Taking Model Uncertainty Seriously: Modeling Autoregressive Distributed Lags via the Bayesian Adaptive Lasso

Taeyong Park*

Abstract

Persistence is a salient characteristic of dynamic political phenomena. Autoregressive distributed lags can be used to estimate persistence of dynamic relationships. However, estimation with lagged variables is complicated by uncertainty in the choice of the number of lags. This uncertainty raises the following issues: 1) restricting the number of lags *a priori* produces underspecification; 2) estimating a parameter-rich general model having many lags avoids underspecification, but overfits the data; 3) it is infeasible to conduct stepwise searching for the "best" lag structure in large-scale multivariate time series models. To resolve these issues, I develop an estimation algorithm to fit time series models via the Bayesian adaptive lasso, a machine learning-based estimator that yields penalized estimation. This algorithm penalizes "unimportant" lagged variables and mitigates overfitting, allowing analysts to employ large-scale time series. This new tool aids in discovering persistence of dynamic relationships or lagged policy effects, which have so far been difficult to theorize or test. These benefits are illustrated by a Monte Carlo simulation analysis and two applications to real-world data.

*Department of Political Science, Washington University in St. Louis, One Brookings Dr., St. Louis MO 63130, `https://graduate.artsci.wustl.edu/tpark`, t.park@wustl.edu.

1. Introduction

Model uncertainty is an issue that analysts regularly face when specifying their statistical models. Determining what variables to include in the model is typically guided by substantive theory. Statistical techniques such as Bayesian model averaging may also help (Montgomery and Nyhan 2010). Yet the problem of model uncertainty is more complicated in time series analysis with lagged variables due to a wide variety of dynamic specifications for the lag structure. Unlike the decision to include certain variables, the choice of how many lagged terms to include in multivariate time series models is rarely guided by substantive theory (Beck 1991; Box-Steffensmeier et al. 2014; Brandt and Freeman 2009; Brandt and Williams 2007; De Boef and Keele 2008; Wilson and Butler 2007). Nevertheless, it is crucial to specify the lag structure appropriately because too few lags produce biased inferences and too many lags overfit the data, leading to spurious estimation.

One existing empirical approach handling uncertainty about the lag structure is the general-to-specific modeling strategy (GSMS, hereafter). The GSMS begins with estimating a *general* autoregressive distributed lag, or ADL, model[1] for stationary time series and exogenous regressors (De Boef and Keele 2008). Since a general ADL model with many lagged terms could overfit the data, the GSMS recommends that analysts impose restrictions on the general lag structure step by step to search for the "best" lag structure. However, it is infeasible to implement this strategy when large numbers of parameters produce a vast model space. For instance, suppose an ADL model examining quarterly presidential approval as a function of lagged dependent variables and six exogenous regressors: if a general ADL model is assumed to have four lags for each of the right-hand side variables, the potential model space will encompass $2^{4+5\times6} \approx 1.7\times10^{10}$ dynamic specifications.[2] As this example illustrates, a general model even with a few lags for a few regressors produces massive numbers of nested specifications, making it infeasible to conduct the stepwise search process.

[1] By a "general" ADL model, De Boef and Keele mean a model that "subsumes the data generating process" (De Boef and Keele 2008, 186). Given that the data generating process is unknown in practice, a general ADL model is typically considered to have many lags. See also Brandt and Williams (2007), Enders (2015), and Hendry (1995).

[2] Alternatively, a prior restriction may assume that low order lags have a larger impact than high order lags and that every model must have all consecutive lags from the lowest all the way up to the highest lag order. Even though this restriction reduces the potential model space substantially, large numbers of nested specifications ($4 \times 5^6 = 62,500$) still make it difficult to execute the GSMS.

Thus, uncertainty about the lag structure in ADL modeling commonly raises the following issues: 1) restricting the lag structure *a priori* produces underspecification; 2) estimating a parameter-rich general model having many lags avoids underspecification, but overfits the data; 3) the GSMS is not a solution when a parameter-rich general model has a multitude of nested specifications.

To resolve these issues, this paper suggests a method that estimates a general ADL model in parallel with mitigating the overfitting problem. Specifically, I develop an estimation algorithm to fit ADL models via the Bayesian adaptive lasso (ADLBL), a penalized regression method.[3] The ADLBL method penalizes or shrinks the coefficient estimates and reduces the variance of the estimated values to make a parameter-rich general ADL model safe from spurious inferences. Consequently, the ADLBL method avoids underspecification and attenuates overfitting, making the GSMS unnecessary. The ADLBL approach is particularly advantageous when substantive theory does not provide precise guidance for dynamic specifications. Yet it is beneficial even when analysts have a strong theoretical justification because it is still useful as a tool for checking the robustness of theory-driven dynamic specifications.

The remainder of this paper follows in five stages. First, I discuss model uncertainty in ADL modeling with an illustrative example. Second, I explain how the Bayesian adaptive lasso alleviates overfitting. Thereafter, I introduce the ADLBL algorithm with an emphasis on its two methodological benefits: 1) the Bayesian framework of ADLBL makes statistical inference straightforward; 2) ADLBL employs adaptive shrinkage to improve estimation accuracy. I also introduce the ECMBL algorithm to extend ADLBL to the error correction model (ECM). Third, I implement a Monte Carlo simulation analysis to verify that ADLBL mitigates overfitting. Fourth, I illustrate the benefits of ADLBL through applications to two empirical examples: the effect of income growth on presidential approval and U.S. policy responses to the Israeli-Palestinian conflict. Finally, I conclude with a discussion of future directions.

[3]Penalized regression methods shrink the coefficient estimates and reduce the variance of the estimated values to improve prediction especially when dealing with large-scale data. The *least absolute shrinkage and selection operator* (lasso) developed by Tibshirani (1996) is one of the most widely used penalized regression methods. Since Tibshirani (1996), a variety of lasso methods have been developed including the Bayesian lasso (Park and Casella 2008) and the Bayesian adaptive lasso (Leng, Tran and Nott 2014).

2. Model Uncertainty in Time Series Analysis

To clarify the problem of model uncertainty in time series analysis, I use De Boef and Keele's (2008, hereafter DBK) replication of Durr, Gilmour, and Wolbrecht's (1997, hereafter DGW) analysis of quarterly congressional approval, 1974-93. Much like DBK's replication, I focus on three dynamic variables including presidential approval, economic expectations, and New York Times coverage of Congress.

Following DBK, I consider the ADL framework. A general model ADL(p, q, k) is displayed as:

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{k} \sum_{i=0}^{q} \beta_{ji} X_{jt-i} + \varepsilon_t, \quad (1)$$

where p refers to the number of lags of Y_t, q denotes the number of lags of X_t, k indicates the number of exogenous regressors, and ε_t is white noise.[4] In ADL modeling, dynamic specifications mostly have to do with determining p and q while the choice of k is usually theory-driven. Given the three dynamic variables used in DGW, let $k = 3$. For simplicity, non-dynamic variables are not reflected in k though they are included in the analysis.

For p and q, DGW and DBK make different choices. DGW's model follows the partial adjustment specification ADL$(1, 0, 3)$. This regression model's right-hand side variables include the dependent variable lagged by one period of time ($p = 1$) and no lagged terms for the exogenous regressors ($q = 0$). DBK argue that the validity of DGW's dynamic specification should be tested. DBK employ the GSMS to search for the appropriate lag structure. They use AIC to compare model specifications,[5] settling on ADL$(1, 1, 3)$ in which $p = q = 1$.[6]

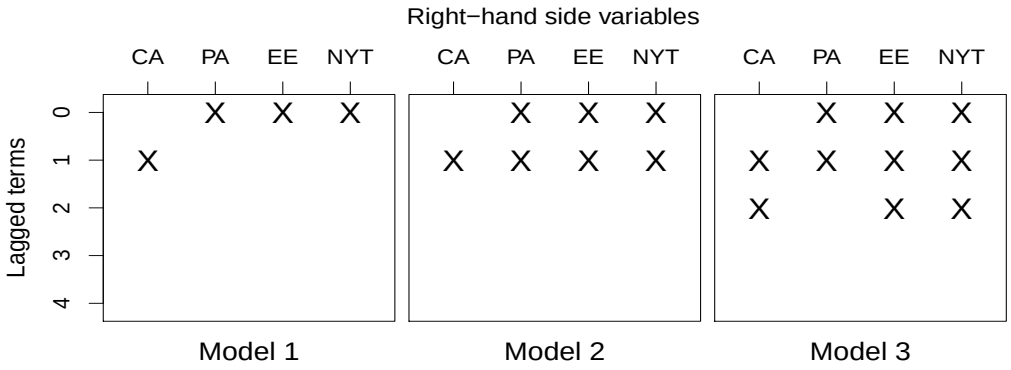
The DGW and DBK models can be graphically illustrated by Model 1 and Model 2 in Figure 1. In the figure, lagged terms in the Y-axis display a 4-quarter lag period: "0" refers

[4]The ADL model is consistently estimated by ordinary least squares (OLS) if white noise error, stationarity process, and weakly exogenous regressors are assumed (Greene 2008). However, the estimation of a lagged dependent variable model with OLS is problematic when residual autocorrelation is present (Achen 2000; Keele and Kelly 2006). Keele and Kelly (2006) show that a common test for residual autocorrelation generally detects the problematic autoregressive model error. In the present article, I conduct the Breucsh-Godfrey LM test to diagnose residual serial correlation, following Keele and Kelly's suggestion.

[5]Akaike's Information Criterion (AIC) is a widely used information criterion for model comparison. A model with a lower value of AIC is considered to have a better model fit.

[6]DBK also deal with an error correction model, but I focus on their ADL model without loss of generality.

Figure 1: Three Examples of Potential Dynamic Specifications Nested in ADL$(4, 4, 3)$



[*Notes*: CA = congressional approval; PA = presidential approval; EE = economic expectations; NYT = New York Times coverage of Congress. Lagged terms in the Y-axis display a 4-quarter lag period: "0" refers to the contemporaneous period, "1" refers to a term lagged by one quarter, and so on. Each cell has "X" if that cell is considered in the dynamic specification. Model 1 is DGW's model and Model 2 is DBK's model. Model 3 represents an alternative dynamic specification.]

to the contemporaneous period, "1" refers to a term lagged by one quarter, and so on. The X-axis represents the right-hand side variables of the regression equation. CA on the X-axis represents a lagged term of the dependent variable, congressional approval. PA stands for presidential approval, EE stands for economic expectations, and NYT stands for New York Times coverage of Congress.[7] "X" indicates that the cell is included in the dynamic specification. For example, the "X" falling under CA in the leftmost panel and the "X" under PA in the same panel mean that Model 1 has variables CA_{t-1} and PA_t. In the same way, the two "X"s under PA in the middle panel correspond to PA_t and PA_{t-1} in Model 2.

DBK's estimation result changes the inference DGW's model draws. Specifically, the short-run effect for PA at $t-1$ is significant in DBK whereas it is assumed to be zero in DGW. This means, as DBK argue, DGW's restriction on PA_{t-1} is invalid.[8] Based on this result, DBK caution against restricting the number of lags in ADL models *a priori*.

DBK's model ADL$(1, 1, 3)$ is selected via AIC. Yet their approach is not without problems. When the GSMS initially considers a long lag period as a general form, the potential model space is often too large to execute stepwise searching.[9] While DBK do not explicitly

[7] As mentioned earlier, other non-dynamic variables used in DGW are included in the analysis but omitted from Figure 1 for the sake of simplicity.

[8] See Table 5 in De Boef and Keele (2008).

[9] Though software packages such as `stepAIC` in R automate the process of comparing models via AIC, the automated stepwise selection method does not cover all possible subsets of a general model. Moreover, it does not necessarily yield the "best" model (James et al. 2013, 209). It is also problematic to use spurious

describe how they dealt with this problem, the GSMS typically restricts the potential model space *a priori*. For example, it is common to employ the same number of lags to different regressors in each step of the GSMS. To clarify the costs of introducing such a restriction, consider a dynamic specification illustrated by Model 3 in Figure 1. This model is grounded in the following theoretical consideration: the three exogenous regressors may have different numbers of lagged terms. This is a reasonable consideration, in that congressional approval could respond to presidential approval in a relatively immediate manner whereas its response to economic expectations and media coverage of Congress could be relatively delayed. This kind of dynamic specification would be omitted from DBK's stepwise search process if the process was restricted to models in which all regressors have the same number of lagged terms.

To assess if it is a cause for concern that DBK's GSMS might omit large numbers of alternative dynamic specifications such as Model 3, I reestimate Model 2 and estimate Model 3.[10] As Table 1 reports, Model 3 has a smaller value of AIC. If DBK considered Model 3 and used AIC for model selection, they could have chosen Model 3. Furthermore, if that were the case, their inference about NYT coverage would differ. The result from Model 3 suggests that NYT coverage of Congress influences congressional approval in a delayed manner: the effect of the NYT coverage variable lagged by two quarters is substantial, which is assumed to be zero in Model 2.[11] This result implies that DBK's restriction on *NY Times* Coverage$_{t-2}$ is possibly incomplete in the same way that DBK's replication suggests DGW's restriction on PA$_{t-1}$ is incomplete.

The takeaway from this replication exercise is that placing restrictions on the potential model space forces us to ignore many specifications that could alter our substantive inferences. The GSMS requires an inordinate amount of tests when large numbers of parameters are under consideration. In such situations, the GSMS has no choice but to place restrictions on the model space. As a consequence, the GSMS may lead to disregarding models

estimation results yielded at a given step to compare models in the subsequent steps (Harrell 2001).

[10]My replication result for Model 2 is virtually identical to Table 5 in DBK (2008, 197). In Table 1, I only report the results for NYT coverage. Other variables from the analysis are included in the estimating equations but omitted from the table. See online appendix A1 for further details of the estimation results.

[11]Graphical representations of impulse response provide further clarification for the difference in statistical inferences between Model 2 and Model 3. See online appendix A2.

Table 1: Replication of De Boef and Keele's Replication of Durr, Gilmour, and Wolbrecht: Effect of Different Dynamic Specifications

	Model 2	Model 3
NY Times Coverage	0.18	0.20
	(0.07)	(0.07)
NY Times Coverage$_{t-1}$	0.04	0.01
	(0.07)	(0.07)
NY Times Coverage$_{t-2}$		0.18
		(0.08)
LM Test p-value	0.67	0.53
AIC	405.32	404.14
Num. obs.	78	78

[*Notes*: Model 2 represents DBK's model and Model 3 represents an alternative ADL model, as specified by Figure 1. Both models are estimated by ordinary least squares. Standard errors are in parentheses. Other variables from the analysis are included in the estimating equations but omitted from the table. See online appendix A1 for further details of the estimation results. The highlighted boxes are for emphasis. The Breucsh-Godfrey LM test using the `lmtest` package in R (Hothorn et al. 2015) provides no evidence of autocorrelated residuals in the three models as the LM test p-values show.]

like Model 3 that appeal to reasonable theoretical considerations.

To solve this problem, I suggest the ADLBL approach: ADL modeling via the Bayesian adaptive lasso. Both the GSMS and ADLBL begin by fitting a general ADL model. However, these two approaches differ on addressing the problem that a parameter-rich general model overfits the data. The GSMS narrows down the general model step by step to search for the best lag structure. On the contrary, ADLBL penalizes the regression coefficients in a principled way to make them safe from spurious inferences. ADLBL allows researchers to make inferences based on the general specification, making the GSMS unnecessary. In the following section, I explain how Bayesian penalized regression attenuates overfitting. Thereafter, I introduce the ADLBL method.

3. Solution: ADLBL

Overfitting and Bayesian penalized regression

As a regression model becomes complex or overparameterized, it becomes likely to overfit the data. The problem with an overfit model is that its extreme specificity reduces generalizability. Since an overfit model capitalizes on the idiosyncratic characteristics of the sample at hand, small changes in that sample data or use of new sample datasets produce dramatically different, and perhaps inaccurate, results. That is, estimation results appearing in an overfit model may be spurious findings that do not exist in the population (Babyak 2004). In this regard, a parameter-rich general ADL model with many lags is susceptible to spurious inferences.

To address the problem of overfitting in general ADL models, I consider a penalized regression approach. Penalized linear regression introduces a penalty term to the ordinary least squares (OLS) loss function. The penalty term shrinks the magnitude of the coefficient estimates and reduces the variance of the estimated values (Tibshirani 1996). While the OLS estimates minimize the prediction error for a single sample, they tend to have large variance from sample to sample. The relatively large variance makes the OLS estimates sensitive to idiosyncrasies of an individual sample. Therefore, permitting some bias, i.e. shrinking the OLS estimates, to reduce the large variance around them produces estimates robust to small changes in the data, and consequently it ensures more generalizable inferences (James et al. 2013; McNeish 2015; Tibshirani 1996).

The lasso (least absolute shrinkage and selection operator) developed by Tibshirani (1996) is one of the most widely used penalized regression methods. Consider the following linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}), \tag{2}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of responses, $\boldsymbol{X}$ is the $n \times p$ matrix of exogenous variables, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of independent and identically distributed normal errors with mean zero and variance σ^2, and $\boldsymbol{I}$ is the identity

matrix. Without loss of generality, $\boldsymbol{y}$ and $\boldsymbol{X}$ are standardized so that the intercept can be omitted. Moreover, by standardizing them, the lasso methods prevent different scales of predictors from affecting penalization.

The lasso estimator penalizes the linear regression coefficients through the L_1-penalized least squares procedure:

$$\hat{\boldsymbol{\beta}}_{Lasso} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \quad (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{k=1}^{p} |\beta_k|, \tag{3}$$

where $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ is the OLS loss function, and $\lambda \sum_{k=1}^{p} |\beta_k|$ is the penalty term. $\lambda \geq 0$ is the shrinkage parameter that determines the amount of shrinking the coefficients. If λ is zero, $\hat{\boldsymbol{\beta}}_{Lasso}$ becomes the OLS estimator. A larger λ causes greater shrinkage of the solutions towards zero.

Since Park and Casella (2008), a number of Bayesian methods for the lasso-type penalization procedure have also been studied (e.g. De Mol, Giannone and Reichlin 2008; Kyung et al. 2010; Leng, Tran and Nott 2014; Ratkovic and Tingley 2017). Prior distributions in the Bayesian framework play the role of the penalty term. For example, Park and Casella (2008) exploit the Laplace prior to yield a Bayesian formulation of the original frequentist lasso displayed in equation (3).

Bayesian approach to inference is advantageous to treatment of the lasso. The Bayesian lasso computes standard errors in a straightforward manner. In a Bayesian analysis researchers can summarize the posterior distribution in any way they like. It is instantaneous to compute standard errors using the posterior distribution. In contrast, the frequentist lasso does not have a general, statistically valid method for obtaining standard errors for the individual regression coefficients (Kyung et al. 2010).[12] For this reason, I employ the Bayesian approach in the present paper. I take advantage of its instantaneous, valid computations of standard errors in order to make my inferential tasks straightforward.

[12]Lockhart et al. (2014) recently developed a method for performing a significance test for the frequentist lasso. This new method does not require resampling or splitting the original data. Hence, it may not have the problem of unstable bootstrapped estimates of standard errors associated with lasso estimator that Kyung et al. (2010) criticize. However, this new method still does not provide a generic calculation of standard errors for the individual regression coefficients. Instead, similar to the likelihood ratio test, it tests whether including the predictor of interest to the model significantly changes the covariance between the observed outcome variable and the predicted outcome variable (McNeish 2015).

Below, I focus on delineating how the Bayesian lasso can alleviate the problem of overfitting, which infrequently appears in the literature.[13] Consider the following posterior distribution of $\boldsymbol{\beta}$ based on Park and Casella (2008):

$$\boldsymbol{\beta} \mid \sigma^2, \{\tau_j^2\}_{j=1}^p, \lambda, \boldsymbol{X}, \boldsymbol{y} \sim N\left((\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_\tau^{-1})^{-1}\boldsymbol{X}\boldsymbol{y}, \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_\tau^{-1})^{-1}\right), \tag{4}$$

where the same terms used in equation (3) are used, except for τ_j^2 and $\boldsymbol{D}_\tau$. These two terms come from the Laplace prior distribution of $\boldsymbol{\beta}$, which is formulated as a scale mixture of a normal distribution (5) with an exponential density (6):

$$\boldsymbol{\beta}|\sigma^2, \tau_1^2, ..., \tau_p^2 \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{D}_\tau), \qquad \boldsymbol{D}_\tau = \text{diag}(\tau_1^2, ..., \tau_g^2) \tag{5}$$

$$\tau_1^2, ..., \tau_p^2 \sim \prod_{j=1}^{p} \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2\tau_j^2}{2}\right) d\tau_j^2, \qquad \tau_1^2, ..., \tau_p^2 > 0. \tag{6}$$

To understand how the Bayesian lasso mitigates overfitting, it is useful to compare the posterior for the Bayesian lasso represented by (4) to the posterior for the standard Bayesian linear regression displayed in (7):

$$\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{X}, \boldsymbol{y} \sim N\left((\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{y}, \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I})^{-1}\right) \tag{7}$$

$$\boldsymbol{\beta}|\sigma^2 \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \tag{8}$$

where $\boldsymbol{I}$ is the identity matrix. A multivariate normal prior on $\boldsymbol{\beta}$ is assumed as (8) indicates.

As large shrinkage λ^2 is implemented, the exponential distribution displayed in (6) tends to produce $\tau_1^2, ..., \tau_g^2$ smaller than 1. Since $\tau_1^2, ..., \tau_g^2$ are the entries along the diagonal in the matrix $\boldsymbol{D}_\tau$, the diagonal entries in $\boldsymbol{D}_\tau^{-1}$ become larger than 1 when $\tau_1^2, ..., \tau_g^2$ are smaller than 1. Consequently, the diagonal entries in $\sigma^2(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_\tau^{-1})^{-1}$ become smaller than $\sigma^2(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I})^{-1}$. That is, the variance term for the posterior distribution of $\boldsymbol{\beta}$ executed by the Bayesian lasso becomes smaller than the variance term executed by the standard Bayesian linear regression when large shrinkage is implemented. In the same manner, the mean term $(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_\tau^{-1})^{-1}\boldsymbol{X}\boldsymbol{y}$ for the Bayesian lasso becomes smaller than the mean term $(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{y}$

[13] I refer the reader to Park and Casella (2008) for details of how to derive the Bayesian lasso.

for the standard Bayesian linear regression when large shrinkage is implemented.

To summarize, the Bayesian lasso produces smaller estimated values and smaller variance around them, compared to the standard Bayesian linear regression. Like other penalization methods, this bias-variance tradeoff enforced by the Bayesian lasso mitigates the problem of overfitting. The ADLBL method introduced below exploits this feature in the context of time series analysis.

Estimating ADL(p, q, k) via the Bayesian adaptive lasso

The ADLBL method is based on a new Bayesian adaptive lasso hierarchical model and an associated Gibbs algorithm designed for this model. Before introducing these, I explain several important features that distinguish the ADLBL method from alternative methods. Above all, the ADLBL method builds on the Bayesian adaptive lasso to enforce varying shrinkage. While the Bayesian lasso developed by Park and Casella (2008) employs the same shrinkage parameter λ for every coefficient, previous research has demonstrated that employing varying shrinkage parameters adaptive to different coefficients improves estimation (Wang, Li and Tsai 2007; Zou 2006). The idea is that if small shrinkage is applied to important predictors and large shrinkage is applied to unimportant predictors, estimation will be more accurate.

Moreover, though a Bayesian adaptive lasso method is provided by Leng, Tran and Nott (2014), the ADLBL method differs from Leng, Tran and Nott's method. The ADLBL method's Gibbs algorithm allows the prior distribution for the shrinkage parameter to vary across lag orders, depending on the analyst's prior belief. Consequently, if the analyst has a prior belief that the parameter spaces for lagged effects are sparse, which is common to time series analysts, she can impose large shrinkage on high order lags. This feature of the ADLBL method is in line with the view that the prior distribution of the parameter space in time series analysis is not flat. For example, the Bayesian structural vector autoregression (B-SVAR) model employing the Sims-Zha prior puts larger shrinkage on the coefficients at higher lags based on the belief that more proximate events are highly predictive of the events today (Brandt and Freeman 2009; Brandt, Colaresi and Freeman 2008; Sattler, Freeman and Brandt 2007).

Both the ADLBL method and the B-SVAR model execute shrinkage for lagged variables

instead of zeroing out lags or deleting variables altogether. In so doing, these methods avoid underspecification caused by deleting lagged variables *a priori*. Furthermore, any restrictions imposed by the prior distribution are adjusted by the data: if a prior belief about the lag structure is inconsistent with the data, the posterior distribution will not reflect this view.

Despite these similarities, the ADLBL method differs from the B-SVAR model in a few important respects. First, while B-SVAR models generally conduct pretests using AIC or a χ^2 test for lag length selection, the ADLBL method does not require pretests for lag selection. The ADLBL method implements specification of the lag structure and estimation simultaneously and in a principled way. Indeed, lasso methods have been used to amend the deficiencies of classical model selection methods. For instance, Wang, Li and Tsai (2007) point out that the statistical performance of AIC and BIC can be unstable, and inferences may differ depending on different selection criteria. They argue their lasso method be a better approach for estimation and autoregressive order selection in time series analysis. Second, the ADLBL Gibbs algorithm imposes shrinkage that not only varies between different lags within a regressor but also varies between regressors. In particular, this algorithm distinguishes lagged dependent variables from other regressors to impose different amounts of shrinkage. This is a desired property since lagged variables and other regressors are grounded in different data generating processes. Finally, the ADLBL method provides an estimation algorithm directly applicable to any single-equation ADL models.[14] By contrast, the B-SVAR model and the Sims-Zha prior are designed to estimate multiple equations.

Now, I derive the Bayesian hierarchical framework for ADLBL and the Gibbs sampler for the method. I continue considering the general ADL(p, q, k) presented in equation (1). But here the general model is expressed as a matrix form as follows:

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}), \tag{9}$$

where $\boldsymbol{y}$ is the $T \times 1$ vector of the standardized dependent variable, $\boldsymbol{Z}$ is the $T \times [p + k(q+1)]$ matrix of standardized regressors, $\boldsymbol{\theta}$ is the $[p + k(q+1)] \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is the $T \times 1$ vector of white noise error.

[14] As discussed later, the ADLBL method is also extended to the error correction model (ECM).

$\boldsymbol{Z}$ denotes all dynamic variables in the right-hand side of the equation including lagged dependent variables and exogenous regressors. $\boldsymbol{Z}$ can include non-dynamic variables in practice. Yet here, I focus on dynamic variables for the sake of simpler discussion.

Building on Park and Casella (2008) and Leng, Tran and Nott (2014), the hierarchical framework of ADLBL is derived as follows:

$$\boldsymbol{y}|\boldsymbol{Z},\boldsymbol{\theta},\sigma^2 \sim N(\boldsymbol{Z\theta},\sigma^2\boldsymbol{I}) \tag{10}$$

$$\boldsymbol{\theta}|\sigma^2,\tau^2_{1(m)},...,\tau^2_{g(m)} \sim N(\boldsymbol{0},\sigma^2\boldsymbol{D}_\tau), \qquad \boldsymbol{D}_\tau = \text{diag}(\tau^2_{1(m)},...,\tau^2_{g(m)}) \tag{11}$$

$$\tau^2_{1(m)},...,\tau^2_{g(m)} \sim \prod_{g=1}^{G} \frac{\lambda^2_{g(m)}}{2} \exp\left(-\frac{\lambda^2_{g(m)}\tau^2_{g(m)}}{2}\right) d\tau^2_{g(m)}, \qquad \tau^2_{1(m)},...,\tau^2_{g(m)} > 0 \tag{12}$$

$$\sigma^2 \sim \pi(\sigma^2)d\sigma^2 \tag{13}$$

$$\lambda^2_{g(m)} \sim \frac{\delta^r_m}{\Gamma(r)}(\lambda^2_{g(m)})^{r-1}\exp(-\delta_l\lambda^2_{g(m)}), \quad \lambda^2_{g(m)}, r, \delta_m > 0, \quad m = 1,...,(p \text{ or } q), \tag{14}$$

where g is in $\{1,...,G\}$, G denotes $p + k(q + 1)$ that is the total number of parameters in ADL(p, q, k), the prior distribution of $\tau^2_{g(m)}$ is exponential, and a non-informative, scale-invariant prior on σ^2 is used. In addition, the prior distribution of $\lambda^2_{g(m)}$ is gamma, where m denotes a lag order.[15] r and δ_m are hyper-prior parameters for $\lambda^2_{g(m)}$. The Laplace prior for $\boldsymbol{\theta}$ is represented by a scale mixture of a normal distribution (11) with an exponential density (12), following Park and Casella (2008).[16]

As subscript m in $\lambda^2_{g(m)}$ reveals, the gamma prior distribution of $\lambda^2_{g(m)}$ varies by lag order depending on hyper-parameter δ_m. As a result, this prior allows analysts to impose larger shrinkage to higher order lags when their prior beliefs tell them to do so. This is the difference between the hierarchical model for ADLBL and the Bayesian adaptive lasso model derived by Leng, Tran and Nott (2014).

Bayesian joint posterior distribution is produced by conditioning the prior distributions from (11) through (14) on the likelihood function given in (10). Full descriptions of the

[15] I treat $\lambda^2_{g(m)}$ as random variables and assume that the prior distribution is gamma. This approach allows easy extension of the Gibbs sampler by conjugacy properties. Alternatively, the shrinkage parameters can also be estimated by maximum marginal likelihood using an Empirical Bayes Gibbs algorithm (Park and Casella 2008).

[16] See also Andrews and Mallows (1974).

derivation of the posterior distributions are in the online appendix, and here I only present the Gibbs algorithm. The following Gibbs algorithm for the ADLBL model produces marginal posterior distributions for the parameters of interest from their full conditional distributions.

1) Choose starting values $\boldsymbol{\theta}^{(0)}$, $\sigma^{2^{(0)}}$, $\{\tau^2_{g(m)}\}^{(0)}$, and $\{\lambda^2_{g(m)}\}^{(0)}$.

2) At the sth iteration, draw

$$
\begin{aligned}
\boldsymbol{\theta}^{(s)} &\sim N([\boldsymbol{Z'Z} + \boldsymbol{D}_{\boldsymbol{\tau}}^{-1^{(s-1)}}]^{-1}\boldsymbol{Z'y}, \sigma^{2^{(s-1)}}[\boldsymbol{Z'Z} + \boldsymbol{D}_{\tau}^{-1^{(s-1)}}]^{-1}), \\
(1/\tau_{g(m)})^{2^{(s)}} &\sim Inv.Gaussian\left(\sqrt{\frac{\lambda^{2^{(s-1)}}_{g(m)}\sigma^{2^{(s-1)}}}{\theta^{2^{(s)}}_{g(m)}}},\ \lambda^{2^{(s-1)}}_{g(m)}\right), \\
\lambda^{2^{(s)}}_{g(m)} &\sim Gamma\left(r+1, \frac{\tau^{2^{(s)}}_{g(m)}}{2} + \delta\right), \\
\sigma^{2^{(s)}} &\sim Inv.Gamma\left(\frac{T-1+G}{2}, \frac{1}{2}[(\boldsymbol{y} - \boldsymbol{Z\theta}^{(s)})'(\boldsymbol{y} - \boldsymbol{Z\theta}^{(s)}) + \boldsymbol{\theta}^{(s)'}\boldsymbol{D}^{(s)}_{\boldsymbol{\tau}}\boldsymbol{\theta}'^{(s)}]\right), \\
&\text{where} \quad g = 1, ..., G, \quad m = 1, ..., (p \text{ or } q).
\end{aligned}
\tag{15}
$$

3) Go to 2) until $s = B + S$, where B is the burn-in sample and S is the desired sample size for convergence.

Estimating ECM(p, q, k) via the Bayesian adaptive lasso: ECMBL

The error correction model or ECM is another class of time series models that has become popular in contemporary research. De Boef and Keele (2008) show that the ECM is equivalent to the ADL, and argue that the ECM is appropriate for use with stationary data as well as cointegrated time series.[17] Consider the following general ECM(p, q, k):[18]

$$
\Delta Y_t = \alpha_0 + \sum_{i=1}^{p}\alpha_i Y_{t-i} + \sum_{j=1}^{k}\beta_j \Delta X_{jt} + \sum_{j=1}^{k}\sum_{i=1}^{q}\gamma_{ji}X_{jt-i} + \varepsilon_t, \tag{16}
$$

[17] There is a counter-argument that the ECM should generally not be used with political data (Grant and Lebo 2016). See the debate between Grant and Lebo (2016) and Keele, Linn and Webb (2016) appeared in the time series symposium at *Political Analysis* 24(1). See also Esarey (2016), Freeman (2016), and Helgason (2016), which offer commentaries on that debate.

[18] Derivations of the general form of the ECM can be seen in Banerjee et al. (1993, 50-54).

where p refers to the number of lags of Y_t, q denotes the number of lags of X_t, k indicates the number of exogenous regressors, and ε_t is white noise. Based on the direct tie between the ECM and the ADL, it is straightforward to derive ECMBL – Estimating ECM(p, q, k) via the Bayesian adaptive lasso. As the general ADL(p, q, k) was converted to a matrix form in equation (9), the general ECM(p, q, k) can be written as a matrix form such as $\boldsymbol{y} = \boldsymbol{Z\theta} + \boldsymbol{\varepsilon}$. Given this matrix form, the hierarchical model and the Gibbs sampler for ECMBL is equivalent to those for ADLBL except that $\boldsymbol{y}$ and $\boldsymbol{Z}$ now refer to the dependent variable and the right-hand side variables displayed in (16).

Uncertainty in the dynamic specifications for the ECM yields the same practical difficulties as for the ADL. ECMBL provides a solution to model uncertainty for the ECM in the same way that ADLBL does for the ADL. Though I focus on ADLBL in the following sections for simulation analysis and empirical applications, both ADLBL and ECMBL are implemented in the second data analysis in the empirical examples section.

4. Monte Carlo Simulation Analysis

This section conducts a Monte Carlo simulation analysis to assess whether ADLBL mitigates the problem of overfitting. Mitigating it is a key property that ADLBL must hold to guarantee that it is reliable to make statistical inferences based on a parameter-rich general ADL model. Recall that the problem with an overfit model is that its extreme specificity reduces generalizability. Hence, my simulation analysis is designed to evaluate the out-of-sample prediction performance. It compares ADLBL with the ADL model estimated via OLS (henceforth ADLOLS) in terms of prediction accuracy measured by the mean squared error.

The simulation exercise is carried out on different sample sizes. I expect that ADLBL performs better than ADLOLS especially under the circumstance where the sample size is small and overparameterization is more likely to occur. Consider the following data generating

process represented by ADL(1,1,4):

$$
\begin{aligned}
&Y_t = \alpha_i Y_{t-1} + \sum_{i=0}^{1} \beta_{1i} X_{1t-i} + \sum_{i=0}^{1} \beta_{2i} X_{2t-i} + \sum_{i=0}^{1} \beta_{3i} X_{3t-i} + \sum_{i=0}^{1} \beta_{4i} X_{4t-i} + u_t \\
&X_t = \rho X_{t-1} + e_t, \\
&u_t,\ e_t \overset{iid}{\sim} N(0, \sigma^2),
\end{aligned}
\tag{17}
$$

where the exogenous regressors $X_1, ..., X_4$ follow the AR(1) process, $\alpha = 0.5$, $\rho = 0.75$, and the coefficients for the four contemporaneous exogenous variables are set to be 1.5 and the coefficients for the four lagged terms are set to be 0.5.

In ADL modeling analysts are recommended to start with a general dynamic specification that subsumes the data generating process (De Boef and Keele 2008). Similarly, my simulation exercise considers an extensive dynamic specification assumed to be general, which is represented by $p = q = 8$. Accordingly, I fit ADLOLS(8,8,4) and ADLBL(8,8,4) to the data generated by the above process (17).

By design, the dynamic specification (8,8,4) is inconsistent with the true data generating process. This setting provides a rigorous test for the performance of ADLBL(8,8,4). To be specific, it allows me to assess whether ADLBL(8,8,4) performs well even though its lag structure does not fit into the data generating process, which is a situation that analysts could encounter frequently in practice. With this in mind, I fit ADLOLS(1,1,4) as a benchmark. Because the dynamic specification for ADLOLS(1,1,4) by design reflects the data generating process shown in (17), the performance of ADLOLS(1,1,4) should be the best. My focus will be on comparing each of ADLOLS(8,8,4) and ADLBL(8,8,4) with the benchmark ADLOLS(1,1,4) in terms of out-of-sample prediction.

The number of parameters is 44 for both ADLOLS(8,8,4) and ADLBL(8,8,4).[19] Considering this number of parameters, I experiment with datasets of size 100, 200, and 400. The statistics literature recommends that analysts fit one parameter for each 10 observations when the data are independent and identically distributed to avoid overfitting (Babyak 2004; Keele, Linn and Webb 2016). This rule of thumb suggests that the data of size 100 is too small for 44 parameters whereas the data of size 400 is large enough. Thus, the predic-

[19] This is calculated as $8 + 9 \times 4 = 44$.

tion of ADLBL(8,8,4) is expected to outperform that of ADLOLS(8,8,4) especially in the experiment with the data of size 100 in which overfitting is highly likely to be problematic.

To run ADLBL(8,8,4) I standardize all variables and plug them into the Gibbs sampler presented by algorithm (15).[20] I also standardize all variables for ADLOLS(8,8,4) and for ADLOLS(1,1,4) for a direct comparison. For each of the three different sizes of data, I fit the three models 50 times to obtain the average of the mean squared errors.

Table 2 reports the simulation analysis results. The average mean squared errors from ADLOLS(1,1,4), the benchmark model, are set to be 1. And then, the average mean squared errors from ADLOLS(8,8,4) and ADLBL(8,8,4) are presented in comparison to the benchmark. For example, the top-left cell shows that the average of the mean squared errors corresponding to ADLOLS(8,8,4) with the data size of 100 is about 20 times larger than the benchmark model's.

Table 2: Comparing Out-of-sample Prediction Performance with Average Mean Squared Errors: 50 Replications

Model	N=100	N=200	N=400	Estimation
ADLOLS(8,8,4)	20.37	1.55	1.20	OLS
ADLBL(8,8,4)	1.10	1.11	1.06	Bayesian adaptive lasso
Benchmark ADLOLS(1,1,4)	1.00	1.00	1.00	OLS

[*Notes*: The relative size of the average mean squared errors from 50 replications is presented. ADLOLS(1,1,4) is the benchmark model that is, by design, consistent with the data generating process presented in (17). The average mean squared errors from ADLOLS(1,1,4) are set to be 1 as a benchmark. Cell entries for ADLOLS(8,8,4) and ADLBL(8,8,4) represent the average mean squared errors in comparison to those from the benchmark model. For example, "20.37" indicates a mean squared error 20 times larger than the benchmark model's. "N" denotes the sample size. Across different experiments using different sample sizes, ADLBL(8,8,4) produces smaller mean squared errors, compared to ADLOLS(8,8,4). In particular, the former performs a lot better than the latter in the small sample size setting.]

These results demonstrate that ADLBL(8,8,4) is less susceptible to overfitting than ADLOLS$(8, 8, 4)$ even though they have the same lag structure. The out-of-sample prediction

[20]The hyper-parameters for the shrinkage parameter are assumed to be 1 and 0.1 for the shape and scale parameters, respectively. This choice of hyper-parameters follows Kyung et al. (2010) and Leng, Tran and Nott (2014). As Leng, Tran and Nott (2014) point out, the choice of hyper-parameters does not have a considerable effect on inferences because they are deeper in the hierarchy. The estimates come from 10,000 iterations of the Gibbs sampler after 1,000 burn-in iterations.

performance of ADLOLS$(8, 8, 4)$ is poor when the size of sample is small, i.e. when overfitting is likely to occur. As the size of sample becomes large, the performance of ADLOLS$(8, 8, 4)$ improves, revealing that it performs fairly well when the problem of overfitting is not severe. On the contrary, ADLBL$(8, 8, 4)$ performs consistently well. Its prediction performance is comparable to the performance of the benchmark model in every experiment. This simulation study shows that the ADLBL approach attenuates overfitting, and hence provides reliable estimation results even with a relatively small sample size. This holds true even when its general dynamic specification does not fit into the data generating process represented by a simple dynamic specification.

5. Empirical Examples

In this section, I apply the ADLBL method to real-world data to illustrate its usefulness. I consider two empirical examples that provide some scope for broad applications of ADLBL: they are engaged in two different research areas, one on American politics and the other on international relations. I also apply the ECMBL method as well as ADLBL to the second example in order to connect my analysis to the ongoing debate on whether the ECM is appropriate for use with political time series data.

Example I: Income growth and presidential approval

My first example is an application to the effect of macro-level income growth on presidential approval in the U.S. context. Income is a widely talked-about issue to gauge the state of the national economy. Ordinary citizens also seriously care about changes in their income. For these reasons, a number of previous studies on the economy and presidential support paid attention to a macro-level income indicator, disposable personal income per capita (De Boef and Nagler 2005; Erikson 1989; Hibbs 1982; Markus 1988).

However, disparate findings exist on this topic. On the one hand, sizable literatures analyzing the presidential vote find evidence supporting substantial effects of income growth. For instance, Erikson (1989) performs a regression analysis for the 1948-84 presidential elections to show that the incumbent party vote share rises as income growth steepens. Markus

(1988) employs a pooling of survey data, 1956-84, augmented by a time series of economic statistics, and finds that income growth increases the probability of a pro-incumbent vote. On the other hand, time series analyses of presidential approval have failed to find such evidence. MacKuen, Erikson, and Stimson's (1992) analysis using quarterly time series data from 1954 to 1988 finds no substantial effect of per capita income growth on presidential approval.[21] De Boef and Nagler (2005) use a 1965-96 quarterly time series dataset to examine the effect of income growth on presidential approval by individual income level, and find a practically null effect.

I revisit the time series analysis of presidential approval to investigate if a general dynamic specification with a long lag length changes the substantive inferences of previous studies. Employing a long lag length appeals to a theoretical justification. It takes some time for people to get a sense of how government policy has influenced their disposable income. Sometimes people do not know how much they are receiving in terms of tax breaks until they pay their taxes. Or a reward for growth at a company can come in an annual bonus that the company workers do not receive until the end of the year. It may also be the case that it takes time for quarterly increases in disposable income to add up to make a real difference. Thus, there may exist sizable effects at high order lags even after controlling for the effects at the contemporaneous period and low lags.

Data

I use the Gallup quarterly presidential approval time series from 1964-2015 for the dependent variable.[22] The key explanatory variable, macro-level income growth is measured by the percentage change in real disposable personal income per capita. This variable comes from the Federal Reserve Bank of St. Louis economic data archive.[23] I use seasonally adjusted data for income growth.

To address the concern that other economic variables drive presidential approval, I control

[21] They found that the income variable performed poorly in all specifications and dropped the variable from the presentation. See Endnote 7 in MacKuen, Erikson, and Stimson (1992, 609).

[22] When there are multiple polls for one quarter, I took the average of all approval rates in that quarter. In addition to this measurement scheme, I employed an alternative scheme that selects one particular approval rating per quarter. This alternative scheme does not change substantive results. See online appendix A4 for further details of the alternative scheme.

[23] The website is `https://research.stlouisfed.org/fred2/`.

for economic growth, unemployment change, inflation, and the interest rate. Economic growth is measured by the quarterly percent change in real gross domestic product (GDP), unemployment change is measured by a first difference in the quarterly unemployment rate, inflation is measured by the quarterly percent change in the Consumer Price Index (CPI), and the interest rate is measured by a first difference in the effective federal funds rate.[24] Additionally, a first difference in the quarterly index of consumer sentiment[25] is included to assess whether the effect of income growth is substantial even after the public's aggregate-level evaluations of the economy is accounted for (MacKuen, Erikson and Stimson 1992).

Following previous research on presidential approval (De Boef and Nagler 2005; Hibbs 1987; MacKuen, Erikson and Stimson 1992), my analysis considers presidential administration dummies, inauguration periods, and political events. The political events under consideration follow MacKuen, Erikson, and Stimson's (1992) coding. I additionally include international "rally" events for the Gulf war and the 9/11 attacks.[26] I employ a dummy variable for the first eight quarters of each administration which are estimated by economic conditions entirely or partially credited to the previous administration.

The ADL model requires stationary time series. I use the augmented Dickey-Fuller (ADF) test to check if each variable has a unit root. All variables used in the analysis rejected the null hypothesis of a unit root at the 95% confidence level.[27]

I aim to examine how income growth influences presidential approval, not instantaneously, but over an extended period. The highest lag order is assumed to be eight. That is, I assume that ADL(8,8,6) is a general model, which accounts for an 8-quarter lag period for the dependent variable and for the six dynamic exogenous regressors. My sample encompasses 205 quarters, from the second quarter of 1964 to the second quarter of 2015.

[24]These macroeconomic indicators are from the Federal Reserve Bank of St. Louis economic data archive. All indicators based on U.S. dollars, such as real GDP, are expressed as billions of chained 2009 dollars. I use seasonally adjusted data for these indicators except for the interest rate.

[25]The University of Michigan's Index of Consumer Sentiment (ICS), `https://data.sca.isr.umich.edu`.

[26]See online appendix A5 for detailed coding scheme.

[27]I used the `adf.test` function available in the `tseries` package (Trapletti and Hornik 2011). In addition, I also used the `ur.df` function available in the `urca` package (Pfaff, Zivot and Stigler 2010), and it was possible to reject the null of a unit root for all variables in the drift and trend test types.

Prior Distribution

ADLBL's prior gamma distribution on the shrinkage parameter is designed to impose disproportionate amounts of shrinkage to different lag orders. In general, more proximate events are believed to be better predictors for the values today. My prior belief is in line with this view, and I employ a prior distribution that imposes larger shrinkage on higher lags.[28]

As the left panel in Figure 2 illustrates, the prior distribution on the shrinkage parameter produces larger values for higher lags. In the ADLBL hierarchical framework, the prior specification for the shrinkage parameter determines the covariance matrix of the multivariate normal prior distribution on the coefficients: a large shrinkage parameter leads to small variance around its associated coefficient. As a result, as the right panel in Figure 2 shows, the coefficients of higher lags have relatively small variance. In so doing, the coefficients of higher lags are forced to have smaller values close to zero.[29] This feature is similar to how the Sims-Zha prior deals with lag coefficients in B-SVAR models. The Sims-Zha prior results in lag coefficients that shrink to zero over time and have smaller variance at higher lags (Brandt and Freeman 2009, 119).

In addition to the main analysis specifying prior distributions as Figure 2, I consider two alternative specifications for a robustness check. First is a prior imposing larger shrinkage on higher lags, but the rate parameters now have different values.[30] Second is a prior imposing non-varying shrinkage across the nine lag orders.[31] While the first specification yields similar prior distributions to those presented in Figure 2, the second specification yields flat prior distributions. The results based on the main analysis are robust to these alternative prior specifications. I only report the main analysis results in the following subsection.

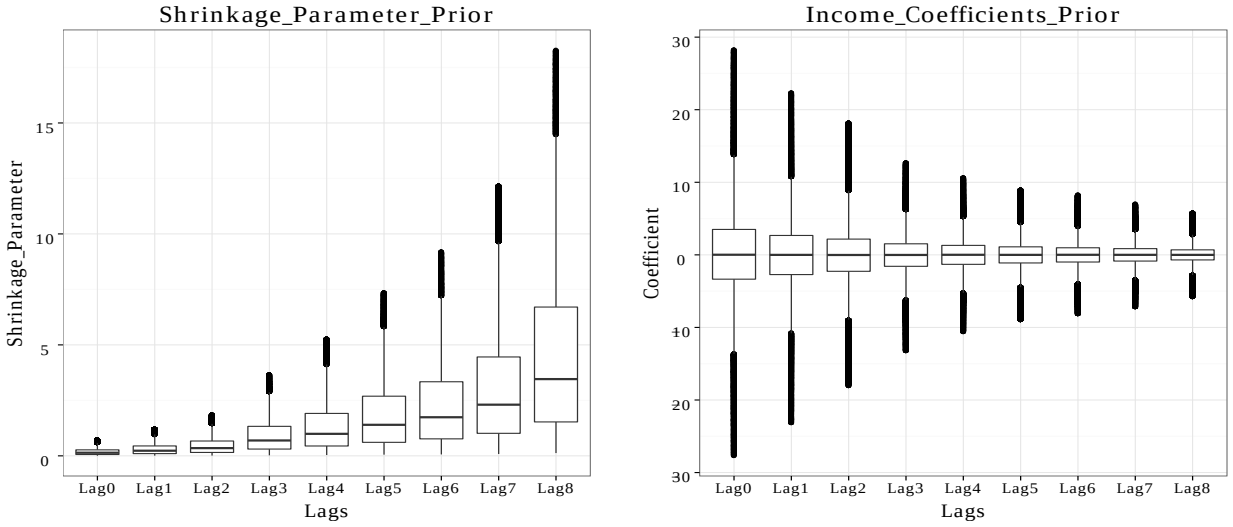
[28]The rate parameters for the gamma distribution are given by $\delta_0 = 5, \delta_1 = 3, \delta_2 = 2, \delta_3 = 1, \delta_4 = .7, \delta_5 = .5, \delta_6 = .4, \delta_7 = .3,$ and $\delta_8 = .2$, where m in δ_m denotes a lag order. The shape parameter is fixed at 1 regardless of lag orders. Given a fixed shape parameter, a larger value for the rate parameter generates a gamma distribution yielding smaller values.

[29]To be clear, the posterior distribution will also reflect the data.

[30]Specifically, the rate parameters for the gamma distribution are given by $\delta_0 = 1.8, \delta_1 = 1.6, \delta_2 = 1.4, \delta_3 = 1.2, \delta_4 = 1, \delta_5 = .8, \delta_6 = .9, \delta_7 = .4,$ and $\delta_8 = .2$, where m in δ_m denotes a lag order.

[31]The shape parameter is fixed at 1 and the rate parameter is fixed at 0.1.

Figure 2: The Prior Distribution of the Shrinkage Parameter and the Prior Distribution of the Income Coefficient: Varying by Lag Order



Results

To interpret the results, I focus on comparing the estimated long-run effects obtained from three different dynamic specifications: a partial adjustment model, or ADL(1,0,6), an ADL model with a 1-quarter lag, or ADL(1,1,6), and a general specification ADL(8,8,6). The long-run effect, or long-run multiplier, is one of the key effects to be estimated in time series analysis, which represents the total causal effect of X_t on Y_t (De Boef and Keele 2008). Given the general model ADL(p, q, k) represented by equation (18), the long-run effect (LRE) of X_j on Y is given by equation (19):

$$Y_t \;=\; \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{k} \sum_{i=0}^{q} \beta_{ji} X_{jt-i} + \varepsilon_t, \tag{18}$$

$$\text{LRE}_{\text{ADL}} = \frac{\sum_{i=0}^{q} \beta_i}{1 - \sum_{i=1}^{p} \alpha_i}. \tag{19}$$

In general, a direct estimate of the standard error for the long-run effect is not provided by ADL models. However, a Bayesian approach provides the posterior distribution that can be used to compute the standard error for the long-run effect. For this reason, I estimate ADL(1,0,6) and ADL(1,1,6) via Bayesian Gaussian linear regression.[32] I estimate

[32]I use the `MCMCregress` function in the `MCMCpack` package (Martin, Quinn and Park 2011) to generate a sample from the posterior distributions for ADL(1,0,6) and ADL(1,1,6). I use diffuse prior distributions

ADL(8,8,6) via the Bayesian adaptive lasso to address overfitting. I call these three models ADLOLS(1,0,6), ADLOLS(1,1,6), and ADLBL(8,8,6) to clarify each model's estimation method. I standardize all variables for a direct comparison.

Figure 3 presents the estimation results for the long-run effects of income growth on presidential approval.[33] The figure also displays the estimated long-run effects of inflation for a comparison. The long-run effect of income growth estimated by ADLBL(8,8,6) is substantively large and statistically reliable.[34] The point estimate from ADLBL(8,8,6), which comes from the posterior mean, indicates that one additional standard deviation of income growth increases approximately one standard deviation of presidential approval. Based on the data at hand, this point estimate is interpreted as a 1% increase in income growth effecting an increase in presidential approval rating by about 12% points.[35] In contrast, the long-run effects obtained from ADLOLS(1,0,6) and ADLOLS(1,1,6) are not statistically distinguishable from zero.

The large long-run effect estimated by ADLBL(8,8,6) results from substantial positive lagged effects at high order lags. The estimated dynamic marginal effects at lags 1, 4, 7, and 8 are positive and statistically reliable at the 90% credible level.[36] This result reveals that a shock in income growth has a long lasting effect on presidential approval and dissipates slowly.

Finally, it is informative to compare the results for income growth and those for inflation. In the case of income growth, the difference in the posterior means between ADLOLS(1,0,6) and ADLBL(8,8,6) is statistically reliable: the 90% highest posterior density for the two models do not overlap each other in Figure 3. Though the difference between ADLBL(8,8,6) and ADLOLS(1,1,6) is less noticeable, the 70% highest posterior densities for the two models

that `MCMCregress` provides as default. This approach is equivalent to OLS estimation.

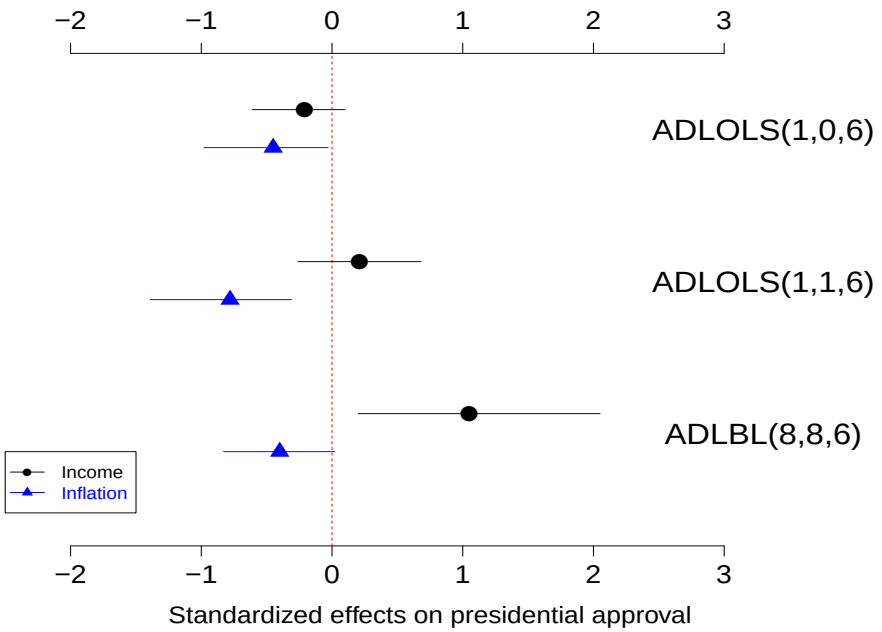
[33]The Breucsh-Godfrey LM test provides no evidence of autocorrelated residuals in the three models I estimate. The ADLBL estimates come from 50,000 iterations of the Gibbs sampler after 10,000 burn-in iterations. The convergence of the Markov Chain Monte Carlo (MCMC) outputs generated by the Gibbs sampler should be diagnosed (Gill 2014). A test for convergence of the MCMC output was performed by the `superdiag` package (Tsai and Gill 2012), which finds no evidence of non-convergence using all of the standard procedures. For example, Geweke's diagnostic returns a small value of $|Z - \text{score}|$ that is less than one, indicating that the values early and late in the MCMC outputs are not statistically distinguishable.

[34]This finding holds based on the 95% credible interval though the figure presents the 90% credible interval.

[35]In the data, one standard deviation of presidential approval rating is approximately 11% points and that of income growth is approximately 0.9% points.

[36]Graphical representations of the dynamic marginal effects are in the online appendix. See A6.

Figure 3: Comparison of Estimated Long-Run Effects on Presidential Approval



[*Notes*: This figure depicts the estimated long-run effects of income growth and their statistical uncertainty obtained from three different models: ADLOLS(1,0,6), ADLOLS(1,1,6), and ADLBL(8,8,6). Each circle point is a point estimate of the effect of one additional standard deviation of income growth, which comes from the posterior mean. For example, the point estimate from ADLBL(8,8,6) reveals that one additional standard deviation of income growth increases approximately one standard deviation of presidential approval. For the purpose of comparison, this figure also displays the estimated long-run effects of inflation and their statistical uncertainty. Each triangle point represents a point estimate of the effect of one additional standard deviation of inflation. Each horizontal line represents a 90% Bayesian credible interval.]

do not overlap each other. On the contrary, we find different results in the case of inflation. the estimation results of ADLBL(8,8,6) are not statistically distinguishable from the results of the two other models. This difference between income growth and inflation indicates that while considering high order lags is important to draw an accurate inference for the effect of income growth, considering them does not change substantive inferences for the effect of inflation. In other words, while a shock in income growth has a long lasting effect on presidential approval, a shock in inflation affects presidential approval in a relatively immediate way and dissipates quickly. This kind of comparison can broaden our understanding of a variety of dynamic relationships between multiple economic indicators and presidential approval.

Example II: U.S. policy responses to the Israeli-Palestinian conflict

The next example uses events data from Brandt, Colaresi, and Freeman (2008, hereafter BCF). I examine how the U.S. government's policy toward Palestine responds to the Israeli government's foreign policy behavior toward the Palestinians, controlling for Palestinian policy toward the Israelis and Jewish people's support for peace. As in the first example, I focus on evaluating lagged relationships. That is, I address the possibility that it takes months for the U.S. government's foreign policy to respond to political dynamics in the Levant.

In order to determine the number of lagged variables, I could use a χ^2 test or information criteria. Indeed, BCF use a χ^2 test and AIC to specify their B-SVAR model. However, this stepwise search process is not without its own shortcomings. Above all, this approach typically restricts the potential model space to apply the same number of lags to different regressors in each step. Otherwise, the model space would be too large to use the stepwise method. As a consequence, model comparisons omit a vast number of specifications. For instance, BCF's pretests compare 12 specifications for 12 different lag lengths, omitting large numbers of specifications like a model having two lags for Israeli behavior, a one lag for Palestinian behavior, and three lags for Jewish people's support for peace.

Alternatively, my analysis executes ADLBL to draw inferences from a general model directly. I consider six lagged terms for each dynamic variable to account for a 6-month lag period. In other words, a general model is specified as ADL(6,6,3). ADLBL mitigates the problem of overfitting caused by the general model's large numbers of parameters.

Data

The events data used in my analysis measure the foreign policy behaviors of the U.S., the Israelis, and the Palestinians. As in BCF's main analysis, I use monthly averages of Goldstein-scaled events, which were extracted from the Kansas Event Data System (KEDS) from Agence France Presse (AFP) news stories. My sample is monthly from November 1996 to December 2015. BCF use monthly averages rather than the total data because monthly averages place the event data on a scale similar to the Jewish public opinion data. The same

reason applies to my analysis since I also employ the public opinion data. Moreover, as BCF point out, it is reasonable to deem that policy makers are concerned with deviations from the average level of an ongoing conflict.

My dependent variable is the directed behavior of the U.S. toward the Palestinians. This variable is denoted by A2P following BCF's variable mnemonics. The primary explanatory variable is Israeli foreign policy behavior toward the Palestinians, denoted by I2P. Control variable P2I refers to Palestinian actions toward the Israelis. A negative value of these policy behavior variables represents a hostile action, and a positive value represents a cooperative action. Another control variable JPI refers to Jewish public support for peace, which comes from polls conducted by the Tami Steinmetz Center for Peace Research (TSC) at Tel Aviv University. In addition, I consider nine trend and dummy variables used in BCF.[37] I assume I2P, P2I, and JPI to be exogenous to A2P. BCF find that U.S. actions drive neither I2P nor P2I whereas they are affected by I2P and P2I. According to this finding, the U.S. government appears to be responsive to, rather than driving, the political dynamics in the Levant.

I used the augmented Dickey-Fuller (ADF) test to check for unit-root of the dynamic variables. I obtained disparate test results. I used the `ur.df` function available in the `urca` R package and it was possible to reject the null hypothesis of a unit root for all variables in the drift and trend test types at the 95% confidence level. However, it was not possible to reject the null of a unit root in the "none" test type that includes neither an intercept nor a trend. Alternatively, I also used the `adf.test` function available in the `tseries` R package, and none of the variables in my analysis rejected the null of a unit root. Given that the data might be individually integrated, I performed tests for cointegration. I conducted both the Engle-Granger two-step method and ECM test following De Boef and Granato (1999), and the tests rejected the null of no cointegration.

These pretests reveal that the data could be either stationary or individually integrated and jointly cointegrated. Thus, I fit both ADL and ECM models and compare the results. By implementing both the ADL and ECM, I also connect my analysis to the ongoing debate on whether the ECM is appropriate for use with political time series data. In the time series symposium at *Political Analysis* 24(1), one of the points of debate between Grant and Lebo

[37] See online appendix A7 for further details about these control variables.

(2016) (GL, hereafter) and Keele, Linn and Webb (2016) (KLW, hereafter) is whether the ADL and ECM produce the same inference for the long-run effect. GL's analysis of the long-run relationship between supreme court approval and congressional approval suggests that the ADL and ECM produce inconsistent inferences. However, KLW point out that it is problematic for GL to base their assessment of the long-run relationship on the statistical significance of individual terms in the model. Instead, KLW argue that the long-run effect be estimated directly. The Bayesian inference used in my analysis provides the estimates of the long-run effects with their statistical uncertainty and permits a direct comparison of the ADL and ECM.

Prior Distribution

As in the first example, I employ a prior gamma distribution on the shrinkage parameter that imposes larger shrinkage on higher lags.[38] Moreover, I consider two alternative specifications of the gamma prior for robustness checks.[39] These alternative specifications do not change the substantive inferences drawn by the main prior specification. I only present the main results.

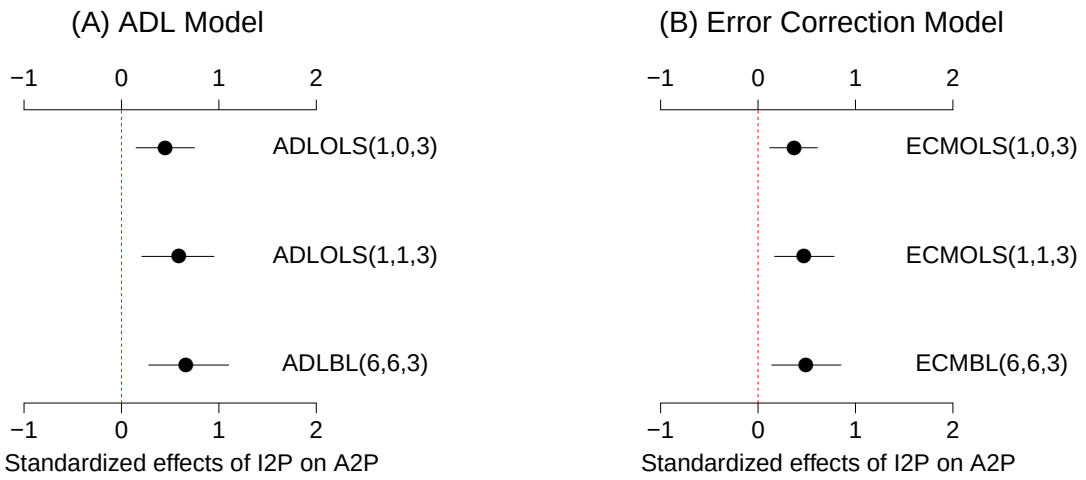
Results

My analysis focuses on comparing the estimated long-run effects obtained from three different dynamic specifications: a partial adjustment model, or ADL(1,0,3), a model with a 1-month lag, or ADL(1,1,3), and a general model ADL(6,6,3). I estimate ADL(1,0,3) and ADL(1,1,3) via Bayesian Gaussian linear regression and estimate ADL(6,6,3) via the Bayesian adaptive lasso. I call these three models ADLOLS(1,0,3), ADLOLS(1,1,3), and ADLBL(6,6,3). I also estimate long-run effects based on three Bayesian ECM models: EC-

[38]The rate parameters for the gamma distribution are given by $\delta_0 = 5, \delta_1 = 3, \delta_2 = 1, \delta_3 = .7, \delta_4 = .5, \delta_5 = .3, \delta_6 = .2$, where m in δ_m denotes a lag order. The shape parameter is fixed at 1 across lag orders. Given a fixed shape parameter, a larger value for the rate parameter generates a gamma distribution that yields smaller values.

[39]First has rate parameters $\delta_0 = 1.8, \delta_1 = 1.6, \delta_2 = 1.4, \delta_3 = 1.2, \delta_4 = 1, \delta_5 = .8, \delta_6 = .9, \delta_7 = .4,$ and $\delta_8 = .2$. Second has a fixed rate parameter $\delta = .1$ across lag orders. For both, the shape parameter is fixed at 1.

Figure 4: Comparison of the Estimated Dynamic Pattern for U.S. Actions toward Palestine: Effects of Israeli Actions toward Palestine



[*Notes*: This figure depicts the estimated long-run effects of Israeli actions toward Palestine (I2P) on U.S. actions toward Palestine (A2P) and their statistical uncertainty. Estimates are obtained from three different ADL models, ADLOLS(1,0,3), ADLOLS(1,1,3), and ADLBL(6,6,3), and three different ECMs, ECMOLS(1,0,3), ECMOLS(1,1,3), and ECMBL(6,6,3). Each dot is a point estimate of the effect of one additional standard deviation of I2P, which comes from the posterior mean. Each horizontal line represents a 95% Bayesian credible interval.]

MOLS(1,0,3), ECMOLS(1,1,3), and ECMBL(6,6,3).[40] I standardize all variables for a direct comparison.

Figure 4 presents the results.[41] The long-run effect of I2P on A2P is substantively large and statistically reliable in all models. The point estimate from ADLBL(6,6,3), which comes from the posterior mean, indicates that one additional standard deviation of I2P increases approximately 0.7 standard deviation of A2P. Based on the data at hand, it is interpreted that a one unit increase in I2P increases A2P by about 0.14 units.[42] This estimated effect is

[40]Given the general ECM(p, q, k) below, the long-run effect (LRE) of X_j on Y is calculated as follows:

$$\Delta Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{k} \beta_j \Delta X_{jt} + \sum_{j=1}^{k} \sum_{i=1}^{q} \gamma_{ji} X_{jt-i} + \varepsilon_t,$$

$$\text{LRE}_{\text{ECM}} = -\frac{\sum_{i=0}^{q} \gamma_i}{\sum_{i=1}^{p} \alpha_i}.$$

[41]The Breucsh-Godfrey LM test provides no evidence of autocorrelated residuals in the three models. The ADLBL and ECMBL estimates come from 50,000 iterations of the Gibbs sampler after 10,000 burn-in iterations. Using the `superdiag` package, I performed a test for convergence of the MCMC output. I found no evidence of non-convergence based on all of the standard procedures.

[42]In the data, one standard deviation of I2P is approximately 2.0, and that of A2P is approximately 0.4.

not statistically different from the effects estimated by the other ADL models.

This inference is consistent with BCF's finding of the positive relationship between I2P and A2P. My analysis also reveals that considering long lag lengths does not change substantive inferences. This indicates that there is no substantial effect at high lags, which is similar to the inflation case in the first empirical example. That is, a shock in I2P affects A2P in a relatively immediate way and dissipates quickly. Nevertheless, estimating the general model using ADLBL is still useful as a tool for checking the robustness of dynamic specifications. Since ADLBL provides a reliable estimation result that suggests no substantial effects of high order lags, I now know that employing a short lag length is valid for the BCF data.

Finally, my analysis offers a commentary on the debate about the use of the ECM. Recall that one of the points of debate between GL and KLW is whether the ADL and ECM produce the same inference: GL argue that the ECM leads to inferential mistakes whereas KLW suggest that the ADL and ECM produce the same inference when the data are stationary. My analysis of the BCF data provides evidence that the ADL and ECM draw statistically identical inferences. For instance, the point estimate of the long-run effect estimated by ADLBL (6,6,3) is 0.66, and that estimated by ECMBL (6,6,3) is 0.54. These point estimates are not statistically different according to the credible intervals. While GL's assessment is based on the significance of individual variables without calculating the long-run effect or its standard error, the Bayesian inference used in my analysis provides the estimates of the long-run effects with their statistical uncertainty.

6. Conclusion

Uncertainty about the lag structure in ADL modeling commonly raises the following issues: 1) restricting the lag structure *a priori* yields underspecification; 2) estimating a parameter-rich general model having many lags avoids underspecification, but overfits the data; 3) the general-to-specific modeling strategy (GSMS) is not a solution in many situations where a parameter-rich general model has large numbers of nested specifications, making it infeasible to search for the "best" model.

I have developed a solution to these issues: ADL modeling via the Bayesian adaptive lasso (ADLBL). As ADLBL fits a general model with sufficiently many lags, the estimation results

are safe from underspecification. Moreover, as penalized regression enacted by ADLBL mitigates overfitting, it is reliable to make inferences based on a general model directly. The ADLBL approach is particularly advantageous when substantive theory does not provide precise guidance for dynamic specifications. Yet it is also beneficial even when analysts have a strong theoretical justification because it is used as a tool for checking the robustness of theory-driven dynamic specifications.

ADLBL opens up new avenues for exploring a broad array of topics in a way that previous studies have been unable to do. The ADLBL approach is applicable to a variety of time series data and research questions as long as the assumptions of stationary time series, weakly exogenous regressors, and white noise error are held. Moreover, potential applications of ADLBL also include time-series cross-sectional (TSCS) data provided that heterogeneity across units in TSCS data is appropriately addressed (Beck and Katz 1995; Shor et al. 2007). Previous research on TSCS analyses published in political science articles find that many TSCS models are sensitive to alternative dynamic specifications (Wilson and Butler 2007). Thus, it will be worth applying the ADLBL approach to TSCS models to take model uncertainty seriously.

Despite the usefulness of ADLBL, I conclude by noting some limitations. First, ADLBL assumes weakly exogenous regressors. Analysts should be careful about this assumption when employing ADLBL. If the assumption of exogenous regressors does not hold, analysts may consider lasso-based approaches for vector autoregression models (Gefang 2014; Hsu, Hung and Chang 2008). Second, ADLBL begins by making an assumption about the nature of the general structure. Researchers whose primary goal is to derive substantive implications must check the robustness of their findings to different assumptions on the general structure.

Even given these last points, this new tool enables political science researchers to analyze the temporal nature of political events in a rigorous and useful way. In addition, this tool addresses the modern problem of modeling big data or high-dimensional data in the context of time series analysis for the social sciences.

Appendix

A1. Complete results for Table 1

Table 1 in the main text only reports the results for New York Time coverage. The table below reports further details of the estimation results. This table also includes the results for Model 1 that is omitted in Table 1.

Table 3: Replication of Table 5 in De Boef and Keele (2008) with An Alternative ADL Model: Effect of Different Model Specifications

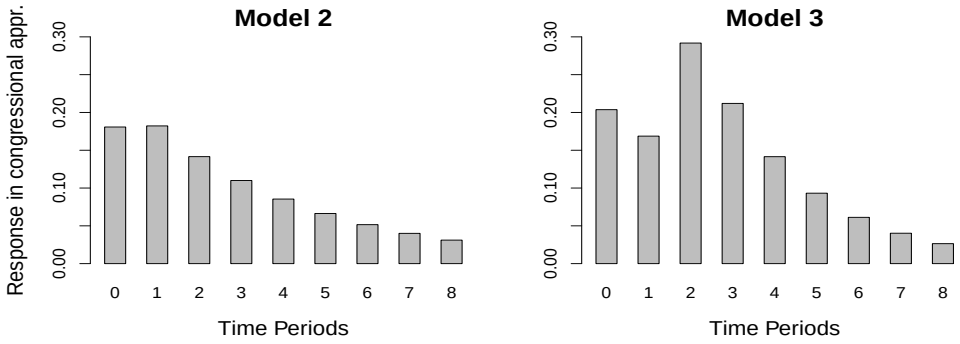
	Model 1	Model 2	Model 3
Congressional Approval$_{t-1}$	0.77	0.78	0.77
	(0.06)	(0.06)	(0.12)
Congressional Approval$_{t-2}$			−0.07
			(0.11)
Presidential Approval	0.05	0.11	0.15
	(0.04)	(0.06)	(0.06)
Presidential Approval$_{t-1}$		−0.09	−0.10
		(0.06)	(0.06)
Economic Expectations	0.08	0.02	0.02
	(0.03)	(0.07)	(0.07)
Economic Expectations$_{t-1}$		0.06	−0.02
		(0.07)	(0.10)
Economic Expectations$_{t-2}$			0.10
			(0.07)
NY Times Coverage	0.20	0.18	0.20
	(0.07)	(0.07)	(0.07)
NY Times Coverage$_{t-1}$		0.04	0.01
		(0.07)	(0.07)
NY Times Coverage$_{t-2}$			0.18
			(0.08)
Constant	10.13	9.93	14.05
	(3.41)	(3.63)	(4.05)
LM Test p-value	0.49	0.67	0.53
AIC	404.35	405.32	404.14
Num. obs.	78	78	78

[*Notes*: Model 1 is Durr, Gilmour, and Wolbrecht's model. Model 2 is De Boef and Keele's model. Model 3 is an alternative model. All of the three models are estimated by ordinary least squares. Other non-dynamic variables from the analysis are included in the estimating equations but omitted from the table. Standard errors are in parentheses. The highlighted boxes are for emphasis. The Breucsh-Godfrey LM test suggests no evidence of autocorrelated residuals in the three models as the LM test p-values show.]

A2. Impulse response function

Graphical representations of impulse response provide further clarifications for the difference in statistical inferences between Model 2 and Model 3 in Table 1 of the main text. Each bar in Figure 5 represents a change in congressional approval at a given time responding to one unit change in New York Times coverage at $t = 0$. The overall effect of media coverage is greater in Model 3 than in Model 2. Moreover, Model 3 finds that the estimated change in congressional approval responding to a shock in media coverage peaks at $t + 2$ and then decays geometrically, which is a finding Model 2 does not capture.

Figure 5: Comparison of Estimated Changes in Congressional Approval Responding to an Impulse Shock in New York Times Coverage of Congress



[*Notes*: Bar plots equivalent to impulse response functions. Each bar represents a change in congressional approval at a given time responding to a shock in New York Times coverage at $t = 0$. The plots are based on the results in Table 1 in the main text. Model 2 and Model 3 correspond to Model 2 and Model 3 in Table 1.]

A3. Deriving the full conditional distributions of the Bayesian adaptive lasso

Consider the following Bayesian adaptive lasso model:[43]

$$\begin{aligned}\pi(\boldsymbol{\theta},\sigma^2|\boldsymbol{y}) &\propto f(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2)\pi(\sigma^2)\prod_{g=1}^{G}\pi(\theta_g|\tau_g^2,\sigma^2)\pi(\tau_g^2)\\ &\propto \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{Z\theta})'(\boldsymbol{y}-\boldsymbol{Z\theta})\right)\\ &\quad\times\frac{\gamma^a}{\Gamma(a)}(\sigma^2)^{-a-1}\exp(-\gamma/\sigma^2)\\ &\quad\times\prod_{g=1}^{G}\frac{1}{(2\pi\sigma^2\tau_g^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2\tau_g^2}\theta_g^2\right)\frac{\lambda_g^2}{2}\exp(-\lambda_g^2\tau_g^2/2),\end{aligned}\tag{20}$$

where $\boldsymbol{y}$ and $\boldsymbol{Z}$ are standardized without loss of generality. The joint posterior becomes proportional to

$$\begin{aligned}&\frac{1}{(\sigma^2)^{(T-1)/2}}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{Z\theta})'(\boldsymbol{y}-\boldsymbol{Z\theta})\right)(\sigma^2)^{-a-1}\exp(-\gamma/\sigma^2)\\ &\times\prod_{g=1}^{G}\frac{1}{(\sigma^2\tau_g^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2\tau_g^2}\theta_g^2\right)\frac{\lambda_g^2}{2}\exp(-\lambda_g^2\tau_g^2/2).\end{aligned}\tag{21}$$

Given this joint posterior density, I derive the full conditional posterior distributions. To begin, derive the full conditional posterior for $\boldsymbol{\theta}$. By conjugacy, the full conditional for $\boldsymbol{\theta}$ is multivariate normal. In order to derive the posterior mean and posterior variance of the multivariate normal distribution, focus on the exponential terms involving $\boldsymbol{\theta}$ as follows:

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{Z},\boldsymbol{y},\sigma^2,\tau_1^2,...,\tau_g^2) &\propto \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{Z\theta})'(\boldsymbol{y}-\boldsymbol{Z\theta})\right)\times\prod_{g=1}^{G}\exp\left(-\frac{1}{2\sigma^2\tau_g^2}\theta_g^2\right)\\ &\propto \exp\left(\frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{Z\theta})'(\boldsymbol{y}-\boldsymbol{Z\theta})+\frac{1}{\sigma^2}\boldsymbol{\theta}'\boldsymbol{D}_\tau^{-1}\boldsymbol{\theta}\right)\\ &\propto \exp\left(\frac{1}{\sigma^2}[\boldsymbol{y}'\boldsymbol{y}-2\boldsymbol{\theta}'\boldsymbol{Z}'\boldsymbol{y}+\boldsymbol{\theta}'\boldsymbol{Z}'\boldsymbol{Z\theta}+\boldsymbol{\theta}\boldsymbol{D}_{\boldsymbol{\tau}}^{-\mathbf{1}}\boldsymbol{\theta}]\right)\\ &\propto \exp\left(\frac{1}{\sigma^2}[\boldsymbol{y}'\boldsymbol{y}-2\boldsymbol{\theta}'\boldsymbol{Z}'\boldsymbol{y}+\boldsymbol{\theta}'(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{D}_\tau^{-1})\boldsymbol{\theta}]\right).\end{aligned}\tag{22}$$

[43]Here, the prior distribution of σ^2 follows a gamma distribution. The posterior distribution of σ^2 based on the gamma prior distribution on σ^2 corresponds to using the non-informative scale-invariant prior $1/\sigma^2$ on σ^2 when $a=\gamma=0$. Without loss of generality I omit subscript m for the gamma prior distribution on the shrinkage parameter, which is used in the main text to reveal that the gamma prior distribution varies by lag order.

The posterior precision comes from the expression between $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$, ending up with $(\sigma^2)^{-1}(\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{D}_\tau^{-1})$. Thus, the posterior variance is $\sigma^2(\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{D}_\tau^{-1})^{-1}$. The posterior mean comes from the quantity premultiplying the coefficient of $-2\boldsymbol{\theta}'$ by the posterior variance. The coefficient of $-2\boldsymbol{\theta}'$ is $(\sigma^2)^{-1}\boldsymbol{Z}'\boldsymbol{y}$, and premultiplying it by the posterior variance ends up the posterior mean $(\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{D}_\tau^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{y}$. Therefore, the full conditional posterior for $\boldsymbol{\theta}$ follows a multivariate normal distribution such that

$$\pi(\boldsymbol{\theta}|\boldsymbol{Z}, \boldsymbol{y}, \sigma^2, \tau_1^2, ..., \tau_g^2) \sim N((\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{D}_\tau^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{y}, \sigma^2(\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{D}_\tau^{-1})^{-1}). \tag{23}$$

To derive the full conditional posterior for σ^2, focus on the terms in the joint posterior involving σ^2:

$$\begin{aligned}
&\pi(\sigma^2|\boldsymbol{Z}, \boldsymbol{y}, \boldsymbol{\theta}, \tau_1^2, ..., \tau_g^2) \\
&\quad \propto \frac{1}{(\sigma^2)^{(T-1)/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\theta})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\theta})\right) (\sigma^2)^{-a-1} \exp(-\gamma/\sigma^2) \\
&\qquad \times \prod_{g=1}^{G} \frac{1}{(\sigma^2\tau_g^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2\tau_g^2}\theta_g^2\right) \\
&\quad \propto \frac{1}{(\sigma^2)^{(T-1)/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\theta})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\theta})\right) (\sigma^2)^{-a-1} \exp(-\gamma/\sigma^2) \\
&\qquad \times \frac{1}{(\sigma^2)^{G/2}} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\theta}'\boldsymbol{D}_\tau^{-1}\boldsymbol{\theta}\right) \\
&\quad \propto (\sigma^2)^{\frac{-(T-1)}{2} - \frac{G}{2} - a - 1} \exp\left(-\frac{\frac{1}{2}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\theta})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\theta}) + \frac{1}{2}(\boldsymbol{\theta}'\boldsymbol{D}_\tau^{-1}\boldsymbol{\theta}) + \gamma}{\sigma^2}\right)
\end{aligned} \tag{24}$$

By conjugacy, the full conditional posterior for σ^2 is inverse-gamma distributed. The superscript of the first part in the last line (24) is the shape parameter, and the numerator of the exponential term is the scale parameter. That is, the full conditional posterior is expressed as

$$\pi(\sigma^2|\boldsymbol{Z}, \boldsymbol{y}, \boldsymbol{\theta}, \tau_1^2, ..., \tau_g^2) \quad \sim \quad \text{IG}\left(\frac{T-1}{2} + \frac{G}{2} + a,\ \frac{1}{2}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\theta})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\theta}) + \frac{1}{2}(\boldsymbol{\theta}'\boldsymbol{D}_\tau^{-1}\boldsymbol{\theta}) + \gamma\right). \tag{25}$$

Now, move to deriving the full conditional for τ_g^2. For each $g = 1, 2, ...G$, the portion of

the joint distribution involving τ^2 is

$$(\tau_g^2)^{-1/2} \exp\left[-\frac{1}{2}\left(\frac{\theta_g^2/\sigma^2}{\tau_g^2} + \lambda_g^2 \tau_g^2\right)\right]. \tag{26}$$

Park and Casella (2008) show that the quantity (26) can be proportional to an inverse-Gaussian distribution. One popular parameterization of the inverse-Gaussian density is

$$f(x) = \sqrt{\frac{\kappa}{2\pi}} x^{-3/2} \exp\left[-\frac{\kappa(x-\varphi)^2}{2\varphi^2 x}\right], \quad x > 0 \tag{27}$$

where $\varphi > 0$ is the mean parameter and $\kappa > 0$ is the scale parameter. Below shows that the full conditional posterior of $1/\tau_g^2$ compares with the above parameterization, and that the full conditional is inverse-Gaussian distributed. Let $\eta_g^2 = 1/\tau_g^2$. Then,

$$\begin{aligned}
\pi(\eta_g^2|\boldsymbol{\theta}, \sigma^2, \boldsymbol{Z}, \boldsymbol{y}) &\propto (\eta_g^2)^{-3/2} \exp\left[-\frac{1}{2}\left(\frac{\theta_g^2 \eta_g^2}{\sigma^2} + \frac{\lambda_g^2}{\eta_g^2}\right)\right] \\
&\propto (\eta_g^2)^{-3/2} \exp\left[-\left(\frac{\theta_g^2 \eta_g^4 + \lambda_g^2 \sigma^2}{2\sigma^2 \eta_g^2}\right)\right] \\
&\propto (\eta_g^2)^{-3/2} \exp\left[-\left\{\frac{\theta_g^2(\eta_g^4 + \frac{\lambda_g^2\sigma^2}{\theta_g^2})}{2\sigma^2\eta_g^2}\right\}\right] \\
&\propto (\eta_g^2)^{-3/2} \exp\left[-\frac{\theta_g^2\left\{\left(\eta_g^2 - \sqrt{\frac{\lambda_g^2\sigma^2}{\theta_g^2}}\right)^2 + 2\eta_g^2\sqrt{\frac{\lambda_g^2\sigma^2}{\theta_g^2}}\right\}}{2\sigma^2\eta_g^2}\right] \\
&\propto (\eta_g^2)^{-3/2} \exp\left[-\frac{\theta_g^2\left(\eta_g^2 - \sqrt{\frac{\lambda_g^2\sigma^2}{\theta_g^2}}\right)^2}{2\sigma^2\eta_g^2}\right].
\end{aligned} \tag{28}$$

Given the parameterization displayed in (27), the last line (28) compares with the inverse-Gaussian density: Let $\sqrt{\lambda_g^2\sigma^2/\theta_g^2} = \varphi$. Then, $\lambda_g^2\sigma^2/\theta_g^2 = \varphi^2$ and $\theta_g^2/\sigma^2 = \lambda_g^2/\varphi^2$. As a result,

the last line (28) is rewritten as

$$(\eta_g^2)^{-3/2} \exp\left[-\frac{\lambda_g^2\left(\eta_g^2-\varphi\right)^2}{2\varphi^2\eta_g^2}\right], \tag{29}$$

and $\pi(\eta_g^2|\boldsymbol{\theta},\sigma^2,\boldsymbol{Z},\boldsymbol{y})$ is proportional to the inverse-Gaussian density with mean parameter $\sqrt{\lambda_g^2\sigma^2/\theta_g^2}$ and scale parameter λ_g^2. Therefore, I can obtain the full conditional posterior of τ_g^2 from the following conditional distribution:

$$\pi(1/\tau_g^2|\boldsymbol{\theta},\sigma^2,\boldsymbol{Z},\boldsymbol{y}) \quad\sim\quad \text{Inverse-Gaussian} \quad \left(\sqrt{\frac{\lambda_g^2\sigma^2}{\theta_g^2}},\ \ \lambda_g^2\right). \tag{30}$$

To summarize, the full conditional posterior of the ADLBL model is give by

$$\begin{aligned}
\pi(\boldsymbol{\theta}|\boldsymbol{Z},\boldsymbol{y},\sigma^2,\tau_1^2,...,\tau_g^2) &\sim N((\boldsymbol{Z'Z}+\boldsymbol{D}_\tau^{-1})^{-1}\boldsymbol{Z'y},\sigma^2(\boldsymbol{Z'Z}+\boldsymbol{D}_\tau^{-1})^{-1}) \\
\pi(\sigma^2|\boldsymbol{Z},\boldsymbol{y},\boldsymbol{\theta},\tau_1^2,...,\tau_g^2) &\sim \text{IG}\left(\frac{T-1}{2}+\frac{G}{2}+a,\ \frac{1}{2}(\boldsymbol{y}-\boldsymbol{Z\theta})'(\boldsymbol{y}-\boldsymbol{Z\theta})+\frac{1}{2}(\boldsymbol{\theta'D}_\tau^{-1}\boldsymbol{\theta})+\gamma\right) \\
\pi(1/\tau_g^2|\boldsymbol{\theta},\sigma^2,\boldsymbol{Z},\boldsymbol{y}) &\sim \text{Inverse-Gaussian} \quad \left(\sqrt{\frac{\lambda_g^2\sigma^2}{\theta_g^2}},\ \ \lambda_g^2\right).
\end{aligned}$$

Note that the conditional posterior distribution for σ^2 presented above is the same as the distribution for σ^2 presented in the Gibbs algorithm in the main text when $a=\gamma=0$. As mentioned earlier, the prior distribution of σ^2 follows a gamma distribution in this formal derivation whereas the main text uses the non-informative scale-invariant prior $1/\sigma^2$ on σ^2. The posterior distribution of σ^2 based on the gamma prior distribution is equivalent to using the non-informative scale-invariant prior $1/\sigma^2$ on σ^2 when $a=\gamma=0$.

In addition, the shrinkage parameter λ_g should be appropriately determined. Following Park and Casella (2008) and Kyung et al. (2010), I employ a diffuse hyperprior for the shrinkage parameter. One important difference is that the previous studies do not implement adaptivity, but the following gamma prior has subscript g to reflect adaptive shrinkage:

$$\pi(\lambda_g^2)=\frac{\delta^r}{\Gamma(r)}(\lambda_g^2)^{r-1}\exp(-\delta\lambda_g^2),\ \ \lambda_g^2>0,\ \ r>0,\ \ \delta>0. \tag{31}$$

When this prior enters the lasso hierarchy, the product of the factors involving λ_g^2 in the joint posterior is

$$\begin{aligned}(\lambda_g^2)^{r-1}\exp(-\delta\lambda_g^2)\times\lambda_g^2\exp(-\lambda_g^2\tau_g^2/2) &= (\lambda_g^2)^{r+1-1}\exp(-\delta\lambda_g^2-\lambda_g^2\tau_g^2/2)\\ &= (\lambda_g^2)^{r+1-1}\exp[-\lambda_g^2(\delta+\tau_g^2/2)].\end{aligned}$$

Consequently, the full conditional distribution of λ_g^2 is gamma such that

$$\lambda_g^2|\boldsymbol{\theta},\sigma^2,\boldsymbol{D}_\tau,\boldsymbol{Z},\boldsymbol{y} \;\sim\; \text{Gamma}\left(r+1,\frac{\tau_g^2}{2}+\delta\right). \tag{32}$$

A4. An alternative scheme to measure presidential approval rates

In the analysis of presidential approval, when there are multiple polls for one quarter, I took the average of all approval rates in that quarter. In addition to this measurement scheme, I implemented an alternative scheme that selects one particular approval rating per quarter. For each of the quarters having multiple polls, I set up the following scheme: (1) For each quarter of the year, February, May, August, and November polls are first considered. (2) January, April, July, and October polls are considered for each quarter of the year unless their counterparts in the previous step are available. (3) Finally, March, June, September, and December are considered for each quarter unless their counterparts in the previous steps are available. (4) For each month, the earliest poll is used. (5) If there is still a missing quarter after steps (1) through (3), then the mean imputation method is applied by using the two immediate previous and later polls. The number of the cases for which the mean imputation is used is 3 out of the total 205 observations. This alternative scheme does not change substantive results.

A5. Control variables in the analysis of presidential approval

Table 4: Coding Scheme

Variable	Time
Moscow treaty	72:Q2
Paris treaty	73:Q1
Watergate scandal	73:Q2 - 73:Q4
Nixon's pardon	74:Q4
Mayaguez incident	75:Q2
Camp David Accords	78:Q4
Assassination attempt	81:Q2
Granada invasion	83:Q4
Iranscam	86:Q4
The Gulf war	91:Q1 - 91:Q4
The 9/11 attacks	01:Q4 - 02:Q4
Carter's Iran crisis	2 for 79:Q4 1 for 80:Q1 -1 for 81:Q2
The Vietnam war	1 for 64:Q1 - 64:Q4 2 for 65:Q1 - 65:Q4 3 for 66:Q1 - 66:Q4 4 for 67:Q1 - 67:Q4 5 for 68:Q1 - 68:Q4

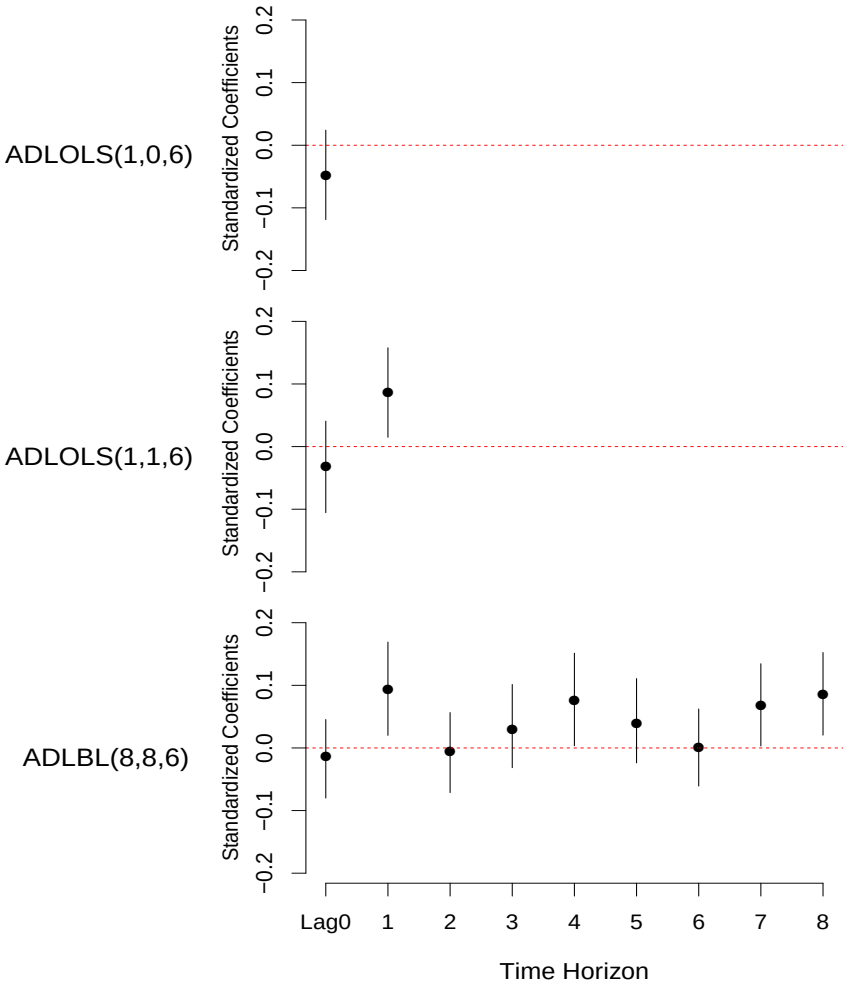
[*Notes*: Year:Quarter for each variable is presented. For example, 72:Q2 means the second quarter of 1972, and 02:Q4 means the fourth quarter of 2002. All variables except for the variables for Carter's Iran crisis and the Vietnam war are dummy variables.]

A6. Estimated dynamic marginal effects for presidential approval

Figure 6 presents the estimated dynamic marginal effects for presidential approval. I compare three different modeling approaches: ADLOLS(1,0,6), ADLOLS(1,1,6), and ADLBL(8,8,6). For the sake of a direct comparison, I standardize all coefficients without loss of generality.

ADLBL$(8, 8, 6)$ captures the dynamic pattern over the course of two years that the restricted models miss. It detects substantial positive lagged effects at lags 1, 4, 7, and 8 that are statistically reliable at the 90% level. This finding of substantial lagged effects suggests that neither the partial adjustment model nor ADL$(1, 1, 6)$ is consistent with the data gen-

Figure 6: Comparison of the Estimated Dynamic Pattern for Presidential Approval: Effects of Income Growth



[*Notes*: ADLOLS(1,0,6) refers to the partial adjustment model. ADLOLS(1,1,6) is an ADL model with the highest lag order of one. Both ADLOLS(1,0,6) and ADLOLS(1,1,6) are estimated by OLS. ADLBL(8,8,6) is a general model with the highest lag order of eight, and it is estimated by the Bayesian adaptive lasso. Standardized coefficients are presented for a direct comparison. Each vertical line corresponds to the 90% confidence (credible in Bayesian analysis) interval.]

erating process because these models assume zero effects at lags higher than 1, which turns out not to be the case in the general model.

A7. Control variables used in the analysis of U.S. policy responses to the Israeli-Palestinian conflict

Following BCF's main analysis, nine dummy and trend variables are included in the estimation equation. To distinguish four prime ministerial regimes, three variables for the identities of the Israeli prime ministers are included. For each regime, four trend variables are employed to capture electorally motivated cooperation and electorally motivated violence: a separate time counter that starts at the value 1 in the month after each Israeli election and increases by one until the time of the next constitutionally mandated election. Finally, two dummy variables account for changes in the trend of the mean level of conflict: from the start of the second Intifada to the start of the Battle of Jenin (October 2000 - April 2002) and the post-Battle of Jenin period (May 2002 - March 2005).

References

Achen, Christopher H. 2000. "Why lagged dependent variables can suppress the explanatory power of other independent variables." Presented at the Annual Meeting of Political Methodology, Los Angeles.

Andrews, David F. and Colin L. Mallows. 1974. "Scale mixtures of normal distributions." *Journal of the Royal Statistical Society. Series B (Methodological)* 36(1):99–102.

Babyak, Michael A. 2004. "What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models." *Psychosomatic Medicine* 66(3):411–421.

Banerjee, Anindya, Juan J Dolado, John W Galbraith, David Hendry et al. 1993. *Co-integration, error correction, and the econometric analysis of non-stationary data*. Oxford University Press.

Beck, Nathaniel. 1991. "Comparing dynamic specifications: The case of presidential approval." *Political Analysis* 3(1):51–87.

Beck, Nathaniel and Jonathan N. Katz. 1995. "What to do (and not to do) with time-series cross-section data." *American Political Science Review* 89(3):634–647.

Box-Steffensmeier, Janet M., John R. Freeman, Matthew P. Hitt and Jon C. Pevehouse. 2014. *Time Series Analysis for the Social Sciences*. Cambridge University Press.

Brandt, Patrick T. and John R. Freeman. 2009. "Modeling macro-political dynamics." *Political Analysis* 17(2):113–142.

Brandt, Patrick T. and John T. Williams. 2007. *Multiple Time Series Models*. Sage.

Brandt, Patrick T., Michael Colaresi and John R. Freeman. 2008. "The dynamics of reciprocity, accountability, and credibility." *Journal of Conflict Resolution* 52(3):343–374.

De Boef, Suzanna and Jim Granato. 1999. "Testing for cointegrating relationships with near-integrated data." *Political Analysis* 8(1):99–117.

De Boef, Suzanna and Jonathan Nagler. 2005. "Do voters really care who gets what? Economic growth, economic distribution, and presidential popularity." Presented at the Annual Meeting of the Midwest Political Science Association, Chicago.

De Boef, Suzanna and Luke Keele. 2008. "Taking time seriously." *American Journal of Political Science* 52(1):184–200.

De Mol, Christine, Domenico Giannone and Lucrezia Reichlin. 2008. "Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?" *Journal of Econometrics* 146(2):318–328.

Durr, Robert H., John B. Gilmour and Christina Wolbrecht. 1997. "Explaining congressional approval." *American Journal of Political Science* 41(1):175–207.

Enders, Walter. 2015. *Applied Econometric Time Series*. Fourth ed. Wiley.

Erikson, Robert S. 1989. "Economic conditions and the presidential vote." *American Political Science Review* 83(2):567–573.

Esarey, Justin. 2016. "Fractionally integrated data and the autodistributed lag model: Results from a simulation study." *Political Analysis* 24(1):42–49.

Freeman, John R. 2016. "Progress in the study of nonstationary political time Series: A Comment." *Political Analysis* 24(1):50–58.

Gefang, Deborah. 2014. "Bayesian doubly adaptive elastic-net Lasso for VAR shrinkage." *International Journal of Forecasting* 30(1):1–11.

Gill, Jeff. 2014. *Bayesian Methods: A Social and Behavioral Sciences Approach*. Third ed. CRC Press.

Grant, Taylor and Matthew J Lebo. 2016. "Error correction methods with political time series." *Political Analysis* 24(1):3–30.

Greene, William H. 2008. *Econometric Analysis*. Sixth ed. Pearson Education.

Harrell, Frank. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer.

Helgason, Agnar Freyr. 2016. "Fractional integration methods and short time series: Evidence from a simulation study." *Political Analysis* 24(1):59–68.

Hendry, David F. 1995. *Dynamic Econometrics*. Oxford University Press.

Hibbs, Douglas A. 1982. "President Reagan's mandate from the 1980 elections: A Shift to the Right?" *American Politics Quarterly* 10(4):387–420.

Hibbs, Douglas H. 1987. *The American Political Economy: Macroeconomics and Electoral Politics in the United States*. Havard University Press.

Hothorn, Torsten, Achim Zeileis, Richard W Farebrother, Clint Cummins, Giovanni Millo, David Mitchell and Maintainer Achim Zeileis. 2015. "lmtest: Testing linear regression models." *R package* .

Hsu, Nan-Jung, Hung-Lin Hung and Ya-Mei Chang. 2008. "Subset selection for vector autoregressive processes using lasso." *Computational Statistics and Data Analysis* 52(7):3645–3657.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2013. *Introduction to Statistical Learning*. Springer.

Keele, Luke and Nathan J. Kelly. 2006. "Dynamic models for dynamic theories: The ins and outs of lagged dependent variables." *Political Analysis* 14(2):186–205.

Keele, Luke, Suzanna Linn and Clayton McLaughlin Webb. 2016. "Treating time with all due seriousness." *Political Analysis* 24(1):31–41.

Kyung, Minjung, Jeff Gill, Malay Ghosh and George Casella. 2010. "Penalized regression, standard errors, and Bayesian lassos." *Bayesian Analysis* 5(2):369–411.

Leng, Chenlei, Minh-Ngoc Tran and David Nott. 2014. "Bayesian adaptive lasso." *Annals of the Institute of Statistical Mathematics* 66(2):221–244.

Lockhart, Richard, Jonathan Taylor, Ryan J Tibshirani and Robert Tibshirani. 2014. "A significance test for the lasso." *Annals of Statistics* 42(2):413.

MacKuen, Michael B, Robert S Erikson and James A Stimson. 1992. "Peasants or bankers? The American electorate and the U.S. economy." *American Political Science Review* 86(3):597–611.

Markus, Gregory B. 1988. "The impact of personal and national economic conditions on the presidential vote: A pooled cross-sectional analysis." *American Journal of Political Science* 32(1):137–154.

Martin, Andrew D., Kevin M. Quinn and Jong Hee Park. 2011. "MCMCpack: Markov chain Monte Carlo in R." *Journal of Statistical Software* 42(9):1–21.

McNeish, Daniel M. 2015. "Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences." *Multivariate Behavioral Research* 50(5):471–484.

Montgomery, Jacob M. and Brendan Nyhan. 2010. "Bayesian model averaging: Theoretical developments and practical applications." *Political Analysis* 18(2):245–270.

Park, Trevor and George Casella. 2008. "The Bayesian lasso." *Journal of the American Statistical Association* 103(482):681–686.

Pfaff, Berhard, Eric Zivot and Matthieu Stigler. 2010. "urca: Unit root and cointegration tests for time series data." *R package* .

Ratkovic, Marc and Dustin Tingley. 2017. "Sparse estimation and uncertainty with application to subgroup analysis." *Political Analysis* DOI: https://doi.org/10.1017/pan.2016.14.

Sattler, Thomas, John R. Freeman and Patrick T. Brandt. 2007. "Political accountability and the room to maneuver: A search for a causal chain." *Comparative Political Studies* 41(9):1212–1238.

Shor, Boris, Joseph Bafumi, Luke Keele and David Park. 2007. "A Bayesian multilevel modeling approach to time-series cross-sectional data." *Political Analysis* 15(2):165–181.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1):267–288.

Trapletti, Adrian and Kurt Hornik. 2011. "tseries: Time series analysis and computational finance." *R package* .

Tsai, Tsung-han and Jeff Gill. 2012. "A comprehensive test suite for Markov chain non-convergence." *The Political Methodologist* 19:12–18.

Wang, Hansheng, Guodong Li and Chih-Ling Tsai. 2007. "Regression coefficient and autoregressive order shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69(1):63–78.

Wilson, Sven E. and Daniel M. Butler. 2007. "A lot more to do: The sensitivity of time-series cross-section analyses to simple alternative specifications." *Political Analysis* 15(2):101–123.

Zou, Hui. 2006. "The adaptive lasso and its oracle properties." *Journal of the American Statistical Association* 101(476):1418–1429.