \mathcal{Z},\ \theta_A \in \Theta_A \text{ and } \theta_B \in \Theta_B,$$

where α is a Bernoulli random variable independent of θ_A and θ_B. The vector $\theta_v \equiv (\alpha, \theta_A, \theta_B)$ parameterizes the mixture model. This is standard notation for mixture models; see Chapter 1, Section 1.1 in Titterington et al. (1985).

We assume that the voter uses the priors π_A and π_B to form a new prior distribution π_v over θ_v. In particular, we assume that under the prior π_v, the parameters $(\alpha, \theta_A, \theta_B)$ are

[14]For the sake of exposition, the definitions are given under the assumption that the support of $\mathcal{Z}$ is finite. They can be easily extended to account for random variables of the absolutely continuous type, by replacing the probability mass function $\Pr(Z|\theta;h)$ by the probability density function $f(Z|\theta;h)$. Recall that also the parameter sets and the ideology set are finite.

independently distributed with marginals $\alpha \sim \text{Bernoulli}(1/2)$, $\theta_A \sim \pi_A$ and $\theta_B \sim \pi_B$.[15] Since $\pi_v(\alpha = 1) = \pi_v(\alpha = 0)$, the voter's mixture model assigns equal *prior* probability to each ideology. Note that the 'mixture' ideology h_v is mathematically equivalent to a 'primitive' ideology with statistical model $\Pr(Z|\cdot; h_v)$ and prior π_v.

- For any ideology h (including those based on mixtures of ideologies, like h_v), we can compute the likelihood of the data as above. We can also compute the posterior distribution of Z given the data

$$\Pr(Z|D; h),$$

and the expected utility of any action a:

$$E[u(a, Z)|D; h].$$

- Let

$$\Lambda(D|h) := \sup_{\theta \in \Theta_h} \Pr(D|\theta; h)$$

be the maximized likelihood of data D in ideology h, where Θ_h denotes the parameter space of the statistical model in ideology h.

- Let $a^*(h, D)$ be the value of $a \in \mathcal{A}$ that maximizes $E[u(a, Z)|D; h]$. It is the optimal action given data D and ideology h.

- Let the voter's posterior expected utility of selecting candidate $i \in \{A, B\}$ be

$$U_i(h_A, h_B, D) := E[u(a^*(h_i, D), Z)|D; (h_A, h_B))].$$

[15] If $\pi_A = \pi_B$, simply define $\pi_v = \pi_A$. Note that the assumption that θ_A and θ_B are mutually independent is made to pin down π_v, but it is not needed

- Let $R(D|h)$ be the *ex-ante accuracy* of ideology h over data D. It is defined as

$$R(D|h) := \frac{\int_{\Theta_h} \Pr(D|\theta; h) d\pi_h}{\Lambda(D|h)}$$

The interpretation of R will be discussed after Theorem 1.

We are now ready for the equilibrium characterization result. We say that ideology h is supported in equilibrium if there is an equilibrium where at least one candidate announces it. The theorem identifies a necessary and sufficient condition:

Theorem 1. *Given data D, ideology h^* is supported in equilibrium if and only if, for any other ideology $h \in H$,*

$$\frac{\Lambda(D|h^*)}{\Lambda(D|h)} \cdot \frac{R(D|h^*)}{R(D|h)} \geq \frac{E\left[u(a^*(h, D), Z) - u(a^*(h^*, D), Z)|D; h\right]}{E\left[u(a^*(h^*, D), Z) - u(a^*(h, D), Z)|D; h^*\right]} \tag{1}$$

We first prove the theorem and provide intuition for the proof. We then turn to interpreting the terms in theorem.

3.1 Proof of Theorem 1

We prove the theorem through two lemmas below. As the lemmas shed light on the intuition for the structure of the equilibrium, we report them in the text. The first lemma deals with the game-theoretic structure of equilibrium, while the latter deals with the statistical side.

The first lemma is:

Lemma 2. *Given data D, ideology h^* is supported in equilibrium if and only if, for any other ideology $h \in H$,*

$$U_A(h^*, h, D) \geq U_B(h^*, h, D).$$

The intuition for the first lemma is that a candidate can win by announcing an ideology that

gives the voter a higher expected payoff than that announced by the opponent. However, such payoff depends – through the voter's prior – on both the ideology announced by the candidate and the ideology announced by the opponent. If there exists an ideology h^* that 'beats' all other ideologies in pairwise comparisons, then that ideology is supported in equilibrium. The lemma shows that this is also a necessary condition (and Section 4 will discuss the possibility that a pure-strategy equilibrium does not exist).

Proof. Let $\Pi_A(h_A, h_B, D)$ be the probability that A is elected if she announces h_A while B announces h_B. This is what A maximizes and what B minimizes. By assumption, if both announce the same ideology h,

$$\Pi_A(h, h, D) = \Pi_B(h, h, D) = \frac{1}{2}.$$

Also,

$$\Pi_A(h_A, h_B, D) = \begin{cases} 1 & \text{if} \quad U_A(h_A, h_B, D) > U_B(h_A, h_B, D) \\ \frac{1}{2} & \text{if} \quad U_A(h_A, h_B, D) = U_B(h_A, h_B, D) \\ 0 & \text{if} \quad U_A(h_A, h_B, D) < U_B(h_A, h_B, D) \end{cases}$$

To prove the "if" part, suppose:

$$U_A(h^*, h, D) \geq U_B(h^*, h, D) \text{ for all } h.$$

Suppose the two candidates are announcing (h^*, h^*) and consider a deviation on the part of B to any other ideology h. The inequality above implies $\Pi_B(h^*, h, D) \leq 1/2$, proving that a deviation is not profitable. A symmetric argument applies to possible deviations of A from h^*. Hence (h^*, h^*) is an equilibrium.

For the "only if" part, suppose the condition fails for some ideology h', namely,

$$U_A(h^*, h', D) < U_B(h^*, h', D).$$

implying $\Pi_B(h^*, h', D) = 1$.

Now suppose for contradiction that there exists an equilibrium where A announces h^* and B announces some h'' (which could be any h, including h^* or h'). As B can guarantee victory by switching to h', it must be that $\Pi_B(h^*, h'', D) = 1$. Hence, in the putative equilibrium A loses for sure. However, if B plays h'', A can guarantee itself a 50% chance of winning by switching to h''. Hence, there is no equilibrium where A plays h^*. The same argument applies to a putative equilibrium where B plays h^*. □

The second lemma for the proof of Theorem 1 maps the game-theoretic characterization of Lemma 2 into a statistical characterization:

Lemma 3. *Given any two ideologies $h^* \in H$ and $h \in H$, the condition $U_A(h^*, h, D) \geq U_B(h^*, h, D)$ holds if and only if*

$$\frac{\Lambda(D|h^*)}{\Lambda(D|h)} \cdot \frac{R(D|h^*)}{R(D|h)} \geq \frac{E\left[u(a^*(h, D), Z) - u(a^*(h^*, D), Z)|D; h\right]}{E\left[u(a^*(h^*, D), Z) - u(a^*(h, D), Z)|D; h^*\right]}.$$

Proof. When candidate A announces ideology h^* and candidate B announces ideology h, we have defined the voter's posterior expected utility of selecting candidate A as

$$U_A(h^*, h, D) := E[u(a^*(h^*, D), Z)|D; (h_A, h_B))],$$

where $a^*(h^*, D)$ is the optimal action given data D and ideology h^*. The first step in the proof of Lemma 3 is to exploit the fact that the voter uses a mixture of the two ideologies to model the distribution of the available data.

Using Lemma 10 and Lemma 11 (proven in the Appendix) we can show that the mixture model

implies that

$$U_A(h^*, h, D) = E[u(a^*(h^*, D), Z)|D; h^*] \Pr(\alpha = 1|D; h^*, h) + E[u(a^*(h^*, D), Z)|D; h] \Pr(\alpha = 0|D; h^*, h).$$

That is, the voter can first compute the expected payoff of the action $a^*(h^*, D)$ given the data, and considering two separate scenarios in which only one of the ideologies is presented. The algebra in Lemma 10 and Lemma 11 shows that $U_A(h^*, h, D)$ can be written as a convex combination between these two payoffs, with weights given by the posterior probability of each ideology.

Given this result, we can then re-write the inequality in Lemma 2

$$U_A(h^*, h, D) \geq U_B(h^*, h, D)$$

as

$$\frac{\Pr(\alpha = 1|D; h^*, h)}{\Pr(\alpha = 0|D; h^*, h)} \geq \frac{E\left[u(a^*(h, D), Z) - u(a^*(h^*, D), Z)|D; h\right]}{E\left[u(a^*(h^*, D), Z) - u(a^*(h, D), Z)|D; h^*\right]}$$

By Bayes' Theorem and because we have assumed that the two ideologies are ex ante equiprobable,

$$\Pr(\alpha = 1|D; h^*, h) = \frac{\Pr(D|\alpha = 1; h^*, h)}{\Pr(D|\alpha = 1; h^*, h) + \Pr(D|\alpha = 0; h^*, h)}$$

Thus,

$$\frac{\Pr(\alpha = 1|D; h^*, h)}{\Pr(\alpha = 0|D; h^*, h)} = \frac{\Pr(D|\alpha = 1; h^*, h)}{\Pr(D|\alpha = 0; h^*, h)}$$

Re-write:

$$\begin{aligned} \Pr(D|\alpha = 1; h^*, h) &= \int_{\Theta_{h^*}} \Pr(D|\theta; h^*) d\pi_{h^*} \\ &= \left(\sup_{\theta \in \Theta_{h^*}} \Pr(D|\theta, h^*) \right) \cdot \frac{\int_{\Theta_{h^*}} \Pr(D|\theta; h^*) d\pi_{h^*}}{\left(\sup_{\theta \in \Theta_{h^*}} \Pr(D|\theta, h^*) \right)}. \end{aligned}$$

Using an analogous decomposition for $\Pr(D|\alpha = 0; h^*, h)$ We therefore have:

$$\frac{\Pr(D|\alpha = 1; h^*, h)}{\Pr(D|\alpha = 0; h^*, h)} = \frac{\Lambda(D|h^*)}{\Lambda(D|h)} \cdot \frac{R(D|h^*)}{R(D|h)}, \tag{2}$$

which completes the proof. □

The second lemma builds on the observation that data combines with the announced ideologies in two ways. First, data is used within each ideology to update the distribution over parameters, which in turn will be used by the candidate who announced that ideology to make decisions if she is elected. The voter is aware of this process and takes it into account in computing the expected value of electing the candidate. Second, data is used by the voter to update the relative probability that the two ideologies, h_A and h_B are correct.

The proof of the lemma proceeds by showing that these two forms of update can be, to a certain extent, decomposed. Namely, the expected utility that the voter receives if he elects candidate A if A announces h^* and B announces h can be written as:

$$\begin{aligned} U_A(h^*, h, D) &= E[u(a^*(h^*, D), Z)|D; h^*, h] \\ &= E[u(a^*(h^*, D), Z)|D; h^*] \Pr(\alpha = 1|D; h^*, h) \\ &\quad + E[u(a^*(h^*, D), Z)|D; h] \Pr(\alpha = 0|D; h^*, h) \end{aligned}$$

The expected utility is a weighted sum of the expected utility under the two ideologies (updated with the data) where the weights are the relative probabilities of the two ideologies (updated with the data).

Armed with this decomposition, one can express the relative advantage of ideology h^* over ideology h as in the inequality in Lemma 3 captures the effect of the expected utility under the two ideologies. The left-hand side captures the relative probabilities of the two ideologies. For the

latter, we show that in our setting the ratio of those probabilities can be decomposed as:

$$\frac{\Pr(D|\alpha=1;h^*,h)}{\Pr(D|\alpha=0;h^*,h)} = \frac{\Lambda(D|h^*)}{\Lambda(D|h)} \cdot \frac{R(D|h^*)}{R(D|h)}.$$

These two terms are interpreted in the next subsection.

We conclude with an observation on the assumption that the voter assigns probability 50% to each ideology:

Remark 1. *Theorem 1 currently assumes that the voter assigns the same prior probability to the two ideologies announced by the candidate. That is, we assume that the* plausibility *of each ideology* $Q_{h^*,h} := P(\alpha = 1; h^*, h)$ *equals 1/2. Suppose that ideology* h^* *is ex ante more convincing than* h*. Then, the right-hand side of Equation* (2) *would have an additional plausibility factor* $Q_{h^*,h}/(1 - Q_{h^*,h})$*. This means that Theorem 1 would then say that ideology* h^* *us supported in equilibrium if and only if, for any other ideology*

$$\frac{\Lambda(D|h^*)}{\Lambda(D|h)} \cdot \frac{R(D|h^*)}{R(D|h)} \geq \frac{1 - Q_{h^*,h}}{Q_{h^*,h}} \cdot \frac{E\left[u(a^*(h,D),Z) - u(a^*(h^*,D),Z)|D;h\right]}{E\left[u(a^*(h^*,D),Z) - u(a^*(h,D),Z)|D;h^*\right]}.$$

Thus, if ideology h^* *is ex-ante very plausible, it is be easier to support it in equilibrium.*

3.2 Discussion of Theorem 1

We now offer an intuitive discussion of Theorem 1. To interpret the theorem, let us state one additional result:

Corollary 4. *If ideology* h^* *is supported in equilibrium, then there exists a symmetric equilibrium where both candidates announce* h^**.*

Proof. If h^* is supported in equilibrium, by Lemma 2,

$$U_A(h^*,h,D) \geq U_B(h^*,h,D) \text{ for all } h.$$

Under that condition, the "if" part of the proof of Lemma 2 shows the existence of a (h^*, h^*) equilibrium. □

The corollary implies that if there exists an asymmetric equilibrium where the two candidates select different ideologies, say (h', h''), then there also exist the two corresponding symmetric equilibria (h', h') and (h'', h''), where the candidates choose the same ideology. This means that to find equilibria it suffices to focus on symmetric equilibria.

As the intuition is more transparent for symmetric equilibria, let's suppose that the condition of the theorem is satisfied and let's focus on the equilibrium where both candidates select ideology h^*. A candidate who deviates from equilibrium (h^*, h^*) and chooses a different ideology will weakly reduce her chance of being elected. Why?

The theorem says that ideology h^* is supported in equilibrium if inequality (1) holds against any other ideology h in the ideology set. The inequality has three terms, each of which can be interpreted as a source of relative advantage for ideology h^* against any alternative ideology h. Let us analyze the three terms one by one.

The *first term*,

$$\frac{\Lambda(D|h^*)}{\Lambda(D|h)},$$

is the ratio of the maximized likelihood for ideology h^* and the maximized likelihood for another ideology h. It measures how well ideology h^* can fit the available data compared to any other ideology h. Everything else equal, in equilibrium candidates will announce ideologies that best fit the data.

To isolate the effect of fit, make two simplifying assumptions. First, assume that the state Z is a scalar and the payoff function u is quadratic ($u(a, Z) = -(a - Z)^2$) ("quadratic payoff assumption"). Second, assume that all ideologies are maximally simple: for every h, the parameter set Θ_h is a singleton ("simple ideologies assumption").

Corollary 5. *Under the quadratic payoff assumption and the simple ideologies assumption, ideology h^* is supported in equilibrium if and only if:*

$$\Lambda(D|h^*) = \max_{h \in H} \Lambda(D|h).$$

Proof. Under the simple ideologies assumptions, the term $R(D|h)$ is always one. Under the quadratic payoff assumption, the third term is always one (see the proof of Proposition 6). □

Under these two assumptions, the winning ideology maximizes fit. In equilibrium, the voter tends to select the ideology that best fits the data as in IMC. It also aligns with the model selection criterion in Schwartzstein and Sunderam (2021).

The *second term*,

$$\frac{R(D|h^*)}{R(D|h)},$$

can be seen as an 'ideology simplicity premium'.

Let us re-visit the example of Section 2.3. We argued that complex ideology had a pro (it fit the data better) and con (because of its flexibility, it carried a risk of overfitting). These two effects are now captured by the two first terms of the theorem:

$$\frac{\Lambda(D|h_C)}{\Lambda(D|h_S)} = \frac{\left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^2}{\left(\frac{1}{2}\right)^6} \approx 1.404 \qquad \frac{R(D|h_C)}{R(D|h_S)} = \frac{\frac{\int_0^1 \theta^4 (1-\theta)^2 d\theta}{\left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^2}}{1} \approx 0.43$$

The first term is greater than one and favors the complex ideology. The second term is smaller than one and favors the simple ideology. The product of the two terms (about 0.6) is smaller than one. Ultimately the simple ideology prevails.[16]

The second term relates to the *Bayesian Occam's Razor Principle*. According to Murphy (2012, p. 183?): "[One way to understand the Bayesian Occam's Razor Principle] is to compare the relative

[16] In this comparison, the third term is not exactly one but very close to. The discussion of the third term is left for later.

predictive abilities of simple and complex models. Since probabilities must sum to one, we have $\sum_D \Pr(D|h) = 1$, where the sum is over all possible datasets [using our paper's notation]. Complex models, which can predict many things, must spread their predicted probability mass thinly, and hence will not obtain as large a probability for any given data set as simpler models. This is sometimes called *the conservation of probability mass principle*.

To see the conservation of probability mass principle in action within our context, let us consider the two extreme cases of the term $R(D, h)$.

One extreme corresponds to the simplest possible type of ideology. Assume that the parameter space Θ_h is a singleton (as in the simple ideology H_S of the Example). Then, the numerator and the denominator of $R(D, h)$ are identical and the ratio attains its maximum value, 1.

At the opposite extreme, the most complex ideology in an Occam Razor's sense is a "theory of everything," that is, a statistical model whose parameters can be adjusted to explain any realized data perfectly. Namely, consider an ideology h such that, for any possible realization of D there exists a value $\theta(D)$ such that $\Pr(D|h, \theta(D)) = 1$. This can be interpreted as a collection of 'ad hoc' models – one for every possible data realization. The ideology will always fit the data perfectly. In that case, the value of R is $1/\#\mathcal{D}$, where $\#\mathcal{D}$ is the number of possible realizations of dataset D. If we let the number of possible realizations go to infinity, we get that $R \to 0$, which is the minimum value of R.

The second term captures a continuum between these two extremes. As the term decreases, the voter becomes more worried that the relative ability of h^* may be spurious and therefore the ideology has limited predictive power.[17]

The *third term* – on the right-hand side of inequality (1) – can be seen as a "fear factor". It captures the relative cost of making mistakes in the two ideologies. The numerator measures the

[17]There is a connection between our simplicity term and Popper's (1959) notion of falsifiability. Popper defines the "simplicity of a theory" as its degree of falsifiability, which depends on the simplicity of the theory. The example he provides is that a theory that postulates a quadratic relation between two variables is less falsifiable than one that postulates a linear relation because it spreads probability mass more thinly. At the extreme, the theory of everything discussed above is not be falsifiable because any data realization is consistent with it.

expected benefit within ideology h of choosing the appropriate action $a^*(h, D)$ rather than the action that is optimal under h^*. The denominator does the converse for ideology h^*.

Thus, the third term favors ideology h^* if the cost of adopting it when h is correct is low and instead the cost of adopting h when h^* is correct is high. An ideology is more likely to win if – within the ideology – it exploits the voter's payoff function to evoke bad outcomes of competing ideologies. It is less likely to win if it receives this treatment from other ideologies.

Going back to the example in Section 2, the alarmist ideology wins because of the third term. Following the simple ideology when the alarmist one is correct is more costly than the converse.[18]

There is a parallel between the third term and Pascal's Wager. Pascal (1670) considers the choice between believing in God and not believing in God and argues that, if God exists and one behaves as if God did not exist, one incurs an infinite loss (perpetual hell), while if God doesn't exist and one behaves as if God existed the loss is finite (because there is no afterlife). For that reason, an agnostic rational person should choose to behave as if God existed. The religious model of the world has a strategic advantage over the non-religious model of the world because it evokes an event (the afterlife) that the non-religious model provides no protection against.

As both the numerator and the denominator of the third term are computed under the assumption that one ideology is correct, the value of the third term is independent of the posterior probabilities of the two ideologies being correct. It only depends on the probability distribution of outcomes within ideologies and the voter's payoff function.

Section 5 explores the role of the third term in more detail.

[18]To apply Theorem 1 to the comparison between H_A and H_S in the example, note that both ideologies are equally simple, so the second term is one. The expected payoff from the alarmist ideology is at least as large as the expected payoff from the simple ideology if the first term is larger than the third term:

$$\frac{P_{AS}}{1-P_{AS}} \geq \frac{E\left[u(a^*(H_A, D), Z) - u(a^*(h_S, D), Z)|D; H_S\right]}{E\left[u(a^*(H_S, D), Z) - u(a^*(H_A, D), Z)|D; H_A\right]},$$

which is satisfied because the left-hand side is approximately $\frac{0.469}{1-0.469} = 0.88$ while the right-hand side is about $2.1/2.6 = 0.80$.

4 Existence of Equilibrium

This section (which can be skipped without affecting the understanding of the remainder of the paper) complements the equilibrium characterization of Theorem 1 with results on existence. Namely, given an ideology set H, a dataset D, and a payoff function u, is there at least one ideology I^* that is supported in equilibrium?

Equilibrium was defined in pure strategies. An equilibrium in mixed strategies always exists.[19] However, as mixed-strategy equilibria of electoral games are difficult to analyze and interpret, we focus on pure-strategy equilibria.

If the set of models has only two elements, then Theorem 1 trivially guarantees existence. But it is easy to think of cases where there are more than two ideologies.

This section begins with a simple three-model example where existence fails. The failure is due to the presence of a cycle. In pairwise competitions, the first model ideology beats the second ideology, the second beats the third, and the third beats the first. The general intuition is that different ideologies evoke different outcomes. The 'strength' of an ideology is therefore not universal. It depends on which ideology it is paired against.

The second part of the section identifies a special environment where ideologies have 'universal strength'. We prove that if the voter's payoff function is quadratic, then for any dataset and any ideology set, there exists a winning model. As we will see, in that case a cycle among models is impossible.

[19] Recall that the action set of each candidate – the set of ideologies H – is assumed to be finite. For each action profile (h_A, h_B), the voter chooses with certainty exactly one of these three actions: elect A, elect B, randomize between the two with equal probability. This reduces the game to a finite-action zero-sum game.

4.1 Non-Existence Example

Consider a situation with three possible states,

$$\mathcal{Z} = \{\text{Rainy}, \text{Sunny}, \text{Windy}\},$$

and three possible actions:

$$\mathcal{A} = \{\text{Shelter}, \text{Canopy}, \text{Wall}\}.$$

The voter's utility function $u(a, z)$ is given by:

	Rainy	Sunny	Windy
Shelter	2	0	1
Canopy	1	2	0
Wall	0	1	2

The actions are mutually exclusive. Each action fits well one state (for example a canopy protects against the sun), is somewhat useful in other state (the canopy provides some protection against the rain), and is useless in the remaining state (the wind would blow the canopy away).

There are three possible ideologies (all three have no parameters – they have zero degrees of freedom):

$$h_R = \{\Pr[\text{Rainy}] = 0.8, \Pr[\text{Sunny}] = 0.1, \Pr[\text{Windy}] = 0.1\}$$

$$h_S = \{\Pr[\text{Rainy}] = 0.1, \Pr[\text{Sunny}] = 0.8, \Pr[\text{Windy}] = 0.1\}$$

$$h_W = \{\Pr[\text{Rainy}] = 0.1, \Pr[\text{Sunny}] = 0.1, \Pr[\text{Windy}] = 0.8\}$$

Thus, each ideology assigns high probability to one of the states and low probability to the remaining two.

Suppose that the data includes three observations: each state has been observed exactly once:

$$D = \{y_1 = \text{Rainy}, y_2 = \text{Sunny}, y_3 = \text{Windy}\}.$$

We can easily see that this game has no pure-strategy equilibrium.

Let's apply Theorem 1 to the three pairwise comparisons. Suppose that Candidate A announces m_R and Candidate B announces m_S. The second coefficient in the theorem is 1 for all models because the models have no parameters. The first coefficient is also 1 because of the symmetry of the data and the ideologies.

The third coefficient is instead different from 1:

$$\begin{aligned}&\frac{E\left[u(a^*(h_R, D), Z) - u(a^*(h_S, D), Z)|h_R, D\right]}{E\left[u(a^*(h_S, D), Z) - u(a^*(m, R), Z)|h_S, D\right]}\\ =&\frac{(0.8)2 + 0.1(0) + 0.1(1) - ((0.8)1 + 0.1(2) + 0.1(0))}{(0.8)2 + 0.1(0) + 0.1(1) - ((0.8)0 + 0.1(2) + 0.1(1))} = \frac{1.7 - 1}{1.7 - 0.3} = \frac{1}{2}\end{aligned}$$

The intuition is that h_S has an advantage over h_R because if the state is Rainy, a Canopy provides a payoff of 1, while if the state is Sunny, a Shelter yields a payoff of zero.

Putting the three coefficients together, we have $1 \times 1 > \frac{1}{2}$. In a pairwise comparison, ideology h_S beats ideology h_R.

Given the symmetry of the models and the data, it is immediate to see that h_W beats h_S and h_R beats h_W. No ideology beats the other two ideologies.

This non-existence because of a cycle bears a superficial connection with the well-known failure of the median voter theorem in the presence of a Condorcet cycle. But the underlying reason is distinct. The failure of the median voter theorem is due to the presence of multiple voters, while in this model there is one voter only. Trivially, the median voter theorem always holds with one voter.

The failure of existence here is because different ideologies have different strengths against different ideologies. For instance, h_S evokes a state that is bad for h_R, but is unable to do the

same against h_W. Section 5 will explore this strategic feature of models in more detail and identify conditions where models win by generating fear. For now, let us look for a condition under which cycles among models cannot occur.

4.2 Sufficient Condition for Existence

We wish to find a condition on the voter's payoff function such that for any dataset and any ideology set, there exists a winning model.

Proposition 6. *Suppose that $a \in \mathbb{R}$, $Z_i \in \mathbb{R}$, and $u(a, Z_{n+1}) = -(a - Z_{n+1})^2$. Then, there exists at least one ideology h^* that is supported in equilibrium.*[20]

Proof. Recall that $E[u(a, Z_{n+1})|(h_A, h_B), D]$ is the voter's expected utility if action a is chosen and the posterior probability distribution is $Pr(Z_{n+1}|(h_A, h_B), D)$.

Take any two elements of H and call them h_1 and h_2. Given D, the quadratic payoff assumption means that the two ideologies lead, respectively, to policies (we drop the subscript $n+1$ from Z_{n+1}):

$$a_1^* = E[Z|D, h_1]$$
$$a_2^* = E[Z|D, h_2]$$

Consider the right-hand side of the inequality in Theorem 1 for any two ideologies h_1 and h_2:

$$\begin{aligned}
& \frac{E\left[u(a_1^*, Z) - u(a_2^*, Z)|D; h_1\right]}{E\left[u(a_2^*, Z) - u(a_1^*, Z)|D; h_2\right]} \\
= & \frac{E\left[(a_1^* - Z)^2 - (a_2^* - Z)^2|D; h_1\right]}{E\left[(a_2^* - Z)^2 - (a_1^* - Z)^2|D; h_2\right]} \\
= & \frac{(a_1^*)^2 - 2(a_1^*)^2 + E\left[Z^2|D; h_1\right] - ((a_2^*)^2 - 2a_2^* a_1^* + E\left[Z^2|D; h_1\right])}{(a_2^*)^2 - 2(a_2^*)^2 + E\left[Z^2|D; h_2\right] - ((a_1^*)^2 - 2a_1^* a_2^* + E\left[Z^2|D; h_2\right])} \\
= & \frac{(a_1^*)^2 - 2a_1^* a_2^* + (a_2^*)^2}{(a_1^*)^2 - 2a_1^* a_2^* + (a_2^*)^2} = 1
\end{aligned}$$

[20] The proposition can be extended to the multidimensional case, where $a \in \mathbb{R}^k$, $Z_i \in \mathbb{R}^k$, and $u(a, Z_{n+1}) = -\sum_{i=1}^k \alpha_i (a^k - Z_{n+1}^k)^2$, where $\alpha_i > 0$ for $i = 1, \ldots, k$.

Theorem 1 now implies that h^* is supported in equilibrium if and only if

$$\Lambda(D|h^*)R(D|h^*) \geq \Lambda(D|h)R(D|h) \quad \text{for all } h \in H.$$

As the ideology set is finite and ideologies can be weakly ranked by the value $\Lambda(D|h)R(D|h)$, there must be at least one ideology h for which the condition above is true. Existence is proven. □

Proposition 6 states that in the quadratic payoff setting there always exists a winning ideology. That holds for any ideology set H and any data D.

The proposition is proven by showing that with a quadratic payoff function the right-hand side of inequality (1) becomes one. Therefore, ideology h^* is supported in equilibrium if and only if

$$\Lambda(D|h^*)R(D|h^*) \geq \Lambda(D|h)R(D|h) \quad \text{for all } h \in H.$$

As the ideology set is finite, there must be an ideology that satisfies this condition.

Under the condition of Proposition 6, the 'fear factor' plays no role. The winning ideology depends on the two other factors only. The intuition as to why the fear factor is neutralized with a quadratic payoff depends on a property that is explored in the next section.

5 Fear

This section focuses on the third term of inequality (1) in Theorem 1. We will show that an ideology has a strategic advantage if it evokes the risk of disastrous outcomes in competing models.

To illustrate the effect, we focus on a simple setting with two ideologies that are identical except that one (the "alarmist ideology") assigns positive probability to an event that has never been observed. This ideology fits the data less well but – if the voter is sufficiently risk-averse – it will always beat the other ideology (the "moderate" one). This setting generalizes the comparison between the simple ideology and the alarmist ideology in the example of Section 2.

The data is composed of n observations of the same realization y_0. There are two competing ideologies:

- The moderate ideology h_{mod} assigns probability 1 to the observed outcome y_0
- The alarmist ideology h_{alarm} assigns probability $1-p$ to the observed outcome y_0 and p to an outcome $\bar{y} = y_0 + h$, where h is a positive number. Outcome $\bar{y}$ has never been observed.

Which of the ideologies models will win?

To answer the question, apply Theorem 1 to this case. As both ideologies have no degrees of freedom (they only have one value of θ), simplicity does not play a role. The alarmist ideology (strictly) beats the moderate ideology if

$$\frac{\Pr(D|h_{\text{alarm}})}{\Pr(D|h_{\text{mod}})} > \frac{E\left[u(a^*(h_{\text{mod}}, D), y) - u(a^*(h_{\text{alarm}}, D), y)|h_{\text{mod}}, D\right]}{E\left[u(a^*(h_{\text{alarm}}, D), y) - u(a^*(h_{\text{mod}}, D), y)|h_{\text{alarm}}, D\right]} \tag{3}$$

If this inequality is satisfied, there is a unique equilibrium where both candidates choose the alarmist ideology. If the inequality is reversed, the moderate ideology wins.

To highlight the role of risk aversion, suppose the utility function is a polynomial of order $\gamma \in [1, 2, 3, ...)$:[21]

$$u(a, y) = -|a - y|^{\gamma}.$$

Proposition 7. *Suppose $\gamma > 2$ ($\gamma < 2$). For any n and any h, if p is low enough, in the only equilibrium both candidates choose the alarmist (moderate) ideology.*

Proof. The left-hand side of (3) is:

$$\frac{\Pr(D|h_{\text{alarm}})}{\Pr(D|h_{\text{mod}})} = (1-p)^n,$$

Thus, for any n, the left-hand side of (3) goes to 1 as $p \to 0^+$.

[21] The result also holds for a continuous γ, but, to keep the proof simple, we focus on integer values.

Without loss of generality, let $y_0 = 0$. The optimal action under the moderate model is zero. Instead the optimal action under the alarmist model is computed as follows

$$\bar{a} = \arg\min_a (1-p)a^\gamma + p(h-a)^\gamma.$$

The first-order condition is

$$(1-p)a^{\gamma-1} = p(h-a)^{\gamma-1},$$

yielding $\bar{a} = k(p)h$ with

$$k(p) = \frac{p^{\frac{1}{\gamma-1}}}{(1-p)^{\frac{1}{\gamma-1}} + p^{\frac{1}{\gamma-1}}}.$$

Note that $\lim_{p\to 0^+} k(p) = 0$.

The right-hand side of (3) is:

$$\begin{aligned}
&\frac{E\left[u(a^*(h_{\text{mod}}, D), y) - u(a^*(h_{\text{alarm}}, D), y) | h_{\text{mod}}, D\right]}{E\left[u(a^*(h_{\text{alarm}}, D), y) - u(a^*(h_{\text{mod}}, D), y) | h_{\text{alarm}}, D\right]} \\
=&\frac{(0) - (-\bar{a}^\gamma)}{((1-p)(-\bar{a}^\gamma) + p(-(h-\bar{a})^\gamma) - (p(-h^\gamma)))} \\
=&\frac{\bar{a}^\gamma}{-(1-p)\bar{a}^\gamma + p(-(h-\bar{a})^\gamma + h^\gamma)} \\
=&\frac{(k(p)h)^\gamma}{-(1-p)(k(p)h)^\gamma + p(-(h-k(p)h)^\gamma + h^\gamma)} \\
=&\frac{k(p)^\gamma}{-(1-p)k(p)^\gamma + p(-(1-k(p))^\gamma + 1)}
\end{aligned}$$

Consider the limit of the reciprocal of the expression above:

$$
\begin{aligned}
&\lim_{p\to 0^+} \frac{-(1-p)k(p)^\gamma + p(-(1-k(p))^\gamma + 1)}{k(p)^\gamma} \\
&= -1 + \lim_{p\to 0^+} \frac{p(-(1-k(p))^\gamma + 1)}{k(p)^\gamma} \\
&= -1 + \lim_{p\to 0^+} \frac{p(-(1-\gamma k(p) + \cdots \pm k(p)^\gamma) + 1)}{k(p)^\gamma} \\
&= -1 + \lim_{p\to 0^+} \frac{p(-(-\gamma k(p) + \cdots \pm k(p)^\gamma))}{k(p)^\gamma} \\
&= -1 + \gamma \lim_{p\to 0^+} \frac{p}{k(p)^{\gamma-1}} \\
&= -1 + \gamma \lim_{p\to 0^+} \frac{p}{\left(\frac{p^{\frac{1}{\gamma-1}}}{(1-p)^{\frac{1}{\gamma-1}} + p^{\frac{1}{\gamma-1}}}\right)^{\gamma-1}} \\
&= -1 + \gamma \lim_{p\to 0^+} \left((1-p)^{\frac{1}{\gamma-1}} + p^{\frac{1}{\gamma-1}}\right)^{\gamma-1} \\
&= -1 + \gamma
\end{aligned}
$$

Thus, in the limit, (3) is satisfied if

$$1 \geq \frac{1}{-1+\gamma},$$

which holds when $\gamma > 2$. □

The intuition for the proposition has to do with the voter's risk aversion. The alarmist ideology evokes the possibility of an unlikely event that deviates from the observed data. To offer protection from this 'catastrophic' event, it picks an action that deviates from the optimal action of the moderate ideology. If the moderate ideology is correct, the cost of the deviation is negligible because the action is almost optimal. If the catastrophic event occurs, the benefit of the deviation is not negligible because the action is far from the optimal one. This fear strategy works if and only the voter is sufficiently risk averse.

Arguably, an example of a successful fear ideology is immigration policy. Available evidence

indicates that immigration to the the United States has mostly had positive social and economic effects and there is no evidence in support of highly negative effects (Abramitsky and Boustan (2022)). However, in his 2016 campaign Donald Trump developed a 'culture of fear' around immigration based on highly negative effects, like a steep surge in crime or a total failure of assimilation (Glassner (2018)). Although these catastrophic effects have not been observed in the past, they are enough to evoke a fear of immigration in some voters and generate a support for anti-immigration policies.

6 The Trade-off between Fit and Simplicity

Theorem 1 shows that the success of an ideology depends on a 'fit term' and a 'simplicity term'. In the discussion after the theorem, we made an intuitive argument that parties may face a tradeoff between these two terms. This section explores this tradeoff in a more analytical manner within the setting of a class of problems that are common in economics: linear regressions with an endogenous set of covariates.

6.1 Set-Up

Often a key ideological choice is the selection of variables that are assumed to determine the outcome under consideration. For instance voters can access almost unbounded information about socio-economic variables. Ideologies often restrict attention to a few key variables. A right-wing populist might explain unemployment with immigration, while her left-wing counterpart might think of fiscal policy. Less populist parties might instead cite multiple causes like technological change and educational gaps. In equilibrium, which set of variables will parties' ideologies highlight?

Within our set-up, we can imagine a situation where the outcome can be explained by a large set of potential covariates, and each ideology corresponds to focusing on a specific subset of covariates.

More specifically, each ideology corresponds to a Bayesian linear regression that includes a subset

J out of k possible covariates. The coefficients of the covariates have a Normal prior distribution. We also allow for the residual variance to be unknown and to have an Inverse-Gamma prior (a standard conjugate prior for Bayesian linear regressions). We now provide formal definitions for the linear regression example:

Data. We assume that the dataset D_n consists of n independent and identically distributed observations of one real-valued outcome variable, y, and also a k-dimensional vector of covariates $x_1, ..., x_k$. In terms of the notation introduced above, the dataset can be denoted as:

$$D_n = \{Z_i\}_{i=1,...,n} \quad \text{where } Z_i \equiv (Y_i, X_i').$$

A true Data Generating Process (DGP), denoted $\mathbb{P}$, determines the joint distribution of the random variables y and x. Neither the candidates nor the voter know $\mathbb{P}$.[22]

To define the ideology set H, we must first define statistical models and prior distributions. Our definitions of statistical models and prior distributions for this regression problem follow the framework of Montiel Olea et al. (2022).

Statistical Models. All available statistical models posit a linear relation between y and the covariates $x \in \mathbb{R}^k$, i.e., conditional on x:

$$y = x'\beta + \epsilon, \quad \text{where} \quad \epsilon | x \sim \mathcal{N}_1(0, \sigma_\varepsilon^2), \quad \beta \in \mathbb{R}^k, \tag{4}$$

that is, they assume that the conditional distribution of y given x is a *homoskedastic linear regression with Gaussian errors* and parameters $(\beta, \sigma_\varepsilon^2) \in \mathbb{R}^k \times \mathbb{R}_+$.[23]

[22] Most results, including Theorem 1, do not refer to the true data generating process. These assumptions are used only in the later part of Section 6 to discuss certain aspects of the tradeoff between fit and complexity.

[23] We use the notation $\mathcal{N}_k(\mu, \Sigma)$ to denote a multivariate normal distribution of dimension k with mean μ and covariance matrix Σ. This notation is standard. See p. 171 of Hogg, Mckean, and Allen (2006) for a textbook reference on this convention.

Agents also assume that the covariates follow a distribution P_Σ which belongs to some parametric family $\{P_\Sigma\}_{\Sigma\in\mathcal{E}}$.[24] This may or may not be the correct distribution. We assume that, for any element in this family, the matrix $E_{P_\Sigma}[xx']$ is positive definite and that the random vector x has finite moments of all orders.[25]

Candidates may consider different explanatory variables in x as relevant for their prediction. We assume that both candidates consider at least one explanatory variable in their models. The following notation will be useful. If $\{1, 2 \ldots, k\}$ label the explanatory variables in x, we denote by $J \subseteq \{1, 2 \ldots, k\}$ the subset that a candidate considers possibly relevant for prediction. For a given vector β, the subvector consisting solely of the components in $J \subseteq \{1, \ldots, k\}$ is denoted by $\beta_J \in \mathbb{R}^{|J|}$. Let x_J be the analogous subvector of x, and X_J the corresponding submatrix of X.

Together, Σ, β_J, and σ^2_ε fully define a joint distribution over y and x with probability density function M_{θ_J} and parameter $\theta_J := (\beta_J, \sigma^2_\varepsilon, \Sigma) \in \Theta_J \equiv \mathbb{R}^{|J|} \times \mathbb{R}_+ \times \mathcal{E}$. In terms of the notation we introduced above, a candidate that only considers the covariates x_J for prediction has a statistical model $\mathcal{M} : \Theta_J \to \mathcal{P}$, where $\mathcal{P}$ is the set of all possible distributions over (y, x'), and $\mathcal{M}(\theta_J) = M_\theta$.

Priors. A model that considers variables J as relevant for prediction is associated with a prior π over Θ_J. It will be convenient to denote by $J(\pi)$ the set of variables that an agent with prior π considers relevant for prediction. Formally, let π_j denote the marginal distribution over β_j corresponding to prior π. If δ_0 denotes a Dirac measure at zero, then

$$J(\pi) := \{j \in 1, \ldots, k : \pi_j(\beta_j) \neq \delta_0\}.$$

In a slight abuse of terminology, we sometimes use $J \subseteq \{1, \ldots, k\}$ to refer to a *model*, which should be understood as the set of explanatory variables that are not exactly equal to zero under the prior

[24] A family of distributions $\mathcal{P}$ is said to be parametric if its elements are indexed by a finite-dimensional vector. One example is $x \sim \mathcal{N}_k(0, \Sigma)$, where Σ is an unknown positive definite matrix.

[25] Positive definiteness of the matrix rules out the case in which one covariate is a linear combination of some of the others.

π. Strictly speaking, though, a statistical model refers to the collection of distributions over data, given parameters as we have defined above.

For the priors, we assume the following standard form.

Definition: Given a statistical model with covariates J, we say that a prior π over $(\beta_J, \sigma_\varepsilon^2)$ has the *Normal-Inverse-Gamma form* with *hyperparameters* (a_0, b_0, γ_0), with $(a_0, b_0, \gamma_0) \gg 0$ and $a_0 > 1$ if:

$$\beta_J | \sigma_\varepsilon^2 \sim \mathcal{N}_{|J|}\left(\mathbf{0}, \frac{\sigma_\varepsilon^2}{\gamma_0 |J|}\mathbb{I}_{|J|}\right), \qquad \sigma_\varepsilon^2 \sim \text{Inv-Gamma}(a_0, b_0),$$

where Inv-Gamma(a_0, b_0) is the Inverse-Gamma distribution with parameters a_0 and b_0.[26]

Two important remarks about the priors used in this example. First, we are assuming that all candidates have common hyperparameters for the residual variance. This means that the main difference between the ideologies is the vector of covariates under consideration. Second, note that we are adjusting the prior variance on the regression coefficients, by the number of covariates used in the statistical model. This normalization is taken from Montiel Olea et al. (2022) and is used to guarantee that the ex-ante prediction errors are the same across different models. This means that, before observing data, no model can be viewed as having a superior predictive performance. The normalization is also helpful to illustrate that the differences in this paper relative to the findings in Montiel Olea et al. (2022) do not arise from differences in the statistical models under consideration and their corresponding priors.

6.2 Analysis

Ideology Set. Having defined models and priors, we are now ready to define the ideology set. Each ideology corresponds to a particular subset J of the k covariates. It can be denoted as $h_J = (\mathcal{M}_J, \beta_J)$, where the model $\mathcal{M}_J$ and the prior β_J were defined above for every subset of

[26]The Inverse-Gamma is a two-parameter family of distributions on the positive real line. For parameters $a_0 > 1, b_0 \geq 0$ it has mean $b_0/(a_0 - 1)$.

covariates J.

The ideology set H is the collection of such ideologies. Note that the ideology set is finite and its cardinality corresponds to the number of elements of the (proper) power set of the k covariates, namely, $2^k - 1$.

Actions, Policies, and Payoff function. There are two equivalent ways to model the actions of the candidate and the payoff of voter.

1. First, we could model the action that the candidate chooses on behalf of the voter as a *prediction function*; i.e., a function $f : \mathbb{R}^k \to \mathbb{R}$ that maps covariates, x, into outcomes. In this case, we would say that $\mathcal{A}$ is the space of all (measurable) functions $f : \mathbb{R}^k \to \mathbb{R}$. The voter's payoff function could then be defined as

$$U(f, z) = -(y - f(x))^2, \quad z \equiv (y, x).$$

   Let $Z_{n+1} \equiv (Y_{n+1}, X_{n+1})$. In this formulation of the problem, there is no need to allow for any additional public signal $X(Z_{n+1})$, so we could set this function to be equal to an arbitrary constant. A policy τ_j for the voter is then equivalent to a choice of prediction function f; and we can simply define τ_j to satisfy $\tau_j(x(z)) = f$ for all $z \in \mathcal{Z}$. Then

$$U(\tau_j(X_j(Z_{n+1})), Z_{n+1}) = (Y_{n+1} - f(X_{n+1}))^2.$$

2. Alternatively, we could model the action that the candidate chooses on behalf of the voter as a *point forecast*; i.e., a scalar $a \in \mathbb{R}$. In this case we would say that $\mathcal{A}$ equals the real line $\mathbb{R}$. The voter's payoff function could then be defined as

$$U(a, z) = (a - y)^2 \quad z \equiv (y, x).$$

If we $Z_{n+1} \equiv (Y_{n+1}, X_{n+1})$ we can then allow the voter to observe $\mathbb{R}^k$-valued public signal $X(Z_{n+1}) = X_{n+1}$. This means that, before choosing a point forecast, the both the candidates and voter observe the covariates associated to Z_{n+1}, but not the outcomes. A policy τ_j for candidate j in this case is a function $\tau_j : \mathbb{R}^k \to \mathbb{R}$. Then

$$U(\tau_j(X_j(Z_{n+1})), Z_{n+1}) = (Y_{n+1} - \tau_j(X_{n+1}))^2.$$

Thus, both formulations are the same.

Candidates can propose linear models that differ in the set of covariates they include. What models will be offered in equilibrium?

We apply Theorem 1 to answer this question. For a statistical model with covariates J and a prior with p.d.f. π we define the *model evidence* at D_n as

$$E(D_n; J, \pi) \equiv \int_{\Theta_J} \Pi_{i=1}^n M_{\theta_J}(y_i, x_i)\pi(\theta_J)d\theta_J. \tag{5}$$

The payoff function we are working with is quadratic prediction error. This simplifies the analysis in two ways. First, we get existence of an equilibrium (through an extension of Proposition 6). Second, the right-hand side of Equation 1 in Theorem 1 equals one. Hence ideology h beats h' if and only if it has a greater evidence, as defined in Equation 5. This means that a model with covariates J' and prior π' is selected over a model with covariates J and prior π in equilibrium if and only if

$$\frac{E(D_n; J', \pi')}{E(D_n; J, \pi)} \geq 1, \tag{6}$$

provided $E(D_n; J, \pi) > 0$.

Throughout the rest of this section, we impose the following assumptions.

Assumption 1. *The prior π over any model $(\beta_J, \sigma^2_\varepsilon)$ is independent of the prior over Σ; that is, $\pi(\theta_J) = \pi(\beta_J, \sigma^2_\varepsilon)\pi(\Sigma)$.*

Assumption 2. *The prior π over any model $(\beta_J, \sigma^2_\varepsilon)$ is of Normal-Inverse-Gamma form with hyperparameters (a_0, b_0, γ_0).*

We first show that under these two assumptions we can characterize the trade-off between fit and simplicity when the dataset contains only one observation of (y, x). In particular, we show that ideologies that use only one covariate have higher model evidence relative to ideologies that use more than one covariate. In light of Equation 6, this means that the only simple ideologies can be supported in equilibrium.

Proposition 8 (Simpler ideologies win when $n = 1$)**.** *Suppose that the Assumptions 1 and 2 hold and that $\gamma_0 \geq 2$. If $n = 1$, and ideology h_J with $|J| > 1$ has at least one covariate different from zero, then there is an ideology $h_{J'}$ with $|J'| = 1$ with higher model evidence than h_J.*

Our result considers the extreme case when the dataset is composed of just one observation, $n = 1$. The main message of the result is that any ideology that uses at least two covariates (and with at least one covariate different from zero) to explain the outcome will have lower model evidence than a simpler ideology that uses only one covariate. The intuition is that model evidence increases with both model fit and model simplicity, and when $n = 1$ any model with two covariates or more has the same fit as a model that excludes one of the covariates. Importantly, our result does not make any assumption about the true data generating process. The proof of this result (given below) uses the closed-form expression available for the ratio of evidence between two models when Assumption 1 and 2 are satisfied (see Lemma 13 in the Appendix)

Proof. Without loss of generality, let J be a statistical model that considers the first k covariates. Thus,

$$X'_J = (x_1, x_2, \ldots, x_k)' \equiv x.$$

Let $\|x\|^2 \equiv (x_1^2 + \ldots + x_k^2)$. Since we have assumed that at least one covariate is different from zero, $\|x\| > 0$. We will show that the model J' that has only the covariate x_j with the largest value of x_j^2 among $(x_1, \ldots, x_k)$ has a strictly larger model evidence.

We first show that

$$\det\left(X_J'X_J + \gamma_0 k\mathbb{I}_k\right) = (\|x\|^2 + \gamma_0 k)(\gamma_0 k)^{k-1}. \tag{7}$$

We do this by showing that the matrix $X_J'X_J + \gamma_0 k\mathbb{I}_k$ has $k-1$ eigenvalues equal to $\gamma_0 k$ and one eigenvalue equal to $(\|x\|^2 + \gamma_0 k)$.

To see this, note that $x/\|x\|$ is an eigenvector of the matrix xx' with eigenvalue $\|x\|^2$. This follows directly from the relation

$$xx'\left(x/\|x\|\right) = x\|x\| = (x/\|x\|)\|x\|^2.$$

The equation above readily implies that $x/\|x\|$ is an eigenvector of the matrix $xx' + \gamma_0 k\mathbb{I}_k$ with eigenvalue $\|x\|^2 + \gamma_0 k$. This follows directly from the relation

$$\left(xx' + \gamma_0 k\mathbb{I}_k\right)\left(x/\|x\|\right) = \left(x/\|x\|\right)\left(\|x\|^2 + \gamma_0 k\right).$$

Note also that any unit-norm vector $z \in \mathbb{R}^k$ such that $x'z = 0$ is an eigenvector of $(xx' + \gamma_0 k\mathbb{I}_k)$ with eigenvalue $\gamma_0 k$.

Let $\Lambda_{n,j}$ be defined as in Lemma 13. Then it follows that

$$\begin{aligned}
\frac{\det\left(\Lambda_{n,J}/|J|\right)}{\det(\Lambda_{n,J'}/|J'|)} &= \frac{(x_1^2 + \ldots + x_k^2 + \gamma_0 k)(\gamma_0 k)^{k-1}k^{-k}}{x_{j'}^2 + \gamma_0} \\
&\geq \frac{1}{k}\frac{(x_{j'}^2 + \gamma_0 k)(\gamma_0)^{k-1}}{x_{j'}^2 + \gamma_0} \\
&= \frac{(x_{j'}^2/k + \gamma_0)}{x_{j'}^2 + \gamma_0}(\gamma_0)^{k-1} \\
&\geq \frac{1}{k}(\gamma_0)^{k-1}.
\end{aligned}$$

where the last line follows from the fact that the function $u \mapsto \frac{u/k+\gamma_0}{u+\gamma_0}$ is decreasing in u (for $u \geq 0$)

and converges to $1/k$ as $u \to \infty$. Note that

$$\begin{aligned} \frac{1}{k}(\gamma_0)^{k-1} > 1 \quad &\iff \quad \gamma_0 > k^{1/(k-1)} \\ &\iff \quad \ln(\gamma_0) > \frac{1}{k-1}\ln(k). \end{aligned}$$

The function $k \mapsto \frac{\ln(k)}{k-1}$ is decreasing in k.[27] Since we have assumed $\gamma_0 \geq 2$, then we conclude that $(\gamma_0)^{k-1}/k \geq 1$ for any $k \geq 2$.

Finally, we provide a simplified expression for $b_{n,J}$. Since $n = 1$, then $Y = y_1 \equiv y$. We can write $b_{n,J}$ and $b_{n,J'}$ as

$$\begin{aligned} b_{n,J} &\equiv b_0 + \frac{y^2}{2}\left(1 - x'(xx' + \gamma_0 k\mathbb{I}_k)^{-1}x\right), \\ b_{n,J'} &\equiv b_0 + \frac{y^2}{2}\left(1 - \frac{x_j^2}{x_j^2 + \gamma_0}\right). \end{aligned}$$

If $y = 0$, both terms are the same and equal to b_0. Thus, we assume that $y \neq 0$. In this case, $b_{n,J} \geq b_{n,J'}$ if and only if

$$x'(xx' + \gamma_0 k\mathbb{I}_k)^{-1}x \leq \frac{x_j^2}{x_j^2 + \gamma_0}.$$

The result is analogous to the proof of Proposition 1 in Montiel Olea, Ortoleva, Pai, Prat (2023) and we reproduce it here for the sake of exposition. Let $z_1, \ldots, z_{k-1}$ denote $k-1$ vectors in $\mathbb{R}^z$ that are orthogonal to x. Using the Gram-Schmidt process we can find vectors $\{v_1, \ldots, v_k\}$ in $\mathbb{R}^d$ such that $v_1 = x/\|x\|$ and for any $j > 1$ we have $v_j'x = 0$, $\|v_j\| = 1$ and $v_j'v_{j'} = 0$, $j' \neq j$. This means that the orthonormal matrix $[v_1, \ldots, v_k]$ can be used to eigendecompose $(x'x + \gamma_0 k\mathbb{I}_k)$ as

$$(xx' + \gamma_0 k\mathbb{I}_k) = \sum_{j=1}^{k} \lambda_j v_j v_j',$$

[27] Its derivative is

$$\frac{1}{k-1}\frac{1}{k} - \frac{1}{(k-1)^2}\ln(k) = \frac{1}{(k-1)^2}\left[1 - \frac{1}{k} - \ln(k)\right],$$

and $1 - (1/k) \leq \ln(k)$ for any k.

where $\lambda_1 = \|x\|^2 + \gamma_0 k > 0$ and $\lambda_j = \gamma_0 k$ for $j > 1$. Thus,

$$x'(xx' + \gamma_0 k \mathbb{I}_k)^{-1} x$$

can be written as

$$x' \left(\sum_{j=1}^{k} \frac{1}{\lambda_j} v_j v_j' \right) x = \frac{1}{\lambda_1} (x' v_1)(v_1' x) = \frac{1}{\lambda_1} \|x\|^2 = \frac{x_1^2 + \ldots + x_k^2}{x_1^2 + \ldots + x_k^2 + \gamma_0 k}.$$

Note that the function $u \mapsto \frac{u}{u+\gamma_0 k}$ is increasing in u. Since $x_1^2 + \ldots + x_k^2 \leq k x_j^2$, then we conclude that

$$x'(xx' + \gamma_0 k \mathbb{I}_k)^{-1} x \leq \frac{x_j^2}{x_j^2 + \gamma_0},$$

and consequently, that $b_{n,J} \geq b_{n,J'}$. The result then follows from Lemma 13.

□

Next, we present an approximation for the relative model evidence between a pair of models, but we assume that the sample size is large.

Proposition 9 (Asymptotic approximation of model evidence when $n \to \infty$)**.** *Suppose that the Assumptions 1 and 2 hold. Suppose further that, under the true distribution $\mathbb{P}$, the random vector $(y, x) \in \mathbb{R} \times \mathbb{R}^k$ has finite moments of all orders and $\mathbb{E}_{\mathbb{P}}[xx']$ is positive definite. Then, for any normal-inverse Gamma priors π and π' and any pair of models J and J',*

$$2 \cdot \ln \left(\frac{E(D_n, J', \pi)}{E(D_n, J, \pi')} \right) = BIC(J, n) - BIC(J', n) + O_{\mathbb{P}}(1), \tag{8}$$

where $BIC(J, n)$ is the Bayesian Information Criterion for the Gaussian linear regression model that uses the covariates with indexes in J:

$$BIC(J, n) \equiv n \ln \left(\widehat{\sigma}_{n,J}^2 \right) + |J| \ln(n), \tag{9}$$

and $\widehat{\sigma}^2_{n,J} \equiv \frac{1}{n}\sum_{i=1}^n (y_i - x'_{i,J}\widehat{\beta}_{OLS\text{-}J})^2$.

Our result says that, when the sample size is large, maximizing model evidence is equivalent to minimizing the Bayes Information Criterion defined in Equation 9.[28]

It is well known that when the sample size is large the log-evidence of a model can be approximated (up to a linear transformation) by the Bayes Information Criterion. The approximation is useful because (9) is explicit about the trade-off between fit and simplicity. For instance, the first term unequivocally rewards model fit: a model that fits the data better generates a smaller estimate of the residual variance (the term $\hat{\sigma}^2_{n,J}$). Thus, the first term in Equation 9 will be smaller for larger models. The second term explictly penalizes model complexity: larger models are penalized by a factor $|J|\log(n)$, where $|J|$ is the number of covariates. In order to minimize the Bayes Information Criterion in (9) a model must trade-off fit and simplicity optimally. The proof of the Proposition uses standard arguments and it is given below.

Proof. Algebra shows that when $\mathbb{E}_{\mathbb{P}}[xx']$ is finite and positive definite then:

$$\begin{aligned}\frac{1}{n}\ln\left(\frac{E(D_n, J', \pi)}{E(D_n, J, \pi')}\right) &= (|J| - |J'|)\,\frac{1}{2}\frac{\ln(n)}{n} \\ &+ \left(\frac{a_0}{n} + \frac{1}{2}\right)\left(\ln\left(\frac{1}{n}b_{n,J}\right) - \ln\left(\frac{1}{n}b_{n,J'}\right)\right) \\ &+ O_{\mathbb{P}}\left(\frac{1}{n}\right).\end{aligned}$$

Let $Y = (y_1, \ldots, y_n) \in \mathbb{R}^{n\times 1}$ and X_J be the $n \times |J|$ matrix that collects the row vector $x'_{i,J}$ for each $i = 1, \ldots, n$. Note that

$$\begin{aligned}\frac{1}{n}b_{n,J} &= \frac{1}{n}b_0 + \frac{1}{2}\left[\frac{1}{n}(Y - X_J\mu_{n,J})'(Y - X_J\mu_{n,J}) + \frac{\gamma_0|J|}{n}\mu'_{n,J}\mu_{n,J}\right] \\ &= \frac{1}{n}b_0 + \frac{1}{2}\left[\arg\min_{\beta\in\mathbb{R}^{|J|}}\left(\frac{1}{n}(Y - X_J\beta)'(Y - X_J\beta) + \frac{\gamma_0|J|}{n}\beta'\beta\right)\right].\end{aligned}$$

[28] Equation (9) is sometimes referred to as the *BIC-cost*; see Equation 5.34 in Section 5.3 of Murphy (2012). It is a linear transformation of the original Bayes Information Criterion introduced in Schwarz (1978). The expression we give in (9) is commonly used in the context of linear regression.

Under the Assumptions of Proposition 9 we have

$$\frac{1}{n}b_{n,J} = \frac{1}{2}\sigma_J^2 + o_{\mathbb{P}}(1), \quad \sigma_{n,J}^2 \equiv \min_{\beta \in \mathbb{R}^{|J|}} \mathbb{E}_{\mathbb{P}}[(y - x'\beta)^2].$$

Let $\widehat{\beta}_{\text{OLS}-J}$ be the ordinary least-squares estimator of β based on the data (Y, X_J). Let

$$\widehat{\sigma}_{n,J}^2 \equiv \frac{1}{n}(Y - X_J\widehat{\beta}_{\text{OLS}-J})'(Y - X_J\widehat{\beta}_{\text{OLS}-J}).$$

Under the Assumptions of the lemma,

$$\frac{1}{2}\widehat{\sigma}_{n,J}^2 = \frac{1}{2}\sigma_J^2 + o_{\mathbb{P}}(1).$$

Therefore, the continuous mapping theorem implies

$$\begin{aligned} 2 \cdot \ln\left(\frac{E(D_n, J', \pi)}{E(D_n, J, \pi')}\right) &= (|J| - |J'|)\ln(n) \\ &+ n\left(\ln\left(\widehat{\sigma}_{n,J}^2\right) - \ln\left(\widehat{\sigma}_{n,J'}^2\right)\right) \\ &+ O_{\mathbb{P}}(1). \end{aligned}$$

□

Implications:

1. Suppose that the candidates have correctly specified models. Then, the probability that the true model is suggested in equilibrium approaches 1 as $n \to \infty$.

2. Suppose that candidates have incorrectly specified models. Suppose also that the true model is a Gaussian linear regression with omitted variables (relative to the k covariates in the random vector X). Then, with probability approaching 1, the model that uses all the available covariates is suggested in equilibrium.

3. The result shows how the tradeoff between fit and simplicity evolves with the accumulation of past data. In the definition of BIC, the number of observations n appears linearly in the first term (fit) and as a logarithm in the second term (simplicity). This means that the relative importance of the second term decreases as n increases. If there is more data, the winning ideology puts relatively more weight on fit than on simplicity. However, this relatively higher weight does not mean that larger models are more likely to be selected in equilibrium. In particular, as mentioned in Point 1 above, when the candidates have correctly specified models, the probability of selecting the true model approaches 1 as $n \to \infty$, even if this model is simpler than its competitor. This stands in contrast with the results in Montiel Olea et al. (2022).

6.3 Numerical Illustration of the Results

In the previous subsection we looked at two extreme cases: when there is one observation only ($n = 1$) and when the number of observations is very large ($n \to \infty$). We focused on these cases because it allows us to provide concrete predictions about the model that will be supported in equilibrium. Here, we present a simple numerical simulation to illustrate these two results and to connect them to one another. The simulation tries to illustrate what happens with different models as the number of past observations grows.

Since this is a numerical simulation, we deviate from the rest of rest of the paper by being explicit about what the true data generating process is. We will use it as a ground truth to determine the frequencies with which differnt ideologies win outside the two extreme studied above.

True Data Generating Process: For the purpose of comparison, we follow closely the set-up in Montiel Olea et al. (2022). Suppose that there are six covariates, $\{x_1, \ldots, x_6\}$, of which only the first five are relevant for prediction in the true linear regression model, i.e., $y = \sum_{j=1}^{5} \beta_j x_j + \epsilon$, and $\epsilon | x \sim \mathcal{N}(0, \sigma^2)$. We assume that each nonzero regression coefficient, β_j, is equal to 1, as is σ^2. Also assume that $x \sim N(0, \mathbb{I}_6)$ under the true data generating process (henceforth, DGP).

Considering all subsets of covariates, there are 63 ideologies, one for each nonempty subset of $\{x_1, \ldots, x_6\}$. By construction, 61 exclude a relevant covariate, one matches the true data generating process, and one has a statistical model of higher dimension compared to the true DGP. For the simulation, we assume that priors for each ideology belong to the *Normal-Inverse Gamma* family described in the previous subsection. We note that, by construction, all ideologies have the same subjective mean-squared prediction error before seeing any data. Specifically, that the priors over the β, and σ^2 have hyperparameters $a_0 = 2$, $b_0 = 10$, and $\gamma_0 = 2$.

Figure 1 plots the frequency (over 1,000 Monte Carlo simulations) that one of the ideologies with k variables is the equilibrium ideology. We let the size of the data to take values $n \in \{1, \ldots, 200\}$.

First, as suggested by Proposition 8, when n is small, an ideology with a low-dimensional statistical model tends to win despite being misspecified. In fact, when $n = 1$, we can see in Figure 1 that the winner is a model with a *single* covariate.

Second, as suggested by Proposition 9 and the well-known properties of the Bayes information criterion, the ideology with covariates $\{x_1, x_2, x_3, x_4, x_5\}$ has the highest model evidence with probability approaching one. Figure 2 reports the posterior probability (averaged over simulations) of this ideology.

What happens when none of the ideologies matches the true data generating process? In the previous simulation, the ideologies under consideration (the linear regression models associated to each nonempty subset of $\{x_1, \ldots, x_6\}$) contained the true data generating process (which corresponds to a linear regression model that excludes x_6 and that has all coefficients equal to one). We now consider a set-up where all ideologies are misspecified. To do so, we maintain the assumption that the true data generating process is a linear regression model, but we consider one that uses all of the six covariates $(x_1, \ldots, x_6)$ plus two additional covariates, $\{x_7, x_8\}$, that are not used by any of 63 ideologies under consideration. Thus, the true data generating process takes the form $y = \sum_{j=1}^{8} \beta_j x_j + \epsilon$, and $\epsilon | x \sim \mathcal{N}(0, \sigma^2)$. We assume that each nonzero regression coefficient, β_j, is equal to 1, as is σ^2. Also assume that $x \sim N(0, \mathbb{I}_8)$. The goal of this simulation is to illustrate that

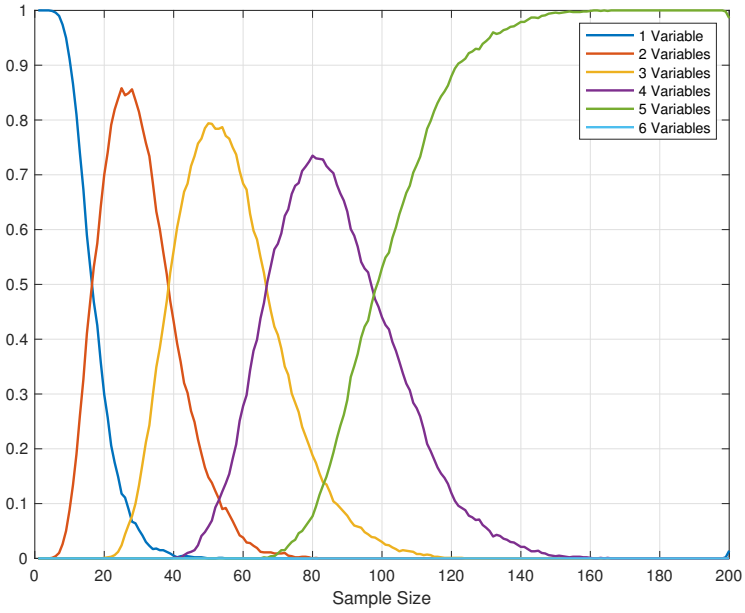


Figure 1: Probability that an ideology with k covariates achieves the highest model evidence.

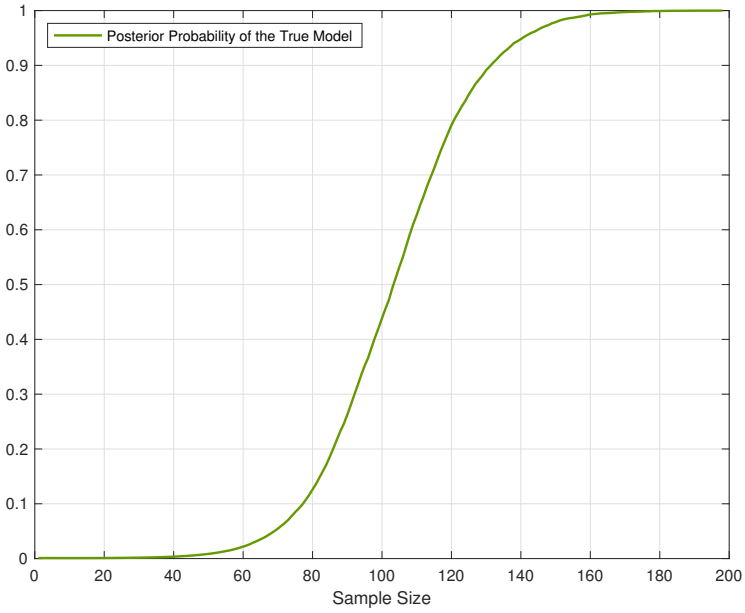


Figure 2: Posterior Probability of the ideology with covariates $\{x_1, x_2, x_3, x_4, x_5\}$ (averaged over 1,000 simulations).

when all ideologies are misspecified, the ideology selected in equilibrium will be the one that best approximates the true DGP. In the case of this simulation, the model that best approximates the true DGP is the model that uses all available covariates.

Figure 3 plots the frequency (over 1,000 Monte Carlo simulations) that an ideology with k variables has the highest model evidence (and is thus suggested in equilibrium). We let the size of the data to take values $n \in \{1, \ldots, 200\}$.

When n is small, an ideology with a low-dimensional statistical model still tends to win despite. However, note that the ideology with largest statistical model tends to be the one with the largest model evidence. Lemma 4 and the well-known properties of the Bayes information criterion can be used to show that the probability that this model has the largest model evidence approaches one as n grows to infinity. Figure 4 presents the posterior probability of the ideology with 6 covariates, averaged over 1,000 simulations.

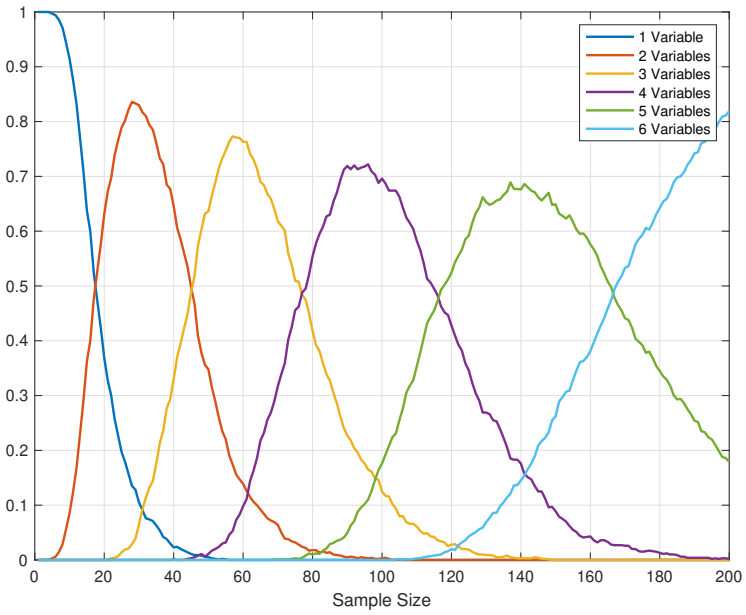


Figure 3: Probability that an ideology with k covariates achieves the highest model evidence. All possible ideologies are misspecified.

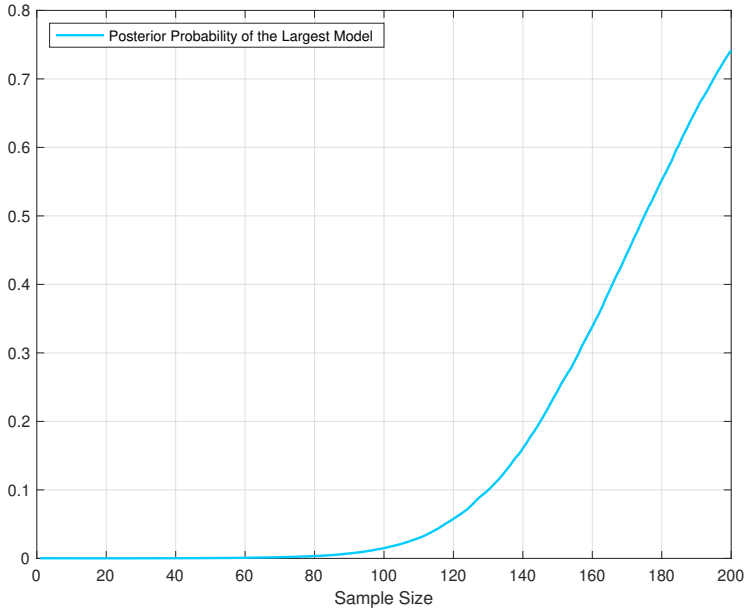


Figure 4: Posterior Probability of the ideology with 6-covariates (averaged over 1,000 simulations).

Remark 2. *As we have argued, the notion of simplicity (or complexity) of an ideology that arises in our set-up connects naturally with the number of parameters that appear in the statistical model. However, it is important to remark that, in general, it is not sufficient to count parameters to understand a model's flexibility (or lack thereof). Examples of this point –along with a definition of the* restrictiveness *of a model in the context of supervised learning problems – are given in the recent work of Fudenberg, Gao, and Liang (2023). A similar observation is part of the literature discussing model complexity from a Bayesian perspective; see, for example, Spiegelhalter, Best, Carlin, and Van Der Linde (2002) and van der Linde (2012). Our somewhat pragmatic view is that there is no universally correct definition of model complexity. Instead, different measures arise naturally (and are useful) in different problems. For instance, the Fundamental Theorem of Statistical Learning (Chapter 6.4 in Shalev-Shwartz and Ben-David (2014)) shows that the right measure of complexity to understand* learnability *in binary classification problems is the Vapnik-Chervonenkis (VC) dimension of the concept (or hypothesis) class generating the outcomes: a concept class is*

learnable if and only if its VC is finite; see Blumer, Ehrenfeucht, Haussler, and Warmuth (1989) and the discussion and references in Montiel Olea and Nesbit (2021). In our context, the comparison of the posterior probability of the ideologies is known to be related to the number of parameters in the model at least since the seminal work of Schwarz (1978).

7 Conclusion

We have represented ideological competition as a game where voters are born without a prior and candidates offer them statistical models to interpret available data. By announcing ideologies, candidates do two things: they commit to policy making rules and they shape the way of thinking of voters. We have considered a general setup where the ideologies set is defined as an abstract collection of statistical models. In our main result, we have characterized the features of ideologies that are selected in equilibrium. These features can be related to ideas regarding statistical fit, model simplicity and the ideology of fear.

In our set-up we have made simplifying assumptions to obtain a sharp characterization of the ideologies selected in equilibrium. For example, we have assumed that all voters are homogeneous. In addition to precluding the possibility of analyzing the polarization of ideologies, it precludes the possibility of zero-sum thinking (Ali, Mihm, and Siga, 2024): namely; that a voter votes against policies supported by others out of fear of what their support could mean for him/her. It would be interesting to extend our model to try to analyze both polarization in the ideologies suggested by the candidates and the implications of voter's zero-sum thinking.

A Additional Derivations

A.1 A characterization of $P(Z|D)$

Lemma 10. *For each $h \in \{h_A, h_B\}$, let*

$$\Pr(\theta_h|D, h)$$

denote the posterior of probability of $\theta_h \in \Theta_h$ if ideology h were known to be true. Let

$$\Pr(\alpha = 1|D; h_v), \quad h_v \equiv (h_A, h_B),$$

denote the posterior probability that the data was generated by ideology h_A given that only ideologies (h_A, h_B) are possible. Then

$$\begin{aligned} \Pr(Z|D; h_v) &= \left[\sum_{\theta_A \in \Theta_A} \Pr(Z|\theta_A; h_A) \Pr(\theta_A|D; h_A)\right] \Pr(\alpha = 1|D; h_v) && (10) \\ &+ \left[\sum_{\theta_B \in \Theta_B} \Pr(Z|\theta_B; h_B) \Pr(\theta_B|D; h_B)\right] \Pr(\alpha = 0|D; h_v). && (11) \end{aligned}$$

Proof. By definition of conditional probability:

$$\Pr(Z|D; h_v) = \Pr(Z|D, \alpha = 1; h_v) \Pr(\alpha = 1|D; h_v) + \Pr(Z|D, \alpha = 0; h_v) \Pr(\alpha = 0|D; h_v).$$

The definition of conditional probability also implies

$$\Pr(Z|D, \alpha = 1; h_v) = \sum_{\theta_A \in \Theta_A} \Pr(Z|D, \theta_A, \alpha = 1; h_v) \Pr(\theta_A|D, \alpha = 1; h_v).$$

Note further that

$$\begin{aligned}
\Pr(Z|D,\theta_A,\alpha=1;h_v) &= \left(\sum_{\theta_B\in\Theta_B}\Pr(Z|D,\theta_B,\theta_A,\alpha=1;h_v)\Pr(\theta_B|D,\theta_A,\alpha=1;h_v)\right)\\
&= \left(\sum_{\theta_B\in\Theta_B}\Pr(Z|\theta_B,\theta_A,\alpha=1;h_v)\Pr(\theta_B|D,\theta_A,\alpha=1;h_v)\right),\\
&\quad \text{(since } Z \text{ is independent of } D \text{ given } \theta_v)\\
&= \left(\sum_{\theta_B\in\Theta_B}\Pr(Z|\theta_A;h_A)\Pr(\theta_B|D,\theta_A,\alpha=1;h_v)\right),\\
&\quad \text{(by definition of the mixture model)}\\
&= \Pr(Z|\theta_A;h_A).
\end{aligned}$$

We now simplify the expression for $\Pr(\theta_A|D,\alpha=1;h_v)$:

$$\begin{aligned}
\Pr(\theta_A|D,\alpha=1;h_v)P(D|\alpha=1;h_v) &= \Pr(D|\theta_A,\alpha=1;h_v)\Pr(\theta_A|\alpha=1;h_v),\\
&= \Pr(D|\theta_A,\alpha=1;h_v)\pi_{h_A}(\theta_A),\\
&\quad \text{(since } \alpha \text{ is independent of } \theta_A \text{ under ideology } h_v),\\
&= \left(\sum_{\theta_B\in\Theta_B}\Pr(D|\theta_A,\theta_B,\alpha=1;h_v)\Pr(\theta_B|\theta_A,\alpha=1;h_v)\right)\pi_{h_A}(\theta_A),\\
&= \left(\sum_{\theta_B\in\Theta_B}\Pr(D|\theta_A;h_A)\Pr(\theta_B|\theta_A,\alpha=1;h_v)\right)\pi_{h_A}(\theta_A),\\
&\quad \text{(by definition of the mixture model)},\\
&= \Pr(D|\theta_A;h_A)\pi_{h_A}(\theta_A).
\end{aligned}$$

This implies

$$\Pr(\theta_A|D,\alpha=1;h_v)=\frac{\Pr(D|\theta_A;h_A)\pi_{h_A}(\theta_A)}{P(D|\alpha=1;h_v)}=\Pr(\theta_A|D;h_A).$$

We conclude that

$$\Pr(Z|D,\alpha=1;h_v)=\sum_{\theta_A\in\Theta_A}\Pr(Z|\theta_A;h_A)\Pr(\theta_A|D;h_A).$$

An analogous argument shows that

$$\Pr(Z|D, \alpha = 0; h_v) = \sum_{\theta_B \in \Theta_B} \Pr(Z|\theta_B; h_B) \Pr(\theta_B|D; h_B),$$

which gives the desired result. □

A.2 Posterior Expected Utility

Lemma 11.

$$\begin{aligned} E[u(a, Z)|D; (h_A, h_B)] &:= \Pr(\alpha = 1|D; (h_A, h_B)) E[u(a, Z)|D; h_A] \\ &+ \Pr(\alpha = 0|D; (h_A, h_B)) E[u(a, Z)|D; h_B]. \end{aligned}$$

Proof. Since

$$E[u(a, Z)|D; (h_A, h_B)] = \sum_{z \in \mathcal{Z}} u(a, z) P(z|D; (h_A, h_B)),$$

Lemma 10 implies

$$\begin{aligned} E[u(a, Z)|D; (h_A, h_B)] &= \sum_{z \in \mathcal{Z}} u(a, z) \left(\left[\sum_{\theta_A \in \Theta_A} \Pr(z|\theta_A, h_A) \Pr(\theta_A|D, h_A) \right] \Pr(\alpha = 1|D; (h_A, h_B)) \right. \\ &+ \left. \left[\sum_{\theta_B \in \Theta_B} \Pr(z|\theta_B, h_B) \Pr(\theta_A|D) \right] \Pr(\alpha = 0|D; (h_A, h_B)) \right). \end{aligned}$$

Collecting terms, and using the definition of $E[u(a, Z)|D; h]$ gives the desired result. □

A.3 Regression Example

Lemma 12 (Evidence Comparison)**.** *Suppose that the priors π and π' satisfy Assumption 1. If the priors over Σ coincide—that is, $\pi(\Sigma) = \pi'(\Sigma)$ for a.e. Σ—and $E(D_n; J, \pi) > 0$, then* (6) *holds if*

and only if

$$\frac{\left(\int_{\mathbb{R}^{|J'|}\times\sigma_\varepsilon^2}\Pi_{i=1}^n\phi(y_i;x_i'\beta_{J'},\sigma_\varepsilon^2)\pi'\left(\beta_{J'},\sigma_\varepsilon^2\right)d\left(\beta_{J'}\times\sigma_\varepsilon^2\right)\right)}{\left(\int_{\mathbb{R}^{|J|}\times\sigma_\varepsilon^2}\Pi_{i=1}^n\phi(y_i;x_i'\beta_J,\sigma_\varepsilon^2)\pi\left(\beta_J,\sigma_\varepsilon^2\right)d\left(\beta_J\times\sigma_\varepsilon^2\right)\right)}\geq 1,$$

where $\phi(\cdot,\mu,\sigma^2)$ denotes the p.d.f of a univariate normal with mean μ and variance σ^2.

Proof. Write $E(D_n;J,\pi)$ as

$$\begin{aligned}
&= \int_{\Theta_J}\Pi_{i=1}^n M_{\theta_J}(y_i|x_i)M_{\theta_J}(x_i)\pi(\theta_J)d\theta_J\\
&= \int_{\Theta_J}\Pi_{i=1}^n \phi(y_i;x_i'\beta_J,\sigma_\varepsilon^2)p_\Sigma(x_i)\pi(\theta_J)d\theta_J\\
&= \int_{\mathcal{E}}\left(\int_{\mathbb{R}^{|J|}\times\sigma_\varepsilon^2}\Pi_{i=1}^n\phi(y_i;x_i'\beta_J,\sigma_\varepsilon^2)\pi\left(\beta_J,\sigma_\varepsilon^2\right)d\left(\beta_J\times\sigma_\varepsilon^2\right)\right)\Pi_{i=1}^n p_\Sigma(x_i)d\Sigma\\
&= \left(\int_{\mathbb{R}^{|J|}\times\sigma_\varepsilon^2}\Pi_{i=1}^n\phi(y_i;x_i'\beta_J,\sigma_\varepsilon^2)\pi\left(\beta_J,\sigma_\varepsilon^2\right)d\left(\beta_J\times\sigma_\varepsilon^2\right)\right)\left(\int_{\mathcal{E}}\Pi_{i=1}^n p_\Sigma(x_i)d\Sigma\right).
\end{aligned}$$

□

Lemma 13 (Evidence Comparison for Conjugate Priors)**.** *Suppose that the Assumptions of Lemma 12 hold. Suppose further that Assumption 2 holds. Then* (6) *holds if and only if*

$$\sqrt{\frac{det\left(\Lambda_{n,J}/|J|\right)}{det(\Lambda_{n,J'}/|J'|)}}\cdot\left(\frac{b_{n,J}}{b_{n,J'}}\right)^{a_n}\geq 1, \tag{12}$$

where

$$\begin{aligned}
a_n &\equiv a_0+\frac{n}{2},\\
b_{n,J} &\equiv b_0+\frac{1}{2}(Y'Y-\mu_{n,J}'\Lambda_{n,J}\mu_{n,J}),\\
\Lambda_{n,J} &\equiv X_J'X_J+\gamma_0|J|\mathbb{I}_{|J|},\\
\mu_{n,J} &\equiv \Lambda_{n,J}^{-1}X_J'Y,\quad Y=(y_1,\ldots,y_n)',
\end{aligned}$$

and X_J is the matrix of dimension $n\times|J|$ that collects the J covariates for the n units in the sample.

Proof. From Lemma 12, it suffices to compute

$$\left(\int_{\mathbb{R}^{|J|} \times \mathbb{R}_+} \Pi_{i=1}^n \phi(y_i; x_i' \beta_J, \sigma_\varepsilon^2) \pi \left(\beta_J, \sigma_\varepsilon^2 \right) d\beta_J \times \sigma_\varepsilon^2 \right),$$

or equivalently

$$\int_{\mathbb{R}_+} \left(\int_{\mathbb{R}^{|J|}} \Pi_{i=1}^n \phi(y_i; x_i' \beta_J, \sigma_\varepsilon^2) \pi \left(\beta_J | \sigma_\varepsilon^2 \right) d\beta_J \right) \pi(\sigma_\varepsilon^2) d\sigma_\epsilon^2.$$

The integrals above represent the "marginal" likelihood in a Gaussian, homoskedastic linear regression model. We present explicit calculations below, although there are well-known expressions for this integral when the priors over $(\beta_J, \sigma_\epsilon^2)$ are conjugate; for example, ADD.

Inner Integral: Let $Y \equiv (y_1, \ldots, y_n)'$ and $X_J \equiv (x_{J,1}', \ldots, x_{J,n}')'$ denote, respectively, the $n \times 1$ vector of outcomes and the $n \times |J|$ matrix containing the covariates in the set J for the n units in the sample. Let $f(Y|X_J; \beta_J, \sigma_\epsilon^2)$ denote the conditional distribution of Y given X_J (for fixed parameters β_J and σ_ϵ^2). Because (y_i, x_i') are i.i.d. across units we have

$$\begin{aligned} f(Y|X_J; \beta_J, \sigma_\epsilon^2) &= \Pi_{i=1}^n \phi(y_i; x_i' \beta_J, \sigma_\varepsilon^2), \\ &= \frac{1}{(2\pi)^{n/2}} \frac{1}{\det(\sigma_\varepsilon^2 \mathbb{I}_n)^{1/2}} \exp\left(-\frac{1}{2\sigma_\epsilon^2} (Y - X_J\beta)'(Y - X_J\beta) \right). \end{aligned}$$

Moreover, under the Normal-Inverse Gamma priors,

$$\beta_J | \sigma_\varepsilon^2 \sim \mathcal{N}_{|J|} \left(\mathbf{0}, \frac{\sigma_\varepsilon^2}{\gamma_0 |J|} \mathbb{I}_{|J|} \right).$$

Therefore:

$$\pi(\beta_J | \sigma_\varepsilon^2) = \frac{1}{(2\pi)^{|J|/2}} \frac{1}{\det\left(\frac{\sigma_\varepsilon^2}{\gamma_0 |J|} \mathbb{I}_{|J|} \right)^{1/2}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2} \beta' (\gamma_0 |J| \mathbb{I}_{|J|}) \beta \right).$$

It is known that under the assumptions of Lemma 13, the posterior distribution of β_J given σ_ε^2

equals

$$\beta_J|Y, X_J, \sigma_\varepsilon^2 \sim \mathcal{N}_{|J|}\left(\mu_{n,J}, \sigma_\epsilon^2 \Lambda_{n,J}^{-1}\right).$$

Consequently,

$$\pi(\beta_J|Y, X_J, \sigma_\varepsilon^2) = \frac{1}{(2\pi)^{|J|/2}} \frac{1}{\det\left(\sigma_\epsilon^2 \Lambda_{n,J}^{-1}\right)^{1/2}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}(\beta - \mu_{n,J})' \Lambda_{n,J} (\beta - \mu_{n,J})\right).$$

Algebra then shows that

$$\begin{aligned} f(Y|X_J; \beta_J, \sigma_\varepsilon^2)\pi(\beta_J|\sigma_\varepsilon^2) &= \frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \frac{\det\left(\sigma_\varepsilon^2 \Lambda_{n,J}^{-1}\right)^{1/2}}{\det\left(\frac{\sigma_\varepsilon^2}{\gamma_0|J|}\mathbb{I}_{|J|}\right)^{1/2}} \exp\left(-\frac{1}{\sigma_\varepsilon^2}(b_{n,J} - b_0)\right) \\ &\cdot \quad \pi(\beta_J|Y, X_J, \sigma_\varepsilon^2). \end{aligned}$$

We conclude that

$$\int_{\mathbb{R}^{|J|}} \Pi_{i=1}^n \phi(y_i; x_i'\beta_J, \sigma_\varepsilon^2) \pi\left(\beta_J|\sigma_\varepsilon^2\right) d\beta_J$$

equals

$$\frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \frac{\det\left(\Lambda_{n,J}^{-1}\right)^{1/2}}{\det\left(\frac{1}{\gamma_0|J|}\mathbb{I}_{|J|}\right)^{1/2}} \exp\left(-\frac{1}{\sigma_\varepsilon^2}(b_{n,J} - b_0)\right). \tag{13}$$

Outer Integral: The prior over σ_ε^2 is an Inverse Gamma with parameters (a_0, b_0). Thus, it has p.d.f. equal to

$$\pi(\sigma_\varepsilon^2) = \frac{b_0^{a_0}}{\Gamma(a_0)} \left(\sigma_\varepsilon^2\right)^{-(a_0+1)} \exp\left(-\frac{b_0}{\sigma_\varepsilon^2}\right).$$

Multiplying this p.d.f by (13) and collecting terms we get

$$\frac{1}{(2\pi)^{n/2}} \frac{\det\left(\Lambda_{n,J}^{-1}\right)^{1/2}}{\det\left(\frac{1}{\gamma_0|J|}\mathbb{I}_{|J|}\right)^{1/2}} \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \cdot \frac{\Gamma(a_n)}{b_{n,J}^{a_n}} \cdot \pi(\sigma_\varepsilon^2|Y, X_J),$$

where $\pi(\sigma_\varepsilon^2|Y, X_J)$ is the p.d.f. of an Inverse Gamma with parameters $(a_n, b_{n,J})$.

We conclude that

$$\int_{\mathbb{R}_+} \left(\int_{\mathbb{R}^{|J|}} \Pi_{i=1}^n \phi(y_i; x_i'\beta_J, \sigma_\varepsilon^2) \pi\left(\beta_J | \sigma_\varepsilon^2\right) d\beta_J \right) \pi(\sigma_\varepsilon^2) d\sigma_\epsilon^2.$$

equals

$$\frac{1}{(2\pi)^{n/2}} \frac{\det\left(\Lambda_{n,J}^{-1}\right)^{1/2}}{\det\left(\frac{1}{\gamma_0|J|}\mathbb{I}_{|J|}\right)^{1/2}} \cdot \frac{b_0^{a_0}}{\Gamma(a_0)} \cdot \frac{\Gamma(a_n)}{b_{n,J}^{a_n}}. \tag{14}$$

The result then follows by noting that

$$\frac{\det\left(\Lambda_{n,J}^{-1}\right)}{\det\left(\frac{1}{\gamma_0|J|}\mathbb{I}_{|J|}\right)} = \frac{1}{\det\left(\Lambda_{n,J}/\gamma_0|J|\right)}.$$

□

Remark 1: We argue that the two terms that appear on the left-hand side of Equation 12 can be interpreted as controlling the effect of "model dimension" and "model fit" on "model evidence". For simplicity, suppose that model J is a larger model than J'. Note first that $\mu_{n,J}$ is the Ridge estimator of β_J. Note then that

$$\begin{aligned} \frac{1}{n} b_{n,J} &= \frac{1}{n} b_0 + \frac{1}{2}\left[\frac{1}{n}(Y - X_J\mu_{n,J})'(Y - X_J\mu_{n,J}) + \frac{\gamma_0|J|}{n}\mu_{n,J}'\mu_{n,J}\right]. \\ &= \frac{1}{n} b_0 + \frac{1}{2}\left[\arg\min_{\beta\in\mathbb{R}^{|J|}}\left(\frac{1}{n}(Y - X_J\beta)'(Y - X_J\beta) + \frac{\gamma_0|J|}{n}\beta'\beta\right)\right]. \end{aligned}$$

We interpret the term in squared brackets as a measure of model fit. In fact, as will be shown later, when $\gamma_0|J|/n$ is large the term in squared brackets approximates the usual frequentist estimate of σ_ε^2, which is smaller for larger models.

However, larger models need not have larger model evidence. The reason is that the term

$$\frac{1}{\det\left(\Lambda_{n,J}/\gamma_0|J|\right)} \tag{15}$$

also depends on the model dimension, and tends to decrease with it. To see this, note that the prior distribution on β_J given σ_ϵ^2 satisfies

$$\beta_J|\sigma_\varepsilon^2 \sim \mathcal{N}_{|J|}\left(\mathbf{0}, \Sigma_{\text{prior}}\right), \quad \Sigma_{\text{prior}} \equiv \frac{\sigma_\varepsilon^2}{\gamma_0|J|}\mathbb{I}_{|J|}.$$

and, under the assumptions of Lemma 2, the posterior distribution of β_J given σ_ε^2 equals

$$\beta_J|Y, X_J, \sigma_\varepsilon^2 \sim \mathcal{N}_{|J|}\left(\mu_{n,J}, \Sigma_{\text{posterior}}\right), \quad \Sigma_{\text{posterior}} \equiv \sigma_\epsilon^2 \Lambda_{n,J}^{-1}.$$

Thus, the term in (15) can be interpreted as

$$\frac{\det\left(\Sigma_{\text{posterior}}\right)}{\det\left(\Sigma_{\text{prior}}\right)}.$$

For illustration, consider the case in which $\Lambda_{n,J}/n \approx \mathbb{I}_{|J|}$ and $\gamma_0 = 1$. Then the ratio of determinants becomes

$$\left(\frac{|J|}{n}\right)^{|J|}.$$

When n is large enough, this term tends to decrease as $|J|$ increases.

B Complete Analysis of the Simple Example

We now present a simple example of our set-up. The goal is to develop intuition on the mechanics of our approach and on the main forces that determine the winning ideology. The rest of the paper will show that this intuition is general.

The data consists of past observations of two states of the world: an outcome variable $Y \in \{-1, 1, 10\}$ and a covariate $X \in \{\text{Green}, \text{Red}\}$. There are six past observations:

i	1	2	3	4	5	6
Y_i	1	-1	1	-1	-1	1
X_i	Green	Red	Red	Green	Red	Green

The data has three features that will matter for the analysis. First, the "extreme outcome" $Y = 10$ has never been observed. Second, the other two outcomes have occurred the same number of times. Third, the outcome Y displays an association with the covariate X: Under Green, $Y = 1$ has appeared 2 times out of 3, while under Red it has appeared only 1 time out of 3.

The election is held at the beginning of time 7. After the election, the winner will observe the realization of the covariate X_7 and will select an action $a \in \Re$. Then Y_7 will be realized, and voter's payoff will be:

$$u(a, Y_7) = -|a - Y_7|^3.$$

The choice of a cubic loss function is meant to capture the voter's aversion to extreme deviations from his policy bliss point (the role of curvature will be investigated in Section 5).

There are three ideologies:

(H_S) *The Simple Ideology*: The two "moderate" outcome realizations have probability 50% each and they do not depend on the covariate.
Namely, $\Pr(Y = 1) = \Pr(Y = -1) = \frac{1}{2}$ and $\Pr(Y = 10) = 0$.

(H_C) *The Complex Ideology*: The probabilities of the two "moderate" outcome realizations may depend on the covariates.
Namely, $\Pr(Y = 1|\text{Green}) = \theta_G$, $\Pr(Y = 1|\text{Red}) = \theta_R$, and $\Pr(Y = 10) = 0$, with θ_G and θ_R independently distributed as uniforms on $[0, 1]$.

(H_A) *The Alarmist Ideology*: It is the same as the simple idology excepts that it assigns a non-zero

probability to the extreme outcome that has never been observed.[29]

Namely, $\Pr(Y = 1) = \Pr(Y = -1) = \frac{1}{2} - 0.01$ and $\Pr(Y = 10) = 0.02$

First, note that the three ideologies can be ranked by how well they can explain the data. The maximum likelihood of the six observations within the simple ideology is just:

$$\Lambda_S = \left(\frac{1}{2}\right)^6$$

The maximum likelihood for the alarmist ideology is smaller because it assigns a lower probability to the events that have been observed:

$$\Lambda_A = \left(\frac{49}{50}\right)^6$$

The maximum likelihood for the complex ideology, which obtains when $\theta_G = \frac{2}{3}$ and $\theta_G = \frac{1}{3}$, is instead greater because it exploits the empirical correlation between the covariate and the outcome to improve explanatory power:

$$\Lambda_C = \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^2$$

If voters used Λ as criterion to rate ideologies, the ranking would be: $H_C \succ H_S \succ H_A$. Under these ranking, office-seeking candidates would both announce the complex ideology.[30]

However, we now show that in our setting the voter ranks the ideologies in the exact opposite order. We do that in two steps.[31]

Does H_S beat H_C?

Suppose candidate A announces the simple ideology while B announces the complex ideology. Who will the voter elect?

[29] For the sake of completeness, assume that all three ideologies assign equal probability to the two realizations of the covariates. Namely, $\Pr(X_i = \text{Green}) = \Pr(X_i = \text{Red}) = 0.5$.

[30] The same ordering would hold if instead of the likelihood we defined explanatory power as least squated residuals.

[31] Here, we present an intuitive summary of the results. The full calculations are available in Appendix ?.

If the voter elects A, policy will be chosen according to H_S. It is immediate to see that the optimal policy in that case will be the midpoint between $Y = -1$ and $Y = 1$, namely $a_S^* = 0$. The voter's payoff is always $U_S = -1$.

If the voter elects B, the policy will be tailored to the realization of the covariate, slightly positive under Green and slightly negative under Red:

$$a_C^*(G) = 5 - 2\sqrt{6} \approx 0.101 \text{ and } a_R^*(G) = -5 + 2\sqrt{6} \approx -0.101$$

This tailoring improves the voter's payoff if the complex ideology is correct, but it actually worsens it – by adding noise to the policy – if the simple ideology is correct.

We assume that the voter assigns equal prior probability to the two ideologies. However, the posterior probability depends on the data too. By Bayes' theorem, the posterior probability of H_C being correct, given that the candidates have announced H_S and H_C is given by:

$$P_{CS} = \frac{\Pr[D|H_C]}{\Pr[D|H_S] + \Pr[D|H_C]},$$

where $\Pr[D|H]$ is the (prior) likelihood of observing data D within ideology H. We have:

$$\begin{aligned}\Pr[D|H_S] =& K\left(\frac{1}{2}\right)^6 = \frac{1}{64}\\ \Pr[D|H_C] =& K\int_0^1\int_0^1 \theta_G^2(1-\theta_G)\theta_R(1-\theta_R)^2 d\theta_R d\theta_G = \frac{1}{144},\end{aligned}$$

where K is a combinatorial constant. The posterior probability of the complex ideology is therefore:

$$P_{CS} = \frac{64}{64+144} = \frac{4}{13} \approx 0.307.$$

In other words, after observing that data, the voter is more skeptical of the complex ideology. That is a manifestation of the Bayesian Occam's Razor, that we will explore in more depth after

the theorem. Given the two additional degrees of freedom (θ_G and θ_R), the complex technology has al large potential to explain observations. Here, H_C underperforms because the association between outcome and color is relatively weak.[32]

We are now ready to compute the voter's expected payoff if he elects the candidate who offers the complex ideology. The payoff depends on which ideology is correct. Given the symmetry of the problem, assume without loss of generality that $X_7 = \text{Green}$. The expected payoff is:

$$U_C = P_{CS} \overbrace{\left(\frac{2}{5}(-(1+a_C^*(G))^3) + \frac{3}{5}(-(1-a_C^*(G))^3)\right)}^{\text{[*] payoff if } H_C \text{ is true } \approx -0.97} - (1-P_{CS}) \overbrace{\left(\frac{1}{2}(-(1+a_C^*(G))^3) + \frac{1}{2}(-(1-a_C^*(G))^3)\right)}^{\text{[**] payoff if } H_S \text{ is true } \approx -1.03} \approx -1.011$$

Recall that the expected payoff under the simple ideology is $U_S = -1$. Expression [*] is greater than -1 because that's when H_C "is right". Instead [**] is smaller than -1, because that's when H_C "is wrong". Ultimately [**] prevails because P_{CS} is well below one half. The simple ideology beats the complex ideology: $U_S > U_C$.

Intuitively, the voter is defending himself against the risk of overfitting. The complex ideology suggests additional patterns in the data. If the ideology is wrong, the patterns are spurious and incorporating them in the policy will lower the voter's expected payoff. Although by design the complex ideology fits the data better, the additional fit does not necessarily benefit the voter. If the posterior probability that the complex ideology is correct is low – as it is in this case – the voter will reject it in favor of the simple one.

This means that there cannot be an equilibrium where both candidates announce H_C or an equilibrium where one announces H_C and the other one announces H_S (recall that if both candidates announce the same ideology they each win with probability 50%). In both those cases, at least one

[32]For instance, if all three 1 outcomes were associated with Green and all three -1 outcomes were associated with Red, the posterior probability of the complex ideology would be much higher: $P_{CS} = \frac{4}{5}$.

candidate would have a strict incentive to switch to H_S.

B.0.1 Does H_S beat H_A?

Now suppose instead that candidate A announces the simple ideology and candidate B announces the alarmist ideology. Who will the voter select?

As in the previous subsection, the policy chosen according to H_S is still $a_S^* = 0$. Instead, the policy chosen under H_A is greater than zero because of the small probability that $Y = 10$. Standard computations show that $a_A^* = 59 - 7\sqrt{(}69) \approx 0.853$.

When H_A and H_S are announced the posterior probabilty of the former is

$$P_{AS} = \frac{49^6}{50^6 + 49^6} \approx 0.469.$$

The posterior probability of H_A is slightly below 50% because H_A predicts an outcome, $Y = 10$, that is never observed.

We can then compute the voter's expected payoff under the two ideologies:

$$\tilde{U}_A = \left(P_{AS} \overbrace{\left(-\frac{49}{100}(1 + a_A^*)^3 - \frac{49}{100}(1 - a_A^*)^3 - \frac{2}{100}(10 - a_A^*)^3 \right)}^{[\#1] \approx -18.4} \right.$$
$$\left. + (1 - P_{AS}) \overbrace{\left(-\frac{1}{2}(1 + a_A^*)^3 - \frac{1}{2}(1 - a_A^*)^3 \right)}^{[\#2] \approx -3.1} \right) \approx -10.34. \tag{16}$$

$$\tilde{U}_S = \left(P_{AS} \overbrace{\left(-\frac{49}{100} - \frac{49}{100} - \frac{2}{100}10^3 \right)}^{[\#3] \approx -21.0} + (1 - P_{AS}) \overbrace{\left(-\frac{1}{2} - \frac{1}{2} \right)}^{[\#4] = -1} \right) \approx -10.38. \tag{17}$$

There are two effects. On one hand, note that the payoff loss if H_S is chosen and H_A rather than H_S is true ($[\#3] - [\#1]$, approximately 2.6) is greater than the payoff loss if H_A is chosen and H_S rather

than H_A is true ([#4] − [#2], approximately 2.1). Choosing the simple ideology when the alarmist one is correct is more costly than the converse. In other words, the alarmist ideology evokes a risk the simple ideology does not protect the voter against.

On the other hand, the alarmist ideology is less likely to be true ($P_{AS} < 0.5$)): that effect militates in favor of H_S, but – in this example – it is not sufficient to offset the risk effect above.

B.0.2 Summing Up

The analysis above has shown that H_A beats H_S and H_S beats H_C – the opposite ordering compared to the a ranking based on explanatory power. This is because there are two additional factors to be considered. In the comparison between H_C and H_S, the Bayesian Occam's Razor favored the simpler model. In the comparison between H_A and H_S, the alarmist ideology won by "making it expensive to trust" the other model.

(To find the equilibrium of the game, we should also ask whether the alarmist ideology beats the complex one.[33] We omit that comparison, but one can show that if A announces H_A and B announces H_C, then A wins. The three comparisons taken together show that the only equilibrium of the game is when both candidates announce the alarmist ideology.)

The example has introduced two effects that play alongside the explanatory power effect to determine the winning ideology. Section 3 shows that these three effects are general and it will derive a simple formula to determine the winning ideology.

References

Abramitsky, R. and L. Boustan (2022): *Streets of Gold: America's Untold Story of Immigrant Success*, PublicAffairs.

[33]This third comparison is not superfluous. Section 4 shows that, except under restrictive conditions, 'beats' is not a transitive relation in our set-up. Cycles are possible.

Aina, C. (2024): "Tailored Stories," *Working paper*.

Ali, S. N., M. Mihm, and L. Siga (2024): "The political economy of zero-sum thinking," *arXiv preprint arXiv:2409.15946*.

Blumer, A., A. Ehrenfeucht, D. Haussler, and M. K. Warmuth (1989): "Learnability and the Vapnik-Chervonenkis dimension," *Journal of the ACM (JACM)*, 36, 929–965.

Chalfin, A. and J. McCrary (2017): "Criminal Deterrence: A Review of the Literature," *Journal of Economic Literature*, 137, 5–48.

Doherty, D., C. M. Dowling, and M. G. Miller (2016): "When Is Changing Policy Positions Costly for Politicians? Experimental Evidence," *Political Behavior*, 38, 455–484.

Eliaz, K., S. Galperti, and R. Spiegler (forthcoming): "False Narratives and Political Mobilization," *Journal of European Economic Association*.

Eliaz, K. and R. Spiegler (2020): "A Model of Competing Narratives," *American Economic Review*, 110, 3786–3816.

Freeden, M. (2003): *Ideology: A very short introduction*, Oxford University Press.

Fudenberg, D., W. Gao, and A. Liang (2023): "How flexible is that functional form? Quantifying the restrictiveness of theories," *Review of Economics and Statistics*, 1–50.

Giné, E. and R. Nickl (2015): *Mathematical foundations of infinite-dimensional statistical models*, vol. 40, Cambridge university press.

Glassner, B. (2018): *The Culture of Fear: Why Americans Are Afraid of the Wrong Things*, Basic Books.

Herman, E. S. and N. Chomsky (1988): *Manufacturing Consent: The Political Economy of the Mass Media*, Pantheon Books.

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

Hogg, R. V., J. W. Mckean, and C. D. Allen (2006): *Introduction To Mathematical Statistics*, Pearson Education India.

Izzo, F., G. Martin, and S. Callander (2023): "Ideological Competition," *American Journal of Political Science*, 67, 697–700.

Kartik, N. and R. P. McAfee (2007): "Signaling Character in Electoral Competition," *American Economic Review*, 97, 852–870.

Kemp, C., A. Perfors, and J. B. Tenenbaum (2007): "Learning overhypotheses with hierarchical Bayesian models," *Developmental Science*, 10, 307–321.

Lakoff, G. (2004): *Don't Think of an Elephant! Know Your Values and Frame the Debate*, Chelsea Green Publishing.

Levy, G., R. Razin, and A. Young (2022): "Misspecified Politics and the Recurrence of Populism," *American Economic Journal*, 112, 928–962.

Luntz, F. (2007): *Words That Work: It's Not What You Say, It's What People Hear*, Hachette Books.

Montiel Olea, J. L. and J. Nesbit (2021): "(Machine) learning parameter regions," *Journal of Econometrics*, 222, 716–744.

Montiel Olea, J. L., P. Ortoleva, M. Pai, and A. Prat (2022): "Competing Models," *Quarterly Journal of Economics*, 137, 2419–2457.

Murphy, K. P. (2012): *Machine Learning: A Bayesian Perspective*, MIT Press.

Pascal, B. (1670): *Pensées*, Port-Royal.

Persson, T. and G. Tabellini (2002): *Political Economics*, MIT Press.

Popper, K. (1959): *The Logic of Scientific Discovery*, Hutchinson.

Schwartzstein, J. R. and A. Sunderam (2021): "Using Models to Persuade," *American Economic Journal*, 111, 697–700.

——— (2024): "Sharing Models to Interpret Data," *Working paper*.

Schwarz, G. (1978): "Estimating the dimension of a model," *The annals of statistics*, 461–464.

Shalev-Shwartz, S. and S. Ben-David (2014): *Understanding machine learning: From theory to algorithms*, Cambridge university press.

Shiller, R. J. (2019): *Narrative Economics*, Princeton University Press.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002): "Bayesian measures of model complexity and fit," *Journal of the royal statistical society: Series b (statistical methodology)*, 64, 583–639.

Stantcheva, S. (2021): "Understanding Tax Policy: How do People Reason?" *Quarterly Journal of Economics*, 136, 2309–2369.

Szeidl, A. and F. Szűcs (2022): *The political economy of alternative realities*, Centre for Economic Policy Research.

Titterington, D. M., A. F. Smith, and U. E. Makov (1985): "Statistical analysis of finite mixture distributions," *(No Title)*.

Van der Linde, A. (2012): "A Bayesian View of Model Complexity," *Statistica Neerlandica*, 66, 253–271.

van der Linde, A. (2012): "A Bayesian view of model complexity," *Statistica Neerlandica*, 66, 253–271.

Vavreck, L. (2009): *The Message Matters: The Economy and Presidential Campaigns*, Princeton University Press.