Perturbation-based Analysis of Compositional Data

Anton Rask Lundborg

Vienna University of Economics and Business – Research seminar

13 April 2024

Outline

1. Introduction to compositional (simplex-valued) data

2. Perturbations on simplices, perturbation effects and semiparametric estimation

3. Applications to datasets on diversity and gut microbiome

Collaborator

Niklas Pfister
University of Copenhagen

What is compositional data?

Aitchison [1982] defines compositional data as proportions of some whole, that is, a random variable is compositional if it takes values in the unit simplex

$$\Delta^{d-1} := \left\{ z = (z^1, \ldots, z^d) \in [0,1]^d \mid \sum_{j=1}^{d} z^j = 1 \right\}.$$

What is compositional data?

Aitchison [1982] defines compositional data as proportions of some whole, that is, a random variable is compositional if it takes values in the unit simplex

$$\Delta^{d-1} := \left\{ z = (z^1, \dots, z^d) \in [0,1]^d \mid \sum_{j=1}^{d} z^j = 1 \right\}.$$

Compositional data occurs in countless applications:

- geochemistry (e.g., mineral compositions)
- ecology (e.g., relative abundances of species)
- biochemistry (e.g., fatty acid proportions)
- sociology (e.g., time budgets)
- geography (e.g., proportions of land use)
- political science (e.g., voting proportions, research on diversity)
- marketing (e.g., brand shares)
- genomics and microbiome research (e.g., proportions of taxonomic units)

Example: 2022 Danish election data

Consider election counts from the 2022 Danish election for each municipality:

municipality	A	B	...	Å	w/o party	not voted
Aabenraa	9695	661	...	359	36	7979
Aalborg	46098	5621	...	3803	155	29843
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$
Vordingborg	9608	566	...	872	84	6476

Example: 2022 Danish election data

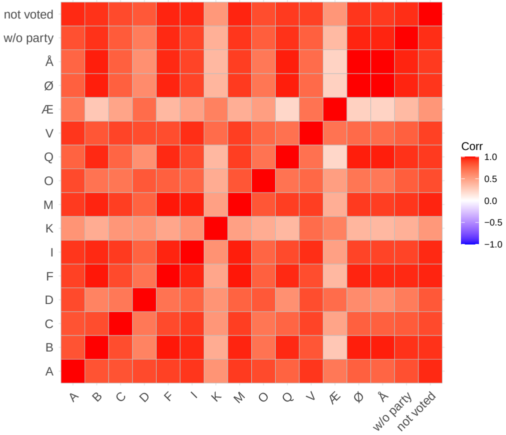
Consider election counts from the 2022 Danish election for each municipality:

municipality	A	B	…	Å	w/o party	not voted
Aabenraa	9695	661	…	359	36	7979
Aalborg	46098	5621	…	3803	155	29843
⋮	⋮	⋮	…	⋮	⋮	⋮
Vordingborg	9608	566	…	872	84	6476

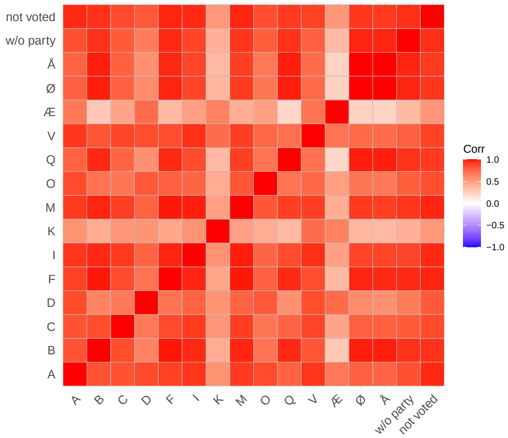
To determine voting patterns, we would like inquire about the relationships between votes for different parties.

Our data analysis might start by estimating correlations between votes for the parties.

2022 Danish election data – count correlations



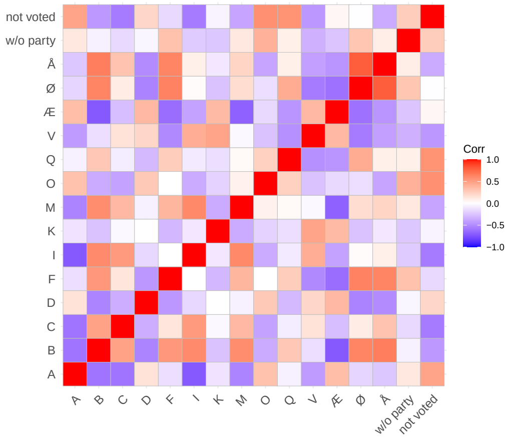
2022 Danish election data – count correlations



All vote counts are highly correlated with the population!

We ignored that the real question is about the proportion of votes.

Example: 2022 Danish election data – proportion correlations



This looks better but is it?

Compositional data and spurious correlations

As early as Pearson [1897], we have known that correlations are not meaningful for compositional data. Pearson argued that even if X, Y and Z are uncorrelated, then $\mathrm{cor}(X/Z, Y/Z) \neq 0$.

Compositional data and spurious correlations

As early as Pearson [1897], we have known that correlations are not meaningful for compositional data. Pearson argued that even if X, Y and Z are uncorrelated, then $\mathrm{cor}(X/Z, Y/Z) \neq 0$.

Let $Z = (Z^1, \dots, Z^d) \in \Delta^{d-1}$. Then, since $\sum_{j=1}^{d} Z^j = 1$,

$$-\mathrm{var}(Z^1) = \sum_{j=2}^{d} \mathrm{cov}(Z^1, Z^j).$$

Compositional data and spurious correlations

As early as Pearson [1897], we have known that correlations are not meaningful for compositional data. Pearson argued that even if X, Y and Z are uncorrelated, then $\mathrm{cor}(X/Z, Y/Z) \neq 0$.

Let $Z = (Z^1, \ldots, Z^d) \in \Delta^{d-1}$. Then, since $\sum_{j=1}^{d} Z^j = 1$,

$$-\mathrm{var}(Z^1) = \sum_{j=2}^{d} \mathrm{cov}(Z^1, Z^j).$$

Similarly, if $Y \in \mathbb{R}$,

$$\sum_{j=1}^{d} \mathrm{cov}(Y, Z^j) = 0.$$

The correlations between components are not meaningful for compositional data!

Example: Effects of presence or absence of microbes

We measure patients and obtain the relative abundance of all gut microbes $Z \in \Delta^{d-1}$ and a binary disease indicator Y.

We want the effect of setting $Z^1 = 0$ on Y.

Example: Effects of presence or absence of microbes

We measure patients and obtain the relative abundance of all gut microbes $Z \in \Delta^{d-1}$ and a binary disease indicator Y.

We want the effect of setting $Z^1 = 0$ on Y.

Regressing Y on $\mathbb{1}_{\{Z^1=0\}} \implies$ misleading if $Z^1 \not\perp\!\!\!\perp Z^{-1}$.

Naively controlling for $Z^{-1} \implies$ the effect is 0 as $Y \perp\!\!\!\perp Z^1 \mid Z^{-1}$ always!

Our proposal explains precisely how to control for the remaining variation in Z.

Summary of existing work

Most existing work on compositional data analysis is based on the work by Aitchison [1982] who proposes a vector space structure on the open simplex by mapping Δ^{d-1} to $\mathbb{R}^{d-1}$ by e.g. the additive log-ratio transform

$$\mathrm{alr}(z)^j := \log(z_j/z_d) \quad \forall j \in \{1, \ldots, d-1\}.$$

Summary of existing work

Most existing work on compositional data analysis is based on the work by Aitchison [1982] who proposes a vector space structure on the open simplex by mapping Δ^{d-1} to $\mathbb{R}^{d-1}$ by e.g. the additive log-ratio transform

$$\mathrm{alr}(z)^j := \log(z_j/z_d) \quad \forall j \in \{1, \ldots, d-1\}.$$

Many modern datasets are high-dimensional, e.g., microbiome or genomics data, and thus require more sophisticated modelling. In particular there is an abundance of zeros which are troublesome for the log-ratio approach.

Can we take a nonparametric perspective in the context of compositional data?

Perturbations (binary)

Let $Y \in \mathbb{R}$ denote a response variable and $Z \in \Delta^{d-1}$ a compositional predictor. We want to summarize changes in the expectation of Y under a pre-specified change in Z.

We specify such changes via perturbations. The simplest perturbations are mappings $\psi : \Delta^{d-1} \times \{0, 1\} \to \Delta^{d-1}$ with $\psi(z, 0) = z$. $\psi(z, 0)$ represents an unperturbed z while $\psi(z, 1)$ represents a perturbed observation.

Perturbations (binary)

Let $Y \in \mathbb{R}$ denote a response variable and $Z \in \Delta^{d-1}$ a compositional predictor. We want to summarize changes in the expectation of Y under a pre-specified change in Z.

We specify such changes via perturbations. The simplest perturbations are mappings $\psi : \Delta^{d-1} \times \{0,1\} \to \Delta^{d-1}$ with $\psi(z,0) = z$. $\psi(z,0)$ represents an unperturbed z while $\psi(z,1)$ represents a perturbed observation.

Letting $f : z \mapsto \mathbb{E}[Y \mid Z = z]$, we define the average (binary) perturbation effect by

$$\lambda_\psi := \frac{\mathbb{E}[f(\psi(Z,1))] - \mathbb{E}[Y]}{\mathbb{P}(Z \neq \psi(Z,1))} = \frac{\mathbb{E}[f(\psi(Z,1))] - \mathbb{E}[f(\psi(Z,0))]}{\mathbb{P}(Z \neq \psi(Z,1))}.$$

The denominator is included to enhance interpretability as one is often interested in how unperturbed points are affected by the perturbation.

The compositional knockout effect (CKE)

Consider a binary perturbation which sets the jth coordinate of z to 0 and rescales the remaining coordinates to lie in the simplex.

Formally, define $\mathsf{C}(z) := z / \sum_{j=1}^{d} z^j$, and let $\psi^j(z, 1)^j := 0$ and $\psi^j(z, 1)^{-j} := \mathsf{C}(z^{-j})$. The compositional knockout effect for the jth feature (CKE^j) is now λ_{ψ^j}.

The CKE summarizes the expected effect on Y of setting Z^j to zero.

The compositional knockout effect (CKE)

Consider a binary perturbation which sets the jth coordinate of z to 0 and rescales the remaining coordinates to lie in the simplex.

Formally, define $\mathsf{C}(z) := z / \sum_{j=1}^{d} z^j$, and let $\psi^j(z,1)^j := 0$ and $\psi^j(z,1)^{-j} := \mathsf{C}(z^{-j})$. The compositional knockout effect for the jth feature (CKE^j) is now λ_{ψ^j}.

The CKE summarizes the expected effect on Y of setting Z^j to zero.

The concept generalizes to settings where some components are set to 0, some are rescaled and some stay fixed.

The compositional knockout effect (CKE)

Consider a binary perturbation which sets the jth coordinate of z to 0 and rescales the remaining coordinates to lie in the simplex.

Formally, define $\mathsf{C}(z) := z / \sum_{j=1}^{d} z^j$, and let $\psi^j(z,1)^j := 0$ and $\psi^j(z,1)^{-j} := \mathsf{C}(z^{-j})$. The compositional knockout effect for the jth feature (CKE^j) is now λ_{ψ^j}.

The CKE summarizes the expected effect on Y of setting Z^j to zero.

The concept generalizes to settings where some components are set to 0, some are rescaled and some stay fixed.

How do we estimate CKE^j from data?

Estimation of average binary perturbation effects

It is helpful to rewrite the estimand slightly. Define $L := \mathbb{1}_{\{Z=\psi(Z,1)\}}$ and $W := \psi(Z,1)$ and note that

$$\mathbb{E}[Y \mid L=1, W] = \frac{\mathbb{E}[YL \mid W]}{\mathbb{E}[L \mid W]} = f(W), \quad \text{so} \quad \lambda_\psi = \frac{\mathbb{E}[\mathbb{E}[Y \mid L=1, W] - Y]}{\mathbb{P}(L=0)}.$$

Estimation of this quantity is well-known in semiparametrics and is related to the estimation of average treatment effects which can utilize machine learning methods.

Estimation of average binary perturbation effects

It is helpful to rewrite the estimand slightly. Define $L \coloneqq \mathbb{1}_{\{Z=\psi(Z,1)\}}$ and $W \coloneqq \psi(Z,1)$ and note that

$$\mathbb{E}[Y \mid L=1, W] = \frac{\mathbb{E}[YL \mid W]}{\mathbb{E}[L \mid W]} = f(W), \quad \text{so} \quad \lambda_\psi = \frac{\mathbb{E}[\mathbb{E}[Y \mid L=1, W] - Y]}{\mathbb{P}(L=0)}.$$

Estimation of this quantity is well-known in semiparametrics and is related to the estimation of average treatment effects which can utilize machine learning methods.

If $f(\psi(Z,1)) - f(Z)$ is constant when $L=0$, then λ_ψ is the coefficient of L in a partially linear model for Y; $\mathbb{E}[Y \mid L, W] = \theta L + h(W)$.

This assumption simplifies estimation of λ_ψ by using debiased/double machine learning requiring just estimates of $\mathbb{E}[Y \mid W]$ and $\mathbb{E}[L \mid W]$ [Chernozhukov et al., 2018].

Semiparametric estimation of perturbation effects

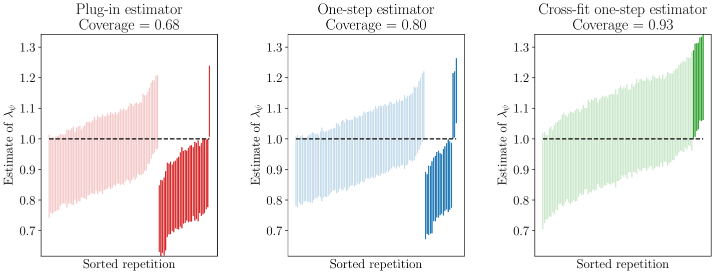
The estimation of functionals is the topic of semiparametric estimation. The primary lessons from this field are:

1. use a one-step corrected estimator (equivalent to Neyman orthogonal),
2. use cross-fitting [Kennedy, 2023].

Semiparametric estimation of perturbation effects

The estimation of functionals is the topic of semiparametric estimation. The primary lessons from this field are:

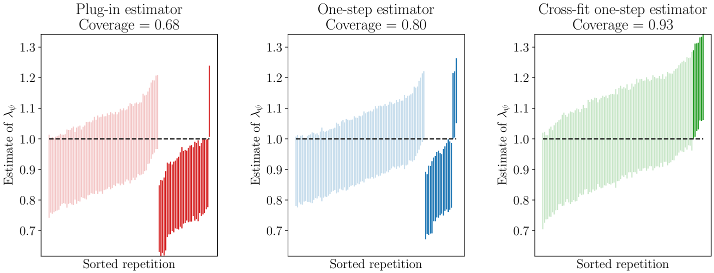
1. use a one-step corrected estimator (equivalent to Neyman orthogonal),
2. use cross-fitting [Kennedy, 2023].



Semiparametric estimation of perturbation effects

The estimation of functionals is the topic of semiparametric estimation. The primary lessons from this field are:

1. use a one-step corrected estimator (equivalent to Neyman orthogonal),
2. use cross-fitting [Kennedy, 2023].



How do we go beyond binary perturbations?

Directional perturbations

We can define perturbations ψ that describe 'local' changes to Z, that is, differences between $\psi(z, 0)$ (doing nothing) and $\psi(z, \epsilon)$ (perturbing slightly) for $\epsilon > 0$.

Directional perturbations

We can define perturbations ψ that describe 'local' changes to Z, that is, differences between $\psi(z, 0)$ (doing nothing) and $\psi(z, \epsilon)$ (perturbing slightly) for $\epsilon > 0$.

Perturbations where $\omega_\psi(z) := \partial_\gamma \psi(z, \gamma) \mid_{\gamma=0}$ exist for all $z \in \Delta^{d-1}$ are directional perturbations. The average (directional) perturbation effect is

$$\tau_\psi := \mathbb{E}\left[\partial_\gamma f(\psi(Z, \gamma)) \mid_{\gamma=0}\right].$$

Directional perturbations

We can define perturbations ψ that describe 'local' changes to Z, that is, differences between $\psi(z,0)$ (doing nothing) and $\psi(z,\epsilon)$ (perturbing slightly) for $\epsilon > 0$.

Perturbations where $\omega_\psi(z) := \partial_\gamma \psi(z,\gamma) \mid_{\gamma=0}$ exist for all $z \in \Delta^{d-1}$ are directional perturbations. The average (directional) perturbation effect is

$$\tau_\psi := \mathbb{E}\left[\partial_\gamma f(\psi(Z,\gamma)) \mid_{\gamma=0}\right].$$

Define the direction and speed of ψ by $v_\psi(z) := \frac{\omega_\psi(z)}{\|\omega_\psi(z)\|_1}$ and $s_\psi(z) := \|\omega_\psi(z)\|_1$, respectively. If ψ and ψ' have the same directions and speeds, then $\tau_\psi = \tau_{\psi'}$.

Thus, it suffices to consider $\psi(z,\gamma) := z + \gamma s_\psi(z) v_\psi(z)$.

The compositional feature influence (CFI)

We can define a directional version of the compositional knockout effect by choosing the direction $v_{\psi^j}(z) := \frac{e_j - z}{\|e_j - z\|_1}$.

We say that the average directional perturbation effect τ_{ψ^j} of any perturbation ψ^j with this direction is a compositional feature influence for the jth feature (CFI^j).

The compositional feature influence (CFI)

We can define a directional version of the compositional knockout effect by choosing the direction $v_{\psi^j}(z) := \frac{e_j - z}{\|e_j - z\|_1}$.

We say that the average directional perturbation effect τ_{ψ^j} of any perturbation ψ^j with this direction is a compositional feature influence for the jth feature (CFI^j).

If $s_\psi = 1$, we say that the perturbation is unit-speed. This is not always an interpretable speed but turns out to be a useful building block. We will consider other speeds later.

Can we rewrite τ_ψ as we did λ_ψ to be able to utilize well-known semiparametric theory?

Derivative-isolating reparametrizations

A reparametrization $\phi : \Delta^{d-1} \to \mathbb{R} \times \mathcal{W}$ is derivative-isolating if

$$\omega_\psi(\phi^{-1}(\ell, w)) = \partial_\ell \phi^{-1}(\ell, w).$$

Derivative-isolating reparametrizations

A reparametrization $\phi : \Delta^{d-1} \to \mathbb{R} \times \mathcal{W}$ is derivative-isolating if

$$\omega_\psi(\phi^{-1}(\ell, w)) = \partial_\ell \phi^{-1}(\ell, w).$$

Using tools from differential geometry, it can be shown that if $(L, W) := \phi(Z)$, then, we have

$$\tau_\psi = \mathbb{E}\left[\partial_\ell \mathbb{E}[Y \mid L = \ell, W] \mid_{\ell = L}\right].$$

Estimating this functional is again well-studied in the semiparametric literature; we can impose a partially linear model and utilize double machine learning once more.

How do we find a derivative-isolating reparametrization?

Unit-speed CFI^j derivative-isolating reparametrization

For the unit-speed CFI^j, the corresponding perturbation is

$$\psi(z,\gamma) = z + \gamma \frac{e_j - z}{\|e_j - z\|_1}$$

and a (somewhat) intuitive choice for a reparametrization $\phi = (\phi^L, \phi^W)$ is given by

$$\phi^L(z) := -\|e_j - z\|_1 \quad \text{and} \quad \phi^W(z) := \frac{e_j - z}{\|e_j - z\|_1}, \quad \text{so that} \quad \phi^{-1}(\ell, w) := \ell w + e_j.$$

Unit-speed CFI^j derivative-isolating reparametrization

For the unit-speed CFI^j, the corresponding perturbation is

$$\psi(z, \gamma) = z + \gamma \frac{e_j - z}{\|e_j - z\|_1}$$

and a (somewhat) intuitive choice for a reparametrization $\phi = (\phi^L, \phi^W)$ is given by

$$\phi^L(z) := -\|e_j - z\|_1 \quad \text{and} \quad \phi^W(z) := \frac{e_j - z}{\|e_j - z\|_1}, \quad \text{so that} \quad \phi^{-1}(\ell, w) := \ell w + e_j.$$

Thus, $\psi(\phi^{-1}(\ell, w), \gamma) = \ell w + e_j + \gamma w$,

$$\partial_\ell \phi^{-1}(\ell, w) = w = \partial_\gamma \psi(\phi^{-1}(\ell, w), \gamma) \mid_{\gamma=0}$$

and therefore ϕ is derivative-isolating. We denote this reparametrization by ϕ_{unit}.

Multiplicative speed

If Z is generated by observing a vector of counts $X \in \mathbb{R}_+^d \setminus \{0\}$ by $Z := \mathrm{C}(X)$, then it could be more natural to look at speeds on the simplex induced by modifying X^j.

Multiplicative speed

If Z is generated by observing a vector of counts $X \in \mathbb{R}^d_+ \setminus \{0\}$ by $Z := \mathrm{C}(X)$, then it could be more natural to look at speeds on the simplex induced by modifying X^j.

We could try an additive perturbation that adds c to X^j, however, since

$$\|\partial_c \mathrm{C}(X + ce_j)\,\big|_{c=0}\,\|_1 = 2\frac{1}{\|X\|_1}\left(1 - \frac{X^j}{\|X\|_1}\right)$$

such a speed is not scale-invariant and therefore ill-defined on the simplex.

Multiplicative speed

If Z is generated by observing a vector of counts $X \in \mathbb{R}_+^d \setminus \{0\}$ by $Z := \mathrm{C}(X)$, then it could be more natural to look at speeds on the simplex induced by modifying X^j.

We could try an additive perturbation that adds c to X^j, however, since

$$\|\partial_c \mathrm{C}(X + ce_j) \mid_{c=0} \|_1 = 2\frac{1}{\|X\|_1}\left(1 - \frac{X^j}{\|X\|_1}\right)$$

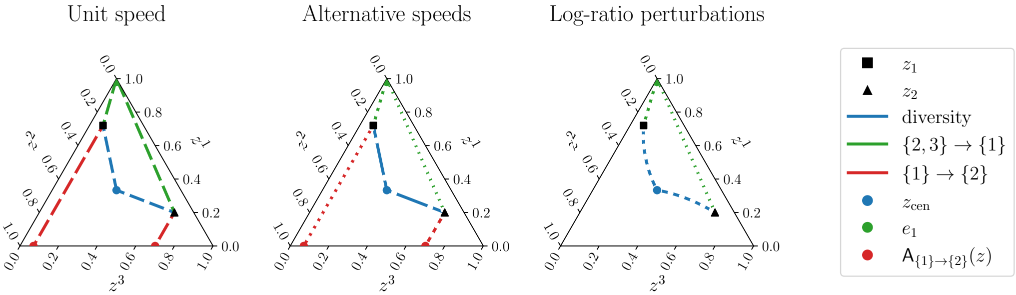
such a speed is not scale-invariant and therefore ill-defined on the simplex.

If we instead consider a multiplicative perturbation that multiplies X^j by $1 + c$, we obtain

$$\|\partial_c \mathrm{C}(X \odot (1 + ce_j)) \mid_{c=0} \|_1 = 2\frac{X^j}{\|X\|_1}\left(1 - \frac{X^j}{\|X\|_1}\right)$$

resulting in the multiplicative speed $s(z) := 2z^j(1 - z^j)$ on the simplex.

Simplex perturbations visualized



Multiplicative speed CFI^j derivative-isolating reparametrization

It turns out that we can use ϕ_{unit} to obtain a reparametrization of CFI^j for any speed.
For a speed s and w (an element of $\mathrm{Im}(\phi_{\mathrm{unit}}^W)$), we define $s_w(\delta) := s(\phi_{\mathrm{unit}}^{-1}(-\delta, w))$.

Multiplicative speed CFI^j derivative-isolating reparametrization

It turns out that we can use ϕ_{unit} to obtain a reparametrization of CFI^j for any speed. For a speed s and w (an element of $\mathrm{Im}(\phi_{\mathrm{unit}}^W)$), we define $s_w(\delta) := s(\phi_{\mathrm{unit}}^{-1}(-\delta, w))$.

If $t_w(\delta)$ solves $-s_w(\delta)^{-1} = \partial_\delta t_w(\delta)$, then ϕ is derivative-isolating;

$$\phi(z) := \left(t_{\phi_{\mathrm{unit}}^W(z)}(\|e_j - z\|_1), \phi_{\mathrm{unit}}^W(z)\right).$$

Multiplicative speed CFI^j derivative-isolating reparametrization

It turns out that we can use ϕ_{unit} to obtain a reparametrization of CFI^j for any speed. For a speed s and w (an element of $\mathrm{Im}(\phi^W_{\mathrm{unit}})$), we define $s_w(\delta) := s(\phi^{-1}_{\mathrm{unit}}(-\delta, w))$.

If $t_w(\delta)$ solves $-s_w(\delta)^{-1} = \partial_\delta t_w(\delta)$, then ϕ is derivative-isolating;

$$\phi(z) := \left(t_{\phi^W_{\mathrm{unit}}(z)}(\|e_j - z\|_1), \phi^W_{\mathrm{unit}}(z) \right).$$

We have $\phi^{-1}_{\mathrm{unit}}(z)^j = \frac{1}{2}\ell + 1$ so $s_w(\delta) = (2-\delta)\delta/2$ and solving;

$$-\frac{2}{(2-\delta)\delta} = \partial_\delta t_w(\delta) \Longrightarrow t_w(\delta) = \log\left(\frac{2-\delta}{\delta}\right) + C.$$

Multiplicative speed CFI^j derivative-isolating reparametrization

It turns out that we can use ϕ_{unit} to obtain a reparametrization of CFI^j for any speed. For a speed s and w (an element of $\mathrm{Im}(\phi^W_{\mathrm{unit}})$), we define $s_w(\delta) := s(\phi^{-1}_{\mathrm{unit}}(-\delta, w))$.

If $t_w(\delta)$ solves $-s_w(\delta)^{-1} = \partial_\delta t_w(\delta)$, then ϕ is derivative-isolating;

$$\phi(z) := \left(t_{\phi^W_{\mathrm{unit}}(z)}(\|e_j - z\|_1), \phi^W_{\mathrm{unit}}(z) \right).$$

We have $\phi^{-1}_{\mathrm{unit}}(z)^j = \frac{1}{2}\ell + 1$ so $s_w(\delta) = (2-\delta)\delta/2$ and solving;

$$-\frac{2}{(2-\delta)\delta} = \partial_\delta t_w(\delta) \Longrightarrow t_w(\delta) = \log\left(\frac{2-\delta}{\delta}\right) + C.$$

Thus (as $\|e_j - z\|_1 = 2(1-z^j)$), $\phi^L(z) := \log\left(\frac{z^j}{1-z^j}\right)$ for multiplicative speed CFI^j.

The compositional diversity influence (CDI)

Another class of perturbations push towards the center of the simplex. We think of these as diversifying perturbations and a unit-speed perturbation is

$$\psi(z, \gamma) = z + \frac{z_{\mathrm{cen}} - z}{\|z_{\mathrm{cen}} - z\|_1}.$$

For any perturbation with the same direction as ψ, we say that τ_ψ is a compositional diversity influence (CDI).

The compositional diversity influence (CDI)

Another class of perturbations push towards the center of the simplex. We think of these as diversifying perturbations and a unit-speed perturbation is

$$\psi(z, \gamma) = z + \frac{z_{\mathrm{cen}} - z}{\|z_{\mathrm{cen}} - z\|_1}.$$

For any perturbation with the same direction as ψ, we say that τ_ψ is a compositional diversity influence (CDI).

We immediately obtain that $\phi_{\mathrm{unit}} = (\phi^L_{\mathrm{unit}}, \phi^W_{\mathrm{unit}})$ given by

$$\phi^L_{\mathrm{unit}}(z) := -\|z_{\mathrm{cen}} - z\|_1 \quad \text{and} \quad \phi^W_{\mathrm{unit}}(z) := \frac{z_{\mathrm{cen}} - z}{\|z_{\mathrm{cen}} - z\|_1}$$

is a derivative-isolating reparametrization for the unit-speed CDI.

CDI with Gini coefficient speed

The conventional way to summarize diversity is by means of a summary statistic, e.g. the Gini coefficient

$$G(z) := \frac{1}{2d} \sum_{j=1}^{d} \sum_{k=1}^{d} |z^j - z^k|.$$

CDI with Gini coefficient speed

The conventional way to summarize diversity is by means of a summary statistic, e.g. the Gini coefficient

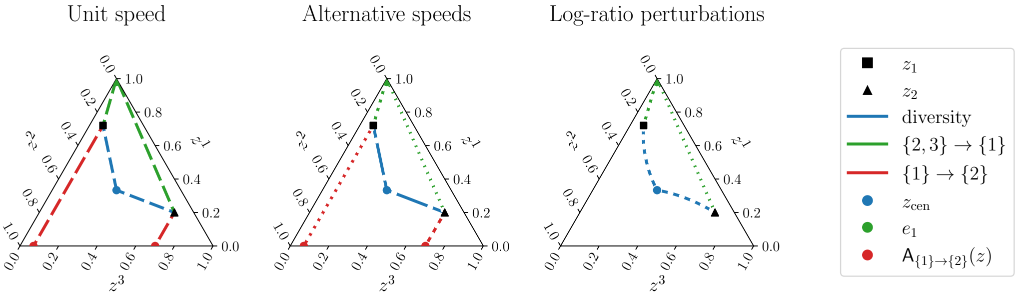
$$G(z) := \frac{1}{2d} \sum_{j=1}^{d} \sum_{k=1}^{d} |z^j - z^k|.$$

It turns out that by using a variant of the argument for CFI^j, we can show that $\phi_{\mathrm{Gini}} = (\phi^L_{\mathrm{Gini}}, \phi^W_{\mathrm{Gini}})$ given by

$$\phi^L_{\mathrm{Gini}}(z) := -G(z) \quad \text{and} \quad \phi^W_{\mathrm{Gini}}(z) := \phi^W_{\mathrm{unit}}(z)$$

is a derivative-isolating reparametrization for a perturbation; the Gini speed CDI.

Simplex perturbations visualized – revisited



Summary of effects

Effect of changes in	Target	$\phi^L(z)$	$\phi^W(z)$
individual components	$\mathrm{CFI}^j_{\mathrm{unit}}$	$-2(1-z^j)$	$\frac{e_j - z}{\lVert e_j - z \rVert_1}$
	$\mathrm{CFI}^j_{\mathrm{mult}}$	$\log(\frac{z^j}{1-z^j})$	$\frac{e_j - z}{\lVert e_j - z \rVert_1}$
	CKE^j	$\mathbb{1}_{\{z^j=0\}}$	$\frac{e_j - z}{\lVert e_j - z \rVert_1}$
diversity	$\mathrm{CDI}_{\mathrm{unit}}$	$-\lVert z_{\mathrm{cen}} - z \rVert_1$	$\frac{z_{\mathrm{cen}} - z}{\lVert z_{\mathrm{cen}} - z \rVert_1}$
	$\mathrm{CDI}_{\mathrm{Gini}}$	$-G(z)$	$\frac{z_{\mathrm{cen}} - z}{\lVert z_{\mathrm{cen}} - z \rVert_1}$
amalgamations	$\mathrm{CAI}^{A\to B}_{\mathrm{unit}}$	$-2\lVert z^A \rVert_1$	see paper
	$\mathrm{CAI}^{A\to B}_{\mathrm{mult}}$	$\log\left(\frac{\lVert z^B \rVert_1}{\lVert z^A \rVert_1}\right)$	see paper
	$\mathrm{CAE}^{A\to B}$	$\mathbb{1}_{\{z^A=0\}}$	see paper

Summary of effects

Effect of changes in	Target	$\phi^L(z)$	$\phi^W(z)$
individual components	$\mathrm{CFI}^j_{\mathrm{unit}}$	$-2(1-z^j)$	$\frac{e_j - z}{\lVert e_j - z \rVert_1}$
	$\mathrm{CFI}^j_{\mathrm{mult}}$	$\log(\frac{z^j}{1-z^j})$	$\frac{e_j - z}{\lVert e_j - z \rVert_1}$
	CKE^j	$\mathbb{1}_{\{z^j=0\}}$	$\frac{e_j - z}{\lVert e_j - z \rVert_1}$
diversity	$\mathrm{CDI}_{\mathrm{unit}}$	$-\lVert z_{\mathrm{cen}} - z \rVert_1$	$\frac{z_{\mathrm{cen}} - z}{\lVert z_{\mathrm{cen}} - z \rVert_1}$
	$\mathrm{CDI}_{\mathrm{Gini}}$	$-G(z)$	$\frac{z_{\mathrm{cen}} - z}{\lVert z_{\mathrm{cen}} - z \rVert_1}$
amalgamations	$\mathrm{CAI}^{A\to B}_{\mathrm{unit}}$	$-2\lVert z^A \rVert_1$	see paper
	$\mathrm{CAI}^{A\to B}_{\mathrm{mult}}$	$\log\left(\frac{\lVert z^B \rVert_1}{\lVert z^A \rVert_1}\right)$	see paper
	$\mathrm{CAE}^{A\to B}$	$\mathbb{1}_{\{z^A=0\}}$	see paper

Use our framework to derive new perturbation effects if your target is not on the list!

Compositional confounding in relationship between income and race

To illustrate the use of CDI, we consider a semisynthetic dataset formed by aggregating individual-level data on income, race and additional predictors into 'communities'. We use the 'Adult' dataset based on the 1994 US census.

Compositional confounding in relationship between income and race

To illustrate the use of CDI, we consider a semisynthetic dataset formed by aggregating individual-level data on income, race and additional predictors into 'communities'. We use the 'Adult' dataset based on the 1994 US census.

The original 48,842 individuals are grouped into 978 observations (by averaging ≈ 50 individuals) and we obtain

Variable	Data	Community aggregation
Y	`compensation`	average
Z	`race` (3 levels)	compute proportions
X	`sex`, `age`, `education`	average/majority vote

Compositional confounding in relationship between income and race

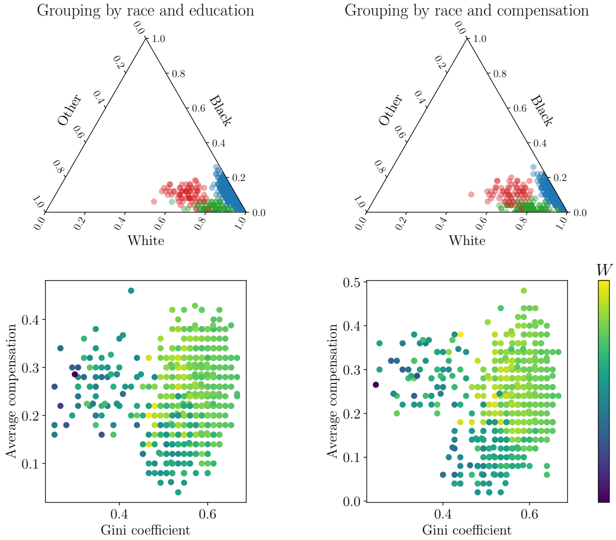
To illustrate the use of CDI, we consider a semisynthetic dataset formed by aggregating individual-level data on income, race and additional predictors into 'communities'. We use the 'Adult' dataset based on the 1994 US census.

The original 48,842 individuals are grouped into 978 observations (by averaging $\approx$ 50 individuals) and we obtain

Variable	Data	Community aggregation
Y	`compensation`	average
Z	`race` (3 levels)	compute proportions
X	`sex`, `age`, `education`	average/majority vote

We group the observations into three different categories to induce compositional confounding; diversity becomes positively associated with compensation, but this is confounded by other aspects of $Z \in \Delta^2$.

Overview of semisynthetic data



Category 1 (green) has average diversity but low education/compensation.

Category 2 (red) has high diversity and high education/compensation.

Category 0 (blue) contains the remaining observations.

The relationship between Gini coefficient and compensation is confounded by $W := \frac{z_{\text{cen}} - Z}{\|z_{\text{cen}} - Z\|_1}$.

Estimated effects of increased diversity

A naive approach to estimating the effect of diversity is computing the Gini coefficient and to compute the effect of the Gini coefficient on Y.

This approach can be modified by controlling for X and/or Z.

Method	Grouping on education		Grouping on compensation	
	Estimate	95% CI	Estimate	95% CI
`naive_diversity`	**−0.082**	(−0.144, −0.019)	**−0.120**	(−0.194, −0.046)
`naive_diversity` $\mid X$	**−0.085**	(−0.142, −0.027)	**−0.106**	(−0.179, −0.034)

Estimated effects of increased diversity

A naive approach to estimating the effect of diversity is computing the Gini coefficient and to compute the effect of the Gini coefficient on Y.

This approach can be modified by controlling for X and/or Z.

Method	Grouping on education		Grouping on compensation	
	Estimate	95% CI	Estimate	95% CI
`naive_diversity`	**-0.082**	$(-0.144, -0.019)$	**-0.120**	$(-0.194, -0.046)$
`naive_diversity` $\mid X$	**-0.085**	$(-0.142, -0.027)$	**-0.106**	$(-0.179, -0.034)$
`naive_diversity` $\mid Z$	-0.219	$(-1.226, 0.788)$	-0.668	$(-1.916, 0.580)$
`naive_diversity` $\mid X, Z$	-0.111	$(-0.902, 0.680)$	0.409	$(-0.403, 1.220)$

Estimated effects of increased diversity

A naive approach to estimating the effect of diversity is computing the Gini coefficient and to compute the effect of the Gini coefficient on Y.

This approach can be modified by controlling for X and/or Z.

Method	Grouping on education		Grouping on compensation	
	Estimate	95% CI	Estimate	95% CI
`naive_diversity`	**−0.082**	(−0.144, −0.019)	**−0.120**	(−0.194, −0.046)
`naive_diversity` $\mid X$	**−0.085**	(−0.142, −0.027)	**−0.106**	(−0.179, −0.034)
`naive_diversity` $\mid Z$	−0.219	(−1.226, 0.788)	−0.668	(−1.916, 0.580)
`naive_diversity` $\mid X, Z$	−0.111	(−0.902, 0.680)	0.409	(−0.403, 1.220)
$\mathrm{CDI}_{\mathrm{Gini}}$	**0.233**	(0.060, 0.406)	**0.614**	(0.429, 0.799)
$\mathrm{CDI}_{\mathrm{Gini}} \mid X$	0.071	(−0.049, 0.191)	**0.611**	(0.455, 0.768)

Only $\mathrm{CDI}_{\mathrm{Gini}}$ correctly captures the sign of the effect!

Variable influence measures when predicting BMI from gut microbiome

To illustrate the use of the CKE and CFI, we analyze the 'American Gut' dataset containing microbiome measurements and metadata from over 10,000 participants.

Our focus is on the relationship between body mass index (BMI) and gut microbiome composition and our goal is to learn which species are important for predicting BMI.

Variable influence measures when predicting BMI from gut microbiome

To illustrate the use of the CKE and CFI, we analyze the 'American Gut' dataset containing microbiome measurements and metadata from over 10,000 participants.

Our focus is on the relationship between body mass index (BMI) and gut microbiome composition and our goal is to learn which species are important for predicting BMI.

After pre-processing, the dataset consists of 4,581 observations of BMI measurements $Y \in \mathbb{R}$ and the relative abundances of 561 microbial species; $Z \in \Delta^{560}$ (on average 60% zeros for each row).

Stability of compositional variable influence measures

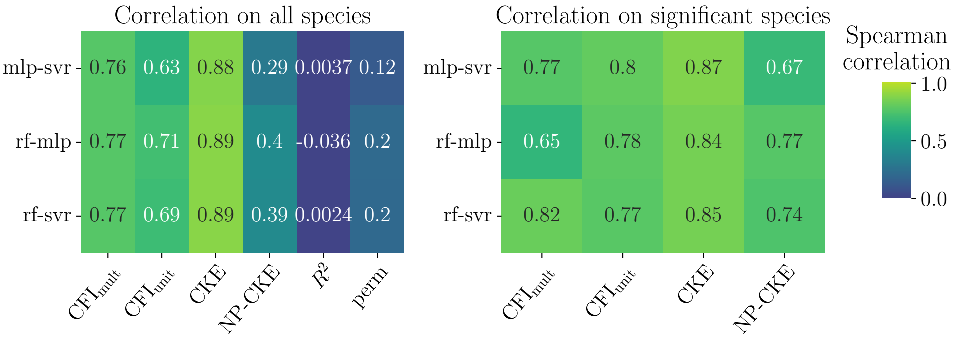
Several nonparametric variable influence measures exist that are regression-agnostic.

Sanity check: results using different ML methods should be similar!

Stability of compositional variable influence measures

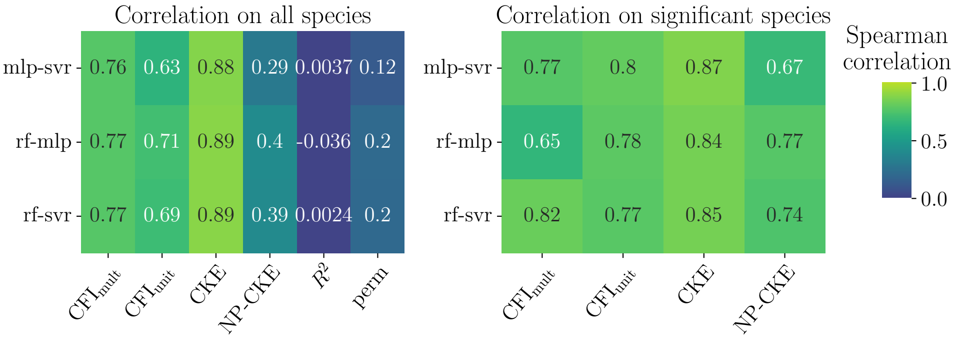
Several nonparametric variable influence measures exist that are regression-agnostic.

Sanity check: results using different ML methods should be similar!



Stability of compositional variable influence measures

Several nonparametric variable influence measures exist that are regression-agnostic.
Sanity check: results using different ML methods should be similar!



It is not possible to ignore the simplex constraint and apply an ordinary partially linear model for Y on Z^j and Z^{-j}; all coefficients exceed 10^{17}!

Comparison of CFI and CKE with log-contrast estimates

The log-ratio-based analysis starts by adding a small positive pseudocount to all observations of Z to remove zeros. The standard pseudocount is the minimum non-zero observation of Z over 2.

A log-contrast regression method is then fit to the data:

$$\mathbb{E}[Y \mid Z] = \sum_{j=1}^{d} \beta^j \log(Z^j) \quad \text{where} \sum_{j=1}^{d} \beta^j = 0.$$

Comparison of CFI and CKE with log-contrast estimates

The log-ratio-based analysis starts by adding a small positive pseudocount to all observations of Z to remove zeros. The standard pseudocount is the minimum non-zero observation of Z over 2.

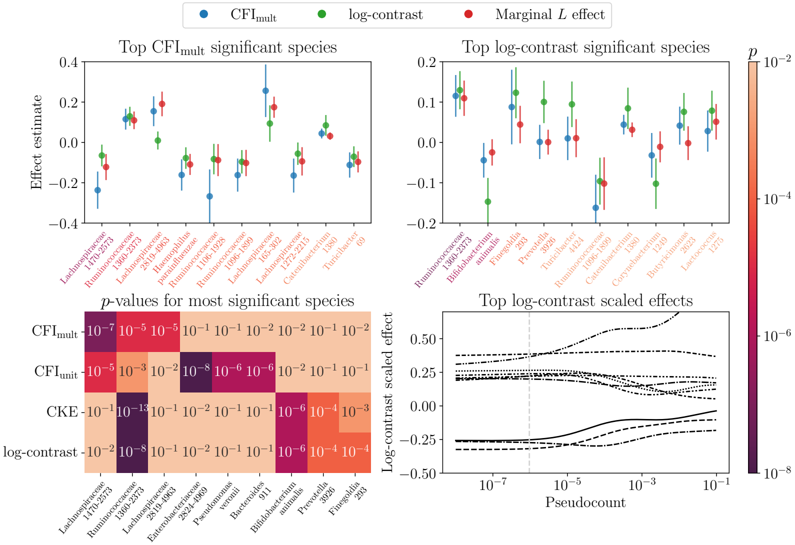
A log-contrast regression method is then fit to the data:

$$\mathbb{E}[Y \mid Z] = \sum_{j=1}^{d} \beta^j \log(Z^j) \quad \text{where} \sum_{j=1}^{d} \beta^j = 0.$$

If Z is high-dimensional, an ℓ^1-penalty can be added and chosen to minimize MSE. These methods predict surprisingly well given their simple structure!

When the true Y on Z model is a log-contrast model, then $\mathrm{CFI}^j_{\mathrm{mult}} = \beta^j$.
How should we interpret log-contrast coefficients when a pseudocount is used?

Comparison of CFI and CKE with log-contrast estimates



Conclusion

- Ignoring compositional structure in data (or, more generally, any ‘manifold’-type constraints) can lead to incorrect conclusions!

Conclusion

- Ignoring compositional structure in data (or, more generally, any 'manifold'-type constraints) can lead to incorrect conclusions!
- Perturbations provide one way of defining interpretable target parameters for compositional variable effect estimates that correct for compositional confounding.

Conclusion

- Ignoring compositional structure in data (or, more generally, any ‘manifold’-type constraints) can lead to incorrect conclusions!
- Perturbations provide one way of defining interpretable target parameters for compositional variable effect estimates that correct for compositional confounding.
- Perturbation effects can be estimated in a regression-agnostic way permitting a trade-off between desired precision and strength of the employed regression estimators.

Conclusion

- Ignoring compositional structure in data (or, more generally, any ‘manifold’-type constraints) can lead to incorrect conclusions!
- Perturbations provide one way of defining interpretable target parameters for compositional variable effect estimates that correct for compositional confounding.
- Perturbation effects can be estimated in a regression-agnostic way permitting a trade-off between desired precision and strength of the employed regression estimators.
- It is an open problem to apply the perturbation framework in other situations with constraints, e.g. directional data.

Conclusion

- Ignoring compositional structure in data (or, more generally, any 'manifold'-type constraints) can lead to incorrect conclusions!
- Perturbations provide one way of defining interpretable target parameters for compositional variable effect estimates that correct for compositional confounding.
- Perturbation effects can be estimated in a regression-agnostic way permitting a trade-off between desired precision and strength of the employed regression estimators.
- It is an open problem to apply the perturbation framework in other situations with constraints, e.g. directional data.

Thank you for listening.

References

J. Aitchison. The statistical analysis of compositional data. Journal of the Royal Statistical Society. Series B (Methodological), 44(2):139–177, 1982. ISSN 00359246. URL http://www.jstor.org/stable/2345821.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097.

Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. arXiv preprint arXiv:2203.06469, 2023.

Anton Rask Lundborg and Niklas Pfister. Perturbation-based analysis of compositional data. arXiv preprint arXiv:2311.18501, 2023.

Karl Pearson. Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London, 60(359-367):489–498, December 1897. doi: 10.1098/rspl.1896.0076.