Micha Andreas Hermann Schneider

Finite Mixtures for the Modelling of Heterogeneity in Ordinal Response

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 7.8.2019

1. Berichterstatter: Prof. Dr. Gerhard Tutz
2. Berichterstatter: Prof. Dr. Matthias Schmid
3. Berichterstatter: Prof. Dr. Christian Heumann

Tag der Disputation: 29.11.2019

Zusammenfassung

Die Modellierung von Heterogenität ist ein entscheidender Aspekt in jeder statistischen Analyse. Um ein geeignetes Modell zu finden, ist es notwendig, möglichst alle relevanten Strukturen und Einflussgrößen einzubeziehen. Die meisten statistischen Modelle können leicht beobachtete Strukturen einbinden, jedoch haben sie oft Schwierigkeiten latente Strukturen abzubilden. Misch-Modelle können Heterogenität berücksichtigen, die aus zugrunde liegenden latenten Strukturen entstehen, wie etwa die unbeobachtete Zugehörigkeit zu verschiedenen Gruppen oder unterschiedliches Antwortverhalten. Mit dieser Doktorarbeit möchte ich einen Beitrag für die Verwendung von Misch-Modellen zur Modellierung von Heterogenität bei ordinalen Zielgrößen leisten und Variablen Selektion in diesem Kontext durchführen.

Zuerst konzentriere ich mich auf Heterogenität, die bei Umfragen auftritt, wenn beispielsweise die Befragten bei der Wahl einer bestimmten geordneten Kategorie unsicher sind. In diesem Fall bestehen die Misch-Modelle üblicherweise aus einer Präferenz-Komponente und einer Unsicherheits-Komponente. Ein Gewicht bestimmt die Neigung jeder Person zu einer dieser beiden Komponenten zu gehören. Das existierende CUB Modell verwendet eine verschobene Binomialverteilung für die erste und eine Gleichverteilung für die zweite Komponente. Im vorgeschlagenem CUP Modell wird die Präferenz-Komponente mit einem beliebigen ordinalen Modell wie dem kumulativen Logit Modell ersetzt, um eine höhere Flexibilität in der Präferenz-Komponente zu erreichen. Im BetaBin Modell wird das Konzept der Unsicherheit als zufällige Wahl einer Kategorie so erweitert, dass Unsicherheit auch die Tendenz zu der zentralen Kategorie und extremen Kategorien erfasst. Auf diese Weise wird die Gleichverteilung des CUP Modells durch einer flexiblere, beschränkte Beta-Binomial Verteilung ersetzt.

Als zweites zeige ich, wie diskrete Cure Modelle verwendet werden können, um in der Survival-Analyse für diskrete Zeit mit Heterogenität umzugehen, die aus der unbeobachteten Zugehörigkeit zu verschiedenen Gruppen entsteht. „Cure“ bezeichnet dabei den Umstand, dass eine Gruppe von Beobachtungen „geheilt ist“ oder als sogenannte Langzeit-Überlebende charakterisiert ist, während die andere Gruppe dem Risiko des Ereignisses wie zum Beispiel „Eintritt von Arbeitslosigkeit“ ausgesetzt ist. Die Zugehörigkeit zu dieser Gruppe ist unbekannt. Cure Modelle schätzen die Wahrscheinlichkeit zur Nicht-geheilten Population zu gehören und die Form der Survival Funktion für die Beobachtungen unter Risiko.

Drittens führe ich Variablen Selektion für das CUB, CUP und das Cure Modell mit Hilfe von Penalisierung und teilweise schrittweise Selektionsverfahren durch. Die Herausforderung liegt insbesondere darin zu entscheiden, welche Variablen in welche Komponente des Misch-Modells aufgenommen werden sollen. Variablen können hier zum einen für die Schätzung der Gewichte der Komponenten und zum anderen für die Form einer oder zwei Misch-Komponenten verwendet werden. Es werden dafür spezifische Bestrafungsterme vorgestellt, die für das jeweilige Modell geeignet sind.

Alle Modelle werden mit dem EM-Algorithmus geschätzt, der die unbekannte Zugehörigkeit zu einer der Komponenten als fehlende Daten behandelt. Es werden auch einige computationale Aspekte besprochen wie etwa mit der Initialisierung und der Konvergenz umzugehen ist. Die penalisierte Likelihood wird mit dem sogenannten FISTA Algorithmus geschätzt, da die Ableitungen der penalisierten Likelihood nicht existieren. Es werden sowohl Simulations-Studien als auch reelle Daten verwendet, um die Nützlichkeit der neuen Ansätze aufzuzeigen.

Abstract

Modelling heterogeneity is a crucial aspect of every statistical analysis. To find a reasonable model, it is necessary to include all relevant structures and explanatory variables. Most statistical models can easily include observed patterns but have often difficulties in dealing with latent structures. Mixture models can account for heterogeneity which arise from latent underlying structures, for example, the unobserved membership to different groups or different response styles. In this thesis, I contribute to the use of mixture models to model heterogeneity in ordinal response and perform variable selection in this context.

First, I focus on heterogeneity, which occurs in surveys when, for instance, respondents are uncertain about choosing a certain ordered category. In this case, the mixture model traditionally consists of a preference component and an uncertainty component. A weight determines the propensity of each person belonging to one of these components. The traditional CUB model uses a shifted binomial distribution for the first and a uniform distribution for the later component. In the proposed CUP model, the preference component is replaced by any ordinal model, such as the cumulative logit model or the adjacent category model, to achieve more flexibility in the preference component. In the BetaBin model, the concept of uncertainty, understood as a random choice of a category, is extended in such a way that uncertainty can also capture the tendency to the middle and extreme categories. Thus, the uniform distribution of the CUP model is replaced by a more flexible restricted beta-binomial distribution.

Second, I show how discrete cure models can be used for dealing with heterogeneity in the survival analysis for discrete time arising from the unobserved membership to different groups. "Cure" refers to the fact that one group of observations is "cured" or characterized as long-term survivors, while the other group is exposed to the risk of the event such as the "occurrence of unemployment". The membership to this group is unknown. Cure models estimate the probability for belonging to the non-cured population and the shape of the survival function of the observations under risk.

Third, I perform variable selection for the CUB, the CUP and the cure model using penalization techniques and to some extend stepwise selection procedures. In particular, the challenge is to decide which variables should be included in which component of the mixture model. On the one hand, variables can be used to estimate the weights of the components and on the other hand, for the shape of one or two mixture components. Therefore, specific penalty terms are presented which are appropriate for the particular model.

All models are estimated with the EM-Algorithm which treats the unknown membership to the components as missing data. I also address some computational issues, for instance, how to deal with initialization and convergence. The penalized likelihood is estimated with the so-called FISTA algorithm since the derivatives of the penalized likelihood do not exist. Both simulation studies and real data applications are used to demonstrate the usefulness of the new approaches.

Acknowledgements

Ich möchte mich bedanken bei ...

- meinem Doktorvater Prof. Dr. Gerhard Tutz für seine hervorragende Betreuung und den konstruktiven und produktiven Austausch zu wissenschaftlichen Fragen,
- Prof. Dr. Matthias Schmid und Prof. Dr. Christian Heumann für die Bereitschaft meine Dissertation zu begutachten und hilfreiche Impulse,
- Prof. Dr. Thomas Augustin für seine offene Tür und Übernahme des Vorsitz in meiner Prüfungskommision und Prof. Dr. Helmut Küchenhoff für die Mitwirkung in meiner Prüfungskommission,
- meinen ehemaligen Kollegen am Lehrstuhl für angewandte Stochastik Gunther Schauberger, Moritz Berger und insbesondere Wolfgang Pößnecker für ihr offenes Ohr bei Fragen und produktiven Diskussionen,
- allen weiteren Mitarbeitern am Institut für Statistik, insbesondere den Mitgliedern der "Mensa"-Gruppe und der Arbeitsgruppe Augustin, für viele Gespräche und entspannte Mittagspausen,
- Ingrid Mauerer und Paul Fink für den kollegialen Austausch,
- allen, die mich während der Promotionszeit unterstützt haben, insbesondere meinen Eltern, die immer für mich da waren.

Overview

1. Introduction

Heterogeneity arises in almost all statistical data. It can be understood as any diverseness of the data. Traditional statistical regression models account for it by using explanatory covariates, such as gender or age, to describe the relationship to a dependent variable, such as income. For example, the simple linear regression assumes a linear association of one covariable to one response variable and a normal distributed error term. More advanced regression models include several explanatory variable, allow for non-parametric relationships between explanatory and dependent variables or are designed for a discrete response. However, they are limited if there is an underlying unobserved, latent structure.

For example, one is interested in modelling the income of athletes. It seems to be reasonable to assume that professional athletes, who live on sports, earn more money than amateurs with respect to the income raised by sports. Thus the income is not homogeneous in the whole population of athletes but is characterized by two groups. The membership of a certain person to one of the two latent groups (professionals vs. amateurs) might not always be observed, but rather estimated by covariates such as "amount of time spent for training". Thus the observed income of athletes can be modelled by a so-called mixture model which is a weighted sum of two income densities. The weights represent the estimated membership to the groups. An introduction to mixture models is given by McLachlan and Peel (2000) and Frühwirth-Schnatter (2006).

In this thesis, I focus on two main application areas: Heterogeneity in social surveys and heterogeneity in the discrete survival analysis. Both cases have in common that the response is measured on an ordinal scale, namely an ordinal Likert scale or as discrete time until an event occurs.

Heterogeneity in social surveys is often interpreted as uncertainty on the side of the respondents. They may have difficulties in choosing a certain category, due to, for instance, lack of time or self-confidence. Thus discrete human choice can be considered as a combination of preference and an uncertainty structure. The components are combined by a mixture weight or propensity which can be determined by covariates. The traditional CUB model, introduced by D'Elia and Piccolo (2005), uses a shifted binomial distribution for the preference component and a uniform distribution for the uncertainty component. Two extensions of this model are proposed. Firstly, Tutz et al. (2017) replace the shifted binomial distribution by ordinal response models, such as the cumulative and adjacent categories model, to achieve more flexibility in the preference component. The so-called CUP model frequently yields a better fit than classical ordinal response models without an uncertainty component. The CUP model is applied to several data sets and a simulation shows that the effect of explanatory variables is underestimated if the uncertainty component

is neglected.

Secondly, Tutz and Schneider (2019) replace the uniform distribution of the CUP model with a restricted beta-binomial distribution to distinguish between the tendency to middle categories and the tendency to extreme categories in the uncertainty component. In the so-called BetaBin model, variables are used to determine the individual shape of the uncertainty distribution. It is demonstrated that severe bias might occur, if inadvertently the uniform distribution is used to model uncertainty. An application to attitudes on the performance of health services illustrates the advantages of the more flexible model. Mauerer and Schneider (2019a) use the same model approach to model the perceived party placements on immigration in the 2017 German election. The BetaBin model is also used by Mauerer and Schneider (2019b) to develop a vote choice model that accounts for uncertainty in issue placements.

In survival analysis, there may be groups of persons who differ in their survival probabilities, but the membership to these groups is unobserved. For example some prisoners released from prison may never be arrested again, but others may lapse back into crime and return to prison. Schneider (2019) presents how discrete cure models are defined, estimated and applied. It is possible to estimate factors, which are associated with being long-term survivor or with the occurence of an event.

Variable selection via group lasso regularization has been applied for the CUB, CUP model and the cure model. The challenge is to decide which variables should be included in which part of the mixture model. There might be variables to estimate the mixture weights and to estimate the shapes of one or two components. Schneider et al. (2019) focus on specific penalty terms for the CUB and CUP model. The method is applied to real data and investigated in a simulation study. The results are compared with a stepwise selection. Some computational issues are addressed, such as initialization and convergence. Schneider (2019) demonstrate how specific penalty terms can be used for variable selection and smoothing the baseline in the discrete survival analysis. The approach is applied to data about criminal recidivism and breast cancer.

The thesis consists of six published articles. The author's contribution is as follows:

- Tutz, G., M. Schneider, M. Iannario and D. Piccolo (2017): Mixture models for ordinal responses to account for uncertainty of choice. *Advances in Data Analysis and Classification 11(2)*, 281–305, doi:10.1007/s11634-016-0247-9.

  The project was set up by Gerhard Tutz, who developed the theoretical framework and investigated the literature. Micha Schneider conducted the simulations and analyses and mainly wrote the sections on the empirical analysis and the EM-estimation. The other authors contributed by providing two data sets, two figures and outlining the presentation.

- Tutz, G., and M. Schneider (2019): Flexible uncertainty in mixture models for ordinal responses. *Journal of Applied Statistics*, 46(9), 1582–1601, doi:10.1080/02664763.2018.1555574.

  Gerhard Tutz conceptualized the theoretical framework and investigated the literature. Micha Schneider implemented the beta-binomial distribution in the mixture model and conducted the simulations and analyses. The article was written in close collaboration by both authors. Both authors contributed to the review process.

- Mauerer, I. and M. Schneider (2019a): Perceived party placements and uncertainty on immigration in the 2017 German election. In Debus, M., M. Tepe and J. Sauermann (Eds.), *Jahrbuch für Handlungs- und Entscheidungstheorie: Band 11*, pp. 117–143. Wiesbaden: Springer, doi:10.1007/978-3-658-23997-8.

  The article was written in very close collaboration by both authors. In particular, Ingrid Mauerer prepared the data and fit the topic into the current political science literature. Micha Schneider conducted the analyses and described the statistical model. Both authors contributed equally to the article and the review process.

- Mauerer, I. and M. Schneider (2019b): Uncertainty in Issue Placements and Spatial Voting. *Technical Report 226*, Department of Statistics, Ludwig-Maximilians-Universität München, doi:10.5282/ubm/epub.68451.

  Both authors developed the approach, and the manuscript was prepared in close collaboration between both authors. Micha Schneider conducted the analysis using the mixture model, computed the adjusted placements, and made the graphics. Ingrid Mauerer performed the data management, fitted the vote choice model, and initially wrote most parts of the manuscript.

- Schneider, M., Pößnecker, W. and G. Tutz (2019): Variable Selection in Mixture Models with an Uncertainty Component. *Technical Report 225*, Department of Statistics, Ludwig-Maximilians-Universität München, doi:10.5282/ubm/epub.68452.

  The manuscript was mainly written by Micha Schneider advised by Gerhard Tutz. Micha Schneider implemented the stepwise variable selection, the parallelized grid search of the penalized model, the current model initialization and convergence checks. He conducted the simulation and analyses. Wolfgang Pößnecker contributed R-Code for the EM-algorithm and an adaption of MRSP to fit the proposed model.

- Schneider, M. (2019): Dealing with Heterogeneity in Discrete Survival Analysis using the Cure Model. *Technical Report 224*, Department of Statistics, Ludwig-Maximilians-Universität München, doi:10.5282/ubm/epub.68455.

The underlying idea arose in a discussion with Gerhard Tutz. Micha Schneider developed the idea further and contributed solely to the paper.

A R-package `DiscMix` that includes the relevant functions is published on `https://github.com/Micha-Schneider/discmix` .

The thesis is structured as follow. First, a short overview of the nature of ordinal data is given and it is explained how ordinal data are treated in statistical models. Then, it is described how mixture models can be used to model heterogeneity in surveys. Here, the new approaches, the CUP model and the BetaBin model, are related to the existing approaches such as the CUB model (Section 3). In the next section, it is demonstrated how to use special mixture models, such as the cure model, in the context of discrete survival analysis. In Section 5 opportunities are outlined to perform variable selection within some of the discussed mixture models. New penalty terms for the CUB, CUP and cure models are introduced. In Section 6 the general estimation procedure is explained. Finally, the key points are summarized and an outlook of future research is given.

2. The Nature of Ordinal Data

Ordinal data arise in several applications. Examples are rating scales capturing opinions (e.g. "strongly disagree", "disagree", "intermediate", "agree", "strongly agree"), attitudes like smoking (e.g. "never", "sometimes", "often") or pain sensation. Another example is the measurement of time by discrete intervals (days, months, years, etc.). All ordinal data have in common that they are defined by categories which can be ordered. Some of them are inherently discrete, and others arise from a metric variable which is discretized, such as age, if it is measured in intervals $(0-18, 19-40, 40-65, 65+)$.

The ordinal scale can be distinguished from other levels of measurement. Nominal data only define the belonging to categories but no ordering such as gender ("male", "female", "non-binary"), marital status (e.g. "single", "married", "divorced", "widowed") or party preference. Using nominal models for ordinal data results in a loss of information because the ordinal nature of the data is not exploited. Sometimes ordinal data are analyzed by models for metric data such as the linear regression. Both the interval and the ratio scales allow calculating differences between the categories. The degree in Celsius is measured on an interval scale, whereas Kelvin belongs to the ratio scale. The difference is that the Kelvin scale is defined by an "absolute" zero value so that ratios are meaningful. However, both differences and ratios do not apply for ordinal data. The distance between ordered categories do not need to be equidistant and ratios are not sensible. German university grades from 1 to 5 are a good example. Receiving the best grade 1, for instance, does not mean that the student knows twice as much as a student with grade 2. Moreover the difference between failing (5) and passing (4) may be different from grade 1 to 2.

Thus, ordinal data should be treated appropriately and one should choose the models in order to use the full information provided by the data. An overview of categorical and ordinal data is given by McCullagh (1980), Agresti (2010), Agresti (2013) and Tutz (2012). In the following, I describe how ordinal variables can be treated as predictor and response in statistical regression models.

2.1. Ordinal Data as Predictors

Many models do not take into account the ordinal nature in predictor variables. The variables are treated either as nominal or metric. In the nominal case, the variables are, for example, dummy-coded so that the respective parameter measures the effect of a specific category compared to the reference category. However, there are some techniques to model the ordinal structure in the predictor in a more appropriate way. In the isotonic regression, order constraints are introduced (see Barlow et al., 1972; Robertson et al., 1988). Some other

researchers use specific splines (see e.g. Helwig, 2017; Leitenstorfer and Tutz, 2007; Ramsey, 1988). Tutz and Gertheiss (2014) apply a penalization approach to penalize the differences between adjacent categories and Tutz and Berger (2018) use trees. Recently, Bürkner and Charpentier (2018) propose a certain transformation which can be included in a Bayesian regression framework.

2.2. Regression Models for Ordinal Response

Several models are available to use the specific nature of ordinal responses. The cumulative model, the sequential model and the adjacent categories model are shortly introduced. Afterward the models are arranged into the context of the generalized linear models. Each model has its legitimated existence depending on the available data and research question. In rating analysis, the cumulative model would be appropriate, whereas a sequential model would be the obvious choice in the context of discrete survival analysis. In general, the aim is to model the ordered response y_i by covariates $\boldsymbol{x}$ in an appropriate way.

2.2.1. The Cumulative Model

The cumulative model can be derived from the idea that the observed categories represent a discrete version of an underlying (continuous) regression model. Let $\tilde{Y}$ be an underlying latent variable that follows a regression model with

$$\tilde{Y} = -\boldsymbol{x}^T\boldsymbol{\gamma} + \epsilon,$$

where ϵ is an noise variable with the distribution function F. Then the observed Y is a discrete version of the latent variable $\tilde{Y}$ determined by

$$Y = r \Leftrightarrow \gamma_{0r-1} \leq \tilde{Y} \leq \gamma_{0r}$$

It follows that

$$P(Y \leq r|\boldsymbol{x}) = P(-\boldsymbol{x}^T\boldsymbol{\gamma} + \epsilon \leq \gamma_{0r}) = P(\epsilon \leq \gamma_{0r} + \boldsymbol{x}^T\boldsymbol{\gamma}) = F(\gamma_{0r} + \boldsymbol{x}^T\boldsymbol{\gamma}).$$

Thus the cumulative model for observation i and categories r has the form

$$P(Y_i \leq r|\boldsymbol{x}_i) = F(\gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}), \quad r = 1, \ldots, k-1,$$

where $F(\eta_i)$ is a cumulative distribution function and the intercepts are defined by $-\infty = \gamma_{00} < \gamma_{01} < \cdots < \gamma_{0k} = \infty$.

The most common model from this model class is the cumulative logit model, which uses the logistic distribution. It is also called *proportional odds model* and is defined by

$$\log\left(\frac{P(Y_i \leq r|\boldsymbol{x}_i)}{P(Y_i > r|\boldsymbol{x}_i)}\right) = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma} = \eta_i, \quad \text{or}$$

$$P(Y_i \leq r) = \frac{\exp(\gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma})}{1 + \exp(\gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma})}, \quad r = 1, \ldots, k-1.$$

The name "proportional odds model" arises from the fact that the cumulative odds $P(Y_i \leq r|\boldsymbol{x}_i)/P(Y_i > r|\boldsymbol{x}_i)$ do not depend on the category. Let us assume that $\boldsymbol{x}_i$ and $\tilde{\boldsymbol{x}_i}$ are two different covariate values, then

$$\frac{P(Y_i \leq r|\boldsymbol{x}_i)/P(Y_i > r|\boldsymbol{x}_i)}{P(Y_i \leq r|\tilde{\boldsymbol{x}_i})/P(Y_i > r|\tilde{\boldsymbol{x}_i})} = \exp((\boldsymbol{x}_i - \tilde{\boldsymbol{x}_i})^T\boldsymbol{\gamma}).$$

Thus, this odds-ratio only depends on the effect $\boldsymbol{\gamma}$ and the difference between $\boldsymbol{x}_i$ and $\tilde{\boldsymbol{x}_i}$, but not on the categories. In this notation, positive $\boldsymbol{\gamma}$-values correspond with a higher probability for smaller categories.

The cumulative model can be fitted for example with the function `vglm` of the R-packages VGAM by Yee (2016), with `MRSP` of the R-package MRSP by Pößnecker (2019) or the function `polr` as part of MASS by Venables and Ripley (2002)[1]. Using alternative link functions leads to different models. The distribution $F(\eta_i) = 1 - \exp(-\exp(\eta_i))$ leads to the cumulative minimum extreme value (Gompertz) model and $F(\eta_i) = \exp(-\exp(\eta_i))$ to the cumulative maximum extreme-value model. Relaxing the proportional odds assumption results in a cumulative model with category-specific effects: $\eta_i = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}_r$.

2.2.2. The Sequential Model

If the response categories are ordered and are reached successively, the sequential model would be a natural choice. This may be the case if the higher category can be only reached if the lower category has been reached already. For example, if a woman gives birth to her third child, she has already born her first and second child. Another common case is when the categories are interpreted as time points. Then one can be only unemployed for 12 months if he or she has been unemployed for $1, 2, \ldots, 11$ months.

The underlying idea is to perform dichotomous decisions between the categories. At category 1 the decision is between category 1 and categories $\{2, \ldots, k\}$ by a dichotomous response model

$$P(Y_i = 1|\boldsymbol{x}_i) = F(\gamma_{01} + \boldsymbol{x}_i^T\boldsymbol{\gamma})$$

If $Y \geq 2$, the second dichotomous decision between category 2 and categories $\{3, \ldots, k\}$ has to be made:

$$P(Y_i = 2|Y_i \geq 2, \boldsymbol{x}_i) = F(\gamma_{02} + \boldsymbol{x}_i^T\boldsymbol{\gamma})$$

This leads to the general form of the sequential model

$$P(Y_i = r|Y_i \geq r, \boldsymbol{x}_i) = F(\gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}), \quad r = 1, \ldots, k-1.$$

The motivation using a latent variable approach can be found in Tutz (1991). In contrast to the cumulative model, the intercepts γ_{0r} can take any value and are not restricted by an ascending ordering. If desired, the parameters can also be modeled category-specific with $\gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}_r$. Based on the logistic distribution

[1]Note that polr uses the parametrization $\eta_i = \gamma_{0r} - \boldsymbol{x}_i^T\boldsymbol{\gamma}$ as default

function $F(\eta_i) = \exp(\eta_i)/(1+\exp(\eta_i))$, the *sequential logit model* is obtained by:

$$\log\left(\frac{P(Y_i = r|Y_i \geq r, \boldsymbol{x}_i)}{P(Y_i > r|Y_i \geq r, \boldsymbol{x}_i)}\right) = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma} = \eta_i, \quad \text{or}$$

$$P(Y_i = r|Y_i \geq r, \boldsymbol{x}_i) = \frac{\exp(\gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma},)}{1+\exp(\gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma},)}, \quad r = 1, \ldots, k-1.$$

Similar to the cumulative model, the odds-ratios are obtained by

$$\frac{P(Y_i = r|\boldsymbol{x}_i)/P(Y_i > r|\boldsymbol{x}_i)}{P(Y_i = r|\tilde{\boldsymbol{x}}_i)/P(Y_i > r|\tilde{\boldsymbol{x}}_i)} = \exp((\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i)^T\boldsymbol{\gamma}).$$

2.2.3. The Adjacent Categories Model

In the adjacent categories model, only adjacent categories are considered:

$$P(Y_i = r+1|Y_i \in \{r, r+1\}, \boldsymbol{x}_i) = F(\gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}), \quad r = 1, \ldots, k-1.$$

The specific model that uses the logistic distribution is the *adjacent categories logit model*

$$\log\left(\frac{P(Y_i = r+1|\boldsymbol{x}_i)}{P(Y_i = r|\boldsymbol{x}_i)}\right) = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}, \quad r = 1, \ldots, k-1.$$

$\exp(\gamma_j)$ shows the effect that compares the odds for categories $r+1$ and r when the j-th variable increase by one unit.

2.2.4. The Generalized Linear Model

Generalized linear models (GLM) are a well-known framework to handle several response distributions which was propose by Nelder and Wedderburn (1972). They are defined by a random and systematic component. The random component specifies that the response y_i's, given some covariates $\boldsymbol{x}_i$, are (conditionally) independent observations from a simple exponential family with density function

$$f(y_i|\theta_i, \phi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i)\right\},$$

where θ_i is the natural parameter of the family, ϕ_i the scale or dispersion parameter and $b(.)$ and $c(.)$ specific functions corresponding to the type of the family (Tutz, 2012). The systematic component describes how the covariates $\boldsymbol{x}_i$ take effect on the response y_i. It consists of the linear term

$$\eta_i = \boldsymbol{x}_i^T\boldsymbol{\gamma}$$

and the relationship of this linear predictor to the response y_i. The conditional expectation $\mu_i = E(y_i|\boldsymbol{x}_i)$ is determined by

$$\mu_i = h(\eta_i) = h(\boldsymbol{x}_i^T\boldsymbol{\gamma})$$

or equivalently by

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\gamma},$$

where h is the so-called response function and g the so-called link function, i.e. the inverse of h. A simple example is the logit model with the two representations:

$$\pi_i = \frac{\exp(\gamma_0 + \boldsymbol{x}_i^T \boldsymbol{\gamma})}{1 + \exp(\gamma_0 + \boldsymbol{x}_i^T \boldsymbol{\gamma})}$$

or

$$\log \frac{\pi_i}{1 - \pi_i} = \gamma_0 + \boldsymbol{x}_i^T \boldsymbol{\gamma},$$

where $h(\eta_i) = \exp(\eta_i)/(1 + \exp(\eta_i))$ and $g(\pi_i) = h^{-1}(\pi_i) = \log(\pi_i/(1 - \pi_i))$. It can be shown the corresponding distribution belongs to a simple exponential family. Other examples are the normal linear regression or the Poisson regression. There are several extensions relaxing the assumption of the linear term by semi or nonparametric methods or the variance assumption by quasi-likelihood approaches.

Tutz (2012) shows that the cumulative and sequential model are specific cases of a multivariate GLM. In cumulative models the probability for a certain category $\pi_{ir} = P(Y_i = r|\boldsymbol{x}_i)$ can be computed by

$$\pi_{ir} = F(\gamma_{0r} + \boldsymbol{x}_i^T \boldsymbol{\gamma}) - F(\gamma_{0r-1} + \boldsymbol{x}_i^T \boldsymbol{\gamma})$$

So that the response function h_r is defined as

$$h_r(\eta_{i1}, \ldots, \eta_{ik-1}) = F(\eta_{ir}) - F(\eta_{ir-1})$$

In the sequential model π_{ir} is calculated by

$$\pi_{ir} = P(Y_i = r|\boldsymbol{x}_i) = F(\gamma_{0r} + \boldsymbol{x}_i^T \boldsymbol{\gamma}) \prod_{r=1}^{k-1} (1 - F(\gamma_{0r} + \boldsymbol{x}_i^T \boldsymbol{\gamma})),$$

which leads to

$$h_r(\eta_{i1}, \ldots, \eta_{ik-1}) = F(\gamma_{0r} + \boldsymbol{x}_i \boldsymbol{\gamma}) \prod_{r=1}^{k-1} (1 - F(\gamma_{0r} + \boldsymbol{x}_i^T \boldsymbol{\gamma})).$$

Thus one has to distinguish between the distribution function F and the response function h of a GLM. However, if one looks only at a certain decision, as in the logistic regression model, e.g. remaining in category r or not, the response function can be used to determine several kinds of logistic regression models. In chapter 4 I follow Tutz and Schmid (2016) and use the terminus "response function" to specify different hazards. But one has to keep in mind that these response functions focus on a binary case rather than on the whole process described by a sequential model with a different response function.

3. Modelling Heterogeneity in Surveys

In most social surveys, attitudes and opinions are measured on an ordinal Likert scale so that respondents $i \in 1, ..., n$ are asked to choose a certain category $r \in 1, ..., k$. This choice is a very subjective process and may be influenced by various factors, such as the own preference or uncertainty due to the lack of information, time and self-confidence. When we think of the positions political parties offer, the parties often blur their positions so that ordinary citizens may have difficulties in perceiving clear stances.

There are several approaches on how to deal with heterogeneity in surveys. They differ not only in their method but also in the kind of heterogeneity which they can examine. Some approaches link heterogeneity with uncertainty, which can be seen as an interpretation of the heterogeneous structure. For example, it could be interpreted that a respondent, who is uncertain about his or her choice, is more likely to choose the middle category, or randomly chooses a category or even refuses to respond. Depending on these different concepts, the models can be adjusted for these scenarios.

The specific uncertainty structures can be also found to some extend in literature referring to "response styles" or "response bias". Vaerenbergh and Thomas (2013) and Baumgartner and Steenkamp (2001) give an overview of several possible response styles and explanations. One popular response style is the tendency to select the middle category. This may lead to a reduction of variance. By contrast, the extreme response style is characterized by choosing the extreme categories, which may be interpreted, for instance, as a "reflection of rigidity" (Baumgartner and Steenkamp, 2001, p. 145). Other mentioned response styles are for example the tendency to avoid the highest and the lowest response category or the tendency to respond randomly.

Response bias in a narrow sense often refers to the definition by Paulhus (1991, p. 17) who defines "response bias" as "a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content". Thus according to this definition response bias is related to repeated measurements or several items. However, if focusing on the decision process it seems reasonable to assume that specific response or rather heterogeneous structures can be also found looking at a single response item. Obviously, in this case it is only possible to make a statement about the decision process of this specific item and model its possible heterogeneous structure.

In this section I show how mixture models contribute to the modelling of heterogeneity in surveys. In general, a mixture model is defined in such a way that the observed density f of the dependent variable can be represented by a combination of a finite set of densities f_g, so that

$$f = \sum_{g=1}^{M} \pi_g f_g,$$

where the densities are weighted by π_g which denotes the mixture proportion or weight. It ranges from 0 to 1 and the sum over all mixture weights $g = 1, \ldots, M$

is equal to one. Traditional mixture model for modelling heterogeneous structures in surveys consists of two components: One component for the preference structure and the other component for the uncertainty structure.

Since all mixture approaches that are discussed here, refer to the CUB model, I describe it in the next section in more details. The approaches can be roughly divided into three groups. Approaches focus on a better modelling of the first component, the preference structure (see CUP and CUBE), the second component, the uncertainty structure (see BetaBin, VCUB and CAUB), or other techniques, which use more than two components or extend the concept in another way (geCUB, LC-CUB, Non-linear-CUB, H-CUB, RCUB). Finally, some non-mixture approaches are presented.

3.1. The CUB Model

The CUB model (formerly called MUB model) is introduced by D'Elia and Piccolo (2005) and further described by Piccolo and D'Elia (2008), Iannario and Piccolo (2010), Piccolo (2006), Piccolo (2003), Iannario and Piccolo (2012), Iannario and Piccolo (2016a), Iannario and Piccolo (2016b), Piccolo et al. (2018) and Piccolo and Simone (2019a). It is defined as a combination of a discrete uniform and a (shifted) binomial distribution:

$$P(R_i = r) = \pi_i P_M(Y_i = r) + (1 - \pi_i) P_U(U_i = r), \tag{3.1}$$

where R_i is the observed response of an individual i with the values $1, \ldots, k$. The first component P_M captures the preference structure and is defined by a shifted binomial distribution

$$p_r^M(\xi_i) = \binom{k-1}{r-1} \xi_i^{k-r} (1 - \xi_i)^{r-1}, \quad r \in \{1, \ldots, k\}.$$

The distribution is determined by the parameter ξ and shifted so that the support is $\{1, \ldots, k\}$ instead of the usual support that includes zero. ξ_i may be linked to covariates $\boldsymbol{x}_i^T$ by the logit link so that

$$\text{logit}\,(\xi_i) \;=\; \boldsymbol{x}_i^T \boldsymbol{\gamma}\,; \qquad\qquad i = 1, 2, \ldots, n\,.$$

This specification can be used to explain the preference by individual characteristics such as gender or age. The (shifted) binomial distribution is motivated by the idea that respondents pairwise compare single items. Thus, choosing the third category out of $k = 6$ categories implies that the respondent rejects category 1 and 2 as well as category 4, 5, and 6. Since one comparison would be Bernoulli distributed, all comparisons are naturally binomially distributed. The binomial distribution has only one parameter so that the mean and variance of the distribution depend on ξ and are not independent from each other.

The heterogenity or in this context usually called "uncertainty" is included by the second component P_U. Iannario and Piccolo (2010) argue that the uncertainty component reflects the "complete indifference" of a persons resulting

in a discrete uniform distribution with the form

$$p_r^U = 1/k, \quad r \in \{1, \ldots, k\}.$$

The strength or propensity of this uncertainty is determined by the size of the mixture weights $1 - \pi_i$. The mixture weight π may be also linked to covariates by

$$\text{logit}\,(\pi_i) \;=\; \boldsymbol{z}_i^T \boldsymbol{\beta}\,; \qquad\qquad i = 1, 2, \ldots, n\,.$$

Then the propensity of each respondent to choose a random category can be modelled by individual characteristics as in the case of ξ. π can take values between zero and one. If $\pi = 1$ no uncertainty component is present and if $\pi = 0$ no preference component is used. The covariates in $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ may be identically, overlap or be completely different. Alternative link functions representing a one-to-one mapping $\mathbb{R}^p \leftrightarrow [0, 1]$ between parameters and covariates would be also possible, but are usually not used in these models.

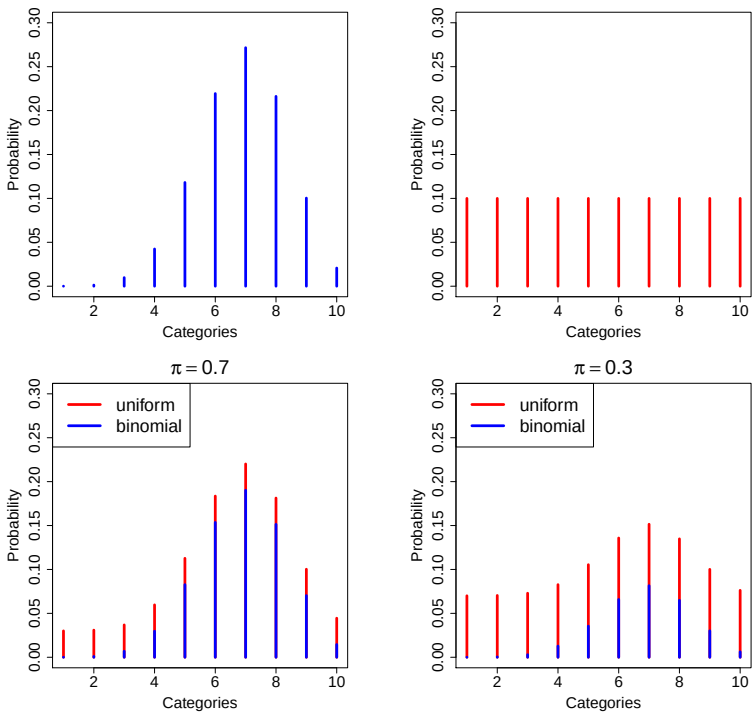


FIGURE 3.1.: *Visualization of the mixture in CUB models*

Figure 3.1 illustrates the composition of the uniform and binomial distribution for two possible mixture distributions. The top panel displays the single distributions for $k = 10$ categories: On the left the shifted binomial distribution and on the right the discrete uniform distribution. Depending on π, two different mixture distributions are computed and displayed in the second row of Figure 3.1. On the left, $\pi = 0.7$ corresponds with a higher importance of the

preference structure than on the right where $\pi = 0.3$. Thus, the mixture weight has a large impact on the shape of the mixture distribution, whereas the single components stay the same. The shape of the binomial distribution depends on ξ and the used covariates. However, the uniform distribution measuring the uncertainty only depends on the number of categories.

3.2. The CUP Model

Tutz et al. (2017) proposed to replace the binomial distribution of the preference structure with a more flexible ordinal cumulative model. The CUP model, as Combination of Uniform and Preference structure, is defined as

$$P(R_i = r|\boldsymbol{x}_i) = \pi_i P_M(Y_i = r|\boldsymbol{x}_i) + (1 - \pi_i)P_U(U_i = r), \tag{3.2}$$

but contrary to the CUB model, P_M can be any ordinal model. We use the proportional odds model

$$\log\left(\frac{P(Y_i \leq r|\boldsymbol{x}_i)}{P(Y_i > r|\boldsymbol{x}_i)}\right) = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}, \quad r = 1, \ldots, k-1,$$

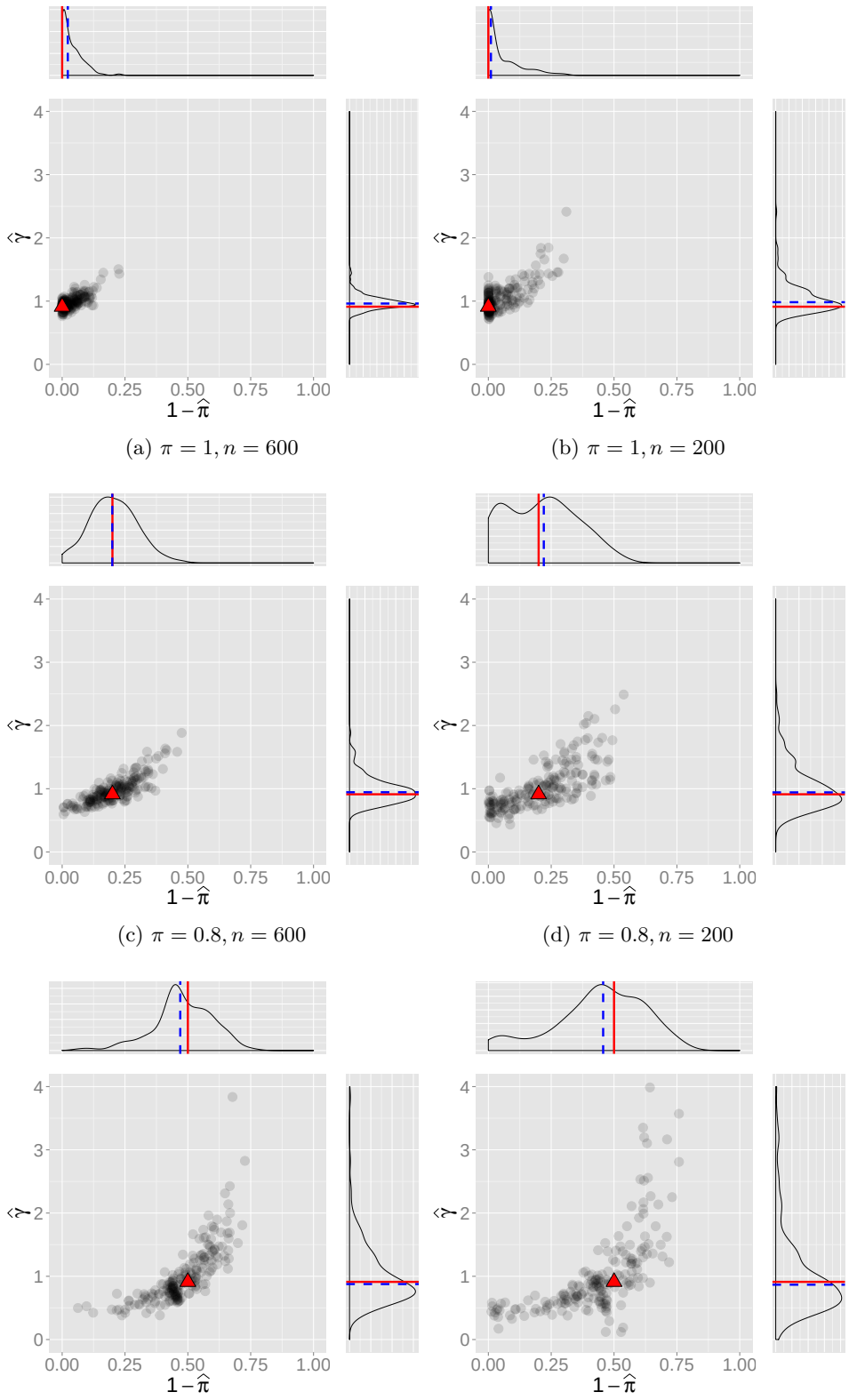
and the adjacent categories model

$$\log\left(\frac{P(Y_i = r+1|\boldsymbol{x}_i)}{P(Y_i = r|\boldsymbol{x}_i)}\right) = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}, \quad r = 1, \ldots, k-1,$$

which are explained in Section 2.2.1 and 2.2.3 in detail. The fundamental advantage of this novel approach is a more flexible model than the CUB model. Using flexible intercepts $\gamma_{01}, \ldots, \gamma_{0k-1}$ leads to more parameters but also a better fit to the data. Tutz et al. (2017) applied the CUP and CUB model to several real data examples. The CUP model outperforms the CUB model clearly in two of three applications by deviance, AIC, and BIC. In the third application, the CUP and CUB models are quite similar.

Figure 3.2 shows the stability analysis for the CUP model. 200 data sets for different settings are generated with $\boldsymbol{\gamma}_0 = (-2.391, -1.221, -0.259)$ and $\gamma = 0.912$ for one continuous variable. For each $\pi = 1, 0.8, 0.5$, two sample sizes, $n = 200$ and $n = 600$, are used. Then, the CUP model was applied to each of the 200 generated data sets. Each point in Figure 3.2 corresponds to one estimated model. The vertical axis displays the estimated γ coefficient with its marginal distribution. The horizontal axis shows $1 - \pi$ with its marginal distribution. The blue dashed line corresponds to the median value of the distribution. The red triangle indicates the simulated values of γ and $1 - \pi$, and the solid red line the corresponding values.

In general, the true parameters could be estimated quite well since the blue dashed line and the red solid line are always close to each other. However, if the sample size is only 200 as at the right column of Figure 3.2, the variance is greater than on the left-hand side with $n = 600$. The π value varies from 1 in the first row of Figure 3.2 over 0.8 to 0.5 in the last row. The larger $1 - \pi$ the more variance is noticeable.



(a) $\pi = 1, n = 600$

(b) $\pi = 1, n = 200$

(c) $\pi = 0.8, n = 600$

(d) $\pi = 0.8, n = 200$

(e) $\pi = 0.5, n = 600$

(f) $\pi = 0.5, n = 200$

FIGURE 3.2.: *Simulation of the CUP model with varying mixture weight* $\pi = 1, 0.8, 0.5$ *and number of observations* $n = 200, 600$

3.3. The BetaBin Model

The CUB and CUP differ in their preference component, but both use the uniform distribution to model the uncertainty structure. Tutz and Schneider (2019) propose to use a more flexible uncertainty component, which can account for the tendency to the middle and extreme categories. This is often found in applications, for example, when persons judge their probabilities of surviving the next 10 years as analyzed by de Bruin and Carman (2012). It may be also more appropriate than assuming that people who are uncertain choose a category at random. This is achieved by replacing the uniform distribution by a specific beta-binomial distribution with a certain constrain so that the mean is always in the middle of support. The random variable U of the uncertainty component with support $\{1, \ldots, k\}$ follows a beta-binomial distribution with

$$f(u) = \begin{cases} \binom{k-1}{u-1} \frac{B(a+u-1, b+k-u+1)}{B(a,b)} & u \in \{1, \ldots, k\} \\ 0 & \text{otherwise,} \end{cases}$$

where $a, b > 0$ are the parameters of the beta-binomial distribution[1]. $B(a, b)$ is the beta function defined as

$$B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt.$$

The specific response styles are obtained by imposing the restriction $a = b$, which lead to $\mu = 0.5$.

Figure 3.3 illustrates the different shapes of the restricted beta-binomial dis-

[1]Note that in Tutz and Schneider (2019) α and β are the parameters of the beta-binomial distribution. Here, this is changed to a and b for readability and consistency

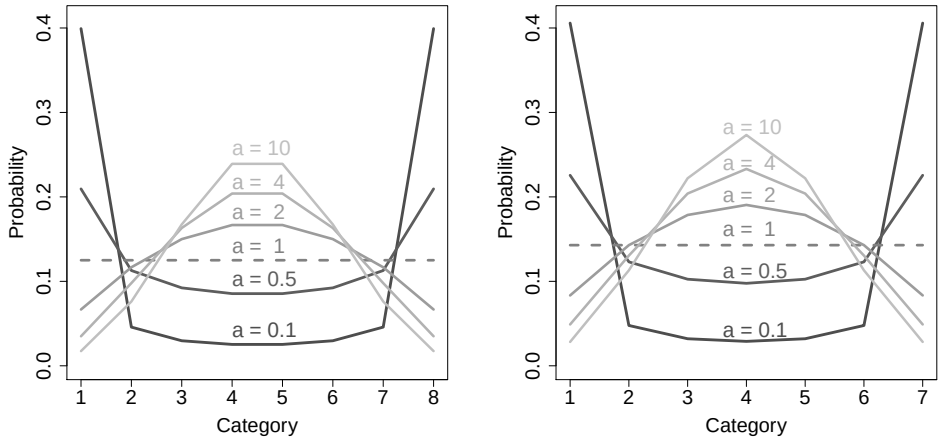


FIGURE 3.3.: *Probability mass on categories for various values of a for $k = 8$ categories (left panel) and $k = 7$ categories (right panel). For illustration the probability of each category are connected by lines. Source: According to Tutz and Schneider (2019, p. 1586)*

tribution for even categories on the left and odd categories on the right. a values smaller than one correspond with a tendency to the extreme categories, while a values larger than one indicate a tendency to the middle categories. For $a = 0$, one obtains the uniform distribution as in the CUP and CUB models.

The parameter a is linked to covariates $\boldsymbol{w}_i$ by

$$a = \exp(\boldsymbol{w}_i^T \boldsymbol{\alpha}) = \exp(\alpha_0) \exp(\alpha_1)^{w_{i1}} ..... \exp(\alpha_m)^{w_{im}},$$

where a is the parameter of the restricted beta-binomial distribution and α_j gives the effect of the j-th covariate linked with exp() to a. Thus, the shape of the beta-binomial distribution depends on individual covariates.

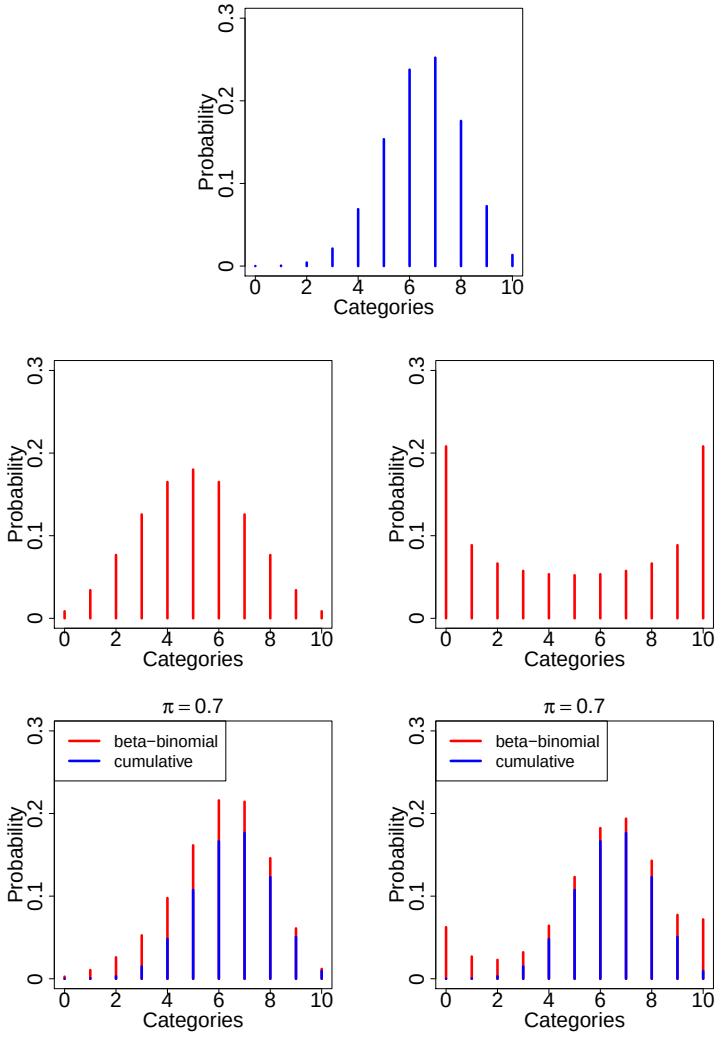


FIGURE 3.4.: *Visualization of the mixture in BetaBin models*

Figure 3.4 visualizes possible combinations of the preference structure with

the response styles captured by the restricted beta-binomial distribution for a 11-point scale, running from zero to ten. The first line present a possible probability distribution of a cumulative model and the second line two possible response style distributions: On the left the tendency to the middle categories by $a = 6$ and on the right the tendency to extreme categories by $a = 0.4$ is displayed. Then, both response styles are combined with the preference component with the weight $\pi = 0.7$ so that the preference structure has a higher strength than the uncertainty component. According to the response styles, two different mixture distributions are obtained.

Mauerer and Schneider (2019a) use the BetaBin model to examine party placements on the immigration issue in the 2017 German Election. In the respective national election study, approximately 2000 participants are asked to state where they preceive the positions of the German parties CDU, CSU, SPD, FDP, Greens, Left Party and AfD on immigration on an eleven point scale from 1 "Immigration should be facilitated" to 11 "Immigration should be restricted". Thus, the party positions are not determined by their manifesto but by individual perceptions. Figure 3.5 displays the estimated uncertainty propensities $(1 - \hat{\pi})$, which clearly depends on the party. The dotted lines correspond with the 2.5% and 97.5% bootstrap quantiles. The weakest uncertainty with 0.05 was found for the AfD. The largest uncertainties were detected for the CDU and the FDP. In both cases, the respondents seem to have more difficulties in placing the parties. However, the performance measures indicate that the mixture models exhibit lower AIC values for all parties as compared to a cumulative model without an uncertainty component. Thus, the BetaBin model improves the understanding and model fit compared to a simple cumulative model.

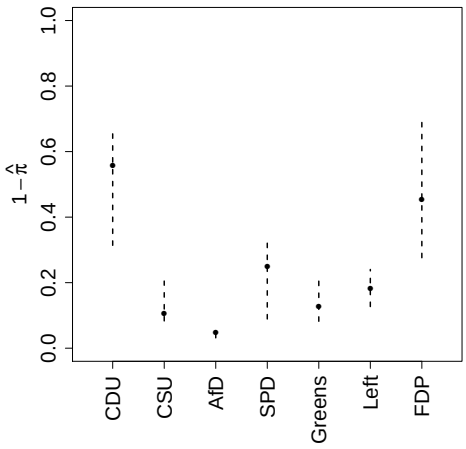


FIGURE 3.5.: *2017 German Election: Uncertainty propensity for different parties. Source: Mauerer and Schneider (2019a, p. 132)*

Mauerer and Schneider (2019b) developed a two-stage voter choice model which uses the BetaBin model to account for uncertainty. Spacial voting approaches assume that citizens vote for parties or candidates that offer positions

that coincide with their preferences. Thus voter choice models frequently use ordinal policy scales such as the liberal-conservative scale to determine the coincidence between voter and perceived party position and use it as a predictor in the model. The closer the positions the more likely a citizen should vote for that party or candidate. For example in the 2016 US presidential election study used by Mauerer and Schneider (2019b) the voters specify their own preference and the perceived candidate position on a 7-point Likert scale from "liberal" (1) to "conservative" (7). The BetaBin model is used to model this placement process and account for possible uncertainty. Since uncertainty can be seen in this context as something which mask or blur the real position, the estimates of the preference component of the BetaBin model ($P_M(Y_i = r)$) are used to estimate adjusted placements. They are obtained for each observation i by first calculating the probabilities of selecting a specific category $r = 1, \ldots, k$ by the differences of two cumulative probabilities

$$P(Y_i = r|\boldsymbol{x}_i) = P(Y_i \leq r|\boldsymbol{x}_i) - P(Y_i \leq r - 1|\boldsymbol{x}_i)$$

and then choosing the category with the highest probability as the most likely placement. Figure 3.6 illustrates how the distribution of self-placement and perceived republican candidate placement on the liberal-conservative scale changes. Generally the variance seems to decrease since the distribution of the self-placement seems to be located more to the middle categories and the distribution of the Republican candidate position concentrates on fewer extreme categories.

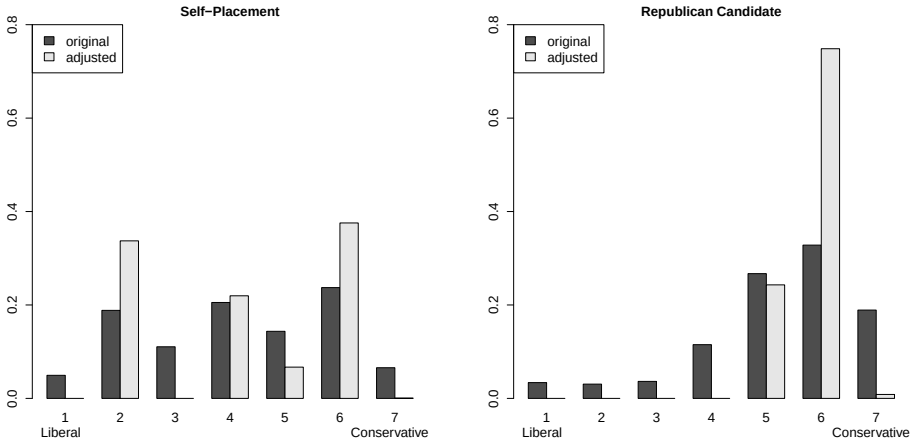


FIGURE 3.6.: *Distributions of self-placement (on the left) and republican candidate placement (on the right) for observed and adjusted placements on the liberal-conservative scale.*

The adjusted placements are then used to calculate the proximity between voter and perceived candidate position and is included as a predictor in a traditional voter choice model. Mauerer and Schneider (2019b) showed that this model performed better than a voter choice model without accounting for uncertainty measured by AIC.

3.4. The CAUB Model

The CAUB model by Simone and Tutz (2018), as combination of adjusted uniform and (shifted) binomial variables, is an approach to improve the flexibility of the uncertainty component of the CUB model, as the BetaBin model is for the CUP model. The main idea is to replace the uniform distribution with a discretized beta distribution. This distribution was introduced by Ursino (2014) and is obtained by dividing the metric range of the beta distribution (see section 3.3) into k equally spaced intervals and integrate over them so that the probability of a discrete response variable D is given by

$$P(D = r|a, b) = P\left(\frac{r-1}{k} \leq X \leq \frac{r}{k}|a, b\right), r = 1, \ldots, k.$$

Imposing the restriction $a = b$ leads to a symmetric version of the discretized beta distribution similar to the restricted beta-binomial distribution of the BetaBin model. However, this specific beta distribution seems to be able to capture a stronger tendency to middle categories. Figure 3.7 displays the probability distributions of both distributions.

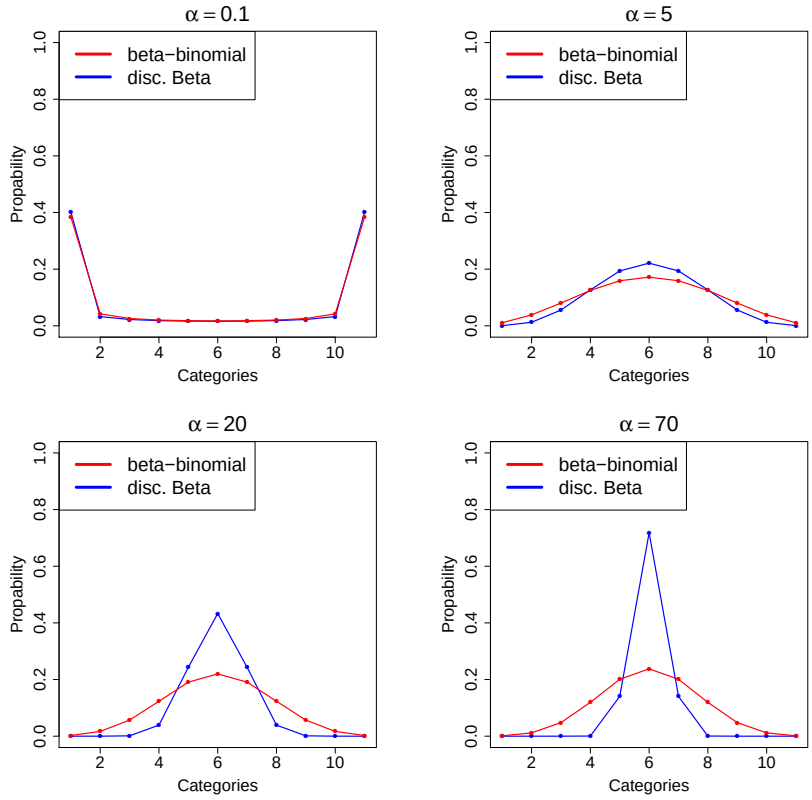


FIGURE 3.7.: *Probability distribution of the restricted beta-binomial distribution (red) and the restricted discretized beta distribution (blue) with different a values*

While the difference is very low for small a values, the probability mass of the beta distribution is higher around the middle categories for larger a

values. The restricted beta distribution converges to a point mass distribution for large a values, while the restricted beta-binomial distribution of the Betabin model converges to a (shifted) binomial distribution so that its probability mass is approximately not as concentrated in the middle categories as in the discretized beta distribution. But, as long as the a-estimates are not too large, the probability masses are quite similar.

One should also keep into mind that the number of response categories influences at which a value the point mass distribution is reached. Figure 3.8 illustrates this case for $k = 5$ and $k = 11$ categories. While in the case of 5 categories $a = 70$ is sufficient, $a = 400$ is needed to reach a similar effect for 11 categories. The case of $a = 70$ for 11 categories can be found in Figure 3.7 where there is a strong tendency to the middle categories but the categories close to the middle category have still a considerable probability mass.

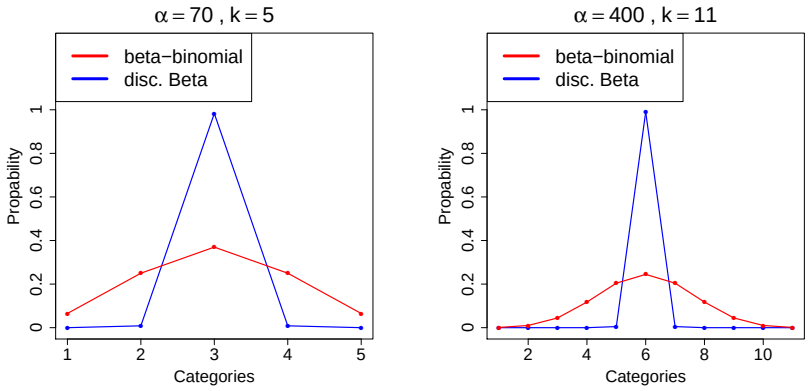


FIGURE 3.8.: *Impact of the number of categories on the convergence of the restricted beta distribution to a point mass distribution*

3.5. Further Extensions of the CUB Model

There are some further mixture approaches in the CUB family. The CUBE model proposed by Iannario (2014) and further developed by Piccolo (2015) stands for the combination of a uniform and a (shifted) beta-binomial random variable. The idea is similar to the CUP model to improve the modelling of the preference component. However, instead of using any ordinal model for the preference structure as in the CUP model, they propose to use the beta-binomial distribution instead of the binomial distribution as in the CUB model. Since the binomial distribution is restricted by one parameter for the mean and variance, the beta-binomial distribution is defined by two parameters. In terms of flexibility, the CUBE model can be ranked between the CUB and the CUP models.

Gottard et al. (2016) propose several other uncertainty distributions instead of the uniform distribution in the CUB model. V-CUB is the acronym for Varying uncertainty in CUB models. They suggest the "trimmed uniform distribution", the "Left/right bounded Uniform distribution", the "triangular distri-

bution" and the "symmetric parabolic distribution". In this way, more specific uncertainty structures can be modelled. On the other hand, each distribution has to be chosen in advance and the shape of the distribution does not depend on covariates. Both aspects are a drawback compared to the BetaBin model, as an extension for CUP model, or the CAUB model, as an extension of the CUB model.

Manisera and Zuccolotto (2014) propose a nonlinear-CUB (NL-CUB) model. They use "transition probabilities" to account for the case that respondents may not consider all ratings on the Likert scale in the same way. For example, considering a four point Likert scale, running from "strong disagree", "disagree", "agree" to "strongly agree". It may be easier for respondents to move from "strongly disagree" to "disagree" than from "disagree" to "agree". The transition probabilities would be higher for the second transition as for the first one. The CUB model can be seen as a special case where the transition probabilities are identical for all transitions. This is called linear transition in the framework of the NL-CUB.

The latent class CUB approach (LC-CUB) by Grilli et al. (2014) extend the number of components of the CUB model. Instead of using one binomial distribution for the preference structure, several binomial distributions are considered. The restricted LC-CUB is given by

$$P(R_i = r, \boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{g}) = \pi \sum_{h=1}^{H} g_h p_r^M(\xi_h) + (1-\pi)\frac{1}{m},$$

where H is the number of (shifted) binomial distributions and g_h the latent class probabilities. The choice of H is not an easy task and identification issues arise.

The generralized CUB models (geCUB) is introduced by Iannario (2012b) as a CUB model with shelter choice and further developed by Iannario and Piccolo (2016b) by including covariates for the shelter choice. This model softens the discrimination between preference and uncertainty component. The idea is to include a so called "shelter effect" to account for a point mass for one specific response category in the sense of inflated models such as the hurdle model or the negative binomial model, which is described by Hilbe (2011). Iannario (2012b) defines the model as

$$P(R_i = r, \xi, \boldsymbol{\pi}) = \pi_1 p_r^M(\xi_i) + \pi_2 P_U(U_i = r) + (1 - \pi_1 - \pi_2) D_r,$$

where

$$D_r = \begin{cases} 1, r = c \\ 0, \text{otherwise.} \end{cases}$$

The probability mass of the third new component D_r is concentrated at $r = c$. If $\pi_1 + \pi_2 = 1$, the model reduces to the traditional CUB model. The parameters π, ξ and δ may be linked to covariates by the logit link. The geCUB model can be also seen as a possibility to model response styles, similar to the BetaBin or the CAUB model by using the shelter effect for the middle category. One important difference is that the BetaBin model and for limited α the CAUB

model rather fit the tendency to the middle category so that also categories close to the middle category have a probability mass larger than zero than focus on one specific category only. Furthermore, the geCUB model can only account for one category, which has to be chosen in advance, but cannot model both the response style to the middle and extreme categories by covariates.

The CUSH model proposed by Capecchi and Piccolo (2017) as an combination of a discrete uniform and a shelter effect, can be seen as a specific geCUB model with no binomial distribution leading to a model with maximum uncertainty. However, the interpretation of a model without a preference component seems to be rather difficult.

Iannario (2012a) proposed using random effects in the preference component of the CUB model. This extension is called hierarchical CUB model (HCUB). Simone et al. (2019) use random effects b_i for the uncertainty propensity to connect multiple items on an individual basis, which they called random-effect CUB model (RCUB). The mixture weight π for item j is defined as

$$\text{logit}\left(\pi_i^{(j)}\right) = \boldsymbol{z}_i^T \boldsymbol{\beta}_j + b_i\,; \qquad i = 1, 2, \ldots, n,\ j = 1, ..., K^*.$$

Thus the mixture weight is determined by covariates $\boldsymbol{z}_i^T$ and an individual effect $b_i \sim N(0, \sigma^2)$, which is the same for one individual i over all j items. The main open question is in the assumption of the RCUB model, namely that the responses are conditionally independent given the individual's uncertainty so that only "weak association structure" (Simone et al., 2019, p. 3) in the responses should be considered. If the different items are too much correlated, as it may be the case of multiple items on the same topic, this assumption may be violated.

Most recent discussions of CUB model and various possible extensions can be found in Tutz (2019), Kenett (2019), Proietti (2019), Grilli and Rampichini (2019), Colombi et al. (2019) and Piccolo and Simone (2019b).

3.6. Some Non-Mixture Approaches to Model Heterogeneity in Surveys

One popular approach, especially in political science, is based on the heteroscedastic model (see Harvey, 1976). The idea is to use a relative simple logit/probit or linear model and explicitly model the variance of the error term with covariates:

$$\begin{aligned} y_i &= \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2) \\ \sigma^2 &= \exp(\boldsymbol{z}_i^T \boldsymbol{\gamma}) \end{aligned}$$

For example, Alvarez and Brehm (1995), Harbers et al. (2013) and de Vries and Steenbergen (2013) follow this approach. Further work was done by Tutz (2018) who link the heterogeneous choice model with the varying-coefficients model. These heteroscedastic models are often not optimized for ordinal response and

they cannot account for certain heterogeneity structures such as the tendency to middle or extreme categories.

Other researchers point out that missing values or "don't know"-categories should be used to evaluate the uncertainty. Both Bartels (1986) and Rozenas (2013) argue that missing data in the response is caused by the uncertainty of the respondents. Another way to allow for "undecided" persons is developed by Plass et al. (2015). They propose to allow indifferent respondents to choose more than one nominal category, which leads to much more possible cases. They use an election study to show their approach and claim that it can be easily extended for ordered categories. Other approaches rely on additional questions to examine how uncertain a respondent is about his or her choice. However, such questions are very rare in questionnaires. It depends on the modeling philosophy if such information should be used to measure uncertainty. Since the data generating process for missing data is usually unclear, missing data can result from various reasons. When missing completely at random can be assumed, it may be also appropriate to exclude missing data and "don't knows" from the analysis. For an overview on missing data in statistical analysis see Little and Rubin (2002).

Tutz and Berger (2016) propose to add additional parameters $\boldsymbol{z}_i^T\gamma$ to a generalized linear model. They use the adjacent categories model to distinguish between "content-related effects" and "heterogeneity in response styles". The first one can be considered as a "preference structure" and later as "uncertainty" in mixture models. For odd categories k the model is defined by:

$$log\frac{P(Y_i = r+1|\boldsymbol{x}_i, \boldsymbol{z}_i)}{P(Y_i = r|\boldsymbol{x}_i, \boldsymbol{z}_i)} = \beta_{0r} + \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\gamma, r = 1, \ldots, m-1,$$
$$log\frac{P(Y_i = r+1|\boldsymbol{x}_i, \boldsymbol{z}_i)}{P(Y_i = r|\boldsymbol{x}_i, \boldsymbol{z}_i)} = \beta_{0r} + \boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{z}_i^T\gamma, r = m, \ldots, k-1,$$

where $m = [k/2]+1$ is the middle category. The content-related effects is represented by $\beta_{0r} + \boldsymbol{x}_i^T\boldsymbol{\beta}$, which are the effects in a traditional adjacent categories model. In contrast, $\boldsymbol{z}_i^T\boldsymbol{\gamma}$ capture the response style. For $\boldsymbol{z}_i^T\boldsymbol{\gamma} \to \infty$ follows $P(Y_i = m|\boldsymbol{x}_i, \boldsymbol{z}_i) \to 1$, which is a strong tendency to the middle category. If $\boldsymbol{z}_i^T\boldsymbol{\gamma} \to -\infty$ one obtains for $P(Y_i = 2|\boldsymbol{x}_i, \boldsymbol{z}_i), \ldots, P(Y_i = k-1|\boldsymbol{x}_i, \boldsymbol{z}_i) \to 0$ and therefore, a strong tendency for extreme categories.

Tutz et al. (2018) propose a parametrization to include response styles in the Partial Credit Model (PCM). Latent trait models (as the PCM) are characterized by person-specific parameters and item difficulty. Other approaches within the framework of item response theory are carried out by Bolt and Johnson (2009), Bolt and Newton (2011) and Johnson (2003). Response styles can be also included in latent class approaches (see e.g. Moors, 2004; Kankaraš and Moors, 2009; Moors, 2009; Rosmalen et al., 2010). Tree type approaches typically assume a nested structure where first a decision about the direction of the response and then about the strength is obtained. Examples are de Boeck and Partchev (2012), Jeon and de Boeck (2016), and Böckenholt (2012).

4. Discrete Survival Analysis

In survival analysis, the time until an event occurs is analyzed. In the following, I focus on discrete survival analysis, where the time is measured on a discrete scale. An overview of such methods can be found in Tutz and Schmid (2016). Earlier introductions are given by Hamerle and Tutz (1989) and Allison (1982). Broström (2012) focuses on the use of the software R for event history analysis in general, and Möst (2014) uses regularization techniques in certain discrete survival models. Willett and Singer (1993a) show how to apply discrete survival analysis for analyzing suicidal ideation and depression, and Willett and Singer (1993b) for analyzing career paths. In many application cases, the time is measured at discrete points in time so that a discrete survival model is the natural choice.

One central issue in survival analysis is censoring, capturing the fact that not all observations are available during the whole studied period. Right-censoring refers to observations who drop out without observing an event. Thus, as long as they are observed, no event was taken place and they might experience the event after dropping out, but not necessarily. Left-censoring means that the entry to the observation period is not known, but the end is observed. The models used here account only for right-censoring, which happens more frequently than left-censoring.

In discrete survival analysis, we can use the available cases at each point in time to calculate the probability of event occurring. The discrete hazard is defined as the probability that an event occurs at time T, given that time T is reached conditional on some covariables $\boldsymbol{x}$. For observation i we obtain:

$$\lambda(t|\boldsymbol{x}_i) = P(T_i = t|T_i \geq t, \boldsymbol{x}_i) = h(\gamma_{0t} + \boldsymbol{x}_i^T\boldsymbol{\gamma}).$$

$h()$ can be various response functions and $g()$ the corresponding link function leading to

$$g(\lambda(t|\boldsymbol{x}_i)) = \gamma_{0t} + \boldsymbol{x}_i^T\boldsymbol{\gamma}.$$

Depending on the choice of the link function, several models are possible. The logit link leads to a logistic discrete hazard model with

$$\log\left(\frac{P(Y_i = r|Y_i \geq r, \boldsymbol{x}_i)}{P(Y_i > r|Y_i \geq r, \boldsymbol{x}_i)}\right) = \log\left(\frac{P(Y_i = r|Y_i \geq r, \boldsymbol{x}_i)}{1 - P(Y_i = r|Y_i \geq r, \boldsymbol{x}_i)}\right) \quad = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}$$

The group proportional hazard model is obtained by using the complementary log-log link:

$$\log(-\log(P(Y_i > r|Y_i \geq r, \boldsymbol{x}_i))) = \gamma_{0t} + \boldsymbol{x}_i^T\boldsymbol{\gamma}.$$

Other well-known models are the probit model with probit link, the gumbel model with log-log link, and the exponential model with log link (see Tutz and

Schmid, 2016).

The survival function is the probability that the event occurs later than at time t, which is the product of $1 - \lambda$ until t:

$$S(t|\boldsymbol{x}_i) = P(T_i > t|\boldsymbol{x}_i) = \prod_{s=1}^{t}(1 - \lambda(s|\boldsymbol{x}_i)) = \prod_{s=1}^{t}(1 - h(\gamma_{0s} + \boldsymbol{x}_i^T\boldsymbol{\gamma}))$$

The unconditional probability of an event is calculated by the product of surviving $t - 1$ time points and experiencing an event at time point t denoted by

$$P(T_i = t|\boldsymbol{x}_i) = \lambda(t|\boldsymbol{x}_i)\prod_{s=1}^{t-1}(1 - \lambda(s|\boldsymbol{x}_i))$$

Heterogeneity arises from the fact that there are individuals or groups with very different survival curves. Figure 4.1 shows the mixture of two groups with different constant hazards.

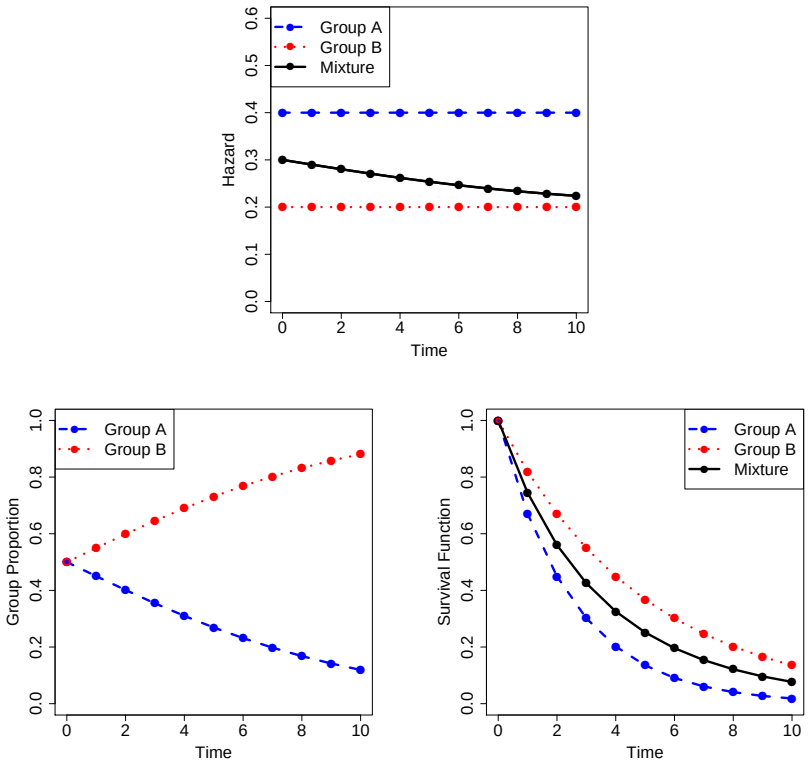


FIGURE 4.1.: *Illustration of the Mixture of two constant hazards*

Group A is characterized by a constant hazard of 0.4, and group B by a constant hazard of 0.2. At the beginning, that is, at $t = 0$, both populations are of the same size. Thus, the proportion for each group is 0.5. Over time more events take place in group A than group B because of the higher probability of

an event in group A. Thus, the proportion of the two groups change dramatically – from an equal weight of 0.5 at $t = 0$ to nearly 0.1 for group A and to 0.9 for group B at $t = 10$. The different hazards also result in different survival curves. The survival function of group B (with higher hazards) is from $t = 1$ on always below of the survival function of group A.

Neglecting the two groups leads to a very different non-constant hazard displayed by the bold line. At the beginning, this "mixture"-hazard is exact in the middle of the two constant hazards because of the equal size of the groups. Since the proportion of the groups changes over time, the estimated "mixture"-hazard changes as well and approximates the hazard of group B. At the end, the estimated "mixture" survival curve is between the survival curve of the true groups A and B. Thus, if there are groups with different survival functions, ignoring this fact exhibits a great impact on the results. I focus on a specific kind of subgroups, known as cure models. Here, one group consists of so-called long-term survivors with a constant survival function of 1 over time, while only the other group is under risk of the event. After describing the discrete cure model, some related approaches are introduced.

4.1. The Discrete Cure Model

Cure models usually consist of two unobserved sub-population – one under risk and one characterized as long-term survivors. An overview of cure models for mainly continuous time is given by Amico and Keilegom (2018) and Maller and Zhou (1996). Proportional hazards cure models were proposed by Kuk and Chen (1992) and Sy and Taylor (2000). Muthén and Masyn (2005) consider discrete survival mixtures. An application with long-term survivors in discrete data can be found in Steele (2003) or Tutz and Schmid (2016).

The Cure-model is defined as a mixture of two survival functions. S_1 denotes the survival function for the non-cured population and $S_2 = 1$ the survival function for the cured or long-term survivors:

$$S(t|\boldsymbol{x}_i) = \pi(\boldsymbol{z}_i) \underbrace{S_1(t|\boldsymbol{x}_i)}_{\text{non-cured}} + (1 - \pi(\boldsymbol{z}_i)) \cdot \underbrace{1}_{\text{cured}}, \tag{4.1}$$

where

$$S_1(t|\boldsymbol{x}_i) = \prod_{s=1}^{t} (1 - h(\gamma_{0s} + \boldsymbol{x}_i^T \boldsymbol{\gamma}))$$

and

$$\pi(\boldsymbol{z}_i) = \frac{\exp(\boldsymbol{z}_i^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{z}_i^T \boldsymbol{\beta})}.$$

For each observation the weight $\pi(\boldsymbol{z}_i)$ can be calculated using covariates $\boldsymbol{z}_i$. Thus, there are two parameter sets $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. For illustration Schneider (2019) use data about criminal recidivism. The aim is to model the discrete time until a released prisoner is arrested again. Some convicts will be arrested again and others never do. Figure 4.2 shows the effect of two chosen variables "work experience" and "financial aid". The vertical axis displays the effect of $\exp(\hat{\boldsymbol{\beta}})$

of being non-cured, while the horizontal axis shows the effect of $\exp(\hat{\gamma})$ on occurrence of the event "arrest". The confidence stars in Figure 4.2 correspond to the 2.5% and 97.5% quantiles of 600 non-parametric bootstrap samples. Receiving financial aid indicates that it might reduce the chance of being non-cured and of being arrested since the factor is smaller than one. However, the effect is not statistically significant as the confidence intervals also cover one, which is equivalent to no multiplicative effect. The covariable "work experience: yes" seems to reduce the chance of an event and seems to increase the chance of being non-cured. The effects are again not statistically significant at the 5% level.

FIGURE 4.2.: *Illustration of parameters estimates in the cure model. Source: Schneider (2019, p. 8)*

4.2. Some Related Approaches

An alternative for modelling heterogeneity in survival models is the so-called frailty model. Here, the individual heterogeneity is included by using random effects. The model allows that each observation is characterized by an individual different hazard function. The subject-specific hazard function is given by

$$\lambda(t|\boldsymbol{x}_i, b_i) = h(b_i + \gamma_{0t} + \boldsymbol{x}_i^T \boldsymbol{\gamma}),$$

where $h()$ is a specific response function, $\boldsymbol{\gamma}$ the effects of the covariables, and b_i the random effects following a specific distribution, such as the normal distribution with $b_i \sim N(0, \sigma^2)$, for instance. Observations with lower b_i are under lower risk, and therefore tend to live longer. Larger b_i values correspond to a higher risk and the tendency to live shorter. Figure 4.3 illustrates these effects. Lines above the bold line with $b_i = 0$ are b_i with smaller values. More details can be found in Tutz and Schmid (2016). In contrast to the cure models, all observations are at-risk, but heterogenetic in their hazards. The cure model assumes that there is a group of subjects who is characterized as long-term survivors with a hazard equal to zero. This situation cannot be captured by a frailty model.

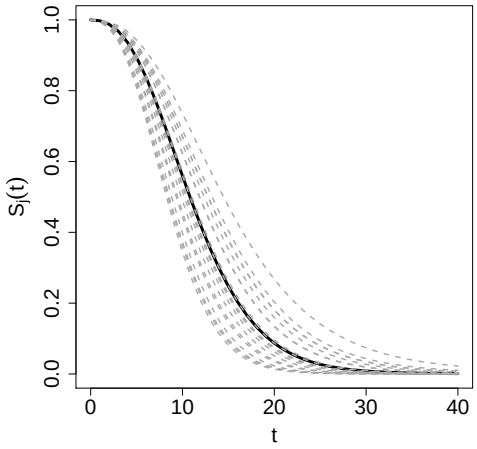


FIGURE 4.3.: *Illustration of the survival function of the Frailty Model*

There are some other approaches which can account for data situations where there is more than one absorbing event as in the cure model. For example, competing risk models can be used to model multiple events such as several types of death. The cause-specific discrete hazard is defined by

$$\lambda_c(t|\boldsymbol{x}_i) = P(T_i = t, C_i = c|T_i \geq t, \boldsymbol{x}_i),$$

where c is a specific event. The overall hazard function results as the sum over all cause-specific discrete hazards so that

$$\lambda(t|\boldsymbol{x}_i) = \sum_{s=1}^{c} \lambda_s(t|\boldsymbol{x}_i).$$

Further information on this can be found in Tutz and Schmid (2016).

Another class of models are the multistate or multiple-spell models, where the event is not irrevocable. For example, a person may be several times in life be employed and unemployed. The event "employment" is not an absorbing event (such as death) but can occur several times during life time. An overview of discrete multiple-spell models can be found in Hamerle and Tutz (1989) and Willett and Singer (1995).

5. Variable Selection

Variable selection is a central task, especially in the proposed mixture models, due to the inflation of number of possible variables, which can occur in the two model components and the weights. Simple strategies are often limited when the complexity of the model increases. A straight-forward method is the all-subset method where all possible models are fitted. This method is very time-consuming and does not work for larger models.

Another strategy are the step-wise procedures which can consist of forward, backward selection or a combination of both. The idea is to include or exclude variables sequentially to the model. However, Breiman (1996), for instance, demonstrated the instability of stepwise regression models. Using backward selection in mixture models may lead to a degenerated model and convergence problems because of too many possibly correlated covariates.

Thus, penalization techniques may be a good solution to receive stable estimates. There, the regular log-likelihood is maximized with respect to a certain side constraint. The penalized log-likelihood for parameter vector $\boldsymbol{\theta}$ is given by

$$l_p(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}),$$

where $l(\boldsymbol{\theta})$ denotes the un-penalized log-likelihood and $J_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ is a specific penalty term. Two popular penalty terms are the ridge (Hoerl and Kennard, 1970) and the lasso regression (Tibshirani, 1996). In the ridge regression, the parameters are shrunk toward zero by using the penalty term

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \lambda \sum_{j=1}^{p} \theta_j^2,$$

while the lasso penalty term ensures also parameter selection by using

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \lambda \sum_{j=1}^{p} |\theta_j|.$$

The larger λ, the larger the shrinkage effect. The proposed penalty terms for CUB, CUP and cure models are based on the group lasso approach by Yuan and Lin (2006), where variables with several categories are selected together instead of selecting only parameters as it is the case in the regular lasso. Thus the vectors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are divided into $\boldsymbol{z}_i^T = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{ig}^T)$ and $\boldsymbol{x}_i^T = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{ih}^T)$ so that each component relates to a single variable. Then $\boldsymbol{z}_i^T\boldsymbol{\beta}$ and $\boldsymbol{x}_i^T\boldsymbol{\gamma}$ are the corresponding predictors and $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_g^T)$ and $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_h^T)$ the corresponding parameter vectors.

5.1. Variable Selection in CUB- and CUP Models

For CUB- and CUP-models Schneider et al. (2019) propose a penalty term which is given by

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda_\beta \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2 + \lambda_\gamma \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2 \tag{5.1}$$

where λ_β and λ_γ are the tuning parameters for the selection of the variables $\boldsymbol{x}$ and $\boldsymbol{z}$, respectively, and $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ the corresponding parameter vectors.

The degrees of freedom $df_{\boldsymbol{\beta}_j}$ are defined as the number of parameters collected in $\boldsymbol{\beta}_j$. $df_{\boldsymbol{\gamma}_j}$ is defined in the same way. $\|\|_2$ is the unsquared L_2-Norm so that the penalty enforces the selection of variables in the spirit of the group lasso (Yuan and Lin, 2006) rather than selecting single parameters. The effective degrees of freedom of each variable are defined by

$$edf(\hat{\boldsymbol{\beta}}_j) = \mathbb{1}(\|\hat{\boldsymbol{\beta}}_j\|_2 > 0) + (df_{\boldsymbol{\beta}_j} - 1)\frac{\|\hat{\boldsymbol{\beta}}_j\|_2}{\|\hat{\boldsymbol{\beta}}_j^{ML}\|_2},$$

$$edf(\hat{\boldsymbol{\gamma}}_j) = \mathbb{1}(\|\hat{\boldsymbol{\gamma}}_j\|_2 > 0) + (df_{\boldsymbol{\gamma}_j} - 1)\frac{\|\hat{\boldsymbol{\gamma}}_j\|_2}{\|\hat{\boldsymbol{\gamma}}_j^{ML}\|_2}$$

If a variable is not penalized, the edf are identical to $df_{\boldsymbol{\beta}_j}$ and $df_{\boldsymbol{\gamma}_j}$, respectively.

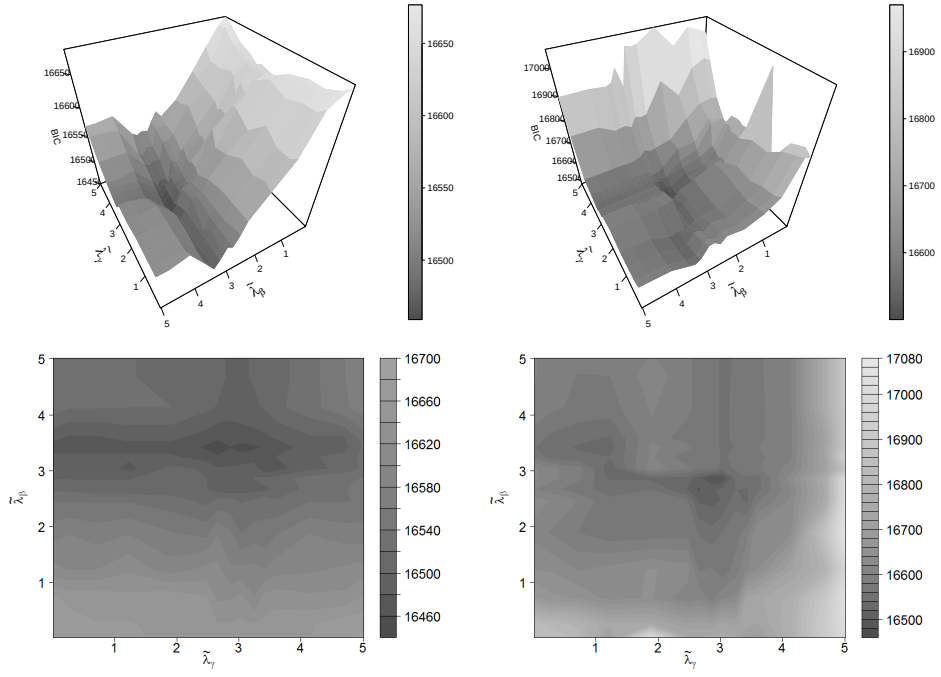


FIGURE 5.1.: *Survey on Household Income and Wealth: Grid of lambda values to find the best model for CUB (left) and CUP model (right). Source: Schneider et al. (2019, p. 15)*

The penalty approach was applied to simulated data as well as to real data. In the Survey on Household Income and Wealth the respondents rate their overall life well-being on a Likert scale, running from "very unhappy" (1) to "very happy" (10). Several covariates, such as marital status, age, or area of living are available. To find the best model according to the lowest BIC value, a 15×15 grid of λ_β and λ_γ values is used. Figure 5.1 shows the results for the CUB and CUP models as surface and contour plot. Dark areas correspond with a low BIC value, whereas white regions are related to high BIC values. The surfaces of both graphs are quite smooth and in both cases a clear area with low BIC values could be detected. However, the surfaces differ from each other so that it can be assumed that the shape depends on the data situation. For both models, a reasonable model was found. For more details I refer to Schneider et al. (2019).

5.2. Variable Selection in Cure Models

For the discrete cure model, Schneider (2019) proposed the following term, which does not only account for variable selection but also for smoothing the baseline hazard:

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda_{\boldsymbol{\beta}} \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2 + \lambda_{\boldsymbol{\gamma}} \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2 + \lambda_0 \sum_{t=1;s>t}^{t^*} \|\gamma_{0t} - \gamma_{0s}\|_2^2.$$

As in the CUB and CUP model, $\lambda_{\boldsymbol{\beta}}$ is the tuning parameter for the mixture weights. $\lambda_{\boldsymbol{\gamma}}$ regulates the amount of shrinkage of the parameters of the hazard function, and λ_0 is the tuning parameter for the baseline hazard. The first two penalty terms, using the unsquared L_2-Norm, enforce the selection of variables in the spirit of group lasso (Yuan and Lin, 2006). By contrast the intercepts of the baseline hazard are smoothed by penalizing the quadratic distances between two neighbouring intercepts. The larger λ_0, the smoother the baseline hazard.

The effective degrees of freedoms of the cure model are calculated by the sum of $edf(\boldsymbol{\beta})$ and $edf(\boldsymbol{\gamma})$ as described in section 5.1 and the $edf(\hat{\boldsymbol{\gamma}}_0)$ given by

$$edf(\hat{\boldsymbol{\gamma}}_0) = 1 + (df_{\boldsymbol{\gamma}_0} - 1) \frac{(R \cdot \hat{\boldsymbol{\gamma}}_0)^T (R \cdot \hat{\boldsymbol{\gamma}}_0)}{(R \cdot \hat{\boldsymbol{\gamma}}_0^{ML})^T (R \cdot \hat{\boldsymbol{\gamma}}_0^{ML})},$$

where for $t^* = 4$ R is given by

$$R = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

If $\lambda_0 \to 0$ the difference of the intercepts are not penalized which may result

in a rough baseline and the penalized $\boldsymbol{\gamma}_0$ is equal to the maximum-likelihood estimate $\boldsymbol{\gamma}_0^{ML}$. In this case, for each intercept one degree of freedom is counted and $edf(\hat{\boldsymbol{\gamma}}_0) = df_{\hat{\boldsymbol{\gamma}}_0}$. In contrast, if $\lambda_0 \to +\infty$ the baseline hazard would be constant with 1 edf. Thus, it is necessary to find a good trade-off between data and smoothing.

The approach was applied to data about breast cancer. To find the combination of tuning parameters with the lowest BIC, a grid search of 15 models are performed. λ_0 was fixed at the value 2 to reduce the model complexity. Figure 5.2 shows the corresponding BIC surface and contour plots. High BIC values correspond with brighter areas, while dark regions represent low BIC values. The size of each region depends on the used gird.

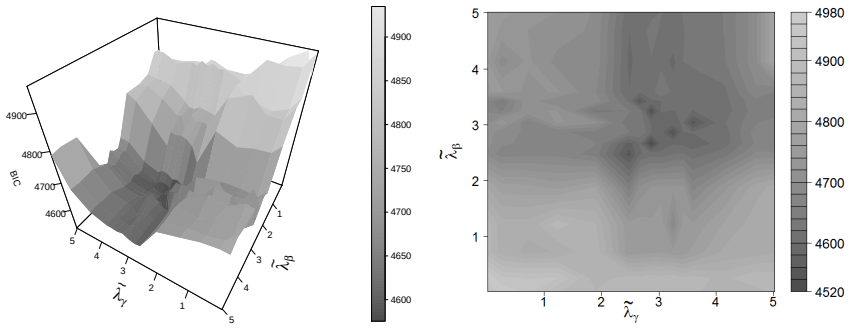


FIGURE 5.2.: *Breast cancer: Grid of λ values to find the best tuning parameter combination according to BIC. Source: Schneider (2019, p. 24)*

Again the proposed penalty term shows a good performance and leads to a clear area with models exhibiting the smallest BIC values.

5.3. Further Remarks

In both settings, the penalization techniques are primarily used for variable selection. In the case of the cure model, one penalty term was designed to smooth the baseline. To avoid biased estimates and to be able to compare the estimates with other selection methods, the models were usually refitted. This leads also to a different value of the degrees of freedom since each unpenalized parameters are counted as one, but penalized parameters may be counted as less than one, too. Even though the shrinkage effect leads to biased estimates, the variance decrease which may result in a smaller mean square error and prediction error. Thus, it may be a question of modelling philosophy whether to use only the selection feature of the penalization technique or to use also the shrank parameters to obtain a model with a lower mean squared error.

Although both the penalization and stepwise selection strategy shows reasonable results, the stepwise selection exhibits some issues such as computational time or degenerated results if variables are highly correlated. However, in each case a certain selection criterion such as AIC, BIC or a statistical test has to be chosen which influence the variable selection process. The likelihood-ratio test exhibits computational difficulties if the p-value of different models lead to

the same result close to zero. In this case the largest deviance difference is used to come to a distinct decision. Furthermore, there is a multiple test problem, where the α level may need to be corrected for. In the applications I usually use the BIC criterion which tend to lead to smaller models than the AIC criterion.

6. Estimation

The mixture models are estimated by an adapted version of the EM algorithm proposed by Dempster et al. (1977). Depending on the specific model, the likelihood is slightly different. In this section I give an overview of the general process. $\log(P_M)$ denotes the log-likelihood of the first component of the mixture model, known as preference component or survival function of the non-cured population. $\log(P_U)$ refer to the log-likelihood of the second component, known as uncertainty or long-term survivors. The general form of the log-likelihood of the mixture model for observation i is given by

$$l_i(\boldsymbol{\theta}) = \{\log(\pi_i(\boldsymbol{z}_i)) + \log(P_M(y_i|\boldsymbol{x}_i))\} + \{\log(1-\pi_i(\boldsymbol{z}_i)) + \log(P_U(y_i|\omega_i))\},$$

where $\boldsymbol{z}_i$ are the variables determining the mixture weights π_i and $\boldsymbol{x}_i$ and $\boldsymbol{\omega}_i$ those connected with the mixture components P_M and P_U, respectively. Using penalization techniques lead to the penalized log-likelihood by adding certain penalty terms J. Each component obtains its own penalty term: J_π, J_M and J_U, so that

$$\begin{aligned} l_i(\boldsymbol{\theta}) = &\log(\pi_i) + \log(P_M(y_i|\boldsymbol{x}_i)) + \log(1-\pi_i) + \log(P_U(y_i|\omega_i)) \\ &- \lambda_\beta J_\pi - \lambda_\gamma J_M - \lambda_\alpha J_U \end{aligned}$$

and for all observations

$$\begin{aligned} l_p(\boldsymbol{\theta}) = &\sum_{i=1}^{n} [\log(\pi_i) + \log(P_M(y_i|\boldsymbol{x}_i)) + \log(1-\pi_i) + \log(P_U(y_i|\omega_i))] \\ &- \lambda_\beta J_\pi - \lambda_\gamma J_M - \lambda_\alpha J_U \end{aligned}$$

The EM algorithm uses the complete likelihood which treats the membership to P_M and P_U as missing data. Let z_i^* take the value 1 if observation i belongs to P_M, and zero if observation i belongs to P_U. Then, the complete penalized

log-likelihood is given by

$$\begin{aligned} l_p(\boldsymbol{\theta}) = &\sum_{i=1}^{n}(z_i^* \left[\log(\pi_i) + \log(P_M(y_i|\boldsymbol{x}_i))\right] \\ &+ \sum_{i=1}^{n}(1 - z_i^*) \left[\log(1 - \pi_i) + \log(P_U(y_i|\omega_i))\right] \\ &- \lambda_\beta J_\pi - \lambda_\gamma J_M - \lambda_\alpha J_U, \end{aligned}$$

Within the EM algorithm, the log-likelihood is iteratively maximized by using an expectation and a maximization step. During the E-step, the conditional expectation of the complete log-likelihood, given the observed data $\boldsymbol{y}$ and the current estimate $\boldsymbol{\theta}^{(s)}$,

$$M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \mathrm{E}(l_p(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}^{(s)})$$

is computed. Because $l_p(\boldsymbol{\theta})$ is linear in the unobservable data z_i^*, it is only necessary to estimate the current conditional expectation of z_i^*. From Bayes's theorem follows

$$\begin{aligned} E(z_i^*|\boldsymbol{y}, \boldsymbol{\theta}) &= P(z_i^* = 1|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= P(y_i|z_i^* = 1, \boldsymbol{x}_i, \boldsymbol{\theta})P(z_i^* = 1|\boldsymbol{x}_i, \boldsymbol{\theta})/P(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= \pi_i P_M(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})/(\pi_i P_M(y_i|\boldsymbol{x}_i) + (1 - \pi_i)P_U(y_i|\omega_i)) = \hat{z^*}_i. \end{aligned}$$

This is the posterior probability that the observation y_i belongs to P_M. For the s-th iteration, one obtains

$$\begin{aligned} M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = &\underbrace{\sum_{i=1}^{n}\left\{\hat{z^*}_i^{(s)} \log(\pi_i) + (1 - \hat{z^*}_i^{(s)}) \log(1 - \pi_i)\right\} - \lambda_\beta J_\pi}_{M_1} \\ &+ \underbrace{\sum_{i=1}^{n}\left\{\hat{z^*}_i^{(s)} \log(P_M(y_i|\boldsymbol{x}_i) + (1 - \hat{z^*}_i^{(s)}) \log(P_U(y_i|\omega_i))\right\} - \lambda_\gamma J_M - \lambda_\alpha J_U}_{M_2} \end{aligned}$$

M_1 and M_2 can be estimated independently from each other. Sometimes M_2 is split into two equations depending on the structure of P_U. Most traditional methods, such as Fisher-Scoring, cannot be used to estimate the penelatized likelihoods because the derivatives do not exist due to the group lasso penalty terms. This problem can be solved with the fast iterative shrinkage-thresholding algorithm (FISTA) by Beck and Teboulle (2009), which is implemented in the MRSP package by Pößnecker (2019) and is used for the maximization problem. It can be formulated as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmax}}\, l_p(\boldsymbol{\theta}) = -\underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}}\, l_p(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}}\; - l(\boldsymbol{\beta}, \boldsymbol{\gamma}) + \sum_{v=1}^{V} \lambda_v J_v. \qquad (6.1)$$

FISTA belongs to the class of proximal gradient methods in which only the unpenalized log-likelihood and its gradient are necessary. The solution for the unknown parameters $\boldsymbol{\theta}$ of the unpenalized log-likelihood in iteration $t+1$ is given by:

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)} + \frac{1}{\nu}\nabla l(\hat{\boldsymbol{\theta}}^{(t)}),$$

where $\nu > 0$ is the inverse stepsize parameter. This estimator converges to the ML estimator so that each update of $\hat{\boldsymbol{\theta}}^{(t)}$ can be considered as a one-step approximation to the ML estimator based on the current iterate. A more detailed description can be found in Tutz et al. (2015).

For given $\boldsymbol{\theta}^{(s)}$, one computes in the E-step the weights $\hat{z^*}_i^{(s)}$ and in the M-step maximizesn $M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$, which leads to the new estimates. The E- and M-steps are repeated alternatingly until the difference $l_p(\boldsymbol{\theta}^{(s+1)}) - l_p(\boldsymbol{\theta}^{(s)})$ is small enough to assume convergence. To account for different sizes of the log-likelihood, we define

$$\Big|\frac{l_p(\boldsymbol{\theta}^{(s+1)}) - l_p(\boldsymbol{\theta}^{(s)})}{rel.tol/10 + |l_p(\boldsymbol{\theta}^{(s+1)})|}\Big| < rel.tol$$

as stopping criteria. *rel.tol* is the relative tolerance which has to be below a certain threshold to assume convergence. The number of penalty terms span a V-dimensional grid of tuning parameter space. Dempster et al. (1977) showed that under weak conditions the EM algorithm finds a local maximum of the likelihood function. Hence it is sensible to use meaningful start values to find a reasonable solution of the maximization problem. Further details can be found in the articles describing the models.

7. Summary and Outlook

Finite Mixtures are a flexible method for modelling heterogeneity in ordinal response. I have demonstrated how to model heterogeneity of ordinal scales in surveys with a flexible preference and uncertainty structure. Furthermore, the method can be also used in the context of discrete survival analysis, where one group is at-risk and the other group is cured, but the membership is unobserved. The proposed variable selection helps to choose an appropriate model and stabilize the estimation procedure. However, there are still some questions which could be examined in more detail in future research. I focus first on some general issues and then continue with questions related to heterogeneity in surveys and in discrete survival analysis.

First, the appropriate computational estimation of the mixture models could be developed further. Since there is no closed form of the likelihoods of the mixture models the search for the best maximum is not an easy task. Multiple starting values, realistic settings, and sanity checks may help to find a reasonable maximum, but there is no guarantee that the found maximum is the global maximum. Second, the more starting values are used and the stronger the stopping criterion is set, the more computational time is necessary. More clarity in this area would improve all results.

The computational time issue becomes especially important when thinking of the grid search in the penalized case or using bootstrap samples, where a mixture model has to be estimated a few hundred times. One promising approach to reduce computational time, would be a model based optimization strategy replacing at least the grid search. Here, the estimation process would start at a certain tuning parameter combination and then find the direction where the penalized likelihood is reduced the most and continue only with this area. Thus, not all possible tuning parameters combinations have to be considered.

From my experience, the use of standard errors is in most cases rather optimistic. For example, if the mixture weight is estimated at the border of the parameter space, standard errors or statistical tests, which do not account for this situation, may be not the best choice. Bootstrap quantiles may be more realistic considering the whole structure and specifies of the model including skewed distribution of the standard errors. If the model needs less computing time for estimation, non-parametric bootstrap errors, which account for the model search also in penalized settings, could be easily applied.

Variable selection is an important issue because of the amount of variables that can be included in the mixture model and the importance of the weights which can be modelled by covariates. Using stepwise selection method and the likelihood-ratio test as criterion exhibits the additional issue of multiple test problem. It might be valuable to discuss if there is need to adjust the

significance level since there are usually twice as much possible models as in the traditional regression model because of the two sets of variables.

Mixture models for survey question provide deeper insights into the mechanism of human decisions. But it is an open research question how to model heterogeneity in survey in the most appropriate way. In the CUB and CUP models variables are used in the mixture weights and the preference structure. In the BetaBin model, the variables are included in the preference structure and the uncertainty structure but not in the weights. Another approach would be to model only the tendency to the middle category by using variables in the preference structure and the mixture weights and use a fixed uncertainty distribution.

So far, the CUP and BetaBin model focus on single response items. Future research could look at the possibilities to extend these models to deal with multiple possible correlated items. Furthermore, one could include the idea of an inflated category in the CUP model. The proposed penalization approach could be also implemented for other mixture models such as the BetaBin to reduce the complexity in the shape of the beta-binomial distribution and the developed vote choice model could be easily extended to a multi-party setting by using a multinomial model.

In survival analysis, the mixture models are restricted to two components: One for long-term survivors and one for the population at-risk. But in some diseases, there may be more than these two groups. There might be patients who are long-term survivors, but also patients with a moderate and a severe course of the disease. Thus, it might be interesting to use more than two groups as it is done in the competing risk models. From a technical point of view, this extension should be straight forward. However, finding a good and stable maximum may be more difficult due to more possible variation.

So far, the smoothing tuning parameter was fixed at a certain level due to computational time. Smoothing the baseline results in a stable estimation even if not at each time an event occurs. The larger the tuning parameter the smoother the baseline hazard. However, the effect of the smoothing tuning parameter has not yet been evaluated in detail, i.e. if the smoothness of the baseline influence the other parameters. Furthermore, the penalization technique could be also compared to a stepwise procedure.

Finally discrete survival models are rarely used although their interpretation is often more intuitive than models for continuous data. There is little research about comparing survival models, especially cure models, for discrete and continuous time. It can be supposed that the more time points or intervals are used, the similar should be the results of both model classes.

In summary, finite mixtures are able to model heterogeneity in ordinal responses. Although there are some open questions and computational challenges, these models help to understand the mechanisms in surveys and in discrete survival analysis.

References

Agresti, A. (2010). *Analysis of Ordinal Categorical Data, 2nd Edition*. New York: Wiley.

Agresti, A. (2013). *Categorical Data Analysis, 3d Edition*. New York: Wiley.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology 13*(1), 61–98.

Alvarez, R. M. and J. Brehm (1995). American ambivalence towards abortion policy: Development of a heteroskedastic probit model of competing values. *American Journal of Political Science 39*(4), 1055–1082.

Amico, M. and I. V. Keilegom (2018). Cure models in survival analysis. *Annual Review of Statistics and Its Application 5*(1), 311–342.

Barlow, R. E., J. M. Brenner, H. D. Brunk, and D. J. Bartholomew (1972). *Statistical inference under order restrictions: The theory and application of isotonic regression*. New York: Wiley.

Bartels, L. M. (1986). Issue voting under uncertainty: An empirical test. *American Journal of Political Science 30*(4), 709–728.

Baumgartner, H. and J.-B. Steenkamp (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research 38*(2), 143–156.

Beck, A. and M. Teboulle (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences 2*(1), 183–202.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods 17*(4), 665–678.

Bolt, D. M. and T. R. Johnson (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement 33*(5), 335–352.

Bolt, D. M. and J. R. Newton (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement 71*(5), 814–833.

Breiman, L. (1996). Heuristics of instability and stabilisation in model selection. *Annals of Statistics 24*(6), 2350–2383.

Broström, G. (2012). *Event History Analysis with R*. New York: CRC Press.

Bürkner, P.-C. and E. Charpentier (2018). Monotonic effects: A principled approach for including ordinal predictors in regression models. *PsyArXiv, doi: 10.31234/osf.io/9qkhj*.

Capecchi, S. and D. Piccolo (2017). Dealing with heterogeneity in ordinal responses. *Quality & Quantity 51*(5), 2375–2393.

Colombi, R., S. Giordano, and A. Gottard (2019). Discussion of "the class of cub models: statistical foundations, inferential issues and empirical evidence". *Statistical Methods & Applications*. online published.

de Boeck, P. and I. Partchev (2012). Irtrees: Tree-based item response models of the glmm family. *Journal of Statistical Software 48*(1), 1–28.

de Bruin, W. B. and K. G. Carman (2012). Measuring risk perceptions: What does the excessive use of 50% mean? *Medical Decision Making 32*(2), 232–236.

de Vries, C. and M. R. Steenbergen (2013). Variable opinions: The predictability of support for unification in European mass publics. *Journal of Political Marketing 12*(1), 121–141.

D'Elia, A. and D. Piccolo (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis 49*(3), 917–934.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B 39*(1), 1–38.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer.

Gottard, A., M. Iannario, and D. Piccolo (2016). Varying uncertainty in cub models. *Advances in Data Analysis and Classification 10*(2), 225–244.

Grilli, L., M. Iannario, D. Piccolo, and C. Rampichini (2014). Latent class cub models. *Advances in Data Analysis and Classification 8*(1), 105–119.

Grilli, L. and C. Rampichini (2019). Discussion of 'the class of cub models: statistical foundations, inferential issues and empirical evidence' by domenico piccolo and rosaria simone. *Statistical Methods & Applications*. online published.

Hamerle, A. and G. Tutz (1989). *Diskrete Modelle zur Analyse von Verweildauern und Lebenszeiten*. Frankfurt/New York: Campus Verlag.

Harbers, I., C. E. de Vries, and M. R. Steenbergen (2013). Attitude variability among Latin American publics: How party system structuration affects left/right ideology. *Comparative Political Studies 46*(8), 947–967.

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica 44*(3), 461–465.

Helwig, N. E. (2017). Regression with ordered predictors via ordinal smoothing splines. *Frontiers in Applied Mathematics and Statistics 3*(1).

Hilbe, J. M. (2011). *Negative Binomial Regression* (2 ed.). Cambridge: Cambridge University Press.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Bias estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Iannario, M. (2012a). Hierarchical CUB models for ordinal variables. *Communications in Statistics-Theory and Methods 41*(16-17), 3110–3125.

Iannario, M. (2012b). Modelling *shelter* choices in a class of mixture models for ordinal responses. *Statistical Methods and Applications 21*(1), 1–22.

Iannario, M. (2014). Modelling uncertainty and overdispersion in ordinal data. *Communications in Statistics-Theory and Methods 43*(4), 771–786.

Iannario, M. and D. Piccolo (2010). A new statistical model for the analysis of customer satisfaction. *Quality Technology & Quantitative Management 7*(2), 149–168.

Iannario, M. and D. Piccolo (2012). CUB models: Statistical methods and empirical evidence. In R. Kennett and S. Salini (Eds.), *Modern Analysis of Customer Surveys: with applications using R*, pp. 231–258. New York: Wiley.

Iannario, M. and D. Piccolo (2016a). A comprehensive framework of regression models for ordinal data. *Metron 74*(2), 233–252.

Iannario, M. and D. Piccolo (2016b). A generalized framework for modelling ordinal data. *Statistical Methods & Applications 25*(2), 163–189.

Jeon, M. and P. de Boeck (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods 48*(3), 1070–1085.

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika 68*(4), 563–583.

Kankaraš, M. and G. Moors (2009). Measurement equivalence in solidarity attitudes in europe insights from a multiple-group latent-class factor approach. *International Sociology 24*(4), 557–579.

Kenett, R. (2019). A review of: The class of cub models: statistical foundations, inferential issues and empirical evidence by domenico piccolo and rosaria simone. *Statistical Methods & Applications*. online published.

Kuk, A. Y. C. and C.-H. Chen (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika 79*(3), 531–541.

Leitenstorfer, F. and G. Tutz (2007). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics 8*(3), 654–673.

Little, R. J. and D. B. Rubin (2002). *Statistical analysis with missing data* (2 ed.). Hoboken: Wiley.

Maller, R. A. and X. Zhou (1996). *Survival analysis with long-term survivors*. New York: Wiley.

Manisera, M. and P. Zuccolotto (2014). Modeling rating data with nonlinear CUB models. *Computational Statistics & Data Analysis 78*(C), 100–118.

Mauerer, I. and M. Schneider (2019a). Perceived party placements and uncertainty on immigration in the 2017 german election. In M. Debus, M. Tepe, and J. Sauermann (Eds.), *Jahrbuch für Handlungs- und Entscheidungstheorie: Band 11*, pp. 117–143. Wiesbaden: Springer.

Mauerer, I. and M. Schneider (2019b). Uncertainty in issue placements and spatial voting. Technical Report 226, Department of Statistics, Ludwig-Maximilians-Universität München.

McCullagh, P. (1980). Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society B 42*(2), 109–127.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: Wiley.

Moors, G. (2004). Facts and artefacts in the comparison of attitudes among ethnic minorities. a multigroup latent class structure model with adjustment for response style behavior. *European Sociological Review 20*(4), 303–320.

Moors, G. (2009). Ranking the ratings: A latent-class regression model to control for overall agreement in opinion research. *International Journal of Public Opinion Research 22*(1), 93–119.

Möst, S. (2014). *Regularization in discrete survival models*. PhD dissertation, Ludwig-Maximilians-Universität München.

Muthén, B. and K. Masyn (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral statistics 30*(1), 27–58.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society A 135*(3), 370–384.

Paulhus, D. L. (1991). Chapter 2 - measurement and control of response bias. In J. P. Robinson, P. R. Shaver, and L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes*, pp. 17–59. Academic Press.

Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica 5*(1), 85–104.

Piccolo, D. (2006). Observed information matrix in mub models. *Quaderni di Statistica 8*(1), 33–78.

Piccolo, D. (2015). Inferential issues on CUBE models with covariates. *Communications in Statistics-Theory and Methods 44*(23), 5023–5036.

Piccolo, D. and A. D'Elia (2008). A new approach for modelling consumers' preferences. *Food Quality and Preference 19*(3), 247–259.

Piccolo, D. and R. Simone (2019a). The class of cub models: statistical foundations, inferential issues and empirical evidence. *Statistical Methods & Applications*. online published.

Piccolo, D. and R. Simone (2019b). Rejoinder to the discussion of "the class of cub models: statistical foundations, inferential issues and empirical evidence". *Statistical Methods & Applications*. online published.

Piccolo, D., R. Simone, and M. Iannario (2018). Cumulative and cub models for rating data: A comparative analysis. *International Statistical Review, online published*.

Plass, J., P. Fink, N. Schöning, and T. Augustin (2015). Statistical modelling in surveys without neglecting "the undecided": Multinomial logistic regression models and imprecise classification trees under ontic data imprecision. Technical Report 179, Department of Statistics, Ludwig-Maximilians-Universität München, Germany.

Pößnecker, W. (2019). MRSP: Multinomial response models with structured penalties. R package version 0.6.11, https://github.com/WolfgangPoessnecker/MRSP.

Proietti, T. (2019). Discussion of the class of cub models: statistical foundations, inferential issues and empirical evidence. *Statistical Methods & Applications*. online published.

Ramsey, J. O. (1988). Monotone splines in action (with discussion). *Statistical Science 3*(4), 425–461.

Robertson, T., F. T. Wright, and R. L. Dykstra (1988). *Order restricted statistical inference*. New York: Wiley.

Rosmalen, J. V., H. V. Herk, and P. Groenen (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research 47*(1), 157–172.

Rozenas, A. (2013). Inferring ideological ambiguity from survey data. In N. Schofield, G. Caballero, and D. Kselman (Eds.), *Advances in Political Economy: Institutions, Modelling and Empirical Analysis*, pp. 369–382. Berlin, Heidelberg: Springer.

Schneider, M. (2019). Dealing with heterogeneity in discrete survival analysis using the cure model. Technical Report 224, Department of Statistics, Ludwig-Maximilians-Universität München.

Schneider, M., W. Pößnecker, and G. Tutz (2019). Variable selection in mixture models with an uncertainty component. Technical Report 225, Department of Statistics, Ludwig-Maximilians-Universität München.

Simone, R. and G. Tutz (2018). Modelling uncertainty and response styles in ordinal data. *Statistica Neerlandica 72*(3), 224–245.

Simone, R., G. Tutz, and M. Iannario (2019). Subjective heterogeneity in response attitude for multivariate ordinal outcomes. *Econometrics and Statistics, online published*.

Steele, F. (2003). A discrete-time multilevel mixture model for event history data with long-term survivors, with an application to an analysis of contraceptive sterilization in bangladesh. *Lifetime Data Analysis 9*(2), 155–174.

Sy, J. P. and J. M. G. Taylor (2000). Estimation in a cox proportional hazards cure model. *Biometrics 56*(1), 227–236.

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

Tutz, G. (1991). Sequential models in ordinal regression. *Computational Statistics & Data Analysis 11*(3), 275–295.

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge: Cambridge University Press.

Tutz, G. (2018). Binary response models with underlying heterogeneity: Identification and interpretation of effects. *European Sociological Review 34*(2), 211–221.

Tutz, G. (2019). Comments on the class of cub models: statistical foundations, inferential issues and empirical evidence by d. piccolo and r. simone. *Statistical Methods & Applications*. online published.

Tutz, G. and M. Berger (2016). Response styles in rating scales: Simultaneous modeling of content-related effects and the tendency to middle or extreme categories. *Journal of Educational and Behavioral Statistics 41*(3), 239–268.

Tutz, G. and M. Berger (2018). Tree-structured modelling of categorical predictors in generalized additive regression. *Advances in Data Analysis and Classification 12*(3), 737–758.

Tutz, G. and Gertheiss (2014). Rating scales as predictors – the old question of scale level and some answers. *Psychometrika 79*(3), 357–376.

Tutz, G., W. Pößnecker, and L. Uhlmann (2015). Variable selection in general multinomial logit models. *Computational Statistics & Data Analysis 82*(1), 207 – 222.

Tutz, G., G. Schauberger, and M. Berger (2018). Response styles in the partial credit model. *Applied Psychological Measurement 42*(6), 407–427.

Tutz, G. and M. Schmid (2016). *Modeling Discrete Time-to-Event Data*. Basel: Springer.

Tutz, G. and M. Schneider (2019). Flexible uncertainty in mixture models for ordinal responses. *Journal of Applied Statistics 46*(9), 1582–1601.

Tutz, G., M. Schneider, M. Iannario, and D. Piccolo (2017). Mixture models for ordinal responses to account for uncertainty of choice. *Advances in Data Analysis and Classification 11*(2), 281–305.

Ursino, M. (2014). *Ordinal data: a new model with applications*. PhD dissertation, Politecnico di Torino, openly available since March/2019.

Vaerenbergh, Y. V. and T. D. Thomas (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research 25*(2), 195–217.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer.

Willett, J. B. and J. D. Singer (1993a). Investigating onset, cessation, relapse, and recovery: Why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *Journal of Consulting and Clinical Psychology 61*(6), 952–965.

Willett, J. B. and J. D. Singer (1993b). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics 18*(2), 155–195.

Willett, J. B. and J. D. Singer (1995). It's déjà vu all over again: Using multiple-spell discrete-time survival analysis. *Journal of Educational and Behavioral Statistics 20*(1), 41–67.

Yee, T. W. (2016). *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.0-3.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B 68*(1), 49–67.

A. Publications

A.1 Tutz, G., M. Schneider, M. Iannario and D. Piccolo (2017): Mixture models for ordinal responses to account for uncertainty of choice. *Advances in Data Analysis and Classification 11(2)*, 281–305, doi:10.1007/s11634-016-0247-9.

A.2 Tutz, G., and M. Schneider (2019): Flexible uncertainty in mixture models for ordinal responses. *Journal of Applied Statistics*, 46(9), 1582–1601, doi:10.1080/02664763.2018.1555574.

A.3 Mauerer, I. and M. Schneider (2019a): Perceived party placements and uncertainty on immigration in the 2017 German election. In Debus, M., M. Tepe and J. Sauermann (Eds.), *Jahrbuch für Handlungs- und Entscheidungstheorie: Band 11*, pp. 117–143. Wiesbaden: Springer, doi:10.1007/978-3-658-23997-8.

A.4 Mauerer, I. and M. Schneider (2019b): Uncertainty in Issue Placements and Spatial Voting. *Technical Report 226*, Department of Statistics, Ludwig-Maximilians-Universität München, doi:10.5282/ubm/epub.68451.

A.5 Schneider, M., Pößnecker, W. and G. Tutz (2019): Variable Selection in Mixture Models with an Uncertainty Component. *Technical Report 225*, Department of Statistics, Ludwig-Maximilians-Universität München, doi:10.5282/ubm/epub.68452.

A.6 Schneider, M. (2019): Dealing with Heterogeneity in Discrete Survival Analysis using the Cure Model. *Technical Report 224*, Department of Statistics, Ludwig-Maximilians-Universität München, doi:10.5282/ubm/epub.68455.

The articles are included in original or as accepted manuscripts if required by license terms and conditions.

A.1. Mixture Models for Ordinal Responses to Account for Uncertainty of Choice

Tutz, G., M. Schneider, M. Iannario and D. Piccolo (2017): Mixture models for ordinal responses to account for uncertainty of choice. ***Advances in Data Analysis and Classification 11(2)*****, 281–305, doi:10.1007/s11634-016-0247-9.**

Adv Data Anal Classif
DOI 10.1007/s11634-016-0247-9

REGULAR ARTICLE

Mixture models for ordinal responses to account for uncertainty of choice

Gerhard Tutz[1] · **Micha Schneider**[1] · **Maria Iannario**[2] · **Domenico Piccolo**[2]

Received: 16 December 2014 / Revised: 14 April 2016 / Accepted: 20 April 2016

Abstract In CUB models the uncertainty of choice is explicitly modelled as a Combination of discrete Uniform and shifted Binomial random variables. The basic concept to model the response as a mixture of a deliberate choice of a response category and an uncertainty component that is represented by a uniform distribution on the response categories is extended to a much wider class of models. The deliberate choice can in particular be determined by classical ordinal response models as the cumulative and adjacent categories model. Then one obtains the traditional and flexible models as special cases when the uncertainty component is irrelevant. It is shown that the effect of explanatory variables is underestimated if the uncertainty component is neglected in a cumulative type mixture model. Visualization tools for the effects of variables are proposed and the modelling strategies are evaluated by use of real data sets. It is demonstrated that the extended class of models frequently yields better fit than classical ordinal response models without an uncertainty component.

Keywords Ordinal responses · Rating analysis · CUP model · CUB model

Mathematics Subject Classification 62-07 · 62H17 · 62J12

1 Introduction

In many applications the responses are measured on an ordinal scale and given in categories. There is a considerable amount of literature devoted to the adequate modelling of such ordered categorical data. In particular the seminal paper of McCullagh (1980)

✉ Gerhard Tutz
tutz@stat.uni-muenchen.de

[1] Ludwig-Maximilians-Universität München, Akademiestraße 1, 80799 Munich, Germany

[2] University of Naples Federico II, Via L.Rodinò 22, 80138 Naples, Italy

Published online: 06 May 2016

stimulated research to find parametric models which should be both parsimonious and well fitted to real data. Overviews on recent research are found, for example, in Agresti (2010, 2013) and Tutz (2012).

Ordered categorical responses typically come in two forms, as *grouped continuous variables* and *assessed ordinal categorical variables* (Anderson 1984). The first type is a mere categorized version of a continuous variable, which in principle can be observed itself. The second type of ordered variable arises when an assessor processes an unknown amount of information, leading to the judgement of the grade of the ordered categorical scale. This sort of variable is found, for example, in preference or evaluation studies and the assessment of pain.

In the following we consider ordinal variables that are generated by judgements. For this type of ordinal response a mixture type model that accounts for the psychological process of human choices has been introduced by Piccolo (2003) and developed in a series of papers by D'Elia and Piccolo (2005), Iannario and Piccolo (2010) and Manisera and Zuccolotto (2014). The basic concept of these so-called CUB models is that the choice of a response category is determined by a mixture of feeling and uncertainty. Feeling refers to the deliberate choice of a response category determined by the preferences of a person while uncertainty refers to the inherent individual's indecision. The first component is modelled by a binomial distribution, the latter by a discrete uniform distribution across response categories. For ordinal response data that reflect opinions or judgements these components, effectively parameterized in a parsimonious manner, allow CUB models to be extremely flexible for capturing the different shapes of ordinal data distributions; in addition, the parameters to be estimated are immediately related to the concept of uncertainty (indecision, fuzziness) and feeling (attraction, preference), which improves the simplicity of the interpretation and makes the comparison among subgroups easier. An introduction and overview was given by Iannario and Piccolo (2012) whereas several generalizations in different fields have been obtained to include objects' covariates multilevel data (Iannario 2012a), and data surveys with *shelter* effects (Iannario 2012b).

Alternative approaches to finite mixtures have been advanced by Wedel and DeSarbo (1995), Greene and Hensher (2003), Grün and Leisch (2008) and Breen and Luijkx (2010) among others. These authors propose convex combinations of probability distributions belonging to the same class of models and assume the existence of subgroups whose responses should be differently modelled. Grilli et al. (2014) proposed a latent class version of CUB models. For general mixture models that are based on latent variables see, for example, Everitt (1988).

In the present paper we consider a mixture model that includes more traditional models for the modelling of preferences than the CUB model. We consider distributions in which the preference part is determined by a cumulative or adjacent categories model, which yields more flexible models. The paper is organized as follows: in the next section we consider uncertainty as a relevant component quite often present in human choices; thus CUB models are briefly reviewed and a new class of models (called CUP) is introduced and estimation concepts are given. For both models a non-parametric measure of heterogeneity may help to understand the weights and the effect of introducing uncertainty in the mixture. In Sect. 3 a deeper discussion is given concerning the effects of the uncertainty component in the interpretation of the model

whereas Sect. 4 deals with the problem of model selection by adequate fitting measures. Section 5 presents some empirical evidence on data sets of different scientific fields and compares standard approaches with mixtures that include an uncertainty component. Some concluding remarks and an appendix devoted to estimation problems conclude the paper.

2 Modelling uncertainty by mixtures

In the following we first sketch the CUB model, which is an abbreviation for Combination of discrete Uniform and shifted Binomial random variables. Then we consider an extended class that contains the CUB as well as standard models for ordinal data as special cases.

2.1 The CUB model

Let in a regression model the response of an individual R_i given explanatory variables $z_i, \boldsymbol{x}_i$ take values from ordered categories $\{1, \ldots, k\}$. Then, the mixture distribution denoted as CUB as considered, for example, by Iannario and Piccolo (2012) has been defined for each subject by

$$Pr(R_i = r|z_i, \boldsymbol{x}_i) = \pi_i \, b_r(\xi_i) + (1 - \pi_i) \, p_r^U, \quad r \in \{1, \ldots, k\}, \tag{1}$$

where the two components of the mixture are specified in the following way. The first component is a shifted binomial distribution given by

$$b_r(\xi_i) = \binom{k-1}{r-1} \xi_i^{k-r} (1 - \xi_i)^{r-1}, \quad r \in \{1, \ldots, k\}.$$

It is a simple binomial distribution determined by the parameter ξ but shifted so that the support is $\{1, \ldots, k\}$ instead of the usual support that includes zero. The component represents the preferences for specific categories, which is captured by the parameter ξ_i.

The second component is a uniform distribution across the response categories,

$$p_r^U = 1/k, \quad r \in \{1, \ldots, k\}.$$

It represents the additional uncertainty arising from factors like amount of time devoted to the response, fatigue, partial understanding, etc. It is explicitly modelled as the indecision component related to the nature of human choices. Iannario and Piccolo (2012) discuss extensively the logical foundations and psychological motivations of the mixture.

In CUB models the parameters π_i and ξ_i are related to the covariates $(\boldsymbol{z}_i^T, \boldsymbol{x}_i^T)$ by the logit links

$$\text{logit}\,(\pi_i) = \boldsymbol{z}_i^T \boldsymbol{\beta}; \qquad \text{logit}\,(\xi_i) = \boldsymbol{x}_i^T \boldsymbol{\gamma}; \qquad i = 1, 2, \ldots, n. \tag{2}$$

In fact, alternative link functions representing a one-to-one mapping $\mathbb{R}^p \leftrightarrow [0, 1]$ between parameters and covariates are also legitimate. It should be noted that, given the parameterization (1), the covariates in $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ may coincide, overlap or be completely different.

The two components, preference/feeling represented by the binomial model, and uncertainty represented by the uniform model, are combined in a mixture with weights $\pi_i, 1 - \pi_i$. The interpretation is that each interviewee has a *propensity* to adhere to a meditated choice (represented by the first component) and to a totally random decision (represented by the uniform distribution) and $\pi_i, 1 - \pi_i$ are just the weights for those propensities. Thus, the quantity $1 - \pi_i$ is interpreted as a measure of uncertainty whereas π_i is seen as a measure of adherence to the structured choice.

In the following we briefly investigate the uncertainty component, which is at the core of this paper. For simplicity we drop the index i for the individual. The first effect of the mixture is that for $\pi < 1$ the distribution of the CUB model is more spread out than the distribution of the binomial model. This can be seen by considering that the variance of the distribution of a CUB model is

$$\text{var}(R) = (k-1)\left[\pi\,\xi\,(1-\xi)\,\{\pi\,(k-1)-(k-2)\} + (1-\pi)\,\frac{3\,\pi\,(k-1)+(k+1)}{12}\right].$$

It is immediate to show that var(R) is monotonically increasing in a linear way with respect to $1 - \pi$ only for $\xi = 1/2$ (a symmetric CUB distribution) whereas it has a minimum for $\pi = 1$ (a shifted binomial model) and a relative maximum for $\pi = 0$ (a discrete uniform model). In fact, the absolute maximum of the parabolic shape happens at

$$\pi = \frac{\left(1 - 6\,\xi + 6\,\xi^2\right)\,(k-2)}{3\,(2\,\xi-1)^2\,(k-1)}, \qquad \text{if} \quad \xi \neq 1/2.$$

As a consequence, as shown in the left panel of Fig. 1, although variance generally increases with uncertainty one cannot conclude that π is strictly related to this aspect of variability.

On the other side, according to the results of Iannario (2012c, pp. 169–170;181), the normalized Gini heterogeneity index increases with uncertainty. It is defined for any discrete distribution $(p_r,\ r = 1, 2, \ldots, k)$ by $G = (1 - \sum_{r=1}^{k}\ p_r^2)\,k/(k-1)$. For the CUB model one obtains

$$G_{CUB} = 1 - \pi^2\,(1 - G_{BIN}),$$

where $G_{BIN} = (1 - \sum_{r=1}^{k}\ b_r(\xi)^2)\,k/(k-1)$ is the Gini index computed for the distribution of the binomial component. From this last result, one can derive that for $\pi < 1$ the Gini index for the mixture model is larger than the Gini index for the binomial model: $G_{CUB} > G_{BIN}$, that is, the heterogeneity of the mixture is greater than that of the binomial component, and heterogeneity is increased if the uniform

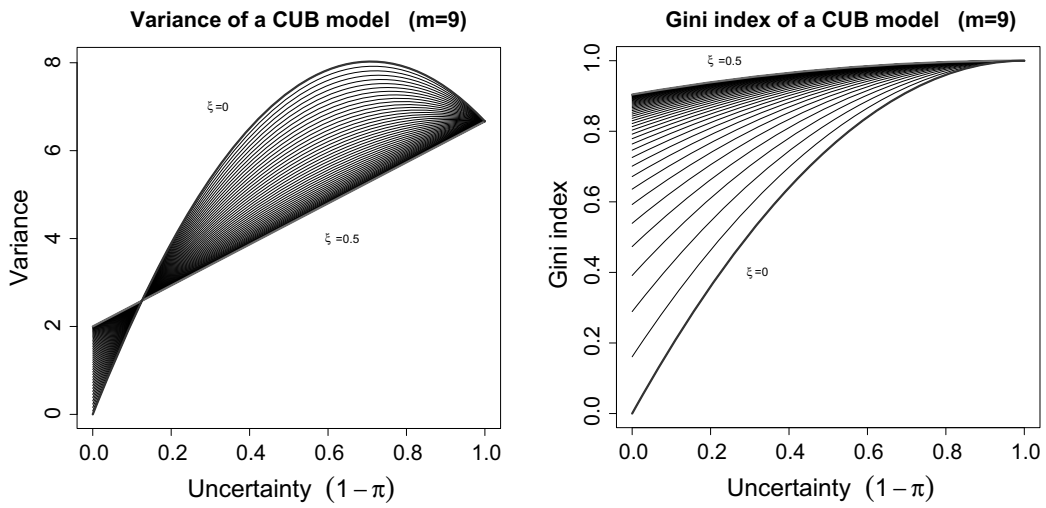


Fig. 1 Variance and Gini heterogeneity measures for CUB models as functions of $1 - \pi$

component can not be neglected. Differently from the variance, the Gini index is monotonically increasing with uncertainty as measured by $(1 - \pi)$ for any given ξ, and this confirms that one should interpret the π parameter as an inverse heterogeneity measure. The behaviour of G_{CUB} with respect to the uncertainty $(1 - \pi)$ is depicted in the right panel of Fig. 1.

Some difficulties arise when the responses are more complex and do not follow a definite pattern as implied by the binomial component (which requires a single mode, for instance). Thus, it seems attractive to extend the standard models for ordinal models by including an uncertainty component, which is the added value of the CUB models framework.

2.2 An extended class of models: CUP

In the CUB model the choice of a binomial distribution and a uniform distribution is mostly based on simplicity criteria although the binomial may be interpreted as a counting process of selection among the k categories and the uniform distribution may be introduced as the most extreme and uninformative case among all discrete alternatives. In a wider class of models proposed here the rather restrictive binomial model is replaced by more flexible ordinal models while the uniform distribution as an uninformative distribution is retained. The general mixture model we consider has the form

$$P(R_i = r|\boldsymbol{x}_i) = \pi_i P_M(Y_i = r|\boldsymbol{x}_i) + (1 - \pi_i) P_U(U_i = r), \tag{3}$$

where R_i represents the observed response and Y_i, U_i are the unobserved random variables taking values from $\{1, \ldots, k\}$. The distribution of Y_i is determined by $P_M(Y_i = r|\boldsymbol{x}_i)$, which can be any ordinal model M, whereas $P_U(U_i = r) = 1/k$ represents the uniform distribution. We refer to the general model (3) as a CUP model for the Combination of Uniform and Preference structures.

For the specification of the latent variable Y_i one can use models that are in common use in ordinal regression, in particular, cumulative type and adjacent categories type models, which have already been considered by McCullagh (1980). The *cumulative model* has the general form

$$P(Y_i \leq r|\boldsymbol{x}_i) = F\left(\gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}\right), \quad r = 1, \ldots, k-1,$$

where $F(.)$ is a given cumulative distribution function and $-\infty = \gamma_{00} < \gamma_{01} < \cdots < \gamma_{0k} = \infty$. This model is obtained by assuming that a latent regression model $\tilde{Y}_i = -\boldsymbol{x}_i^T\boldsymbol{\gamma} + \epsilon$ holds, where ϵ is a noise variable with distribution function F. If we consider the link between the observable categories and the latent variable given by

$$Y_i = r \;\Leftrightarrow\; \gamma_{0,r-1} < \tilde{Y}_i \leq \gamma_{0r}, \quad r = 1, 2, \ldots, k$$

it is straightforward to derive the model. The underlying response variable approach has been widely used to model ordinal response data, and many extensions have been proposed, see, for example, Cox (1995), Brant (1990), Peterson and Harrell (1990), Nair (1987) and Liu and Agresti (2005).

The most widely used model from this class of models is the cumulative logit model, which uses the logistic distribution $F(.)$ It is also called *proportional odds model* and has the form

$$\log\left(\frac{P(Y_i \leq r|\boldsymbol{x}_i)}{P(Y_i > r|\boldsymbol{x}_i)}\right) = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}, \quad r = 1, \ldots, k-1.$$

An alternative choice is the *adjacent categories model* given by

$$P(Y_i = r+1|Y_i \in \{r, r+1\}, \boldsymbol{x}_i) = F\left(\gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}\right), \quad r = 1, \ldots, k-1.$$

The specific model that uses the logistic distribution is the *adjacent categories logit model*

$$\log\left(\frac{P(Y_i = r+1|\boldsymbol{x}_i)}{P(Y_i = r|\boldsymbol{x}_i)}\right) = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}, \quad r = 1, \ldots, k-1.$$

Also sequential models or other ordinal models could be useful. For a discussion of these classes of ordinal models, see, for example, Tutz (2012).

It should be noted that the CUB model is a special case that uses the binomial distribution in the preference part. The use of models like the cumulative or adjacent categories model is attractive because it adds flexibility to the model. For example, the probability distribution of the binomial model is strictly unimodal, in contrast to the cumulative and the adjacent categories model, which allow for all forms of distributions by including the flexible intercepts $\gamma_{01}, \ldots, \gamma_{0k}$. Moreover, cumulative and adjacent categories models are the most widely used models for ordinal data, but an additional uncertainty component seems not to have been used for these models

before. As will be shown parameter estimates are biased if the uncertainty component is ignored. In the following we will use the abbreviations CUP(c) and CUP(a) if the structural response model in the mixture is the cumulative or the adjacent categories model, respectively.

In both models, the effect of the explanatory variables in the model that specifies preference is contained in the linear predictors, which have the form $\eta_{ir} = \gamma_{0r} + \boldsymbol{x}_i^T \boldsymbol{\gamma}$. Therefore, the specification of the linear predictor replaces the assumption logit $(\xi_i) = \boldsymbol{x}_i^T \boldsymbol{\gamma}$, which specifies the dependence of CUB model parameters on covariates. The dependence of the uncertainty component on covariates is modelled in the same way as in CUB models, namely by logit$(\pi_i) = \boldsymbol{z}_i^T \boldsymbol{\beta}$, where $\boldsymbol{z}_i$ can be identical to or partially overlap with $\boldsymbol{x}_i$.

For CUB models the link between the uncertainty component and heterogeneity measured by the Gini index was systematically investigated by Iannario (2012c). A similar link holds for the CUP model. The Gini index for any mixture model is given by $G_{MIX} = \pi^2 G_M + 1 - \pi^2$, where M denotes the ordinal model used in the mixture and the mixture model itself is given by (3). The maximal heterogeneity is obtained for the uniform distribution, that is, $G_{UNI} = 1$. Thus the Gini index can also be given by

$$G_{MIX} = G_{UNI} - \pi^2 (1 - G_M).$$

Considered as a function with argument π it decreases quadratically with increasing probability π from the maximal value to G_M. Therefore, the mixture model has an heterogeneity index between the uniform model and model M, but for $\pi < 1$ is larger than the Gini index for the model M. That means, in the mixture model the probabilities of response categories are more evenly distributed than in model M. By assuming a mixture the basic ordinal model M is shrunk toward the uniform model.

For illustration Fig. 2 shows the Gini index as a function of the weight of the uncertainty component $1 - \pi$. The underlying model is a simple cumulative model with ten categories and a binary predictor with coefficient γ. In the left panel the

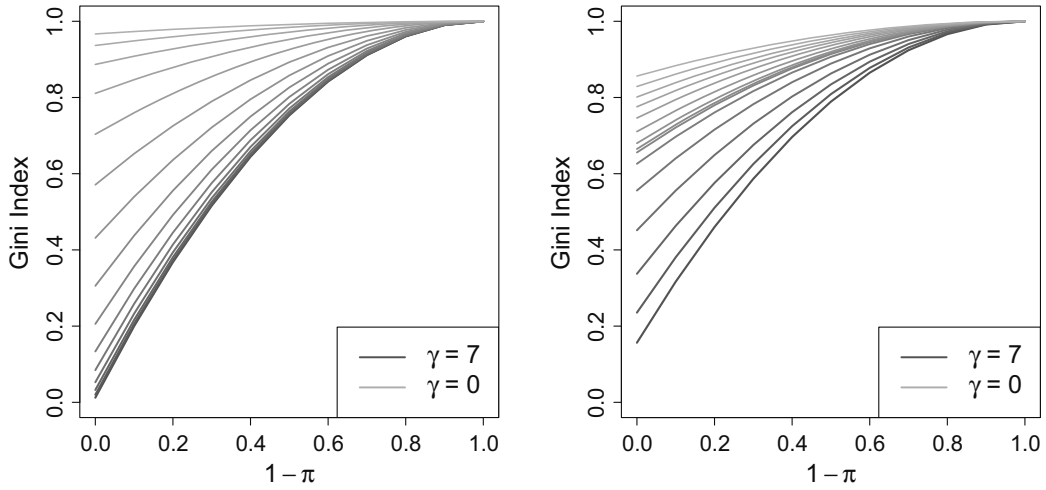


Fig. 2 Gini heterogeneity measures for CUP models as functions of $1 - \pi$ for two sets of thresholds and several values of effect strength γ

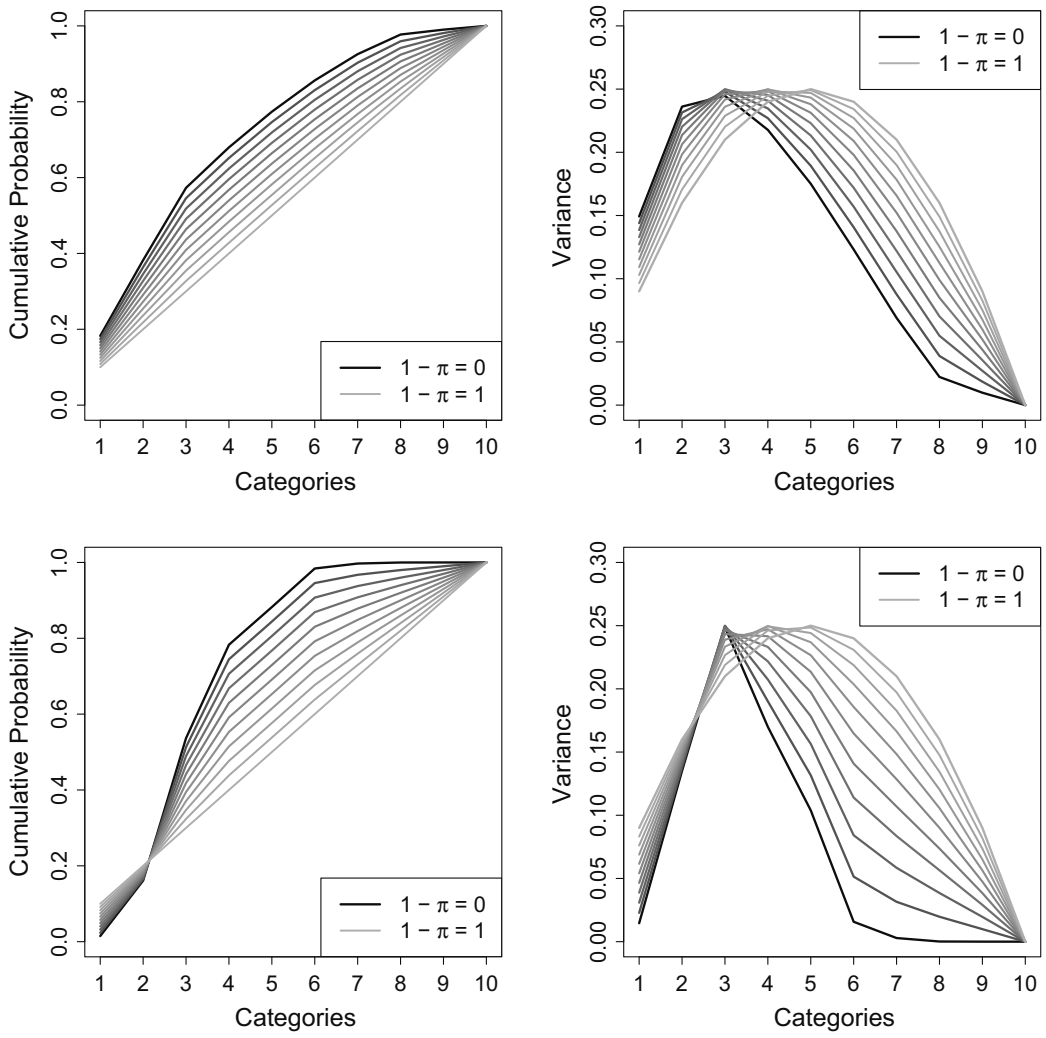


Fig. 3 Cumulative probabilities and variances for CUP models with ten categories for two sets of thresholds ($\gamma = 0.3$)

thresholds were $-1.799, -0.780, -0.006, 0.452, 0.929, 1.484, 2.219, 3.458, 4.304$ in the right panel $-4.497, -1.950, -0.153, 0.983, 1.714, 3.820, 5.547, 8.645, 10.760$. It is seen that the Gini index increases with growing uncertainty $(1 - \pi)$, which also holds if other parameter values are chosen. The increase tends to be strong for strong effects of the predictor and weak if the predictor is less influential.

When considering the effect of uncertainty on the variance it can not be recommended to examine the variance of $Y \in \{1, \ldots, k\}$ itself if one takes the ordinal scale level of Y seriously. Therefore, we consider the variances of the binary variables $Y_r = I(Y \leq r), r = 1, \ldots, k$, which are explicitly modelled within the cumulative model framework. Figure 3 shows the cumulative probabilities $P(Y \leq r)$ (left panels) and the variances of the corresponding variables Y_r (right panels) for the two examples with ten categories considered in Fig. 2. We chose $\gamma = 0.3$, for other values of γ one obtains similar pictures. For increasing uncertainty $1 - \pi$ the cumulative probabilities tend to lie on a straight line. In the same way the variances have a fixed symmetric shape if $1-\pi = 1$. For $1-\pi$ close to 0 one obtains different curves, which depend on the thresh-

olds. The variance curves are not monotone with a peak that depends on the thresholds. In the second set of thresholds (second row) the probability for categories above 5 is very small and therefore the curves decrease strongly above category 3. As for the CUB model the variances considered as a function of $1 - \pi$ (for fixed category) have not to be monotone, however, for small and large categories they typically are monotone.

In mixture models typically identifiability issues arise, see, for example, Follmann and Lambert (1991) and Grün and Leisch (2008). For the CUB model identifiability has been investigated by Iannario (2010). Since CUB models use in the structured part the rather restrictive binomial distribution identifiability is obtained under weak conditions. The extended class of models uses a much more flexible specification in the structured preference part. As a consequence, if no covariates are present, for example, the cumulative model is a saturated model and therefore the mixture model is not identifiable. For the extended model a certain richness of the covariate structure is needed to be identifiable. We consider in the appendix the cumulative CUP model and show that it is identifiable if continuous covariates are in the predictor.

2.3 Estimation

In mixture models, estimation issues can be pursued by exploiting the EM algorithm as proposed by Dempster et al. (1977) and used with special reference to mixtures by McLachlan and Peel (2000). In this context, estimation and tests are obtained by asymptotically efficient procedures based on maximum likelihood methods. For readability we give the used EM algorithm in the appendix. Specific results for CUB models were given by Piccolo (2006).

3 Effect strength in mixture models

If the response is affected by an additional random component that is modelled by a uniform distribution within the mixture framework, the effects of explanatory variables will differ from the effects found by the fitting of a traditional response model. For simplicity we consider in the following the binary logit model. In this case the cumulative, the sequential and the adjacent categories models are equivalent. Then the probability of response category 1, denoted by $p(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x})$, is given by

$$p(\boldsymbol{x}) = \pi p_M(\boldsymbol{x}) + (1 - \pi)/2,$$

where $p_M(\boldsymbol{x}) = \exp(\gamma_0 + \boldsymbol{x}^T\boldsymbol{\gamma})/(1 + \exp(\gamma_0 + \boldsymbol{x}^T\boldsymbol{\gamma}))$ denotes the probability of the logit model. If $\pi < 1$, that is, in the presence of the uncertainty component, one obtains

$$|p(\boldsymbol{x}) - 0.5| = |\pi p_M(\boldsymbol{x}) + (1 - \pi)/2 - 0.5| = \pi \ |p_M(\boldsymbol{x}) - 0.5| < |p_M(\boldsymbol{x}) - 0.5| . \tag{4}$$

That means the true probabilities $p(\boldsymbol{x})$ are closer to 0.5 than the probabilities in the structured component $p_M(\boldsymbol{x})$. This shrinkage toward 0.5 means that the effect strength $\boldsymbol{\gamma}$ tends to be underestimated if the uncertainty component is ignored. More concrete, Eq. (4) shows that the distance between the probability $p(\boldsymbol{x})$ of the data generating process and 0.5 is equal to $\pi|p_M(\boldsymbol{x}) - 0.5|$. Therefore, the distance reduces by the

factor π. It is essential that the reduction is proportional to the distance between the probability and 0.5. That means that a value $p_M(\boldsymbol{x}_1)$ that is farer away from 0.5 changes stronger than a value $p_M(\boldsymbol{x}_1)$ that is closer to 0.5. The consequence is that one observes a weaker effect strength in the mixture model than is present in the model M. In the simplest case one has a binary explanatory variable $x \in \{0, 1\}$. Then both models M and the mixture model are saturated and one can compute the parameter β for the model M and the corresponding parameter $\tilde{\beta}$ that is found when using probabilities $p(x)$. Because $|p(x) - 0.5| = \pi |p_M(x) - 0.5|$ the increase (or decrease) from $p(0)$ to $p(1)$ is always larger than the increase (or decrease) from $p_M(0)$ to $p_M(1)$. Therefore, one obtains $|\tilde{\beta}| < |\beta|$. The case of binary explanatory variables is not interesting by itself, but the tendency to underestimate the effect strengths holds in the general case.

Before considering the effect in the general model with ordered categories it should be noted that the inclusion of an uncertainty component has one other effect. Since the probability $p(\boldsymbol{x})$ of the data generating process is closer to 0.5 than the probability $p_M(\boldsymbol{x})$ also the variance is larger than in the logit model M. Therefore, the inclusion of a uniform component is one way of modelling *overdispersion*.

For ordinal models with $k > 2$ and a cumulative logit model one gets similar results for the cumulative probability $p_r(\boldsymbol{x}) = P(Y \leq r|\boldsymbol{x})$, which is given by

$$p_r(\boldsymbol{x}) = \pi p_{M,r}(\boldsymbol{x}) + (1 - \pi) r/k,$$

where $p_{M,r}(\boldsymbol{x}) = \exp(\gamma_{0r} + \boldsymbol{x}^T \boldsymbol{\gamma})/(1 + \exp(\gamma_{0r} + \boldsymbol{x}^T \boldsymbol{\gamma}))$ specifies a binary logit model. That means one obtains a shrinkage toward r/k. It is easily derived that now $|p_r(\boldsymbol{x}) - r/k| = \pi |p_{M,r}(\boldsymbol{x}) - r/k|$ with the same consequence as in the binary model, namely that the effect strength γ tends to be underestimated if the mixture component is neglected. What differs from the binary model is that one does not necessarily model overdispersion. Of course, for k even and $r = k/2$ one has the same effect as in the binary model considered previously: one has stronger variability than assumed in the model M and therefore models overdispersion. But this has not to hold for all values of r. For example, if $k = 10$, one obtains for $r = 1$ shrinkage toward 0.1. If $p_{M,r}(x)$ is larger than 0.1, then the shrinkage toward 0.1 means that the variance is smaller than in the model without a uniform mixture component. Therefore, in terms of the cumulative probabilities one might model underdispersion in the sense that the mixture model allows to model smaller variance than the pure model with $\pi = 1$.

Although estimation procedures will be considered later, we consider a small example to illustrate the shrinkage effect. Table 1 shows data that have been analysed previously by Mehta et al. (1984). For patients with acute rheumatoid arthritis a new agent was compared with an active control. Each patient was evaluated on a five-point assessment scale ranging from "much improved" to "much worse." Table 2 shows the corresponding estimates for the mixture model with a cumulative logit model as the structuring component CUP(c) and the simple cumulative logit model. It is seen that the effect strength is 0.291 for the cumulative model but 0.394 for the cumulative mixture model. Thus if the mixture component is omitted one obtains a weaker effect of treatment. The difference between effect strengths is rather large because the uniform distribution is included with the rather large probability 0.294.

Table 1 Clinical trial of a new agent and an active control (Mehta et al. 1984).

Drug	Global assessment				
	Much improvement	Improvement	No change	Worse	Much worse
New agent	24	37	21	19	6
Active control	11	51	22	21	7

Table 2 Model fits for arthritis data with explanatory variable drug; fitted models are mixture with a cumulative model [CUP(c)] and a simple cumulative model without uncertainty

	CUP(c)	Cumulative model
Intercept: 1	−1.945	−1.802
Intercept: 2	0.371	0.115
Intercept: 3	1.385	1.008
Intercept: 4	7.668	2.631
Drug	0.394	0.291
Prob (uniform)	0.294	0

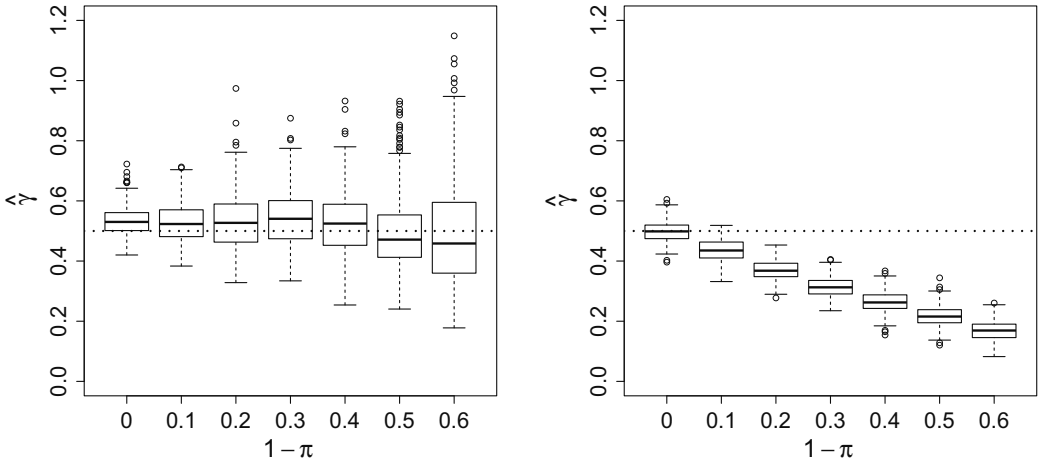


Fig. 4 Simulation with the data generating model being a cumulative mixture model; *left panel* shows the parameter estimates when a cumulative mixture model is fitted, *right panel* shows the estimates if a simple cumulative model is fitted

To further investigate the bias of the estimate if the mixture component is neglected we give the results of a small simulation study. Let the data be generated from a mixture model with the cumulative model in the mixture given by the model fitted for the clinical trial data. We use the thresholds given in Table 2 with effect strength 0.5 and vary the probability $1 - \pi$, that is, the probability of uncertainty in the mixture. The left panel of Fig. 4 shows the estimated parameters when a cumulative CUP model is fitted. The true parameter value is included as a horizontal line. It is seen that the estimates are almost unbiased. Overall the estimation works well with increasing variability if the uncertainty component gets stronger. The results change dramatically if one fits a cumulative model and therefore ignores the uncertainty component (right

panel of Fig. 4). It is seen that the true parameter is strongly underestimated with the bias getting stronger with increasing importance of the uncertainty component.

The main point of the illustrations is that when the data generating model is a mixture model of the form considered here one tends to underestimate the effects of explanatory variables. The effect is similar to what is found in binary (and ordinal) random intercept models. Let repeated measurements on individual i be given by $y_{i1}, \ldots, y_{im}, y_{it} \in \{0, 1\}$ with covariates $\boldsymbol{x}_i$. Then the random intercept model assumes $P(y_{it} = 1|\boldsymbol{x}_i, b_i) = h(b_i + \boldsymbol{x}_i^T \boldsymbol{\gamma})$, where $h(.)$ is a response function and b_i is a subject-specific random effect, typically assumed to be normally distributed, $b_i \sim N(0, \sigma_{\boldsymbol{b}}^2)$. The parameter $\boldsymbol{\gamma}$ contains the *conditional* effect of the explanatory variable *given the random effect* b_i. If one considers the *marginal* model $P(y_{it} = 1|\boldsymbol{x}_{it}) = \int P(y_{it} = 1|\boldsymbol{x}_{it}, b_i)\, p(b_i) db_i$, effects tend to be weaker, see, for example, Caffo et al. (2007). Because the models are non-linear the omission of the random effects yields estimates of parameters that are closer to zero than the actual parameters. Marginal effects are attenuated as compared to the conditional effects.

Interpretation of the parameters in the mixture model is not so straightforward but effects can be interpreted in a similar way as conditional effects in random effects models. Let us consider the proportional odds model for the structured response in the mixture model. Let C denote the latent class; $C = 1$ denotes that the choice made by the individual is deliberate and determined by the proportional odds model M; $C = 0$ means that the choice is made in random mode, determined by the uniform distribution. The mixture is determined by the weights $\pi, 1 - \pi$. Then the parameters of the proportional odds model determine the response given $C = 1$, that is, $P(Y \leq r|\boldsymbol{x}, C = 1) = p_M(\boldsymbol{x}) = \exp(\gamma_{0r} + \boldsymbol{x}^T \boldsymbol{\gamma})/(1 + \exp(\gamma_{0r} + \boldsymbol{x}^T \boldsymbol{\gamma}))$. If one compares two individuals that differ in the variable x_j by one unit one obtains for the cumulative odds given $C = 1$

$$\frac{P(Y \leq r|x_j + 1, \boldsymbol{x}_J, C = 1)/P(Y > r|x_j + 1, \boldsymbol{x}_J, C = 1)}{P(Y \leq r|x_j, \boldsymbol{x}_J, C = 1)/P(Y > r|x_j, \boldsymbol{x}_J, C = 1)} = \exp(\gamma_j),$$

where $\boldsymbol{x}_J$ denotes a vector of covariates without the $j-$th component . Thus the parameter contains the effect of explanatory variable x given both individuals make a deliberate choice, that is, $C = 1$. In that sense the effect is conditional on the action mode $C = 1$. Given this action mode the interpretation is the same as in the common proportional odds models. Ignoring the uncertainty component yields attenuated effects.

4 An illustrative example

To illustrate the effects in a cumulative CUP we consider data from the Survey on Household Income and Wealth (SHIW) by the Bank of Italy, for earlier use of the data see Gambacorta and Iannario (2013). In the analysis presented in Table 3 the response is the happiness index indicating the overall life well-being measured on a Likert Scale from 1 (very unhappy) to 10 (very happy). As covariates the following factors were chosen: the marital status, the place of living, the general degree of confidence in other

Table 3 Parameter estimates and standard errors based on bootstrap (BS.SE) for the SHIW study

Covariates	CUP(c)		Cumulative		CUB	
	est.	BS.SE	est.	se	est.	BS.SE
Constant (β_0)	0.375	0.146			0.419	0.460
Marital status: single	0.579	0.182			0.604	0.214
Marital status: separated	0.866	0.192			1.224	0.256
Marital status: widow	0.954	0.177			1.261	0.212
Living: centre of Italy	0.809	0.171			1.039	0.177
Living: south of Italy	0.425	0.132			0.487	0.156
Confidence in people	0.092	0.024			0.097	0.025
Interview atmosphere	−0.162	0.028			−0.185	0.054
Marital status: single	1.208	0.173	0.356	0.089	0.460	0.066
Marital status: separated	1.340	0.178	0.276	0.108	0.509	0.066
Marital status: widow	1.442	0.168	0.327	0.085	0.567	0.057
Living: centre of Italy	−0.585	0.140	−0.762	0.075	−0.240	0.050
Living: south of Italy	0.347	0.127	−0.087	0.068	0.124	0.047
Confidence in people	−0.107	0.044	−0.080	0.012	−0.041	0.012
Income sufficient	−0.301	0.050	−0.094	0.024	−0.110	0.017
Interview atmosphere	−0.277	0.044	−0.092	0.020	−0.094	0.014
Citizenship: foreign	0.845	0.368	0.243	0.153	0.342	0.123
Age (centered)	0.019	0.005	0.004	0.002	0.006	0.002
Prob(uniform)	0.464		0		0.458	

In the upper part the parameters for the mixture probabilities π_i are shown, the lower part shows the parameters of the cumulative model used to model the preference structure

people (1 to 10), the atmosphere the interview took place in [1 (low) to 10 (high)], the citizenship and the age. The respondents were also asked about their assessment if the household income is sufficient to see the family through to the end of the month rated from 1 (with great difficulty) to 5 (very easily). The analysis is based on a subset with 3816 respondents of the SHIW of 2010. We fitted a cumulative CUP model with logit link and explanatory variables in the cumulative part as well as in the logistic model that determines the mixture probability. In addition we fitted a simple cumulative logit model (proportional odds model) without a mixture component and the CUB model. The standard errors of the coefficients are obtained by 500 non-parametric bootstrap samples (Efron and Tibshirani 1994). We also used the parametric bootstrap and found that it tends to yield smaller standard errors. Therefore a more conservative strategy is to use the non-parametric bootstrap. For non-mixture models as, for example, the cumulative model we always use the usual standard errors obtained from the inverse of the Fisher matrix. The results are given in Table 3.

It is seen that the uncertainty component is very strong with $1 - \bar{\pi} = 0.458$ for the CUB model and $1 - \bar{\pi} = 0.464$ for the cumulative mixture model, where $\bar{\pi} = 1/n \sum_{i=1}^{n} 1/(1 + e^{-z_i^T \boldsymbol{\beta}})$ is the mean value over all the probabilities of the observations. In the tables $1 - \bar{\pi}$ is always denoted by Prob(uniform). As in the

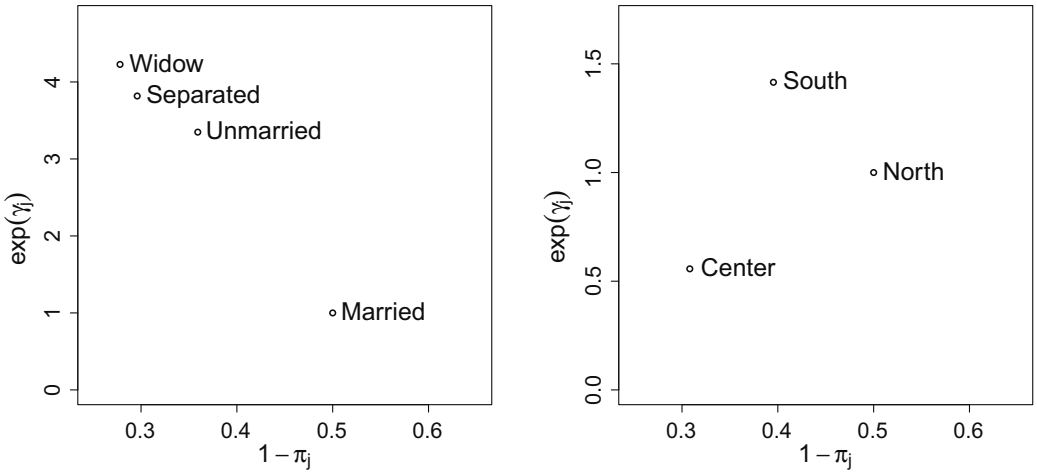


Fig. 5 Effects of the categorical covariates marital status (*left*) and area of living (*right*) in the structure and uncertainty component

previous example it is seen that the estimated effects in the cumulative model part of the mixture model are much stronger than the effects found in the simple cumulative model. We only included effects that have been found to be influential in previous studies (Gambacorta and Iannario 2013) and do not give the threshold parameters.

A tool that has also been used for CUB models is the visualization of effects of explanatory variables that are included in the preference part of the model and also determine the uncertainty. However, alternative specifications are needed to link the effects on the preference part with the effects on uncertainty. Therefore, a specific form of the cumulative logistic model that determines the preference component of the mixture has to be used. A form of the model that has been used for visualization of effects by Tutz and Schauberger (2013) and allows for easy interpretation of the effects on the response is

$$\frac{P(Y \leq r|\boldsymbol{x})}{P(Y > r|\boldsymbol{x})} = \exp(\gamma_{0r}) \exp(\boldsymbol{x}^T \boldsymbol{\gamma}) = e^{\gamma_{0r}} (e^{\gamma_1})^{x_1} \cdots (e^{\gamma_p})^{x_p}, \quad r = 1, \ldots, k-1.$$

That means that the odds of preferring categories $\{1, \ldots, r\}$ over categories $\{r + 1, \ldots, k\}$ are modified by the factor e^{γ_j} if the jth variable is increase by one unit. It is important that the factor is the same for all categories and therefore characterizes the effect of the covariates in a unique way. This is essentially the proportional odds assumption that gives the model its name, see McCullagh (1980) for an extensive discussion of the proportional odds property. Since e^{γ_j} contains the effect of the preference part it can be used to visualize the effect together with the uncertainty, which is contained in the model $\text{logit}(\pi) = \boldsymbol{z}^T \boldsymbol{\beta}$. In Fig. 5 the factor e^{γ_j} is plotted against the strength of the uncertainty $1 - \pi_j$ for the explanatory variables marital status and area of living. To obtain a scale for the uncertainty the other variables are set to fixed values, in particular, confidence and atmosphere are set to category 1, income to 3 and all other variables to zero. Since in the cumulative model large values of $\exp(\boldsymbol{x}^T \boldsymbol{\gamma})$ indicate preference for low response categories, large values indicate unhappiness. It is seen from Fig. 5 that the marital status "widow" corresponds to

high values of unhappiness and high certainty (small $1 - \pi_j$). In contrast, the status "married" indicates happiness but a large amount of uncertainty in the response. From the plot for the variable area it is seen that the people living in the north have large uncertainty and medium happiness whereas people from the south tend to categories that indicate unhappiness with a middle level of uncertainty.

In this application for simplicity we used ordinal covariates like the Likert scale for the degree of confidence in other people as metric covariates. Therefore, it is assumed that respondents answer on a metric scale level. This assumption can be avoided by using dummy variables for each level of the covariate. However, then one obtains 9 parameters for a variable with 10 categories and estimation gets unstable. More recently penalty approaches to account for the ordinal nature of the covariates have been proposed (Gertheiss and Tutz 2009; Tutz and Gertheiss 2014). The principle is to use penalized likelihood methods to reduce the variation of effects of adjacent categories. Although the methods yield stable estimates a tuning parameter has to be chosen. Moreover, special software is needed, which is available for GLMs but not yet for mixture models.

5 Comparison of models

In this section we consider the usefulness of including the uncertainty component in the traditional cumulative models and also compare CUP and CUB proposals by use of real data sets. First we briefly discuss criteria for the comparison of models.

5.1 Criteria

Comparison of models is not straightforward since in general the models are not nested. But even for nested models, for example, when comparing a cumulative mixture model and a pure cumulative model, one can not simply use likelihood ratio tests because one is at the boundary of the parameter space. So one cannot expect the likelihood ratio tests to have the usual $\chi^2_{(1)}$-distribution, compare Böhning et al. (1994).

Alternatives are information criteria as the AIC and BIC given by

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

where m is the number of model parameters, n is the number of observations and $l(\hat{\boldsymbol{\theta}})$ is the log-likelihood function computed at the maximum of the estimated parameter vector $\boldsymbol{\theta}$. AIC and BIC have the advantage that they can be used for non-nested models, therefore they allow to compare, for example, an adjacent or cumulative model to the CUB model. Information criteria are in common use in mixture models although no strong foundation seems available. Leroux (1992) gave some justification for the use of information criteria but it refers to very special cases only. Therefore alternative ways to compare models seem warranted.

A more data driven strategy is the evaluation of the predictive performance. In particular we will consider the deviance as a measure of the discrepancy between data and fit. For the multinomially distributed response one can distinguish two cases. One can group all observations for a fixed value of the explanatory variables obtaining

the distribution $\boldsymbol{r}_i^T = (r_{i1}, \ldots, r_{ik}) \sim \mathrm{M}(n_i, \boldsymbol{p}_i)$, $i = 1, \ldots, N$, where N is the number of distinct values of the explanatory variables, n_i is the number of observations for the i-th value of the explanatory variables. The true underlying probabilities are $\boldsymbol{p}_i^T = (p_{i1}, \ldots, p_{ik})$ and the corresponding estimates without assuming a model are the relative frequencies $(f_{i1}, \ldots, f_{ik})$. Then the deviance for the multinomial distribution has the general form

$$D = 2 \sum_{i=1}^{N} n_i \sum_{r=1}^{k} f_{ir} \log \left(\frac{f_{ir}}{\hat{p}_{ir}} \right),$$

where $\hat{p}_{ir}$ is the estimated probability of category r. In this grouped form it uses that n_i observations are available for a fixed value of the explanatory variable and for GLMs asymptotic distributions are available for $(n_i/N \to \lambda_i \in (0, 1))$ (Fahrmeir and Tutz 2001). If one does not group data, but works with single observations one uses $\boldsymbol{r}_i^T \sim \mathrm{M}(1, \boldsymbol{p}_i)$, $i = 1, \ldots, n$ and obtains the form

$$D = 2 \sum_{i=1}^{n} \sum_{l=1}^{k} r_{il} \log \left(\frac{r_{il}}{\hat{p}_{il}} \right) = -2 \sum_{i=1}^{n} \log(\hat{p}_{i R_i}),$$

where R_i denotes the observation in the categories, that is, $R_i \in \{1, \ldots, k\}$.

In both forms, grouped or un-grouped, the deviance measures the discrepancy between data and fit. It can be used as a predictive measure. Let the data be split into a learning set and a validation set. The model is fitted on the learning set and then one computes the deviance for all the observations in the validation set $(R_i^{(V)}, \boldsymbol{x}_i^{(V)})$, $i = 1, \ldots, n_V$. In the un-grouped form one obtains for the averaged deviance

$$D/n_V = -2 \sum_{i=1}^{n_V} \log(\hat{p}_{i R_i^{(V)}})/n_V,$$

where $\hat{p}_{il}$ is the estimated probability of category l at value $\boldsymbol{x}_i^{(V)}$. It is also known as the logarithmic score. A criticism of scores like the logarithmic score is that the predictive distribution $\hat{\boldsymbol{p}}$ is only evaluated at the value of the observation. Therefore, it takes not the whole predictive distribution into account. In the case of an ordinal response measures that make use of the whole predictive distribution can be derived from the continuous ranked probability score approach discussed by Gneiting and Raftery (2007). For categorical responses one obtains the averaged value

$$L_{RPS}/n_V = \sum_{i=1}^{n_V} \sum_{r} (\hat{p}_i(r) - I(R_i \le r))^2/n_V, \tag{5}$$

where $\hat{p}_i(r) = \hat{p}_{i1} + \cdots + \hat{p}_{ir}$ is the estimated cumulative probability at value $\boldsymbol{x}_i^{(V)}$ and $I(.)$ is the indicator function. It is a sum over quadratic (or Brier) scores for binary data and takes the closeness between the whole estimated distribution and the observed

value into account. For a discussion of measures for the closeness of data and fit see also, with the focus on categorical data, Tutz (2012), Chapter 15.

In the applications to follow the splitting into training and test data sets is done in the form of 10-fold cross-validation. The data set was split into ten sub sets, consecutively nine data sets were used to fit the models and the left out data set was used to evaluate the predictive performance.

5.2 Empirical studies

The models that are used in the applications are

- The cumulative model (without uncertainty),
- CUP(c): the cumulative model with uncertainty component,
- The adjacent categories model model (without uncertainty),
- CUP(a): the adjacent categories with uncertainty component,
- CUB: the binomial with uncertainty component.

5.2.1 Income and wealth

For the Survey on Household Income and Wealth (SHIW) considered in the previous section the performance measures for selected models are given in Table 4. Since data were split into training and test data by 10-fold cross validation one obtains the logScore and RankedScore for each observation of the data set. We give the mean of the observed values and in brackets the standard deviations (sd). In addition, p-values are given. They refer to the hypothesis that the corresponding score is the same for the model under consideration and the CUB model. Therefore, the CUB model is used as a reference model. The used test statistic is the t-test. It is seen that the cumulative mixture model performs best in terms of AIC, BIC, deviance and logScore. The ranked score is the same for CUP(c) and CUP(a). The relevance of the mixture component is underlined by the strong reduction of AIC; the value of the AIC for the mixture model (16218) is much smaller than the AIC for the simple cumulative model

Table 4 Results for the SHIW study

Covariates	CUP(c)	Cumulative	CUP(a)	Adjacent	CUB
Probability(uniform)	0.464	0	0.491	0	0.458
Deviance	16164	16434	16185	16497	16311
AIC	16218	16472	16239	16535	16349
BIC	16387	16591	16408	16654	16467
logScore	4.250	4.307	4.256	4.323	4.284
(sd)	(0.050)	(0.051)	(0.051)	(0.046)	(0.041)
(p value)	(0.000)	(0.051)	(0.000)	(0.002)	
RankedScore	1.306	1.310	1.306	1.311	1.307
(sd)	(0.034)	(0.033)	(0.033)	(0.030)	(0.031)
(p value)	(0.202)	(0.411)	(0.181)	(0.245)	

Table 5 Qualitative assessment of subjective survival probabilities Pr(S)

Category	Expressed probability	Interpretation of the perception
1	$0.00 \leq \Pr(S) \leq 0.05$	Impossible/almost impossible
2	$0.05 < \Pr(S) \leq 0.25$	Low
3	$0.25 < \Pr(S) \leq 0.45$	Moderately low
4	$0.45 < \Pr(S) \leq 0.55$	About fifty/fifty
5	$0.55 < \Pr(S) \leq 0.75$	Moderately high
6	$0.75 < \Pr(S) \leq 0.95$	High
7	$0.95 < \Pr(S) \leq 1.00$	Sure/almost sure

(16472). The same reduction is found when an uncertainty component is included in the adjacent categories model. For the ranked score the performance of all models is very similar and there is no significant difference to the CUB model. However, there is a significant difference in terms of the logScore, in particular CUP(c) and CUP(a) show better performance. Overall, the cumulative mixture model with a substantial amount of uncertainty is to be preferred.

5.2.2 PLUS study

In the Participation, Labour and Unemployment Survey (PLUS) carried out by ISFOL (Institute for training of workers, Ministry of Labour and Welfare, Italy), the participants were asked to rate their probability to reach the age of 75. They chose a value between 0 for a impossible event and 100 for a certain event. Because of rounding effects ordered categories instead of the observed continuous values are to be preferred as suggested by Iannario and Piccolo (2010): see Table 5. The data consists of 20,184 individuals from the survey wave of 2006 and includes several more covariates such as gender (1: female, 0: male), age, marital status (widowed, divorced, married/single) and employment status (1: worker, 0: no-worker). Table 6 shows the results with all of the explanatory variables. In this application the uncertainty component is rather weak ($1 - \bar{\pi} = 0.100$ for the cumulative mixture model, 0.099 for the adjacent categories mixture model and 0.137 for the CUB). Nevertheless the inclusion of the uncertainty component reduces the AIC and the BIC distinctly. It is seen that the cumulative and the adjacent categories mixture models perform better than the models without uncertainty and the CUB with regard to all performance measures. This is also supported by the predictive measures, which show that the models perform significantly better than the CUB model.

5.2.3 Allbus

In the German General Social Survey ALLBUS data on behavior, attitudes and social structure in Germany are collected. 3480 persons answered the questionnaire in 2012. In the present study the respondents rated their trust in the health care system on a scale from 1 (not at all) to 7 (very much). In addition they give their assessment of the own state of health from 1 (very good) to 5 (poor) and their overall life satisfaction from

Table 6 Results for the PLUS study

Covariates	CUP(c)		Cumulative		CUP(a)		Adjacent		CUB	
	est.	BS.se	est.	se	est.	BS.se	est.	se	est.	BS.se
Intercept(β_0)	1.811	0.085			1.823	0.082			1.511	0.078
Female	0.046	0.083			0.092	0.085			0.032	0.080
Age	−1.064	0.144			−1.044	0.132			-1.310	0.125
Age^2	2.713	0.344			2.550	0.334			2.534	0.350
Female	0.128	0.030	0.105	0.027	0.082	0.016	0.038	0.011	0.104	0.024
Divorced	0.308	0.103	0.276	0.079	0.159	0.052	0.100	0.032	0.246	0.076
Widowed	0.360	0.139	0.412	0.114	0.192	0.073	0.179	0.044	0.290	0.108
Work	−0.074	0.032	−0.049	0.028	−0.031	0.018	−0.013	0.012	−0.047	0.025
Age	−0.242	0.037	−0.085	0.033	−0.098	0.021	0.026	0.015	−0.226	0.033
Age^2	−0.716	0.099	−0.925	0.092	−0.426	0.056	−0.458	0.040	−0.559	0.085
Prob(uniform)	0.100		0		0.099		0		0.137	
Deviance	59601		59729		59612		59716		60381	
AIC	59633		59753		59644		59740		60403	
BIC	59759		59847		59771		59835		60490	
logScore	2.9545		2.9592		2.9551		2.9585		2.9927	
(sd)	(0.029)		(0.027)		(0.029)		(0.028)		(0.029)	
(p value)	(0.000)		(0.000)		(0.000)		(0.000)			
RankedScore	0.6726		0.6734		0.6727		0.6733		0.6757	
(sd)	(0.014)		(0.015)		(0.014)		(0.014)		(0.014)	
(p value)	(0.000)		(0.000)		(0.000)		(0.000)			

Table 7 Model results for the Allbus data

Covariates	CUP(c)		Cumulative		CUP(a)		Adjacent		CUB	
	est.	BS.se	est.	se	est.	BS.se	est.	se	est.	BS.se
Intercept(β_0)	4.976	0.895			3.461	0.615			3.663	0.461
Poor health	−1.051	0.222			−0.701	0.168			−0.749	0.144
German	1.116	0.236	0.850	0.162	0.598	0.135	0.321	0.072	0.542	0.111
Income	0.058	0.047	0.005	0.020	0.030	0.024	0.015	0.008	0.031	0.016
Age (centred)	−0.007	0.047	0.005	0.002	−0.004	0.001	−0.002	0.001	−0.004	0.001
Region: east	−0.372	0.093	−0.318	0.071	−0.208	0.050	−0.125	0.029	−0.190	0.042
Life satisfaction	−0.193	0.034	−0.168	0.019	−0.104	0.021	−0.061	0.008	−0.092	0.015
Prob(uniform)	0.122		0		0.172		0		0.164	
Deviance	9925		9976		9928		9990		9942	
AIC	9951		9998		9954		10012		9958	
BIC	10029		10064		10032		10078		10005	
logScore	3.380		3.389		3.381		3.394		3.383	
(sd)	(0.068)		(0.072)		(0.067)		(0.070)		(0.064)	
(p value)	(0.354)		(0.299)		(0.553)		(0.092)			
RankedScore	0.751		0.752		0.751		0.752		0.751	
(sd)	(0.034)		(0.035)		(0.033)		(0.034)		(0.033)	
(p value)	(0.909)		(0.718)		(0.694)		(0.495)			

0 (very unhappy) to 10 (very happy). Other covariates are the respondents age, net income (in 1000 Euros) and citizenship. The variable region specify if the interview took place in the former east part of Germany.

Table 7 shows the results for CUB and the CUP model using a cumulative model or an adjacent category model for the preference structure. It is again found that the inclusion of uncertainty reduces AIC and BIC strongly. Among the models with an uncertainty component there is not much difference in terms of AIC and BIC, BIC even favors the CUB. Also the logscore and the ranked score are very similar for all models. For both predictive measures there is no significant difference to the CUB model.

6 Concluding remarks

It has been shown that the basic concept to include an uncertainty component in the model, as has been done in CUB models before, can be extended to the familiar classes of ordinal models. In our examples the models typically show better fit and better performance in terms of AIC, BIC and prognostic measures than ordinal models without a mixture component and the traditional CUB model. If the uncertainty component is neglected the strength of the explanatory variables tends to be underestimated. An advantage of the models is that the effects of covariates can be easily visualized.

Acknowledgments This work has been partially supported by FIRB2012 project (Code RBFR12SHVV) at University of Perugia and the frame of Programme STAR (CUP E68C13000020003) at University of Naples Federico II, financially supported by UniNA and Compagnia di San Paolo. ISFOL survey data has been used under the agreement ISFOL/PLUS 2006/430.

7 Appendix

7.1 Identifiability

We assume that the number of categories is greater than 2 ($k > 2$) and that there is an effect of a continuous covariate x, that is $\gamma \neq 0$. Let the CUP model with the cumulative logit model in the preference part be represented by two parameterizations, that is, for all x and r one has

$$\pi F(\gamma_{0r} + x\gamma) + (1-\pi)r/k = \tilde{\pi} F(\tilde{\gamma}_{0r} + x\tilde{\gamma}) + (1-\tilde{\pi})r/k.$$

There are values Δ_{0r}, Δ such that $\tilde{\gamma}_{0r} = \gamma_{0r} + \Delta_{0r}$, $\tilde{\gamma} = \gamma + \Delta$. With $\eta_r(x) = \gamma_{0r} + x\gamma$ one obtains for all x and r

$$\pi F(\eta_r(x)) - \tilde{\pi} F(\eta_r(x) + \Delta_{0r} + x\Delta) = (\pi - \tilde{\pi})r/k.$$

Let us consider now the specific values $x_z = -\gamma_{0r}/\gamma + z/\gamma$ yielding for all values z and r

$$\pi F(z) - \tilde{\pi} F(z + \Delta_{0r} + x_z\Delta) = (\pi - \tilde{\pi})r/k.$$

By building the difference between these equations for values z and $z-1$ one obtains for all values z

$$\pi(F(z) - F(z-1)) = \tilde{\pi}(F(z + \Delta_{0r} + x_z\Delta) - F(z - 1 + \Delta_{0r} + x_z\Delta)). \quad (6)$$

The equation has to hold in particular for values $z = 1, 2, \ldots$. Since the logistic distribution function $F(\eta) = \exp(\eta)/(1 + \exp(\eta))$ is strictly monotonic and the derivative $F'(\eta) = \exp(\eta)/(1+\exp(\eta)^2)$ is different for all values η it follows that $\Delta_{0r} = \Delta = 0$ and $\pi = \tilde{\pi}$.

If the support of the covariate is finite one can consider different z- values. If $x \in [l, u]$ (γ positive) one considers the transformed values $z_i = \gamma l + \gamma_{0r} + \gamma(u-l)i/M$, for $i = 1, \ldots, M$, where M is any natural number. Then for all transformed values $x_{z_i} = -\gamma_{0r}/\gamma + z_i/\gamma$ one has $x_{z_i} \in [l, u]$. Thus, Eq. (6) has to hold for M different values z_i. Since M can be any natural number the same argument as before yields $\Delta_{0r} = \Delta = 0$ and $\pi = \tilde{\pi}$.

7.2 Estimation

The general CUP model is determined by the probability

$$P(r_i|\boldsymbol{x}_i) = \pi_i P_M(r_i|\boldsymbol{x}_i) + (1 - \pi_i) P_U(r_i),$$

where the first mixture component follows an ordinal model and the second represents the discrete uniform distribution.

For given data $(r_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, and collecting all parameters of the ordinal model used in the first mixture component in the parameter $\boldsymbol{\theta}$, the log-likelihood to be maximized is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log(\pi_i P_M(r_i|\boldsymbol{x}_i) + (1 - \pi_i) P_U(r_i)).$$

The usual way to obtain estimates is to consider it as a problem with incomplete data and solve the maximization problem by using the EM algorithm. Therefore, let z_i denote the unknown mixture components with $z_i = 1$ indicating that observation i is from the first mixture component, $z_i = 0$ indicates that it is from the second mixture component. Then the complete density for (r_i, z_i) is

$$P(r_i, z_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = P(r_i|z_i, \boldsymbol{x}_i, \boldsymbol{\theta}) P(z_i) = P_M(r_i|\boldsymbol{x}_i)^{z_i} P_U(r_i)^{z_i - 1} \pi_i^{z_i} (1 - \pi_i)^{z_i - 1}$$

yielding the complete log-likelihood

$$\begin{aligned} l_c(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log(P(r_i, z_i|\boldsymbol{x}_i, \boldsymbol{\theta})) \\ &= \sum_{i=1}^{n} z_i(\log(P_M(r_i|\boldsymbol{x}_i)) + \log(\pi_i)) + (1 - z_i)(\log(P_U(r_i)) + \log(1 - \pi_i)). \end{aligned}$$

The EM algorithm treats z_i as missing data and maximizes the log-likelihood iteratively by using an expectation and a maximization step. During the E-step the conditional expectation of the complete log-likelihood given the observed data $\boldsymbol{r}$ and the current estimate $\boldsymbol{\theta}^{(s)}$,

$$M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \mathrm{E}\left(l_c(\boldsymbol{\theta})|\boldsymbol{r}, \boldsymbol{\theta}^{(s)}\right)$$

has to be computed. Because $l_c(\boldsymbol{\theta})$ is linear in the unobservable data z_i, it is only necessary to estimate the current conditional expectation of z_i. From Bayes's theorem follows

$$\begin{aligned} E(z_i|\boldsymbol{y}, \boldsymbol{\theta}) &= P\,(z_i = 1|r_i, \boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= P\,(r_i|z_i = 1, \boldsymbol{x}_i, \boldsymbol{\theta})\; P(z_i = 1|\boldsymbol{x}_i, \boldsymbol{\theta})/P(r_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= \pi_i\, P_M(r_i|\boldsymbol{x}_i, \boldsymbol{\theta})/P(r_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = \hat{z}_i. \end{aligned}$$

This is the posterior probability that the observation r_i belongs to the first component of the mixture. For the s-th iteration one obtains

$$\begin{aligned} M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) &= \sum_{i=1}^{n} \hat{z}_i^{(s)}\,(\log(\pi_i) + \log(P_M(r_i|\boldsymbol{x}_i, \boldsymbol{\theta})) \\ &\quad + (1 - \hat{z}_i^{(s)})\,(\log(1 - \pi_i) + \log(P_U(r_i)) \\ &= \underbrace{\sum_{i=1}^{n} \hat{z}_i^{(s)} \log(\pi_i) + (1 - \hat{z}_i^{(s)})(\log(1 - \pi_i))}_{M_1} \\ &\quad + \underbrace{\sum_{i=1}^{n} \hat{z}_i^{(s)} \log(P_M(r_i|\boldsymbol{x}_i, \boldsymbol{\theta})) + (1 - \hat{z}_i^{(s)}) \log(P_U(r_i))}_{M_2}. \end{aligned}$$

Thus, for given $\boldsymbol{\theta}^{(s)}$ one computes in the E-step the weights $\hat{z}_i^{(s)}$ and in the M-step maximizes $M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$ (or rather M_1 and M_2). If the mixture probabilities do not depend on covariates, that is, $\pi_i = \pi$, one obtains

$$\pi^{(s+1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_i^{(s)} \quad \text{and} \quad \boldsymbol{\theta}^{(s+1)} = \mathrm{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \hat{z}_i^{(s)} \log(P_M(r_i|\boldsymbol{x}_i, \boldsymbol{\theta})).$$

The E- and M-steps are repeated alternatingly until the difference $L(\boldsymbol{\theta}^{(s+1)}) - L(\boldsymbol{\theta}^{(s)})$ is small enough to assume convergence. Computation of $\boldsymbol{\theta}^{(s+1)}$ can be based on familiar maximization tools, because one maximizes a weighted log-likelihood of an ordinal model with known weights. In the case where only intercepts are component-specific,

the derivatives are very similar to the score function used in a Gauss-Hermite quadrature and a similar EM algorithm applies with an additional calculation of the mixing distribution (see Aitkin 1999).

Dempster et al. (1977) showed that under weak conditions the EM algorithm finds a local maximum of the likelihood function $L(\boldsymbol{\theta})$. Hence it is sensible to use different start values $\boldsymbol{\theta}^{(0)}$ to find the solution of the maximization problem.

If covariates determine the probability that observation i belongs to the first mixture component in the form of a logit model, $\pi_i(\boldsymbol{\beta}) = 1/(1 + \exp(-\boldsymbol{z}_i^T\boldsymbol{\beta}))$, M_1 is the weighted log-likelihood of a binary logit model. Then M_1 and M_2 are maximized separately to obtain the next iteration. The simple update $\pi^{(s+1)} = \sum_{i=1}^{n} \hat{z}_i^{(s)}/n$ is replaced by

$$\boldsymbol{\beta}^{(s+1)} = \text{argmax}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \hat{z}_i^{(s)} \log(\pi_i(\boldsymbol{\beta})) + (1 - \hat{z}_i^{(s)})(\log(1 - \pi_i(\boldsymbol{\beta}))).$$

As default value for the stopping of the iterations we used the difference in two consecutive likelihoods; if it was below 10^{-6} the algorithm was stopped.

References

Agresti A (2010) Analysis of ordinal categorical data, 2nd edn. Wiley, New York

Agresti A (2013) Categorical data analysis, 3d edn. Wiley, New York

Aitkin M (1999) A general maximum likelihood analysis of variance components in generalized linear models. Biometrics 55:117–128

Anderson JA (1984) Regression and ordered categorical variables. J Royal Stat Soc B 46:1–30

Böhning D, Dietz E, Schaub R, Schlattmann P, Lindsay BG (1994) The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. Ann Inst Stat Math 46:373–388

Brant R (1990) Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics 46:1171–1178

Breen R, Luijkx R (2010) Mixture models for ordinal data. Sociol Methods Res 39:3–24

Caffo B, An M-W, Rhode C (2007) Flexible random intercept models for binary outcomes using mixtures of normals. Comp Stat Data Anal 51:5220–5235

Cox C (1995) Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. Stat Med 14:1191–1203

D'Elia A, Piccolo D (2005) A mixture model for preference data analysis. Comp Stat Data Anal 49:917–934

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc B 39:1–38

Efron B, Tibshirani RJ (1994) An introduction to the bootstrap, vol 57. CRC Press, London

Everitt BS (1988) A finite mixture model for the clustering of mixed-mode data. Stat Prob Lett 6(5):305–309

Fahrmeir L, Tutz G (2001) Multivariate statistical modelling based on generalized linear models. Springer, New York

Follmann DA, Lambert D (1991) Identifiability of finite mixtures of logistic regression models. J Stat Plan Infer 27(3):375–381

Gambacorta R, Iannario M (2013) Measuring job satisfaction with CUB models. Labour 27(2):198–224

Gertheiss J, Tutz G (2009) Penalized Regression with Ordinal Predictors. Int Stat Rev 77:345–365

Gneiting T, Raftery A (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102:359–376

Greene W, Hensher D (2003) A latent class model for discrete choice analysis: contrasts with mixed logit. Trans Res Part B 39:681–689

Grilli L, Iannario M, Piccolo D, Rampichini C (2014) Latent class CUB models. Adv Data Anal Class 8(1):105–119

Grün B, Leisch F (2008) Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. J Class 25:225–247
Iannario M (2010) On the identifiability of a mixture model for ordinal data. Metron 68(1):87–94
Iannario M (2012a) Hierarchical CUB models for ordinal variables. Commun Stat Theory Methods 41:3110–3125
Iannario M (2012b) Modelling shelter choices in a class of mixture models for ordinal responses. Stat Methods Appl 21:1–22
Iannario M (2012c) Preliminary estimators for a mixture model of ordinal data. Adv Data Anal Class 6:163–184
Iannario M, Piccolo D (2010) Statistical modelling of subjective survival probabilities. Genus 66:17–42
Iannario M, Piccolo D (2012) CUB models: Statistical methods and empirical evidence. In: Kennett SSR (ed) Modern analysis of customer surveys: with applications using R. Wiley, New York, pp 231–258
Leroux BG (1992) Consistent estimation of a mixing distribution. Ann Stat 20:1350–1360
Liu Q, Agresti A (2005) The analysis of ordinal categorical data: An overview and a survey of recent developments. Test 14:1–73
Manisera M, Zuccolotto P (2014) Modeling rating data with nonlinear CUB models. Comp Stat Data Anal 78:100–118
McCullagh P (1980) Regression model for ordinal data (with discussion). J Royal Stat Soc B 42:109–127
McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York
Mehta CR, Patel NR, Tsiatis AA (1984) Exact significance testing to establish treatment equivalence with ordered categorical data. Biometrics 40:819–825
Nair VN (1987) Chi-squared-type tests for ordered alternatives in contingency tables. J Am Stat Assoc 82:283–291
Peterson B, Harrell FE (1990) Partial proportional odds models for ordinal response variables. Appl Stat 39:205–217
Piccolo D (2003) On the moments of a mixture of uniform and shifted binomial random variables. Quaderni di Stat 5:85–104
Piccolo D (2006) Observed information matrix in MUB models. Quaderni di Stat 8:33–78
Tutz G (2012) Regression for categorical data. Cambridge University Press, Cambridge
Tutz G, Gertheiss (2014) Rating scales as predictors—the old question of scale level and some answers. Psychometrika 79:357–376
Tutz G, Schauberger G (2013) Visualization of categorical response models - from data glyphs to parameter glyphs. J Comp Graph Stat 22(1):156–177
Wedel M, DeSarbo W (1995) A mixture likelihood approach for generalized linear models. J Class 12:21–55

A.2. Flexible Uncertainty in Mixture Models for Ordinal Responses

Tutz, G., and M. Schneider (2019): Flexible uncertainty in mixture models for ordinal responses. *Journal of Applied Statistics*, 46(9), 1582–1601, doi:10.1080/02664763.2018.1555574.

Flexible Uncertainty in Mixture Models for Ordinal Responses

Gerhard Tutz & Micha Schneider

[1] Ludwig-Maximilians-Universität München, Akademiestraße 1, 80799 München

gerhard.tutz@stat.uni-muenchen.de, Micha.Schneider@stat.uni-muenchen.de

Abstract

In classical mixture models for ordinal data with an uncertainty component the Uniform distribution is used to model indecision. In the approach proposed here the discrete Uniform distribution is replaced by a more flexible distribution, which is centered in the middle of the response categories. The resulting model allows to distinguish between a tendency to middle categories and a tendency to extreme categories. By linking these preferences to explanatory variables one can investigate which persons show a tendency to these response styles. It is demonstrated that severe bias might occur if inadvertently the Uniform distribution is used to model uncertainty. An application to attitudes on the performance of health services illustrates the advantages of the more flexible model.

Keywords: Ordinal responses, response styles, rating scales, mixture models, CUP model, CUB model

1 Introduction

In recent years a class of mixture models for ordinal data has been introduced that considers the choice of a response category as resulting from a mixture of a deliberate choice and uncertainty. In the original CUB model (for Combination of discrete Uniform and shifted Binomial random variables), see D'Elia and Piccolo (2005), the deliberate choice is modelled by a Binomial distribution and the uncertainty by a discrete Uniform distribution. Various models with different specifications of the distributions of the deliberate choice and the uncertainty part have been proposed since then, see, for example, Iannario and Piccolo (2010), Iannario et al. (2012), Iannario and Piccolo (2012), Iannario (2012a), Iannario (2012b), Manisera and Zuccolotto (2014), Capecchi and Piccolo (2016), and Tutz et al. (2017). Overviews on the modelling approaches were given by Iannario and Piccolo (2016a) and Iannario and Piccolo (2016b).

The basic assumption of most of these extensions is that uncertainty follows a discrete Uniform distribution. However, the assumption that all categories, including middle and extreme categories, share the same degree of uncertainty is rather strong. In particular it excludes the preference of middle or extreme categories, which is a response style that is often found in applications. In the present paper we propose a more flexible uncertainty component which is able to capture response styles.

The presence of response styles has been found in many studies, see, for example, Clarke (2000), Van Herk et al. (2004), Marin et al. (1992) and Meisenberg and Williams (2008). Several modelling approaches have been proposed for repeated measurements within the framework of item response models, see Bolt and Johnson (2009), Bolt and Newton (2011), Johnson (2003), Eid and Rauber (2000). More recently tree type approaches have been considered. They typically assume a nested structure where first a decision about the direction of the response and then about the strength is obtained, see, for example, De Boeck and Partchev (2012), Jeon and De Boeck (2016), and Böckenholt (2012). Mixture modelling of response styles by use of latent class models has been investigated by Moors (2004), Kankaraš and Moors (2009), Moors (2010), and Rosmalen et al. (2010).

The mixture considered here does not assume that responses on several items are available as is usually assumed in item response theory. We aim at separating the deliberate choice from the tendency to middle or extreme categories by using a mixture model in the tradition of CUB models. However, in contrast to these models we consider an uncertainty component that can account for response styles. By linking the uncertainty component to covariates, the model is able to uncover which person characteristics determine the response style.

The paper is organized as follows: In Section 2 we consider uncertainty as a relevant component quite often present in human choices. Thus CUB models and models with alternative parameterizations are briefly reviewed. Then the new class of models with more flexible uncertainty components is introduced. In Section 3 we investigate the consequences of fitting misspecified models in a simulation study. Section 4 gives the details of the fitting algorithm and in Section 5 the model is used to investigate the satisfaction with the Health Service in European Countries.

2 Mixture Models for Ordinal Responses

In the following we briefly consider an extended form of the CUB model. Then we consider alternative specifications of the uncertainty component.

2.1 Mixture Models for the Combination of Uncertainty and Preference

Let in a regression model the response of an individual R_i given explanatory variables take values from ordered categories $\{1, \ldots, k\}$.

The general mixture model we consider is the CUP model, which is an acronym for Combination of Uncertainty and Preference. It has the form

$$P(R_i = r|\boldsymbol{x}_i) = \pi_i P_M(Y_i = r|\boldsymbol{x}_i) + (1 - \pi_i)P_U(U_i = r), \tag{1}$$

where R_i is the ordinal response variable, Y_i denotes the unobserved random variable that represents the deliberate choice, that is, the preference on the ordinal scale and U_i is the unobserved uncertainty component, $\boldsymbol{x}_i$ is a vector of covariates and π_i represents the mixture probability, which measures the importance of the structured component in the mixture. Thus the observed response results from a discrete mixture of the preference and the uncertainty component. Both variables Y_i and U_i take values from $\{1, \ldots, k\}$.

In model (1) the distribution of Y_i is determined by $P_M(Y_i = r|\boldsymbol{x}_i)$, which can be any ordinal model M. In CUP models the uncertainty component is specified by the Uniform distribution, $P_U(U_i = r) = 1/k$. The assumption of a more flexible distribution than the Uniform distribution is the central issue here but postponed to the next section. Instead we consider briefly the ordinal models that can be used in the preference part.

In traditional CUB models the distribution of Y_i is specified as a shifted Binomial distribution, that is,

$$P_M(Y_i = r|\boldsymbol{x}_i) = \binom{k-1}{r-1}\xi_i^{k-r}(1 - \xi_i)^{r-1}, \quad r \in \{1, \ldots, k\}.$$

In extended versions (Tutz et al., 2017) more general models as the cumulative or the adjacent categories models are used. Cumulative models have the form

$$P(Y_i \leq r|\boldsymbol{x}_i) = F(\gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}), \quad r = 1, \ldots, k-1,$$

where $F(.)$ is a cumulative distribution function and $-\infty = \gamma_{00} < \gamma_{01} < \cdots < \gamma_{0k} = \infty$. The most widely used model from this class of models is the cumulative logit model, which uses the logistic distribution $F(.)$. It is also called *proportional odds model* and has the form

$$\log\left(\frac{P(Y_i \leq r|\boldsymbol{x}_i)}{P(Y_i > r|\boldsymbol{x}_i)}\right) = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}, \quad r = 1, \ldots, k-1.$$

An alternative choice is the *adjacent categories model* given by

$$P(Y_i = r + 1|Y_i \in \{r, r + 1\}, \boldsymbol{x}_i) = F(\gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}), \quad r = 1, \ldots, k-1.$$

If the probability $P(Y_i = r|Y_i \geq r, \boldsymbol{x}_i)$ represents the probability of failure in (time) category r given category r is reached it can be seen as a discrete hazard. The specific model that uses the logistic distribution is the *adjacent categories logit model*

$$\log\left(\frac{P(Y_i = r+1|\boldsymbol{x}_i)}{P(Y_i = r|\boldsymbol{x}_i)}\right) = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}, \quad r = 1, \ldots, k-1.$$

A general discussion of ordinal models is found in McCullagh (1980), Agresti (2010), Agresti (2013) and Tutz (2012).

2.2 Models with a Flexible Uncertainty Component

The Uniform distribution as uncertainty component has the advantage of simplicity. However, it implies that uncertainty is uniformly distributed over the response categories. A more flexible concept allows that uncertainty may express itself in a stronger tendency toward middle or extreme categories. In particular persons who are undecided or have no strong opinion may have a tendency to choose middle categories and not choose at random from the whole spectrum of categories. Therefore, instead of the Uniform distribution we use a specific version of the Beta-Binomial distribution.

A random variable U with support $\{1, \ldots, k\}$ follows a Beta-Binomial distribution, $U \sim \textit{Beta-Binomial}(k, \alpha, \beta)$, if the mass function is given by

$$f(u) = \begin{cases} \binom{k-1}{u-1}\frac{B(\alpha+u-1,\beta+k-u+1)}{B(\alpha,\beta)} & u \in \{1, \ldots, k\} \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha, \beta > 0$ and $B(\alpha, \beta)$ is the beta function defined as

$$B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt.$$

With $\mu = \alpha/(\alpha+\beta)$ and $\delta = 1/(\alpha+\beta+1)$ one obtains the expected value $E(U)$ and the variance $var(U)$

$$E(U) = (k-1)\mu + 1, \quad \text{var}(U) = (k-1)\mu(1-\mu)[1 + (k-2)\delta].$$

As $\delta \to 0$, the Beta-Binomial distribution converges to the (shifted) Binomial distribution $B(k, \mu)$ with mean μ and support $\{1, \ldots, k\}$.

Since we aim at modelling a tendency to middle categories we choose a fixed value $\mu = 0.5$ and therefore $\alpha = \beta, \delta = 1/(2\alpha+1)$ to obtain

$$\text{E}(U) = (k+1)/2.$$

For the variance one obtains

$$\text{var}(U) = ((k-1)/4)\frac{2\alpha + k - 1}{2\alpha + 1}.$$

The restricted Beta-Binomial distribution is determined by the parameters α and k. An interesting extreme case is $\alpha = 0$, which yields

$$\text{var}(U) = ((k-1)^2/4),$$

and corresponds to a two point distribution on 1 and k. If α tends to infinity one obtains

$$\text{var}(U) = ((k-1)/4).$$

Therefore, the parameter α determines the concentration of the distribution in the middle, for small values the probability mass is concentrated in the end points, for $\alpha = 1$ one obtains the discrete Uniform distribution and for $\alpha \to \infty$ one obtains a (shifted) Binomial distribution, which is symmetric around its mean $(k-1)/2$.

Figure 1 shows the Beta-Binomial distribution for selected values of α. The mode is always in the middle of the support, in the case of an odd number of categories the mode is represented by one of the categories. We use the Beta-Binomial distribution in its symmetric version. This restriction is warranted by the aim to model the specific response style that is characterized by a tendency to middle or extreme categories. It makes the uncertainty component a one parameter distribution, which implies that expectation and variance are related.

It should be mentioned that the Beta-Binomial distribution has also been used in ordinal data models to allow for more dispersion of the feeling/preference component, see Iannario (2014) and Piccolo (2015).

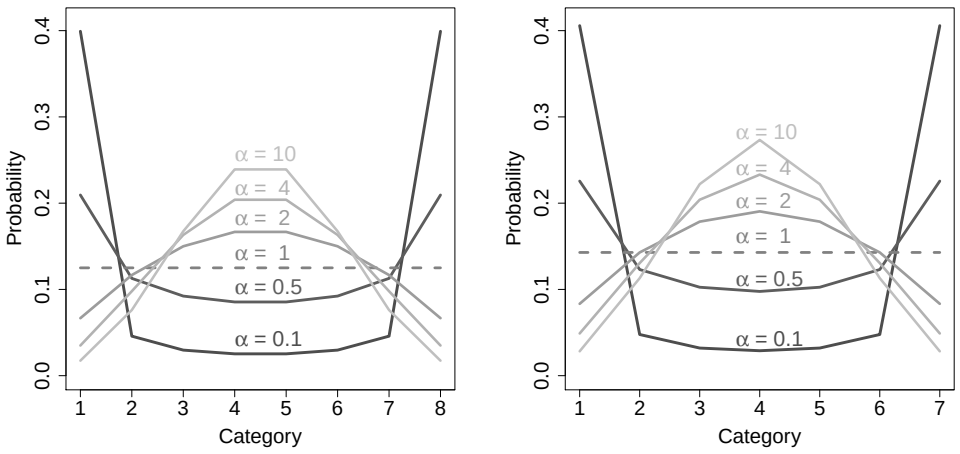


FIGURE 1: *Probability mass on categories for various values of α for $k = 8$ categories (left panel) and $k = 7$ categories (right panel).*

While mixture models in the tradition of CUB models use the Uniform distribution, the Beta-Binomial distribution provides a wider concept of uncertainty

in mixture models. Models with alternative uncertainty distributions have also been proposed by Simone and Tutz (2018) and Gottard et al. (2016). The latter allow, among other distributions, that the uncertainty is given by a parabolic or a triangular distribution. However, one has to choose the mode of the triangular distribution, therefore a priori information is needed. Moreover, the uncertainty distribution is not linked to explanatory variables in contrast to the approach proposed here (see next section).

The mixture model we consider accounts for response styles in the uncertainty component and therefore in the indecision part of the mixture. Similar concepts have been used in psychometrics, for example, by Rost (1991), von Davier and Rost (1995), von Davier and Yamamoto (2007). In these mixture models it is assumed that respondents come from different latent classes. Different item response models are fitted within these classes, some may represent the substantive trait, some may represent response style behaviour. However, in these models the latent classes are left unspecified. Thus one obtains classes that have different parameters but these do not necessarily correspond to response styles. Moreover, one has to choose the number of classes and differing numbers of classes yield quite different results. In contrast, all CUB type models use just two components, which correspond to the classes, and the components are specified to yield a clear interpretation. In our approach the two components are the preference and the uncertainty component with the latter containing the response style. It should be noted that alternatively one could also consider the response style to be part of the preference component. This is the concept used by the Shelter CUB (Iannario, 2012b) and the Nonlinear CUB (Manisera and Zuccolotto, 2014). For example, the tendency to middle categories can be seen as a refuge choice or shelter effect rather than indecision. As a reviewer commented it is question of the modelling philosophy in which part one wants to include the response style. If one includes it in the uncertainty part, as is done in our approach, one gives up the Uniform distribution in the uncertainty part. However, although the Uniform distribution has the advantage of being very simple it is a very strong assumption that uncertainty is determined by the same probability for all categories wherever the preference is located. Therefore a relaxation of this assumption seems sensible.

Nevertheless, one should be aware that the model, as all statistical models, provides a specific view into the data, only structures are detected that are specified in the model, and the model usually is a simplified version of the data generating mechanism. In particular in mixture models, which contain unobservable components, it is difficult to ensure that the model is correctly specified. For example, the presence of two extreme modal values may also be the result of the presence of two clusters, one composed of people that are favourable to the item, and the other of unfavourable ones. Then the appropriate mixture model would be a quite different one. In this sense, the model that is used represents a choice, it may be seen as a working hypothesis that allows to identify structures

in the data.

It should be mentioned that in mixture models identifiability problems may occur, for CUB type models some results are available, see Iannario (2010), Tutz et al. (2017). In the model considered here, in particular if the preference distribution is symmetric one has a mixture of two symmetric distributions. Then, it certainly takes large data sets to be able to distinguish between the two distributions, although they should be identifiable given their different form. Nevertheless, general results are not yet available.

2.3 Parametrization

In the general mixture model (1) the preference for categories is determined by the covariates $\boldsymbol{x}_i$ within the ordinal model that is used in the preference part. However, also the strength of the tendency to middle or extreme categories may depend on covariates. Therefore, we let the parameter α, which determines the distribution in the uncertainty part, depend on covariates $\boldsymbol{w}_i^T = (1, w_{i1}, \ldots, w_{im})$, which can be different, identical to or partially overlapping with the covariates $\boldsymbol{x}_i$. A simple link is given by

$$\alpha = \exp(\boldsymbol{w}_i^T \boldsymbol{\alpha}) = \exp(\alpha_0) \exp(\alpha_1)^{w_{i1}} ..... \exp(\alpha_m)^{w_{im}},$$

where $\boldsymbol{\alpha}^T = (\alpha_0, \ldots, \alpha_m)$. The parameter α_j contains the effect of the j-th covariate. The parameter α changes by the factor $\exp(\alpha_j)$ if w_{ij} increases by one unit given all other variables are kept constant. The parameters determine how a variable influences the tendency to middle or extreme categories. It should be noted that in the case without covariates one has the simple reparameterization $\alpha = \exp(\alpha_0)$.

The model (1) with a Beta-Binomial mixture component is called the BetaBin model. Although it is a generalization of CUP models the intention of the modelling approach is quite different. In CUP models the uncertainty is specified by a discrete Uniform distribution and the underlying assumption is that a person is torn between his/her preference and uncertainty. The uncertainty is such that each category has the same probability. The BetaBin model is composed of a preference model and a model that represents a tendency to middle or extreme categories. It allows to model not only the preference as a function of covariates but also the tendency to middle or extreme categories as a function of covariates. One may see, for example, differences in the preference of middle or extreme categories induced by covariates like gender. Therefore, response patterns induced by explanatory variables can be identified.

The family of models considered here can be specified by Mix(structured part, uncertainty part). The structured part indicates which model is used to model the deliberate choice, and the uncertainty part indicates which distribution is used to model the uncertainty. Examples are

Mix(Binomial, Uniform) (or CUB), which means that the structured response follows Binomial distribution and uncertainty is determined by the Uniform distribution

Mix(Cumulative, Uniform) (or CUP), which means that the structured response is determined by a cumulative model, the uncertainty is the same as in the previous example

Mix(Cumulative, Beta-Binomial), which means that the uncertainty is determined by the Beta-Binomial distribution

Mix(Binomial, Beta-Binomial), which means that the structured response follows Binomial distribution and uncertainty is determined by the Beta-Binomial distribution

The model Mix(Cumulative, Beta-Binomial($\alpha = 1$)) is equivalent to the CUP cumulative model, Mix(Binomial, Beta-Binomial($\alpha = 1$)) is equivalent to the CUB model and Mix(Beta-Binomial, Uniform) denotes the CUBE model.

3 Estimation

The likelihood contribution of observation i when category y_i is observed is determined by

$$P(R_i = y_i|\boldsymbol{w}_i, \boldsymbol{x}_i) = \pi_i \, P_M(Y_i = y_i|\boldsymbol{x}_i) + (1 - \pi_i) \, P_U(U_i = y_i|\boldsymbol{w}_i) \qquad (2)$$

yielding the log-likelihood contribution

$$l_i(\boldsymbol{\gamma}, \boldsymbol{\alpha}) = \log(\pi_i \, P_M(Y_i = y_i|\boldsymbol{x}_i) + (1 - \pi_i) \, P_U(U_i = y_i|\boldsymbol{w}_i))$$

A way to obtain stable estimates is to consider it as a problem with incomplete data and use the EM-algorithm (Dempster et al., 1977). Therefore, let z_i^* denote the unknown mixture components that indicate whether y_i belongs to the first or second component of the mixture

$$z_i^* = \begin{cases} 1, \text{observation } y_i \text{ is from the first mixture component} \\ 0, \text{otherwise.} \end{cases}$$

The corresponding complete log-likelihood is given by

$$l_c(\boldsymbol{\gamma}, \boldsymbol{\alpha}) =$$
$$\sum_{i=1}^{n} z_i^* \left\{\log(\pi_i) + \log(P_M(Y_i = y_i|\boldsymbol{x}_i))\right\} + (1 - z_i^*) \left\{\log(1 - \pi_i) + \log(P_U(U_i = y_i|\boldsymbol{w}_i))\right\}.$$

The EM-algorithm treats z_i^* as missing data and maximizes the log-likelihood iteratively by using an Expectation step (E-step) and a Maximization step

(M-step). During the E-step the conditional expectation of the complete log-likelihood given the observed data $\boldsymbol{y}^T = (y_1, \ldots, y_n)$ and the current estimate $\boldsymbol{\theta}^{(s)} = (\boldsymbol{\gamma}^{(s)}, \boldsymbol{\alpha}^{(s)})$,

$$M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \mathrm{E}(l_c(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}^{(s)})$$

has to be computed. Because $l_c(\boldsymbol{\theta})$ is linear in the unobservable data z_i^*, it is only necessary to estimate the current conditional expectation of z_i^*. From Bayes's theorem follows

$$\begin{aligned} E(z_i^*|\boldsymbol{y}, \boldsymbol{\theta}) &= P(z_i^* = 1|y_i, \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\theta}) = P(R = y_i|z_i^* = 1, \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\theta})/P(R = y_i|\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\theta}) \\ &= \pi_i \, P_M(Y_i = y_i|\boldsymbol{x}_i, \boldsymbol{\theta})/(\pi_i \, P_M(Y_i = y_i|\boldsymbol{x}_i) + (1 - \pi_i) \, P_U(U_i = y_i|\boldsymbol{w}_i)) \\ &= \hat{z}_i^* = \hat{z}^*. \end{aligned}$$

This is the posterior probability that the observation y_i belongs to the first component of the mixture. Because there are no individual covariates determining the propensity to the structure component $\hat{z}_i^*$ the expectation $E(z_i^*|\boldsymbol{y}, \boldsymbol{\theta})$ is the same for all observations. For the s-th iteration one obtains

$$\begin{aligned} M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) &= \sum_{i=1}^{n} \hat{z}^* \left\{\log(\pi) + \log(P_M(Y_i = y_i|\boldsymbol{x}_i))\right\} \\ &\quad + (1 - \hat{z}^*) \left\{\log(1 - \pi) + \log(P_U(U_i = y_i|\boldsymbol{w}_i))\right\} \\ &= \underbrace{\sum_{i=1}^{n} \hat{z}^* \log(\pi) + (1 - \hat{z}^*) \log(1 - \pi)}_{M_1} \\ &\quad + \underbrace{\sum_{i=1}^{n} (1 - \hat{z}^*) \log(P_U(U_i = y_i|\boldsymbol{w}_i))}_{M_2} \\ &\quad + \underbrace{\sum_{i=1}^{n} \hat{z}^* \log(P_M(Y_i = y_i|\boldsymbol{x}_i))}_{M_3}. \end{aligned}$$

The maximization in the M-step uses the decomposition into M_1, M_2 and M_3. M_2 corresponds to the uncertainty component and M_3 to the structure component. M_1, M_2 and M_3 can be maximised separately with traditional software. For M_1 and the shifted Binomial distribution (M_3 in CUB-models) we use the R-package MRSP by Poessnecker (2015). For the Beta-Binomial distribution (M_2) and the cumulative model (M_3 in CUP-models) we used the R-package VGAM by Yee (2016). For the CUB-models there is also the package CUB (Iannario et al., 2018) available. In the s-th EM iteration M_1, M_2 and M_3 are not maximised until convergence is reached but only a few iterations in the sense of the generalized EM-Algorithm. So for given $\boldsymbol{\theta}^{(s)}$ one computes in the E-step the weights $\hat{z}^{*(s)}$ and in the M-step maximizes $M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$, which yields the new estimates.

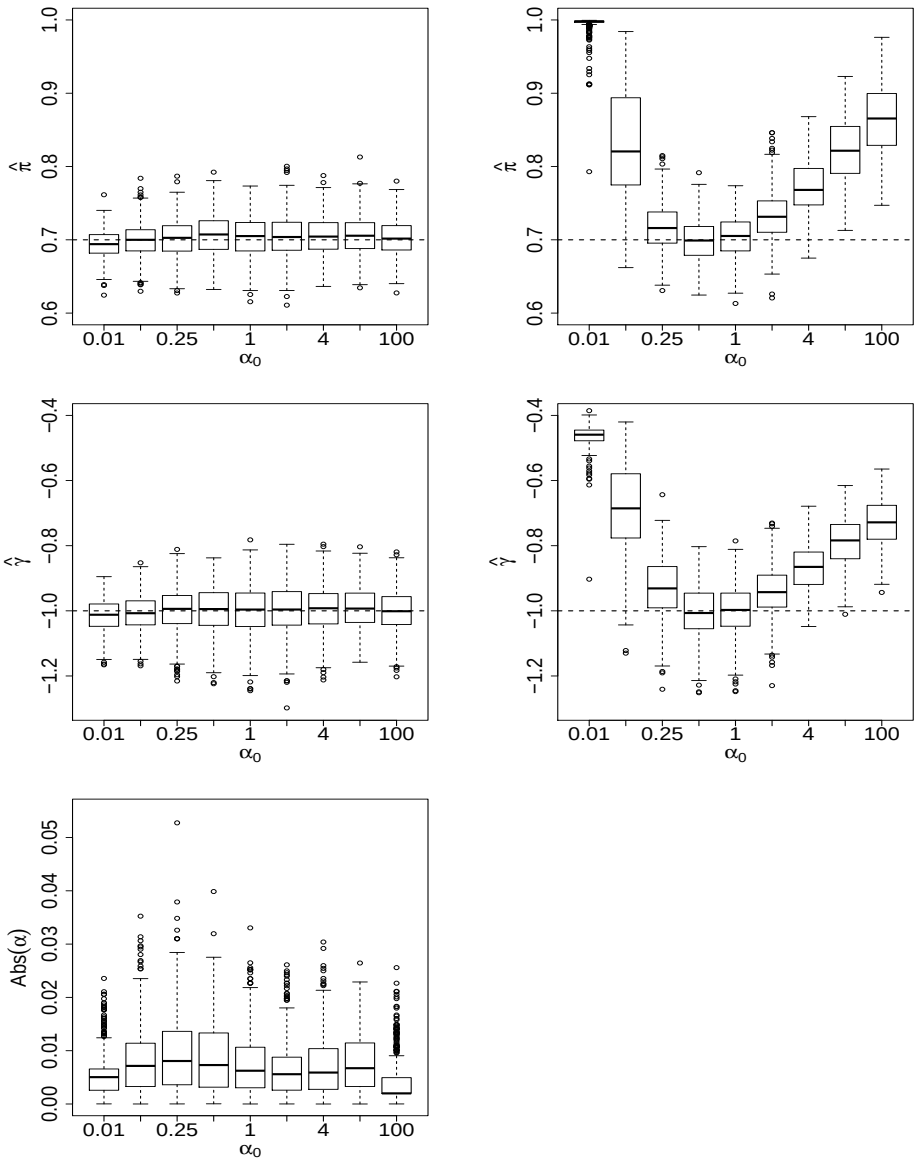


FIGURE 2: *Estimated parameters $\hat{\pi}$, $\hat{\gamma}$ for the Mix(Cumulative, Beta-Binomial) on the left and the Mix(Cumulative, Uniform) model on the right (true values are $\pi = 0.7$ and $\gamma = -1$). The true α-values are {0.01, 0.1, 0.25, 0.5, 1, 2, 4, 10, 100}. For the Mix(Cumulative, Beta-Binomial) model also the Abs_α is given (it is only reasonable for the Mix(Cumulative, Beta-Binomial) model).*

4 Simulations

In the following we investigate the consequences of fitting misspecified models in a simulation study. In particular we investigate the consequences if the proposed

model is the data generating model and one fits a model that uses a Uniform distribution in the uncertainty part. First we compare the Mix(Cumulative, Uniform) (or cumulative CUP) and the Mix(Cumulative, Beta-Binomial), and then CUB (or Mix(Binomial, Uniform)) and Mix(Binomial, Beta-Binomial).

We use a response with $k = 7$ categories, $n = 2000$ observations and one explanatory variable, which follows a Uniform distribution with support $[-4, 4]$. The data were simulated from a mixture model with different values for π, α and γ. For the mixture weights π the values we used are $0.5, 0.7$ and 0.8. In the structure component we use the cumulative model and the shifted Binomial model. In the cumulative model the predictor has the form $\eta_i = \gamma_{0r} + x_i\gamma$. The effect of the explanatory variable, γ, was fixed at -1 and -2. The intercepts γ_{0r} were set to $-4, -3, -2, -1, 0, 1$. In the shifted Binomial model the parameter ξ_i is parameterized by $\text{logit}(\xi_i) = \gamma_0 + x_i\gamma$. We used $\gamma_0 = 1$ and for γ again -1 and -2. The range of the α-values, which represent the response style, was $\{0.01, 0.1, 0.25, 0.5, 1, 2, 4, 10, 100\}$ so that both the tendency to the middle categories with $\alpha > 1$ and the tendency to extreme categories with $\alpha < 1$ are covered. Also the special case $\alpha = 1$, in which the uncertainty components of CUP and Mix(Cumulative, Beta-Binomial) are identical, is included. For each parameter combination 500 data sets were simulated from the model with the Beta-Binomial distribution. The BetaBin model as well as the model with Uniform distribution were fitted. Then the performance of the new proposed model is compared to the performance of the misspecified model with a Uniform distribution.

Before given detailed tables for all used combinations of π, α and γ we show some illustrative box plots. Figure 2 displays the estimated parameters for different α-values for both models with π set to 0.7 and γ set to -1. Each boxplot consists of 500 samples. The results of the BetaBin model are displayed on the left hand side and the results of the CUP model on the right hand side. The top row shows the π-estimates and the middle row the γ-estimates. For the BetaBin model all the estimates are close to the true parameters regardless which response style is true. The model is able to capture both a strong tendency to the middle category as well as a strong tendency to extreme categories. On the right hand side the different response styles are neglected and it is always assumed that the uncertainty component follows a Uniform distribution. The results show that estimates are strongly biased if the model is unable to account for the response style. If the true α-value is far away form $\alpha = 1$, which is assumed by the CUP model, there is a large discrepancy between the true parameter values and the estimated parameters. For example, if $\alpha = 0.01$, which indicates a strong tendency to the extreme categories, the CUP model estimates a π-value which is close to one. Thus, one would falsely infer that no uncertainty component is needed. At the same time the strength of the effect of the variable is underestimated. If there is a strong tendency to the middle categories the results are similar. So by using the Uniform distribution as a possible response style not only the π-values but also the γ-values are strongly biased if the data generating model contains a

specific response style.

To investigate the accuracy of estimates we consider the mean squared error. For the comparison we use the log proportions

$$\text{lp} = \frac{1}{S}\sum_{i=1}^{S} \log \frac{\text{MSE(Uniform)}_i}{\text{MSE(Beta-Binomial)}_i},$$

where S is the number of simulations and $\text{MSE(Beta-Binomial)}_i$ denotes the mean squared error in the ith sample if the BetaBin model is fitted and MSE(Uniform) the mean squared error if the Uniform model is fitted. Positive values of lp indicate that the Uniform model yields estimates that are worse than the estimates obtained by the BetaBin model. We compare only models that differ in the uncertainty component since when comparing models that differ in both components one can not link poor performance to the type of misspecification.

Table 1 and 2 show the log proportions for γ and π for several parameter combinations. In the case of $\alpha = 1$ the log proportions are close to zero so that both models fit equally. But there is a strong monotone increase when the true α-values are more and more away from $\alpha = 1$. For example, one obtains for $(\pi, \gamma, \alpha) = (0.5, -1, 4)$ $lp = 0.6509$, which means that the MSE of the Uniform model is 1.92 times the MSE of the BetaBin model, for $(\pi, \gamma, \alpha) = (0.5, -2, 4)$ one has $lp = 1.4235$ denoting that the MSE of the Uniform model is 4.15 times the MSE of the BetaBin model. It is also seen that for small values of π the proportions of γ-values are larger than for large values of π (close to 1), therefore for small values of π a wrong response style has stronger impact on the γ-parameters. For larger value of γ one obtains larger log proportions.

For the accuracy of the estimated response style we do not use the mean squared errors of the α-values. The reason is the scaling of the parameter. For very large α-values the Beta-Binomial distribution is close to the Binomial distribution, which is obtained if α-values is infinitely large. Consequently very large α-values may be different in their absolute value but lead to nearly the same distribution function. Therefore, we use the absolute differences of the estimated distributions

$$Abs_\alpha = \frac{1}{S}\sum_{i=1}^{S}\left(\frac{1}{k}\sum_{r=1}^{k}|Pr_i(U = r|\hat{\alpha}) - Pr_i(U = r|\alpha)|\right),$$

where k is the number of categories and S the number of simulations. As seen from Table 3 in all settings the Abs_α is less than 0.012 so that whatever response style is present the model is able to estimate it very well. The differences slightly increase with higher π-values. The last panel in Figure 2 shows the corresponding box plots, which are all close to zero. One can see that the BetaBin model is able to fit the true response style very well.

π	γ	α								
		0.01	0.1	0.25	0.5	1	2	4	10	100
0.5	-1	1.1120	1.0887	1.0061	0.2167	0.0007	0.2717	0.6509	0.8644	0.9398
0.7	-1	0.8909	0.5479	0.1319	-0.0138	0.0006	0.1043	0.2375	0.3690	0.4783
0.8	-1	0.2506	0.0925	-0.0080	-0.0334	-0.0008	0.0660	0.1343	0.2060	0.2711
0.5	-2	6.1968	5.6751	5.3883	1.0003	-0.0014	0.3862	1.4235	3.5169	4.8718
0.7	-2	6.3268	4.1285	1.2782	-0.0426	-0.0231	0.1646	0.2528	1.0761	1.8724
0.8	-2	2.4443	0.8960	0.2808	-0.0005	-0.0036	0.1269	0.3452	0.5010	0.7295

TABLE 1: *Log proportions of γ-values. Positive values indicate that γ estimates of the CUP model are further away from the true γ-values than the estimates of the Mix(Cumulative, Beta-Binomial) model.*

π	γ	α								
		0.01	0.1	0.25	0.5	1	2	4	10	100
0.5	-1	7.5542	7.0457	6.1898	0.9142	-0.0923	2.2108	4.9798	6.2124	6.7589
0.7	-1	6.6484	4.3927	0.7165	-0.1107	0.0312	1.0513	2.9954	4.1365	4.9370
0.8	-1	2.0368	0.8023	0.4547	-0.0873	-0.0320	0.5957	1.7226	2.6127	3.5027
0.5	-2	7.7898	7.6342	6.9940	2.2677	0.0124	0.4676	1.0724	3.3313	5.8254
0.7	-2	7.0814	5.0817	2.5478	0.4908	-0.0218	0.2590	0.4474	0.4942	0.8461
0.8	-2	3.6844	2.3575	0.9614	0.1400	-0.0077	0.1195	0.2181	0.5144	0.4748

TABLE 2: *Log proportions of π-values. Positive values indicate that π estimates of the CUP model are further away from the true π-values than the estimates of the BetaBin model.*

Similar results are obtained if the shifted Binomial distribution is used in the preference part and therefore the CUB is obtained if the uncertainty part is determined by the Uniform distribution. Now we compare Mix(Binomial, Uniform) (or CUB) with Mix(Binomial, Beta-Binomial). Figure 3 and 4 show the same setting as before, they compare the Beta-Binomial distribution with the Uniform distribution in the uncertainty part, but now the shifted Binomial distribution determines the preference component of both models. The figures show the results for $\gamma = -1$ as well as $\gamma = -2$. The well specified model can deal with different α and γ-values. But there are clear discrepancies in the misspecified models. For extreme α-values the estimates of γ and π in the misspecified models are poor. In the case of $\gamma = -1$ the π-values are underestimated for α-values smaller than one and overestimated for α-values greater than one. But for $\gamma = -2$ the opposite behaviour is observed. In both cases the γ estimates show the same trend. In Table 4 and 5 the results for all combinations are displayed. In general, there is clear discrepancy in the misspecified models but the direction (i.e. over or underestimation of the parameter) can vary. If the Uniform distribution is the true uncertainty component the CUB-model seems to be a bit closer to the true π-values than the model with the Beta-Binomial distribution. But the log proportions are close to zero so that the differences of the π-estimates in both models are very small. Moreover, in the BetaBin model the uncertainty component has

π	γ	α								
		0.01	0.1	0.25	0.5	1	2	4	10	100
0.5	-1	0.0035	0.0063	0.0068	0.0060	0.0051	0.0048	0.0051	0.0056	0.0032
0.7	-1	0.0055	0.0082	0.0094	0.0089	0.0075	0.0065	0.0071	0.0081	0.0042
0.8	-1	0.0076	0.0116	0.0114	0.0110	0.0097	0.0095	0.0098	0.0091	0.0052
0.5	-2	0.0022	0.0046	0.0049	0.0051	0.0043	0.0040	0.0040	0.0043	0.0029
0.7	-2	0.0028	0.0057	0.0067	0.0070	0.0058	0.0056	0.0057	0.0060	0.0033
0.8	-2	0.0039	0.0072	0.0086	0.0084	0.0075	0.0071	0.0077	0.0073	0.0043

TABLE 3: *Absolute Differences that measure the discrepancy between the estimated and the true Beta-Binomial distribution.*

to be estimated which is more difficult than assuming that α is exactly fixed at 1 as in the CUB model. In all other cases the BetaBin model clearly outperforms the CUB-model in terms of accuracy of the parameter estimates.

π	γ	α								
		0.01	0.1	0.25	0.5	1	2	4	10	100
0.5	-1	7.9908	6.9918	5.6090	3.8647	-0.2223	3.1404	4.3818	4.9940	5.4806
0.7	-1	6.3887	5.6636	4.4604	2.7079	-0.2538	2.0550	3.3342	4.0105	4.2475
0.8	-1	5.3687	4.7483	3.7107	2.0522	-0.0751	1.4637	2.4287	3.1594	3.3904
0.5	-2	6.7311	6.9559	6.2788	3.5601	-0.1111	1.8015	3.2645	3.2836	3.4449
0.7	-2	5.6132	4.9745	3.6663	1.7816	-0.0643	0.7877	1.5349	2.0155	2.2496
0.8	-2	4.9915	4.1609	3.0496	1.0264	-0.0259	0.6279	0.9362	1.3346	1.4698

TABLE 4: *Log proportions of γ-values. Positive values indicate that γ estimates of the CUB model are further away from the true γ-values than the estimates of the Mix(Binomial, Beta-Binomial) model.*

π	γ	α								
		0.01	0.1	0.25	0.5	1	2	4	10	100
0.5	-1	2.0685	1.7739	0.9497	-0.0189	-0.0027	0.6860	1.8735	3.2835	4.0075
0.7	-1	1.2640	0.7599	0.7672	0.3249	0.0407	0.3471	0.8965	1.4374	1.9017
0.8	-1	1.6844	1.2142	0.7175	0.4247	-0.0611	0.0645	0.2233	0.8416	1.1060
0.5	-2	4.4434	5.5704	5.9981	3.1680	0.0910	1.0716	2.1484	2.7990	3.2827
0.7	-2	3.8922	3.2800	2.2376	0.6130	-0.1299	0.6281	1.4431	2.1902	2.7078
0.8	-2	3.3097	2.3723	1.4503	0.3256	0.1253	0.5703	1.1688	1.6480	1.9694

TABLE 5: *Log proportions of π-values. Positive values indicate that π estimates of the CUB model are further away from the true π-values than the estimates of the Mix(Binomial, Beta-Binomial) model.*

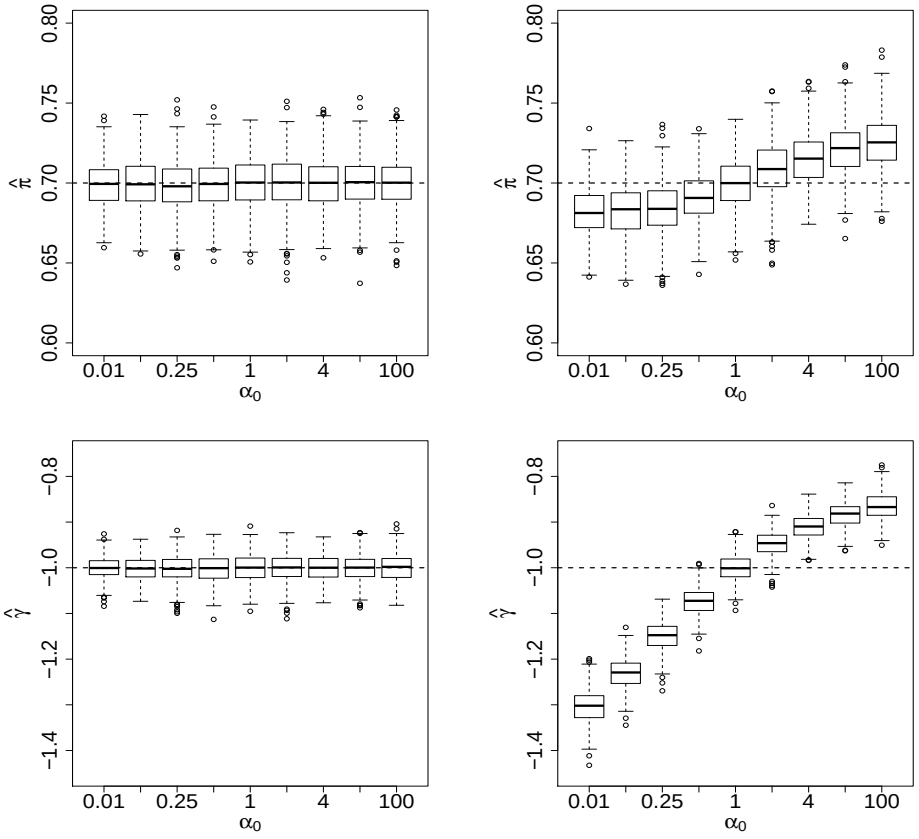


FIGURE 3: *Comparison of the estimated parameters $\hat{\pi}$, $\hat{\gamma}$ between the Mix(Binomial, Beta-Binomial) model on the left and the Mix(Binomial, Uniform) (or CUB) model on the right for $\pi = 0.7$ and $\gamma = -1$. The true α-values are $\{0.01, 0.1, 0.25, 0.5, 1, 2, 4, 10, 100\}$. The MSE of α is only reasonable for the BetaBin model.*

5 Application: Satisfaction with the Health Service in European Countries

To illustrate the new model we use the European Social Survey which measures the behaviour, attitudes and beliefs of populations in various European countries. We use the data of the 7th round in 2014, which is available at `http://www.europeansocialsurvey.org`. We focus on the attitude concerning the state of the health services measured on a Likert Scale from 0 "extremely bad" to 10 "extremely good". The scale was shifted to $1 \ldots 11$ to meet the requirements of the models. The covariates are gender (1: female), the age in decades (centered at 50), citizenship, the area of living (1: "big city" as reference, 2: "suburbs or

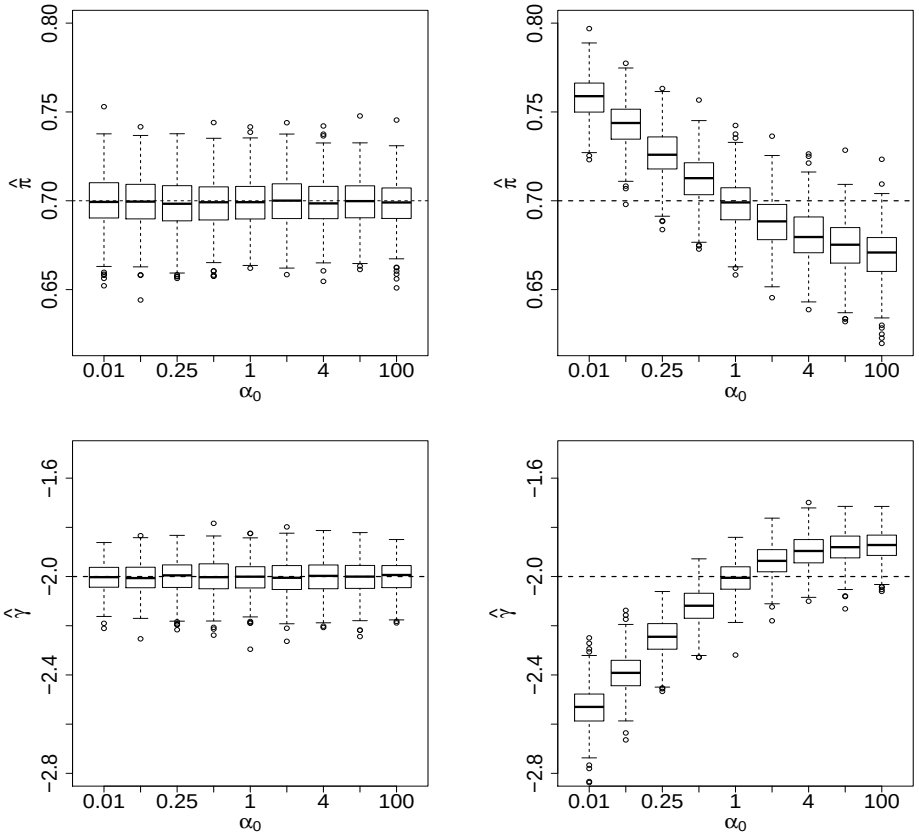


FIGURE 4: *Comparison of the estimated parameters $\hat{\pi}$, $\hat{\gamma}$ between the Mix(Binomial, Beta-Binomial) model on the left and the Mix(Binomial, Uniform) (or CUB) model on the right for $\pi = 0.7$ and $\gamma = -2$. The true α-values are $\{0.01, 0.1, 0.25, 0.5, 1, 2, 4, 10, 100\}$. The MSE of α is only reasonable for the BetaBin model.*

outskirts of a big city", 3: "town or small city", 4: "country village", 5: "farm or home in the countryside"), the smoke behaviour (1: "I smoke daily", 2: "I smoke but not every day", 3: "I don't smoke now but I used to", 4: "I have only smoked a few times", 5: "I have never smoked" as reference) and if the person is handicapped in its daily activities in any way by any longstanding illness, disability, infirmity or mental health problem (1: "yes a lot", 2: "yes to some extent", 3: "no" as reference).

An identical model with the same covariates is fitted separately for several countries. We give detailed results for Germany and compare the estimated uncertainty propensity and gender effects across countries.

Table 6 shows the estimates of the Beta-Binomial model for Germany with

	estimate	BS.sd	BS.2.5	BS.97.5	
female	0.2778	0.0751	0.1486	0.4385	
age	0.0677	0.0239	0.0237	0.1181	
age^2	-0.1009	0.0122	-0.1283	-0.0798	
German citizen: No	-1.3709	0.2270	-1.8828	-0.9374	
domicile: suburb	0.1442	0.1405	-0.1212	0.4177	
domicile: town	0.2566	0.1082	0.0574	0.4792	
domicile: village	0.2402	0.1106	0.0366	0.4747	$\hat{\gamma}$
domicile: countryside	0.0925	0.2153	-0.3162	0.5384	
handicapped: a lot	0.4302	0.1752	0.1254	0.7786	
handicapped: to some extent	0.4212	0.0999	0.2319	0.6397	
smoke: daily	0.3879	0.1175	0.1900	0.6403	
smoke: not every day	0.3936	0.2157	-0.0041	0.8214	
smoke: no, but used to	0.1042	0.0994	-0.0715	0.3067	
smoke: only a few times	-0.2471	0.1279	-0.4953	0.0035	
(Intercept)	3.8184	1.5363	1.8803	8.1662	
female	-2.3892	1.1699	-5.1968	-0.7173	
age	-0.6522	0.4172	-1.9058	-0.1083	
age^2	0.2528	0.1510	-0.0707	0.5546	$\hat{\boldsymbol{\alpha}}$
handicapped: a lot	-3.5315	1.5560	-6.6147	-1.0599	
handicapped: to some extent	-1.8433	1.2455	-3.7856	0.2417	
$1-\hat{\pi}$	0.1177	0.0349	0.0995	0.2123	

TABLE 6: *Estimates for state of health services in Germany, first group of estimates indicates effects on preference, second group indicates effects on uncertainty; BS.sd, BS.2.5, BS.97.5 refer to the bootstrap standard error and the quantiles for 2.5% and 97.5%, respectively.*

a cumulative model in the structure part. In the upper panel the effects on the preference part are displayed. Positive values indicate less satisfaction with the health services. One can see that females are less satisfied with the health services in Germany than men. Persons who are not German citizen are happier with the health services than German citizens. It is often discussed if there is a difference between urban and rural health service supply. According to the model, responders living in a town or in a village are significantly less happy with the health services than people living in a big city. For people living in the countryside or suburbs the difference to people living in a big city is non-significant. Also handicapped persons are less satisfied with the health services than non-handicapped persons. In the lower part the response style effects are displayed. Positive values indicate a tendency to the middle, negative values indicate a tendency to extreme categories. This follows from the parametrization of the α-values of the Beta-Binomial distribution, because for positive estimates one obtains exp(estimate) > 1 and therefore α increases. One sees that females tend to choose more extreme categories than men. Handicapped persons also prefer more extreme categories than non-handicapped persons.

In addition to giving estimates we use visualization tools to make the found effects easily accessible. In particular we use two-dimensional plots of the effects found in the preference part and the uncertainty part of the model. In the latter we use the response style parameters. More concrete, we plot the $\hat{\alpha}$ and

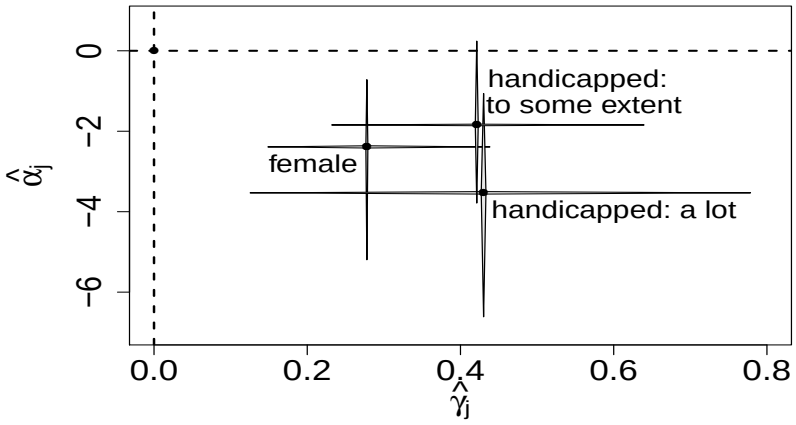


FIGURE 5: *State of health services in Germany: Gender and Handicap Effects*

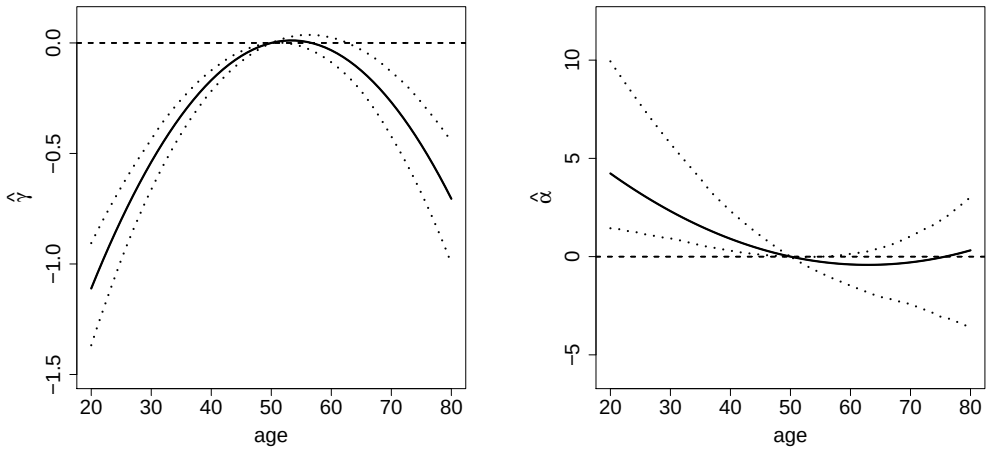


FIGURE 6: *State of health services in Germany: Age effects on preference (left) and uncertainty (right).*

$\hat{\gamma}$ values together with the confidence intervals obtained by bootstrap to obtain a star for each binary variable and several stars for multi-categorical variables. Figure 5 shows the estimated effects $(\hat{\gamma}, \hat{\alpha})$ of gender and being handicapped. Positive values in the γ-dimension indicate a tendency to negative statements concerning the state of the health services, positive values in the α-dimension indicate a tendency to middle categories. Females tend to see the health servi-

ces more sceptically and tend to choose more extreme categories. The effect of being handicapped is stronger than the gender effect in terms of a preference to categories indicating scepticism. The effects of being handicapped are almost the same in the preference part but differ in the uncertainty part. If a person is more handicapped it tends to choose more extreme categories. The effects are all significant except of "handicapped: to some extent" in the uncertainty component α. We used the 2.5% and 97.5% quantiles of the bootstrap samples instead of the bootstrap standard errors, because the distribution of the bootstrap standard errors may be skewed.

The effect of age is displayed in Figure 6. The dotted lines correspond to point-wise 95% bootstrap confidence intervals. They are constructed in such a way that in every bootstrap sample the age curve is calculated. Then the point-wise 2.5% and 97.5% quantiles are used to draw the dotted lines. On the left hand side the effect of age on the satisfaction of the health services is shown. It becomes obvious that younger and older persons are more satisfied with the health services than persons in their 50s. The response style shows a different picture. Young persons below 50 years of age show a significant tendency to middle categories whereas for persons older than 50 years of age no significant tendency to middle or extreme categories can be detected.

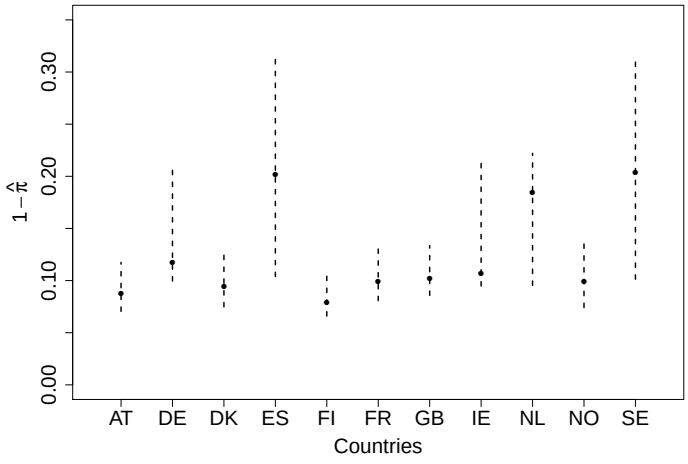


FIGURE 7: *State of health services: Importance of the uncertainty component in different countries*

For the comparison of countries we consider the performance of the BetaBin model, the estimates $1 - \hat{\pi}$ and the effect of gender across countries. The countries considered are Austria (AT), Germany (DE), Denmark (DK), Spain (ES), Finland (FI), France (FR), Great Britain (GB), Ireland (IE), Netherlands (NL), Norway (NO) and Sweden (SE). This set of countries is used as a representation

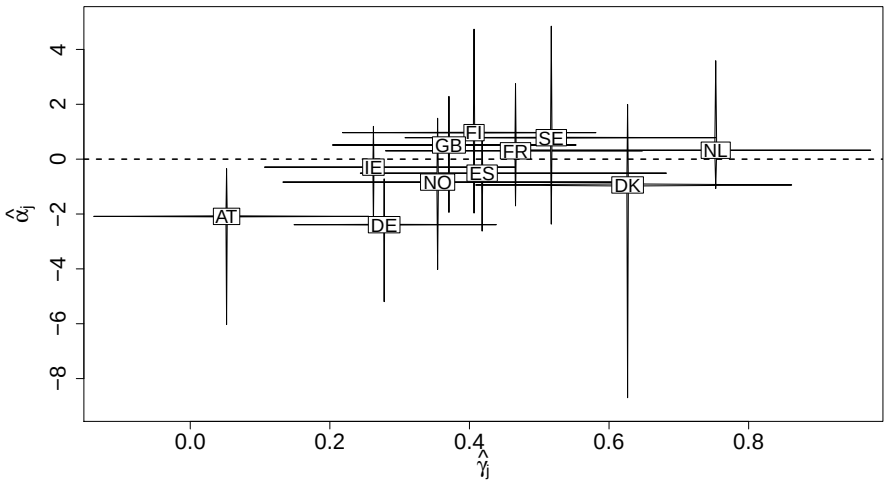


FIGURE 8: *State of health services: Influence of gender in different countries*

of European countries including large and small countries and from all parts of Europe.

There are some differences in the estimates of $1 - \hat{\pi}$, which is a measure of the importance of the uncertainty component. Large values indicate that the uncertainty component is strong. Figure 7 shows the proportions of the response styles. The dotted lines correspond to the 2.5% and 97.5% bootstrap quantiles. In Germany (DE) the tendency to response styles is in the middle range. In Spain (ES) and Sweden (SE) the model estimates show higher proportions of the response style. The lowest estimated proportions are found for Austria (AT) and Finland (FI), with values 0.0876 and 0.0793, respectively.

Figure 8 displays the effect of gender across the different countries. As in Figure 5 the x-axis corresponds to the effect on the preference structure and the y-axis to the effect of the response style. The confidence intervals are again obtained by bootstrap samples. For all countries the γ-parameters are positive which indicates that women are less satisfied with the health services of their country than men. The strongest effect can be found for the Netherlands (NL) and Denmark (DK) and the smallest for Austria (AT). The effects are significant for all countries with the exception of Austria (AT), for which the 95% bootstrap confidence interval contains zero.

In contrast, the gender effect in the response style is not homogeneous across countries. Positive α-parameters for Great Britain (GB), Finland (FI), France (FR), Netherlands (NL) and Sweden (SE) indicate that women show a weak ten-

dency to the middle category. In the other countries the estimated α-parameters are negative. However, except for Austria (AT) and Germany (DE) the effects are not significant.

Table 7 compares the performances of the proposed BetaBin model and the simple CUP model when fitting the models with all covariates included for each country. For all countries the deviance for the BetaBin model is smaller than for the CUP model. Also, for all countries except for Denmark the AIC values are smaller when fitting the BetaBin model. The AIC is defined by

$$AIC = -2l(\hat{\boldsymbol{\theta}}) + 2m\,,$$

where m is the number of model parameters, n is the number of observations and $l(\hat{\boldsymbol{\theta}})$ is the log-likelihood function computed at the maximum of the estimated parameter vector $\boldsymbol{\theta}$, which comprises all parameters of the model. The largest reduction can be found for Germany (reduction by 42 in the deviance and 30 in the AIC).

Countries	Deviance Uniform	Deviance BetaBin	AIC Uniform	AIC BetaBin
AT	7358	7342	7408	7404
DE	12864	12822	12914	12884
DK	6078	6070	6128	6132
ES	8553	8532	8603	8594
FI	8126	8112	8176	8174
FR	7797	7778	7847	7840
GB	9684	9665	9734	9727
IE	10354	10336	10404	10398
NL	7611	7594	7661	7656
NO	5677	5657	5727	5719
SE	7421	7393	7471	7455

TABLE 7: *Comparison of CUP and BetaBin models*

6 Concluding Remarks

It has been shown that the modelling of the uncertainty component by a Beta-Binomial distribution yields a more flexible model than traditional mixture models. The shape of the response style is allowed to depend on personal attributes and leads to a better understanding of the concept of uncertainty. The inclusion of covariate effects on the uncertainty also increases the interpretability of the model parameters. A simulation study showed that ignoring the response style may yield biased estimates. The applications demonstrate that the more flexible model outperforms the traditional model in most cases in terms of goodness-of-fit and AIC. Some of these findings have been demonstrated before in a Technical Report (Tutz and Schneider, 2017). Mauerer and Schneider (2019) used the

proposed model to examine response patterns to party placements on the immigration issue.

References

Agresti, A. (2010). *Analysis of Ordinal Categorical Data, 2nd Edition*. New York: Wiley.

Agresti, A. (2013). *Categorical Data Analysis, 3d Edition*. New York: Wiley.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods 17*(4), 665–678.

Bolt, D. M. and T. R. Johnson (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement 33*(5), 335–352.

Bolt, D. M. and J. R. Newton (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement 71*(5), 814–833.

Capecchi, S. and D. Piccolo (2016). Investigating the determinants of job satisfaction of italian graduates: a model-based approach. *Journal of Applied Statistics 43*(1), 169–179.

Clarke, I. (2000). Extreme response style in cross-cultural research: An empirical investigation. *Journal of Social Behavior & Personality 15*(1), 137–152.

De Boeck, P. and I. Partchev (2012). Irtrees: Tree-based item response models of the glmm family. *Journal of Statistical Software 48*(1), 1–28.

D'Elia, A. and D. Piccolo (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis 49*(3), 917–934.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B 39*(1), 1–38.

Eid, M. and M. Rauber (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment 16*(1), 20.

Gottard, A., M. Iannario, and D. Piccolo (2016). Varying uncertainty in CUB. *Advances in Data Analysis and Classification 10*(2), 225–244.

Iannario, M. (2010). On the identifiability of a mixture model for ordinal data. *Metron 68*(1), 87–94.

Iannario, M. (2012a). Hierarchical CUB models for ordinal variables. *Communications in Statistics-Theory and Methods 41*(16-17), 3110–3125.

Iannario, M. (2012b). Modelling *shelter* choices in a class of mixture models for ordinal responses. *Statistical Methods and Applications 21*(1), 1–22.

Iannario, M. (2014). Modelling uncertainty and overdispersion in ordinal data. *Communications in Statistics-Theory and Methods 43*(4), 771–786.

Iannario, M., M. Manisera, D. Piccolo, and P. Zuccolotto (2012). Sensory analysis in the food industry as a tool for marketing decisions. *Advances in Data Analysis and Classification 6*(4), 303–321.

Iannario, M. and D. Piccolo (2010). Statistical modelling of subjective survival probabilities. *Genus 66*(2), 17–42.

Iannario, M. and D. Piccolo (2012). Investigating and modelling the perception of economic security in the survey of household income and wealth. In C. Perna and M. Sibillo (Eds.), *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, pp. 237–244. Berlin: Springer.

Iannario, M. and D. Piccolo (2016a). A comprehensive framework of regression models for ordinal data. *Metron 74*(2), 233–252.

Iannario, M. and D. Piccolo (2016b). A generalized framework for modelling ordinal data. *Statistical Methods & Applications 25*(2), 163–189.

Iannario, M., D. Piccolo, and R. Simone (2018). *CUB: A Class of Mixture Models for Ordinal Data*. R package version 1.1.2.

Jeon, M. and P. De Boeck (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods 48*(3), 1070–1085.

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika 68*(4), 563–583.

Kankaraš, M. and G. Moors (2009). Measurement equivalence in solidarity attitudes in europe insights from a multiple-group latent-class factor approach. *International Sociology 24*(4), 557–579.

Manisera, M. and P. Zuccolotto (2014). Modeling rating data with nonlinear CUB models. *Computational Statistics & Data Analysis 78*, 100–118.

Marin, G., R. J. Gamba, and B. V. Marin (1992). Extreme response style and acquiescence among hispanics the role of acculturation and education. *Journal of Cross-Cultural Psychology 23*(4), 498–509.

Mauerer, I. and M. Schneider (2019). Perceived party placements and uncertainty on immigration in the 2017 german election. In M. Debus, J. Sauermann, and M. Tepe (Eds.), *Jahrbuch für Handlungs- und Entscheidungstheorie*, Volume 11. Springer. forthcoming.

McCullagh, P. (1980). Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society B 42*(2), 109–127.

Meisenberg, G. and A. Williams (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences 44*(7), 1539–1550.

Moors, G. (2004). Facts and artefacts in the comparison of attitudes among ethnic minorities. a multigroup latent class structure model with adjustment for response style behavior. *European Sociological Review 20*(4), 303–320.

Moors, G. (2010). Ranking the ratings: A latent-class regression model to control for overall agreement in opinion research. *International Journal of Public Opinion Research 22*(1), 93–119.

Piccolo, D. (2015). Inferential issues on CUBE models with covariates. *Communications in Statistics-Theory and Methods 44*(23), 5023–5036.

Poessnecker, W. (2015). *MRSP: Multinomial Response Models with Structured Penalties*. R package version 0.6.11.

Rosmalen, J. V., H. V. Herk, and P. Groenen (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research 47*(1), 157–172.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology 44*(1), 75–92.

Simone, R. and G. Tutz (2018). Modelling uncertainty and response styles in ordinal data. *Statistica Neerlandica*, doi: 10.1111/stan.12129.

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge: Cambridge University Press.

Tutz, G. and M. Schneider (2017). Mixture models for ordinal responses with a flexible uncertainty component. Technical Report 203, Department of Statistics LMU Munich.

Tutz, G., M. Schneider, M. Iannario, and D. Piccolo (2017). Mixture models for ordinal responses to account for uncertainty of choice. *Advances in Data Analysis and Classification 11*(2), 281–305.

Van Herk, H., Y. H. Poortinga, and T. M. Verhallen (2004). Response styles in rating scales evidence of method bias in data from six eu countries. *Journal of Cross-Cultural Psychology 35*(3), 346–360.

von Davier, M. and J. Rost (1995). Polytomous mixed rasch models. In *Rasch Models*, pp. 371–379. Berlin: Springer.

von Davier, M. and K. Yamamoto (2007). Mixture-distribution and HYBRID rasch models. In M. von Davier and C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models*, pp. 99–115. Berlin: Springer.

Yee, T. W. (2016). *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.0-3.

A.3. Perceived Party Placements and Uncertainty on Immigration in the 2017 German Election

Mauerer, I. and M. Schneider (2019a): Perceived party placements and uncertainty on immigration in the 2017 German election. In Debus, M., M. Tepe and J. Sauermann (Eds.), *Jahrbuch für Handlungs- und Entscheidungstheorie: Band 11*, pp. 117–143. Wiesbaden: Springer, doi:10.1007/978-3-658-23997-8.

Perceived Party Placements and Uncertainty on Immigration in the 2017 German Election

Ingrid Mauerer, Micha Schneider*

Abstract

Almost all national election studies contain policy scales that are intended to measure where respondents perceive parties or candidates on central campaign issues. These placements form the basis for models of survey responses, party perceptions, and voter choice. It is well known that the placements might be affected by uncertainty. We use the finite mixture model 'BetaBin' to study response patterns to party placements on policy issues. The model consists of a placement part and an uncertainty part. Whereas the placement part of the model accounts for lower and higher placements on the ordinal scales, the uncertainty component accounts for tendencies to locate the parties on the middle or at the extremes of the policy scales. We use the 2017 German national election (Study-No. ZA6800, GLES 2017, Cologne: GESIS Data Archive) and apply the model to the immigration issue. Our results demonstrate that uncertainty strongly influences the respondents' perceptions of most parties. Neglecting this structure leads to worse models as indicated by performance measures.

Keywords: Party Placements, Uncertainty, Mixture Models, BetaBin, 2017 German National Election

*The order of the authors is in alphabetical order. Both authors contributed equally.

1 Introduction

Almost all national election studies around the world contain policy scales. These policy scales were developed to assess the policy preferences of citizens and the positions of political figures on central campaign issues. A lot of theoretical and empirical concepts rely on the respondents' stated policy preferences and political perceptions. They form the basis for models of survey responses, party perceptions, policy choices, and voter choice. In public opinion research, several studies explore citizens' policy preferences (i.e., self-placements) on policy scales and test rival explanations of variability in attitudes due to uncertainty, ambivalence or equivocation. Alvarez and colleagues focus on policy choices of American citizens on public policies such as abortion (Alvarez and Brehm, 1995), racial policies (Alvarez and Brehm, 1997) or Internal Revenue Services (Alvarez and Brehm, 1998) (see also Alvarez and Brehm, 2002). In a similar vein, Harbers et al. 2013 explore response variability in Left-Right placements among the Latin American electorate, and De Vries and Steenbergen (2013) examine European citizens' ambivalence in attitudes toward European integration.

Citizens' policy preferences and their perceptions about party platforms also play a central role in spatial voter choice models. These models assume that parties take stances on issues and that voters can perceive these stances (Downs, 1957; Davis et al., 1970; Campbell et al., 1960). However, Shepsle (1972) and Enelow and Hinich (1981) reasoned that voters might be uncertain about the positions parties or candidates take on policies due to limited information on the side of the voters or position blurring on the side of the parties. They propose to represent party or candidate positions by probability distributions instead of single points. A few empirical studies account for voter uncertainty and incorporate it into the choice rule (e.g., Bartels, 1986; Gill, 2005; Berinsky and Lewis, 2007). These studies highlight that electoral decisions are not only systematically related to spatial distance but also to uncertainty about party platforms. Another recent study found that also voters are quite uncertain about their policy preferences and show a considerable degree of inconsistency (Stoetzer, 2017). In addition, not all parties put the same emphasis on the same issues and the reliance on issues when voting is party-specific (Mauerer et al., 2015; Mauerer, 2016). As a result, citizens might be uncertain what position they should ascribe when it comes to specific issues and parties.

If ordinary citizens do elect parties or candidates that best represent their preferences on public policies, it is necessary to understand in the first place how they perceive the parties' policy platforms and what role uncertainty plays in these perceptions. As uncertainty about party platforms can hamper democratic representation, we need models of survey responses that are able to detect the impact of uncertainty and to account for different response patterns due to uncertainty. Such insights will help us to understand how the electorate incorporates uncertainty into their political perceptions. This will then add to our understanding of the impact of policy-oriented decision making and its electoral consequences.

The approach we develop in this paper is a model of political perceptions that allows detecting how the perceptions of party platforms among ordinary citizens are structured. It is a model of survey responses to ordinal policy scales where specific response styles capture the uncertainty structure. The model belongs to the class of mixture models for ordinal data that are able to account for both the placement structure and the uncertainty structure of responses. We apply the so-called BetaBin model (Tutz and Schneider, 2017) that can handle different response patterns when citizens are uncertain where to place the parties on the policy scales: a tendency to the middle category and a tendency to extreme categories. Whereas the concentration in the middle category is widely known (see, e.g., Aldrich et al., 1982; Alvarez and Franklin, 1994), the approach we develop does not only account for this tendency but also for response styles to extreme categories, which has been discussed by Baumgartner and Steenkamp (2001) for example. Both the placement and the uncertainty structure of responses can be related to covariates. Since there are little theoretical and empirical insights into the mechanisms of how respondents' political perceptions are structured, we exploratively evaluate different sets of explanatory variables. We examine predictors that relate to cognitive processes or the respondents' information costs, to the relationship between the respondent and the party to be located at the policy, to issue characteristics, and standard demographics.

Compared to existing approaches, we see three main advantages of the proposed model. First, instead of modeling the variance by additional scale parameters as done in the heteroscedastic regression model (e.g., Harvey, 1976; Alvarez and Brehm, 1995), our approach can model specific response styles, namely the tendency to the middle, random choice or to extreme categories. Thus, we are able to detect particular structures of uncertainty which can be explained by covariates. Using models with scale parameters can only model high or low variance, but the variability is still rather unstructured. Second, the model is designed for ordinal responses, as compared to previous work relying on the logit/probit model for binary response (e.g., Alvarez and Brehm, 1995) or on the linear regression model (e.g., Harbers et al., 2013; De Vries and Steenbergen, 2013). It is well known that these models are not the best choice for modeling ordinal response data. Third, we do not need additional survey questions that directly ask respondents how certain they are about party or candidate positions, which are very rare in surveys.

We use the 2017 German national election study (Rossteutscher et al., 2017) to demonstrate the advantages of the proposed model. The election study contains typical eleven-point issue scales to measure the positions of parties on issues of current concern, such as immigration, taxes and climate change. On these ordinal scales, respondents were asked to place the parties. We apply the model to the immigration issue that played a significant role in the 2017 election campaign with different parties being more or less clear or ambiguous in the position they offered on it. We examine where the respondents perceive the major German parties on this central campaign issue and what role uncertainty plays in these perceptions. Our results show that the BetaBin model provides fruitful and new

insights into the perceptions of party platforms and outperforms traditional cumulative models without uncertainty structure. Including the uncertainty structure leads to better model performances, and therefore increases our understanding of political perceptions. Uncertainty strongly determines the respondents' perceptions of most parties on the immigration issue. Whereas the respondents expressed a clear preference where to place the AfD, they exhibit major difficulties in locating the CDU and the FDP.

2 Measuring and Modeling Uncertainty

There are some empirical approaches in the literature on how to measure and model uncertainty in party platforms. Here, we give a very brief overview of the most important models and approaches. One way to deal with variability and uncertainty in survey responses is to rely on range formats that adjust the traditional seven-point or eleven-point policy scales (see, e.g., Tomz and Van Houweling, 2009; Aldrich et al., 1982; Alvarez, 1999). Another approach is to stick to the original policy scales and to design additional survey questions to measure and examine uncertainty variability of survey responses (e.g., Alvarez and Franklin, 1994). These questions directly ask respondents how certain they are about candidate or party positions after they have provided these placements. However, only a few electoral studies have included self-reports on uncertainty yet.

Instead of relying on survey-based measures of uncertainty using self-assessments of respondents, Bartels (1986) proposes the following two-stage procedure to examine the impact of issue uncertainty on individual voting behavior. First, he develops a model of survey responses to assess the respondents' uncertainty in party placements. He takes refused answers as an indicator of uncertainty. The basic idea is that respondents who are uncertain are not able to give a placement at all. If a respondent refuses to place a candidate, this is interpreted as uncertainty which can be modeled as a function of observable characteristics of the respondent, the candidate, and the political environment. Based on the estimated probabilities of non-response that should capture the variance of candidate perceptions, in the second stage, he estimates a voter choice model to assess the importance of uncertainty in individual voting decisions. In both stages, he uses a linear probability approach.

Another idea is to model the variance by a heteroscedastic regression model introduced by Harvey (1976). In this case, the variance of the disturbance is modeled by covariates. Alvarez and Brehm (1995) apply a heteroscedastic binary probit model to analyze attitudes toward abortion in the U.S. electorate. Harbers et al. (2013) and De Vries and Steenbergen (2013) use a heteroscedastic linear regression model. These approaches have the disadvantage that they are not designed for small ordinal response scales. Using heteroscedastic linear regression for ordinal responses can lead to several difficulties. The error terms might not be normally distributed, and the linear regression might predict values lower, in between or above the response scale. Furthermore, it is not designed to

measure specific response styles such as a tendency to middle or extreme categories.

The model by Rozenas (2013) can be seen as a combination of non-response (Bartels, 1986) and variance heterogeneity (Harvey, 1976) which lead to a complex model with hyper-parameters and the necessity of choosing appropriate prior distributions. Another approach was developed by Gill (2005) who combines uncertainty with the concept of entropy. The entropy approach is based on an aggregate measure of issue uncertainty that uses information on the survey question, the issue to be evaluated, attributes of the candidates or parties as well as aggregated responses by the whole sample. In contrast to Bartels (1986) who imposes a homogeneous uncertainty threshold across respondents to model uncertainty, the entropy approach is based on an uncertainty term that is still the same across respondents but varies across issues and candidates.

The existing approaches deal very differently with missing values. The crucial question when relying on missing data in the response to measure uncertainty is whether there is a particular mechanism for generating the missing data. Empirical applications based on pure heteroscedastic models (as, e.g., Harvey, 1976) do not make use of any missing data and rely only on observed values. Contrarily, Bartels (1986) and Rozenas (2013) argue that missing data in the response is caused by the uncertainty of the respondents and related to covariates. This might be the case but maybe not the only or major process of generating missing values in the response structure. Respondents might have a clear position but do not want to report it because of social desirability, which is quite probable when it comes to delicate questions or policies. Another reason might be that respondents just skip the question because of time limitations or lack of motivation. In such cases, missing values consist of both uncertain and certain placements. Since we usually do not know the true data generating process of missing data, we prefer to exclude the missing values (including 'don't know' answers) from the analysis instead of assuming that missing data in the response is directly linked to the uncertainty of the respondents.

3 Response Styles and Variability in Uncertainty across Parties

The literature is in agreement that uncertainty is inherently subjective and that particular segments are more certain or uncertain about party placements. Previous research on response patterns mainly suspected that respondents show a tendency to the middle of the scale due to limited information, when they are not politically interested or involved. On the policy scales, the middle categories reflect moderate positions. Whereas the concentration in the middle category is widely known (see, e.g., Aldrich et al., 1982; Alvarez and Franklin, 1994), the approach we develop in this paper can account for several kinds of uncertainty – especially the tendency to middle or extreme categories. Particularly the response style to extreme categories, i.e., a tendency to ascribe parties extreme policy stances, seems to be very promising because the response patterns might not only be the

result of cognitive processes of citizens but might also stem from how parties behave on issues, what strategies parties pursue on particular issues. Parties might take ambiguous stances, blur their positions that induce uncertainty on the side of the voters where the party actually positions itself. However, parties also often overshoot their positions (Kedar, 2005a,b). Therefore, one might also expect that respondents show a tendency to place the parties at the extremes when they are uncertain.[1]

In addition, we expect differences in uncertainty patterns across parties. In general, parties often pursue different strategies on different issues, yielding different levels of uncertainty in the position respondents ascribe to parties. We argue that this variation relates to the underlying party system and resulting dynamics of party competition that reduce or increase uncertainty in perceived party platforms. We apply the model to the issue of immigration. Why should there be different levels of uncertainty in the position respondents ascribe to parties on the immigration issue? We expect that there is systematic variation in uncertainty on the immigration issue due to party family.

Immigration lies at the core of Inglehart's (1997) post-materialist dimension, and therefore represents an important 'new politics' issue. Immigrants issues grew increasingly salient in Western Europe in the last decades and gave rise to the emergence of new competitors on the radical right of the political spectrum. A considerable amount of scholarly attention has been paid to the explanation of the electoral fortunes of populist radical right parties (see, e.g., Givens, 2005; Kitschelt and McGann, 1995; Mudde, 2007; Ignazi, 2003; Art, 2011). Also, much scholarly work has been devoted to clarifying our understanding of the dynamics of party competition on immigrants issues (see, e.g., Meguid, 2005; Ivarsflaten, 2008; Norris, 2005; Bale, 2003, 2008; Abou-Chadi, 2015; Pardos-Prado et al., 2014). The recent studies point out that right-of-center and populist radical right parties have a strategic advantage over their competitors on immigrants issues. These parties lay particular emphasis on their core issue immigration by profoundly polarizing on them. In the light of their issue portfolio and long-term ideological backgrounds, center-right and radical right parties increase the saliency of immigrants issues by strongly politicizing it and taking unambiguous restrictive stances. As a result, one might expect that citizens are quite certain in their perceptions about what positions radical right and right-of-center parties offer on immigration.

Other studies focus on the mainstream left and investigate the electoral strategies of this party family concerning immigration (see, e.g., Bale et al., 2010; Alonso and Da Fonseca, 2012). For instance, Bale et al. (2010) examine the strategic responses of social-democratic parties. They find that these parties face a 'triple challenge' due to the rise and the success of extreme right parties: (1) Populist radical right parties mainly campaign on immigrants issues – and therefore increase the saliency of issues – that are traditionally owned by right-of-center parties; (2) the extreme right attempts to mobilize the working-

[1]Baumgartner and Steenkamp (2001) provide other reasons. For instance, extreme response styles (ERS) can be seen as a 'reflection of rigidity'. Baumgartner and Steenkamp (2001) and Vaerenbergh and Thomas (2013) give an overview of different response styles.

class who is habitually linked to left-of-center parties; and (3) populist radical right parties ease to form center-right governments (see also Bale, 2003, 2008). Based on this line of reasoning, one might expect that citizens have difficulties in locating the mainstream left on immigrants issues.

4 A Model of Perceived Party Platforms under Uncertainty

Our modeling approach is based on the idea that variability in survey responses can be modeled by mixture models. In the framework of mixture models, any density f can be represented as a combination of a finite set of densities so that

$$f = \sum_{m=1}^{M} \pi_g f_g, \tag{1}$$

where $0 \leq \pi_g \leq 1$ is the mixture proportion or weight, and M is the number of densities used to describe the density f. Mixture models are widely used. An introduction to this model class is given by McLachlan and Peel (2000). Iannario and Piccolo (2016a) and Iannario and Piccolo (2016b) provide an overview of mixture models for ordinal data.

When studying perceived party placements, several requirements have to be considered. First, the number of densities M can be restricted to two, as we are only interested in the placement structure and uncertainty structure of survey responses. Second, we have ordinal responses so that we need density functions that are appropriate for this data type; any continuous densities cannot be considered. Third, we would like to use densities that are the best choice for modeling both components. However, the best choice is not always to use densities from the same type for both components.[2] By including an uncertainty component, we can account for specific response styles. One basic response style can be represented by a uniform distribution corresponding with a random choice of response category as done for example by D'Elia and Piccolo (2005) or Tutz et al. (2017). We use the so-called BetaBin model proposed by Tutz and Schneider (2017), which is characterized by flexible modeling of the response style and placement structure.

4.1 Model Formulation

Let R_i be the observed response of an individual i to an ordinal policy scale taking the values $\{1, \ldots, k\}$. Y_i denotes an unobserved random variable that presents the deliberate choice, i.e., the real party placement. Let U_i be the unobserved uncertainty component modeling the type of response style. Both Y_i and U_i take ordered values from $\{1, \ldots, k\}$.

[2]Choosing two binomial distributions are considered as the same type.

Given explanatory variables $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$, the mixture model 'BetaBin' has the form

$$P(R_i = r|\boldsymbol{x}_i, \boldsymbol{w}_i) = \pi_i P_M(Y_i = r|\boldsymbol{x}_i) + (1 - \pi_i) P_U(U_i = r|\boldsymbol{w}_i). \tag{2}$$

$\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ are vectors of covariates for both components that can be identical, overlapping or completely different. π_i represents the mixture probability that measures the importance of the structured component in the mixture model. Thus, the observed response results from a discrete mixture of the placement and the uncertainty component.

For the placement component $P_M(Y_i = r|\boldsymbol{x}_i)$, any ordinal model would be possible. We use the following widely known cumulative logit model (see Tutz, 2012), also known as ordered or ordinal logit model:

$$\log\left(\frac{P(Y_i \leq r|\boldsymbol{x}_i)}{P(Y_i > r|\boldsymbol{x}_i)}\right) = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}, \quad r = 1, \ldots, k-1,$$

where γ_{0r} are the intercepts or thresholds and $\boldsymbol{\gamma}$ the estimated effects independent of r. Note that in the literature different notations of the cumulative logit model are used. Here a positive value of γ indicates that a lower category is more probable. The response style U follows a beta-binomial distribution, $U \sim \textit{Beta-binomial}(k|\alpha, \beta)$ with the mass function

$$f(u) = \begin{cases} \binom{k-1}{u-1} \frac{B(\alpha+u-1, \beta+k-u+1)}{B(\alpha,\beta)} & u \in \{1, \ldots, k\} \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha, \beta > 0$ are the parameters of the beta-binomial distribution. $B(\alpha, \beta)$ is the beta function defined as

$$B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt.$$

Since α and β are not identical with the location and scale of the distribution, the following reformulation is necessary:

$$\mu = \alpha/(\alpha + \beta), \quad \delta = 1/(\alpha + \beta + 1).$$

Now the expected value $E(U)$ and the variance $var(U)$ are given by

$$E(U) = (k-1)\mu + 1, \quad \text{var}(U) = (k-1)\mu(1-\mu)[1 + (k-2)\delta].$$

As $\delta \to 0$, the beta-binomial distribution converges to the (shifted) binomial distribution $B(k, \mu)$ with mean μ and support $\{1, \ldots, k\}$. The specific response styles characterized by a tendency to middle or extreme categories are determined by imposing the restriction $\alpha = \beta$, which lead to $\mu = 0.5$ and $\delta = 1/(2\alpha+1)$. Thus, the location of the distribution is always fixed at the middle of support. The only flexible parameter is δ or rather α. The smaller α, the larger δ and therefore the variance. Figure 1 shows the different shapes of this restricted beta-binomial distribution.

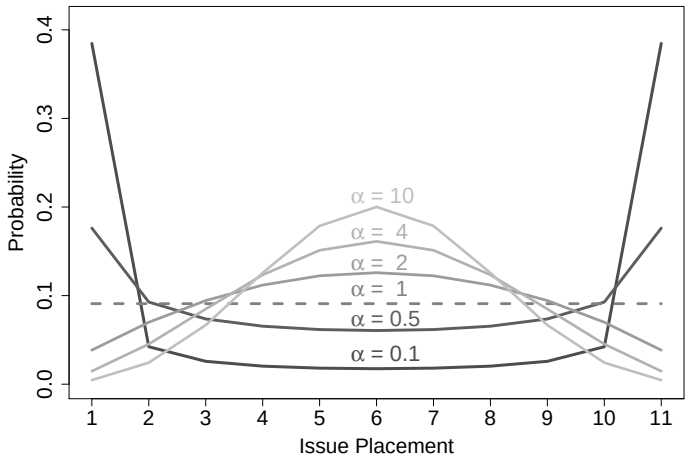


Figure 1: Probability mass on categories for various values of α for $k = 11$ categories.

α values larger than one correspond with a tendency to the middle categories, while α values smaller than one indicate a tendency to extreme categories. The distribution ranges from a (shifted) binomial distribution with the mode in the middle of support to almost a distribution with two equal point mass at the border of support. The first distribution corresponds with a strong tendency to the middle categories, the latter with a strong tendency to both extreme categories given by the minimum and maximum of k. All gradations between the two extreme cases are possible. The parameter α is linked to covariates $\boldsymbol{w}_i$ by

$$\alpha = \exp(\boldsymbol{w}_i^T \boldsymbol{\alpha}) = \exp(\alpha_0) \exp(\alpha_1)^{w_{i1}} ..... \exp(\alpha_m)^{w_{im}},$$

where α is the parameter of the restricted beta-binomial distribution and α_j gives the effect of the j-th covariate linked with exp() to α. Thus, α changes by the factor $\exp(\alpha_j)$ when w_{ij} increases by one unit, given all other variables in the model are kept constant. The exponential function ensures that α is always positive, although the effects of the covariates may be positive or negative. Positive α_j values lead to α values larger than one and indicate a tendency to middle categories. Negative α_j values lead to α-values smaller than one and indicate the tendency to extreme categories. For example, an effect of $\alpha_j = 1$ leads to $\alpha = \exp(1) = 2.71$ showing a tendency to the middle categories for the j-th covariate. For $\alpha = 1$ ($\alpha_j = 0$) one obtains the discrete uniform distribution (dashed line in Figure 1) corresponding with a random choice of a category.

4.2 Survey Responses to Party Placements

We apply the model of party perceptions to the immigration issue contained in the 2017 German national election study (Rossteutscher et al., 2017). The respondents were asked to state where they perceive the parties on an eleven-point scale with the following endpoints: 1 “Immigration should be facilitated” (pro) and 11 “Immigration should be re-

stricted" (contra). We restrict our analysis to the seven most important German parties: the Christian Democratic Union (CDU), its Bavarian sister party the Christian Social Union (CSU), the Social Democratic Party (SPD), the Liberal Party (FDP), the Greens, the Left Party, and the Alternative for Germany (AfD). We also excluded respondents that provided no answer or opted for the 'don't know' category. The stated positions of these parties on the immigration issue present the observed response R_i in Equation 2.

Figure 2 illustrates the distribution of party perceptions on the immigration issue. Since not all survey respondents were able to locate all the seven parties, the number of observations (out of 2179 total respondents) slightly differs. The minimum number of observations is 1387 for the FDP, and the maximum is 1949 for the CDU. For each of the eleven categories, the percentages are reported in Figure 2 so that the shape between the different parties are comparable even though the absolute numbers are not identical. We observe that the shape of the distributions is very different between the seven parties. Unsurprisingly, almost 70% of the respondents locate the AfD at the far right, resulting in a very skewed distribution. Also, the CSU is perceived as taking a rather contra-immigration stance, whereas the respondents place its sister party CDU closer to the middle categories without a clear modal value. The distributions of the perceptions for the FDP and SPD are more symmetric with modal values at 6 and 5, respectively. The perceptions of the Greens are skewed toward the pro-immigration pole, and the Left Party is perceived as taking the most pro-immigration stance. This data situation demands a flexible model that can handle all these different distributions, and therefore presents an ideal situation to demonstrate the benefits of the proposed mixture model.

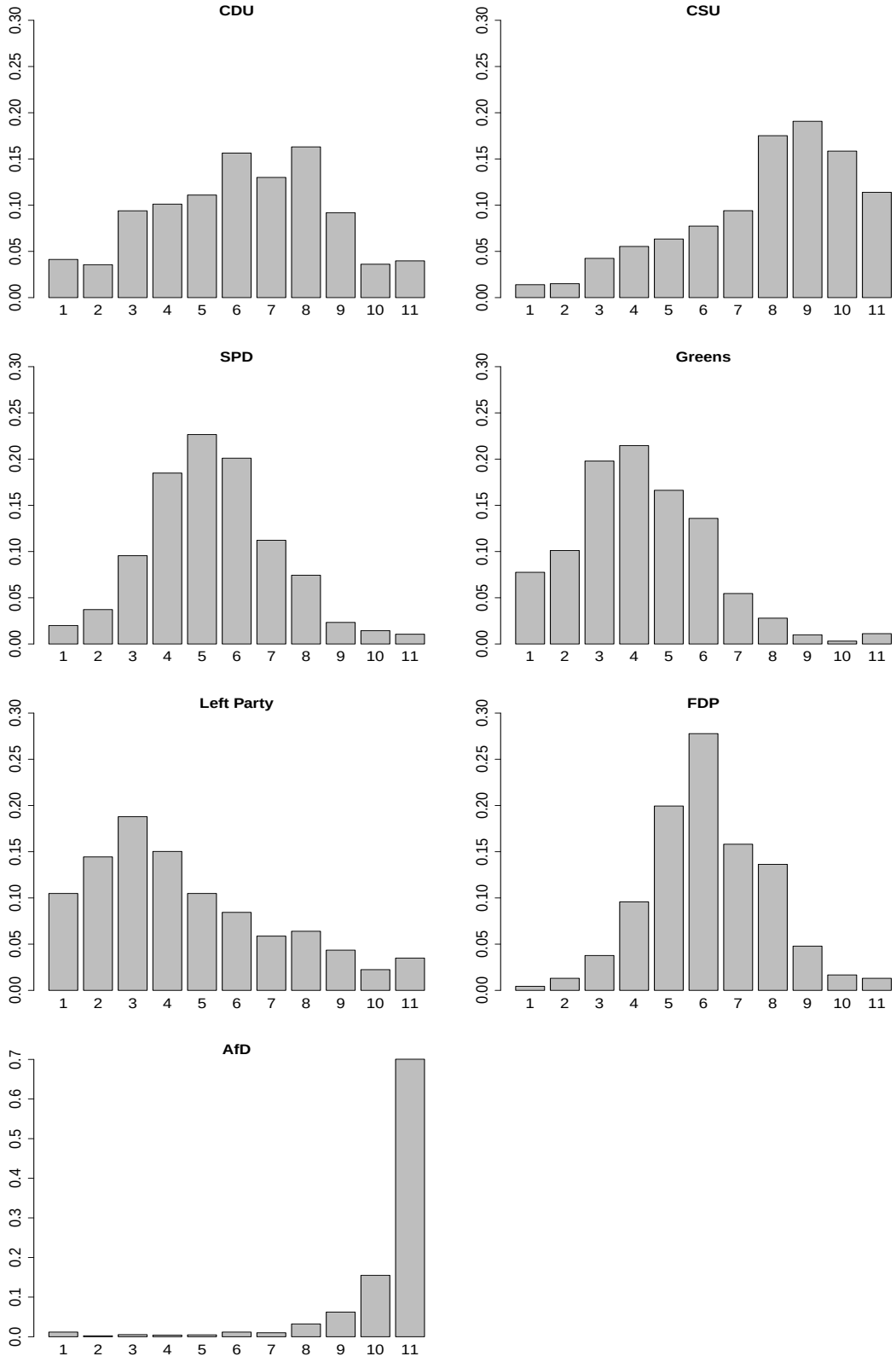
4.3 Predictors for Issue Placements and Issue Uncertainty

As outlined above, the approach can model the placement and uncertainty structure of political perceptions by covariates $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$. For both components, we use the same set of predictors we describe next.

4.3.1 Political Sophistication

The first set of explanatory variables accounts for cognitive processes or information costs and relates to the concept of political sophistication (e.g., Luskin, 1987, 1990; Delli Carpini and Keeter, 1993). Going back to Downs (1957), it is frequently argued that citizens who possess lower information costs tend to be more informed about the positions parties offer on central policies. Therefore, individuals with higher levels of political information are presumed to be less uncertain about party platforms. To operationalize the concept of political sophistication and to identify segments that might exhibit different response styles due to uncertainty or place parties into a particular direction, we explore a subjective and an objective measure: the strength of political interest, and political knowledge.

The level of political interest is usually measured by relying on respondents' self-reports. The 2017 German election study includes a question in which respondents were



Note: 1 "Immigration should be facilitated" (pro) and 11 "Immigration should be restricted" (contra)

Figure 2: Distribution of party perceptions on the immigration issue

asked to state their level of political interest on a five-point scale. We recoded the variable so that one gives the response "not interested at all" and five "very interested". To mea-

sure the respondents' level of political knowledge we rely on factual knowledge questions with right or wrong answers. Several studies have shown that factual political knowledge questions present good empirical indicators for the concept of political knowledge (Luskin, 1987, 1990; Delli Carpini and Keeter, 1993).[3] Based on the replies to five questions, we generated an additive knowledge score in which for each correct answer a value of one is assigned, whereas wrong answers and "don't know/no answer" responses give a value of zero. The first two questions concern the German electoral system.[4] In addition, the respondents were confronted with pictures showing three politicians, and they were asked to state the party each politician belongs to.[5] The answers are aggregated by counting the number of times a respondent correctly answered all five questions, resulting in a six-categorical variable running from zero to five (0 none correct, 5 all answers correct).

4.3.2 Party Attributes

One might also expect that respondents are more confident about where to place the party due to the relationship they have established with the respective party. The second set of covariates intends to capture the relationship between the respondent and the party to be located. One might expect that a long-standing leaning toward a party influences both where to place the party and the response patterns. So that those respondents who identify themselves with the party to be located are more certain about the position the party offers. On the contrary, when respondents do not identify themselves with the respective party, specific response styles due to uncertainty might be likely to occur. The same argument might apply to the sympathy of the parties' candidates. Party identification is a dummy variable with one indicating that the respondent identifies with the party to be placed and zero otherwise (i.e., no party identification or identification with any other party). For each of the seven parties, we generated such a dummy variable. As candidate evaluations, we consider feeling thermometers on eleven-point scales (1 very negative; 11 very positive).[6]

4.3.3 Issue Importance

Also, the issue itself might influence the respondents' perceived party placements and uncertainty. Respondents who consider the policy as important might have a clearer understanding of where to place the parties. When the issue is of personal importance, the respondent might have considered in more detail what the parties actually offer on it.

[3] For a recent comparative assessment, see Rapeli (2013).

[4] "Which one of the two votes is decisive for the relative strengths of the parties in the German parliament?"; "What is the percentage of the second vote a party needs to be able to definitely send delegates to the German parliament?".

[5] These politicians are Martin Schulz (SPD), Katrin Göring-Eckardt (Greens), and Christian Lindner (FDP).

[6] The candidates are Angela Merkel for the CDU, Horst Seehofer for the CSU, Martin Schulz for the SPD, Christian Lindner for the FDP, Cem Özdemir for the Greens, Sahra Wagenknecht for the Left, and Frauke Petry for the AfD.

Therefore, we might expect that as the level of personal issue importance increases, the respondents show less uncertainty response patterns. To identify the level of importance respondents ascribe to the immigration issue, we employ a common measure, self-reports. The 2017 German election study includes a question in which respondents are asked to state the importance of the immigration issue on five-point scales running from "not at all important" to "very important".[7]

4.3.4 Standard Demographics

Finally, the models account for the effects of standard demographics, including age and gender. Since East/West Germany constituted for a long time a major explanation for differences in public opinion in Germany, the models also control for this east-west divide. The variables are coded as follows: Age: centered around the sample mean, measured in decades; Gender: 1 (female), 0 (male); Former West/East Germany: 1 (West Germany), 0 (East Germany). As with the party placements, we excluded missing values on all these variables.

5 Empirical Results

For each party, we specified a separate BetaBin model. The models are estimated with the EM-Algorithm as described by Tutz and Schneider (2017). The result presentation is divided into three parts: We begin by examining the role uncertainty plays in the party perceptions on the immigration issue. Then, we present the estimates for the placement and the uncertainty part of the model. Finally, we systematically compare the proposed BetaBin model with the traditional cumulative model without uncertainty component based on performance measures.

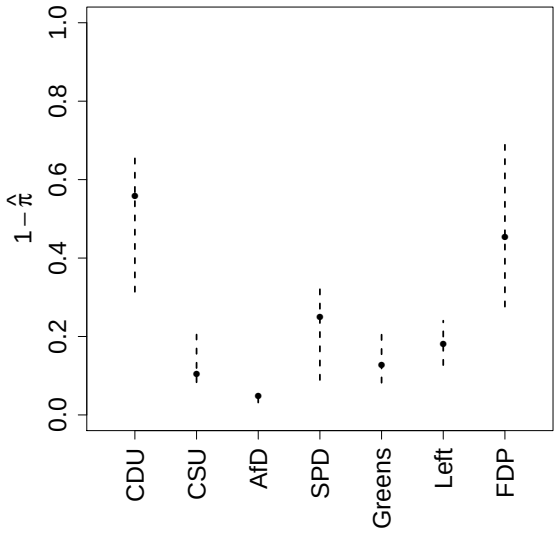
5.1 The Role of Uncertainty in Perceived Party Platforms

Let us first examine the mixture probability that measures the importance of the uncertainty part in the mixture model. $1 - \hat{\pi}$ is an estimate for the weight of the uncertainty component and can be interpreted as a measure of how clear or unambiguous the respondents perceive the party positions. The weight takes values between zero and one. A value of zero represents the traditional cumulative model without mixture and indicates the absence of any response styles, i.e., no tendency to the middle of the scale, no tendency to extremes, and no random party perception. A value of one gives a model without placement structure, i.e., no structure in placing the parties on the policy scales is detected. The higher the value of $1 - \hat{\pi}$, the stronger the uncertainty.

Figure 3 displays the estimated $1 - \hat{\pi}$ values for the seven German parties. The dotted lines correspond to the 2.5% and 97.5% bootstrap quantiles. As expected, we detect

[7] Corresponding question: "How important is this issue to you personally?".

the weakest uncertainty weight (0.05) for the location of the AfD. This indicates that respondents expressed a clear preference where to place the AfD on the immigration issue. Regarding the two Christian Democratic parties, the respondents exhibit much more uncertainty, 0.56 and 0.11, respectively. A higher level of uncertainty for these two right-of-center parties seems plausible because of the internal divisions on the immigration issue. The chancellor and leader of the CDU Angela Merkel, Thomas de Maizière (Minister of the Interior and member of the CDU), and Horst Seehofer (leader of the CSU) expressed quite different opinions about migrants and refugees. Thus, the respondents are more uncertain about both positions. However, the CSU could offer a much clearer position, as suggested by the considerable difference in uncertainty between the CDU and CSU. The behavior on the side of the parties is reflected in the uncertainty weights we estimate. Our model also estimates a large uncertainty weight for the FDP. Apparently, respondents show enormous difficulties in placing both the CDU and the FDP on the immigration issue, whereas the strength of uncertainty is in the middle range for the remaining parties, except for the AfD. Note that the uncertainty weight illustrates the strength of the uncertainty component regardless of whether the perceived party position is pro or contra immigration.



Note: Dotted lines correspond to the 2.5% and 97.5% bootstrap quantiles.

Figure 3: Importance of the uncertainty component $(1 - \hat{\pi})$

5.2 Party Placements and Response Style Effects

Table 1 and Table 2 give the results for the placement and uncertainty components of the models. The estimates for the placement part ($\hat{\boldsymbol{\gamma}}$) are displayed at the top, the estimates for the uncertainty response style effects ($\hat{\boldsymbol{\alpha}}$) at the bottom. We report for each effect the 2.5% and 97.5% quantiles of 500 non-parametric bootstrap samples. An estimate is considered as significant at the 5%-level when the bootstrap confidence intervals cover

the estimate but not zero. The estimates for the preference part ($\hat{\boldsymbol{\gamma}}$) are interpreted in the following way. Positive estimates indicate that lower categories become more likely meaning that the respondents tend to place the party toward pro immigration. Negative values indicate a tendency to locate the party in a higher category on the issue scale corresponding with contra-immigration stances.

We observe that the effects of political sophistication vary between the different parties. When the respondents' political interest increases, they tend to locate the SPD, the Greens and the Left closer toward pro-immigration stances. Higher political knowledge leads to a location of both Christian Democrat parties closer to contra-immigration positions, whereas the opposite is the case for the Greens and the Left. These results suggest that for the majority of the parties political sophistication does provide an explanation of placement structures. Regarding our second set of predictors, we find that candidate images are significant for all parties, except for the Left and FDP. However, the estimates differ in direction. Whereas an increase in candidate sympathy leads respondents to locate the CSU and the AfD toward pro-immigration positions, it yields placements closer to contra-immigration stances for the CDU, SPD, and the Greens. By contrast, party identification only impact on the placement of two parties, namely the CSU and the Greens. When respondents identify themselves with the Greens, they show a tendency to locate the party closer at pro-immigration stances. In the case of the CSU, the effect is negative so that CSU-party identifiers tend to place their party closer to contra-immigration positions. Concerning issue importance, there are almost no significant findings. By contrast, the demographics exhibit some interesting findings. We observe significant negative age effects for several parties. This implies that the older a person, the stronger the tendency to ascribe the parties more contra-immigration positions. In addition, females show a tendency to locate the Left closer at contra-immigration stances, whereas respondents based in former West Germany tend to shift the SPD, the Greens and the Left toward contra-immigration positions.

The uncertainty component contains the estimates for the shape of the uncertainty distribution (determined by $\hat{\boldsymbol{\alpha}}$), displayed at the bottom of Tables 1 and 2. Here, positive values indicate a tendency to the middle categories, whereas negative values suggest a tendency to the extremes of the scales. As in the placement structure, measures of political sophistication also show some significant effects on the response style dimension. The higher the political knowledge, the more respondents tend to locate the CDU, the CSU, the SPD, and the Left in middle categories. By contrast, political interest does not seem to influence the response styles. Examining the predictors that relate to the relationship between the respondent and the party to be located, we find the strongest effects for candidate images, whereas party identification shows no significant effects. We observe that an increased satisfaction with the candidate leads to a tendency to locate the CDU, the AfD, and the Left in middle categories. Also, issue importance only marginally impacts on response styles. Surprisingly, most demographic variables do not exhibit any

significant effects. We observe that older people tend to locate the CDU toward more extreme positions, females show a tendency to place the Greens in middle categories, and respondents from former West Germany ascribe extreme positions to both the AfD and the Left.

Table 1: Parameter estimates of BetaBin model I

	CDU				CSU				AfD			
	Estimate	BS.2.5	BS.97.5	sig	Estimate	BS.2.5	BS.97.5	sig	Estimate	BS.2.5	BS.97.5	sig
Placement Part												
Political Interest	-0.281	-0.454	0.049		-0.131	-0.278	0.018		-0.147	-0.299	0.006	
Political Knowledge	-0.234	-0.374	-0.064	*	-0.340	-0.462	-0.233	*	-0.079	-0.198	0.029	
Party Identification	0.423	-0.120	0.909		-0.771	-1.712	-0.127	*	0.055	-0.560	0.784	
Candidate Images	-0.183	-0.372	-0.065	*	0.164	0.113	0.238	*	0.250	0.198	0.296	*
Issue Importance	-0.091	-0.521	0.212		-0.068	-0.206	0.067		-0.127	-0.277	0.016	
Age	-0.167	-0.285	-0.053	*	-0.194	-0.283	-0.150	*	-0.145	-0.226	-0.065	*
Gender	-0.414	-0.748	0.006		-0.218	-0.468	0.024		-0.201	-0.471	0.050	
West Germany	-0.044	-0.463	0.314		-0.050	-0.245	0.204		0.183	-0.068	0.468	
Uncertainty Part												
Political Interest	0.127	-0.293	0.611		-0.123	-3.236	1.381		0.843	-2.930	6.211	
Political Knowledge	0.385	0.191	1.183	*	1.038	0.043	3.974	*	-0.417	-3.420	1.613	
Party Identification	0.230	-0.381	3.891		10.80	-1.063	255.78		1.176	-10.38	8.248	
Candidate Images	0.163	0.071	0.448	*	0.254	-0.697	1.761		0.545	0.768	2.711	*
Issue Importance	-0.264	-1.574	0.029		-0.485	-4.050	1.577		-1.088	-5.934	1.100	
Age	-0.192	-0.572	-0.069	*	-0.378	-1.872	0.633		0.202	-1.091	2.119	
Gender	-0.049	-0.845	0.704		0.818	-2.451	5.068		-0.365	-6.158	4.866	
West Germany	0.323	-0.532	1.083		-0.009	-3.813	3.082		-1.763	-8.680	-0.979	*
$1 - \hat{\pi}$	0.558	0.314	0.657	*	0.106	0.083	0.205	*	0.047	0.032	0.065	*
N		1949				1872				1715		

Note: Cut points of the placement part and intercept of the uncertainty part are not displayed.

Table 2: Parameter estimates of BetaBin model II

	SPD				Greens				Left				FDP			
	Estimate	BS.2.5	BS.97.5	sig	Estimate	BS.2.5	BS.97.5	sig	Estimate	BS.2.5	BS.97.5	sig	Estimate	BS.2.5	BS.97.5	sig
Placement Part																
Political Interest	0.347	0.185	0.544	*	0.426	0.310	0.588	*	0.405	0.276	0.584	*	0.039	-0.353	0.309	
Political Knowledge	0.100	-0.004	0.204		0.266	0.180	0.380	*	0.240	0.140	0.333	*	-0.214	-0.545	0.047	
Party Identification	0.211	-0.039	0.511		0.375	0.034	0.681	*	-0.195	-0.525	0.233		0.396	-0.860	1.505	
Candidate Images	-0.080	-0.156	-0.022	*	-0.184	-0.254	-0.118	*	-0.010	-0.068	0.042		0.135	-0.004	0.613	
Issue Importance	0.014	-0.133	0.246		0.095	-0.061	0.281		0.085	-0.049	0.240		-0.325	-1.020	-0.047	*
Age	-0.135	-0.198	0.054		-0.015	-0.074	0.050		-0.187	-0.253	-0.127	*	-0.090	-0.279	0.023	
Gender	-0.264	-0.493	0.001		-0.134	-0.368	0.097		-0.442	-0.677	-0.231	*	-0.067	-0.616	0.667	
West Germany	-0.257	-0.572	-0.014	*	-0.342	-0.603	-0.095	*	-0.870	-1.144	-0.615	*	0.249	-0.416	0.857	
Uncertainty Part																
Political Interest	0.116	-0.938	2.048		-0.259	-2.110	3.162		0.638	-0.541	2.482		-0.468	-1.908	0.100	
Political Knowledge	0.410	0.130	2.724	*	0.840	-0.683	2.598		1.325	0.321	2.129	*	0.530	-0.086	1.426	
Party Identification	-0.338	-3.615	2.891		-3.867	-7.972	13.73		-1.065	-6.663	3.798		7.921	-1.045	90.53	
Candidate Images	0.150	-1.061	1.105		0.440	-0.767	1.251		0.895	0.343	1.564	*	0.402	-0.273	1.466	
Issue Importance	-1.156	-4.253	1.466		-3.505	-4.907	1.490		-3.818	-5.044	-0.870	*	-0.022	-1.532	2.345	
Age	-0.461	-1.740	0.005		-0.728	-1.853	0.239		-0.234	-1.219	0.480		-0.234	-0.710	0.113	
Gender	-0.546	-3.498	3.071		5.846	0.505	8.671	*	0.341	-2.642	2.402		-0.102	-1.580	1.550	
West Germany	0.778	-2.612	3.635		1.927	-1.805	8.687		-4.534	-7.096	-1.843	*	0.736	-0.323	4.328	
$1 - \hat{\pi}$	0.249	0.089	0.336	*	0.128	0.082	0.205	*	0.181	0.127	0.240	*	0.454	0.276	0.711	*
N		1813				1621				1527				1387		

Note: Cut points of the placement part and intercept of the uncertainty part are not displayed.

5.3 Model Comparisons

Finally, we demonstrate that the mixture model outperforms the traditional cumulative model based on performance measures. In Table 3, we compare the performances of the proposed BetaBin model with the cumulative model without uncertainty component. We measure model performance by the Log-Likelihood and the AIC. The latter is defined by

$$AIC = -2l(\hat{\boldsymbol{\theta}}) + 2m,$$

where m is the number of model parameters and $l(\hat{\boldsymbol{\theta}})$ is the log-likelihood function computed at the maximum of the estimated parameter vector $\hat{\boldsymbol{\theta}}$. We see that the mixture model improves all considered performance measures as compared to the cumulative model without any uncertainty component. All AIC values are lower for the mixture than for the pure cumulative model. While the pure cumulative model is based on 18 parameters (10 intercepts and 8 covariables), the mixture model is based on a total of 28 parameters: 18 parameters for the placement part, which is identical with the pure cumulative model, 9 parameters for modeling the shape of the uncertainty distribution (1 intercept and 8 covariables) and 1 parameter to estimate the mixture weight π. Even though the mixture model is much more complex, the performance measures indicate that it yields not only a better likelihood but also to a better model fit measured by AIC. Since the number of observations differs among parties, the values can only be compared across the different models but not across parties.

Table 3: Model comparisons based on performance measures

	N	LogL		AIC	
		Mixture	Cumulative	Mixture	Cumulative
Model					
CDU	1949	-4314.096	-4370.631	8684.191	8777.262
CSU	1872	-3926.739	-3941.070	7909.478	7918.140
AfD	1715	-1706.051	-1723.742	3468.102	3483.484
SPD	1813	-3620.930	-3637.966	7297.860	7311.933
Greens	1621	-3153.747	-3171.438	6363.494	6378.875
Left	1527	-3299.113	-3326.538	6654.227	6689.075
FDP	1387	-2652.197	-2675.003	5360.394	5386.006

6 Discussion and Outlook

Political perceptions play an important role in the decision-making process. In this paper, we have applied a special mixture model, the so-called BetaBin model, to the perception of party placements on ordinal policy scales. The model consists of two components, a

placement part and an uncertainty part. The latter enables us to model response styles to the middle categories as well as to extreme categories. For the placement part, a cumulative model is used, and for the uncertainty part, we rely on a restricted beta-binomial distribution. We applied the model to the immigration issue in the occasion of the 2017 German national election. Our results demonstrate that the respondents' perceptions of most parties on the immigration issue are strongly influenced by uncertainty. We detect the lowest uncertainty in locating the AfD and the highest in placing the FDP and the CDU. Regarding the predictors we examined, we find that particularly political sophistication and candidate images influence both where to place the party and the uncertainty response patterns. Especially interesting are also the age effects we detect. The older the respondents, the more they tend to locate the parties toward the contra-immigration pole. Finally, our model outperforms traditional cumulative models without uncertainty structure based on model performance.

Next steps will be to apply the model to other policy scales and contexts. Another interesting aspect would be to examine all parties simultaneously in a multivariate model. In the proposed model, we use covariates to model the uncertainty structure and the placement part, but not to model the mixture weights π. There are other approaches (e.g., Tutz et al., 2017) which use covariates in the placement part and the mixture weights, but not in the uncertainty part. Including covariates in all three components of the mixture model may lead to identifiability issues which have not been discussed yet. In future research, we also intend to develop a voter choice model that relies on the model of survey responses we proposed here. A voter choice model that is based on the party placement and uncertainty estimates of the mixture model will then add to our understanding of how uncertainty impacts on policy-oriented decision making and its electoral consequences. Then it will also be possible to examine and test for behavioral implications of uncertainty in political perceptions.

References

Abou-Chadi, Tarik. 2015. Niche party success and mainstream party policy shifts – how green and radical right parties differ in their impact. *British Journal of Political Science* 46 (2): 417–436.

Aldrich, John H., Richard G. Niemi, George Rabinowitz, and David W. Rohde. 1982. The measurement of public opinion about public policy: A report on some new issue question formats. *American Journal of Political Science* 26 (2): 391–414.

Alonso, Sonia and Sara Claro Da Fonseca. 2012. Immigration, left and right. *Party Politics* 18 (6): 865–884.

Alvarez, R. Michael. 1999. *Information and Elections*. Ann Arbor: University of Michigan Press.

Alvarez, R. Michael and John Brehm. 1995. American ambivalence towards abortion policy: Development of a heteroskedastic probit model of competing values. *American Journal of Political Science* 39 (4): 1055–1082.

Alvarez, R. Michael and John Brehm. 1997. Are Americans ambivalent towards racial policies? *American Journal of Political Science* 41 (2): 345–374.

Alvarez, R. Michael and John Brehm. 1998. Speaking in two voices: American equivocation about the internal revenue service. *American Journal of Political Science* 42 (2): 418–452.

Alvarez, R. Michael and John Brehm. 2002. *Hard Choices, Easy Answers: Values, Information, and American Public Opinion*. Princeton: Princeton University Press.

Alvarez, R. Michael and Charles H. Franklin. 1994. Uncertainty and political perceptions. *The Journal of Politics* 56 (3): 671–688.

Art, David. 2011. *Inside the Radical Right: The Development of Anti-Immigrant Parties in Western Europe*. New York: Cambridge University Press.

Bale, Tim. 2003. Cinderella and her ugly sisters: The mainstream and extreme right in Europe's bipolarising party systems. *West European Politics* 26 (3): 67–90.

Bale, Tim. 2008. Turning round the telescope. Centre-right parties and immigration and integration policy in Europe. *Journal of European Public Policy* 15 (3): 315–330.

Bale, Tim, Christoffer Green-Pedersen, André Krouwel, Kurt Richard Luther, and Nick Sitter. 2010. If you can't beat them, join them? Explaining social democratic responses to the challenge from the populist radical right in Western Europe. *Political Studies* 58 (3): 410–426.

Bartels, Larry M. 1986. Issue voting under uncertainty: An empirical test. *American Journal of Political Science* 30 (4): 709–728.

Baumgartner, Hans and Jan-Benedict Steenkamp. 2001. Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research* 38 (2): 143–156.

Berinsky, Adam J. and Jeffrey B. Lewis. 2007. An estimate of risk aversion in the U.S. electorate. *Quarterly Journal of Political Science* 2 (2): 139–154.

Campbell, Angus, Philip E. Converse, E. Warren Miller, and Donald E. Stokes. 1960. *The American Voter*. Chicago: University of Chicago Press.

Davis, Otto A., Melvin J. Hinich, and Peter C. Ordeshook. 1970. An expository development of a mathematical model of the electoral process. *The American Political Science Review* 64 (2): 426–448.

De Vries, Catherine and Marco R. Steenbergen. 2013. Variable opinions: The predictability of support for unification in European mass publics. *Journal of Political Marketing* 12 (1): 121–141.

D'Elia, Angela and Domenico Piccolo. 2005. A mixture model for preference data analysis. *Computational Statistics & Data Analysis* 49 (3): 917–934.

Delli Carpini, Michael X. and Scott Keeter. 1993. Measuring political knowledge: Putting first things first. *American Journal of Political Science* 37 (4): 1179–1206.

Downs, Anthony. 1957. *An Economic Theory of Democracy*. New York: Harper & Row.

Enelow, James M. and Melvin J. Hinich. 1981. A new approach to voter uncertainty in the Downsian spatial model. *American Journal of Political Science* 25 (3): 483–493.

Gill, Jeff. 2005. An entropy measure of uncertainty in vote choice. *Electoral Studies* 24 (3): 371–392.

Givens, Terri E. 2005. *Voting Radical Right in Western Europe*. New York: Cambridge University Press.

Harbers, Imke, Catherine E. De Vries, and Marco R. Steenbergen. 2013. Attitude variability among Latin American publics: How party system structuration affects left/right ideology. *Comparative Political Studies* 46 (8): 947–967.

Harvey, A. C. 1976. Estimating regression models with multiplicative heteroscedasticity. *Econometrica* 44 (3): 461–465.

Iannario, Maria and Domenico Piccolo. 2016a. A comprehensive framework of regression models for ordinal data. *Metron* 74 (2): 233–252.

Iannario, Maria and Domenico Piccolo. 2016b. A generalized framework for modelling ordinal data. *Statistical Methods & Applications* 25 (2): 163–189.

Ignazi, Piero. 2003. *Extreme Right Parties in Western Europe*. Oxford: Oxford University Press.

Inglehart, Ronald. 1997. *Modernization and Postmodernization: Cultural, Economic, and Political Change in 43 Societies*. Princeton: Princeton University Press.

Ivarsflaten, Elisabeth. 2008. What unites right-wing populists in Western Europe? Re-examining grievance mobilization models in seven successful cases. *Comparative Political Studies* 41 (1): 3–23.

Kedar, Orit. 2005a. How diffusion of power in parliaments affects voter choice. *Political Analysis* 13 (4): 410–429.

Kedar, Orit. 2005b. When moderate voters prefer extreme parties: Policy balancing in parliamentary elections. *The American Political Science Review* 99 (2): 185–199.

Kitschelt, Herbert and Anthony McGann. 1995. *The Radical Right in Western Europe: A Comparative Analysis*. Ann Arbor: University of Michigan Press.

Luskin, Robert C. 1987. Measuring political sophistication. *American Journal of Political Science* 31 (4): 856–899.

Luskin, Robert C. 1990. Explaining political sophistication. *Political Behavior* 12 (4): 331–361.

Mauerer, Ingrid. 2016. A party-varying model of issue voting. A cross-national study (doctoral dissertation). *University of Munich (LMU), Germany* .

Mauerer, Ingrid, Paul W. Thurner, and Marc Debus. 2015. Under which conditions do parties attract voters' reactions to issues? Party-varying issue voting in German elections 1987-2009. *West European Politics* 38 (6): 1251–1273.

McLachlan, Geoffrey J. and David Peel. 2000. *Finite Mixture Models*. New York: Wiley.

Meguid, Bonnie M. 2005. Competition between unequals: The role of mainstream party strategy in niche party success. *The American Political Science Review* 99 (3): 347–359.

Mudde, Cas. 2007. *Populist Radical Right Parties in Europe*. Cambridge: Cambridge University Press.

Norris, Pippa. 2005. *Radical Right: Voters and Parties in the Electoral Market*. New York: Cambridge University Press.

Pardos-Prado, Sergi, Bram Lancee, and Iñaki Sagarzazu. 2014. Immigration and electoral change in mainstream political space. *Political Behavior* 36 (4): 847–875.

Rapeli, Lauri. 2013. *The Conception of Citizen Knowledge in Democratic Theory*. Basingstoke: Palgrave Macmillan.

Rossteutscher, Sigrid, Harald Schoen, Rüdiger Schmitt-Beck, Bernhard Wessels, Christof Wolf, Ina Bieber, Lars-Christopher Stövsand, and Melanie Dietz. 2017. Pre-election cross section (GLES 2017). ZA6800 Data file Version 3.0 Cologne: GESIS Data Archive.

Rozenas, Arturas. 2013. Inferring ideological ambiguity from survey data. In *Advances in Political Economy: Institutions, Modelling and Empirical Analysis*, eds. Norman Schofield, Gonzalo Caballero, and Daniel Kselman, 369–382. Berlin, Heidelberg: Springer.

Shepsle, Kenneth A. 1972. The strategy of ambiguity: Uncertainty and electoral competition. *The American Political Science Review* 66 (2): 555–568.

Stoetzer, Lukas F. 2017. A matter of representation: Spatial voting and inconsistent policy preferences. *British Journal of Political Science* Advance online publication.

Tomz, Michael and Robert P. Van Houweling. 2009. The electoral implications of candidate ambiguity. *American Political Science Review* 103 (1): 83–98.

Tutz, Gerhard. 2012. *Regression for Categorical Data*. Cambridge: Cambridge University Press.

Tutz, Gerhard and Micha Schneider. 2017. Mixture models for ordinal responses with a flexible uncertainty component. Technical Report 203, Department of Statistics, University of Munich (LMU), Germany.

Tutz, Gerhard, Micha Schneider, Maria Iannario, and Domenico Piccolo. 2017. Mixture models for ordinal responses to account for uncertainty of choice. *Advances in Data Analysis and Classification* 11 (2): 281–305.

Vaerenbergh, Yves Van and Troy D. Thomas. 2013. Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research* 25 (2): 195–217.

A.4. Uncertainty in Issue Placements and Spatial Voting

Mauerer, I. and M. Schneider (2019b): Uncertainty in Issue Placements and Spatial Voting. *Technical Report 226*, Department of Statistics, Ludwig-Maximilians-Universität München, doi:10.5282/ubm/epub.68451.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Ingrid Mauerer and Micha Schneider

Uncertainty in Issue Placements and Spatial Voting

Technical Report Number 226, 2019
Department of Statistics
University of Munich

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

Uncertainty in Issue Placements and Spatial Voting

August 6, 2019

Ingrid Mauerer
Department of Political Science, LMU Munich
Ingrid.Mauerer@gsi.uni-muenchen.de

Micha Schneider
Department of Statistics, LMU Munich
Micha.Schneider@stat.uni-muenchen.de

Abstract

Empirical applications of spatial voting approaches frequently rely on ordinal policy scales to measure the policy preferences of voters and their perceptions about party or candidate platforms. Even though it is well known that these placements are affected by uncertainty, only a few empirical voter choice models incorporate uncertainty into the choice rule. In this manuscript, we develop a two-stage approach to further the understanding of how uncertainty impacts on spatial issue voting. First, we model survey responses to ordinal policy scales where specific response styles capture the uncertainty structure in issue placements. At the second stage, we model voter choice and use the placements adjusted for the detected uncertainty as predictors in calculating spatial proximity. We apply the approach to the 2016 US presidential election and study voter preferences and perceptions of the two major candidate platforms on the traditional liberal-conservative scale and three specific issues. Our approach gives insights into how voters attribute issue positions and spatial voting behavior, and performs better than a voter choice model without accounting for uncertainty measured by AIC.

Keywords: Ordinal Policy Scales, Issue Placements, Uncertainty, Spatial Voting, Mixture Models

1 Introduction

Spatial voting approaches assume that citizens elect parties or candidates that offer policy platforms that coincide with their preferences. Empirical applications frequently rely on ordinal policy scales to determine the citizens' policy preferences and their perceptions about party platforms. This practice presupposes that (1) voters have well-defined individual preferences about public policy issues, and (2) parties take certain policy positions and voters perceive these platforms. It is well recognized that uncertainty influences these placements. Within spatial voting approaches, uncertainty is mainly considered as one that stems from candidate or party platforms (Shepsle 1972; Enelow and Hinich 1981). The literature argues that ambiguous or vague position taking (or campaigning) and limitation in voter information cause uncertainty in the positions a candidate or party represents and therefore in the decision of voters. Recently, it has been reasoned that voters might not be equipped with consistent and well-structured policy preferences as well (Stoetzer 2017). As a consequence, uncertainty seems to play a central role in both perceptions of party platforms and voters' policy preferences. However, there are only a few neo-Downsian empirical models that incorporate voter uncertainty into the choice rule (Bartels 1986; Gill 2005; Berinsky and Lewis 2007).

The purpose of this paper is twofold. First, we aim to understand what drives survey response variability in political perceptions and policy preferences: how are the perceptions of party platforms and policy preferences of voters structured and what role does uncertainty play in these perceptions and placements? Second, we want to further the understanding of how uncertainty impacts on spatial issue voting behavior. We develop an approach that allows studying the behavioral implications of uncertainty in political perceptions and policy preferences and its consequences for political representation. The approach consists of two analysis steps. First, we model survey responses to ordinal policy scales where specific response styles capture the uncertainty structure in issue placements. We use the so-called BetaBin model (Tutz and Schneider 2019; Mauerer and Schneider 2019), which belongs to the class of mixture models for ordinal responses. The model permits accounting for both the placement and uncertainty structure of survey responses, which can be modeled by covariates. In addition, it allows modeling specific response

patterns, such as the tendency to select the middle category (see, e.g., Aldrich et al. 1982; Alvarez and Franklin 1994) or the tendency to choose extreme categories (Baumgartner and Steenkamp 2001; Vaerenbergh and Thomas 2013). At this first stage, we determine the positions on the policy scales accounting for uncertainty. At the second stage, we model voter choices and use the adjusted placement values estimated by the mixture model. This procedure allows us to improve the vote choice model by accounting for individual uncertainty in issue placements.

The empirical application uses survey data from the 2016 US presidential election and examines how voters' perceptions of candidate platforms are structured on the traditional liberal-conservative dimension and specific policy issues. The results indicate that voter show much less uncertainty in placing themselves than in attributing positions to the candidates. Our findings also suggest, for instance, that voter who identify themselves with the Democratic or Republican party, respectively, tend to push their self-placements toward the perceived candidate platforms. Furthermore, our approach improves model performance measures at all stages.

2 Uncertainty in Policy Preferences and Platforms

Survey responses to ordinal policy scales are frequently used to measure the policy preferences of the electorate and perceptions of party or candidate policy platforms. In public opinion research, several studies assess variability in policy preferences and examine competing explanations based on uncertainty, ambivalence or equivocation. Some studies explore specific attitudes towards, for instance, abortion, racial policies or European integration (Alvarez and Brehm 1995, 1997, 1998, 2002; De Vries and Steenbergen 2013), others explore variability in Left-Right placements (Harbers, De Vries and Steenbergen 2013).

However, there is also work that argues that relying on individual placements and perceptions of party locations might cause difficulties due to interpersonal incomparability of survey responses or rationalization processes. The first difficulty arises when respondents have a subjective understanding of issue scales, the so-called differential-item functioning (Brady 1985), which distorts the placements. Starting with the Aldrich-McKelvey sca-

ling method (Aldrich and McKelvey 1977), a considerable amount of research proposes statistical procedures to correct for the interpersonal incomparability of survey responses to issue scales (see, e.g., Hare et al. 2015; Poole et al. 2016; Poole 1998). Based on issue scale data, these approaches provide estimates for self-placements and party locations to construct common underlying latent policy dimensions.

The second difficulty stems from rationalization processes that induce distortions in attributing issue positions to parties or candidates. Markus and Converse (1979) already introduced the concepts of persuasion and projection. Persuasion means that voters are persuaded by the parties or candidates so that they change their positions to bring them closer to the position of favored parties. Projection means that voters project their own positions onto parties they favor, i.e., a tendency to adjust the policy location of parties they prefer. Drawing from balance theory (Heider 1946, 1958) and the social judgment-involvement approach (Sherif and Hovland 1961), two types of projection effects can be distinguished: assimilation and contrast (see, e.g., Merrill III, Grofman and Adams 2001; Merrill III and Grofman 1999; Conover and Feldman 1982, 1981; Granberg and Brown 1992; Granberg and Brent 1980; Granberg and Jenks 1977; Granberg 1987; Feldman and Conover 1983). The first effect is based on the argument that respondents assimilate the stances of parties they prefer by reducing the perceived distance between their policy preferences and the party they favor to move them closer to their own preferences. The latter refers to the effect that respondents tend to contrast the positions of parties they dislike, i.e., respondents project parties they dislike away by exaggerating the ideological distance to those. To evaluate these effects, Merrill III, Grofman and Adams (2001), for instance, divide the respondents into two groups, supporters and non-supporters of a particular party. Then, they relate the self-placement to the median candidate placement, separately for the two groups at the population level. Their results indicate that in many cases, the group of supporters behaves differently than non-supporters in placing the candidates. For instance, the more conservative the supporters place themselves on average, the higher the median placement of the supported candidate.

The existing literature offers a few approaches to measure and model variability and uncertainty in issue placements. One approach is to directly measure uncertainty by

asking respondents to report how certain they are about party or candidate platforms (e.g., Alvarez and Franklin 1994), or to adjust the 7-point or 11-point policy scales by range formats (see, e.g., Tomz and Van Houweling 2009; Aldrich et al. 1982; Alvarez 1999).

Another way to measure uncertainty is to rely on indirect methods. Harvey (1976) introduced the heteroscedastic regression framework, which models the variance of the disturbance by predictors and is applied, for instance, in Harbers, De Vries and Steenbergen (2013); De Vries and Steenbergen (2013). Alvarez and Brehm (1995), for example, use a heteroscedastic binary probit model. Bartels (1986) infers uncertainty from patterns in missing data, based on the idea that respondents who are uncertain are not able to provide placements at all. In a two-stage procedure, he first relies on a model of survey responses where non-responses indicate uncertainty and are a function of attributes of the candidate, the voter, and the political setting. In the second stage, the estimated probabilities of non-response are used to examine the impact of uncertainty on voting behavior, applying a linear probability model in both analysis steps. Aldrich et al. (2018) follow a similar approach. First, they estimate the probability of not placing themselves or at least two parties on ordinal scales. Then, they use these probabilities as well as other covariates to evaluate the variability in the difference between the individual-specific party placement and the sample mean party placement.

Campbell (1983*b*,*a*) also uses an indirect measure by using sample standard deviations of placements. Gill (2005) connects uncertainty with the entropy concept. He develops an approach that provides an aggregate measure of uncertainty by relying on aggregated responses and information on candidate characteristics, the issue to be assessed, and the respective survey questions. His uncertainty term is more flexible than the one by Bartels (1986)'s by allowing it to vary across candidates and issues, but it still assumes homogenous uncertainty across voters. Rozenas (2013) offers an approach that integrates variance heterogeneity (Harvey 1976) and non-response (Bartels 1986), resulting in a quite difficult model with hyper parameters for whom appropriate prior distributions need to be selected.

The handling of missing values also plays a central role in the study of uncertainty.

Current approaches treat missing values in diverse ways. Some studies use observed values only and do not rely on any missing data, such as applications of the pure heteroscedastic models (e.g., Harvey 1976; Alvarez and Brehm 1995). Others reason that uncertainty induces missing data in the survey responses (e.g., Bartels 1986; Rozenas 2013). We believe that the crucial issue here is whether a particular underlying mechanism generates missing data. Missing values in the response structure might reflect uncertainty, but also other processes might cause missing data. For instance, respondents might show clear preferences and political perceptions but refuse to report them due to social desirability. A lack of motivation or time might also cause that respondents do not provide placements. In such cases, missing values would embody both certain and uncertain placements. Our survey response model does not include any missing data (including 'don't know' replies). Usually, we do not know the true missing-data generating process. Therefore, we assume that missing data in survey responses to policy scales is not directly linked to uncertainty.

The model of survey responses we develop in this paper, which then forms the basis for the voter choice model, differs from existing approaches in the following aspects. First, our approach does not require additional survey questions in which respondents state how uncertain they are about policy platforms (Alvarez and Franklin 1994) nor does it adjust the original 7-point or 11-point policy scales (e.g., Tomz and Van Houweling 2009; Aldrich et al. 1982; Alvarez 1999). Second, the model explicitly takes into account the ordinal nature of policy scales, which is in contrast to previous studies that use the linear regression model (e.g., Harbers, De Vries and Steenbergen 2013; De Vries and Steenbergen 2013) or binary outcome-models based on logit/probit link functions (e.g., Alvarez and Brehm 1995). Especially when dealing with limited ordinal policy scales, it is not clear whether the distance between each category is equal, which is assumed in the linear regression framework. In addition, the error terms might be not normally distributed, and the linear regression might predict values lower, in between or above the limited ordinal response scale. Third, the model can handle three specific response styles: a random choice, a tendency to moderate, and a tendency to extreme placements on the policy scales. This allows detecting particular uncertainty structures that can be modeled by explanatory variables, in contrast to models such as the heteroscedastic regression

model (e.g., Harvey 1976; Alvarez and Brehm 1995) where additional scale parameters are used to model only low or high variance, and therefore rather unstructured variability.

3 Modeling Issue Placements and Spatial Voting under Uncertainty

Our approach proceeds in two steps. In the first stage, we develop a model of survey responses for ordinal policy scales. Here, we estimate the positions on the policy scales corrected by uncertainty. In the second stage, we specify a voter choice model that is based on these adjusted values as the key predictors.

3.1 Stage 1: Survey Response Model

The model of survey responses belongs to the class of mixture models (McLachlan and Peel 2000; Iannario and Piccolo 2016; Piccolo and Simone 2019) which can be used to model variability in ordinal response data. As human choices or political perceptions can be understood as a combination of placement and uncertainty, we rely on a mixture model with two components

$$f = \sum_{g=1}^{2} \pi_g f_g, \tag{1}$$

where the mixture proportion or weight π_g can take values between 0 and 1, and $\sum_{g=1}^{2} \pi_g = 1$. The density f can be described by the combination of f_1 and f_2. We only consider density functions that are in accord with the nature of ordinal data. Examples for the placement component are the cumulative logit model or the adjacent categories model (Tutz et al. 2017). The uncertainty component allows taking into account specific response styles. D'Elia and Piccolo (2005) or Tutz et al. (2017), for instance, rely on the uniform distribution, which reflects a random choice of the response category. We use the BetaBin model (Tutz and Schneider 2019) that enables us to model both the response styles and the placement structure in a flexible way. In contrast to other possible approaches, this model can handle response styles to the middle as well as to extreme categories to model uncertainty in policy placements.

The mixture model BetaBin assumes that we observe the response of an individual i to an ordinal policy scale, denoted by R_i. Let Y_i be the unobserved random variable that gives the placement on the ordinal policy scale. U_i is the unobserved uncertainty component which models the type of response style. All these variables take the ordered values $\{1, \ldots, k\}$. The mixture model BetaBin has the form

$$P(R_i = r|\boldsymbol{x}_i, \boldsymbol{w}_i) = \pi_i P_M(Y_i = r|\boldsymbol{x}_i) + (1 - \pi_i) P_U(U_i = r|\boldsymbol{w}_i), \tag{2}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ are vectors of explanatory variables. Both the placement and the uncertainty part can be modeled by the same, overlapping or entirely distinct covariates. π_i is the mixture probability that indicates the weight of the structural component in the mixture. Consequently, $1 - \pi_i$ represents the strength of the uncertainty component. As a result, the observed response R_i stems from a discrete mixture of the uncertainty and the placement part.

Any ordinal model can be used for the placement part $P_M(Y_i = r|\boldsymbol{x}_i)$ of the model. We rely on the cumulative logit model (aka ordered/ordinal logit model, proportional odds model) (see Tutz 2012):

$$\log\left(\frac{P(Y_i \leq r|\boldsymbol{x}_i)}{P(Y_i > r|\boldsymbol{x}_i)}\right) = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}, \quad \text{or}$$
$$P(Y_i \leq r) = \frac{\exp(\gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma})}{1 + \exp(\gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma})}, \quad r = 1, \ldots, k-1.$$

γ_{0r} denote the thresholds or intercepts and $\boldsymbol{\gamma}$ the estimated effects that do not depend on r. In our notation, positive values increase the ratio $\log\left(\frac{P(Y_i \leq r|\boldsymbol{x}_i)}{P(Y_i > r|\boldsymbol{x}_i)}\right)$ and connote that lower categories are more likely than higher ones. Regarding the uncertainty part $P_U(U_i = r|\boldsymbol{w}_i)$, the model assumes that the random variable U follows a Beta-Binomial distribution: $U \sim \textit{Beta-Binomial}(k|\alpha, \beta)$

$$f(u) = \begin{cases} \binom{k-1}{u-1} \frac{B(\alpha+u-1, \beta+k-u+1)}{B(\alpha,\beta)} & u \in \{1, \ldots, k\} \\ 0 & \text{otherwise.} \end{cases}$$

$\alpha, \beta > 0$ are the parameters of the distribution, and $B(\alpha, \beta)$ gives the beta function:

$$B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt.$$

By assuming that $\mu = \alpha/(\alpha + \beta)$ and $\delta = 1/(\alpha + \beta + 1)$[1], the expected value $E(U)$ and the variance $var(U)$ are

$$E(U) = (k-1)\mu + 1, \quad \text{var}(U) = (k-1)\mu(1-\mu)[1 + (k-2)\delta].$$

The beta-binomial distribution converges to the (shifted) binomial distribution $B(k, \mu)$ with mean μ and categories $\{1, \ldots, k\}$ when δ approaches 0. We aim to model two response styles: a tendency to middle or extreme categories. This is achieved by setting $\alpha = \beta$ so that $\mu = 0.5$ and $\delta = 1/(2\alpha + 1)$. As a result, μ, which gives the location of the distribution, is set at the middle of the policy scale. α and δ are not fixed: smaller α values result in larger δ values, and therefore greater variance. Figure 1 depicts the restricted beta-binomial distribution for different α values. For $\alpha = 1$, one obtains the discrete

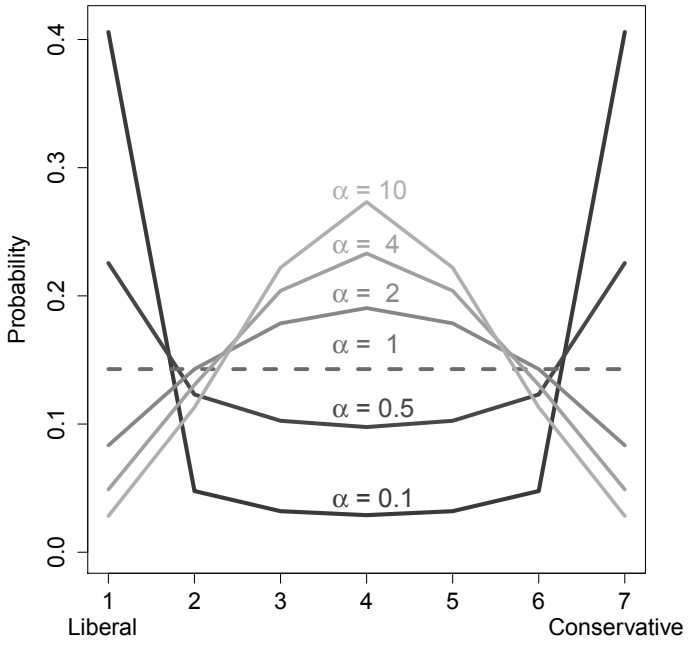


Figure 1: Probability mass on 7-point liberal-conservative scale for different α values.

uniform distribution. $\alpha > 1$ indicates a tendency to the middle categories and $\alpha < 1$ a

[1]Note that this reformulation is required because α and β do not correspond with the location and scale of the distribution.

tendency to extreme categories. Given different α values, the distribution encompasses a (shifted) binomial distribution with the mode in the middle of the scale, which reflects a strong tendency to middle categories, and one with almost equal point mass at the endpoints of the scale, which corresponds with a strong tendency to extreme categories (i.e., minimum and maximum of k). Between these two extremes, any gradations are feasible.

The coefficient α, the parameter of the restricted beta-binomial distribution, ascertains the shape of the distribution in the uncertainty component and is connected to the explanatory variables $\boldsymbol{w}_i$ by

$$\alpha = \exp(\boldsymbol{w}_i^T \boldsymbol{\alpha}) = \exp(\alpha_0) \exp(\alpha_1)^{w_{i1}} ..... \exp(\alpha_m)^{w_{im}}.$$

The parameter α_j contains the effect of the explanatory variable w_{ij}. Since the exponential function links the explanatory variables to α, the coefficient α changes by the factor $\exp(\alpha_j)$ for every one-unit change in w_{ij}, holding all other variables constant. The parameters indicate how a variable impacts on the tendency to middle or extreme placements: $\alpha_j > 0$ results in $\alpha > 1$ and imply a tendency to middle categories; $\alpha_j < 0$ results in $\alpha < 1$ and imply a tendency to extreme placements.

3.2 Stage 2: Voter Choice Model

Following the classical proximity model (Downs 1957; Davis, Hinich and Ordeshook 1970; Enelow and Hinich 1984), the voter choice model is decision theoretical and focuses on the impact of spatial considerations on voting. To identify each candidate's amount of utility, voters are assumed to compare candidates' policy proposals on several issues and choose the one that offers issue positions that are closest to the voters' most preferred issue positions. The model also accounts for nonpolicy factors (e.g., Adams, Merrill III and Grofman 2005; Adams and Merrill III 1999; Thurner 2000), such as voters' socioeconomic characteristics.

For voter $i \in \{1, \ldots, n\}$ and candidate or party $j \in \{1, \ldots, J\}$, define V_{ij} as a linear predictor for each candidate j that accumulates the systematic determinants of the vote choice in a scalar quantity. V_{ij} consists of voter-party proximity measures

$z_{ijk}, k \in \{1, \ldots, K\}$, that represent the proximity between voter i and party j on each issue k. The model is based on respondent-specific perceptions of party positions and applies linear utility losses in the calculation of issue distances. Let $s_{il},\ l \in \{1, , \ldots, p\}$ refer to voter characteristics. The deterministic part of utility takes the form:

$$V_{ij} = \beta_{j0} + \sum_{k=1}^{K} z_{ijk}\alpha_k + \sum_{l=1}^{p} s_{il}\beta_{jl} = \beta_{j0} + \boldsymbol{z}_{ij}^T\boldsymbol{\alpha} + \boldsymbol{s}_i^T\boldsymbol{\beta}_j. \tag{3}$$

The parameters $\beta_{10}, \ldots, \beta_{J0}$ represent alternative-specific constants (ASCs). These coefficients contain the unmeasured utility components. $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K$ is a k-dimensional vector related to the voter-party proximity measures $\boldsymbol{z}_{ij}$. $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$ is a p-dimensional coefficient vector related to voter attributes contained in the covariate vector $\boldsymbol{s}_i$. The corresponding coefficients indicate segment-specific evaluations of parties. The utility expression V_{ij} is linked to voter choice by a logit link function:

$$P(Y = j|\boldsymbol{z}_{ij}, \boldsymbol{s}_i) = \frac{\exp(\beta_{j0} + \boldsymbol{z}_{ij}^T\boldsymbol{\alpha} + \boldsymbol{s}_i^T\boldsymbol{\beta}_j)}{\sum_{r=1}^{J} \exp(\beta_{r0} + \boldsymbol{z}_{ij}^T\boldsymbol{\alpha} + \boldsymbol{s}_i^T\boldsymbol{\beta}_r)}, \tag{4}$$

where $Y \in \{1, ..., J\}$ denotes the j-categorical, probabilistic response variable.

4 Empirical Application

We apply our approach to the 2016 US presidential election and focus on the two major party candidates, the Democratic nominee Hillary Clinton and the Republican opponent Donald Trump. The empirical application examines how self-placements and political perceptions are structured on both the traditional liberal-conservative scale and three specific policy issues (Spending and Services, Defense Spending, Health Insurance).[2] The respondents were asked to state where they place themselves and perceive each of the candidates on seven-point scales. The liberal-conservative scale runs from (1) "extremely liberal" to (7) "extremely conservative". The first specific issue measures attitudes and political perceptions on public spending and services, with (1) representing "Government should provide many fewer services" and (7) "Government should provide many more

[2]Note that the 2016 American National Election Study (ANES) includes additional position issues that we do not consider. Our analysis is based on the cross-sectional pre-election survey.

services". The second scale captures attitudes on the amount of the budget spent on defense, running from (1) "Government should decrease defense spending" to (7) "Government should increase defense spending". The third taps positions on public versus private medical support (1 "Government insurance plan", 7 "Private insurance plan"). We restrict our analysis to those respondents that provided self-placements and party placements for both the Democrat and the Republican, and reported voting for one of the two major candidates.

4.1 Survey Responses to Placements

The stated positions give the observed response R_i in Equation 2. Figure 2 depicts the distribution of issue placements on the liberal-conservative scale and the three specific issue scales. All bar plots show the percentages for each category on the ordinal scales. While the distributions of the self-placements are in most cases rather unstructured, with a small tendency to middle categories (except for the issue of health insurance), the densities of the perceived candidate positions are mostly skewed. The modal value of the candidate positions is in all cases at the opposite sides, except for the issue of defense spending. Inspecting, for instance, the liberal-conservative scale reveals that the majority of the probability mass for the Republican candidate is located at high categories (5,6,7), while the majority of the probability mass for Democratic candidate is located at the opposite side (1,2,3). The tendency to the left or right side of the scales depends on the policy and the scale coding. Thus, voters perceive the Republican candidate as offering more conservative positions, favoring the increase of defense spending and private health insurance. By contrast, voters ascribe the Democratic candidate more liberal positions and perceive the candidate as favoring government health insurance. The only exception to this pattern is observed for the issue of defense spending, where the voters perceive the Democratic candidate as taking a more moderate position. The same opposite tendency is noticeable for the issue of spending and services, but in reversed corners of the scale because of the different coding. Here, the distribution of Democratic candidate placements is left-skewed corresponding with more services and higher categories, whereas the distribution of perceived stances for the Republican candidate is right-skewed corresponding

with fewer services at smaller categories.

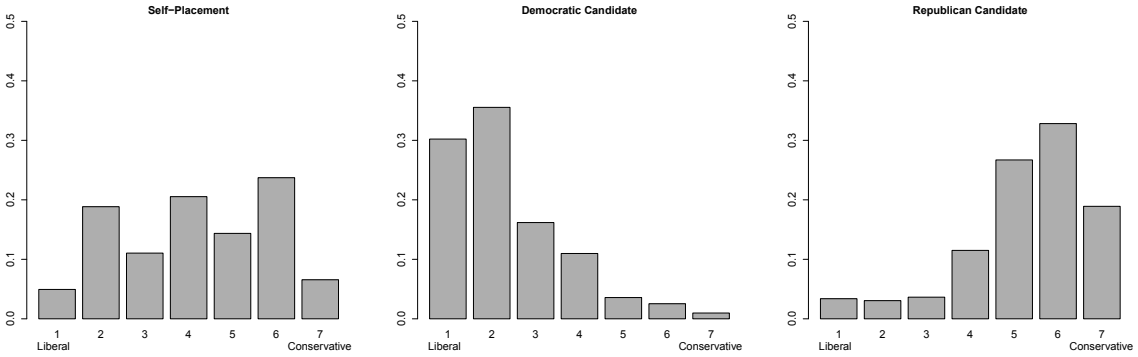
4.2 Stage 1: Predictors for Uncertainty and Placements

The model of survey responses can link both the placement and uncertainty structure of self-placements and candidate placements to explanatory variables. Note that we model the shape of the uncertainty structure by covariates but not the uncertainty weights. We examine two sets of covariates which enter both components. The first three variables relate to cognitive processes or information costs. Voters vary in political sophistication or awareness (e.g., Luskin 1987, 1990; Delli Carpini and Keeter 1993; Rapeli 2013). It is frequently reasoned in the literature that voters with lower information costs are more informed about the stances parties or candidates take on public policies. We hypothesize, therefore that voters who are equipped with higher levels of political information are more certain about their own placements and party platforms. We rely on three measures to explore whether different levels of political sophistication yield special response patterns due to uncertainty or placements in a particular direction. The first is the level of education, measured in 8 categories from 1 (high school degree or less) to 8 (doctorate). The second variable captures the strength of political interest and employs self-reports on how much attention the respondent pays to politics and elections. The original five-point scale was reversed so that 1 represents the response "never" and 5 "always". To distinguish segments with different political knowledge, we use factual knowledge questions with correct and incorrect responses. The respondents were asked to recognize the job or political office the following persons hold: Vice-President Joe Biden, Speaker of the House Paul Ryan, Chancellor of Germany Angela Merkel, President of Russia Vladimir Putin, US Supreme Ct Chief Justice John Roberts. We computed an additive knowledge score where each incorrect reply gives a value of 0 and correct answers a value of 1. We counted the number of times each respondent reported right answers yielding a six-categorical variable (0 none correct, 5 all answers correct).

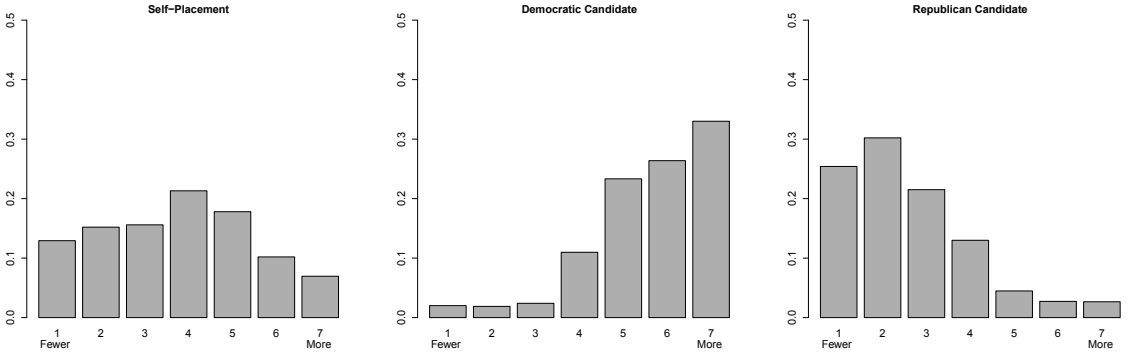
The second set of predictors are partisan variables and assessments of personal or character qualities of the candidates. We hypothesize that voters are more certain where to locate the candidates on the policy scales when they have a long-standing leaning

Figure 2: Distribution of Issue Placements

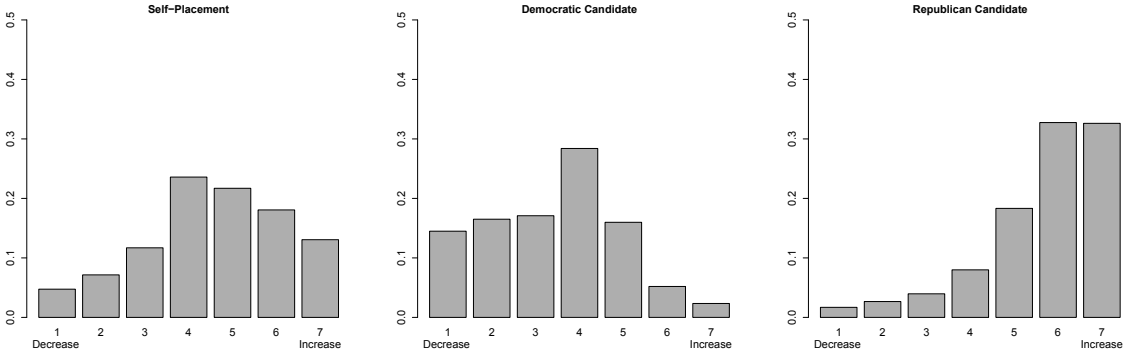
(a) Liberal-Conservative



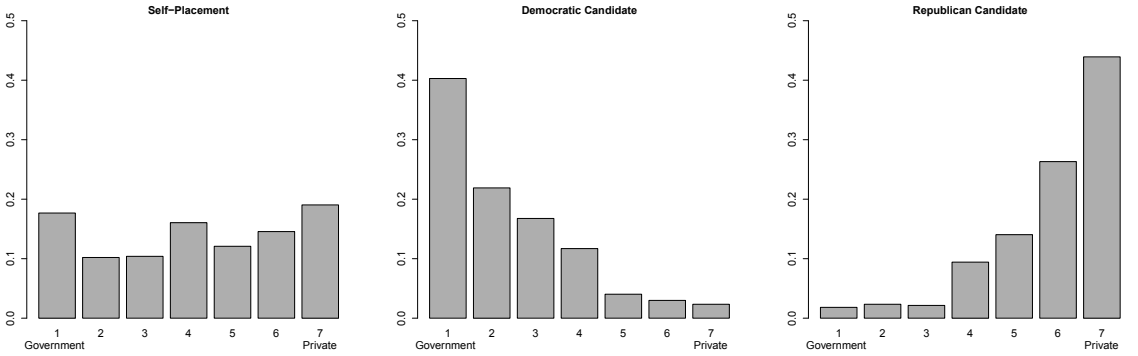
(b) Spending and Services



(c) Defense Spending



(d) Health Insurance



Source: 2016 ANES, N=1539.

toward the party whose candidate they place. By contrast, specific response styles due to uncertainty might be likely to observe when the voter does not identify with the respective party. The same expectation could also be formulated for candidate images. To capture the relationship between the voter and the candidate to be located, we consider party identification, which enters the models by two dummy-coded variables. For each of the two major parties, we generated a variable that takes the value of 1 when the respondent identifies as Democrat or Republican, respectively, and 0 otherwise (i.e., no preference, any other party identification or Independents). We also explore candidate images measured by character traits. The respondents were asked to assess the candidates on six traits (strong leadership, really cares, knowledgeable, honest, speaks mind, and even-tempered), each measured on a five-point scale running from "not well at all" to "extremely well". For each of the two candidates, an index of the overall evaluation was generated by adding all trait evaluations and dividing it by the number of traits.

4.3 Stage 2: Predictors for Vote Choice

The voter choice model is based on the placement and uncertainty estimates of the underlying survey response models. In addition to these spatial considerations, we also account for standard voter characteristics such as age (centered around the sample mean, measured in decades), gender (1 female, 0 male), regional differences (North Central, Southern and Western part of the US, with Northeast as reference), economic considerations (evaluation of the country-level economy in the past year, ranging from 1 "much worse" to 5 "much better"), two race variables (self-identifications as being Black or Latino), and the level of education. Also here we exclude all missing values.

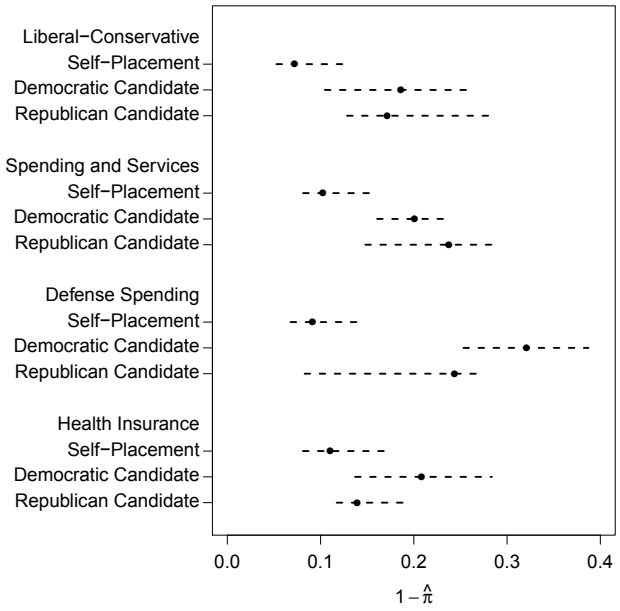
5 Results

The result presentation proceeds as follows: First, we discuss the findings of the survey response models. Here, we begin by assessing how uncertainty impacts on voters' self-placements and the perceived party platforms. Then, we present the estimates for the placement and the uncertainty part of the models, followed by a comparison of the BetaBin models with the ordinal models without uncertainty component based on performance

measures. In the second part of the analysis, we use the estimated placements to predict voter choice between the two major candidates.

5.1 Issue Placements and Issue Uncertainty

We specify for each of the placements separate BetaBin models.[3] Figure 3 illustrates the role of uncertainty in the placements. The mixture probability $\hat{\pi}$ indicates the importance of the structural component in the mixture models. Therefore, $1 - \hat{\pi}$ measures the weight of the uncertainty component and can be understood as an indicator of how certain voters are about their own placements and how clear or unambiguous they perceive the candidate platforms. The closer the weight $\hat{\pi}$ is to 1, the weaker the uncertainty so that $1 - \hat{\pi} = 0$ yields the pure cumulative model without any response styles. The closer the weight $\hat{\pi}$ is to 0, the stronger the uncertainty and the weaker the placement structure. The shape of the uncertainty structure is modeled by covariates, which may lead to a tendency to the middle categories, extreme categories or any graduation between these two extremes. The dotted lines in Figure 3 represent the 2.5% and 97.5% bootstrap quantiles.[4]



Note: Dotted lines correspond to the 2.5% and 97.5% bootstrap quantiles.

Figure 3: Importance of the uncertainty component $(1 - \hat{\pi})$

[3]Details on the EM-Algorithm to estimate the models can be found in Tutz and Schneider (2019). We also used the R packages MRSP (Pößnecker 2019) and VGAM (Yee 2016).

[4]Note that it is not unusual when the intervals are non-symmetric around the estimate because no distribution assumption is made.

We obtain the weakest uncertainty weights for the self-placements, ranging from 0.07 to 0.11. This result indicates that respondents exhibit clear positions on the ideological and policy scales. Much higher uncertainty levels are detected for the placements of the two presidential candidates. Voters are the most certain where to place both candidates on the liberal-conservative scale. We also observe comparatively moderate uncertainty weights for the Republican candidate on the issue of health insurance. A little higher uncertainty weights are estimated for the issue of spending and services. Regarding the issue of defense spending, the quite large weight for the Democratic candidate (0.32) suggests that voters exhibit much more difficulties in placing Hillary Clinton as compared to Donald Trump (0.24).

Tables 1 - 3 display the results of the mixture models. In each of the tables, the upper part gives the estimates for the placement component ($\boldsymbol{\gamma}$), the lower part the estimates for uncertainty response style effects ($\boldsymbol{\alpha}$). We also report the 2.5% and 97.5% quantiles of 300 non-parametric bootstrap samples for each effect. We consider an estimate as significant at the 5%-level when the bootstrap confidence intervals cover the estimate but not zero. The interpretation is as follows: Positive coefficients in the preference part suggest that lower categories are more likely, negative coefficients that higher categories are more likely. Regarding the uncertainty part, which contains the estimates for the shape of the uncertainty distribution, positive values suggest a tendency to locations at middle categories, that is, moderate positions. Negative values indicate a tendency to the extremes of the scales.

Let us first focus on the results for the self-placements, displayed in Table 1. Regarding the placement effects, we observe positive effects for education on self-placements on the liberal-conservative scale and the defense spending scale. The higher the level of education, the more the voters position themselves toward liberal stances or favor the decrease of defense spending. Political interest exhibits an effect on the liberal-conservative scale and political knowledge on spending and services. The negative coefficient for political interest suggests that the more attention the voters pay to politics and elections, the more they tend to have conservative views. The positive effect for political knowledge indicates that the higher the level of political knowledge, the more they favor fewer services.

Regarding the partisan variables, we obtain very plausible results. Here, one should keep in mind that the placements perceived as 'democratic' correspond with lower categories on the liberal-conservative scale and health insurance, and with higher categories on the issue of spending and services (see Figure 2). The perceived Republican positions are associated with higher categories on the liberal-conservative scale, defense spending and health insurance, and with lower categories on spending and services. Thus, the effects of party identification and candidate traits correspond with these tendencies. In particular, we observe in three out of the four cases a negative effect for Republican identifiers, with a particularly large one on the liberal-conservative scale, which is consistent with the perceived candidate position that is located at higher categories. Nevertheless, the positive effect on spending and services is also in line with this interpretation since it is the only scale where the perceived Republican position corresponds with lower categories. Accordingly, Democrats have more liberal stances and favor more services.

Concerning the Democratic candidate traits, we obtain significant effects on all scales: the higher the assessment of the Democratic candidate, the more liberal is the self-placement, the more they favor the increase of services, the decrease of defense spending and government health insurance. The Republican candidate traits also significantly impact on all self-placements, with reversed effects: the higher voters assess the qualities of the Republican candidate, the more conservative the attitudes, the more they favor fewer services, the increase in defense spending and private insurance.

As can be seen at the bottom of Table 1, we obtain only four significant response style effects for uncertainty and small uncertainty weights. When the voters' political interest increases, they tend to favor extreme positions, which is only statistically significant on the liberal-conservative and the spending and services scales. The positive effects for political knowledge indicate a tendency to favor a moderate position, which is only statistically significant for spending and services. Since the uncertainty weights are small, the estimated effects of the uncertainty part should be interpreted with caution.

Table 2 contains the placement and uncertainty estimates for the perceived Democratic candidate platforms. Voters with a higher level of political interest tend to ascribe the Democratic candidate more liberal positions and in favor of increasing spending and ser-

vices. Those with higher political knowledge scores locate the candidate towards offering a stance in favor of government health insurance. The more liberal the voters, the more liberal positions they ascribe to the candidate. The positive coefficient for the candidate traits suggest that the higher the voters assess the qualities of the candidate, the more they perceive the Democratic candidate as taking a fewer-services stance.

The estimates for the uncertainty response styles suggest that the own positions have a strong impact on the shape of the uncertainty component. The strongest impact is found on the liberal-conservative scale. The more conservative the voters, the more they tend to locate the Democratic candidate towards extreme stances on the liberal-conservative, defense spending and health insurance scales. One may interpret this behavior as 'contrast'. However, extreme stances include both sides of the scale. The positive effect indicates that voters with more conservative views tend to ascribe the candidate a moderate position on spending and services. The same tendency is estimated for Democratic party identifiers and the assessment of candidate traits for the issue of services and spending and the issue of health insurance.

Table 3 reports the results for the Republican candidate placements. Here, most of the estimated coefficients are not in line with the tendencies detected for the candidate positions in Figure 2. For instance, voters who identify with the Republican party and assign the Republican candidate higher quality traits, tend to place the candidate toward positions that correspond with more spending and services. Thus, the estimates differ from the perceived candidate tendency. Regarding the uncertainty estimates, only one significant effect is obtained: The higher the voters assess the candidate traits, the more they tend to perceive the candidate as taking moderate positions on the liberal-conservative scale.

Table 1: Parameter estimates of the BetaBin model: Self-Placements

	Liberal-Conservative				Spending and Services				Defense Spending				Health Insurance			
	Estimate	BS.2.5	BS.97.5	sig.	Estimate	BS.2.5	BS.97.5	sig.	Estimate	BS.2.5	BS.97.5	sig.	Estimate	BS.2.5	BS.97.5	sig.
Placement Part																
Education	0.084	0.035	0.144	*	0.009	-0.046	0.055		0.139	0.086	0.195	*	-0.042	-0.111	0.011	
Political Interest	-0.136	-0.266	-0.014	*	-0.001	-0.121	0.122		-0.188	-0.309	-0.080		-0.021	-0.147	0.083	
Political Knowledge	0.003	-0.117	0.116		0.138	0.043	0.244	*	0.016	-0.083	0.151		0.071	-0.023	0.211	
Democratic Party Identification	0.979	0.712	1.404	*	-0.769	-1.143	-0.467	*	0.000	-0.312	0.327		0.048	-0.306	0.409	
Republican Party Identification	-1.789	-2.233	-1.351	*	0.477	0.177	0.848	*	-0.388	-0.700	-0.120	*	-0.861	-1.213	-0.513	
Democratic Candidate Traits	1.073	0.954	1.279	*	-1.043	-1.268	-0.926	*	0.633	0.467	0.805	*	0.929	0.779	1.208	*
Republican Candidate Traits	-0.967	-1.212	-0.794	*	0.665	0.448	0.908	*	-0.938	-1.169	-0.728	*	-0.824	-1.085	-0.637	*
Uncertainty Part																
Education	0.570	-0.400	1.735		0.486	-0.373	1.415		0.239	-0.656	0.933		0.559	-0.627	1.614	
Political Interest	-3.039	-4.501	-0.213	*	-2.494	-5.367	-0.611	*	-0.940	-4.651	0.231		-0.452	-3.998	0.868	
Political Knowledge	1.378	-1.732	2.070		1.836	0.103	3.282	*	1.379	-1.003	3.605		0.486	-1.437	2.763	
Democratic Party Identification	3.855	-1.664	8.513		0.134	-6.631	6.262		1.816	-8.753	6.791		-2.949	-7.899	8.589	
Republican Party Identification	7.371	-0.789	12.064		1.176	-3.898	8.408		2.153	-2.416	6.277		5.815	-2.803	12.543	
Democratic Candidate Traits	1.933	-1.514	3.198		1.834	-0.965	4.676		1.587	0.299	5.924	*	2.112	-2.398	4.252	
Republican Candidate Traits	3.294	-1.046	7.924		2.366	-1.748	5.617		1.717	-3.325	4.646		-2.468	-5.761	2.861	
$1 - \hat{\pi}$	0.072	0.053	0.127		0.102	0.081	0.161		0.091	0.068	0.144		0.110	0.081	0.174	

Source: 2016 ANES. Note: Cut points of the placement part and intercept of uncertainty part are not displayed. An estimate is considered as significant at the 5%-level when the bootstrap confidence intervals cover the estimate but not zero, N=1539.

Table 2: Parameter estimates of the BetaBin model: Democratic Candidate Placements

	Liberal-Conservative				Spending and Services				Defense Spending				Health Insurance			
	Estimate	BS.2.5	BS.97.5	sig.	Estimate	BS.2.5	BS.97.5	sig.	Estimate	BS.2.5	BS.97.5	sig.	Estimate	BS.2.5	BS.97.5	sig.
Placement Part																
Education	0.019	-0.080	0.087		0.023	-0.043	0.087		-0.033	-0.128	0.037		-0.009	-0.085	0.071	
Political Interest	0.165	0.011	0.313	*	-0.330	-0.484	-0.215	*	-0.086	-0.256	0.099		0.163	0.000	0.332	
Political Knowledge	0.096	-0.098	0.272		0.017	-0.110	0.166		0.041	-0.182	0.222		0.196	0.026	0.341	*
Self-Placement	0.272	0.060	0.517	*	0.026	-0.120	0.164		-0.111	-0.265	0.060		0.089	-0.039	0.220	
Democratic Party Identification	-0.229	-0.777	0.313		0.321	-0.004	0.696		-0.531	-1.145	-0.052	*	-0.625	-1.076	-0.231	*
Democratic Candidate Traits	-0.850	-1.107	-0.573	*	1.067	0.863	1.328	*	-1.823	-2.190	-1.553	*	-0.693	-0.923	-0.489	*
Uncertainty Part																
Education	0.078	-0.756	0.537		-0.180	-0.577	0.181		0.064	-0.498	0.419		-0.235	-1.083	0.548	
Political Interest	0.199	-2.011	1.147		-0.191	-0.721	0.572		-0.741	-2.432	0.341		-0.593	-2.560	0.112	
Political Knowledge	0.475	-0.731	1.770		0.157	-0.504	0.599		0.854	-0.280	2.165		0.505	-0.506	2.118	
Self-Placement	-2.823	-3.891	-0.504	*	0.524	0.139	1.685	*	-1.263	-3.111	-0.163	*	-0.775	-2.528	-0.037	*
Democratic Party Identification	2.316	-1.966	5.242		2.576	0.060	13.610	*	0.710	-0.916	11.703		1.789	-0.762	5.166	
Democratic Candidate Traits	1.179	-0.077	4.739		2.120	1.140	3.472	*	0.770	-0.439	2.539		1.867	0.519	4.416	*
$1 - \hat{\pi}$	0.186	0.105	0.259		0.201	0.161	0.237		0.321	0.253	0.387		0.208	0.137	0.284	

Source: 2016 ANES. Note: Cut points of the placement part and intercept of uncertainty part are not displayed. An estimate is considered as significant at the 5%-level when the bootstrap confidence intervals cover the estimate but not zero, N=1539.

Table 3: Parameter estimates of the BetaBin model: Republican Candidate Placements

	Liberal-Conservative				Spending and Services				Defense Spending				Health Insurance			
	Estimate	BS.2.5	BS.97.5	sig.	Estimate	BS.2.5	BS.97.5	sig.	Estimate	BS.2.5	BS.97.5	sig.	Estimate	BS.2.5	BS.97.5	sig.
Placement Part																
Education	-0.066	-0.135	0.014		-0.010	-0.076	0.082		0.055	-0.030	0.111		-0.032	-0.097	0.051	
Political Interest	-0.077	-0.237	0.144		0.009	-0.115	0.107		-0.254	-0.403	-0.110		-0.096	-0.245	0.053	
Political Knowledge	0.079	-0.079	0.413		0.093	-0.042	0.270		0.037	-0.107	0.138		-0.015	-0.138	0.105	
Self-Placement	0.413	0.239	0.554	*	0.085	-0.111	0.206		-0.152	-0.375	-0.092		0.054	-0.027	0.154	
Republican Party Identification	-0.522	-0.952	0.046		-0.400	-0.810	-0.043	*	0.751	0.184	1.026	*	0.408	0.109	0.805	*
Republican Candidate Traits	-0.388	-0.698	0.029		-0.556	-0.921	-0.287	*	0.335	-0.071	0.562		0.417	0.126	0.677	*
Uncertainty Part																
Education	0.202	-0.362	1.187		0.106	-0.314	1.590		0.123	-0.316	1.720		0.795	-0.663	1.846	
Political Interest	0.616	-0.754	2.442		-0.070	-1.856	1.293		-0.453	-2.595	0.766		-0.317	-2.008	0.957	
Political Knowledge	0.454	-0.188	2.534		0.700	-0.098	3.644		-0.059	-0.865	2.392		-0.423	-1.553	0.932	
Self-Placement	0.489	-0.262	1.393		-0.381	-2.616	1.505		-3.026	-3.434	2.404		0.546	-2.840	2.434	
Republican Party Identification	1.564	-0.356	9.685		1.032	-0.283	11.849		-0.110	-2.246	11.412		3.137	-1.564	15.647	
Republican Candidate Traits	2.783	1.343	6.417	*	0.838	-1.444	4.606		-2.106	-2.811	5.899		3.254	-0.634	5.839	
$1 - \hat{\pi}$	0.172	0.128	0.285		0.238	0.148	0.286		0.244	0.083	0.275		0.139	0.117	0.189	

Source: 2016 ANES. Note: Cut points of the placement part and intercept of uncertainty part are not displayed. An estimate is considered as significant at the 5%-level when the bootstrap confidence intervals cover the estimate but not zero, N=1539.

Table 4 contrasts performance measures of the BetaBin models with the cumulative models without uncertainty component. Model performance is measured by the Log-Likelihood (LogL) and the AIC.[5] The values indicate that the mixture models outperform the traditional ordinal models. The model fit described by the Log-Likelihood is better for the mixture model than the pure cumulative model in all settings. Furthermore, almost all AIC values for the pure cumulative models are larger than for the mixture models, except for the self-placement on the issue of health insurance. The performance measures suggest that the mixture models give a better model fit, although the mixture model is much more complex. The pure cumulative models are based on 13 parameters for the self-placements (6 intercepts and 7 covariates) and 12 parameters for the candidate placements (6 intercepts and 6 covariates). The corresponding mixture models are based on a total of 22 and 20, respectively: the identical number of parameters enters the placement part (13 and 12, respectively), parameters to model the shape of the uncertainty distribution (1 intercept, 7 and 6 covariates, respectively), and the parameter for the mixture weight estimate $\hat{\pi}$.

5.2 Spatial Voting under Uncertainty

Next, we compare the voter choice models based on the original placements with the ones we predict based on the mixture models. These survey response models adjust for special response styles due to uncertainty, which leads to somehow biased observed placements. Thus, we correct the observed positions by using the estimates of the structural component to generate positions which are adjusted by the detected uncertainty. The difference of two cumulative probabilities gives the probability π_{ir} for a particular response category r

$$\pi_{ir} = P(Y_i \leq r) - P(Y_i \leq r-1),$$

so that we obtain for each observation i the probability for choosing category $\{1, \ldots, k\}$ based on the estimates of the structural component of the model and the considered predictors. The category with the highest probability is chosen as the most likely position

[5] The AIC is defined by $AIC = -2l(\hat{\boldsymbol{\theta}}) + 2m$, where $l(\hat{\boldsymbol{\theta}})$ is the log-likelihood function computed at the maximum of the estimated parameter vector $\boldsymbol{\theta}$ and m is the number of model parameters, comprising all model parameters.

Table 4: Model comparisons based on performance measures

	LogL		AIC	
	Cumulative	Mixture	Cumulative	Mixture
Liberal-Conservative Scale				
Self-Placements	-2114.413	-2102.812	4254.827	4249.624
Democratic Candidate Placements	-2042.724	-2010.927	4109.449	4061.854
Republican Candidate Placements	-2437.803	-2408.940	4899.607	4857.880
Spending and Services				
Self-Placements	-2477.834	-2461.440	4981.668	4966.880
Democratic Candidate Placements	-2150.124	-2047.078	4324.248	4134.157
Republican Candidate Placements	-2466.211	-2439.574	4956.421	4919.149
Defense Spending				
Self-Placements	-2527.637	-2512.997	5081.274	5069.993
Democratic Candidate Placements	-2412.246	-2325.138	4848.491	4690.276
Republican Candidate Placements	-2341.586	-2305.823	4707.172	4651.647
Health Insurance				
Self-Placements	-2567.289	-2559.251	5160.578	5162.501
Democratic Candidate Placements	-2196.097	-2153.841	4416.193	4347.683
Republican Candidate Placements	-2194.387	-2171.660	4412.774	4383.320

for each observation i. These adjusted values are used as explanatory variables in the vote choice model. At least 50% of the adjusted values are different from the original observed values. With almost 70%, most values are adjusted for self-placement on the issue of spending and services.

Table 5 compares the voter choice models based on the original placements with the ones we predicted based on the mixture models. The estimates for the spatial proximities are displayed at the top, followed by the estimates for the voter attributes. The constant and the parameters related to voter attributes are set to zero for the Republican candidate to ensure model identification. Thus, the interpretation of these coefficients is always relative to Donald Trump. When inspecting the proximities, we observe that the effects are positive in both models so that the larger the proximity between the candidates and the voter, the more likely it is vote for this candidate. However, the effect sizes differ between both models. Based on the original placements, the liberal-conservative scale has the largest impact, followed by attitudes toward defense spending. Spatial proximities on

the issue of health insurance show the weakest effect. In the voter choice model based on the adjusted placements, the liberal-conservative scale does not significantly impact on voting anymore, and also the remaining issues differ in effect strength. The effects for both the issues of spending and services and health insurance are more than twice the size of the ones we obtained for the unadjusted placements. We also identify some interesting individual-specific effects, indicating that some segments are more likely to vote for a particular candidate. In the vote choice model based on original placements Blacks and Latinos tend to favor the Democratic candidate Clinton. The same pattern is observed for higher education segments and those that positively evaluate the economy. In the vote choice model with adjusted placements, we observe the same direction of effects, but only the effects for Latinos and economic considerations remain statistically significant. An inspection of some goodness-of-fit measures, reported at the bottom of Table 5, reveals that the vote choice model that accounts for uncertainty in the issue placements performs better according to the Log-Likelihood and AIC than the model that relies on the original, unadjusted placements. In particular, the AIC is reduced by around 27% with the same number of parameters.

6 Discussion and Concluding Remarks

In this manuscript, we developed a vote choice model that accounts for the uncertainty in issue placements, which arises from the difficulty to select a particular category on ordinal policy scales. Our approach consists of two stages. First, the perceived party platforms and policy preferences are adjusted for uncertainty. Then, these values are used to estimate voter choices. Drawing on the 2016 US presidential election and examining voting for one of the two major candidates, we showed that our approach outperforms the traditional models at both stages: the cumulative model without uncertainty at the first stage and the vote choice model without uncertainty correction at the second stage. So far, we focus on goodness-of-fit measures based on the likelihood of the fitted models. However, it might be useful to consider additionally predictive measures to compare the models. One strategy would be to use k-cross-validation, where the data is split into k sets. $k-1$ parts are used for estimation and the kth part to evaluate how good the model

Table 5: Voter Choice Models

	Original Placements			*Adjusted Placements*		
Predictors	coef.	se	p-value	coef.	se	p-value
Liberal-Conservative	0.757	0.093	0.000	0.047	0.149	0.751
Spending and Services	0.419	0.086	0.000	0.949	0.228	0.000
Defense Spending	0.604	0.086	0.000	0.635	0.167	0.000
Health Insurance	0.207	0.055	0.000	0.392	0.132	0.003
Age	0.002	0.087	0.980	-0.200	0.105	0.057
Gender	0.157	0.289	0.586	-0.335	0.350	0.338
Black	2.806	0.768	0.000	0.725	0.678	0.285
Latino	1.736	0.716	0.015	2.047	0.836	0.014
North Central	-0.409	0.386	0.289	-0.727	0.484	0.134
South	-0.648	0.426	0.128	-0.948	0.485	0.051
West	-0.751	0.471	0.111	0.381	0.588	0.517
(Ref: Northeast)						
Economy	0.772	0.167	0.000	0.539	0.206	0.009
Education	0.244	0.074	0.001	0.050	0.089	0.572
Constant	-9.040	1.674	0.000	-4.465	1.877	0.017
LogL		-172.881			-121.825	
AIC		373.762			271.649	
Pseudo R^2		0.838			0.886	
df		14			14	

Source: 2016 ANES. Notes: The response variable is binary and gives the vote intention for either the Democratic or Republican candidate. The interpretation of voter attributes refers to Clinton as compared to Trump. N=1539.

performs. In our application, it may be appropriate to evaluate how many times the predicted choice is identical to the observed choice behavior. There are measures, such as the Brier score, which are appropriate to evaluate discrete responses. Although our empirical application rests on a binary choice model, the approach can be easily extended to a multi-party setting by replacing the binary choice model with a multinomial one. Likewise, the number of issue dimensions can be extended as well.

Acknowledgements: This manuscript was presented at the Workshop "Recent Developments of Spatial Models of Party Competition" at the Mannheim Centre for European Social Research (MZES), July 26-27, 2019. We thank the participants for highly valuable comments, and especially Bernard Grofman for discussing our manuscript.

References

Adams, J., and S. Merrill III. 1999. "Party policy equilibrium for alternative spatial voting models: An application to the Norwegian Storting." *European Journal of Political Research* 36(2): 235–255.

Adams, J., S. Merrill III, and B. Grofman. 2005. *A Unified Theory of Party Competition: A Cross-national Analysis Integrating Spatial and Behavioral Factors.* New York, NY: Cambridge University Press.

Aldrich, J. H., G. S. Schober, S. Ley, and M. Fernandez. 2018. "Incognizance and Perceptual Deviation: Individual and Institutional Sources of Variation in Citizens' Perceptions of Party Placements on the Left–Right Scale." *Political Behavior* 40(2): 415–433.

Aldrich, J. H., and R. D. McKelvey. 1977. "A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections." *The American Political Science Review* 71(1): 111–130.

Aldrich, J. H., R. G. Niemi, G. Rabinowitz, and D. W. Rohde. 1982. "The Measurement of Public Opinion about Public Policy: A Report on Some New Issue Question Formats." *American Journal of Political Science* 26(2): 391–414.

Alvarez, R. M. 1999. *Information and Elections.* Ann Arbor, MI: University of Michigan Press.

Alvarez, R. M., and C. H. Franklin. 1994. "Uncertainty and Political Perceptions." *The Journal of Politics* 56(3): 671–688.

Alvarez, R. M., and J. Brehm. 1995. "American Ambivalence Towards Abortion Policy: Development of a Heteroskedastic Probit Model of Competing Values." *American Journal of Political Science* 39(4): 1055–1082.

Alvarez, R. M., and J. Brehm. 1997. "Are Americans Ambivalent Towards Racial Policies?" *American Journal of Political Science* 41(2): 345–374.

Alvarez, R. M., and J. Brehm. 1998. "Speaking in Two Voices: American Equivocation about the Internal Revenue Service." *American Journal of Political Science* 42(2): 418–452.

Alvarez, R. M., and J. Brehm. 2002. *Hard Choices, Easy Answers: Values, Information, and American Public Opinion.* Princeton, NJ: Princeton University Press.

Bartels, L. M. 1986. "Issue Voting Under Uncertainty: An Empirical Test." *American Journal of Political Science* 30(4): 709–728.

Baumgartner, H., and J.-B. Steenkamp. 2001. "Response Styles in Marketing Research: A Cross-National Investigation." *Journal of Marketing Research* 38(2): 143–156.

Berinsky, A. J., and J. B. Lewis. 2007. "An Estimate of Risk Aversion in the U.S. Electorate." *Quarterly Journal of Political Science* 2(2): 139–154.

Brady, H. E. 1985. "The Perils of Survey Research: Inter-Personally Incomparable Responses." *Political Methodology* 11(3/4): 269–291.

Campbell, J. E. 1983*a*. "Ambiguity in the Issue Positions of Presidential Candidates: A Causal Analysis." *American Journal of Political Science* 27(2): 284–293.

Campbell, J. E. 1983*b*. "The Electoral Consequences of Issue Ambiguity: An Examination of the Presidential Candidates' Issue Positions from 1968 to 1980." *Political Behavior* 5(3): 277–291.

Conover, P. J., and S. Feldman. 1981. "The Origins and Meaning of Liberal/Conservative Self-Identifications." *American Journal of Political Science* 25(4): 617–645.

Conover, P. J., and S. Feldman. 1982. "Projection and the Perception of Candidates' Issue Positions." *Western Political Quarterly* 35(2): 228–244.

Davis, O. A., M. J. Hinich, and P. C. Ordeshook. 1970. "An Expository Development of a Mathematical Model of the Electoral Process." *The American Political Science Review* 64(2): 426–448.

De Vries, C., and M. R. Steenbergen. 2013. "Variable Opinions: The Predictability of Support for Unification in European Mass Publics." *Journal of Political Marketing* 12(1): 121–141.

D'Elia, A., and D. Piccolo. 2005. "A Mixture Model for Preference Data Analysis." *Computational Statistics & Data Analysis* 49(3): 917–934.

Delli Carpini, M. X., and S. Keeter. 1993. "Measuring Political Knowledge: Putting First Things First." *American Journal of Political Science* 37(4): 1179–1206.

Downs, A. 1957. *An Economic Theory of Democracy.* New York, NY: Harper & Row.

Enelow, J. M., and M. J. Hinich. 1981. "A New Approach to Voter Uncertainty in the Downsian Spatial Model." *American Journal of Political Science* 25(3): 483–493.

Enelow, J. M., and M. J. Hinich. 1984. *The Spatial Theory of Voting: An Introduction.* Cambridge, England: Cambridge University Press.

Feldman, S., and P. J. Conover. 1983. "Candidates, Issues and Voters: The Role of Inference in Political Perception." *The Journal of Politics* 45(4): 810–839.

Gill, J. 2005. "An Entropy Measure of Uncertainty in Vote Choice." *Electoral Studies* 24(3): 371–392.

Granberg, D. 1987. "A Contextual Effect in Political Perception and Self-Placement on an Ideology Scale: Comparative Analyses of Sweden and the U.S." *Scandinavian Political Studies* 10(1): 39–60.

Granberg, D., and E. Brent. 1980. "Perceptions of Issue Positions of Presidential Candidates: Candidates Are Often Perceived by Their Supporters as Holding Positions on the Issues That Are Closer to the Supporters' Views than They Really Are." *American Scientist* 68(6): 617–625.

Granberg, D., and R. Jenks. 1977. "Assimilation and Contrast Effects in the 1972 Election." *Human Relations* 30(7): 623–640.

Granberg, D., and T. A. Brown. 1992. "The Perception of Ideological Distance." *The Western Political Quarterly* 45(3): 727–750.

Harbers, I., C. E. De Vries, and M. R. Steenbergen. 2013. "Attitude Variability Among Latin American Publics: How Party System Structuration Affects Left/Right Ideology." *Comparative Political Studies* 46(8): 947–967.

Hare, C., D. A. Armstrong, R. Bakker, R. Carroll, and K. T. Poole. 2015. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59(3): 759–774.

Harvey, A. C. 1976. "Estimating Regression Models with Multiplicative Heteroscedasticity." *Econometrica* 44(3): 461–465.

Heider, F. 1946. "Attitudes and Cognitive Organization." *The Journal of Psychology* 21(1): 107–112.

Heider, F. 1958. *The psychology of interpersonal relations*. Hoboken, NJ: John Wiley & Sons.

Iannario, M., and D. Piccolo. 2016. "A Generalized Framework for Modelling Ordinal Data." *Statistical Methods & Applications* 25(2): 163–189.

Luskin, R. C. 1987. "Measuring Political Sophistication." *American Journal of Political Science* 31(4): 856–899.

Luskin, R. C. 1990. "Explaining Political Sophistication." *Political Behavior* 12(4): 331–361.

Markus, G. B., and P. E. Converse. 1979. "A Dynamic Simultaneous Equation Model of Electoral Choice." *The American Political Science Review* 73(4): 1055–1070.

Mauerer, I., and M. Schneider. 2019. Perceived Party Placements and Uncertainty on Immigration in the 2017 German Election. In *Jahrbuch für Handlungs- und Entscheidungstheorie: Band 11*, edited by M. Debus, M. Tepe, and J. Sauermann, 117–143. Wiesbaden, Germany: Springer Fachmedien Wiesbaden.

McLachlan, G. J., and D. Peel. 2000. *Finite Mixture Models*. New York, NY: Wiley.

Merrill III, S., and B. Grofman. 1999. *A Unified Theory of Voting: Directional and Proximity Spatial Models*. New York, NY: Cambridge University Press.

Merrill III, S., B. Grofman, and J. Adams. 2001. "Assimilation and Contrast Effects in Voter Projections of Party Locations: Evidence from Norway, France, and the USA." *European Journal of Political Research* 40(2): 199–221.

Piccolo, D., and R. Simone. 2019. "The class of cub models: statistical foundations, inferential issues and empirical evidence." *Statistical Methods & Applications* . online published.

Poole, K. T. 1998. "Recovering a Basic Space From a Set of Issue Scales." *American Journal of Political Science* 42(3): 954–993.

Poole, K. T., J. Lewis, H. Rosenthal, J. Lo, and R. Carroll. 2016. "Recovering a Basic Space from Issue Scales in R." *Journal of Statistical Software* 69(7): 1–21.

Pößnecker, W. 2019. "MRSP: Multinomial Response Models with Structured Penalties.". R package version 0.6.11, https://github.com/WolfgangPoessnecker/MRSP.

Rapeli, L. 2013. *The Conception of Citizen Knowledge in Democratic Theory*. Basingstoke, England: Palgrave Macmillan.

Rozenas, A. 2013. Inferring Ideological Ambiguity from Survey Data. In *Advances in Political Economy: Institutions, Modelling and Empirical Analysis*, edited by N. Schofield, G. Caballero, and D. Kselman, 369–382. Heidelberg, Germany: Springer.

Shepsle, K. A. 1972. "The Strategy of Ambiguity: Uncertainty and Electoral Competition." *The American Political Science Review* 66(2): 555–568.

Sherif, M., and C. I. Hovland. 1961. *Social judgment: Assimilation and contrast effects in communication and attitude change*. New Haven, CT: Yale University Press.

Stoetzer, L. F. 2017. "A Matter of Representation: Spatial Voting and Inconsistent Policy Preferences." *British Journal of Political Science,* 49(3): 941–956.

Thurner, P. W. 2000. "The Empirical Application of the Spatial Theory of Voting in Multiparty Systems with Random Utility Models." *Electoral Studies* 19(4): 493–517.

Tomz, M., and R. P. Van Houweling. 2009. "The Electoral Implications of Candidate Ambiguity." *American Political Science Review* 103(1): 83–98.

Tutz, G. 2012. *Regression for Categorical Data*. Cambridge, England: Cambridge University Press.

Tutz, G., and M. Schneider. 2019. "Flexible uncertainty in mixture models for ordinal responses." *Journal of Applied Statistics* 46(9): 1582–1601.

Tutz, G., M. Schneider, M. Iannario, and D. Piccolo. 2017. "Mixture Models for Ordinal Responses to Account for Uncertainty of Choice." *Advances in Data Analysis and Classification* 11(2): 281–305.

Vaerenbergh, Y. V., and T. D. Thomas. 2013. "Response Styles in Survey Research: A Literature Review of Antecedents, Consequences, and Remedies." *International Journal of Public Opinion Research* 25(2): 195–217.

Yee, T. W. 2016. *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.1-1.

A.5. Variable Selection in Mixture Models with an Uncertainty Component

Schneider, M., Pößnecker, W. and G. Tutz (2019): Variable Selection in Mixture Models with an Uncertainty Component. *Technical Report 225*, Department of Statistics, Ludwig-Maximilians-Universität München, doi:10.5282/ubm/epub.68452.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Micha Schneider, Wolfgang Pößnecker, Gerhard Tutz

Variable Selection in Mixture Models with an Uncertainty Component

Technical Report Number 225, 2019
Department of Statistics
University of Munich

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

Variable Selection in Mixture Models with an Uncertainty Component

Micha Schneider, Wolfgang Pößnecker, Gerhard Tutz
Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München

August 6, 2019

Abstract

Mixture Models as CUB and CUP models provide the opportunity to model discrete human choices as a combination of a preference and an uncertainty structure. In CUB models the preference is represented by shifted binomial random variables and the uncertainty by a discrete uniform distribution. CUP models extend this concept by using ordinal response models as the cumulative model for the preference structure. To reduce model complexity we propose variable selection via group lasso regularization. The approach is developed for CUB and CUP models and compared to a stepwise selection. Both simulated data and survey data are used to investigate the performance of the selection procedures. It is demonstrated that variable selection by regularization yields stable parameter estimates and easy-to-interpret results in both model components and provides a data-driven method for model selection in mixture models with an uncertainty component.

Keywords: Mixture Models; Variable Selection; lasso, CUB model; CUP model

1 Introduction

Mixture models are widely used to model heterogeneity in populations. D'Elia and Piccolo (2005) proposed a mixture type model for ordinal responses that accounts for the psychological process of human choices. The model has been investigated and extended in a series of papers for example by Piccolo and D'Elia (2008), Iannario and Piccolo (2012b) and Iannario and Piccolo (2012a). The basic concept of the so-called CUB model is that the choice of a response category is determined by a mixture of feeling and uncertainty. Feeling refers to the deliberate choice of a response category determined by the preferences of a person

while uncertainty refers to the inherent individual's indecision. The first component is modelled by a binomial distribution, the latter by a discrete uniform distribution across response categories. An introduction and overview is given in Piccolo and Simone (2019). The CUP model described in Tutz et al. (2017) and further developed by Tutz and Schneider (2019) extends this concept by using any ordinal model as the cumulative model for the preference structure.
In this type of models the right choice of covariates is essential to get sensible models. Even for a moderate number of covariates simple methods as all-subset selection are too time consuming so that other techniques are in demand. Using penalization techniques as lasso by Tibshirani (1996) can overcome this issue. Previous work on variable selection in mixtures focused on mixtures of normal densities and mixtures where the weights do not depend on covariates. Khalili and Chen (2007) used the lasso approach for mixture models and chose a penalty function which is proportional to the mixture weight. Further work was done by Luo et al. (2008) who propose to penalize the coefficients within and between Gaussian components and Städler et al. (2010) focus on high dimensional settings where $p >> n$. But regularization has not been used to investigate the structure of CUB and CUP models with a specific discrete component and weights that depend on individual-specific covariates. In the following we show how to adopt the lasso framework to CUB and CUP models and compare the approach to a forward selection procedure.

The article is organized as follows. First, in section 2 the models are briefly described. In section 3 we discuss variable selection by a step procedure and the proposed lasso method, followed by section 4 about computational aspects of estimation, initialization and convergence. In section 5 we provide results of a simulation study and in section 6 we use the SHIW and ALLBUS survey to show the applicability of the methods on two real data problems. Finally the results are summarized.

2 Model Class

Let the probability that an individual i chooses the category r from ordered categories $\{1, \dots, k\}$ given explanatory variables $\boldsymbol{z}_i, \boldsymbol{x}_i$ be composed of the individual's propensity towards uncertainty and preference structure. The mixture distribution has the general form

$$P(R_i = r|\boldsymbol{x}_i) = \pi_i P_M(Y_i = r|\boldsymbol{x}_i) + (1 - \pi_i) P_U(U_i = r), \tag{1}$$

where π_i is the propensity or mixture weight, $P_M(Y_i = r|\boldsymbol{x}_i)$ is a model for the preference, and the uncertainty component $P_U(U_i = r)$ is determined by a uniform distribution with probability $1/k$ for each response category. The uncertainty is assumed to include all kinds of indecision related to the nature of human choices like willingness to respond, lack of time, partial understanding

etc. The probability π_i is assumed to be linked to covariates by the logit model

$$\text{logit}\,(\pi_i) \;=\; \boldsymbol{z}_i^T\boldsymbol{\beta}\,, \qquad\qquad i = 1, 2, \ldots, n\,. \tag{2}$$

The CUB and CUP models, used in this article, only vary in the choice of the preference component. The preference structure in CUB models (combination of uncertainty and binomial) is modelled by a shifted binomial distribution $b_r(.)$ with parameter ξ, that is,

$$b_r(\xi_i) = \binom{k-1}{r-1}\xi_i^{k-r}(1-\xi_i)^{r-1}, \quad r \in \{1, \ldots, k\},$$

where ξ_i is linked to the covariates $\boldsymbol{x}_i^T$ by

$$\text{logit}\,(\xi_i) \;=\; \gamma_0 + \boldsymbol{x}_i^T\boldsymbol{\gamma}\,, \qquad\qquad i = 1, 2, \ldots, n\,. \tag{3}$$

The so called CUP model (combination of uncertainty and preference), described in Tutz et al. (2017), uses any ordinal model. A traditional model is the cumulative logit model

$$\log\left(\frac{P(Y_i \leq r|\boldsymbol{x}_i)}{P(Y_i > r|\boldsymbol{x}_i)}\right) = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}, \quad r = 1, \ldots, k-1.$$

(see Agresti, 2013; Tutz, 2012). The CUP models are more flexible and can handle complex ordinal data structures. However, the intercept parameters depend on the number of categories k so that more parameters have to be estimated.
Both models use covariates to model the preference structure and the weights. In general the covariates $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ may be identical, completely different or overlap. It should be mentioned that the omission of the uncertainty component typically yields biased parameter estimates.

3 Variable Selection

Since there are two sets of covariates, variable selection is an major issue in mixture models. Let $\mathcal{X}$ contain all possible variables which can be selected for the two independent sets of $\boldsymbol{z}$ and $\boldsymbol{x}$, which are linked to the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. It is typically not known which variables are relevant for the weights ($\boldsymbol{z}$) and which for the preference structure ($\boldsymbol{x}$) so that variable selection has to handle two separate effect structures. We propose a variable selection based on penalty terms that are tailored to the problem of selecting variables in two components and compare it with a stepwise procedure.

3.1 Stepwise Variable Selection

Two traditional methods are the forward and backward selection. The latter allows that all available explanatory variables are included in both components and the model complexity is reduced stepwise. Especially in mixture models too many possibly correlated covariates can lead to model degeneracy and convergence problems so that the estimates in the fit are hardly trustworthy or the complete model can not be fitted.
Alternatively, one might use a forward search procedure. Here the selection process starts with a basic model as the intercept model. In the first step all models with one covariate in any part of the model are fitted. Then the model with the strongest improvement in terms of a specific criterion is selected. In the next step the procedure continues with this selected model and all remaining covariates are evaluated. The procedure continues until no improvement is detected. In each step a covariate is assigned to only one of the two variable sets $\boldsymbol{z}$ and $\boldsymbol{x}$. If a covariate is selected for one of the two sets, it is still possible that the same covariate is selected for the other variable set later. Several criteria can be used:

$$AIC(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + 2df(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}),$$
$$BIC(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + log(n)df(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}),$$

or the likelihood-ratio test with

$$lq = -2[l_0(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) - l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})] \overset{a}{\sim} \chi^2(|df(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) - df_0(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})|),$$

where the likelihood of the previous model is compared to the likelihood of the enlarged model. Since the likelihood-ratio test uses the difference of deviances we refer to it also as "deviance" criterion. That variable is selected that yields the largest improvement in AIC or BIC or the smallest p-value of the likelihood-ratio test. If there are several p-values that are numerically close to zero, the model with the largest deviance difference is selected. When the AIC/BIC does not improve or the p-value of the likelihood-ratio test is larger than 0.05 the forward selection is terminated. The estimation of these models is performed as described in Section 4.1. The initializations and convergence checks are described in detail in section 4.2.
Backward/forward strategies have the disadvantage that they are rather variable. The instability of stepwise regression models was demonstrated, for example, by Breiman (1996). Moreover, the standard errors computed for the final model are not trustworthy because they simply ignore the model search. The larger the available number of variables the more models have to be estimated so that these techniques may not work well for very large data sets.

3.2 Variable Selection by Penalization

We propose to use a version of the lasso (Tibshirani, 1996) that is adapted to the mixture models to obtain a procedure that is not limited by the number of variables and produces stable results. The penalized log-likelihood that is to be maximized is given by

$$l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = l(\boldsymbol{\beta}, \boldsymbol{\gamma}) - J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}),$$

where $l(\boldsymbol{\beta}, \boldsymbol{\gamma})$ denotes the un-penalized log-likelihood and $J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is a specific penalty term that enforces the selection of variables in both model components. Let the vectors $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ be partitioned into $\boldsymbol{z}_i^T = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{ig}^T)$ and $\boldsymbol{x}_i^T = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{ih}^T)$ such that each components refer to a single variable. For example, the vector $\boldsymbol{z}_{ij}$ can represent all the dummy variables that are linked to the j-th variable, or represent the power functions of the j-th variable if one includes polynomial terms. The corresponding predictors are $\boldsymbol{z}_i^T\boldsymbol{\beta}$ and $\boldsymbol{x}_i^T\boldsymbol{\gamma}$ with corresponding partitioning of the parameter vectors, $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_g^T)$ and $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_h^T)$, respectively. Then the proposed penalty has the form

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda_\beta \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2 + \lambda_\gamma \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2, \tag{4}$$

where λ_β and λ_γ are the tuning parameters for the selection of $\boldsymbol{x}$ and $\boldsymbol{z}$ variables, respectively. The weights $df_{\boldsymbol{\beta}_j}$ are defined as the number of parameters collected in the corresponding parameter vector $\boldsymbol{\beta}_j$, the weights $df_{\boldsymbol{\gamma}_j}$ are defined in the same way. $\|\|_2$ is the unsquared L_2-Norm so that the penalty enforces the selection of variables in the spirit of the group lasso (Yuan and Lin, 2006) rather than selection of single parameters.

All covariables have to be standardized to ensure that the selection of variables does not depend on their scale. Categorical variables have to be orthonormalized. The parameters $\lambda_\beta, \lambda_\gamma$ can be used to enforce specific selection properties. If $\lambda_\beta \to \infty$ no explanatory variables are included in the mixture component and selection is restricted to the effect of explanatory variables on the structured response. If $\lambda_\gamma \to \infty$ no explanatory variables are included in the structured response part and selection is confined to the mixture component. If no specific structure is pre-specified $\lambda_\beta, \lambda_\gamma$ can take any value and can be chosen in a data driven way. A simplification that is tempting is to set $\lambda_\beta = \lambda_\gamma$. It might be sufficient in some applications but it should be used with care.

To select a certain model the use of a selection criterion is needed. In mixture models cross validation can be very time consuming so that we propose the use of AIC or BIC,

$$AIC(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + 2edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}),$$
$$BIC(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + log(n)edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}),$$

where $edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ is the effective degrees of freedoms of the mixture model. For each parameter set $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ the effective degrees of freedoms are calculated separately by

$$\begin{aligned} edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) &= edf(\hat{\boldsymbol{\beta}}) + edf(\hat{\boldsymbol{\gamma}}) \\ &= 1 + \sum_{j=1}^{g} edf(\hat{\boldsymbol{\beta}}_j) + I + \sum_{j=1}^{h} edf(\hat{\boldsymbol{\gamma}}_j), \end{aligned}$$

where 1 refers to the intercept β_0 and I to the number of intercepts γ_0. The CUB-model consist of $1 + 1$-intercepts and the CUP-model of $1 + (k-1)$-intercepts. g and h denote the number of the penalized variables. Following Yuan and Lin (2006) the effective degrees of freedom of each variable are computed by

$$\begin{aligned} edf(\hat{\boldsymbol{\beta}}_j) &= \mathbb{1}(\|\hat{\boldsymbol{\beta}}_j\|_2 > 0) + (df_{\boldsymbol{\beta}_j} - 1)\frac{\|\hat{\boldsymbol{\beta}}_j\|_2}{\|\hat{\boldsymbol{\beta}}_j^{ML}\|_2}, \\ edf(\hat{\boldsymbol{\gamma}}_j) &= \mathbb{1}(\|\hat{\boldsymbol{\gamma}}_j\|_2 > 0) + (df_{\boldsymbol{\gamma}_j} - 1)\frac{\|\hat{\boldsymbol{\gamma}}_j\|_2}{\|\hat{\boldsymbol{\gamma}}_j^{ML}\|_2}. \end{aligned}$$

If a variable is not penalized the edf are identical to $df_{\boldsymbol{\beta}_j}$ and $df_{\boldsymbol{\gamma}_j}$, respectively.

To find the best model the procedure has to be optimized with reference to all sensible combinations of the tuning parameters λ_β and λ_γ. We focus on the BIC criterion to find the best model with the lowest BIC value. A two-dimensional grid of λ-values is investigated and parallelized in the following way. One dimension is kept fixed while the other dimension is varied. By repeating this line search all combinations of tuning parameters are covered. For example, using a 15×15 grid results in a 15 times 1×15 line. The advantage of this approach is that we can use parallized computing architecture but also include the results of the previous model for the initialisation of the current model. This saves computing time and leads to non-degenerated results because the fit of the current model should be close to the fit of the previous model with a slightly different tuning parameter. Nevertheless we still use several random initialisations which are described in Section 4.2 to ensure that the fit is not conditioned on the previous results.
Using a complete random choice of tuning parameter combinations can be parallelized even better, but previous knowledge about model results can not be included easily. Another promising approach is the use of model based optimization as described in Bischl et al. (2017) to replace the more time consuming grid search.

4 Computational Aspects

4.1 Estimation with the EM-Algorithm

The mixture models considered in the previous sections can be estimated by an adapted version of the EM algorithm proposed by Dempster et al. (1977). Given the observed category y_i the likelihood contribution of observation i is

$$Pr(y_i|\boldsymbol{z}_i, \boldsymbol{x}_i) = \pi_i \, P_M(y_i|\boldsymbol{x}_i) + (1 - \pi_i) \, P_U(y_i) \quad y_i \in \{1, \ldots, k\} \tag{5}$$

yielding the log-likelihood

$$l_i(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \{\log(\pi_i) + \log(P_M(y_i|\boldsymbol{x}_i))\} + \{\log(1 - \pi_i) + \log(1/k)\}$$

The corresponding penalized log-likelihood is obtained by including the proposed penalty term yielding

$$\begin{aligned} l_i(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \{\log(\pi_i) + \log(P_M(y_i|\boldsymbol{x}_i))\} + \{\log(1 - \pi_i) + \log(1/k)\} \\ &\quad - \lambda_\beta \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2 - \lambda_\gamma \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2, \end{aligned}$$

and for all observations

$$\begin{aligned} l_c(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \sum_{i=1}^{n} \left[\{\log(\pi_i) + \log(P_M(y_i|\boldsymbol{x}_i))\} + \{\log(1 - \pi_i) + \log(1/k)\}\right] \\ &\quad - \lambda_\beta \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2 - \lambda_\gamma \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2. \end{aligned}$$

The EM algorithm uses the complete likelihood treating the membership to the uncertainty or structure component as missing data. Let z_i^* take the value 1 if observation i belongs to the structure component and zero if observation i belongs to the uncertainty component. Then the complete penalized log-likelihood is given by

$$\begin{aligned} l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \sum_{I=1}^{n} z_i^* \{\log(\pi_i) + \log(P_M(y_i|\boldsymbol{x}_i))\} + (1 - z_i^*) \{\log(1 - \pi_i) + \log(1/k)\} \\ &\quad - \lambda_\beta \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2 - \lambda_\gamma \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2, \end{aligned}$$

where the probability π_i depends on the individual characteristics by

$$\pi_i = 1/(1 + e^{-\boldsymbol{z}_i^T \boldsymbol{\beta}}).$$

Within the EM algorithm the log-likelihood is iteratively maximized by using an expectation and a maximization step. During the E-step the conditional

expectation of the complete log-likelihood given the observed data $\boldsymbol{y}$ and the current estimate $\boldsymbol{\theta}^{(s)} = (\boldsymbol{\beta}^{(s)}, \boldsymbol{\gamma}^{(s)})$,

$$M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \mathrm{E}(l_p(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}^{(s)})$$

has to be computed. Because $l_p(\boldsymbol{\theta})$ is linear in the unobservable data z_i^*, it is only necessary to estimate the current conditional expectation of z_i^*. From Bayes's theorem follows

$$\begin{aligned} E(z_i^*|\boldsymbol{y}, \boldsymbol{\theta}) &= P(z_i^* = 1|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= P(y_i|z_i^* = 1, \boldsymbol{x}_i, \boldsymbol{\theta}) P(z_i^* = 1|\boldsymbol{x}_i, \boldsymbol{\theta}) / P(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= \pi_i P_M(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) / (\pi_i P_M(y_i|\boldsymbol{x}_i) + (1 - \pi_i) 1/k) = \hat{z_i}^*. \end{aligned}$$

This is the posterior probability that the observation y_i belongs to the structure component of the mixture. For the s-th iteration one obtains

$$\begin{aligned} M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = & \underbrace{\sum_{i=1}^{n} \left\{ \hat{z^*}_i^{(s)} \log(\pi_i) + (1 - \hat{z^*}_i^{(s)}) \log(1 - \pi_i) \right\} - \lambda_\beta \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2}_{M_1} \\ & + \underbrace{\sum_{i=1}^{n} \left\{ \hat{z^*}_i^{(s)} \log(P_M(y_i|\boldsymbol{x}_i) + (1 - \hat{z^*}_i^{(s)}) \log(1/k) \right\} - \lambda_\gamma \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2}_{M_2} \end{aligned}$$

M_1 and M_2 can be estimated independently from each other but most traditional methods, such as Fisher-Scoring, can not be used because the derivatives do not exist. This problem can be solved with the fast iterative shrinkage-thresholding algorithm (FISTA) of Beck and Teboulle (2009) which is implemented in the MRSP package by Pößnecker (2019) and is used for the maximisation problem of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, which can be formulated generally as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmax}} \, l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = -\underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}} \, l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}} \; -l(\boldsymbol{\beta}, \boldsymbol{\gamma}) + J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}). \quad (6)$$

FISTA belongs to the class of proximal gradient methods in which only the unpenalized log-likelihood and its gradient is necessary. The solution for the unknown parameters $\boldsymbol{\theta}$ of the unpenalized log-likelihood in iteration $t + 1$ is given by:

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)} + \frac{1}{\nu} \nabla l(\hat{\boldsymbol{\theta}}^{(t)}),$$

where $\nu > 0$ is the inverse stepsize parameter. This estimator converges to the ML estimator so that each update of $\hat{\boldsymbol{\theta}}^{(t)}$ can be considered as an one-step approximation to the ML estimator based on the current iterate. This can be

used to define a searchpoint $\boldsymbol{u}$. To motivate the procedure with penalty the equation (6) is reformulated by Lagrange duality to

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathcal{C}}{\operatorname{argmin}}(-l(\boldsymbol{\theta})),$$

where $\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^d | J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) \leq \lambda\}$ is the constraint region corresponding to $J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Given $\boldsymbol{u}$, the proximal operator associated with the penalty $J_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ is then defined by

$$\mathcal{P}(\boldsymbol{u}) = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}} \left(\frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{u}\|^2 + J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \right)$$

and leads to

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \mathcal{P}_{\frac{\lambda}{\nu}}(\hat{\boldsymbol{\theta}}^{(t)} + \frac{1}{\nu}\nabla l(\hat{\boldsymbol{\theta}}^{(t)}).$$

In a first step the penalty is ignored and a step toward the ML estimator via first-order methods creates a search point. Then this search point is projected onto the constraint region C to account for the penalty term. A detailed description is given in Tutz et al. (2015).

For given $\boldsymbol{\theta}^{(s)}$ one computes in the E-step the weights $\hat{z^*}_i^{(s)}$ and in the M-step maximizes $M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$ (or rather M_1 and M_2), which yields the new estimates

$$\boldsymbol{\beta}^{(s+1)} = \text{argmax}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{ \hat{z^*}_i^{(s)} \log(\pi_i) + (1 - \hat{z^*}_i^{(s)}) \log(1 - \pi_i) \right\} - \lambda_\beta \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2$$

$$\boldsymbol{\gamma}^{(s+1)} = \text{argmax}_{\boldsymbol{\gamma}} \sum_{i=1}^{n} \hat{z^*}_i^{(s)} \log(P_M(y_i|\boldsymbol{x}_i)) - \lambda_\gamma \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2.$$

The E- and M-steps are repeated alternatingly until the difference $l_p(\boldsymbol{\theta}^{(s+1)}) - l_p(\boldsymbol{\theta}^{(s)})$ is small enough to assume convergence. To account for different sizes of the log-likelihood we define

$$\left| \frac{l_p(\boldsymbol{\theta}^{(s+1)}) - l_p(\boldsymbol{\theta}^{(s)})}{rel.tol/10 + |l_p(\boldsymbol{\theta}^{(s+1)})|} \right| < rel.tol$$

as stopping criteria. *rel.tol* is the relative tolerance which has to below a certain value, such as $1e-6$, to assume convergence. λ_β and λ_γ span a two-dimensional grid of tuning parameter space. Dempster et al. (1977) showed that under weak conditions the EM algorithm finds (only) a local maximum of the likelihood function. Hence it is sensible to use meaningful start values to find a good solution of the maximization problem, which is described in the next section 4.2.

4.2 Initialization and Convergence

Using meaningful starting values is a crucial point in mixture models. Misspecified starting values can lead to degenerated results, can be time consuming and can lead to poor estimation results. In the literature several methods were proposed as described in Baudry and Celeux (2015) and Karlis and Xekalaki (2003). In the random setting several random start values are chosen and all models are run until convergence. Then the best fit is selected. In the small EM strategy a large number of short runs are evaluated which do not have to converge completely. Only the model with the best fit is run until full convergence.

We use a special version of the small EM that refers to the model class considered here so that we use several different configurations. The mixture model components are restricted to two components so that for every observation only π_i and its complement $1 - \pi_i$ need to be chosen which has to sum up to 1. From experience we know that the mean weight for the uncertainty component $(1 - \bar{\pi})$ is in most cases between 0.1 and 0.4. By using this information we are able to create meaningful scenarios which are more likely to be close to a realistic solution. The first strategy is to use a fixed weight for all $\pi_i, i = 1, \ldots, n$. Here we chose $\pi_i = 0.9$ and $\pi_i = 0.7$ which correspond with a realistic weight for the uncertainty component $(1 - \pi_i)$ of 0.1 and 0.3, respectively.

The second strategy is drawing the weights π_i so that they are not constant for all observations. For example if we choose the value 0.7 and its complement 0.3 we assign randomly one of this two values to π_i. Because of the randomness we repeat the sample strategy at least two times for the chosen value resulting in two weight vectors $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$. To ensure that we have obtained different realizations we calculate for each observation the quadratic difference between $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ and compute the sum over all observations. If $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are identical the computed sum is zero so that $\boldsymbol{\pi}_2$ would be replaced by a new random sample. As a rule of thumb the overall sum has to be larger than $0.1 \cdot n$ to accept $\boldsymbol{\pi}_2$ as a valid initialization. Thus, the sample strategy produce several weight vectors for one chosen value. Here we used 0.9 as well as 0.7 leading to four different initializations. Together with the two constant initializations we obtain at least six configurations which are run until small convergence defined as rel.tol < 0.01 or until the maximal numbers of em-iterations equal to 60 depending on which criteria is reached first. The one with the best result is selected and is run until complete convergence (rel.tol $< 1e-6$ or maximal numbers of em-iterations equal to 200). One E- and one M-step is defined as one em-iteration.

Every time the model is called we use at least these six configurations regardless if we use the stepwise selection or the penalization. In the latter we may also include another weight initialization. As described in section 3.2 we use a line search to find the best tuning parameter combination. Thus from the second position onwards we can use the computed weights of the previous tuning parameter combination as initialization for the current weights. Since at the be-

ginning of each line search less information about a realistic model is available, we use more configurations for initialization. It consists of the constant choice and two samples of the values $0.6, 0.7, 0.8$ and 0.9.

Dempster et al. (1977) showed that the EM-algorithm converges to a local maximum which is measured in this models by a small difference in the (penalized) Likelihood. A priori we have little information about the exact geometrical shape of the likelihood so that in practice several problems may be occur.

It is well known that the speed of convergence is slow near to the maximum. If the density close to the maximum is very flat we experienced that the difference criterion in the (penalized) likelihood may be too strong. So the rule of likelihood difference is supplemented by a maximum number of em-iterations which can be used. Since the number of em-iterations is in most cases a backstop rule, we usually use a higher number of possible em-iterations which we think should usually not be reached. An exception is the initialization part of the algorithm where the algorithm should not run until complete convergence.

In some cases the (penalized) likelihood may jump between several values without approaching a maximum. This can be solved by adjusting the step-size or, if necessary, taking the best values even if the criterion of small differences in likelihood is not completely reached.

If the starting values are too close to the maximum it may happen that the algorithm diverges from the maximum or a good solution. For this case we implemented some checks to ensure that the best composition is used instead of using a solution which is worse but satisfying the criterion of small difference in (penalized) likelihood. During the EM-algorithm we keep the last ten results to be able to jump back to a previous solution. If this problem occurs between different starting values we select the next best solution. On the other hand we also want to allow the algorithm to search for a better solution. So we allow the algorithm to carry on after a dis-improvement of the likelihood in the first six em-iterations. If the algorithm still does not detect a better likelihood we jump back to the best solution found so far.

On rare occasions the parameters found may be close to the edge of the parameter space. Especially if almost all estimated mixture weights are close to zero or one. In this case we imposed a threshold of $1e-06$ to prevent the weights of being exact zero or exact one. Nevertheless if all mixture weights are close to one for one of the two components a mixture model may be questionable. In case of doubt we recommend to have a look at the estimated mixture weights.

The difference in the (penalized) likelihood is the main criteria of convergence. Only in the case of non-regular behaviour other criteria may be used. Different starting values not only help to find the best maximum but also help to avoid degenerated results.

5 Simulation

To illustrate whether the two selection methods are able to select the "true" covariates we use simulated data with effects and white noise variables with no effects. For $n = 3000$ observations and $k = 5$ response categories we generate five metric covariates from a standard normal distribution ($N(0, 1)$) and six categorical covariates. We use the same 11 covariates for $\boldsymbol{x}$ and $\boldsymbol{z}$, but the effects differ. The first two columns of Table 1 contain the exact values for β and γ used in the simulation. We want to use almost all possible combinations so that some effects of β and γ are identical and some differ. Also the covariates with no effect are sometimes identical (e.g. `Continous_5`) and in other cases there is an effect for only one of the parameters β and γ (e.g. `Continous_1`+4). In both parameter sets there are two continuous and three categorical covariates with no effects.

We use also relative small parameter values to create a realistic setting and to examine whether the size of the effect may have an impact on the different selection methods. The effects of the continuous covariates are 0.2, 0.3, 1, -1 and 2. Three categorical covariates are binary with the effect strength -0.2, 1 and 0. The other three categorical covariates consist of four, four and five categories. Only for the first of them we use effect sizes different from zero namely 0.2, 0.4 and 0.8. The other multi-categorical variables are white noise. The constants in the CUP model are $-2.391, -1.221, -0.259, 1.023$ and in the CUB model -1.5.

We generate $S = 20$ samples from the CUP- and CUB-model each following the described structure and selected variables with the penalization approach and forward selection. For the CUB and CUP model we present in Table 1 the number of times the covariate was selected depending on the used selection technique and the model. The last row includes the π-deviations, which measure the difference between the estimated individual mixture weights π and the true values, defined by

$$\pi\text{-Deviation} = \frac{1}{S}\sum_{j=1}^{S}\left(\frac{1}{n}\sum_{i=1}^{n}|\pi_{ij} - \hat{\pi}_{ij}|\right),$$

where S is the number of simulated data sets, n the number of observations in each data set, π_{ij} is the true mixture weight of the i-th observation in the j-th simulation, and $\hat{\pi}_{ij}$ is the corresponding estimated mixture weight. We compute the absolute differences on each individual mixture weight and use the average over all observations and all samples as a measurement of discrepancy.

Both the penalization and forward selection technique show good results. Both techniques selected covariates with clear effects (-1,1 and 2) in almost 100% and show worse performance with smaller effect size of the parameters. But the penalization technique selected more often covariates with smaller effect size than the forward selection. For example looking at `Categorical_1` the penalization technique selected these covariates in 30% and 95% of the cases in the CUB model compared with only 10% or 65% of the cases using forward selection. The

TABLE 1: *Result of simulated data*

	Simulated		Selected CUB				Selected CUP			
			Penalize		Forward		Penalize		Forward	
Covariates	β	γ	β	γ	β	γ	β	γ	β	γ
`Continous_1`	0	0.3	0%	100%	0%	100%	0%	100%	0%	100%
`Continous_2`	-1	1	100%	100%	100%	100%	100%	100%	100%	100%
`Continous_3`	2	2	100%	100%	100%	100%	100%	100%	100%	100%
`Continous_4`	0.2	0	60%	0%	20%	0%	40%	25%	15%	0%
`Continous_5`	0	0	5%	0%	0%	0%	5%	5%	0%	0%
`Categorical_1`	-0.2	-0.2	30%	95%	10%	65%	10%	95%	0%	25%
`Categorical_2`	1	1	100%	100%	100%	100%	95%	100%	90%	100%
`Categorical_3`	0	0	0%	0%	0%	0%	0%	15%	0%	0%
`Categorical_4:2`	0.2	0.2	20%	100%	0%	100%	0%	100%	0%	95%
`Categorical_4:3`	0.4	0.4	20%	100%	0%	100%	0%	100%	0%	95%
`Categorical_4:4`	0.8	0.8	20%	100%	0%	100%	0%	100%	0%	95%
`Categorical_5:2-4`	0	0	0%	0%	0%	0%	0%	5%	0%	0%
`Categorical_6:2-5`	0	0	0%	5%	0%	0%	0%	0%	0%	0%
π-Deviation	0		0.044		0.033		0.056		0.037	

same behaviour applies for the CUP model. The selection of the β-parameters, which are linked to the mixture weights, seem to be more difficult for both selection techniques than the selection of the γ-parameters. The `Continous_1` and `Continous_4` are characterized by nearly the same effect size (0.3 and 0.2), but differ very much in their selection frequency. While `Continous_1` was selected for γ in 100% correctly, the covariate `Continous_4` was only selected in 60% at the most correctly for the β-parameter. Similar consequences can be drawn from the covariate `Categorical_4`. The covariate was selected in almost 100% of the cases for γ, but very rarely for β.

Table 2 summarizes the results of Table 1 by investigating how often effects that are zero and effects that are different from zero are detected correctly by the two selection methods. The forward selection technique never selected covariates with a true effect of zero while the penalization approach shows small false positive rates. However, the penalization approach performs distinctly better in detecting variables that have a non-zero effect. Both methods show lower rates in detecting effects for β than for γ.

The computed π-deviations displayed in Table 1 are very small for both selection methods given the average size of the simulated $\pi = 0.8011$ and that both selection methods are not always able to select all covariates correctly. Figure 1 displays the original deviations for all samples resulting in $60,000$ observations in each boxplot. Most of them are very close to zero. The penalty approach shows

TABLE 2: *Summary of simulated data*

Type	Parameters	Selected CUB		Selected CUP	
		Penalize	Forward	Penalize	Forward
Zero effects	β	1%	0%	1%	0%
	γ	1%	0%	10%	0%
Non-zero effects	β	68%	55%	57%	51%
	γ	99%	94%	99%	87%

higher variability and forward selection seems to yield lower discrepancies than the penalization approach.[1]

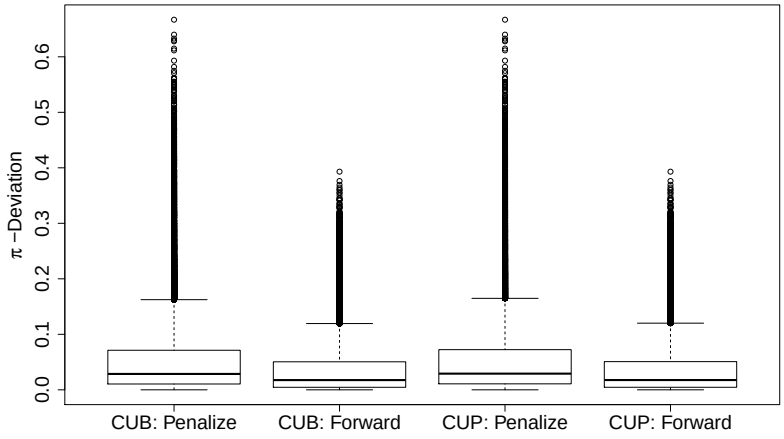


FIGURE 1: *Simulation: Boxplots of π-Deviations for the different selection methods.*

6 Applications

6.1 Life Well-Being in the Survey on Household Income and Wealth

In the following, the methods are applied to the data from the Survey on Household Income and Wealth (SHIW) by the Bank of Italy, which are earlier used by Gambacorta and Iannario (2013). The data set consists of 3816 respondents from the wave of 2010. The response is the happiness index indicating the overall life well-being measured on a Likert Scale from 1 (very unhappy) to 10 (very happy). 25 covariates as, for example, age, marital status, area of living and educational degree are included in the model selection.

[1]Note that the π_{ij}-differences of the penalization approach based on the penalized estimates.

First we describe the used penalization approach and then the forward selection. Then both techniques are compared and some parameter interpretations are given.

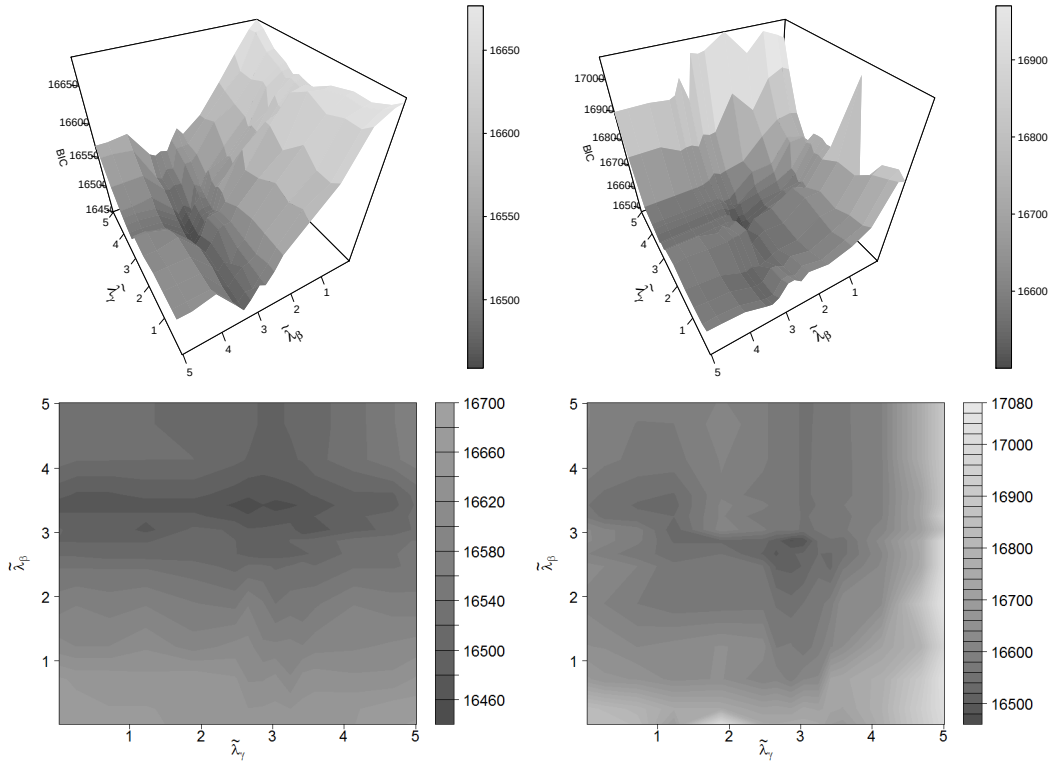


FIGURE 2: *SHIW: Grid of lambda values to find the best model for CUB (left) and CUP model (right).*

6.1.1 Penalization

To illustrate the proposed penalization we use both the CUB and the CUP-model. A 15 times 15 grid of λ_β and λ_γ-values is used to find the best combination of the tuning parameters regarding to the lowest BIC-value. The tuning parameters are transformed by $\log(\lambda + 1)$ because they were created on a logarithm scale and to avoid very large negative values when λ-values are close to zero. Figure 2 shows the results of the 225 models each for the CUB-model on the left hand side and for the CUP-model on the right. If both tuning parameters are zero (right corner) an unpenalized model is estimated. In this case all available covariates are included. On the opposite corner (left) the model is close to an intercept model.

In this application the BIC-surfaces for the two models are quite different. In the CUB-model the choice of the tuning parameter for the $\boldsymbol{\beta}$-covariates seems to be more important than the choice of the preference covariates. So it is advisable

to use a smaller grid to find the λ_β-value than the λ_γ-value. In the CUP-model both dimensions of the tuning parameters seem to be more equally important.

The lowest BIC value was found at 16450 with $\log(\lambda_\beta+1) \approx 3.42$ and $\log(\lambda_\gamma+1) \approx 2.66$ in the CUB-model and at 16478 with $\log(\lambda_\beta+1) = \log(\lambda_\gamma+1) \approx 2.86$ for the CUP-model. The tuning parameters are not the same but are found in a similar region. Choosing only identical λ-values leads to a slightly worse BIC of 16462 in the CUB-model but with the same selected variables. This is consistent with the nature of lasso regularization which not only selects covariables but also shrinks variables towards zero. It is not unusual that new variables do not enter the model at every grid point in both model components. In general a grid of several tuning parameters should be used, but in this application the restriction on $\lambda_\beta = \lambda_\gamma$ would be sufficient.

To get a better understanding of the mechanism of the variable selection we cut Figure 2 into slices and look at the development of both coefficient sets β and γ. Because of the two-dimensional grid one dimension is fixed to the selected λ-value and the other varies from high penalty (5.02) to low penalty (1.89). The lower the penalty the more parameters enter the model. Each line type in the coefficient paths stands for one parameter group. Because of the penalty term there are some parameters which are selected in both parameter set as for example marital status or area of living and others which are only selected in one of the two sets.

Figures 3 and 4 display the results for the CUB- and CUP-model, respectively. In the first and second row the development of the γ- and β-parameter are displayed. In the third row the resulting boxplots of the weights π are shown. The weights are calculated by using the individual characteristics and estimated β-coefficients. In the first column λ_γ is fixed to the best λ_γ-value and λ_β varies. So the effect of penalization of the β-parameters specifying the weights are shown for β, γ and the weights. In the second column λ_β is fixed and λ_γ varies so that the penalty for the parameters determining the weights do not change.

In the CUB-model two different λ-values are found at $\log(\lambda_\beta+1) \approx 3.42$ and $\log(\lambda_\gamma+1) \approx 2.66$ to receive the lowest BIC value. On the left column in Figure 3 the λ_γ-parameter is fixed at 2.66 and the penalty for the β varies.

Looking at the β-coefficients in the left column shows that at 5.02 no covariates are selected and the model for the weights only consists of the intercept. The π_i-values are 0.534 for all observations because no individual covariable is present. By adding covariables to the model the weights π_i are adjusted by the individual characteristics of persons and change individually. However the median of the distribution stays almost the same. The more covariables enter the model the variance increase so that the discriminatory power increase, too. But as we can see from Figure 2 using much variables in the β-part increase the BIC-values so that in this case the better discriminatory power does not compensate the higher number of variables. The best trade off between number of variables and model fit according to BIC is found at 3.42. While the β-coefficients are changing the

γ-parameters, displayed in the upper left corner, stay nearly constant.

When λ_β is fixed at 3.42 and only the penalty for the structure component λ_γ changes, as displayed in the right column, the graphs are swapped. Now the coefficients for β are nearly constant while more and more γ-coefficients enter the model. The weights are almost constant. Note that at the maximum of λ_γ already parameters are non-zero. In contrary to a flexible λ_β there is not a pure intercept model for $log(\lambda_\gamma + 1) = 5.02$.

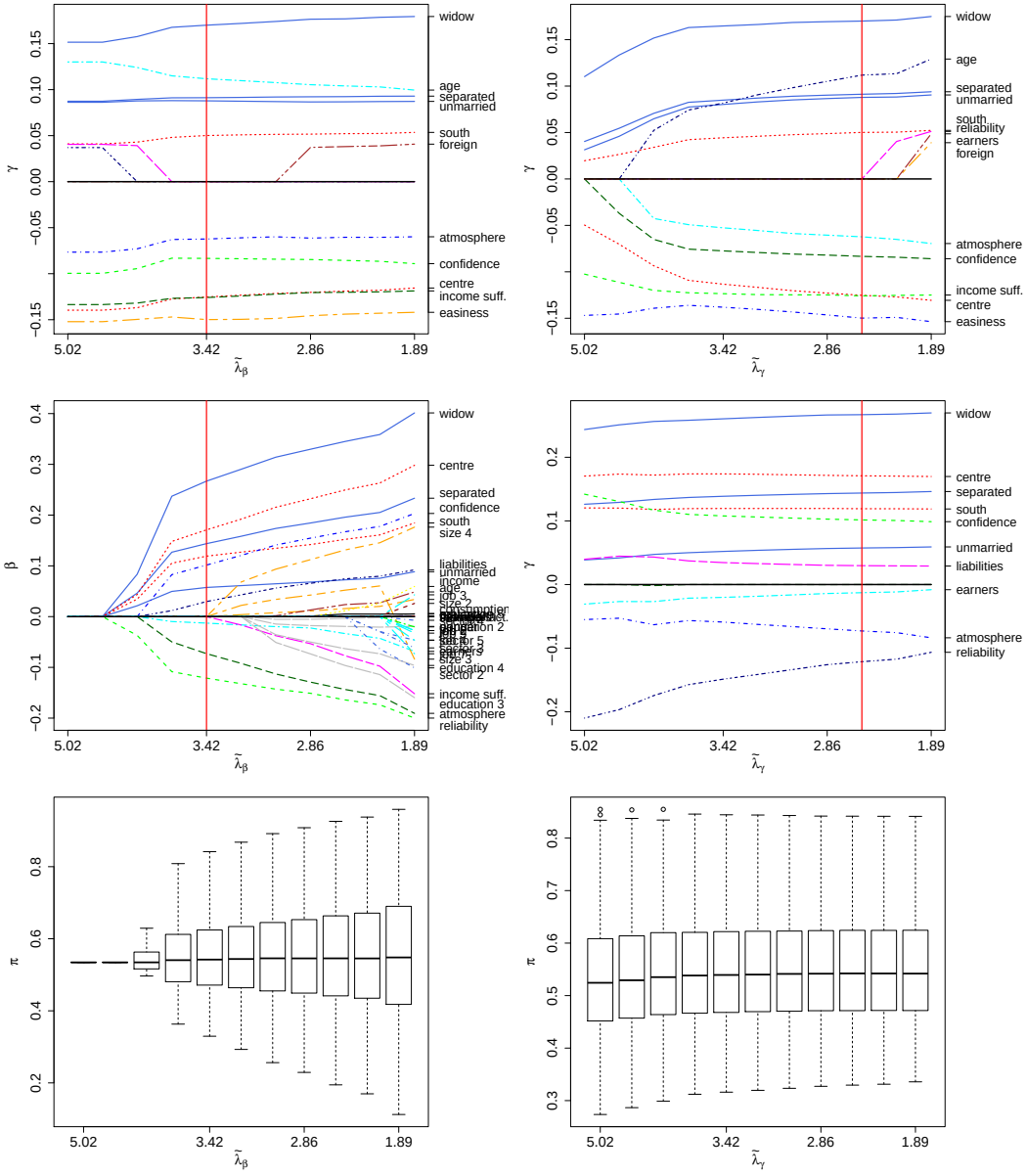


FIGURE 3: *SHIW: Standardized coefficient paths of β and γ and π for fixed lambda (left) and fixed c.lambda (right) in the CUB model.*

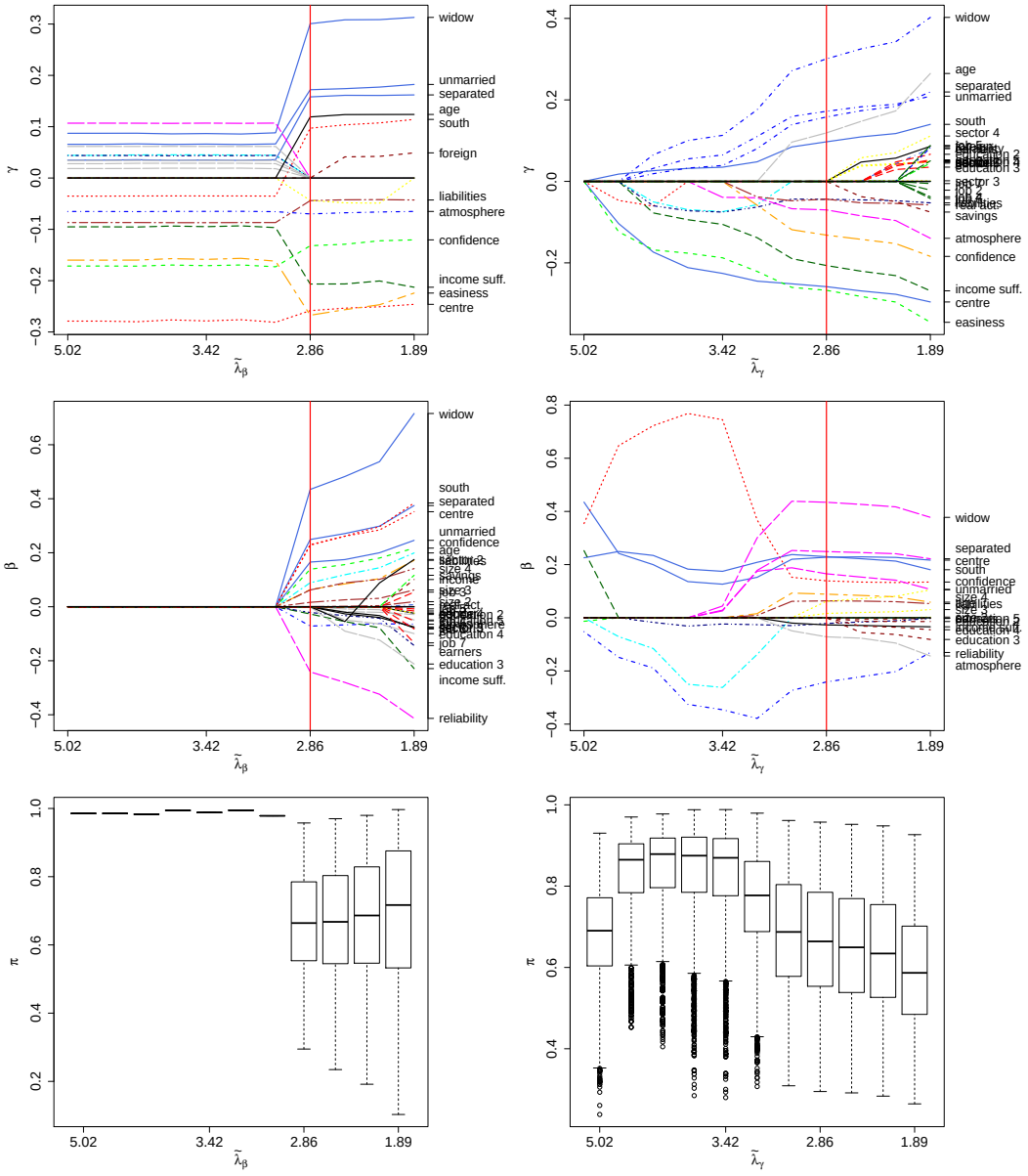


FIGURE 4: *SHIW: Standardized coefficient paths of β and γ and π for fixed lambda (left) and fixed c.lambda (right) in the CUP model.*

The behavior in the CUP model is different. The left column of Figure 4 shows the results for λ_γ fixed at 2.86 and a flexible λ_β-parameter. The first time β-Parameter entering the model is much later than in the CUB models. Until 2.86 a pure intercept model is fitted where nearly no uncertainty component is present, because the π-values are close to 1. At 2.86 some parameters are non-zero and the marginal median weight declines to 0.664. Then again the variance enlarge with more covariates but the marginal median does not change much. At 2.86 the coefficients for γ also change even if the penalty is not changed for γ. That's may be the result of the very different weights which are used for the structured component. Before and after this cutpoint the coefficients of γ are nearly constant.

The results for a fix λ_β at 2.86 is displayed in the right column of Figure 4. With less penalty more and more γ-coefficients enter the model. Even though λ_β does not change, the β-coefficients are not constant and consequently also the weights π_i change substantially.
Both the CUB- and CUP-model detect a reasonable combination of parameter and the CUB-model seem to be more stable than the CUP-model in this application.

6.1.2 Forward Selection

Using forward selection no choice of tuning parameters is necessary. Figure 5 displays the forward selection process for the CUB (left) and CUP-model (right). The y-axis shows the value of the used criteria and the x-axis the selected variables. In the case of the likelihood-ratio test we display the estimated deviance as well as the corresponding p-values. The selected variable is the result of estimating several models and choosing the variable with the greatest impact at that stage of the selection process. The last covariate on the x-axis on the right is the first one which is not selected and where the algorithm stopped. At the beginning the reduction is mostly the highest. The criteria seem to have a great impact on how and which variables are chosen. In the CUB model on the left hand side referring to BIC results in a sparer model than using the likelihood-ratio test or deviance. Not only the number of variables but also the order of selected variables are different. The model constructed by the deviance includes also all variables from the smaller model selected by the BIC. In the CUP case the deviance criterion surprisingly results in a sparer model than using the BIC criterion. However, there are some variables which are only included in one of the models. For example gender and income is only selected in the model with the deviance criterion. The selection process between the CUP and CUB model seems to be also different. Some covariates are selected in both models by both criteria and some are only available in a certain model.

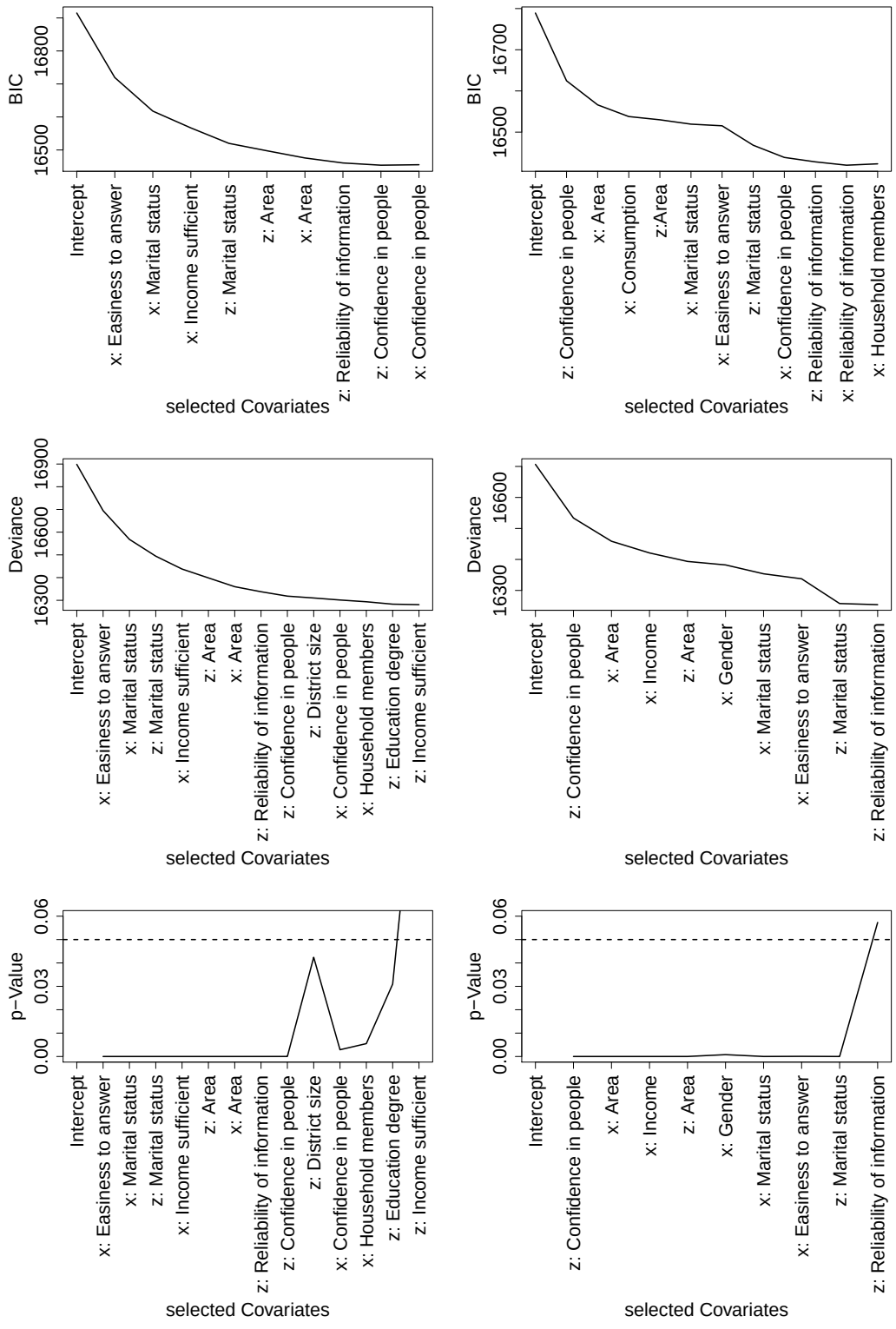


FIGURE 5: *SHIW: Forward selection for the CUB (left) and CUP model (right).*

6.1.3 Comparison of the Selection Approaches

Table 3 compares both selection methods concerning different selection criteria. For each criterion the value and (effective) degrees of freedom are given. The first entry 16450 is the BIC value which results of a variable selection via the penalization approach for the CUB model with the BIC as optimization criterion followed by the effective degrees of freedom. The next entry 16288 is the AIC value of the same selection technique but optimized according to AIC. Thus each column represents a different model search. In five of the six settings the penalization approach reach a lower value of the selection criteria than the forward selection. In all cases the penalization methods selects larger models than the forward selection.

TABLE 3: *SHIW: Comparison of selection methods*

model	method	criteria					
		BIC		AIC		Deviance	
		value	(e)df	value	(e)df	value	(e)df
CUB	penalize	**16450**	21.33	**16288**	42.46	**16192**	58.99
	forward	16453	16	16335	21	16283	25
CUP	penalize	16479	35.41	**16178**	64.27	**16038**	78.44
	forward	**16420**	26	16389	24	16257	24

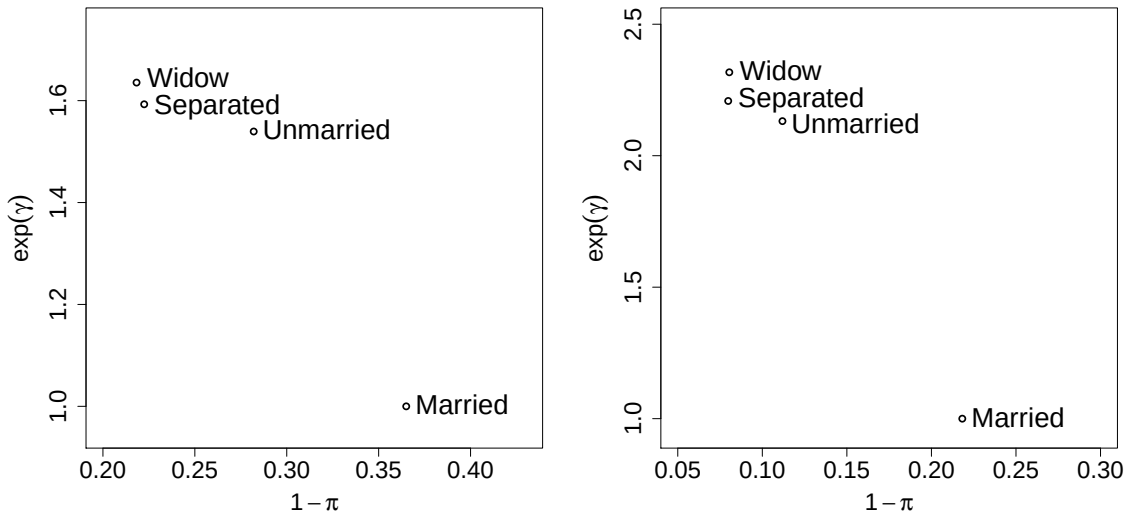


FIGURE 6: *SHIW: Effects of the categorical covariates marital status in CUB- (left) and CUP-Modell (right) in the structure and uncertainty component.*

6.1.4 Parameter Interpretation

For illustration we use both models and selection techniques optimized according to BIC. Using the penalization approach we refitted the models to avoid shrinked coefficients. Note that in this case the goodness-of-fit measurements may be slightly changed, too. Table 4 shows the result for the CUB model and Table 5 for the CUP model. As already mentioned the number of variables are smaller using forward selection than the penalization approach in both models. The effect sizes are similar and show always the same direction. In both components the CUP-model select more variables than the CUB-model.

Figure 6 illustrates the effects of marital status in the CUP- and CUB-model. It is not possible to compare the values of the γ-parameter directly because the models are too different. But for both models the marital status "widow" corresponds to high values of unhappiness and high certainty (small $1 - \pi$). In contrast, the status "married" indicates happiness but a large amount of uncertainty in the response. The order of the marital categories is almost the same in CUB- and CUP-model, but the connected uncertainty is for them higher in the CUB-model than in the CUP-model. This is consistent with the overall behavior of the CUB-model predicting a higher uncertainty than the CUP-model.

TABLE 4: *SHIW: Coefficients of the chosen (refitted) CUB model*

	Refitted Penalized model		Forward Selection	
Covariates	Concomitant(β)	Structure(γ)	Concomitant(β)	Structure(γ)
Constant	0.554	0.734	0.538	0.586
Marital status: Unmarried	0.381	0.431	0.489	0.368
Marital status: Separated	0.698	0.466	0.834	0.400
Marital status: Widow	0.722	0.492	1.174	0.560
Area: Centre of Italy	0.528	-0.259	0.936	-0.255
Area: South of Italy	0.273	0.100	0.412	0.071
Confidence in people	0.042	-0.042	0.093	
Interview atmosphere	-0.050	-0.038		
Income sufficient		-0.113		-0.126
Age (centered)		0.005		
Easiness to answer		-0.088		-0.133
Income earners	-0.018			
Reliability of information	-0.073		-0.193	
Financial liabilities	0.004			

TABLE 5: *SHIW: Coefficients of the chosen (refitted) CUP model*

	Refitted Penalized model		Forward Selection	
Covariates	Concomitant(β)	Structure(γ)	Concomitant(β)	Structure(γ)
Constant	1.276		0.682	
Marital status: Unmarried	0.796	0.757	0.801	0.402
Marital status: Separated	1.169	0.792	1.088	0.361
Marital status: Widow	1.160	0.840	1.429	0.649
Area: Centre of Italy	0.783	-0.514	0.930	-0.680
Area: South of Italy	0.514	0.230	0.633	0.180
Confidence in people	0.055	-0.056	0.128	-0.208
Interview atmosphere	-0.051	-0.048		
Income sufficient	-0.023	-0.170		
Age (centered)	0.006	0.008		
Easiness to answer		-0.167		-0.282
Income earners	-0.029			
Reliability of information	-0.139		-0.211	0.046
Financial liabilities	0.040	-0.012		
Foreign		0.203		
Real activity		-0.002		
District size Cat2	0.045			
District size Cat3	0.027			
District size Cat4	0.271			
Family Consumption				-0.009

6.2 Enrichment of Cultural Life by Foreigners in the German General Social Survey

The German General Social Survey (ALLBUS) provided by the GESIS-Leibniz-Institut für Sozialwissenschaften (2017) collects data on behavior, attitudes and social structure in Germany. In 2016 a big focus was on attitudes towards migrants, foreigners and religious groups. The 3490 participants were asked to rate on a 7-point scale whether foreigners enrich the German cultural life from "Completely disagree" (1) to "Completely agree" (7). The data set consist of over 700 possible variables. We restricted ourself to the 43 most meaningful variables which would still result in over 200 parameters for the complete model because of a large number of categorical variables and the two parameter sets $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

We applied both proposed methods. For the penalization we used a 19 times 19 grid of λ_β and λ_γ-values to deduce the best combination of the tuning parameters regarding to the lowest BIC-value. The result of this procedure is displayed in Figure 7. White areas in the contour plots correspond with higher BIC-values than being able to be displayed in this figure. In this application the surface of

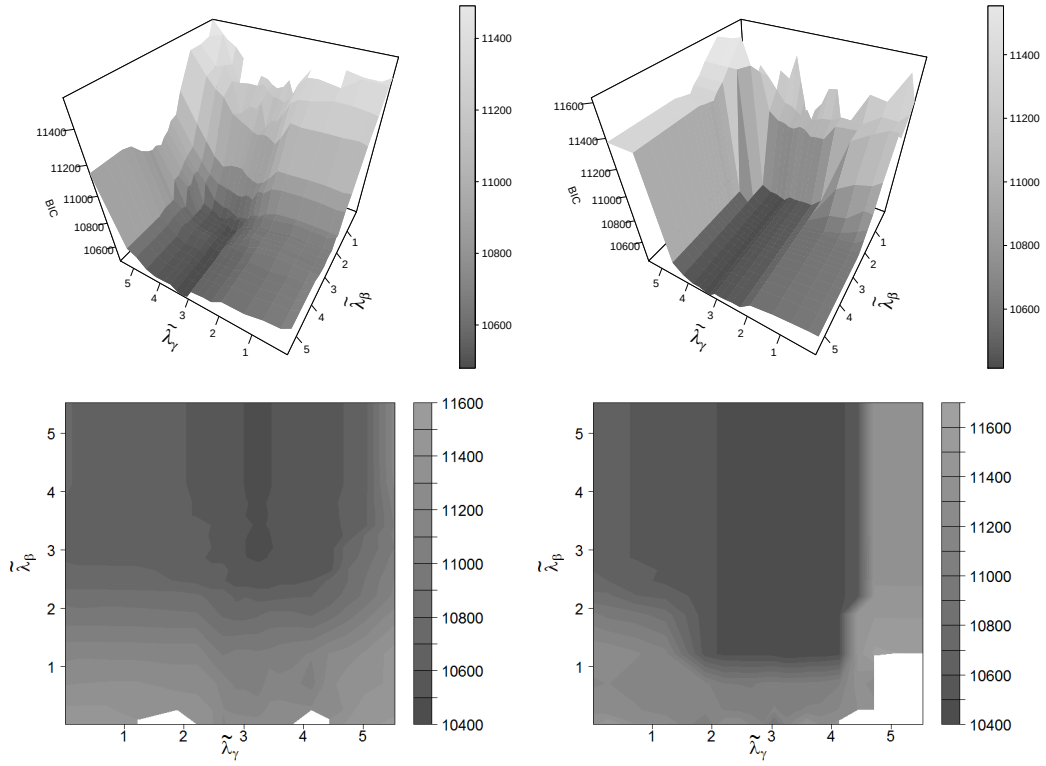


FIGURE 7: *ALLBUS: Grid of lambda values to find the best model for CUB (left) and CUP model (right).*

both models are quite similar. But in the CUP-model the transition from low to higher BIC-values is sharper than in the CUB-model even though in both models the same grid is used. In the CUB-model the lowest BIC was found at 10471 for $\log(\lambda_\beta + 1) \approx 5.02$ and $\log(\lambda_\gamma + 1) \approx 3.245$. The CUP-model detected the lowest BIC-value at 10408 with $\log(\lambda_\beta + 1) \approx 3.25$ and $\log(\lambda_\gamma + 1) \approx 3.42$. In both the CUB and the CUP models no covariables are selected in the $\boldsymbol{\beta}$-component. This results in a pure intercept model for the weights which are constant for all individuals. The mean mixture weight $(1 - \bar{\pi})$ is 0.0004 for the CUP-model and 0.33 for the CUB-model. If there are no covariables in $\boldsymbol{\beta}$ selected, the intercepts of the cumulative model γ_{0r} in the CUP-model seem to be able to capture the constant probability of the uniform distribution for all individuals resulting in a mixture weight for the uncertainty component close to zero. Moreover the BIC is lower than in the CUB-model with a much higher weight for the uncertainty component.

Using the forward selection leads to models with covariates in both mixture components. Figure 8 displays the selection process for the CUB and CUP model, respectively. Furthermore the selected covariates are quite different between the CUB and the CUP model. In the first case "foreign literature", "age", "household income" and "party membership" are selected for $\boldsymbol{z}$ whereas in the CUP

model only "age" and "eastwest" were chosen which also results in quite different mixture weights π. The questions of the selected covariates can be found in the appendix.

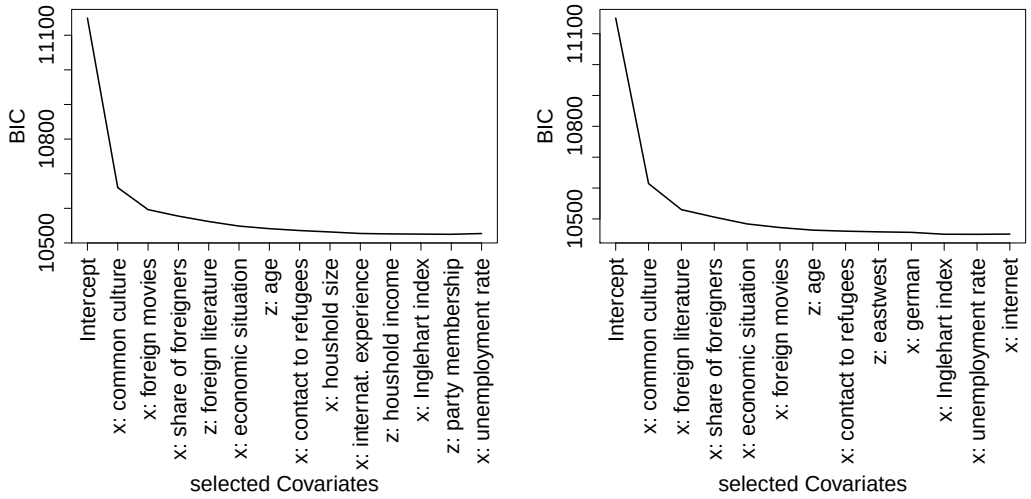


FIGURE 8: *Allbus: Forward selection for the CUB (left) and CUP model (right).*

Table 6 summarizes the results for this application. For both models the BIC value is smaller when using the penalization approach than the forward selection. Also both selection techniques differ not very much in the estimated average mixture weight $\bar{\pi}$ especially in the CUB model, the models are quite different. Using penalization results in larger models but without β effects except of the intercept. On the other hand the selected β-coefficients using forward selection seem to have not enough impact to reduce the BIC in an reasonable way. The lowest BIC value was detected for the penalized CUP model with mixture weight of 0.9996 which is almost a pure cumulative model without uncertainty component.

TABLE 6: *Allbus: Comparison of selection methods*

model	method	BIC	No β	No γ	$\bar{\pi}$
CUB	penalize	**10470**	0	26	0.6747
	forward	10524	4	15	0.6273
CUP	penalize	**10408**	0	27	0.9996
	forward	10450	2	17	0.8742

This application shows that the penalization approach leads here to lower BIC

values as the forward selection and stable results even if no β-effects are selected. Furthermore the best combination of tuning parameters is quite different from the previous application so that the best tuning parameter combination has to be estimated for each application separately.

7 Concluding Remarks

We have shown how to adapt the group lasso framework for mixture models with an uncertainty component and compared it to the forward selection. As demonstrated in the simulation section both methods show good performance in selecting the true covariates. The methods allow to decide which variables should be included in the uncertainty part of the model and/or in the preference part of the model. Since often covariates are only included in one of the model components, the model complexity can be reduced substantially. Although forward selection often yields sparser models variable selection via stepwise procedures has some drawbacks. The procedure is rather variable and time-consuming when the number of covariates increases, and often yields higher goodness-of-fit measurements than the penalization approach. Penalization is more flexible and can be used in very high dimensional settings.

It is seen from the applications to real data problems that the choice of the selection method and the optimization criterion determine which final model is chosen. In the Survey on Household Income and Wealth some variables as "marital status" and "area of living" were always selected. Regularization methods yield information on the importance of covariates by visualization of coefficient paths. Also nonparametric bootstrap samples might be a possibility to evaluate how often a covariate is selected. However, including the search for the best tuning-parameter combination without restrictions will lead to huge computing time. One possibility to save computing time would be the restriction on the tuning parameters to be equal. In the first application this restriction would have been sufficient. However, further research is necessary to derive a general rule.

Acknowledgements: Thanks to Maria Iannario and Domenico Piccolo who provided the SHIW data (STAR Programme at University of Naples Federico II).

References

Agresti, A. (2013). *Categorical Data Analysis, 3d Edition*. New York: Wiley.

Baudry, J.-P. and G. Celeux (2015). Em for mixtures - initialization requires special care. https://hal.inria.fr/hal-01113242.

Beck, A. and M. Teboulle (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences 2*(1), 183–202.

Bischl, B., J. Richter, J. Bossek, D. Horn, J. Thomas, and M. Lang (2017). mlrmbo: A modular framework for model-based optimization of expensive black-box functions. *ArXiv e-prints 1703.03373*.

Breiman, L. (1996). Heuristics of instability and stabilisation in model selection. *Annals of Statistics 24*, 2350–2383.

D'Elia, A. and D. Piccolo (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis 49*, 917–934.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B 39*, 1–38.

Gambacorta, R. and M. Iannario (2013). Measuring job satisfaction with CUB models. *Labour 27*(2), 198–224.

GESIS-Leibniz-Institut für Sozialwissenschaften (2017). *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2016*, Volume 2.1.0. GESIS Datenarchiv Köln.

Iannario, M. and D. Piccolo (2012a). CUB models: Statistical methods and empirical evidence. In R. Kennett and S. Salini (Eds.), *Modern Analysis of Customer Surveys: with applications using R*, pp. 231–258. New York: Wiley.

Iannario, M. and D. Piccolo (2012b). Investigating and modelling the perception of economic security in the survey of household income and wealth. In M. S. C. Perna (Ed.), *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, pp. 237–244. Berlin: Springer.

Karlis, D. and E. Xekalaki (2003). Choosing initial values for em algorithm for finite mixtures. *Computational Statistics and Data Analysis 41*, 577–590.

Khalili, A. and J. Chen (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association 102*(479), 1025–1026.

Luo, R., H. Wang, and C. Tsai (2008). On mixture regression shrinkage and selection via the mr-lasso. *International Journal of Pure and Applied Mathematics 46*, 403–414.

Piccolo, D. and A. D'Elia (2008). A new approach for modelling consumers' preferences. *Food Quality and Preference 19*, 247–259.

Piccolo, D. and R. Simone (2019). The class of cub models: statistical foundations, inferential issues and empirical evidence. *Statistical Methods & Applications. online published*.

Pößnecker, W. (2019). MRSP: Multinomial response models with structured penalties. R package version 0.6.11, https://github.com/WolfgangPoessnecker/MRSP.

Städler, N., P. Bühlmann, and S. van de Geer (2010). L1-penalization for mixture regression models. *Test 19*, 209–256.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B 58*, 267–288.

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge: Cambridge University Press.

Tutz, G., W. Pößnecker, and L. Uhlmann (2015). Variable selection in general multinomial logit models. *Computational Statistics & Data Analysis 82*, 207 – 222.

Tutz, G. and M. Schneider (2019). Flexible uncertainty in mixture models for ordinal responses. *Journal of Applied Statistics 46*(9), 1582–1601.

Tutz, G., M. Schneider, M. Iannario, and D. Piccolo (2017). Mixture models for ordinal responses to account for uncertainty of choice. *Advances in Data Analysis and Classification 11*, 281–305.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B 68*, 49–67.

A Appendix

Variable description of some selected covariates of the ALLBUS data:

- `Common culture`: It is better for a country, if all persons belong to a common culture?
  "completely agree", "rather agree", "rather disagree", "completely disagree"

- `Economic situation`: How do you evaluate the current economic situation in Germany?
  "very good", "good","partly good/ partly bad", "bad", "very bad"

- `Foreign literature`: Do you read - at least occasionally - newspapers, magazines or books in a foreign language?
  "yes", "no"

- `Foreign movies`: Do you watch - at least occasionally - television broadcast or movies in a foreign language without subtitles?
  "yes", "no"

- `Contact to refugees`: Have you had direct personal contact with refugees?
  "yes", "no"

- `Internat. experience`: Have you stayed during your life for more than three months in a foreign country?
  "yes", "no"

- `Internet`: ... Do you use at least occasionally the internet for private purposes?
  "yes", "no"

- `Household size`: ... Do other persons than you live in this household?
  "yes", "no, I live alone"

- `Party membership`: ... Are you member of a political party?
  "yes", "no"

- `German`: German citizen
  "yes, only", "yes, too","no"

- `Eastwest`: Living region
  "Old Federal states", "Newly-formed German states"

- `Inglehart Index`: Computed from several questions:
  "postmaterialist", "postmaterialist mix", "materialist mix", "materialist"

- `Age`: Age of the respondent

- `Household income`: Equivalised disposable income

- `Share of foreigners`: Share of foreigners in living region

- `Unemployment rate`: Unemployment rate in living region

A.6. Dealing with Heterogeneity in Discrete Survival Analysis using the Cure Model

Schneider, M. (2019): Dealing with Heterogeneity in Discrete Survival Analysis using the Cure Model. *Technical Report 224*, Department of Statistics, Ludwig-Maximilians-Universität München, doi:10.5282/ubm/epub.68455.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Micha Schneider

Dealing with Heterogeneity in Discrete Survival Analysis using the Cure Model

Technical Report Number 224, 2019
Department of Statistics
University of Munich

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

Dealing with Heterogeneity in Discrete Survival Analysis using the Cure Model

Micha Schneider
Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München

August 6, 2019

Abstract

Cure models are able to model heterogeneity which arises from two subgroups with different hazards. One subgroup is characterized as long-term survivors with a hazard equal to zero, while the other subgroup is at-risk of the event. While cure models for continuous time are well established, cure models for discrete time points are rarely prevalent. In this article I describe discrete cure models, how they are defined, estimated and can be applied to real data. I propose to use penalization techniques to stabilize the model estimation, to smooth the baseline and to perform variable selection. The methods are illustrated on data about criminal recidivism and applied to data about breast cancer. As one result patients with no positive lymph nodes, a very small tumor, which can be well differentiated from healthy cells and with ethnicity which is neither black or white have the best estimated chances to belong to the long-term survivors of breast cancer.

Keywords: Cure Model, Discrete, Survival Analysis, Variable Selection, lasso

1 Introduction

In traditional survival analysis it is assumed that all analyzed subjects may be affected by the event of interest at sometime. Thus all subjects are at-risk of that event. But it happens frequently that a certain subgroup of the population never experience the event of interest. This subjects are called "cured", "long-term survivors" (LTS) or "not-at-risk".

Traditional examples can be found in clinical studies where some patients are long-term survivors of a severe disease as cancer and never suffer from the

recurrence of it. In the social sciences one could be interested in analyzing the recurrence of released prisoners (see Rossi et al., 1980). Some of the released prisoners will be arrested again and others never do. Another example can be found in the educational sphere. Some students may be never able to solve a certain task, because it is too difficult for them, while others can solve the problem.

While cure models for continuous time are widely used and described for example by Amico and Keilegom (2018), Sy and Taylor (2000), Kuk and Chen (1992) and Maller and Zhou (1996), cure models for discrete time points are rarely prevalent. Tutz and Schmid (2016) give an overview about discrete time modelling and Muthén and Masyn (2005) about discrete-time survival mixtures. Actually in a lot of settings the time is not measured in continuous time but in discrete time points. In most cases a study ask their participants at fixed time points as months or years if they are still cured by the disease or still not in jail. If it is a retrospective study, the respondents may have also difficulties in remembering the exact time, but give an approximated response. Furthermore discrete survival analysis has the advantage that the interpretation may be easier since the hazard can be interpreted as probability and time depended variables can be introduced quite easily. The model used in this article is not designed for re-occurrence of an event (see Willett and Singer, 1995) or competing events (see Tutz and Schmid, 2016).

In this article I describe discrete cure models, how they are defined, estimated and how variable selection and smoothing can be performed. Thus we get a very flexible and easy-to-interpret tool for understanding complex discrete survival data situations. The discrete cure model has been considered by Tutz and Schmid (2016). Steele (2003) also applied a discrete-time mixture model with long-term survivors, but uses a different estimation method.

The article is organized as follows: First the discrete cure model is described and an overview of the discrete data structure is given. Then the model is illustrated by an application about criminal recidivism (Section 4). In Section 5 variable selection with an adopted version of lasso is proposed, followed by the description of the estimation of the (penalized) discrete cure model. In Section 7 the proposed selection technique is used to improve the model for criminal recidivism, followed by a further application about breast cancer (in Section 8). After some comments to the identifiability of discrete cure models the article is concluded.

2 The Discrete Cure Model

The cure model is defined as a finite mixture of survival functions. Typically it consists of two latent classes: One sub-population at risk and one sub-population characterized as long-term survivors or “cured”. The survival function of the

cured remains at 1 whereas the survival function of the non-cured population decrease over time t so that the observed survival function of the cure model is defined as

$$S(t|\boldsymbol{x}) = \pi(\boldsymbol{z})S_1(t|\boldsymbol{x}) + (1 - \pi(\boldsymbol{z})) \cdot 1, \tag{1}$$

where $\pi(\boldsymbol{z})$ is the weight for the non-cured population determining for each observation the probability belonging to this group. The weights can be calculated using individual specific covariates $\boldsymbol{z}$ by

$$\pi(\boldsymbol{z}) = \frac{\exp(\boldsymbol{z}^T\boldsymbol{\beta})}{1 + \exp(\boldsymbol{z}^T\boldsymbol{\beta})}.$$

The discrete survival function is the probability that the event has not been occurred at time point t:

$$S(t|\boldsymbol{x}) = P(T > t|\boldsymbol{x}) = \prod_{s=1}^{t}(1 - \lambda(s|\boldsymbol{x})),$$

which can be expressed by the discrete hazard $\lambda(t|\boldsymbol{x})$. It is defined as the probability that an event occurs at time T, given that time T is reached conditional on some covariables $\boldsymbol{x}$:

$$\begin{aligned}\lambda(t|\boldsymbol{x}) = P(T = t|T \geq t, \boldsymbol{x}) &= h(\gamma_{0t} + \boldsymbol{x}^T\boldsymbol{\gamma}) \\ &= \frac{\exp(\gamma_{0t} + \boldsymbol{x}^T\boldsymbol{\gamma})}{1 + \exp(\gamma_{0t} + \boldsymbol{x}^T\boldsymbol{\gamma})}, \quad t = 1, \ldots, t^*.\end{aligned}$$

γ_{0t} is the parameter of the so called baseline hazard. The logistic distribution function $h() = \exp()/(1 + \exp())$ leads to the logistic discrete hazard model. However, one may also choose other link functions as the clog-log link to obtain the group proportional hazard model (see Tutz and Schmid, 2016).

There are two covariable sets $\boldsymbol{x}$ and $\boldsymbol{z}$ in the cure model. They can be identical, overlap or completely different. But they have very different functions, $\boldsymbol{x}^T\boldsymbol{\gamma}$ is used to estimate the survival function of the non-cured population so that this predictor influence the probability of an event in the non-cured population. On the other hand $\boldsymbol{z}^T\boldsymbol{\beta}$ determine the probability of being cured or not. In Section 5 I propose variable selection via penalization to decide which variables should be included in which part of the model.

3 Data Structure in Discrete Survival Analysis

In discrete survival analysis a certain data structure is usually very helpful. Let y_{is} be an indicator of the occurrence of an event so that

$$y_{is} = \begin{cases} 1, \text{if individual fails at time } s \\ 0, \text{if individual survives time } s \end{cases}$$

Thus each observation i generates a specific vector $(y_{i1}, \ldots, y_{it_i})$ with the entries 0 or 1 and the length t_i. For a non-censored observation the vector has the form $(0, \ldots, 0, 1)$ because at time t_i the event occurs. Censored observations can be individuals who drop out during the study without observing an event or the study concludes when some participants have not experienced an event yet. For the censored observations the vector contains only zeros until the individual is censored: $(0, \ldots, 0)$. The length t_i is variable and depends on how long each individual is observed. If the person drops out of the study in the first time interval the length of y_{is} is one. Table 1 illustrates the data structure for $T = 3$ time points and three individuals i. The first individual is observed for all three time points and experience the event at time point 3. Consequently, y_i has the form $(0, 0, 1)$ with $t_i = 3$. Each row contains the information about one specific person at one specific time point. Thus observations have as many rows as observed time points. The second observation $i = 2$ drops out of the study after two time points. Thus, there are only two rows for observation 2 and $y_i = (0, 0)$, because no event take place. Since x_{i1} is a time-constant variable the value is the same for one person and different time points[1].

i	y_i	$t = 1$	$t = 2$	$t = 3$	$x_{i1} = \text{Age}$	t_i
1	0	1	0	0	20	$t_1 = 3$
1	0	0	1	0	20	
1	1	0	0	1	20	
2	0	1	0	0	30	$t_2 = 2$
2	0	0	1	0	30	
3	0	1	0	0	55	$t_3 = 1$

TABLE 1: *Example for data structure in long format*

In Section 6.1 it will be shown that the likelihood by using y_{is} is equivalent to the likelihood of a binary response model with observations y_{is}.

To include time-varying covariables for the population under risk in the discrete cure model we just have to add a new column x_{i2} to the data structure. While the value of the time-constant covariables is repeated for observation i for each row, the values of time-varying covariables can change with each row of the same observation i. In Table 2 the time-varying covariable "employment" is added by x_{i2}. If the person has a job at time t the value is one otherwise zero. For example person 1 is unemployed at time $t = 1$ and gets hired at $t = 2$. At time $t = 3$ person 1 is unemployed again.

[1]Note that this data structure may be adjusted for the need of the software which is used. For example MRSP by Pößnecker (2019) requires that y_i has always the length T and missing values are filled up with NA. In this case y_2 would be $(0, 0, NA)$ and $y_3 = (0, NA, NA)$

i	y_i	$t=1$	$t=2$	$t=3$	$x_{i1}=$ Age	$x_{i2}=$ Emp	t_i
1	0	1	0	0	20	0	$t_1=3$
1	0	0	1	0	20	1	
1	1	0	0	1	20	0	
2	0	1	0	0	30	1	$t_2=2$
2	0	0	1	0	30	1	
3	0	1	0	0	55	0	$t_3=1$

TABLE 2: *Example for data structure with time-depending covariable*

4 Illustrative Example: Criminal Recidivism

For illustration I use data about criminal recidivism, which is available in the R-package RcmdrPlugin.survival by Fox and Carvalho (2012). The data was generated within the scope of the "Transitional Aid Research Project" and described by Rossi et al. (1980). The aim of this project was to reduce the recidivism of prisoners and to examine the effect of financial aid. The data set used here consist of 432 released prisoners, who were observed during one year after release.

We know for each week if the person has been rearrested or not, which leads to 52 time points. Since there are not events at every time point, the time is reduced to 49. Half of the convicts received financial aid. Other variables are the age of the person at the time of release, the race ("black", "others"), the marital status ("married", "not married") and the level of education ("6th grade or less", "7th to 9th grade", "10th to 11th grade", "12th grade or higher"). Furthermore it was reported if the convicts worked full-time before incarceration ("no", "yes"), if they were released on parole ("no", "yes") and the number of convictions prior to the current incarceration. An overview of the available variables can be found in Table 3 and Table 4.

	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
Age (at release)	17	20	23	25	27	44
Prior convictions	0	1	2	3	4	18

TABLE 3: *Descriptive statistics of quantitative explanatory variables for the recidivism data*

First I will focus on a few important variables which are included in both parts of the model. In Section 5 we will see how this model can be further improved by using variable selection and smoothing techniques. Financial aid is one of the main variables in this setting. If financial aid has a positive effect, one can assume that it increases the probability of being cured and decreases the probability of an event. If someone has enough money for his/her basic needs, it may be less

	Category	observations	Proportions (in %)
Financial aid	No	216	50
	Yes	216	50
Race	Black	379	88
	Others	53	12
Work experience	No	185	43
	Yes	247	57
Married	Yes	53	12
	No	379	88
On parole	No	165	38
	Yes	267	62
Education	$\leq$6th	239	55
	7-9th	24	6
	10-11th	119	28
	12th+	50	12

TABLE 4: *Descriptive statistics of discrete explanatory variables for the Recidivism data*

probable that the person commits a crime. Similar applies for work experience. Someone, who has work experience, should be hired easier than someone without any work experience. So the hypothesis is that work experience reduces the probability of being arrested. In contrary the number of prior convictions may increase the probability of being non-cured and the probability of an event after release, since multiple offender may have more difficulties than first offender to change their lifestyle. Finally, age is included to account for demographic effects.

The result of the model, which includes these variables, can be found in Table 5. The standard errors are calculated by 600 bootstrap samples. Although the same variables are used for both parts of the model the meaning is completely different. The parameters in the upper part correspond with the probability that the person is part of the non-long-term survivors. If the person received financial aid the chance to be non-cured compared to be cured is reduced by the multiplicative factor $\exp(-0.2147) = 0.8068$. Thus the probability to be long-term survivor seems to be increased by financial aid. The number of prior convictions shows a positive effect so that the more prior convictions someone has committed the higher the probability of being non-cured. However, none of the estimates are statistically significant, since all confidence intervals include zero, so that the coefficients need to be interpreted with care.

In the lower part of the table the effects on the hazard function are displayed. Positive values correspond with a higher (and earlier) risk of arrest while negative values reduce the risk of recidivism. Here financial aid and prior work experience seem to coincide with a lower risk of an event. The number of prior convictions and a greater age seem to increase the probability of recidivism at any time t compared to an event later than t. Although these effects are again statistically

non-significant the results are consistent with the hypotheses.

	Estimates	BS.sd	BS.2.5	BS.97.5	
Intercept	0.1319	0.5188	-0.0753	1.4505	
Financial aid: yes	-0.2147	0.1312	-0.3167	0.1794	
Age	-0.0522	0.0240	-0.0538	0.0355	$\hat{\boldsymbol{\beta}}$
Work experience: yes	0.2426	0.2079	-0.1257	0.7782	
Number prior convictions	0.1023	0.0556	-0.0154	0.1793	
Financial aid: yes	-0.1186	0.2605	-0.8841	0.1261	
Age	0.0154	0.0362	-0.1237	0.0306	$\hat{\gamma}$
Work experience: yes	-0.9839	0.4536	-1.7538	0.1102	
Number prior convictions	0.0412	0.0444	-0.0167	0.1615	

TABLE 5: *Model 1 - Estimates for recidivism. First group of estimates indicates effects on being non-long-term survivor, second group indicates effects on the event fall-back. BS.sd, BS.2.5, BS.97.5 refer to the bootstrap standard error and the quantiles for 2.5% and 97.5%, respectively.*

Figure 1 illustrates some parameter estimates. On the left hand side the effect of "financial aid" and "work experience" is displayed in the two-dimensional space of non-cured on the y axis and risk of an event on the x axis. The stars correspond to 0.95 confidence intervals using the 2.5% and 97.5% quantiles of the bootstrap samples. At the dashed lines no effect is found, because $\exp(0) = 1$. Since each confidence intervals cover this lines, it is easy to see that none of the effects is statistically significant. However, since the effect of financial aid is in both dimensions below 1, it indicates that there might be reduction of the chances in both dimensions.

On the right hand side of Figure 1 the estimated effect of financial aid on the survival function of the cure model is displayed. In this figure the variable work experience is set to "no" and the other two variables to their median value of 23 for age and 2 for prior convictions. Thus financial aid increases the survival function and leads to a higher survived proportion at the end of the study.

The discrete cure model is a very helpful tool to gain better insights in this complex data situation and can be easily interpreted. In contrast to cure models for continuous time the hazard can be always interpreted as probability. However, there may be also some challenges. First the variable selection is an crucial point and it might be difficult to decide which variables should be included in which part of the model. Second the baseline hazard may need very much parameters and may result in a quite rough function. Furthermore time points where no event take place may cause difficulties in the estimation process since the corresponding intercept should be minus infinity. All this issues can be addressed by the proposed penalization technique in the next section.

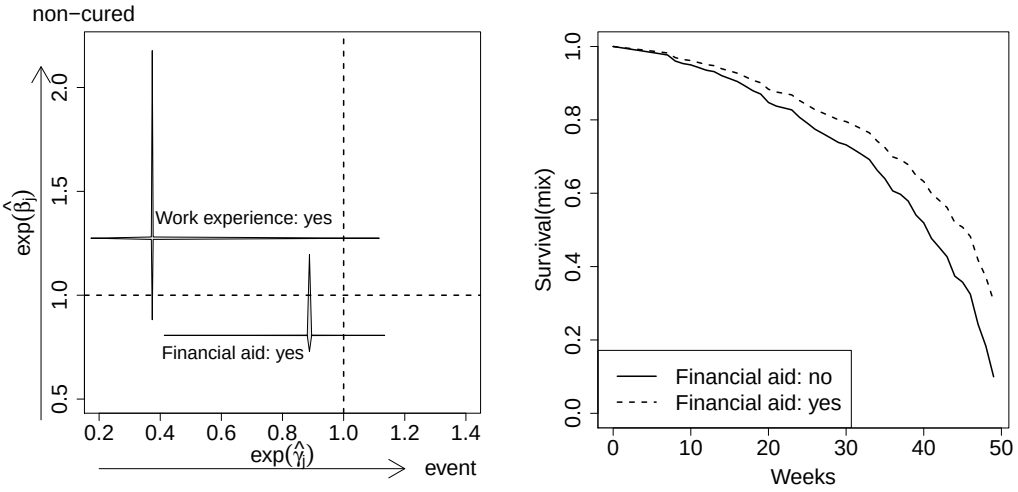


FIGURE 1: *Illustration of parameters estimates in model 1*

5 Penalization for Variable Selection and Smoothing

Penalization in discrete cure models can fulfill two main goals. First it is possible to select variables in a data driven way. Usually it is not obvious which covariates should be included in which part of the model. Using the proposed version of lasso (Tibshirani, 1996) for cure models can solve this issue. Second penalization can reduce the degrees of freedom concerning the intercepts. In discrete cure models there are intercepts for each transition from time t to $t+1$. This may result in a large number of parameters which may not be necessary, in a quite rough baseline function and in computational difficulties if no event take place. Thus it is proposed to penalize the squared distances of two neighbouring intercepts. The penalized likelihood is given by

$$l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = l(\boldsymbol{\beta}, \boldsymbol{\gamma}) - J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}),$$

where $l(\boldsymbol{\beta}, \boldsymbol{\gamma})$ denotes the unpenalized log-likelihood and $J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ a specific penalty term.

Let the vectors $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ refer to the effect of j-th variable so that $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_g)$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_h)$. The corresponding vectors $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ are partitioned into $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{ig})$ and $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ih})$ such that each components refer to a single variable. For example $\boldsymbol{x}_{ij}$ can represent for observation i all dummy variables that are linked to the j-th variable. $df_{\boldsymbol{\beta}_j}$ and $df_{\boldsymbol{\gamma}_j}$ are defined as the number of parameters collected in the corresponding parameter vector $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$, respectively. So if the j-th $\boldsymbol{x}$-variable is marital status with the 4 categories "single", "married", "divorced" and "widowed", the length of $\boldsymbol{x}_{ij}$ and the degrees of freedom $df_{\boldsymbol{\beta}_j}$ would be both 3. To ensure that the selection does not depend

on the scale of the variables, all continuous and categorical variables need to be standardized.

The proposed penalty term is given by

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda_{\boldsymbol{\beta}} \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \left\| \boldsymbol{\beta}_j \right\|_2 + \lambda_{\boldsymbol{\gamma}} \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \left\| \boldsymbol{\gamma}_j \right\|_2 \tag{2}$$

$$+ \lambda_0 \sum_{t=1;s>t}^{t^*} \left\| \gamma_{0t} - \gamma_{0s} \right\|_2^2. \tag{3}$$

It consists of three summands connected to the parameters $\boldsymbol{\beta}$ of the mixture weights, $\boldsymbol{\gamma}$ of the hazard function and $\boldsymbol{\gamma}_0 = (\gamma_{01}, \ldots, \gamma_{0(t^*)})$ of the baseline hazard. Each component posseses its own tuning parameter $\lambda_{\boldsymbol{\beta}}$, $\lambda_{\boldsymbol{\gamma}}$ and λ_0, which regulate the amount of shrinkage. $\|\|_2$ is the unsquared L_2-Norm so that the penalty enforces the selection of variables in the spirit of the group lasso (Yuan and Lin, 2006) rather than selection of single parameters. A large λ value corresponds with large shrinkage, which may also lead to more parameters set to zero. On the other hand a λ value closer to zero results in a an estimate closer to the unpenalized ML-estimate with low shrinkage and less variable selection since less parameter groups are set to zero.

The first two penalty terms are constructed to shrink and select variables for the model components and refer to the values of each parameter vector. The aim of the third penalty term is the smoothing of the baseline hazard so that this term penalizes the squared distances of two neighbouring intercepts and not the intercepts itself. This penalty term can be also defined by matrices, which leads to

$$\lambda_0 \sum_{t=1;s>t}^{t^*} \left\| \gamma_{0t} - \gamma_{0s} \right\|_2^2 = \lambda_0 (R \cdot \boldsymbol{\gamma}_0)^T (R \cdot \boldsymbol{\gamma}_0)$$

and with $t^* = 4$ one obtains the following matrix:

$$R = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

An alternative strategy for smoothing the baseline hazard may be the use of splines as illustrated for example by Berger and Schmid (2018). However, using the squared distances is purely discrete and does not need any underlying continuous assumption about time. Furthermore there is no limitation at the borders of time space.

Since cross validation can be computational time consuming in mixture models it is proposed to use AIC or BIC as selection criteria. To account for the fit as well as for the complexity of the model it is necessary to define them in an

appropriate way. Parameters, which are shrank should counted less than unpenalized parameters, which is captured by the effective degrees of freedom. The AIC and BIC are defined as

$$AIC(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + 2edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}),$$
$$BIC(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + \log(n)edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}),$$

where $edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ is the effective degrees of freedoms of the cure model. For each parameter set $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ the effective degrees of freedoms are calculated separately by

$$\begin{aligned} edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) &= edf(\hat{\boldsymbol{\beta}}) + edf(\hat{\boldsymbol{\gamma}}) \\ &= 1 + \sum_{j=1}^{g} edf(\hat{\boldsymbol{\beta}}_j) + edf(\hat{\boldsymbol{\gamma}}_0) + \sum_{j=1}^{h} edf(\hat{\boldsymbol{\gamma}}_j), \end{aligned}$$

where 1 refers to the intercept β_0 and $edf(\hat{\boldsymbol{\beta}}_j)$ to the effective degrees of freedom of the j-th parameter group of $\hat{\boldsymbol{\beta}}$. $edf(\hat{\boldsymbol{\gamma}}_0)$ denotes the effective degrees of freedom of the baseline and $edf(\hat{\boldsymbol{\gamma}}_j)$ to the j-th parameter group of $\hat{\boldsymbol{\gamma}}$. Following Yuan and Lin (2006) the effective degrees of freedom of each parameter group are given by

$$edf(\hat{\boldsymbol{\beta}}_j) = \mathbb{1}(\|\hat{\boldsymbol{\beta}}_j\|_2 > 0) + (df_{\boldsymbol{\beta}_j} - 1)\frac{\|\hat{\boldsymbol{\beta}}_j\|_2}{\|\hat{\boldsymbol{\beta}}_j^{ML}\|_2},$$
$$edf(\hat{\boldsymbol{\gamma}}_j) = \mathbb{1}(\|\hat{\boldsymbol{\gamma}}_j\|_2 > 0) + (df_{\boldsymbol{\gamma}_j} - 1)\frac{\|\hat{\boldsymbol{\gamma}}_j\|_2}{\|\hat{\boldsymbol{\gamma}}_j^{ML}\|_2}$$
$$edf(\hat{\boldsymbol{\gamma}}_0) = 1 + (df_{\boldsymbol{\gamma}_0} - 1)\frac{(R \cdot \hat{\boldsymbol{\gamma}}_0)^T(R \cdot \hat{\boldsymbol{\gamma}}_0)}{(R \cdot \hat{\boldsymbol{\gamma}}_0^{ML})^T(R \cdot \hat{\boldsymbol{\gamma}}_0^{ML})},$$

The idea is to relate the penalized estimates to the unpenalied maximum likelihood estimates (ML). For example, if the baseline parameters $\boldsymbol{\gamma}_0$ are not penalized, $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_0^{ML}$ will be identical, which lead to $edf(\hat{\boldsymbol{\gamma}}_0) = df_{\boldsymbol{\gamma}_0}$. If the baseline parameters are penalized at most, the baseline hazard is almost constant and only one degree of freedom remains. In general if a variable is not penalized the *edf* are identical to $df_{\boldsymbol{\beta}_j}$ and $df_{\boldsymbol{\gamma}_j}$, respectively.

Since there are three independent tuning parameters there would be a three-dimensional grid for selection the best combination of tuning parameters. Since the smoothing is less crucial it can be recommended to fix $\boldsymbol{\lambda_0}$ at some medium level to reduce the model complexity and computing time.

6 Estimation

6.1 Construction of the log-likelihood

The likelihood of the discrete cure model can be derived from the unconditional probability of the occurrence of an event

$$P(T = t|\boldsymbol{x}) = \lambda(t|\boldsymbol{x}_i) \prod_{s=1}^{t-1} (1 - \lambda(s|\boldsymbol{x}_i))$$

If an observation is not censored and no event is observed the contribution is $(1-\lambda(s|\boldsymbol{x}_i))$ for at least $t_i - 1$ time points. If an event take place the contribution is $\lambda(t|\boldsymbol{x}_i)$. Using the information provided by y_{is} (introduced in Section 3) the likelihood of the discrete survival model of one specific observation i can be written as

$$L_i^{disc} = \prod_{s=1}^{t_i} \lambda(s|\boldsymbol{x}_i)^{y_{is}} (1 - \lambda(s|\boldsymbol{x}_i))^{1-y_{is}}$$

This likelihood is equivalent to the likelihood of a binary response model with observations y_{is}. As long as $y_{is} = 0$ the contribution to the likelihood function is $1 - \lambda(s|\boldsymbol{x}_i)$. If an event is observed $\lambda(s|\boldsymbol{x}_i)$ is added to the log-likelihood. In the cured population the probability of an event is zero so that the likelihood of the long-term survivors can be simplified to[2]

$$L_i^{LTS} = \prod_{s=1}^{t_i} 0^{y_{is}} (1 - 0)^{1-y_{is}}$$

The likelihood of the cure model combines L_i^{LTS} and L_i^{disc} to

$$\begin{aligned} L_i = \pi(\boldsymbol{z}_i) \left(\prod_{s=1}^{t_i} \lambda(s|\boldsymbol{x}_i)^{y_{is}} (1 - \lambda(s|\boldsymbol{x}_i))^{1-y_{is}} \right) \\ + (1 - \pi(\boldsymbol{z}_i)) \left(\prod_{s=1}^{t_i} 0^{y_{is}} 1^{1-y_{is}} \right) \end{aligned} \tag{4}$$

Note that this equation only holds for modelling the failure time. One could also include the contribution of the censoring process itself as shown in Tutz and Schmid (2016).

[2]Note that $0^0 := 1$, $1^0 := 1$ and $\log(0) \to -\infty$

The complete log-likelihood for all observations is given by

$$
\begin{aligned}
l_c(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \Bigg(\log(\pi(\boldsymbol{z}_i)) + \log\Big(\prod_{s=1}^{t_i} \lambda(s|\boldsymbol{x}_i)^{y_{is}} (1-\lambda(s|\boldsymbol{x}_i))^{1-y_{is}} \Big) \\
&\quad + \log(1-\pi(\boldsymbol{z}_i)) + \log\Big(\prod_{s=1}^{t_i} 0^{y_{is}} 1^{1-y_{is}} \Big) \Bigg) \\
&= \sum_{i=1}^{n} \Bigg(\log(\pi(\boldsymbol{z}_i)) + \sum_{s=1}^{t_i} \Big(\log\big(1-\lambda(s|\boldsymbol{x}_i)\big) + y_{is} \log\Big(\frac{\lambda(s|\boldsymbol{x}_i)}{1-\lambda(s|\boldsymbol{x}_i)}\Big) \Big) \\
&\quad + \log(1-\pi(\boldsymbol{z}_i)) + \sum_{s=1}^{t_i} \big(y_{is} \log(0) \big) \Bigg) \\
&:= \sum_{i=1}^{n} \Bigg(\log(\pi(\boldsymbol{z}_i)) + \log(S(y_i|\boldsymbol{x}_i)) + \log(1-\pi(\boldsymbol{z}_i)) + \log(S^{LTS}(y_i)) \Bigg), \quad (5)
\end{aligned}
$$

where $\boldsymbol{\theta}$ includes all parameters. $l_c(\boldsymbol{\theta})$ can be estimated using the EM-algorithm described in the next section. For readability reason only the last line is used for the further description.

6.2 Estimation via EM-Algorithm

The EM-algorithm by Dempster et al. (1977) is used to estimate $l_c(\boldsymbol{\theta})$ by treating the unknown class membership as a problem with incomplete data. ζ_i denote the unknown mixture component that indicate whether observation i belongs to the non-cured population

$$
\zeta_i = \begin{cases} 1, \text{observation } i \text{ is from the non-cured population} \\ 0, \text{observation } i \text{ is from the cured population} \end{cases}
$$

With equation 5 follows

$$
l_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Big(\zeta_i \big\{ \log(\pi(\boldsymbol{z}_i)) + \log(S(y_i|\boldsymbol{x}_i)) \big\} + (1-\zeta_i) \big\{ \log(1-\pi(\boldsymbol{z}_i)) + \log(S^{LTS}(y_i)) \big\} \Big)
$$

In case of penalization the proposed penalty terms are added to $l_c(\boldsymbol{\theta})$. The penalized log-likelihood is

$$
\begin{aligned}
l_p(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \Big(\zeta_i \big\{ \log(\pi(\boldsymbol{z}_i)) + \log(S(y_i|\boldsymbol{x}_i)) \big\} + (1-\zeta_i) \big\{ \log(1-\pi(\boldsymbol{z}_i)) + \log(S^{LTS}(y_i)) \big\} \Big) \\
&\quad - \lambda_{\boldsymbol{\beta}} \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \big\| \boldsymbol{\beta}_j \big\|_2 - \lambda_{\boldsymbol{\gamma}} \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \big\| \boldsymbol{\gamma}_j \big\|_2 - \lambda_0 \sum_{t=1;s>t}^{t^*} \big\| \gamma_{0t} - \gamma_{0s} \big\|_2^2.
\end{aligned}
$$

If the estimation is not penalized, the penalty terms can be omitted.

Within the EM algorithm the log-likelihood is iteratively maximized by using an expectation and a maximization step. During the E-step the conditional expectation of the complete log-likelihood given the observed data $\boldsymbol{y}$ and the current estimate $\boldsymbol{\theta}^{(s)} = (\boldsymbol{\beta}^{(s)}, \boldsymbol{\gamma}^{(s)})$,

$$M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \mathrm{E}(l_p(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}^{(s)})$$

has to be computed. Because $l_p(\boldsymbol{\theta})$ is linear in the unobservable data ζ_i, it is only necessary to estimate the current conditional expectation of ζ_i. From Bayes's theorem follows

$$\begin{aligned} E(\zeta_i|\boldsymbol{y}, \boldsymbol{\theta}) &= P(\zeta_i = 1|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= P(y_i|\zeta_i = 1, \boldsymbol{x}_i, \boldsymbol{\theta}) P(\zeta_i = 1|\boldsymbol{x}_i, \boldsymbol{\theta}) / P(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= \pi_i S(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) / \{\pi_i S(y_i|\boldsymbol{x}_i) + (1 - \pi_i) S^{LTS}(y_i)\} = \hat{\zeta}_i. \end{aligned}$$

This is the posterior probability that the observation y_i belongs to the non-long-term survivor component of the mixture. In general it is permitted that an observation, for which an event is observed, might have a $\hat{\zeta}_i$ lower than one to account for all possible data structures including events by mistake. However, since the log-likelihood contribution of $S^{LTS}(y_i)$ would be close to minus infinity if an event take place this would occur very rarely and the algorithm usually avoids to assign such values for observations with observed events.

For the s-th iteration one obtains

$$\begin{aligned} M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) \quad & \left.\begin{aligned} &= \textstyle\sum_{i=1}^{n} \left\{ \hat{\zeta}_i^{(s)} \log(\pi_i) + (1 - \hat{\zeta}_i^{(s)}) \log(1 - \pi_i) \right\} \\ &- \lambda_\beta \textstyle\sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2 \end{aligned} \right\} M_1 \\ & \left.\begin{aligned} &+ \textstyle\sum_{i=1}^{n} \hat{\zeta}_i^{(s)} \log(S(y_i|\boldsymbol{x}_i)) \\ &- \lambda_\gamma \textstyle\sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2 - \lambda_0 \textstyle\sum_{t=1;s>t}^{t^*} \left\|\gamma_{0t} - \gamma_{0s}\right\|_2^2 \end{aligned} \right\} M_2 \\ & \left.\begin{aligned} &+ \textstyle\sum_{i=1}^{n} (1 - \hat{\zeta}_i^{(s)}) \log(S^{LTS}(y_i)) \end{aligned} \right\} M_3 \end{aligned}$$

M_1, M_2 and M_3 can be estimated independently from each other. The R-package MRSP by Pößnecker (2019) contains functions to estimate M_1 and M_2 including the mentioned penalty terms. Not every package would be suitable since the derivatives of M_1 and M_2 do not exist because of the group lasso penalty term. This problem can be solved with the fast iterative shrinkage-thresholding algorithm (FISTA) of Beck and Teboulle (2009) which is implemented in the MRSP package and is used for the maximisation problem of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. It can be generally formulated as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmax}}\, l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = -\underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}}\, l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}}\, -l(\boldsymbol{\beta}, \boldsymbol{\gamma}) + J_\lambda(\boldsymbol{\beta}, \boldsymbol{\gamma}). \quad (6)$$

FISTA belongs to the class of proximal gradient methods in which only the unpenalized log-likelihood and its gradient is necessary. A detailed description can be found in Schneider et al. (2019). For given $\boldsymbol{\theta}^{(s)}$ one computes in the E-step the weights $\hat{\zeta}_i^{(s)}$ and in the M-step maximizes $M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$. The E- and M-steps are repeated alternatingly until the relative tolerance

$$\Big|\frac{l_p(\boldsymbol{\theta}^{(s+1)}) - l_p(\boldsymbol{\theta}^{(s)})}{rel.tol/10 + |l_p(\boldsymbol{\theta}^{(s+1)})|}\Big| < rel.tol$$

is small enough to assume convergence. λ_β, λ_γ and λ_0 span a three-dimensional grid of tuning parameter space. Dempster et al. (1977) showed that under weak conditions the EM algorithm finds a local maximum of the likelihood function. Hence it is always advisable to use meaningful start values to find a good solution of the maximization problem.

7 Illustrative Example: Penalization for Recidivism Data

Here I demonstrate, how the proposed penalization technique from Section 5 works and how it can improve the model of recidivism of prisoners. In Section 4 the chosen variables are used in both parts of the model. While most estimates were in line with the hypotheses, none of them were statistically significant. Now all those variables mentioned in Table 3 and 4 are included in the selection process. In addition to the previous variables marital status, race, released on parole and the level of education are available. The penalty terms ensure that only complete variables can be chosen but not single categories of one variable. The tuning parameter for the baseline hazard of the non-long-term survivor component λ_0 is set to 2, while the other two tuning parameters span a two-dimensional grid with λ_β and λ_γ range from 150 to 0.01 using 15 discrete values, respectively. λ_β and λ_γ are transformed by $\tilde{\lambda} = \log(\lambda + 1))$ to obtain a logarithm scale.

Figure 2 shows the results of the selection process using $15 \times 15 = 225$ tuning parameter combinations. If $\tilde{\lambda}_\beta = \tilde{\lambda}_\gamma \approx 5$ a pure intercept model is fitted. If both tuning parameters are close to zero an almost unpenalized model is estimated. The highest BIC values are detected in the corners of the graph in which at least one $\tilde{\lambda}$ is close to zero. That implies that models where all available variables are included in at least one component are not an appropriate choice according to BIC. It is possible to detect a clear region of very low BIC values. The minimum is found for $\tilde{\lambda}_\beta \approx 2.48$ and $\tilde{\lambda}_\gamma \approx 1.89$ at BIC = 1372.70. This is a strong reduction compared to the unpenalized model 1 with BIC value of 1673.36.

To get more insights in the mechanism of the variable selection Figure 2 is cut into slices and we look at the development of both coefficient sets β and γ. For that matter one λ value is fixed at the chosen value while the other λ varies from high penalty (5.02) to low penalty (0.01). Each line type in the coefficient path represent one parameter group. In the first row of Figure 4 the γ estimates

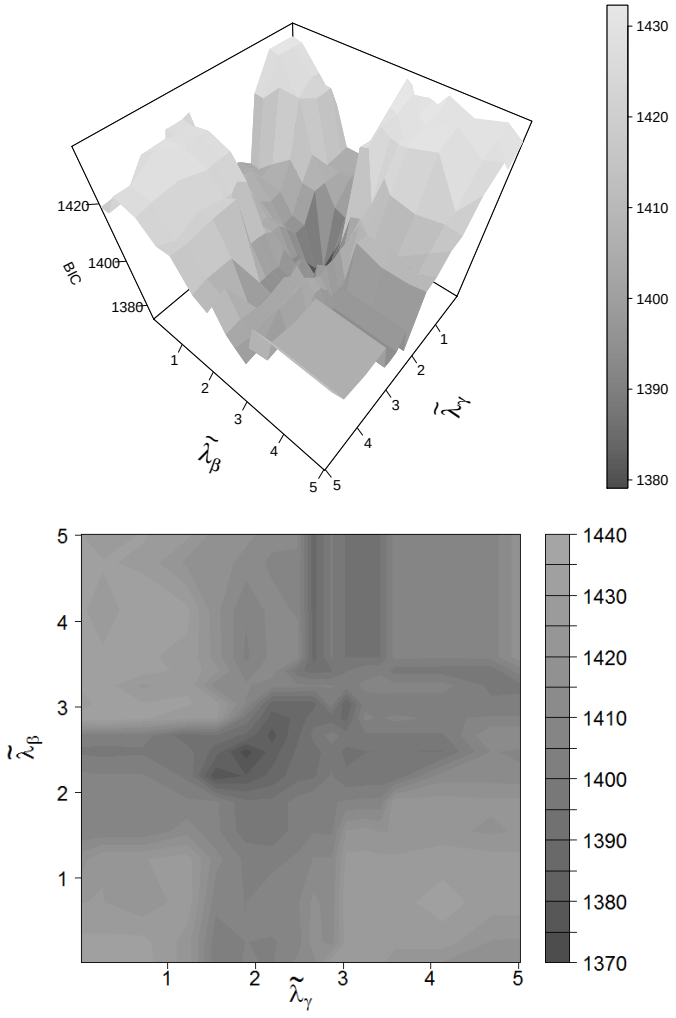


FIGURE 2: *Criminal recidivism: Grid of λ values to find the best tuning parameter combination according to BIC*

for the standardized covariates are displayed. In the second row the β coefficients can be found and the last row contains the boxplots of the estimated π. On the left hand side $\tilde{\lambda}_\gamma$ is set to 1.89 and $\tilde{\lambda}_\beta$ varies. On the right hand side $\tilde{\lambda}_\beta$ is fixed at 2.48 and $\tilde{\lambda}_\gamma$ is changing.

Usually the lower the tuning parameter the more coefficients are different from zero. But one should keep into mind that the estimates of β and γ are not completely independent from each other. Looking at the left hand side of Figure 4 one can see that the γ estimates are quite unsteady although the corresponding $\tilde{\lambda}_\gamma$ is fixed. But the mixture weights determined by the β coefficients change. At $\tilde{\lambda}_\beta = 5.02$ no β coefficient is selected and for all observations a constant π around 0.38 is estimated. Then π increases to 0.53, before the first β coefficient is selected and the weights become more and more individual specific. In this case a high variation in the π boxplots can be seen as a higher individual differentiation

which is desirable. However, each additional coefficient not only needs to improve the model fit but also reduces the BIC value to be selected. Thus the BIC is used to find a trade-off between model fit and number of parameters.

On the right hand side of Figure 4 $\tilde{\lambda}_\beta$ is fixed and $\tilde{\lambda}_\gamma$ varies. Here the weights π and β coefficients are almost constant. At the time when "not married" is selected in the γ part of the model two β coefficients are set to zero. Thus the interdependence works in both ways.

Since the graph illustrates the coefficient paths for the standardized covariates, we can also compare the absolute values of the estimates. In case of the β coefficients at the left hand side of Figure 4 it is obvious that age and "prior convictions" are the first parameters which are selected. At $\tilde{\lambda}_\beta \approx 2.48$ the parameter "financial aid" is the smallest one out of the three coefficients. Thus age and "prior convictions" have a stronger impact than "financial aid". On the right "work experience" seem to have the greatest effect in the γ dimension followed by "not married".

	Penalized		Refit	
Covariates	Non-LTS(β)	Hazard(γ)	Non-LTS(β)	Hazard(γ)
Constant	0.0075		0.1067	
Financial aid: yes	-0.1543		-0.2857	
Age	-0.0410		-0.0477	
No prior convictions	0.0556		0.1064	
Work experience: yes		-0.6216		-0.8843
Married: No		0.5461		0.9952

TABLE 6: *Comparison of penalized and upenalized coefficients of the cure model for recidivism data*

Covariates	Estimates	BS.sd	BS.2.5	BS.97.5	
Constant	0.1067	0.3518	0.0323	1.3252	
Financial aid: yes	-0.2857	0.2667	-0.9565	0.1124	$\hat{\boldsymbol{\beta}}$
Age	-0.0477	0.0202	-0.0939	-0.0162	
Number prior convictions	0.1064	0.0607	0.0291	0.2550	
Work experience: yes	-0.8843	0.3422	-1.4758	-0.0793	$\hat{\boldsymbol{\gamma}}$
Married: No	0.9952	0.4367	0.1772	1.8567	

TABLE 7: *Model 2: Refit of the cure model for recidivism data with penalized intercepts*

Table 6 gives the estimates of the selected model. For β only "financial aid", age and "prior convictions" are selected. "Work experience" and "not married" are chosen for modeling the hazard. It is an coincidence that in this case none of the variables is selected in both parts of the model. The first two estimation

columns show the penalized estimates while the last two column contain the unpenalized estimates of the refit. The disadvantage of penalized estimates is that they are not unbiased but on the other hand they may lead to a smaller variance. Usually the unpenalized absolute estimates for one parameter group are larger than the penalized. However, if someone wants to have traditional standard errors and confidence intervals, it is plausible to refit the model without penalization to obtain unpenalized estimates and to be able to calculate standard errors. Table 7 contains the unpenalized estimates with Bootstrap standard errors and confidence intervals. They are obtained by 600 non-parametric samples of the data. Someone should keep into mind that these Bootstrap results ignore the model search and that the intercepts γ_0 are not displayed but their differences are still penalized. Now all coefficients are statistically significant to 5% level except of "financial aid". I would recommend to use the calculated bootstrap confidence intervals determined by the 2.5% and 97.5% quantiles of the bootstap distribution, because according to my experience the sampled distributions are often very skewed and the estimated coefficient value do not need to be in the middle of the sampled distribution. If this interval contains zero the corresponding coefficient is non-significant to the level 5%, which only applies for "financial aid".

The interpretation of the coefficients is the same as in Section 4. Age and financial aid reduce the probability to be non-cured while the number of prior convictions increase the probability. If someone is married and has work experience the probability of an event in the non-cured population is reduced.

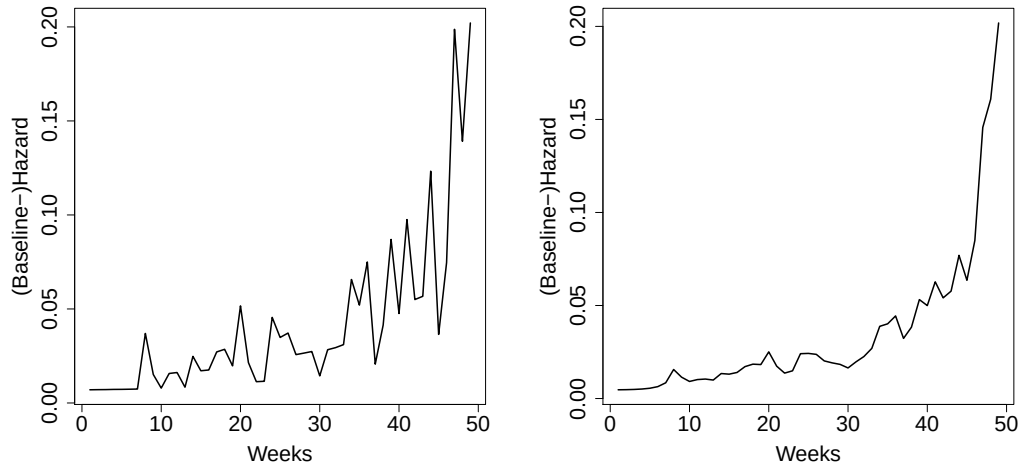


FIGURE 3: *Criminal recidivism: Comparison of the baseline hazard of the unpenalized model and the refitted penalized model*

Finally Figure 3 illustrates the effect of smoothing the baseline by penalizing the difference between neighbouring intercepts. On the left the baseline hazard of the unpenalized model is displayed. It is a quite rough function with many ups and downs. On the right the penalized baseline hazard is shown, which is much

smoother, but keep the nature of the original function at the same time. The tuning parameter for smoothing can be enlarged to get a even smoother curve.

The proposed penalization technique could improve the original model substantially and results in an easy-to-interpret model.

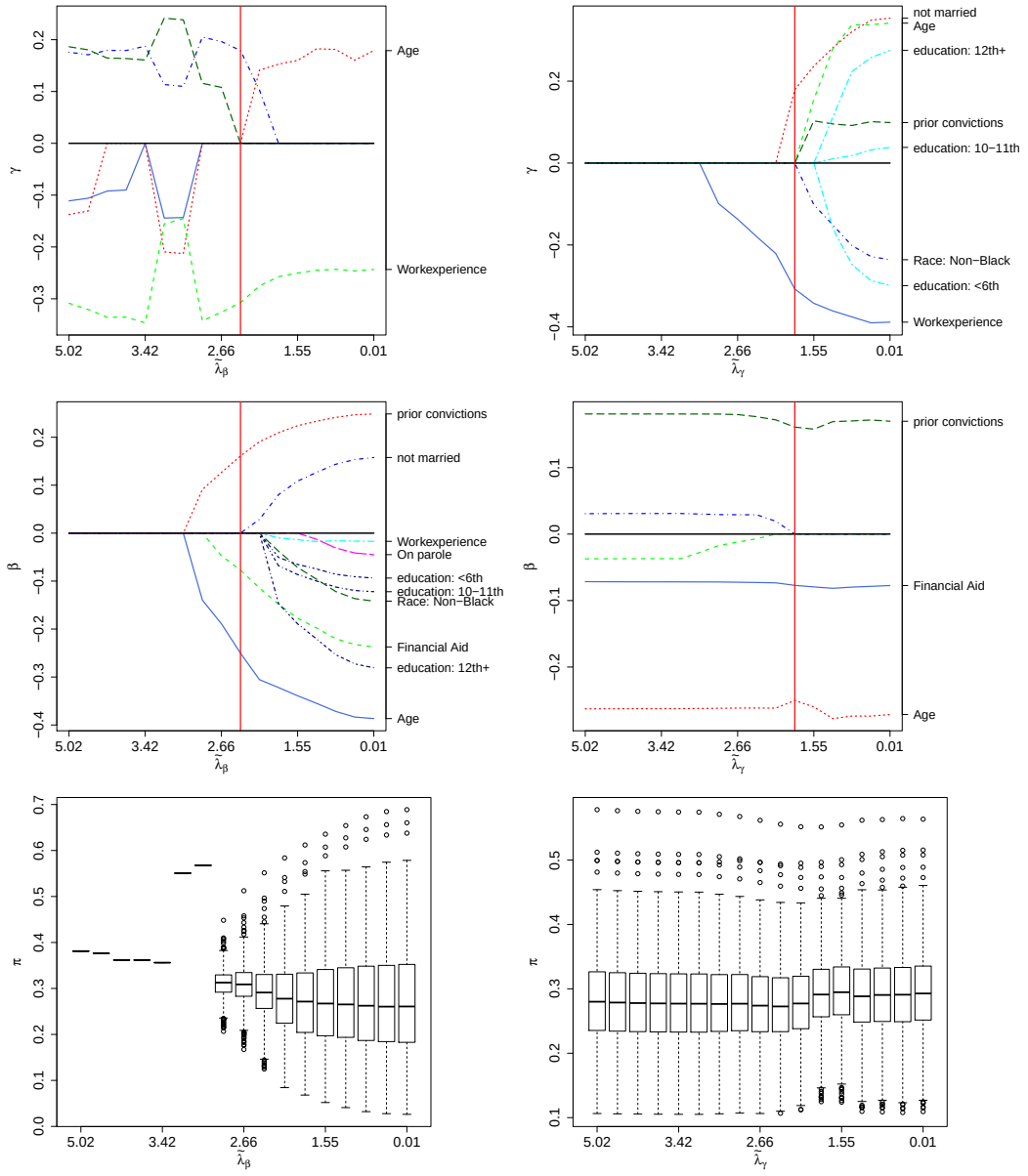


FIGURE 4: *Criminal recidivism: Standardized coefficient paths of β and γ and π for fixed λ_γ (left) and fixed λ_β (right) in the cure model*

8 Application: Breast Cancer

Breast cancer is the most common cancer for women in developed countries. The average risk for a American woman to develop breast cancer sometime in her life is around 12% (see Akram et al., 2017). Thus, it is extremely relevant, which variables may be associated with being a long-term survivor from breast cancer and how variables are associated with the survival time of the patients. I use data of the SEER data base and the proposed methods to evaluate these questions.

SEER is the "Surveillance, Epidemiology, and End Results" Program (`www.seer.cancer.gov`), which collects information on cancer in the U.S. population on an individual basis. The time from diagnosis to death from breast cancer in years is given and I draw a random sample of $6,000$ breast cancer patients who entered the SEER data base between 1997 and 2011 (using SEER $1973-2011$ Research Data, version of November 2013). Since only the time span matters, the year of diagnosis can vary between the persons. The observed time may be also right-censored, when an event has not been observed (yet). Furthermore only female patients, younger than 76 years with first malignant tumor and without distant metastases were included so that there is a realistic chance to be a long-term survivor. Events can take place from the first until the 15th year.

	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
Age at diagnosis (years)	18	48	56	56	64	75
Tumorsize (mm)	1	10	16	21	25	230
Number examined nodes	1	3	7	9	14	57

TABLE 8: *Descriptive statistics of quantitative explanatory variables for the breast cancer data (SEER)*

Table 8 and 9 shows the covariates, which might be selected. Most of the variables are related to the medical data. The primary site denotes where the breast cancer was found. The most frequent locations are C504 which is the upper outer quadrant of the breast and C508 which is the overlapping lesion of breast. The tumor grade specifies how well the tumor can be differentiated from healthy cells ranging from "well" over "moderately" to "poorly". It is known which radiation therapy and in which order was applied. Then it is reported how many lymph nodes were examined and how many positive lymph nodes were found. The latter variable has four categories: None, one to three, four to six and seven or more positive lymph nodes. The T-stage variable classify the tumor according to AJCC 6th in four categories relying mainly on the size of the tumor and its extension. Further variables are the hormone receptor status (positive or negative) of estrogen (ER) and progesterone (PR), the laterality (right or left), the tumorsize (in mm), the age at diagnosis (in years), the race (white, black, others) and the marital status (single, married, separated, divorced, widowed).

	Category	Observations	Proportions (in %)
Marital status	single	803	13
	married	3898	65
	separated	50	1
	divorced	717	12
	widowed	532	9
Race	white	4836	81
	black	536	9
	others	628	10
Primary Site	C500 areolar	27	0
	C501 subareolar	289	5
	C502 Upper inner	718	12
	C503 Lower inner	356	6
	C504 Upper outer	2201	37
	C505 Lower outer	437	7
	C506 Axillary tail	41	1
	C508 Overlapping lesion	1203	20
	C509 Entire breast	728	12
Laterality	right	2877	48
	left	3123	52
Tumor Grade	1 well	1300	22
	2 moderately	2569	43
	3 poorly	2131	36
Radiation therapy	1 None	2106	35
	2 Beam	3715	62
	3 Implants	82	1
	4 Combinations	42	1
	5 Other	55	1
Radiation Sequence	1 None	2170	36
	2 Other	37	1
	3 Rad. after surgery	3793	63
ER status	positive	4760	79
	negative	1240	21
PR status	positive	4241	71
	negative	1759	29
Number positive nodes	0	4018	67
	1-3	1416	24
	4-6	274	5
	7+	292	5
T-Stage	I	3922	65
	II	1701	28
	III	281	5
	IV	96	2

TABLE 9: *Descriptive statistics of discrete explanatory variables for the breast cancer data (SEER)*

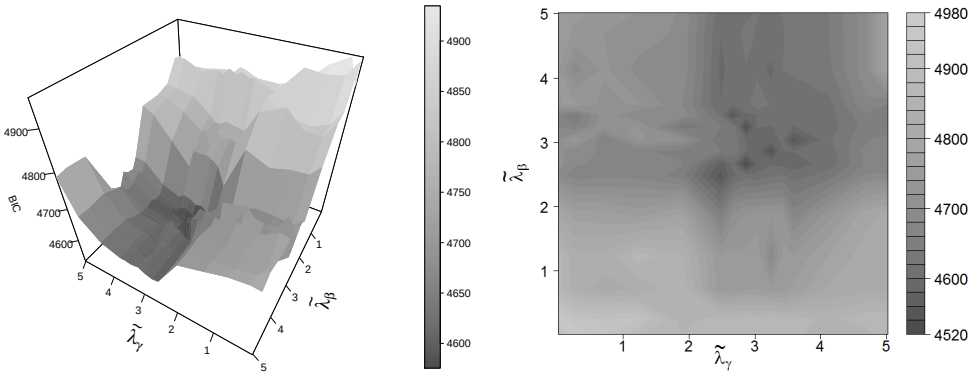


FIGURE 5: *Breast cancer: Grid of λ values to find the best tuning parameter combination according to BIC*

Figure 5 shows the result of the grid search for 15×15 tuning parameter combinations. On the left the surface is illustrated and on the right the corresponding contour plot. As in the illustrative example the tuning parameter for the baseline hazard is fixed at 2. The transformed tuning parameters for the other two dimensions $\tilde{\lambda} = log(\lambda + 1)$ vary between 5.02 (high penalty) and 0.01 (low penalty). Including all variables in both parts of the model using a very low penalty leads to the highest BIC displayed in the right corner of the surface in Figure 5. But also using a very high penalty for both dimensions (left corner of the surface) does not lead to a desirable result. Although both tuning parameters are important to find the lowest BIC value, a too low $\tilde{\lambda}_\beta$ leads to higher BIC values regardless of $\tilde{\lambda}_\gamma$. Thus specifying the probability of being a long-term survivor seems to be more relevant.

The lowest BIC was found at 4525.62 with the tuning parameters $\tilde{\lambda}_\beta \approx 2.66$ and $\tilde{\lambda}_\gamma \approx 2.87$. After selecting the variables the model was refitted using only a penalized baseline, but no penalization term for the other coefficients. The parameter estimates of this refitted model are displayed in Table 10. The bootstrap confidence intervals rely on the bootstap 2.5% (BS.2.5) and 97.5% (BS.97.5) quantiles of 600 non-parametric bootstrap samples. Note that these bootstrap samples do not account for the selection process since only the selected variables are included.

The result of the proposed variable selection is a selection of only 20 out of 68 possible coefficients related to covariates. Moreover it can be decided which covariate effects the probability of being a non-long-term survivor captured by β, which covariate is important for the occurrence of an event modeled by γ and which covariates are necessary in both components. Here only the race and the number of positive nodes are selected for modeling non-LTS. The tumor grade, size of tumor and T-stage are chosen in both components and the laterality, ER and PR status are only chosen for the event occurrence.

Positive estimates in the upper part of the table are related with an increase

Covariates	Estimates	BS.sd	BS.2.5	BS.97.5	
Constant	-2.7980	0.1133	-3.0630	-2.6120	
Race: Black	0.3157	0.0928	0.1446	0.5036	
Race: Others	-0.4800	0.0857	-0.6744	-0.3351	
Number of pos. nodes: 1-3	0.5516	0.0938	0.3994	0.7743	
Number of pos. nodes: 4-6	0.9564	0.1524	0.7339	1.3237	
Number of pos. nodes: 7+	1.8370	0.1791	1.5765	2.2828	$\hat{\boldsymbol{\beta}}$
Tumor Grade: II	0.1317	0.1069	0.0243	0.4291	
Tumor Grade: III	0.5968	0.1012	0.4458	0.8494	
Size of tumor	0.0141	0.0031	0.0076	0.0197	
T-Stage: II	0.2178	0.0712	0.0688	0.3496	
T-Stage: III	-0.1057	0.0875	-0.2926	0.0533	
T-Stage: IV	0.5393	0.1146	0.3595	0.8198	
Tumor Grade: II	0.3625	0.2727	-0.1719	0.8821	
Tumor Grade: III	0.9388	0.2791	0.4019	1.4753	
Size of tumor	0.0051	0.0057	-0.0022	0.0204	
T-Stage: II	0.3293	0.1640	-0.0188	0.6223	
T-Stage: III	0.5239	0.4156	-0.5143	1.1599	$\hat{\gamma}$
T-Stage: IV	1.4500	0.3698	0.6414	2.2179	
Laterality: Left	0.3762	0.1279	0.2050	0.7069	
ER status: negative	0.8662	0.1617	0.4472	1.0762	
PR status: negative	0.5684	0.1485	0.3306	0.8956	
$1 - \bar{\pi}$	0.8498	0.0066	0.8313	0.8571	

TABLE 10: *Parameter estimates of the refitted cure model for breast cancer. Only the baseline is penalized. The standard errors and confidence intervals are obtained by bootstrap samples*

of the probability of being a non-LTS and in the lower part with an increase of the probability of an event namely death by breast cancer. Thus the number of positive nodes have an positive effect of being a non-LTS. The more positive nodes are found the higher the probability that the person is non-cured. If one to three nodes are positive the chance to be non-LTS compared to be LTS is increased by the factor $\exp(0.5516) = 1.74$ compared to patients without positive nodes. If the number of positive nodes are seven or more the multiplicative factor is with $\exp(1.8370) = 6.28$ much higher. Compared to white ethnic black people have a higher chance of being non-LTS while "others" have a lower chance.

Figure 6 illustrates the effect of tumor grade and T-stage in both dimensions. On the y axis the effect of being non-LTS is marked. The x axis shows the effect of an event. Generally the higher the category of tumor grade and T-stage the higher the chances in both dimensions. The only exception is tumor grade III which reduce the chance of non-LTS compared to tumor grade I. However the main driven factor of the T-stages categories I to III is the size of the tumor so that the negative effect of T-stage III can be compensated to some extend by the effect of tumorsize. The highest category of T-stage and tumor grade show the

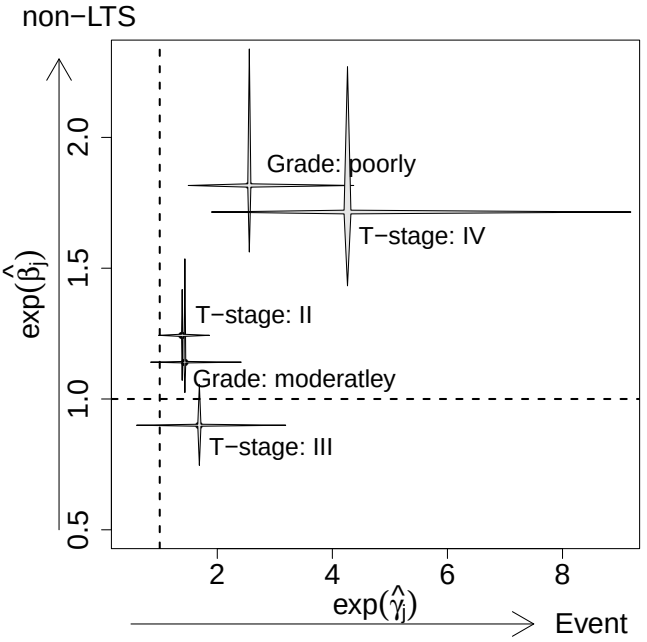


FIGURE 6: *Breast cancer: Effect of tumor grade and T-stage including confidence intervals computed by the* 2.5% *and* 97.5% *quantiles of non-parametric bootstrap samples*

strongest effects in both dimensions.

If the cancer is detected on the left the chance of the event "death by breast cancer" at time t (compared to an event death by breast cancer later than t) is increased by $\exp(0.3762) = 1.46$ compared to laterality right. One reason may be that it is more difficult to treat cancer on the left side since the cancer is closer to the heart so that the radiation therapy for example need to be applied with more care than on the right. The negative status of both hormone receptors ER and PR increase the risk of the event, too. As long as the status of one of the hormone receptors is positive, it is possible to use drugs to fight the cancer. If the status is negative, hormone therapy does not work. The so called triple-negative breast cancers are defined by negative ER, PR and HER2. This type of cancer usually grows and spreads faster than other types of breast cancer and hormone therapy can not be applied. Because HER2 is only reported for observations from the year of diagnosis of 2010 onwards, the parameter could not be considered in this application.

According to the model the best chances of belonging to the long-term survivor group have patients with no positive lymph nodes, a very small tumor, which can be well differentiated from healthy cells and with ethnicity which is neither black or white. If the person does not belong to the long-term survivors the best survival chances are estimated for patients with a small tumor, which can be well

differentiated from healthy cells, located at the right hand side and characterized by a positive ER and PR status. However, one should keep in mind that these results are not based on a randomized trial.

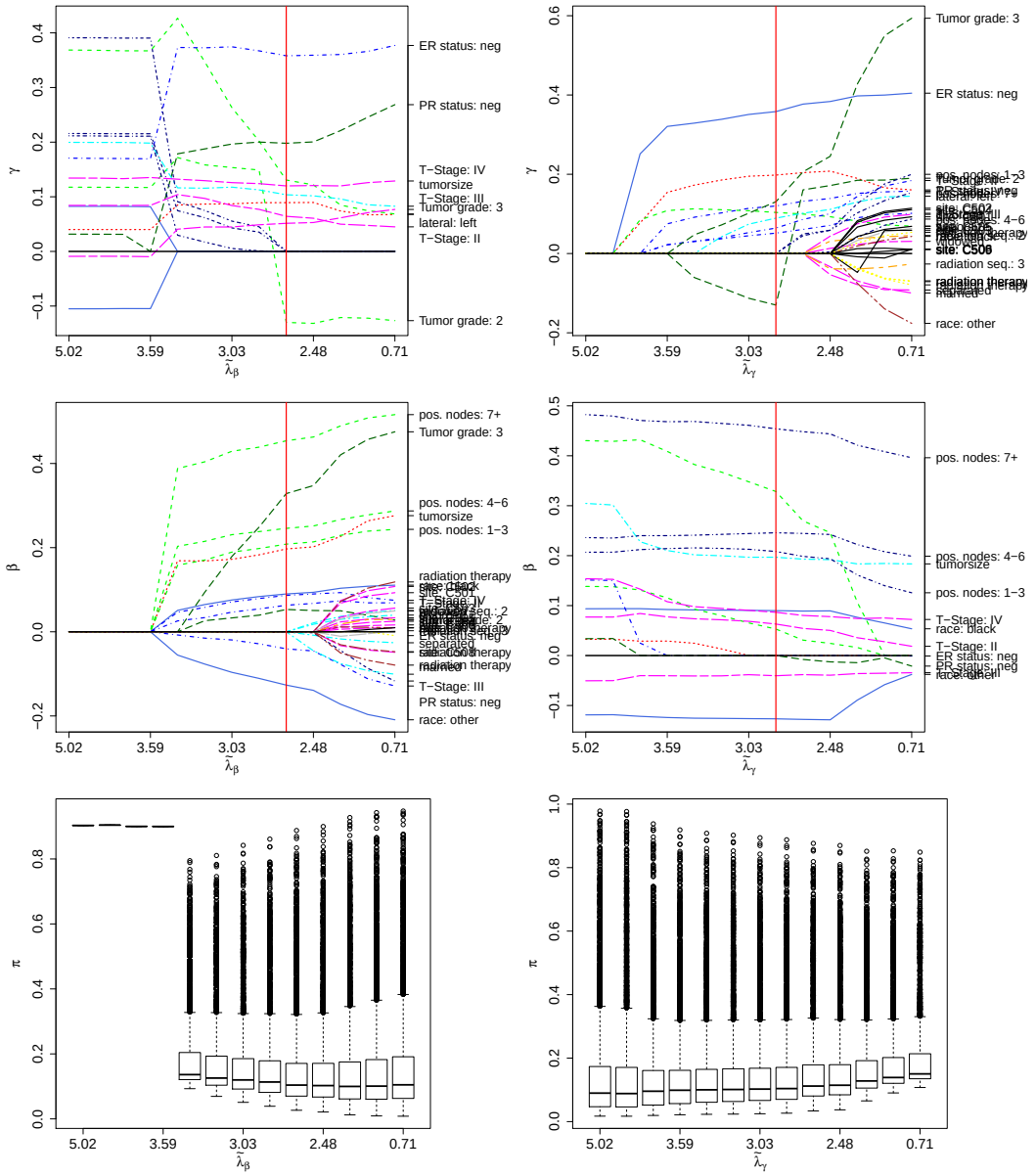


FIGURE 7: *Breast cancer: Standardized coefficient paths of β and γ and π for fixed λ_γ (left) and fixed λ_β (right) in the cure model*

Figure 7 illustrates the standardized coefficient paths for this model. Since there are two varying tuning parameters it is necessary to introduce some constraints. On the left the coefficient paths are displayed when $\tilde{\lambda}_\gamma$ is hold constant at 2.87. On the right $\tilde{\lambda}_\beta$ is fixed at 2.66 and $\tilde{\lambda}_\gamma$ varies. The first row contains the estimates of γ, the second the estimates of β and the last one the boxplots

of π. Each line type correspond with one covariable which can consist of more than one coefficient as T-Stage for example. On the left the effect of entering β coefficients is remarkable. At $\tilde{\lambda}_\beta = 3.42$ some β coefficients enter the model so that the median π drops dramatically. From there on the π are calculated for each observation individually. The values of γ coefficients change as well at this point although $\tilde{\lambda}_\gamma$ is kept constant. The estimated weights defined by β may have a strong influence on the γ estimates. On the right hand side $\tilde{\lambda}_\beta$ is fixed and $\tilde{\lambda}_\gamma$ varies. Here the effect of changing γ has less effect on β because γ has no direct relation to π which stay almost constant. However, it can be seen that the values of γ and β sometimes change the sign or become smaller with smaller penalty. This might be caused by inter dependencies between γ and β or by the data structure, when covariates influence each other by correlation.

9 Identifiability

Identifiability of cure models for continuous time was shown by Li et al. (2001) and Hanin and Huang (2014). It is assumed that there are at least three discrete time points ($t \geq 3$) and there is an effect $\gamma \neq 0$ of a continuous covariate x. Let the cure model be represented by two parameterizations

$$\pi_\beta S(\gamma_{0t} + \boldsymbol{x}^T\gamma) + (1 - \pi_\beta) = \pi_{\tilde{\beta}} S(\tilde{\gamma}_{0t} + \boldsymbol{x}^T\tilde{\gamma}) + (1 - \pi_{\tilde{\beta}})$$

There are values δ_{0r}, δ such that $\tilde{\gamma}_{0r} = \gamma_{0r} + \delta_{0r}$, $\tilde{\gamma} = \gamma + \delta$. With $\eta_r(x) = \gamma_{0r} + x\gamma$ one obtains for all x and r

$$\pi S(\eta_r(x)) - \tilde{\pi} S(\eta_r(x) + \delta_{0r} + x\delta) = (\pi - \tilde{\pi}).$$

Let us consider now the specific values $x_z = -\gamma_{0r}/\gamma + z/\gamma$ yielding for all values z and r

$$\pi S(z) - \tilde{\pi} S(z + \delta_{0r} + x_z\delta) = (\pi - \tilde{\pi}).$$

By building the difference between these equations for values z and $z - 1$ one obtains for all values z

$$\pi(S(z) - S(z-1)) = \tilde{\pi}(S(z + \delta_{0r} + x_z\delta) - S(z - 1 + \delta_{0r} + x_z\delta)).$$

The equation has to hold in particular for values $z = 1, 2, \ldots$. Since the logistic distribution function $F(\eta) = \exp(\eta)/(1 + \exp(\eta))$ is strictly monotonic and the derivative is different for all values η it follows that $\delta_{0r} = \delta = 0$ and $\pi = \tilde{\pi}$.

10 Concluding Remarks

It has been shown that the discrete cure model can be used to model heterogeneity which arises from long-term survivors and patients at-risk in a discrete time

setting. In the discrete survival analysis the hazard can be always interpreted as probability which makes any interpretation more intuitive. The instabilities of the model as no event occurrence at a certain time point or the number of parameters to estimate a rather rough baseline hazard can be overcome by the proposed penalization techniques. Furthermore it is possible to carry out variable selection so that there is a data driven way to decide which variable should be included in which part of the model. The variables can be chosen for one of the model components as well as for both model components. The proposed methods show stable and easy-to-interpret results in the applications. Thus it is possible to reduce the number of coefficients substantially and evaluate which covariates are associated with long-term survivors and the event of risk.

In case of breast cancer patients with no positive lymph nodes, a very small tumor, which can be well differentiated from healthy cells and with ethnicity which is neither black or white have the best chances to belong to the long-term survivors. The best survival chances in the group of non-LTS are estimated for patients with a small tumor, which can be well differentiated from healthy cells, located at the right hand side and characterized by a positive ER and PR status.

However, further research is necessary to evaluate the effect of the smoothing parameter on the general results and to develop computational efficient bootstrap samples which take the model search into account. In general, discrete cure models are the appropriate method, if the time is discrete and if there are two subgroups where one is characterized as long-term survivors.

Acknowledgements: Thanks to Gerhard Tutz and Paul Fink for fruitful discussions.

References

Akram, M., M. Iqbal, M. Daniyal, and A. U. Khan (2017). Awareness and current knowledge of breast cancer. *Biological Research 50*(33), 1–23.

Amico, M. and I. V. Keilegom (2018). Cure models in survival analysis. *Annual Review of Statistics and Its Application 5*(1), 311–342.

Beck, A. and M. Teboulle (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences 2*(1), 183–202.

Berger, M. and M. Schmid (2018). Semiparametric regression for discrete time-to-event data. *Statistical Modelling 18*(3-4), 322–345.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B 39*, 1–38.

Fox, J. and M. S. Carvalho (2012). The rcmdrplugin.survival package: Extending the r commander interface to survival analysis. *Journal of Statistical Software 49*(7), 1–32.

Hanin, L. and L.-S. Huang (2014). Identifiability of cure models revisited. *Journal of Multivariate Analysis 130*, 261–274.

Kuk, A. Y. C. and C.-H. Chen (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika 79*(3), 531–541.

Li, C.-S., J. M. G. Taylor, and J. P. Sy (2001). Identifiability of cure models. *Statistics & Probability Letters 54*(4), 389–395.

Maller, R. A. and X. Zhou (1996). *Survival analysis with long-term survivors*. Wiley New York.

Muthén, B. and K. Masyn (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral statistics 30*(1), 27–58.

Pößnecker, W. (2019). MRSP: Multinomial response models with structured penalties. R package version 0.6.11, https://github.com/WolfgangPoessnecker/MRSP.

Rossi, P. H., R. A. Berk, and K. J. enihan (1980). *Money, Work, and Crime: Some Experimental Results*. New York: Academic Press.

Schneider, M., W. Pößnecker, and G. Tutz (2019). Variable selection in mixture models with an uncertainty component. Technical Report 225, Department of Statistics, Ludwig-Maximilians-Universität München.

Steele, F. (2003). A discrete-time multilevel mixture model for event history data with long-term survivors, with an application to an analysis of contraceptive sterilization in bangladesh. *Lifetime Data Analysis 9*(2), 155–174.

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) (2014). *Research Data (1973-2011)*. National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2014, based on the November 2013 submission.

Sy, J. P. and J. M. G. Taylor (2000). Estimation in a cox proportional hazards cure model. *Biometrics 56*(1), 227–236.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B 58*, 267–288.

Tutz, G. and M. Schmid (2016). *Modeling Discrete Time-to-Event Data*. Springer.

Willett, J. B. and J. D. Singer (1995). Its déja vu all over again: Using multiple-spell discrete-time survival analysis. *Journal of Educational and Behavioral Statistics 20*(1), 41–67.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B 68*, 49–67.

Eidesstattliche Versicherung

Hiermit erkläre ich, Micha Schneider, an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 19.12.2019