A log-ratio approach to cluster analysis of count data when the total is irrelevant

M. Comas-Cufí[1]
marc.comas@udg.edu

J.A. Martín-Fernández[1]
josepantoni.martin@udg.edu

G. Mateu-Figueras[1]
gloria.mateu@udg.edu

J. Palarea-Albaladejo[2]
javier.palarea@bioss.ac.uk

[1] Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona, Girona

[2] Biomathematics and Statistics Scotland, Edinburgh

Compositional data analysis

- **Compositional data** (CoDa), $(p_1, \ldots, p_D)$, are quantitative descriptions of the parts of some whole, conveying **relative information**. CoDa are commonly expressed in proportions, percentages, or ppm (Aitchison, 1986).
- The simplex

$$S^D = \left\{ (p_1, \ldots, p_D) \mid p_i > 0, \sum_{i=1}^{D} p_i = \kappa \right\},$$

is the sample space of CoDa.

- **Log-ratios** of parts handle relative information and satisfy desirable properties such as scale invariance and subcompositional coherence:

$$\log\left(\frac{p_i}{p_j}\right), \log\left(\frac{p_j}{\sqrt[D]{\prod_{\ell=1}^{D} p_\ell}}\right), \sqrt{\frac{j}{j+1}} \log\frac{\sqrt[j]{\prod_{\ell=1}^{j} p_\ell}}{p_{j+1}}, \ldots$$

CoDa and the zero problem

- Zeros prevent from using log-ratios. Most proposals have been focused on the continuous case (**zCompositions** R package; Palarea-Albaladejo & Martín-Fernández, 2015).
- **Compositional count data sets:** discrete vectors of number of outcomes falling into mutually exclusive categories.

Municipality	jxsi	psc	pp	catsp	cs	cup
S. Jaume de F.	14	1	0	2	0	5
Gisclareny	20	0	0	0	1	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
L'Hosp. de Llob.	23843	28947	14336	16855	29773	7528
Barcelona	326376	100806	80529	85841	155361	87774

- **Assumption 1:** The relative information is relevant, the total is not.
- **Assumption 2:** The probability of a part different of zero is not zero. Zeros due to sampling limitations.

Parametric approaches to cluster compositional count data set (1)

- **Compositional and count variability not taken into account.**
- **Count variability taken into account, but not compositional variability.** Mixtures of multinomial distributions.
- **Compositional variability taken into account, but not count variability.** Zero multiplicative replacement methods.
  - Dirichlet prior (Martín-Fernández *et al.*, 2015)
  - Log-ratio normal prior (Comas-Cufí *et al.*, 2019)

Parametric approaches to cluster compositional count data set (2)

- **Compositional and count variability taken into account**.
  - *Topic models*. Mixture of multinomials where mixing proportions are modelled in the Simplex.
    - Latent Dirichlet Allocations (Blei *et al.* 2003)
    - Correlated Topic Models (Blei & Lafferty, 2007)
  - *Mixtures of compounding distributions*.
    - Mixtures of Dirichlet-multinomial distributions (Holmes et al., 2012)
    - Mixtures of log-ratio-normal-multinomial distributions (Comas-Cufí et al., 2017)

Parametric approaches to cluster compositional count data set (2)

- **Compositional and count variability taken into account.**
  - *Topic models.* Mixture of multinomials where mixing proportions are modelled in the Simplex.
    - Latent Dirichlet Allocations (Blei *et al.* 2003)
    - Correlated Topic Models (Blei & Lafferty, 2007)
  - *Mixtures of compounding distributions.*
    - Mixtures of Dirichlet-multinomial distributions (Holmes et al., 2012)
    - Mixtures of log-ratio-normal-multinomial distributions (Comas-Cufí et al., 2017)

Main limitations

- Dirichlet-based approaches have **modelling issues**.
- Gaussian-based approaches have **estimation issues**.

Our proposal

Classical clustering approaches applied to cluster count data

1. **Dealing with zeros**.
    - Expected values of a Dirichlet-multinomial (DM) distribution seems to be conservative in keeping the covariance structure observed in the original count data set. The regression toward the mean is moderate.
    - Zero replacement is even more conservative in keeping covariance structure observed in the count data set. But counts with small parts tend to define clusters by themselves.
2. **Compositional variability**. Model your data using a generic distribution defined on the Simplex. Gaussian mixtures are easy to estimate.
3. **Count variability**. Create B new samples using the posterior distribution, and find clusters using classical methods on its log-ratio coordinates.
4. **Consensus clustering**. Use a clustering ensemble method to build a final cluster (e.g. majority voting (Dudoit & Fridlyand, 2003)).

Example: 2017 Catalan regional election

Multivariate count data set

- 947 municipalities. To illustrate the approach we only consider three parts obtained with the following amalgamations.
  - **Pro-independence parties (ind)**: CUP (cup), Esquerra Repúblicana de Catalunya (erc), Junts per Catalunya (jxcat).
  - **Anti-independence parties (esp)**: Ciutadans (cs), Partit Popular (pp), Partit Socialista de Catalunya (psc).
  - **Mixed opinions (other)**: Catalunya si que es pot (catsp), others (other).

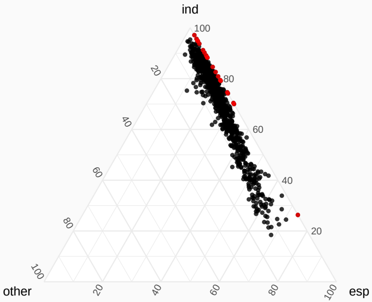
mun	catsp	cs	cup	erc	jxcat	other	pp	psc
Abella de la Conca	4	9	8	30	50	0	0	13
Abrera	815	2559	198	1411	634	158	321	1487
Agramunt	80	472	100	1148	956	7	125	161
Aguilar de Segarra	1	12	38	37	85	0	3	9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Multivariate count data set

- 947 municipalities. To illustrate the approach we only consider three parts obtained with the following amalgamations.
    - **Pro-independence parties (ind)**: CUP (cup), Esquerra Repúblicana de Catalunya (erc), Junts per Catalunya (jxcat).
    - **Anti-independence parties (esp)**: Ciutadans (cs), Partit Popular (pp), Partit Socialista de Catalunya (psc).
    - **Mixed opinions (other)**: Catalunya si que es pot (catsp), others (other).

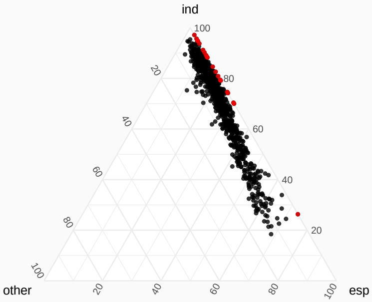
mun	ind	esp	other
Abella de la Conca	88	22	4
Abrera	2243	4367	973
Agramunt	2204	758	87
Aguilar de Segarra	160	24	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Dealing with zeros



- Most municipalities lie between **ind** and **esp** parties.
- Some municipalities have some zero (see ●).

Dealing with zeros



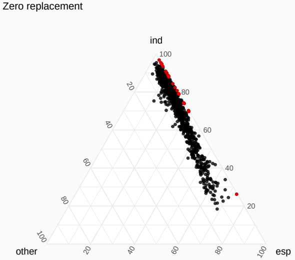
- Most municipalities lie between **ind** and **esp** parties.
- Some municipalities have some zero (see ●).

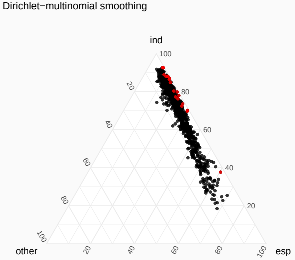
→ **We will deal with zeros first.**

Dealing with zeros

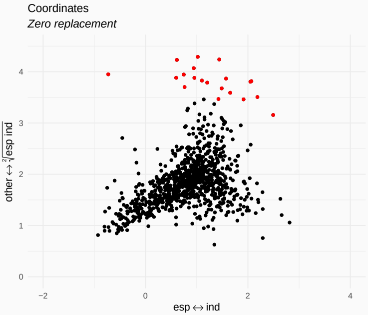
Here, we consider two different approaches:

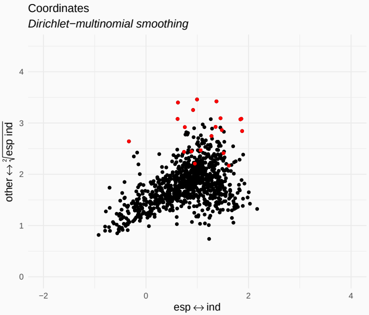
- Geometric Bayesian multiplicative (Martín-Fernández, 2015), and
- Dirichlet-multinomial smoothing after replacing by the expected posterior probabilities.





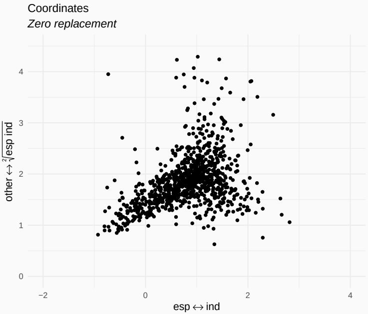
Dealing with zeros

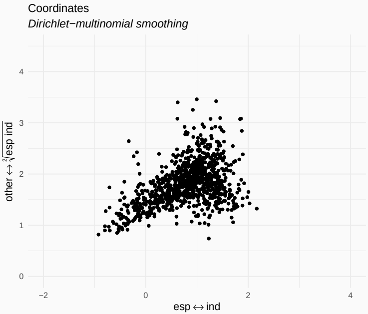




Basis $\mathcal{B}$	ind	esp	other
B1	1	-1	0
B2	1	1	-1

Clustering directly in count data

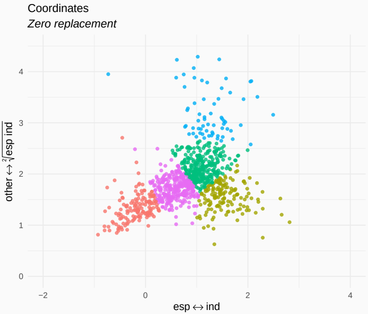


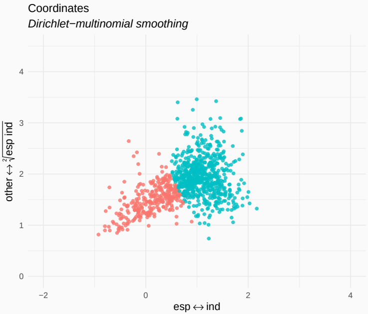


We can cluster our compositional data for example using k-means.

- Duda-Hart test was used to discard one cluster.
- Calinski-Harabasz index was used to select k between 2 and 10.

Clustering directly in count data

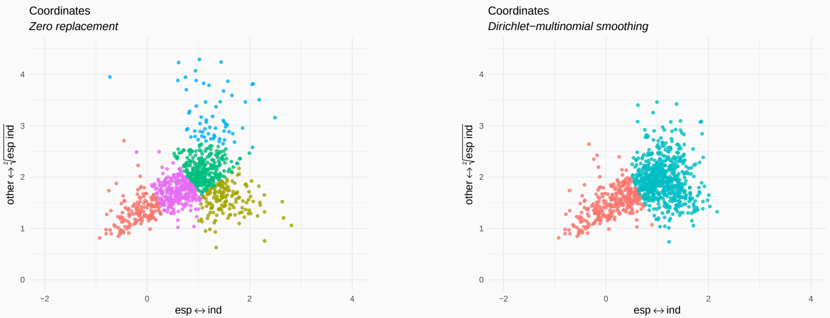




We can cluster our compositional data for example using k-means.

- Duda-Hart test was used to discard one cluster.
- Calinski-Harabasz index was used to select k between 2 and 10.

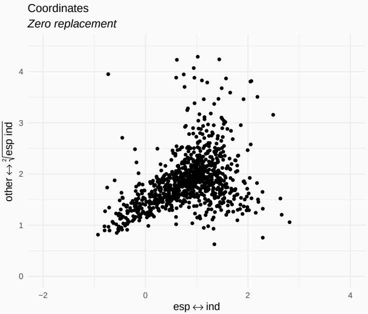
Clustering directly in count data

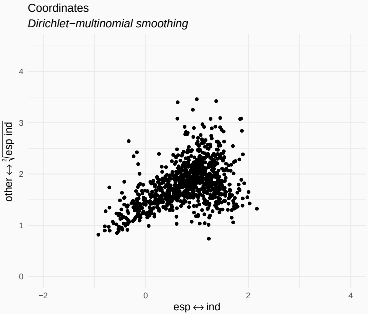


Limitations

- In the zero-replacement approach, observations with a small amount of counts tend to create clusters.
- In DM smoothing results can be affected by the Dirichlet prior.

Compositional variability



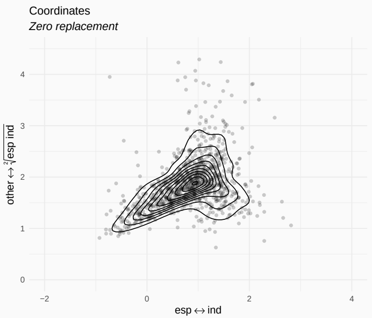


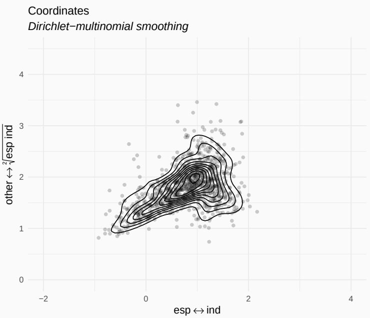
Modelling using Gaussian mixtures

Find a distribution to model the original sample. Mixtures of Gaussian distribution are a good option (Nguyen & McLachlan, 2018)

$\rightarrow$ We estimate a mixture of *ten* Gaussian distributions with equal volume.

Compositional variability



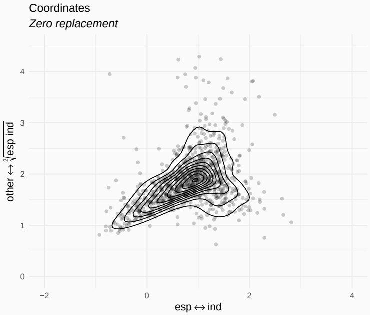


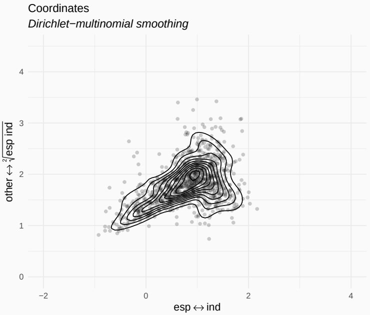
Modelling using Gaussian mixtures

Find a distribution to model the original sample. Mixtures of Gaussian distribution are a good option (Nguyen & McLachlan, 2018)

$\rightarrow$ We estimate a mixture of *ten* Gaussian distributions with equal volume.

Count variability



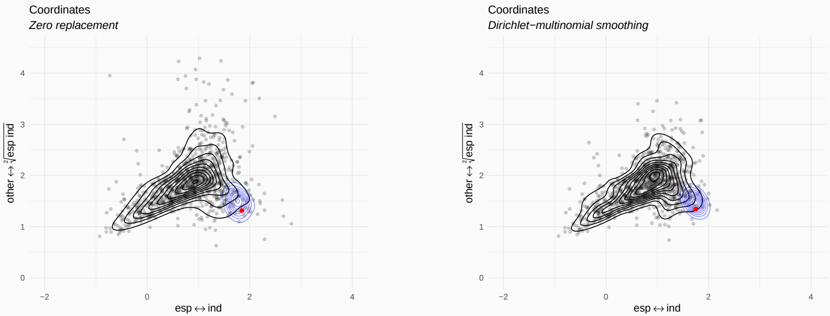


Sampling from the posterior distribution

For each count observation $\mathbf{x}_i$, we sample from the posterior distribution $P(\mathbf{h} \mid \mathbf{x}_i; f)$, where f is the mixture of Gaussian distributions.

- Metropolis–Hastings using $g(\mathbf{h}) = f(\mathbf{h}) \cdot \text{Mult}(\mathbf{x}_i; \mathbf{ilr}_{\mathcal{B}}^{-1}(\mathbf{h}))$, and proposal step using the Laplace approximation of $P(\mathbf{h} \mid \mathbf{x}_i; f)$ centred at zero.

Count variability

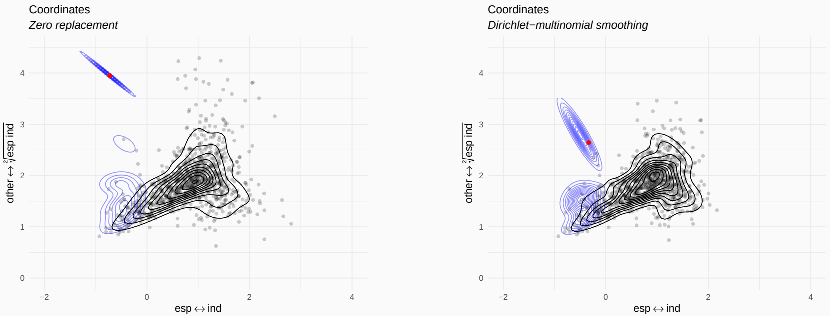


Sampling from the posterior distribution

For each count observation $\mathbf{x}_i$, we sample from the posterior distribution $P(\mathbf{h} \mid \mathbf{x}_i; f)$, where f is the mixture of Gaussian distributions.

- Metropolis–Hastings using $g(\mathbf{h}) = f(\mathbf{h}) \cdot \text{Mult}(\mathbf{x}_i; \mathbf{ilr}_{\mathcal{B}}^{-1}(\mathbf{h}))$, and proposal step using the Laplace approximation of $P(\mathbf{h} \mid \mathbf{x}_i; f)$ centred at zero.
- $i =$ "Argelaguer", $\mathbf{x}_i = (\text{ind: } 259, \text{esp: } 19, \text{other: } 14)$

Count variability

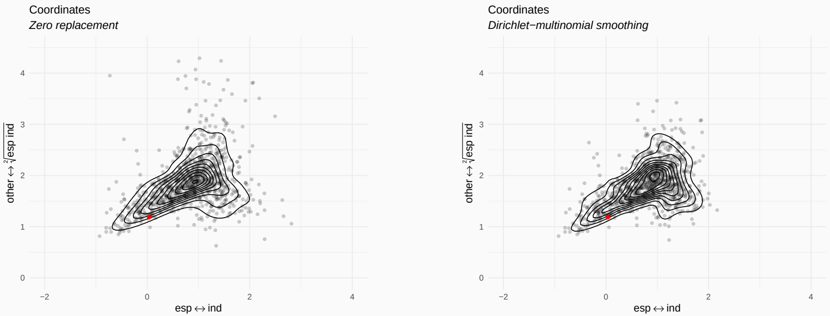


Sampling from the posterior distribution

For each count observation $\mathbf{x}_i$, we sample from the posterior distribution $P(\mathbf{h} \mid \mathbf{x}_i; f)$, where f is the mixture of Gaussian distributions.

- Metropolis–Hastings using $g(\mathbf{h}) = f(\mathbf{h}) \cdot \mathrm{Mult}(\mathbf{x}_i; \mathbf{ilr}_{\mathcal{B}}^{-1}(\mathbf{h}))$, and proposal step using the Laplace approximation of $P(\mathbf{h} \mid \mathbf{x}_i; f)$ centred at zero.
- $i =$ "Arres", $\mathbf{x}_i = (\text{ind: } 10, \text{esp: } 28, \text{other: } 0)$

Count variability

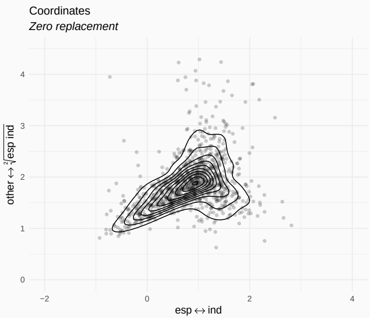


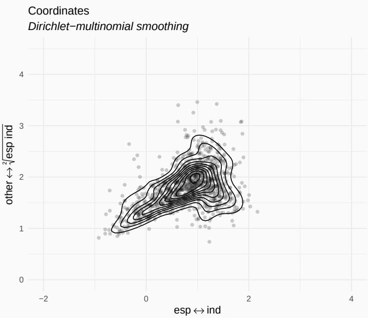
Sampling from the posterior distribution

For each count observation $\mathbf{x}_i$, we sample from the posterior distribution $P(\mathbf{h} \mid \mathbf{x}_i; f)$, where f is the mixture of Gaussian distributions.

- Metropolis–Hastings using $g(\mathbf{h}) = f(\mathbf{h}) \cdot \text{Mult}(\mathbf{x}_i; \mathbf{ilr}_{\mathcal{B}}^{-1}(\mathbf{h}))$, and proposal step using the Laplace approximation of $P(\mathbf{h} \mid \mathbf{x}_i; f)$ centred at zero.
- $i =$ "Barcelona", $\mathbf{x}_i = (\text{ind: } 429\,782, \text{esp: } 405\,924, \text{other: } 96\,748)$

Compositional and count variability

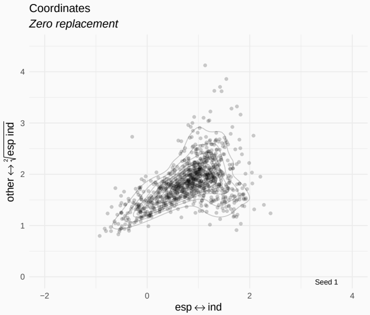


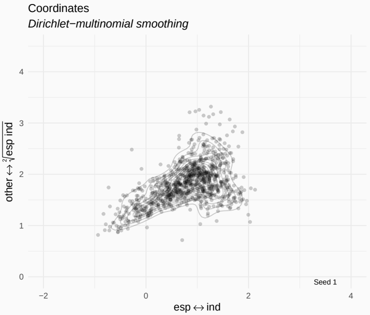


Creating new samples

→ Using the posterior distribution we can create B new samples ($B = 100$),

Compositional and count variability

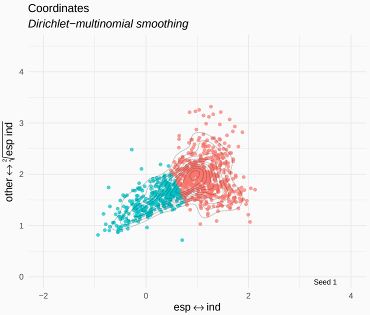




Creating new samples

- $\rightarrow$ Using the posterior distribution we can create B new samples ($B = 100$),
- $\rightarrow$ and applying a clustering algorithm (k-means, $k \in \{2, \dots, 10\}$, Calinski-Harabasz index).

Compositional and count variability

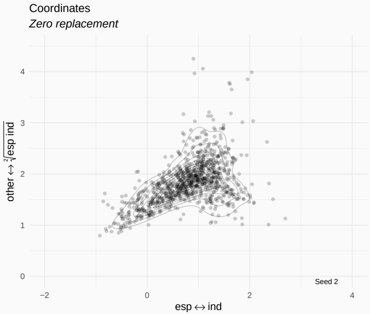


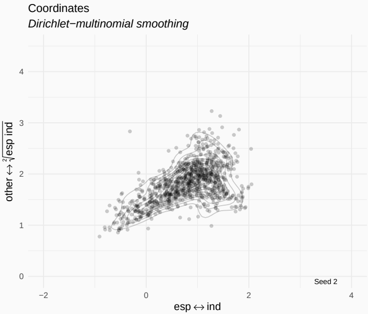
Creating new samples

→ Using the posterior distribution we can create B new samples ($B = 100$),

→ and applying a clustering algorithm (k-means, $k \in \{2, \ldots, 10\}$, Calinski-Harabasz index).

Compositional and count variability



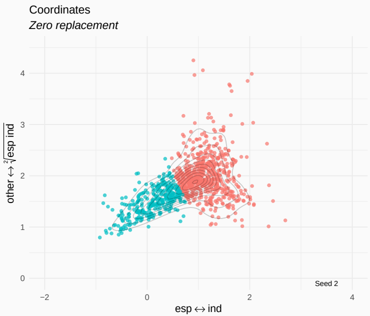


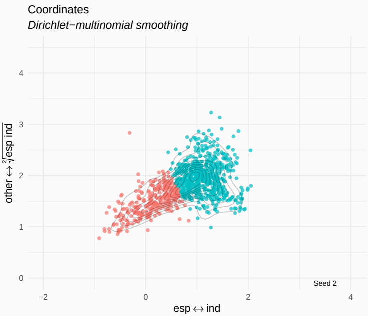
Creating new samples

→ Using the posterior distribution we can create B new samples ($B = 100$),

→ and applying a clustering algorithm (k-means, $k \in \{2, \ldots, 10\}$, Calinski-Harabasz index).

Compositional and count variability



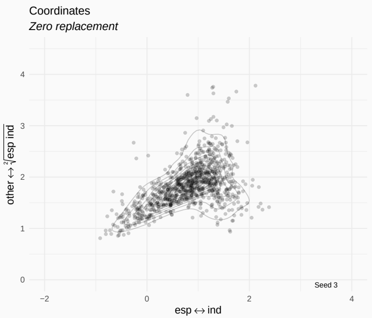


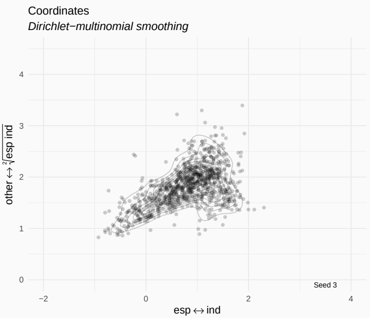
Creating new samples

→ Using the posterior distribution we can create B new samples ($B = 100$),

→ and applying a clustering algorithm (k-means, $k \in \{2, \ldots, 10\}$, Calinski-Harabasz index).

Compositional and count variability



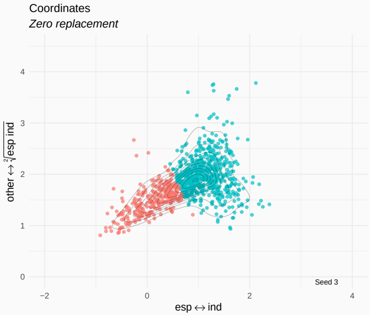


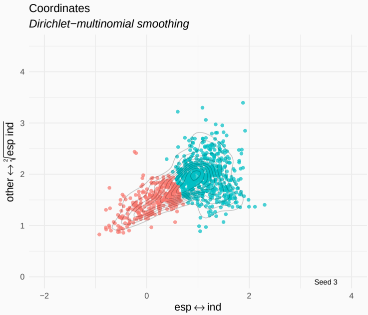
Creating new samples

→ Using the posterior distribution we can create B new samples ($B = 100$),

→ and applying a clustering algorithm (k-means, $k \in \{2, \ldots, 10\}$, Calinski-Harabasz index).

Compositional and count variability



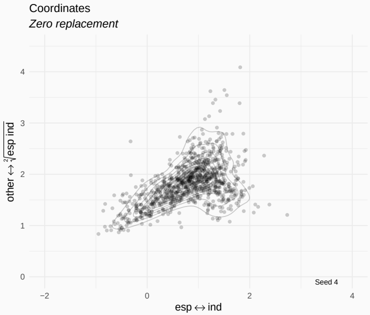


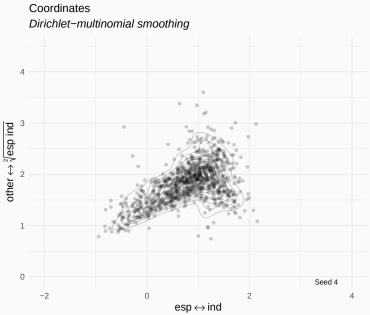
Creating new samples

→ Using the posterior distribution we can create B new samples ($B = 100$),

→ and applying a clustering algorithm (k-means, $k \in \{2, \ldots, 10\}$, Calinski-Harabasz index).

Compositional and count variability



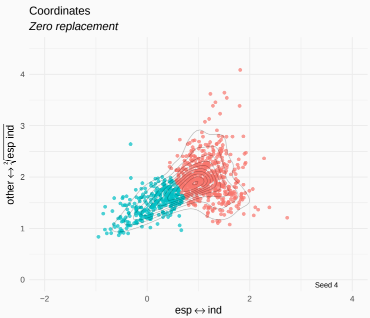


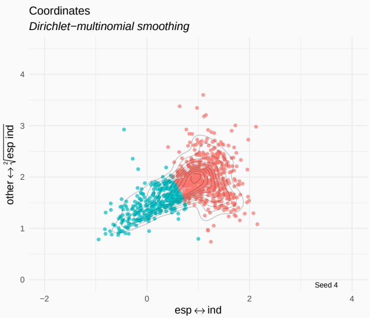
Creating new samples

→ Using the posterior distribution we can create B new samples ($B = 100$),

→ and applying a clustering algorithm (k-means, $k \in \{2, \dots, 10\}$, Calinski-Harabasz index).

Compositional and count variability





Creating new samples

→ Using the posterior distribution we can create B new samples ($B = 100$),

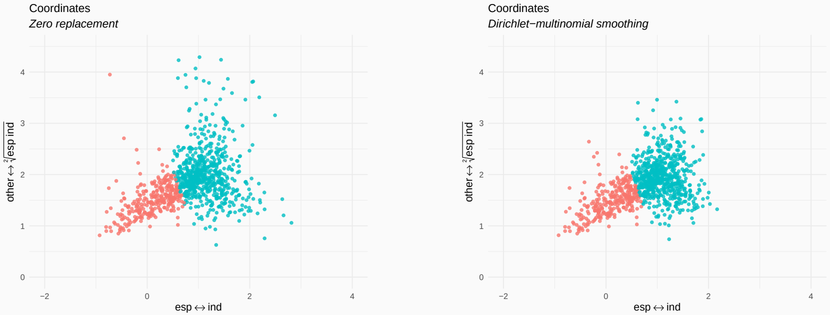
→ and applying a clustering algorithm (k-means, $k \in \{2, \ldots, 10\}$, Calinski-Harabasz index).

Consensus clustering

- Consensus clustering can be applied to summarise all B clusterings.
  → Majority voting (Dudoit and Fridlyand, 2003)).

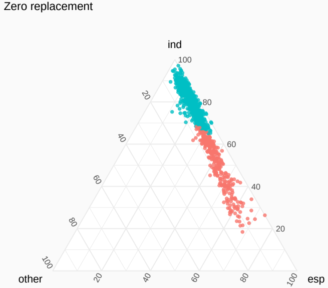
Consensus clustering

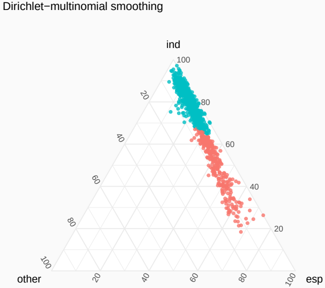
- Consensus clustering can be applied to summarise all B clusterings.
  - → Majority voting (Dudoit and Fridlyand, 2003)).



Consensus clustering

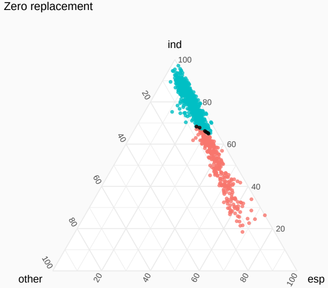
- Consensus clustering can be applied to summarise all B clusterings.
  → Majority voting (Dudoit and Fridlyand, 2003)).

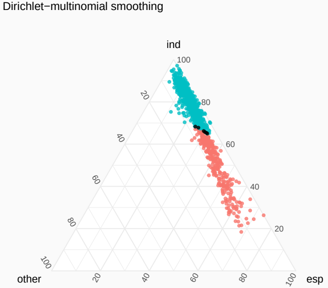




Consensus clustering

- Consensus clustering can be applied to summarise all B clusterings.
  → Majority voting (Dudoit and Fridlyand, 2003)).
- Only *six* municipalities differ in the final clustering.





Final remarks

- An approach to cluster count data when only the relative relation between parts is of interest has been presented.
- A parametric approach can be constructed in such a way that the variability coming from a multinomial counting process can be incorporated to the observed compositional variability.
- To obtain a final clustering, consensus clustering algorithms can be applied to clustering obtained from each sample.