Semiparametric Analysis of Compositional Data

Anton Rask Lundborg

Aalborg DSTS two-day meeting

9 May 2023

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

$$\Delta^{d-1} := \{x \in \mathbb{R}^d : x_j \geq 0, \sum_{j=1}^{d} x_j = 1\}.$$

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

$$\Delta^{d-1} := \{x \in \mathbb{R}^d : x_j \geq 0, \sum_{j=1}^{d} x_j = 1\}.$$

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)
- marketing (e.g., brand shares)
- genomics and microbiome research (e.g., proportions of taxonomic units)

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

2022 Danish election data

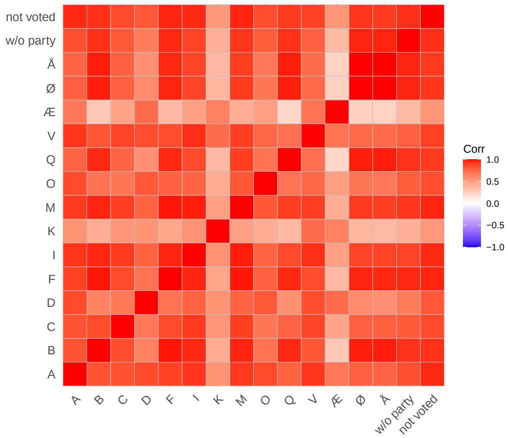
Consider election counts from the 2022 Danish election for each municipality:

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

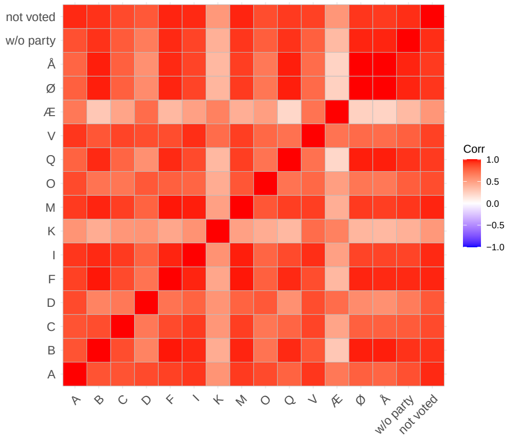
To determine voting patterns, we would like inquire about the relationships between votes for different parties.

Our data analysis might start by looking at the correlation between votes for different parties.

2022 Danish election data – count correlations



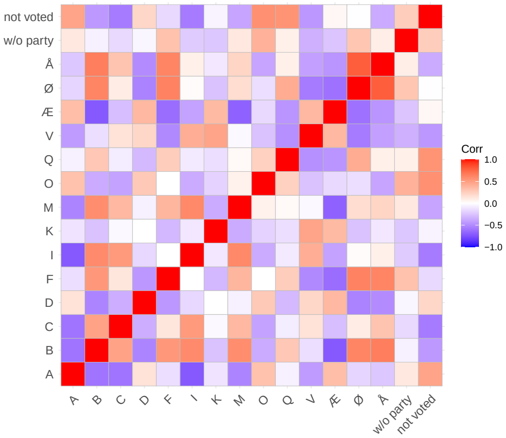
2022 Danish election data – count correlations



All vote counts are highly correlated with the population of the municipality!

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

2022 Danish election data – propotion correlations



This looks better but is it?

Compositional data and spurious correlations

As early as Pearson [1897], it has been noted that correlations are not meaningful for proportional data. Pearson argued that even if X, Y and Z are uncorrelated, then X/Z and Y/Z will always be correlated.

Compositional data and spurious correlations

As early as Pearson [1897], it has been noted that correlations are not meaningful for proportional data. Pearson argued that even if X, Y and Z are uncorrelated, then X/Z and Y/Z will always be correlated.

Let $Z \in \Delta^{d-1}$. 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], it has been noted that correlations are not meaningful for proportional data. Pearson argued that even if X, Y and Z are uncorrelated, then X/Z and Y/Z will always be correlated.

Let $Z \in \Delta^{d-1}$. 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 is a real-valued response,

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

The correlations between components are not meaningful for compositional data!

Log-ratios

Aitchison [1982] proposes a vector space structure on the open simplex by mapping Δ^{d-1} to $\mathbb{R}^{d-1}$ by the additive log-ratio transform

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

Log-ratios

Aitchison [1982] proposes a vector space structure on the open simplex by mapping Δ^{d-1} to $\mathbb{R}^{d-1}$ by the additive log-ratio transform

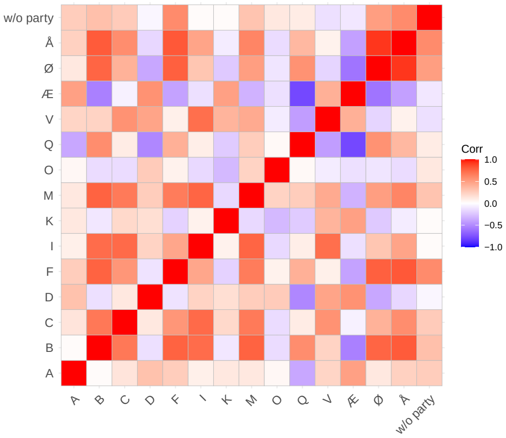
$$\mathrm{alr}(z)_j \mapsto \log(z_j/z_d) \quad \forall j \in \{1, \ldots, d-1\}.$$

Aitchison expanded on this idea to propose the log-contrast regression model

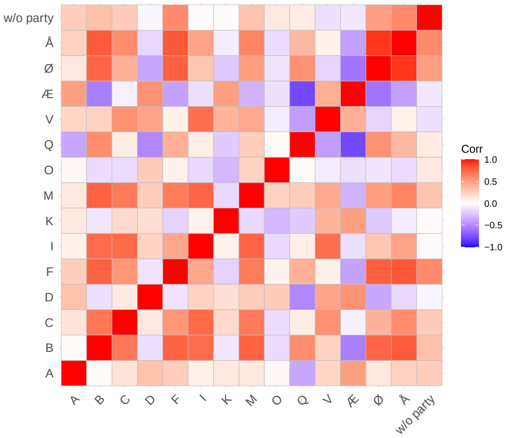
$$Y = \sum_{j=1}^{d} \beta_j \log(Z_j) + \varepsilon, \quad \sum_{j=1}^{d} \beta_j = 0.$$

The techniques proposed by Aitchison are applied extensively in geology and others fields under the CoDA-brand.

2022 Danish election data – log-ratio correlations



2022 Danish election data – log-ratio correlations



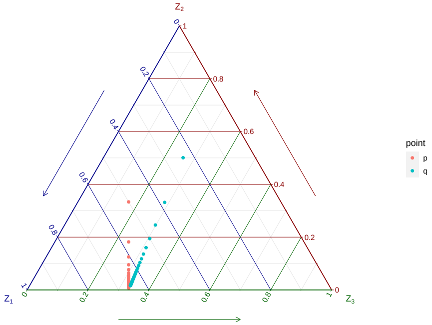
We see that there are many fewer negative correlations between the log-ratios than the raw proportions.

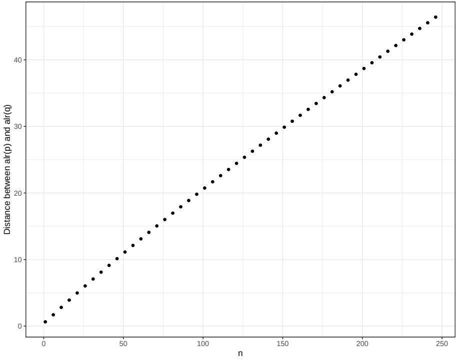
We were forced to add 1 to all counts to ensure each row was in the open simplex.

Problems with log-ratios – zeros

Log-ratio transforms require all data to be strictly positive. It is sometimes argued that adding a small constant is harmless. Is it?

Consider $p_n = (\frac{2}{3} - \frac{1}{n}, \frac{2}{n}, \frac{1}{3} - \frac{1}{n})$ and $q_n = (\frac{2}{3} - \frac{6}{n^{1.1}}, \frac{7}{n^{1.1}}, \frac{1}{3} - \frac{1}{n^{1.1}})$ in Δ^2 [Park et al., 2022].

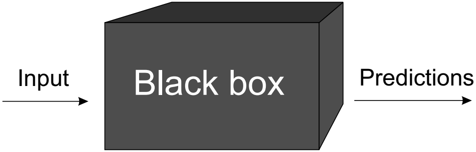




Problems with log-ratios – high-dimensional data and nonparametrics

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.

Black box methods (random forests, boosted trees, neural networks) display superior predictive performance on such datasets.



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

Intermezzo: causal estimation and testing

To provide a causally interpretable analysis, we usually:

1. Define a causal estimand of interest $\psi^* : \mathcal{P}^* \to \mathbb{R}$ defined on $\mathcal{P}^*$, the space of (hypothetical) interventional distributions.
2. Define an observational estimand $\psi : \mathcal{P} \to \mathbb{R}$ defined on $\mathcal{P}$, the space of observational distributions.
3. Provide assumptions under which $\psi^*(P^*) = \psi(P)$.

Intermezzo: causal estimation and testing

To provide a causally interpretable analysis, we usually:

1. Define a causal estimand of interest $\psi^* : \mathcal{P}^* \to \mathbb{R}$ defined on $\mathcal{P}^*$, the space of (hypothetical) interventional distributions.
2. Define an observational estimand $\psi : \mathcal{P} \to \mathbb{R}$ defined on $\mathcal{P}$, the space of observational distributions.
3. Provide assumptions under which $\psi^*(P^*) = \psi(P)$.

The goal of statistics is to provide efficient estimators of $\psi(P)$. In particular, estimators where

$$\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

are desirable to be able to test hypotheses.

Subcompositional irrelevance

For a real-valued response Y and predictors $X \in \mathbb{R}^d$, we are used to testing $Y \perp\!\!\!\perp X_j \,|\, X_{-j}$ when determining the relevance of certain features.

If $Z \in \Delta^{d-1}$, it will always be true that $Y \perp\!\!\!\perp Z_j \,|\, Z_{-j}$; Z_{-j} completely determines Z_j.

Subcompositional irrelevance

For a real-valued response Y and predictors $X \in \mathbb{R}^d$, we are used to testing $Y \perp\!\!\!\perp X_j \,|\, X_{-j}$ when determining the relevance of certain features.

If $Z \in \Delta^{d-1}$, it will always be true that $Y \perp\!\!\!\perp Z_j \,|\, Z_{-j}$; Z_{-j} completely determines Z_j.

We propose making Z_j variation independent from Z_{-j} by projecting into a smaller simplex:

$$\mathbb{C}(Z_{-j}) := \frac{1}{1 - Z_j} Z_{-j}.$$

Subcompositional irrelevance

For a real-valued response Y and predictors $X \in \mathbb{R}^d$, we are used to testing $Y \perp\!\!\!\perp X_j \mid X_{-j}$ when determining the relevance of certain features.

If $Z \in \Delta^{d-1}$, it will always be true that $Y \perp\!\!\!\perp Z_j \mid Z_{-j}$; Z_{-j} completely determines Z_j.

We propose making Z_j variation independent from Z_{-j} by projecting into a smaller simplex:

$$\mathbb{C}(Z_{-j}) := \frac{1}{1 - Z_j} Z_{-j}.$$

We say that Z_j is subcompositionally irrelevant for predicting Y (or just subcompositionally irrelevant) if $Y \perp\!\!\!\perp Z_j \mid \mathbb{C}(Z_{-j})$.

Can we quantify subcompositional irrelevance? Can we test for it?

Compositional feature influence (CFI)

Suppose that we had access to data on the absolute scale X and corresponding compositional variable $Z := \mathbb{C}(X)$.

A natural perturbation on X_j is multiplication with $c \geq 0$. What happens to the composition formed from X under this perturbation?

$$\phi(Z, c) = \frac{1}{1 - z_j + cz_j}(Z_1, \ldots, cZ_j, \ldots, Z_d) \quad \text{but} \quad \mathbb{C}(Z_{-j}) = \mathbb{C}(\phi(Z, c)_{-j})$$

Compositional feature influence (CFI)

Suppose that we had access to data on the absolute scale X and corresponding compositional variable $Z := \mathbb{C}(X)$.

A natural perturbation on X_j is multiplication with $c \geq 0$. What happens to the composition formed from X under this perturbation?

$$\phi(Z, c) = \frac{1}{1 - z_j + cz_j}(Z_1, \ldots, cZ_j, \ldots, Z_d) \quad \text{but} \quad \mathbb{C}(Z_{-j}) = \mathbb{C}(\phi(Z, c)_{-j})$$

Define $W := \mathbb{C}(Z_{-j})$, $f(z, w) := \mathbb{E}[Y \mid Z_j = z, W = w]$, then

$$\mathrm{CFI}_j := \mathbb{E}\left[\left(\frac{\partial}{\partial c} f\left(\frac{cZ_j}{1 - Z_j + cZ_j}, W\right)\right)\Big|_{c=1}\right] = \mathbb{E}\left[Z^j(1 - Z^j)\frac{\partial}{\partial z^j} f(Z_j, W)\right].$$

Compositional feature influence (CFI)

Suppose that we had access to data on the absolute scale X and corresponding compositional variable $Z := \mathbb{C}(X)$.

A natural perturbation on X_j is multiplication with $c \geq 0$. What happens to the composition formed from X under this perturbation?

$$\phi(Z, c) = \frac{1}{1 - z_j + cz_j}(Z_1, \ldots, cZ_j, \ldots, Z_d) \quad \text{but} \quad \mathbb{C}(Z_{-j}) = \mathbb{C}(\phi(Z, c)_{-j})$$

Define $W := \mathbb{C}(Z_{-j})$, $f(z, w) := \mathbb{E}[Y \mid Z_j = z, W = w]$, then

$$\mathrm{CFI}_j := \mathbb{E}\left[\left(\frac{\partial}{\partial c} f\left(\frac{cZ_j}{1 - Z_j + cZ_j}, W\right)\right) \Big|_{c=1}\right] = \mathbb{E}\left[Z^j(1 - Z^j)\frac{\partial}{\partial z^j} f(Z_j, W)\right].$$

$\mathrm{CFI}_j = 0$ corresponds to subcompositional irrelevance if the distribution of Z_j has no atoms. We allow zeros!

Compositional knockout effect (CKE)

In many instances it may be of interest to determine the effect of 'knocking out' a particular part of a composition.

We cannot consider a perturbation that modifies Z by setting $Z_j = 0$, since the resulting vector is no longer in the simplex – we choose to fix $W := \mathbb{C}(Z_{-j})$.

Compositional knockout effect (CKE)

In many instances it may be of interest to determine the effect of 'knocking out' a particular part of a composition.

We cannot consider a perturbation that modifies Z by setting $Z_j = 0$, since the resulting vector is no longer in the simplex – we choose to fix $W := \mathcal{C}(Z_{-j})$.

Let $B := \mathbb{1}_{\{Z_j > 0\}}$ and $f(b, w) := \mathbb{E}[Y \mid B = b, W = w]$, then

$$\mathrm{CKE}_j := \mathbb{E}\left[f(0, W) - f(1, W)\right].$$

$\mathrm{CKE}_j = 0$ corresponds to subcompositional irrelevance if the effect of Z_j is only through B. Naive estimator:

$$\widetilde{\mathrm{CKE}}_j := \frac{1}{n} \sum_{i=1}^{n} \hat{f}(0, W_i) - \hat{f}(1, W_i).$$

Cross-fitting and plug-in bias

Problem: The naive estimator uses the same data for the fitting of $\hat{f}$ and estimating CKE_j. This induces a bias!

Cross-fitting and plug-in bias

Problem: The naive estimator uses the same data for the fitting of $\hat{f}$ and estimating CKE_j. This induces a bias!

Solution: Cross-fit the estimator! Split the data indices into K folds $I_1, \ldots, I_K$, compute $\hat{f}_k$ on I_{-k} and compute

$$\widehat{\mathrm{CKE}}_j := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|I_k|} \sum_{i \in I_k} \hat{f}_k(0, W_i) - \hat{f}_k(1, W_i).$$

Cross-fitting and plug-in bias

Problem: The naive estimator uses the same data for the fitting of $\hat{f}$ and estimating CKE_j. This induces a bias!

Solution: Cross-fit the estimator! Split the data indices into K folds $I_1, \dots, I_K$, compute $\hat{f}_k$ on I_{-k} and compute

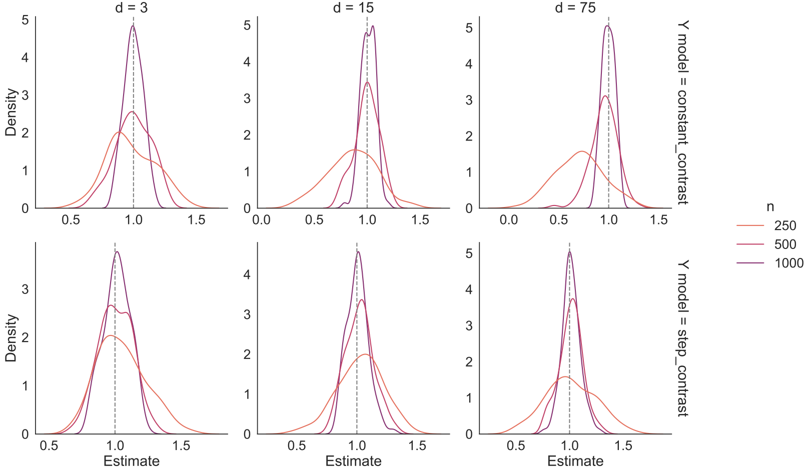
$$\widehat{\mathrm{CKE}}_j := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|I_k|} \sum_{i \in I_k} \hat{f}_k(0, W_i) - \hat{f}_k(1, W_i).$$

We cross-fit all estimators with $K = 2$. In simulations we compare:

`constant_contrast` $Y = \mathbb{1}_{\{B=0\}} + \mathcal{N}(0,1)$

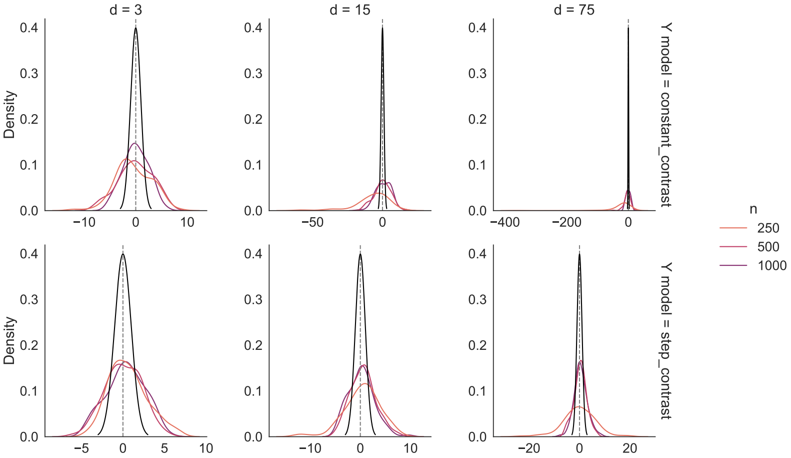
`step_contrast` $Y = \mathbb{1}_{\{B=0\}} \mathbb{1}_{\{W_1 > \mathrm{median}(W_1)\}} + \mathcal{N}(0,1)$

CKE plug-in estimator



Looks good! But what about asymptotic distribution of $\frac{\sqrt{n}}{\hat{\sigma}}(\widehat{\mathrm{CFI}}_j - \mathrm{CFI})$?

CKE plug-in estimator asymptotic distribution



Looks bad! Bias dominates variance of the plug-in estimator even with cross-fitting.

Partially linear double machine learning estimator

Suppose we make the assumption that

$$Y = \theta(1 - B) + h(W) + \varepsilon \quad \mathbb{E}[\varepsilon \mid B, W] = 0.$$

Then $\theta = \mathrm{CKE}_j$ and also

$$\theta = \frac{\mathbb{E}[\mathrm{Cov}(Y, 1 - B \mid W)]}{\mathbb{E}[\mathrm{Var}(1 - B \mid W)]}.$$

Partially linear double machine learning estimator

Suppose we make the assumption that

$$Y = \theta(1 - B) + h(W) + \varepsilon \quad \mathbb{E}[\varepsilon \mid B, W] = 0.$$

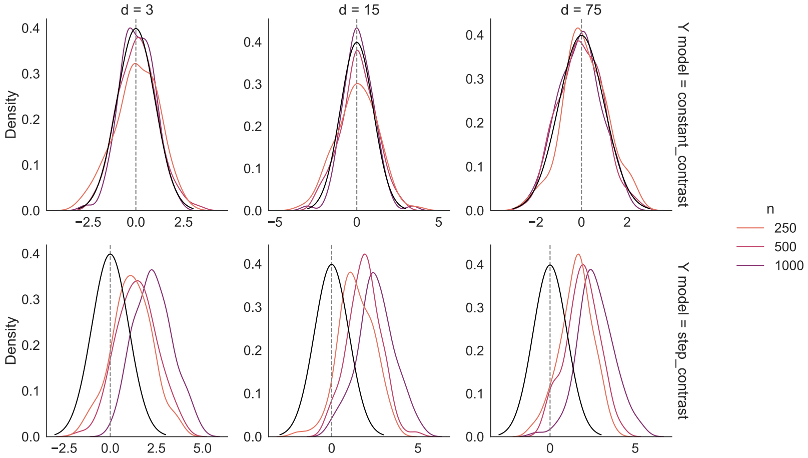
Then $\theta = \mathrm{CKE}_j$ and also

$$\theta = \frac{\mathbb{E}[\mathrm{Cov}(Y, 1 - B \mid W)]}{\mathbb{E}[\mathrm{Var}(1 - B \mid W)]}.$$

Chernozhukov et al. [2018] provide an efficient estimator of θ under conditions on $g(w) := \mathbb{E}[Y \mid W = w]$ and $\pi(w) := \mathbb{E}[1 - B \mid W = w]$:

$$\hat{\theta} = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{i \in I_k} \{Y_i - \hat{g}_k(W_i)\}\{1 - B_i - \hat{\pi}_k(W_i)\}}{\sum_{i \in I_k} \{1 - B_i - \hat{\pi}_k(W_i)\}^2}.$$

CKE DML estimator asymptotic distribution



Looks good if model assumptions hold! Can we do better?

One-step theory

Using semiparametric theory [Kennedy, 2023], we can give a general approach to correcting the bias of a functional ψ.

Using these principles, we get a new estimator:

$$\widehat{\mathrm{CKE}}_j := \frac{1}{K}\sum_{k=1}^{K} |I_k|^{-1} \sum_{i \in I_k} \hat{f}_k(0, W_i) - \hat{f}_k(1, W_i) + \frac{Y_i - \hat{f}_k(B_i, W_i)}{\hat{\pi}_k(B_i \mid W_i)}(1 - 2B_i)$$

One-step theory

Using semiparametric theory [Kennedy, 2023], we can give a general approach to correcting the bias of a functional ψ.

Using these principles, we get a new estimator:

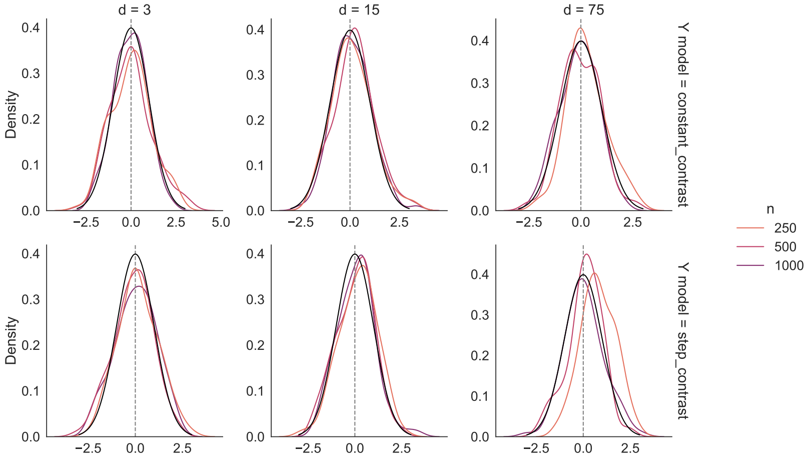
$$\widehat{\mathrm{CKE}}_j := \frac{1}{K}\sum_{k=1}^{K} |I_k|^{-1} \sum_{i \in I_k} \hat{f}_k(0, W_i) - \hat{f}_k(1, W_i) + \frac{Y_i - \hat{f}_k(B_i, W_i)}{\hat{\pi}_k(B_i \mid W_i)}(1 - 2B_i)$$

Closely related to the augmented inverse propensity weighted (AIPW) estimator.

We prove, under conditions that,

$$\frac{\sqrt{n}}{\hat{\sigma}}(\widehat{\mathrm{CKE}}_j - \mathrm{CKE}) \xrightarrow{d} \mathcal{N}(0, 1).$$

One-step sims



Looks good!

Conclusion

- Compositional data are data that lie in a unit simplex.
- The analysis of compositional data using conventional techniques can lead to misleading results.
- A nonparametric perspective is essential when dealing with complicated, high-dimensional (compositional) data.
- Semiparametric theory allows us to construct efficient estimators of causally interpretable quantities.

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. URL https://doi.org/10.1111/ectj.12097.

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

Junyoung Park, Changwon Yoon, Cheolwoo Park, and Jeongyoun Ahn. Kernel methods for radial transformed compositional data with many zeros. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 17458–17472. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/park22d.html.

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. URL https://doi.org/10.1098/rspl.1896.0076.