A Note on Data Simulations for Voting by Evaluation

Antoine Rolland, Jean-Baptiste Aubin, Irène Gannaz, Samuela Leoni

A Note on Data Simulations for Voting by Evaluation

Antoine Rolland[1], Jean-Baptiste Aubin[2], Irène Gannaz[3], and Samuela Leoni[2]

[1]ERIC EA 3083, Université de Lyon, Université Lumière Lyon 2,
5 Pierre Mendès France, 69596 Bron Cedex, France
[2]Univ Lyon, INSA Lyon, UJM, UCBL, ECL, ICJ, UMR5208,
69621 Villeurbanne, France
[3]Univ. Grenoble Alpes, CNRS, Grenoble INP*, G-SCOP,
38000 Grenoble, France

December 6, 2022

Abstract

Traditionally, probabilistic analysis of voting rules supposes the use of simulation models to generate preferences data, like the Impartial Culture (IC) or Impartial and Anonymous Culture (IAC) models. Voting rules based on evaluation inputs rather than preference orders have been recently proposed, like majority judgment, range voting or approval voting. These voting rules deserve specific data out of the traditional framework. We propose in this paper several simulation models for generating evaluation-based voting inputs. These models can cope with dependent and non identical marginal distributions of the evaluations received by the candidates. A last part is devoted to fitting these models to real data sets.
Keywords: voting rules, Evaluation based voting rules, Simulation, IC model.

1 Introduction

Voting rules can be seen as functions which aim at determining a winner in a set of candidates considering preferences of a set of voters. Both social and mathematical approaches consider positive or negative properties satisfied by a given voting rule as a matter of interest. Studying the properties of voting rules

*Institute of Engineering Univ. Grenoble Alpes

can be done either in an axiomatic approach or in a probabilistic approach. The axiomatic approach supposes to determine which properties characterize a specific voting process, *i.e.* which properties are to be observed, and which are not, via formal theorems. The probabilistic approach aims at determining whether these properties are likely to be observed, *i.e.* determining the frequency of occurrence of such properties using a voting rule. This pragmatic approach is based on simulations. Several models exist, which have been well studied for years. One can refer to Diss and Kamwa [12] for a recent state-of-the-art of simulation techniques for a probabilistic approach of voting theory. Tideman and Plassmann [33, 34], Plassmann and Tideman [29], Green-Armytage et al. [19] contain examples of simulation-based studies of voting rules. All these models are appropriate to study voting rules using preference orders on candidates as input. But other voting rules have been recently proposed based on evaluations given by the voters about the candidates, using preference intensities, *i.e.* evaluations on a given scale, and not only preference orders. A complete review of both order preference-based and evaluation-based voting processes can be found in Felsenthal and Machover [16].

We propose in this paper to investigate several methods to simulate evaluation/notation data for studying evaluation-based voting rules. The aim is double. First, defining simulations, one can study the occurrence of statistical properties or paradoxes for different voting rules in a general context. Second, one can model real life votes, to study the behavior of voting rules in specific frameworks.

We first introduce evaluation-based voting methods in Section 2. We discuss about the relevance of Impartial Culture (IC) framework in evaluations. In this context, voters' preferences on each candidate are independent and identically distributed (i.i.d.). Section 3 presents i.i.d. models on evaluation, which hence yields IC models on preference orders. Dealing with evaluation rather than preferences, the IC is not the key assumption. It is more appropriate to discuss about the relaxation of either the identity of the distributions or the independence assumptions. Section 4 is devoted to voter's preferences on each candidate differently distributed (but still independent). In Section 5, some models where these distributions are dependent are explored. We introduce among others copula-based evaluations models and spatial models. A real data application is given in Section 6. This works ends with a discussion on relative benefits and drawbacks of these simulation models.

2 Evaluation based methods

Classical voting methods are based on preference rankings. Traditional use of these methods has a clear historical justification: before the computers, further preferences of voters couldn't be taken in consideration for practical reasons. Nevertheless, the information contained in these rankings is very limited. As a consequence, these classical methods based on preferences are vulnerable to numerous "paradoxes" [16] and impossibility theorems. The most famous is Arrow's impossibility theorem [2].

Alternative methods based on evaluations exist, which are more nuanced and keep more information. Obtained results are promising and stretch some limits of the classical methods. The three most famous methods based on evaluations are the approval voting, the range voting and the majority judgment. The approval voting (see [9] for a complete study) is maybe the most famous of these methods: each voter evaluates candidates on a scale of 2 grades, which is the simplest possible scale. The voter grades 1 if the candidate is acceptable, else 0. The voter can then votes for several candidates, even all of them or none of them, accordingly to his/her convictions. Note that this method is very simple to apply in practice. The two other methods are based on more nuanced classes of grading, which can be continuous or on a discrete scale. With the range voting, proposed by Smith [31], the winner is the candidate with the highest average grade. With the majority judgment, introduced by Balinski and Laraki [4, 5], the winner is the candidate with the highest median grade. Tie-break situations are, here, a matter of importance and are taken into account for example in [6] and [15].

Note that approval voting can be seen either as a range voting or as a majority judgment, with a grading scale reduced to a binary scale 0 or 1.

These evaluation-based voting methods can either be seen as particular cases of a more general voting methods family which is the deepest voting family. Deepest voting is a new promising family of social decision functions based on evaluations, which has been introduced and studied in [3]. Let us consider n voters and d candidates. Each voter can be seen as a point in $\mathbb{R}^d$ whose components are the grades for each candidate. The set of all the voters' grades is then a point cloud. The key idea of deepest voting is to consider the grades of the *most central* voter of the cloud, which can be find by maximizing a depth function [36]. The associated social decision function simply gives the grades of this innermost voter as output.

The recent interest in evaluation-based votes yields a need of simulation methods to study their properties. Hence, our objective is to extend preference simulation procedures to evaluations.

Notations

In the following, we will consider situations with n voters and d candidates. Each voter associates a grade in a set $\mathcal{E}$ to each candidate. Evaluation of voter j for candidate i will be denoted e_{ij}, for $i = 1, \ldots, d$, $j = 1, \ldots, n$. Observations $\{(e_{ij})_{i=1,\ldots,d},\, j = 1, \ldots, n\}$ are n independent realizations of a random variable $E = (E_1, \ldots, E_d)$, which takes values in $\mathcal{E}^d$. Defining a simulation setting can be seen as defining a multivariate probability distribution on $\mathcal{E}^d$. We will consider two cases with respect to the amount of information contained in the set $\mathcal{E}$:

- continuous grades: without loss of generality, $\mathcal{E} = [0, 1]$;
- discrete grades: without loss of generality, $\mathcal{E} = \{0, \ldots, K\}$ with $K \in \mathbb{N} \setminus \{0\}$.

Observe that preference orders can be deduced from evaluations: when there is no *ex-aequo*, a strict preference relation on the candidates can be obtained from the evaluations for each voter. In case of *ex-aequo*, a total weak order can be obtained from evaluations. Hence, simulating evaluations data lead naturally to data on orders of preference.

Evaluations versus preference orders

In voting framework based on preference orders, the Impartial Culture model (IC model) seems to be both the oldest and the most widely used in simulation model. Introduced by Guilbaud [20] in 1952, IC model supposes that each preference order on the candidates is equally likely to be selected by each voter. Each individual randomly and independently chooses their preferences, with a Uniform probability distribution on all orders. The Impartial and Anonymous Culture (IAC) model was introduced by Gehrlein and Fishburn [17] and Kuga and Hiroaki [21], and supposes that all preferences on the candidates concerning the whole set of voters are equally likely to appear. As Diss and Kamwa [12] noticed, "both models are based on a notion of equi-probability, but the elementary events are preference orders under IC and voting situations under IAC".

Both IC and IAC models are relevant for voting processes taking a set of preferences as input data. By contrast, evaluation-based voting processes need a vector of numerical evaluations on the set of candidates. An extension of IC and IAC models to evaluations is, hence, necessary.

IAC models. Let us first highlight the distinction between IC and IAC models. The difference is that IC models suppose that the voters' preferences are independent and identically distributed, whereas IAC models suppose that each configuration of preferences has the same probability of occurrence.

Let us consider an example with two voters and two candidates. Denote r_i the preferred candidate of voter v_i, with $r_i = j$ if voter v_i chooses candidate c_j, $i, j = 1, 2$. IC models consider that events $(r_1, r_2) \in \{(1,1),(1,2),(2,1),(2,2)\}$ are equally likely, with probability $1/4$ for each configuration. By contrast, IAC models suppose that preferences $(1,1),(1,2),(2,2)$ all occur with the same probability, $1/3$. This is due to the fact that events $(r_1, r_2) = (1,2)$ and $(r_1, r_2) = (2,1)$ give the same preferences.

The counterpart of IAC with evaluation-based models is less tractable than IC. When $\mathcal{E}$ is continuous, the evaluations may be associated with a continuous distribution. In that case, a given situation with d candidates has a null probability to appear. Therefore, the definition of IC and IAC are not relevant. When $\mathcal{E}$ is discrete, generating random evaluation situations *e.g.* with a Uniform probability seems intractable as the number of possibilities grow exponentially with the number of candidates, evaluation levels and/or voters. For example, with $n = 100$ voters giving an evaluation on a scale with $K = 7$ grades, for $d = 3$ candidates, there are $(7^3)^{100} = 3.4 \cdot 10^{253}$ possibilities. As stated in [27], generating random preference orders in an IAC framework is very difficult when the number of candidates is greater

than 3. Generating evaluation situations in an IAC framework is even more difficult. Therefore this issue deserves special studies that overcome the aim of this paper and should be treated in a forthcoming work.

IC models. Based on the above considerations, IC models seems much more appropriate to generalize to evaluation-based voting processes. The natural extension is to consider independent and identically distributed (i.i.d.) distributions of the d random variables $E_1, \dots E_d$. Indeed, in such a case, the preferences resulting from the evaluations will satisfy an IC model. It seems obvious that differently distributed variables $E_i, i = 1, \dots, d$, yields non IC model on preferences. One can then wonder if Impartial Culture is obtained if and only if $E_1, \dots E_d$ are identically distributed. We will show that dependent and identically distributed variables are not always leading to IC model on preferences. Hence, reasoning on the resulting model on preference orders seems too reducing for evaluation-based processes. It appears more appropriate to discuss about the independence hypothesis and the identity of marginal distributions, which provides finer information on the evaluations.

3 Independent and identically distributed evaluation models

The simplest case of IC modeling is when voters' preference orders on each candidate are i.i.d. A natural extension to evaluation-based voting process is to consider that voters' evaluations of each candidate are i.i.d., which we will call an Independent and Identically Distributed evaluations (Ev-IID) model.

Definition 1. Ev-IID model

The Independent Identically Distributed evaluations (Ev-IID) model based on a distribution $\mathcal{D}$ on $\mathcal{E}$ is such that random variables $E_1, \dots, E_d$ are independent and identically distributed.

For the resulting preference orders, the i.i.d. character is still satisfied.

Proposition 1. *Preference orders obtained from Ev-IID models follow an IC model.*

The proof of Proposition 1 is obvious and, thus, omitted.

The Ev-IID model does not make any assumption on the distribution used in the model.

3.1 Ev-IID continuous models

When the evaluations are continuous, that is, $\mathcal{E} = [0, 1]$, we propose in the following to use Uniform, truncated Normal or Beta distributions. Other distributions can also be considered, but will not be studied in this paper for the sake of brevity.

- **Ev-IID Uniform model**

  The Ev-IID Uniform model is an Ev-IID model with $E_i \sim \mathcal{U}[0, 1]$ for all $i = 1, \dots, d$, where $\mathcal{U}[0, 1]$ is the Uniform distribution on $[0, 1]$.

- **Ev-IID truncated Normal model**

  The Ev-IID truncated Normal model is an Ev-IID model with $E_i \sim \mathcal{N}_{\mathcal{T}}(\mu, \sigma)$ for all $i = 1, \dots, d$, where $\mathcal{N}_{\mathcal{T}}(\mu, \sigma)$ is the Normal distribution, with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$, truncated between 0 and 1 (see [30] for details).

- **Ev-IID Beta model**

  The Ev-IID Beta model is an Ev-IID model with $E_i \sim \mathcal{B}(\alpha, \beta)$ for all $i = 1, \dots, d$, where $\mathcal{B}(\alpha, \beta)$ is the Beta distribution of parameters α and β, with $\alpha > 0$ and $\beta > 0$.

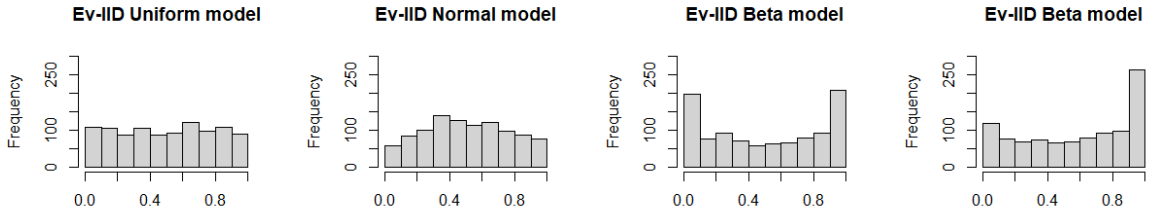

Figure 1: Simulation examples of Ev-IID evaluations on one candidate for $n = 1000$ voters with respectively Uniform distribution, truncated Normal distribution (with $\mu = 0.5$ and $\sigma = 0.35$) and Beta distribution (respectively with $\alpha = \beta = 0.5$ and with $\alpha = 0.7$ and $\beta = 0.5$).

Examples of simulations using Ev-IID Uniform, truncated Normal and Beta models are presented in Figure 1 for $d = 1$ candidate and $n = 1000$ voters. It can be seen that these three families of distributions cover a large scope of distributions. Each one has a different shape. Note that the Beta distribution, in particular, allows asymmetric distributions.

3.2 Ev-IID discrete models

When the evaluations are discrete, i.e. $\mathcal{E} = \{0, \dots, K\}$, we propose in the following to use discrete Uniform, Binomial or Beta-Binomial distributions, which can be seen as the discrete counterpart of the Uniform, truncated Normal and Beta continuous distributions.

- **Ev-IID discrete Uniform model**

  The Ev-IID discrete Uniform model is an Ev-IID model with $E_i \sim \mathcal{U}\{0, \dots, K\}$ for all $i = 1, \dots, d$, where $\mathcal{U}\{0, \dots, K\}$ is the discrete Uniform distribution on $\{0, \dots, K\}$.

- **Ev-IID Binomial model**

  The Ev-IID Binomial model is an Ev-IID model with $E_i \sim \mathcal{B}(K, p)$ for all $i = 1, \dots, d$, where

$\mathcal{B}(K, p)$ is the Binomial distribution with parameters K and $p \in (0, 1)$.

- **Ev-IID Beta-Binomial model**

  The Ev-IID Beta-Binomial model is an Ev-IID model with $E_i \sim \mathcal{BB}(K, \alpha, \beta)$ for all $i = 1, \ldots, d$, where $\mathcal{BB}(K, \alpha, \beta)$ is the Beta-Binomial distribution of parameters K, α and β, with $\alpha > 0$ and $\beta > 0$. The Beta-Binomial distribution is the Binomial distribution in which the probability of success at each of K trials follows a Beta distribution with parameters α and β.

Examples of simulations using Ev-IID discrete Uniform, Binomial and Beta-Binomial models are presented in Figure 2 for $d = 1$ candidate and $n = 1000$ voters. It illustrates that discrete Uniform, Binomial and Beta-Binomial models can, indeed, be seen as discrete versions of respectively continuous Uniform, truncated Normal and Beta models.

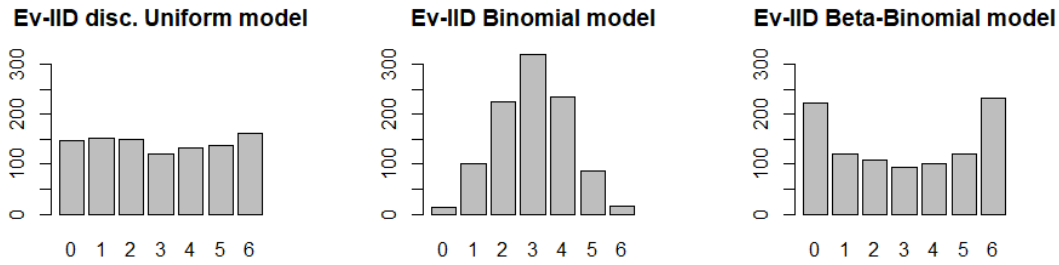

Figure 2: Simulation examples of Ev-IID discrete evaluations with $K = 7$ levels and one candidate for $n = 1000$ voters, using respectively Uniform distribution, Binomial distribution (with $p = 0.5$) and Beta-Binomial distribution (with $\alpha = \beta = 0.5$).

4 Independent and differently distributed evaluation models

As far as the evaluations are i.i.d., Ev-IID models yield IC models when considering only preference orders. We propose hereafter alternatives to Ev-IID models. The first possibility is to deal with independent and not identically distributed distributions. Such models will be called Independent and Differently Distributed evaluations (Ev-IDD) Models. We present these models in this Section.

The second alternative deals with dependent distributions between candidates. This is the object of Section 5.

In social choice, there occur situations where some candidates divide the voters population between strongly pro and strongly cons voters, whereas some other candidates are less polarizing. This later can

be expressed by a smaller dispersion of the evaluations. Figure 3 displays an example of such a situation, with two different profiles of evaluations. In this Section, the evaluation distribution for each candidate is independent from the others, but the distributions are not identical. Such models, even if independence holds, allows different evaluation distributions for each candidates, and they will be denoted as Ev-IDD models.

Definition 2. Ev-IDD model

The Evaluation Independent and Differently Distributed (Ev-IDD) model based on a distribution $\mathcal{D}$ on $\mathcal{E}$ is such that the random variables $E_1, \ldots, E_d$ are independent and non identically distributed.

The marginal distributions can be different by changing parameters of a given distribution family, or by changing the distribution family. We choose here to present only models with the same family for all candidate, changing only parameters. This choice is motivated by the simplicity of use of such models.

4.1 Ev-IDD continuous models

For all $i = 1, \ldots, d$, let E_i be independent random variables on a continuous set $\mathcal{E} = [0, 1]$.

We propose two distributions.

- **Ev-IDD truncated Normal model**

  The Ev-IDD truncated Normal model is defined by

  $$E_i \sim \mathcal{N}_{\mathcal{T}}(\mu_i, \sigma_i), \text{ for all } i = 1, \ldots, d,$$

  where $\mathcal{N}_{\mathcal{T}}(\mu_i, \sigma_i)$ is the Normal distribution with mean μ_i and standard deviation σ_i, truncated to the interval $[0, 1]$, with $\mu_i \in \mathbb{R}$, $\sigma_i > 0$, for $i \in \{1, \ldots, d\}$.

- **Ev-IDD Beta model**

  The Ev-IDD Beta model is defined by

  $$E_i \sim \mathcal{B}(\alpha_i, \beta_i), \text{ for all } i = 1, \ldots, d,$$

  where $\mathcal{B}(\alpha_i, \beta_i)$ is the Beta distribution of parameters α_i and β_i, with $\alpha_i > 0$, $\beta_i > 0$, for $i \in \{1, \ldots, d\}$.

With these models, the shape of the distributions are, hence, the same as with Ev-IID truncated Normal and Ev-IID Beta models, displayed in Figure 1. Yet, each candidate can have a different distribution.

4.2 Ev-IDD discrete models

Similarly, for all $i = 1, \dots, d$, let E_i be independent random variables on a discrete set $\mathcal{E} = \{0, \dots, K\}$. We propose two distributions.

- **Ev-IDD Binomial model**
  The Ev-IDD Binomial model is defined by

$$E_i \sim \mathcal{B}(K, p_i), \text{ for all } i = 1, \dots, d,$$

  where $\mathcal{B}(K, p_i)$ is the binomial distribution with parameters K and p_i, with $(p_i)_{i=1,\dots,d} \in (0,1)^d$.

- **Ev-IDD Beta-Binomial model**
  The Ev-IDD Beta-Binomial model is defined by

$$E_i \sim \mathcal{BB}(K, \alpha_i, \beta_i), \text{ for all } i = 1, \dots, d,$$

  where $\mathcal{BB}(K, \alpha_i, \beta_i)$ is the Beta-Binomial distribution of parameters α_i and β_i, with $\alpha_i > 0$, $\beta_i > 0$, for $i \in \{1, \dots, d\}$.

The shape of the marginal distributions are, hence, the same as with Ev-IID Binomial and Ev-IID Beta-Binomial models, displayed in Figure 2.

As noticed previously, Beta (discrete) distributions and Beta-Binomial (continuous) distributions enable to consider various shapes of distributions. As an example, the evaluations of polarizing and non polarizing candidates simulated in Figure 3 result from two Beta-Binomial distributions, with different parameters.

5 Dependent evaluation models

The second alternative to extend Ev-IID models (and Ev-IDD models) is to remove the independence hypothesis between the evaluations of each candidate. In such a case, the marginal distributions can be either identical or different. Hence, these models can be considered as Dependent and Independently Distributed evaluations models (Ev-DID). But considering non identical distributions provide Dependent and Differently Distributed evaluations models (Ev-DDD).

Definition 3. Ev-DID model

The Dependent and Identically Distributed evaluations (Ev-DID) model based on a distribution $\mathcal{D}$ on $\mathcal{E}$ is such that random variables $E_1, \dots, E_d$ are dependent and identically distributed.

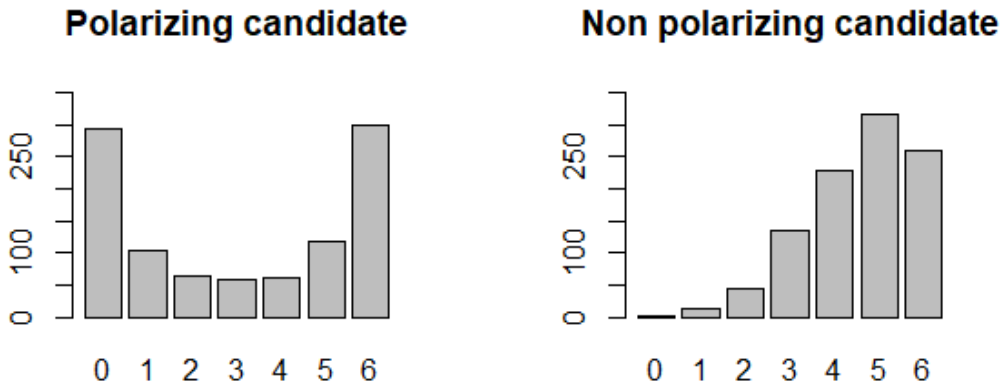

Figure 3: Simulation examples of polarizing and non polarizing candidates on $n = 1000$ voters obtained with Beta-Binomial distributions, with respectively $\alpha = 0.3$ and $\beta = 0.3$ for the polarizing candidate and $\alpha = 10$ and $\beta = 3$ for the non polarizing candidate.

Definition 4. Ev-DDD model

The Dependent and Differently Distributed evaluations (Ev-DDD) model based on a distribution $\mathcal{D}$ on $\mathcal{E}$ is such that the random variables $E_1, \ldots, E_d$ are dependent and non identically distributed.

Such models are detailed below. Note that contrary to Ev-IID models, Ev-DID models do not imply IC models on preference orders as stated in Proposition 2.

Proposition 2. *Preference orders obtained from Ev-DID models do not always follow an IC model.*

The counterexample in Table 1 shows two candidates having identical marginal distribution, but $P(E_1 > E_2) = \frac{1}{2} \neq P(E_2 > E_1) = \frac{1}{4}$, where E_1 is the evaluation of candidate 1 and E_2 is the evaluation of candidate 2.

5.1 Multinomial and Dirichlet models

Ev-IID and Ev-IDD models suppose that each voter evaluates all the candidates independently, without any constraint on the evaluations vector. However, an alternative evaluation process consists in dividing a total score on the candidate evaluations, such that the evaluations sum on the set of candidates is the same for each voter. The evaluations of each candidate are identically distributed, as the model is totally

E_2 \ E_1	0	1	2	$P(E_2)$
0	0	1/4	0	1/4
1	0	1/4	1/4	1/2
2	1/4	0	0	1/4
$P(E_1)$	1/4	1/2	1/4	

Table 1: Example of a joint distribution form a Ev-DID model leading to a preference order non-IC Model

symmetric, but they are not independent as there is a link between the evaluations of candidates given by a voter.

Both discrete and continuous models are available, using multinomial distribution in the discrete case and Dirichlet distribution for the continuous case.

Ev-DDD multinomial models. When the evaluations are discrete, i.e. $\mathcal{E} = \{0, \dots, K\}$, the Ev-DDD multinomial model is the following.

Definition 5. Ev-DDD multinomial model

The Ev-DDD multinomial model is defined by $(E_1, \dots, E_d) \sim \mathcal{M}\{K, p_1, \dots, p_d\}$ *for all* $i = 1, \dots, d$*, where* $\mathcal{M}\{K, p_1, \dots, p_d\}$ *is the multinomial distribution of parameters* K *and* $p_1, \dots, p_d$*, with for all* $i = 1, \dots, d$*,* $p_i \geq 0$ *and* $\sum_{i=1}^{d} p_i = 1$*.*

If parameters p_i are different for different i, then the evaluations are not identically distributed, the associated multinomial model is then an Ev-DDD model. If for all $i = 1, \dots, d$, $p_i = 1/d$ then the associated multinomial model is an Ev-DID model.

An example of simulation with 3 candidates using different probabilities for the candidates is shown in Figure 4.

Ev-DDD Dirichlet models. The continuous counterpart of the multinomial distribution is obtained through the use of a Dirichlet distribution on $\mathcal{E} = [0, 1]^d$ as follows (see [26] for details about Dirichlet distribution).

Definition 6. Ev-DDD Dirichlet model

The Ev-DDD Dirichlet model is defined by $(E_1, \dots, E_d) \sim \mathcal{D}ir\{p_1, \dots, p_d\}$ *for all* $i = 1, \dots, d$*, where* $\mathcal{D}ir\{p_1, \dots, p_d\}$ *is the Dirichlet distribution of parameters* $p_1, \dots, p_d$*, with for all* $i = 1, \dots, d$*,* $p_i \geq 0$ *and* $\sum_{i=1}^{d} p_i = 1$*..*

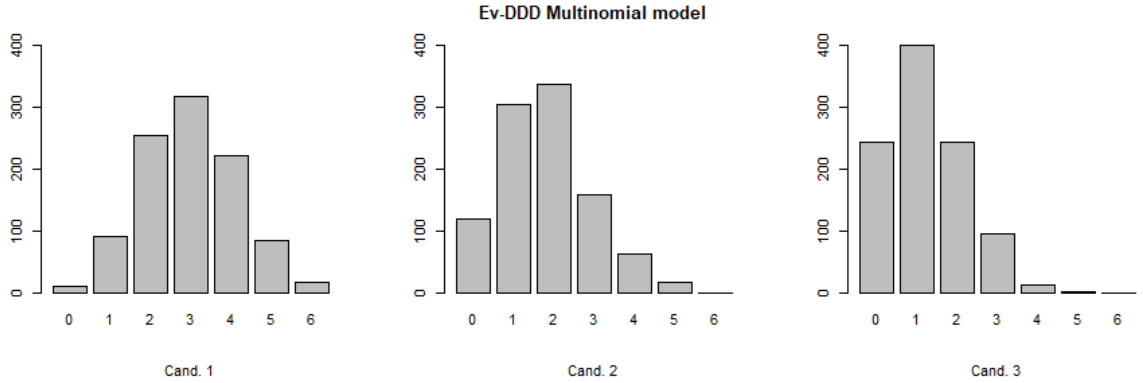

Figure 4: Histograms of evaluations for each of $d = 3$ candidates ($n = 1000$ voters) in a Ev-DDD Dirichlet model with probability vector (0.5, 0.3, 0.2).

If parameters p_i are different for different i, then the Dirichlet model is an Ev-DDD model. If for all $i = 1, \ldots, d$, $p_i = 1/d$, then the Dirichlet model is an Ev-DID model.

An example of simulation with 3 candidates using different probabilities for the candidates is shown in Figure 5. The links between the evaluations of the three candidates is shown in Figure 6.

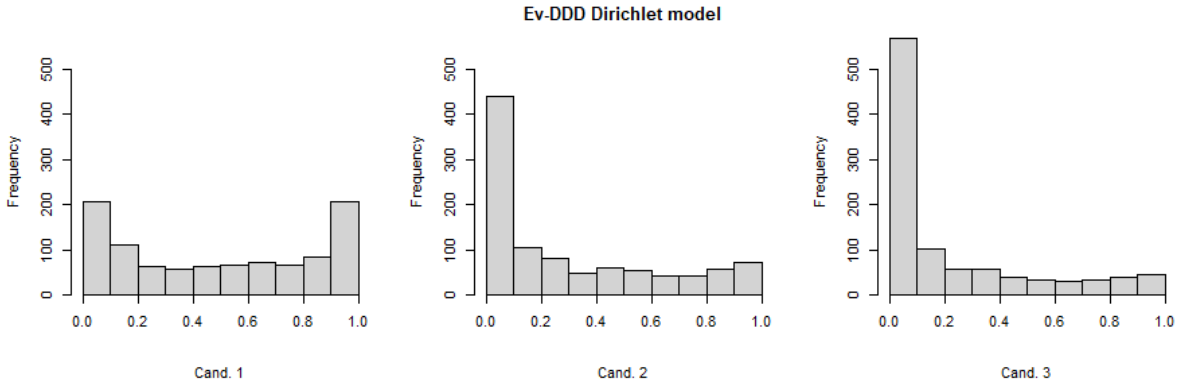

Figure 5: Histograms of evaluations for each of $d = 3$ candidates ($n = 1000$ voters) in a Ev-DDD Dirichlet model with probability vector (0.5, 0.3, 0.2).

Ev-DID Cumulative Dirichlet model. Suppose that the candidates are so different that each voter should agree with at least one candidate, and strongly disagree with at least another one. Therefore, evaluations should be set to 1 for the best candidate score, and 0 for the the worst candidate score. In

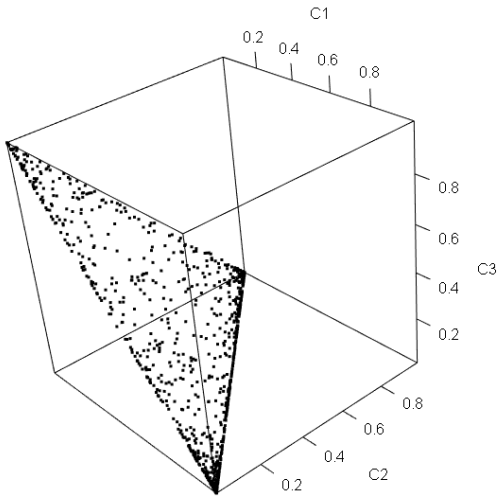

Figure 6: Histograms of evaluations for each of $d = 3$ candidates ($n = 1000$ voters) in a Ev-DDD Dirichlet model with probability vector (0.5, 0.3, 0.2).

this case, the distribution is a cumulative Dirichlet distribution.

Definition 7. Ev-DID Cumulative Dirichlet model
Let $\mathcal{E} = [0,1]^d$. The Cumulative Dirichlet model is defined by

$$\begin{cases} E_{\pi(1)} = 0 \\ E_{\pi(i)} = \sum_{k=1}^{i-1} \delta_k, \quad \textit{for all } i = 2, \dots, d. \end{cases}$$

where $(\delta_1, \dots, \delta_{d-1})$ are independent variables from a Dirichlet distribution $Dir(1, \dots, 1)$ on $[0,1]^{d-1}$, and π a random permutation on $\{1, \dots, d\}$.

Note that, even if the cumulative Dirichlet model is not based on independent distributions for evaluations of candidates, the related distributions for the preferences are i.i.d. That is, the cumulative Dirichlet model yields an IC model on preferences.

An example of evaluations from a Cumulative Dirichlet model is displayed on Figure 7 and on Figure 8. $d = 4$ candidates are considered, with $n = 1000$ voters. Figure 7 shows that all the marginal distributions, that is, the distributions of the evaluations for each candidate, are the same. They have a symmetric shape, with a high occurrence of extreme values 0 and 1. On Figure 8, the evaluations of the

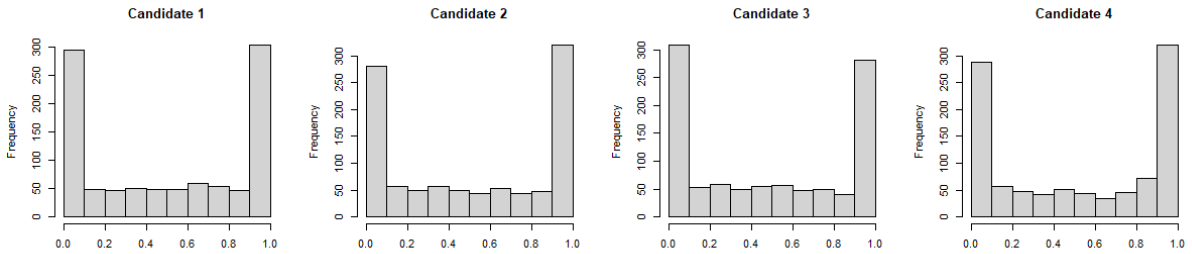

Figure 7: Histograms of evaluations for each of $d = 4$ candidates ($n = 1000$ voters) in a Ev-DID Cumulative Dirichlet model.

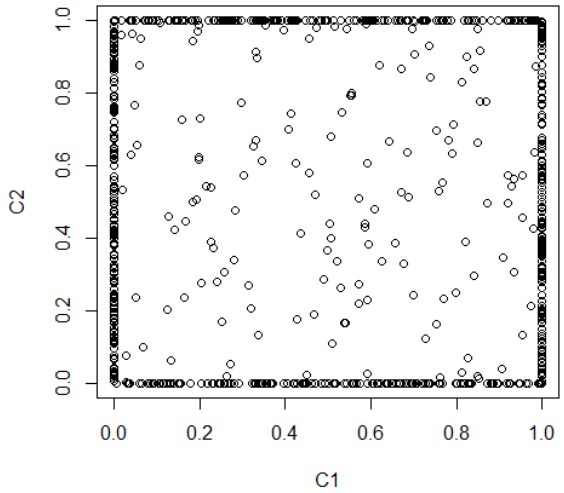

Figure 8: Plot of evaluations of Candidate 2 vs evaluations of Candidate 1 ($n = 1000$ voters) in a Cumulative Dirichlet Ev-DDD model.

candidates do not look correlated. Nevertheless, the evaluations are not independent by definition, since $P(E_1 = 1 \cap E_2 = 1) = 0$ and $P(E_1 = 1) = P(E_2 = 1) \neq 0$.

5.2 Ev-DDD truncated Normal model

The idea now is to extend General Ev-IDD truncated Normal models to Ev-DDD models by introducing correlations between the evaluations for each candidate. This corresponds to correlations between variables E_i, $i = 1, \ldots, d$. It comes naturally with the Normal distribution, using a covariance matrix.

Definition 8. *The Ev-DDD truncated Normal model is defined by*

$$E = (E_1, \ldots, E_d) \sim \mathcal{N}_{\mathcal{T}}(\mu, \Sigma),$$

where $\mathcal{N}_{\mathcal{T}}(\mu, \Sigma)$ is the d-multivariate Normal distribution with mean μ and covariance matrix Σ, with each component truncated to the interval $[0, 1]$, with $\mu \in \mathbb{R}^d$, and Σ a positive definite matrix in $\mathbb{R}^{d \times d}$.

Observe that Ev-IDD truncated Normal model is a subclass of the model of Definition 8, obtained when Σ is a diagonal matrix, $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_d)$. Ev-IID truncated Normal models correspond to diagonal matrices Σ with the same entry in all the diagonal.

Figure 9 presents three different simulations based on the same marginal distributions for two candidates (truncated Normal distributions with a mean equal to 0.7 and a standard deviation equal to 0.3): the first one with independent (uncorrelated) evaluations for candidate 1 and candidate 2, the second one with a positive correlation with $\rho = 0.8$ and the third one with a negative correlation with $\rho = -0.8$.

This model has the advantage to be easy to simulate and to enable to define in a very comprehensive way the dependence between the evaluations. Nevertheless, this model implies strong restrictions on the shape of the marginals. To get rid of these restrictions, we propose below to use copula.

5.3 Copula-based Ev-DID and Ev-DDD models

To generalize Ev-DDD truncated Normal model to any distributions (including discrete and continuous sets $\mathcal{E}$) and introduce dependencies between candidates, one can use copulas. If the distributions of the evaluations of each candidate are not independent, a multivariate copula can be used to take into account their dependencies. In a nutshell, a copula is a multivariate cumulative distribution function which has all its margins uniformly distributed on the unit interval. It can also be applied on transform of random variables to generate dependence with non uniform marginals. See [18, 25] for a formal presentation of the subject.

Definition 9. Copula Ev-DDD models

The Copula Ev-DDD models are defined by

$$E = (E_1, \ldots, E_d) \sim C(\delta_1, \ldots \delta_d),$$

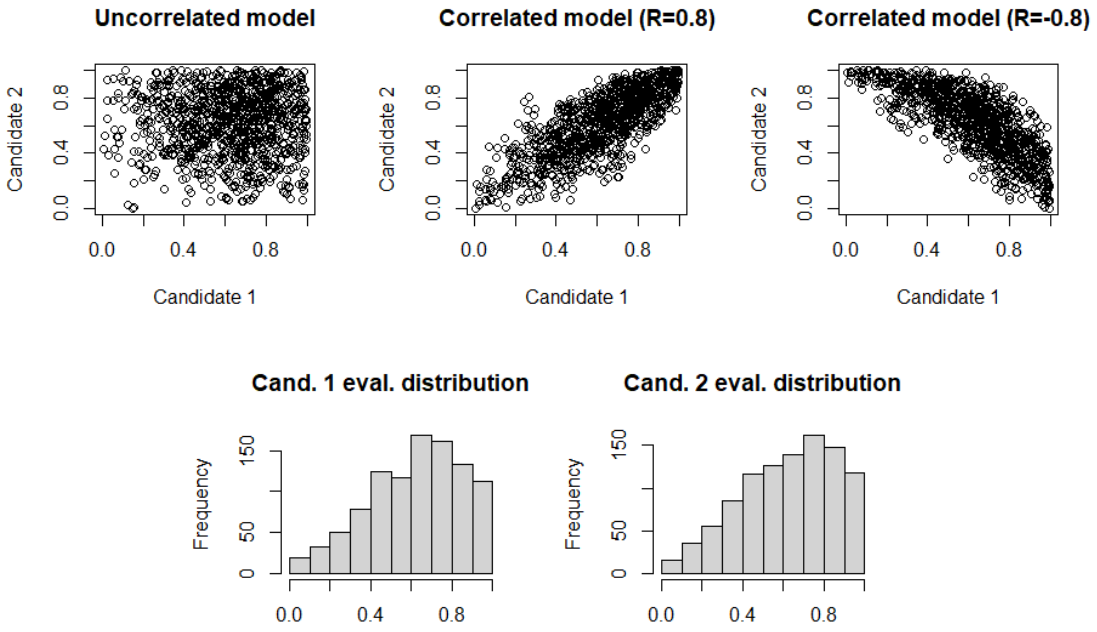

Figure 9: Simulation examples of Ev-DDD truncated Normal evaluations of $d = 2$ candidates by $n = 1000$ voters in 3 situations : uncorrelated (up left), positive correlated (up center) and negative correlated (up right) evaluations. Marginal distributions are given on the bottom line.

where C is a multivariate copula and $\delta_1, \ldots, \delta_d$ are distributions on $\mathcal{E}$.

Figure 10 presents a simulation based on the use of two different marginal distributions for two candidates (Beta distribution with parameters (0.7,0.5) for the first candidate and (0.5,0.7) for the second one), using a Gaussian copula of parameter 0.8 (see below for details).

A strength of copulas is that they allow any marginal distributions. Therefore, the model should specify both the marginal distribution for each candidate (*e.g.* with continuous or discrete distributions of Section 2), and the copula used to model the dependencies between variables. In the following, we distinguish the cases of a continuous set $\mathcal{E}$ and a discrete set $\mathcal{E}$.

5.3.1 Copula Ev-DID and Ev-DDD continuous models

First, one has to choose the copula. Gaussian copulas offer a simple way to model dependencies between each pairs of candidates, through the correlation coefficients. The dependencies of the copula between

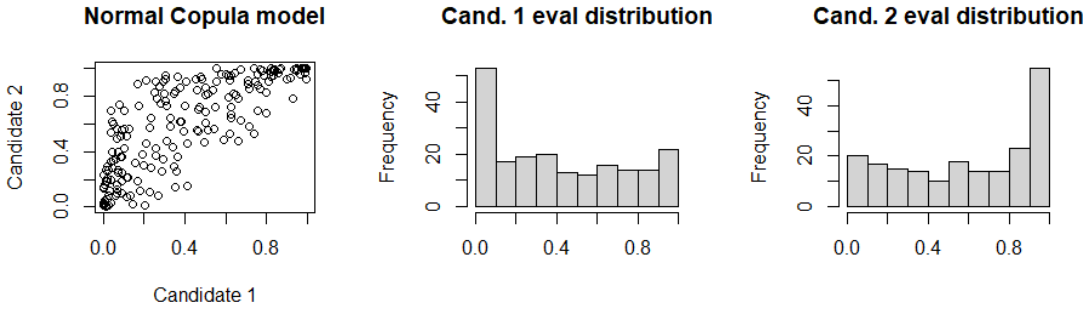

Figure 10: Simulation examples of Ev-DDD Copula evaluations for $d = 2$ candidates and $n = 200$ voters

two variables are exactly characterized by the correlation coefficients. Another interesting copula class is the checkerboard copula class [10], which represents a good compromise between the richness of the expression and the complexity of the model. See *e.g.* [14] for a discussion on the choice of a copula.

Next, marginal distributions must be chosen. One can consider the same distribution for each candidate and obtain Ev-DID models, or different distributions following Ev-DDD models. As before, we will focus on the case where each marginal belongs to the same family, with possibly different parameters, but different families can be used for different marginals in the same model. We propose three distributions.

- **Copula Ev-DID Uniform model**

  The Copula Ev-DID Uniform model is defined by

  $$E = (E_1, \dots, E_d) \sim C(\delta, \dots \delta),$$

  where C is a multivariate copula on $[0,1]^d$ and δ is the Uniform distribution on $\mathcal{E} = [0,1]$.

- **Copula Ev-DDD truncated Normal model**

  The Copula Ev-DDD truncated Normal model is defined by

  $$E = (E_1, \dots, E_d) \sim C(\delta_1, \dots \delta_d),$$

  where C is a multivariate copula on $[0,1]^d$ and δ_i is the truncated Normal distribution $\mathcal{N}_{\mathcal{T}}(\mu_i, \sigma_i)$ with mean μ_i and standard deviation σ_i, truncated to the interval $[0,1]$, with $\mu_i \in \mathbb{R}$, $\sigma_i > 0$, for $i \in \{1, \dots, d\}$.

- **Copula Ev-DDD Beta model**

  The Copula Ev-DDD Beta model is defined by

  $$E = (E_1, \dots, E_d) \sim C(\delta_1, \dots \delta_d),$$

where C is a multivariate copula on $[0,1]^d$ and δ_i is the Beta distribution $\mathcal{B}(\alpha_i, \beta_i)$ of parameters α_i and β_i, with $\alpha_i > 0$, $\beta_i > 0$, for $i \in \{1, \dots, d\}$.

Note that the copula-based Uniform models are Ev-DID models. Indeed, they belong to the Impartial Culture setting since all candidates have the same distributions. With Copula Ev-DDD truncated Normal and Ev-DDD Beta models, the marginal distributions are the same as with the associated Ev-IDD models. Yet, dependence between the evaluation has been added through the copula C.

5.3.2 Copula Ev-DID and Ev-DDD discrete models

Evaluations on discrete scales need the use of specific discrete copulas for simulation, as copulas are different for discrete and continuous cases. Among others, pair-copulas [28] and Gaussian copulas [7] have been proposed to simulate dependent discrete data, and therefore can be used also to model discrete evaluations in a social choice framework.

Following previous discussions, we propose three models based on the same distributions as above.

- **Copula Ev-DID Uniform model**
  The Copula Ev-DID Uniform model is defined by
  $$E = (E_1, \dots, E_d) \sim C(\delta, \dots \delta),$$
  where C is a multivariate copula on $[0,1]^d$ and δ is the Uniform distributions on $\mathcal{E} = \{0, \dots, K\}$.

- **Copula Ev-DDD Binomial model**
  The Copula Ev-DDD Binomial model is defined by
  $$E = (E_1, \dots, E_d) \sim C(\delta_1, \dots \delta_d),$$
  where C is a multivariate copula on $[0,1]^d$ and δ_i is the Binomial distribution $\mathcal{B}(K, p_i)$ with parameters K and p_i, $i = 1, \dots, d$, with $(p_i)_{i=1,\dots,d} \in (0,1)^d$.

- **Copula Ev-DDD Beta-Binomial model**
  The Copula Ev-DDD Beta-Binomial model is defined by
  $$E = (E_1, \dots, E_d) \sim C(\delta_1, \dots \delta_d),$$
  where C is a multivariate copula on $[0,1]^d$ and δ_i is the Beta-Binomial distribution $\mathcal{BB}(K, \alpha_i, \beta_i)$ of parameters α_i and β_i, with $\alpha_i > 0$, $\beta_i > 0$, for $i \in \{1, \dots, d\}$.

Except for the Copula Ev-DID Uniform discrete and continuous models, each of these models allows for different marginal distributions, for the evaluations of each candidate. Additionally to this non identical setting, the dependence modeling, through the copula, yields a large scope of models. These copula-based models appear very rich and adapted for covering much framework of simulations of evaluations.

5.4 Spatial models

Spatial voting simulations have been developed following the early work of [13]. The model is based on the use of an euclidean distance between the candidates and the voters, which are described in the same uni- or multi-dimensional space: the smaller the distance, the greater the preference. Tideman and Plassmann [32] conclude that a spatial model "describes the observations in data sets much more accurately" than other models.

Let k be a given dimension parameter. Parameter k should be seen as the number of latent characteristics which are used to build an opinion on the candidates. Typically, $k = 2$ or 3, see [1] for a discussion on the choice of k. Voters $v_1, \ldots, v_n$ and candidates $c_1, \ldots, c_d$ are then randomly uniformly generated as points inside the hypercube $[0,1]^k$. Spatial voting is next based on the distances between the generated points. The closer a voter to a candidate, the higher their evaluation of this candidate.

Definition 10. Ev-DDD Spatial models

Let $k \in \mathbb{N} \setminus \{0\}$, $c_1, \ldots, c_d, v_1, \ldots, v_n$ be independent realizations from a distribution on $[0,1]^k$. The Ev-DDD spatial model for the evaluation e_{ij} of candidate c_i by voter v_j is defined as

$$\forall i \in \{1, \ldots, d\}, \forall j \in \{1, \ldots, n\}, \quad e_{ij} = f(d(c_i, v_j))$$

where d is a distance between c_i and v_j and f a non-increasing function mapping $\mathbb{R}^+$ to $[0,1]$.

Typically, an intuitive spatial simulation model is given by

- $c_1, \ldots, c_d, v_1, \ldots, v_n$ obtained with a Uniform distribution on $[0,1]^k$,
- $\forall i \in 1, \ldots, d, \forall j \in 1, \ldots, n$, $e_{ij} = max\{0, (1 - \ell \times d_e(c_i, v_j))\}$ with d_e the euclidean distance. The parameter ℓ defines the decreasing rate of the evaluations with respect to the distance. For example, ℓ greater than 2 ensures that a voter being on the frontier of the unit cube will give a null score to a candidate who is on the center of the unit cube.

Other choices than the Uniform distribution on $[0,1]^k$ are possible, as proposed for example in [24], as well as other choices than the euclidean distance. We choose here to focus on the Uniform distribution and the euclidean distance for clarity. Other functions f are also possible, as for example the sigmoïd: $\forall i \in 1, \ldots, d, \forall j \in 1, \ldots, n$, $e_{ij} = (1 + e^{\lambda(\beta d_e(c_i, v_j) - 1)})^{-1}$, $\lambda > 0$ and $\beta > 0$. Figure 11 presents an example of such a function for $\lambda = 5$ and $\beta = 2$.

For a given position of candidates in $[0,1]^k$, the spatial model is clearly an Ev-DDD model, since the marginal distributions are different (see for example Figure 12) and potentially correlated (see for example Figure 13).

An example of evaluations obtained through a spatial model is presented in Figure 14. Two candidates and $n = 100$ voters have been randomly generated in $[0,1]^2$, that is, with $k = 2$. Evaluations are then

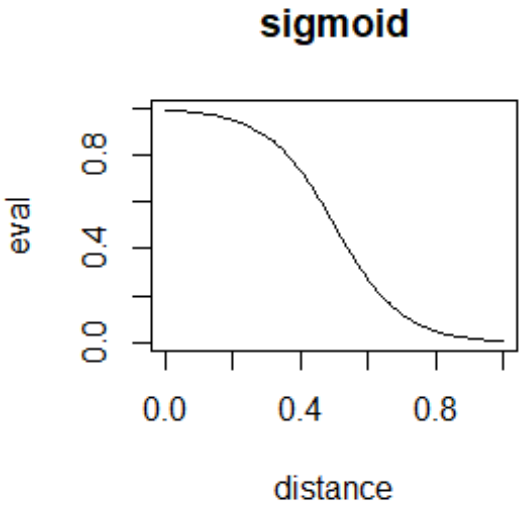

Figure 11: Example of sigmoïd transformation with $\lambda = 5$ and $\beta = 2$.

obtained from the euclidean distance of each voter to each candidate. We propose in Figure 12 histograms of three different ways to obtain evaluations, based on the same spatial situation (the one described in Figure 14), for each candidate:

- evaluations in model 1 are obtained using $e_{ij} = max\{0, (1 - 2 \times d_e(c_i, v_j))\}$,
- evaluations in model 2 are obtained using $e_{ij} = (1 + e^{\lambda(\beta d_e(c_i,v_j)-1)})^{-1}$ with $\lambda = 5$ and $\beta = 2$,
- evaluations in model 3 are obtained using $e_{ij} = (1 + e^{\lambda(\beta d_e(c_i,v_j)-1)})^{-1}$ with $\lambda = 2$ and $\beta = 2$.

Spatial models with discrete evaluations on $\{0, \ldots, K\}$ can easily be obtained from continuous models by dividing the $[0, 1]$ interval onto the $K + 1$ intervals. Then if the continuous grade obtained with the spatial model belongs to the l^{th} interval, the discrete evaluation is set to $l - 1$. That is, for a continuous evaluation e_{ij} on $[0, 1]$, we consider $\lfloor (K + 1)e_{ij} \rfloor$ as the resulting discrete evaluation, where $\lfloor u \rfloor$ denotes the smallest integer lower than u. The spatial interpretation of such a process is to determine K spheres centered on the candidate and to give the evaluation of K if the voter is into the smallest sphere, $K - 1$ for the second smallest sphere and so on, until 0.

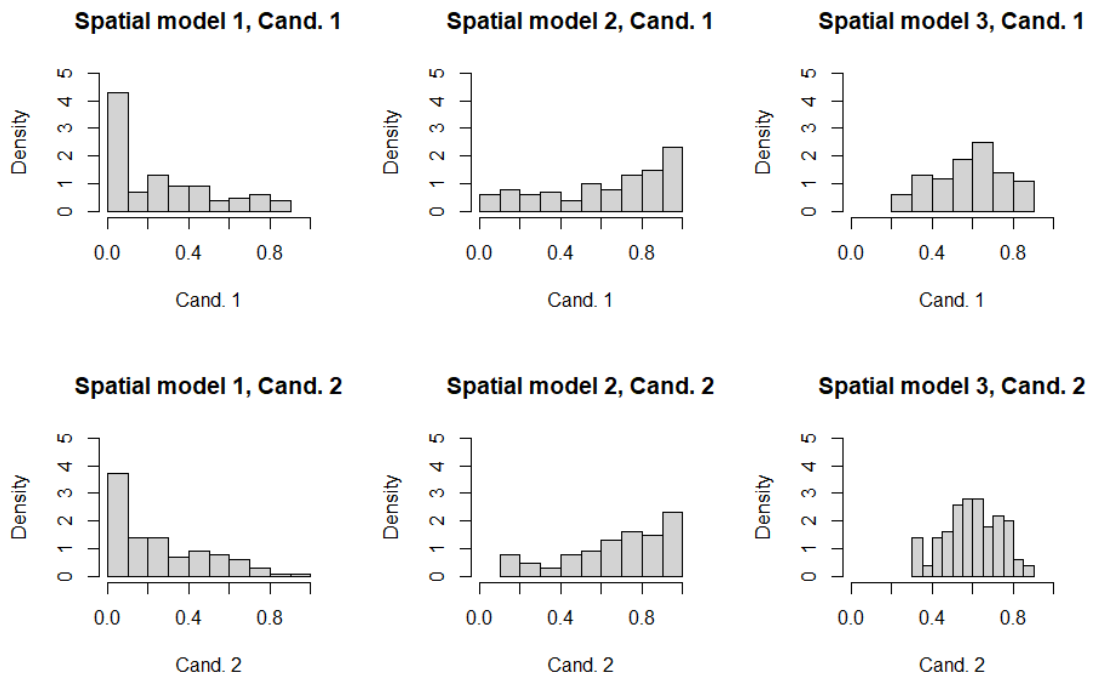

Figure 12: Histograms of evaluations obtained through the spatial model of Figure 14 and three different calculation models.

6 Fitting real data

Simulation models proposed above are based on theoretical considerations. Real voting situations do not follow pure theoretical models. Voters evaluations in real life depend of many latent factors that cannot be easily modeled. Moreover, the voters do not always have homogeneous behaviors. There may exist several groups of voters, with different distributions of evaluations of the candidates. Therefore, it is an illusion to think that a single model can capture a real vote situation.

A guideline for the choice of a specific model could be the following:

1. Fit the marginal distributions of the evaluations of each candidate, $E_1, \ldots, E_d$.
2. Test if these marginal distributions can be considered as identical or not.
3. Test the independence of $E_1, \ldots, E_d$.

The results of the distributions equality (id. distrib.) and independence tests lead to the cases summarized in Table 2.

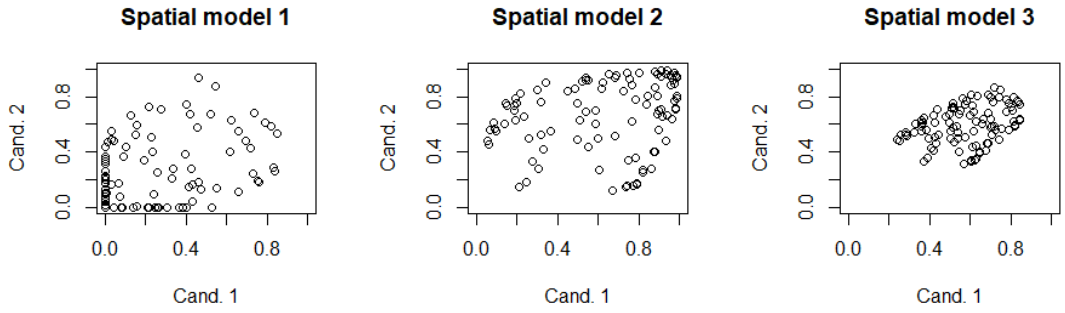

Figure 13: Plot of evaluations of candidate 2 vs candidate 1 given by each voter, with three different calculation models.

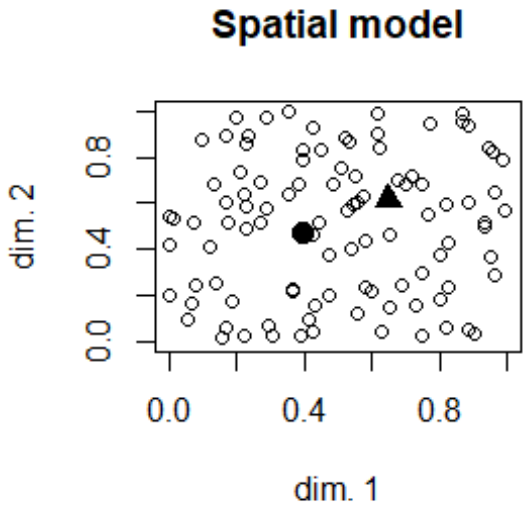

Figure 14: Simulation example of spatial model, with $d = 2$ candidates (▲ is candidate 1 and • is candidate 2) and $n = 100$ voters in a 2-dimension space ($k = 2$), obtained through an Uniform distribution on $[0, 1]^2$.

	Identical Distributions = True	Identical Distributions = False
Independence = True	Ev-IID	Ev-IDD
Independence = False	Ev-DID	Ev-DDD

Table 2: Correspondence between results of statistical tests and models on evaluations.

These guidelines are relevant only in case of small amount of data, as it is well-known that statistical tests are inclined to systematically reject the null hypothesis when the sample size is too large [22, 23].

As a matter of illustration, we propose in the following two examples of fitting real voting situations through the proposed models. The first one deals with continuous evaluations, whereas the second one focuses on discrete evaluations. Note that the high number of voters prevents from the pertinence of goodness-of-fit tests, as any χ^2 test or Kolmogorov test should conduct to reject any regular hypothesis on the distributions.

6.1 Continuous case

The first example, in a continuous framework, is based on the use of a survey concerning the 2017 presidential election in France. Data are available in [8], and deal with $d = 5$ candidates with an evaluation in $\{0, \dots, 100\}$ by $n = 20210$ voters. We transform this 0-100 scale into a continuous scale, replacing any value $n = 0, \dots, 99$ given by a voter to a candidate by a random value uniformly distributed between n and $n + 1$. Values 99 and 100 are replaced by a random value uniformly distributed between 99 and 100. These values are then scaled to the $[0, 1]$ interval.

We focus on three candidates: François Fillon (FF), Benoit Hamon (BH) and Emmanuel Macron (EM) (who finally won the election). Illustrations of the observed distributions are presented in Figure 15.

Marginal distributions. We first propose to model the marginal distributions, using either uniform distributions, or Beta distributions. The Truncated Normal distribution does not seem adequate with these data, since, as illustrated in Figure 15, the shapes of the distributions are far from a Gaussian curve. Note that the estimation of mean and standard deviation of a Truncated Normal distribution leads to incoherent values, and therefore the marginal distributions can't be fitted by any Truncated Normal distribution.

The Uniform distribution does not need any parameter estimation. The Beta distribution (Definition 3.1) depends on two parameters α and β. These parameters can be estimated using the method of moments. Let m denote the sample mean and s the sample standard deviation. Then parameters α and β can be estimated respectively by $\hat{\alpha} = m\left(\frac{m(1-m)}{s^2} - 1\right)$ and $\hat{\beta} = (1-m)\left(\frac{m(1-m)}{s^2} - 1\right)$.

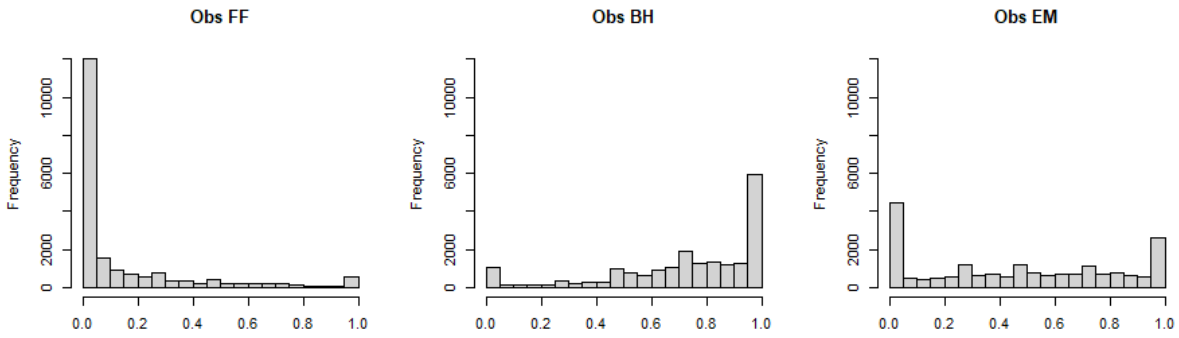

Figure 15: Histograms of observed evaluations of $d = 3$ candidates given by $n = 20\,210$ voters at the French 2017 election.

Table 3 gives the Kolmogorov-Smirnov statistic, corresponding to the distance between the distribution of observed evaluation and the distribution of simulated evaluation. The smaller the distance, the more the model fits the sample distribution. One can see that the Beta model is the most adequate in any case, and that the distribution of the evaluations of the candidate EM better fits any models than other candidates. Note that the critical value for $\alpha = 0.05$ is equal to 0.009 and the critical value for $\alpha = 0.01$ is equal to 0.011.

	FF	BH	EM
Uniform marginals, Ev-IID or Ev-DID models	0.572	0.359	0.180
Beta marginals, Ev-IDD or Ev-DDD models	0.265	0.113	0.113
Marginals from nonparametric Copula Ev-DDD model (40)	0.270	0.151	0.113
Marginals from nonparametric Copula Ev-DDD model (101)	0.006	0.007	0.008

Table 3: Kolmogorv-Smirnov statistic. The number of intervals is provided in parenthesis for nonparametric Copula Ev-DDD models.

Dependence Let us now focus on the independence between the evaluations. Evaluations given to FF and BH are negatively correlated, whereas evaluations given to FF and EM are slightly positively correlated, and evaluations given to BH and EM are not correlated, as one can see in Table 4. Hence, the independence assumption is not realistic and we consider Copula-based Ev-DDD models in order to capture the coupling between the 3 evaluations. We consider two Copula Ev-DDD models:

Parametric Copula Ev-DDD model. We consider first a Normal copula, with marginal distributions fitted by Beta distributions as established above, and correlation coefficients equal to the observed correlation coefficients shown in Table 4. It has the advantage to be parametric and to provide a reproducible modeling.

Non parametric Copula Ev-DDD model. We also consider a checkerboard copula model, which does not require any assumption on the marginal distributions. The marginal distributions are simply the sample distributions, divided into k classes. We tried $k = 101$ to recover the initial discrete values on the 0-100 scale, and, as suggested in [10], we also tried $k = 40$. In practice it is often efficient to fit a dataset but not appropriate to generate a predictive modeling.

Since the parametric Copula model is based on Beta distributions, the Kolmogorov-Smirnov statistics for the fitting of the marginals has already been calculated previously. Concerning the non parametric approach, the marginal distributions are the empirical distributions with a division in 40 or 101 classes (*i.e.* a Uniform distribution in each class, with the empirical probability to belong to each class). The resulting statistics of the Kolmogorov-Smirnov test are presented in Table 3.

One can see that the non parametric Copula model better fits the data than the parametric Copula model, at least when the number of class is big enough. This is not surprising due to the parametric/non parametric nature of the models.

Table 5 displays the correlation coefficients obtained with the two modelings. One can also see that the correlation coefficients obtained by the non-parametric approach are closer to the observed correlation coefficients than with the parametric approach.

	FF	BH	EM
FF	1	-0.41	0.30
BH	-0.41	1	0.0008
EM	0.30	0.0008	1

Table 4: Correlations between candidates at the French 2017 presidential election.

As a conclusion, Copula DDD-models are more appropriate. A non-parametric approach better fits the data, but a parametric approach seems more appropriate to generate a simulation model.

6.2 Discrete case

We study data from the Comparative Studies of Electoral Systems project (`https://cses.org/`), and especially from the "module 5" which consists in surveys about 38 elections between 2016 and 2020 worldwide. Future voters are invited to evaluate several candidates competing at each election. The

	FF	BH	EM
FF	1	-0.33	0.23
BH	-0.33	1	0.002
EM	0.23	0.002	1

	FF	BH	EM
FF	1	-0.41	0.30
BH	-0.41	1	0.01
EM	0.30	0.01	1

Table 5: Correlations between candidates obtained with copula Ev-DDD models at the French 2017 presidential election. On the left, correlations obtained with a Normal copula and Beta marginals, on the right correlations obtained through the use of a checkerboard copula with empirical marginal distributions and 40 classes.

CSES - module 5 dataset includes evaluations of candidates by voters on a discrete 0-10 scale. Hence, with a re-numeration, $\mathcal{E} = \{1, \ldots, K\}$ with $K = 11$. As a matter of example, we selected candidates B, F and G from the 2019 general elections in Denmark. A basic treatment has been necessary to remove missing answers, leaving 1108 voters (from 1345) who have given an evaluation to the three candidates.

Marginal distributions. We first propose to model the marginal distributions on the 3 candidates using three different models: Uniform distributions, Binomial distributions and Beta-Binomial distributions.

The Binomial distribution (Definition 3.2) needs the estimation of a parameter p, which can be estimated by $\hat{p} = m/K$, where m is the sample mean and K is the number of scales in the evaluations. The Beta-Binomial distribution (Definition 3.2) needs the estimation of parameters α and β, obtained as follows: $\hat{\alpha} = \frac{Km - m^2 - s^2}{K(s^2/m - 1) + m}$ and $\hat{\beta} = \frac{(K-m)(K - m - s^2/m)}{K(s^2/m-1)+m}$, with m the sample mean and s the sample standard deviation [35].

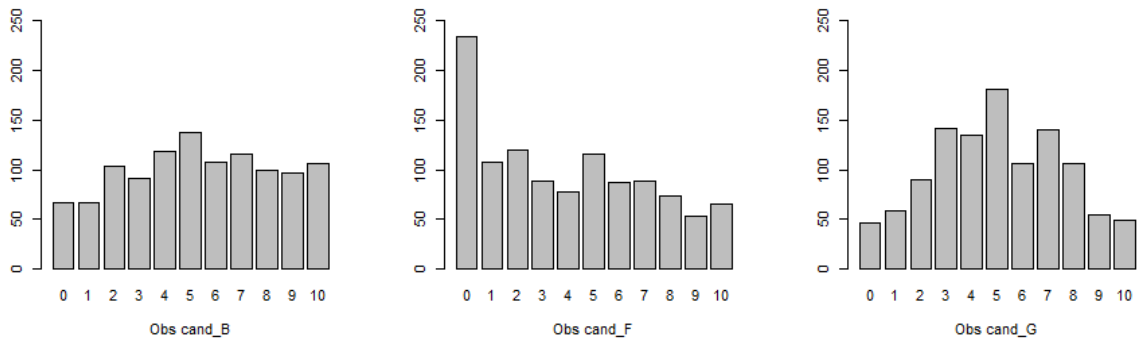

Figure 16: Bar plots of observed values for $d = 3$ candidates at the Danish 2019 election.

Table 6 gives the χ^2 statistics of the distances between the distribution of observed evaluation and the distribution of simulated evaluation. The smallest the distance, the more the observed distribution is fitted by the model. One can see that the Beta-Binomial distribution better fits the data for candidates F and G, whereas the Uniform distribution is the best for candidate B. Figure 16 shows bar plots for observed distributions.

	B	F	G
Uniform	43.8	236.3	204.8
Binomial	14476.3	53236.7	4607.2
Beta-Binomial	120.3	78.5	126.8

Table 6: χ^2 distance between the observed distributions of the evaluations and the simulated distributions of $d = 3$ candidates and 3 simulation models at the Danish 2019 election.

Dependence. As in the continuous case, the evaluations are not independent, as one can see in Table 7. We use a discrete copula to capture the dependence between teh evaluations of candidates. First the marginal distributions are fitted as described above by a Beta-Binomial distribution. Then, a discrete copula is used to model the correlation between candidates. The obtained correlations are presented in Table 7, and bar plots are presented in Figure 17.

	B	F	G
B	1	-0.58	0.63
F	-0.58	1	-0.40
G	0.63	-0.40	1

	B	F	G
B	1	-0.57	0.65
F	-0.57	1	-0.41
G	0.65	-0.41	1

Table 7: Correlations between the evaluations of each candidates at the Danish 2019 election. On the left the observed correlations, on the right the correlations obtained with a Copula Ev-DDD Beta-Binomial model.

The discrete copula captures the dependence structure, but may introduce overfitting of the dependence, when generating simulated observations. No parametric modeling, among the ones proposed, are well adapted in such context.

6.3 Spatial representation

The spatial model introduced in Section 5.4 can also be used as a representation of candidates and voters in the same space. However, the experiments show that it is difficult to simulate new data from a spatial

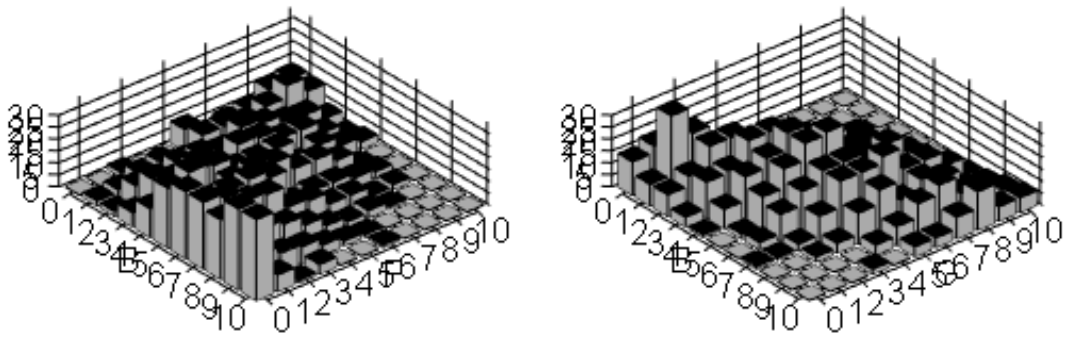

Figure 17: Barplots of observed evaluations between candidates B and F (left) and candidates B and G (right) at the Danish 2019 election.

representation: even if the spatial representation of the voting situation is accurate, it is not an easy task to identify the latent space (its dimension k and its metric) nor the distribution of the voters in the voting space. Suppose the dimension k known, that the metric is given, and let us focus on the distributions of voters in the latent space. We propose a two steps process, to simulate new data:

1. estimate both candidates and voters positions into a k-dimensions space, for example by the use of the SMACOF method [11].

2. estimate the distribution of voters into the k-dimensions space, in order to generate new voters with the same distribution.

The second step needs to fit a multidimensional distribution. The fitting is not done directly on the data but on the latent positions obtained with the first step. This can be done for example with a Copula model, as previously.

We choose hereafter to consider $k = 2$, and the euclidean distance. The spatial representations obtained by the SMACOF algorithm respectively on the continuous and discrete data are shown in Figure 18.

Hence, from real data, we have calibrated the parameters of a spatial model, which allows generating new data. Note that the intrinsic dimension and the used distance are to be set by the user. Some post-hoc measures of the quality of the adjustment are available for comparing different choices [11, chapter

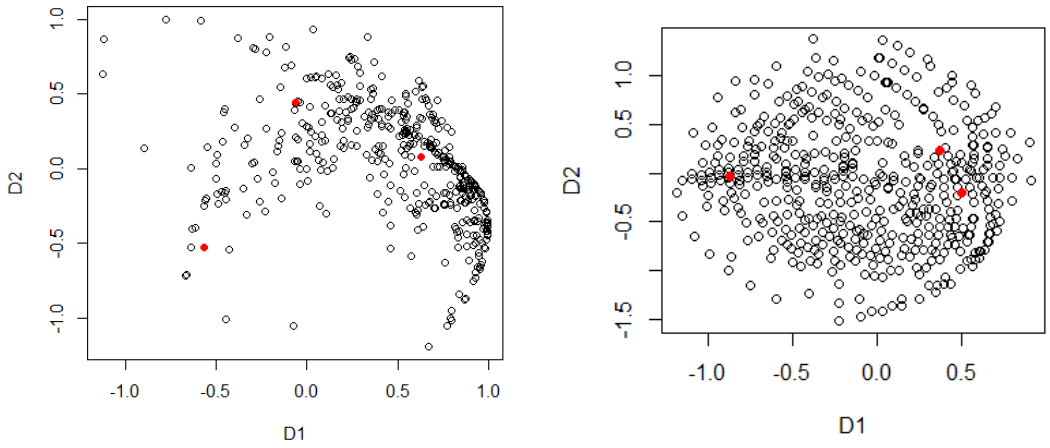

Figure 18: Spatial representation of candidates (filled bullets) and voters of the French 2017 election (left) and the Danish 2019 election (right) with by the SMACOF algorithm.

4].

7 Conclusive recommendation

7.1 Synthesis

Let us recall the four setting of simulations which are proposed here:

- Ev-IID models, Definition 1, where the evaluations are i.i.d.,
- Ev-IDD models, with independent non identically distributed distributions, Section 4,
- Ev-DID and Ev-DDD distributions, with correlations, Section 5, which includes:
  - the multinomial and the (cumative or not) Dirichlet models,
  - the truncated Normal model,
  - copula-based models, where the choice of the copula is necessary,
- the Spatial models, where the parameters to fix are the latent dimension, the distribution in this latent dimension, the distance and the mapping from the distance to the set of evaluations.

Next, for Ev-IID models, Ev-IDD models and Copula Ev-DDD models, the marginal distributions of the evaluations have to be chosen. The different families proposed here are summarized in Table 8.

Marginal continuous distributions	Parameters
Continuous Uniform $\mathcal{U}[0,1]$	$\cdot$
Truncated Normal $\mathcal{N}_T(\mu,\sigma^2)$	(μ,σ)
Beta $\mathcal{B}(\alpha,\beta)$	(α,β)

Marginal discrete distributions	Parameters
Discrete Uniform $\mathcal{U}\{0,\dots,K\}$	$\cdot$
Binomial $\mathcal{B}(K,p)$	p
Beta-Binomial $\mathcal{B}(K,\alpha,\beta)$	(α,β)

Table 8: Marginal distributions proposed for the distribution of the evaluations in Ev-IID models, Ev-IDD models and Copula Ev-DDD models.

7.2 Conclusion

As explained in the preamble, simulations can be done in two different settings.

- On the one hand, simulations can be done without any specific context, and the tuning of the distribution of the evaluations is let free or determined by external considerations. One has therefore to choose a model and set the parameters to arbitrary values. Examples of such simulation settings have been proposed with the description of each model above (Sections 3 to 5.4).
- On the other hand, one can wish to simulate observations in harmony with real data. In that case, an adjustment of the model to the observed data is necessary. The aim is therefore 1) to choose the appropriate model 2) to infer the model parameters from the available data. This situation is detailed in Section 6.

We introduced in this paper several models to simulate evaluation-based voting data in a probabilistic-based analysis perspective of evaluation-based voting rules. Three main families of distributions were proposed for the marginal distributions of the evaluations, in a continuous setting and in a discrete setting. On the contrary to preference orders models, where the key notion is the impartiality, a more refine discussion is needed for evaluation-based processes. Independent and identically distributed modeling (Ev-IID models, Section 3) yields Impartial Culture on preferences, but there are two possibilities for relaxing this assumption. We propose first to distinguish either the marginal distributions are identical or not (Ev-IDD models, Section 4). Such models do not imply Impartial Culture on preferences. Next, introducing dependence (Ev-DDD models, Section 5) creates more complex models. We give examples of dependent distributions with identical marginals (Ev-DID models) which provide Impartial Culture on preferences. In particular, we introduce Copula Ev-DID and Ev-DDD models which allow to model the dependence between the evaluations. The variety of modeling described here offers the possibility of

studying the properties of evaluation-based voting processes with an extensive probabilistic approach. It also provides new IC and non IC simulation approaches for preferences, since preferences orders can be deduced from evaluations. Finally, as some proposed settings are parametric, they can be fitted to real dataset to deduce more realistic frameworks. We present examples of such an approach on real data for continuous and discrete evaluations (Section 6).

References

[1] David A. Armstrong, Rayan Bakker, Roice Carroll, Christopher Hare, Keith T. Poole, and Howard Rosenthal. *Analyzing Spatial Models of Choice and Judgment (2nd ed.)*. CRC Press, New York, 2020.

[2] Kenneth J. Arrow. *Social choice and individual values*. Cowles Foundations and Wiley, New York, 1951.

[3] Jean-Baptiste Aubin, Irène Gannaz, Samuela Leoni, and Antoine Rolland. Deepest voting: A new way of electing. *Mathematical Social Sciences*, 116:1–16, 2022. ISSN 0165-4896. doi: https://doi.org/10.1016/j.mathsocsci.2021.12.006. URL https://www.sciencedirect.com/science/article/pii/S0165489621001232.

[4] Michel Balinski and Rida Laraki. A theory of measuring, electing and ranking. *Proceedings of the National Academy of Sciences USA*, 104(21):8720–8725, 2007.

[5] Michel Balinski and Rida Laraki. *Majority Judgment; Measuring, Ranking, and Electing*. MIT Press, Cambridge, MA, 2011.

[6] Michel Balinski and Rida Laraki. Majority judgment vs. majority rule. *Social Choice and Welfare*, 54(2):429–461, 2020.

[7] Alessandro Barbiero and Pier Alda Ferrari. An R package for the simulation of correlated discrete variables. *Communications in Statistics - Simulation and Computation*, 46(7):5123–5140, 2017.

[8] Sylvain Bouveret, Renaud Blanch, Antoinette Baujard, François Durand, Herrade Igersheim, Jérôme Lang, Annick Laruelle, Jean-François Laslier, Isabelle Lebon, and Vincent Merlin. Voter autrement 2017 - online experiment, July 2018. URL https://doi.org/10.5281/zenodo.1199545.

[9] Steven Brams and Peter C. Fishburn. *Approval voting*. Springer, New York, 2007.

[10] Andrés Cuberos, Esterina. Masiello, and Véronique Maume-Deschamps. Copulas checker-type approximations: Application to quantiles estimation of sums of dependent random variables. *Communications in Statistics - Theory and Methods*, 49(12):3044–3062, 2020.

[11] Jan de Leeuw and Patrick Mair. Multidimensional scaling using majorization: Smacof in r. *Journal of Statistical Software*, 31(3):1–30, 2009.

[12] Mostapha Diss and Eric Kamwa. Simulations in models of preference aggregation. *Œconomia*, 10 (2):279–308, 2020.

[13] Antony Downs. An economic theory of political action in a democracy. *Journal of Political Economy*, 65(2):135–150, 1957.

[14] Valdo Durrleman, Ashkan Nikeghbali, and Thierry Roncalli. Which copula is the right one? *Available at SSRN 1032545*, 2000.

[15] Adrien Fabre. Tie-breaking the highest median: alternatives to the majority judgment. *Social Choice and Welfare*, 56(1):101–124, 2021.

[16] Dan S. Felsenthal and Moshé Machover. *Electoral Systems; Paradoxes, Assumptions, and Procedures.* Springer Berlin, Heidelberg, 2012.

[17] William V. Gehrlein and Peter C. Fishburn. Condorcet's paradox and anonymous preference profiles. *Public Choice*, 26(1):1–18, 1976.

[18] Christian Genest, Anne-Catherine Favre, et al. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, 12(4):347–368, 2007.

[19] James Green-Armytage, T. Nicolaus Tideman, and Rafael Cosman. Statistical evaluation of voting rules. *Social Choice and Welfare*, 46(1):183–212, 2016.

[20] Georges-Théodule Guilbaud. Les théories de l'intérêt général et le problème logique de l'agrégation. *Economie Appliquée*, 5(4):501–584, 1952.

[21] Kiyoshi Kuga and Nagatani Hiroaki. Voter antagonism and the paradox of voting. *Econometrica*, 42(6):1045–1067, 1974.

[22] Björn Lantz. The large sample size fallacy. *Scandinavian Journal of Caring Sciences*, 27(2):487–492, 2013.

[23] Mingfeng Lin, Henry C. Lucas, and Galit Shmueli. Research commentary: Too big to fail: Large samples and the p-value problem. *Information Systems Research*, 24(4):906–917, 2013.

[24] Anghel Negriu and Cyrille Piatecki. On the performance of voting systems in spatial voting simulations. *Journal of Economic Interaction and Coordination*, 7(1):63–77, 2012.

[25] Roger B. Nelsen. *An Introduction to Copulas*. Springer, New York, second edition, 2006.

[26] Kai Ng, Guo-Liang Tian, and Man-Lai Tang. *Dirichlet and Related Distributions: Theory, Methods and Applications*. Wiley, Chichester, 2011.

[27] Abdelhalim El Ouafdi, Issofa Moyouwou, and Hatem Smaoui. IAC Probability Calculations in Voting Theory: Progress Report. In Mostapha Diss and Vincent Merlin, editors, *Evaluating Voting Systems with Probability Models*, Studies in Choice and Welfare, pages 399–416. Springer, 2021.

[28] Anastasios Panagiotelis, Claudia Czado, and Harry Joe. Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, 107(499):1063–72, 2012.

[29] Florenz Plassmann and T. Nicolaus Tideman. How frequently do different voting rules encounter voting paradoxes in three-candidate elections? *Social Choice and Welfare*, 42(1):31–75, January 2014.

[30] Christian Robert. Simulation of truncated normal variables. *Stat Comput*, 5:121–125, 1995.

[31] Warren D. Smith. Range voting. http://rangevoting.org/RangeVoting.html, 2000. Accessed: 2014-10-12.

[32] T. Nicolaus Tideman and Florenz Plassmann. The structure of the election-generating universe. unpublished, 2010.

[33] T. Nicolaus Tideman and Florenz Plassmann. Developing the empirical side of computational social choice. In *International Symposium on Artificial Intelligence and Mathematics, ISAIM 2012, Fort Lauderdale, Florida, USA, January 9-11, 2012*, 2012.

[34] T. Nicolaus Tideman and Florenz Plassmann. Which voting rule is most likely to choose the "best" candidate? *Public Choice*, 158(3/4):331–357, 2014.

[35] Ram C. Tripathi, Ramesh C. Gupta, and John Gurland. Estimation of parameters in the beta binomial model. *Annals of the Institute of Statistical Mathematics*, 46(2):317–331, 1994.

[36] Yiun. Zuo and Robert Serfling. General notions of statistical depth function. *Annals of Stat.*, 28: 461–482, 2000.