The continuous categorical: a novel simplex-valued exponential family

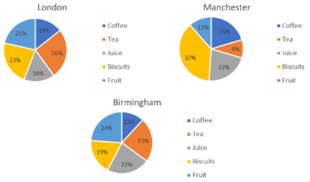
Elliott Gordon-Rodríguez, Gabriel Loaiza-Ganem, John P. Cunningham

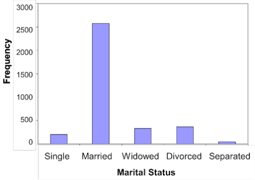
https://arxiv.org/abs/2002.08563

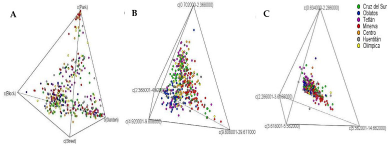
ICML 2020

Columbia University in the City of New York
layer6

Motivation: compositional data







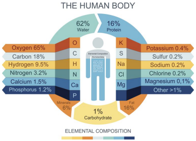
Definition (simplex): $\mathbb{S}^K := \{\mathbf{x} \in \mathbb{R}_+^K : \sum_{i=1}^K x_i = 1\}$

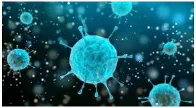
Gordon-Rodriguez, E., Loaiza-Ganem, G., & Cunningham, J. P. (2020). The continuous categorical: a novel simplex-valued exponential family.

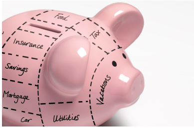
Motivation: compositional data

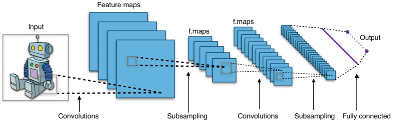
Examples:

- Geology
- Chemistry
- Microbiology
- Genetics
- Economics
- Politics
- **Machine learning**



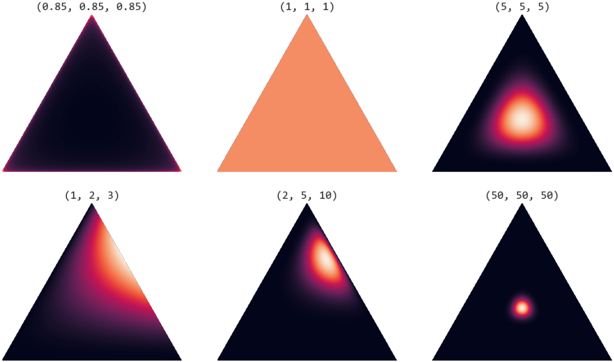


Shortcomings of the Dirichlet

Definition: $\mathbf{x} \sim \mathit{Dirichlet}(\boldsymbol{\alpha})$ if $\mathbf{x} \in \mathbb{S}^K$ with density:

$$p(\mathbf{x}; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}. \tag{1}$$



Shortcomings of the Dirichlet

Definition: $\mathbf{x} \sim \textit{Dirichlet}(\boldsymbol{\alpha})$ if $\mathbf{x} \in \mathbb{S}^K$ with density:

$$p(\mathbf{x}; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}. \tag{1}$$

- **Extrema.** $\log p(\mathbf{x}; \boldsymbol{\alpha}) \to \pm\infty$ as $x_j \to 0$.
  $\therefore$ log-likelihood is undefined in the presence of zeros.

Shortcomings of the Dirichlet

Definition: $\mathbf{x} \sim \mathit{Dirichlet}(\boldsymbol{\alpha})$ if $\mathbf{x} \in \mathbb{S}^K$ with density:

$$p(\mathbf{x}; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}. \tag{1}$$

- **Extrema.** $\log p(\mathbf{x}; \boldsymbol{\alpha}) \to \pm\infty$ as $x_j \to 0$.
  $\therefore$ log-likelihood is undefined in the presence of zeros.
- **Bias.** Re-write the density in canonical form
  $p(\mathbf{x}; \boldsymbol{\alpha}) = h(\mathbf{x}) \exp\left(\sum_{i=1}^{K} \alpha_i \log x_i - A(\boldsymbol{\alpha})\right)$.
  By theory of exponential families, MLE is unbiased for $\mathbb{E} \log x_j$.
  $\therefore$ MLE is biased for the mean $\mu_j = \mathbb{E} x_j$.

Shortcomings of the Dirichlet

Definition: $\mathbf{x} \sim \mathit{Dirichlet}(\boldsymbol{\alpha})$ if $\mathbf{x} \in \mathbb{S}^K$ with density:

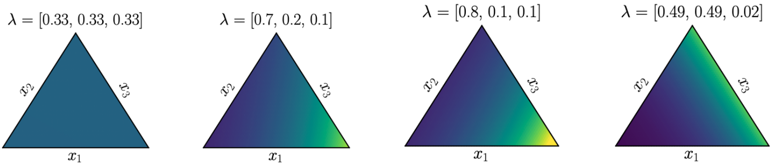
$$p(\mathbf{x}; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}. \tag{1}$$

- **Extrema.** $\log p(\mathbf{x}; \boldsymbol{\alpha}) \to \pm\infty$ as $x_j \to 0$.
  $\therefore$ log-likelihood is undefined in the presence of zeros.
- **Bias.** Re-write the density in canonical form
  $p(\mathbf{x}; \boldsymbol{\alpha}) = h(\mathbf{x}) \exp\left(\sum_{i=1}^{K} \alpha_i \log x_i - A(\boldsymbol{\alpha})\right)$.
  By theory of exponential families, MLE is unbiased for $\mathbb{E} \log x_j$.
  $\therefore$ MLE is biased for the mean $\mu_j = \mathbb{E} x_j$.
- **Flexibility.** If $\mathbf{x}_0 \in \mathbb{S}^K$ is a single datapoint, then
  $\log p(\mathbf{x}_0; \boldsymbol{\alpha}) \to \infty$ as $\boldsymbol{\alpha} \to \infty$ along $\boldsymbol{\alpha} = k\mathbf{x}_0$.
  $\therefore$ the Dirichlet log-likelihood is ill-behaved under flexible predictive models (e.g. GLMs, neural networks).

Solution: a new exponential family

Definition: $\mathbf{x} \in \mathbb{S}^K$ follows a *continuous categorical* ($\mathcal{CC}$) distribution with parameter $\boldsymbol{\lambda} \in \mathbb{S}^K$ if:

$$\mathbf{x} \sim \mathcal{CC}(\boldsymbol{\lambda}) \iff p(\mathbf{x}; \boldsymbol{\lambda}) \propto \prod_{i=1}^{K} \lambda_i^{x_i}$$



Gordon-Rodriguez, E., Loaiza-Ganem, G., & Cunningham, J. P. (2020). The continuous categorical: a novel simplex-valued exponential family.

Solution: a new exponential family

Definition: $\mathbf{x} \in \mathbb{S}^K$ follows a *continuous categorical* ($\mathcal{CC}$) distribution with parameter $\boldsymbol{\lambda} \in \mathbb{S}^K$ if:

$$\mathbf{x} \sim \mathcal{CC}(\boldsymbol{\lambda}) \iff p(\mathbf{x}; \boldsymbol{\lambda}) \propto \prod_{i=1}^{K} \lambda_i^{x_i}$$

- **Extrema.** $\log p(\mathbf{x}; \boldsymbol{\lambda})$ is finite at the extrema of the simplex. $\therefore$ log-likelihood is well-defined in the presence of zeros.

Solution: a new exponential family

Definition: $\mathbf{x} \in \mathbb{S}^K$ follows a *continuous categorical* ($\mathcal{CC}$) distribution with parameter $\boldsymbol{\lambda} \in \mathbb{S}^K$ if:

$$\mathbf{x} \sim \mathcal{CC}(\boldsymbol{\lambda}) \iff p(\mathbf{x}; \boldsymbol{\lambda}) \propto \prod_{i=1}^{K} \lambda_i^{x_i}$$

- **Extrema.** $\log p(\mathbf{x}; \boldsymbol{\lambda})$ is finite at the extrema of the simplex.
  $\therefore$ log-likelihood is well-defined in the presence of zeros.
- **Bias.** Re-write the $\mathcal{CC}$ density in canonical form
  $p(\mathbf{x}; \boldsymbol{\lambda}) \propto \exp\left(\sum_{i=1}^{K} \log(\lambda_i) \cdot x_i\right)$.
  $\therefore$ by theory of exponential families, MLE is unbiased for the mean $\mu_j = \mathbb{E}x_j$.

Solution: a new exponential family

Definition: $\mathbf{x} \in \mathbb{S}^K$ follows a *continuous categorical* ($\mathcal{CC}$) distribution with parameter $\boldsymbol{\lambda} \in \mathbb{S}^K$ if:

$$\mathbf{x} \sim \mathcal{CC}(\boldsymbol{\lambda}) \iff p(\mathbf{x}; \boldsymbol{\lambda}) \propto \prod_{i=1}^{K} \lambda_i^{x_i}$$

- **Extrema.** $\log p(\mathbf{x}; \boldsymbol{\lambda})$ is finite at the extrema of the simplex.
  $\therefore$ log-likelihood is well-defined in the presence of zeros.
- **Bias.** Re-write the $\mathcal{CC}$ density in canonical form $p(\mathbf{x}; \boldsymbol{\lambda}) \propto \exp\left(\sum_{i=1}^{K} \log(\lambda_i) \cdot x_i\right)$.
  $\therefore$ by theory of exponential families, MLE is unbiased for the mean $\mu_j = \mathbb{E}x_j$.
- **Flexibility.** The $\mathcal{CC}$ density is convex in $\mathbf{x}$.
  $\therefore$ cannot represent interior modes, cannot concentrate mass on interior points and log-likelihood does not diverge.

Solution: a new exponential family

Definition: $\mathbf{x} \in \mathbb{S}^K$ follows a *continuous categorical* ($\mathcal{CC}$) distribution with parameter $\boldsymbol{\lambda} \in \mathbb{S}^K$ if:

$$\mathbf{x} \sim \mathcal{CC}(\boldsymbol{\lambda}) \iff p(\mathbf{x}; \boldsymbol{\lambda}) \propto \prod_{i=1}^{K} \lambda_i^{x_i}$$

Where did this come from?

- A probabilistic cross-entropy loss for compositional data.
- Multivariate generalization of the *continuous Bernoulli* distribution (Loaiza-Ganem & Cunningham, NeurIPS 2019): $x \sim \mathcal{CB}(\lambda) \iff p(x|\lambda) \propto \lambda^x (1-\lambda)^{1-x}$, for $x \in [0,1] = \mathbb{S}^1$.
- A continuous relaxation of the categorical distribution.
- Switching the role of the parameter and the argument in the Dirichlet density.
- Restricting independent exponential RVs to the simplex.

Normalizing constant

Theorem: Write $C(\boldsymbol{\lambda})$ for the normalizing constant of the $\mathcal{CC}(\boldsymbol{\lambda})$ distribution, i.e.

$$\int_{\mathbb{S}^K} C(\boldsymbol{\lambda}) \prod_{i=1}^{K} \lambda_i^{x_i} d\mu(\mathbf{x}) = 1. \tag{2}$$

Then

$$C(\boldsymbol{\lambda}) = \left((-1)^{K+1} \sum_{k=1}^{K} \frac{\lambda_k}{\prod_{i \neq k} \log \frac{\lambda_i}{\lambda_k}} \right)^{-1},$$

Normalizing constant

Theorem: Write $C(\boldsymbol{\lambda})$ for the normalizing constant of the $\mathcal{CC}(\boldsymbol{\lambda})$ distribution, i.e.

$$\int_{\mathbb{S}^K} C(\boldsymbol{\lambda}) \prod_{i=1}^{K} \lambda_i^{x_i} d\mu(\mathbf{x}) = 1. \tag{2}$$

Then

$$C(\boldsymbol{\lambda}) = \left((-1)^{K+1} \sum_{k=1}^{K} \frac{\lambda_k}{\prod_{i \neq k} \log \frac{\lambda_i}{\lambda_k}} \right)^{-1},$$

Remark:

- Closed-form in terms of elementary functions only.
- Can compute moments, MGF, and more, directly from $C(\cdot)$.

Related distributions

Beta

Continuous
Bernoulli

Dirichlet

Continuous
Categorical

Related distributions

$x^{\alpha-1}(1-x)^{\beta-1}$

$\lambda^x(1-\lambda)^{1-x}$

$\prod_{i=1}^K x_i^{\alpha_i-1}$

$\prod_{i=1}^K \lambda_i^{x_i}$

Related distributions

$$x^{\alpha-1}(1-x)^{\beta-1} \xrightarrow{\text{Generalize to simplex}} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

$$\lambda^{x}(1-\lambda)^{1-x} \xrightarrow{\text{Generalize to simplex}} \prod_{i=1}^{K} \lambda_i^{x_i}$$

Related distributions

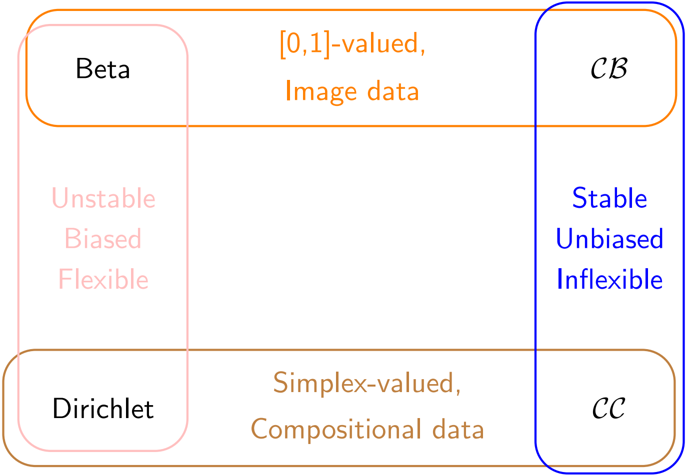
$$
\begin{tikzcd}
x^{\alpha-1}(1-x)^{\beta-1} \arrow[r, "\text{Switch parameter and argument}"] \arrow[d, "\text{Generalize to simplex}"'] & \lambda^{x}(1-\lambda)^{1-x} \arrow[d, "\text{Generalize to simplex}"'] \\
\prod_{i=1}^{K} x_i^{\alpha_i-1} \arrow[r, "\text{Switch parameter and argument}"] & \prod_{i=1}^{K} \lambda_i^{x_i}
\end{tikzcd}
$$

Gordon-Rodriguez, E., Loaiza-Ganem, G., & Cunningham, J. P. (2020). The continuous categorical: a novel simplex-valued exponential family.

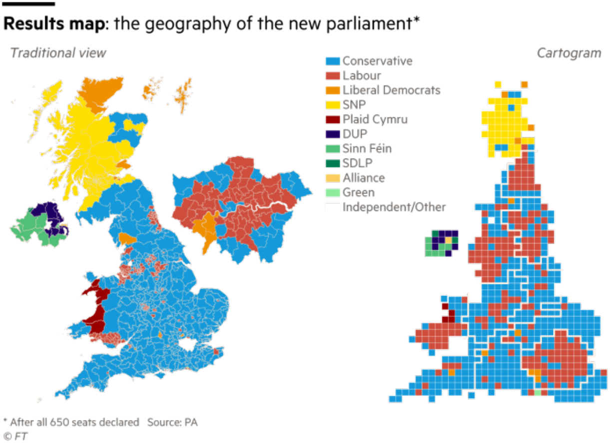
Related distributions

$$
\begin{tikzcd}
\text{Beta} \arrow[r, "\text{Switch parameter and argument}"] \arrow[d, "\text{Generalize to simplex}"'] & \mathcal{CB} \arrow[d, "\text{Generalize to simplex}"'] \\
\text{Dirichlet} \arrow[r, "\text{Switch parameter and argument}"] & \mathcal{CC}
\end{tikzcd}
$$

Related distributions

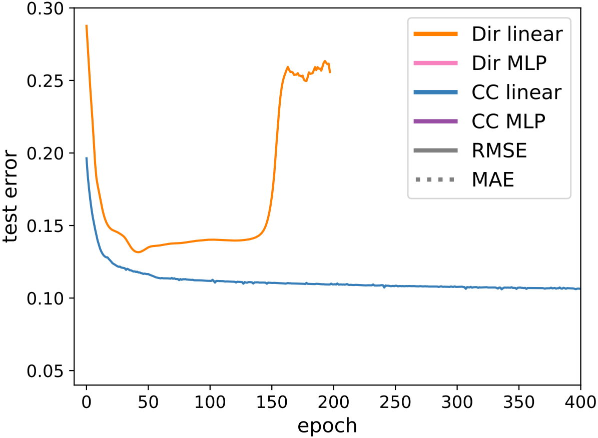


Application: UK 2019 general election

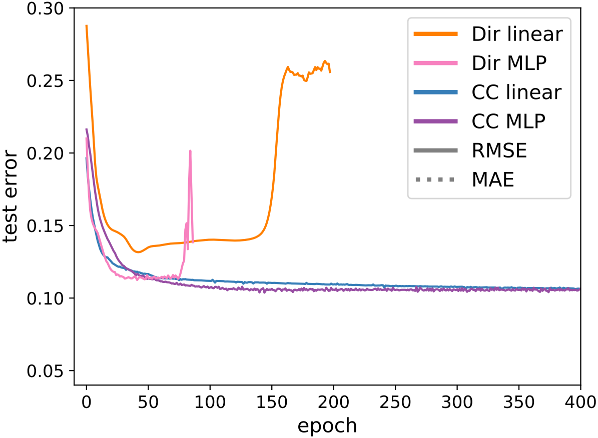


Constituency-level predictors $\xrightarrow{\text{regression function (linear or MLP)}}$ Voting outcomes

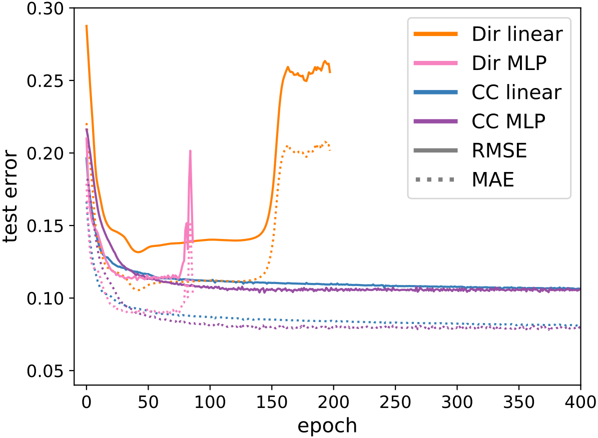
Election data: results



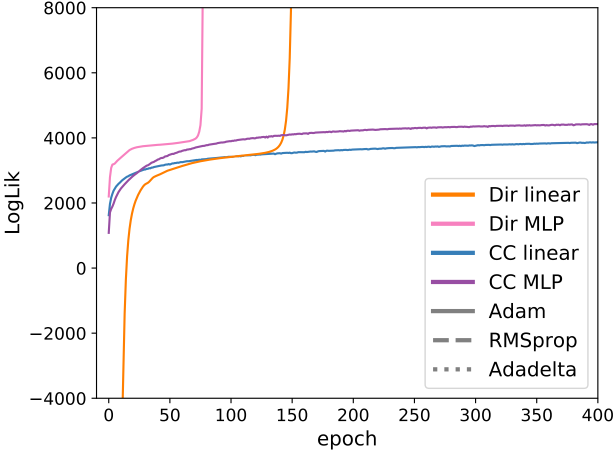
Election data: results



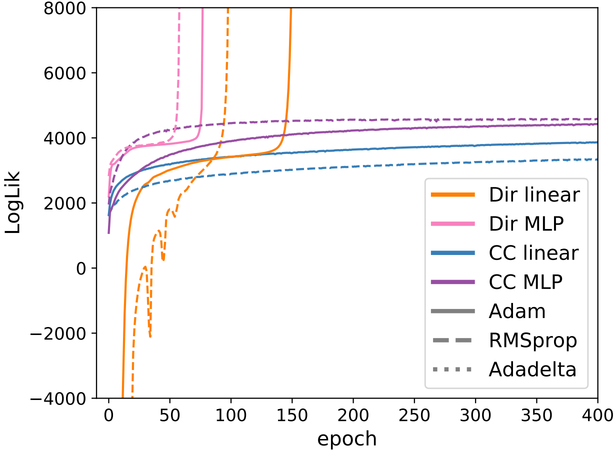
Election data: results



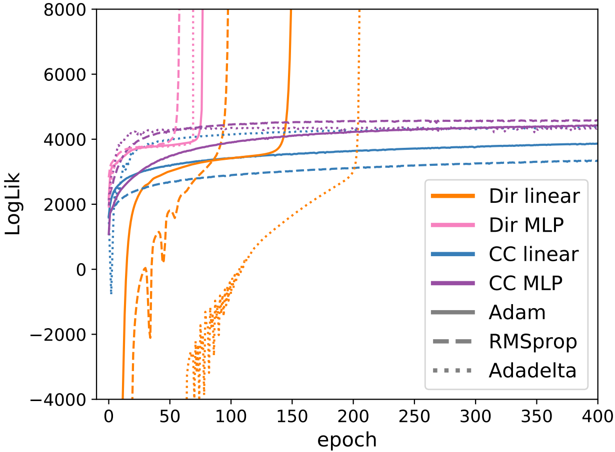
Election data: optimizers



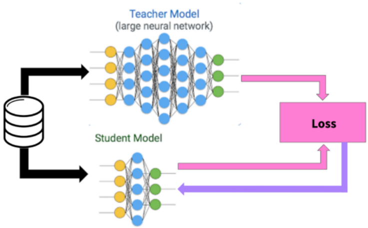
Election data: optimizers

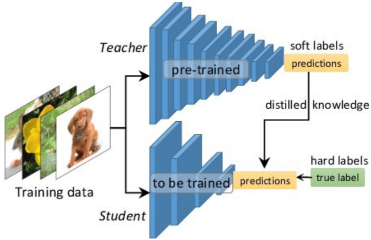


Election data: optimizers



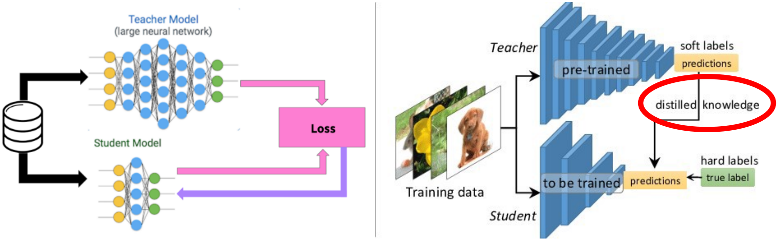
Model compression (knowledge distillation)





Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).

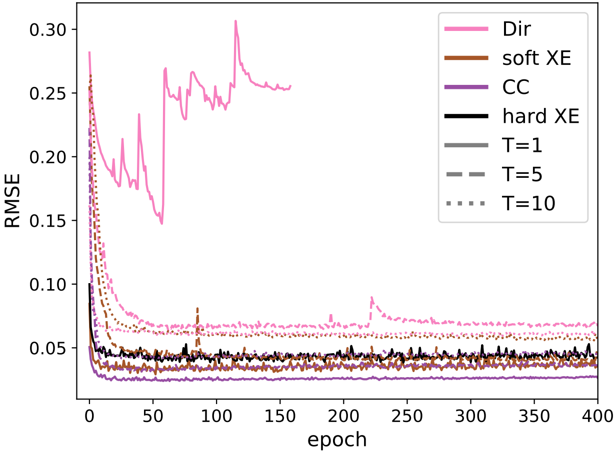
Model compression (knowledge distillation)



Student network learns from (soft) outputs of teacher model, via (soft) cross-entropy loss $\longrightarrow$ replace with $\mathcal{CC}$ log-likelihood.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).

Model compression: results on MNIST



Conclusion

- Novel exponential family of distributions.
- Attractive mathematical properties.
- Outperforms the Dirichlet in regression models of compositional outcomes.