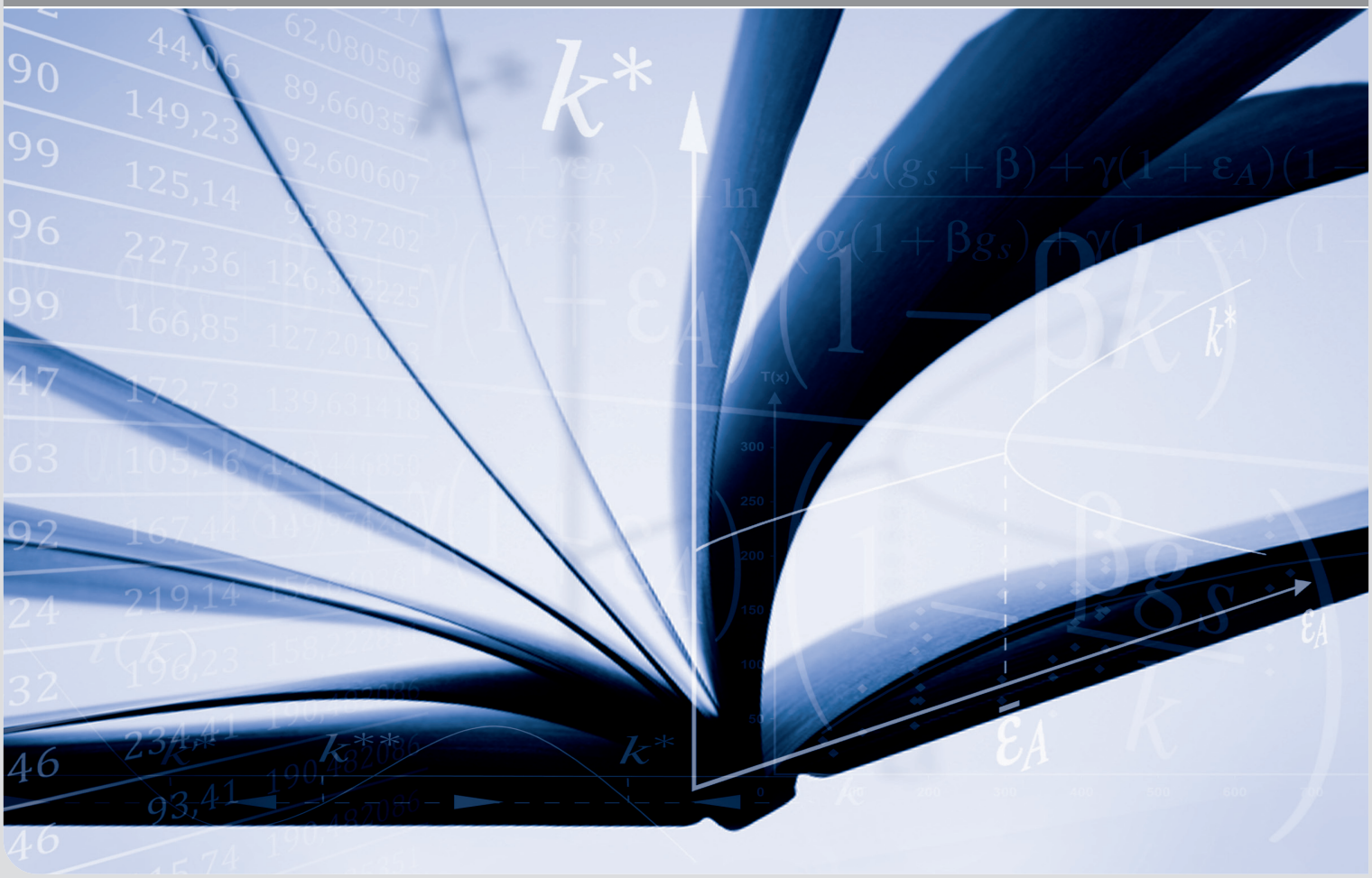
Karlsruher Institut für Technologie

Selection of questions for VAAs and the VAA-based elections

by Andranik S. Tangian

No. 100 | JANUARY 2017

WORKING PAPER SERIES IN ECONOMICS



Impressum

Karlsruher Institut für Technologie (KIT)
Fakultät für Wirtschaftswissenschaften
Institut für Volkswirtschaftslehre (ECON)

Schlossbezirk 12
76131 Karlsruhe

KIT – Die Forschungsuniversität in der Helmholtz-Gemeinschaft

Working Paper Series in Economics
No. 100, January 2017

ISSN 2190-9806

econpapers.wiwi.kit.edu

Institut für Wirtschaftstheorie und Operations Research

Karlsruhe Institute of Technology

Selection of questions for VAAs and the VAA-based elections

Andranik S. Tangian

Working paper Nr. 100

January 2017

E-mail: andranik-tangian@boeckler.de
Tel: +49 211 7778-259
Fax: +49 211 7778-190

Kollegium am Schloss 76128 Karlsruhe Deutschland

Abstract

During the 2016 election to the Student Parliament of the Karlsruhe Institute of Technology (KIT), an experiment on 'The Third Vote' was conducted. The goal was to test an alternative election method based on the idea of internet voting advice applications (VAAs). Under the election method tested, the voters cast no direct votes for candidate parties; rather, they are asked about their preferences on the policy issues as declared in the party manifestos. These embedded referenda measure the degree to which the parties' positions match the policy preferences of the electorate. The parliament seats are then distributed among the parties in proportion to their indices of representativeness: popularity (the average percentage of the population represented on all the issues) and universality (frequency in representing a majority).

The Third Vote Experiment reveals that the critical point is the selection of questions: unless they draw sufficient distinctions between the parties, it can cause a malfunction of both the VAA and the VAA-based election method. To solve this problem, this paper develops a model for contrasting as much as possible between the parties by maximizing the total distance between the party policy profiles while simultaneously reducing the number of questions. The guaranteed best solution is obtained by means of an exhaustive search on all the possible combinations of m out of n initial questions. However, since this search is cumbersome, a stepwise removal of questions is proposed. This alternative is shown to offer a good compromise between formal rigor and computational efficiency.

Keywords: Policy representation, elections, theory of voting, feature selection, variable selection.

JEL Classification: D71

Acknowledgements

The author is grateful to Jan de Leeuw, Professor Emeritus, UCLA Statistics, for the valuable information about the current research in feature selection.

Contents

1 Introduction

During the 2016 election to the Student Parliament of the Karlsruhe Institute of Technology (KIT), an experiment on 'The Third Vote' was conducted. The goal was to test the election method based on the idea of internet voting advice applications (VAAs), like the German *Wahl-O-Mat*.[1] Under the election method tested, the voters cast no direct votes for candidate parties; rather, they are asked about their preferences on the policy issues as declared in the party manifestos. Thereby, the balance of public opinion on each issue is revealed. These embedded referenda measure the degree to which the parties' positions match the policy preferences of the electorate. The parliament seats are then distributed among the parties in proportion to their indices of representativeness: popularity (the average percentage of the population represented on all the issues) and universality (frequency in representing a majority); see [Tangian 2014, 2017a–b].

The Third Vote Experiment was organized in the following way. In addition to the official electoral ballot with the names of the seven student parties, each voter was offered an experimental ballot to be filled in on a voluntary basis; see Figure 1. The experimental ballot is called 'The Third Vote' because it complements the German two-vote system[2] with an additional vote in the form of a questionnaire. The preamble to the ballot explains the goal of the experiment and assures that it does not impact the official election. For analysis purposes, the voter is asked to indicate the party he/she voted for in the official ballot and whether the *StuPa-O-Mat* — the KIT voting advice application analogous to the Wahl-O-Mat — influenced the choice. A table contains ten questions on university policies, which were heuristically selected by the experiment organizers from the 27 StuPa-O-Mat questions (shown in Table 1) as being most important and discriminating between the parties. The positive party responses are coded by 1s, the negative by -1s, and the neutral or missing responses by 0s. From the 3671 registered voters, 1069 valid experimental ballots were received; the results of the experiment are described in [Diemer and Eßwein 2016, KIT 2016, Tangian 2016].

The Third Vote Experiment shows that the alternative election method can increase the representativeness of a parliament. At the same time, it makes clear that there are some bottlenecks that are also inherent in the VAAs. One of most critical points is the selection of questions and their wordings. Currently they are the responsibility of a supposedly neutral official commission, providing that certain criteria are met. Since it is nearly impossible to fulfill this task impartially, there is a risk of manipulating electoral outcomes by posing questions favorably for one candidate party and unfavorably for others. To avoid this, the questions could be drawn up by the parties themselves either implicitly, within the party manifestos, or explicitly, by announcing a list of program policy issues. The questions formulated by one party could be shared with all other parties, giving them an opportunity to make their positions comparable. Furthermore, competing parties could negotiate on the formulation of questions in order to prevent misinterpretations. This process, if considered part of the electoral campaign, would exclude all claims of partiality in the selection and formulation of questions.

However, allowing the candidates (parties) to propose the questions themselves has three shortcomings. Firstly, they could be too numerous for inclusion into electoral ballots. For instance, if each of 30 German parties is entitled to five questions, their total number rises to 150 — then most VAA users and/or voters would likely just skip most of them. Secondly, if the questions are numerous, some, though formulated differently, could in fact treat the same topic,

[1] A VAA asks the user a number of questions on topical policy issues (Introduce nationwide minimum wage? Yes/No; Introduce a speed limit on the motorways? Yes/No, etc.). The computer program, drawing on all the parties' answers, finds for the user the best-matching party, the second-best-matching party, etc., 'advising' thereby the optimal choice; see [Bundeszentrale für politische Bildung 2014]

[2] The first vote is for an individual representative of the constituency and the second vote is for a party. Since the latter determines the proportion of parliament factions, the second vote is decisive.

resulting in its overweight compared with others. Thirdly, certain questions can be redundant, like the StuPa-O-Mat Question 16 in Table 1, which received the same answer from all seven student parties. If such redundant questions or those which poorly discriminate between the parties (like the StuPa-O-Mat Questions 2, 17 and 25 in Table 1) are numerous, then the parties' indices of representativeness are too close to each other, causing a malfunction of both the VAA and the VAA-based election method. Indeed, in this case the parties seem almost equally representative for the VAA users, and the third vote results in party factions of almost equal size. This effect has already been observed in the Third Vote Experiment. As follows from [Tangian 2016, Figure 3], the direct vote discriminates between the most and the least successful parties by a factor of 6 (FiPS with 33.7% and Rosa with 5.6% of the votes), whereas the mean indices of the most and the least representative parties differ by a factor of 2 (Juso with a mean index of 63% and Rosa with 34%).

To surmount these shortcomings, the questions included in the electoral ballots should be rather few yet maximally distinguish between the parties. A similar problem emerges in testing products, where evaluation criteria should highlight the differences in their quality. If the criteria poorly discriminate between the products — for instance, if the power consumption of electric devices is equal, the noise is of the same level, and the size of the units is the same, then a test based on these features is ill-designed. Likewise, a survey questionnaire should also reveal differences, because nearly-identical responses are of little use.

As for selecting few questions, this task can be formulated in terms of reduction of the number of variables with little loss of information. In the machine learning and data mining literature, this problem is known as 'feature selection' or 'variable subset selection' [Feature selection 2017]. In combinatorial mathematics, the goal is formulated as the reduction of matrices while preserving most of the column data, which is called 'the column subset selection'; for surveys see [Kumar and Schneider 2016, Zheng et al 2010]. In the 'principal component variable selection', which is a sub-domain of the principal component analysis, one finds the principal components that are linear combinations of the initial variables and then reduces the number of variables while preserving the 'most important' components. The first results date back to 1970s; see [Jolliffe 1972, 1973, 2002] and [McCabe 1975, 1984]. They were developed, for instance, by [Al Kandari and Jolliffe 2005, Armstrong et al 2014, Husson et al 2011, Krzanowski 1987, Kuroda et al 2011, McKay and Campbell 1982a, McKay and Campbell 1982b, Mori et al 2007, Mori et al 2016, Pacheco et al 2013]. For the particular case of binary variables see [Broadbent et al 2010, De Leeuw 2006].

The known methods are however of limited applicability for our purposes. Being primarily designed for big data, they are based on approximations which are not necessarily optimal. They also attempt to maintain the initial distances between the 'observations', whereas our goal is to accentuate the differences by increasing them. Above all, these methods cannot be easily explained to non-mathematicians, which is critical in convincing the general public to apply them in the context of elections. Therefore, we introduce a simple direct procedure to optimally select the questions both for VAAs and the VAA-based election method. This procedure selects a subset of questions that best contrasts between the parties by maximizing the total distance between the party policy profiles. The distance between policy profiles is defined in several ways, and the results are compared. All computations are performed with MATLAB 2016a, with the output in LaTeX, running a PC with Intel Core i7 CPU (3.5 GHz) and 16 GB RAM.

In Section 2, 'Measuring the degree of discrimination between parties', we consider the Euclidean, Manhattan and Hamming distances between the party policy profiles as well as their correlation. These measures allow us to optimally select the ten out of 27 StuPa-O-Mat questions that provide the highest degree of differentiation between the seven student parties. As follows from our comparisons, the heuristical selection method used by the experiment organizers was quite good but ultimately led to some unforeseen shortcomings.

– EXPERIMENT –
"The Third Vote"

In this experiment, we wish to test the idea of Prof. Andranik Tangian aimed at making representative democracy more representative. With this alternative election method, the electorate's policy profile is measured using a third vote. The policy profile of the electorate is compared with that of the candidate parties, and the degree to which they match determines the election result. In this way, we endeavor to overcome irrational behavior and voting partiality.

Participation in the survey is completely **voluntary, anonymous** and has **NO** influence on the official election. Results of our analysis will be be made available on *www.studierendenwahl.econ.kit.edu*. For further questions, please do not hesitate to ask the election coordinators at the ballot boxes.

What party did you vote for on the official ballot?

- ☐ Liberale Hochschulgruppe (LHG)
- ☐ RCDS - Ring christlich-demokratischer Studenten
- ☐ Liste für basisdemokratische Initiative, Studium, Tierzucht und Elitenbeförderung (LISTE) / Liste unabhängiger studierender Tierzüchter (LUST)
- ☐ FiPS - Fachschaftserfahrung im Parlament der Studierenden
- ☐ Die Linke.SDS
- ☐ Rosa Liste
- ☐ Juso - studentisch, demokratisch, solidarisch

Did you use the StuPa-O-Mat to help you make your choice?

- ☐ yes
- ☐ no

Please answer these selected StuPa-O-Mat questions to help us define your policy profile:

	+	o	−	#
Baden-Württemberg-wide off-peak ticket with the semester fee				1
More video surveillance in insecure areas of campus, e.g. lockers				2
More vegan choices in the cafeteria, even if it limits meat meals				3
Abolish admission restrictions for courses of study				4
Sexism is a current problem at the KIT				5
Abolish the maximum duration of study				6
Promote gender-neutral restroom facilities on campus				7
Heavily restrict commercial advertising on campus				8
Special deals on tickets to cultural events with the semester fee				9
Replace low-attendance lectures with recordings and exercise classes				10

+ agree o neutral − against

Figure 1: English translation of the experimental electoral ballot

Table 1: Evaluation of four selections of 10 questions from StuPa-O-Mat questionnaire

Questions	Party positions (matrix $\boldsymbol{B}$)							Questions selection criterion				
	LHG	RCDS	LISTE	FiPS	Linke	Rosa	Juso	Selection by organizers	Euclidian distance	Manhattan distance	Hamming distance	correlation
1 Financing the student body. The student body should be financed exclusively by voluntary contributions	1	0	−1	−1	−1	−1	−1					X
2 Room for children and infants. There should be a room at the KIT for child and infant care that students can use	0	1	1	1	1	1	1				X	
3 State wide transport ticket. A Baden-Württemberg-wide transport ticket for evenings and weekends, funded through the mandatory semester fee, should be introduced	−1	0	1	−1	1	1	−1	X	X	X		
4 Military research. Military research should be heavily restricted at the KIT. Possible answers: 'Military research should be completely prohibited'; 'Research for purely military objectives should be prohibited'; 'Military research should be allowed with no restrictions'	−1	−1	0	0	1	1	0					
5 Dealing with the KIT past. The student body should take up a debate accounting for the past of the KIT and its predecessors	0	1	0	1	1	1	0					
6 Video surveillance. There should be more video surveillance in security-sensitive areas (eg. lockers) on campus	−1	1	−1	−1	−1	−1	−1	X			X	
7 Vegan meals in the canteen. The canteen should offer more vegan and sustainable options, even if this means limiting the offer of meals containing meat	−1	−1	−1	0	1	1	0	X		X		
8 Career launch. Courses of study at KIT should be designed to promote quick entry into a career	−1	1	−1	−1	−1	−1	−1				X	
9 University competition. Competition between universities should be reduced	−1	−1	−1	0	1	1	0		X	X		
10 Child care places for students. There should be more places in daycare facilities near the KIT for the children of students	0	0	0	1	1	1	0					
11 Religion room. The KIT should provide a room that is always open for the exercise of religion	−1	0	1	−1	0	1	−1		X	X		X
12 BAFöG. The BAFöG (student financial aid in Germany) should be independent of parental income	1	1	1	1	1	1	−1				X	
13 Admission restrictions. Admission restrictions for courses of study should be abolished	−1	−1	1	−1	0	1	−1	X	X	X		X
14 Sexism. Sexism is a current problem at the KIT	−1	0	0	1	1	1	0	X				
15 Maximum studies duration. The maximum duration of study should be abolished	1	−1	1	−1	1	1	−1	X	X	X		
16 Committees of the student body. The Student Parliament and the Conference of Faculties should be merged together	−1	−1	−1	−1	−1	−1	−1				X	

Table 1: Evaluation of four selections of 10 questions from StuPa-O-Mat questionnaire (continued)

Questions	Party positions (matrix $\boldsymbol{B}$)							Questions selection criterion				
	LHG	RCDS	LISTE	FiPS	Linke	Rosa	Juso	Selection by organizers	Euclidian distance	Manhattan distance	Hamming distance	correlation
17 Sponsoring. The student body should make use of sponsors at events like the University festival and other cultural events	1	1	1	1	0	1	1				X	
18 Gender-neutral restrooms. The student body should campaign for gender-neutral restroom facilities on campus	−1	−1	−1	0	1	1	−1	X	X	X		X
19 Payments for AStA speakers. Students who get involved at AStA should do so on a strictly unpaid basis	1	−1	0	−1	−1	−1	−1					X
20 Dormitory construction. The expansion of dormitory facilities should be paid for by student grants	−1	0	−1	−1	1	−1	−1				X	X
21 Subtitles in lecture videos. All recorded courses should be uploaded with subtitles (for inclusion of hearing-impaired students)	1	0	0	1	1	1	−1		X			
22 fzs. The student body should become a member of the fzs (Freier Zusammenschluss von Studenten). Explanation: the fzs is a nationwide and politically neutral alliance of student bodies. It represents students at the federal level and is a member of European Student Union (ESU). Currently the member fee is 40 ct. per student per semester	−1	−1	−1	−1	1	1	0		X	X		X
23 Advertisments on campus. Promotion and advertisements from companies should be heavily restricted on campus	−1	−1	−1	−1	1	1	−1	X	X	X	X	X
24 Cultural events. The student body should advocate special deals on entrance fees and cultural events by introducing a mandatory semester fee	−1	−1	−1	−1	1	1	0	X	X	X		X
25 Accessibility. All areas of the KIT should be accessible without restrictions	1	1	1	0	1	1	1				X	
26 Poor attended lectures. Lectures with low attendance rates should be replaced by recordings and exercise classes	−1	0	0	−1	−1	1	−1	X				X
27 Political mandate. The student body should participate in the general political debate. Explanation: the coalition agreement of the latest green-black (Green-CDU/CSU) state government intends to limit the political mandate of student bodies, restricting them to issues of university policy only	1	−1	1	1	1	1	1				X	
Total Euclidian distance between party 10-profiles								26.19	26.75	26.74	18.32	25.29
Total Manhattan distance between party 10-profiles								20.40	21.80	22.20	10.20	19.60
Total Hamming distance between party 10-profiles								8.00	7.10	6.90	14.70	8.10
Total correlation between party 10-profiles								0.71	0.16	−1.30	11.82	−2.46

In Section 3, 'Stepwise removal of questions', the set of 27 questions is distilled to just those that most sharply discriminate between the parties. The ten questions finally selected are the same as those chosen under the exhaustive search, but the computation time is drastically curtailed. While this procedure does not *guarantee* the optimal output, it does work fairly well, providing an acceptable compromise between formal rigor and computational efficiency.

In Section 4, 'Conclusions', the results of the paper are recapitulated and put into context.

2 Measuring the degree of discrimination between parties

Let us consider the task performed by the organizers of The Third Vote Experiment, selection of ten out of 27 StuPa-O-Mat questions. The heuristical choice made by the organizers is shown by Xs in the first column of the right-hand section of Table 1.

Our goal is to reduce the number of questions while accentuating the differences between the parties. We define the latter to be the total distance between the party policy profiles displayed in the middle section of Table 1 (denoted as matrix $\boldsymbol{B}$):

$$\begin{aligned} \text{Total Euclidean distance} &= \sum_{i<j} d[\boldsymbol{B}(:,i), \boldsymbol{B}(:,j)] \\ &= \sum_{i<j} \sqrt{\sum_{q} [\boldsymbol{B}(q,i) - \boldsymbol{B}(q,j)]^2} \ , \end{aligned}$$

where

i, j are indices of the columns of matrix $\boldsymbol{B}$, associated with the parties,

: denotes the full range of the matrix rows; in our case the set of row numbers $\{1, 2, \ldots, 27\}$,

$\boldsymbol{B}(:,i)$ is the ith column and $\boldsymbol{B}(:,j)$ is the jth column of matrix $\boldsymbol{B}$,

q are indices of the rows of matrix $\boldsymbol{B}$, associated with the questions.

To select ten questions that maximize the total Euclidian distance between the columns of the remainder of matrix $\boldsymbol{B}$, we perform an exhaustive search on all the $8,436,285$ combinations of ten out of 27 questions. In other words, we solve the maximization problem

$$\max_{Q:\ Q \subset 1:27,\ |Q|=10} \ \sum_{i<j} \sqrt{\sum_{q \in Q} [\boldsymbol{B}(q,i) - \boldsymbol{B}(q,j)]^2} \ . \tag{1}$$

The selection of ten out of 27 questions that maximize the total Euclidean distance between the party profiles is shown by Xs in the second column of the right-hand section of Table 1.

Since we consider both the full-sized matrix $\boldsymbol{B}$ and its reduced versions with fewer rows (fewer questions), it is difficult to see the gain in the total distance between the matrix columns. To make the measurements comparable, we consider the *total normalized distance*, that is, divide the sum of the squared distances by $m =$ vertical size of $\boldsymbol{B}$ (number of questions considered):

$$\begin{aligned} \text{Total normalized Euclidean distance} &= \sum_{i<j} \sqrt{\frac{\sum_{q \in Q} [\boldsymbol{B}(q,i) - \boldsymbol{B}(q,j)]^2}{m}} \\ &= \frac{1}{\sqrt{m}} \sum_{i<j} \sqrt{\sum_{q \in Q} [\boldsymbol{B}(q,i) - \boldsymbol{B}(q,j)]^2} \ . \end{aligned}$$

Additionally to the Euclidean distance between policy profiles, we use three other discrimination measures, which are also normalized (which is unnecessary for the total correlation):

$$\begin{aligned} \text{Total Manhattan distance (sum of absolute differences)} &= \sum_{i<j}\sum_{q\in Q}\frac{|\boldsymbol{B}(q,i)-\boldsymbol{B}(q,j)|}{m} \\ \text{Total Hamming distance (number of mismatches)} &= \sum_{i<j}\sum_{q\in Q}\frac{\operatorname{sign}|\boldsymbol{B}(q,i)-\boldsymbol{B}(q,j)|}{m} \\ \text{Total correlation} &= \sum_{i<j}\rho[\boldsymbol{B}(Q,i),\boldsymbol{B}(Q,j)] \ . \end{aligned}$$

By analogy with (1), we solve similar optimization problems using the enumerated discrimination measures as objective functions. (For the total correlation, the maximization should be replaced by minimization.) Thereby, we obtain different selections of ten questions shown by Xs in the right-hand section of Table 1.

The bottom section of Table 1 shows the evaluation of the heuristical and the four model-based selections of ten questions. The framed values highlight the evaluation of the optimal selection with the use of the corresponding measure. As one can see, the heuristical selection made by the experiment organizers is optimal with respect to no formal criterion.

It should be taken into account that the optimal selection of ten out of 27 questions is not unique. For instance, the optimal selection with the use of total Euclidian distance could contain Question 7 instead of Question 9. For other discrimination measures, there are also multiple optimal selections of questions of the same size, and Table 1 displays only one of each type. However, their evaluations at the bottom of the table are common to all optimal selections of the given size for the given discrimination measure.

It should be also noted that neither the Hamming distance nor the correlation are good discrimination measures for our purpose. The former is too inaccurate, responding only to instances of mismatching. The correlation, even if inverted, lacks distance properties. Nevertheless, these two measures are often used in applications and are thus included for the generality of our consideration.

3 Stepwise removal of questions

As one can see from the last column of Tables 2–5, finding the optimal selection of m out of 27 questions can be time consuming. To enhance the computational efficiency, we apply a stepwise procedure analogous to the backward stepwise regression, which is also practiced in factor analysis [Hogarty et al 2004, Kano and Harada 2000]. We remove questions one-by-one, finding the least important question at each step. To be specific, we illustrate this procedure through the total Euclidean distance. The steps are traced in Table 2.

First, we find Question r such that, after its removal, the total normalized Euclidian distance is maximal. In other words, we solve the optimization problem

$$\max_r \frac{1}{\sqrt{26}}\sum_{i<j}\sqrt{\sum_{q\neq r}[\boldsymbol{B}(q,i)-\boldsymbol{B}(q,j)]^2} \ .$$

At this step Question 16 is removed, which increases the total normalized Euclidean distance from 22.96 to 23.39.

At the next step we remove Question s, solving the optimization problem

$$\max_s \frac{1}{\sqrt{25}}\sum_{i<j}\sqrt{\sum_{q\neq r,s}[\boldsymbol{B}(q,i)-\boldsymbol{B}(q,j)]^2} \ .$$

Now Question 2 is removed, which increases the total Euclidean distance normalized from 23.39 to 23.75. As one can see from the second section of Table 2, the questions remaining after these two steps are the same as under the exhaustive search on all combinations of 25 out of 27 questions. Removing questions one-by-one in the same way (maximizing the total distance), their number is reduced to ten. Again, the questions remaining after 16 steps are the same as under the exhaustive search on all combinations of 10 out of 27 questions. The question removed first is 16 — on which all seven parties were in agreement — and next those which least discriminate between the parties: 2, 17 and 25. The total normalized Euclidian distance for the remaining questions increases at each step, meaning a gradual increase in the discrimination between the parties. This trend ends as the number of remaining questions becomes small — three in this case. Below the table, the optimal selection of ten questions is shown with the questions also chosen by the experiment organizers in boxes.

As there can be multiple optimal selections of m out of 27 questions, the removal of a question at each step is not unique either. There can be several questions, after removal of each the total discrimination between the party policy profiles is maximal, which results in a branching process. Table 2 traces only one out of 24 possible successions of removed questions which is compared with most close optimal selections of questions obtained with an exhaustive search (they are also not unique!). Here, the stepwise procedure outputs the same selection of questions as the exhaustive search, but this is not *guaranteed* in the general case. Therefore, we call the stepwise search 'suboptimal' — to oppose it to the really optimal exhaustive search.

The use of other three discrimination measures is analogous. Tables 3–5 trace the stepwise removal of questions with their use. The selections obtained with the use of Manhattan and Hamming distances coincide with the ones obtained by means of the exhaustive search. The results with the use of correlation are not that perfect.

4 Conclusions

The Third Vote Experiment reveals certain particularities to be taken into account while designing VAAs and organizing VAA-based elections. The most critical point is the selection of questions, which should be rather few in number and must maximally discriminate between the parties, otherwise there is a risk of malfunction of both VAAs and the VAA-based elections.

Since it is nearly impossible to fulfill this task impartially, the questions could be drawn up by the parties themselves and shared with all other parties, giving them an opportunity to make their positions comparable. Furthermore, competing parties could negotiate on the formulation of questions in order to prevent misinterpretations. This process, if considered part of the electoral campaign, would exclude all claims of partiality in the selection and formulation of questions.

As for reducing the number of initial questions while maximally discriminating between the parties, we suggest to perform this task by maximizing the total distance between the party policy profiles. The guaranteed optimal solution can be obtained by means of an exhaustive search on all the combinations of m out of n initial questions. Since this search is cumbersome and resource-intensive, a stepwise removal of questions is proposed. We show that this alternative provides a good compromise between formal rigor and computational efficiency. It is also shown that the questions selected by our formal models discriminate between the parties better than the questions selected heuristically by the experiment organizers.

Table 2: Suboptimal and optimal selection of questions with the use of Euclidian distance

Number of retained questions	Suboptimal selection of questions by their stepwise removal		Optimal selection of questions by exhaustive search of all combinations				
	Question removed at the given step	Total Euclidian distance normalized	Total Euclidian distance normalized	Questions in optimal selection but not in suboptimal selection	Questions in suboptimal selection but not in optimal selection	Number of combinations of questions	Processing time in seconds
27	None	22.96	22.96	None	None	1	0
26	16	23.39	23.39	None	None	27	0
25	2	23.75	23.75	None	None	351	0
24	17	24.13	24.13	None	None	2925	0
23	25	24.52	24.52	None	None	17550	0
22	10	24.85	24.85	None	None	80730	2
21	5	25.19	25.19	None	None	296010	7
20	14	25.32	25.32	None	None	888030	23
19	6	25.47	25.47	None	None	2220075	56
18	4	25.61	25.61	None	None	4686825	117
17	8	25.75	25.75	None	None	8436285	211
16	12	25.87	25.87	None	None	13037895	324
15	26	26.00	26.00	None	None	17383860	430
14	19	26.16	26.16	None	None	20058300	497
13	27	26.31	26.31	None	None	20058300	496
12	1	26.45	26.45	None	None	17383860	430
11	20	26.60	26.60	None	None	13037895	322
10	7	26.75	26.75	None	None	8436285	207
9	21	26.93	26.93	None	None	4686825	115
8	9	27.14	27.14	None	None	2220075	54
7	23	27.34	27.34	None	None	888030	22
6	11	27.52	27.52	None	None	296010	7
5	22	27.74	27.74	None	None	80730	2
4	24	27.81	27.81	None	None	17550	0
3	13	28.02	28.02	None	None	2925	0
2	3	27.05	27.05	None	None	351	0
1	18	24.00	24.00	None	None	27	0

Ten questions retained (selected by organizers are in boxes): [3] 9 11 [13] [15] [18] 21 22 [23] [24]

Table 3: Suboptimal and optimal selection of questions with the use of Manhattan distance

Number of retained questions	Suboptimal selection of questions by their stepwise removal		Optimal selection of questions by exhaustive search of all combinations				
	Question removed at the given step	Total Manhattan distance normalized	Total Manhattan distance normalized	Questions in optimal selection but not in suboptimal selection	Questions in suboptimal selection but not in optimal selection	Number of combinations of questions	Processing time in seconds
27	None	16.07	16.07	None	None	1	0
26	16	16.69	16.69	None	None	27	0
25	2	17.12	17.12	None	None	351	0
24	17	17.58	17.58	None	None	2925	0
23	25	18.09	18.09	None	None	17550	0
22	5	18.36	18.36	None	None	80730	2
21	6	18.67	18.67	None	None	296010	7
20	8	19.00	19.00	None	None	888030	21
19	10	19.37	19.37	None	None	2220075	51
18	12	19.78	19.78	None	None	4686825	108
17	27	20.24	20.24	None	None	8436285	194
16	1	20.50	20.50	None	None	13037895	297
15	19	20.80	20.80	None	None	17383860	393
14	20	21.14	21.14	None	None	20058300	457
13	14	21.38	21.38	None	None	20058300	458
12	21	21.67	21.67	None	None	17383860	397
11	26	22.00	22.00	None	None	13037895	296
10	4	22.20	22.20	None	None	8436285	190
9	23	22.44	22.44	None	None	4686825	106
8	7	22.50	22.50	None	None	2220075	50
7	9	22.57	22.57	None	None	888030	20
6	11	22.67	22.67	None	None	296010	7
5	13	22.80	22.80	None	None	80730	2
4	18	23.00	23.00	None	None	17550	0
3	22	23.33	23.33	None	None	2925	0
2	24	24.00	24.00	None	None	351	0
1	3	24.00	24.00	None	None	27	0

Ten questions retained (selected by organizers are in boxes): [3] [7] 9 11 [13] [15] [18] 22 [23] [24]

Table 4: Suboptimal and optimal selection of questions with the use of Hamming distance

Number of retained questions	Suboptimal selection of questions by their stepwise removal		Optimal selection of questions by exhaustive search of all combinations				
	Question removed at the given step	Total Hamming distance normalized	Total Hamming distance normalized	Questions in optimal selection but not in suboptimal selection	Questions in suboptimal selection but not in optimal selection	Number of combinations of questions	Processing time in seconds
27	None	9.93	9.93	None	None	1	0
26	4	10.12	10.12	None	None	27	0
25	7	10.32	10.32	None	None	351	0
24	9	10.54	10.54	None	None	2925	0
23	11	10.78	10.78	None	None	17550	0
22	3	11.00	11.00	None	None	80730	2
21	14	11.24	11.24	None	None	296010	7
20	13	11.45	11.45	None	None	888030	21
19	18	11.68	11.68	None	None	2220075	53
18	21	11.94	11.94	None	None	4686825	111
17	22	12.24	12.24	None	None	8436285	199
16	24	12.56	12.56	None	None	13037895	307
15	26	12.93	12.93	None	None	17383860	408
14	5	13.21	13.21	None	None	20058300	469
13	10	13.54	13.54	None	None	20058300	468
12	15	13.92	13.92	None	None	17383860	405
11	1	14.27	14.27	None	None	13037895	302
10	19	14.70	14.70	None	None	8436285	197
9	20	15.22	15.22	None	None	4686825	108
8	23	15.75	15.75	None	None	2220075	51
7	2	15.86	15.86	None	None	888030	20
6	6	16.00	16.00	None	None	296010	7
5	8	16.20	16.20	None	None	80730	2
4	12	16.50	16.50	None	None	17550	0
3	17	17.00	17.00	None	None	2925	0
2	25	18.00	18.00	None	None	351	0
1	27	21.00	21.00	None	None	27	0

Ten questions retained (selected by organizers are in boxes): 2 [6] 8 12 16 17 20 [23] 25 27

Table 5: Suboptimal and optimal selection of questions with the use of correlation

Number of retained questions	Suboptimal selection of questions by their stepwise removal		Optimal selection of questions by exhaustive search of all combinations				
	Question removed at the given step	Total correlation normalized	Total correlation normalized	Questions in optimal selection but not in suboptimal selection	Questions in suboptimal selection but not in optimal selection	Number of combinations of questions	Processing time in seconds
27	None	6.59	6.59	None	None	1	0
26	16	5.85	5.85	None	None	27	0
25	25	5.26	5.26	None	None	351	0
24	2	4.55	4.48	16	17	2925	1
23	17	3.72	3.72	None	None	17550	4
22	12	3.08	3.07	16	5	80730	18
21	5	2.28	2.09	16	27	296010	65
20	27	1.27	1.27	None	None	888030	196
19	10	0.56	0.56	None	None	2220075	494
18	21	−0.10	−0.10	None	None	4686825	1083
17	14	−0.80	−0.80	None	None	8436285	1845
16	15	−1.13	−1.13	None	None	13037895	2842
15	4	−1.43	−1.43	None	None	17383860	3791
14	6	−1.66	−1.66	None	None	20058300	4377
13	3	−1.83	−1.83	None	None	20058300	4369
12	8	−2.08	−2.08	None	None	17383860	3783
11	11	−2.30	−2.30	None	None	13037895	2837
10	19	−2.42	−2.46	11 19	7 9	8436285	1831
9	23	−2.54	−2.73	19 23	7 9	4686825	1017
8	20	−2.70	−2.88	20 23	7 9	2220075	482
7	7	−2.80	−2.99	20	9	888030	193
6	9	−3.10	−3.10	None	None	296010	64
5	26	−3.11	−3.12	19	13	80730	18
4	22	−3.00	−3.00	19 20	1 13	17550	4
3	18	−2.50	−2.50	18 19 20	1 13 24	2925	1
2	13	−2.00	−2.00	19 20	1 24	351	0
1	1	0.00	0.00	27	24	27	0

Questions retained (selected by organizers are in boxes): 1 11 [13] [18] 19 20 22 [23] [24] [26]

References

[Al Kandari and Jolliffe 2005] Al Kandari NM, Jolliffe IT (2005) Variable selection and interpretation of correlation principal component. Environmetrics, 16, 659–72

[Armstrong et al 2014] Armstrong II DA, Bakker R, Carroll R, Hare Ch, Poole KT, Rosenthal H (2014) Analyzing Spatial Models of Choice and Judgment with R. CRC Press, Boca Raton FL

[Broadbent et al 2010] Broadbent ME, Brown M, Penner K (2010) Subset Selection Algorithms: Randomized vs Deterministic. SIAM Undergraduate Research Online, 3, May 2010. (Faculty advisors: I.C.F. Ipsen and R. Rehman).

[Bundeszentrale für politische Bildung 2014]Bundeszentrale für politische Bildung (2014). Wahl-O-Mat. `http://www.bpb.de/methodik/XQJYR3`. Cited 7 Feb 2014

[De Leeuw 2006] De Leeuw J (2006) Principal component analysis of binary data by iterated singular value decompositions. Computational Statistics and Data Analysis, 50(1: 2nd special issue on matrix computations and statistics), 21–39

[Diemer and Eßwein 2016] Diemer A, Eßwein B (2016) The Third Vote (Videofilm). `https://www.youtube.com/watch?v=TCkSYpF5es8`

[Feature selection 2017] Feature selection (2017). Wikipedia. `https://en.wikipedia.org/wiki/Feature_selection`

[KIT 2016] KIT (2016) The Third Vote — Die Stimme für Ihre politische Meinung. `http://studierendenwahl.econ.kit.edu/`

[Hogarty et al 2004] Hogarty KY, Kromrey JD, Ferron JM, Hines CV (2004) Selection of variables in exploratory factor analysis: an empirical comparison of a stepwise and traditional approach. Psychometrika 69, 593–611

[Husson et al 2011] Husson F, Lê S, Pagès J (2011) Exploratory Multivariate Analysis Using R. CRC Press, Boca Raton FL

[Jolliffe 1972] Jolliffe IT (1972) Discarding variables in a principal component analysis-I: artificial data. Applied Statistics 21, 160–173

[Jolliffe 1973] Jolliffe IT (1973) Discarding variables in a principal component analysis-II: real data. Applied Statistics 22, 21–31

[Jolliffe 2002] Jolliffe IT (2002) Principal Component Analysis, 2nd ed. Springer, New York

[Kano and Harada 2000] Kano Y, Harada A (2000) Stepwise variable selection in factor analysis. Psychometrika 65, 7–22

[Krzanowski 1987] Krzanowski WJ (1987). Selection of variables to preserve multivariate data structure using principal components. Applied Statistics 36, 22–33

[Kumar and Schneider 2016] Kumar NK, Schneider J(2016) Literature survey on low rank approximation of matrices. arXiv:1606.06511 [math.NA] `https://arxiv.org/abs/1606.06511`

[Kuroda et al 2011] Kuroda M, Iizuka M, Mori Y, Sakakihara M (2011) Principal components based on a subset of qualitative variables and its accelerated computational algorithm. Proc. 58th World Statistical Congress, 2011, Dublin. Int Statistical Inst, The Hague, December 2012

[McCabe 1975] McCabe GP (1975) Computations for variable selection in discriminant analysis. Technometrics 17, 103–109

[McCabe 1984] McCabe GP (1984) Principal variables. Technometrics 26 (2), 137–144

[McKay and Campbell 1982a] McKay RJ, Campbell NA (1982a) Variable selection techniques in discriminant analysis: I. Description. British Journal of Mathematical and Statistical Psychology 35, 1–29

[McKay and Campbell 1982b] McKay RJ, Campbell NA (1982b) Variable selection techniques in discriminant analysis: II. Allocation. British Journal of Mathematical and Statistical Psychology 35, 30–41

[Mori et al 2007] Mori Y, Iizuka M, Tarumi T, Tanaka Y (2007) Variable selection in principal component analysis. In: Härdle W, Mori Y, Vieu P (Eds) Statistical Methods for Biostatistics and Related Fields. Springer-Verlag, Berlin, 265–284

[Mori et al 2016] Mori Y, Kuroda M, Makino N (2016) Nonlinear Principal Component Analysis and Its Applications. Springer, Singapore

[Pacheco et al 2013] Pacheco J, Casado S, Porras S (2103) Exact methods for variable selection in principal component analysis: Guide functions and pre-selection. Computational Statistics and Data Analysis 57, 95–11

[Tangian 2014] Tangian A (2014) Mathematical Theory of Democracy. Springer, Berlin–Heidelberg.

[Tangian 2016] Tangian A (2016) The third vote experiment: VAA-based election to enhance policy representation of the KIT student parliament. Working paper series in economics No. 93 (September 2016). Karlsruhe Institute of Technology, Karlsruhe. `http://econpapers.wiwi.kit.edu/`

[Tangian 2017a] Tangian A (2017a) An election method to improve policy representation of a parliament. Group Decision and Negotiation (Forthcoming; available online 24.09.2016 `http://link.springer.com/article/10.1007/s10726-016-9508-4`)

[Tangian 2017b] Tangian A (2017b) Policy representation of a parliament: The case of the German Bundestag 2013 elections. Group Decision and Negotiation (Forthcoming; available online 09.09.2016 `http://link.springer.com/article/10.1007/s10726-016-9507-5`)

[Zheng et al 2010] Zheng Z et al (2010) Advancing Feature Selection Research – ASU Feature Selection Repository. Computer Science & Engineering, Arizona State University

Working Paper Series in Economics

recent issues

No. 100 *Andranik S. Tangian:* Selection of questions for VAAs and the VAA-based elections, January 2017

No. 99 *Dominik Rothenhäusler, Nikolaus Schweizer and Nora Szech:* Guilt in Voting and Public Good Games, November 2016

No. 98 *Eckhardt Bode, Stephan Brunow, Ingrid Ott and Alina Sorgner:* Worker personality: Another skill bias beyond education in the digital age, November 2016

No. 97 *Clemens Puppe:* The single-peaked domain revisited: A simple global characterization, November 2016

No. 96 *David Burka, Clemens Puppe, Laszlo Szepesvary and Attila Tasnadi:* Neural networks would `vote' according to Borda's rule, November 2016

No. 95 *Vladimir Korzinov and Ivan Savin:* Pervasive enough? General purpose technologies as an emergent property, November 2016

No. 94 *Francesco D'Acunto, Daniel Hoang and Michael Weber*: The effect of unconventional fiscal policy on consumption expenditure, October 2016

No. 93 *Andranik S. Tangian:* The third vote experiment: VAA-based election to enhance policy representation of the KIT student parliament, September 2016

No. 92 *Clemens Puppe and Arkadii Slinko:* Condorcet domains, median graphs and the single-crossing property, June 2016

No. 91 *Markus Höchstötter, Mher Safarian, Anna Krumetsadik:* Analysis of stochastic technical trading algorithms, June 2016

No. 90 *Nikolaus Schweizer and Nora Szech:* Optimal revelation of life-changing information, May 2016