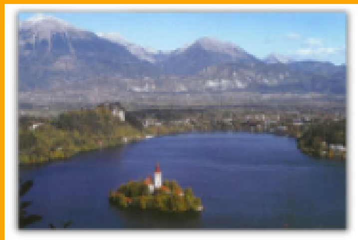
17th Applied Statistics 2021
International Conference



On-line format, Slovenia
September 20 - 22, 2021

Combinatorial regression model in abstract simplicial complexes

Invited session – Compositional data analysis, Organizer: Michael Greenacre, Universitat Pompeu Fabra, Barcelona, Spain

Andrej Srakar, Asst. Prof., Institute for Economic Research (IER), Ljubljana and School of Economics and Business, University of Ljubljana, Slovenia

Joint work with Miroslav Verbič (School of Economics and Business, University of Ljubljana and Institute for Economic Research (IER))

Outline of the presentation

- Definition of the problem and basic overview of the idea
- Symplectic (CoDA) regression models
- Abstract simplicial complexes and algebraic topology
- Multivariate Distance Matrix Regression approach
- Estimation for Jensen-Shannon type divergences
- Properties of the estimator
- Applications
- Conclusion and extensions

Regressions for diversity and economic inequality

Tabela 5 a: Deleži bruto dohodka, akontacije dohodnine, socialnih prispevkov in neto dohodka, podatkovni vir A

Leto	delež bruto dohodka	delež akontacije dohodnine	delež socialnih prispevkov	delež »neto« dohodka
1993	1,000	0,140	0,218	0,642
1994	1,000	0,142	0,205	0,654
1995	1,000	0,143	0,200	0,658
1996	1,000	0,146	0,198	0,656
1997	1,000	0,145	0,198	0,657
1998	1,000	0,147	0,202	0,652
1999	1,000	0,148	0,202	0,649
2000	1,000	0,150	0,204	0,647
2001	1,000	0,150	0,204	0,646
2002	1,000	0,151	0,204	0,645
2003	1,000	0,152	0,204	0,644
2004	1,000	0,152	0,203	0,645
2005	1,000	0,142	0,201	0,657

Tabela 5 b: Ginijev koeficient ter koeficienti koncentracije za akontacijo dohodnine, socialne prispevke in neto dohodek, podatkovni vir A

Leto	Ginijev koeficient za bruto dohodek	koeficient koncentracije za akontacijo dohodnine	koeficient koncentracije za socialne prispevke	koeficient koncentracije za »neto« dohodek
1993	0,282	0,389	0,279	0,259
1994	0,285	0,464	0,282	0,248
1995	0,295	0,472	0,293	0,257
1996	0,299	0,476	0,295	0,261
1997	0,302	0,480	0,297	0,265
1998	0,305	0,485	0,302	0,266
1999	0,313	0,492	0,309	0,273
2000	0,312	0,490	0,310	0,272
2001	0,314	0,491	0,312	0,273
2002	0,310	0,486	0,308	0,269
2003	0,311	0,486	0,309	0,270
2004	0,308	0,480	0,303	0,269
2005	0,308	0,514	0,304	0,264

Howie and Kleczyk's full-factorial attraction model

- We will develop a large extension of a decade and half ago developed transformation of the MCI model, called Full-Factorial Attraction Model, as developed in Howie and Kleczyk (2007).
- The approach is based on a reconceptualization of any market share variable for each brand as a series of two-product markets (in this way, the number of units grows to $^{I!}/_{2!}$ (see Howie and Kleczyk, 2007; 2008a; 2008b – I is the number of units/brands) which gains quite a lot of degrees of freedom for the analysis).
- The final equation for Full-Factorial Attraction Model is provided below:

$$m_{ijt} = \alpha_i + \beta X_{ijt} + \varepsilon_{it}$$

- Where
- $m_{ijt} = \frac{M_{it}}{(M_{it}+M_{jt})}$ where $i = 1, \dots, I-1; j = 1, \dots, I-1$ and $i \neq j, t = 1, \dots, T;$
- $X_{ijt} = x_{it} - x_{jt}$ where $i = 1, \dots, I; j = 1, \dots, I$ and $i \neq j, t = 1, \dots, T;$
- t is a time variable and T is the maximal time;
- α_i is a parameter for the constant influence of brand i;
- ε_i is a random error term.

Combinatorial regression

	region1	region2	year	gdp	emptot	emparts	innov	birth	sh	preb	panelid	nrpanel
1	Gorenjska	Goriska	2005	-1000	26.81	-.48	-1	915	.370313	.039408	Gorenjska-Goriska	Gorenjska-Goriska
2	Gorenjska	Goriska	2006	-1200	26.8	-.74	.	1113	.442553	.039662	Gorenjska-Goriska	Gorenjska-Goriska
3	Gorenjska	Goriska	2007	-1600	27.11	-1.26	-13	982	.416296	.040005	Gorenjska-Goriska	Gorenjska-Goriska
4	Gorenjska	Goriska	2008	-2000	26.55	-1.48	7	1033	.401254	.039946	Gorenjska-Goriska	Gorenjska-Goriska
5	Gorenjska	Goriska	2009	-2400	25.75	-1.24	4	1025	.396307	.04096	Gorenjska-Goriska	Gorenjska-Goriska
6	Gorenjska	Goriska	2010	-2000	26.44	-1.24	.	1119	.398765	.040949	Gorenjska-Goriska	Gorenjska-Goriska
7	Gorenjska	Goriska	2011	-1700	27.1	-1.01	11	987	.554591	.041109	Gorenjska-Goriska	Gorenjska-Goriska
8	Gorenjska	Goriska	2012	-1300	26.96	-1.04	31	980	.291569	.041266	Gorenjska-Goriska	Gorenjska-Goriska
9	Gorenjska	Goriska	2013	-900	27.65	-1.1	12	1005	.300999	.041277	Gorenjska-Goriska	Gorenjska-Goriska
10	Gorenjska	Goriska	2014	-500	28.21	-.87	11	961	.246305	.041493	Gorenjska-Goriska	Gorenjska-Goriska
11	Gorenjska	Goriska	2015	-700	28.84	-.86	15	1004	.275862	.041525	Gorenjska-Goriska	Gorenjska-Goriska
12	Gorenjska	Jugovzhodna Slovenija	2005	-1300	17.84	1.34	.	634	.560284	.02966	Gorenjska-Jugovzhodna Slovenija	Gorenjska-Jugovzhodna Slovenija
13	Gorenjska	Jugovzhodna Slovenija	2006	-1900	17.31	1.3	-15	719	.627767	.02962	Gorenjska-Jugovzhodna Slovenija	Gorenjska-Jugovzhodna Slovenija
14	Gorenjska	Jugovzhodna Slovenija	2007	-2000	17.5	1.09	1	657	.589099	.029668	Gorenjska-Jugovzhodna Slovenija	Gorenjska-Jugovzhodna Slovenija
15	Gorenjska	Jugovzhodna Slovenija	2008	-2200	17.8	1.1	.	697	.558952	.029472	Gorenjska-Jugovzhodna Slovenija	Gorenjska-Jugovzhodna Slovenija
16	Gorenjska	Jugovzhodna Slovenija	2009	-2500	17.26	1.23	7	802	.629797	.029824	Gorenjska-Jugovzhodna Slovenija	Gorenjska-Jugovzhodna Slovenija
17	Gorenjska	Jugovzhodna Slovenija	2010	-2200	17.58	1.17	5	767	.640873	.029707	Gorenjska-Jugovzhodna Slovenija	Gorenjska-Jugovzhodna Slovenija
18	Gorenjska	Jugovzhodna Slovenija	2011	-2100	18.41	1.15	2	671	.682443	.029727	Gorenjska-Jugovzhodna Slovenija	Gorenjska-Jugovzhodna Slovenija
19	Gorenjska	Jugovzhodna Slovenija	2012	-1800	18.45	1.15	30	644	.650131	.02986	Gorenjska-Jugovzhodna Slovenija	Gorenjska-Jugovzhodna Slovenija
20	Gorenjska	Jugovzhodna Slovenija	2013	-1700	18.59	1.29	8	688	.602857	.029859	Gorenjska-Jugovzhodna Slovenija	Gorenjska-Jugovzhodna Slovenija
21	Gorenjska	Jugovzhodna Slovenija	2014	-1500	18.87	1.46	8	517	.584795	.029834	Gorenjska-Jugovzhodna Slovenija	Gorenjska-Jugovzhodna Slovenija
22	Gorenjska	Jugovzhodna Slovenija	2015	-1500	18.32	1.57	17	608	.55102	.029801	Gorenjska-Jugovzhodna Slovenija	Gorenjska-Jugovzhodna Slovenija
23	Gorenjska	Koroska	2005	1100	50.15	1.67	.	1349	.699115	.062327	Gorenjska-Koroska	Gorenjska-Koroska
24	Gorenjska	Koroska	2006	1200	49.95	1.6	.	1458	.781955	.062561	Gorenjska-Koroska	Gorenjska-Koroska
25	Gorenjska	Koroska	2007	1500	51.08	1.37	2	1474	.751337	.062816	Gorenjska-Koroska	Gorenjska-Koroska
26	Gorenjska	Koroska	2008	1500	52.09	1.36	12	1494	.737752	.062956	Gorenjska-Koroska	Gorenjska-Koroska
27	Gorenjska	Koroska	2009	1200	50.89	1.41	8	1534	.708122	.06362	Gorenjska-Koroska	Gorenjska-Koroska
28	Gorenjska	Koroska	2010	1500	51.36	1.52	.	1721	.708333	.063552	Gorenjska-Koroska	Gorenjska-Koroska
29	Gorenjska	Koroska	2011	1100	50.96	1.52	.	1599	.774697	.063864	Gorenjska-Koroska	Gorenjska-Koroska
30	Gorenjska	Koroska	2012	800	49.5	1.56	18	1550	.752266	.064069	Gorenjska-Koroska	Gorenjska-Koroska
31	Gorenjska	Koroska	2013	1000	49.82	1.79	17	1578	.661442	.064058	Gorenjska-Koroska	Gorenjska-Koroska
32	Gorenjska	Koroska	2014	1400	50.65	1.95	-5	1431	.701754	.064213	Gorenjska-Koroska	Gorenjska-Koroska
33	Gorenjska	Koroska	2015	1300	51.47	2.06	18	1497	.72	.064253	Gorenjska-Koroska	Gorenjska-Koroska
34	Gorenjska	Obalno-kraska	2005	-2600	32.79	-.11	1	1111	.722561	.046713	Gorenjska-Obalno-kraska	Gorenjska-Obalno-kraska
35	Gorenjska	Obalno-kraska	2006	-3100	31.4	-.16	0	1272	.707483	.046649	Gorenjska-Obalno-kraska	Gorenjska-Obalno-kraska

Combinatorial regression

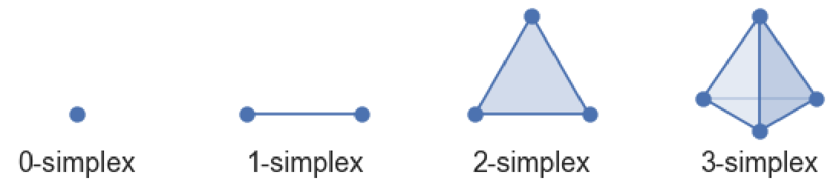
	nr1	nr2	nr3	gender1	gender2	gender3	age1	age2	age3	eduy1	eduy2	eduy3
1	Decile1	Decile2	Decile3	1.57913	1.69152	1.63851	65.872	70.278	69.554	9.37037	8.34576	9.23311
2	Decile1	Decile2	Decile4	1.57913	1.69152	1.63176	65.872	70.278	67.9696	9.37037	8.34576	10.0304
3	Decile1	Decile2	Decile5	1.57913	1.69152	1.55593	65.872	70.278	66.9424	9.37037	8.34576	10.0441
4	Decile1	Decile2	Decile6	1.57913	1.69152	1.54882	65.872	70.278	66.5993	9.37037	8.34576	10.5455
5	Decile1	Decile2	Decile7	1.57913	1.69152	1.51186	65.872	70.278	66.7356	9.37037	8.34576	11.0814
6	Decile1	Decile2	Decile8	1.57913	1.69152	1.5	65.872	70.278	65.1858	9.37037	8.34576	11.2703
7	Decile1	Decile2	Decile9	1.57913	1.69152	1.49662	65.872	70.278	65.9696	9.37037	8.34576	12.3885
8	Decile1	Decile2	Decile10	1.57913	1.69152	1.55254	65.872	70.278	62.8102	9.37037	8.34576	12.339
9	Decile1	Decile3	Decile4	1.57913	1.63851	1.63176	65.872	69.554	67.9696	9.37037	9.23311	10.0304
10	Decile1	Decile3	Decile5	1.57913	1.63851	1.55593	65.872	69.554	66.9424	9.37037	9.23311	10.0441
11	Decile1	Decile3	Decile6	1.57913	1.63851	1.54882	65.872	69.554	66.5993	9.37037	9.23311	10.5455
12	Decile1	Decile3	Decile7	1.57913	1.63851	1.51186	65.872	69.554	66.7356	9.37037	9.23311	11.0814
13	Decile1	Decile3	Decile8	1.57913	1.63851	1.5	65.872	69.554	65.1858	9.37037	9.23311	11.2703
14	Decile1	Decile3	Decile9	1.57913	1.63851	1.49662	65.872	69.554	65.9696	9.37037	9.23311	12.3885
15	Decile1	Decile3	Decile10	1.57913	1.63851	1.55254	65.872	69.554	62.8102	9.37037	9.23311	12.339
16	Decile1	Decile4	Decile5	1.57913	1.63176	1.55593	65.872	67.9696	66.9424	9.37037	10.0304	10.0441
17	Decile1	Decile4	Decile6	1.57913	1.63176	1.54882	65.872	67.9696	66.5993	9.37037	10.0304	10.5455
18	Decile1	Decile4	Decile7	1.57913	1.63176	1.51186	65.872	67.9696	66.7356	9.37037	10.0304	11.0814
19	Decile1	Decile4	Decile8	1.57913	1.63176	1.5	65.872	67.9696	65.1858	9.37037	10.0304	11.2703
20	Decile1	Decile4	Decile9	1.57913	1.63176	1.49662	65.872	67.9696	65.9696	9.37037	10.0304	12.3885
21	Decile1	Decile4	Decile10	1.57913	1.63176	1.55254	65.872	67.9696	62.8102	9.37037	10.0304	12.339
22	Decile1	Decile5	Decile6	1.57913	1.55593	1.54882	65.872	66.9424	66.5993	9.37037	10.0441	10.5455
23	Decile1	Decile5	Decile7	1.57913	1.55593	1.51186	65.872	66.9424	66.7356	9.37037	10.0441	11.0814
24	Decile1	Decile5	Decile8	1.57913	1.55593	1.5	65.872	66.9424	65.1858	9.37037	10.0441	11.2703
25	Decile1	Decile5	Decile9	1.57913	1.55593	1.49662	65.872	66.9424	65.9696	9.37037	10.0441	12.3885
26	Decile1	Decile5	Decile10	1.57913	1.55593	1.55254	65.872	66.9424	62.8102	9.37037	10.0441	12.339
27	Decile1	Decile6	Decile7	1.57913	1.54882	1.51186	65.872	66.5993	66.7356	9.37037	10.5455	11.0814

What is a simplex?

- In geometry, a simplex (plural: simplexes or simplices) is a generalization of the notion of a triangle or tetrahedron to arbitrary dimensions. The simplex is so-named because it represents the simplest possible polytope in any given space.

For example:

- a 0-simplex is a point,
- a 1-simplex is a line segment,
- a 2-simplex is a triangle,
- a 3-simplex is a tetrahedron,
- a 4-simplex is a 5-cell.



- A *composition* is defined as a vector of D positive components $x = (x_1, x_2, \dots, x_D)$ summing up to a given constant κ
- It is generally – although not universally - agreed that the appropriate sample space for compositional data is the standard simplex (also called the "unit simplex"). It is defined as

$$S^D = \left\{ x = [x_1, x_2, \dots, x_D] \,\middle|\, x_i > 0, i = 1,2, \dots, D; \sum_{i=1}^{D} x_i = \kappa \right\}$$

Compositional regression models – on one simplex

- In regression analysis of the market share data four main (parametric) type models are prevalent: multinomial logistic regression, attraction models of various types, Dirichlet covariance models, and compositional regression (Morais et al., 2017).
- Market-share models were developed in the 80's, mainly by Cooper and Nakanishi (e.g. Cooper and Nakanishi, 1988). They are inspired from an aggregated version of the conditional multinomial logit (MNL) models. For individual data, conditional MNL models, widely used in econometrics, model discrete choices of individuals, i.e. the probability that an individual i chooses an alternative j.
- Multiplicative competitive interaction model (MCI), also called the "attraction" model has the following general two-part structure:

$$M_i = \frac{\mathcal{A}_i}{\sum_{j=1}^{m} \mathcal{A}_j}$$

$$\mathcal{A}_i = \prod_{k=1}^{K} f_k(X_{ki})^{\beta_k}$$

- The Dirichlet distribution is the distribution of a composition obtained as the closure of a vector of D independent gamma-distributed variables with the same scale parameter. This, it is another distribution adapted for variables lying in the simplex. Let $S = (S_1, \dots, S_D) \sim \mathcal{D}(\alpha_1, \dots, \alpha_D)$ where $S_j > 0$ and $\sum_{j=1}^{D} S_j = 1$, $\alpha_j > 0$ and $\sum_{j=1}^{D} \alpha_j = \alpha_0$. α_0 is called the precision parameter. Then, $\mathbb{E}(S_j) = \frac{\alpha_j}{\alpha_0}$. Two parametrizations exist for the Dirichlet regression model, the common and the alternative one.

Compositional regression models – on one simplex

- CoDA models can be expressed either in terms of the initial compositional observations in the simplex or alternatively in terms of the corresponding transformed coordinates in the Euclidean space, as below.

- Linear CODA model in the simplex (in terms of compositions):

$$S_t = a \oplus_{k=1}^{K} B_k \boxdot Z_{kt} \oplus \varepsilon_t$$

- with $S, a, Z_k, \varepsilon \in S^D$ and $B_k \in \mathbb{R}_{D\times D}$ such that row and column sums are equal to zero, and the following operations are used in the simplex:

- $\oplus$ is the perturbation operation, corresponding to the addition operation in the simplex: $x \oplus y = \mathcal{C}(x_1 y_1, \dots, x_D y_D)'$ with $x, y \in S^D$, and $\oplus_{k=1}^{K}$ corresponds to $\sum_{k=1}^{K}$.
- $\odot$ is the power transformation, corresponding to the multiplication operation in the simplex: $x \odot \lambda = \mathcal{C}(x_1^{\lambda}, \dots, x_D^{\lambda})'$ with $\lambda \in \mathbb{R}, x \in S^D$
- $\boxdot$ is the compositional matrix product, corresponding to the matrix product in the simplex: $B \boxdot x = \mathcal{C}(\prod_{j=1}^{D} x_j^{b_{1j}}, \dots, \prod_{j=1}^{D} x_j^{b_{Dj}})'$ with $B \in \mathbb{R}_{D\times D}, x \in S^D$

- For any vector of D real positive components $z = [z_1, z_2, \dots, z_D] \in \mathbb{R}_+^D$ ($z_i > 0$ for all $i = 1, 2, \dots, D$), the closure of z is defined as

$$\mathcal{C}(z) = \left[\frac{\kappa \cdot z_1}{\sum_{i=1}^{D} z_i}, \frac{\kappa \cdot z_2}{\sum_{i=1}^{D} z_i}, \dots, \frac{\kappa \cdot z_D}{\sum_{i=1}^{D} z_i}\right]$$

Compositional regression models – on one simplex

- Linear CoDA model in the Euclidean space in terms of isometric log-ratio (ilr) coordinates:

$$S_{jt}^{*} = a_j^{*} + \sum_{k=1}^{K} \sum_{m=1}^{D-1} b_{kjm}^{*} X_{kmt}^{*} + \varepsilon_{jt}^{*}, \forall j \in 1, \dots, D-1$$

- where j is the index of S's ilr coordinates, m is the index of X's ilr coordinates and $\varepsilon_j^{*} \sim \mathcal{N}(0, \sigma^2)$. Above equation corresponds to a system of $D-1$ linear models, one for each ilr coordinate of S.

- Nonparametric regression in CoDA – local polynomial (Di Marzio et al., 2015); simplicial splines (Machalová, Hron and Talská, 2019); simplicial wavelets (Srakar and Fry, 2019).

- We extend this arsenal of possibilities with a novel regression perspective (applicable to simplicial complexes, i.e. to sets of simplexes), labelled *combinatorial regression*, based on combining n-tuplets of sampling units into groups. This perspective is based on broad generalization of the Full-Factorial Attraction model from marketing (Howie & Kleczyk, 2007; 2008). We extend the Howie and Kleczyk perspective by considering instead of pair of "brands" (regions, etc.) triplets, quadruplets, indeed, any *combinatorial variation* of units as the basis for constructing new regression units.

Topological data analysis (TDA)

- TDA is a data analysis method that provides information about the „shape" of data.
- It has been developed within the last twenty years and is rooted in the mathematical field of algebraic topology („*Topology is the branch of mathematics that studies shape, and algebraic topology is the application of tools from abstract algebra to quantify shape.*")
- Homology: $H_n(X) = \frac{Ker(\partial_n)}{Im(\partial_{n+1})}$
- Two elements, $\alpha, \beta \in Ker(\partial_n)$, are homologous if $\alpha = \beta + x$, where $x \in Im(\partial_{n+1})$
- Dimension of a homology: Betti Number, $\beta_n = \dim(H_n(X))$

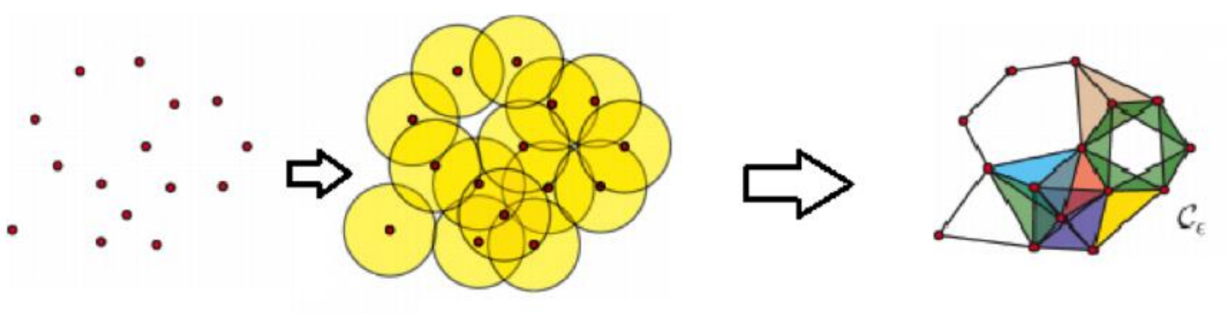
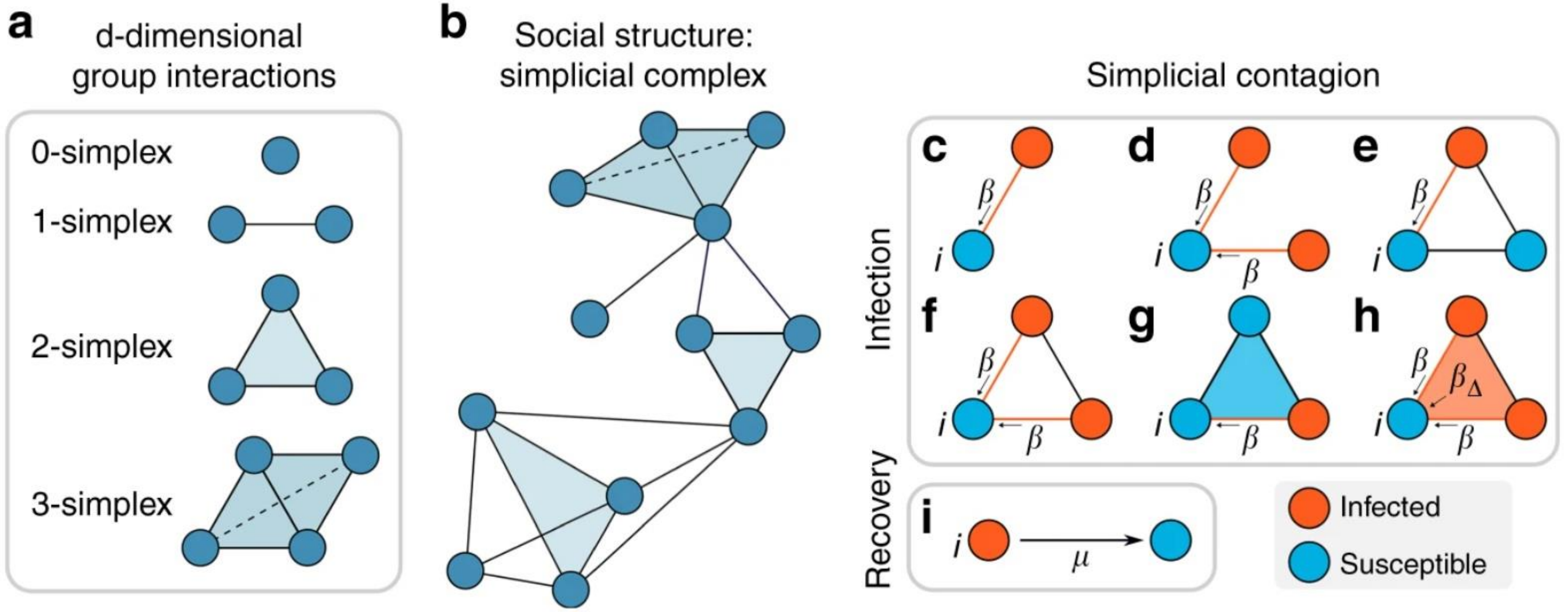


Figure 3.14: Building the Čech complex [16]

Abstract simplicial complexes

Fig. 1

From: Simplicial models of social contagion



Abstract simplicial complexes

- A simplicial complex K consists of:
- A set of objects, $V(K)$, i.e. vertices
- A set, $S(K)$, of finite non-empty subsets of $V(K)$, i.e. simplices such that simplices satisfy the following conditions:
- If $\sigma \subset V(K)$ is a simplex and $\tau \subset \sigma, \tau \neq 0$, then τ is also a simplex;
- Every singleton $\{v\}, v \in V(K)$, is a simplex.
- We say τ is a face of σ. If $\sigma \in S(K)$ has $p + 1$ elements it is said to be a p-simplex. The set of p-simplices of K is denoted by K_p. The dimension of K is the largest p such that K_p is non-empty.
- A map of simplicial complexes $K \to L$ is a function $f: V(K) \to V(L)$ such that whenever $\sigma \subseteq V(K)$ belongs to $S(K)$, the image $f(\sigma)$ belongs to $S(L)$.
- Definition: The standard (topological) p-simplex is taken to be the convex hull of the basis vectors $e_1, e_2, \dots, e_{p+1}$ in $\mathbb{R}^{p+1}$.

Abstract simplicial complexes

- $|K|$ is the set of all functions from $V(K)$ to the closed interval $[0,1]$ such that
- If $\alpha \in |K|$, the set $\{v \in V(K) | \alpha(v) \neq 0\}$ is a simplex of K;
- For each $\alpha \in |K|$,

$$\sum_{v \in V(K)} \alpha(v) = 1$$

- *Metric topology*:
- We put a metric d on $|K|$ (labelled $|K|_d$) by:

$$d(\alpha, \beta) = (\sum_{v \in V(K)} \big(\alpha(v) - \beta(v)\big)^2)^{\frac{1}{2}}$$

- *Coherent topology*:
- Each geometric simplex $|s|$ consists of all $\alpha \in |K|$ supported in s, and is given the subspace topology inherited as a subset of $|K|_d$; then the coherent topology on $|K|$ is the largest topology for which all inclusions $|s| \to |K|$ are continuous. This topological space is normally denoted just $|K|$, reflecting the fact that the coherent topology is regarded as the default topology to put on the set $|K|$.

Abstract simplicial complexes

- **Geometric simplicial complex**
- A finite collection of simplices K called the faces of K such that:

- $\forall \sigma \in K, \sigma$ is a simplex
- $\sigma \in K, \tau \subset \sigma \Rightarrow \tau \in K$
- $\forall \sigma, \tau \in K$, either $\sigma \cap \tau = \emptyset$ or $\sigma \cap \tau$ is a common face of both

- **Abstract simplicial complex**
- Given a finite set of elements P, an abstract simplicial complex K with vertex set P is a set of subsets of P such that:

- $\forall p \in P, p \in K$
- If $\forall \sigma \in K$ and $\tau \subseteq \sigma$, then $\tau \in K$

- The elements of K are called the (abstract) simplices or faces of K
- The dimension of a simplex σ is $\dim(\sigma) = \#vert(\sigma) - 1$

Multivariate Distance Matrix Regression perspective

- Anderson (2001) and McArdle & Anderson (2001) proposed a nonparametric regression approach, based on pairwise distances between vectors of scores on the outcome variables.
- Multivariate Distance Matrix Regression (MDMR) quantifies structure in the data based on similarities between subjects rather than similarities between variables. Distance between two vectors of scores on a multivariate outcome is defined as the result of a function $d(Y_i', Y_j')$ that quantifies the dissimilarity of the response profiles of subjects i and j, i.e. distance between their vectors of scores.
- MDMR differs from the standard linear model in its representation of the sum of squares of the outcome. The standard linear model can be viewed as a variable-centered approach to regression that is used to partition the sum of squared Euclidean distances between each subject's vector of scores on Y and the mean vector of Y. MDMR facilitates a person-centered approach that instead partitions the sum of squared distances between all pairs of individuals.

$$d_E(Y_i', Y_j') = \sqrt{\sum_{k=1}^{q} (Y_{ik} - Y_{jk})^2} \qquad d_M(Y_i', Y_j') = \sum_{k=1}^{q} |Y_{ik} - Y_{jk}| \qquad d_D(Y_i', Y_j') = \sum_{k=1}^{q} \mathbb{I}(Y_{ik} \neq Y_{jk})$$

$$SSE \to SSD = \sum_{i<j} D_{ij}^2 = \sum_{j<i} D_{ij}^2$$

- Gower's G: $G = \left(I - \frac{1}{n}J\right) A \left(I - \frac{1}{n}J\right)$ where J is a square n-dimensional matrix of 1's and $A = \{a_{ij}\} = \{-\frac{1}{2} d_{ij}^2\}$
- $G = U \Lambda U'$ where Λ is the diagonal $n \times n$ matrix whose columns $\lambda_k (k = 1, \dots, n)$ are the eigenvalues of G, and U is the $n \times n$ matrix whose columns u_k are the orthogonal eigenvectors of G corresponding to λ_k.

Multivariate Distance Matrix Regression perspective

- MDMR test statistic:

$$\tilde{F} = \frac{tr[Z'HZ]/p}{tr[Z'(I-H)Z]/(n-p-1)} = \frac{tr[HGH]/p}{tr[(I-H)G(I-H)]/(n-p-1)}$$

$$= \frac{tr[(H-H_0)G(H-H_0)]/r}{tr[(I-H)G(I-H)]/(n-p-1)}$$

$$\tilde{F} = \frac{\sum_{k=1}^{n} \lambda_k \, tr\, [\hat{u}_k' \hat{u}_k]}{\sum_{k=1}^{n} \lambda_k \, tr\, [r_k' r_k]} = \frac{\sum_{k=1}^{n} \lambda_k \, \hat{u}_k' \hat{u}_k}{\sum_{k=1}^{n} \lambda_k \, r_k' r_k}$$

$$\hat{u}_k' \hat{u}_k / \sigma_k^2 \sim \chi^2(p)$$

$$r_k' r_k / \sigma_k^2 \sim \chi^2(n-p-1)$$

$$p_p(\tilde{F}) = \frac{\sum_{m=1}^{n!} \mathbb{I}(\tilde{F} \le \tilde{F}_m)}{n!}$$

$$P(\tilde{F} \le \tilde{f}) = P\left(\frac{tr[HGH]}{tr[(I-H)H(I-H)]} \le \tilde{f}\right) = P\left(\frac{\sum_{k=1}^{n} \lambda_k \, \hat{u}_k' \hat{u}_k}{\sum_{k=1}^{n} \lambda_k \, r_k' r_k} \le \tilde{f}\right)$$

$$= P\left(\sum_{k=1}^{n} \lambda_k \, \hat{u}_k' \hat{u}_k - \tilde{f} \sum_{k=1}^{n} \lambda_k \, r_k' r_k \le 0\right)$$

Combinatorial regression

- The basic form of the combinatorial regression model is provided below:

$$m_{ijklt} = \alpha_{ijklt} + \beta X_{ijklt} + \varepsilon_{ijklt}$$

- where

$$m_{ijkl} = \frac{M_i}{M_i + M_j + M_k + M_l};\ m_{jikl} = \frac{M_j}{M_i + M_j + M_k + M_l};\ m_{kijl} = \frac{M_k}{M_i + M_j + M_k + M_l};\ m_{lijk} = \frac{M_l}{M_i + M_j + M_k + M_l};$$

$$m_{ijkl} + m_{jikl} + m_{kijl} + m_{lijk} = 1$$

- $X_{ijklt} = d(x_{im_{ijkl}}, x_{im_{jikl}}, x_{im_{kijl}}, x_{im_{lijk}})$, where $i = 1, \dots, I$; $j = 1, \dots, I$ and $i \neq j$, $t = 1, \dots, T$;
- t is a time variable and T is the maximal time;
- α_i is a parameter for the constant influence of brand i;
- ε_i is a random error term.
- There are many possibilities to construct a dependent variable in this case, above is a basic one where it consists of a share of the first brand/region/category in the total pair/triplet/quadruplet/etc., as above (example of quadruplets):

- It also allows applications to very small datasets as the number of units in the new model can be expressed in terms of generalized factorial products (Dedekind numbers) of units of original sample.

k		k	
		4	168
0	2	5	7581
1	3	6	7828354
2	6	7	2414682040998
3	20	8	56130437228687500000000

Jensen-Shannon and generalized Jensen-Shannon divergence measures

- Consider the set $M_+^1(A)$ of probability distributions where A is a set provided with some σ-algebra of measurable subsets. In particular, we can take A to be a finite or countable set with all subsets being measurable.

- The Jensen-Shannon divergence (JSD) $M_+^1(A) \times M_+^1(A) \to [0, \infty)$ is a symmetrized and smoothed version of the Kullback-Leibler divergence $D(P \parallel Q)$. It is defined by:

$$JSD(P \parallel Q) = \frac{1}{2} D(P \parallel M) + \frac{1}{2} D(Q \parallel M)$$

- where $M = \frac{1}{2}(P + Q)$

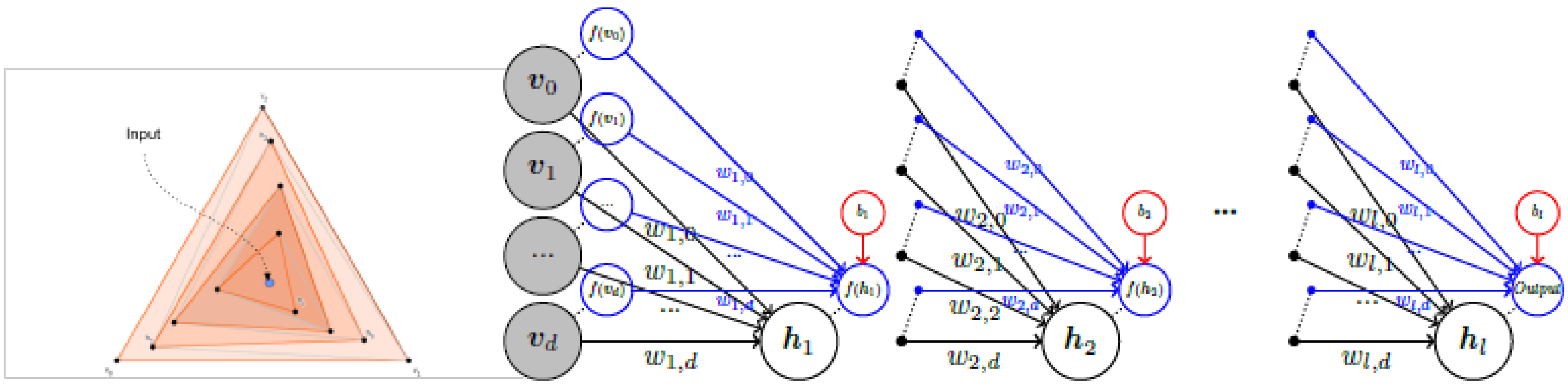
- A generalization of the Jensen-Shannon divergence using abstract means (e.g. geometric, harmonic) was proposed in Nielsen (2019). The geometric Jensen-Shannon divergence (G-Jensen-Shannon) yields a closed-form formula for divergence between two Gaussian distributions by taking the geometric mean.

- A more general definition (allowing more than two probability distributions):

$$JSD_{\pi_1,\dots,\pi_n}(P_1, P_2, \dots, P_n) = H\left(\sum_{i=1}^{n} \pi_i P_i\right) - \sum_{i=1}^{n} \pi_i H(P_i)$$

- where $\pi_1, \dots, \pi_n$ are weights that are selected for the probability distributions $P_1, P_2, \dots, P_n$ and $H(P)$ is the Shannon entropy for distribution P.

Estimation – simplicial complex networks



- Firuzzi et al. (2020) introduce the notion of automatic subdivisioning and devise a particular type of neural networks for regression tasks: Simplicial Complex Networks (SCNs). SCN's architecture is defined with a set of bias functions along with a particular policy during the forward pass which alternates the common architecture search framework in neural networks.

- Left: without applying any transformation to its input, an SCN locates the position of input in the input space through a set of nested simplexes.
- Right: architecture of an SCN: v_i are a given set of visible vertices of a primary simplex in the input space that the sample falls inside. A network of hidden vectors h_i are then used to parameterize a sequence of nested simplexes for locating the input. Each h_i is a convex combination of a subset of its preceding vectors. In parallel, another network is used to generate SCN's output utilizing the output of SCN at all v_i and h_i, combined with a group of bias values b_i.

Estimation – simplicial complex networks

- **Definition (barycentric subdivision):** Barycentric subdivision (BCS) of a d-simplex K consists of $(d+1)!$ d-simplexes. Each d-simplex $[v_o, \ldots, v_d]$ out of these $(d+1)!$ simplexes is associated with a permutation $p_o, p_1, \ldots, p_d$ of the vertices of K such that v_i denotes the barycenter (centroid) of $p_o, p_1, \ldots, p_i$ where $1 \leq i \leq n$

- **Simplicial Approximation Theorem:** Let X and Y be two simplicial complexes and $f: X \rightarrow Y$ be a continuous function. Then for arbitrary ϵ, there exist sufficiently large k and l and a simplicial mapping $g: X^{(k)} \rightarrow Y^{(l)}$ approximating f such that $sup_{x \in X} \|f(x) - g(x)\| < \epsilon$. $X^{(k)}$ and $Y^{(l)}$ represent the k-th and l-th barycentric subdivision of X and Y, respectively.

Estimation – simplicial complex networks

Algorithm 1 Generating a simplex in barycentric subdivision of a d-simplex

input: d-simplex $\sigma = [v_0, v_2, ..., v_d]$, permutation $P = (p_0, p_1, ..., p_d)$

initialize: $N_0 = \sigma, w = \mathbb{1}^d, j = 0$

repeat

compute $u = \sum_{i=0}^{d} \frac{w_i}{(d+1)-j} v_i$
Set $w_{p_j} = 0$
Set $N_{j+1} = N_j$ with p_j-th vertex replaced by u
$j = j + 1$

until $j = d$

return: N_d

Algorithm 2 One gradient step in automatic subdivisioning of a d-simplex using one data sample

input: d-simplex $\sigma = [v_0, v_2, ..., v_d]$, sample $x = \sum_{i=0}^{d} w_{x_i} v_i, \Theta = \{w_1, ..., w_l\}$, Loss L
initialize: $N_0 = \sigma, j = 0$
repeat
Compute $u = \sum_{i=0}^{d} w_{j_i} \boldsymbol{N}_{j_i}$
Set $k = arg\min_i \frac{w_{x_i}}{w_{j_i}}$
Set $N_{j+1} = N_j$ with k-th vertex replaced with u
Set w_x as convex combination weights of x represented with vertices in N_{j+1}
$j = j + 1$
until $j = l$
$\Theta = \Theta - \alpha \nabla_\Theta L(\Theta)$
Project each $w_i \in \Theta$ on the standard d-simplex

Estimation – simplicial complex networks

Algorithm 3 training procedure for a general SCN

$S = [v_0, ..., v_d]$, l = depth, θ_b, θ_W (bias function, and weight params), P (network policy), $(x = \sum_{i=0}^{d} w_{x_i} S_i, y)$ (input/output pair), α (learning rate)

// forward pass

for $m \in \{1, ..., l\}$ **do**

  Permute S using P

  $h_m = \sum_{i=0}^{d} w_{m,i}.S_i$

  $f(h_m) = \sum_{i=0}^{d} w_{m,i} f(S_i) + b_m(S; \theta_b)$

  Extract j and update w_x using x, S, and lemma 1

  $S_j = h_m$

end for

$f(x) = \sum_{i=0}^{d} w_{x_i} f(S_i)$

// backward pass and parameter updates

$\theta_b = \theta_b - \alpha \nabla_{\theta_b} \mathcal{L}(f(x), y)$

$\theta_W = \theta_W - \alpha \nabla_{\theta_W} \mathcal{L}(f(x), y)$

for $m \in \{1, ..., l\}$ **do**

  project w_m on the standard d-simplex

end for

Estimation – simplicial complex networks

Reformulation of a linear regression problem using simplicial complex network:

- A real valued linear function $f: \Delta^d \rightarrow R$ from a d-dimensional simplex $\Delta^d = [v_o, \ldots, v_d]$, can be specified by the values of f at each v_i. These values are represented by $f(v_i)$.
- Assume a data matrix $X \in \mathbb{R}^{N \times d}$ of N samples within Δ^d, and their corresponding output in a vector y. We formulate the linear regression problem with training a weight w that minimizes a regresion optimization problem, e.g. $\|Xw - y\|_2^2$.
- We present the coefficients of representation samples in X as a convex combination of $v_0, \ldots, v_d$ in a matrix $C \in \mathbb{R}^{N \times (d+1)}$ with a rank of at most d, where i-th row indicates the corresponding coefficients for i-th sample. Then the linear regression problem above can be reformulated as,

$$\|Cf - y\|_2^2$$

- where f is a $(d+1)$ dimensional vector representing the function value at v_i as its i-th element. With a straightforward computation, one can verify that the optimal w or f can be computed from the optimal value of the other one.

Estimation – simplicial complex networks

- Preliminary results for the asymptotics of the approach:
- Theorem 1: If $d \geq 0$ and λ varies monotonically with respect to dimensions, i.e. $\dot{\lambda}(d) \triangleq d\lambda(d)$ /dd≥ 0 or $\dot{\lambda}(d) \leq 0$, then the state of the neural network is asymptotically stable for the arbitrary initial state, i.e. $\forall v(0) \in v$, $\lim_{d\to\infty} v(d) = \bar{v}$, where $\bar{v}$ is a steady state of the neural network.
- Theorem 2: Let the evaluation function be defined as previously and $\nabla_v E[v(d), \lambda(d)] \not\equiv 0$. Let the subdivision function be defined as earlier and $q\big(v(d)\big) = p(v(d))/c_u$. If $\forall d \geq 0$, $\lambda(d) >$ 0, $p\big(v(d)\big) > 0$ implies that $\dot{\lambda}(d) \geq 0$, and $\dot{\lambda}(d) \not\equiv 0$, then the stable state of the neural network represents a feasible solution of the reformulated problem, i.e. $\forall v(0) \geq V$, $\lim_{d\to\infty} p(v(d)) = 0$ or $\lim_{d\to\infty} v(d) = \bar{v} \in \hat{V}$.
- Theorem 3: Let the evaluation function be defined as previously, where $f(v)$ and $p(v)$ are convex, and $\nabla_v E[v(t), \lambda(t)] \not\equiv 0$. Let the subdivision function be defined as earlier and $q\big(v(d)\big) = p(v(d))/c_u$. If $v(0) \in V$ and $v(0) \notin \hat{V}$, $\forall d \geq 0$, $\lambda(d) > 0$, then the steady state of the neural network represents an optimal solution to the optimization problem.

Generalized Bradley-Terry estimation approach

- Consider an experiment where N judges are asked to rank K items, and assume no ties. The outcome of the experiment is a set of N rankings $\{y^{(n)} \equiv (y_1^{(n)}, \dots, y_K^{(n)}) | n = 1, \dots, N\}$ where a ranking is defined as a permutation of the K rank indices. Each ranking has an associated ordering $\omega^{(n)} \equiv (\omega_1^{(n)}, \dots, \omega_K^{(n)})$ where an ordering is defined as a permutation of the K item indices; judge n puts item $\omega_i^{(n)}$ in position i. Rankings and orderings are related by $\omega_{y_i} = i, y_{\omega_i} = i$.
- The Plackett Luce (PL) model is a distribution over rankings y which is best described in terms of the associated ordering ω. It is parameterised by a vector $v = (v_1, \dots, v_n)$ where $v_i \geq 0$ is associated with item index i:

$$PL(\omega|v) = \prod_{k=1,\dots,K} f_k(v)$$

- where:

$$f_k(v) \equiv f_k(v_{\omega_k}, \dots, v_{\omega_K}) \triangleq \frac{v_{\omega_k}}{v_{\omega_k} + \cdots + v_{\omega_K}}$$

Generalized Bradley-Terry estimation approach

- MM algorithm for estimation of PL model:

$$P_A[\pi(1) \to \cdots \to \pi(k)] = \prod_{i=1}^{k} \frac{v_{\pi(i)}}{v_{\pi(i)} + \cdots + v_{\pi(k)}}$$

- Assuming independent rankings, the log-likelihood may be written as:

$$l(v) = \sum_{j=1}^{N} \sum_{i=1}^{m_j - 1} [\ln v_{a(j,i)} - \ln \sum_{s=i}^{m_j} v_{a(j,s)}]$$

$$Q_k(v) = \sum_{j=1}^{N} \sum_{i=1}^{m_j - 1} [\ln v_{a(j,i)} - \frac{\sum_{s=i}^{m_j} v_{a(j,s)}}{\sum_{s=i}^{m_j} v_{a(j,s)}^{(k)}}]$$

- $Q_k(v)$ minorizes the log likelihood $l(v)$ at $v^{(k)}$, up to a constant.

Properties of the estimator – short simulation evidence

			DGP_1		
Data	**OLS**	**LL**	**MRMD-JS**	**MRMD-GJS**	**GBT**
Gaussian	0.8703	0.8790	0.8730	0.8785	0.9196
10	0.1765	0.1747	0.1808	0.1783	0.1808
	0.9066	0.9157	0.8818	0.8964	0.9289
20	0.1234	0.1172	0.1271	0.1281	0.1126
	0.9215	0.8849	0.9204	0.9252	0.9368
50	0.0979	0.0989	0.0876	0.0909	0.0723
	0.9338	0.9244	0.9279	0.9307	0.9438
100	0.0823	0.0782	0.0850	0.0874	0.0693
Log normal	0.8687	0.8921	0.8376	0.8797	0.9029
10	0.1708	0.1725	0.1746	0.1725	0.1648
	0.9049	0.9001	0.8817	0.8977	0.9213
20	0.1212	0.1176	0.1247	0.1254	0.1190
	0.9296	0.9026	0.9200	0.9321	0.9382
50	0.0979	0.0999	0.0962	0.1022	0.0787
	0.9305	0.9238	0.9314	0.9446	0.9451
100	0.0823	0.0807	0.0859	0.0881	0.0614

Properties of the estimator – short simulation evidence

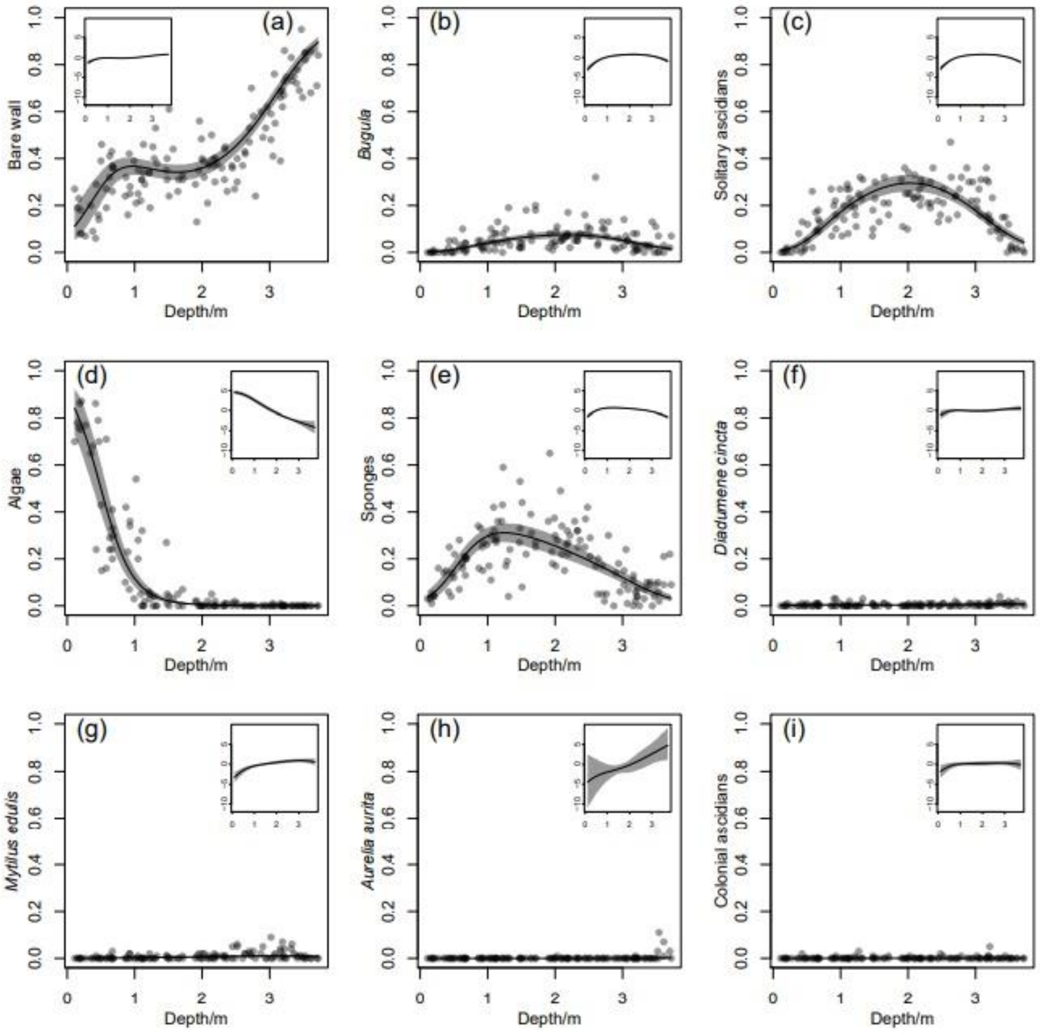
			DGP_2		
Data	**OLS**	**LL**	**MRMD-JS**	**MRMD-GJS**	**GBT**
Gaussian	0.8355	0.8439	0.8555	0.8345	0.9012
10	0.1747	0.1677	0.1754	0.1765	0.1736
	0.8885	0.8699	0.8377	0.8605	0.9103
20	0.1222	0.1137	0.1220	0.1243	0.1092
	0.8754	0.8407	0.8928	0.9067	0.9087
50	0.0940	0.0979	0.0858	0.0891	0.0701
	0.8964	0.8966	0.9001	0.9028	0.8966
100	0.0807	0.0743	0.0808	0.0857	0.0672
Log normal	0.8600	0.8654	0.8209	0.8622	0.8668
10	0.1623	0.1656	0.1711	0.1691	0.1582
	0.8868	0.8911	0.8376	0.8708	0.8937
20	0.1188	0.1164	0.1222	0.1204	0.1166
	0.8924	0.8575	0.8832	0.9041	0.9007
50	0.0960	0.0969	0.0952	0.1002	0.0748
	0.9119	0.9146	0.9035	0.8974	0.9262
100	0.0782	0.0790	0.0842	0.0872	0.0602

Application: out of pocket health care expenses of the elderly

- In a small application that I show below we constructed a combinatorial data set from the Survey of Health, Aging and Retirement in Europe (SHARE) dataset (Waves 5 and 6), combining 8-tuples.
- Dependent variable: out-of-pocket expenditures for medicines as a dependent variable (deciles of the distribution).
- Independent variables: gender, age, years of education, number of chronic diseases and the number of different types of drugs taken by the survey respondent.

	OLS			Local Linear		MRMD-JS		MRMD-GJS	
	Coef.	*(Boot.) SE*	*Sig.*	*Coef.*	*Sig.*	*Coef.*	*Sig.*	*Coef.*	*Sig.*
drugsdif	0.09	0.09		0.08		0.08		0.06	
genddif	1.51	0.38	***	1.12	***	1.65	**	1.71	**
agedif	0.06	0.02	***	0.05	**	0.05	**	0.06	*
eduydif	0.16	0.04	***	0.18	***	0.20	***	0.17	**
chrondif	0.16	0.11		0.20		0.12		0.19	

Applications: sessile hard-substrate marine organisms image data from Italian coast areas



Applications: sessile hard-substrate marine organisms image data from Italian coast areas

$$y_i \sim \text{multinomial}\,(n_i, \rho_i)$$

$$\rho_i = \text{ilr}^{-1} x_i$$

$$x_i = \beta_0 + \beta_1 z_i + \beta_2 {z_i}^2 + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \Sigma)$$

	OLS			Local Linear		MRMD-JS		MRMD-GJS	
	Coef.	*(Boot.) SE*	*Sig.*	*Coef.*	*Sig.*	*Coef.*	*Sig.*	*Coef.*	*Sig.*
x1	0.88	0.74		0.83		0.89		0.96	
x2	1.45	0.38	***	1.46	***	1.45	**	1.48	**
x3	0.69	0.48		0.80	*	0.78		0.81	
x4	1.16	0.04	***	1.13	***	1.24	***	1.47	**
x5	1.07	0.11	***	0.95	**	1.13	***	1.34	***
x6	0.69	0.13	***	0.55	*	0.61	**	0.56	**
x7	0.89	0.32	**	1.01	**	1.21	**	0.98	*
x8	1.31	0.29	***	1.15	**	1.29	***	1.41	**
x9	0.67	0.94		0.65		0.59		0.57	

Conclusion and extensions

- New and unexplored regression perspective, to our knowledge second one on simplicial complexes, opening up vast area for future research with most of the options the approach provides still unexplored, for example:

1) Statistical criteria for the selection of combinations to be included in the combinatorial regression analysis and model fit criteria
2) Extension of the estimation approach and analysis of the properties (as the likelihood is hard to compute – likelihood free approaches: ABC, indirect inference and others)
3) Parametric, semi- and nonparametric perspectives – distributional perspectives remain to be addressed
4) Combinations with other approaches in mathematical statistics and econometrics, for example Bayesian approaches of many types, causal inference, additional combinations with machine learning methods
5) Time series and panel data perspectives
6) Probabilistic perspectives: stochastic processes on simplicial complexes (random walks on simplicial complexes; lattice models, e.g. Ising)
7) Extension of the perspectives from algebraic topology and algebraic statistics – regression models on other topological objects (Vietoris-Rips and Čech complexes, matroids, greedoids, and many other)

Thank you for listening and opportunity to present!

srakara@ier.si

miroslav.verbic@ef.uni-lj.si