Moving Beyond the Collapsed Campaign: Tools and Techniques for Studying Non-Normal, Heterogeneous Dynamics

by

Bryce Edward Corrigan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Political Science)
in The University of Michigan
2013

Doctoral Committee:

Associate Professor Ted Brader, Co-Chair
Professor John Jackson, Co-Chair
Associate Professor Edward Ionides
Professor Walter Mebane

For my grandparents, Annette and Robert Burke, and Margery and Tom Corrigan, who have inspired me (and my family) with their service to family and humanity, their emphasis on science and truth, and the certain streak of maverick in our personalities that I hope never to lose.

ACKNOWLEDGEMENTS

I would like to thank my committee, Ted Brader, Walter Mebane, John Jackson, and Edward Ionides, for helping me to understand how both my methodological and substantive work can be an important scholarly contribution. Ted was an important mentor and collaborator throughout my graduate school years, giving me advice on every aspect of my scholarly work and the dissertation and publication processes. Walter always made himself available, was always a friend when I needed a friend, and unflinchingly assumed a mentoring role when my work turned in a methodological direction. Time and again I rely on his understanding of both technical and professional opportunities. John has been a valuable colleague as I learned to teach methodology courses, and did me many small and large favors by telling me point-blank when something I claimed did not make sense. Edward provided essential advice about an oversight in the empirical model specification as well as the path to publication in statistics.

My family remained a rock-solid foundation, ensuring that I could always choose normalcy (though I never did) and that I could always retreat to a safe space (though I had to promise more frequent visits after completion). My stepfather John Surber and stepmother Lizette Lewis, and my sisters Erin and Kelsey Corrigan, have been continual sources of love and encouragement. My uncle John and aunt Angela Corrigan and their kids Tim and Abby welcomed me into their home and helped me to make aspirations each year to make progress. My mother Roberta, an academic, has

been a stalwart supporter of the principles of basic research, quality teaching, and academic freedom throughout my life, and inspires me to believe in these things and to try my best to set these examples for others. My father Edward, an entrepreneur (and many other things), continually reminded me of who I am and want to be, and through his example he demonstrates the vast opportunities to successfully invent and build institutions through hard work and perserverance. My brother Zach believes in me and be proud of me, but I think we shall have a competition in this regard, for it is his work that has inspired my desire to work towards the betterment of the world around me.

Most of all, I would like to thank my partner of many years, Adam Seth Levine. His tireless counsel was instrumental to my ability to navigate the many intricacies of graduate school, the dissertation process, and my first job. He accepts me for the challenges that I put in front of myself, which made it possible to do so again and again without giving up on the basic and big-problems research that I hold dear. He always focuses my attention on the many small achievements and this brightens my darkest days. Even at the moments where everything seemed overwhelming, he stood with me for who I am and believed in me.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

CHAPTER I

Introduction

Every four years a juggernaut of journalists and insider commentariat pore over the vast and ever-expanding elements of the modern Presidential campaign. Responding to the breathless cues of their media hosts (and supplied with the latest polls sponsored by their employers), pundits describe to viewers precisely how each candidate appearance or ad buy, prime-time convention speech, economic news release, or debate sophistry can be expected to decisively shift the course of that year's competition. Scholars who study campaigns and elections typically have a strongly counterveiling message—that voters are minimally influenced by campaigns—the so-called minimal effects perspective. In brief, political science research has suggested that people instead have an easy route to their voting decisions via longstanding partisan inclinations and widely reported economic news. Insofar as they are effectual, campaign dynamics may reflect the shifts of voters learning more details that effectively enrich their heuristic starting point. This perspective is not at all shocking, but the premise of this dissertation is roughly, the "devil is in the details." If bringing people into line with their predispositions only comes as the result of punctuated episodes that accord with some possible framing of candidate roles or capabilities, or that prime people to emphasize some considerations over another, and this can

be electorally effectual, then understanding these episodes and their dynamics offers important information above and beyond predictive heuristics.

Beyond the Collapsed Campaign: Understanding Heterogeneity in Campaign Effects

To identify the limited role of campaigns in voter attitudes and decision-making, studies tend to use one of two major approaches—measuring the impact of cumulative measures of exposure to certain types of campaign activities across campaigns and/or voters (e.g. Shaw, 2006) or tracing the over-time dynamics of responses to campaigns (e.g. Wlezien & Erikson, 2012). To study cumulative campaign effects one can examine the total change in expressed voting preference, or in other survey measures of attitudes or beliefs, between timepoints over the course of a campaign. Variation across individuals can furthermore be attributed to differences in the amounts of localized intensity of campaigns competing to activate voters and present candidates in a maximally favorable light (Holbrook & McClurg, 2005). Furthermore, individual characteristics may influence how much new information is received and whether processing leads to resistance or persuasion (Zaller, 1992). The category of campaign advertising effects provides a useful example of the fruits of these past efforts (see, e.g., Corrigan & Brader, 2010). For example, after accounting for the confounding effects of other campaign efforts that are most intense in battleground states, Huber & Arceneaux (2012) estimate that ad purchases of roughly 1500 gross ratings points (GRPs) increased candidate support during the 2000 election campaign by roughly 9 points on average (with a confidence interval ranging from zero to twice this effect). This accords with Shaw's (1999a) estimate from studies of statewide advertising totals during the Bush and Clinton years attributing roughly 2 to 3 points to candidate

support for each 500 GRPs, controlling for other electioneering activities.

On the other hand, to trace the over-time dynamics of the public's responses to political campaigns, scholars have generally examined aggregate measures of opinion or conformity of candidate preferences to expectations based on economic conditions or prior predispositions. Many of these studies eshew claims of direct partisan effects from electioneering efforts, but instead identify the apparent second-order process of citizens sorting themselves. For example, Gelman & King (1993) find that a set of important considerations such as party and ideology yield a prediction of electoral outcomes towards which trial heat polls appear to converge during the final months of the campaign. Wlezien & Erikson (2012) find that polling aggregates and economic indicators suddenly increase quickly in their accuracy as forecasts of election results during the final one hundred days of U.S. Presidential campaigns. Aggregate studies have focused on the persistence of effects too, generally concurring that shifts in public opinion during campaigns have very short-term effects (but c.f. Box-Steffensmeier & Darmofal, 2009). Finally, using a randomized field experiment, Gerber et al. (2011) finds little evidence that advertising effects in a Gubernatorial race persisted for longer than one week. While by definition such transient campaign effects do not affect the equilibrium of opinion, they could be critical when they occur in close proximity to an election.

Although much past research emphasizes the individual heterogeneity in responses to campaigns, as well as important dynamic phenomena, these studies typically collapse one dimension to focus on the other. Though accounts of cumulative effects are able to explain the extent to which various features of campaigns (e.g. campaign intensity, appearances, advertisements) appear to move citizens, and even attempts to distinguish psychological processes like learning, priming, or persuasion,

these studies have little to say about the dynamic evolution of such processes. On the other hand, research focusing on over-time dynamics tends to use only aggregates, revealing little about the time-course whereby campaigns are influential. For example, do relatively less, medium, or well-informed people respond in a more permanent fashion as we might expect from theories of persuasion and survey response (see Chapter 2)? To date, no published work has been brought to bear on this question.

An overriding theme of the present research is that studies of campaigns try to move beyond such "collapsed campaign" studies to better integrate important individual and temporal heterogeneity and dynamics. I now discuss a set of more specific, related goals.

Accounting systematically for influential events

Some campaign activities or events are much more salient than others, a factor evident to anyone who remembers Bush's Willie Horton ad in 1988 or his failure to recall the price of milk in 1992 (Holbrook, 1996). Scholars can point to many particular instances of surprising or widely-discussed economic news, advertisements, and candidate statements just like these, relying on their careful attention to campaigns or documentary evidence. In some cases, a cursory examination of the data may yield a structural break in opinion at a specific moment in time that seems attributable to such an event, making it worthy of considering whether these events are quite different from, say, everyday reminders of party positions in media reports, or failed attempts to scandalize a candidate. A recent example includes the brief buoyancy in support for President Obama after the death of Osama bin Laden in 2011. From the appearance of large shifts, we can conclude that many voters formed a lasting impression of these events, which influenced their willingness to support the

President. However, many other notable events during Obama's campaign and first term—the collapse of financial markets in 2008, the passage of healthcare reform in 2010, the 2011 debt crisis—are much more difficult to identify by eye in polling data.

Past research on campaign effects often recognizes critical campaign events and uses them to structure an interpretation of the campaign. However, an unsatisfactory aspect of this strategy is the *ad hoc* method in which such key episodes or structural breaks are recognized. This is not to suggest that researchers who formulate a conjecture about a turning-point based on a pilot analysis or media reports and other sources, and explicitly build their analysis around this assumption, are doing something wrong. Nonetheless, a more systematic understanding of campaigns could be achieved if every event at some attainable granularity can be modeled, and its significance reflected empirically through the model.

Identifying and distinguishing permanent and transient responses

One difficulty with accounting for influential events is that shifts in opinion are not always immediate and permanent. Thus, for example, the decline in support for McCain that was apparent during the collapse of financial markets in 2008 may well be interpretable as the tail end of a transient bounce in support he received due to the Republican National Convention and McCain's selection of Palin. Some information flows, such as that leading to the public repudiation of Congress during the 2011 debt crisis, may appear as gradual trends over an extended period of time, suggesting a slow diffusion of information into the public's awareness.

These short-term dynamics in aggregate responses to campaigns have important implications for understanding both the way in which events come to public attention and the way in which voters process information. The former might be understand-

able in terms of a wider conceptualization of informative, rare events as producing waves that ramp up to a fixed set of new implications for voter decisions, or spikes that ramp up the salience of some bits of information and then gradually pass from the media spotlight (e.g. Shaw, 1999b)—rather than the binary shocks assumed by most scholars. The resulting opinion dynamics can equivalently be modeled in isolation by higher-order dynamics encompassing the velocity and acceleration of public opinion change. Furthermore, at least some part of each innovation can be specified as permanent, as in the popular Beveridge-Nelson decomposition (Beveridge & Nelson, 1981).

When explicating campaign dynamics in the aggregate, we can only indirectly distinguish these (elite, communications) phenomena from related psychological phenomena that have similar implications. A psychological interpretation postulating a mixture of permanent and transient responses to campaign information is that while some part of any given signal may drive formation of impressions or the acquisition of new considerations (or the revision of previous ones), another part may reflect information that simply becomes less available in memory for most voters over time. Some scholars have gone further in interpreting results suggesting a very short timespan to the effectiveness of campaign stimuli as evidence of memory-based processing (Gerber et al., 2011), though this requires the assumption that we can distinguish these processing phenomena from the way in which information is diffused, i.e., how a person comes to learn about and respond to a Gubernatorial ad.

In summary, most recent literature characterizes campaign effects as transient or permanent, but there is clearly reason to believe that both types of dynamics are manifested, at times under different conditions—or even for individuals with different charcteristics, as discussed in the next section. A specific goal of the present work is

to model voters as exhibiting both types of response, and thereby provide a composite portrait of permanent and transient campaign effects.

Characterizing individual and temporal heterogeneity in dynamic responses to campaigns

Past research emphasizes people's responses to particular campaign events and activities insofar as these drive them towards or away from one or the other candidate, or in the minimal effects tradition, in terms of the resulting learning or priming effects. However, just as people differ in the strengths of their prior attitudes, they may also vary in the amount of campaign information they receive and process due to interest and the richness of their existing schemata, or even in the particular processing mechanism at work as alluded to in the past section. One way to validate a rich psychological model of the processing of political stimuli would be to examine the implications of these differences for campaign dynamics as a function of individual characteristics. Based on methodological accounts, it would seem that a true panel with granularity that can detect highly transient dynamics in responses—that is, one we rarely have—might be required to study individual differences in processing (e.g. Brady & Johnston, 2006). However, as long as public opinion data represent a randomized sample in time as well as in the cross-section, a sufficiently large dataset enables us to study heterogeneous dynamics. The issue is primarily methodological in character, as despite the emphasis on multi-level data structures it remains difficult to represent such heterogeneity, and is addressed in the following section.

Campaign shocks have also been found to induce different dynamics at different points in time prior to Presidential elections (Gelman & King, 1993; Wlezien & Erikson, 2002). In particular, the residual variability of survey responses in models of citizens preferences is observed to decrease. A clear reason to expect such differences

is that people's attitudes are crystallizing, and this leads them to be less responsive to new information. A different explanation with experimental support is that people may respond differently to negative information during the pre-decision versus the post-decision period (Krupnikov, 2011).

Another important reason to seek to capture heterogeneous dynamics is in attempting to address the concern raised in the previous subsection about being able to distinguish the dynamics characterizing ebbs and flows of elite communications from aspects of voter information processing. If measures of the types of individuals more likely to exhibit a certain kind of processing (i.e., well-informed and having chronic availability of strong considerations, versus poorly informed and highly primable or persuadable), or the types of information that are likely to reverberate in particular ways (i.e., as originating in different types of messages or events, as in Shaw, 1999b), then we may be better able to distinguish the mechanisms underlying the dynamics.

Representing Heterogeneous Dynamics in a Multi-Level Model

Representing a hybrid cross-sectional and over-time structure—with allowance for individual-level heterogeneity—requires models that differ substantially from classical forms. In this dissertation, I develop two flexible classes of multi-level models—the linear mode-separable model and a generalized variant—along with a Bayesian computation strategy for simulating from the posterior distribution of the parameters in such models. These developments enable practical estimation of a wide variety of substantive specifications, including the empirical analysis of heterogeneous campaign effects presented in Chapter 2. In this section, I discuss the most general aspects of multi-level modeling and classical and Bayesian strategies for inference. I then introduce a more restricted class of models that facilitates the

development of Bayesian computational strategies because it encompasses both efficient quadratic approximations to non-linear models, and an adaptive strategy for Markov Chain Monte Carlo introduced in Chapter 4.

Conditional models and fluent expressions of dynamics

The latter two of the three substantive goals described in the previous section require the ability to specify dynamics in a multi-level model. First, we can imagine that observations of the dependent variable are directly dependent on their past observations, as in a lagged-dependent variable model. Second, we can imagine that a latent variable, typically referred to as a "random effect" in political science, is not defined by the usual assumption of being independent and identically distributed. Instead, the random effect can be stipulated to obey some time-series model, with its conditional distribution at time t depending on its value at time $t-1$. Third, we can imagine that the time-series model obeyed by the random effects has certain parameters, and that these parameters could depend on features of the individual. The combination of these elements proves fatal to basic econometric strategies for time-series modeling, where we seek to transform every model into a lagged-dependent variable model, and apply a single step OLS. Such strategies were never adequate for a dynamic specification even as simple as ARIMA with fixed parameters, which always requires a recursive or non-linear estimation procedure.

Fortunately, the Wold representation theorem tells us that complex, covariance-stationary dynamics can always be specified by some weighted combination of residuals from the infinite past:

$$v_t = \sum_{s=-\infty}^{t} r_{t-s} u_s$$

Here the input random effects u_s are assumed to be independent, and we truncate

the series at some point in the finite past to make it possible to explicitly evaluate v_t.[1] This means that the unnormalized joint density for the case of a linear specification with common response and random effects distributions has the following simple representation as a conditionally-linear model:

$$f\left(y,\,\beta,\,u|\,\theta\right)=\prod_{i=1}^{N}f_{\theta}\left(\left.y_i-x_i'\beta-\sum_{s=s_0}^{t}r_{it-s}\left(\theta\right)u_s\right|\,\theta\right)\prod_{s=s_0}^{s_T}f_{\theta}\left(u_s|\,\theta\right)$$

Here the first term is the product of the unvariate densies of each observation, and the second term is the product of the univariate densities of each innovation u_s.[2] The variables in θ are "hyperparameters" that describe the distribution of the random effects. Note that there is no assumption here that the dynamic parameters or weights from the Wold decomposition are the same in each individual's linear predictor. Incorporating individual heterogeneity is thus only a complication in computing this predictor—we must simply evaluate each unit's contribution to the density in terms of the dynamic model for that individual.

Even in the case of homogeneous dynamics, the problem reduces to that of a complex pattern of cross-classified random effects. Each unit covaries to some extent with every other unit. However, for heterogeneous dynamics, the above unit-specific linear representation appears to be the strongest simplification possible: in other words, various attempts to represent the model in matrix or operator notation seem to yield no conceptual simplification, and typically no computational simplification. Nonetheless, representing the model in this way avoids the need to collapse the data in any way—we already know how to estimate multi-level models with linear predictors!

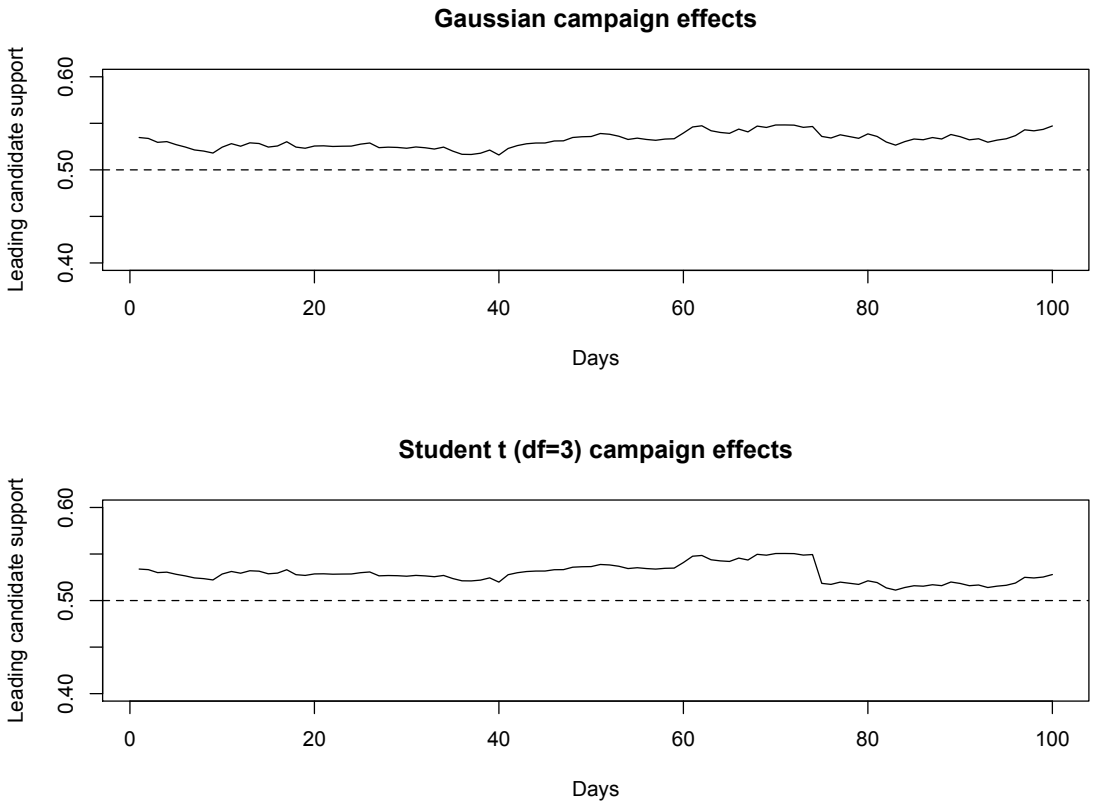
[1] I mostly set aside tricky decisions about initialization; exact initialization will certainly be intractable for the class of models proposed here, but the Bayesian solution I will adopt is to parameterize the initial conditions.

[2] Sometimes these are not the residuals that are of interest for predictions, but this complication is not relevant for estimation.

Non-normality, mode-separability, and tail-boundedness

In order to draw inferences about the parameters, we need to express the likelihood in terms only of the observations. The primary difficulty posed by multi-level models in general is that they stipulate two sets of random quantities, the latent variables and the observed data. Since the latent variables have some distribution, we must form the likelihood by forming the joint density of all random quantities, and integrating this with respect to all random quantities except the observations themselves. In the case of the linear multi-level model with normal errors, the likelihood can be expressed explicitly and inferences can be obtained by maximizing this likelihood.

However, a straightforward way to attain an additional substantive goal described above—accounting for relatively influential events in a systematic way—is to consider alternative assumptions about the distribution of the innovations. Compare for example, the following tableau:



The top series was generated from the inverse normal cumulative distribution function applied to a set of uniform random variables to produce a particular ARIMA time-series based on normal innovations. The second series was generated from the same uniform variables via the inverse Student cumulative distribution function ($df = 3$), yielding the same ARIMA model but with Student-t innovations. The series with Student innovations seems to match the the current consensus among campaign researchers that salient events are rare. Thus, modeling such data using the Student distribution would seemingly allow the data to better "speak for themselves" about the occurrence of important events, rather than imposing this. Of course, the relative frequencies of "important" innovations is precisely determined by the distributional assumption, and we still have to make a decision guided by what constitutes a salient event and how often we think that happens. Nonetheless, the decision to use some

heavy-tailed density whose kurtosis is imposed *a priori* or (with probably much greater difficulty) estimated, seems worthy of consideration as a way to address this form of heterogeneity in campaign dynamics.

I now place two further restrictions on the distributions that can be selected for the response and random effects in such dynamic models. First, I require that the observations and each random effect has a known transformation, to a density that is unimodal[3] in the random effects. These are typically also dependent on the hyperparameters θ such as the log of the random effects standard deviation. As a result, the joint density factors into the product of unimodal functions of the random effects, conditional on the hyperparameters. The joint density of the observations and random effects, and its log, have the following most general form:

$$f\left(y,\,\gamma|\,\theta\right) = C \times \left(\prod_{k=1}^{K} g_{\theta}^{(k)}\left(\gamma\right)\right) \text{ with } g^{(k)}:\,\mathbb{R}^{J} \to \mathbb{R} \text{ unimodal}$$

$$\ln\, f\left(y,\,\gamma|\,\theta\right) = \ln C + \left(\sum_{k=1}^{K} \ln g_{\theta}^{(k)}\left(\gamma\right)\right) \text{ with } g^{(k)}:\,\mathbb{R}^{J} \to \mathbb{R} \text{ unimodal}$$

Here C is a normalizing constant allowing us to disregard whether the component functions $g^{(k)}$ are normalized. When thinking of these densities as functions of the random effects I will refer to these functions as components. Together, any structural parameters β and random effects u upon which the kth component depends are in γ_k. Also, J is defined to be the size of the space of γ.

Second, I require that each of the densities $g_{\theta_k}^{(k)}$ is tail-bounded in γ_k with known $(d_{left},\, d_{right})$. This means that letting γ_0 be the mode of $g_{\theta_k}^{(k)}$ the here exists constants $(c_{left},\, c_{right})$ and $(d_{left},\, d_{right})$ such that $g_{\theta_k}^{(k)}$ is bounded above by $c_{left}/x^{d_{left}}$ for all $\gamma \ll \gamma_0$ and bounded above by $c_{right}/x^{d_{right}}$ for all $\gamma \gg \gamma_0$.

I call this class of models *mode-separable* because of the idea that we can separate

[3]The technical requirement is of linearly-strict quasi-concavity as defined in the Appendix.

the modes of the component, leading to a readily available approximation to the mode of the entire joint density, discussed in Chapter 3 and in the Appendix. A further property that I expect to leverage in the development of an Adaptive Importance Sampler discussed in Chapter 4 and in the Appendix is that any multivariate quasi-concave function can be re-expressed as a product of univariate, unimodal functions of deviations about some point in the space of transformed latent variables.

Non-linear models and blockwise linear models

Naturally, we might suppose that different people are differently susceptible to rare shocks and therefore act as if they are responding to different densities of innovations. A straightforward way to address this is to note that heavy-tailed innovations can be generated by normal-scale mixtures. That means we could construct the joint density as the product of terms specifying normal random effects, conditional upon their log standard deviation, with a set of terms specifying an unconditional model of the log standard deviations. This yields no advantage, but the model no longer has a single linear predictor. By adding additional random effects to the predictors of the variance components, we can generalize the overall model to yield members of the ARCH family of dynamic models for the random effects. A second source of non-linearity (see Wlezien & Erikson, 2012) would be if initial information that drives deviations from prior belief tends to provoke elite counterattacks in a way that limits the volume of new information and drives response dynamics towards an equilibrium. This type of non-linear dynamics can be represented using interactions between two latent state variables—one representing the prior belief and the second representing the innovation. Notably, both of these types of non-linearity can be added to our conditionally linear model as an additional conditionally-linear layer,

or "block" of random effects. The random variables in one block serve as nonlinear hyperparameters in every other block, and this is expected to be useful for Bayesian computation.

Examples of Joint Densities Satisfying these Criteria

I now discuss a set of examples of joint densities exhibiting the characteristics defined above. I express each model in terms of a representation of the log of the joint density as the sum of log-components, plus an additional term $\varphi(\sigma, \theta)$ that involves only the response standard deviation and/or hyperparameters. Evidently, many common distributions—the Gaussian, exponential, Poisson, gamma, and beta distributions—lead to joint log-densities with terms that are unimodal in their location parameter.[4] Further, if the location parameters of such models are instead expressed as a linear combination of linearly dependent variables, then the resulting joint density functions will also be unimodal when they are identified.[5]

For binomial-logistic joint densities, an individual observation has a probability mass function that is concave but not unimodal. However, the product of pairs of binomial-logistic components including at least one success and one failure yields a log-concave and unimodal joint density when the parameters are identified.[6]

Note that the intent in the following discussion of Figures 1.1-1.3 is not to offer toy models, but general forms that encompass many useful applications.

Figure 1.1 gives the Gaussian linear mixed effects model in terms of its components. Since the Gaussian distribution has squared-exponential tails, any power-

[4] As noted above, the point is that these distributions are log-concave in their location parameters—not that they are log-concave distributions (though that is often true too).

[5] There are some technical details about identification of such models outside of the normal case. Binomial-logistic response models must not be separated or quasi-separated.

[6] To detect the cases of separation or quasi-separation in binomial-logistic response models, one can utilize a recently-developed sequential quadratic programming algorithm to detect a lack of identification in advance (see Konis, 2007).

Figure 1.1: Gaussian linear mixed-effects model components

$$\overbrace{-\frac{1}{2\sigma}\left(y - X\beta - R_\gamma u\right)'\left(y - X\beta - R_\gamma u\right) + -\frac{1}{2}u'u}^{g^{(0)}(\beta,\, u),\, y,\, u \sim \text{Gaussian}} + \overbrace{-n\ln\sigma}^{\varphi(\sigma,\,\theta)}$$

Figure 1.2: Gaussian linear model with Student random effects components

$$\overbrace{-\frac{1}{2\sigma}\left(y - X\beta - R_\gamma u\right)'\left(y - X\beta - R_\gamma u\right)}^{g^{(0)}(\beta),\, y \sim \text{Gaussian}} + \sum_1^K \left[\overbrace{-\left(\frac{\nu+1}{2}\right)\ln\left(1 + \frac{u_k^2}{\nu}\right)}^{g^{(k)}(u_k),\, u_k \sim \text{Student's } t_\nu}\right] + \overbrace{-n\ln\sigma}^{\varphi(\sigma,\,\theta)}$$

distributed tail component will dominate.

Figures 1.2 and 1.3 give models combining Gaussian and Binomial-logistic responses with Student-t random effects, respectively. This enables specifications in which random effects are more often smaller than Gaussian random effects, with occasional large shocks that could represent the influence of relatively rare events or structural breaks together with noisy responses. The logistic density has exponential tails which also overwhelm the Student-t tails. Therefore, the random effects components in both cases are dominated by the density, in the tails.

Figures 1.4 and 1.5 give the structure of two models which are not mode-separable, but which are mode-separable conditionally within blocks defined as random effects that enter components specifying the mean and variance, respectively.

Figure 1.3: Logistic linear model with Student random effects components

$$\overbrace{\sum_i \left(y_i\left(-\ln\left(1 + \exp\left(-X\beta - R_\gamma u\right)\right)\right) + \left(1 - y_i\right)\ln\left(X\beta + R_\gamma u\right)\right)}^{g^{(0)}(\beta),\, y \sim \text{Bernoulli}} + \sum_1^K \left[\overbrace{-\left(\frac{\nu+1}{2}\right)\ln\left(1 + \frac{u_k^2}{\nu}\right)}^{g^{(k)}(u_k),\, u_k \sim \text{Student's } t_\nu}\right]$$

Figure 1.4: Gaussian linear model with GARCH-t random effects components

$$\sum_{t=1}^{T}\left[\overbrace{-\frac{1}{2\sigma}\left(y_t - X_t\beta - r_t s_{\mathcal{F}_\gamma^{(t-1)}}\right)^2}^{g_t^{(0)}(\beta,\, r_t),\, y,\sim\text{Gaussian}} + \overbrace{-\left(\frac{\nu+1}{2}\right)\ln\left(1+\frac{u_t^2}{\nu}\right)}^{g^{(1)}(r_t),\, r_t\sim\text{Student's } t_\nu}\right] + \overbrace{-n\ln\sigma}^{\varphi(\sigma,\,\theta)}$$

Figure 1.5: Gaussian linear model w/ crossed Gaussian & GARCH-t re's components

$$\sum_{t=1}^{T}\left[\overbrace{-\frac{1}{2\sigma}\left(y_t - X_t\beta - R_{\gamma t}u + r_t s_{\mathcal{F}_\gamma^{(t-1)}}\right)^2 + -u_t^2}^{g_t^{(0)}(\beta,\, u,\, r_t),\, y,\, u\sim\text{Gaussian}} + \overbrace{-\left(\frac{\nu+1}{2}\right)\ln\left(1+\frac{u_t^2}{\nu}\right)}^{g^{(1)}(r_t),\, r_t\sim\text{Student's } t_\nu}\right] + \overbrace{-n\ln\sigma}^{\varphi(\sigma,\,\theta)}$$

Multi-Level Statistical Inference

Approaches to estimating multi-level models tend to differ radically between the likelihoodist and Bayesian computational approaches. In the likelihoodist approach, the most prevalent approach is to directly evaluate the marginal likelihood by marginalizing (exactly or approximately) over the set of random effects, and then maximizing this true likelihood. This is also sometimes known as the residual maximum likelihood, or REML approach. One could instead treat all of the random effects as true parameters and maximize the conditional density of the response, given the parameters, but then there would be no benefit to specifying random effects having a distribution that better reflects our expectations about the generating process.

Recent work in applied statistics advocates the use of modal approximations to the marginal likelihood, provided that the bulk of mass under the integrand lies near its mode. The most popular approximation for multi-level models seems to be the Laplace approximation. The Laplace approximation is based on directly integrating a Taylor expansion of the likelihood at the mode. The log of the second-order

Laplace approximation has a simple expression in terms of the normal approximation at the mode: it is the sum of half of the joint log-density evaluated at the conditional MLE of the latent variables (the first-order approximation), minus the log of the norm of the conditional MLE plus half the log of the determinant of the observed information matrix (the second-order approximation). When the fixed effects are conditioned upon, this is the loglikelihood in the case of the normal linear model. However, it is conventional now to also integrate away the fixed effects as if they were nuisance parameters, yielding the so-called REML estimates. This further marginalized likelihood is the quantity that is maximized in R's glmer package for generalized linear mixed effects regression (Bates, et al., 2012).

A non-computational Bayesian strategy looks very much like the likelihoodist method: the log prior is added to the loglikelihood and a maximum a posteriori estimate (MAP) is identified, with uncertainty characterized using a local approximation to the Hessian computed just like the observed information matrix. This technique is asymptotically useful due to the Bayesian Central Limit Theorem, which ensures that, fixing the size of the parameter space, any posterior will be increasingly well-approximated by the multivariate normal distribution corresponding to the MAP and its local variance-covariance. For greater accuracy, some form of higher-order Laplace approximation or higher order polar or rectangular cubature can be used to incorporate information from additional points regularly or randomly spaced around the mode of the likelihood or posterior (Rue et al., 2009; Monahan and Genz, 1997; Genz & Kass, 1997). Unfortunately, regardless of whether tractable likelihood, REML, or Bayesian criterion, the variances of the random effects can have multiple modes in multi-level models. In non-normal models, the situation is worse, as the likelihood remains intractable, but there is no guarantee in such models that a sin-

gle or dominant mode of the latent variables exists upon which to base the Laplace approximation.

The computational Bayesian approach is first and foremost the focus of this project because it yields estimates that are non-asymptotic and can potentially deal with multi-modality. As I discuss in Chapter 3, the approach typically relies on a blockwise Gibbs or Metropolis sampling: alternating draws targeting different sets of parameters. There is no need for explicit marginalization—hyperparameters and random effects are all conceptually the same thing. Though there is then no equivalent of selecting a criterion that concentrates out nuisance parameters, one can always impose prior distributions aiming to smooth a problem. The typical computational Bayesian strategy is to design a sampler that samples from a Markov chain converging to the posterior over all parameters simultaneously.

However, despite these appealing features, Bayesian computation is still difficult for practitioners to apply to multi-level models, especially with non-standard choices of distributions. Thus, another goal of this project is to design strategies that facilitate the Bayesian strategy for a wider audience within political science. In particular, recent work focuses on leveraging high-quality approximations to build proposal distributions that avoid local random walks as much as possible, instead moving rapidly about the entire parameter space (e.g. Ardia et al., 2009; Bornkamp, 2011). To advance the project, I sought to automate the process of producing a high-quality approximation based on the model structure. However, using such approximations, an initial naive (but practical) strategy of alternating between draws that converge simultaneously to the conditional posterior of the hyperparameters given the random effects, and vice-versa, placing the sets of parameters on exactly equal footing, led to very poor results. In Chapters 3 and 4, I explain why this is likely to be so, suggest

an improvement based on the Pseudo-Marginal Method, and demonstrate that it outperforms the naive method.

CHAPTER II

Heterogeneous Dynamics of Citizens' Responses to U.S. Presidential Campaigns

A number of recent studies conclude that campaigns have only very limited influences on patterns of individual voting and electoral outcomes (e.g. Gelman & King, 1993; Holbrook, 1996; Wlezien & Erikson, 2002). A potential problem with these studies is their assumption that voters are uniform in their responses to information, collapsing across both time and individual-differences. Such an assumption contradicts a body of observations and contemporary theory related to the mechanisms of opinion formation. In this chapter, I review past work and test a set of new hypotheses regarding individual differences in the dynamics of candidate appraisal. The results are important to our understanding of civic competence, broadly speaking, and specifically the soundness of conventional claims about "October surprises" and other purported threats to democratic outcomes.

Studies of presidential campaigns suggest that candidate efforts to drive public attention towards preferred issues, or persuade the persuadable, have only a small electoral yield. Further, election results are predictable well in advance of the fall campaign period, whereas poll results remain divergent until the end of the campaign (Gelman & King, 1993; Holbrook, 1996). The apparent drift of public preferences

among presidential candidates towards scholars' prior expectations appears to take place roughly during the last one hundred days of the fall campaign (Wlezien & Erikson, 2002).

However, there is little work that directly addresses the mechanism at work in these manifestations of campaign timing effects, and their broader implications. Research on the impact of political communications concurs that the ability of elite actors to persuade is greatly limited, and that members of the public are not all equally likely to be responsive to the information flows. Specifically, people heed the salient issues raised by the news media, and adopt the views of elites based on cues of shared identity or preferences (Iyengar & Kinder, 1987; Zaller, 1992; Lupia & McCubbins, 1991). These mechanisms operate conditionally, based on characteristics of the individual as well as features of the message and context.

In this paper I draw from the latter approach to further investigate the mechanisms underlying the public's dynamic responses to campaigns. In particular, I evaluate hypotheses about how the *persistence* of responses to campaign information might vary across individuals and over time. By persistence, I mean the extent to which effects are long-lasting or permanent—a crucial factor in determining electoral implications (see, e.g., Shaw 1999; Wlezien & Erikson, 2002). Existing work suggests several competing hypotheses for who is likely to exhibit persistent responses, and how this is influenced by campaign timing or information. I briefly review the relevant theory, detail the hypotheses under consideration, and discuss how the required statistical models can be estimated using the available observational data. I conclude with a discussion of the normative implications and proposed directions for future inquiry.

Perspectives on campaign dynamics

A handful of broad generalizations can be drawn at present about the impact of campaigns on voter decisions. Most simply, certain kinds of campaign events and efforts on the part of campaigns and parties produce demonstrable returns (e.g. Shaw, 1999). A clear example is the tendency of each nominating convention to attract voters to its party's presidential candidate, presumably due to several days of overwhelmingly positive media attention for that candidate. However, the effects of campaign information on electoral outcomes tend to be small in absolute magnitude (Holbrook, 1996). Insofar as citizens' electoral outcomes do deviate from pre-campaign expectations, the magnitude in recent elections has been only a few percentage points. In terms of the dynamics of citizens responses, poll results assessing citizens' preferences tend to converge to electoral outcomes themselves—which in turn tend to match pre-campaign forecasts based on static features of the electorate and political conditions (Gelman & King, 1993). This convergence itself is hypothesized to be a permanent drift in opinion that begins roughly during the last one hundred days of a presidential campaign (Wlezien & Erikson, 2002).

The above process has been referred to speculatively as the "enlightenment model" because one interpretation of the observed dynamics is that they represent how citizens learn how their true preferences map onto the candidates. This is only one possible dynamic progression, however, and we do not know whether all citizens are subject to it. A more onerous possibility is that opinion change that only lasts a few days could swing an election. For example, it is often speculated in the last several Presidential elections that one side or the other could somehow manufacture an "October surprise." The concern about "October" suggests an implicit assumption that

campaigns can induce temporary opinion change, that does not represent the fully enlightened view of the public and due to its instability reverts to a prior equilibrium. The term conjures an election in which the electorate, having voted for candidate A in response to event E in late October, would have supported candidate B if E had not happened and will support B within a short time period, *ceteris paribus*. While game-changing events may be far more likely in September, these may be seen as less deliterious if the electorate has the time to evaluate and reconsider the new information, along with counter-arguments presented by elites.

Individual differences in the dynamics of candidate evaluations

Despite a moderate amount of work exploring the aggregate-level dynamics of campaigns—and the immediate and important follow-on of understanding when and for who campaign effects are strongest and most persistent—there are virtually no studies addressing the latter questions in tandem. On one hand, panel studies typically do not interview respondents repeatedly over the short timepsans anticipated for campaign effects. However, non-panel data are now widely available. As I will show, there are fairly obvious (if crude) ways of using them to explore levels of persistence of campaign effects for different groups of citizens. A more serious issue may be the theoretical rigidity of frameworks generally employed to model dynamics. The most widespread models for representing dynamics—the ARIMA and error-correction models—enforce orthogonality between forces that produce permanent or nonstationary and those that produce short-term or stationary dynamics. An alternative recently promoted in political science—fractionally-integrated dynamics—are claimed to be empirically widespread, but are motivated as a reduced-form for ag-

gregated data, rather than as consequences of substantive phenomena intrinsic to the units under study.

In what follows, I describe a functional model of individual-level candidate evaluation that will serve as the starting point for developing a set of hypotheses about individual differences in the dynamics of candidate evaluations. These hypotheses are derived from past literature on campaign dynamics as well as theories of political communication and psychology. To anticipate, these represent (1) the fact that certain types of people are relatively more likely to receive persuasive messages—so that even if shocks are common, they are not expected to be equally felt, (2) that people are sometimes motivated to selectively accept or reject new information in ways that bolster or maintain existing attitudes, (3) that a person's mode of processing for updating judgments (online- or memory-based) is expected to vary, with a propensity for each that varies across individuals, with implications when the short-term component of opinion dynamics can be isolated, and (4) that individual/candidate partisanship shapes the dynamics of responses to focal events like the conventions and debates, and (5) that certain types of people may be more susceptible to priming due to shifting media emphasis.

General framework

In the following, I use a terminology for the process of political decision-making broadly compatible to Zaller (1992). Denote individual i's hypothetical evaluation of candidate C at time t as $y_{i,t}^{C}$. Henceforth I omit reference to the candidate and drop that superscript. By stating that the evaluation is "hypothetical," I mean that it is how the individual would typically judge the candidate if asked by a surveyor or otherwise motivated to do so, but that it doesn't always exist outside of that

context. The evaluation is modeled as a measure of a set of considerations that are stored in an individual's memory and which can be used to evaluate the candidate. For simplicity, I take the set of considerations to be finite with k evaluative implications $\left\{\Delta^{(0)} \dots \Delta^{(k)}\right\}$. I then specify candidate evaluations to be formed via a linear combination of these evaluative implications:

$$y_{i,t} = \sum_{j=0}^{k} w_{i,t}^{(j)} \Delta_{i,t}^{(j)} \tag{2.1}$$

Aside from the indexing over individual and time, this specification of candidate evaluation matches typical models of candidate evaluation and relatedly, electoral choice. An important note about is that, while theoretically the weights $w_{i,t}^{(j)}$ and all of the considerations yielding the quantities $\Delta_{i,t}^{(j)}$ determining candidate evaluations could vary amonst individuals, our ability to leverage the model empirically requires that some of these are either measured or shared in common for the population of interest. In other words, we must assume a population with characteristics sufficiently well-defined by a common model. Most commonly, some of the considerations may be proxied by a set of respondent characteristics that are thought to be related to long-term predispositions, and one or more others may be latent signals indexed by time, hypothetically representing new, common information received by each individual over the course of a campaign. The simplest version of such a specification is:

$$y_{i,t} = \left(\sum_{j=0}^{k-1} w_{i,t}^{(j)} \Delta_{i,t}^{(j)} \right) + \tau_{i,t} \Delta_{t}^{(k)} \tag{2.2}$$

Here $\Delta_{t}^{(k)}$ is the evaluative implication communicated by the common campaign signal at time t, and $\tau_{i,t}$ is a weight determining its standard deviation. This is a trivial instance of campaign dynamics, in which people respond in a transient way to

common information, but the evaluative implications of this information potentially die out by the next time period. In contrast, persistent campaign dynamics can be represented in this specification by supposing that some of the other considerations $\Delta_{i,t}^{(j)},\ j < k$, are equal to or related to past shocks $\Delta_{i,s}^{(k)},\ s < t$. Explicitly, for example, past considerations introduced to the individual by the campaign may lead to evaluative implications that are weighed in a way that decays with time in accordance with an autoregressive model:

$$y_{i,t} = \left(\sum_{j=0}^{k-1} w_{i,t}^{(j)} \Delta_{i,t}^{(j)} \right) + \sum_{s=\infty}^{1} \left(\alpha_{i,t}^{s} \Delta_{t-s}^{(k)} \right) + \tau_{i,t} \Delta_{t}^{(k)} \tag{2.3}$$

These weights may decay in other ways. The substantive reasons why we might often expect such decay is discussed in the sections below. If for a particular individual at a certain time, past shocks have effects that do not decay, but add up over time, then $\alpha_{i,t} = 1$. Such shocks might for example arise due to a person learning something important about a candidate, or due to a change in economic conditions. In the next sections, I talk about how individual differences among respondents lead us to predict different levels of dynamic variance and different degrees of persistence.

Individual differences in the dynamic variance of candidate evaluations

Zaller's (1992) RAS (receive, accept, sample) model of judgment can be instantiated in the above form, holding that people construct their opinions when necessary based on considerations recalled from memory (see also Zaller & Feldman, 1992). An implication is that individuals with either relatively low or relatively high levels of political sophistication will have attitudes that are more stable and less subject to campaign effects and information. In particular, low sophisticates are expected to be exposed to less political information in the first place, and hence few oppor-

tunities for persuasion. In terms of the above model, $\Delta_t^{(k)}$ may often be zero, and its expected size expressed via $\tau_{i,t}$ is is therefore smaller than those with middling political sophistication. On the other hand, even though high sophisticates are exposed to more information, they already possess the ability to recognize a partisan or ideological context, and strong predispositions and many considerations enabling them to construct a counterargument. Their rejection of considerations can again be translated into a smaller $\tau_{i,t}$. Together, these observations suggest that, to be consistent with Zaller's model and findings, middle levels of political sophistication ought to lead to the greatest levels of responsiveness to campaign new information.

An alternative way to operationalize Zaller's contribution would be to explicitly model the way in which predispositions and extant considerations lead individuals to interpret the evaluative implications of campaign information differently, for example by allowing for an interaction between the dynamic $\Delta_t^{(k)}$ and other considerations to represent partisan bias. See, e.g., Bartels, (2002).

Individual differences in the persistence of candidate evaluations

Like Zaller's (1992) RAS model, Lodge & Taber's (2000) hot cognition model can be instantiated in the above form. Both models lead to hypotheses concerning the relationship between persistence of the dynamics of candidate evaluation and citizens' levels of political knowledge or sophistication. The RAS model stipulates memory-based processing—that each evaluation requires a set of considerations in memory to be sampled and processed (Hastie & Park, 1986). In terms of the above specification, the$\Delta_{i,t}^{(j)}$'s are never derived from $y_{i,t-1}$ but instead come from specific considerations that are repeatedly brought to bear in evaluating the candidate.

The memory-based model is not explicitly autoregressive since past values of eval-

uations are never involved. However, the dependence of evaluations on the activation of concepts in memory is expected to produce dynamic effects such as persistence effect illustrated in equation 2.3 above. One reason is that the availability of a recently activated consideration may make it more likely to play a greater role in evaluation, as observed in the process of priming (Iyengar & Kinder, 1987). Indeed, a primary role of political communications during campaigns is seen to be their influence on the salience or importance of particular bits of information (Iyengar & Simon, 2000). These kinds of effects are expected to be temporary, since—unless it becomes chronically available—a concept in memory must be repeatedly activated or other concepts will become the locus of activation. [1] A different source of short-term persistence in the limitations of memory imply that only a limited number of bits of information can be incorporated into the judgment, and therefore updating the set of considerations over the course of moving from time t to time $t+1$ must involve effectively downweighting less salient considerations any time the campaign brings to light a new and useful parcel of information about a candidate.

In contrast, the hot cognition model implies that evaluations are produced online—each is stored as an affective tally associated with the target object in memory and any new information is immediately used to update one's associated affect (Lodge, McGraw, & Stroh, 1990; Lodge & Taber, 2000). In simplest form, this is explicitly a model in which all effects are permanent. In particular, a simple form of the online-processing model is instantiated by positing that $k = w_0 = 1$, $\Delta_t^{(0)} = y_{i,t-1}$, and $\tau_{i,t}\Delta_t^{(1)}$ is the incremental increase or decrease in candidate evaluation implied by some new piece of information. This leads to candidate evaluations produced by

[1] This need not always be the case—a permanent shift in salience can have effects akin to that of learning a piece of information whose implications are not forgotten. For example, an emerging economic crisis may lead people to increase the weight on a perceived quality of a candidate as being a good, or poor, economic manager, and this could persist for the remainder of the campaign.

the non-stationary model $y_{i,t} = y_{i,t-1} + \tau_{i,t}\Delta_t$. Here candidate evaluations are stored and updated repeatedly, and that they always reflect the entire past history even if no information about the contributing considerations is recalled.

Psychologists working on judgment tend to identify the cognitive processing involved as potentially involving either memory-based or online- processing, depending on attributes of the individual and context (Hastie & Park, 1986; Redlawsk & Lau). When a person knows in advance that an evaluation will be required—which may be true for many during a political campaign—he or she may be motivated to continually process new information. To express an opinion when an online attitude is available, an individual needs only to recall it from memory. However, the effort required to be attentive to new information is typically thought of as only worth the effort for those of very high political knowledge or sophistication (Lodge, et al. 1995). In contrast, individuals with relatively low levels of sophistication may reconstruct their judgments from memory each time an evaluation is required. This is consistent with the scenario investigated in Byers, Davidson, & Peel (1997) for the responsiveness of public opinion polls: that the population is heterogeneous with regard to how it processes information and combines elements of the above two processes. Some, relatively committed individuals exhibit relatively high persistence and little responsiveness to new information. Other, relatively uncommitted individuals exhibit relatively low persistence and high responsiveness to new information.

Based on these findings of individual differences, and the identification of online-processing with dynamic persistence via the incorporation of stored past evaluations into present evaluations, I posit first that those with relatively high political knowledge or sophistication will tend to exhibit greater persistence in their opinion change. However, to be clear, this hypothesis is not entirely unambiguous. Although the sim-

ple form of the online-processing model described above is obviously an integrated (non-stationary) time-series model in terms of the evaluative implications of new information, a memory-based model of campaign learning could also be non-stationary if people learn new considerations and do not forget them. However, as I have argued, cognitive limitations would mean that some bits of information are forgotten by the time an individual is called upon to make a judgment. If we assume these limitations are least among those of high sophistication, then they ought to exhibit the greatest persistence.[2]

If people engage in a mixture of these processes, and also learn new information during campaigns, the implications differ depending on how and when this mixing occurs. If, with a certain probability, an affective tally is stored, and otherwise the tally is discarded and available considerations in memory are used, then we can expect stationary dynamics but with a higher level of dependence than we would expect due to the occasional use of perfectly persistent tallies.[3]On the other hand, if the tally is retained but not always used, then a person's evaluations will consist of

[2]Nor does Zaller's memory-based account provide much explicit guidance. Zaller distinguishes an equilibrium attitude from context effects (p. 69), but seems to provide no mechanism whereby some pieces of information are accepted permanently, but others are allowed to be forgotten. Three ways to resolve this are: allow all attitudes to be forgotten, suggest that nothing is ever forgotten, or provide an "integration" mechanism which reinforces and remodels some memories as citizens gain expertise. Zaller himself notes that the model lacks an "integration" step (as Zaller notes on page 281) to explain long-term learning

[3]Formally, define a person's processing mode in terms of a random variable $A_{i,t}$ with $A_{i,t} = 1$ implying forming or updating a stored tally, and $A_{i,t} < 1$ memory-based processing as meaning $A_{i,t} < 1$, where $A_{i,t}$ is random and its realization is the autoregressive coefficient $\alpha_{i,t}$, and suppose people receive and process only one consideration of new information at each time-period whose evaluative implication is$\tau_{i,t}\Delta_{i,t}$ in order to update $y_{i,t}$. Without loss of generality, suppose that under memory-based processing, *all* prior considerations decay geometrically as in equation 2.3, so that under memory-based processing, $y_{i,t} = \alpha_{i,t}y_{i,t-1} + \Delta_{i,,t}$To represent the fact that people might engage in one or the other form of processing, and their propensities depend on their characteristics, let $B_{i,t}$ be Bernoulli with a probability $p_{i,t}$ of online processing depending on these characteristics. Then we have

$$y_{i,,t} = \begin{cases} y_{i,,t} + \Delta_{i,t} & \text{if } B_{i,t} = 1 \\ \alpha_{i,t}y_{i,t-1} + \Delta_{i,t} & \text{o. w.} \end{cases}$$

and therefore marginalizing over $B_{i,t}$,

a component that is permanent and a component that is transitory. At the end of this chapter, I consider a model that allows for such a possibility by incorporating both stationary and non-stationary components.

Individual differences in measurement error

A basic source of heterogeneity is that we expect relatively more sophisticated citizens to exhibit more stable responses and less measurement error (Basilli & Krosnick, 2000). This will be reflected by a lower variance for the "innovations" term $\Delta_{i,t}$ relative to less politically sophisticated citizens.

Current Study

Based on the above discussion, I consider the following three hypotheses in the present study.

H1: Relatively more politically sophisticated citizens ought to exhibit greater persistence in the impact of past evaluations.

H2: Relatively more politically sophisticated citizens ought to exhibit less measurement error in their opinions.

$$\begin{aligned} \mathrm{E}\left[y_{i,t}\,|y_{i,t-1}\right] &= p_{i,t}\left(y_{i,t-1}-\mu_i+\Delta_{i,t}\right)+\left(1-p_{i,t}\right)\left(\alpha_{i,t}\left(y_{i,t-1}-\mu_i\right)+\Delta_{i,t}\right) \\ &= \beta_{i,t}\left(y_{i,t-1}-\mu_i\right)+\Delta_{i,t} \end{aligned}$$

with $\beta_{i,t} \equiv p_{i,t}+\alpha_{i,t}-\alpha_{i,t}p_{i,t} = \alpha_{i,t}+\left(1-\alpha_{i,t}\right)p_{i,t}$. Based on this appearance of $\beta_{i,t}$, we see explicitly that groups of individuals who are more likely to engage in online processing in any given time-period (a larger p_{it}) will exhibit a larger $\beta_{i,t}$ with the limiting cases $\beta_{i,t}=1$ for $p_{i,t}=1$ and $\beta_{i,t}=\alpha_{i,t}$ for $p_{i,t}=0$.

Research Methods

Data & Measurement

The two main data sources available for evaluating the detailed dynamics of attitudes towards candidates are the Continuous Time Monitoring Survey of the 1984 American National Election Studies (ANES) (Miller, 2000), and the rolling cross-sections collected as part of the National Annenberg Election Studies (NAES) (Romer, et al, 2004). The critical measures within these surveys are questions that assess the favorability of respondents towards presidential candidates over the course of the campaign periods in which they were conducted. Beyond the explanatory factors that are hypothesized to be related to the dynamics of judgments, determinants of favorability towards candidates potentially include any of the respondent characteristics and judgments thought to be relevant to vote choice.

The NAES studies include (or are expected to include) representative repeated cross-sections of the U.S. population, collected at daily randomly spaced time-points during the 2000, 2004, and 2008 election campaigns. Other recent work (Shaw, 1999) finds persistence in campaign effects over the course of periods from one to ten days apart. That is, he shows that the effect of a particular event is completely realized within the span of several days. I utilized three day spans because at this level of aggregation, every combination of political knowlege, party identification, and the the time indicator, contained at least one observation.

Inferences about dynamic responses are possible using the NAES data using pseudo-panel estimation methods. The basic idea is that, although we lack past observations for each individual, we may still be able to impute characteristics of their potential responses at past time points based on the set of all other observations observed at that time point. This idea has been used to develop the so-called

pseudo-panel estimation strategy (e.g. Deaton, 1985; Moffitt, 1993; Varbeek & Vella 2005; Varbeek, 2008; Wooldridge, 2008) that I propose for the observational analyses in the following section. Further, the use of one-hundred point thermometer rating scales implies an additional complication of measurement that I discuss in the future directions section below.

Basic Model & Estimation

When we substitute information concerning the responses of a distinct sample of individuals for a genuine history of each individual, we inevitably introduce violations of the basic moment conditions on which small-sample estimators rely. For example, the OLS 'pseudo-panel' estimator I develop in this paper based on projecting past responses onto the characteristics of present individuals, will be biased insofar as the estimated lagged terms deviate from the true lagged terms. However, unlike many instances of measurement error, the estimator is consistent provided we can make certain assumptions, such as that determinants of present response (other than the lagged dependent variable) are serially independent (Varbeek & Vella, 2008). Further, if we wish to model the response in terms of other dynamic covariates, a simple extension of the pseudo-panel estimation strategy leads to a valid 2SLS estimator.

Consider a standard AR(1) model of persistence in an individual's evaluations of a candidate. To start with, I assume a fixed autoregressive parameter α, and ignore all constant individual-level considerations. The resulting model is:

$$y_{i,t} = \alpha y_{i,t-1} + \Delta_{i,t} \tag{2.4}$$

When a true panel is available, we observe all quantities except the autoregressive parameter α_t (determining the quantity of interest, the persistence of shocks to the

individuals' evaluations from one time period to the next), and the innovations $\Delta_{i,t}$ (representing the individual's updated evaluations with the new information available at time t). Estimation of these unknowns would be relatively straightforward. If the innovations were uncorrelated, an OLS regression estimator is unbiased for α_t. If the above model is properly specified, we might still expect $\Delta_{i,t}$ to exhibit serial correlation, since campaign information sources are likely to deliver fairly similar messages over any short timespan. In the case of serial autocorrelation in the innovations, the OLS estimator would be biased due to the simultaneous correlation between $y_{i,t-1}$ and $\Delta_{i,t}$. That could be dealt with by adding to the model additional lagged realizations $y_{i,t-s}$, $s > 1$, and then estimating it using OLS—if that were sufficient to account for the more extended persistence. Alternatively, a FGLS estimation strategy could be used to account for a more complicated form of persistence.

In the following, I follow with some modification the developments of Varbeek & Vella (2008) and Franklin, (1989). When we do not have a true panel, we do not observe $y_{i,t-1}$ for each individual. We may still be able to adequately estimate the equation (2.4) above with a pseudo-panel estimator that utilizes information available from other individuals who were measured at time $t-1$. We begin by forming an estimate of the true, unobserved lagged responses. The expected value $\mathrm{E}[y_{i,t-1}]$ minimizes the sum of squared deviations from these lagged responses. We have also that $\mathrm{E}[y_{i,t-1}] = \mathrm{E}[y_{j,t-1}]$. Therefore, a good estimate of the population mean of observed lagged responses might provide an estimate of the unobserved lagged responses with desirable properties. Given no information other than the set of measured responses at time $t-1$, an estimate is given by the OLS regression on time dummies, backcasted one period. This estimate, the lagged mean $\sum_j y_{j,t-1} \equiv \bar{y}_{t-1}$, is the best linear unbiased prediction of the population mean lagged response, and

therefore the best linear predictor of the true, unobserved lagged response.

The lagged response can be expressed in terms of the estimate by $y_{i,t-1} = \bar{y}_{t-1} + d_{i,t-1}$. Substituting in equation (2.4) above, we obtain:

$$y_{i,t} = \alpha\bar{y}_{t-1} + \alpha d_{i,t-1} + \Delta_{i,t} \tag{2.5}$$

Here the unobserved measurement error $d_{i,t-1}$ is correlated with the included regressor $\bar{y}_{t-1}$, but asymptotically the correlation diminishes to zero (Varbeek & Vella, 2005). Therefore, provided that $\Delta_{i,t}$ is uncorrelated, an OLS estimator of α based on equation (2.5) will be consistent. As above, additional estimates of the lagged responses, or FGLS, could be utilized in the case that $\Delta_{i,t}$ exhibits serial correlation.

Recall that above I ignored any constant individual-specific considerations, such that each individual would tend towards zero in the absence of new information. Suppose in fact that the correct specification includes a baseline judgment μ_i that reflects the evaluative implications of considerations that are not changed over the course of the campaign. When true panel data are available including measures proxying such a baseline, then direct estimation of μ_i is important to ensure that non-zero autoregressive parameters or autocorrelation in the innovations convey the persistence of responses to new information, rather than cross-sectional variation in predispositions. In the absence of true panel data, the availability of measurements of individuals' predispositions related to μ_i allows us to obtain a better estimate of the unobservable lagged response than the sample mean $\bar{y}_{t-1}$, in a way that will be explained below. However, for the same reason of deconfounding individual-specific effects from correlated dynamics as described above, we must begin by specifying these constant factors in the response model.

In particular, suppose we observe a set of constant background characteristics for

each individual $\vec{z}_i$ such that $\mu_i = \vec{z}_i^T \vec{\beta}$ for some $\vec{\beta}$—that is, the variables in $\vec{z}_i$ are characteristics that are thought to be related to candidate evaluations. For example, we might observe the party identification and race of each individual, and these are strong indicators of his or her propensity to like a candidate (assuming the party identification of the candidate is known). Assume that individuals update their evaluations in the same way as specified above, so that we can write:

$$y_{i,t} - \mu_i = \alpha \left(y_{i,t-1} - \mu_i \right) + \Delta_{i,t} \tag{2.6}$$

$$y_{i,t} = \vec{z}^T \vec{\beta} + \alpha \left(y_{t-1} - \vec{z}^T \vec{\beta} \right) + \Delta_{i,t} \tag{2.7}$$

The logic we use to develop an estimator using added covariates parallels that which justifies the use of the lagged mean $\bar{y}_{t-1}$ above. The conditional expectation $\mathrm{E}\left[y_{i,t-1} \,|\vec{z}_i \right]$ minimizes the sum of squared deviations from the unobserved lagged responses. Since $\mathrm{E}\left[y_{i,t-1} \,|\vec{z}_i \right] = \mathrm{E}\left[y_{j,t} \,|\vec{z}_i \right]$, a good estimate of the population conditional mean response in the previous period—*conditional upon the population having the characteristics of the individuals observed in the current period*—might provide an estimator of the lagged responses with desireable properties. Not surprisingly, an estimate can be formed via OLS regression on time dummies as well as the observed covariates, backcasted one period.

There are two important types of individual characteristics that determine how $\vec{z}_i$ can best be structured and hence the nature of our regression-type estimator for the lagged reponse. First, there may be a small number of mutually exclusive groups of individuals expected to be similar in their responses, and which are determined exogeneously from the dependent variables and error terms. For example, party membership is a strong determinant of individuals' responses to campaigns, and we can expect members of each party to exhibit similar temporal patterns of candi-

date evaluations during a presidential campaign. Note that we must assume that identifying with one party or another (or no party) is exogeneous here, such that people are not changing parties in response to the candidates or the campaign, and accept any bias that results to the extent that this isn't true.[4] The temporal patterns of party categorization are expected to be good predictors of lagged individual responses $y_{i,t-1}$ which we cannot observe. To make use of mutually exclusive groups such as party membership, we include an interaction of time and group dummies in the backcasting regression used to estimate $\mathrm{E}\left[y_{i,t-1}\,|\vec{z}_i\right]$, a time-group dummy is used to identify the separate trajectory for each group.

On the other hand, there may be time-constant covariates that distinguish individual prior evaluations but which do not influence how individuals respond to the campaign. For example (though perhaps not in 2008), we might suppose that racial differences primarily influence affinity for the Republican or Democratic candidate, but do not lead to different types of responses to the campaign. However, race is still an important predictor of candidate evaluations that might help us to form a better estimate of lagged responses via $\mathrm{E}\left[y_{i,t-1}\,|\vec{z}_i\right]$. We include covariates that are thought to have constant effects across groups and over time additively rather than interactively.

Suppose $\vec{z}^{\star T}_{i,t}$ represents for a given individual at a given time a row vector of indicator variables for every cohort-time combination, along with other covariates, and $Z^\star$ represents the matrix stacking all these row vectors. Then the backcasted OLS estimate $\bar{y}^\star_{i,t-1}$ of the lagged value $y_{i,t-1}$ is:

$$\bar{y}^\star_{i,t-1} = \vec{z}^{\star T}_{i,t}\left(Z^{\star T}Z^\star\right)^{-1}Z^{\star T}\vec{y}, \tag{2.8}$$

[4]This is a reason not to employ the full seven-point scale that incorporates a relatively malleable measure of partisan strength.

The estimate $y^{\star}_{i,t-1}$ deviates from the true but unobservable lagged response with $y_{i,t-1} = \bar{y}^{\star}_{i,t-1} + d^{\star}_{i,t-1}$. Substituting in the response model with covariates given by equation (2.7), we obtain:

$$y_{i,t} = \lambda + \vec{z}^{\star T}_{i,t}\beta + \alpha\bar{y}^{\star}_{i,t-1} + \alpha d_{i,t-1} + \Delta_{i,t} \tag{2.9}$$

The pseudo-panel estimator can be formed by an OLS regression of this equation provided that the error terms are not serially correlated. When the number of cross-sectional samples is large, the measurement error $d_{i,t-1}$ should have a relatively trivial correlation with the included regressor $\bar{y}^{\star}_{i,t-1}$ (Varbeek & Vella, 2005). On the other hand, if serial autocorrelation of the innovations $\Delta_{i,t}$ is present, $\bar{y}^{\star}_{i,t-1}$ will be correlated with $\Delta_{i,t}$ in a non-vanishing way. Again, additional estimates of the lagged responses could be included in the model, in the event that $\Delta_{i,t}$ exhibits serial correlation, in order to ensure that the serial autocorrelation is eliminated.[5]

In this initial pilot study, I assume that, aside from a set of fixed individual considerations $\mu_i = \left(\sum_{j=0}^{k-1} w_i^{(j)}\Delta_i^{(j)}\right)$ that are proxied by a set of observed covariates, the dynamics of candidate evaluations can be captured by an AR(5) stationary model, spanning 5 lags of 3 days each for a total of fifteen days. If in fact the true dynamics are non-stationary, at least for some individuals, I expect approximate non-parametric estimates of the impulse response over this time-period to exhibit the failure of dynamic responses to decay.[6]

It seems possible that candidate evaluations will change systematically over relatively short time frames during political campaigns, requiring a more general model than that described above. Indeed, such episodic shifts involving non-stationary dy-

[5] Alternatively, some kind of FGLS estimator might be constructed to address such correlation, though to my knowledge the effectiveness of this procedure has not been discussed.

[6] I say "approximate" here because, although the estimates are strictly parametric, the use of such a high-order model is used solely to estimate the impulse response and calculate summary statistics, rather than to estimate specific parameters.

namics are an important part of responses, for example, to many debates (e.g., Shaw, 1999; Wlezien & Erikson, 2012). To improve the model in these cases, covariates that specify the amplitude of the campaign impulse ought to be included as predictors, or alternatively the model must incorporate a non-stationary latent component as described at the end of this chapter.

Generalization: Relating Persistence to Political Sophistication

Although I have suggested using a simple autoregressive model, the exact dynamic structure that is appropriate for changes in candidate evaluations is unknown. In order to assess the relationship between persistence and political sophistication, I instead estimate a high-order autoregressive model and then invert the estimates to produce an approximate non-parametric estimate of the impulse response function (IRF).[7] The impulse response function (IRF) measures for each j the effect on a variable in a future period $t+j$ of a unit change in a variable that affects $y_{i,t}$ during the present period t. By definition, $IRF(0) = 1$, and $IRF(j)$ declines steeply with j depending on the persistence of the response of $y_{i,t}$ to the variables that affect it. A variable that exhibits strongly persistent responses will take a long time to return to its prior value in response to even a very temporary shock.

Researchers have sought to summarize the extent of persistence using a measure called the mean lag length, defined for a strictly positive IRF. The mean lag length summarizes the persistence of effects on $y_{i,t+j}$ by treating the IRF as an unnormalized probability distribution and computing the expected lag with respect to this distribution. This amounts to a weighted average of the possible lag between present and future effect, j, weighting by the size of the impulse response function $IRF(j)$

[7] An improvement to the procedure used here might be the method of flexible local regressions as described by Jorda (2004), but in the present version I use the standard method of transforming the estimated AR coefficients to the corresponding MA coefficients to find the IRF.

. That is, $\mu_L = \sum_{j=0}^{\infty} j \times IRF(j) / \sum_{j=0}^{\infty} IRF(j)$, so that bigger values of μ_L imply that a variable exhibits longer periods of adjustment in its responses. As estimated impulse responses can be (and often are) negative, $IRF(j) / \sum_{j=0}^{\infty} IRF(j)$ is not always a proper probability distribution. In order to assess persistence, I restrict this computation to the set of coefficients prior to the first negative IRF coefficient.

From this, it is trivial to estimate the mean (or median) lag lengths. These measure the typical amount of time a unit shock to candidate evaluation persists in influencing responses. Thus, a mean lag length of zero implies that shocks influence citizens' judgments of candidates immediately and there is no persistence over time. A larger value implies that shocks have an effect that accumulates over time. In order to address the issue of negative impulse responses (see De Boef & Keele, 2005), I treat these as "missing data", setting them to zero.

Results

My initial results assess the dynamics of Bush and Gore approvals during roughly the last one hundred days of the 2000 presidential general election campaign, using the 2000 NAES rolling cross sections (Romer et al., 2004). My dependent variables are 100pt approval scales for the two candidates.

A challenging aspect of the application of the current techniques to this particular dataset is that the available political knowledge questions (factual questions related to policy) were only asked during the primary season. To measure political sophistication during the general eleciton, I constructed an auxillary penalized linear regression of an index formed from these responses on a large number of demographic variables that are likely to be related to political knowledge.[8] The fitted values from

[8]The penalized regression technique is known as LARS (Efron et al., 2003) and is employed here as a way to effectively impute political knowledge, due to the fact that the ideal model is unknown

this measure for respondents during the final 100 days were divided into quintiles and used to assign respondents to cohorts by political knowledge and party-id for the pseudo-panel estimation of the general election campaign. A replication of the analysis using education rather than imputed political knowledge found substantively similar results. Below are the primary and auxilliary models, respectively, I used to specify an entirely separate dynamic process for members of each party. I constrain the dynamic parameter to be the same for members of each party, and include other important covariates of candidate evaluations.

$$y_{i,t} = \sum \alpha_j \bar{y}^{\star}_{i,t-j} + \gamma_1 Party_i + \gamma_2 Socialissues_i + \gamma_3 Econissues_i + \gamma_4 age_i + \gamma_5 age_i^2 + \gamma_6 income_i + \gamma_7 female_i + \gamma_8 married_i + \gamma_9 south_i + \gamma_{10} black_i + \gamma_{11} edu_i$$

$$\bar{y}^{\star}_{i,t-1} = \overrightarrow{Time_i \times Party_i} \cdot \vec{\Theta} + \varphi_1 Social\ issues_i + \varphi_2 Econ\ issues_i + \varphi_3 age_i + \varphi_4 age_i^2 + \varphi_5 income_i + \varphi_6 female_i + \varphi_7 married_i + \varphi_8 south_i + \varphi_9 black_i + \varphi_{10} edu_i$$

I employ OLS pseudo-panel estimators at three-day lags. This specification includes lags up to 5×3 day periods, as well as several covariates. Figures 2.1 and 2.2 show comparisons of the mean lag length for evaluations of Bush, consistent with the idea of hypothesis H1 that persistence generally increases with levels of political knowledge. The displayed 95% confidence envelopes were computed using a non-parametric bootstrap procedure, based on the percentile method (thereby avoiding either imposing an arbitrary transformation or intervals containing inadmissable parameter values). However, the confidence interval for each of the impulse response coefficients is large and includes zero, suggesting little evidence of any dynamic response whatsoever. For Gore, the summary measure seems to suggest the that persistence decreases with levels of political knowledge, inconsistent with H1. A closer look at the impulse response functions for Gore, Figures 2.5 - 2.9, among those with high levels of knowledge indicates that part of this might be due to the particular choice of mean-lag length measure that cuts off any positive portion of the

and the full model would be expected to produce overly variable imputations.

lag after the first negative impulse response coefficient. For this group, the IRF is positive (but not significant) at long lags, hinting that there may be some long-term dependence. However, again, the confidence interval for each of the impulse response coefficients includes zero. Figures 2.3 and 2.4 illustrate how the estimated regression standard errors consistently diminish with increased political knowledge—consistent with hypothesis H2.

Future Directions

Response Scale

The favorability ratings used in the NAES—like an ANES thermometer scale—attracts many responses near the minimum, midpoint, and maximum of the scale. An anchoring-and-adjustment process might account for the prevalence of certain responses. That is, people may first construct a response that is either the minimum, midpoint, or maximum scale points using their available attitude. Then if they find that potential response mismatches their attitude to a degree that is unsatisfactory, use that as an anchor but adjust their response in the direction of their perceived attitude. [9]

Parametric Model

While we often think of the benefits of the multilevel model in terms of "sharing strength" for inferences among observations with similar attributes (c.f. Bartels,

[9]Several attempts have been made to account for the problem of interpersonal incomparability of the thermometer scales—see, e.g. Brady, 1985. Based on Brady's formulation of the problem, some of my work on random coefficients modeling might be helpful if the only problem were biased estimates of the variance-covariance matrix. Alternatively, the semantic definition given in these types of questions may be taken to imply that the minimum, midpoint, and maximum of the scale can be treated as comparable ordinal measures and the data ought to be coarsened to a 3-point scale. This would only be justifiable if the correlation in measurement errors between the thermometer and other individual-level variables or cohort indicators were a more severe problem than the attenuation bias that coarsening would likely produce.

1996), in this case it is absolutely essential that we recognize some commonality among individuals interviewed at the same time. Because of the strong differentiating role of party identification in terms of both what information people receive and how they process it (Bartels, 2002), I have above assumed that individuals identifying with a common party label exhibit common responses to information at any given time point, so that $\Delta_{i,t}$ is effectively $\Delta_{Party_i,t}$. From here on out, I assume that we are dealing with only members of one party and omit reference to those subscripts.

In addition to sharing the same informative signal due to their common party affiliation, I assume that subgroups of individuals with the same level of political knowledge exhibit the same degree of responsiveness and persistence in their opinions. That is, α_i is effectively $\alpha_{K(i)}$ and σ_i is effectively $\sigma_{K(i)}$. Given these assumptions, I re-express the above specification as a multilevel dynamic model whose individual-level specification includes a separate dynamic random effect for each political knowledge group, multiplied by a dummy variable k_{ij} indicating whether respondent i belonged to knowledge group j. I collect the dynamic random effects λ_{tj} in a vector$\vec{\lambda}_t$ and re-express the model as follows:

$$y_{i,t} = \mu_i + \sum_j k_{ij}\lambda_{tj} \tag{2.10}$$

$$\vec{\lambda}_{1t} = A\vec{\lambda}_{1(t-1)} + \vec{\xi}_t \tag{2.11}$$

Here the time level equation is a vector autoregression of the random effects with a matrix of autoregressive parameters A and correlated errors $\vec{\xi}_t$. To express the above assumptions of common persistence and information, I let $A = Diagonal\,(\alpha_i)$, so that in this case, the random effects are seemingly unrelated AR(1) time-series. I assume perfect correlation but unequal variance of the innovations $\vec{\xi}_t$ across groups,

so $Var\left(\vec{\xi_t}\right) = \vec{\sigma}_\xi \vec{\sigma}_\xi^T$.

The remainder of this document discusses a set of preliminary results illustrating how political knowledge conditioned the dynamics of NAES respondents' evaluations of Gore and Bush during the 2000 Presidential general election. Initial suspicions of non-stationary dynamics for both dependent variables and in most subsets of the data, as well as heteroskedasticity across levels of political knowledge, led me to further extend the above model to incorporate two additional sets of random effects: a time-level random walk and time-level random intercepts. I subscript these with j because I assume these are nested within political knowledge groups just like the short term dynamics described above. Substantively, these can be interpreted as terms that capture persuasion that appeared to permanently shift respondents' evaluations, and contextual fluctuations (or sampling variability). Again, I incorporate these into the specification by multiplying each by a dummy indicating an individual's political knowledge, and assume the errors are correlated but with potentially different variances across levels of political knowledge. Collecting the dynamic random effects π_{tj} and λ_{tj} into vectors $\vec{\pi}_t$ and $\vec{\lambda}_t$, I give the form of the model that represents over-time dynamics in candidate evaluations as a function of political knowledge[10]:

[10]Evaluations of Bush or Gore among either Republicans, Democrats, or Independents are modeled separately. Attributes of the individual consisted of the same regressors included in the earlier analysis.

$$y_{i,t} = \mu_i + \vec{k}_i^T \vec{\lambda}_{1t} + \vec{k}_i^T \vec{\pi}_{2t} + \vec{k}_i^T \vec{\xi}_{1t} \tag{2.12}$$

$$\vec{\lambda}_t = A\vec{\lambda}_{1(t-1)} + \vec{\xi}_{2t} \tag{2.13}$$

$$\vec{\pi}_t = \vec{\pi}_t + \vec{\xi}_{3t} \tag{2.14}$$

$$Var\left(\vec{\xi}_t\right) = \vec{\sigma_\xi}\vec{\sigma_\xi}^T \tag{2.15}$$

While the specification looks quite complex, it is input into 'hetdyn' as a response model and four simple calls to set up each error component (individual-level, AR(1), time-level, and unit root). For example, below is the command that sets up the multilevel model of Gore evaluations among Independents only using the 2000 NAES rolling cross sections:

```
hetdyn(response = evalgore~1, Independent( ~ 1 ),
Structured( list( ~ polknowfactor - 1, dynamics=''autoregressive'' ),
list( ~ polknowfactor - 1, dynamics = ''integrated'' ),
list( ~ polknowfactor - 1, dynamics = ''none'' ), index = ~ threedaycount - 1,
covariances = ~  U(3) + P(5),
window = 15 ), data = gen[(gen$isD == 0) & (gen$isR == 0), ])
```

Maximum likelihood estimates were computed for the parameters in the loglikelihood, for each of the combinations of party identification and target candidate (each included approximately 9000 to 10000 cases over 39 three-day time-periods during the fall Presidential campaign period). I omit further discussion of the results, as I found no evidence of either non-stationary dynamics nor transient dynamics in any of these results, consistent with the analysis earlier in this chapter.

Figure 2.1: Semiparametric estimate of mean lag length for evaluations of Bush by levels of political knowledge.

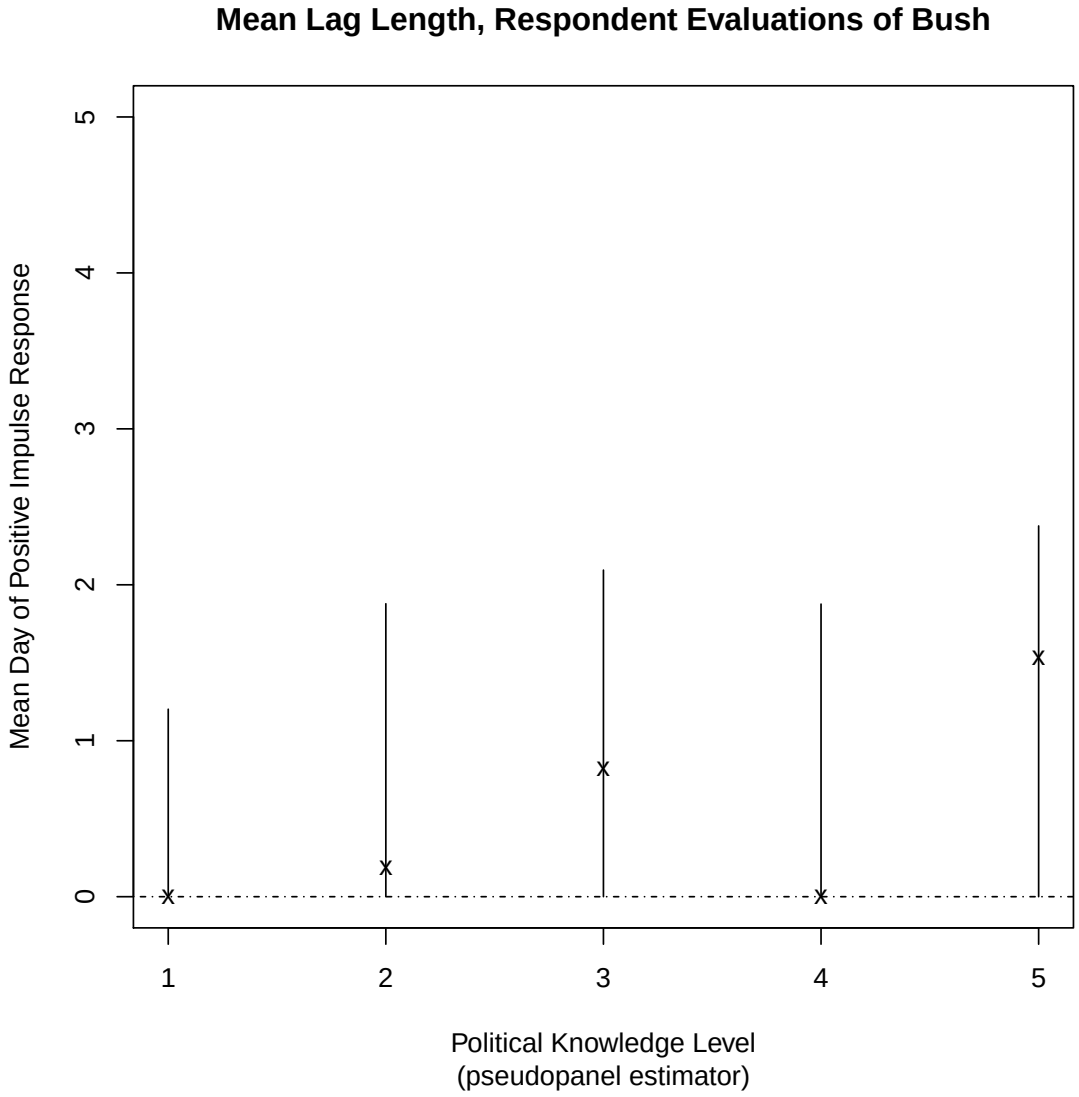


Figure 2.2: Semiparametric estimate of mean lag length for evaluations of Gore by levels of political knowledge.

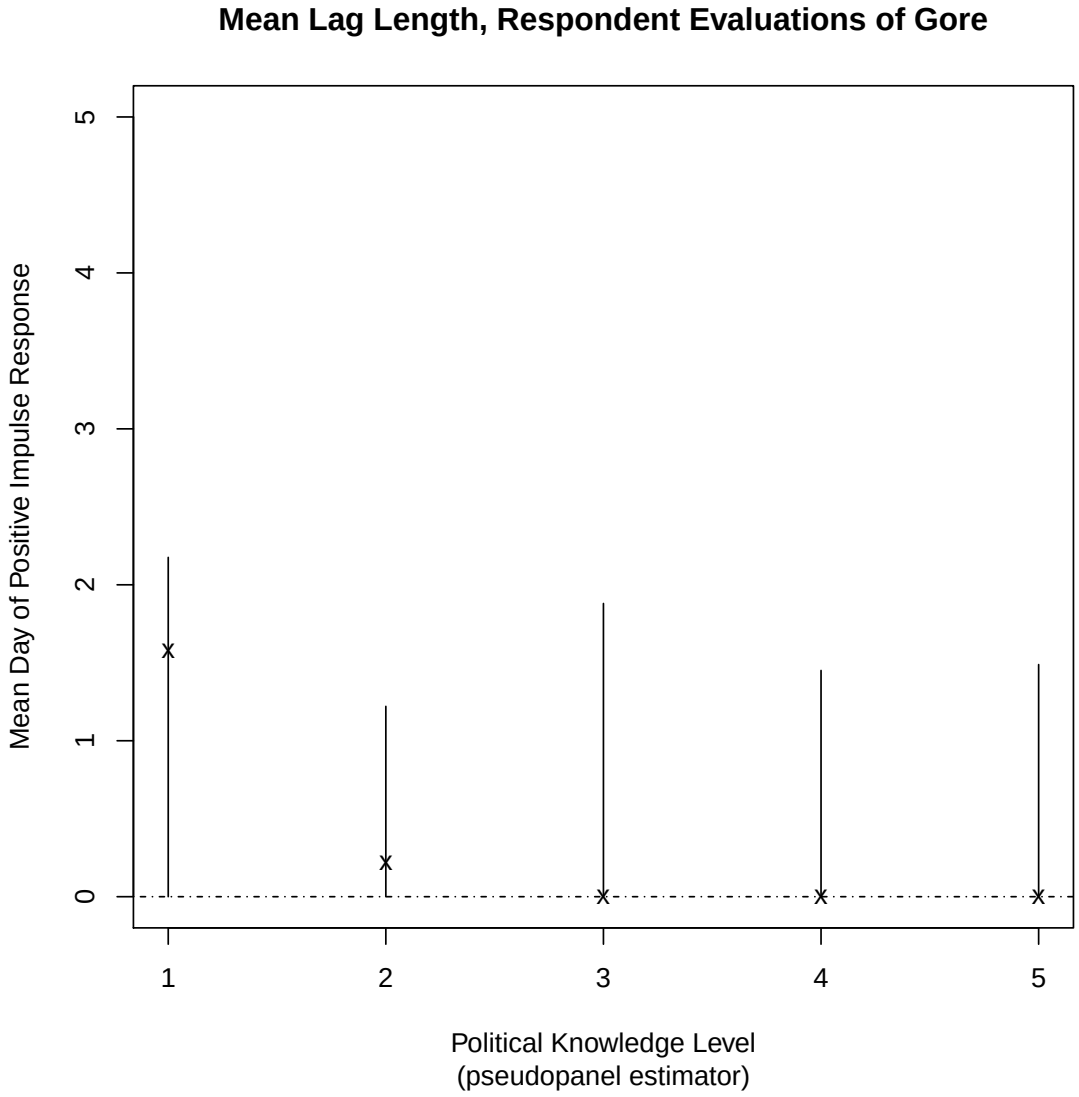


Figure 2.3: Regression standard error of pseudo-panel regression of evaluations of Bush by levels of political knowledge.

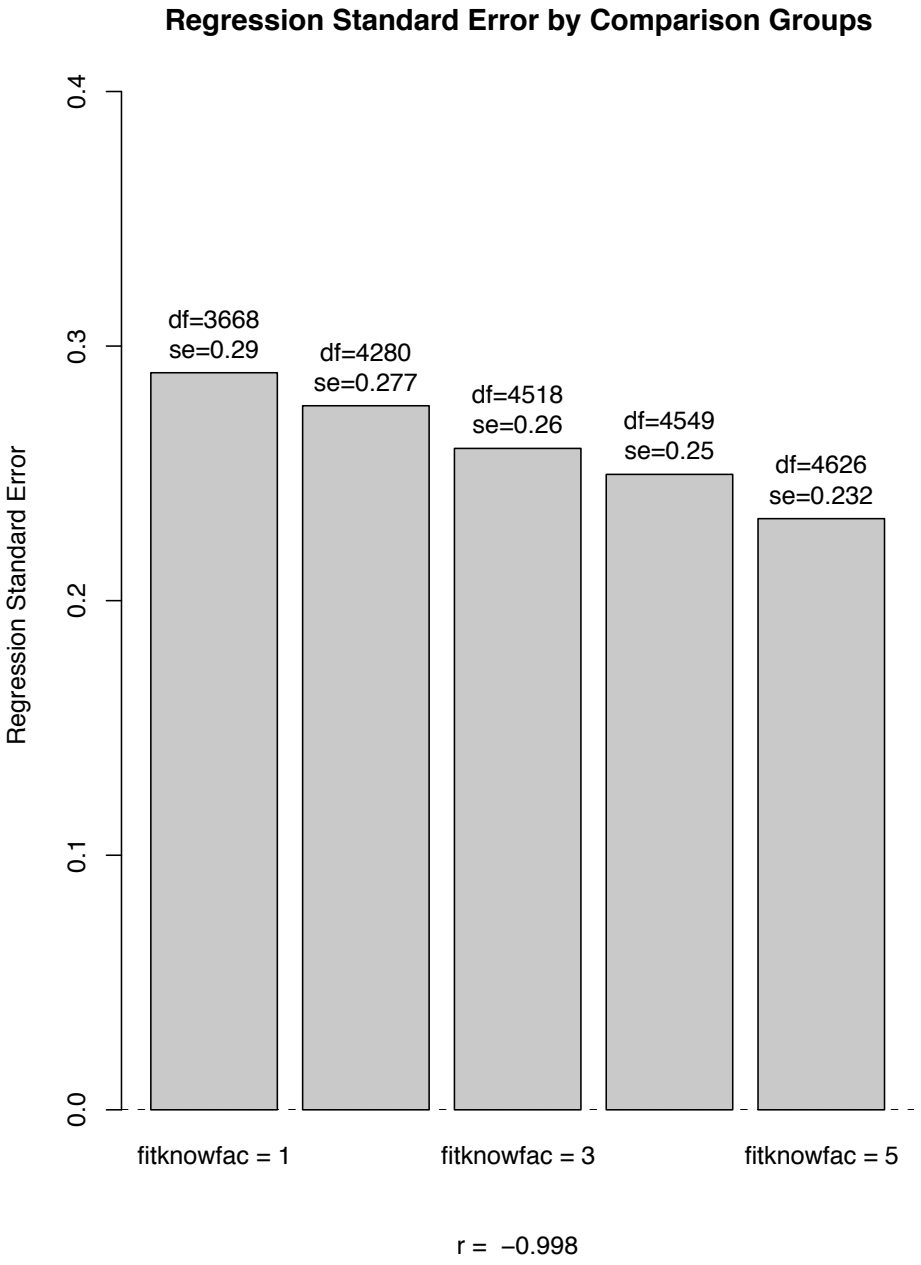


Figure 2.4: Regression standard error of pseudo-panel regression of evaluations of Gore by levels of political knowledge.

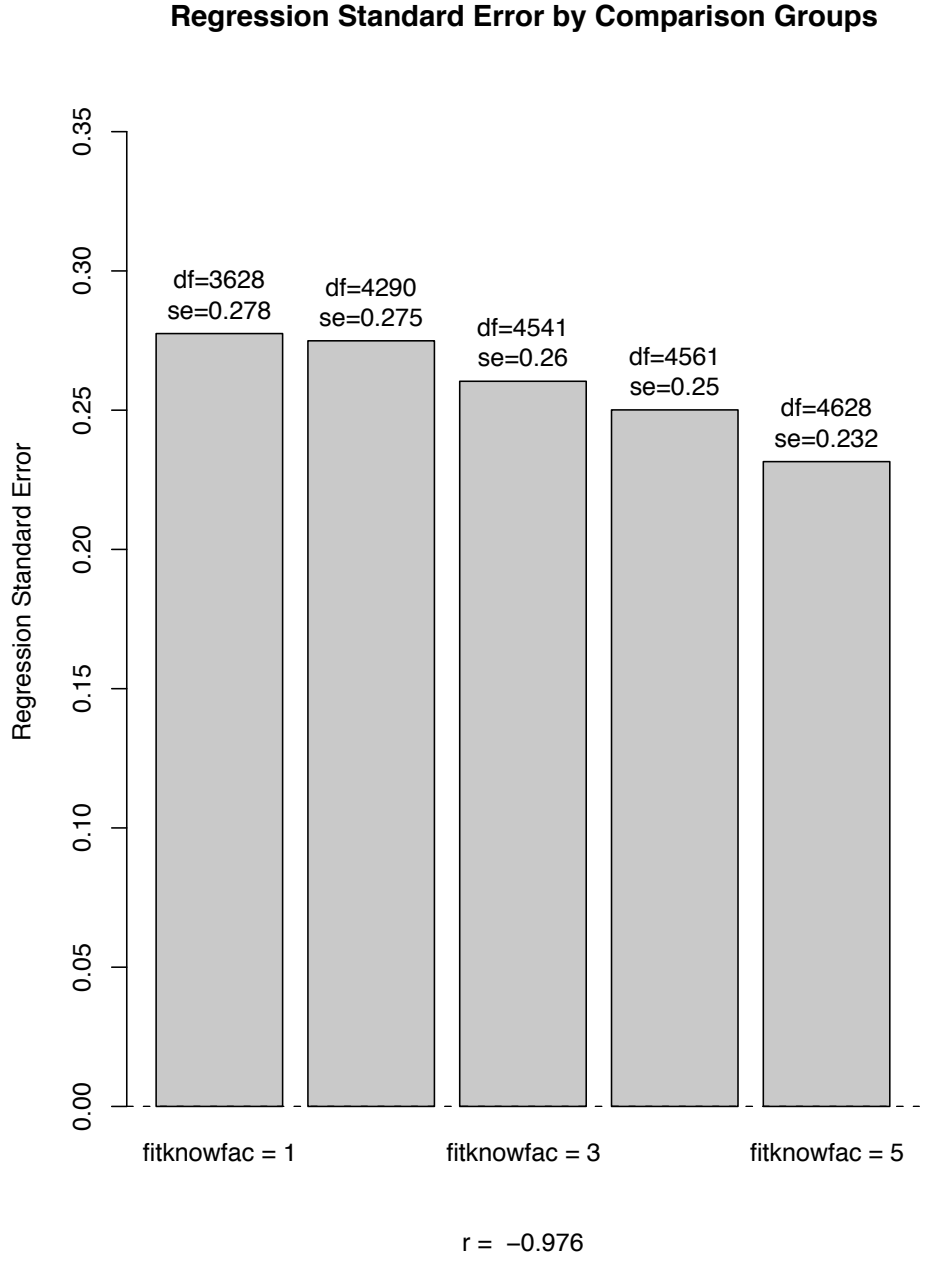


Figure 2.5: Nonparametric impulse response functions for evaluations of Gore over 3 day periods at political knowledge level 1

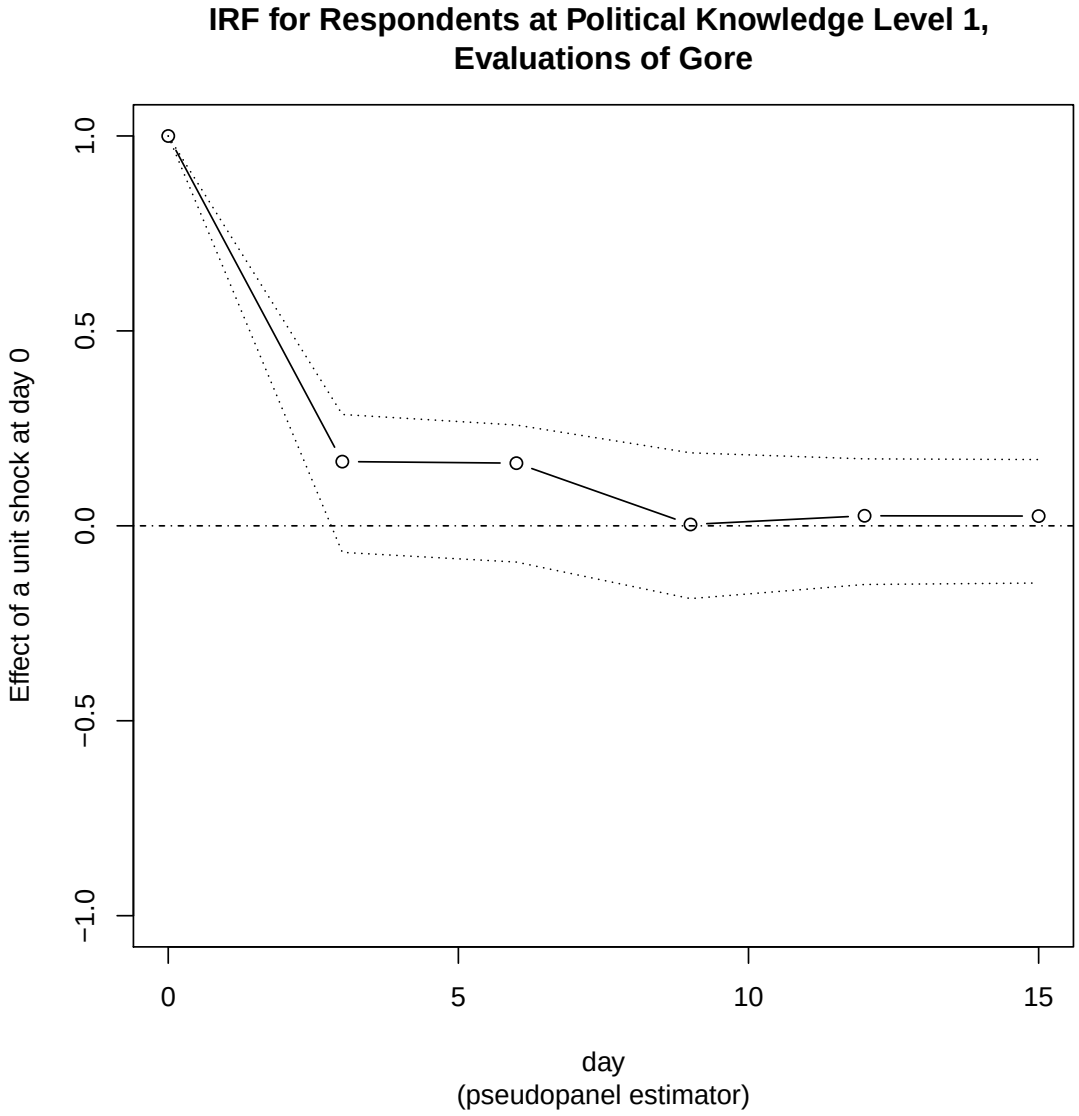


Figure 2.6: Nonparametric impulse response functions for evaluations of Gore over 3 day periods at political knowledge level 2

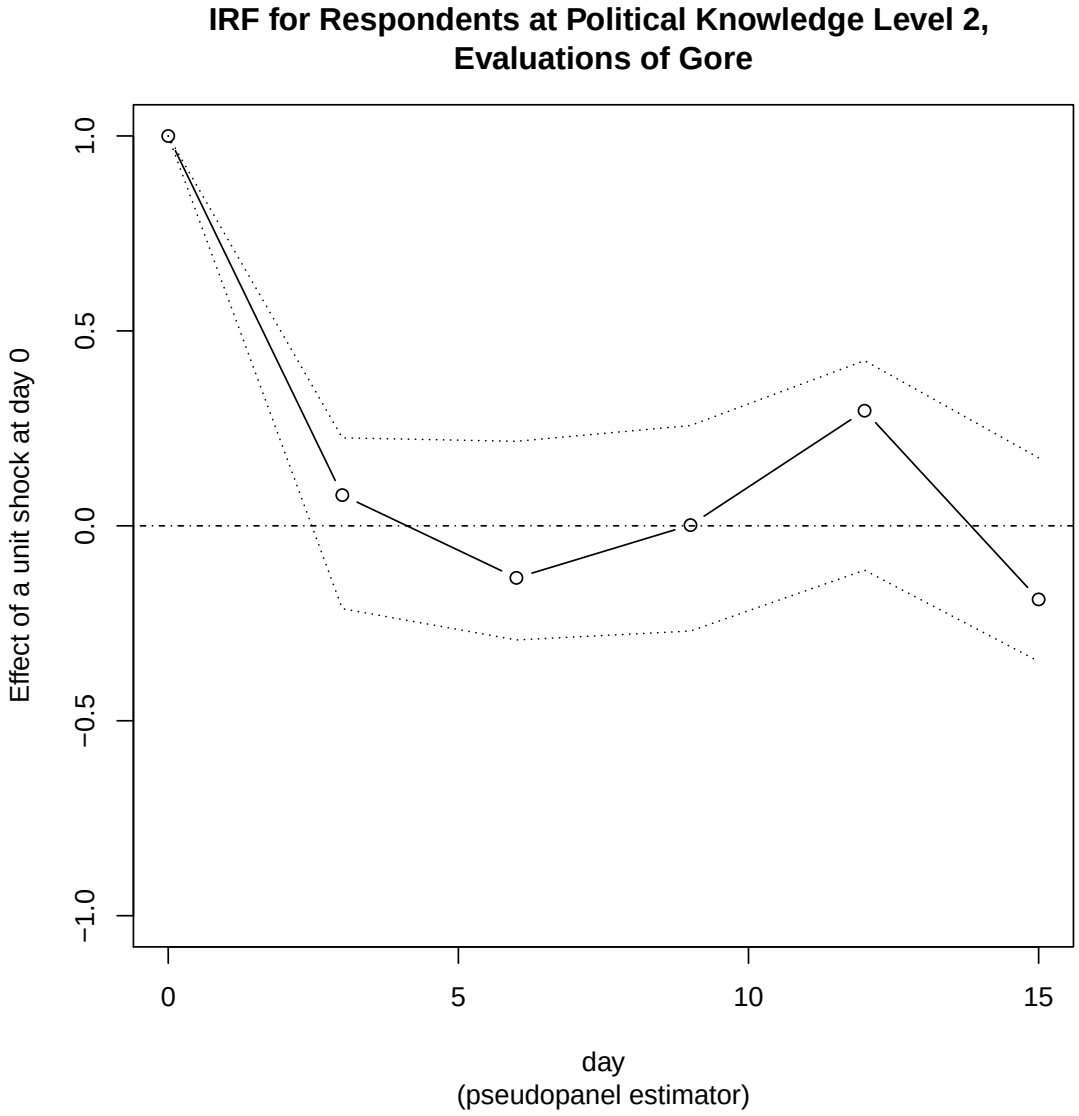


Figure 2.7: Nonparametric impulse response functions for evaluations of Gore over 3 day periods at political knowledge level 3

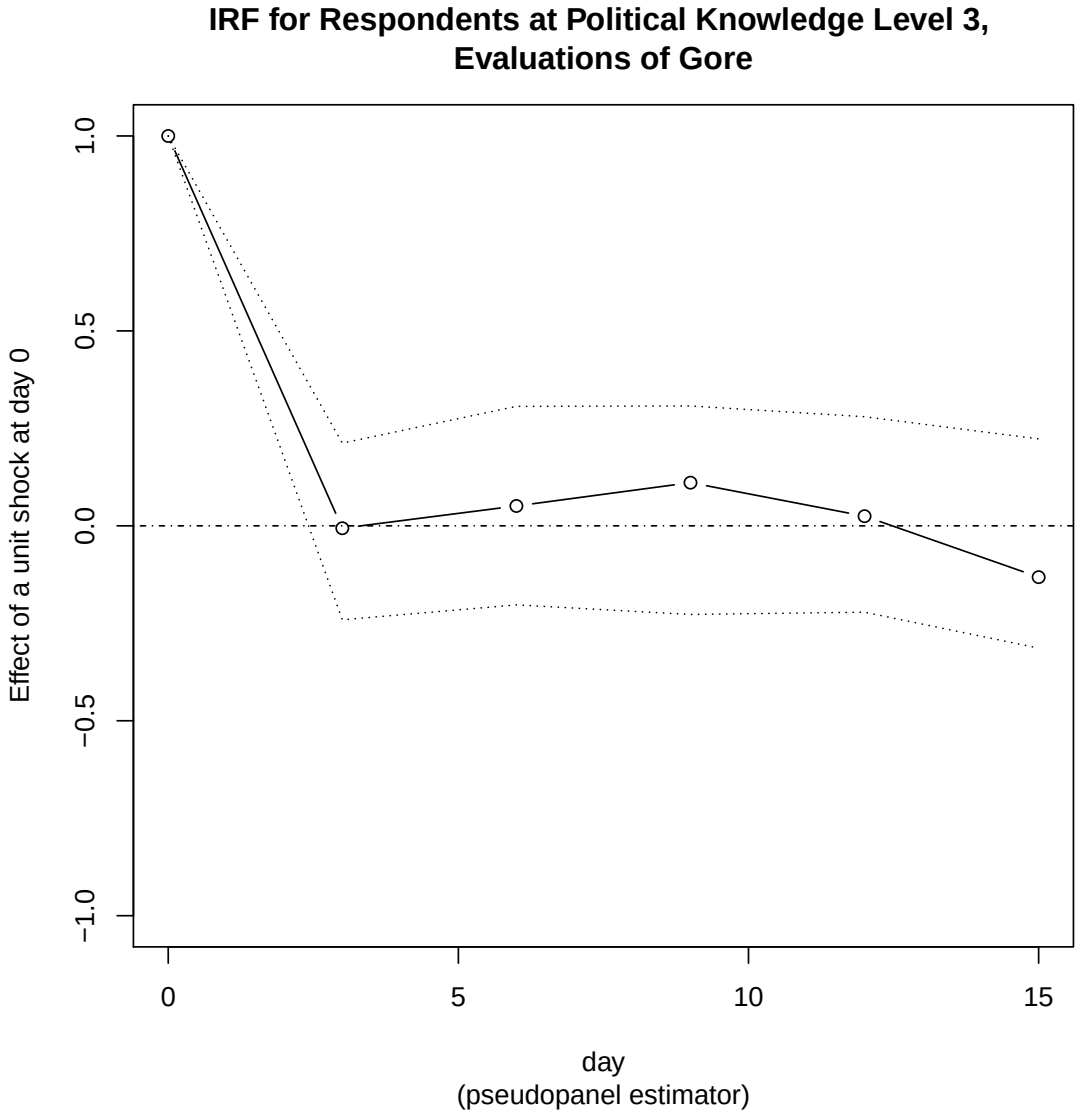


Figure 2.8: Nonparametric impulse response functions for evaluations of Gore over 3 day periods at political knowledge level 4

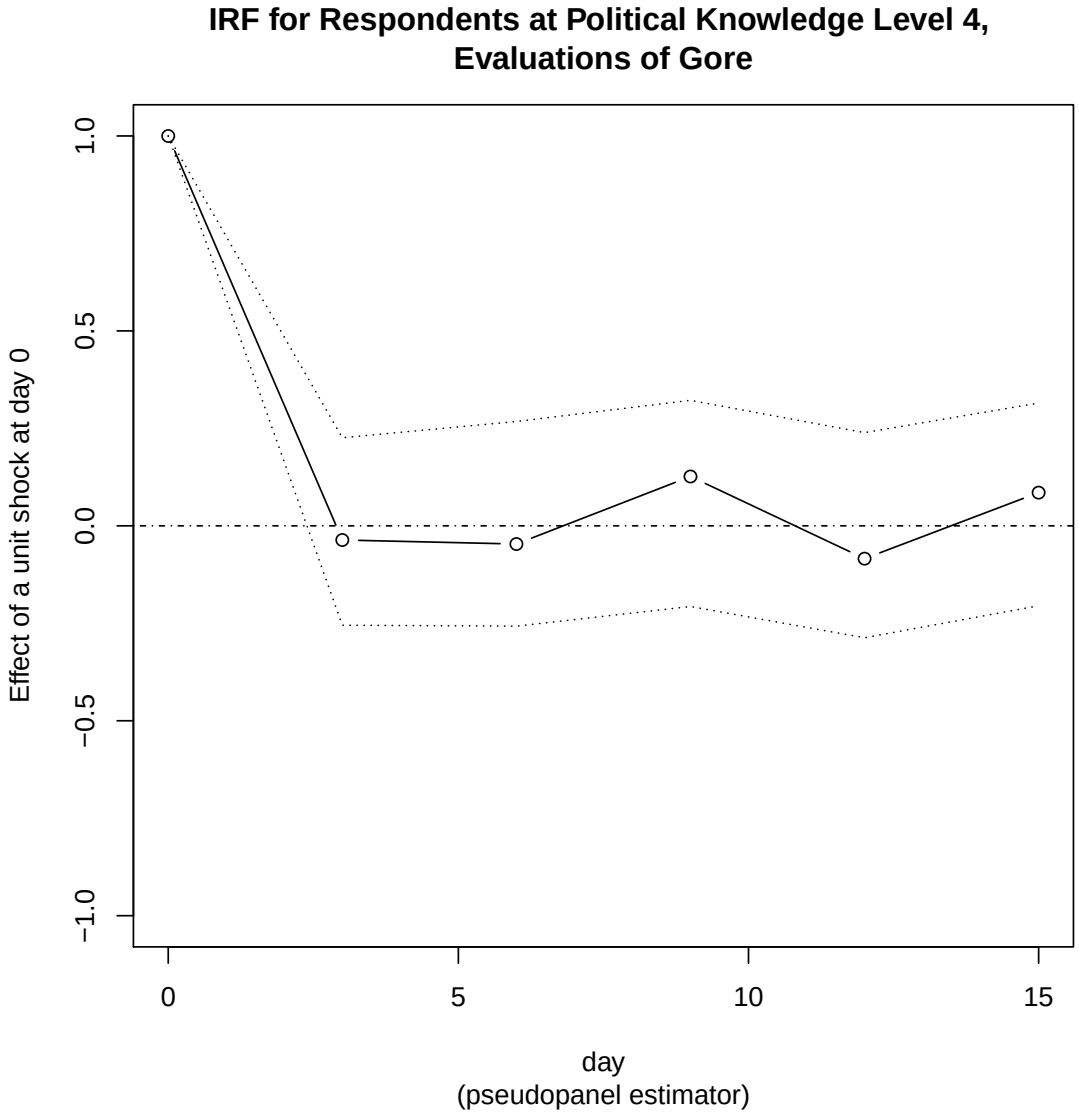


Figure 2.9: Nonparametric impulse response functions for evaluations of Gore over 3 day periods at political knowledge level 5

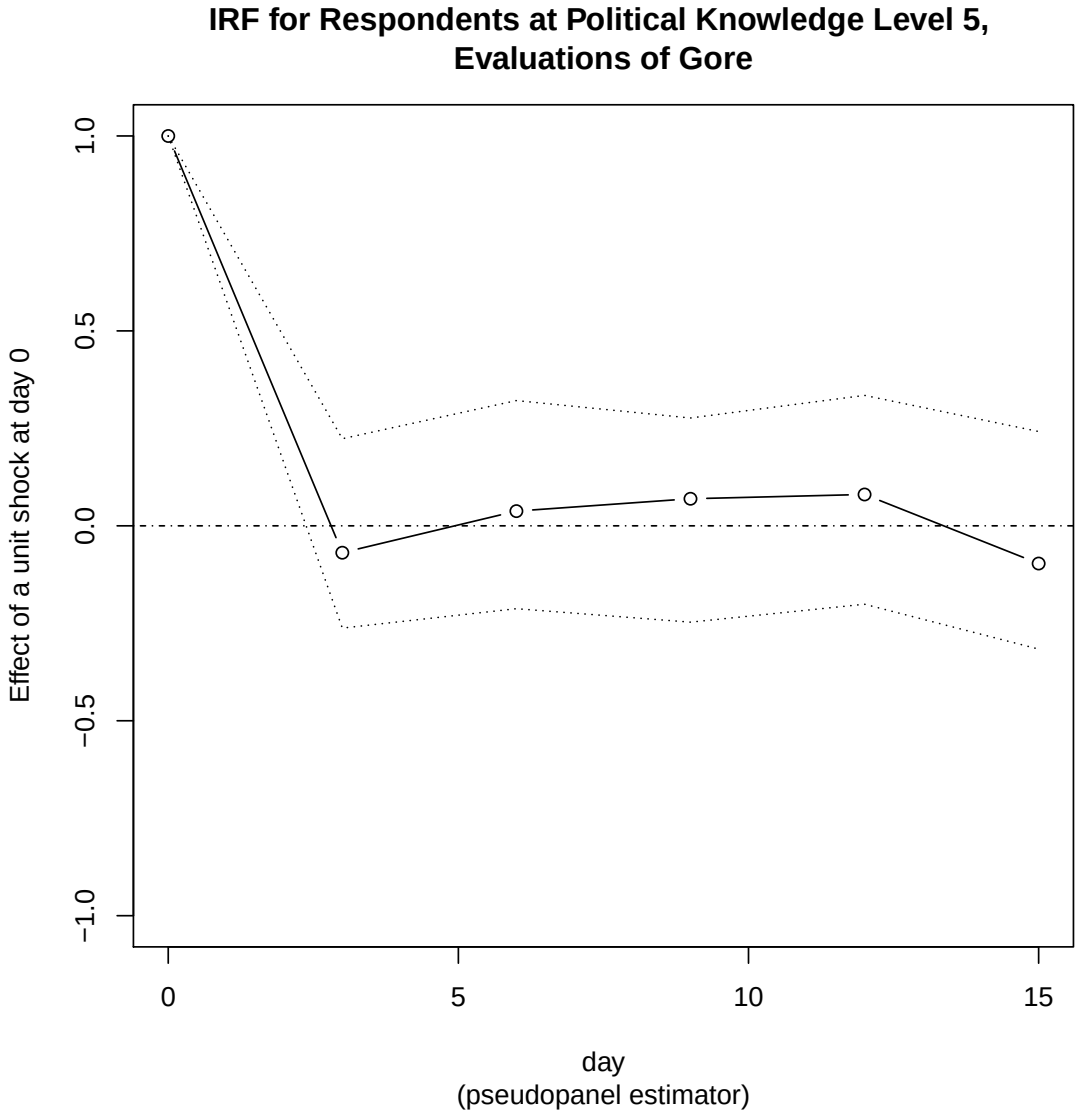


Figure 2.10: Nonparametric impulse response functions for evaluations of Bush over 3 day periods at political knowledge level 1

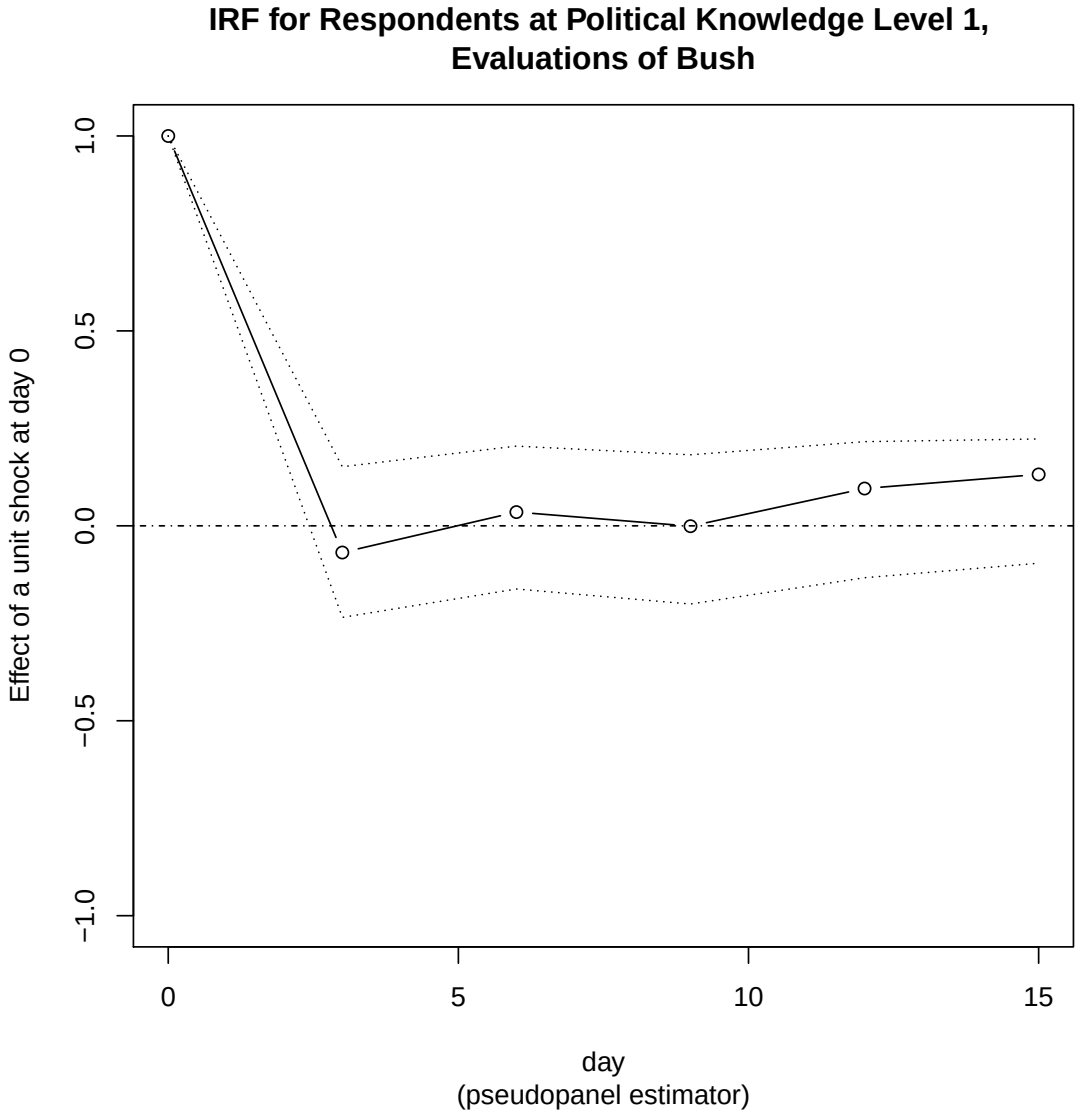


Figure 2.11: Nonparametric impulse response functions for evaluations of Bush over 3 day periods at political knowledge level 2

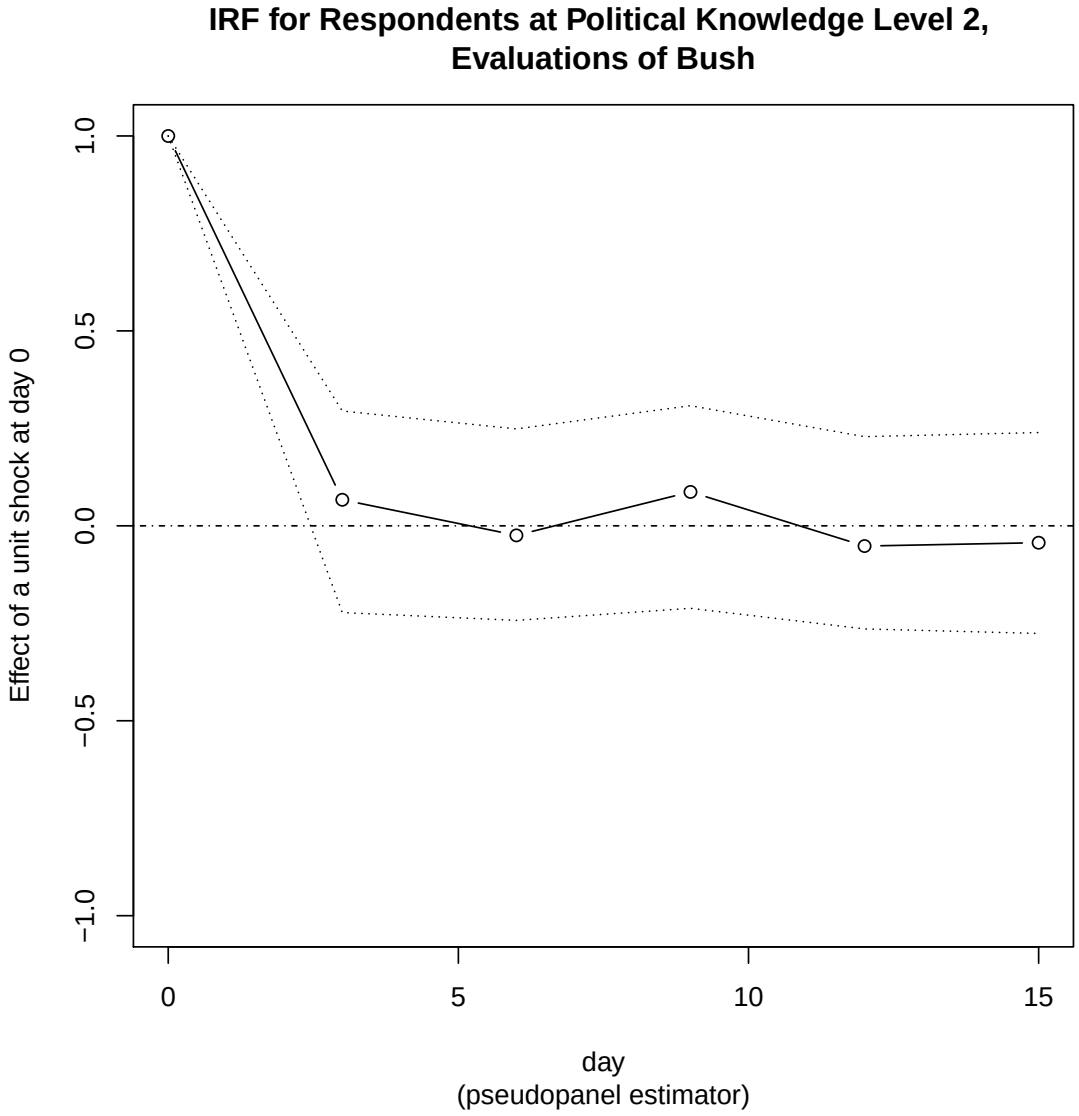


Figure 2.12: Nonparametric impulse response functions for evaluations of Bush over 3 day periods at political knowledge level 3

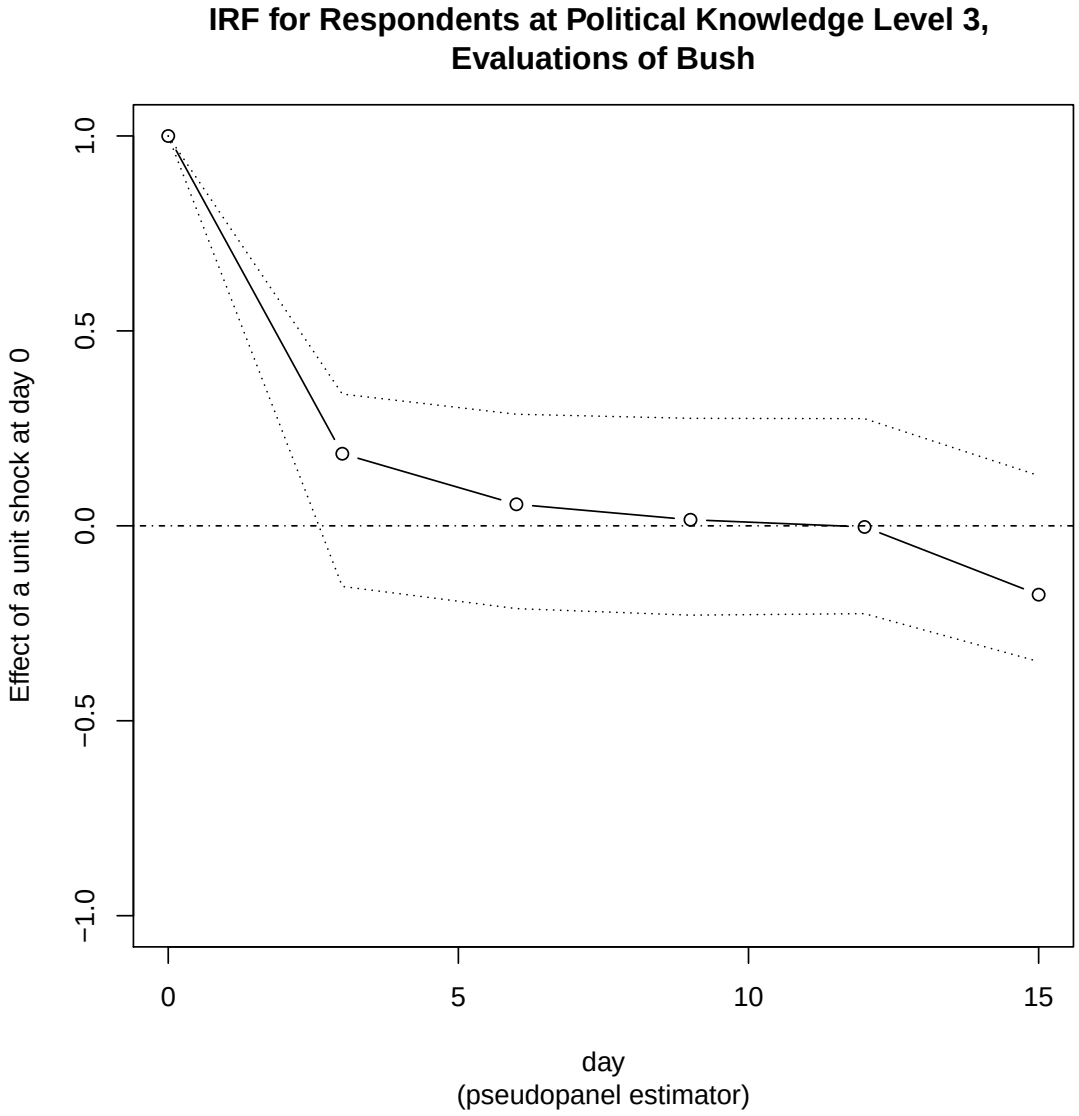


Figure 2.13: Nonparametric impulse response functions for Independent evaluations of Bush over 3 day periods at political knowledge level 4

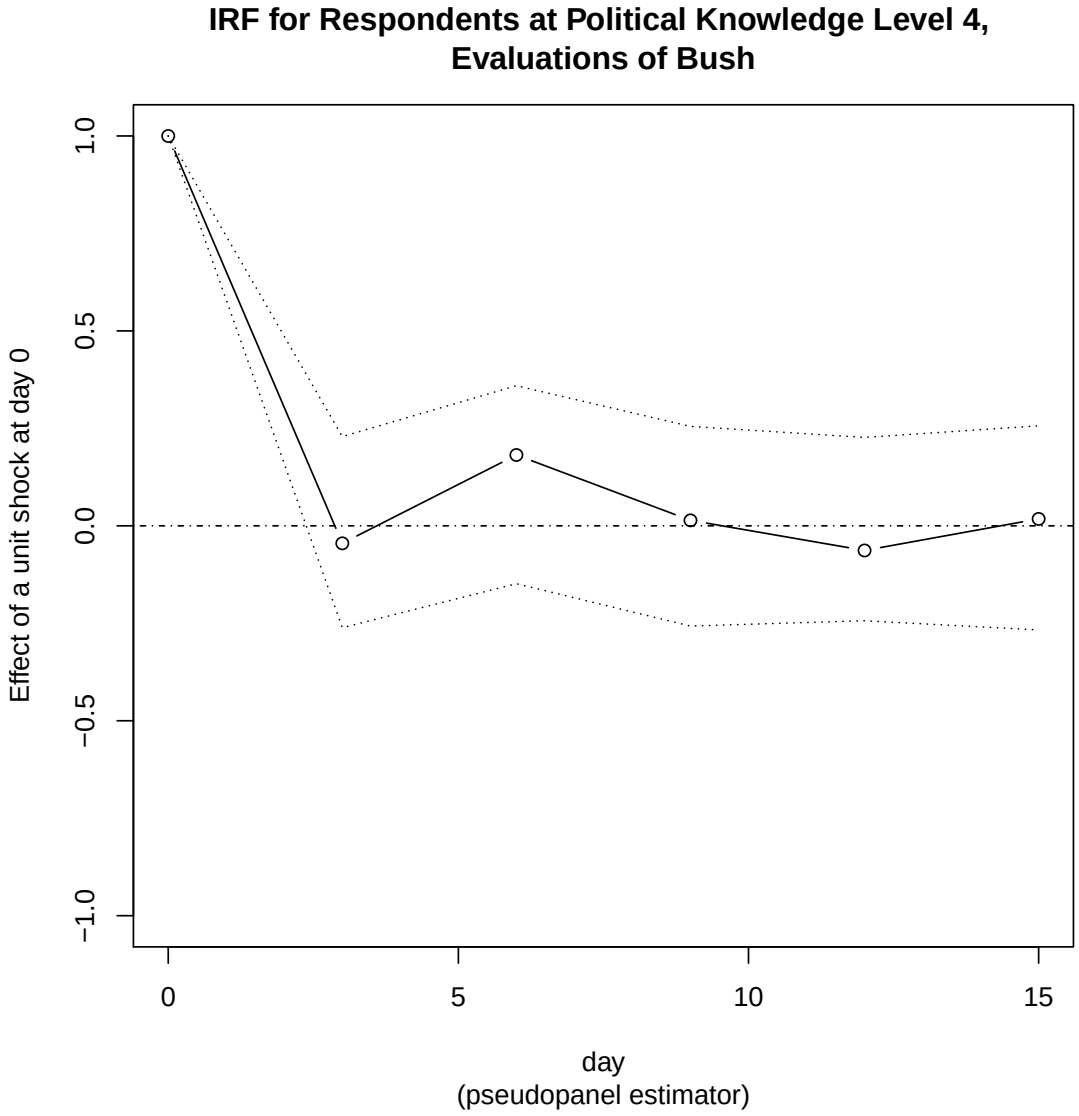
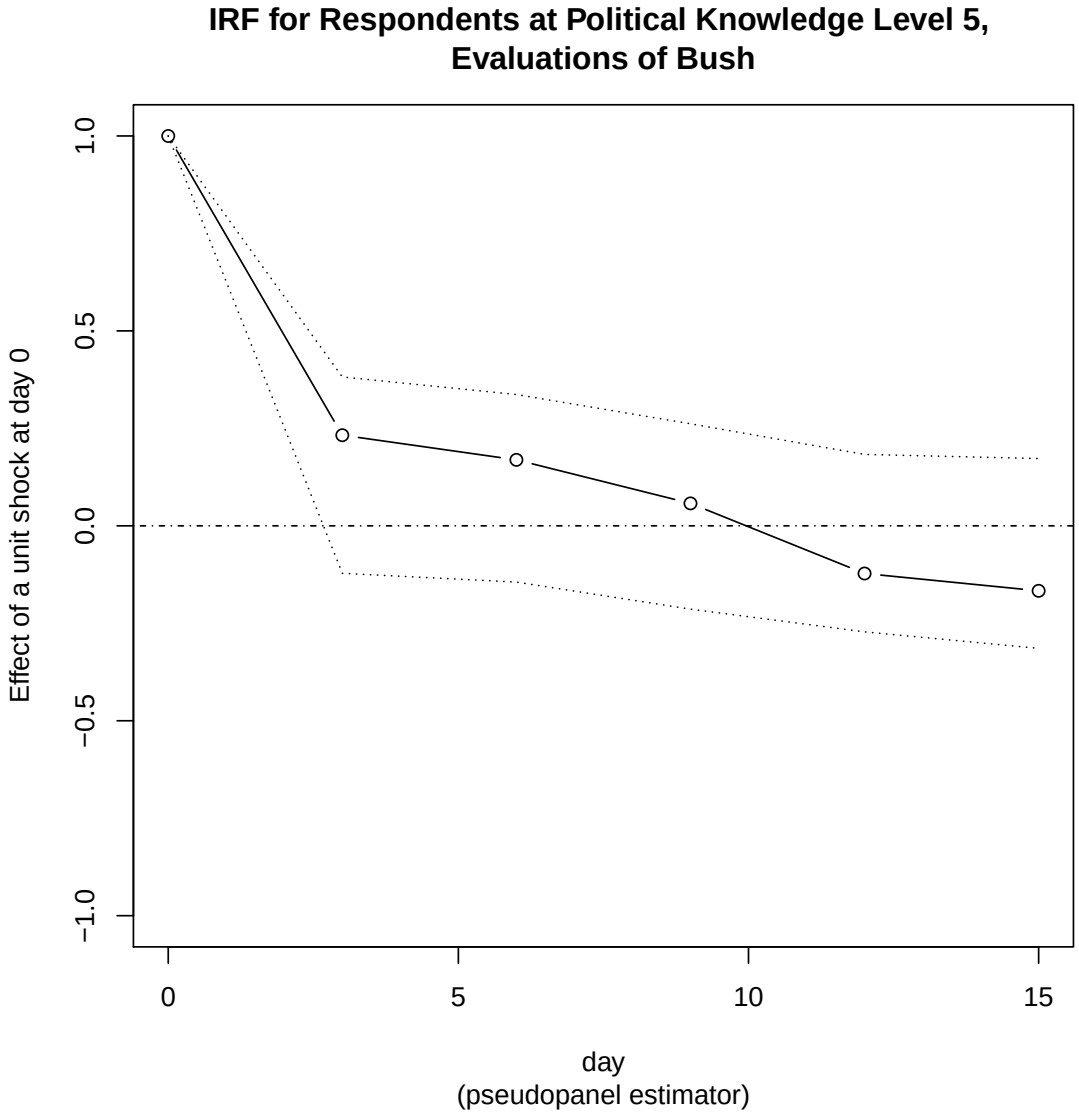


Figure 2.14: Nonparametric impulse response functions for evaluations of Bush over 3 day periods at political knowledge level 5



CHAPTER III

Approaches to Practical Bayesian Inference for Large-scale Random Effects Models

The multi-level dynamic model introduced in the previous chapter is useful because it represents many subtleties of campaign effects without throwing away important cross-sectional information. Political analysts might apply such models in circumstances involving rare events and occasional structural breaks, and/or in which units are subject to these phenomena on a basis contingent on their characteristics. While a classical approach such as maximum likelihood estimation provides a feasible option for inference, these models can exhibit multi-modality (Liu et al., 2003). Furthermore, the required sample size for asymptotic assumptions to be adequate (especially with correlated data) is unclear.

Motivated by this concern about classical approaches, I devote this chapter and the following one to the development of a Bayesian inference framework for general multi-level models and the investigation of computationally efficient (and therefore practical) inference strategies. Within the constraint of blockwise linearity and mode-separability, such models can include non-normal effects, non-normal responses, and even certain forms of non-linearity. The implication is that we can construct a global quadratic approximation for these models based on the least squares estimator. This

in turn is used in Gibbs and Pseudo-marginal Metropolis samplers for large-scale non-normal random effects models. In the remainder of this chapter, I detail the theory of and motivations behind the proposed framework.

Bayesian Computational Inference by Monte Carlo and Markov-Chain Monte Carlo

Bayesian inference methods are now widely used by political scientists in quantitative research, both to incorporate prior information or induce smoothing, and as a means for inference in cases where asymptotic approaches are intractable or untrusted. Bayesians augment likelihood models that express political theories with a set of priors that represent uncertain beliefs about the state of the world. Using these two ingredients, Bayesian inference applies Bayes Law to produce a posterior distribution that summarizes beliefs about the state of the world after exposure to the observed data:

$$f\left(\theta,\, \gamma \,|y\right) = \frac{f\left(y\,|\theta,\, \gamma\right) f(\theta,\, \gamma)}{f\left(y\right)}$$

In certain pairings of prior distributions and likelihoods—possessing conjugacy—the posterior of the parameters is in the same family of densities as the priors. In these cases, the posterior can be given in closed form so that its parameters can be exactly specified in terms of summary statistics of the data. Then inference can proceed using formulae much like the celebrated ordinary least squares formula of classical inference. It is tempting to choose a prior to be conjugate to achieve this simplification of the posterior distribution. However, prior beliefs are often difficult to elicit, and/or we wish to express prior ignorance as a matter of scientific method. Unfortunately, uninformative priors under conjugacy are often distributions near the edge

of an acceptable parameter space. The limiting distribution typically doesn't exist, and it's not clear how close to the edge we need to be sufficiently ignorant.

More principled strategies utilize priors that are selected via criteria other than conjugacy, and often lead to non-conjugate models. For example, one might select an uninformative prior by a principle such as maximum entropy, or—in the case of random effects which are actually nuisance parameters—based on an elicitation concerning the nature of the underlying events and common covariance (see Chapter 2). Prior distributions must be selected with care—for example, the function $\pi(\theta) = 1$ or some other function that conveys our beliefs without being a true density can sometimes serve as a prior—in which case it is called improper. However, if this also yields an improper posterior that does not integrate to one as a probability density must, then numerical or graphical summaries and model comparison are impossible.

With a non-conjugate model, we must find some way to draw inferences without having a closed form for the posterior. Conceptually, we are almost there just by computing the joint distribution, since Bayes' Law states that the conditional differs only by a normalizing constant. However, to develop a useful summary of the parameters, we must either obtain a sample from them or normalize the posterior. Usually, neither are easily obtained. It is hypothetically possible to normalize the entire posterior by numerical integration, but the number of function evaluations grows exponentially with dimensionality. No computational scheme claims to accurately integrate problems with dimensionality greater than about 100. Features can become become unimaginably difficult to measure in problems of this size. For example, the volume of a unit hypersphere in that dimension is on the order of 10^{-70} times smaller than that of the volume of a unit cube (Hahn, 2006). These are problems that hypothetically can be solved by Monte Carlo and quasi-Monte Carlo methods,

which benefit from estimation errors that theoretically decline with the square root of the sample size, in the case of Monte Carlo—or better, for quasi-Monte Carlo.

In recent years, political scientists and others using Bayesian inference have focused instead on Markov-Chain Monte Carlo (MCMC) sampling methods. Such methods provide well-known ways to exploit specific model structures in constructing posterior samplers. A conventional MCMC algorithm produces a dependent sequence of draws in which memory is only one time-step in duration: that is, conditioning on the previous time-step eliminates all dependency on prior time-steps. Asymptotically as the sequence grows to infinite length, each draw is a dependent draw from the true posterior.

Many political scientists use BUGS (Bayesian inference Using Gibbs Sampling) or variants such as JAGS (Just Another Gibbs Sampler) for MCMC and related sampling methods. Given a statement of a model in the form of a symbolic (or even diagrammatic) representation of the data-generating-process, these programs deduce posterior dependencies among the parameters and then use an expert system to construct a sampling algorithm to draw (or approximately draw) from the posterior distribution. Usefully, these tools can also automate many aspects of tuning and visualization.

BUGS or JAGS will identify conjugacy and deduce part of the posterior it can sample from whenever possible, but usually at least some part of the model requires a more general-purpose MCMC algorithm—e.g. one of the Gibbs, Metropolis, Metropolis-Hastings, or slice-sampling algorithms. Gibbs sampling is possible whenever one can sample from the complete conditional densities (or discrete probabilities) of every unobserved quantity represented in the posterior. These are the conditional posterior distributions given all of the other parameters and the observed data. A

simple Gibbs sampling update, given in Algorithm III.1, simply alternates draws from these densities, in some sequence, and replaces the current value with the sampled value. Amazingly, repeated alternation of such draws leads to a sequence that converges eventually to a sequence of draws from the joint posterior density.

Whenever the complete conditional distributions are unavailable, two major alternatives are the Metropolis and Metropolis-Hastings algorithms. The updates are given in Algorithms III.2 and III.3 (Metropolis et al., 1963; Hastings, 1970). (Metropolis-Hastings actually serves as a general form encompassing most Bayesian computational methods.) The basic idea of these algorithms is that an arbitrary proposal can be used for each parameter, but an acceptance step must be used to correct for discrepancies between the proposal distribution and the target posterior. MCMC algorithms must preserve a property called detailed balance to ensure that draws from the target distribution are only updated in a way that leads to another such draw. The difference between the two updates is that, to preserve detailed balance for non-symmetric proposals, the Metropolis-Hastings acceptance ratio includes a ratio correcting for any imbalance between the forward and reverse prpoosals. In contrast, the Metropolis algorithm applies to symmetric proposals and preserves detailed balance without any correction.

These basic strategies turn out to be difficult to apply to the very large, non-standard models discussed in the previous chapter. They can take many hours to yield an inference. Much of the problem is that, while chains produced by simple updates are guaranteed to converge under weak conditions, and often exhibit the essential property of ergödicity yielding an efficient sampling mechnism, the practicality of these methods depends on careful tuning and supervision for each specific data analysis. As a result, Bayesian methodologists have sought to develop globally

adaptive methods that automatically tune the samplers. Arguably with more success, they have developed alternative methods that can adapt locally to the many features of the posterior density at each location, such as the slice-sampler and diffusions such as the No-U-Turn sampler.

As a starting point, I make use of the mode-separable representation developed in Chapter 1 to produce a Blockwise Metropolis-Hasings sampler that alternately updates the hyperparameters and samples from an approximation to the conditional posterior distribution of the latent states. The updates for the hyperparameters are a symmetric random walk based on Algorithm III.2, whereas the updates for the random effects are non-symmetric and based on Algorithm III.3. The approximation is a multivariate elliptical draw whose mean and covariance match the normal-linear case, but having a Student tail to bound the true conditional posterior. Further details are available in the Appendix.[1]

In accordance with suggestions in the literature on Bayesian computation for sampling from random effects models, the Blockwise Metropolis-Hastings algorithm targets both the hyperparameters and random effects simultaneously. However, targeting both is not strictly necessary—and potentially problematic—when deriving inferences on the hyperparameters is the central goal. It is well-known that dependence between the hyperparameters and random effects leads to deficiencies that

[1]This sampling strategy is designed for a wide class of models: those whose random effects and likelihood are log-quasi-concave as a function of the random effects—conditioning blockwise on other parameters—and such that each are bounded by a known power function. In the simplest case, all random effects have quasi-concave densities that are a priori independent, and the likelihood is a linear or a monotonic transformation of the random effects. This implies that the linear-Gaussian approximation has a unique solution. To attain further generality such as a GARCH specification, to improve on the linear-Gaussian approximation via the method of scale mixtures, or to represent any other hierarchy of parameters, such conditions need only be met blockwise. In that case, all random effects can be initialized at random, optimized sequentially to locate a joint mode (a fixed point must exist due to conditional quasi-concavity), and then drawn sequentially from the elliptical approximation to that mode.

can delay convergence from poor starting points. In simple terms, small variance parameters conditionally imply small random effects and vice-versa (Gelman, 2008). A more subtle point, apparently unnoticed in past work, is that we should expect the computation of the Hastings ratio for updating the latent parameters to be computationally expensive, as it requires solving an additional inverse problem.

Based on my initial experiences with this sampler, I propose an alternative to Blockwise Metropolis-Hastings, the Pseudo-Marginal Metropolis strategy, which seeks convergence only for the hyperparameters despite the intractability of their marginal distribution.

Pseudo-Marginal Metropolis (PMM)

Algorithm III.1 Gibbs Update

for $i = 1$ to I **do**
  Draw $\vartheta_i' \sim f\left(\vartheta_i' \middle| \vartheta_{-i}\right)$
  $\vartheta_i \leftarrow \vartheta_i'$
end for

Algorithm III.2 Metropolis Update

Draw $\vartheta' \sim q\left(\vartheta' \middle| \vartheta\right)$ with $q\left(\vartheta' \middle| \vartheta\right) = q\left(\vartheta \middle| \vartheta'\right)$
if $\text{Unif}\left(0,1\right) < \max\left(1, \frac{f(\vartheta')}{f(\vartheta)}\right)$ **then**
  $\vartheta \leftarrow \vartheta'$
end if

Algorithm III.3 Metropolis-Hastings Update

Draw $\vartheta' \sim q\left(\vartheta' \middle| \vartheta\right)$
if $\text{Unif}\left(0,1\right) < \max\left(1, \frac{q(\vartheta|\vartheta')f(\vartheta')}{q(\vartheta'|\vartheta)f(\vartheta)}\right)$ **then**
  $\vartheta \leftarrow \vartheta'$
end if

The Pseudo-Marginal Metropolis approach is introduced in recent papers pursuing hybrids of Monte Carlo and MCMC methods in challenging substantive applications (Beaumont, 2003; Andrieu, et al., 2010). The basic idea is an extension of

Algorithm III.4 Pseudo-Marginal Metropolis (General) Update

Draw $\vartheta' \sim q(\vartheta' | \vartheta)$ with $q(\vartheta' | \vartheta) = q(\vartheta | \vartheta')$
$J' \leftarrow f(\vartheta') + \epsilon$, with $f(\vartheta') = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\vartheta', \gamma)\, d\gamma$, $\mathrm{E}[\epsilon] = 0$
if $\mathrm{Unif}(0,1) < \max\left(1, \frac{J'}{J}\right)$ **then**
    $\vartheta \leftarrow \vartheta'$
    $J \leftarrow J'$
end if

Algorithm III.5 Pseudo-Marginal Metropolis (IS) Update

Draw $\vartheta' \sim q(\vartheta' | \vartheta)$ with $q(\vartheta' | \vartheta) = q(\vartheta | \vartheta')$
Draw $\gamma' \sim r(\gamma' | \vartheta)$
$J' \leftarrow f(\vartheta', \gamma) / r(\gamma' | \vartheta)$
if $\mathrm{Unif}(0,1) < \max\left(1, \frac{J'}{J}\right)$ **then**
    $\vartheta \leftarrow \vartheta'$
    $J \leftarrow J'$
end if

the Metropolis algorithm targeting a marginal distribution by replacing evaluations of the posterior density with unbiased estimates (Andrieu & Roberts, 2009; credited to Beaumont, 2003 as Grouped Independence Metropolis Hastings). A basic Pseudo-Marginal Metropolis algorithm is listed in Algorithm III.4. It is straightforward to show that the resulting Markov chain still converges in distribution to f at stationarity.

An important technicality should be noted, as it is crucial to achieving a convergent Pseudo-Marginal Metropolis simulation. In the three listings referenced above, each time a draw is accepted, the new approximate marginal likelihood J' is stored in J. In the Metropolis-Hastings algorithm, such a scheme is traditionally used for computational efficiency—though one might contend that it mars the theoretical clarity of the presentation. However, if one replaces posterior evaluations with approximate posterior evaluations to form a Pseudo-Marginal Metropolis-Hastings algorithm, each new proposal must be compared to the same approximate value associated with the first acceptance of the present state. Andrieu & Roberts (2009) show

that, if we generate a similar algorithm in which the approximations for the initial state are regenerated at each step, the resulting chain only approximately converges to the target stationary distribution.

The motivation for the Pseudo-Marginal Metropolis algorithm is that a convergent simulation can be developed under particular conditions even for posterior distributions that are only available in approximate form. This applies to multi-level modeling, in which one can evaluate the posterior of a joint model including all of the random effects parameters, but one often cannot evaluate the marginal distribution directly. This difficulty of marginalization is one of the basic motivations for utilizing the random effects representation of the multi-level model. However, rather than sampling the random effects, Pseudo-Marginal Metropolis only requires an unbiased approximation to the integral.

Where Pseudo-Marginal Metropolis is possible, it remains to be seen whether it leads to a useful standalone strategy in the Bayesian toolbox. Scientists might consider marginalization of parameters that are uninteresting or whose costs of the regeneration from a well-mixed chain at stationarity are small relative to the costs involved with obtaining that chain, and when the benefits of reducing correlations in the posterior outweigh the costs of marginalization. Multi-level models with large numbers of random effects could therefore be an exemplary application of Pseudo-Marginal Metropolis, because the random effects are nuisance parameters and are often highly correlated with the random effects. At present, it has been applied to an intractable posterior distribution in modeling geneological histories (Beaumont, 2003) and variable selection problems (Andriu & Roberts, 2009). It also serves as a central step in a technique denoted Particle Markov-Chain Monte Carlo (Andrieu et al., 2010)—a strategy for propagating a sequence of sample-based approximations

corresponding to an evolving, globally intractable posterior (e.g., in a non-Gaussian state-space model) in a way that ensures they eventually evolve in accordance with that posterior. Another reason to consider Pseudo-Marginal Metropolis is the possibility that past computations can be recycled in models with certain structure, an idea that has already been exploited in the particle filtering and population Monte Carlo literatures.

Finally, two important and subtle computational simplifications are available, which are exemplified in Algorithm III.5. Here an importance sampler is used to construct the unbiased approximation to the marginal, as will be detailed in the next section. However, one aspect is immediately evident: the approximation to the marginal can be based on auxillary draws that only depend on the current state of the hyperparameters. There is thus no need to find a good approximation to the conditional posterior for every new value of the hyperparameters that is proposed. This yields a vital computational simplification over the Blockwise Metropolis-Hastings. Furthermore, because Blockwise Metropolis-Hastings seeks to converge in the state variables γ' based on the modal approximation, and this particular draw is non-symmetric and uses III.3, the Hastings ratio involving a computationally expensive inverse of the previous state variable γ is required. in contrast However, in contrast to the above instances, in multi-level models the joint posterior itself is generally available, ensuring that any form of Metropolis-Hastings can be applied.

The question seems to be, do these advantages pan out in terms of accelerating convergence, either based on computational simplification, or larger steps in a smaller space? An empirical comparison of the two methods examines this question in Chapter 4.

Monte Carlo methods and Importance Sampling

The way in which Algorithm III.5 works to approximate the marginal in an unbiased way is as a Monte Carlo approximation, specifically an importance sampler. Monte Carlo methods seek to approximately evaluate integrals by using the fact that they can be estimated by weighted averages of random functions. The most general set of efficient methods are importance samplers, which use weights that are the inverses of the sampling density to estimate the expectation of a function. To be efficient, an importance sampling density must concentrate samples in places where the function to be integrated, the integrand, is large. Monte Carlo does perform better than the most naive cubature strategy, in which univariate numerical integration (quadrature) is performed for every combination of all variables of integration (in this case, the fixed and random effects γ) except the innermost one. This is a strawman but it illustrates an important consideration when designing or selecting these methods. A nested set of quadrature steps leads to an exponential growth in computing costs—an instance of the "curse of dimensionality." The number of equal-sized volume units expands exponentially with dimension. Amazingly, Monte Carlo with a fixed proposal exhibits geometric convergence—as with sample means—that does not depend on the dimensionality of the problem, provided the estimates have a finite variance.

Numerical methods such as quadrature can easily miss features of the integrand where sampling is insufficient. A more systematic description of such failures is available in Monte Carlo in cases in which the integrand has an infinite domain or range and decays very slowly (e.g. polynomially). If the importance function decays very quickly in comparison (e.g. exponentially), the tails of the target function will not be adequately explored, and very large importance weights will occur frequently.

Formally, the variance of importance weights and importance sampling estimates is infinite in this case, and a variety of methods have been proposed to stabilize importance sampling estimators (see, e.g. Owen & Zhou, 2000).

The potential for infinite importance weights arises only with unbounded integrands. Transformation of the domain of the integrand is a popular strategy in both quadrature and Monte Carlo methods for yielding a domain that is more convenient. Unfortunately, for unbounded intervals in the variable of integration x, transformation of the domain of integration to a bounded domain is not sufficient to ensure that Monte Carlo importance weights are stabilized. The reason is that the transformed function remains unbounded, and hence a poor choice of importance distribution can still produce infinite variance weights. This arises, for example, in trying to compute the integral of $x^{-\frac{1}{2}}$ on [0, 1] by uniform sampling, a method that does not work because even though the integral exists, the importance weights remain unbounded and diverge as $x \to 0$ (Evans & Swartz, 2000).

The importance of tail-boundedness in computational strategies for estimation in multi-level models is thus to absolutely ensure finite importance sampling estimates. This means that we can select an importance function such that the ratio of the integrand to the importance function—the importance weight—has a finite limit at the ends of the domain of integration. In our case, the Student-t proposal distribution or a power distribution with the same power parameter setting as the largest tail bound will satisfy the required criterion.

CHAPTER IV

Comparing MCMC Methods for Non-Normal Multi-Level Models

In the previous chapter, I discussed the role of modal approximations in facilitating the application of Bayesian computational methods to large, non-normal multi-level models. Past work suggests that classical algorithms often performed poorly in such models, and this motivates consideration of other alternatives. A variety of samplers may be credible once we consider designs that focus on the hyperparameters and treat the random effects merely as auxillary parameters (see, e.g., Storvik, 2011). The Pseudo-Marginal Metropolis sampler, even with an importance sample of size 1, places a less stringent detailed balance requirement on the MCMC chain, and therefore may gain both speed and accelerated convergence. However, as noted in the previous chapter, both samplers have asymptotically the same computational compelexity. In this chapter, I undertake an empirical comparison of the performance of the two samplers in posterior draws from a medium-scale Student-t random effects model. To anticipate, I find favorable results for the convergence of the Pseudo-Marginal Metropolis sampler relative to the Blockwise Metropolis-Hastings sampler. However, I also observe that both samplers mix poorly, the Pseudo-Marginal Metropolis sampler especially poorly. I discuss two algorithms,

Deterministic Mixture Importance Sampling and Annealed Importance Sampling, that may better tackle the problem of mixing. In the latter case, the algorithm also more fully leverages the underlying mode-separable model structure.

Research Design: Sampler Convergence on a Medium-Scale Non-Normal Random Effects Model

I study the performance of the Blockwise Metropolis-Hastings and Pseudo-Marginal Metropolis samplers on an instantiation of a non-normal random effects model—a specification with an intercept, a unit-level factor, and two independent residuals in each observation equation, one Student-t group effect and one normal observation error. The medium-scale model consisted of 80 random effects, associated with five observations each. The data generating process is summarized as follows:

$$y_{i,t} = \beta_0 + \beta_1 x_i + \gamma v_t + \sigma \epsilon_i$$

$$v_t = v_{t-1} + \theta u_t$$

$$\epsilon_i \sim N\left(0,\, 1\right)$$

$$u_t \sim Student_3$$

Since the challenging aspect of conventional sampling the hyperparameters in multi-level models is escaping an area of the posterior where the random effects variance is very small, I deliberately sampled the initializing log standard deviation parameters uniformly on $[-3, 2]$, even though the true log standard deviations in both models were equal to 1. This implied that the starting points were about three times too often smaller than the true value than larger than the true value, making the sampling process often challenging even for a univariate posterior. All samplers were initialized with the same starting values for each of their chains.

Examining the performance of these samplers on a dataset simulated from this model requires the exploration of a large number of potential settings of the tuning parameters, using multiple chains per setting to obtain a convergence diagnosis. Each chain must be run until convergence is evident, and possibly long enough to resolve any ambiguity in the relative performance of the two samplers. As a result, the selected specification is simpler than that of many applied contexts. Reassuringly, I have observed that the two samplers behaved similarly when additional random slopes are added, with larger or smaller datasets and numbers of random effects, and with differing distributions for the random effects.

To enable a fair comparison between the two samplers, each is coded as similarly as possible using the 'R' software environment for statistical computing (R Core Team, 2012). Note, however, that as discussed in the previous chapter, the Hastings ratio is simplified in the Pseudo-Marginal Metropolis computation. That is, while both samplers require the computation of the probability of each state draw conditional on the current set of hyperparameters, only the Blockwise Metropolis-Hastings sampler requires the computation of the density of the previous auxillary state under each new hyperparameter proposal. Concretely, this means that each cycle of draws from the Blockwise Metropolis-Hastings sampler requires an additional 'backsolve' operation in comparison to the Pseudo-Marginal method.

Another important aspect of the design was tuning. Both algorithms featured a tunable scale for the proposal distribution for sampling each of the two hyperparameters, the log response scale and the log random effects scale. To reduce the search-space of such settings, I constrained the proposal standard deviation for the log response scale to be one-half of the proposal standard deviation for the log random effects scale. This roughly matched the best tuning located during pilot runs. A sep-

arate tuning parameter scaled the eliptical distribution of the random effects drawn at each iteration. For the Blockwise Metropolis-Hastings algorithm, this was set (by definition) to one, whereas for the Scaled Blockwise Metropolis-Hastings Algorithm, this was allowed to vary freely. I allowed this same tunable scaling of the auxillary variables playing the corresponding role of random effects in the Pseudo-Marginal Metropolis algorithm.

To measure convergence, I employed the Potential Scale Reduction Factor $\sqrt{\hat{R}}$ (Gelman & Rubin, 1992) for the log random effects scale hyperparameter. This is the square root of the ratio between two estimates of the combined variance across the ten chains: the mean of all of the estimates of the within-chain variances, and the cross-chain estimate based on the law of total variance. At convergence, the chains are independent and the square root of the ratio of the estimators converges to one, whereas prior to convergence the ratio is greater than 1. The R package 'coda' is used to compute the estimate and a 95% upper credible interval (based on the upper bound of a two-sided 90% credible interval) will be used to examine the evidence that the chains produced by one algorithm have converged more rapidly than those produced by another. Typically a value $\sqrt{\hat{R}} < 1.1$ is considered indicative of convergence.

Many studies of computational performance consider only whether an objective has been achieved by a lesser number of iterations for one algorithm versus another. However, it was quickly evident that the three samplers, even when coded as comparably as possible, did not complete a single iteration in the same amount of CPU time, evidently due to the extra cost described above of computing the 'backsolve' operation used in the Hastings Ratio computation of the Blockwise Metropolis algorithms. On the other hand, the Pseudo-Marginal-Metropolis algorithm tends to

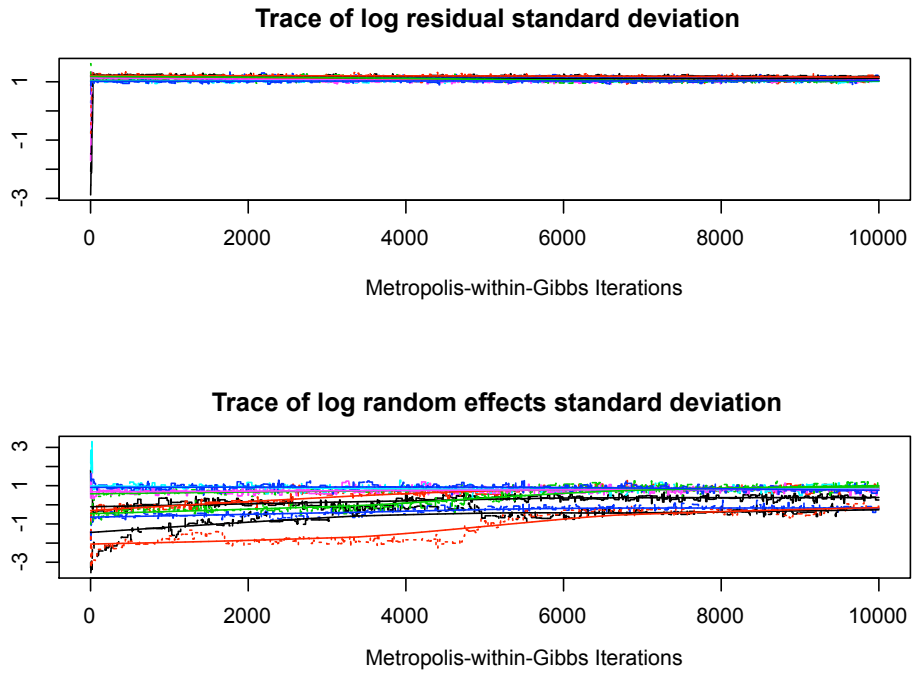


Figure 4.1: Parameter trace, Blockwise Metropolis-Hastings chains

exhibit much lower acceptance rates for its hyperparameters. Therefore, in addition to assessing the convergence of the samplers on an iteration by iteration basis, the samplers were compared in their progress towards convergence as a function of CPU time based on $\sqrt{\hat{R}}$.

Results and Discussion

Figures 4.1, 4.2, and 4.3 show the evolution of the ten chains of the best performing samplers over the 10000 iteration simulations. All three samplers quickly generated values for the log response scale that were near the mode. However, the convergence of the three algorithms on the log random effects scale parameter

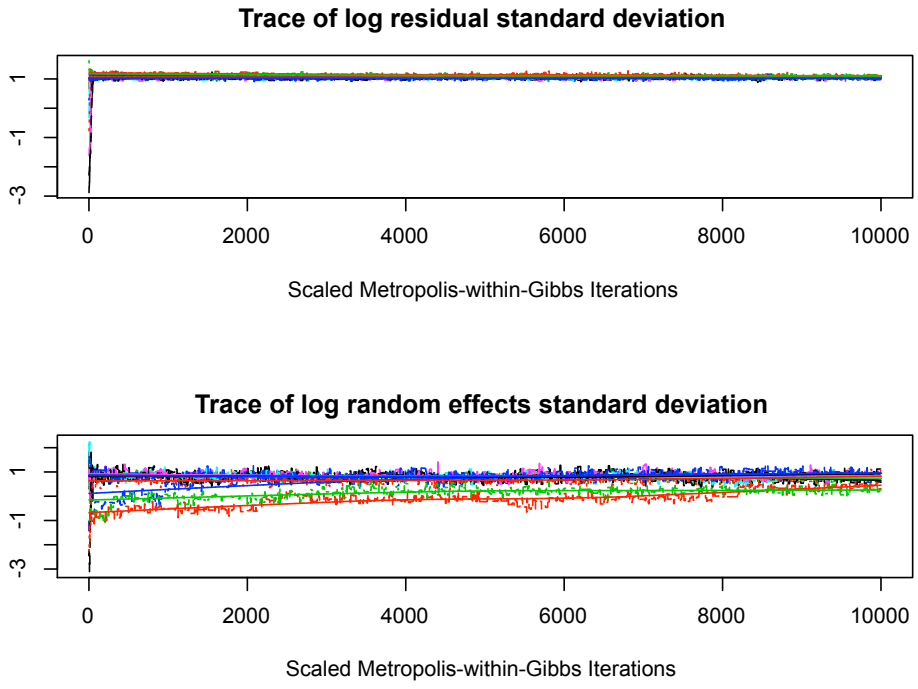


Figure 4.2: Parameter trace, Scaled Blockwise Metropolis-Hastings chains

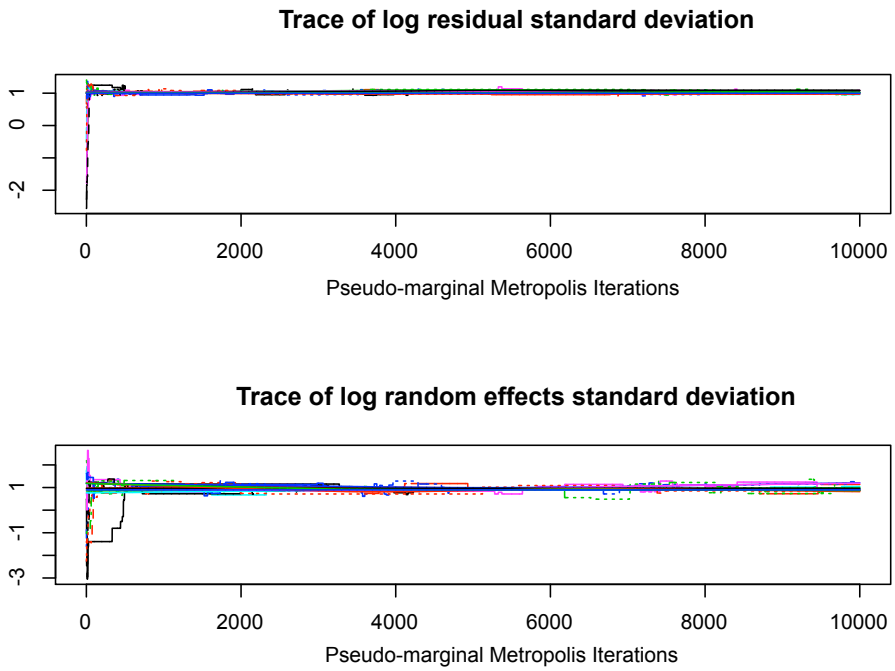


Figure 4.3: Parameter trace, Pseudo-Marginal Metropolis-Hastings chains

was markedly different. Those chains generated by the optimally-tuned conventional Blockwise Metropolis-Hastings sampler that started at parameter values far from the true value typically made an initial jump towards the mode, and then converged very slowly for a large number of iterations, consistent with the reported convergence difficulties in the literature. Only chains that were initialized quite close to the mode appeared to converge quickly. The optimally-tuned Scaled Blockwise Metropolis-Hastings sampler performed somewhat better—these chains initially jumped about halfway closer to the mode than the conventional sampler. Nonetheless, like the chains generated by the conventional sampler, the chains generated by the scaled sampler converged slowly for a large number of iterations. In contrast, all of the chains generated by the Pseudo-Marginal Metropolis algorithm jumped very quickly to the region near the mode.

Figure 4.4 shows the convergence of the chains generated by the three optimally tuned samplers as a function of CPU time. It is clear that, despite its poorer acceptance rate, the rapid convergence on an iteration-by-iteration basis as well as the savings in computational costs for the Pseudo-Marginal-Method led to a much better performing sampler in terms of its computational efficiency. None of the potential scale reduction factors of the three 10-chain samples reach a typical threshold like 1.1 within their 10000 iteration spans. However, the chains generated by the Pseudo-marginal Metropolis algorithm converge most quickly in computational time, with their PSRF upper credible bound contained below 2 within 18 seconds, a level never reached in the over two minute runs of other samplers. To be clear, both of the lesser-performing algorithms eventually converged when the simulations were repeated using a much larger number iterations.

Convergence and mixing behavior for chains produced by the three samplers are

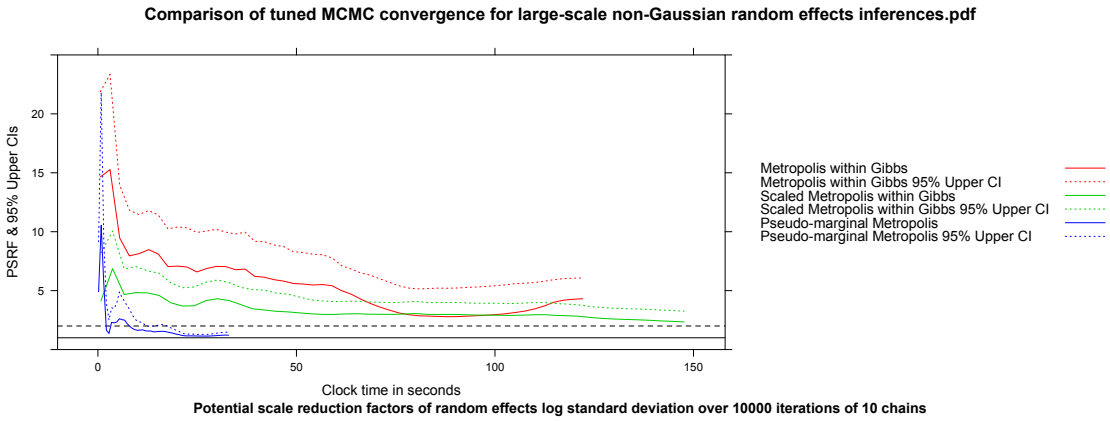


Figure 4.4: Convergence of chains, by algorithm, under optimal tuning

Chain	Random Effects Draws	Effective Yield	Acceptance Rate	ESS/s
Blockwise Metropolis-Hastings	1	0.094	0.356	3.198
	4	0.076	0.382	1.993
Scaled Blockwise Metropolis-Hastings	1	0.063	0.161	4.219
	4	0.037	0.082	2.010
Pseudo-Marginal Metropolis	1	0.014	0.004	4.286
	4	0.021	0.011	2.055

Table 4.1: Mixing properties of chains, by algorithm, under optimal tuning

very different (this is reported by other researchers using other algorithms, e.g. Hoffman & Gelman, 2011), and I found that the mixing of the optimally converging chains was quite poor. Responding to the poor mixing, I investigated further by initializing chains near equilibrium, and confirmed that the Pseudo-Marginal Metropolis algorithm was accepting fewer than 1% of the draws, as shown in Table 4.1.

In an attempt to stabilize the algorithms, I expanded the space of tuning parameters to include allowing each algorithm to use more than one replication per iteration. I considered the cases of either one or four importance draws per iteration for Pseudo-Marginal Metropolis, and either one or four averaged chains of the random effects for Blockwise Metropolis-Hastings. The Pseudo-Marginal Metropolis numerator ought to more closely mimic the true marginal distribution with an increasing number of replications. More speculatively, the Blockwise Metropolis-Hastings algorithm might perform better if conditional draws for the hyperparameter were based on the less variable mean over multiple random effects. Summaries of mixing for settings allowing four replications are also shown in Table 4.1. Here the effective yield is the effective sample size of each iteration, whereas the effective yield per second is the effective sample size per iteration, divided by the number of seconds it took to sample one iteration.

For the best of the tunings over single chains of 1000 iterations, I found that, despite a relatively low acceptance rate and yield in effective samples per iteration, the pseudo marginal method was still competitive in effective samples per second. An expanded study design would be required to see whether the pseudo-marginal method is definitively more effective at mixing than the other two samplers, since the best setting of the tuning parameter was on the edge of the simulation parameter space. Despite the very low acceptance rates of the Pseudo-Marginal Metropolis chain,

attempts to improve its mixing (or that of the Blockwise Metropolis-Hastings chains) via multiple replications of random effects draws during each iteration produced no improvement in computational efficacy.

Improving the Pseudo-Marginal sampler: Deterministic Mixture Importance Sampling

The Pseudo-Marginal Metropolis method described above converges rapidly despite making use of a very crude importance sampling scheme. To reduce the number of inverse problems that need to be solved, and hence the constant in front of the np^2 in its computational complexity, importance sampling is based on the conditional posterior mode of the random effects, given the *current* settings of the hyper-parameters. As noted above, the sampler exhibited poor mixing properties, and drawing repeated importance samples did not stabilize the sampler. This is contrary to the literature suggesting that better importance sampling estimates improve the performance of the Pseudo-Marginal samplers (Andrieu & Roberts, 2009), but it may be that the benefits are so slight as to be undetectable without a much better proposal distribution or importance sample size.

While it is impressive that the single importance draw version of the Pseudo-Marginal sampler quickly located regions of high marginal mass, it remains to be seen whether other features can be exploited to better balance the tremendous economy of the simple version above (i.e., allocating minimal resources to proposals that are likely to be rejected anyway) with improved mixing (i.e., by increasing the quality of the proposals, and thus improving importance sampling and attaining the benefits of dimensionality reduction).

As Storvik (2011) notes, pseudo-marginal methods can allow dependence among

auxillary draws. This leads to the possibility that the sampler can be improved by redefining the notion of a replication. Rather than draw j repeated sets of random effects at each iteration, as in the replicated version used in the mixing experiment above, instead $j-1$ past sets and 1 present set could be stored. Then, at each proposal step, the marginal distribution would be estimated using all j sets. I have labeled this a deterministic mixture importance sample since it is based on the idea that an importance sampler can be built by simply retaining the weights and drawing from a set of importance samplers. In this case, the set of importance samplers would be draws from a sequence of approximations to the conditional posterior over multiple steps. The latter ought to increasingly approximate the marginal distribution of the auxillary variables, and might thereby facilitate jumps of the hyperparameters through a wider range of the space of the random effects than any single auxillary draw. Once a hyperparameter proposal is accepted, the oldest set of random effects can be replaced by the newest. To correct some of the bias in the empirical distribution estimator for the random effects, each set of stored random effects can be given a weight corresponding to the number of times the state was retained.

The asymptotic computational complexity of this deterministic mixture importance sampling (DMIS) version of the pseudo-marginal metropolis algorithm is compared to other approaches in Table 1 of Chapter 3. A promising feature is that, if indeed recycling past draws of the random effects yields a more global importance distribution, it does so very frugally. The additional cost is only that of the additional function evaluations that would be required in pairing these random effects with new hyperparameters. Despite the additional dependency between the estimated marginal posterior values used in the Metropolis acceptance ratio, the numerator and denominator of the ratio remain unbiased. Technically, the sampler

remains non-adaptive because dependence on past draws disappears conditionally on the past j sets of random effects. It remains to be proven that such a scheme converges to the correct marginal distribution.

Improving the Pseudo-Marginal Sampler: Annealed Importance Sampling

The previous methods employ a normal approximation to the conditional posterior as a basis for importance sampling. When a model contains a large number of non-normal random effects relative to the amount of data, there is reason to expect the effectiveness of the approximation to decline. The Bayesian Central Limit Theorem implies that the posterior approaches normality only as the number of terms in the likelihood grows relative to the size of the parameter space. Indeed, a central motivation for applying Bayesian estimation to large non-normal random effects models, rather than a simpler method, is the expectation that the non-normality makes a critical difference to inference!

Given the concern that the conditional posterior under the normal approximation is expected to deviate substantially from that under the true model, we might ask whether there is a way to improve the approximation at minimal cost. Annealed importance sampling (Neal, 2001) is a method for importance sampling from potentially multi-modal distributions. It is readily applicable in the present case bacause of the assumptions of mode-separability and tail-boundedness introduced in Chapter 1. In particular, because we can readily identify the modes of the components along any direction in the parameter space, we can construct lower and upper quadrature grids in each direction, as well as bounding functions for the tails. The combination of the grids and bounding functions for the tails enable us to both draw approximate

samples from the conditional posterior along each dimension, and compute an unbiased importance sampling estimate of the integral in that direction as well as an upper bound on the variance of this estimate. The accuracy of the approximation is determined by the bounding grid, so the "annealed" part of the algorithm is our ability to progressively refine the grid until the approximate marginal is sufficiently precise.

Of course, using this hybrid of cubature and random sampling to design an annealed importance sampler will only be effective if it is computationally efficient relative to the equivalent number of poor quality, but cheap, Metropolis-Hastings steps. A reason for cautious optimism is that for unimodal component functions (as we have assumed), the quadrature grids can be reused during each iteration of the algorithm. Furthermore, the targeted precision of the AIS will be its tuning parameter, and this seems like it could lead to an effective algorithm. In particular, we might pick a level of precision that suffices for areas of relatively high posterior density in the marginal posterior of a multi-level model, and thereby spend very little computational time refining the grids when we are far from the mode.

The Appendix lists a prototype algorithm for Annealed Importance Sampling.

CHAPTER V

Conclusion

In this dissertation, I first examined how to go beyond the collapsed campaign to conduct a detailed study of heterogeneity in citizens' responses to political campaigns. In Chapter 2, I examined substantive theories of response variance and response persistence, and derived hypotheses relating levels of political sophistication to the persistence of citizens responess. I then tested these hypotheses using one edition of the National Annenberg Election Study (Romer 2000), using pseudo-panel methods to identify the mean lag length of the dynamics of Gore and Bush evaluations for respondents at various levels of political knowledge. I also conducted a fully parametric investigation of these relationships. These analyses generally found no relationship, but represent ground-breaking efforts to simultaneously estimate spatio-temporal heterogeneity and dynamics. In particular, though the present case was limited to cohorts consisting of five levels of political knowledge, the author's Hetdyn software package for R used for the final parametric analysis at the end of Chapter 2 allows for responses to depend on multiple types of dynamics and individual-specific parameterizations.

A major shortcoming of the empirical analyses in Chapter 2, in my view, is the reliance on semi-parametric estimators or the assumption that dynamics are nor-

mally distributed. These lead to tractable inferences but are at least inefficient, and probably very misleading if campaign effects typically look more like rare structural breaks. Chapters 3 and 4 investigate the pseudo-marginal method as a strategy for accelerating parametric Bayesian inference for non-normal models. Dropping the assumption of normality is expected to enable automatic estimation of rare but large structural breaks, possibly enabling us to better capture campaign effects without a priori specification of which points in time led to substantial persuasive impact. In Chapter 4, I find that the pseudo-marginal method, which replaces evaluations of the marginal with an unbiased approximation thereof, is substantially faster (in CPU time and iteration counts) than a generic Blockwise Metropolis-Hastings algorithm that alternates between draws of random effects and hyperparameters on the entire parameter space. The ultimate goal of these efforts, currently part of the author's QuasiModal package for R, is to develop Bayesian techniques that are fast and flexible enough for widespread use by political scientists.

BIBLIOGRAPHY

Andrieu & Roberts, 2009. "The Pseudo-Marginal Approach for Efficient Monte Carlo Computations." *The Annals of Statistics,* 37(2):697-725.

Andrieu, C., Doucet, A. & Holenstein, R., 2010. "Particle Markov chain Monte Carlo methods." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72: 269-342.

Ardia, David, Lennart F. Hoogerheide, & Herman K. Van Dijk, 2009. "Adaptive Mixture of Student-t Distributions as a Flexible Candidate Distribution for Efficient Simulation: The R Package AdMit." *Journal of Statistical Software.* 29(3), 1-32.

Bartels, Larry, 2002. "Beyond the Running Tally: Partisan Bias in Political Perceptions." *Political Behavior.* 24(2): 117-150.

Beattie, Christopher, 2007. "Matrix Factorizations, and Direct Solution of Linear Systems." In Leslie Hogben (ed.), *Handbook of Linear Algebra.* Chapman & Hall/CRC Press, Boca Raton, FL.

Beaumont, M., 2003. "Estimation of Population Growth or Decline in Genetically Monitored Populations." *Genetics.* 164(3): 1139-1160.

Bornkamp, Björn, 2011. "Approximating Probability Densities by Iterated Laplace Approximations." *Journal of Computational and Graphical Statistics*. 20: 656-669.

Box-Steffensmeier, Janet M., Darmofal, David, & Christian A. Farrell, 2009. "The Aggregate Dynamics of Campaigns." *The Journal of Politics*. 71(1), 309–323.

Brady, Henry E., 1985. "The Perils of Survey Research: Inter-Personally Incomparable Responses." *Political Methodology.* 11:269-291.

Brady, Henry E., and Richard G. C. Johnston. 2006. *Capturing Campaign Effects.* University of Michigan Press.

Hahn, 2006. "Multidimensional numerical integration with Cuba." In J. Blümlein, W. Friebel, T. Naumann, D. Perret-Gallix, T. Riemann, P. Wegner (eds.), *Advanced Computing and Analysis Techniques in Physics Research: Proceedings of the X International Workshop on Advanced Computing and Analysis Techniques in Physics Research.* 559(1).

De Boef, Suzanna, & Luke Keele, 2005. "Taking Time Seriously." *American Journal of Political Science.* 52(1):184-200.

Deaton, A., 1985. "Panel Data from Time Series of Cross Sections," *Journal of Econometrics*. 30:109-126.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. "Least Angle Regression." *Annals of Statistics* 32(2): 407–499.

Evans, Michael John, & T. Swartz. 2000. *Approximating Integrals Via Monte Carlo and Deterministic Methods.* Oxford University Press.

Finkel, Steven E., 1993. "Reexamining the 'Minimal Effects' Model in Recent Presidential Elections." *Journal of Politics.* 55(2):1-21.

Franklin, Charles, 1989. "Estimation across Data Sets: Two-Stage Auxillary Instrumental Variables Estimation." *Political Analysis.* 1(1):1-23.

Gelman, Andrew, & Gary King, 1993. "Why Are American Presidential Election Polls So Variable When Votes Are So Predictable?" British Journal of Political Science 23(3):409-51.

Genz, Alan, & Robert E. Kass. 1997. "Subregion-Adaptive Integration of Functions Having a Dominant Peak." *Journal of Computational and Graphical Statistic.* 6(1): 92-111.

Gerber, Alan S., Gimpel, James G., Donald P. Green, & Daron R. Shaw, 2011. "How Large and Long-lasting Are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment." *American Political Science Review.* 105(1): 135-150.

Grimmett, Geoffrey, and David Stirzaker. 2009. *Probability and Random Processes.* Oxford University Press.

Hastie, Reid, & Park, Bernadette. 1986. The relationship between memory and judgment depends on whether the task is memory-based or on-line. *Psychological Review.* 93:258-268.

Hoffman, Matthew D., and Andrew Gelman. 2011. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." arXiv e-print (arXiv:1111.4246)

Holbrook, Thomas M. 1996. *Do Campaigns Matter?* Sage. Thousand Oaks, CA.

Holbrook, Thomas M., & Scott D. McClurg. 2005. "The Mobilization of Core Supporters: Campaigns, Turnout, and Electoral Composition in United States Presidential Elections." *American Journal of Political Science.* 49(4):689-703.

Huber, Gregory A., & Kevin Arceneaux, 2012. "Identifying the Persuasive Effects of Presidential Advertising." *American Journal of Political Science.* 51(4):961–981.

Iyengar, Shanto, & Donald Kinder, 1987. *News that Matters: Television and American Opinion.* University of Chicago Press.

Jorda, Òscàr. 2004 "Model Free Impulse Responses." Working Papers 06-8, University of California at Davis, Department of Economics.

Konis, Kjell. 2007. Linear programming algorithms for detecting separated data in binary logistic regression models. DPhil, University of Oxford.

Lodge, Milton, Kathleen M. McGraw, & Patrick Stroh, 1989) An impression-driven model of candidate evaluation. American Political Science Review 83:399-420.

Liu, Jiannong, & Hodges, James S. 2003. "Posterior bimodality in the balanced one-way random-effects model." *Journal of the Royal Statistical Society B*, 65(1), 247–255.

Lodge, Milton, 1995. "Toward a Procedural Model of Candidate Evaluation." Lodge, Milton & Kathleen McGraw (eds.), *Political Judgment: Structure and Process.* Ann Arbor, 110-140.

Lodge, Milton, Marco Steenbergen, & Shawn Brau, 1995. "The responsive voter: Campaign information and the dynamics of candidate evaluation." *American Political Science Review.* 89(2):309-326.

Lodge, Milton, & Charles Taber, 2000. "Three Steps toward a Theory of Motivated Political Reasoning." in Lupia et al. (eds.), *Elements of Reason.* pp. 183-213. Cambridge University Press, New York.

Lupia, Arthur, & Matthew McCubbins. 1998. *The democratic dilemma: can citizens learn what they need to know?* Cambridge University Press.

Krupnikov, Yanna, 2011. "When Does Negativity Demobilize? Tracing the Conditional Effect of Negative Campaigning on Voter Turnout." *American Journal of Political Science*. 55: 797–813.

McLeish (2010). A general method for debiasing a Monte Carlo estimator.

Miller, Warren E., & National Election Studies. *American National Election Study, 1984.* Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].

Moffitt, R., 1993. "Identification and Estimation of Dynamic Models with a Time Series of Repeated Cross-Sections." *Journal of Econometrics*. 59:99-123.

Monahan, J. & Alan Genz, 1997. Spherical-radial integration rules for Bayesian computation. *Journal of the American Statistical Association.* 92: 664-674.

Owen, A., & Y. Zhou, 2000. Safe and effective importance sampling. *Journal of the American Statistical Association* 95, 135–143.

Prior, Marcus, & Arthur Lupia. 2008. Money, Time, and Political Knowledge: Distinguishing Quick Recall and Political Learning Skills. *American Journal of Political Science*, 52: 169–183.

Romer, Daniel, Kathleen Hall Jamieson, Christopher Adasiewicz, Kate Kenski, Paul Waldman. 2004. *Capturing Campaign Dynamics: The National Annenberg Election*

Study: Design, Method, & Data. Oxford University Press.

Rue, H., S. Martino, & N. Chopin. "Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion)." *Journal of the Royal Statistical Society, Series B*, 71(2):319-392.

Shaw, Daron R. 1999. "A Study of Presidential Campaign Event Effects from 1952 to 1992." *Journal of Politics.* 61(2): 387-422.

Shaw, Daron R. 1999. "The Effect of TV Ads and Candidate Appearances on Statewide Presidential Votes, 1988-1996." *American Political Science Review* 93:345–361.

Shaw, Daron R. 2006. *The Race to 270: The Electoral College and the Campaign Strategies of 2000 and 2004.* Chicago, IL: University of Chicago Press.

Storvik, Geir. 2011. On the Flexibility of Metropolis–Hastings Acceptance Probabilities in Auxiliary Variable Proposal Generation. Scandinavian Journal of Statistics, 38: 342–358.

Thompson Madeleine B. (2011). *Slice Sampling with Multivariate Steps.* Ph.D. thesis, University of Toronto.

Varbeek, 2008. "Pseudo-Panels and Repeated Cross-Sections." Mátyás, L. & P. Sevestre, (eds.), *The Econometrics of Panel Data.* Springer-Verlag. Berlin.

Varbeek & Vella, 2005. "Estimating Dynamic Models from Repeated Cross-Sections." *Journal of Econometrics*. 127:83-102.

Wooldridge, 2008. "Minimum Distance Estimation Using Pseudo-Panel Data." Unpublished manuscript.

Wlezien & Erikson, 2012. *The Timeline of Presidential Elections: How Campaigns Do (and Do Not) Matter.* University of Chicago Press. Chicago.

Zaller & Feldman, 1992. "A simple theory of the survey response: Answering questions versus revealing preferences." *American Journal of Political Science.* 36(3):579-616.

Zaller, 1992. *The Nature and Origins of Mass Opinion.* Cambridge University Press.