LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Ingrid Mauerer, Wolfgang Pößnecker, Paul W. Thurner, Gerhard Tutz

Modeling electoral choices in multiparty systems with high-dimensional data: A regularized selection of parameters using the Lasso approach

Technical Report Number 158, 2014
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Modeling electoral choices in multiparty systems with high-dimensional data: A regularized selection of parameters using the Lasso approach

Ingrid Mauerer*[1], Wolfgang Pößnecker†[2], Paul W. Thurner‡[1], and Gerhard Tutz§[2]

[1] Geschwister-Scholl-Institute of Political Research, University of Munich (LMU)
[2] Department of Statistics, University of Munich (LMU)

April 17, 2014

Abstract

The increasing popularity of the Spatial Theory of Voting has given rise to the frequent usage of multinomial logit/probit models with alternative-specific covariates. The flexibility of these models comes along with one severe drawback: the proliferation of coefficients, resulting in high-dimensional and difficult-to-interpret models. In particular, choice models in a party system with J parties result in maximally $J-1$ parameters for chooser-specific attributes (e.g., sex, age). For the specification of alternative-specific attributes (e.g., issue distances), maximally J parameters can be estimated. Thus, a model of party choice with five parties based on three issues and ten voter attributes already produces 59 possible coefficients. As soon as we allow for interaction effects to detect segment-specific reactions to issues, the situation is even aggravated. In order to systematically identify relevant predictors in spatial voting models, we derive and use for the first time Lasso-type regularized parameter selection techniques that take into account both individual- and alternative-specific variables. Most importantly, our new algorithm can handle the alternative-wise specification of issue distances. Applying the Lasso method to the 2009 German Parliamentary Election, we demonstrate that our approach massively reduces the model's complexity and simplifies its interpretation. Lasso-penalization clearly outperforms the simple ML estimator.

Keywords: Spatial Theory of Voting, Multinomial Logit Model, Regularization, Variable Selection, Lasso.

*Ingrid.Mauerer@gsi.uni-muenchen.de
†Wolfgang.Poessnecker@stat.uni-muenchen.de
‡Paul.Thurner@gsi.uni-muenchen.de
§Gerhard.Tutz@stat.uni-muenchen.de

1 Introduction

The Spatial Theory of Voting has become increasingly important and influential both for our understanding of the mechanisms of party competition and individual choice behavior (e.g., Alvarez and Nagler, 1998; Thurner, 2000; Dow and Endersby, 2004; Adams, Merrill, and Grofman, 2005; Schofield et al., 1998). Since Downs' seminal work on democracy as a political market (Downs, 1957), a growing number of scholars theoretically as well as technically developed this approach. A central point in this regard has been the appropriate translation of the spatial voting theory into a statistical model. Alvarez and Nagler (1998) have shown that the Spatial Theory of Voting can be adequately applied by utilizing multinomial logit or probit models which allow us to consider both attributes of voters (i.e., individual- or chooser-specific variables such as age, sex, etc.) as well as characteristics of the parties/candidates (i.e., alternative- or choice-specific variables such as issue positions/distances).[1] In addition, Dow and Endersby (2004) provide a comparison of multinomial logit and probit models, and King, Tomz, and Wittenberg (2000) offer a series of helpful tools for the substantial interpretation of such complex models. Finally, Adams, Merrill, and Grofman (2005) theoretically combine policy and non-policy factors in an integrated model of party competition.

The increased popularity of these discrete choice models in the analysis of multiparty elections is due to their flexibility. For example, it is possible to specify choice-specific attributes in an alternative-specific way. That is, instead of assuming that voters are equally sensitive with regard to all parties when they evaluate issue distances towards all parties (an assumption generally implied by estimating only one issue distance parameter for all J considered parties), the specification of alternative-specific coefficients allows estimating as many issue distance parameters as parties are competing.[2] By removing this statistical "restriction of equality", the electoral researcher is able to identify for which specific party choice issues are relevant predictors. Consider, for instance, voters' choice among multiple parties and suppose that these parties offer stances on the issue of environment. Due to their attributed competence or "ownership" on environmental issues (Budge and Farlie, 1983; Petrocik, 1996), we would suppose that the choice of Green Parties is more strongly determined by this issue as compared to other parties. As a result, we would expect that the issue distance coefficients vary considerably across parties, and that only subsets of these coefficients prove to be relevant determinants of party choice in spatial models of voting.[3]

[1] In the following we use the general term "multinomial logit model" to refer to multinomial logit models including both alternative- and individual-specific variables. In addition, we use the terms "alternative-specific", "choice-specific" and "party-specific" covariates interchangeably to refer to attributes of alternatives (i.e., parties/candidates). The same applies to the terms "chooser-specific", "voter-specific" or "individual-specific" variables to refer to voter characteristics.

[2] This specification, which has received little attention and is greatly unexamined in the empirical study of spatial issue voting (see as an exception Thurner, 2000), is widely used and applied in transportation economics and econometrics. See Ben-Akiva and Lerman (1985); Train (1978), or Louviere, Hensher, and Swait (2000).

[3] Using the suggested models and the alternative-wise specification, Thurner (2000) showed in his

This flexibility gives also rise to new research questions and hypotheses, which will improve our understanding of multiparty competition and individual vote choice considerably. For instance, in a recent contribution, Mauerer, Thurner, and Debus (2014) explicitly questioned why we should expect that not every party is equally successful in attracting voters based on their position-taking, and therefore observe so-called "party-specific issue reactions" with quite different spatial equilibria.[4]

These recent developments demonstrate not only the theoretically promising advantage of multinomial logit and probit models, but also highlight one challenging drawback: the proliferation of possible coefficients, and hence the need for sophisticated parameter selection techniques. To be precise, party choice in a party system with J parties results in maximally $J-1$ parameters in the case of chooser-specific attributes (e.g., sex, age). For the specification of alternative-specific attributes (e.g., issue distances), maximally J parameters can be estimated. Consequently, the amount of possible individual- and party-specific coefficients increases rapidly, resulting in highly complex and difficult-to-interpret models. Moreover, as soon as we allow for (theoretically derived) interaction effects (e.g., to test for segment-specific reactions to issue distances and to identify so-called issue publics), the situation is even aggravated. For these reasons, the following question needs to be addressed: How can electoral researchers systematically identify relevant predictors and parameters in models for which very many predictors are available? A solution to this problem would enormously reduce model complexity and facilitate the interpretation. We propose the Lasso approach which is a regularized parameter selection technique that guarantees – in contrast to classical subset selection approaches – continuous, stable and computationally efficient variable selection.

Our objective in this paper is to illustrate the benefits of regularization methods in the statistical analysis of the Spatial Theory of Voting. In particular, we introduce and derive for the first time a Lasso-type regularization technique in the estimation of multinomial logit models (MNLs) which takes into account both individual- and alternative-specific variables. Most importantly, it allows us to handle the alternative-wise specification of choice-specific covariates (e.g., issue distances). These Lasso-type regularization methods penalize the L_1-norm of the coefficients, which enforces parameter selection and reduction of the predictor space. Actually, we demonstrate that our proposed approach massively reduces the complexity of spatial voting models and facilitates their interpretation by selecting the most important effects. We show that Lasso-penalization clearly outperforms

analysis of the 1990 German Parliamentary Election why we should split up the issue distance coefficients and how this specification strategy provides more detailed insights into the dynamics of multiparty competition. His results suggest that only several party-specific issue distance parameters turn out to be statistically significant. For example, with regard to the issue of German unification, the author found that on this particular dimension only the Christian Democratic Party proved to be a significant objective of issue voting.

[4] Applying the alternative-wise specification to position issues included in German Parliamentary Elections from 1987 to 2009, the authors discovered that issue reactions at the level of the voters vary substantially across parties and that in particular with regard to niche parties offering polar stances issue effects are much more likely.

the simple ML estimator.

The paper is structured as follows: To assess the advantage of regularized parameter selection within the spatial modeling of voter choices in multiparty systems with high-dimensional data, we first provide a short formal outline of the Spatial Theory of Voting. Second, we briefly examine classical variable selection procedures and their limitations and demonstrate why and how the usage of regularization methods in the study of multiparty elections enables us to efficiently identify important predictors, and therefore to improve spatial models of voting, and to facilitate their interpretation. For the illustration of the usefulness of the proposed approach, we provide a regularized analysis of party choice in the 2009 German Parliamentary Election.

2 The Formal Set-Up of the Spatial Theory of Voting

The Spatial Theory of Voting is based on the following theoretical assumptions:

a) The policy proposals of parties and the policy preferences of voters can be arrayed on K policy dimensions, which are assumed to be continuous and separable. These dimensions constitute the K-dimensional political space (Davis, Hinich, and Ordeshook, 1970, p. 430).

b) The voter's most preferred policy position on each k dimension, denoted by x_{ik} and called ideal point, is defined as a finite point of maximum utility.

c) Each party j takes policy positions on K policy dimensions, denoted by p_{jk}.

d) According to the principle of utility maximization, voter i compares the parties' policy proposals p_{ijk} and identifies each party's supplied amount of utility, denoted by $U_{ij}, j = 1, \dots, J$. The voter chooses the alternative (i.e., party/candidate) that provides the highest level of utility, which is the party whose policy positions are closest to the voter's ideal points:

$$U_{ij} > U_{ih} \quad \forall j \neq h,$$

where[5]

$$U_{ij} = -\left(\sum_{k=1}^{K} \alpha_k \left|x_{ik} - p_{ijk}\right|\right). \tag{1}$$

α_k presents the weight or saliency of each kth policy dimension. These coefficients are the utility parameters indicating the value attached to the dimensions.

e) Random Utility Maximization (Manski, 1973, 1977): The level of utility provided by each alternative is known with certainty by the voter, but not all determinants of the decision can be observed. This limitation at the level of the researcher is

[5] For a legitimation of the disaggregate City-Block metric, see Singh (2014).

captured by dividing the overall utility U_{ij} into two parts. The first part, denoted by V_{ij}, represents the contributions that are measured by the analyst. The second part defines the sources of utility that the researcher cannot observe, denoted by ϵ_{ij}:

$$U_{ij} = V_{ij} + \epsilon_{ij}.$$

As a result, the choice is probabilistic and cannot be predicted exactly.

f) The observed part of utility V_{ij} consists of two components that are connected additively: individual-specific and alternative-specific variables. Individual-specific or chooser-specific variables s_i refer to voter characteristics that vary over individuals i, but are constant over alternatives. Alternative-specific or choice-specific variables z_{ij} represent attributes of the alternatives (e.g., issue distances) and vary across both alternatives and individuals:

$$V_{ij} = z_{ij} + s_i.$$

Based on these assumptions, the Spatial Theory of Voting is translated into a statistical model by the following steps:

a) Data: A sample of n voters is available in which, for each voter i, a set of individual-specific variables $s_{il},\ l = 1, \ldots, p$, as well as alternative-specific variables $z_{ijk},\ k = 1, \ldots, K$, are observed. Variable z_{ijk} represents the distance between voter i and party j on policy dimension k and is defined by $z_{ijk} = -|x_{ik} - p_{ijk}|$.

b) Parameterization: Individual-specific variables s_{il} are always used with alternative-specific coefficients $\beta_{jl},\ j = 1, \ldots, J$. By contrast, the effect of alternative-specific variables can be specified in two different ways: constant (also called generic) and alternative-specific. Equation (1) is based on the constant specification which results in estimating one coefficient for all alternatives. Therefore, it assumes that the variable has an equal/constant effect on all alternatives. In contrast to the restrictive set-up with constant coefficients, we apply the alternative-wise specification which allows estimating J different parameters α_{jk} in the case of J alternatives, so that different effects on alternatives are possible (Ben-Akiva and Lerman, 1985; Train, 2009; Louviere, Hensher, and Swait, 2000; Thurner, 2000). Hence, we explicitly relax the assumption that all voters react identically to the position-taking of all parties.

c) Linear predictors: For each alternative j, a so-called linear predictor η_{ij} accumulates the observable determinants of the vote decision process in a scalar quantity, formalizing the deterministic part of utility.[6] Based on the discussion from b), alternative-specific predictors are equipped with alternative-specific parameters.

[6] In fact, V_{ij} and η_{ij} denote the same quantity. Since "linear predictor" is the common denomination in the statistical literature, we prefer this term in discussions about formal model-setup.

Thus, for $i = 1, \dots, n$ and $j = 1, \dots, J$, these linear predictors take the following form:

$$\eta_{ij} = \beta_{j0} + \sum_{l=1}^{p} s_{il}\beta_{jl} + \sum_{k=1}^{K} z_{ijk}\alpha_{jk} = \beta_{j0} + \boldsymbol{s}_i^T\boldsymbol{\beta}_j + \boldsymbol{z}_{ij}^T\boldsymbol{\alpha}_j.$$

The parameters $\beta_{10}, \dots, \beta_{J0}$ denote alternative-specific constants (ASCs). Each $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_J$ is a p-dimensional coefficient vector related to the p-dimensional individual-specific covariate vector $\boldsymbol{s}_i$. The coefficient vectors $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_J$ contain the utility parameters that indicate the value attached to the K policies/issues in the alternative-specific covariate vector $\boldsymbol{z}_{ij}$.

d) Link function and model: Given all the previous assumptions and considering both individual-specific predictors and the alternative-wise specification of choice-specific variables, the multinomial logit model (MNL) in its generic form can be stated as follows (see Tutz, 2012):[7]

$$\pi_{ij} = P(Y = j|\boldsymbol{s}_i, \boldsymbol{z}_{ij}) = \frac{\exp(\eta_{ij})}{\sum\limits_{r=1}^{J}\exp(\eta_{ir})} = \frac{\exp(\beta_{j0} + \boldsymbol{s}_i^T\boldsymbol{\beta}_j + \boldsymbol{z}_{ij}^T\boldsymbol{\alpha}_j)}{\sum\limits_{r=1}^{J}\exp(\beta_{r0} + \boldsymbol{s}_i^T\boldsymbol{\beta}_r + \boldsymbol{z}_{ij}^T\boldsymbol{\alpha}_r)}, \tag{2}$$

where $Y \in \{1, ..., J\}$ denotes the j-categorical, probabilistic response variable, i.e. $Y = j$ indicates that party j is chosen.[8]

Note that (2) refers to the MNL in its generic form, which means that the parameters $\beta_{10}, ..., \beta_{J0}$ and $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J$ are not identifiable. In order to identify the model, a side constraint, such as defining a reference category or using a symmetric side constraint, has to be introduced.[9]

At this point it is important to emphasize the inherent complexity of the MNL from (2). As the following example indicates, the number of possible individual- and party-specific coefficients increases rapidly, resulting in highly complex and difficult-to-interpret models. Consider a model of party choice in a system with five major parties based on three choice-specific variables and ten chooser-specific attributes. This specification already produces 59 possible parameters (ASCs included).[10] The following section introduces the Lasso method, a parameter selection strategy that systematically and efficiently reduces

[7] Logit models assume that the random part of utility, ϵ_{ij}, follows an iid maximum extreme value distribution. See also McFadden (1973, 1984).

[8] For later use, we define J-dimensional response vectors $\boldsymbol{y}_i = (0, ..., 0, 1, 0, ..., 0)^T$ with 1 on jth position indicating the chosen alternative (i.e. $Y = j$). Additionally, let $\boldsymbol{\pi}_i$ denote the J-dimensional vector of choice probabilities. Conditional on the covariates $\boldsymbol{s}_i$ and $\boldsymbol{z}_{ij}$, $\boldsymbol{y}_i$ can be considered as independent realizations of drawing from a multinomial distribution: $\boldsymbol{y}_i \mid \boldsymbol{s}_i, \boldsymbol{z}_{ij} \sim M(1, \boldsymbol{\pi}_i),\ i = 1, \dots, n$.

[9] However, in Section 3.2.3, we will show an interdependency between the particular choice of an identifiability constraint and the Lasso method. Therefore, the discussion of identifiability is delayed to Section 3.2.3, that is, until the Lasso method has been formally introduced.

[10] If a reference category as side constraint is used, we obtain 4 ASCs + 4*10 coefficients for individual-specific variables + 3*5 parameters for party-specific variables, resulting in a total of 59 parameters. In the case of a symmetric side constraint, 70 nominal parameters result, while still obtaining only 59 degrees of freedom. See Section 3.2.3.

this high-dimensional predictor space, and thus ensures more parsimonious spatial voting models.

3 The Lasso Approach: Parameter Selection and Regularization of MNLs with alternative-specific Covariates

The previous section highlighted the inherent complexity of MNLs, and therefore the practical need for sophisticated parameter selection procedures in the statistical analysis of the Spatial Theory of Voting. In this section, we introduce and derive for the first time a Lasso-type regularization technique for the estimation of MNLs that considers both individual- and alternative-specific variables. Additionally, this new algorithm allows us to handle the alternative-wise specification of issue distances. Based on a brief review of classical subset selection procedures and their weaknesses and limitations, we outline the general idea of regularization and penalty approaches. Then, the Lasso approach is presented, including a technical discussion of its computation, its interdependency with the choice of identifiability constraint in MNLs, and how its variable selection properties can be improved.

3.1 Classical Subset Selection Techniques

How can electoral researchers select the most parsimonious model out of a large set of possible models (based on a large number of potential predictor variables) while simultaneously reducing the model's complexity? Since choosing a model that maximizes some goodness-of-fit measure (e.g., pseudo R^2 or loglikelihood) causes overfitting and low predictive accuracy, the task of variable selection is typically tackled by introducing optimality criteria that approximate a given model's expected performance on future observations. Two of the most popular optimality criteria are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).[11] Classical subset selection techniques, such as best-subset[12] or stepwise selection techniques[13], investigate the influence of the inclusion or exclusion of individual predictors on minimizing these optimality criteria. As

[11] For the specified MNL considering the alternative-wise specification of issue distances, these optimality criteria are given by

$$\mathrm{AIC}(\hat{\boldsymbol{\theta}}) = -2l(\hat{\boldsymbol{\theta}}) + 2\mathrm{df}(\hat{\boldsymbol{\theta}}); \qquad \mathrm{BIC} = -2l(\hat{\boldsymbol{\theta}}) + \log(n)\mathrm{df}(\hat{\boldsymbol{\theta}}), \tag{3}$$

where $\hat{\boldsymbol{\theta}}^T = (\hat{\beta}_{10}, \ldots, \hat{\beta}_{J0}, \hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_J, \hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_J)$ denotes the estimator of the model's overall parameter vector $\boldsymbol{\theta}$, $l(\boldsymbol{\theta})$ the loglikelihood, and $\mathrm{df}(\hat{\boldsymbol{\theta}})$ the degrees of freedom, equalling here the number of parameters.

[12] Best-subset refers to the choice of the, in terms of optimality criteria, best possible subset of variables out of all possible variable combinations.

[13] Stepwise selection approaches include forward selection, backward selection, or a combination thereof. Stepwise approaches can be seen as an attempt to approximate best-subset selection with lower computational burden. For an overview of variable selection based on subset techniques, see Hastie, Tibshirani, and Friedman (2009) and references therein.

a result, the combination of covariates yielding the smallest value of the optimality criteria is chosen.

However, these frequently used classical subset selection approaches exhibit several weaknesses. With regard to best-subset selection procedures, it is rarely possible to compute exactly the optimal set of variables due to the associated computational burden. Take, for example, a model of party choice consisting of five parties, three issues and ten chooser-specific attributes (see Section 4). Applying the best-subset selection technique to this model, 55 parameters could be set to zero[14], resulting in a total of $2^{55} \approx 3.6 \times 10^{16}$ possible models. This example demonstrates a side-effect of the flexibility of the MNL framework: Since the number of parameters is the product of the number of alternatives and predictors, it is impracticable to fit all possible models, unless the number of alternatives and predictors is extremely small.

In order to obtain a satisfying subset of important predictors in reasonable time, one typically applies stepwise approaches in which variables are added or removed from the current state/model until the optimality criteria cannot be improved any more. However, since subset selection is a discrete process and all optimality criteria have multiple local optima, these stepwise approaches suffer from considerable instability (Hastie, Tibshirani, and Friedman, 2009). Thus, starting stepwise variable selection either from a full model or starting it from an ASCs-only model can lead to completely different results. Consequently, even the slightest change in the data or in the starting point can produce significantly different outcomes of subset selection. Due to this instability, stepwise approaches cannot be recommended, but also an exhaustive all-subset search is usually impossible for the considered model class. In the next section, we present the Lasso approach. In contrast to classical subset selection techniques, the Lasso is a parameter selection method guaranteeing continuous, stable and computationally efficient variable selection.

3.2 The Lasso

In order to motivate the Lasso method that efficiently identifies relevant predictors, we briefly outline the general idea of regularization techniques, of which the Lasso is a special case. Regularization implies, inter alia, introducing penalty terms that restrict the estimated coefficients (see Tutz, 2010). Therefore, penalization refers to the formulation of side constraints on the values of the parameters which are taken into account in the estimation. Penalty approaches based on L_p-norm of the parameter vector aim to penalize the size or the length of these parameters. Their goal is to shrink the coefficients and to set, ideally, coefficients of weak predictors exactly to zero, yielding continuous, stable and computationally efficient variable selection.

In the following, we first introduce the Lasso approach, including its definition and basic properties. Second, we outline how the Lasso's variable selection properties can

[14] This number of parameters implies that ASCs always remain in the model.

be improved by using adaptive weights. Then, we demonstrate its interdependency with the choice of identifiability constraint in MNLs, followed by a technical discussion on the computation of the Lasso estimator and the choice of the tuning parameters λ.

3.2.1 Definition and Basic Properties

The Lasso (Least Absolute Shrinkage and Selection Operator), introduced by Tibshirani (1996), is a penalty approach to variable selection in regression models. For a general model with loglikelihood $l(\cdot)$ and parameter vector $\boldsymbol{\theta}$[15], penalty approaches introduce a penalty term $P(\boldsymbol{\theta})$ that is subtracted from the loglikelihood, resulting in the penalized loglikelihood $l_{\text{pen}}(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - P(\boldsymbol{\theta})$. That is, instead of maximizing the loglikelihood in the estimation process, we maximize the penalized loglikelihood. If the penalty term is chosen appropriately, the penalized parameter estimator can have superior properties compared to the unpenalized maximum likelihood (ML) estimator, such as a reduced variance or a lower dimensionality. The p-norm penalized parameter estimator is, in its most general form, defined by

$$\hat{\boldsymbol{\theta}}^{pen}(\lambda, p) = \underset{\boldsymbol{\theta}}{\text{argmax}}\ l(\boldsymbol{\theta}) - \lambda||\boldsymbol{\theta}||_p = \underset{\boldsymbol{\theta}}{\text{argmin}}\ - l(\boldsymbol{\theta}) + \lambda||\boldsymbol{\theta}||_p, \qquad p \geq 0, \tag{4}$$

where $\lambda \geq 0$ denotes the tuning parameter controlling the degree of penalization. The Lasso estimator is obtained by penalizing the L_1-norm of the parameter vector: $P_{\text{Lasso}}(\boldsymbol{\theta}) = \lambda||\boldsymbol{\theta}||_1 = \lambda \sum_i |\theta_i|$.[16]

In the following, we explicitly derive for the first time the Lasso for the MNL based on the predictor structure given in (2) and discuss its properties. The Lasso has previously been extended to MNLs by Friedman, Hastie, and Tibshirani (2010). However, their work mainly focuses on algorithms and only briefly mentions the MNL as a possible application for their algorithm. Additionally, Friedman, Hastie, and Tibshirani (2010) exclusively consider individual-specific predictors $\boldsymbol{s}_i$, whereas we explicitly derive the Lasso for a MNL containing both individual- and choice-specific predictors. Therefore, and in contrast to previous work, which only briefly mentions the MNL as a possible application and exclusively considers individual-specific predictors, we explicitly derive the Lasso for a MNL based on both individual- and alternative-specific predictors and in which the alternative-specific variables are specified as alternative-specific effects.

[15] For the specified model including issue distances with alternative-specific effects, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$, where $\boldsymbol{\beta} = (\beta_{10}, \dots, \beta_{J0}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_J)$ is the vector of all β-parameters and, accordingly, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_J)$.

[16] Penalizing the squared L_2-norm results in ridge regression (Hoerl and Kennard, 1970). Note that best subset selection based on AIC or BIC is also contained in (4) by using the L_0-pseudo-norm and particular values of λ.

Applying the general form of the Lasso to our model yields[17]

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) = \underset{\boldsymbol{\beta},\boldsymbol{\alpha}}{\operatorname{argmax}}\; l(\boldsymbol{\beta}, \boldsymbol{\alpha}) - \lambda \sum_{j=1}^{J} \left(\sum_{l=1}^{p} |\beta_{jl}| + \sum_{k=1}^{K} |\alpha_{jk}| \right). \tag{5}$$

The Lasso solutions may contain exact zeros, so that the corresponding effects are effectively removed from the model. Thus, the Lasso implicitly performs variable selection as a by-product of the estimation process. The larger λ, the stronger is the penalization, and therefore the sparser is the estimated coefficient vector. Setting $\lambda = 0$ leads to the ML estimator. A maximal value $\lambda_{\max}$ can be derived so that all penalized parameters are estimated to be zero whenever $\lambda \geq \lambda_{\max}$. Since the Lasso estimator is the maximizer of a continuous and concave objective, it is continuous itself, unique for any fixed λ[18] and less sensitive to noise in the data than the discrete best subset method. In fact, the choice $p = 1$ in (4), which defines the Lasso, is the only choice for which the L_p-norm penalized log-likelihood is concave and, at the same time, yields a sparse estimator; i.e., is able to contain exact zeros.[19] Consequently, the Lasso can perform variable selection just like subset selection, but avoids its drawbacks by being continuous in the data and computationally efficient.

More insights into the Lasso can be gained by considering the following alternative representation. Based on standard results from optimization theory (cf. Boyd and Vandenberghe, 2004), the Lasso can equivalently be defined by a constrained optimization:[20]

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) = \underset{\boldsymbol{\beta},\boldsymbol{\alpha}}{\operatorname{argmax}}\; l(\boldsymbol{\beta}, \boldsymbol{\alpha}) \qquad \text{subject to} \quad \sum_{j=1}^{J} \left(\sum_{l=1}^{p} |\beta_{jl}| + \sum_{k=1}^{K} |\alpha_{jk}| \right) \leq t. \tag{6}$$

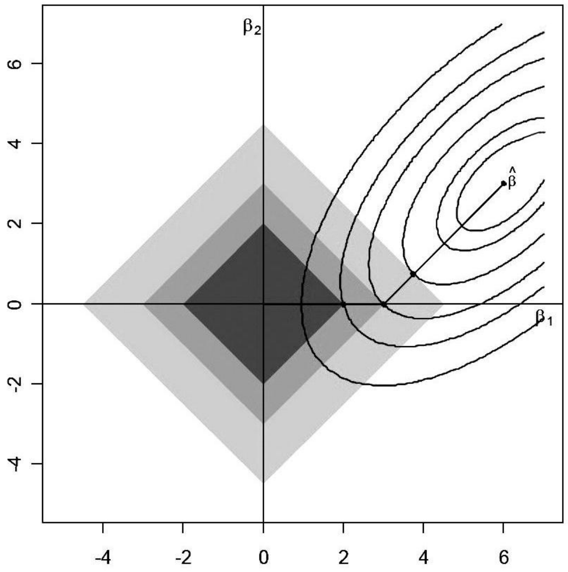
The tuning parameter $t \geq 0$ of the Lasso's constrained definition is connected to λ in its penalized form by a one-to-one mapping, but deriving a closed form for their relationship is difficult. By defining $t_{\max} = \sum_{j=1}^{J} \left(\sum_{l=1}^{p} |\hat{\beta}_{jl}^{ML}| + \sum_{k=1}^{K} |\hat{\alpha}_{jk}^{ML}| \right)$, the Lasso estimator is equal to the ML estimator for any $t \geq t_{\max}$. For $t < t_{\max}$, the constraint in (6) induces shrinkage of the parameters. As this constrained definition of the Lasso is highly useful to illustrate why the Lasso is able to shrink coefficients to exactly zero, consider the geometric illustration in Figure 1. The figure shows the contour lines of the log-likelihood function

[17] Note that if alternative J is chosen as reference to achieve identifiability of the MNL, we set $\beta_{J0} = 0$, $\boldsymbol{\beta}_J = \mathbf{0}$. Hence, summing up over all J alternatives in (5) is equivalent to penalizing only $J-1$ β-vectors that actually have to be estimated. Therefore, formula (5) is applicable to all MNLs, regardless of the chosen identifiability constraint. Also note that we do not penalize the ASCs. To keep the notation readable, we omit the dependence of Lasso estimates on a particular choice of the tuning parameter λ.

[18] The objective function is concave, and thus has a unique maximum. Assuming a sensible design, i.e. no avoidable multicollinearity, uniqueness of the parameter vector that attains this maximum is ensured if $n > J(1 + p + K)$. This is virtually always the case for applications in political and social sciences.

[19] Formally, this follows from two facts: 1.) The L_p-norm is nonconvex for $p < 1$, leading to local maxima of the penalized log-likelihood, and thus to a discrete variable selection method. 2.) For $p > 1$, the L_p-norm is differentiable at zero, and hence cannot yield solutions containing exact zeros.

[20] More specifically, equivalence between the constrained and penalized definition of the Lasso follows from the so-called Karush-Kuhn-Tucker conditions, which extend the well-known Langrange multiplier method to optimization under inequality constraints.



Note: Lasso estimates are given by the points of contact between log-likelihood contour lines and Lasso constraint regions.

Figure 1: Lasso geometry: constraint regions for the Lasso with log-likelihood contour lines for two predictors in a simple logistic regression model.

and the Lasso constraint regions for different values of t in a two-dimensional predictor space; i.e., based on a simple logistic regression model with two predictors and without an intercept. The Lasso solutions must necessarily lie at the contact points of the contour lines with the constraint regions, indicated by the black line. The ML estimator is at the center of the log-likelihood contour lines. As the inspection of Figure 1 demonstrates, the diamond-shape of the Lasso constraint causes the corresponding Lasso solutions to move in a straight line towards the axis when t is successively reduced, eventually shrinking $\hat{\beta}_2$ to zero.

The tendency of the Lasso to produce sparse solutions can also be demonstrated by considering the special case of a linear model with orthonormal design: Let $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. By assuming that $\boldsymbol{y}$ and all predictors $\boldsymbol{x}_l$ $(l = 1, \ldots, p)$ are mean-centered, no intercept is used, and that $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$, the Lasso has an analytical solution: $\hat{\beta}_l^{Lasso} = \text{sign}(\hat{\beta}_l^{ML}) \cdot \max\left(|\hat{\beta}_l^{ML}| - \lambda, 0\right), \; l = 1, \ldots, p$. Hence, the lasso solution for each predictor in this particular setting is obtained by shrinking the respective ML estimator towards zero by a value of λ. Since the MNL is a nonlinear model, even an orthonormal design matrix does not allow us to derive a similar result (or any analytical solution), but the functioning of the Lasso in MNLs is very similar.

Additionally to yielding sparse solutions and due to the fact that the Lasso is a shrinkage estimator, the Lasso estimates have reduced size and variability. These aspects, as well as the computation of the effective degrees of freedom are discussed and summarized in Appendix A.

3.2.2 Improving the Lasso using adaptive Weights

In an $n \to \infty$ setting (i.e., with arbitrarily large sample size), a variable selection method should guarantee the selection of the correct model, that is to assign nonzero estimates to truly nonzero effects and to set the coefficients of irrelevant predictors to zero. In addition, it is desirable that the method asymptotically performs as well as the ML estimator, applied to the correct set of variables. If an estimator possesses these two properties, it is said to be as good as an "oracle" that knows the correct set of variables ahead of time. As shown by Zou (2006), the ordinary Lasso from (5) does not possess this oracle property because it applies the same degree of penalization / shrinkage to all parameters, regardless of their specific size. Choosing λ as large as it is necessary to remove all irrelevant predictors implies too much bias on the estimated coefficients of the selected variables. How can we avoid that the same degree of penalization is applied to all parameters, regardless of their specific size, and therefore improve the Lasso solutions? The remedy proposed by Zou (2006) is the so-called adaptive Lasso, which is defined as follows:

$$\begin{aligned} (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) &= \underset{\boldsymbol{\beta},\boldsymbol{\alpha}}{\operatorname{argmax}}\; l(\boldsymbol{\beta}, \boldsymbol{\alpha}) - \lambda \sum_{j=1}^{J} \left(\sum_{l=1}^{p} w_{jl}^{b}\, |\beta_{jl}| + \sum_{k=1}^{K} w_{jk}^{a}\, |\alpha_{jk}| \right), \\ w_{jl}^{b} &= \frac{1}{|\hat{\beta}_{jl}^{ML}|}, \quad w_{jk}^{a} = \frac{1}{|\hat{\alpha}_{jk}^{ML}|}, \qquad l = 1, \ldots, p;\; k = 1, \ldots, K;\; j = 1, \ldots, J. \end{aligned} \tag{7}$$

As formula (7) shows, the adaptive Lasso incorporates an individual weight for the penalty of each coefficient. Note that each weight (w_{jl}^{b} and w_{jk}^{a}) consists of the inverse of the size of the corresponding ML estimate. By using these weights, the adaptive Lasso is able to adapt more adequately to the varying importance of different variables for different alternatives / parties. Since the ML estimator is consistent, the ML estimates of irrelevant effects asymptotically turn to zero, yielding an infinite (or at least very large) penalization. In the case of nonzero effects, the adaptive weights converge to a finite, constant value. In general, the larger the ML estimate of a particular parameter, the less penalization is applied to this parameter by the adaptive Lasso, and vice versa. This logic also holds in non-asymptotic settings as long as the ML estimator is stable. Indeed, our experience has shown that the use of adaptive weights provides a considerable improvement of the Lasso's performance, both in terms of variable selection and prediction accuracy – even for datasets with a moderate number of observations. For these reasons and even though the adaptive Lasso was originally motivated by theoretical, asymptotic considerations, we strongly recommend its usage in virtually all practical applications.

3.2.3 Lasso and its Interdependency with the Choice of Identifiability Constraint

As we noted in Section 2, not all individual-specific variables (and ASCs) in (2), giving the MNL in its generic form, are identifiable. In order to identify the model, a side constraint has to be introduced. In this section we point out how the choice of the identifiability constraint in MNLs and the result of the Lasso interdepend. In particular, we show that the use of a symmetric identifiability constraint prevents that the arbitrary choice of a reference category influences the results of the Lasso method.

The non-identifiability of all individual-specific covariates in MNLs can be expressed in the following way: for $l = 0, 1, \dots, p$, let $\delta_l \in \mathbb{R}$ denote a shift which is applied to all coefficients belonging to the same individual-specific predictor, so that the new coefficients $\tilde{\beta}_{jl} = \beta_{jl} + \delta_l$ result. By gathering this shift across all p individual-specific predictors into a shift vector $\boldsymbol{\delta}^T = (\delta_1, \dots, \delta_p)$, the model based on $\tilde{\beta}$-parameters, for $j = 1, \dots, J$, can be expressed as follows:[21]

$$\begin{aligned} P(Y = j|\boldsymbol{s}, \boldsymbol{z}) &= \frac{\exp(\tilde{\beta}_{j0} + \boldsymbol{s}^T\tilde{\boldsymbol{\beta}}_j + \boldsymbol{z}_j^T\boldsymbol{\alpha}_j)}{\sum\limits_{r=1}^{J} \exp(\tilde{\beta}_{r0} + \boldsymbol{s}^T\tilde{\boldsymbol{\beta}}_r + \boldsymbol{z}_r^T\boldsymbol{\alpha}_r)} \\ &= \frac{\exp(\beta_{j0} + \delta_0 + \boldsymbol{s}^T(\boldsymbol{\beta}_j + \boldsymbol{\delta}) + \boldsymbol{z}_j^T\boldsymbol{\alpha}_j)}{\sum\limits_{r=1}^{J} \exp(\beta_{r0} + \delta_0 + \boldsymbol{s}^T(\boldsymbol{\beta}_r + \boldsymbol{\delta}) + \boldsymbol{z}_r^T\boldsymbol{\alpha}_r)} \\ &= \frac{\exp(\delta_0 + \boldsymbol{s}^T\boldsymbol{\delta}) \cdot \exp(\beta_{j0} + \boldsymbol{s}^T\boldsymbol{\beta}_j + \boldsymbol{z}_j^T\boldsymbol{\alpha}_j)}{\exp(\delta_0 + \boldsymbol{s}^T\boldsymbol{\delta}) \cdot \sum\limits_{r=1}^{J} \exp(\beta_{r0} + \boldsymbol{s}^T\boldsymbol{\beta}_r + \boldsymbol{z}_r^T\boldsymbol{\alpha}_r)}. \end{aligned} \tag{8}$$

Equation (8) shows that the absolute level of the β-parameters is not relevant for the MNL, only the differences $\beta_{jl} - \beta_{ql},\ j \neq q,\ l = 0, 1, \dots, p$ are meaningful. Since the Lasso penalty is based on this absolute value of the coefficients, the choice of the identifiability constraint in MNLs influences the results of the Lasso method, and therefore the selection of the most promising model.

In general, scholars choose as identifiability constraint a particular reference category. By using $J = 1$ as reference/baseline category, β_{10} and $\boldsymbol{\beta}_1$ are set to zero. This ensures identifiability, and the effects of individual-specific covariates are interpreted relative to the first alternative/party. In order to highlight the interplay between the choice of a particular reference category as identifiability constraint and the Lasso, consider an unpenalized model with five parties, in which the first party is chosen as reference. Assume that the ML-coefficients of variable s_l for this model are given by $\beta_l = (0, 3, 0.2, -0.5, 2.5)^T$, so that $||\beta_l||_1 = 6.2$. By defining the second party as reference instead, it follows from (8) that β_l changes to $(-3, 0, -2.8, -3.5, 0.5)$, and thus $||\beta_l||_1 = 9.8$. As this simple example demonstrates, the arbitrary change of the reference category causes to increase both the

[21] Note that this kind of parameter shifting does not change the model.

overall L_1-norm and the number of coefficients that are relatively far away from zero. Consequently, the sparsity of the Lasso solutions also decreases – at least for some values of λ. Using real data, we illustrate in Section 4.3 that the sparsity and predictive performance of the Lasso indeed vary considerably across different reference categories.

Hence, how can we prevent that the arbitrary choice of a reference category influences the results of the Lasso method? We propose to use a symmetric identifiability constraint, which results from imposing that

$$\sum_{j=1}^{J} \beta_{jl} = 0, \qquad l = 0, 1, \ldots, p. \tag{9}$$

Using this symmetric side constraint yields J ASCs and $J{*}p$ parameters in the case of individual-specific predictors. However, only $J-1$ and $(J-1){*}p$ of these parameters can be estimated freely. As Zou, Hastie, and Tibshirani (2007) showed, the effective degrees of freedom of the Lasso estimator are generically given by the number of nonzero parameters. By combining this general template with the symmetric side constraint (i.e., for the estimator from (5) coupled with (9)), the following formula for the Lasso's effective degrees of freedom results:[22]

$$\hat{\mathrm{df}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) = \sum_{l=0}^{p} \max\left(\left(\sum_{j=1}^{J} I(|\hat{\beta}_{jl}| > 0)\right) - 1,\ 0\right) + \sum_{k=1}^{K}\sum_{j=1}^{J} I(|\hat{\alpha}_{jk}| > 0). \tag{10}$$

Finally, in the next section we discuss the computation of the Lasso estimator and the choice of the tuning parameter λ.

3.2.4 Estimation and Choice of Tuning Parameters

In order to compute the Lasso estimator for any given λ, we use the Fast Iterative Shrinkage Thresholding Algorithm (FISTA) introduced by Beck and Teboulle (2009). After adjusting the formulas for the loglikelihood and its derivatives, the FISTA algorithm can be directly applied to our Lasso specification ((5) or (7)). For the specified MNL including issue distances with alternative-specific effects, the loglikelihood is computed as

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij} \log\left(\pi_{ij}(\boldsymbol{\beta}, \boldsymbol{\alpha})\right), \tag{11}$$

where π_{ij} is given in formula (2) and is written here as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. For $j = 1, \ldots, J$, the derivatives of the loglikelihood with respect to the parameters are

[22] $I(\cdot)$ denotes an indicator function.

obtained as follows:

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \beta_{jl}} &= \sum_{i=1}^{n} s_{il}(y_{ij}-\pi_{ij}) = \boldsymbol{s}_l^T(\boldsymbol{y}_j-\boldsymbol{\pi}_j), \quad l=0,1,\ldots,p,\\ \frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \alpha_{jk}} &= \sum_{i=1}^{n} z_{ijk}(y_{ij}-\pi_{ij}) = \boldsymbol{z}_{jk}^T(\boldsymbol{y}_j-\boldsymbol{\pi}_j), \quad k=1,\ldots,K,\end{aligned} \tag{12}$$

where $\boldsymbol{s}_l^T = (s_{1l},\ldots,s_{nl})$, $\boldsymbol{z}_{jk}^T = (z_{1jk},\ldots,z_{njk})$, $\boldsymbol{y}_j^T = (y_{1j},\ldots,y_{nj})$, and $\boldsymbol{\pi}_j^T = (\pi_{1j},\ldots,\pi_{nj})$ denote vectors pooling the corresponding quantities across n observations. To sum up, we offer the Lasso specification as well as the computation of the loglikelihood and its derivatives for a MNL including alternative-specific variables with alternative-specific effects.

Since the Lasso penalty is not invariant to the scale and the variance of the covariate vectors $\boldsymbol{s}_l$ and $\boldsymbol{z}_{jk}$, we ensure that all variables have the identical chance of being selected by standardizing these to zero mean and unit variance before applying the Lasso. This leads to Lasso-penalized regression coefficients for the standardized covariates from which the corresponding coefficients belonging to the original covariates can easily be reconstructed.

In the following, we outline the choice of the tuning parameter λ, the parameter controlling the degree of penalization. To choose an appropriate λ, the Lasso solutions are computed over a grid of different λ-values. By m denoting the number of possible λs, this grid is given by a sequence $\lambda_1 > \lambda_2 > \ldots > \lambda_m$. For practical reasons, the lower endpoint λ_m of this grid usually consists of a very small, but positive value (e.g., $\lambda_m = 0.01$.) The corresponding upper endpoint is chosen as $\lambda_1 = \lambda_{\max}$, which is the smallest value of the tuning parameter for which all penalized coefficients are set to zero. Using standard results from convex optimization theory, it is easy to show that $\lambda_{\max}$ can be computed as

$$\lambda_{\max} = \max\left(\max_{j,l}\left|\boldsymbol{s}_l^T\left(\boldsymbol{y}_j-\hat{\boldsymbol{\pi}}_j^0\right)\right|,\ \max_{j,k}\left|\boldsymbol{z}_{jk}^T\left(\boldsymbol{y}_j-\hat{\boldsymbol{\pi}}_j^0\right)\right|\right),$$

where $\hat{\boldsymbol{\pi}}_j^0$ presents the estimated choice probabilities of a model containing only unpenalized predictors. Note that in our case this is an ASCs-only model. In order to select a concrete λ and thus a concrete estimator / model, the Lasso solutions for different values of the tuning parameter are evaluated by model selection criteria. Usually, these optimality criteria are either crossvalidation (CV), AIC or BIC (which can be computed applying (10)). According to our experience, the selection of λ based on BIC leads to the most sparse and CV to the least sparse model. Since the AIC presents a desirable compromise, we strongly prefer to use this optimality criterion for the choice of λ.

Here it is important to note that variable selection using the Lasso solely requires a one-dimensional search over a grid of λ-values, whereas the complexity of variable selection based on best subset techniques or test-based procedures increases exponentially with the number of covariates and alternatives.

4 Application: Regularized Analysis of Party Choice in the 2009 German Parliamentary Election

This section illustrates the advantages of the Lasso-type regularization method in the analysis of multiparty competition using individual survey data from the 2009 German Parliamentary Election.[23] Party choice is operationalized via the stated vote intention. The set of alternatives includes parties having a vote share total of at least 5% in the 2009 election. The respective parties are: the Christian-Democratic Party (CDU), the Social-Democratic Party (SPD), the Liberal Party (FPD), the Green Party (Greens), and the Leftist Party (Leftists). The MNL of party choice is based on the following explanatory variables: As alternative-specific predictors we include three issues, measured as the absolute distance between the individually perceived party positions and the self-reported positions of voters (ideal points) on 11-point scales.[24] In order to fully exploit the Lasso's potential to effectively reduce the predictor space, we consider 10 individual-specific covariates, including sex, age, West/East Germany, union membership, high school degree, unemployment, political interest, satisfaction with democracy, and religious denomination.[25]

The presentation of the regularized analysis is divided into three parts: We begin by fitting a Lasso-penalized MNL based on solely main effects. In the second part we additionally include interaction effects. Finally, we provide a systematic comparison of the penalized and unpenalized models.

4.1 Main Effects Model

As a first step of our analysis, we estimated a Lasso-penalized MNL of party choice containing the main effects of all variables introduced above. As we argued in Section 3.2.3, we applied a symmetric side constraint to ensure identifiability of the model and to avoid that the arbitrary choice of a reference category influences the results of the Lasso

[23] German Longitudinal Election Study (GLES) 2009, Pre-election Cross-Section. The data set is available under http://www.gesis.org/wahlen/gles/daten-und-dokumente/daten/. Identification Number: ZA5300, Version 5.0.0.

[24] These issues are: Taxes: "1" = Lower taxes, even if that means less government spending on health, education and social benefits, "11" = More government spending on health, education and social benefits, even if that means higher taxes; Immigration: "1" = Laws on immigration should be relaxed, "11" = Laws on immigration should be tougher; Nuclear Energy: "1" = More nuclear power stations, "11" = Close down all nuclear power stations immediately.

[25] These variables are coded as follows: sex: 1 (male), 0 (female); age: centered around the sample mean of 50.5 years, measured in decades; West/East Germany: 1 (former West Germany), 0 (former East Germany); union membership: 1 (union members) , 0 (otherwise); high school degree: 1 (yes), 0 (no); unemployment: 1 (currently unemployed), 0 (otherwise); political interest: 1 (less interested), 0 (very interested); satisfaction with democracy: 1 (not satisfied), 0 (satisfied); religious denomination: (Protestant, Roman-Catholic, otherwise).

method.[26] We fit this model over a grid of 100 λ-values and chose the best one according to the AIC-criterion. Following the discussion from Section 3.2.2, adaptive weights are included in the Lasso penalty. Since there exists no analytical formula for standard errors of Lasso estimates, we computed them via bootstrap.[27] Parameter estimates as well as approximate p-values (based on the bootstrapped standard errors) of this main effects model are summarized in Table 1.[28] As can be seen immediately from Table 1, a considerable amount of effects is set to zero by the Lasso, yielding an enormous reduction of the complexity of the estimated model. In particular, 32 out of 70 nominal parameters are set to zero, and thus only 38 coefficients are selected; i.e., remain in the model.[29] Note that all issue effects are selected by the Lasso and most of them are also highly significant.[30] On the contrary, the variables sex, unemployment status and having a religious affiliation other than Protestant or Catholic are entirely removed from the model. In addition, most of these individual-specific effects have the expected sign. Notice, e.g., the strength of the Leftist Party among voters in former East Germany, or the highly significant age effect on the vote for the Christian-Democratic Party CDU and the Greens. As to be expected, the Leftist Party still finds its main support in former East Germany, and the CDU is strong

	CDU		**SPD**		**FDP**		**Greens**		**Leftist**	
ASC	1.08***	(0.000)	-0.11	(0.666)	-0.39	(0.093)	-0.14	(0.635)	-0.44	(0.154)
Sex	0		0		0		0		0	
West Germany	-0.10	(0.506)	0.42*	(0.035)	0		0.22	(0.307)	-0.53**	(0.005)
Age	0.20***	(0.000)	0		0		-0.20***	(0.000)	0	
Union Membership	-0.62	(0.067)	0.03	(0.882)	0		0		0.59*	(0.020)
High School	0		0		0		0.31	(0.164)	-0.31	(0.179)
Unemployment	0		0		0		0		0	
Pol. Interest	0		0.32*	(0.042)	0		0		-0.32	(0.066)
Democracy	-0.73***	(0.000)	-0.21	(0.207)	0		0		0.94***	(0.000)
ReligionCatholic	0.16	(0.204)	-0.16	(0.192)	0		0		0	
ReligionOther	0		0		0		0		0	
Taxes	0.17**	(0.006)	0.24***	(0.000)	0.20**	(0.004)	0.24**	(0.002)	0.33***	(0.000)
Immigration	0.26***	(0.000)	0.07	(0.198)	0.10	(0.147)	0.20***	(0.000)	0.16**	(0.002)
Nuclear Energy	0.17***	(0.000)	0.30***	(0.000)	0.29***	(0.000)	0.27***	(0.000)	0.11	(0.073)

Numbers in parentheses show approximate p-values based on simple, two-sided t-tests using bootstrapped standard errors.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1: Lasso-regularized coefficient estimates of the main effects model, German Parliamentary Election 2009.

[26] More details on how the Lasso's sparsity and predictive performance vary across different reference categories will be given below in Section 4.3.

[27] Our bootstrapped standard error estimates are based on $B = 500$ bootstrap samples.

[28] Note that the p-values of coefficients which are set to zero by the Lasso are necessarily 1. Therefore, we do not display these p-values.

[29] These 70 nominal parameters result from the symmetric side constraint, but still only 59 degrees of freedom are obtained.

[30] Also note that all issue effects are positive. Therefore, since the issue distance covariates z_{ijk} are defined as $z_{ijk} := -|x_{ik} - p_{ijk}|$, a larger difference between the individually perceived party positions and the voter's ideal points on these issues indicates that this party is less likely to be chosen by this voter. That is, if the perceived distance increases by one unit, the corresponding issue variable decreases by one unit due to the negative sign.

among older voters, whereas the Greens seem to be attractive among younger voters (see, e.g., Jun, 2011; Elff and Rossteutscher, 2011). A series of further interesting effects have been selected by the Lasso, including the strength of both the Leftist Party and the Social-Democratic Party SPD among voters being members of a labor union, or the positive effect of having a Catholic affiliation on supporting the CDU. These two findings are particularly relevant for researchers focusing on the impact of traditional cleavages on party choice. In line with recent work (see, e.g., Elff, 2009, 2007), these selected effects support their claim that social and class divisions continue to be relevant determinates of electoral behavior in Germany. However, note that only the selected effect of the variable union membership on the Leftist vote proves to be statistically significant, whereas the one on the SPD vote is not. This result may indicate that the SPD has become less effective in attracting and mobilizing its traditional target and that the class cleavage is more effectively represented by the Leftist Party. The reason why the social cleavage seems to find less expression in the SPD can be seen as the result of unpopular labor market reforms (Hartz IV) initiated by Chancellor Schröder in 1998 (see, e.g., Elff and Rossteutscher, 2011; Anderson and Hecht, 2012; Lees, 2012). Also note that our results suggest that the persistent relevance of traditional cleavages does not as strongly apply to religious versus secular divisions. Although the Lasso selected a positive effect of having a Catholic affiliation on supporting the CDU, and a negative effect on the SPD vote, they do not prove to be significant, indicating an increasing secularisation of the German society.[31] Since Christian affiliation constituted for a long time a major cleavage in the German party system, this is an important result. Finally, also note the interesting and highly significant effect of voter's satisfaction with democracy. Voters who are unsatisfied with the democracy in Germany are much less likely to vote for the CDU, while they are much more likely to vote for the Leftists. Thus, out of the five parties considered here, the Leftist Party seems to be the strongest attractor for "protest voters".

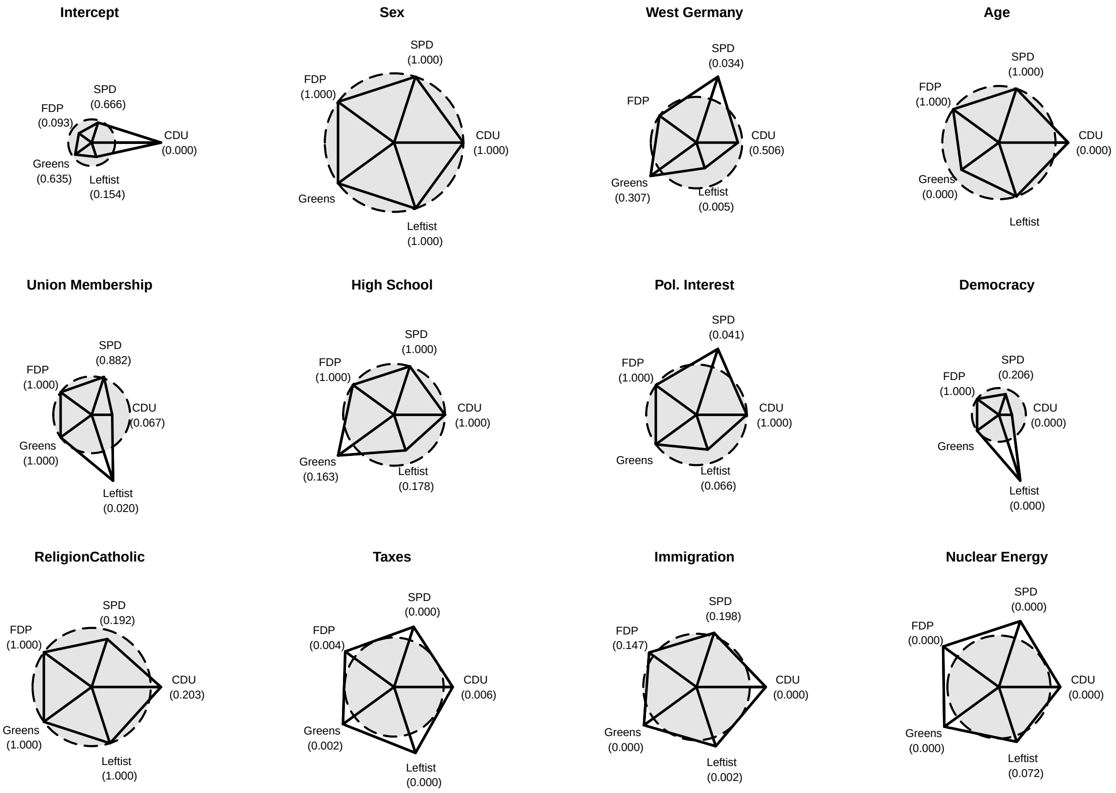
Also note the following two points emerging from the results in Table 1. Even though the coefficients of the covariate political interest have the same absolute value for both the Social-Democratic Party SPD and the Leftist Party (due to the symmetric side constraint), the effect on the SPD vote is significant while the effect on the Leftists vote is not. This is no contradiction because the variance of the effect of a dummy-coded binary predictor on a particular party depends on the percentage of observations having chosen this party, and the corresponding predictor simultaneously took the value coded with 1. Consequently, these two estimated effects can differ in variance. In general, it should be noted that the p-values and significance levels shown in Table 1 are based on simple t-tests, and thus ignore the correlation between predictors. Therefore, a selected parameter not gaining significance according to this test should not cause concern.

The coefficients from Table 1 can be interpreted as usual. For example, for voters from former West Germany relative to those living in former East Germany, the odds of voting

[31] However, note here that we operationalize the religious-secular cleavage by religious denomination without additionally considering the frequency of church attendance.

for the CDU relative to voting for the SPD change by a factor of $\exp(-0.10 - 0.42) = 0.594$, if everything else remains fixed. Thus, the odds of voting for CDU instead of SPD are roughly 40% lower in former West Germany, ceteris paribus. In contrast to the individual-specific variables and as the following example demonstrates, the alternative-specific issue distance coefficients have to be interpreted slightly differently. If the perceived distance between a voter and the Green Party on the issue of immigration increases by x units, the odds of voting for the Green Party relative to voting for any other party change by a factor of $\exp(-0.2 \cdot x) = 0.81^x$, given the other variables in the model are held constant.[32] Hence, ceteris paribus, a one-unit increase in the distance on the issue of immigration decreases the odds of voting for the Greens by 19%, a two-unit increase decreases the odds by 33% etc.[33]

To summarize the main effects model from Table 1 in a concise and easy-to-read fashion, we suggest the use of so-called effect stars (Tutz and Schauberger, 2013). Figure 2 depicts these effect stars for each covariate. Each ray of a particular effect star corresponds to one party and has length proportional to $\exp(\hat{\beta}_{jl})$ or $\exp(\hat{\alpha}_{jk})$, respectively. The dashed



Note: The effect stars are based on the main effects model in Table 1.

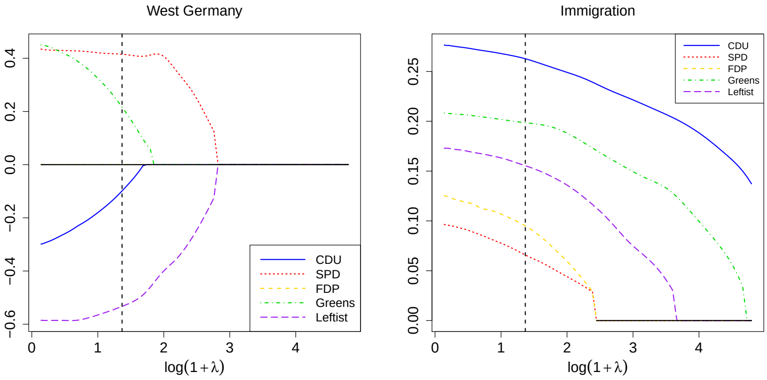
Figure 2: Effect stars, German Parliamentary Election 2009.

[32] Note again that due to the negative sign of the issue distances, an increase in the perceived distance corresponds to a decrease in the issue variable.

[33] $\exp(-0.2 \cdot 2) = 0.67$; $1 - 0.67 = 0.33 = 33\%$.

circles correspond to a null effect, and therefore indicate a length of $\exp(0) = 1$. If the ray of an effect star lies outside of this circle, the corresponding variable exhibits a positive effect on the respective party; if the ray lies inside, the effect is negative. The numbers in parentheses show approximate p-values.[34] Since the overall size of all effect stars is the same, small circles correspond to predictors whose maximal effect is large, whereas large circles indicate a small maximal effect. For instance, visual inspection shows that the covariate sex does not affect party choice, all rays of its effect star lie on the null circle. Also note the striking large effect of dissatisfaction with democracy on the Leftist Party, or the strong negative effect of union membership on the CDU vote.[35]

Next, let us turn to the visualization of the Lasso's shrinkage property as outlined in Section 3.2.1. This shrinkage effect can be illustrated by so-called coefficient paths. Figure 3 depicts these coefficient paths for the variable West Germany and the issue of immigration. Each path indicates the Lasso estimates over the chosen grid of the tuning parameter λ. In particular, the paths illustrate how the estimated coefficients move towards zero when the penalization is increased. Therefore, they show for which specific party these two variables have the most persistent effect. With regard to the covariate West Germany, these parties are the SPD and the Leftist Party: Only by applying a strong penalization, the corresponding coefficients turn to zero. The horizontal black line indicates a zero effect, the dashed vertical line visualizes the λ chosen via AIC. Hence, the height of the intersection between coefficient paths and the dashed line is equal to the estimators from Table 1. Take, for instance, the path of the SPD for the variable West Germany. This path intersects the dashed line at about .4, visualizing the Lasso-regularized coefficient estimate for the SPD vote. In order to facilitate the readability of the coefficient paths, we use a log-scale for the x-axis. By applying the transformation



Note: The coefficient paths are based on the main effects model in Table 1.

Figure 3: Coefficient paths for the variables West Germany and the issue immigration, German Parliamentary Election 2009.

[34] Note again that the p-values of coefficients which are set to zero by the Lasso are necessarily 1. Therefore, we do not display these p-values. For instance, see the effect of sex on the Greens vote.

[35] Due to space restrictions, we do not display the irrelevant variables unemployment and having a religious affiliation other than Protestant or Catholic in Figure 2.

"log$(1+\lambda)$", the paths begin at $\lambda = 0$, which corresponds to the ML estimator. Note that the left panel of Figure 3 does not contain a yellow path representing the Liberal Party (FPD). This means that the effect of the variable West Germany on the FDP vote was removed by the Lasso even for $\lambda = 0.01$, reflecting the smallest λ in our grid. The complete presentation of coefficient paths can be found in Figure 4 in Appendix B.

4.2 Including Interactions between Issues and Voter Attributes

Next, we consider a model allowing for interaction effects between issues and voter attributes to test segment-specific reactions to issue distances. In particular, by using this much more complex specification we are able to determine the issue effects for specific voter demographics. Note that each issue distance parameter of the main effects model estimated in the previous section shows the effect of an issue on a party for **all** kinds of voters; i.e., the marginal issue effect across the whole population. Although the alternative-specific specification of the issue variables already allows us to detect which particular issue proves to affect which particular party choice, the main effects model does not allow us to infer which specific voter segments place a differential emphasis on the issues when casting their ballots. By segmenting the population into subgroups, we are also able to identify so-called issue publics (Converse, 1964). According to the issue public hypothesis, the population can be divided into issue publics, each consisting of voters who intensively care about particular issues. Instead of assuming that voters are homogenously sensitive to the whole spectrum of issues, the issue public hypothesis suggests that specific voter segments can be distinguished by their differing sensitivities towards issues based on, e.g., their personal interests or demographic characteristics (see, e.g., Krosnick, 1990; Thurner, 2000; Mebane, Jackson, and Wall, 2014).

While the use of interactions is attractive from both a theoretical and practical perspective, it massively increases the model's complexity and aggravates its interpretation. For instance, our application case consists of three issue variables, ten voter attributes and five parties, resulting in a total of 150 possible interaction terms. This high-dimensional interaction model cannot be properly handled by unpenalized ML estimation. In particular, fitting a model of this size without penalization yields highly unstable estimates with poor predictive performance.[36] Thus, the main challenge is once again continuous, stable and computationally efficient variable selection. In addition, variable selection should ensure that only those interaction effects are kept in the model that yield an improvement over the more simple main effects model from Section 4.1.

Using the same methodology and steps as in the previous section, we fit a Lasso-penalized model that additionally includes all possible interactions between the issues and voter characteristics. Table 2 contains the Lasso-penalized parameter estimates of this interaction model.

The shrinkage and regularization effect of the Lasso stabilizes this extremely large

[36] See Section 4.3 for a more detailed discussion on estimate stability and prediction accuracy.

	CDU		**SPD**		**FDP**		**Greens**		**Leftist**	
ASC	1.26***	(0.000)	0.34	(0.187)	-1.12***	(0.000)	0.29	(0.338)	-0.77*	(0.017)
Sex	0		0		0		0		0	
West Germany	-0.19	(0.296)	0.19	(0.313)	0		0		0	
Age	0		0		0		0		0	
Union Membership	-0.56	(0.140)	0		0		-0.29	(0.428)	0.84*	(0.023)
High School	0		-0.04	(0.831)	0		0		0.04	(0.795)
Unemployment	0		0		-0.43	(0.215)	0		0.43	(0.208)
Pol. Interest	0		0		0		0		0	
Democracy	-0.50	(0.114)	0		0.04	(0.874)	0		0.46	(0.163)
ReligionCatholic	-0.68*	(0.022)	0		0.68*	(0.026)	0		0	
ReligionOther	-0.80*	(0.017)	0		0.80*	(0.017)	0		0	
Taxes	0.18	(0.089)	0.38**	(0.002)	0		0.36*	(0.038)	0.38*	(0.023)
Immigration	0.28	(0.051)	0		0		0.31**	(0.001)	0.07	(0.419)
Nuclear Energy	0.28**	(0.004)	0.34**	(0.005)	0.21*	(0.011)	0.14	(0.102)	0	
Taxes X Sex	0		0		0.08	(0.331)	0		0	
Taxes X West Germany	0		0		0		0		0.18	(0.099)
Taxes X Age	0		0		0		0.04	(0.223)	-0.05	(0.129)
Taxes X Union Membership	-0.12	(0.375)	0		0		0		0	
Taxes X High School	0		0		0		0		0	
Taxes X Unemployment	0.24	(0.160)	0		-0.52	(0.051)	0		0	
Taxes X Pol. Interest	0		0		0		0.09	(0.418)	0	
Taxes X Democracy	0		0		0.12	(0.150)	0		-0.18	(0.266)
Taxes X ReligionCatholic	0		-0.06	(0.544)	0.32*	(0.020)	-0.13	(0.396)	0	
Taxes X ReligionOther	0		-0.25	(0.070)	0.11	(0.282)	-0.23	(0.142)	0	
Immigration X Sex	0		0.04	(0.443)	-0.06	(0.364)	0		0	
Immigration X West Germany	0.06	(0.467)	0		0		-0.13	(0.139)	0	
Immigration X Age	0		0		0		0.06*	(0.026)	0	
Immigration X Union Membership	0		0		0		0.06	(0.665)	0	
Immigration X High School	0		0		0		0		0.09	(0.346)
Immigration X Unemployment	0		0		0		0		0	
Immigration X Pol. Interest	0.14	(0.152)	0		0		0		0.14*	(0.048)
Immigration X Democracy	0		0		0		0		0	
Immigration X ReligionCatholic	-0.32*	(0.011)	0		0.14	(0.183)	0		0.05	(0.386)
Immigration X ReligionOther	-0.14	(0.279)	0.21*	(0.016)	0.15	(0.164)	0		-0.05	(0.359)
Nuclear Energy X Sex	0		0		0		0		-0.12	(0.122)
Nuclear Energy X West Germany	0		0		0		0		0.16	(0.135)
Nuclear Energy X Age	-0.04	(0.061)	-0.02	(0.449)	0		0		0.01	(0.545)
Nuclear Energy X Union Membership	0		0		0.48	(0.111)	0		0	
Nuclear Energy X High School	0		0		0		0		0	
Nuclear Energy X Unemployment	0		0		1.18***	(0.000)	0		0	
Nuclear Energy X Pol. Interest	-0.12	(0.126)	-0.09	(0.405)	0		0		0	
Nuclear Energy X Democracy	0.12	(0.102)	0		0		0		0.07	(0.375)
Nuclear Energy X ReligionCatholic	-0.09	(0.318)	0		0		0.16	(0.177)	0	
Nuclear Energy X ReligionOther	-0.08	(0.342)	0		0		0.13	(0.205)	0	

Numbers in parentheses show approximate p-values based on simple, two-sided t-tests using bootstrapped standard errors.

* p < 0.05, ** p < 0.01, *** p < 0.001

Table 2: Lasso-regularized coefficient estimates of the interaction model, German Parliamentary Election 2009.

model and greatly reduces the number of parameters that have to be interpreted. Therefore, the Lasso overcomes the two major drawbacks of this interaction model – instability and exuberant complexity. In particular, 107 out of 150 possible interaction terms are efficiently removed from the model, and thus only 43 relevant interactions remain. Indeed, several interesting, large and highly significant interaction terms have been selected by the Lasso, including the interaction effect between unemployment and the issue of nuclear energy on FDP vote, the interaction effect between political interest and the issue of immigration on Leftist vote, the interaction effect between religion and the issue of immigration on both CDU and SPD vote, as well as the interaction effect between being Catholic and the issue of taxes on FDP vote. For instance, the issue of immigration seems to influence the vote decision in favor of the Greens only in the case of older voter segments, whereas for young voters this particular issue plays no central role.

Also note that the tax issue exhibits a zero main effect for the Liberal Party FDP. Since the FDP is known to be very successful in promoting the reduction of taxes and efficiently politicizing this issue, this result may seem surprising at first glance. However, when interpreting this finding and the particular meaning of the estimates in interaction models, we have to take into account that main and interaction effects cannot be interpreted separately.[37] Consider a voter for whom all individual-specific variables take the value of zero and let us call a voter with these attributes the "reference voter". In the present application, this reference voter is defined by the following characteristics: 50.5 years old, female, Protestant, based in former East Germany, no union member, no high school degree, currently not unemployed, very interested in politics, and satisfied with the democracy. With z_1 denoting the tax issue, z_2 the immigration issue, and z_3 nuclear energy issue, it follows from (2) that the linear predictor for this particular reference voter is given by $\hat{\eta}_{ij} = \hat{\beta}_{0j} + z_{ij1}\hat{\alpha}_{j1} + z_{ij2}\hat{\alpha}_{j2} + z_{ij3}\hat{\alpha}_{j3}$. Therefore, the estimated zero main effect of the issue taxes on the FDP vote indicates that the issue of taxes does not influence FPD voters having this specific combination of demographic characteristics. If we take instead a Catholic voter being otherwise identical to the reference voter, the linear predictor of the FDP ($j = 3$) increases by $0.68 + z_{i31} \cdot 0.32 + z_{i32} \cdot 0.14$. To give a second example, each unit-increase in the distance between the reference voter and the CDU on the issue of nuclear energy decreases the linear predictor of the CDU by 0.28. If the voter is 70.5 years old instead, it only decreases by 0.20.[38]

4.3 Comparison of Models

In this section, we demonstrate how the Lasso outperforms the ML estimator with regard to all considered performance measures. For this purpose, we compare the Lasso-penalized main effects and interaction models as well as the corresponding unpenalized models. Additionally, we show how the sparsity and predictive performance of the Lasso vary across different reference categories.

The models' predictive performance is measured by the cross-validated deviance (CV), the AIC, and the BIC. The models' complexity is determined by the effective degrees of freedom.[39] The results of this comparison are presented in Table 3. Parameter tables for the ML models are given in Tables 5 and 6 in Appendix B.

As Table 3 shows, Lasso-penalization clearly outperforms the simple ML estimator. For both the main effects and the interaction model, the Lasso considerably improves

[37] See, e.g., Brambor, Clark, and Golder (2006); Kam and Franzese (2007); Berry, Golder, and Milton (2012); Berry, DeMeritt, and Esarey (2010); Ai and Norton (2003); Norton, Wang, and Ai (2004); Braumoeller (2004).

[38] The translation of changes in the linear predictors to changes in odds or odds ratios is unaffected by the presence of interactions and works as usual.

[39] For an unpenalized MNL with symmetric side constraint, effective degrees of freedom equal the number of nominal parameters minus the number of individual-specific variables.

Model	CV	AIC	BIC	$\widehat{\text{df}}$
ML main effects	203.44	2013.44	2291.00	59
Lasso main effects	195.28	1982.19	2127.06	31
ML interaction	223.74	2115.66	3098.88	209
Lasso interaction	187.42	1935.71	2237.37	72

Table 3: Comparison of models based on predictive performance and complexity.

all four considered predictive performance and complexity measures compared to the ML estimator. Using ML estimation, the inclusion of interactions leads to a remarkable deterioration of the model. This finding again demonstrates that, within the ML context, one would have to select important interactions manually or via elaborate testing procedures. By contrast, the Lasso approach easily allows including all possible interactions. The Lasso-penalized interaction model exhibits by far the best crossvalidation score and AIC. Since the ML interaction model includes 209 potential effects, compared to the 59 of the corresponding main effects model, it comes as no surprise that the former uses more parameters than the latter, even after the Lasso has culled irrelevant effects. Therefore, the Lasso main effects model outperforms the Lasso interaction model in terms of the BIC, which assigns a heavy penalty to each parameter. In the end, the choice between these two models will depend on the researcher's preference and the study's objective: the interaction model offers more detailed insights into the importance of political issues for specific voter segments, but at the same time it is much more complex and difficult to interpret than the main effects model whose strength lies in its simplicity.

To support our arguments from Section 3.2.3, Table 4 summarizes the performance measures for the Lasso-penalized main effects model based on the six possible identifiability contraints (i.e., five choices of reference category and symmetric side constraint). It is immediately seen that the sparsity of the models varies substantially across different choices of the reference category, indicated by the amount of nonzero parameters. Therefore, the Lasso's predictive performance – as CV, AIC and BIC illustrate – also varies considerably across reference categories. Applying the symmetric side constraint provides in this application the most parsimonious model that also performs best in terms of AIC and BIC. This observed pattern, however, cannot be generalized – the relative performance of a model applying a symmetric side constraint compared to one using a reference category depends on the dataset at hand. Nonetheless, the symmetric side constraint provides a universal solution to the identifiability issue in MNLs while the choice of a specific reference category is always arbitrary.

5 Conclusion

Multinomial logit models (MNLs) are a powerful tool to translate the Spatial Theory of Voting into statistical models. Although offering the possibility to address new aspects of

Identifiability Contraint	CV	AIC	BIC	$\widehat{\mathrm{df}}$
CDU	189.77	1999.40	2238.64	51
SPD	189.96	2002.38	2244.19	52
FDP	190.30	1988.09	2188.89	43
Greens	189.74	1990.99	2211.16	47
Leftist	190.37	2002.56	2237.27	50
Symmetric	195.28	1982.19	2127.06	31

Table 4: Comparison of performance measures for Lasso main effects model based on different choices of the identifiability constraint.

issue voting, such as party-specific issue reactions, the flexibility of these models implies at the same time an enormous proliferation of coefficients, highlighting the practical need for sophisticated parameter selection techniques.

In this paper, we outline the benefits of regularization methods in the statistical analysis of the Spatial Theory of Voting. In particular, we explicitly derive for the first time Lasso-type regularization techniques of MNLs that take into account both individual- and alternative-specific variables, and that are able to incorporate the alternative-wise specification of choice-specific covariates (e.g., issue distances). Since the Lasso is able to set parameters of irrelevant predictors to exactly zero, the corresponding effects are effectively removed from the model. Thus, the Lasso implicitly performs variable selection as a by-product of the estimation process, and therefore enforces parameter selection and reduction of the predictor space. Hence and in contrast to classical selection techniques, the Lasso guarantees continuous, stable and computationally efficient variable selection. We also illustrate that the usage of adaptive weights yields a considerable improvement of the Lasso's performance, both in terms of variable selection and prediction accuracy.

By applying the Lasso method to the 2009 German Parliamentary Election, we demonstrate that our proposed approach massively reduces the model's complexity and simplifies its interpretation by selecting highly interesting and theoretically promising effects. In addition, it improves the model's predictive performance. More specifically, Lasso-penalization clearly outperforms the simple ML estimator, for both the main effects and the interaction model. Finally, we show that using a symmetric identifiability constraint prevents that the arbitrary choice of a reference category influences the selection of the most promising model. In particular, we demonstrate that the models' sparsity and complexity substantially vary across different choices of the reference category. Therefore, we strongly recommend the usage of symmetric side constraints as identifiability constraint in MNLs. Naturally, other interpretation techniques as well as the derivation of equilibria can build on these results.

A Properties of Lasso Estimates

Let $\hat{\boldsymbol{\theta}}_\lambda$ denote the Lasso estimator of the model's overall parameter vector $\boldsymbol{\theta}$ for tuning parameter λ. Let $\lambda_{\max} > \lambda_2 > \lambda_1 > 0$ and let $||\cdot||_1$ denote the L_1-norm.

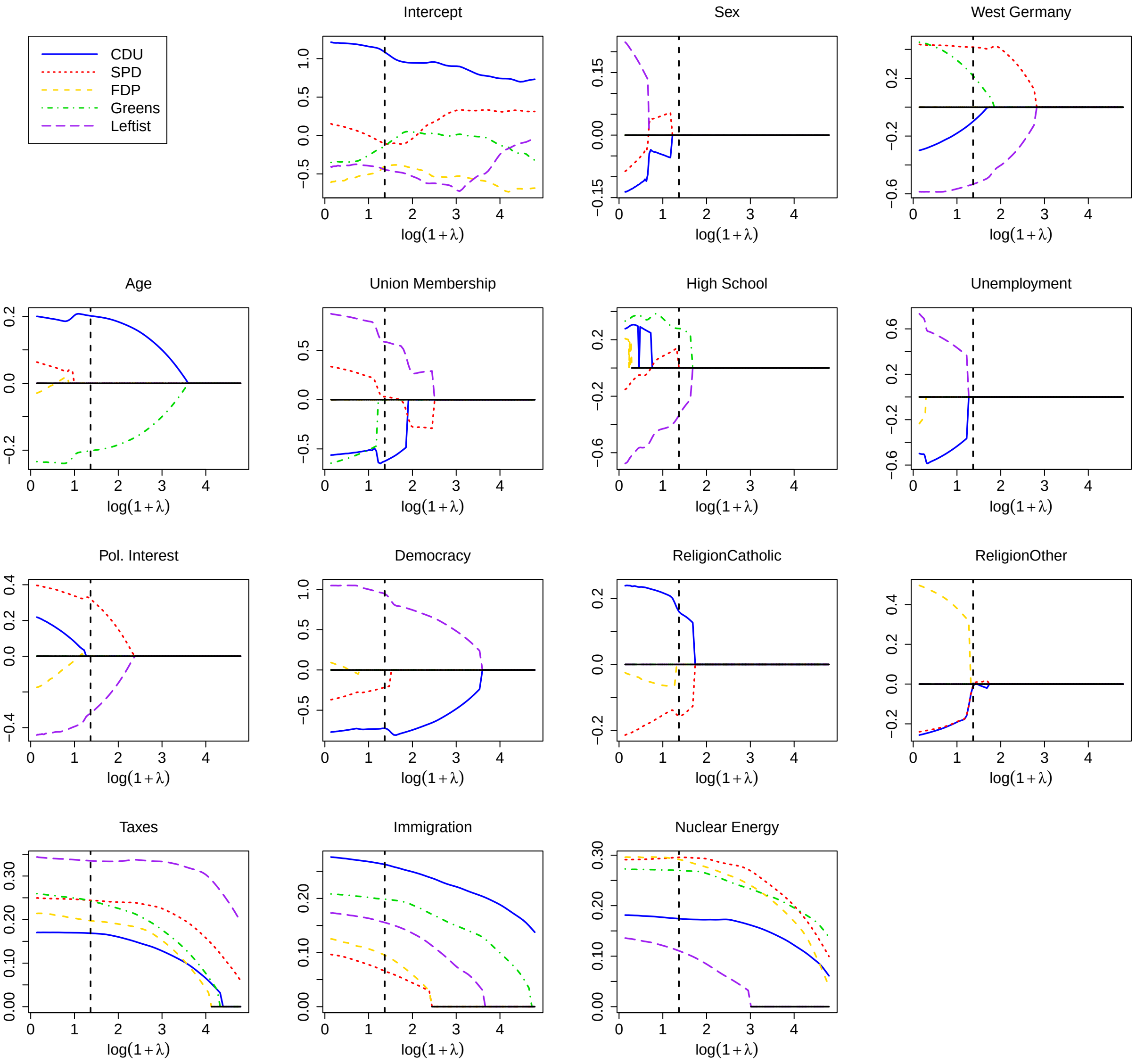
Then, it follows from the constraint definition of the Lasso that the L_1 norm of the Lasso is smaller than that of the ML estimator and that it decreases with increasing penalty level: $||\hat{\boldsymbol{\theta}}^{ML}||_1 = ||\hat{\boldsymbol{\theta}}_0||_1 > ||\hat{\boldsymbol{\theta}}_{\lambda_1}||_1 > ||\hat{\boldsymbol{\theta}}_{\lambda_2}||_1$. Note, however, that not every single coefficient is shrunk compared to its ML counterpart. For groups of highly correlated predictors, the Lasso typically only selects one of them while the others are set to zero. In such cases, the coefficient for the selected predictor partially subsumes the effects of the removed correlated predictors. Thus, in the presence of correlation among the predictors, single coefficients can get larger when λ is increased.

Due to its shrinkage property, the Lasso is a biased estimator. Its variance, by contrast, becomes smaller for a larger degree of penalization. It can be shown that $\text{Var}(\hat{\boldsymbol{\theta}}^{ML}) > \text{Var}(\hat{\boldsymbol{\theta}}_{\lambda_1}) > \text{Var}(\hat{\boldsymbol{\theta}}_{\lambda_2})$.[40]

The last property of the Lasso that we consider is its effective degrees of freedom. Since the Lasso is a shrinkage estimator, its effective degrees of freedom will intuitively be smaller than the number of estimated parameters. In Zou, Hastie, and Tibshirani (2007), it was shown that its effective degrees of freedom can be estimated by the number of nonzero parameters and that this df-estimator is unbiased and consistent: $\hat{\text{df}}(\hat{\boldsymbol{\theta}}_\lambda) = \sum_i I(|\hat{\theta}_{i,\lambda}| > 0)$. However, this general formula has to be handled with care due to the identifiability constraints used in MNLs. We will give an explicit formula for the df of the Lasso for our model in Section 3.2.3

[40] Here, $\text{Var}(\hat{\boldsymbol{\theta}})$ is a covariance matrix, so that the notation "$\text{Var}(\hat{\boldsymbol{\theta}}_{\lambda_1}) > \text{Var}(\hat{\boldsymbol{\theta}}_{\lambda_2})$" means that the difference of these matrices positive definite.

B Extended Figures and Tables



Note: The coefficient paths are based on the main effects model in Table 1.

Figure 4: Coefficient paths for all predictors, German Parliamentary Election 2009.

	CDU		**SPD**		**FDP**		**Greens**		**Leftist**	
ASC	1.22***	(0.000)	0.17	(0.597)	-0.65	(0.080)	-0.32	(0.504)	-0.42	(0.267)
Sex	-0.21	(0.194)	-0.17	(0.279)	0.13	(0.547)	0.08	(0.698)	0.17	(0.394)
West Germany	-0.33	(0.056)	0.42*	(0.029)	0.07	(0.786)	0.45	(0.114)	-0.61**	(0.004)
Age	0.19***	(0.000)	0.06	(0.204)	-0.05	(0.476)	-0.24***	(0.000)	0.03	(0.580)
Union Membership	-0.48	(0.073)	0.43*	(0.029)	-0.34	(0.303)	-0.57	(0.084)	0.96***	(0.000)
High School	0.30	(0.140)	-0.16	(0.488)	0.22	(0.385)	0.34	(0.189)	-0.69*	(0.015)
Unemployment	-0.48	(0.289)	0.04	(0.916)	-0.29	(0.678)	-0.07	(0.906)	0.79*	(0.022)
Pol. Interest	0.22	(0.266)	0.38*	(0.033)	-0.20	(0.340)	0.07	(0.780)	-0.46*	(0.033)
Democracy	-0.74***	(0.000)	-0.35*	(0.037)	0.15	(0.490)	-0.14	(0.530)	1.08***	(0.000)
ReligionCatholic	0.39*	(0.037)	-0.08	(0.653)	0.14	(0.609)	-0.22	(0.418)	-0.22	(0.371)
ReligionOther	-0.34	(0.076)	-0.33	(0.099)	0.43	(0.088)	0.06	(0.801)	0.18	(0.410)
Taxes	0.17**	(0.004)	0.25***	(0.000)	0.22**	(0.001)	0.26**	(0.001)	0.34***	(0.000)
Immigration	0.28***	(0.000)	0.10	(0.064)	0.12	(0.075)	0.21***	(0.000)	0.17**	(0.001)
Nuclear Energy	0.18***	(0.000)	0.29***	(0.000)	0.29***	(0.000)	0.27***	(0.000)	0.14*	(0.031)

Numbers in parentheses show approximate p-values based on simple, two-sided t-tests using bootstrapped standard errors.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: ML coefficient estimates of the main effects model, German Parliamentary Election 2009.

	CDU		**SPD**		**FDP**		**Greens**		**Leftist**	
ASC	2.34**	(0.004)	0.51	(0.521)	-2.13*	(0.014)	-0.43	(0.645)	-0.28	(0.747)
Sex	-0.41	(0.340)	0.31	(0.471)	0.28	(0.595)	0.21	(0.677)	-0.39	(0.432)
West Germany	-0.49	(0.335)	0.82	(0.158)	0.33	(0.567)	-0.10	(0.886)	-0.56	(0.337)
Age	0.20	(0.146)	-0.07	(0.585)	-0.10	(0.551)	-0.03	(0.856)	0.01	(0.958)
Union Membership	-1.22	(0.168)	0.23	(0.734)	0.76	(0.574)	-1.33	(0.077)	1.55*	(0.031)
High School	0.37	(0.530)	-1.00	(0.119)	0.63	(0.317)	0.62	(0.347)	-0.63	(0.363)
Unemployment	0.49	(0.735)	1.13	(0.379)	-4.18*	(0.011)	1.09	(0.448)	1.48	(0.254)
Pol. Interest	0.24	(0.613)	-0.09	(0.853)	-0.89	(0.087)	0.86	(0.139)	-0.13	(0.802)
Democracy	-0.70	(0.130)	-0.69	(0.106)	1.09	(0.051)	-0.65	(0.261)	0.95	(0.147)
ReligionCatholic	-0.98	(0.105)	-0.51	(0.284)	1.63*	(0.019)	0.17	(0.802)	-0.30	(0.647)
ReligionOther	-1.53**	(0.007)	-0.50	(0.346)	1.37*	(0.031)	0.69	(0.284)	-0.03	(0.960)
Taxes	0.54*	(0.037)	0.67*	(0.027)	-0.06	(0.809)	0.52	(0.159)	0.44	(0.178)
Immigration	0.37	(0.180)	0.06	(0.797)	0.04	(0.882)	0.26	(0.356)	0.34	(0.195)
Nuclear Energy	0.42	(0.056)	0.37	(0.212)	0.16	(0.598)	0.15	(0.576)	0.19	(0.533)
Taxes X Sex	-0.18	(0.252)	-0.06	(0.697)	0.32	(0.086)	-0.18	(0.475)	0.04	(0.836)
Taxes X West Germany	-0.21	(0.224)	0.20	(0.310)	-0.13	(0.463)	0.11	(0.652)	0.19	(0.344)
Taxes X Age	0.02	(0.765)	0.01	(0.770)	-0.05	(0.360)	0.09	(0.208)	-0.08	(0.248)
Taxes X Union Membership	-0.37	(0.140)	-0.22	(0.371)	0.40	(0.482)	-0.36	(0.274)	0.24	(0.424)
Taxes X High School	0.16	(0.462)	-0.14	(0.573)	0.13	(0.568)	0.11	(0.681)	-0.04	(0.896)
Taxes X Unemployment	0.47	(0.455)	-0.40	(0.548)	-1.53**	(0.007)	0.41	(0.547)	-0.12	(0.826)
Taxes X Pol. Interest	-0.14	(0.462)	-0.11	(0.538)	-0.09	(0.644)	0.28	(0.284)	-0.03	(0.869)
Taxes X Democracy	0.12	(0.434)	-0.24	(0.156)	0.22	(0.212)	-0.14	(0.612)	-0.25	(0.443)
Taxes X ReligionCatholic	-0.04	(0.860)	-0.26	(0.205)	0.51	(0.057)	-0.28	(0.351)	0.04	(0.882)
Taxes X ReligionOther	-0.27	(0.187)	-0.39	(0.100)	0.02	(0.903)	-0.36	(0.216)	-0.03	(0.887)
Immigration X Sex	0.07	(0.667)	0.12	(0.363)	-0.15	(0.405)	0.16	(0.299)	0.02	(0.971)
Immigration X West Germany	0.10	(0.557)	-0.11	(0.513)	0.19	(0.332)	-0.29	(0.112)	-0.24	(0.177)
Immigration X Age	0.05	(0.219)	-0.01	(0.862)	0.03	(0.598)	0.07	(0.178)	-0.03	(0.550)
Immigration X Union Membership	-0.02	(0.930)	0.27	(0.224)	-0.26	(0.541)	0.32	(0.196)	-0.02	(0.949)
Immigration X High School	-0.17	(0.334)	-0.07	(0.754)	-0.22	(0.349)	-0.10	(0.654)	0.12	(0.623)
Immigration X Unemployment	-0.08	(0.911)	0.35	(0.549)	0.06	(0.901)	-0.14	(0.856)	0.26	(0.582)
Immigration X Pol. Interest	0.27	(0.115)	-0.01	(0.926)	-0.19	(0.342)	0.07	(0.698)	0.12	(0.448)
Immigration X Democracy	-0.22	(0.106)	-0.01	(0.944)	0.10	(0.585)	-0.01	(0.954)	-0.12	(0.557)
Immigration X ReligionCatholic	-0.46*	(0.023)	0.01	(0.957)	0.10	(0.666)	0.11	(0.572)	0.18	(0.403)
Immigration X ReligionOther	-0.19	(0.403)	0.19	(0.303)	0.27	(0.194)	0.18	(0.343)	-0.12	(0.532)
Nuclear Energy X Sex	-0.03	(0.805)	0.12	(0.521)	-0.07	(0.627)	-0.01	(0.976)	-0.38	(0.053)
Nuclear Energy X West Germany	0.02	(0.896)	0.14	(0.510)	-0.05	(0.783)	0.01	(0.996)	0.20	(0.389)
Nuclear Energy X Age	-0.05	(0.227)	-0.05	(0.317)	0.02	(0.734)	-0.03	(0.597)	0.07	(0.179)
Nuclear Energy X Union Membership	0.17	(0.550)	-0.01	(0.984)	1.08	(0.164)	-0.22	(0.449)	0.29	(0.367)
Nuclear Energy X High School	0.05	(0.765)	-0.23	(0.349)	0.28	(0.268)	0.11	(0.568)	-0.12	(0.629)
Nuclear Energy X Unemployment	0.14	(0.870)	0.53	(0.495)	2.05**	(0.010)	0.34	(0.727)	-0.08	(0.903)
Nuclear Energy X Pol. Interest	-0.19	(0.175)	-0.25	(0.232)	-0.02	(0.926)	-0.08	(0.674)	-0.04	(0.825)
Nuclear Energy X Democracy	0.15	(0.261)	0.09	(0.642)	0.14	(0.502)	-0.09	(0.593)	0.27	(0.218)
Nuclear Energy X ReligionCatholic	-0.20	(0.224)	-0.04	(0.843)	0.13	(0.610)	0.26	(0.249)	-0.28	(0.355)
Nuclear Energy X ReligionOther	-0.22	(0.186)	-0.10	(0.645)	-0.06	(0.749)	0.27	(0.155)	-0.10	(0.699)

Numbers in parentheses show approximate p-values based on simple, two-sided t-tests using bootstrapped standard errors.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: ML coefficient estimates of the interaction model, German Parliamentary Election 2009.

References

Adams, James F., Samuel Merrill, and Bernard Grofman (2005). *A unified theory of party competition: A cross-national analysis integrating spatial and behavioral factors*. Cambridge and New York: Cambridge University Press.

Ai, Chunrong and Edward C. Norton (2003). “Interaction terms in logit and probit models”. In: *Economics Letters* 80.1, pp. 123–129.

Alvarez, R. Michael and Jonathan Nagler (1998). “When politics and models collide: Estimating models of multiparty elections”. In: *American Journal of Political Science* 42.1, pp. 55–96.

Anderson, Christopher J and Jason D Hecht (2012). "Voting when the economy goes bad, everyone is in charge, and no one is to blame: The case of the 2009 German election". In: *Electoral Studies* 31.1, pp. 5–19.

Beck, Amir and Marc Teboulle (2009). "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". In: *SIAM Journal on Imaging Sciences* 2.1, pp. 183–202.

Ben-Akiva, Moshe E. and Steven R. Lerman (1985). *Discrete choice analysis: Theory and application to travel demand.* Cambridge Massachusetts: MIT Press.

Berry, William D., Jacqueline H. R. DeMeritt, and Justin Esarey (2010). "Testing for interaction in binary logit and probit models: Is a product term essential?" In: *American Journal of Political Science* 54.1, pp. 248–266.

Berry, William D., Matt Golder, and Daniel Milton (2012). "Improving tests of theories positing interaction". In: *The Journal of Politics* 74.03, pp. 653–671.

Boyd, Stephen P. and Lieven Vandenberghe (2004). *Convex optimization.* Cambridge: Cambridge University Press.

Brambor, Thomas, William R. Clark, and Matt Golder (2006). "Understanding interaction models: Improving empirical analyses". In: *Political Analysis* 14.1, pp. 63–82.

Braumoeller, Bear F. (2004). "Hypothesis testing and multiplicative interaction terms". In: *International Organization* 58.4, pp. 807–820.

Budge, Ian and Dennis J. Farlie (1983). "Party competition – selective emphasis or direct confrontation? An alternative view with data". In: *Western European Party Systems. Continuity and Change.* Ed. by Hans Daalder and Peter Mair. Beverly Hills, London, New Delhi: Sage, pp. 267–305.

Converse, Philip E. (1964). "The nature of belief systems in mass publics". In: *Ideology and Discontent.* Ed. by David Apter. New York: Free Press, pp. 240–268.

Davis, Otto A., Melvin J. Hinich, and Peter C. Ordeshook (1970). "An expository development of a mathematical model of the electoral process". In: *The American Political Science Review* 64.2, pp. 426–448.

Dow, Jay K. and James W. Endersby (2004). "Multinomial probit and multinomial logit: A comparison of choice models for voting research". In: *Electoral Studies* 23.1, pp. 107–122.

Downs, Anthony (1957). *An economic theory of democracy.* New York: Harper & Row.

Elff, Martin (2007). "Social structure and electoral behavior in comparative perspective: The decline of social cleavages in Western Europe revisited". In: *Perspectives on Politics* 5.02, pp. 277–294.

— (2009). "Social divisions, party positions, and electoral behaviour". In: *Electoral Studies* 28.2, pp. 297–308.

Elff, Martin and Sigrid Rossteutscher (2011). "Stability or decline? Class, religion and the vote in Germany". In: *German Politics* 20.1, pp. 107–127.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2010). "Regularization paths for generalized linear models via coordinate descent". In: *Journal of Statistical Software* 33.1, pp. 1–22.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The elements of statistical learning: Data mining, inference and prediction.* New York: Springer.

Hoerl, Arthur E. and Robert W. Kennard (1970). "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1, pp. 55–67.

Jun, Uwe (2011). "Volksparteien under pressure: challenges and adaptation". In: *German Politics* 20.1, pp. 200–222.

Kam, Cindy D. and Robert J. Franzese (2007). *Modeling and interpreting interactive hypotheses in regression analysis.* University of Michigan Press.

King, Gary, Michael Tomz, and Jason Wittenberg (2000). "Making the most of statistical analyses: Improving interpretation and presentation". In: *American Journal of Political Science* 44.2, pp. 347–361.

Krosnick, Jon A. (1990). "Government policy and citizen passion: A study of issue publics in contemporary America". In: *Political Behavior* 12.1, pp. 59–92.

Lees, Charles (2012). "The paradoxical effects of decline: Assessing party system change and the role of the catch-all parties in Germany following the 2009 federal election". In: *Party Politics* 18.4, pp. 545–562.

Louviere, Jordan J., David A. Hensher, and Joffre D. Swait (2000). *Stated choice methods: Analysis and applications.* Cambridge University Press.

Manski, Charles F. (1973). *The analysis of qualitative choice: Ph.D. Dissertation. Department of Economics.* Cambridge: Mass.

— (1977). "The structure of random utility models". In: *Theory and Decision* 8.3, pp. 299–254.

Mauerer, Ingrid, Paul W. Thurner, and Marc Debus (2014). "Party-varying issue voting: Identifying the impact of issue campaign strategies". In: *Ms. LMU Munich [under review].*

McFadden, Daniel (1973). "Conditional logit analysis of qualitative choice behavior". In: *Frontiers in Econometrics.* Ed. by Paul Zarembka. New York: Academic Press, pp. 105–142.

— (1984). *Econometric analysis of qualitative response models.* Ed. by Z. Griliches and M. D. Intriligator.

Mebane, Walter R., John E. Jackson, and Jonathan Wall (2014). "Preference heterogeneities in models of electoral behavior". In: *Working paper presented at the 2014 Annual Meeting of the Midwest Political Science Association, Chicago, April 3-6, 2014.*

Norton, Edward C., Hua Wang, and Chunrong Ai (2004). "Computing interaction effects and standard errors in logit and probit models". In: *Stata Journal* 4, pp. 154–167.

Petrocik, John R. (1996). "Issue ownership in presidential elections, with a 1980 case study". In: *American Journal of Political Science* 40.3, pp. 825–850.

Schofield, Normal et al. (1998). “Multiparty electoral competition in the Netherlands and Germany: A model based on multinomial probit”. In: *Empirical Studies in Comparative Politics*. Ed. by Melvin J. Hinich and Michael C. Munger. Springer US, pp. 39–75.

Singh, Shane (2014). “Linear and quadratic utility loss functions in voting behavior research”. In: *Journal of Theoretical Politics* 26.1, pp. 35–58.

Thurner, Paul W. (2000). “The empirical application of the spatial theory of voting in multiparty systems with random utility models”. In: *Electoral Studies* 19.4, pp. 493–517.

Tibshirani, Robert (1996). “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society B* 58.1, pp. 267–288.

Train, Kenneth (1978). “A validation test of a disaggregate mode choice model”. In: *Transportation Research* 12.3, pp. 167–174.

— (2009). *Discrete choice methods with simulation*. 2nd ed. Cambridge University Press.

Tutz, Gerhard (2010). “Guest Editorial: Regularisation methods in regression and classification”. In: *Statistics and Computing* 20.2, pp. 117–118.

— (2012). *Regression for categorical data*. Cambridge University Press.

Tutz, Gerhard and Gunther Schauberger (2013). “Visualization of categorical response models - from Data Glyphs to Parameter Glyphs”. In: *Journal of Computational and Graphical Statistics* 22, pp. 156–177.

Zou, H. (2006). “The adaptive lasso and its oracle properties”. In: *Journal of the American Statistical Association* 101.476, pp. 1418–1429.

Zou, Hui, Trevor Hastie, and Robert Tibshirani (2007). “On the ”degrees of freedom” of the lasso”. In: *The Annals of Statistics* 35.5, pp. 2173–2192.